nature genetics

ARTICLES

https://doi.org/10.1038/s41588-022-01118-8

Systematic identification of cell-fate regulatory programs using a single-cell atlas of mouse development

Lijiang Fei[1,6], Haide Chen[1,2,6], Lifeng Ma[1,6], Weigao E[1,6], Renying Wang[1,6], Xing Fang[1,6], Ziming Zhou[1,6], Huiyu Sun[1], Jingjing Wang[2], Mengmeng Jiang[2], Xinru Wang[1], Chengxuan Yu[1], Yuqing Mei[1], Danmei Jia[1], Tingyue Zhang[2], Xiaoping Han[1] ✉ and Guoji Guo[1,2,3,4,5] ✉

Waddington's epigenetic landscape is a metaphor frequently used to illustrate cell differentiation. Recent advances in single-cell genomics are altering our understanding of the Waddington landscape, yet the molecular mechanisms of cell-fate decisions remain poorly understood. We constructed a cell landscape of mouse lineage differentiation during development at the single-cell level and described both lineage-common and lineage-specific regulatory programs during cell-type maturation. We also found lineage-common regulatory programs that are broadly active during the development of invertebrates and vertebrates. In particular, we identified *Xbp1* as an evolutionarily conserved regulator of cell-fate determinations across different species. We demonstrated that *Xbp1* transcriptional regulation is important for the stabilization of the gene-regulatory networks for a wide range of mouse cell types. Our results offer genetic and molecular insights into cellular gene-regulatory programs and will serve as a basis for further advancing the understanding of cell-fate decisions.

The robustness of the developmental process for multicellular organisms suggests a dedicated regulatory program that governs the trajectories of cell-fate decisions[1–3]. According to Waddington's epigenetic landscape theory, differentiated cell types arise from an unstable stem/progenitor state and eventually fall into stable cell-fate attractors[4]. The emerging concept of the state manifold derived from single-cell data has further enhanced our understanding of lineage progression[5]. State manifolds, as a more general and data-driven representation of a Waddington landscape, reflect the high-dimensional nature of cell-fate decisions and provide high-resolution descriptions of dynamic cell trajectories[6]. What are the gene-regulatory programs underlying these state manifolds? How are they regulated? These are two central questions that are puzzling those working in the field.

Transcription factors (TFs) and gene-regulatory networks (GRNs) are known to govern cell-fate decisions[7]. For example, the *GATA1*/*PU.1* system makes the binary choice between the erythroid/megakaryocyte and myeloid lineages in the process of hematopoietic differentiation[8] and the Myo*D* system has critical roles in myogenic cell-lineage specification during development and trans-differentiation[9]. *Oct4*-*Cdx2* makes the decisions between inner cell mass and trophectoderm cells during embryogenesis[10]. These studies demonstrate the importance of lineage-specific transcriptional regulations in different cellular systems. However, these focused analyses of a cell type's regulatory network modules cannot offer a global view of the complex GRNs operating during organism development.

With breakthroughs in single-cell RNA-sequencing (scRNA-seq), single-cell atlases of various developmental stages have been profiled at the organism level[11–16]. Single-cell datasets offer unprecedented opportunities to systematically unravel the nature of cell-fate regulatory programs[17,18]. A systematic and global view of multi-lineage, multi-species, cell-fate gene-regulatory modules may help us to understand cellular lineage specification and maturation. In the present study, we determined the molecular content of lineage-common and lineage-specific regulatory programs through multi-lineage and cross-species analysis. We constructed a time-series mouse cell differentiation atlas (MCDA) to reveal the GRNs that govern cell-fate decisions (Fig. 1a). We characterized a general feature of decreased entropy with less complexity in most lineages along with development. Through cross-species analysis, we identified conserved features of cellular differentiation, one of which was that ribosomal genes are universally expressed at high levels in stem/progenitor cells. Importantly, we experimentally verified *Xbp1* as a lineage-common master regulator that was involved in core fate-determining circuits in mice.

Results

Construction of MCDA. We performed single-cell transcriptomic analysis on mice at seven life stages ranging from the early embryonic stage to the mature adult stage: embryonic day (E) 10.5, E12.5, E14.5, postnatal day (P) 0, P10, P21 and adult. Altogether, we profiled more than 520,000 single cells (Fig. 1b and Supplementary Tables 1–3). The profiled organs, including the brain, heart, intestine, kidneys, liver, lungs, pancreas, stomach, testes and uterus, spanned

[1]Center for Stem Cell and Regenerative Medicine and Bone Marrow Transplantation Center of the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. [2]Liangzhu Laboratory, Zhejiang University Medical Center, Hangzhou, China. [3]Zhejiang University–University of Edinburgh Institute, Zhejiang University School of Medicine, Zhejiang University, Hangzhou, China. [4]Zhejiang Provincial Key Lab for Tissue Engineering and Regenerative Medicine, Dr. Li Dak Sum & Yip Yio Chin Center for Stem Cell and Regenerative Medicine, Hangzhou, China. [5]Institute of Hematology, Zhejiang University, Hangzhou, China. [6]These authors contributed equally: Lijiang Fei, Haide Chen, Lifeng Ma, Weigao E, Renying Wang, Xing Fang, Ziming Zhou. ✉e-mail: xhan@zju.edu.cn; ggj@zju.edu.cn

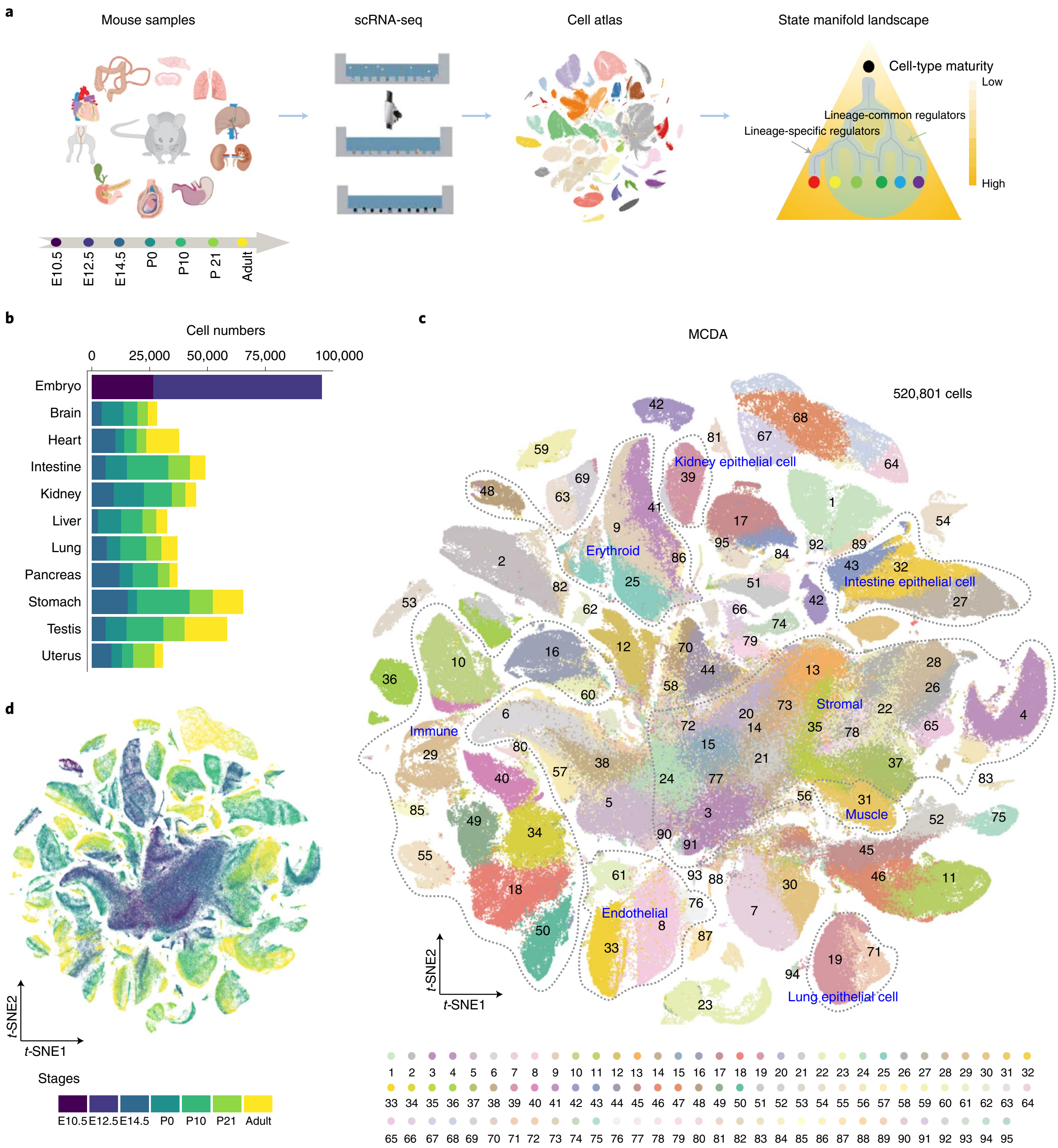


Fig. 1 | Single-cell transcriptional atlas of mouse differentiation. **a**, Overview of the experimental and bioinformatics analysis workflow. **b**, A total of ten organs were analyzed at seven different timepoints. The barplot shows the number of sequenced cells per organ per stage prepared by Microwell-seq. **c**, The *t*-SNE visualization of 520,801 single cells from the MCDA, colored by cluster identity. The gray dashed lines mark the cell types and lineages. **d**, The *t*-SNE visualization of 520,801 single cells from different developmental stages of mice, colored by stages. Parts **b** and **d** share the same color legend of stages.

diverse systems. Previously published E14.5 and adult data[11,14] represented approximately 30% of the cells in the entire dataset. Systemic mouse single-cell atlases of P0, P10 and P21 have not been depicted thus far. We projected all single cells on a *t*-distributed stochastic neighbor embedding (*t*-SNE) plot and obtained 95 transcriptionally distinct cell populations (Fig. 1c and Supplementary Table 4). Clusters that were composed of multiple tissues included immune cells (C9, C16, C18 C25, C29, C34), stromal cells (C13, C20, C22, C26, C28), muscle cells (C31) and endothelial cells (C8), whereas epithelial cells differed across tissues and formed separate clusters

(C19, C27, C39) (Extended Data Fig. 1a–c and Supplementary Table 5). Moreover, the clusters were arranged in chronological order, showing projections from fetal progenitors toward adult mature cell types (Fig. 1d). Analysis of differentially expressed genes (DEGs) in neighboring stages for each tissue showed that the critical period of tissue maturity varied across different stages. The transition from E14.5 to P0 led to dramatic changes during development (Extended Data Fig. 1d). Changes from P0 to P10 were dominated by energy metabolism on account of the different energy sources[19], whereas changes from P10 to adulthood focused on pathways of response, transport and metabolism (Extended Data Fig. 1e and Supplementary Table 6). We observed lots of distinct clusters from the P0 and P10 samples, indicating continuing cellular transitions after birth. We have provided an interactive website, http://bis.zju.edu.cn/MCA, to enable public access to this systematic single-cell atlas of mouse lineage differentiation from embryogenesis through to mature adult.

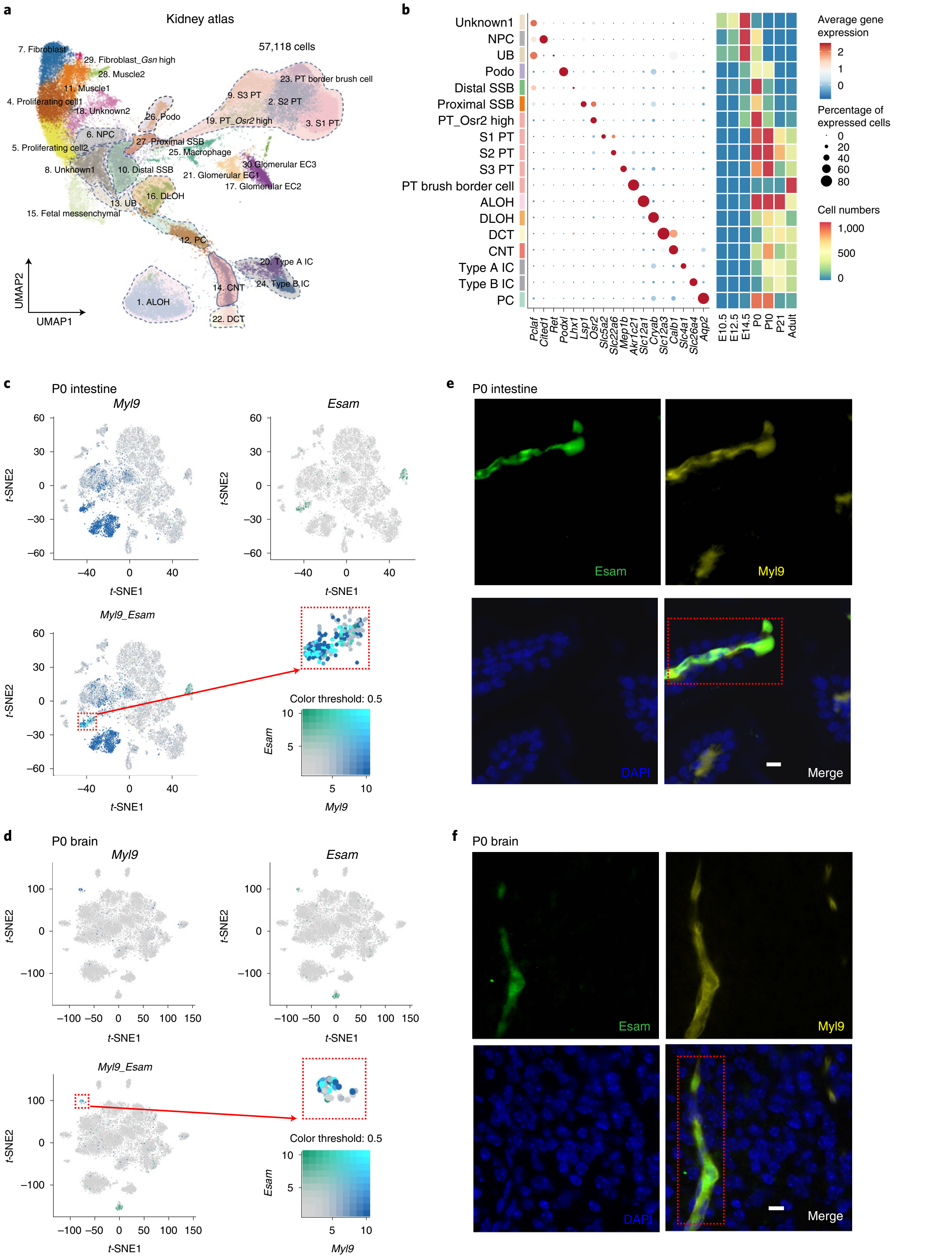
Cellular changes during mouse development. The tissue effect gave rise to 31.9% of the global variance, which is much more than variance from the stage and sex effects (Extended Data Fig. 1f,g). We studied dynamic changes in the kidneys as a representative. After analyzing kidney samples from the E10.5 to adult stages, we defined 30 clusters with canonical markers[20,21] which included stromal cells, nephron epithelial cells, fenestrated endothelial cells and immune cells (Fig. 2a, Extended Data Fig. 1h and Supplementary Table 7). Cells from diverse developmental stages of nephrogenesis were well captured in our single-cell data, with ureteric bud (UB) cells (*Ret*$^+$, *Gata3*$^+$), nephron progenitor cells (NPCs, *Cited1*$^+$, *Gdnf*$^+$, *Six2*$^+$), proximal S-shaped body (SSB) cells (*Lsp1*$^+$, *Tmem100*$^+$), distal SSB cells (*Lhx1*$^+$), podocytes (Podos, *Podxl*$^+$), five types of proximal tubule (PT) cells, ascending and descending loop of Henle cells (ALOH and DLOH), connecting nephron tubule (CNT) cells, distal collecting tubule (DCT) cells, two subsets of intercalated cells (ICs) and principal cells (PCs). Notably, UB cells and NPCs included cells at the P0 stage, whereas distal and proximal SSB cells included cells at the P10 stage (Fig. 2b). This result indicated that nephrogenesis continued postnatally instead of being completed before birth in the mice. Moreover, the maturation of renal function continued until the adult stage with gradual physiological changes (Extended Data Fig. 1i).

To reveal cellular heterogeneity in mouse tissues during the development, we performed *t*-SNE and differential gene expression analysis for each tissue at different stages (Extended Data Figs. 2 and 3 and Supplementary Table 8). We then uncovered 37 previously unrecognized cell populations with interesting gene expression patterns with regard to mouse development (Supplementary Table 9). For example, several cell types were found co-expressing markers of two cell types. We identified cells that co-expressed makers of myocytes (*Myl9*, *Acta2*) and endothelial cells (*Esam*, *Gng11*) in both intestine and brain at the P0 stage (Fig. 2c,d). The co-immunofluorescence of *Myl9* and *Esam* further confirmed the scRNA-seq results (Fig. 2e,f). These myoendothelial cell types may be endowed with multi-lineage potential similar to human myoendothelial cells[22]. In the P10 lung, we verified a special club cell type (*Scgb1a1*, *Scgb3a1*) expressing goblet cell markers (*Tff2*, *Muc5b*), which may be an intermediate cell type during airway epithelial differentiation (Extended Data Fig. 4a,b). In addition, some tissue-specific markers showed ectopic expression in other tissues. For example, we discovered hepatocyte-like cells (*Afp*, *Alb*) in the pancreas at both the P0 and the P10 stages, and immunofluorescence assays confirmed their existence (Extended Data Fig. 4c,d). They displayed different expression patterns from liver hepatocytes and showed high expression of early hepatic stem or progenitor marker *Hnf4a*[23,24] (Extended Data Fig. 4e,f). Together, progenitor pools with co-expression or ectopic expression patterns may widely present in developing organs, suggesting the complexity of the mammalian state manifolds before terminal differentiation.

Characterization of regulatory programs in MCDA. High-resolution MCDA offers a powerful resource for studying the molecular basis of cell-fate decisions through various lineages. To reveal organism-wide characteristics, we applied different potency models based on entropy to qualify the state manifold landscape[25–29]. Entropy decreased continuously along with organ maturation in the most assayed lineages using different computational methods (Fig. 3a,b and Extended Data Fig. 5a–d), revealing a decrease in transcriptional plasticity and an increase in transcriptional stability. Based on the principles of these methods, we inferred that cell-type maturation appears to be an event associated with more singular transcriptomes and biological processes.

