

# Construction of a human cell landscape at single-cell level

<https://doi.org/10.1038/s41586-020-2157-4>

Received: 23 October 2018

Accepted: 12 March 2020

Published online: 25 March 2020

 Check for updates

Xiaoping Han<sup>1,2,19</sup>✉, Ziming Zhou<sup>1,19</sup>, Lijiang Fei<sup>1,19</sup>, Huiyu Sun<sup>1,19</sup>, Renying Wang<sup>1,19</sup>, Yao Chen<sup>3,19</sup>, Haide Chen<sup>1,4,19</sup>, Jingjing Wang<sup>1,4,19</sup>, Huanna Tang<sup>5</sup>, Wenhao Ge<sup>6</sup>, Yincong Zhou<sup>7</sup>, Fang Ye<sup>1</sup>, Mengmeng Jiang<sup>1</sup>, Junqing Wu<sup>1</sup>, Yanyu Xiao<sup>1</sup>, Xiaoning Jia<sup>8</sup>, Tingyue Zhang<sup>1</sup>, Xiaojie Ma<sup>9</sup>, Qi Zhang<sup>10</sup>, Xueli Bai<sup>10</sup>, Shujing Lai<sup>1</sup>, Chengxuan Yu<sup>1</sup>, Lijun Zhu<sup>6</sup>, Rui Lin<sup>11</sup>, Yuchi Gao<sup>12</sup>, Min Wang<sup>13</sup>, Yiqing Wu<sup>3</sup>, Jianming Zhang<sup>14</sup>, Renya Zhan<sup>15</sup>, Saiyong Zhu<sup>9</sup>, Hailan Hu<sup>8</sup>, Changchun Wang<sup>16</sup>, Ming Chen<sup>7</sup>, He Huang<sup>2,17,18</sup>, Tingbo Liang<sup>10</sup>, Jianghua Chen<sup>5</sup>, Weilin Wang<sup>6</sup>, Dan Zhang<sup>3</sup> & Guoji Guo<sup>1,2,4,17,18</sup>✉

Single-cell analysis is a valuable tool for dissecting cellular heterogeneity in complex systems<sup>1</sup>. However, a comprehensive single-cell atlas has not been achieved for humans. Here we use single-cell mRNA sequencing to determine the cell-type composition of all major human organs and construct a scheme for the human cell landscape (HCL). We have uncovered a single-cell hierarchy for many tissues that have not been well characterized. We established a ‘single-cell HCL analysis’ pipeline that helps to define human cell identity. Finally, we performed a single-cell comparative analysis of landscapes from human and mouse to identify conserved genetic networks. We found that stem and progenitor cells exhibit strong transcriptomic stochasticity, whereas differentiated cells are more distinct. Our results provide a useful resource for the study of human biology.

Individual cells are fundamental units of life. Breakthroughs in single-cell mRNA sequencing have greatly enhanced our ability to identify the transcriptomes of individual types of cell<sup>2–5</sup>. Using high-throughput barcoding strategies, it is now possible to profile thousands of single cells at the same time<sup>6,7</sup>. These methods have allowed the mapping of cell atlases for whole organisms<sup>8–15</sup>. For example, cell atlases for the mammalian system have been generated by analysing both fetal and adult mouse tissues<sup>16–19</sup>. Despite extensive efforts in dissecting the cellular compositions of various human tissues<sup>20–32</sup>, to our knowledge a comprehensive cell landscape for humans has not been achieved.

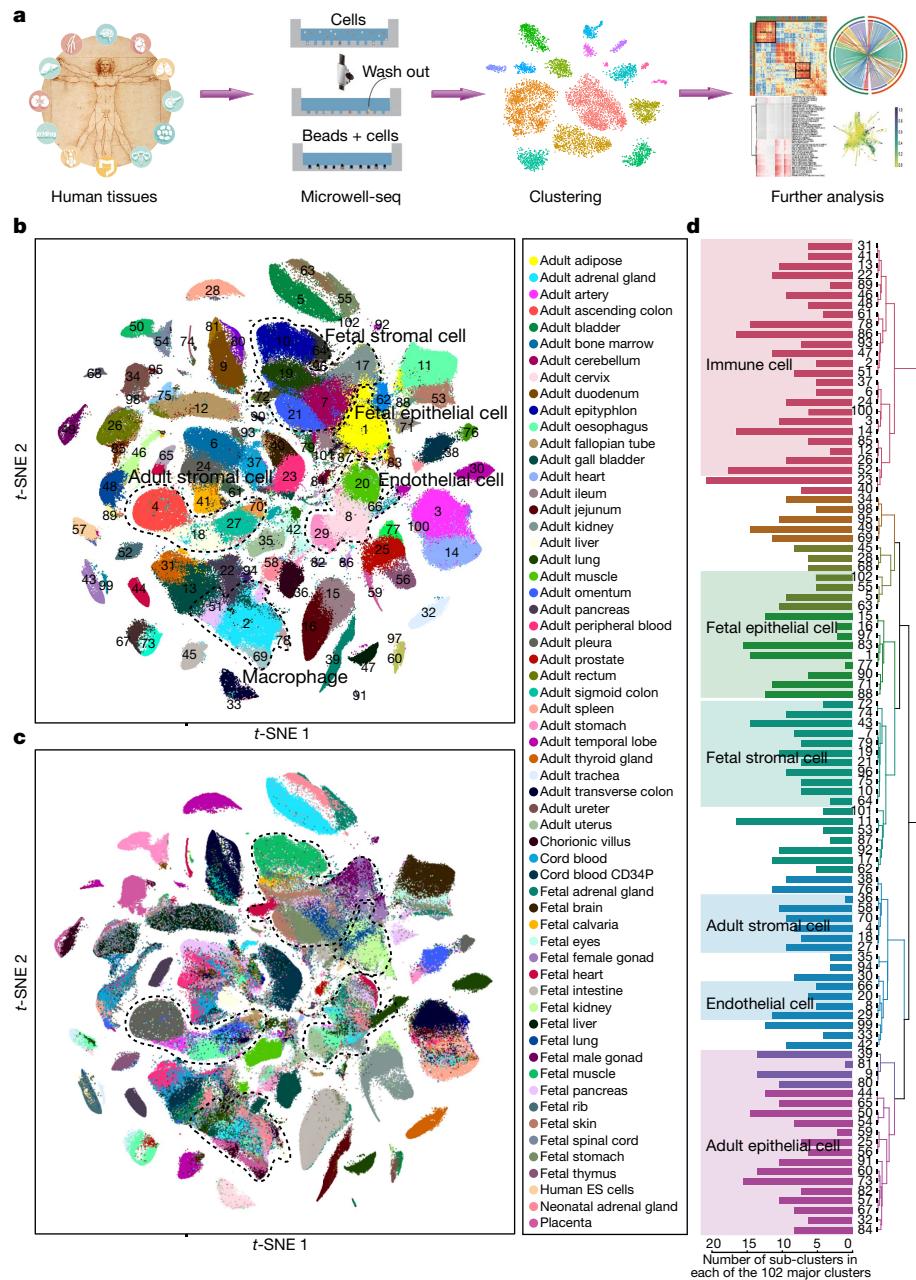
## Constructing an HCL using microwell-seq

Microwell-seq is a cost-effective single-cell mRNA sequencing technology that offers advantages over existing technologies in doublet rate and cell-type compatibility<sup>16</sup>. Sequencing titration experiments and cross-platform comparison suggest that this method can robustly detect rare populations even at low sequencing depth (Supplementary Table 1, Extended Data Fig. 1a). Using microwell-seq, we embarked upon the creation of a basic landscape of major human cell types using

samples from a Chinese Han population. Donated tissues were perfused or washed and prepared as single-cell suspensions using specific protocols (Supplementary Table 1). Our analyses included samples of both fetal and adult tissue and covered 60 human tissue types (two to four replicates per tissue type in general; Extended Data Fig. 1b). We also analysed seven types of cell culture, including induced pluripotent stem (iPS) cells, embryoid body cells, haematopoietic cells derived from co-cultures of human H9 and mouse OP9 cells<sup>33</sup>, and pancreatic beta cells derived from H9 cells using a seven-stage protocol<sup>34</sup> (Extended Data Fig. 1b). Single cells were processed using microwell-seq<sup>16</sup> and sequenced at around 3,000 reads per cell; data were then processed using published pipelines<sup>35</sup> (Fig. 1a). Altogether, 702,968 single cells passed our quality control tests (Supplementary Table 1).

In a global view, the complete human tissue data set is grouped into 102 major clusters (Fig. 1b; Supplementary Table 2). Multiple tissues, including artery, trachea, pleura, omentum, oesophagus and fallopian tube, contributed to the defined adult stromal/mesenchymal cells, such as cluster 4 (C4), C18, C27, and C70. Other clusters with substantial multi-tissue contributions correspond to fetal stromal cells (C7, C10, C17, C19, C21, C64, and C72), endothelial cells (C8, C20, C29, and C66), macrophages (C2, C51, C69, and C78), and fetal epithelial cells (C1)

<sup>1</sup>Center for Stem Cell and Regenerative Medicine, Zhejiang University School of Medicine, Hangzhou, China. <sup>2</sup>Bone Marrow Transplantation Center, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. <sup>3</sup>Department of Reproductive Endocrinology, Women’s Hospital, Zhejiang University School of Medicine, Hangzhou, China. <sup>4</sup>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. <sup>5</sup>Kidney Disease Center, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. <sup>6</sup>Department of Hepatobiliary and Pancreatic Surgery, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. <sup>7</sup>College of Life Sciences, Zhejiang University, Hangzhou, China. <sup>8</sup>Center for Neuroscience, Zhejiang University School of Medicine, Hangzhou, China. <sup>9</sup>Life Sciences Institute, Zhejiang University, Hangzhou, China. <sup>10</sup>Department of Hepatobiliary and Pancreatic Surgery, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. <sup>11</sup>Hangzhou Repugene Technology, Hangzhou, China. <sup>12</sup>Annoroad Gene Technology, Beijing, China. <sup>13</sup>Veritas Genetics Asia, Hangzhou, China. <sup>14</sup>Department of Neurosurgery, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. <sup>15</sup>Department of Neurosurgery, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. <sup>16</sup>Institute of Cancer Research and Basic Medical Sciences of Chinese Academy of Sciences, Cancer Hospital of University of Chinese Academy of Sciences, Zhejiang Cancer Hospital, Hangzhou, China. <sup>17</sup>Institute of Hematology, Zhejiang University, Hangzhou, China. <sup>18</sup>Stem Cell Institute, Zhejiang University, Hangzhou, China. <sup>19</sup>These authors contributed equally: Xiaoping Han, Ziming Zhou, Lijiang Fei, Huiyu Sun, Renying Wang, Yao Chen, Haide Chen, Jingjing Wang. ✉e-mail: xhan@zju.edu.cn; ggj@zju.edu.cn



**Fig. 1 | Constructing an HCL using microwell-seq.** **a**, Illustration of the experimental workflow using the microwell-seq platform. **b**, t-SNE analysis of 599,926 single cells from the HCL. Differentiated cell culture data and granulocyte-colony stimulating factor (G-CSF)-mobilized peripheral blood data were not included. In the t-SNE map, 102 cell-type clusters are labelled by different colours. Cell cluster markers are listed in Supplementary Table 2. **c**, t-SNE analysis of 599,926 single cells from the HCL. Differentiated cell culture

data and G-CSF mobilised PB data were not included. In the t-SNE map, tissues are labelled by different colours. Tissue contributions to each cluster are listed in Supplementary Table 2. **d**, Dendrogram showing relationships among 102 cell types. The bar chart on the left represents the number of sub-clusters in each major cluster. A total of 843 sub-clusters were predicted from 102 major clusters.

(Fig. 1b, c; Supplementary Table 2). We then performed sub-clustering analysis for each of the 102 major clusters and predicted a total of 843 cell-type sub-clusters in the hierarchy (Fig. 1d). Through correlation analysis between bulk and single-cell mRNA sequencing as well as cell number sub-sampling analysis, we estimated a high gene and cell-type coverage of HCL (Extended Data Fig. 1c, d). Multi-donor analysis of representative tissues indicates that there are limited donor or batch effects on cell-type discovery (Extended Data Fig. 1e). To our knowledge, these data represent the most comprehensive cell-type repertoire yet described for the human species. By applying the concept of

a pseudo-cell<sup>36</sup>, we can aggregate data from the same cluster to increase gene representation and improve cluster separation (Extended Data Fig. 2a, b). This strategy enabled us to interpret transcription factor (TF) function and generate a correlation network that covers 91% of all human TFs (Extended Data Fig. 2c; Supplementary Table 2). A highlight of the network suggests that, in the HCL, master TFs work in discrete modules to specify major human cell types such as neuron, erythroid cell, and acinar cell (Extended Data Fig. 2d). The resource is publicly available at <http://bis.zju.edu.cn/HCL/> (with a mirror website for international users at <https://db.cngb.org/HCL/>).

## Cellular heterogeneity in human tissues

We performed *t*-distributed stochastic neighbour embedding (*t*-SNE) and differential gene expression analyses for each specific organ and uncovered previously unrecognized cell heterogeneity in a wide range of human tissues (Supplementary Table 3, Extended Data Figs. 3, 4). After analysing human kidney tissues at the fetal and adult stages, we defined 21 fetal clusters (FCs) and 22 adult clusters (ACs) with specific molecular markers (Extended Data Fig. 5a, b, Supplementary Table 3), including epithelial, endothelial, stromal and tissue-resident immune cells. We found previously undescribed types of S-shaped body cell (FC1, FC9 and FC15) in the fetal kidney and a new intercalated cell (IC)-tran-principal cell (PC) cell type (AC19) in the adult kidney. In the lung, we identified 22 fetal clusters and 25 adult clusters (Extended Data Fig. 5c, d, Supplementary Table 3). We found distal progenitors (FC4) and proximal progenitors (FC9) in the fetal lung, as well as alveolar type 2 (AT2) cells (AC1), alveolar type 1 (AT1) cells (AC3), two subsets of club cells (AC16 and AC20), and ciliated cells (AC22) in the adult lung. The alveolar bipotent/intermediate cells (AC14) express *KRT8*, cell cyclin genes, and lineage markers from both AT1 and AT2. We constructed ligand–receptor maps using CellPhoneDB<sup>37</sup> to reveal cell–cell interactions in kidney and lung tissues (Extended Data Fig. 5e–h). Stromal and endothelial cells were at the centre of the network for both organ types. In fetal tissues, our data predict that stromal and endothelial cells interact with epithelial cell progenitors to support tissue development. However, in adults, they are predicted to interact with immune cells such as T cells and macrophages.