Cell types represent high-dimensional attractor states of GRNs[30]. TFs function as important regulators in GRNs to specify cell types and differentiation patterns[31]. To identify critical TFs of cell identity, we took the advantages of both data-driven (SCENIC)[32] and database-derived (VIPER-DOROTHEA)[33] methods to estimate the activities of TFs. We achieved >75% sensitivity to detect tissue-specific TFs based on single-cell datasets (Extended Data Fig. 5e). Over 900 TFs were identified with confidence levels ranging from A (high confidence) to C (low confidence) (Supplementary Table 10). Aggregated heatmaps were constructed to display the specific and common relationships of the TFs and their enriched lineages during development (Fig. 3c and Extended Data Fig. 5f). The neural lineage was characterized by *Dlx1*, *Pou3f3* and *Sox10*. The *Cebpa* and interferon regulatory transcription factor genes marked the immune lineage, whereas the endothelial lineage exhibited prominent *Sox17* and *Sox18* expression. Strikingly, hierarchical clustering analysis showed two modules of lineage-sharing TFs, which were enriched in adult tissues and fetal tissues, respectively (Fig. 3c and Extended Data Fig. 5g). Enrichment and occupancy of Hox and zinc-finger families in fetal tissues have previously been associated with embryonic development[34,35]. The ubiquitous expression cluster in adult tissues was shared for a wide range of lineages, with extensive representation of *Xbp1*, genes of the activator protein-1 (AP-1) family and other molecules. Only 78 out of 268 TFs in this adult multi-lineage cluster were housekeeping genes[36] (Extended Data Fig. 5h). *Jun* and *Fos* gene families can dimerize and form AP-1, which has been reported to act as a regulator in the differentiation of various cell types[37,38]. In addition, AP-1 family members have been recently suggested to act as central regulators of somatic cell fate[39,40]. These highlighted the important roles of AP-1 family members in cell-type differentiation and

Fig. 2 | Cellular heterogeneity in mouse tissues. a, UMAP visualization of 57,118 single cells in the kidneys at 7 different timepoints, colored by cluster identity. **b**, Dot-plot visualization of expression levels of representative markers in each cell type in the kidney single-cell data. The size of the dot encodes the percentage of cells within the cell type and the color encodes the average expression level. Heatmap showing the cell number of corresponding cell types at each timepoint. **c**,**d**, Feature plots in the *t*-SNE map of P0 intestine (**c**, $n = 9{,}265$ cells) and P0 brain (**d**, $n = 9{,}101$ cells). Cells are colored according to the expression of the indicated marker genes or two genes. The red boxes magnify the co-expressed cell types in the tissues. **e**,**f**, Immunofluorescence assay for the cells that co-expressed makers of myocytes (*Myl9*) and endothelial cells (*Esam*) in both intestine (**e**) and brain (**f**) at the P0 stage. The blue marks the cell nucleus using DAPI. The red boxes indicate the co-expressed locations. The experiment was replicated three times with similar results. Scale bar, 20 µm.


a
Kidney atlas
57,118 cells
7. Fibroblast
29. Fibroblast_Gsn high
28. Muscle2
11. Muscle1
4. Proliferating cell1
18. Unknown2
26. Podo
9. S3 PT
2. S2 PT
23. PT border brush cell
19. PT_Osr2 high
3. S1 PT
6. NPC
5. Proliferating cell2
27. Proximal SSB
25. Macrophage
8. Unknown1
10. Distal SSB
21. Glomerular EC1
30. Glomerular EC3
17. Glomerular EC2
13. UB
16. DLOH
15. Fetal messenchymal
12. PC
20. Type A IC
24. Type B IC
14. CNT
1. ALOH
22. DCT
UMAP2
UMAP1
b
Unknown1
NPC
UB
Podo
Distal SSB
Proximal SSB
PT_Osr2 high
S1 PT
S2 PT
S3 PT
PT brush border cell
ALOH
DLOH
DCT
CNT
Type A IC
Type B IC
PC
Pclaf
Cited1
Ret
Podxl
Lhx1
Lsp1
Osr2
Slc5a2
Slc22a6
Mep1b
Akr1c21
Slc12a1
Cryab
Slc12a3
Calb1
Slc4a1
Slc26a4
Aqp2
E10.5
E12.5
E14.5
P0
P10
P21
Adult
Average gene expression
2
1
0
Percentage of expressed cells
0
20
40
60
80
Cell numbers
1,000
500
0
c
P0 intestine
Myl9
Esam
60
30
0
−30
−60
t-SNE2
−40
0
40
t-SNE1
Myl9_Esam
Color threshold: 0.5
10
5
Esam
5
10
Myl9
d
P0 brain
Myl9
Esam
100
0
−100
t-SNE2
−100 −50 0 50 100 150
t-SNE1
Myl9_Esam
Color threshold: 0.5
10
5
Esam
5
10
Myl9
e
P0 intestine
Esam
Myl9
DAPI
Merge
f
P0 brain
Esam
Myl9
DAPI
Merge

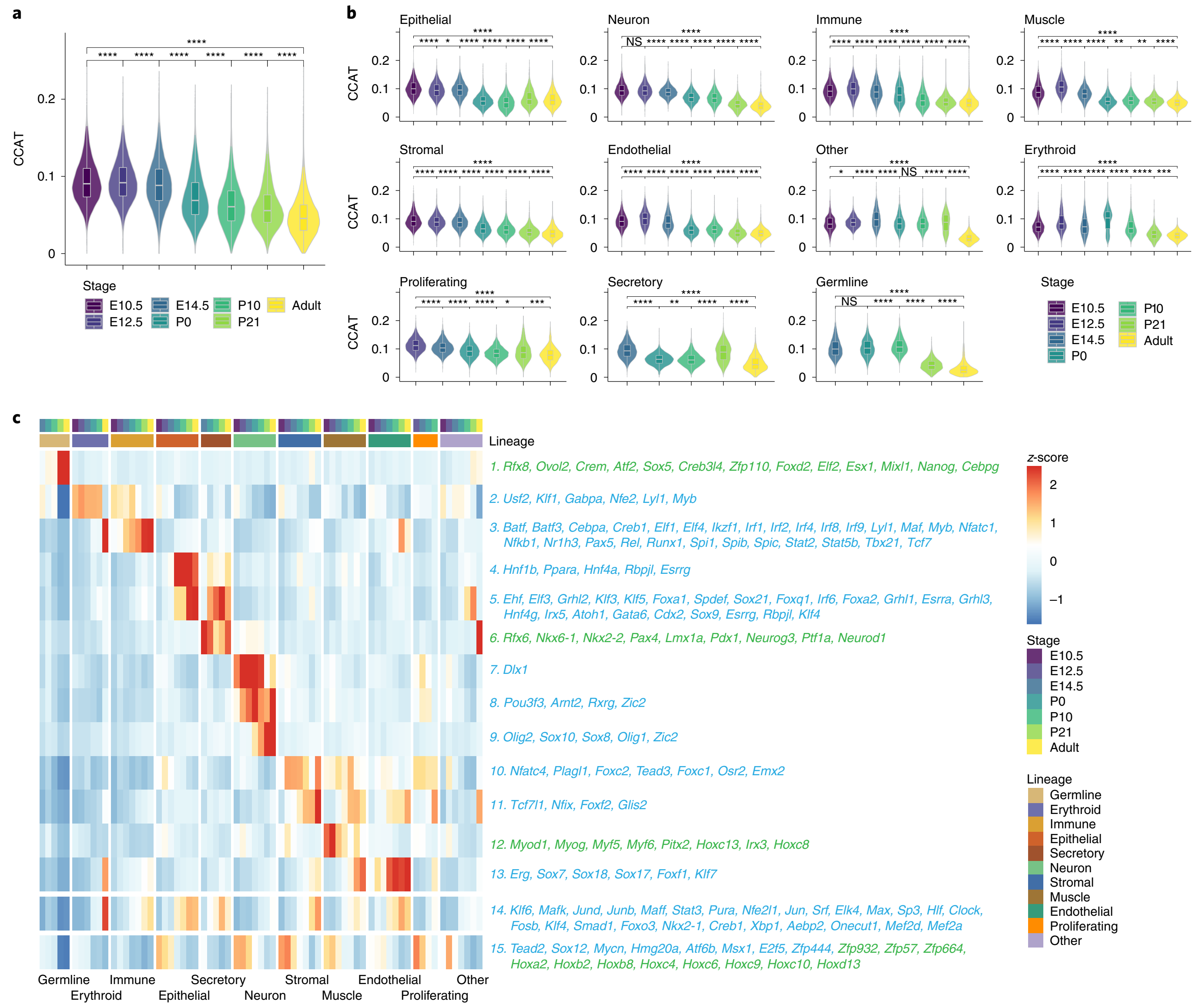


Fig. 3 | Analysis of regulatory programs in MCDA. **a**, Entropy measurement of MCDA using the CCAT method in different development stages ($n = 520{,}801$ cells). P values are from a two-sided Wilcoxon's rank-sum test comparing entropies of two different development stages. NS, not significant; $P > 0.05$, $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$, $^{****}P \leq 0.0001$. The exact P values have been displayed in the Source data. Boxplots: center line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within 1.5 × interquartile ratio (IQR). The same statistical analysis was performed for **a** and **b**. **b**, Entropy measurement of each lineage in MCDA using the CCAT method in different developmental stages (epithelial: $n = 116{,}436$ cells; neuron: $n = 41{,}342$ cells; immune: $n = 75{,}433$ cells; muscle $n = 17{,}909$ cells; stromal: $n = 106{,}955$ cells; endothelial: $n = 23{,}243$ cells; other: $n = 30{,}575$ cells; erythroid: $n = 41{,}683$ cells; proliferating: $n = 16{,}567$ cells; secretory: $n = 15{,}161$ cells; germline: $n = 35{,}497$ cells). **c**, Heatmap of aggregated module activities of TFs clustered by fuzzy c-means showing variations by stage and lineage from SCENIC. The representative TFs of each lineage in the MCDA are listed. The blue marks the TFs in collection A (high confidence) and the green marks the TFs in collection B (medium confidence).

cell-identity maintenance. Moreover, these TFs exhibited increasingly upregulated gene expression levels during lineage maturation (Extended Data Fig. 5i), which coincided with decreased entropy in most lineages (Fig. 3a,b and Extended Data Fig. 5a–d). Taken together, these results suggest that these lineage-common TFs function as vital regulators during maturation across a range of mouse cell types.

Global features during cell-fate decisions across species. Given that the suite of regulatory genes that control development is ancient[41], we wondered whether GRNs are conserved in invertebrates and vertebrates. We decided to investigate the lineage-specific and lineage-common regulatory elements during evolution. First, we performed a comparative analysis of gene regulation during development in seven species with varying evolutionary distances at single-cell resolution. Development atlases of four invertebrates and three vertebrates were collected, including *Schmidtea mediterranea*[12], *Caenorhabditis elegan*[15], *Ciona intestinalis*[16], *Hydra vulgaris*[13], *Danio rerio*[42], *Mus musculus*[11] and *Homo sapiens*[14]. More than 1,100,000 cells were categorized into 665 cell-type pairs for relatively differentiated states and undifferentiated states (Extended Data Fig. 6a and Supplementary Table 11). Partition-based graph abstraction (PAGA)[43] was applied to map cell types along the developmental branch for invertebrates (Extended Data Fig. 6b–d). For

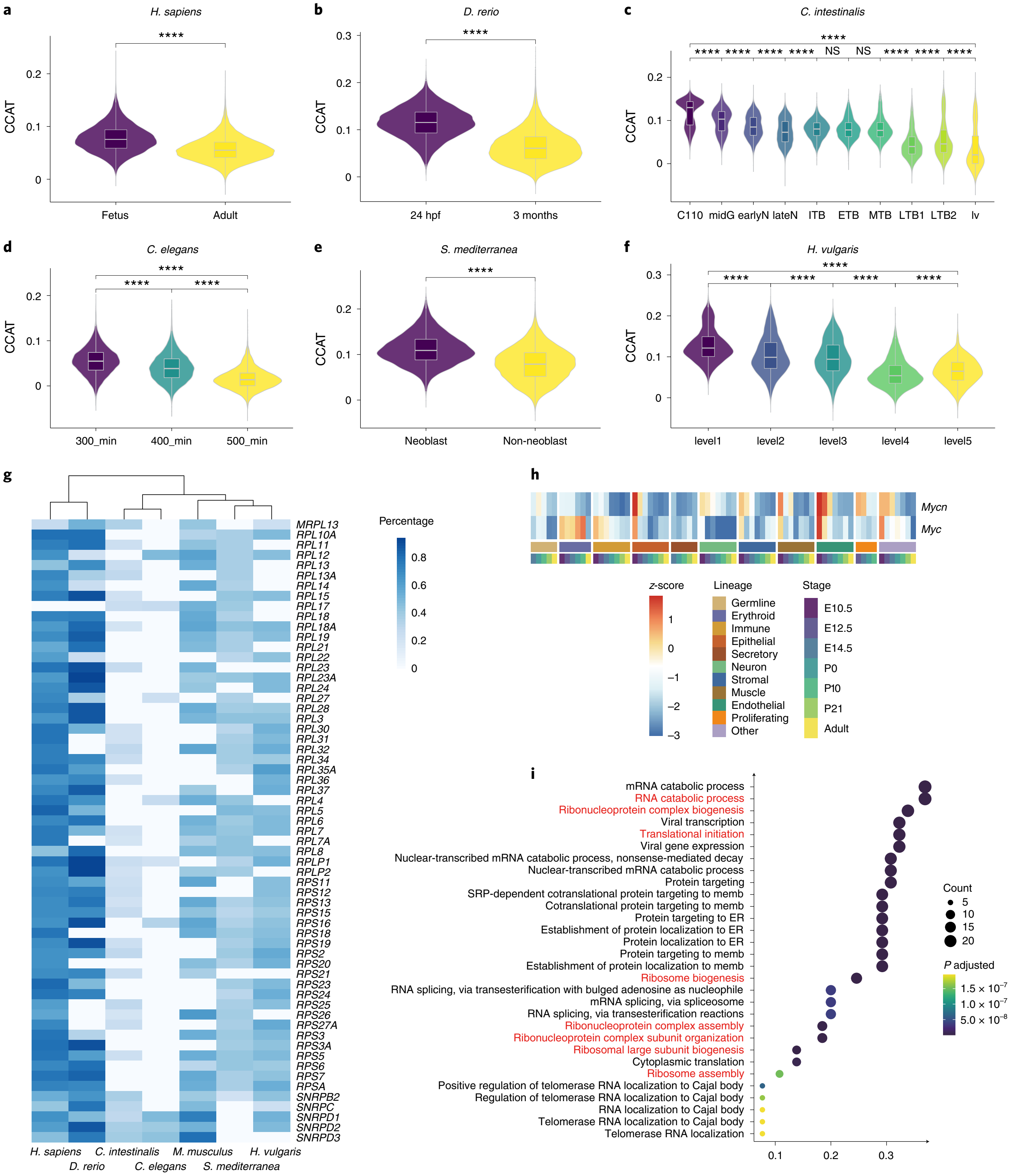


Fig. 4 | Global characteristics of cell differentiation across species. a–f, Entropy measurement of cells in *H. sapiens* (**a**, $n = 85{,}181$ cells), *D. rerio* (**b**, $n = 76{,}838$ cells), *C. intestinalis* (**c**, $n = 90{,}579$ cells), *C. elegans* (**d**, $n = 61{,}810$ cells), *S. mediterranea* (**e**, $n = 21{,}612$ cells) and *H. vulgaris* (**f**, $n = 25{,}052$ cells) using the CCAT methods. The color represents the stage. *P* values are from a two-sided Wilcoxon's rank-sum test comparing entropies of two different development stages. NS, not significant; $P > 0.05$, $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$, $^{****}P \leq 0.0001$. The exact *P* values have been displayed in Source data. Boxplots: center line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within 1.5 × IQR. **g**, Heatmap showing the cell-type frequencies of commonly downregulated ribosomal protein genes, mitochondrial ribosomal proteins and small nuclear ribonucleoprotein genes in at least four species. More genes are included in Supplementary Table 13. **h**, Heatmap showing the activity scores of *Mycn* and *Myc* in different stages and lineages in MCDA. **i**, Gene enrichment analysis of the driving genes in the CCAT method. The top 30 enriched biological processes were displayed. Red marks the GO terms related to the ribosome biogenesis. ER, endoplasmic reticulum; memb, membrane.

vertebrates, to minimize the impact of tissue effects, we connected cell states of the same tissue across time based on gene expression similarity[44], and cell hierarchies of the human lung were shown as an example (Extended Data Fig. 6e).

To explore the common changes in cross-species development, we performed entropy analysis and found that entropy decreased in all seven species along with development, which suggested that the increase in transcriptional stability was evolutionarily conserved (Fig. 4a–f and Extended Data Fig. 7a,b). For the molecular changes, we performed differential gene expression analysis between corresponding cell-type pairs and mapped homologous genes to the human gene symbols to find commonly regulated genes in multiple species (Extended Data Fig. 7c,d and Supplementary Tables 12 and 13). For all species, the numbers of conserved downregulated genes were greater than those of conserved upregulated genes, which suggests that stem/progenitors have more convergent expression patterns than differentiated cell types[45] (Extended Data Fig. 7e). Both commonly downregulated and upregulated genes in at least three species tended to have more protein–protein interactions (PPIs) than other conserved and not conserved genes in at least three species, which indicated that the common regulators were evolutionarily older[46] (Extended Data Fig. 7f,g). The genes downregulated during development were enriched with ribosomal protein genes, mitochondrial ribosomal protein genes and small nuclear ribonucleoprotein genes (Fig. 4g, Extended Data Fig. 7h and Supplementary Table 14). Notably, *Myc* and *Mycn*, as regulators of ribosome biogenesis[47,48], showed high activity scores in the early stages of mouse development (Fig. 4h). They were classified in the common (fetal) module (Supplementary Table 10). These findings were highly consistent with recent studies, which reported that ribosomal protein genes as central network hubs are robust markers of differentiation potency[5,49]. In our cross-species entropy analysis, the conserved driving genes (Methods) in cells with high differentiation potential were also enriched in ribosomal biogenesis[50] (Fig. 4i). Ribosomal protein genes are also suppressed during zebrafish hematopoiesis[51,52]. Our results suggest that ribosomal protein genes are a conserved feature of stemness and they are downregulated during cell-type differentiation. On the other hand, the upregulated genes were highly enriched for immunity pathways (Extended Data Fig. 7i,j and Supplementary Table 15), which was consistent with recent reports on human and mouse adult tissues[14,53]. Together, we present a catalog of common features during lineage development from invertebrates to vertebrates; particular ribosomal protein genes are enriched in the less differentiated cells.

Gene regulation networks of cell-fate decisions across species. To search for lineage-specific regulators among different species, we systematically aligned homologous pairs of cell lineages from each species across large evolutionary distances. Two methods, SAMap[54] and MetaNeighbor[55], were applied with different calculation principles and homologous gene-mapping methods. SAMap enables mapping single-cell transcriptomes between phylogenetically remote species based on the gene expression similarity whereas MetaNeighbor has high replicability in cell-type matching using homologous weighted gene matrices. High confidence thresholds (alignment scores with >0.5 in SAMap and Mean_AUROC >0.8 in MetaNeighbor) were adopted to obtain complementarily reliable cell-type matches across species. Some 47 of the 60 cell lineages from 7 species were characterized into 8 meta-lineages (Extended Data Fig. 8a). The Uniform Manifold Approximation and Projection (UMAP) embedding based on pseudo-bulk cells per species proved the rationality of meta-lineages, in which pseudo-bulk cells from the same meta-lineage were more intensively clustered (Extended Data Fig. 8b,c). Then, the specificity of TFs was characterized with the modified regulon-specific scores with TF expression count matrices as input per species[56,57]. Lineage-specific TFs displayed sequence similarity within the meta-lineage across species (Extended Data Fig. 8d–j). Vertebrates tended to have more conserved species-specific TFs than invertebrates.

For lineage-common regulators among different species, we found that several commonly upregulated TFs exhibited remarkable convergence, including *XBP1*, *JUND*, *FOSB*, *JUN*, *BHLHE40* and others (Fig. 5a), consistent with the enriched TFs in various adult mouse tissues (Fig. 3c and Supplementary Table 10). These TFs also displayed strong negative correlations with TFs that were enriched in lineage-specific progenitor cells (*GATA1*, *PAX6*, *NKX6-2*, *NEUROD1*, *SOX10*, *OLIG2*) in the Human Cell Landscape (HCL), a comprehensive cell landscape for humans generated by Microwell-seq[14] (Fig. 5b and Supplementary Tables 16 and 17). We suspect that these TFs may function as evolutionarily conserved regulators to guide multi-lineage cells to differentiation and maturity. We found that only one TF, *Xbp1*, stands out in all seven species (Fig. 5a and Extended Data Fig. 7d). Therefore, we attempted to further characterize the role of *Xbp1* in cell-type maturation. Previous work has emphasized functions of the basic helix–loop–helix TF *Xbp1* for cell differentiation in various cell types, including secretory cells, plasma cells, T cells, neurons, hepatocytes and other cell types[58–62]. As a putative common regulator, *Xbp1* showed an upregulated expression pattern in most lineages of MCDA (Fig. 5c). We further dissected its regulatory role from a cell atlas perspective and found that stem regulators such as *SOX4*, *SON* and *HES1* are the most negatively correlated with *XBP1* in the HCL (Fig. 5d). In addition, the *XBP1*-binding motif in hematopoietic progenitors and neural progenitors was less enriched than their corresponding mature cell types in the single-cell assay for transposase accessible chromatin using sequencing (scATAC-seq) data of the mouse and human[63–65] (Fig. 5e,f).

***Xbp1* as a common regulator in multi-lineage progression.** To dissect the mechanistic roles of the potential lineage-common regulators *Xbp1*, we used clustered regularly interspaced short palindromic repeats (CRISPR)–Cas9 to disrupt the *Xbp1* locus in mice (Fig. 6a, Extended Data Fig. 9a,b and Supplementary Table 18). As most *Xbp1*$^{-/-}$ embryos died at E13.5, we applied scRNA-seq to analyze embryos at E12.5 from *Xbp1*$^{+/-}$ heterozygous crosses before massive embryonic lethality[62] (Fig. 6b,c, Extended Data Fig. 9c and Supplementary Table 19). We found that increased cell groups after *Xbp1* disruption were all related to progenitor and immature cells (for example, fetal mesenchymal progenitors, early primitive erythroid progenitor, muscle progenitors, radial glia, oligodendrocyte progenitors and immature neurons) (Extended Data Fig. 9d). In addition, when compared with wild-type (WT) cells, *Xbp1*$^{-/-}$ cells displayed higher entropy in a broad range of lineages, which may be linked to the eventual failure of cell-type maturation (Fig. 6d and Extended Data Fig. 9e,f). Then we performed differential expression analysis and observed that a group of ribosomal protein genes (for example, *Rps3a1* and *Rps7*) were specifically upregulated in *Xbp1*$^{-/-}$ cells. Moreover, progenitor markers such as *Sox4*, *Id2*, *Son* and the imprinted gene *H19* were enriched in *Xbp1*$^{-/-}$ cells. The lineage-common regulators *Fosb* and *Jun* were downregulated in *Xbp1*$^{-/-}$ cells (Fig. 6e and Supplementary Table 20). Thus, disruption of *Xbp1* caused mouse embryos to acquire a more progenitor state.

To characterize the loss-of-function changes at protein levels, we performed liquid chromatography–mass spectrometry (LC–MS) proteomic analysis on both WT and knockout (KO) embryos (Supplementary Table 21). *Xbp1*$^{-/-}$ embryos exhibited higher expression level of pluripotency-related proteins such as Lin28a, Lin28b[66], Pcgf6 (ref. [67]) and Jarid2 (ref. [68]) and lower expression level of cell type-specific proteins such as Snca in neural cells, Clu in stromal cells, Afp in hepatocytes, C1qb in macrophage and Blvrb in erythroid cells (Fig. 6f and Extended Data Fig. 10a).

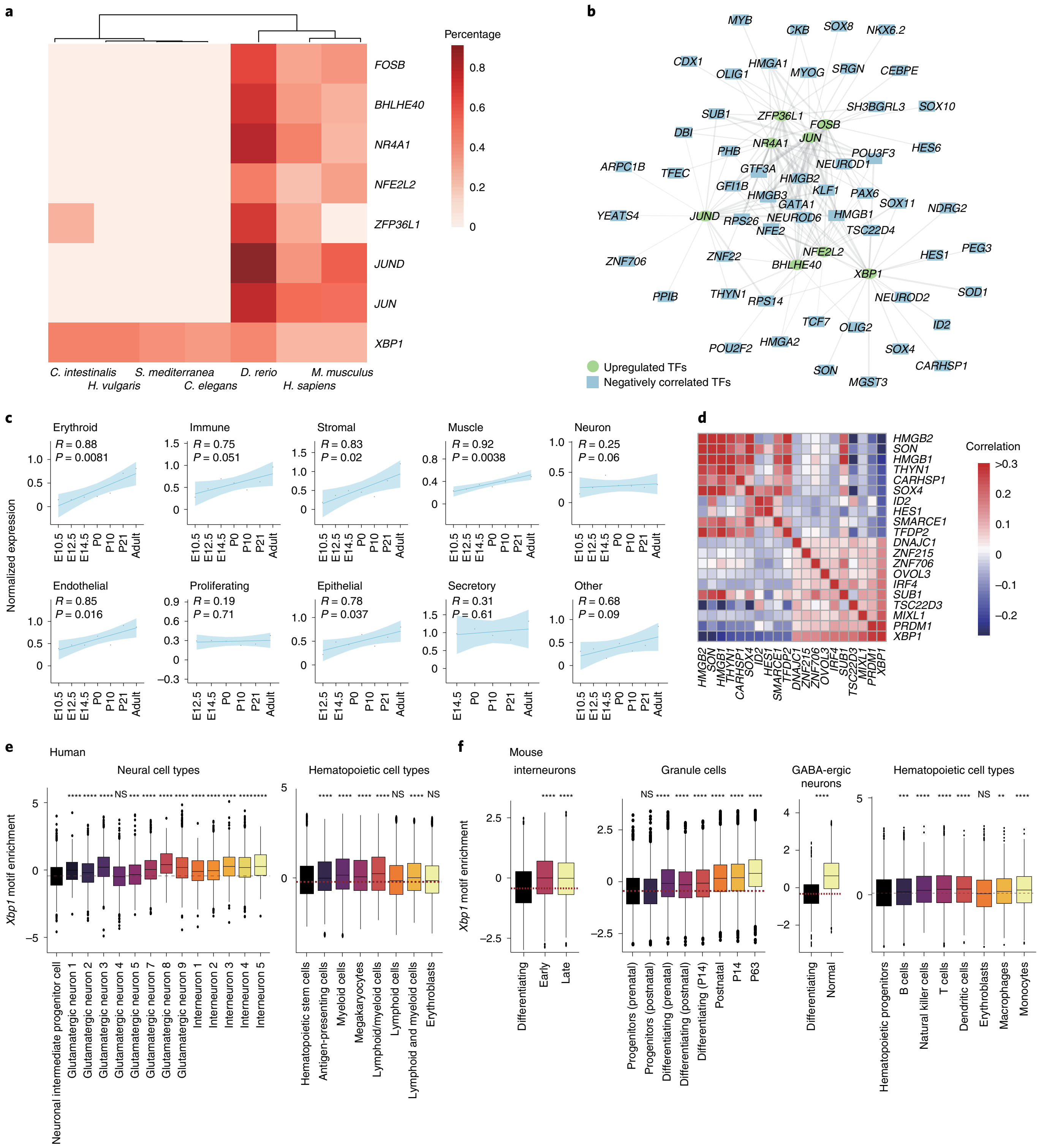


Fig. 5 | Inference of gene regulation during cell-fate decisions across species. a, Heatmap showing the cell-type frequencies of commonly upregulated TFs in seven species. **b**, Regulatory network showing the top 20 most negatively relevant TFs in the HCL for the commonly upregulated TFs (Pearson's correlation $P \leq 0.05$). **c**, Scatter plot showing aggregated *Xbp1* expression patterns in MCDA per lineage. Lines were estimated through linear regression and the 95% confidence interval is shown in blue with the mean value in gray points. **d**, Heatmap showing the top 10 TFs most correlated with *XBP1* in the HCL. **e**,**f**, Boxplot showing the *z*-scores of *Xbp1* motif enrichment in neural cell types and hematopoietic cell types in the human (**e**) and the mouse (**f**) in scATAC-seq data (human neural cell types: $n = 22,075$ cells; human hematopoietic cell types: $n = 16,133$ cells; mouse interneurons: $n = 5,134$ cells; mouse granule cells: $n = 25,155$ cells; mouse γ-aminobutyric acid (GABA)-ergic neurons: $n = 2,041$ cells; mouse hematopoietic cell types: $n = 24,125$ cells). *P* values are from a two-sided Wilcoxon's rank-sum test comparing the *Xbp1* enrichment score between the progenitor cell types (the first box) and other cell types. NS, not significant; $P > 0.05$, $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$, $^{****}P \leq 0.0001$. The exact *P* values have been displayed in Source data. Boxplots: center line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within 1.5 × IQR.

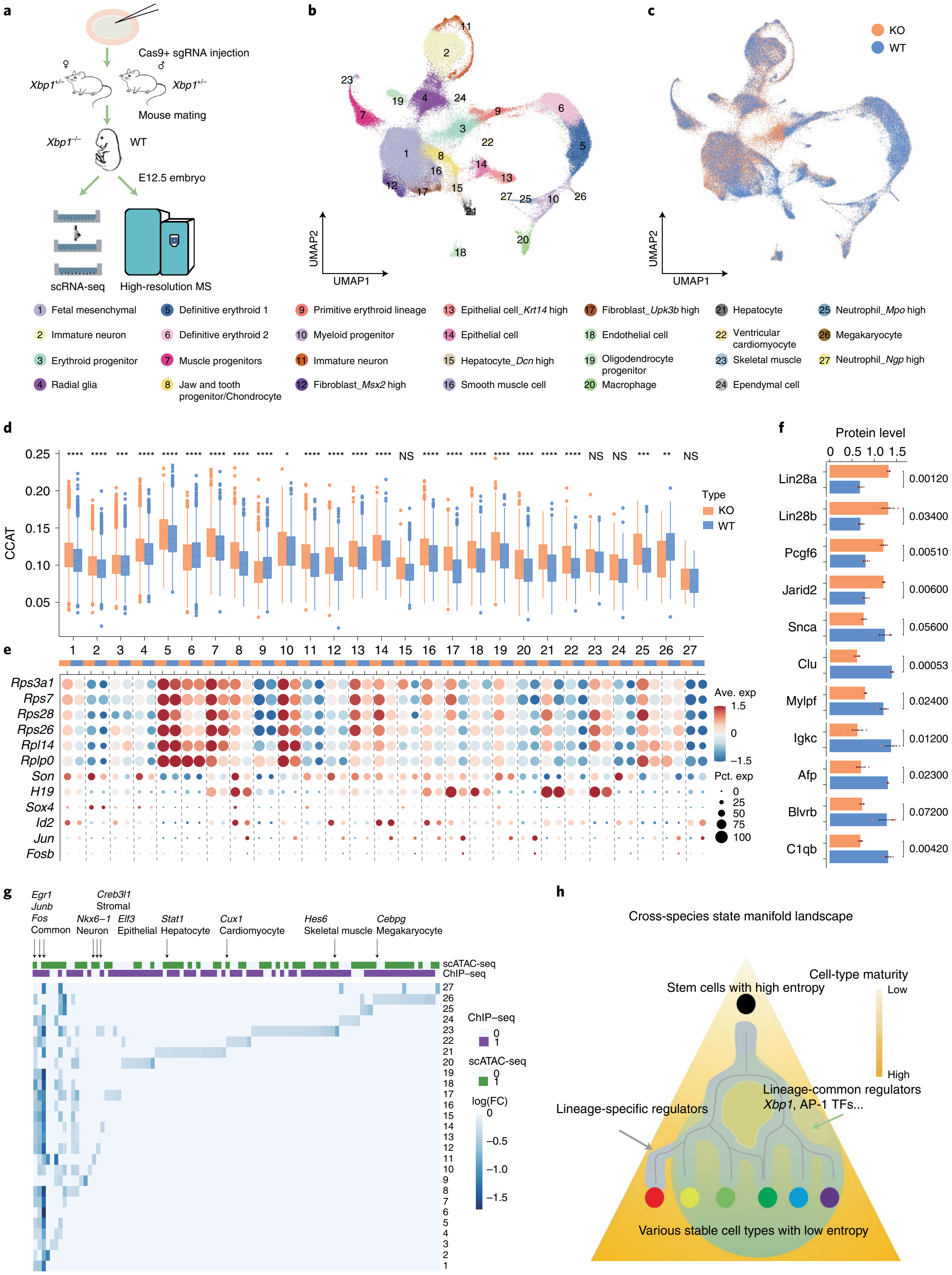

a
Cas9+ sgRNA injection
♀ Xbp1+/−
♂ Xbp1+/−
Mouse mating
Xbp1−/−
WT
E12.5 embryo
scRNA-seq
High-resolution MS
b
UMAP2
UMAP1
c
KO
WT
UMAP2
UMAP1
1 Fetal mesenchymal
2 Immature neuron
3 Erythroid progenitor
4 Radial glia
5 Definitive erythroid 1
6 Definitive erythroid 2
7 Muscle progenitors
8 Jaw and tooth progenitor/Chondrocyte
9 Primitive erythroid lineage
10 Myeloid progenitor
11 Immature neuron
12 Fibroblast_Msx2 high
13 Epithelial cell_Krt14 high
14 Epithelial cell
15 Hepatocyte_Dcn high
16 Smooth muscle cell
17 Fibroblast_Upk3b high
18 Endothelial cell
19 Oligodendrocyte progenitor
20 Macrophage
21 Hepatocyte
22 Ventricular cardiomyocyte
23 Skeletal muscle
24 Ependymal cell
25 Neutrophil_Mpo high
26 Megakaryocyte
27 Neutrophil_Ngp high
d
CCAT
0.25
0.20
0.15
0.10
0.05
**** **** *** **** **** **** **** **** **** * **** **** **** **** NS **** **** **** **** **** **** **** NS NS *** ** NS
Type
KO
WT
e
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
Rps3a1
Rps7
Rps28
Rps26
Rpl14
Rplp0
Son
H19
Sox4
Id2
Jun
Fosb
Ave. exp
1.5
0
−1.5
Pct. exp
0
25
50
75
100
f
Protein level
0 0.5 1.0 1.5
Lin28a 0.00120
Lin28b 0.03400
Pcgf6 0.00510
Jarid2 0.00600
Snca 0.05600
Clu 0.00053
Mylpf 0.02400
Igkc 0.01200
Afp 0.02300
Blvrb 0.07200
C1qb 0.00420
g
Egr1
Junb
Fos
Common
Nkx6−1
Neuron
Creb3l1
Stromal
Elf3
Epithelial
Stat1
Hepatocyte
Cux1
Cardiomyocyte
Hes6
Skeletal muscle
Cebpg
Megakaryocyte
scATAC-seq
ChIP–seq
27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
ChIP–seq
0
1
scATAC-seq
0
1
log(FC)
0
−0.5
−1.0
−1.5
h
Cross-species state manifold landscape
Stem cells with high entropy
Cell-type maturity
Low
High
Lineage-common regulators
Xbp1, AP-1 TFs...
Lineage-specific regulators
Various stable cell types with low entropy

Fig. 6 | ScRNA-seq and high-resolution MS revealed gene and protein changes in $Xbp1^{-/-}$ embryos. a, Overview of the CRISPR–Cas9 experiment. $Xbp1^{-/-}$ and WT embryos at E12.5 were prepared and processed by Microwell-seq and LC–MS. **b**,**c**, UMAP visualization of 93,246 single cells from $Xbp1^{-/-}$ and WT embryos at E12.5, colored by cluster identity (**b**) and genotype (**c**) (KO: $n = 49,498$; WT: $n = 43,748$). **d**, Entropy measurement of each cluster in $Xbp1^{-/-}$ and WT embryos using the CCAT method ($n = 93,246$ cells). The color represents the genotype. *P* values are from a two-sided Wilcoxon's rank-sum test comparing entropies of two groups with different genotypes from the same cluster. NS, not significant; $P > 0.05$, $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$, $^{****}P \leq 0.0001$. The exact *P* values have been displayed in Source data. Boxplots: center line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within 1.5 × IQR. **e**, Dot-plot showing representative DEGs (ribosomal protein genes, progenitor marker genes and lineage-common regulators) of each cluster in KO and WT cells. Ave. exp, Average expression; Pct. exp, Percentage of expressed cells. **f**, Barplot showing representative protein expression levels (pluripotency-related proteins and cell type-specific proteins) between $Xbp1^{-/-}$ and WT mice (KO: $n = 3$; WT: $n = 3$, mean ± s.d.). A two-sided Student's *t*-test was performed to determine the statistical significance. The illustrative genes were manually selected from the full heatmap, which is shown in Extended Data Fig. 10a. **g**, Heatmap showing significantly variable TFs in $Xbp1^{-/-}$ samples. Green and purple indicate the chromatin accessibility of the *Xbp1*-binding motif as determined by scATAC-seq and ChIP–seq, respectively. The *Xbp1*-binding motif of the mouse was from the CisBP database. Representative TFs are marked and were manually selected from Supplementary Table 23. Wilcoxon rank-sum test (two-sided) was performed to identify significantly variable TFs and *p*-value adjustment was performed using bonferroni correction (*p* adjusted values < 0.05 and fold change > 1.25). **h**, Schematic of cross-species state manifold landscape.

In addition, canonical *Xbp1* targets related to the unfolded protein response (UPR)[69–71] displayed no significant changes at the protein level (Extended Data Fig. 10b). Furthermore, *Xbp1* disruption in mouse embryonic stem cells (mECSs) did not alter stem cell culture and pluripotent gene expression, indicating that *Xbp1* transcriptional regulation of lineage decisions is not downstream of the UPR (Extended Data Fig. 10c,d and Supplementary Table 22).

We applied VarID[72] to qualify lineage-determining factor changes in scRNA-seq datasets. These significantly variable TFs in $Xbp1^{-/-}$ samples displayed an *Xbp1*-binding motif in both scATAC-seq and chromatin immunoprecipitation sequencing (ChIP–seq) data (Fig. 6g and Supplementary Table 23). Our results indicate a direct role of *Xbp1* in lineage maturation via transcriptional regulation, during which *Xbp1* functions through a mechanism that is independent of the UPR.

Discussion

Overall, our comprehensive MCDA atlas of mouse cell differentiation and maturation offers a powerful resource for investigating cell-fate decisions. We characterize a general feature of decreased entropy in most lineages during development. Our analysis of GRN dynamics reveals both lineage-common and lineage-specific regulators that contribute to cell-fate decisions. We highlight that *Xbp1* is a critical and conserved transcriptional regulator of cell-type differentiation in many lineages, as shown in our multi-omic analysis of *Xbp1* KO mouse embryos. However, the regulatory mechanisms of lineage-common regulators still require further research and functional validation in other settings such as in vitro differentiation, de-differentiation and trans-differentiation.

In the present study, we propose a systematic view of the cross-species state manifold landscape. Cells gradually progress from a stem/progenitor state toward specific cell fates with decreased entropy. During the process, divergent GRNs function following cell differentiation, including lineage-specific and lineage-common regulators. Lineage-specific TFs probably direct cell fate as potential regulators for the emergence of each cell type, whereas lineage-common ones probably represent general regulators to stabilize cell fates across various cell types, such as gravity through the process of state manifold[18,73]. We identify examples of common GRNs as conserved regulators of cell-fate stabilization (Fig. 6h). Thus, our work introduces a new functional classification of gene-regulatory programs to improve state manifold representations.

Tissue development and maturation atlases can provide global views of the cell-fate decision process. Using our data, we identified new cell types with co-expression and ectopic expression patterns during mouse development. We verified a myoendothelial cell type that co-expressed makers of myocytes and endothelial cells, a cell type that co-expressed makers of club and goblet cells and hepatocyte-like cell types in the pancreas during mouse development. We hypothesize that early transitional cell types may serve as a pool of progenitors to broadly support the normal progression of much functional tissue formation. We also observe new cell types in neonatal mice that will require further verification and characterization. By integrating developmental atlases across species, we describe common characteristics at varying evolutionary distances during development. Entropy is a concise, independent and robust measurement for differentiation potential and we further associate it with ribosomal protein gene expression in evolutionarily distant species.

Our single-cell analysis of KO mice provides a systematic insight into gene function at the organism level. Similar strategies can be applied to a series of KO embryos for the dissection of a functional GRN during development. It would also be interesting to compare quantitative gene function across different model systems and species. Specific combinations of functional regulatory networks across species may hint at evolutionary regularities of cell types. It is worth noting that limitations of single-cell technologies such as the digestion process, batch effects and sequencing depth should be taken into consideration during such analyses.

In conclusion, we constructed an MCDA and systematically characterized the cell-fate regulatory programs during development across species, which will lead to new understandings of cell-fate decisions and the cellular state manifolds.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01118-8.

Received: 19 July 2021; Accepted: 1 June 2022;
Published online: 11 July 2022

References

1. Mathis, L. & Nicolas, J.-F. Cellular patterning of the vertebrate embryo. *Trends Genet.* **18**, 627–635 (2002).
2. Heinäniemi, M. et al. Gene-pair expression signatures reveal lineage control. *Nat. Methods* **10**, 577–583 (2013).
3. Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).
4. Waddington, C. H. *The Strategy of the Genes* (Routledge, 2014).
5. Teschendorff, A. E. & Feinberg, A. P. Statistical mechanics meets single-cell biology. *Nat. Rev. Genet.* **22**, 459–476 (2021).
6. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).