During our comprehensive analysis, we noted the presence of populations with interesting expression signatures, particularly in tissues that have not previously been well characterized (summarized in Supplementary Table 4). In adult pleura, we identified three clusters of mesothelial cells (C2, C13 and C14), as well as an unknown cell cluster (C12) that expressed high levels of interferon-induced proteins (Extended Data Fig. 6a, b). In adult omentum, we identified a group of immune-related mesothelial cells that expressed *CCL2* (Extended Data Fig. 6c). In fetal muscle, we noted two groups of tendon cells expressing glucagon (*GCG*; Extended Data Fig. 6d). Other previously undescribed populations include *MCP<sup>+</sup>* progenitor cells in fetal brain, stomach cells that co-expressed a D and X/A cell signature, and *CCL2<sup>+</sup>* enteric nerve cells in the colon as well as 18 cell clusters that could not be annotated on the basis of known markers (Supplementary Table 4).

Notably, we identified endothelial cell populations that expressed major histocompatibility complex (MHC) class II genes such as *HLA-DPA1* and *HLA-DRA* in a variety of adult tissues, including bladder (Fig. 2a) and kidney. Immunofluorescence assays for the endothelial cell marker *PECAM1* (also known as *CD31*) and the MHC class II marker *HLA-DR* further confirmed the existence of antigen-presenting endothelial cells in the adult human bladder (Fig. 2b). This professional antigen-presenting signature indicates that these endothelial cells might have an immune function. To estimate the global involvement of non-immune cells in regional immunity in humans, we performed a cross-stage, cross-tissue comparison of endothelial, epithelial and stromal cells for their expression of immune-related genes (Fig. 2c). We identified MHC class II<sup>+</sup> endothelial, interleukin-expressing stromal cells and CXCL<sup>+</sup> epithelial cells in about half of human adult tissues (Extended Data Fig. 6e, f). This widespread immune activation of non-immune cells represents a new layer of cellular regulation for tissue-specific immunity.

To characterize the global cellular hierarchy of endothelial cells, we integrated endothelial cell data across diverse tissues and identified 14 major clusters (Fig. 2d, e, Supplementary Table 4). In the tSNE map, C2, C3, and C9 contained contributions from multiple fetal tissues, including the heart, skin and kidney. Whereas C8, C11, C12 and C13 endothelial cells are tissue-specific, C6 and C10 are shared by different adult organs. Notably, there is a distinct group of endothelial cells,

including C1, C5 and C7, that show high expression of immune-related genes. C1 endothelial cells can be found in a wide range of adult organs such as bladder, kidney, artery, thyroid and omentum; they express *HLA-DRA*, *CD74* and *HLA-DRB1*. C5 endothelial cells come from adult uterus and express *CXCL8*, *IL6* and *HMOX1* at high levels. C7 endothelial cells come from adult kidney and express high levels of *HLA-DQA1*, *HLA-DPA1* and *EMCN*. In a similar way, we performed cross-tissue analysis for stromal cells and identified four immune-active clusters (C5, C8, C12 and C15) in the HCL stromal cell hierarchy (Extended Data Fig. 7a, b). Through donor contribution analysis, we showed that these stromal and endothelial subtypes contained cells from multiple donors (Extended Data Fig. 7c).

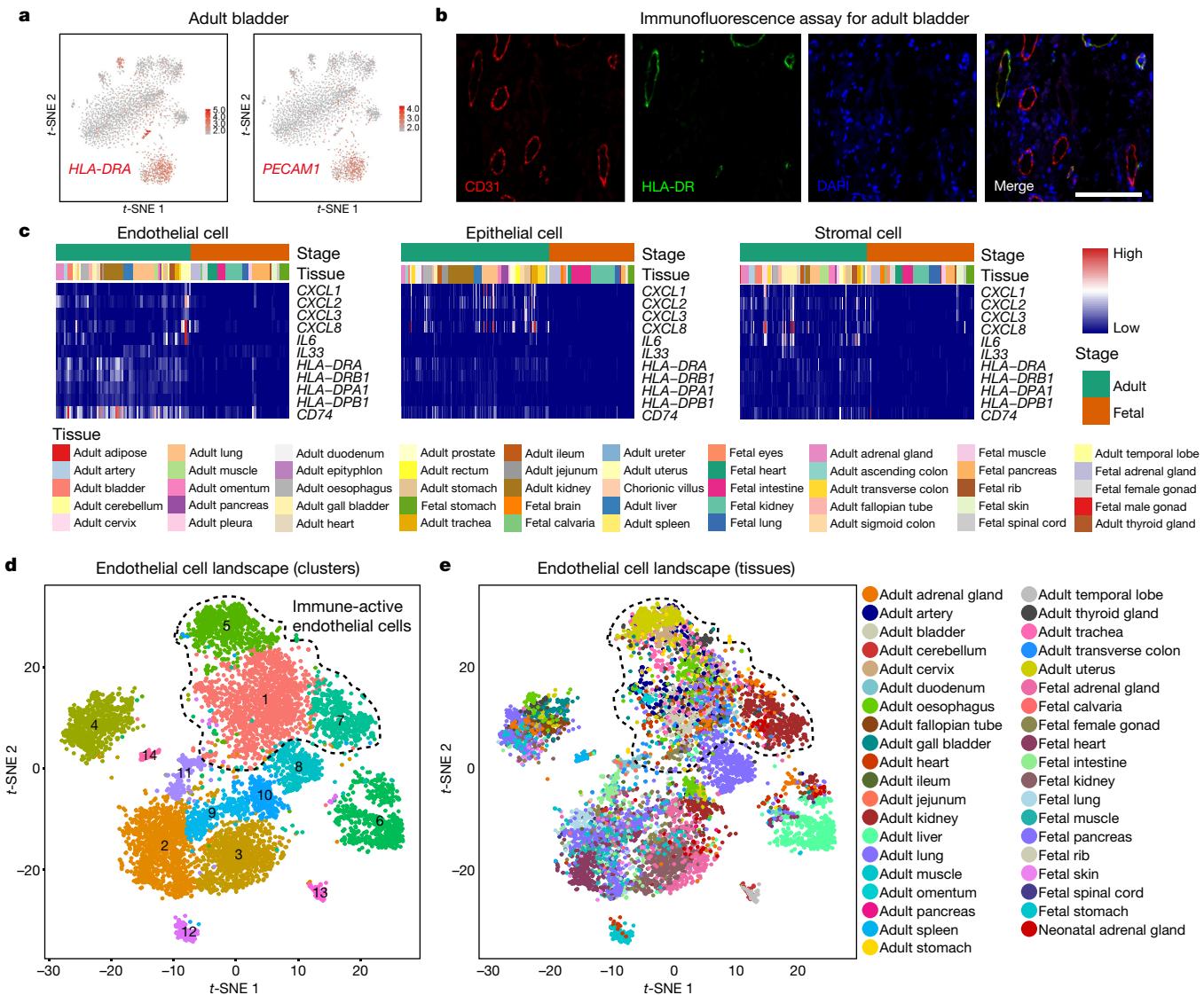
## Fetal-to-adult cell-type transitions

We next studied human organ development by assessing the similarity of cell types between fetal and adult tissues. In the kidney and lung, the gene expression patterns of epithelial, mesenchymal, endothelial and immune cells were correlated between the two stages; tissue-resident immune and stromal cells appear early during organogenesis (Extended Data Fig. 8a, b). To visualize the process of global multi-lineage specification, we performed trajectory analysis for fetal and adult HCL data using the partition-based graph abstraction (PAGA) method<sup>38</sup>. We obtained a landscape showing projections from fetal progenitors towards adult mature cell types (Extended Data Fig. 8c, d). On the landscape, we defined 36 cell groups arranged into more than 10 lineage branches. Notably, fetal cells reside at the centre of the landscape; they are less separated than adult cells, which spread out from the middle cloudy region towards different destinations. This is consistent with the observation in Fig. 1b that stromal and epithelial cells from various fetal tissues form an interconnected cell group (C1, C7, C17, C19 and C21). However, they become highly separated at the adult stage. Moreover, single cells exhibit higher entropy in the fetal stage than in the adult stage, suggesting that fetal cells possess higher transcriptional plasticity, whereas adult cells possess more stable transcriptomes (Extended Data Fig. 8e).

To understand the genetic regulation behind this fetal-to-adult transition, we analysed changes in gene expression between all specific pairs of fetal and adult cell types and ranked the most commonly enriched genes for the fetal and adult stages, respectively (Supplementary Table 4). Markers that were commonly enriched in fetal cells included a large group of ribosomal protein genes (such as *RPS17*, *RPS18*, and *RPS24*), imprinted genes (such as *IGF2* and *H19*), and stem cell regulators (such as *SOX4* and *SON*). These signatures suggest a high metabolic rate, low epigenetic restriction and the presence of multi-potent machinery, which may help to explain the high entropy and transcriptional stochasticity found in fetal cells. By contrast, markers that were commonly enriched in adult cells include MHC class I cell surface receptors (such as *HLA-A* and *HLA-B*), MHC class II cell surface receptors (such as *HLA-DPA1* and *HLA-DPB1*), and immune-related cytokines (such as *IFI27*, *CXCL2* and *CCL5*).

## Mapping human cell types using scHCL

Using the HCL database, we build a single-cell mapping pipeline (scHCL) for the classification of human cell types. We integrated our HCL with other published human data sets (Supplementary Table 5) and made transcriptome references for all available human cell-type clusters from single-cell studies. Input digital gene expression (DGE) data were compared to each transcriptome reference to provide a match score based on gene expression correlation (Fig. 3a). By mapping bulk RNA sequencing data to our HCL reference, we can robustly define cell lineages of cultured cell populations<sup>39</sup> or cancer cell organoids<sup>40</sup> (Extended Data Fig. 9a, b). We then processed single-cell data from liver bud organoid cells<sup>41</sup> and cerebral organoid cells<sup>42</sup>. While confirming the existence of



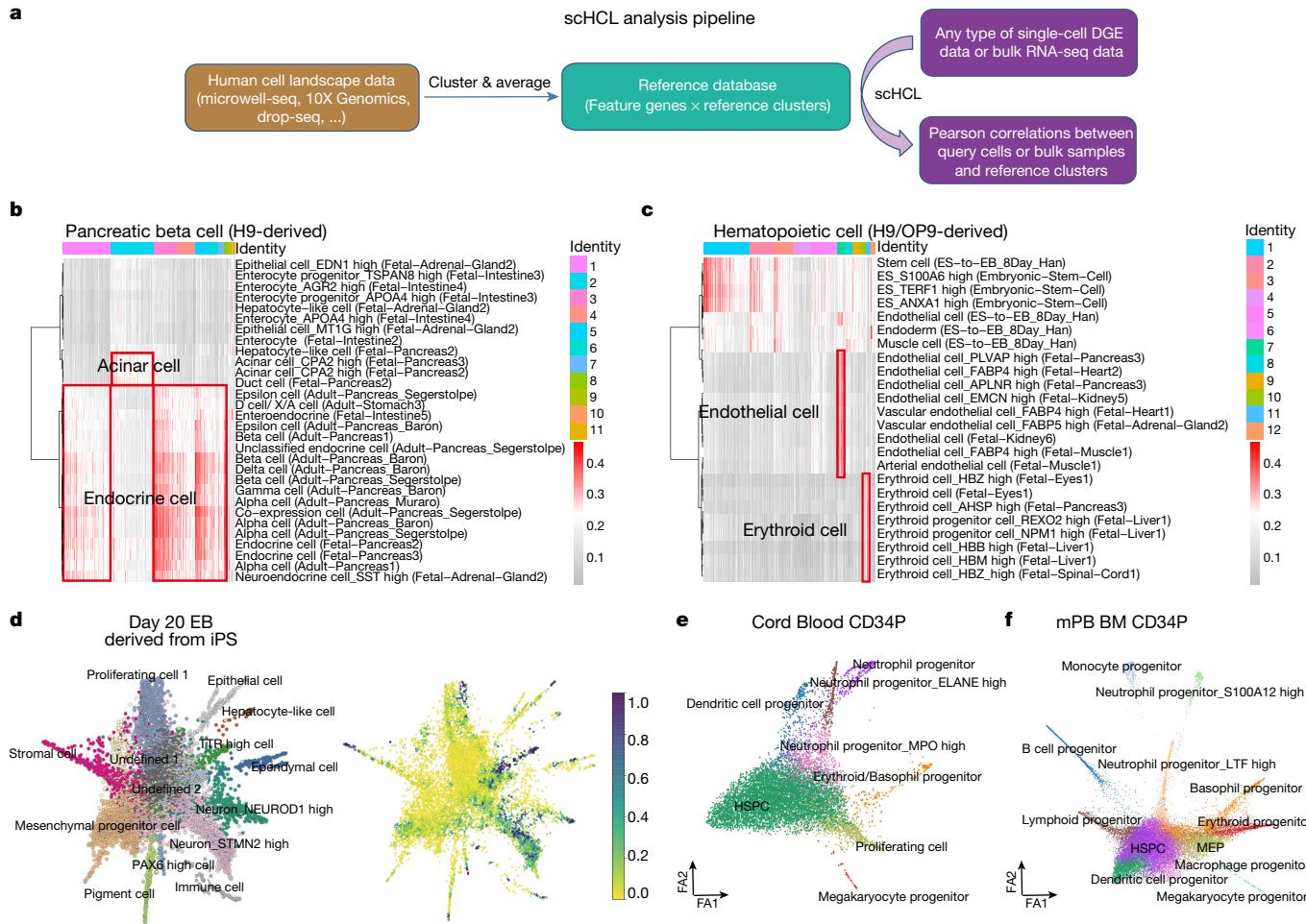
**Fig. 2 | Immune activation of non-immune cells in the HCL.** **a**, Feature plot in the t-SNE map of single-cell data from adult bladder cluster 2 ( $n=2,750$  cells). Cells are coloured based on the expression of *HLA-DRA* or *PECAM1*. The experiment was replicated twice with similar results. **b**, Immunofluorescence assay for the endothelial cell marker *PECAM1* (CD31) and the MHC class II marker *HLA-DR* in human adult bladder tissue. Scale bar, 100  $\mu$ m. The

experiment was replicated three times with similar results. **c**, Heat maps illustrating the expression of immune-related genes in endothelial, epithelial and stromal cells. x-axis represents clusters; y-axis represents genes. **d**, **e**, t-SNE maps of single-cell data for human tissue-specific endothelial cells (cell number  $n=7,140$ ). Cells are coloured according to endothelial cell subtype (**d**) or tissue type (**e**).

different parenchymal and mesenchymal cells (Extended Data Fig. 9c, 9d), we found that these organoids lacked important cell types, such as tissue-resident immune cells and endothelial cells.

We performed an scHCL analysis to evaluate a well-established seven-stage protocol for deriving pancreatic beta cells from H9 cells<sup>34</sup>. Many cells present in the differentiated culture on day 24 exhibited a strong correlation with pancreatic cell types (Fig. 3b). Cells of C3, C5 and C8 highly expressed *GCG*, *INS* and *SST*, respectively, representing a tendency towards differentiation into alpha, beta and delta cells. However, a large proportion of these cells co-expressed islet signatures that correspond to various islet cell types, suggesting a remaining problem related to cell-type maturation in the differentiation protocol. We then used scHCL to analyse haematopoietic cells derived from H9/OP9 co-cultures<sup>33</sup>. Only cells from C7 and C11, which contained endothelial cells and erythroid cells, had high scHCL scores for mature human cell types (Fig. 3c). The few C7 and C11 cells show low conversion efficiency in the H9/OP9 co-culture system. We did not detect any cell that was correlated with human haematopoietic stem and progenitor cells.

We next examined single-cell data for day 20 embryoid bodies differentiated from human iPS cells. The single cells of the embryoid body were divided into 15 clusters (Extended Data Fig. 9e, Supplementary Table 3). The scHCL results for the embryoid body cells showed clusters mapped to mesenchymal, neuronal, ependymal and immune cells (Extended Data Fig. 9f). A large group of cells in the middle of the t-SNE map, including C2, C3, C4 and C9, could not be associated with a specific lineage in the scHCL analysis. Moreover, the cell–cell correlation network of embryoid body cells showed complex and unspecialized phenotypes for these populations (Extended Data Fig. 9g). For example, C3 (Undefined-2) co-expressed the endothelial cell-related gene *EDNRB* and the neuron-related gene *MPZ*. Trajectory and RNA velocity analysis suggested that the undefined cell clusters were at the root of the differentiation hierarchy, with high RNA turnover rate, whereas the defined cell types are at the endpoint of the trajectory and have relatively stable transcriptomes (Fig. 3d). After differentiation for 20 days, many embryoid body cells appear to be still in a primarily undetermined state, whereas some cells have gone through these high turnover states and



**Fig. 3 | Application of scHCL analysis for stem cell biology.** **a**, Diagram showing the pipeline for scHCL analysis. **b, c**, scHCL results for pancreatic islet cells (**b**;  $n=4,156$  cells) derived from seven-stage differentiation of H9 cells (day 24) and haematopoietic cells (**c**;  $n=1,115$  cells) derived from H9/OP9 cell co-cultures (day 9). Each row represents one cell type in our reference. Each column represents data from a single cell. Pearson correlation coefficient was used to evaluate cell-type gene expression similarity. Red indicates a high correlation; grey indicates a low correlation. Some cell-type data come from published works, as denoted by first author name: Segerstolpe<sup>27</sup>, Muraro<sup>28</sup>, Baron<sup>29</sup>, Han<sup>37</sup>. **d**, Branching gene expression trajectory analysis of day 20

embryoid bodies differentiated from human iPS cells. Left, trajectory of different lineages coloured by cluster identities; right, model of the transition probabilities derived from RNA velocity using a Markov process. The colour scale represents the density of the end points of the Markov process, from yellow (low) to blue (high). **e**, Branching gene expression trajectory analysis of cells in cord blood (CD34 $^{+}$ ) using PAGA, coloured by cell lineages. **f**, Branching gene expression trajectory analysis of cells in mPB (CD34 $^{+}$ ) using PAGA, coloured by cell lineages. Force atlas 1 (FA1) and Force atlas 2 are used to present a continuous graph layout.

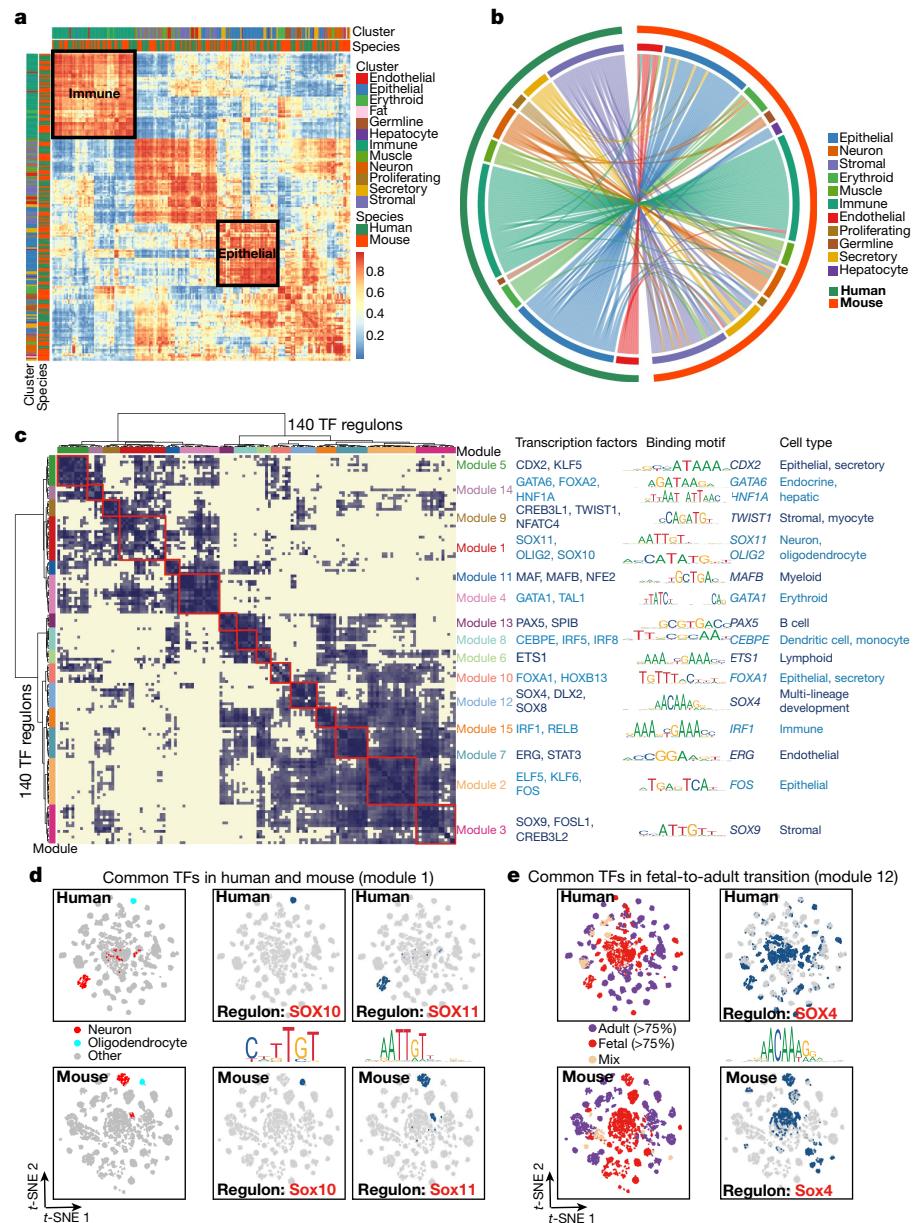
exhibit homeostasis. This *in vitro* stem cell differentiation hierarchy mimics the *in vivo* fetal-to-adult cell-type landscape (Extended Data Fig. 8c, d). We also found similar patterns for the human haematopoietic differentiation system using single-cell data from CD34 $^{+}$  cord blood and CD34 $^{+}$  mobilized peripheral blood (mPB; Fig. 3e, f).

## Comparison of mammalian cell landscapes

Single-cell transcriptomics offers the opportunity to compare cell types across species. To compare the HCL and the mouse cell landscape, we re-clustered the updated data sets from the Mouse Cell Atlas (MCA), removed batch gene background, and performed analysis using the same pipeline and parameters as for the HCL (Extended Data Fig. 10a, b). To attenuate the effects of noise and outliers, we calculated pseudo-cells for each cell cluster to proceed with network construction (Extended Data Fig. 10c). Orthologous genes were extracted from the data to enable cross-species analysis. A correlation heatmap among 102 human and 104 mouse cell types showed that cell-type similarity in orthologous gene expression overrides species differences,

particularly for immune and epithelial cells (Fig. 4a). As revealed by the circos plot, the gene expression patterns of major mammalian cell types are conserved; about 95% of cell types have a strong correlation (area under the receiver operating characteristics (AUROC) score  $>0.9$ ) between species (Fig. 4b). Tissue-specific cell-type pair analyses also suggested that the major cell types in mammalian organs are similar (Extended Data Fig. 10d, e). Our finding that mammalian cell types are conserved is consistent with the results of single-cell comparative genomics studies<sup>29,43,44</sup>.

We next investigated the genetic network that underlies mammalian cell-type conservation. We extracted TF regulons from both HCL and MCA data using SCENIC<sup>45</sup> and identified 140 orthologous TF regulons grouped into 15 major modules across the mammalian cell landscapes (Fig. 4c, Supplementary Table 5). For example, M1, M6/8/11/13/15, M7 and M9 are associated with neuron, immune, endothelial and stromal cell types, respectively. These modules are enriched with not only lineage-specific transcription factors but also conserved binding motifs that lead to co-ordinated module activation in the lineage establishment (Fig. 4c).



**Fig. 4 | Cross-species comparison of cell landscapes.** **a**, Correlation of orthologous gene expression between human and mouse cell types. AUROC scores were used to measure the similarity of cell types: red, high correlation; blue and yellow, low correlation, based on the Spearman correlation between all human and mouse pseudo-cells ( $n = 46,793$  pseudo-cells). Cluster and species information is marked by different colours. **b**, Circos plot showing the similarity of cell types in human and mouse. Paired cell types with average AUROC scores greater than 0.9 are connected by lines. **c**, Identification of regulon modules based on the regulon matrix of the HCL and the MCA. The

To understand the specific genetic regulation of human and mouse cell types, we performed regulon activity analysis with all TFs (Extended Data Fig. 11a, b). We present a list of species-specific TF regulons enriched with basic helix-loop-helix, Cys<sub>2</sub>-His<sub>2</sub> zinc finger and homeodomain proteins (Supplementary Table 5, Extended Data Fig. 11c–e). Nevertheless, most lineage-specific regulons are conserved (Supplementary Table 5, Extended Data Fig. 11f). For example, the SOX10 regulon for oligodendrocytes and SOX11 regulon for neurons are shared between human and mouse (Fig. 4d). Notably, the development-related SOX4 regulon dominates the unseparated fetal cell clusters; it shows a broad and stochastic distribution for both human and mouse (Fig. 4e).

network shows 140 orthologous TF regulons grouped into 15 major modules, along with representative TF regulons, corresponding binding motifs and associated cell types. **d**, **e**, Binary regulon activity scores (RASs) for regulons SOX10, SOX11 and SOX4 in the human and mouse regulon activity t-SNE map; dark blue dots represent one and grey dots represent zero. Other colours are used to mark specific cell types. **d**, Representative regulons in module 1. **e**, representative regulons in module 12. The regulon activity t-SNE maps were based on the binary regulon activity matrix of 17,028 human pseudo-cells and 16,740 mouse pseudo-cells.

## Discussion

We have used microwell-seq to perform single-cell transcriptomic analysis for a wide range of human tissues. The method proved to be compatible with nearly all cell types; microwell-seq data generated from different systems showed good comparability. The strength of

this landscape study is its broad coverage of both fetal and adult human tissue types. However, the scale of the current analysis is limited in sequencing depth and cell number for each individual tissue.

From the HCL, we identify MHC class II<sup>+</sup> endothelial cells, CXCL<sup>+</sup> epithelial cells, and interleukin-expressing stromal cells as major cell-type categories that lie between classical immune cells and non-immune cells. By comparing different developmental systems in the HCL, we propose a landscape model that is conserved in mammals: stem and progenitor cells are transcriptionally indistinct and stochastic; differentiated cells are transcriptionally distinct and stable; and the wiring of the regulons in the genome predetermines the terminal steady cellular states. By integrating different human single-cell data sets, we provide the HCL website and the associated scHCL pipeline for human cell-type identification. This pilot study will make contributions to the eventual completion of the international Human Cell Atlas.

## 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/s41586-020-2157-4>.

1. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
2. Tang, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
3. Ramsköld, D. et al. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
4. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
5. Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
6. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
7. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
8. Plass, M. et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaoq1723 (2018).
9. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
10. Sebe-Pedros, A. et al. Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-seq. *Cell* **173**, 1520–1534.e1520 (2018).
11. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
12. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* **360**, eaaoq1736 (2018).
13. Farrell, J. A. et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).
14. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
15. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
16. Han, X. et al. Mapping the mouse cell atlas by microwell-seq. *Cell* **172**, 1091–1107.e1017 (2018).
17. Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
18. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
19. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
20. Zhong, S. et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).
21. Young, M. D. et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361**, 594–599 (2018).
22. Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
23. Gao, S. et al. Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat. Cell Biol.* **20**, 721–734 (2018).
24. Velten, L. et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017).
25. Li, L. et al. Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* **20**, 858–873.e854 (2017).
26. Guo, J. et al. Chromatin and single-cell RNA-seq profiling reveal dynamic signaling and metabolic transitions during human spermatogonial stem cell development. *Cell Stem Cell* **21**, 533–546.e536 (2017).
27. Segerstolpe, Å. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
28. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**, 385–394.e383 (2016).
29. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360.e344 (2016).
30. Blakeley, P. et al. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3151–3165 (2015).
31. Yan, L. et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
32. Aizarni, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).
33. Vodyanik, M. A., Bork, J. A., Thomson, J. A. & Slukvin, I. I. Human embryonic stem cell-derived CD34+ cells: efficient production in the coculture with OP9 stromal cells and analysis of lymphohematopoietic potential. *Blood* **105**, 617–626 (2005).
34. Rezania, A. et al. Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. *Nat. Biotechnol.* **32**, 1121–1133 (2014).
35. 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).
36. Tosches, M. A. et al. Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**, 881–888 (2018).
37. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).
38. 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).
39. Chu, L. F. et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* **17**, 173 (2016).
40. Roerink, S. F. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457–462 (2018).
41. Camp, J. G. et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533–538 (2017).
42. Camp, J. G. et al. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl. Acad. Sci. USA* **112**, 15672–15677 (2015).
43. La Manno, G. et al. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167**, 566–580.e519 (2016).
44. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
45. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

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

© The Author(s), under exclusive licence to Springer Nature Limited 2020

# Article

## Methods

### Ethics statement

The collection of human samples and research conducted in this study were approved by the Research Ethics Committee of the Zhejiang University School of Medicine, the Research Ethics Committee of the First Affiliated Hospital, the Research Ethics Committee of the Second Affiliated Hospital and the Research Ethics Committee of Women's Hospital at Zhejiang University (approval numbers: 20170029, 20180017, 20190034, 2018015, 2018507, 2018766 and 2018185). Informed consent for fetal tissue collection and research was obtained from each patient after her decision to legally terminate her pregnancy but before the abortive procedure was performed. Informed consent for collection of and research using surgically removed adult tissues was obtained from each patient before the operation. Informed consent for the collection and research of tissues from deceased organ donors was obtained from the donor family after the cardiac death of the donor. Details on donor information are provided in Supplementary Table 1. All the protocols used in this study were in strict compliance with the legal and ethical regulations of Zhejiang University School of Medicine and Affiliated Hospitals. All the protocols used in this study complied with the 'Interim Measures for the Administration of Human Genetic Resources' administered by The Ministry of Science and Technology and The Ministry of Public Health (approval number: 2020BAT0007). Mouse experiments in this study were approved by the Animal Ethics Committee of Zhejiang University; experiments conformed to the regulatory standards at Zhejiang University Laboratory Animal Center.