7. Huang, S., Eichler, G., Bar-Yam, Y. & Ingber, D. E. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.* **94**, 128701 (2005).
8. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem. *Cell Biol. Cell* **132**, 631–644 (2008).
9. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987–1000 (1987).
10. Niwa, H. et al. Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell* **123**, 917–929 (2005).
11. Han, X. et al. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091–1107.e17 (2018).
12. Plass, M. et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018).
13. Siebert, S. et al. Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. *Science* **365**, eaav9314 (2019).
14. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
15. Packer, J. S. et al. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**, eaax1971 (2019).
16. Cao, C. et al. Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature* **571**, 349–354 (2019).
17. Mittnenzweig, M. et al. A single-embryo, single-cell time-resolved model for mouse gastrulation. *Cell* https://doi.org/10.1016/j.cell.2021.04.004 (2021).
18. Qiu, C. et al. Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nat. Genet.* **54**, 328–341 (2022).
19. Ferre, P., Decaux, J.-F., Issad, T. & Girard, J. Changes in energy metabolism during the suckling and weaning period in the newborn. *Reprod. Nutr. Dev.* **26**, 619–631 (1986).
20. Stewart, B. J. et al. Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–1466 (2019).
21. Chen, L. et al. Renal-tubule epithelial cell nomenclature for single-cell RNA-sequencing studies. *J. Am. Soc. Nephrol.* **30**, 1358–1364 (2019).
22. Zheng, B. et al. Prospective identification of myogenic endothelial cells in human skeletal muscle. *Nat. Biotechnol.* **25**, 1025–1034 (2007).
23. Chaudhari, P., Tian, L., Deshmukh, A. & Jang, Y.-Y. Expression kinetics of hepatic progenitor markers in cellular models of human liver development recapitulating hepatocyte and biliary cell fate commitment. *Exp. Biol. Med.* **241**, 1653–1662 (2016).
24. Willnow, D. et al. Quantitative lineage analysis identifies a hepato-pancreato-biliary progenitor niche. *Nature* **597**, 87–91 (2021).
25. Banerji, C. R. S. et al. Cellular network entropy as the energy potential in Waddington's differentiation landscape. *Sci. Rep.* **3**, 3039 (2013).
26. Grün, D. et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266–277 (2016).
27. Guo, M. SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res.* **45**, 14 (2017).
28. Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat. Commun.* **8**, 15599 (2017).
29. Teschendorff, A. E., Maity, A. K., Hu, X., Weiyan, C. & Lechner, M. Ultra-fast scalable estimation of single-cell differentiation potency from scRNA-Seq data. *Bioinformatics* **37**, 1528–1534 (2021).
30. Kauffman, S. Homeostasis and differentiation in random genetic control networks. *Nature* **224**, 177–178 (1969).
31. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
32. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
33. Holland, C. H. et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* **21**, 36 (2020).
34. Parker, H. J. Mammalian embryo: *Hox* genes. *eLS* 1–15 (2020).
35. Cassandri, M. et al. Zinc-finger proteins in health and disease. *Cell Death Discov.* **3**, 17071 (2017).
36. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
37. Jochum, W., Passegué, E. & Wagner, E. F. AP-1 in mouse development and tumorigenesis. *Oncogene* **20**, 2401–2412 (2001).
38. Velazquez, F. N., Caputto, B. L. & Boussin, F. D. c-Fos importance for brain development. *Aging* **7**, 1028 (2015).
39. Liu, J. et al. The oncogene c-Jun impedes somatic cell reprogramming. *Nat. Cell Biol.* **17**, 856–867 (2015).
40. Madrigal, P. & Alasoo, K. AP-1 takes centre stage in enhancer chromatin dynamics. *Trends Cell Biol.* **28**, 509–511 (2018).
41. Hinman, V. & Cary, G. The evolution of gene regulation. *eLife* **6**, e27291 (2017).
42. Li, J. *et al*. Inferring predictive genetic models and regulatory elements by deep learning of cross-species single-cell gene expression landscapes. Preprint at https://www.researchsquare.com/article/rs-1544073/v1 (2022).
43. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
44. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
45. Chakraborty, C. & Agoramoorthy, G. Stem cells in the light of evolution. *Indian J. Med. Res.* **135**, 813 (2012).
46. Saeed, R. & Deane, C. M. Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinform.* **7**, 128 (2006).
47. Boon, K. et al. *N-myc* enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis. *EMBO J.* **20**, 1383–1393 (2001).
48. Van Riggelen, J., Yetil, A. & Felsher, D. W. *MYC* as a regulator of ribosome biogenesis and protein synthesis. *Nat. Rev. Cancer* **10**, 301–309 (2010).
49. Shi, J., Teschendorff, A. E., Chen, W., Chen, L. & Li, T. Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. *Brief. Bioinform.* https://doi.org/10.1093/bib/bby093 (2018).
50. Farley-Barnes, K. I. et al. Diverse regulators of human ribosome biogenesis discovered by changes in nucleolar number. *Cell Rep.* **22**, 1923–1934 (2018).
51. Macaulay, I. C. et al. Single-cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep.* **14**, 966–977 (2016).
52. Athanasiadis, E. I. et al. Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nat. Commun.* **8**, 2045 (2017).
53. Krausgruber, T. et al. Structural cells are key regulators of organ-specific immune responses. *Nature* **583**, 296–302 (2020).
54. Tarashansky, A. J. et al. Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *eLife* **10**, e66747 (2021).
55. Fischer, S., Crow, M., Harris, B. D. & Gillis, J. Scaling up reproducible research for single-cell transcriptomics using MetaNeighbor. *Nat. Protoc.* **16**, 4031–4067 (2021).
56. Van de Sande, B. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276 (2020).
57. Wang, J. et al. Tracing cell-type evolution by cross-species comparison of cell atlases. *Cell Rep.* **34**, 108803 (2021).
58. Lee, A.-H., Chu, G. C., Iwakoshi, N. N. & Glimcher, L. H. XBP-1 is required for biogenesis of cellular secretory machinery of exocrine glands. *EMBO J.* **24**, 4368–4380 (2005).
59. Todd, D. J. et al. XBP1 governs late events in plasma cell differentiation and is not required for antigen-specific memory B cell development. *J. Exp. Med.* **206**, 2151–2159 (2009).
60. Pramanik, J. et al. Genome-wide analyses reveal the IRE1a-XBP1 pathway promotes T helper cell differentiation by resolving secretory stress and accelerating proliferation. *Genome Med.* **10**, 76 (2018).
61. Masaki, T., Yoshida, M. & Noguchi, S. Targeted disruption of CRE-binding factor *TREB5* gene leads to cellular necrosis in cardiac myocytes at the embryonic stage. *Biochem. Biophys. Res. Commun.* **261**, 350–356 (1999).
62. Reimold, A. M. et al. An essential role in liver development for transcription factor XBP-1. *Genes Dev.* **14**, 152–157 (2000).
63. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324.e18 (2018).
64. Di Bella, D. J. et al. Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature* **595**, 554–559 (2021).
65. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
66. Viswanathan, S. R. & Daley, G. Q. Lin28: a microRNA regulator with a macro role. *Cell* **140**, 445–449 (2010).
67. Yang, C.-S., Chang, K.-Y., Dang, J. & Rana, T. M. Polycomb group protein Pcgf6 acts as a master regulator to maintain embryonic stem cell identity. *Sci. Rep.* **6**, 26899 (2016).
68. Pasini, D. et al. JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature* **464**, 306–310 (2010).
69. Yoshida, H., Matsui, T., Yamamoto, A., Okada, T. & Mori, K. XBP1 mRNA Is Induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell* **107**, 881–891 (2001).
70. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882.e21 (2016).
71. Acosta-Alvear, D. et al. XBP1 controls diverse cell type- and condition-specific transcriptional regulatory networks. *Mol. Cell* **27**, 53–66 (2007).
72. Grün, D. Revealing dynamics of gene expression variability in cell state space. *Nat. Methods* **17**, 45–49 (2020).
73. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* https://doi.org/10.1038/s41586-019-0933-9 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Mouse experiments to supplement the MCDA database. WT C57BL/6J mice were ordered from Shanghai Laboratory Animal Center. All mice were housed at Zhejiang University Laboratory Animal Center in a specific pathogen-free facility with individually ventilated cages. The room had a controlled temperature (20–22 °C), humidity (30–70%) and light program (12 h light:dark cycle). The mice were provided free access to a regular rodent chow diet.

To obtain embryonic samples (E10.5 embryos, E12.5 embryos), C57BL/6 mice were mated. Noon on the day the vaginal plug was visible was considered to be E0.5. Sex was not determined before tissue pooling for E10.5, E12.5 and P0 samples (except for the gonads). Embryos were collected from at least three independent litters (in total three to nine embryos) per development stage. For P10 and P21 samples, testes were collected from male mice and all the other tissues were collected from female mice.

All experiments performed in the present study were approved by the Animal Ethics Committee of Zhejiang University. All experiments conformed to the relevant regulatory standards at Zhejiang University Laboratory Animal Center.

Generation of *Xbp1* KO mESC and mouse models. SgRNAs targeting exon 2 of *Xbp1* were designed using the Zhang laboratory CRISPR design website tool (http://crispr.mit.edu). Oligonucleotides were synthesized and then cloned into an epiCRISPR–Cas9 vector[74]. The vector was extracted using an EndoFree Mini Plasmid Kit II (Tiangen Biotech, catalog no. 4992422) following the manual. Approximately 4×10^5 E14 mESCs were transfected with 2 µg of the vector with Lipofectamine 3000 (Life Technologies, catalog no. L3000001) based on an online protocol. At days 2–10, cells were selected with puromycin (0.5–1.0 µg ml^{-1}). Then, single cells were reseeded in a 6-well plate and cultured in mESC media for 7–10 d. Individual colonies were picked and genotyped. The genomic RNA target sites and oligonucleotides used in the present study can be found in Supplementary Table 22.

Xbp1 KO C57BL/6J mice were generated by Nanjing Gempharmatech. Mice were genotyped by PCR using genomic tail DNA. To obtain live KO embryos at E12.5 for scRNA-seq, we used a Scientific Phire Animal Tissue Direct PCR Kit (Thermo Fisher Scientific, catalog no. F140WH) to genotype embryos quickly. All primers used for KO and genotyping are listed in Supplementary Table 18.

Immunofluorescent staining. Fresh mouse tissues were frozen in disposable molds containing optimal cutting temperature compound. Frozen sections were cut at 10 µm in CryoStar NX50 (Thermo Fisher Scientific), mounted on microscope slides and stored at −80 °C. Before staining, the sections were thawed for 20 min and 4% formaldehyde in phosphate-buffered saline (PBS) was added to cover the sections. Tissues were fixed for 15 min at room temperature. After fixation, sections were washed three times with PBS. Cells were permeabilized by covering the sections with 0.1% Triton X-100 in PBS for 10 min. Then, the sections were washed three times with PBS and blocked with 3% bovine serum albumin in PBS for 1 h at room temperature. Primary antibodies (anti-ESAM (1:50; Thermo Fisher Scientific, catalog no. MA5-24072), anti-Myl9 (1:400; Abcam, catalog no. ab187152), anti-Scgb1a1 (1:50, R&D, catalog no. MAB4218-SP), anti-tff2 (1:200; ProteinTech, catalog no. 13681-1-AP) and anti-AFP (1:200, Affinity, catalog no. AF5134)) diluted in blocking solution were added to cover the sections. The slides were placed in a wet box and incubated overnight at 4 °C. Relevant Alexa Fluor-488/594-conjugated secondary antibodies (1:1,000; Thermo Fisher Scientific, catalog nos. A-21208, A-21206 and A-11037) were used for labeling. The slides were then washed three times with blocking solution and stained with DAPI. Glass coverslips were then attached to the slides using mounting media. Immunofluorescence images were obtained using Olympus VS200.

Western blot. The mouse embryos were solubilized in radioimmunoprecipitation assay lysis buffer (20 mg per 200 µl; Beyotime, catalog no. P0013D). The mixture was lysed using a homogenizer for 5 min on ice. Tissue lysates were then cleared by centrifugation at 14,000*g* for 10 min at 4 °C. Equal amounts of total protein were used for experimental and control. Samples were fractionated using sodium dodecylsulfate (SDS)–polyacrylamide gel electrophoresis and transferred to a poly(vinylidene fluoride) membrane. After blocking with 5% milk in tris-buffered saline + Tween (TBST) for 1 h at room temperature, the membranes were probed with the corresponding primary and secondary antibodies. Primary antibodies (anti-Xbp1 (1:1,000; Abcam, catalog no. ab37152), anti-β-tubulin (1:3,000; HUABIO, catalog no. EM0103)) and secondary antibodies (anti-mouse immunoglobulin (Ig)G (1:5,000; TransGen Biotech, catalog no. HS201-01), anti-rabbit IgG (1:5,000; Multi Science, catalog no. GAR007)) diluted in TBST were used.

Cell preparation. Mouse tissues were minced into pieces of ~1 mm on ice using scissors. The tissue pieces were transferred to a 15-ml centrifuge tube, rinsed twice with cold Dulbecco's (D)PBS and suspended in 5 ml of a solution containing dissociation enzymes. The samples were treated with various enzymes for different periods of time (Supplementary Table 3). During dissociation, the tissue pieces were pipetted up and down gently several times until no tissue fragments were visible. The dissociated cells were centrifuged at 300*g* for 5 min at 4 °C and then resuspended in 3 ml of cold DPBS. After passage through a 40-µm strainer (Biologix), the cells were washed twice, centrifuged at 300*g* for 5 min at 4 °C and resuspended at a density of 1×10^5 cells ml^{-1} in cold DPBS containing 2 mM EDTA.

ScRNA-seq. Single-cell complementary DNA libraries were prepared using the Microwell-seq[11]. Briefly, cells were loaded on the microwell plate and extra cells were washed away gently using ice-cold PBS. Then bead suspension (sequences listed in Supplementary Table 2) was loaded on the plate and extra beads were washed away on a magnet. The plate was covered using cold lysis buffer (0.1 M Tris-HCl, pH 7.5, 0.5 M LiCl, 1% SDS, 10 mM EDTA and 5 mM dithiothreitol (DTT)) and incubated on ice for 12 min. Then, beads were collected and washed using 6 × saline sodium citrate and 50 mM Tris-HCl, pH 8.0. After washing, beads were resuspended in reverse transcription (RT) mix and incubated at 42 °C for 90 min. After RT, beads were washed in TE–TW (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 0.01% Tween20) and 10 mM Tris-HCl, pH 8.0. Beads were resuspended in exonuclease I mix and incubated at 37 °C for 30 min. Then, beads were washed in TE–SDS (1XTE + 0.5% sodium dodecyl sulfate), TE–TW and 10 mM Tris-HCl, pH 8.0. Beads were resuspended in PCR mix with TSO (template switch oligo) primer to amplify the cDNA. After PCR, beads were removed and cDNA products were purified using 0.8 × VAHTS DNA Clean Beads (Vazyme. catalog no. N411-01). A more detailed version of the Microwell-seq protocol is available in Han et al.[14]. Then, the purified cDNA libraries were fragmented using a customized transposase that carries two identical insertion sequences. The customized transposase was included in the TruePrep Homo-N7 DNA Library Prep Kit for Illumina (Vazyme, catalog no. TD513) or TruePrep Homo-N7 DNA Library Prep Kit for MGI (Vazyme, catalog no. L-N7E461L0). The fragmentation reaction was performed according to the instructions provided by the manufacturer. We used customized P5 primer (listed in Supplementary Table 2) and VAHTS RNA Adapters set3-set6 for Illumina (Vazyme, catalog no. N809/N810/N811/N812) or our MGI P7 primers (N8XX, listed in Supplementary Table 2) to specifically amplify fragments that contain the 3′-ends of transcripts. Other fragments will form self-loops, impeding their binding to PCR primers. The PCR program was as follows: 72 °C for 3 min; 98 °C for 1 min; 5 cycles of 98 °C for 15 s, 60 °C for 30 s and 72 °C for 3 min; 72 °C for 5 min; and a 4 °C hold. The PCR product was purified using 0.9 × VAHTS DNA Clean beads (Vazyme, catalog no. N411-01). Then, a 25-µl PCR mix (1 × HiFi HotStart Readymix and 0.2 µM 2100 primer) was added to each sample. The PCR program was as follows: 95 °C for 3 min; 5 cycles of 98 °C for 20 s, 60 °C for 15 s and 72 °C for 15 s; 72 °C for 3 min; and a 4 °C hold. To eliminate primer dimers and large fragments, 0.55–0.15× VAHTS DNA Clean beads were then used to purify the cDNA library. The size distribution of the products was analyzed on an Agilent 2100 bioanalyzer, and a peak in the range 400–700 bp was observed. Finally, the samples were subjected to sequencing on the Illumina HiSeq (data for MDCA) or MGI DNBSEQ-T7 (data for *Xbp1* KO experiment). For MGI sequencing, we applied the protocol provided by the VAHTS Circularization Kit for MGI (Vazyme, catalog no. NM201-01) to obtain single-stranded circular cDNA available for DNB (DNA Nanoball) generation. We also replaced the official R1 sequencing primers with our customized R1 sequencing primers A and B (listed in Supplementary Table 2) to ensure the completion of the sequencing.

Processing of Microwell-seq data. Microwell-seq datasets were processed as described[11]. Reads were aligned to the Mus_musculus. GRCm38.88 genome using STAR[75] (v.2.5.2a). The digital gene expression (DGE) data matrices were obtained using the Drop-seq core computational protocol (available at website http://mccarrolllab.org/dropseq) with the default parameters. For quality control, we filtered out cells with detection of < 500 transcripts. Cells with a high proportion of transcript counts (> 20%) derived from mitochondria-encoded genes were also excluded. Cells were also corrected for RNA contamination and background-removed DGE data were constructed[14]. The SCANPY[76] (v.1.6.0) python package and Seurat[77] (v.3.2.2) R package were used to load the cell-gene count matrix and perform downstream analysis.

Clustering of the single-cell data matrix. For clustering of the complete mouse tissue dataset (520,801 cells), qualified cells were processed using SCANPY (v.1.6.0) in a Python (v.3.6.9) environment. Background-removed DGE data for cells analyzed in each tissue and genes expressed in at least 20 cells were used as inputs[14]. Then, DGE data were ln(c.p.m./(100 + 1)) transformed (where c.p.m. is counts min^{-1}). We selected approximately 3,000 highly variable genes according to their average expression and dispersion. We then regressed out unique molecular identifiers and gene numbers and scaled each gene to unit variance, and the values beyond an s.d. of 10 were clipped. For the mouse tissue dataset, we chose PCs for principal component analysis (PCA) according to elbow plots and 50 PCs were used to create a neighborhood graph for the cells. We then used Leiden clustering to cluster with resolution = 8 and $k = 25$. Marker genes were calculated using Wilcoxon's rank-sum test (two-sided) and p-value adjustment was performed using the Benjamini–Hochberg correction. For visualization, *t*-SNE was used.

For kidney data, bbknn[78] (v.1.4.0) was performed by using ridge regression to remove batch effects. For clustering of single tissues, the Seurat pipeline was used with the default parameters for fewer cells. Cell type and lineage information of each cell type were manually annotated according to the marker genes reported in a previous paper[11]. A hierarchical tree of the MCDA was computed using the correlations of average gene expression of 95 clusters with highly variable genes.

Estimation of the variance of the MCDA. To estimate the variance in the data depending on age, tissue or sex, we first aggregated the gene expression for each tissue at multiple time points. Using the above metadata as input, we performed principal variance component analysis (PVCA) using R Package pvca (v.1.26.0, https://www.bioconductor.org/packages/release/bioc/html/pvca.html) with the default parameters. It leverages the strengths of two popular data analysis methods PCA and variance components analysis and integrates them into a new algorithm. It also uses the eigenvalues associated with their corresponding eigenvectors as weights, to quantify the magnitude of each source of variability. All factors as well as their interaction terms are treated as random effects in the mixed model for variance component estimation. It fits a linear mix-effects model to data. Items such as 'tissue' and 'gender' are variances explained by interactions of two factors instead of the union of two factors.

Inference of the TFs for MCDA. As a proof of principle, we applied experimentally verified, tissue-specific TFs from the literature[79] as the gold standard. We included both tissue-restricted TFs and nonuniformly expressed TFs in different tissues as tissue-specific TFs. For datasets used, we selected high-quality cells with > 800 gene numbers as single-cell datasets, and also aggregated every 20 single cells in each cell type to produce pseudo-cells to enrich our choices of input datasets. We compared SCENIC[32] (v.0.10.0) and VIPER-DOROTHEA[33] (viper v.1.28.0 and dorothea v.1.6.0) for inferring specific TFs in the tissues. The DOROTHEA database provided TFs from different types of evidence with a different confidence. We used ABCDE (1,113 TFs) categories of DOROTHEA TFs in our comparison. Regulon specificity scores (RSSs)[56] were calculated to represent TF specificity in the tissue for both VIPER-DOROTHEA and SCENIC. Then we employed the youden index (sensitivity + specificity − 1) to find the best performance of VIPER-DOROTHEA and SCENIC in classifying tissue-specific TFs in both sensitivity and specificity. These TFs were compared with the gold standard lists for four aspects: sensitivity, specificity, false-positive rate and area under the precision-recall curve.

To define regulatory programs in MCDA, SCENIC and VIPER-DOROTHEA were applied first to infer the GRN with default parameters using high-quality single cells with > 800 genes. For VIPER-DOROTHEA, ABCDE (1,113 TFs) categories of DOROTHEA TFs were used. Second, z-scaled RSSs for VIPER-DOROTHEA and z-scaled TF activity scores of SCENIC in each stage lineage were calculated as a TF-by-lineage matrix. Then, fuzzy c-means clustering was performed on the TF-by-lineage matrix calculated by SCENIC and VIPER-DOROTHEA, resulting in a TF-by-module 'membership matrix' and a lineage-by-module 'centers matrix'. The centers matrix with 15 modules was used to generate the heatmap. We defined a threshold membership score (threshold = 0.2) in which TFs were assigned to a module. With the fuzzy c-means heatmap, we identified which modules/TFs were lineage specific and which were lineage sharing. We assigned TFs into specific lineages according to the aggregated patterns of modules manually and the resulting TFs were classified into three collections with high to low confidence: collection A consisted of TFs from both methods, collection B were TFs only from SCENIC and collection C TFs only from VIPER-DOROTHEA (Supplementary Table 10).

Analysis of time-related genes during cell-type maturation. Early organ formation in mice begins at E10.5 and cells undergo differentiation to reach maturity during development[80]. Thus, we identified time-related genes that showed upregulation patterns at the expression levels during the developmental processes by using Spearman's rank correlation analysis for different lineages in each tissue[81]. Spearman's rank correlation coefficient, which has low requirements on data distribution and a high tolerance for outliers, can directly reflect the monotonous relationship between variables, so we adopted it. We treated the seven-stage information (E10.5, E12.5, E14.5, P0, P10, P21 and adult) as the vectors labeled (1, 2, 3, 4, 5, 6 and 7), and then calculated the correlation between the gene expression levels across seven development stages and the vectors for each stage. The larger the absolute value of the correlation coefficient, the stronger the monotonicity of the gene expression level and timepoints. The TFs with Spearman's rank correlation coefficient ≥ 0.8 in at least four lineages in five tissues with a $P \leq 0.05$ were retained as the common time-upregulated TFs during lineage maturation.

Single-cell entropy analysis. Single-cell entropy estimation was performed using three methods: CCAT[29] (SCENT v.1.0.2), SLICE[27] (v.0.99.0) and StemID[26] (RaceID v.0.2.2). To obtain the best performance, normalization was dependent on the computational methods. For CCAT, it is an approximation of network entropy. We applied CCAT to compute the correlations with the connectome and transcriptome based on the 'net13Jun12.m' PPIs. We performed CCAT analysis by using a weighted matrix to leverage all the homology genes between human and other species. The weighted matrix was obtained by converting the gene homology relationship (one-to-one, one-to-many, many-to-one and many-to-many) into a binary matrix and normalized it to one human gene. In StemID, it estimates the Shannon entropy of a cell's transcriptome directly based on the expression of each gene. We used StemID to infer entropies with default parameters. For SLICE, it established a kappa matrix of gene ontology (GO) annotations of the human or mouse to evaluate the probability distribution of the functional activation of each cell. SLICE was performed as a deterministic calculation of scEntropy of individual cells over the GO cluster activation profile with iter = 50. Cells were downsampled to 2,000 per tissue per stage to cut the calculation burden of SLICE and StemID. In summary, CCAT calculates the entropy-related values from the perspective of the network entropy of the gene interaction network. SLICE and StemID calculate the entropy values by using the activation of the gene pathway and the gene expression as probabilistic events, respectively. Although the principles of the three methods are different, their central idea is to couple entropy with developmental potential. They evaluate biological systems using physical concepts and reflect the physical properties of biological systems.

Construction of a cell-type hierarchy across species and gene regulation analysis. For invertebrates, to infer the topological relations of cell-type development, we first constructed a PAGA graph[43] per lineage using SCANPY (v.1.6.0). We processed the data following the steps suggested by SCANPY, including total count normalization, log(1P) transformation, highly variable gene extraction, potential regression of confounding factors of genes and counts, scaling to z-scores and PCA. Then, we computed a neighborhood graph among data points and used UMAP for topologically faithful embedding with min_dist = 0.1. Then, PAGA was performed with iter = 1,000. The cell-type tree layout was based on a minimum spanning tree fitted to edges weighted by inverse connectivity. Edges in an abstracted graph with a probability > 0.0005 were considered as possible connections of cell-type hierarchies. For *S. mediterranea*, cell-type hierarchies were obtained from the consolidated lineage tree, which was provided in a paper[12] and, for *C. intestinalis*, lineage and stage information were directly from a paper[16]. For complex vertebrates, we connected cell states across time according to gene expression similarity[44]. For each tissue, we asked each adult cluster to 'vote' on its most likely ancestor cluster from the fetal stage. To eliminate the influence of cell number, we randomly sampled 150 cells to embed them into the PCA space learned from the second timepoint only and kept nontrivial PCs as defined above. Then, in this embedding, for each cluster in the late timepoint, the cluster identities of the five nearest neighbors of each constituent cell from the previous timepoint were determined using a Euclidean distance metric. The percentages of votes cast for each possible ancestor were calculated and the maximum frequencies of votes (20–100%) of the cells in the cell group decided the ancestor group. For zebrafish datasets, we integrated the data using Seurat (v.4.0.1) which anchors integration functions to do the batch correction before the PCA. Sankey plots were generated using the networkD3 (v.0.4, https://christophergandrud.github.io/networkD3) R package. For atlas projects across species, we performed the same differential expression analysis for cells in each tissue-cell type/lineage-cell type separately according to the cell-type hierarchy using FindMarkers function in Seurat (v.4.0.1). Wilcoxon's rank-sum test was performed to determine the statistical significance and the Benjamini–Hochberg correction was used for the p-value adjustment. The top common DEGs (20–100% of total cell-type pairs, mean 60–94% lineages, p adjusted values < 0.1, $\log_2$(fold-change) ($\log_2$(FC)) ≥ 0.25, min_pct ≥ 0.1) were estimated according to the frequency of differential expression in all unstable-to-stable cell-type pairs across species. To match the two timepoints (fetal and adult) of humans, only the E14.5 and adult stages of mice and 24-h post-fertilization and 3-month stages of zebrafish were considered for cross-species analysis. Genes that display consistent patterns in at least three species were defined as commonly upregulated and downregulated genes. Genes that were either 'up-' or 'down-'regulated were excluded in the analysis.

The top 20 most negative TFs of the upregulated TFs were determined by Pearson's correlations based on single-cell datasets and visualized by Cytoscape (v.3.5.0)[82].

Collection and prediction of orthologous genes and TFs. For *H. sapiens, M. musculus*, *D. rerio* and *C. elegans*, orthologous pairs were obtained from Ensembl v.96 by BioMarkt. The transcriptome of *S. mediterranea* was downloaded from the PlanMine database[83] (*S. mediterranea*, dd_Smed_v6). The transcriptome of *H. vulgaris* was downloaded from the website https://research.nhgri.nih.gov/hydra/download/?dl=tr. The transcriptome of *C. intestinalis* was downloaded from http://ghost.zool.kyoto-u.ac.jp/download_kh.html. Then, the protein-coding sequence (CDS) was predicted by TransDecoder[84] (v.5.3.0) with the default parameters. Orthologous pairs were predicted by OrthoFinder[85] (v.2.2.6) with CDS files as the input. In the present study, we considered only one-to-one orthologous pairs with humans for commonly regulated genes. As for species-specific TFs, TFs of *H. sapiens, M. musculus*, *D. rerio* and *C. elegans* were downloaded from the AnimalTFDB 3.0 database[86]. Other species-specific TFs except *H. vulgaris* were obtained from a paper[57]. Genes from *H. vulgaris* were obtained with Swiss-Prot IDs of best hits. Thus, the TFs of *H. vulgaris* were defined by the genes annotated with the GO terms downloaded from the uniport website: DNA-binding TF activity or TF binding. Those Swiss-Prot IDs of best hits were also checked for TFs from AnimalTFDB 3.0 and used as a supplement to TFs.

Lineage-specific TFs analysis across species. We applied two methods to calculate the lineage evolution relationship across species with the pseudo-cell as inputs (aggregated every 20 cells from each cell type): SAMap[54] (v.0.3.0) and MetaNeighbor[55] (pyMN v.0.1.0). SAMap enables mapping single-cell transcriptomes between phylogenetically remote species based on the expression similarity whereas MetaNeighbor has high replicability in cell-type matching using homologous weighted gene matrices. For SAMap, it constructs a gene–gene bipartite graph with cross-species edges connecting homologous gene pairs, weighted by protein sequence similarity. For MetaNeighbor, we constructed weighted matrices to leverage all the homology genes between humans and other species. The weighted matrices were obtained by converting the gene homology relationship (one-to-one, one-to-many, many-to-one and many-to-many) into a binary matrix and normalized it to one human gene each. Lineage pairs with high confidence thresholds (alignment scores with > 0.5 in SAMap and Mean_AUROC > 0.8 in MetaNeighbor) were considered as highly reliable and biologically plausible matches from different aspects. The combined projection of seven species was obtained from the function 'AMAP.scatter' of SAMap. The specificity of TFs was characterized using modified regulon specificity scores in SCENIC with TF expression count matrices as input[56,57]. We then calculated the z-score-normalized TF specificity score to predict the essential TFs in each lineage. Development-related, lineage-specific TFs were intersected with upregulated genes across species. The sequence similarity score was determined by the National Center for Biotechnology Information's (NCBI's) BLAST with transcriptome or proteome data as inputs. An E-value threshold of 1×10^6 was set. It was also integrated into SAMap.

Pathway enrichment analysis. We used clusterProfiler[87] (v.3.14.3) to perform GO biological pathway enrichment analysis and orthologous genes were taken as the universe. Hypergeometric test was performed to identify significant GO terms and the Benjamini–Hochberg correction was used to adjust p-values. We considered biological pathways with p adjusted values < 0.05. We used REVIGO[88] to visualize the enrichment results. For Extended Data Fig. 1e, we used clusterProfiler to perform GO biological pathway enrichment analysis for DEGs at neighboring stages. We considered biological pathways with p adjusted values ≤ 0.01. For each stage, the enrichment terms, as determined by clusterProfiler, were used to manually combine into 13 'super terms' for biological processes. For Extended Data Fig. 1i, GO enrichment analysis was first computed using the DEGs of the kidneys. Then, the enrichment scores of the terms were calculated and aggregated for each stage using AUCell[32].

PPI analysis. We downloaded the PPI resource of human genes from STRING[89] (v.11). Experimentally validated interactions from humans and transferred by homology from other species were used for the analysis. Then, we compared the $\log_{10}$(PPI no.) of four groups, the upregulated genes in at least three species, downregulated genes in at least three species, other conserved genes in at least three species and all other genes in the PPI resource. We also downloaded the gene functional assignments from the eggNOG database (v.5.0) and used the mammals' nonsupervised orthologous groups (maNOG) to assign genes into 26 categories[90]. The 26 gene categories were arranged by their average number of PPIs in ascending orders. Statistical analyses were done with R package ggpubr (v.0.4.0, https://rpkgs.datanovia.com/ggpubr) for two-tailed Wilcoxon's rank-sum test to determine the statistical significance of the differences between two groups.

Analysis of the CCAT-driving gene across species. CCAT directly measures the correlation between transcriptome and connectome and will therefore be positive if most of the network hubs are overexpressed in more potent cells[29]. Thus, we used the number of adjacent edges to evaluate the degree of each gene in the PPI network and the top 20% of genes were regarded as network hubs. We intersected them with the commonly downregulated genes we found in the manuscript (highly expressed in undifferentiated cells, Supplementary Table 13) in each species as CCAT-driving genes in more potent cells. Genes that appeared in at least five species were regarded as conserved CCAT-driving genes. We performed gene enrichment analysis using clusterProfiler on those conserved CCAT-driving genes. The biological processes related to ribosome biogenesis were marked red according to a previous paper[50].

Analysis of *Xbp1* expression pattern in MCDA. Given the low detection rate of TFs in the single-cell experiment, we chose high-quality cells with > 800 genes and calculated the average expression of *Xbp1* by normalization to a group of stably expressed gene sets generated from scMerge R package (v.1.2.0, https://bioconductor.org/packages/release/bioc/html/scMerge.html). We used linear regression to measure the expression trend of *Xbp1* with a 95% confidence interval.

Cell-type composition analysis. Significant differences in cell-type composition between groups were assessed using a propeller test from the speckle R package (v.0.0.1, https://github.com/Oshlack/speckle/). We considered groups with false discovery rate (FDR) ≤ 0.01 to represent significantly changed cell types.

Gene expression variability analysis. To detect sensitive changes in weakly expressed genes, we calculated the gene expression variability using VarID[72] (RaceID, v.0.2.2). We ran VarID with regNB=FALSE, $k = 10$ for the pruning step, no_cores=10 and default parameters otherwise.

Analysis of global proteomics data. LC–MS proteomic analysis was carried out by PTM Bio[91]. Briefly, mouse embryos were ground into powder in liquid nitrogen and suspended in an ice-cold lysis buffer with 1% Triton X-100 and 1% protease inhibitor based on occasional sonication. The cell lysates were centrifuged at 12,000g and 4 °C for 15 min. The supernatants were collected and the protein concentration was measured. Proteins were precipitated using 20% trichloroacetic acid for 2 h at 4 °C and then centrifuged at 4,500g for 5 min. The precipitate was washed three times with cold acetone. The dried protein pellets were resuspended within 200 mM tetraethylammonium bromide based on occasional sonication and then digested with trypsin overnight. DTT was added to a final concentration of 5 mM and the supernatants were incubated at 56 °C for 30 min. Iodoacetamide was added to a final concentration of 11 mM and the supernatants were incubated in the dark for 15 min. Peptides were separated using NanoElute and analyzed using timsTOF Pro. The resulting MS–MS data were processed using the MaxQuant search engine (v.1.5.2.8, https://www.maxquant.org) and mapped to the Mus_musculus_10090 database. The FDR was adjusted to < 1% and the minimum score for modified peptides was set to >40. Trypsin/P was defined as the cleavage enzyme, and up to two missing cleavages were allowed. For proteomic analysis, the first search range was set to 5 p.p.m. for precursor ions and the main search range was set to 5 p.p.m. and 0.02 Da for fragment ions. Carbamidomethylation of cysteines was defined as the fixed modification and oxidation on methionine was defined as the variable modification. The quantification method used was label-free quantification, the FDR was adjusted to <1% and the minimum score for modified peptides was > 40.

ScATAC-seq and ChIP–seq data analysis. We used ChromVAR[92] (v.1.12.0) to calculate the accessibility of the *Xbp1* motif in scATAC-seq datasets for comparing the *Xbp1* motif enrichment between differentiated states and undifferentiated states in both the human and the mouse. The mouse scATAC-seq data were downloaded from two papers[63,64] and human scATAC-seq data from another paper[65]. The motif PWM was downloaded from the CisBP database (http://cisbp.ccbr.utoronto.ca). For better visualization, we arranged the cells according to their differentiated states. This comparison was restricted to the cell-type annotations provided. As shown, the *Xbp1* motif was less opened in undifferentiated cells in both human and mouse tissues in neuron cell types and hematopoietic cell types. ChIP–seq data for *Xbp1* were downloaded from previous studies[60,71,93,94]. The target genes were binarized and integrated for visualization.

Statistics and reproducibility. No statistical methods were used to predetermine sample size; 520,801 single cells were analyzed in total for a time-series MCDA construction. A total of 52 mouse tissues from different development stages were analyzed. Two to four replications were done for different tissues. The results of major cell-type clusters are reproducible. Experimental mice and embryos were randomized before sample preparation. Different single cells were randomly captured before analysis. For all experiments, investigators were blinded to group allocation during the data collection and analysis. All related statistical methods and sample size are described in the figure legends and Methods.

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

Data availability

The data generated in the present study can be downloaded from the NCBI's Gene Expression Omnibus under accession nos. GSE176063 and GSE178217. The raw and processed files of MCDA are at accession no. GSE176063. The raw and processed files of WT and *Xbp1* KO embryos are at accession no. GSE178217. Processed count matrices and cell annotations are provided on the figshare website (https://figshare.com/s/340e8e7f349559f61ef6), including the development stage, tissue of origin, lineage information and cell-type annotations. We have provided separate datasets for each tissue and the merged datasets for the MCDA. We have also provided an interactive website (http://bis.zju.edu.cn/MCA) to enable public access to the data. The proteomics data was provided in the Proteomics Identifications Database (PRIDE) under accession no. PXD032847. The following publicly available datasets were used in the study: Mus_musculus. GRCm38.88 genome, Mus_musculus_10090 database, AnimalTFDB 3.0 database, STRING database (v.1.1), eggNOG database (v.5.0), Ensembl v.96; the *S. mediterranea* dataset generated by Plass et al.[12] (accession no. GSE103633), the *C. elegan* dataset generated by Packer et al.[15] (accession no. GSE126954.); the *C. intestinalis* dataset generated by Cao et al.[16] (accession no. GSE131155); the *H. vulgaris* dataset generated by Siebert et al.[13] (accession no. GSE121617); the *D. rerio* dataset generated by Li et al.[42] (GSE178151); the *H. sapiens* dataset generated by Han et al.[14] (GSE134355); and part of the *M. musculus* dataset (E14.5 and adult) generated by Han et al.[11] (accession nos. GSE108097 and GSE134355). The mouse scATAC-seq dataset was generated by Cusanovich et al.[63] (accession no. GSE111586, https://atlas.gs.washington.edu/mouse-atac/data) and Di Bella et al.[64]

(accession no. GSE153164), and the human scATAC-seq dataset by Domcke et al.[65] (descartes.brotmanbaty.org).

Code availability

Detailed code is available at GitHub (https://github.com/ggjlab/MCDA) and Zenodo (https://zenodo.org/record/6548256#.Yn92F-hBw2w)[95].

References

74. Xie, Y. et al. An episomal vector-based CRISPR/Cas9 system for highly efficient gene knockout in human pluripotent stem cells. *Sci. Rep.* **7**, 2320 (2017).
75. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
76. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
77. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
78. Park, J.-E. et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, eaay3224 (2020).
79. Zhou, Q. et al. A mouse tissue transcription factor atlas. *Nat. Commun.* **8**, 15089 (2017).
80. Lambert, L. J., Muzumdar, M. D., Rideout III, W. M. & Jacks, T. Basic mouse methods for clinician researchers: harnessing the mouse for biomedical research. in *Basic Science Methods for Clinical Researchers* 291–312 (Elsevier, 2017).
81. Teschendorff, A. E. & Wang, N. Improved detection of tumor suppressor events in single-cell RNA-Seq data. *NPJ Genom. Med.* **5**, 43 (2020).
82. Saito, R. et al. A travel guide to Cytoscape plugins. *Nat. Methods* **9**, 1069–1076 (2012).
83. Rozanski, A. et al. PlanMine 3.0—improvements to a mineable resource of flatworm biology and biodiversity. *Nucleic Acids Res.* **47**, D812–D820 (2019).
84. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
85. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
86. Hu, H. et al. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.* **47**, D33–D38 (2019).
87. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: J. Integr. Biol.* **16**, 284–287 (2012).
88. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
89. Mering, Cvon et al. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).
90. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
91. Song, Y. et al. Screening of potential biomarkers for gastric cancer with diagnostic value using label-free global proteome Analysis. *Genom. Proteom. Bioinform.* **18**, 679–695 (2020).
92. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
93. Argemí, J. et al. X-box binding protein 1 regulates unfolded protein, acute-phase, and DNA damage responses during regeneration of mouse liver. *Gastroenterology* **152**, 1203–1216. e15 (2017).
94. Khetchoumian, K. et al. Pituitary cell translation and secretory capacities are enhanced cell autonomously by the transcription factor Creb3l2. *Nat. Commun.* **10**, 3960 (2019).
95. Fei, L. ggjlab/MCDA: v1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.6423564 (2022).

Acknowledgements

G.G. is a participant of the Human Cell Atlas Project. We thank J. Wu, H. Wu, J. Liu, D. Wang, S. Xiong, Y. Huang and P. Xu for support on the project. We thank: G-BIO (Hangzhou), BGI (Shenzhen) and CNGB (Shenzhen) for supporting this project; Vazyme (Nanjing) for supplying the customized enzymes in the study; the Core Facilities of Zhejiang University School of Medicine for technical support; the Center of Cryo-Electron Microscopy at Zhejiang University for computational support; PTM Biolab (Hangzhou) for performing the LC–MS proteomic analysis; and Gempharmatech (Nanjing) for producing the KO mice. Funding was provided by: National Natural Science Foundation of China (grant nos. 31930028 to G.G., 31871473 to X.H., 31922049 to X.H., 91842301 to G.G., 32001068 to H.C. and 32000461 to J.W.), the Funds for Creative Research Groups of China (grant no. T2121004 to G.G.), National Key Research and Development Program (grant nos. 2018YFA0107804 to G.G., 2018YFA0107801 to X.H. and 2018YFA0800503 to G.G.); and Fundamental Research Funds for the Central Universities (to G.G.).