### Fabrication of the microwell device

The diameter and depth of the microwells were 28 and 35 µm, respectively. First, a silicon plate containing 100,000 microwells was manufactured by Suzhou Research Materials Microtech Co., Ltd (Suzhou). The silicon microwell plate was then used as a mould to make a poly-dimethylsiloxane (PDMS) plate with the same number of micropillars. Before the experiments, a disposable agarose microwell plate was made by pouring 5% agarose solution onto the surface of the PDMS plate. Both the silicon and the PDMS plates are reusable. One silicon microwell plate allows almost permanent use.

### Synthesis of barcoded beads

Magnetic beads (20–25 mm in diameter) coated with carboxyl groups were provided by Suzhou Knowledge & Benefit Sphere Tech. Co., Ltd (Suzhou; <http://www.kbspheretech.com/>). The barcoded oligonucleotides on the surface of the beads were synthesized by three rounds of split-pool. All the sequences used are the same as those reported previously<sup>10</sup>.

For each batch of bead synthesis, 300–350 µl carboxyl magnetic beads (50 mg/ml) were washed twice with 0.1 M 2-(*N*-morpholino) ethanesulfonic acid (MES). The beads were then suspended in a final volume of 635 µl of 0.1 M MES. 1-Ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC; 3.08 mg) was added to the beads, and 6.2 µl beads was then placed in each well of a 96-well plate. Amino-modified oligonucleotides (2.5 µl, 50 µM in 0.1 M MES) were then added to each well. After vortexing the mixture and incubating it for 20 min at ambient temperature, we distributed a 0.5-µl mix (6 mg EDC in 100 µl of 0.1 M MES) into each well. After an additional round of vortexing and incubation for 20 min at ambient temperature, an additional 0.5 µl mix (6 mg EDC in 100 µl of 0.1 M MES) was distributed into each well. After vortexing and incubation for 80 min at ambient temperature, the beads were collected in 1 ml of 0.1 M phosphate-buffered saline (PBS) containing 0.02% Tween 20. After centrifugation, the supernatant was carefully removed. The beads were then washed twice in 1 ml TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0).

In the second split-pool, the beads were washed with water and divided among the wells of another 96-well plate containing

polymerase chain reaction (PCR) mix (1 × Phanta Master Mix, Vazyme) and 5 µM oligonucleotides. The oligonucleotides in each tube encoded a sequence with reverse complementarity to linker 1, a unique barcode and a linker 2 sequence. The PCR program was as follows: 94 °C for 5 min; five cycles of 94 °C for 15 s, 48.8 °C for 4 min, and 72 °C for 4 min; and a 4 °C hold. The third split-pool procedure was the same as the second one. The PCR program was as follows: 94 °C for 5 min, 48.8 °C for 20 min and 72 °C for 4 min, and a 4 °C hold. The beads were mixed sufficiently between denaturation (95 °C) and primer annealing (48.8 °C) in every cycle. The oligonucleotides used in each tube encoded a linker 2 reverse-complementary sequence, a unique barcode, a UMI sequence and a poly-T tail. All oligonucleotides were synthesized by Sangon Biotech Co., Ltd using high-performance liquid chromatography purification. To remove the chains without the third barcoded sequence, the beads were collected and suspended in 200 µl exonuclease I mix (containing 1 × exonuclease I buffer and 1 U/µl exonuclease I) and incubated at 37 °C for 15 min (the beads were mixed by a rotary mixer). After being washed with 200 µl TE-TW (10 mM Tris (pH 8.0), 1 mM EDTA, 0.01% Tween 20) and 200 µl of 10 mM Tris-HCl (pH 8.0), the beads were resuspended in 1 ml double-distilled water. To remove complementary chains, the beads were placed in a 95 °C water bath for 6 min and separated using a magnet, removing the supernatant quickly, twice. The beads could be stored in TE-TW for 4 weeks at 4 °C. All oligonucleotide sequences are provided in Supplementary Table 1.

### Mouse experiments to supplement the MCA database

Wild-type C57BL/6J mice were ordered from Shanghai SLAC Laboratory Animal. 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 h light–dark cycle). Mice were provided ad libitum access to a regular rodent chow diet.

Mouse adult adrenal gland, omentum, pleura and stomach tissues were collected from 8–10 week-old female mice. Mouse fetal pancreas and stomach tissues were collected from mouse embryos at embryonic day (E)14.5. Single-cell collections for these mouse tissues followed the same protocols as for the corresponding human tissues (Supplementary Table 1). Around 10,000 single cells were sampled for each tissue from 2–4 randomly chosen replicates. The data generated from these tissues were used to supplement the MCA database to enable human–mouse comparison.

### Cell preparation

Samples from surgically removed adult tissues and aborted fetal tissues were put into Dulbecco's modified Eagle's medium plus 10% fetal bovine serum (FBS) immediately after resection. Samples from donations after cardiac death were resected after perfusion, and the time between cardiac death and tissue collection was controlled to within 1 h. Sample age (gestational age for fetal tissues), gender, cause of death, medical history were documented in Supplementary Table 1. Donated tissues were transported on ice from the hospital to the laboratory in less than 1 h, immediately transferred to cold Dulbecco's PBS (DPBS), and minced into pieces (-1 mm) on ice using scissors. The tissue pieces were transferred to a 15-ml centrifuge tube, rinsed twice with cold DPBS, and suspended in 5 ml solution containing dissociation enzymes. The samples were treated with various enzymes for different amounts of time (Supplementary Table 1). During dissociation, the tissue pieces were pipetted up and down gently several times until no tissue fragments were visible. The methods used for single-cell isolation from different tissues are listed in Supplementary Table 1. The dissociated cells were centrifuged at 300g for 5 min at 4 °C and then resuspended in 3 ml cold DPBS. After passage through a 40-µm strainer (Biologix), the cells were washed twice, centrifuged at 300g for 5 min at 4 °C, and resuspended at a density of  $1 \times 10^5$  cells/ml in cold DPBS containing

2 mM EDTA. CD34<sup>+</sup> cord blood and CD34<sup>+</sup> mPB were enriched using a human CD34 selection kit (StemCell Technologies).

### Embryoid body differentiation

Human iPS cells (Sidansai) were cultured in hES cell medium with basic fibroblast growth factor (bFGF), dissociated using dispase and resuspended in embryoid body medium (hESC medium without bFGF). Cell clusters were seeded in six-well ultra-low adherent plates (Corning) at a density of  $2 \times 10^5$  cells per well. The medium was changed every 3 days. On days 9, 18 and 20, embryoid body single cells were harvested by trypsinization and resuspended at a density of  $1 \times 10^5$  cells/ml in cold DPBS containing 2 mM EDTA.

### Cell collection and lysis

No fluorescence-activated cell sorting (FACS) step was used in any of the HCL experiments. All the analyses were single-cell analyses from fresh samples. Cell concentration was carefully controlled using a haemocytometer before microwell-seq. The proper cell concentration is ~100,000/ml (with 10% of the wells occupied by single cells). The proper bead concentration is ~1,000,000/ml (with every well occupied by single beads). An evenly distributed cell suspension was pipetted onto the microwell array, and extra cells were washed away. To eliminate cell doublets, the plate was inspected under a microscope. Cell doublets were reduced by pipetting over the region of high cell density. The bead suspension was then loaded into the microwell plate, and the plate was placed on a magnet. Excess beads were washed away slowly. Cold lysis buffer (0.1 M Tris-HCl (pH 7.5), 0.5 M LiCl, 1% sodium dodecyl sulfate (SDS), 10 mM EDTA, and 5 mM dithiothreitol) was pipetted over the surface of the plate and removed after 12 min incubation. The beads were then collected, transferred to an RNase-free tube, and washed once with 1 ml 6 × SSC (6 × saline sodium citrate solution: 0.9 M sodium chloride and 0.09 M sodium citrate), once with 500 µl 6 × SSC, and once with 200 µl of 50 mM Tris-HCl pH 8.0. Finally, ~50,000 beads were collected in a 1.5-ml tube.

### Reverse transcription

In this procedure, the instructions from the Smart-seq2 protocol were followed. In brief, 20 µl reverse transcription (RT) mix was added to the collected beads. The RT mix contained 200 U SuperScript II reverse transcriptase, 1× Superscript II first-strand buffer (Takara), 20 U RNase inhibitor (Sangon Biotech), 1 M betaine (Sigma), 6 mM MgCl<sub>2</sub> (Ambion), 2.5 mM dithiothreitol, 1 mM deoxynucleoside triphosphate, and 1 µM TSO primer. The beads were incubated at 42 °C for 90 min with mixing on a rotary mixer and then washed with 200 µl TE-SDS (1 × TE + 0.5% SDS) to inactivate reverse transcriptase. All oligonucleotide sequences are provided in Supplementary Table 1.

### Exonuclease I treatment

The beads were washed with 200 µl TE-TW and 200 µl of 10 mM Tris-HCl (pH 8.0), resuspended in 100 µl exonuclease I mix containing 1× exonuclease I buffer and 1 U/µl exonuclease I (NEB), and incubated at 37 °C for 60 min with mixing on a rotary mixer to remove oligonucleotides that did not capture mRNA. The beads were then pooled and washed once with TE-SDS, once with 1 ml TE-TW, and once with 200 µl of 10 mM Tris-HCl (pH 8.0).

### cDNA amplification

The beads were distributed into four PCR tubes. To each tube, 12.5 µl PCR mix (1× HiFi HotStart Readymix (Kapa Biosystems) and 0.1 µM TSO\_PCR primer) was added (Supplementary Table 1). The PCR program was as follows: 98 °C for 3 min; six cycles of 98 °C for 20 s, 65 °C for 45 s, and 72 °C for 6 min; 72 °C for 10 min; and a 4 °C hold. After all PCR products were pooled, AMPure XP beads (Beckman Coulter) were used to purify the cDNA samples from the PCR mix. Then, 25 µl PCR mix (1× HiFi HotStart Readymix and 0.1 µM TSO\_PCR primer) was

added to each DNA sample. The PCR program was as follows: 10 to 12 cycles of 98 °C for 3 min, 98 °C for 20 s, 67 °C for 20 s, and 72 °C for 6 min; 72 °C for 10 min; and a 4 °C hold. AMPure XP beads were used to purify the cDNA library.

### Transposase fragmentation and selective PCR

The purified cDNA library was fragmented using a customized transposase that carries two identical insertion sequences. The customized transposase was included in the TruePrep DNA Library Prep Kit V2 for Illumina (Vazyme). The fragmentation reaction was performed according to the instructions provided by the manufacturer. We replaced the index 2 primers (N5 × ×) in the kit with our P5 primer 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 30 s; five 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 AMPure XP beads. Then, 25 µl PCR mix (1× HiFi HotStart Readymix and 0.1 µM 2100 primer) was added to each sample. The PCR program was as follows: 95 °C for 3 min; five cycles of 98 °C for 20 s, 60 °C for 15 s, and 72 °C for 15 s; 72 °C for 35 min; and a 4 °C hold. To eliminate primer dimers and large fragments, AMPure XP beads were then used to purify the cDNA library. The size distribution of the products was analysed on an Agilent 2100 bioanalyser, and a peak in the 400–700-bp range was observed. Finally, the samples were subjected to sequencing on the Illumina HiSeq systems. All oligonucleotide sequences are provided in Supplementary Table 1.

### Immunohistochemistry

Donated human bladder and lung tissues were fixed in 4% paraformaldehyde overnight at 4 °C. Then, 30% sucrose/PBS was used to dehydrate the samples for 3 days at 4 °C. Immunohistochemistry was performed by Servicebio (Wuhan, China). In brief, the tissues were cut into 7-µm-thick frozen sections and mounted on precleaned slides. The sections were washed three times with PBS and blocked with 5% FBS in PBS for 1 h at room temperature. Primary antibodies (anti-HLA-DR (1:200, EPR3692; Abcam), anti-CD31 (1:100, 66065-1-Ig; Proteintech), anti-Krt17 (1:50, MAS-13539; Thermo Fisher), and anti-CXCL2 (1:100, AHP773; Bio-rad)) diluted in blocking solution (5% FBS in PBS) were added to cover the sections. The slides were placed in a wet box and incubated at 4 °C. Relevant AlexaFluor488/594-conjugated secondary antibodies (1:1,000, Invitrogen) were used for labelling. The slides were then washed three times with blocking solution and stained with DAPI. Glass coverslips were then attached to the slides using mounting medium. Immunofluorescence images were obtained using confocal microscopy.

### Processing of microwell-seq data

Microwell-seq data sets were processed as described<sup>16</sup>. Reads from HCL data were aligned to the *Homo sapiens* GRCh38 genome using STAR<sup>46</sup> and the DGE data matrices were obtained using the Dropseq core computational protocol (available at <http://mccarrolllab.org/dropseq/>) with default parameters. For quality control, we filtered out cells with detection of fewer than 500 transcripts. Cells with high proportion of transcript counts derived from mitochondria-encoded genes were also excluded.

### Clustering of single-cell data matrix

Seurat<sup>35</sup> was used to perform clustering analysis of single-cell data from different tissues. DGE data were used as inputs. Cells from the pre-processed data and genes expressed in more than three cells were selected for further analysis. Filtered data were ln(CPM/100 + 1) transformed, and the number of UMI and the percentage of mitochondrial gene content were regressed out according to the published method<sup>47</sup>. About 2,000 genes with an average expression of more than 0.01 and

# Article

a dispersion greater than 0.5 were used as inputs for initial principal component analysis (PCA) and the number of principal components (PCs) used for nonlinear dimensional reduction (*t*-SNE) analysis was chosen according to the PCElbowPlot function and JackStrawPlot function. For clustering, we set different resolution parameters between 0.6 and 4 in FindAllCluster function and narrowed down to certain cluster numbers by distinguishing differential genes among clusters. The heat map produced by DoHeatmap function is one basis for judging the quality of clustering. These parameters, including the resolution and number of PCs, were adjusted on per-tissue basis. The default Wilcoxon rank-sum test was used by running FindAllMarkers function in Seurat to find differentially expressed markers in each cluster. Finally, we annotate each cell type by extensive literature reading and searching for the specific gene expression pattern.

## Batch removal for cross-tissue comparison

To improve the presentation, we strictly removed the batch gene background for cross-tissue comparison. We assumed that, for each batch of experiments, cell barcodes with fewer than 300 UMIs correspond to empty beads exposed to free RNA during cell lysis, RNA capture and washing steps. Genes with extensive expression in all beads were considered batch genes. The batch gene background value was defined as the average gene detection for all cellular barcodes with fewer than 300 UMIs multiplied by the median of the fold difference between the detected gene expression of a cell and the average detected gene expression for beads with fewer than 300 UMIs; this value was then rounded to the nearest integer. We subtracted the batch gene background for each cell from the digital expression matrix before performing the cross-tissue comparison. We used the background-removed matrix to perform cross-tissue analyses such as the cross-tissue comparison of endothelial cells and stromal cells over the whole body.