Author contributions

G.G. conceived the project. X.H., H.C., X.F., Z.Z., R.W. and L.F. performed the experiments. L.F., L.M., W.E., H.S., J.W., X.W., C.Y. and Y.M. performed the single-cell data processing, clustering analyses, gene-regulated analyses and cell-type annotation. G.G., L.F., H.C., L.M., W.E. and X.F. wrote the paper. M.J., D.J. and T.Z. performed the sequencing experiments. L.F., L.M. and W.E. preserved and made available the data, code and materials on publication. G.G., X.H., H.C. and J.W. acquired the funds.

Competing interests

The authors declare no competing interests.

Additional information

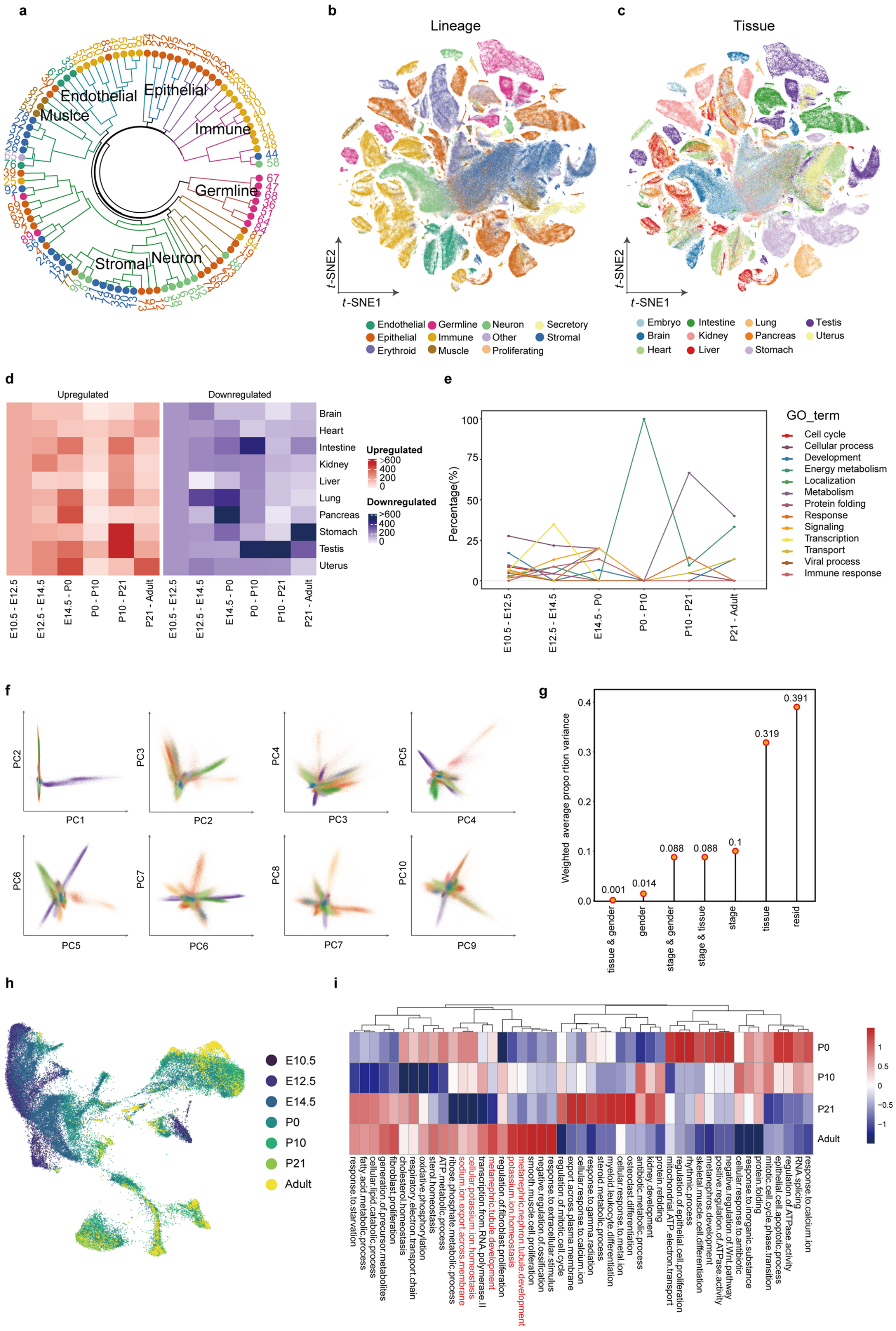
Extended data Extended data are available for this paper at https://doi.org/10.1038/s41588-022-01118-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01118-8.

Correspondence and requests for materials should be addressed to Xiaoping Han or Guoji Guo.

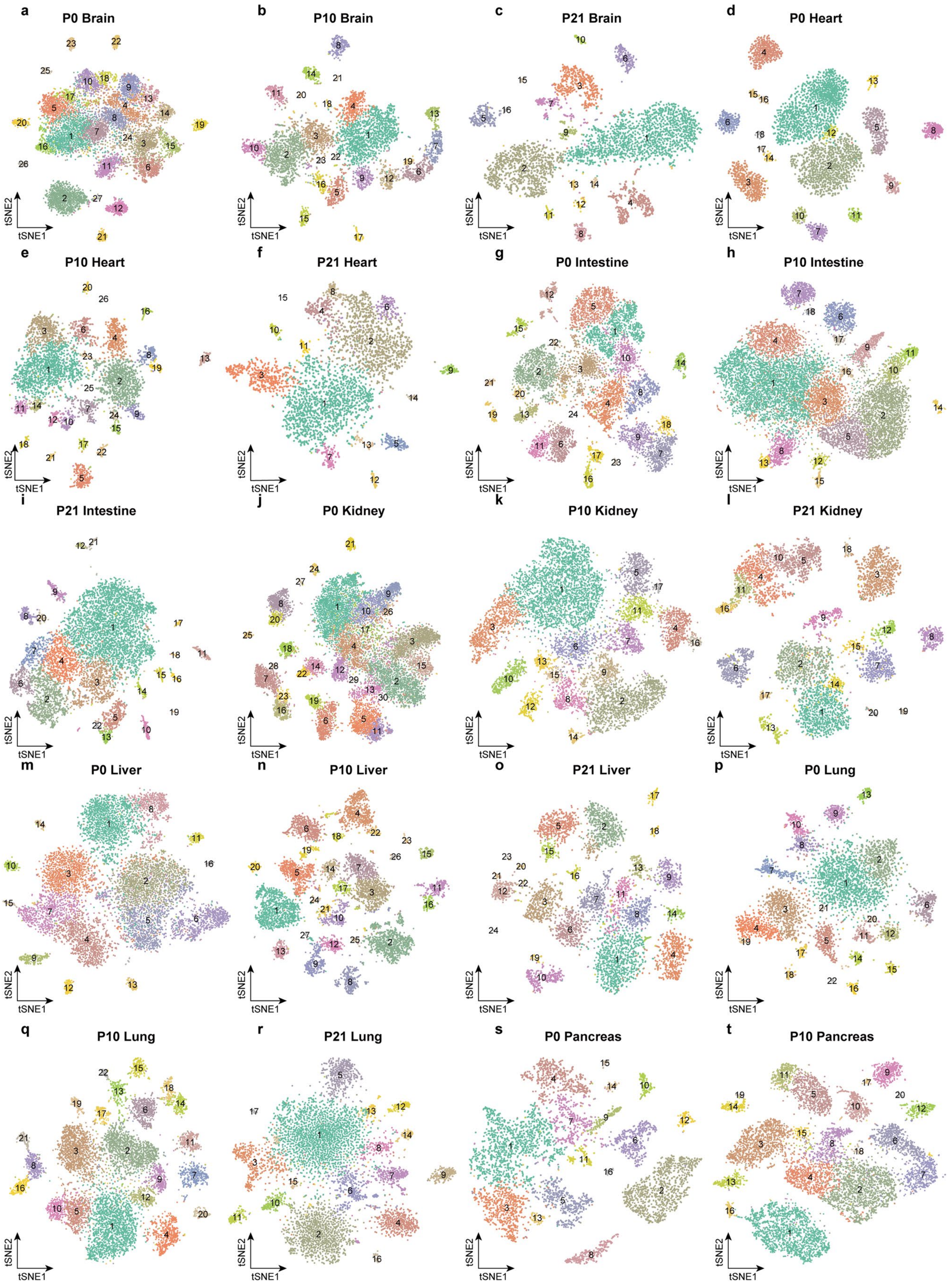
Peer review information *Nature Genetics* thanks Malte Spielmann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.



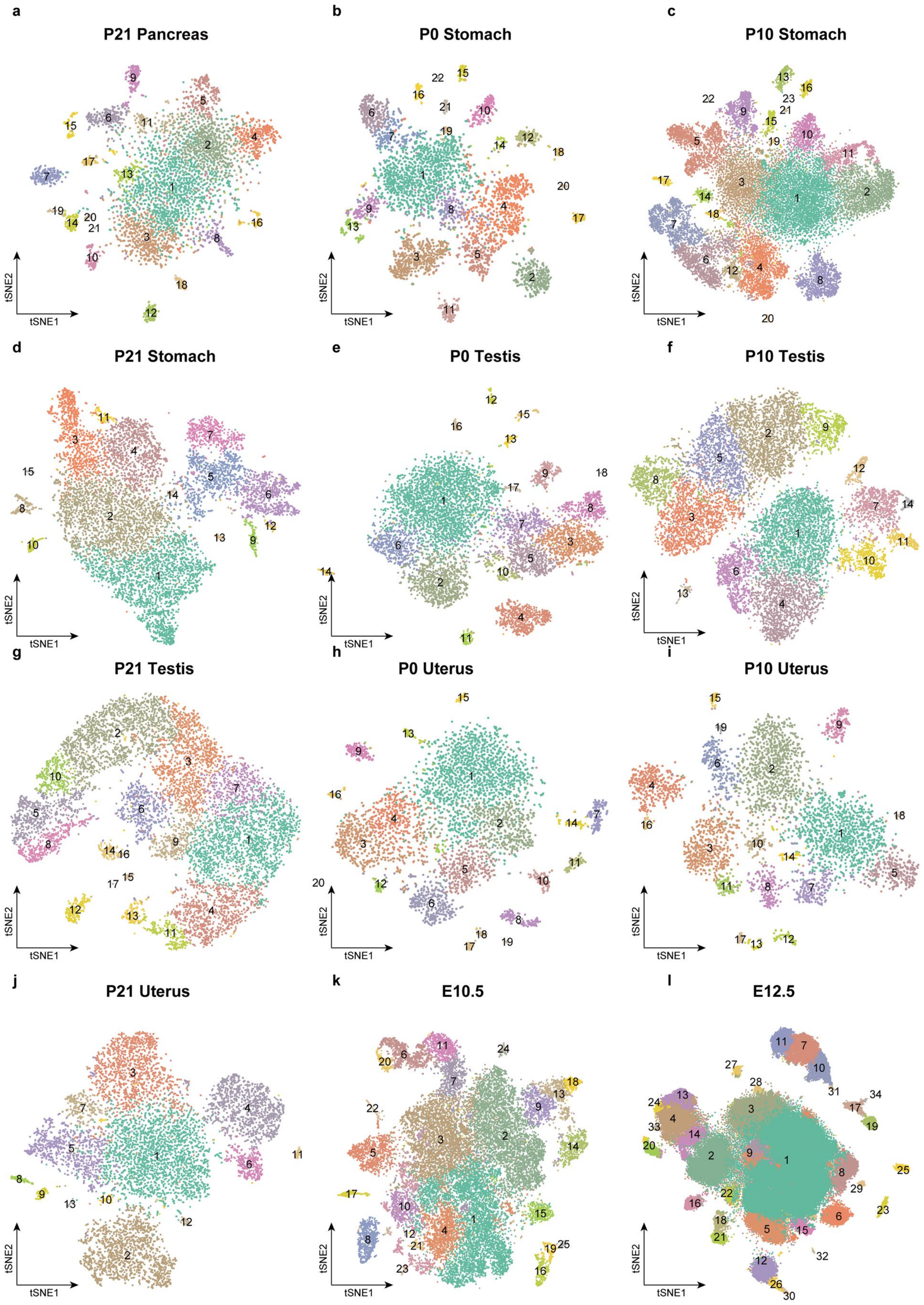
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Construction of the MCDA. a, Hierarchical trees showing the relationship between 95 cell types in MCDA, colored by lineage. **b**, t-SNE visualization of 520,801 single cells from seven developmental stages of mice, colored by lineage. They share the same color legend of lineages. **c**, t-SNE visualization of 520,801 single cells from different developmental stages of mice, colored by tissue. **d**, Heatmaps showing the number of differentially expressed genes (DEGs) in each developmental stage across the ten tissues of mice. DEGs between two stages of cells were identified using a Wilcoxon rank sum test. **e**, Summary of the GO enrichment analysis performed on the DEGs in each developmental stage. **f**, Visualization of the top 10 principal components of PCA in MCDA. Colors represent tissues, which is the same in Extended Data Fig. 1c. **g**, Lollipop chart displaying the gene expression variance explained by residuals (that is, biological and technical noise) or experimental factors such as tissue, stage, gender, and their respective combinations. Items like "tissue and gender" are variances explained by interactions of two factors instead of the union of two factors. **h**, UMAP visualization of 57,118 single cells in the kidneys at 7 different time points, colored by stage. **i**, Summary of the GO enrichment analysis performed on the DEGs in the kidneys across different stages. The red marks the go terms related to physiological functions of renal functions.



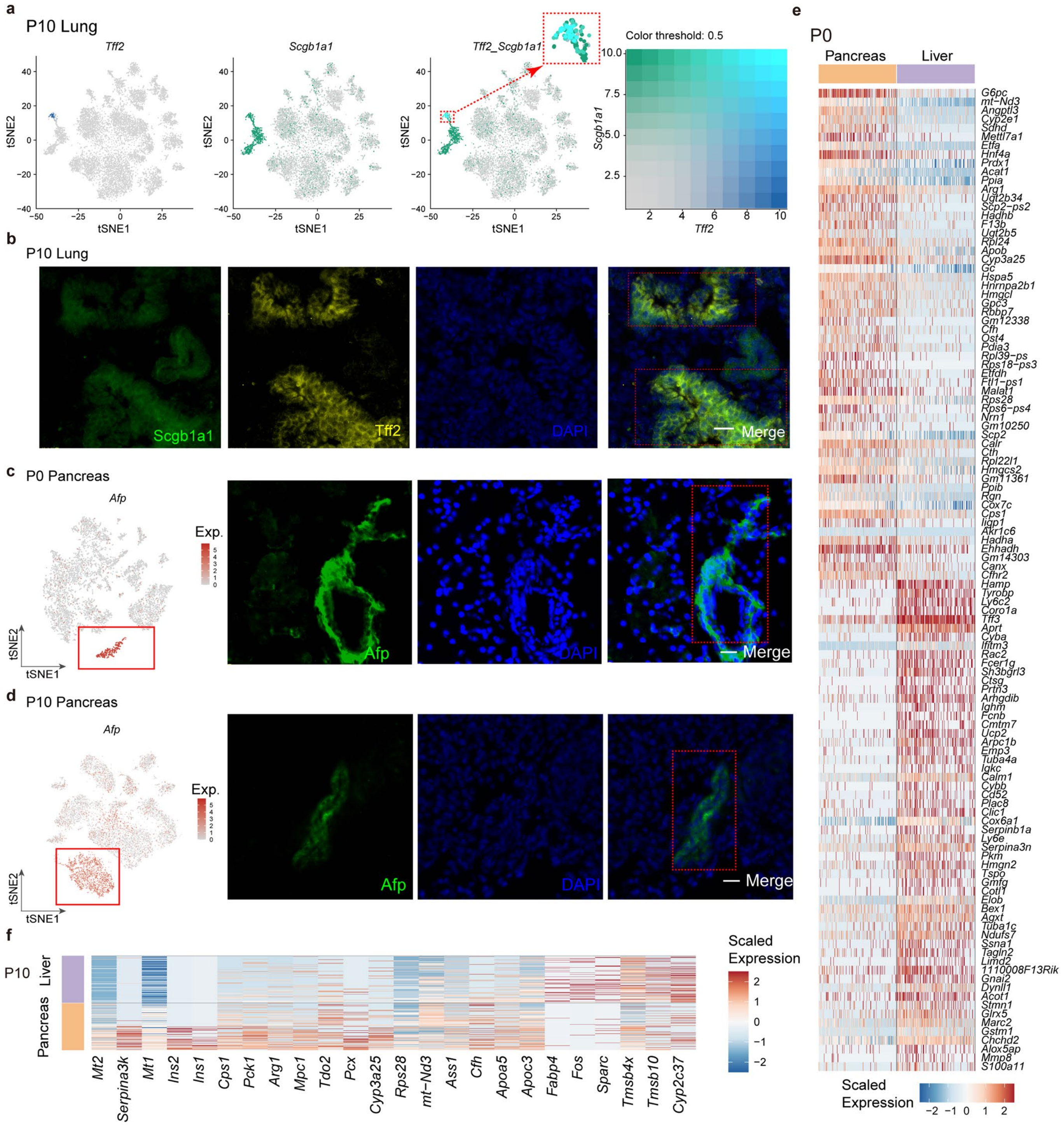
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | t-SNE maps for examples of analyzed tissues in MCDA. t-SNE maps for single-cell data from brain at P0 (**a**, n =9,265 cells), P10 (**b**, n = 6,100 cells), P21 (**c**, n = 4,433 cells) stages, heart at P0 (**d**, n = 3,948 cells), P10 (**e**, n = 5,383 cells), P21 (**f**, n = 4,054 cells) stages, intestine at P0 (**g**, n = 9,101 cells), P10 (**h**, n = 17,909 cells), P21 (**i**, n = 9,365 cells) stages, kidney at P0 (**j**, n = 13,155 cells), P10 (**k**, n = 12,129 cells), P21 (**l**, n = 5,700 cells) stages, liver at P0 (**m**, n = 9,980 cells), P10 (**n**, n = 9,259 cells), P21 (**o**, n = 5,867 cells) stages, lung at P0 (**p**, n = 5,906 cells), P10 (**q**, n = 11,314 cells), P21 (**r**, n = 6,391 cells) stages, and pancreas at P0 (**s**, n = 5,639 cells), P10 (**t**, n = 11,007 cells) stages.

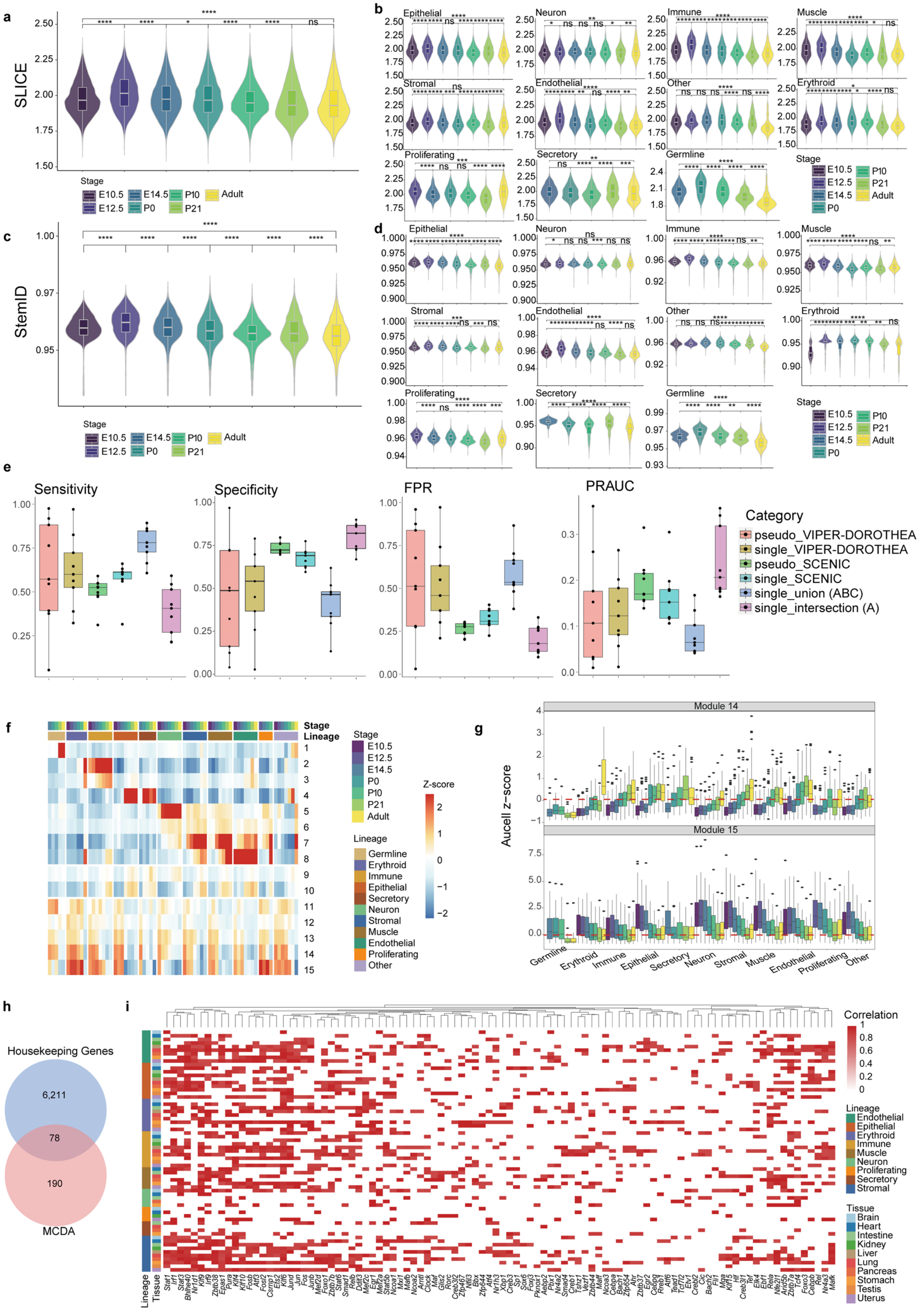


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | t-SNE maps for examples of analyzed tissues in MCDA. t-SNE maps for single-cell data from pancreas at P21 (**a**, n = 4,858 cells) stages, stomach at P0 (**b**, n = 4,073 cells), P10 (**c**, n = 22,599cells), P21(d, n = 9,945 cells) stages, testes at P0 (**e**, n = 9,034 cells), P10 (**f**, n = 15,808 cells), P21 (**g**, n = 9,095 cells) stages, uterus at P0 (**h**, n = 4,561 cells), P10 (**i**, n = 4,841 cells), P21 (**j**, n = 9,077 cells) stages, and embryo at E10.5 (**k**, n = 26,551 cells) and E12.5 (**l**, n = 72,792 cells) stages.