## Landscape construction

For clustering of the complete human tissue data set in Fig. 1, we excluded data from differentiated cell cultures and G-CSF-mobilized PB to ensure natural status of the cell landscape. A total of 599,926 cells were selected and processed using Scanpy<sup>48</sup> in python environment. Background-removed DGE data with cells analysed in each tissue and genes expressed in at least 20 cells were used as inputs for Scanpy. Then, DGE data were  $\ln(\text{CPM}/100 + 1)$  transformed. We selected about 3,000 highly variable genes according to their average expression and dispersion. We then regressed out UMI and gene numbers and scaled each gene to unit variance with clip values exceeding a standard deviation of 10. We chose 50 PCs for PCA and computed the neighbourhood graph of cells. We then used the Louvain clustering to cluster cells with resolution of 3.5 and  $k=15$ . Finally, 102 clusters for the human landscape were produced and marker genes were calculated by the Wilcoxon rank-sum test. For sub-clustering of 102 clusters, we used the Seurat pipeline described previously with the default resolution of 0.8 to process cells from each cluster and predicted a total of 843 sub-clusters. The dendrogram in Fig. 1d was drawn using Pearson correlation coefficient (PCC) with mean gene expression for each cluster (CPM/100 values of all genes) with the R package denextend<sup>49</sup>.

## Pseudo-cell analysis

To increase the gene number and gene expression correlation from high-throughput single-cell mRNA data, we aggregated data from multiple cells in the same cell cluster to make pseudo-cells for genetic network interpretation<sup>50</sup>. To justify the use of pseudo-cells, we compared the performance of different levels of pseudo-cells in three mouse lung data sets from three different technologies: microwell-seq, 10X Genomics and Smart-seq2. Gene number, gene expression correlation, and silhouette value were used to evaluate the performance (Extended Data Fig. 2a). A high silhouette value indicates a high degree of separation

among cell types using pseudo-cell data. The global *t*-SNE map based on pseudo-cells indicates improved separation among cell types (Extended Data Fig. 2b). On the basis of human pseudo-cell 20 data, we calculated the TF-TF correlation using PCC and generated a correlation heat map that covers 1,521 human TFs (Extended Data Fig. 2c). A correlation threshold of 0.5 was used to construct a high-confidence TF-TF correlation network (Extended Data Fig. 2d).

## Receptor–ligand pairing analysis

Analysis of potential receptor–ligand pairings was performed by applying the recently published method CellPhoneDB<sup>37</sup>. First, we aggregated the gene expression levels of 20 cells from each cluster in the adult lung, adult kidney, fetal lung and fetal kidney. To eliminate the effect of variable cell numbers in each cluster, we randomly sampled three pseudo-cells for analysis. Only receptors and ligands expressed in more than 10% of the cells in the specific cluster were considered. By permuting cluster labels randomly 1,000 times to calculate the mean expression values of ligands and receptors, interaction was constructed as a receptor–ligand pairing matrix. Then, we used pairwise comparisons between all cell types and obtained a likelihood of *P* value to filter the false-positive interaction. The cutoff was set with the mean expression greater than 0.1 and *P* values smaller than 0.1. We used the sum of the number of receptor–ligand pairs in each cell–cell pairing to indicate the strength of the cell–cell interactions. Finally, the network was set to degree sorted circle layout for visualization in Cytoscape<sup>51</sup>.

## RNA velocity analysis

We used velocyto<sup>52</sup> to calculate RNA velocity on data from the human embryoid body on day 20. Velocyto estimates the rate of transcriptional changes of each cell based on the ratio of spliced and unspliced reads. We used the ‘velocyto run’ function described on the website (<http://velocyto.org/velocyto.py/tutorial/index.html>) with default parameters. Then, we imported the loom file from the last step and discarded cells that either did not have enough UMIs after the new mapping or did not have unspliced reads. Based on the coefficient of variation and average expression, we selected 2,013 genes to perform a PCA. Sixty-five PCs were used to impute with  $k = 5$  nearest neighbours. The plot was visualized with *t*-SNE embedding and the differentiation start and end points were estimated using a Markov process. We mainly followed the steps from the open repository ([https://github.com/rajewsky-lab/planarian\\_lineages](https://github.com/rajewsky-lab/planarian_lineages)).

## Differential gene expression analysis between fetal and adult cell lineages

For differential gene expression analysis in Supplementary Table 4, we aggregated data from 20 cells in the same cluster to make pseudo-cells for each cell type. Then we used MetaNeighbour<sup>53</sup> to find the related cell-type pairs between fetal and adult tissues. In brief, MetaNeighbour was obtained through neighbour voting based on the Spearman correlation between all fetal and adult pseudo-cells. Then the mean AUROC scores were obtained from MetaNeighbour. Cell type pairs with AUROC score  $> 0.9$  were regarded to have strong relationships in the development. Next, we performed the Wilcoxon rank-sum test to find the differential expressed genes within these cell type pairs. Genes with adjusted  $P < 0.05$  for each cell-type pair were then labelled. Common top differential genes were estimated by the frequency of differential expression in all fetal-to-adult cell-type pairs.

## Single-cell trajectory analysis

We used PAGA in scanpy<sup>48</sup> to infer the lineage tree of single-cell data from non-immune cell types in fetal and adult tissue as well as CD34<sup>+</sup> haematopoietic data sets. The graph abstraction algorithm reconciles clustering and trajectory inference by explaining data variability in terms of both discrete and continuous latent variables. First, we

processed the data following the steps suggested by scanpy, including total count normalization, log1p logarithmization, highly variable genes extraction, a potential regression of confounding factors of genes and counts, a scaling to z-scores and PCA analysis. Then, we computed a neighbourhood graph among data points and used UMAP to generate a topologically faithful embedding with `min_dist = 0.1`. Then, PAGA was performed; with `iter = 1,000`, the trajectory was constructed using layout ‘fa’. Differential gene expression analysis was performed with the `tl.rank_gens_groups()` function in scanpy with Wilcox rank-sum test.

### Single-cell entropy analysis

We applied single cell lineage inference using cell expression similarity and entropy (SLICE) for quantitative measurements of cell differentiation states based on the calculation of single-cell entropy (`scEntropy`)<sup>54</sup>. We performed deterministic calculation of `scEntropy` of individual cells in fetal and adult tissues with default parameters according to the SLICE pipeline. Similar results can also be obtained using the StemID method<sup>55</sup>.

### HCL website construction

The main HCL website uses a bootstrap framework to improve overall adaptability and interactivity. Its back-end is completed by PHP, R language and MySQL. The main functions of the HCL website are divided into four parts: gallery, landscape, search and scHCL. Gallery provides interactive t-SNE maps for more than 80 data sets to show the distribution of different clusters. Specific markers for each cluster are listed in a data table. Landscape achieves better visualizations for global view of 102 clusters using both single cells and pseudo-cells. Search describes the expression of a given gene in different clusters from any selected tissue. scHCL provides the function of single-cell correlation analysis with the HCL database. After users upload their own DGE files, the data are processed by the R script and compared to the HCL reference file. The scHCL result is returned in JSON format and presented as an interactive heat map.

### scHCL analysis

Similar to our previously published single-cell mapping pipeline<sup>16</sup>, the scHCL analysis was conducted through the following steps. We downloaded published drop-seq, in-drop, seq-well and 10X Genomics data from human samples<sup>11,20,22,24–31,37,41,43,56–62</sup> (Supplementary Table 5). We redid the clustering for each data set, combined the cell-type clusters with our HCL clusters and generated a total of 1,841 cell-type clusters (cell-type redundancies exist in these clusters). Then, for each cell-type cluster, we randomly sampled 100 single cells without replacement (all cells for clusters with fewer than 100 cells), calculated the average expression normalized to 100,000 transcripts, and rounded the number down to the nearest integer. We constructed the averaged cell-type transcriptome data three times for each cell cluster. This resulted in main transcriptome references in our scHCL pipeline. We then performed differential gene expression analysis for each cell type against all the other cell types and selected the top 20 marker genes for each cell type ( $\log_{2}\text{foldchange} > 1$ ). Markers for each cell type were merged to create the combined feature gene list. The PCCs of the given single-cell data against each HCL cell-type reference were then calculated using the combined feature gene list. To properly display the mapping results of scHCL, the number of top cell-type hits for each single cell (`nCluster`) and the lowest correlation coefficient threshold (`corThreshold`) can be manually adjusted. For all scHCL results shown in this paper, `nCluster` was set to be 1; only the top 1 mapping hit was shown for each data point. The single-cell fragments per kilobase of transcript per million mapped reads (FPKM), reads per kilobase of transcript per million mapped reads (RPKM), transcripts per million (TPM) digital gene expression (DGE) and bulk RNA DGE matrix can also be applied to the scHCL pipeline.

### Cross-species transcriptome comparison

Before transcriptome comparison, we supplemented the MCA database with new microwell-seq data set from adult adrenal gland, omentum, pleura and stomach tissues, as well as fetal pancreas and stomach tissues. We followed the HCL pipeline to re-cluster the updated MCA data ([https://figshare.com/articles/MCA\\_DGE\\_Data/5435866](https://figshare.com/articles/MCA_DGE_Data/5435866)) ( $n = 333,778$ ) into 104 major clusters (Extended Data Fig. 10a, b). We then divided the 102 HCL and 104 MCA cell clusters into 12 major cell lineages. To make the gene expression profiles of cross-species cell types comparable, we downloaded the homology correspondences between human and mouse provided by dmod-ENCODE<sup>63</sup>. The gene expression profiles for human (this study) and mouse<sup>16</sup> were normalized to the total number of transcripts and multiplied by 100,000. To attenuate the effects of noise and outliers, we used pseudo-cells<sup>36</sup> for further analysis; each pseudo-cell was an average of 20 cells randomly selected from the same cell type. To compare cross-species transcriptomes, we performed MetaNeighbour<sup>53</sup> analysis through neighbour voting based on the spearman correlation between all human and mouse pseudo-cells. Then the mean AUROC scores were obtained from MetaNeighbour. The Circlize<sup>64</sup> package was used to view the similarity of cell-type pairs between species, whose AUROC scores were higher than 0.9 or match types were ‘Reciprocal\_top\_hit’.

### Cross-species regulatory network comparison

Human and mouse gene regulatory network analysis was performed using SCENIC<sup>45</sup> with default parameters. In brief, the gene regulatory network was based on co-expression and DNA motif analyses, and then the AUCell algorithm was used to score the activity of each TF regulon in each pseudo-cell. From SCENIC analysis, we obtained a total of 259 human TF regulons and 248 mouse TF regulons, among which we defined 119 human unique TF regulons, 108 mouse unique TF regulons and 140 orthologous TF regulons (Supplementary Table 5). Based on 140 orthologous TF regulons, we obtained a merged AUCell score matrix from both human and mouse pseudo-cells. Then the PCC of AUCell score was calculated for each pair of orthologous TF regulons. From the analysis, we identified 15 orthologous TF regulon modules based on the connection specificity index (CSI)<sup>65</sup> using the ‘ward.D’ clustering method. For a fixed pair of regulons, a and b, the corresponding CSI was defined as the fraction of regulons whose PCC of AUCell scores with a and b were lower than the PCC between a and b themselves. The distribution of mean AUCell scores across the cell types was used to present the relationship between orthologous TF regulon modules and cell types.

### Regulon activity analysis

To determine the ‘on/off’ activity of each regulon (259 TF regulons for human and 248 TF regulons for mouse) in each cell type, we used ‘0.5\*max (AUC scores)’ for each regulon as a threshold to binarize the regulon activity scores and created the ‘binary regulon activity matrix’ from 17,028 human pseudo-cells and 16,740 mouse pseudo-cells. The values of the matrix that corresponded to ‘on’ regulons in a given pseudo-cell were 1, and 0 for ‘off’ regulons. The regulon activity t-SNE maps were created using the function `tsneAUC(..., nPCs = 50, perplexity = 50, aucType = "binary")` in R package SCENIC with the binary regulon activity matrix. Binary RASs for regulons were projected to the regulon activity t-SNEs with the function `runSCENIC_4_aucell_binarize()` in R package SCENIC. To connect regulons with cell types, we used the Wilcoxon rank-sum test to identify cell-type-specific regulons with AUC score matrices (Extended Data Fig. 11, see Source Data), and the corresponding binding motifs for regulons were obtained from JASPAR, CIS-BP or the HOCOMOCO database.

## Statistics and reproducibility

In this study, we analysed around 10,000 single cells for each tissue type to ensure reproducible detection of different cell populations. Whenever donors were available, we collected 2–4 biological replicates to evaluate donor effects for different tissues. For the boxplots in Extended Data Figs. 2a, 8e, the top of the rectangle indicates the third quartile, a horizontal line near the middle of the rectangle indicates the median, and the bottom of the rectangle indicates the first quartile. A distance of 1.5 times the interquartile range (IQR) is measured out and a whisker is drawn up to the largest observed point from the data set that falls within this distance. Similarly, a distance of 1.5 times the IQR is measured out below the lower quartile and a whisker is drawn up to the lower observed point from the data set that falls within this distance. All other observed points are plotted as outliers.

## Reporting summary

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

## Data availability

The CNGB Nucleotide Sequence Archive accession number is CNP0000325 (<https://db.cngb.org/search/?q=CNP0000325>). The GEO accession number is GSE134355. The human DGE data are available at [https://figshare.com/articles/HCL\\_DGE\\_Data/7235471](https://figshare.com/articles/HCL_DGE_Data/7235471). The mouse DGE data are available at [https://figshare.com/articles/MCA\\_DGE\\_Data/5435866](https://figshare.com/articles/MCA_DGE_Data/5435866). Source Data for Figs. 2, 4 and Extended Data Figs. 2, 5, 7, 8, 9, 10, 11 are provided with the paper. HCL data can also be accessed at <http://bis.zju.edu.cn/HCL/> or <https://db.cngb.org/HCL/>.

## Code availability

Detailed codes for figures are provided at <https://github.com/ggjlab/HCL/>. An online R package is available for scHCL (<https://github.com/ggjlab/scHCL/>).

46. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
47. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
48. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
49. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).
50. Suo, S. et al. Revealing the critical regulators of cell identity in the mouse cell atlas. *Cell Rep.* **25**, 14436–1445.e3 (2018).
51. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
52. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
53. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).

54. Guo, M., Bao, E. L., Wagner, M., Whitsett, J. A. & Xu, Y. SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res.* **45**, e54 (2017).
55. Grün, D. et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266–277 (2016).
56. Guo, J. et al. The adult human testis transcriptional cell atlas. *Cell Res.* **28**, 1141–1157 (2018).
57. Han, X. et al. Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biol.* **19**, 47 (2018).
58. Karamitos, D. et al. Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nat. Immunol.* **19**, 85–97 (2018).
59. Nguyen, Q. H. et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* **9**, 2028 (2018).
60. Tsang, J. C. H. et al. Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics. *Proc. Natl Acad. Sci. USA* **114**, E7786–E7795 (2017).
61. Villani, A. C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
62. Wang, M. et al. Single-cell RNA sequencing analysis reveals sequential cell fate transition during human spermatogenesis. *Cell Stem Cell* **23**, 599–614.e4 (2018).
63. Celiker, S. E. et al. Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
64. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
65. Fuxman Bass, J. I. et al. Using networks to measure similarity between genes: association index selection. *Nat. Methods* **10**, 1169–1176 (2013).

**Acknowledgements** We thank G-BIO (Hangzhou), Sidansai Biotechnology (Shanghai), BGI (Shenzhen) and CNGB (Shenzhen) for supporting this project; Vazyme 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; and Y. Zhu, J. Zhu, L. Huang, L. Shao, Z. Wang, H. Huang, X. Wu, W. Lin, M. Bai, Q. Sun, X. Wu, M. Yao, F. Zhu, Z. Li, L. Huang, L. Shao, Z. Wang and X. Chen for help with sample collection. This publication is part of the Human Cell Atlas: [www.humancellatlas.org/publications/](http://www.humancellatlas.org/publications/). G.G. is a participant of the Human Cell Atlas Project (International), the Alliance for Atlas of Blood Cells (China), and the Cell Atlas Project (Zhejiang University Stem Cell Institute). This work was supported by the National Natural Science Foundation of China (grants 91842301, 81770188, 31722027, 31922049, 31701290, and 31871473), the National Key Research and Development Program (grants 2018YFA0107804, 2018YFA0107801, 2018YFA0800503, and 2018YFC1005003), the Zhejiang Provincial Natural Science Foundation of China (grant R17H080001), and the Fundamental Research Funds for the Central Universities (G.G.).

**Author contributions** The project was conceived by G.G. Tissue digestion experiments were performed by X.H., Z.Z., R.W. and Y.C. Microwell-seq experiments were performed by X.H., Z.Z., R.W., H.C., F.Y., M.J., J. Wu and S.L. Single-cell data processing, clustering and trajectory analyses were performed by L.F., H.S., J. Wang, Y.X. and C.Y. scHCL analyses and website construction were performed by H.S., Y.Z. and M.C. Cross-species and gene regulation analyses were performed by J. Wang and H.S. Immunostaining experiments were performed by H.C. Stem cell differentiation experiments were performed by H.C., X.M. and S.Z. Sequencing experiments were performed by R.L., Y.G. and M.W. Fetal tissue collections were conducted by Y.C., Y.W. and D.Z. Adult brain tissue collections were conducted by X.J., J.Z., R.Z. and H. Hu. Adult hematopoietic cell collections were conducted by H. Huang. Other adult tissue collections were conducted by H.T., W.G., T.Z., Q.Z., X.B., L.Z., C.W., T.L., J.C. and W.W.. The paper was written by G.G., X.H., Z.Z., L.F., H.S., R.W., Y.C., H.C. and J. Wang. Funding was acquired by G.G. and X.H.

**Competing interests** The authors declare no competing interests.

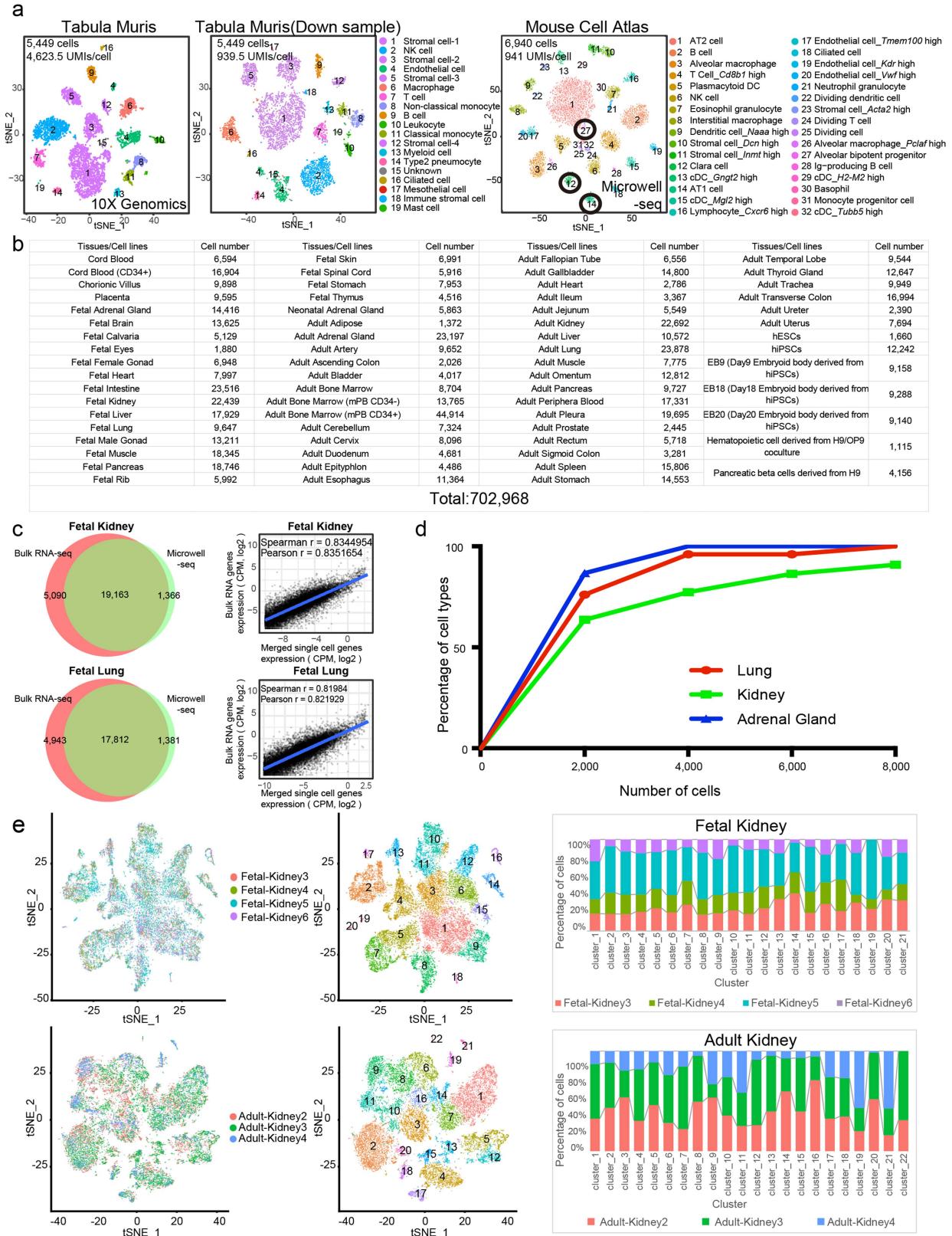
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-020-2157-4>.

**Correspondence and requests for materials** should be addressed to X.H. or G.G.

**Peer review information** *Nature* thanks Berthold Gottgens, Rosario Isasi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

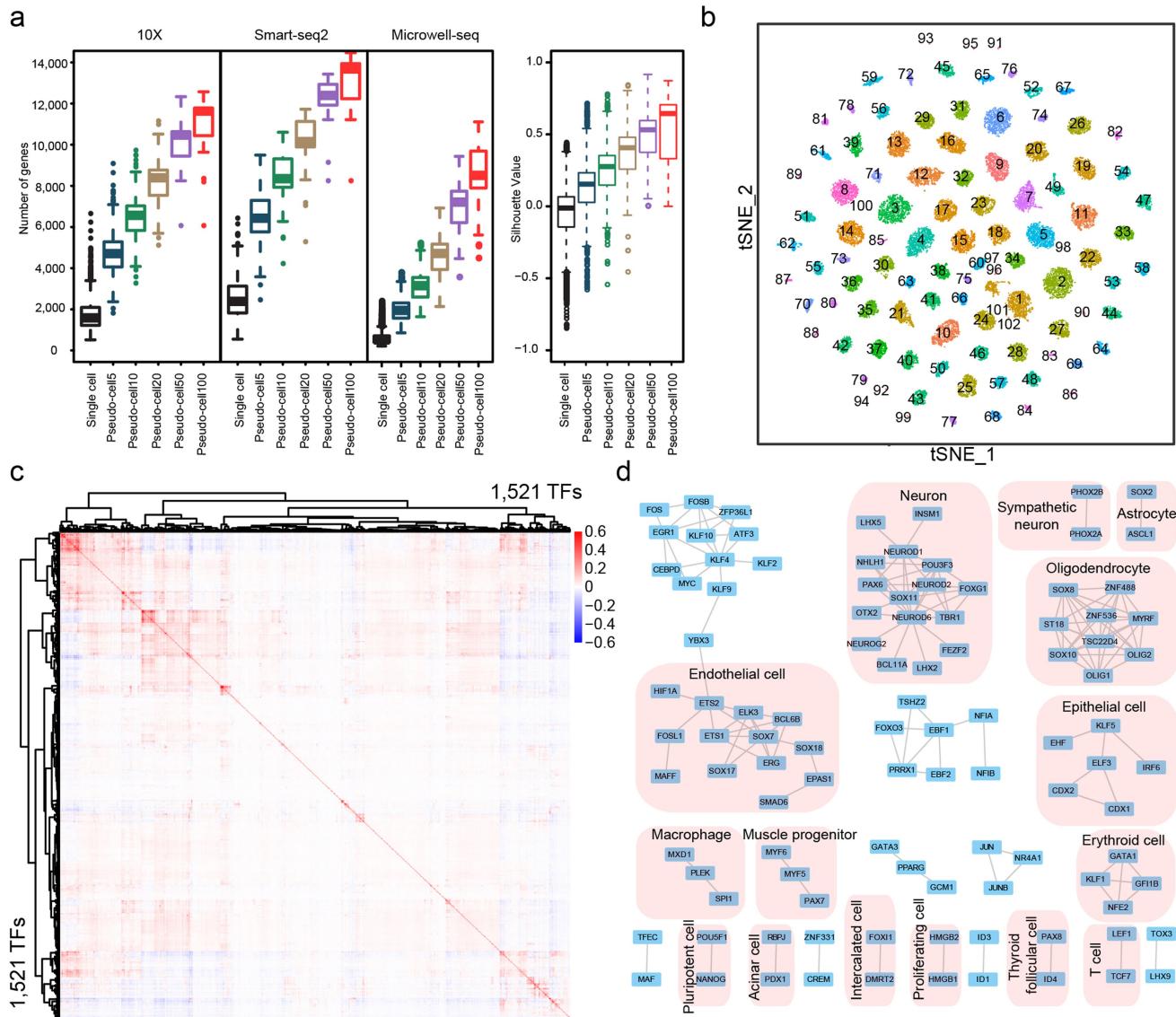


**Extended Data Fig. 1** | See next page for caption.

## Article

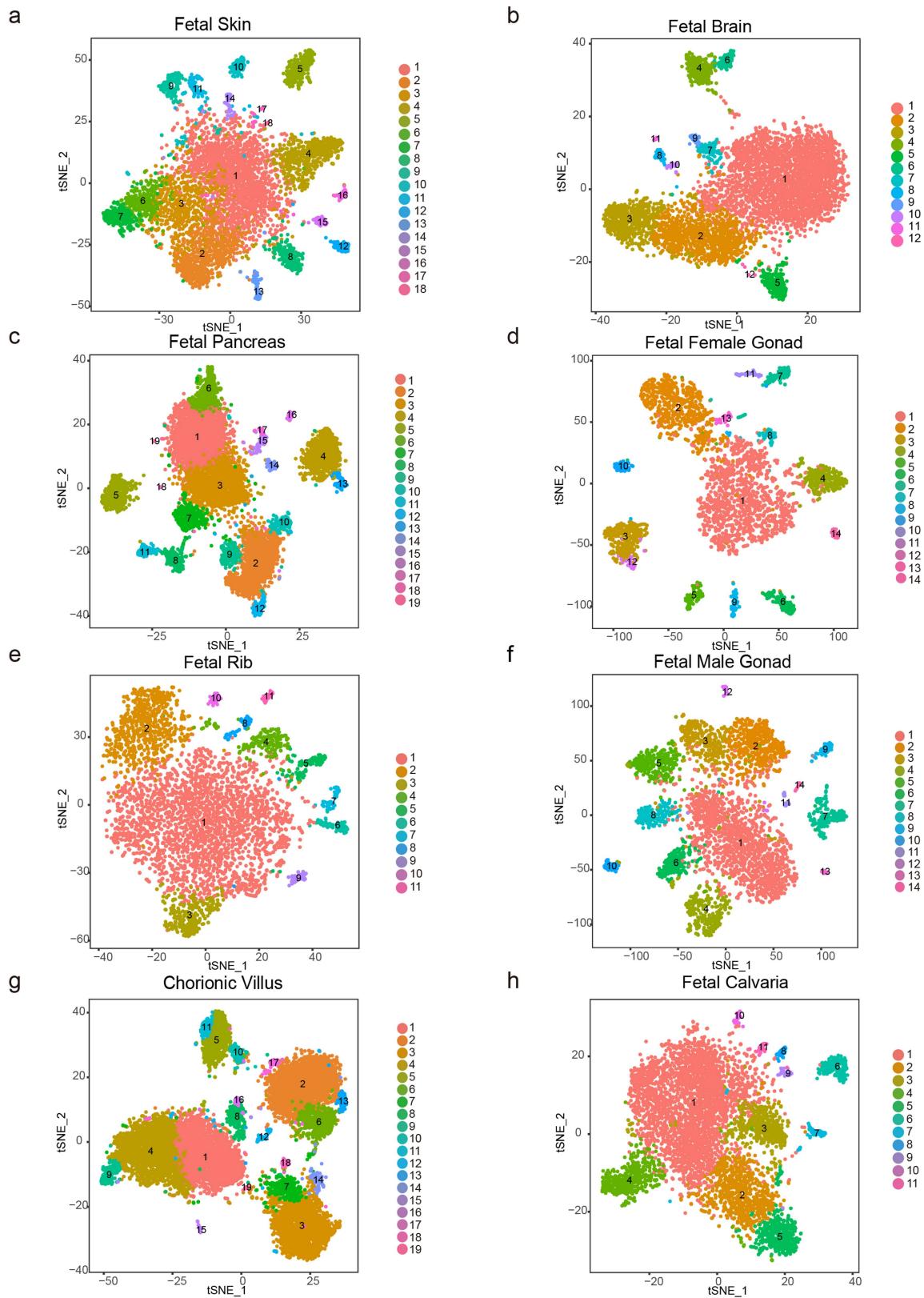
**Extended Data Fig. 1 | Construction of the HCL.** **a**, Comparison of t-SNE maps for lung data sets from Tabula Muris (10X Genomics), down-sampled Tabula Muris data (10X Genomics), and MCA data (microwell-seq). Note that downsampling in sequencing depth does not affect cell-type clusters in the Tabula Muris data. Notably, lung data from MCA (sequenced at lower depth) detects more cell-type clusters, including important lung epithelial cells such as AT1 cells, club cells and biopotent progenitors. **b**, Numbers of cells processed by 31 December 2019 at the HCL for each tissue type. **c**, Venn diagrams of gene numbers detected in bulk RNA sequencing and microwell-seq (genes with fewer than three counts are excluded). Scatter plots on the right show high correlations (more than 0.8) of average gene expression between

bulk RNA sequencing and microwell-seq. We analysed 17,058 genes for kidney and 16,910 genes for lung. **d**, The percentage of cell types recovered in sub-samples of adult lung, kidney and adrenal gland single-cell data. The major cell-type numbers in representative tissues are near plateau at around 8,000 cells; we collected more than 10,000 cells per tissue on average. **e**, A Seurat analysis of donor batch effect from four fetal kidney samples ( $n=22,439$  cells) and three adult kidney samples ( $n=22,692$  cells). The mixing of different donor single cells in each cell-type cluster suggests a relatively low batch effect in the data. The cluster contribution bar charts on the right suggest that one of the fetal kidney donors lacks C19 and one of the adult kidney donors lacks C22.