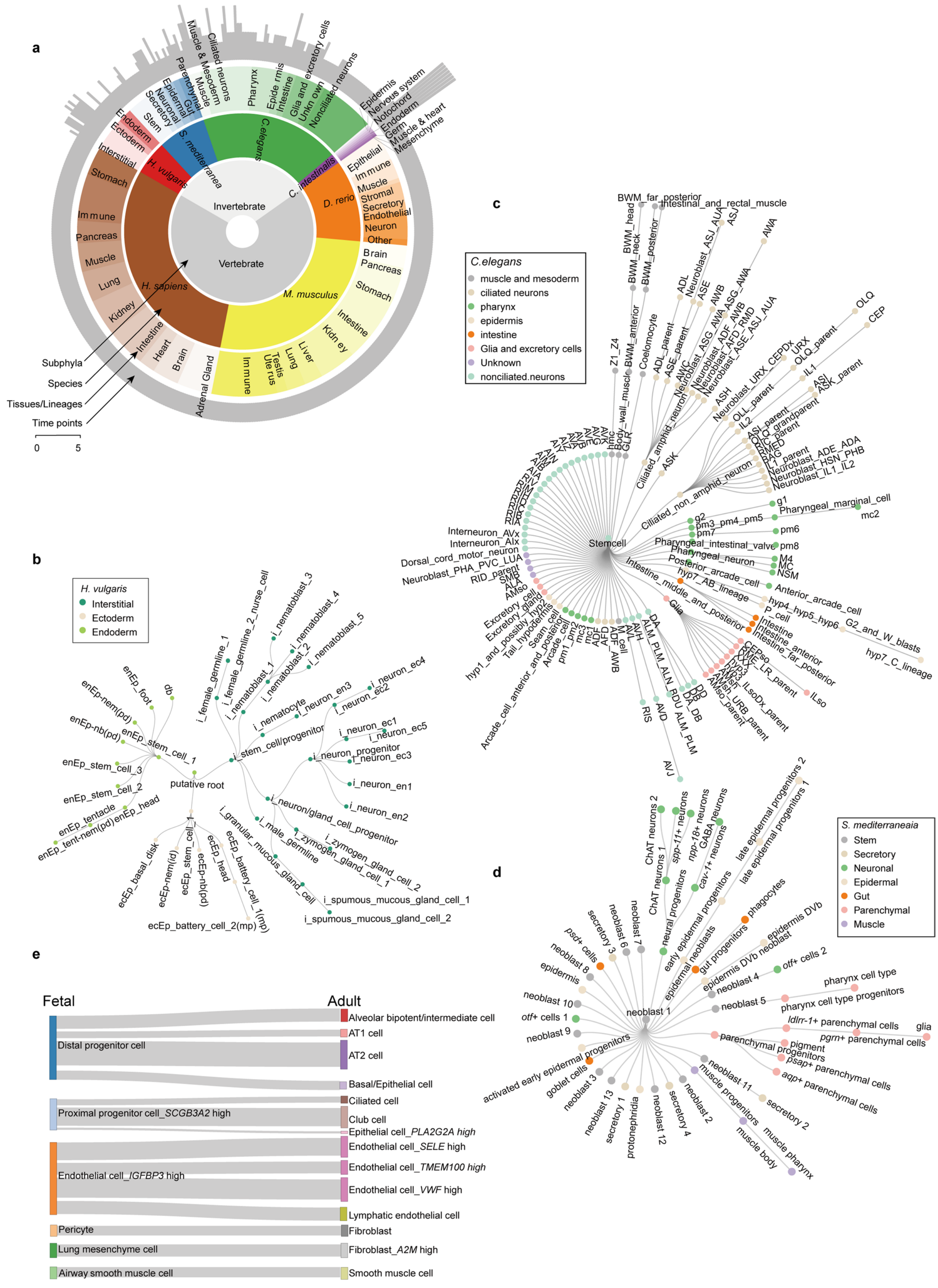


Extended Data Fig. 4 | Examples of novel cell populations. a, Feature plots in the t-SNE map of P10 lung (n = 11,314 cells). Cells are colored according to the expression of the indicated marker genes or two genes. The red boxes magnify the co-expressed cell types in the tissues. **b**, Immunofluorescence assay for the club cell marker gene Scgb1a1 (green) and goblet cell marker gene Tff2 (yellow) in P10 lung. The red boxes indicate the co-expressed locations. The experiment was replicated three times with similar results. Scale bar, 20 µm. **c**, **d**, Left: feature plots of Afp in the t-SNE map of P0 pancreas (**c**, n = 5,639 cells), P10 pancreas (**d**, n = 11,007 cells). Cells are colored according to the expression of Afp. Right: immunofluorescence assay for the hepatocyte marker gene Afp (green) in P0 (**c**) pancreas and P10 (**d**) pancreas. The experiment was replicated three times with similar results. Scale bar, 20 µm. **e**, Heatmap shows the differentially expressed genes between liver hepatocytes and pancreas hepatocyte-like cells at the P0 stage. Wilcoxon rank-sum test (two-sided) was performed to identify differentially expressed genes and p-value adjustment was performed using bonferroni correction (p adjusted values < 0.05, fold change >= 2). **f**, Heatmap shows the differentially expressed genes between liver hepatocytes and pancreas hepatocyte-like cells at the P10 stage. Wilcoxon rank-sum test (two-sided) was performed to identify differentially expressed genes and p-value adjustment was performed using bonferroni correction (p adjusted values < 0.05, fold change > 2).



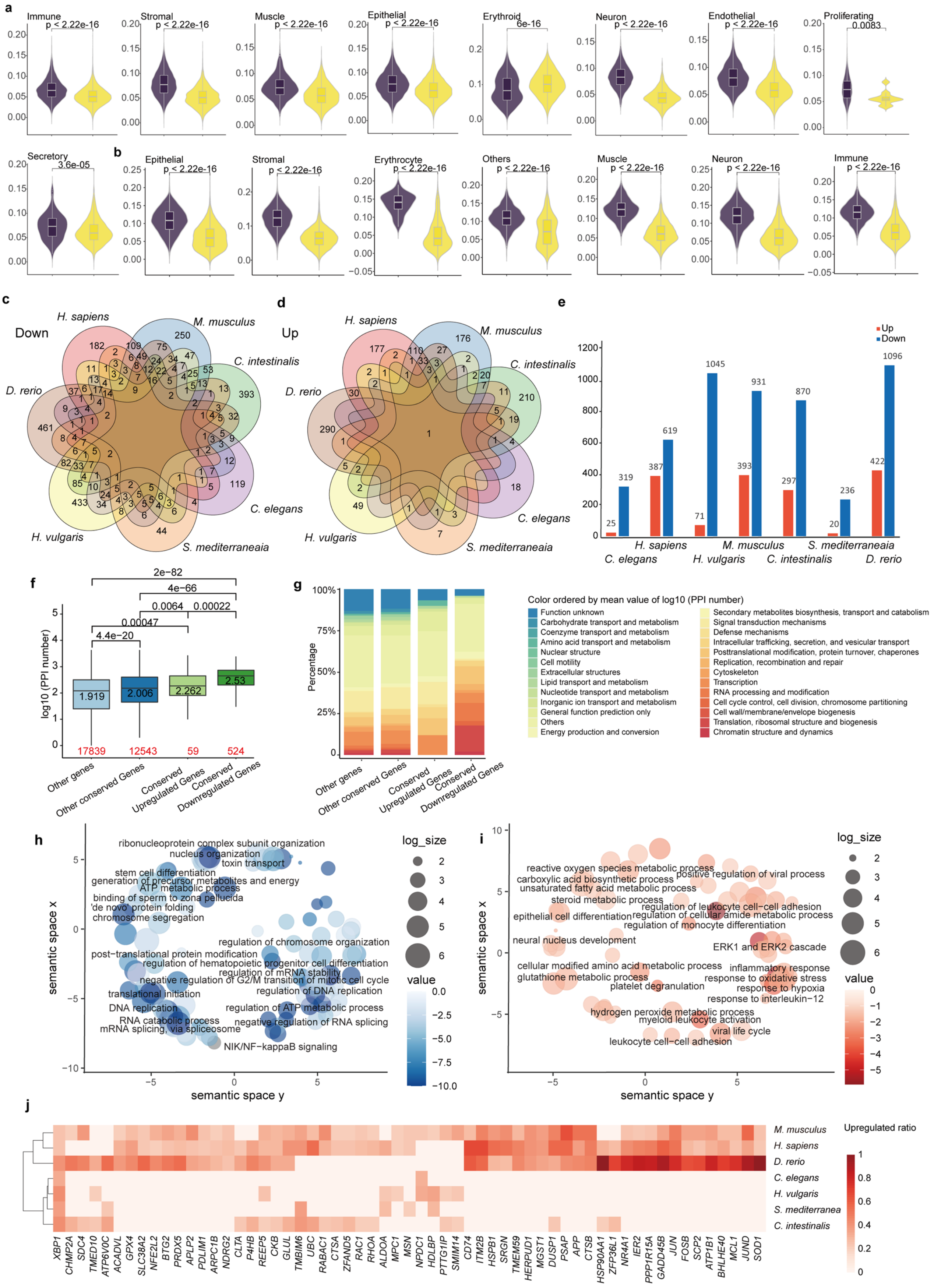
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Entropy estimations of the MCDA using. a, Entropy measurement of cells in MCDA using the SLICE method. The color represents the stage. P-values are from a two -sided Wilcoxon rank sum test comparing entropies of two different development stages (n = 60,065 cells, ns: not significant, p-value > 0.05, * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001, **** p-value ≤ 0.0001). The exact p-values were displayed in the Source Data. Box plots: center line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within 1.5 × IQR. The same statistical analysis was performed for Extended Data Fig. 5a-d. **b**, Entropy measurement of each lineage in MCDA using the SLICE method. The color represents the stage (epithelial: n = 13,642 cells, neuron: n = 3,638 cells, immune: n = 15,719 cells, muscle n = 2,592 cells, stromal: n = 8,541 cells, endothelial: n = 4,528 cells, other: n = 2,626 cells, erythroid: n = cells, proliferating: n = 3,442 cells, secretory: n = 2,892 cells, germline: n = 5,480 cells). **c**, Entropy measurement of cells in MCDA using the StemID method (n = 60,065 cells). The color represents the stage. **d**, Entropy measurement of each lineage in MCDA using the StemID method. The color represents the stage (epithelial: n = 13,642 cells, neuron: n = 3,638 cells, immune: n = 15,719 cells, muscle n = 2,592 cells, stromal: n = 8,541 cells, endothelial: n = 4,528 cells, other: n = 2,626 cells, erythroid: n = cells, proliferating: n = 3,442 cells, secretory: n = 2,892 cells, germline: n = 5,480 cells). **e**, Boxplots displaying the sensitivity, specificity, FPR (False Positive Rate), and PRAUC (Precision-Recall Area Under Curve) of two methods with different inputs to detect tissue-specific TFs in MCDA (n = 9 tissues per box). Methods represented are running VIPER-DOROTHEA with pseudo cells (pseudo_VIPER-DOROTHEA) or single cells (single_VIPER-DOROTHEA), running SCENIC with pseudo cells (pseudo_SCENIC) or single cells(single_SCENIC). The union of the two methods with single cells (single_union (ABC)) was the union of collection ABC. And the intersection of the two methods with single cells (single_intersection (A)) is the collection A. Box plots: center line, median; boxes, first and third quartiles of the distribution; point, tissues in MCDA. The results indicate SCENIC with single-cell datasets performs better in specificity and PRAUC than VIPER-DOROTHEA. The union of two methods achieves over 75% sensitivity in identifying regulatory programs while the intersection of two methods achieves the highest specificity. **f**, Heatmap of aggregated module activities of TFs clustered by fuzzy c-means showing variation by stage and lineage from VIPER-DOROTHEA. **g**, Boxplot showing the module activity scores in module 14 (n = 56 TFs) and module 15 (n = 36 TFs) per lineage per stage in SECNIC. Red lines mark the zero line. Colors from blue to yellow represent the 7 development stages from E10.5 to adult stage. Box plots: center line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within 1.5 × IQR. **h**, Venn diagrams of the numbers of overlapping genes between housekeeping TFs and commonly upregulated TFs (TFs in module 14, collection ABC) in MCDA. **i**, Heatmap showing commonly upregulated TFs (TFs in module 14, collection ABC) with regard to expression levels in MCDA. The color displays the Spearman correlation between aggregated TF expression levels in tissue-lineage against development stages (labeled as 1 to 7 to represent E14.5 to adult). Red blocks indicate the TFs display the upregulated expression patterns in the specific lineages of tissues.



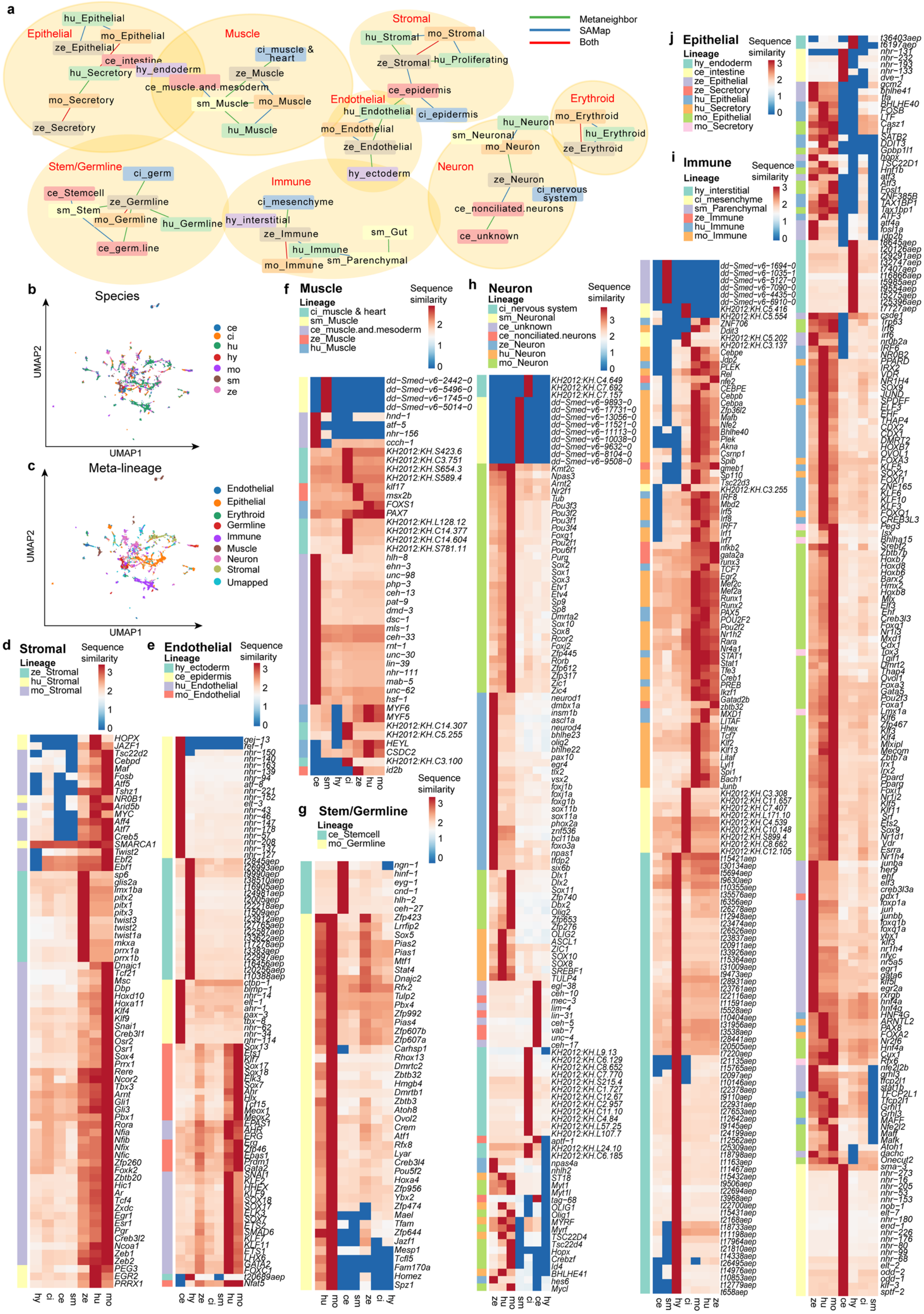
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Analysis of the developmental branch across species. **a**, Circos plot showing the subphyla, species, tissues/lineages, and time points of the single-cell dataset used in the cross-species analysis. **b**–**d**, Radial network plot showing the inferred relationships among cell types of invertebrates (**b**, *H. vulgaris* **c**, *C. elegans* **d**, *S. mediterraneaia*). Dot representing cell types, colored by lineage. **e**, Sankey plot showing the inferred relationships among cell types in fetal and adult human lungs.