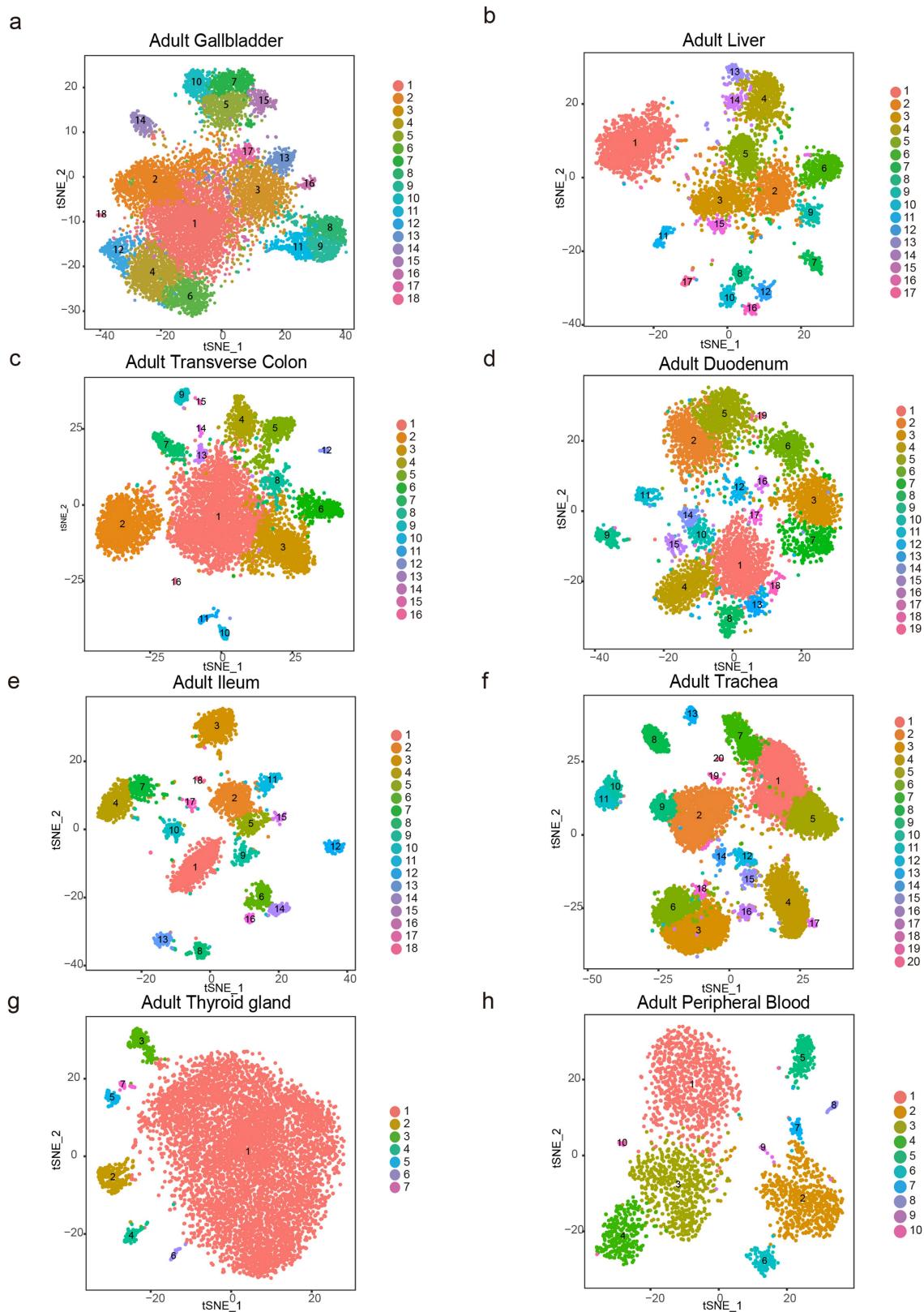
**Extended Data Fig. 2 | Genetic network analysis of the HCL.** **a**, Verification of pseudo-cell analysis for network interpretation from HCL data. Left, numbers of genes in single cell, pseudo-cell 5, pseudo-cell 10, pseudo-cell 20, pseudo-cell 50 and pseudo-cell 100 from mouse lung single-cell data<sup>14,15</sup> generated by 10X Genomics, Smart-seq2 and microwell-seq. Genes were calculated in each cell or pseudo-cell. Sample sizes for each box from left to right were: 10X Genomics: 5,449, 1,089, 540, 272, 112 and 62; Smart-seq2: 1,620, 324, 158, 83, 33 and 20; microwell-seq: 6,940, 1,390, 686, 349, 142 and 81. Right, silhouette value in single cell, pseudo-cell 5, pseudo-cell 10, pseudo-cell 20, pseudo-cell 50 and pseudo-cell 100 from mouse lung single-cell data. A high

silhouette value represents good separation. Sample sizes for each box from left to right were: 6,940, 1,390, 686, 349, 142 and 81. Box plots: centre line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within  $1.5 \times \text{IQR}$ . **b**, t-SNE map of HCL pseudo-cell data showing improved cell-type clustering ( $n = 30,053$  pseudo-cells). **c**, TF-TF correlation heat map covering 1,521 human TFs generated using HCL pseudo-cell data. The correlation data are listed in Supplementary Table 2. **d**, Representative TF network in the HCL ( $\text{PCC} > 0.5$ ). Note that the HCL TF network is highly related within small modules but discrete among different modules.



**Extended Data Fig. 3 | t-SNE maps for examples of analysed fetal tissues in the HCL.** t-SNE maps for single-cell data from fetal skin 2 (**a**; n=5,294 cells), fetal brain 5 (**b**; n=5,096), fetal pancreas 2 (**c**; n=6,939), fetal female gonad

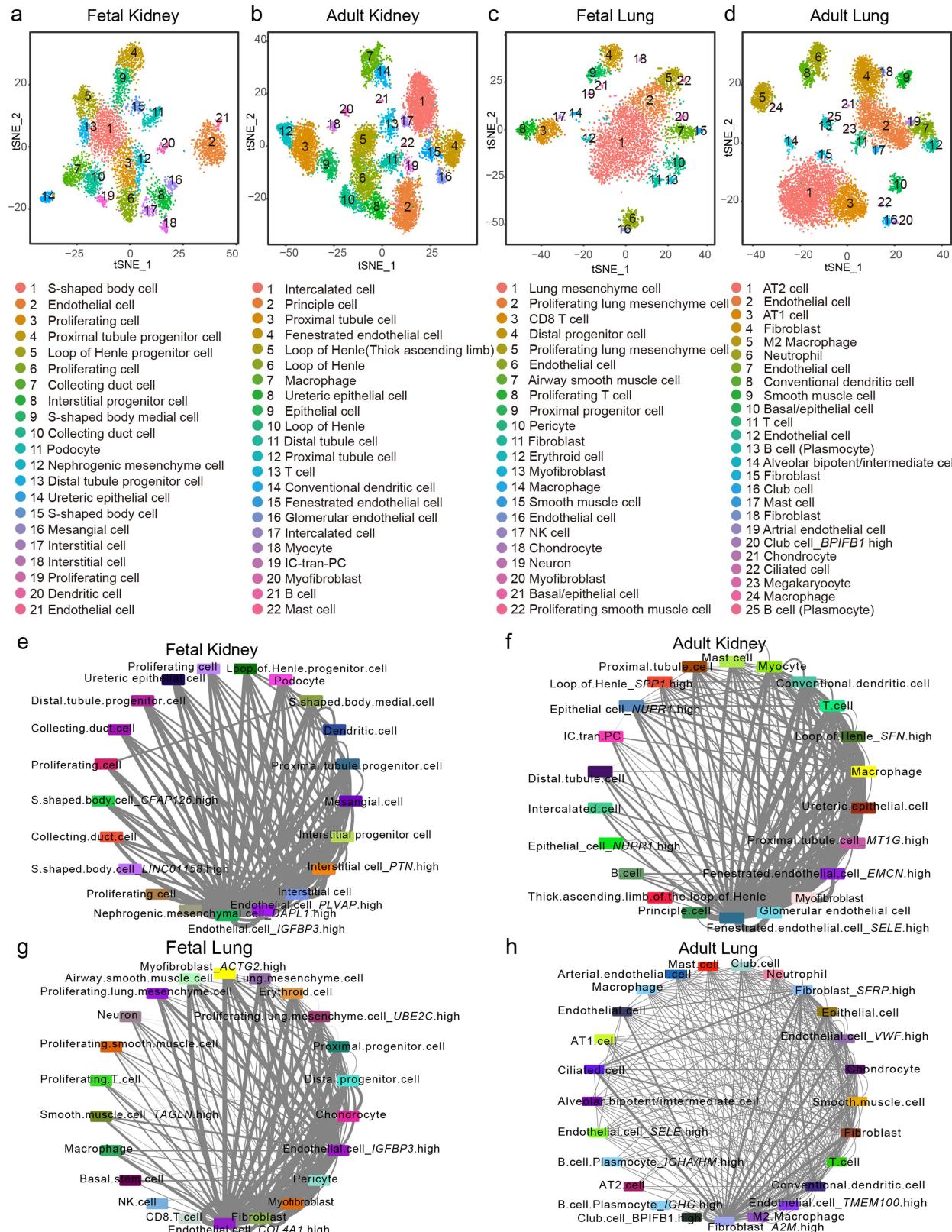
(**d**; n=2,710), fetal rib 3 (**e**; n=4,560), fetal male gonad 1 (**f**; n=3,358), chorionic villus 1 (**g**; n=9,898), and fetal calvaria 1 (**h**; n=5,129). Cells are coloured by cell-type cluster.



**Extended Data Fig. 4 | t-SNE maps for examples of analysed adult tissues in the HCL.** t-SNE maps for single-cell data from adult gallbladder 1 (a,  $n = 9,769$  cells), adult liver 4 (b,  $n = 4,384$ ), adult transverse colon 1 (c,  $n = 5,765$ ), adult

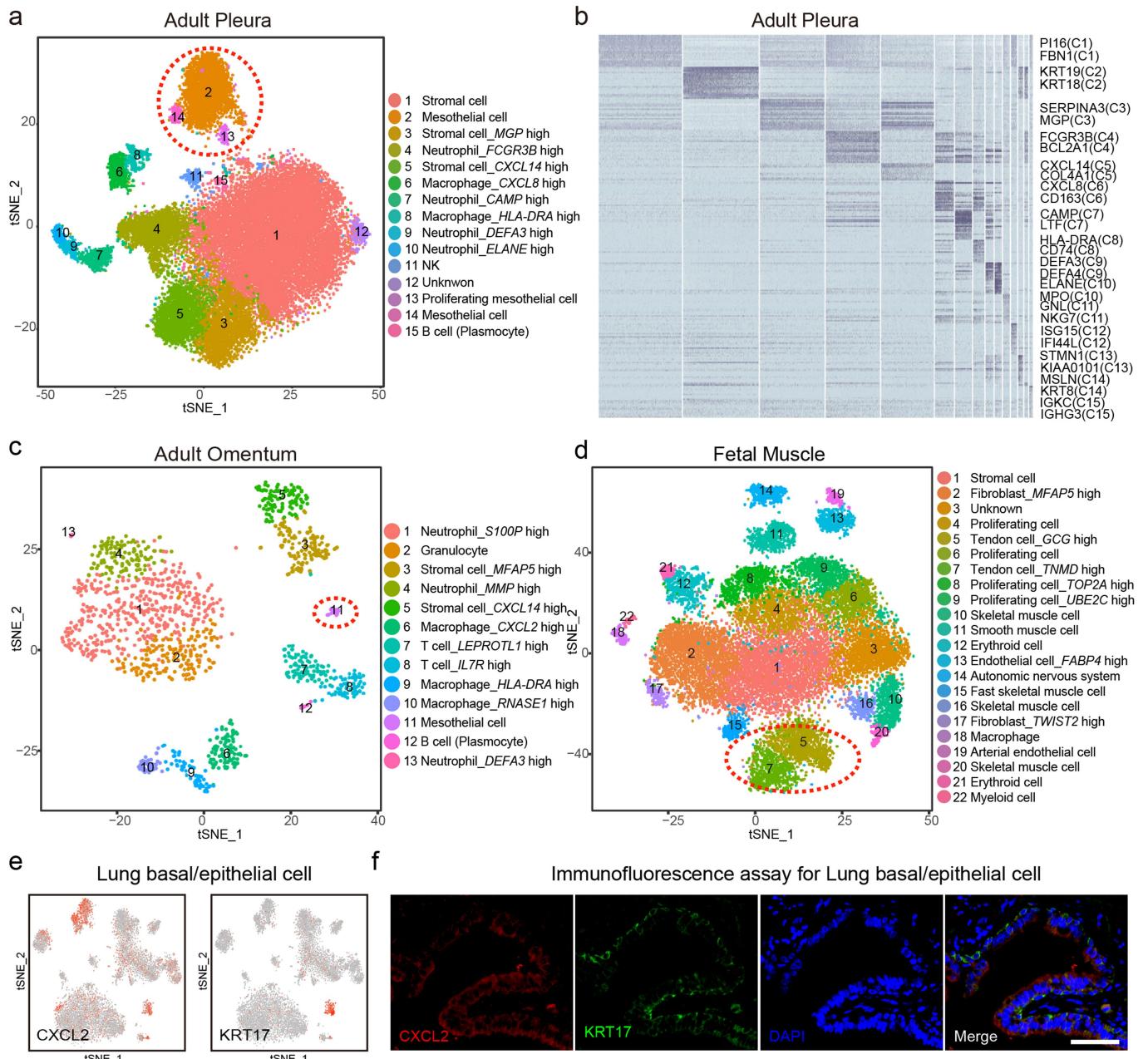
duodenum 1 (d,  $n = 4,681$ ), adult ileum 2 (e,  $n = 3,367$ ), adult trachea 2 (f,  $n = 9,949$ ), adult thyroid gland 1 (g,  $n = 6,319$ ), and adult peripheral blood 1 (h,  $n = 2,719$ ). Cells are coloured according to cell-type cluster.