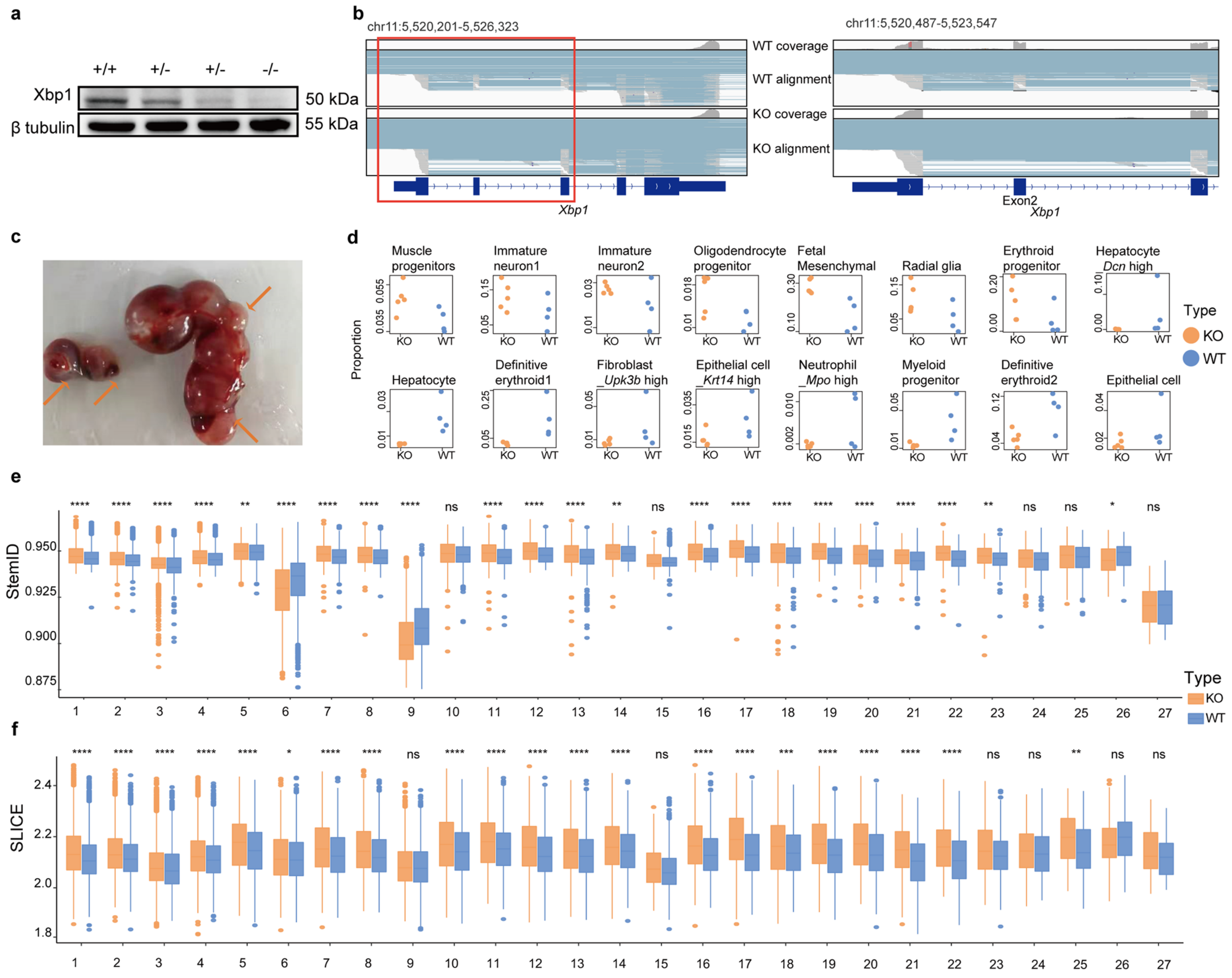
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Cross-species analysis of commonly upregulated and downregulated genes. a, **b**, Entropy measurement of each lineage in *H. sapiens* (**a**) and *D. rerio* (**b**) using the CCAT methods (*H. sapiens*: immune, n = 26,976 cells, stromal, n = 11,278 cells, muscle, n = 5,450 cells, epithelial, n = 20,347 cells, erythroid, n = 1,897 cells, neuron, n = 4,659 cells, endothelial n = 7,475 cells, proliferating, n = 3,421 cells, secretory, n = 3,708 cells; *D. rerio*: epithelial, n = 36,243 cells, stromal, n = 8,801 cells, erythroid, n = 693 cells, others, n = 3,454 cells, muscle, n = 4,140, neuron: n = 10,363 cells, immune: n = 10,104 cells). The color represents the stage. P-values were from a two-sided Wilcoxon rank sum test comparing entropies of two different development stages. Box plots: center line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within 1.5 × IQR. **c**, **d**, Venn plots showing the downregulated (**c**) and upregulated (**d**) genes in 7 species (homologous genes of humans, p-adj < 0.1). **e**, Bar plot showing the numbers of conserved upregulated and conserved downregulated genes per species, which were homologous genes of humans. **f**, Boxplots showing the number of log10 protein–protein interactions of commonly upregulated genes (at least 3 species, n = 59), commonly downregulated genes (at least 3 species, n = 524), other conserved genes (at least 3 species and homologous to human genes, n = 12,543), and other genes (n = 17,839). P-values were from a twosided Wilcoxon rank sum test comparing log10 PPI numbers of two different gene types. Box plots: center line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within 1.5 × IQR. **g**, Bar plot showing the gene composition of conserved upregulated genes (at least 3 species, n = 59 genes), conserved downregulated genes (at least 3 species, n = 524 genes), other conserved genes (in at least 3 species and homologous to human genes, n = 12,543 genes), and other genes (n = 17,839 genes). Gene categories were colored by mean values of log10 PPI number (blue: less PPIs, red: more PPIs). **h**, **i**, Bubble plot showing the GO terms of commonly downregulated (**h**) and upregulated (**i**) genes. The bubble color indicates the value representing the proportion of selected GO term in the EBI GOA database for the human. Higher value implies more general terms, lower implies more specific ones. The bubble size indicates the frequency of the GO term in the underlying GOA database. Hypergeometric test was performed to identify significant go terms and benjamini-hochberg correction was used to adjust p-values. **j**, Heatmap showing the cell type frequencies of commonly upregulated genes in 7 species.

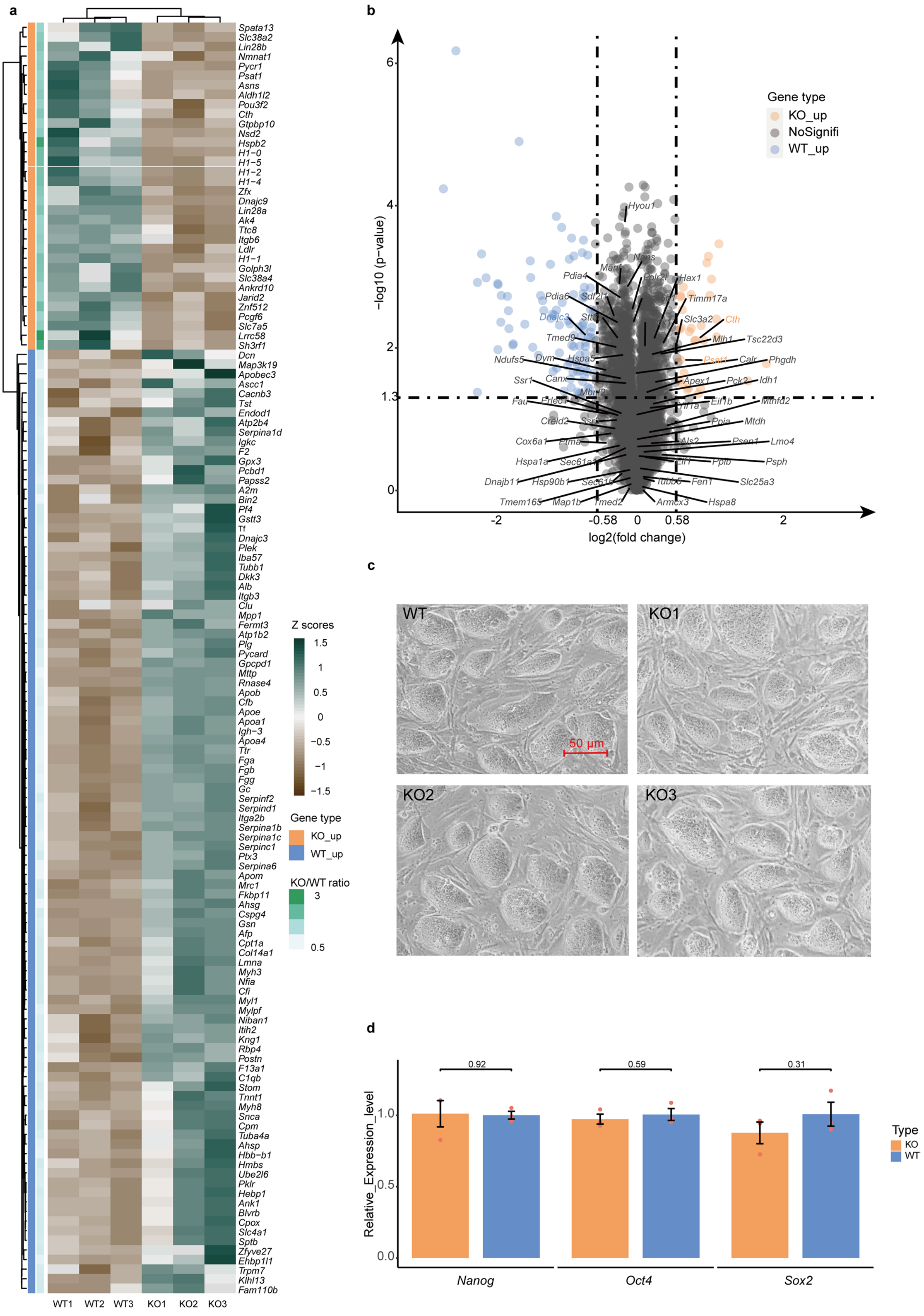


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Lineage-specific regulators among different species. a, Network plot showing the reliable and biologically plausible matches of lineages from 7 species using Metaneighbor and SAMap (sm: *S. mediterranea*, ce: *C. elegans*, hy: *H. vulgaris*, ci: *C. intestinalis*, ze: *D. rerio*, mo: *Mus M. musculus*, hu: *H. sapiens*, the abbreviations are the same in Extended Data Fig. 8). **b**, UMAP showing the combination projection of seven species based on pseudo-bulk cells, colored by species. **c**, UMAP showing the combination projection based on pseudo-bulk cells, colored by meta-lineages. **d**–**j**, Heatmaps showing the sequence similarities (log values) of development-related lineage-specific TFs within the meta-lineage across species: stromal (**d**), endothelial (**e**), muscle (**f**), stem/germline (**g**), neural (**h**), immune (**i**), and epithelial (**j**).



Extended Data Fig. 9 | scRNA-seq revealed the changes in *Xbp1*$^{-/-}$ embryos. a, Western blot for the knockout experiment. The molecular weight markers were labeled. The experiment was replicated three times with similar results. **b**, A igv view of mapped reads in the *Xbp1* gene in the sequencing data of the WT and KO embryos. The left one shows the entire *Xbp1* gene. The right one shows the marked red region which is the exon1 and exon2 region of *Xbp1*. The exon2 region shows no read coverage, which indicates that the exon2 (97 bp) has been completely disrupted in KO embryos. The blue lines link the different parts of reads that, by definition, map on several exons. The left and right genome browser tracks share the same y axis. **c**, *Xbp1*$^{-/-}$ embryos at E12.5. The arrows represent dead embryos. **d**, Scatter plot showing the cell composition proportions of differential cell types between KO and WT embryos on E12.5 (WT: n = 4, KO: n = 5, FDR < 0.01). **e**–**f**, Entropy measurement of each cluster in Fig. 6b using the StemID (**e**, n = 93,246 cells) and SLICE (**f**, n = 93,246 cells) methods. They share the same text in the x coordinates. P-values are from a two-sided Wilcoxon rank sum test comparing entropies of two different groups from each cluster (ns: not significant, p-value > 0.05, * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001, **** p-value ≤ 0.0001). The exact p values were displayed in the Source Data. Box plots: center line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within 1.5 × IQR.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | High-resolution MS revealed the protein changes in *Xbp1$^{-/-}$* embryos. a, Heatmap illustrating the proteins that were differentially expressed in *Xbp1$^{-/-}$* embryos and wild-type embryos (the colors represent the z-scores of the protein expression). A two-sided t-test is performed for comparing protein levels of KO embryos to WT embryos (p-value ≤ 0.05, fold change ≥ 1.5). **b**, Volcano plot showing the differentially expressed proteins in *Xbp1$^{-/-}$* embryos and WT embryos. The lines mark thresholds for log values of the p-value and fold change. The dots of text annotations are genes that are canonical *Xbp1* targets related to the unfolded protein response (UPR). The yellow and blue dots are genes with significantly upregulated genes in KO embryos and WT embryos respectively. **c**, mESCs and *Xbp1$^{-/-}$* mESCs grown in mESCs medium for 3 days and showing no visible differences in cell morphology. The experiment was replicated three times with similar results. Scale bar, 50 μm. **d**, qPCR analysis of *Nanog*, *Oct4*, and *Sox2* expression in mESCs and *Xbp1$^{-/-}$* mESCs showing no significant differences (normalized by the expression level of *Gapdh*, n = 3 per box). A two-sided Wilcoxon rank sum test is performed for comparing gene expression levels of wild-type and knockout mESCs (p-value ≥ 0.05: not significant, mean ± s.d.).

nature portfolio

Corresponding author(s): Guoji Guo, Xiaoping Han

Last updated by author(s): Apr 3, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Reads from single cell datasets were aligned to the Mus_musculus.GRCm38.88 using STAR 2.5.2a and the DGE data matrices were obtained using the Dropseq Core Computational Protocol (available at http://mccarrolllab.org/dropseq/) with default parameters. Downstream standard procedures for filtering, variable gene selection, dimensionality reduction and clustering were performed using the Seurat 3.2.2 in R3.6.3. Scanpy 1.6.0 was used for single cell gene expression analysis, such as clustering analysis and lineage trajectory analysis. bbknn 1.4.0 was applied to removed batch effects in the kidney tissues. pvca (v1.26.0) was employed to evaluate the variance of MCDA. pySCENIC 0.10.0 (available at https://github.com/aertslab/pySCENIC) and VIPER-DOROTHEA (viper 1.28.0 and dorothea 1.6.0) were used to infer gene regulatory networks. Cytoscape 3.5.0 was used for network visualization. Orthofinder 2.2.6 was used to infer orthologs. SLICE 0.99.0, RaceID 0.2.2 (StemID), SCENT 1.0.2 (CCAT) , and CytoTRACE 0.3.3 were used for single cell entropy analysis. VarID, as a part of RaceID 0.2.2, was used for differentially variable genes detection. ClusterProfiler 3.14.3 was used for Gene Ontology biological pathway enrichment analysis and REVIGO (http://revigo.irb.hr/, latest updated on November 16, 2021) was used for visualization. SAMap 0.3.0 and MetaNeighbor (pyMN 0.1.0) were used to infer lineage relationships across species. Differential expressed genes between cell type pairs across species were calculated using FindMarkers function in Seurat 4.0.1. Speckle 0.0.1 (propeller) was used for Cell type composition analysis. scMerge 1.2.0 was used to evaluate stably expressed gene sets in MCDA. MaxQuant search engine (v.1.5.2.8) was used for the Analysis of global proteomics data and the reference was Mus_musculus_10090 database. ChromVAR 1.12.0 was used to calculate the accessibility of the Xbp1 motif in scATAC-seq datasets. R package ggpubr 0.4.0 was used to determine the statistical significance of the differences between two groups. Detailed codes for figures are provided in github (https://github.com/ggjlab/MCDA).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data generated in this study can be downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) under accession numbers GSE176063 and GSE178217. The raw and processed files of MCDA were in GSE176063 The raw and processed files of wild-type and Xbp1 knockout embryos were in GSE178217. Processed count matrices and cell annotations were provided on the figshare website (https://figshare.com/s/340e8e7f349559f61ef6/), including the development stage, tissue-of-origin, lineage information, cell-type annotations. We provided separate datasets for each tissue and the merged datasets for the MCDA. We also provide an interactive website (http://bis.zju.edu.cn/MCA/) to enable public access to the data. The following publicly available datasets were used in the study: Mus_musculus. GRCm38.88 genome, Mus_musculus_10090 database, AnimalTFDB 3.0 database, STRING database (v11), eggNOG database (v5.0), Ensembl v96; the Schmidtea mediterranea dataset generated by Plass et al. (GSE103633), the Caenorhabditis elegan dataset generated by Packer et al. (GSE126954.); the Ciona intestinalis dataset generated by Cao et al, (GSE131155); the Hydra vulgaris dataset generated by Siebert et al. (GSE121617); the Danio rerio dataset generated by Li et al. (GSE178151); the Homo sapiens dataset generated by Han et al. (GSE134355), and part of Mus musculus dataset (E14.5 and adult) generated by Han et al. (GSE108097 and GSE134355). The mouse scATAC-seq dataset generated by Cusanovich et al (https://atlas.gs.washington.edu/mouse-atac/data/) and Di Bella et al. (GSE153164), and the human scATAC-seq dataset generated by Domcke et al. (descartes.brotmanbaty.org).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No statistical methods used to predetermine sample size. 520,801 single cells were analyzed in total for a time-series mouse cell differentiation atlas construction. A total of 52 mouse tissues from different development stages were analyzed. In our previous study (Han et al., Nature, 2020), we estimate that the major cell-type discovery in representative tissues are near plateau at around 8000 cells. Therefore we collected more than 10000 cells per tissue on average. That makes a total of 520,801 single cells. For wild-type and knockout embryos, we collected more than 40,000 cells respectively with more than 3 experimental replicates per genotype for single-cell experiment. 3 experimental replicates per genotype were used for mass spectrometry proteomic analysis and mESC related experiments.

Data exclusions

Data points with fewer than 500 UMI were excluded. The detected transcript from a single live mammalian cell under our sequencing depth (3000 reads/cell) should be more than 500 UMI, as we have exemplified in our previous Mouse Cell Atlas paper (Han et al., Cell, 2018). Cell barcodes with less than 500 UMI usually correspond to empty beads exposed to free RNA during cell lysis, RNA capture and washing steps.

Replication

2-4 replications were done for different tissues when samples were available. The results of major cell type clusters are reproducible.

Randomization

Experimental mice and embryos were randomized before sample preparation. Different single cells were randomly captured before analysis.

Blinding

For all experiments, investigators were blinded to group allocation during the data collection and analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
□	☒ Antibodies
□	☒ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
□	☒ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Antibodies

Antibodies used

anti-ESAM (MA5-24072; Thermo);
https://www.thermofisher.cn/cn/zh/antibody/product/ESAM-Antibody-clone-340236-Monoclonal/MA5-24072
anti-Myl9 (ab187152; Abcam);
https://www.abcam.com/myl9-antibody-epr130132b-ab187152.html
anti-Scgb1a1 (MAB4218-SP; R&D);
https://www.rndsystems.com/cn/products/human-uteroglobin-scgb1a1-antibody-394324_mab4218
anti-tff2 (13681-1-AP; ProteinTech);
https://www.ptglab.com/products/TFF2-Antibody-13681-1-AP.htm
anti-AFP (AF5134; Affinity));
http://www.affbiotech.com/goods-4441-AF5134-AFP_Antibody.html
anti-Xbp1 (ab37152; Abcam);
https://www.abcam.com/xbp1-antibody-ab37152.html
anti-β-tubulin (EM0103; HUABIO);
https://www.huabio.com/products/beta-tubulin-antibody-clone-1-b11-monoclonal-em0103
anti-mouse IgG (HS201-01; TransGen Biotech);
https://www.transgenbiotech.com/secondary_antibody/proteinfind_goat_anti_mouse_igg_h_l_hrp_conjugate.html
anti-rabbit IgG (GAR007; MultiSciences);
http://www.liankebio.com/product-736524.html
Donkey anti-Rat IgG (H+L) Highly Cross-Adsorbed, Alexa Fluor 488 (A-21208; Thermo) ;
https://www.thermofisher.cn/cn/zh/antibody/product/Donkey-anti-Rat-IgG-H-L-Highly-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-21208
Goat anti-Rabbit IgG (H+L) Highly Cross-Adsorbed, Alexa Fluor 594 (A-11037; Thermo) ;
https://www.thermofisher.cn/cn/zh/antibody/product/Goat-anti-Rabbit-IgG-H-L-Highly-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-11037
Donkey anti-Rabbit IgG (H+L) Highly Cross-Adsorbed, Alexa Fluor 488 (A-21206; Thermo);
https://www.thermofisher.cn/cn/zh/antibody/product/Donkey-anti-Rabbit-IgG-H-L-Highly-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-21206

Validation

Validation are available for all antibodies from the manufacturer. Please refer to references contained in the provided links.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	mESC is from George Q. Daley Lab. RRID: CVCL_C320.
Authentication	The mESC cell line is authenticated by morphology, karyotyping, and immunostaining with Sox2/Nanog/Oct4.
Mycoplasma contamination	The cell line is negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

Wild-type C57BL/6J mice were ordered from Shanghai SLAC Laboratory Animal. Xbp1 knockout mice were generated by Nanjing Gempharmatech. Seven development stages of mice (E10.5, E12.5, E14.5, P0, P10, P21 and 6-10weeks) were used for single-cell experiments. Testes were collected from male mice and all the other tissues were collected from female mice. For E10.5 and E12.5 samples, both female and male embryos were used. Wild-type Danio rerio strain AB was raised and maintained in standard zebrafish units at Core Facilities, Zhejiang University School of Medicine. Two development stages (24hpf and 3 month) were used for single-cell experiments. Both female and male strains were contained.

Wild animals

The study did not involve wild animals.

Field-collected samples	All mice were housed at Zhejiang University Laboratory Animal Center in a Specific Pathogen Free (SPF) facility with individually ventilated cages. The room has controlled temperature (20-22°C), humidity (30%–70%) and light (12 hour light-dark cycle). Mice were provided ad libitum access to a regular rodent chow diet. No no field-collected samples were used in the study.
Ethics oversight	All experiments performed in this study were approved by the Animal Ethics Committee of Zhejiang University. All experiments conformed to the relevant regulatory standards at Zhejiang University Laboratory Animal Center.

Note that full information on the approval of the study protocol must also be provided in the manuscript.