# Article



**Extended Data Fig. 5 | Analysis of human lung and kidney.** **a**, t-SNE map of fetal kidney 4 single-cell data ( $n=4,511$  cells). The experiment was replicated four times with similar results. **b**, t-SNE map of adult kidney 2 single-cell data ( $n=8,877$  cells). The experiment was replicated three times with similar results. **c**, t-SNE map of fetal lung 1 single-cell data ( $n=4,526$  cells). The experiment was replicated twice with similar results. **d**, t-SNE map of adult lung 1 single-cell data

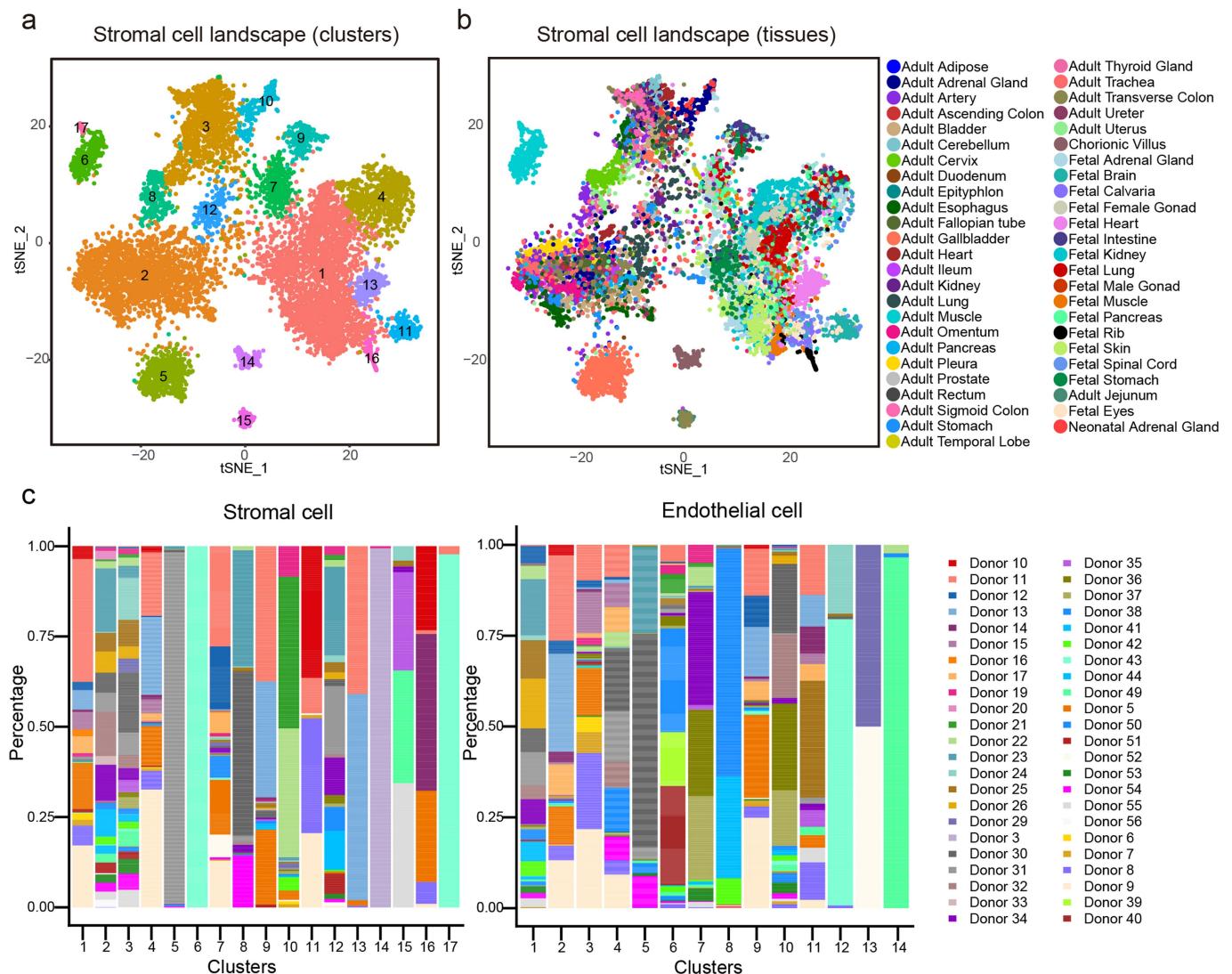
( $n=8,426$  cells). The experiment was replicated three times with similar results. All cells in **a-d** are coloured according to cell-type cluster. **e-h**, Ligand and receptor analysis of fetal kidney 4 (**e**), adult kidney 2 (**f**), fetal lung 1 (**g**) and adult lung 1 (**h**) using the method CellPhoneDB. The colours represent cell types; line thickness indicates the degree of association between cell types.



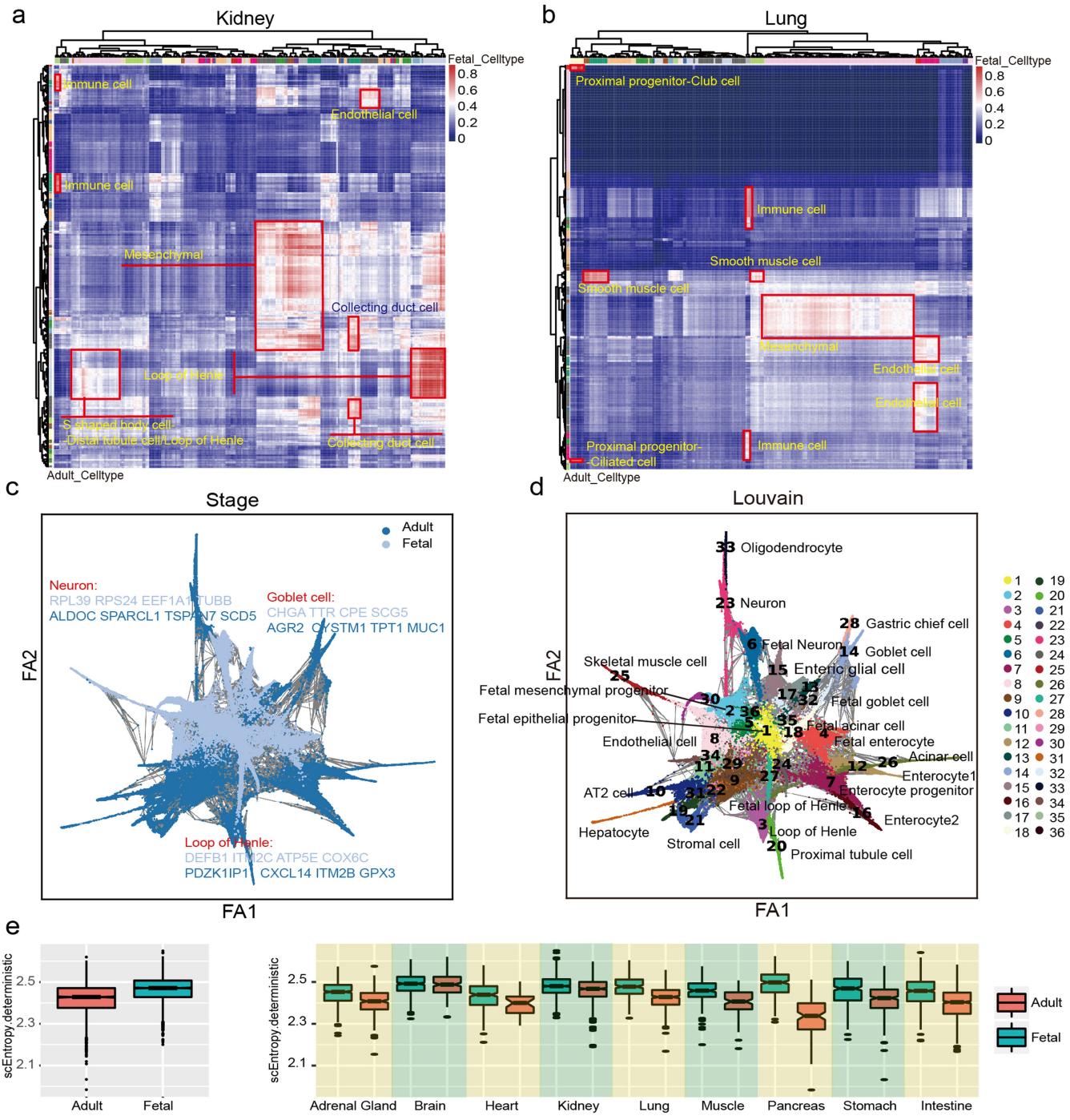
**Extended Data Fig. 6 | Examples of novel populations.** **a**, t-SNE map of adult pleura 1 single-cell data ( $n=19,695$  cells). Cells are coloured according to cell-type cluster in **a**, **c** and **d**. **b**, Gene expression heatmap showing the top 20 differentially expressed genes for each cell cluster in adult pleura 1. Dark blue, high expression; light blue, low expression. Representative genes are labelled in the corresponding area on the right. **c**, t-SNE map of adult omentum

3 single-cell data ( $n=1,354$  cells). **d**, t-SNE map of fetal muscle single-cell data ( $n=18,345$  cells). **e**, Feature plot in the t-SNE map of adult lung 1 single-cell data ( $n=8,426$  cells). Cells are coloured according to the expression of the indicated marker genes. **f**, Immunofluorescence assay for the epithelial cell marker KRT17 and the CXC chemokine CXCL2 in human adult lung tissue. Scale bar, 50  $\mu$ m. The experiment was replicated three times with similar results.

# Article



**Extended Data Fig. 7 | Cross-tissue cellular network.** **a, b**, t-SNE maps of single-cell data for human tissue-specific stromal cells ( $n=9,452$  cells). Cells are coloured according to stromal cell subtype (a) or tissue type (b). **c**, Bar plots showing the contributions of donors to each of the stromal and endothelial cell clusters.



**Extended Data Fig. 8 | Analysis of fetal-to-adult transition.** **a**, Heat map showing the similarity (PCC) between cell types of adult kidney and fetal kidney. Blue, low similarity; red, high similarity. **b**, Heat map showing the similarity (PCC) between cell types of adult lung and fetal lung. Blue, low similarity; red, high similarity. **c**, **d**, Branching gene expression trajectory analysis of non-immune cells in fetal and adult human tissues using PAGA. **c**, Coloured by developmental stages; **d**, coloured by cell lineages. Differential gene expression analysis was performed for representative lineage

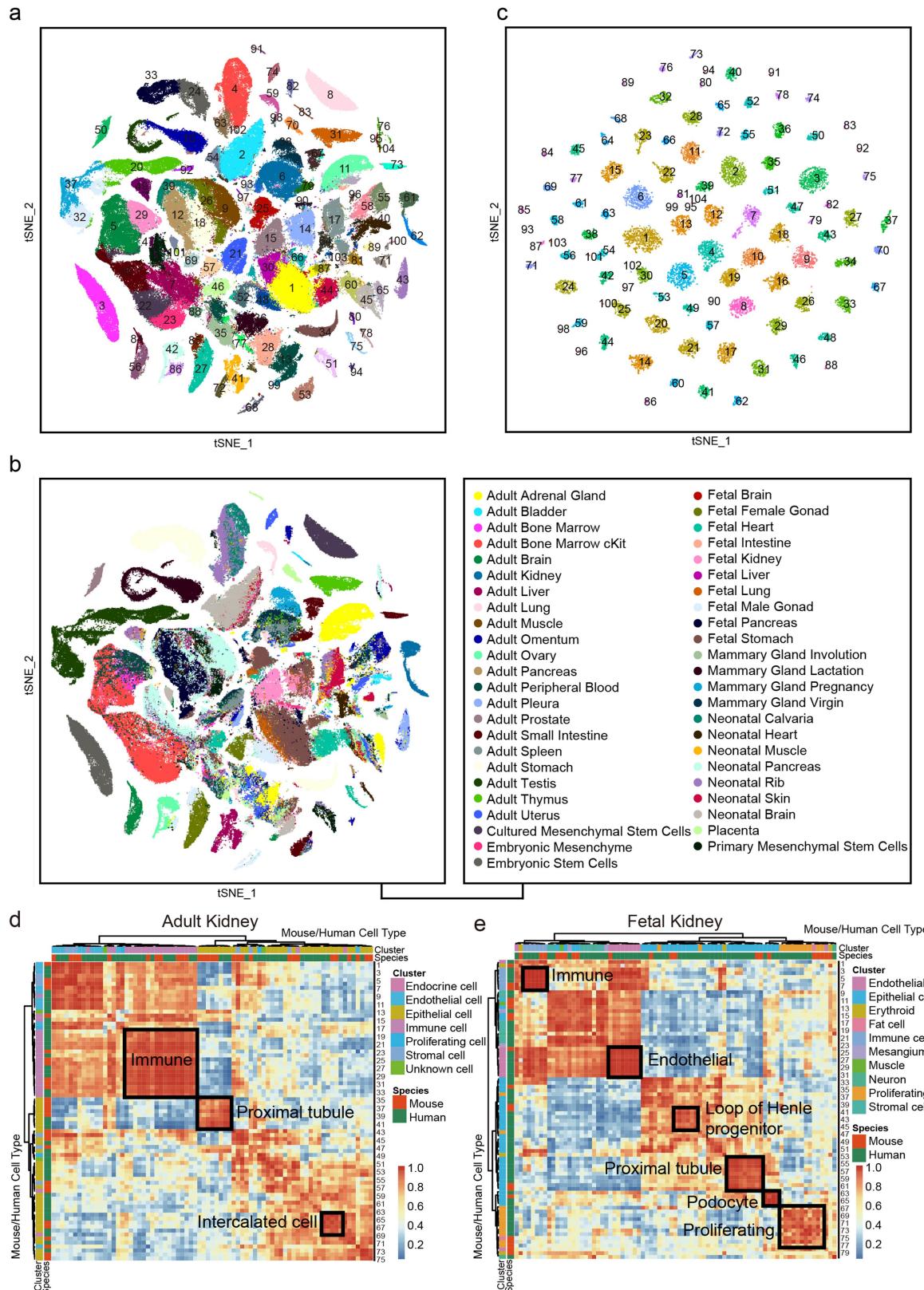
progression; top markers for fetal (light blue) and adult (dark blue) cells are shown. Marker lists are provided in the source data. **e**, Single-cell entropy of non-immune cells in each fetal and adult tissue pair (right) and for combined adult and fetal data (left) measured by SLICE<sup>54</sup>. Box plots: centre line, median; boxes, first and third quartiles of the distribution; whiskers, highest and lowest data points within  $1.5 \times \text{IQR}$ . Sample sizes from left to right: 2,770, 3,557, 522, 279, 340, 193, 270, 38, 885, 631, 208, 353, 187, 111, 268, 112, 294, 314, 583, 739.

# Article



**Extended Data Fig. 9 | scHCL analysis.** **a**, scHCL results for bulk human ES cell data ( $n=17$  samples). Each row represents one cell type in our reference. Each column represents a bulk sample. PCC was used to evaluate cell-type gene expression similarity. Red, high correlation; grey, low correlation. Some cell-type data come from published works, as denoted by first author name: Zhong<sup>20</sup>, Li<sup>25</sup>, Segerstolpe<sup>27</sup>, Baron<sup>29</sup>, VentoTormo<sup>37</sup>, Camp<sup>42</sup>, LaManno<sup>43</sup>, Han<sup>57</sup>. **b**, scHCL results for colorectal cancer organoid data ( $n=15$  samples), as in **a**. **c**, scHCL results for liver bud organoid data ( $n=465$  cells), as in **a** except that each

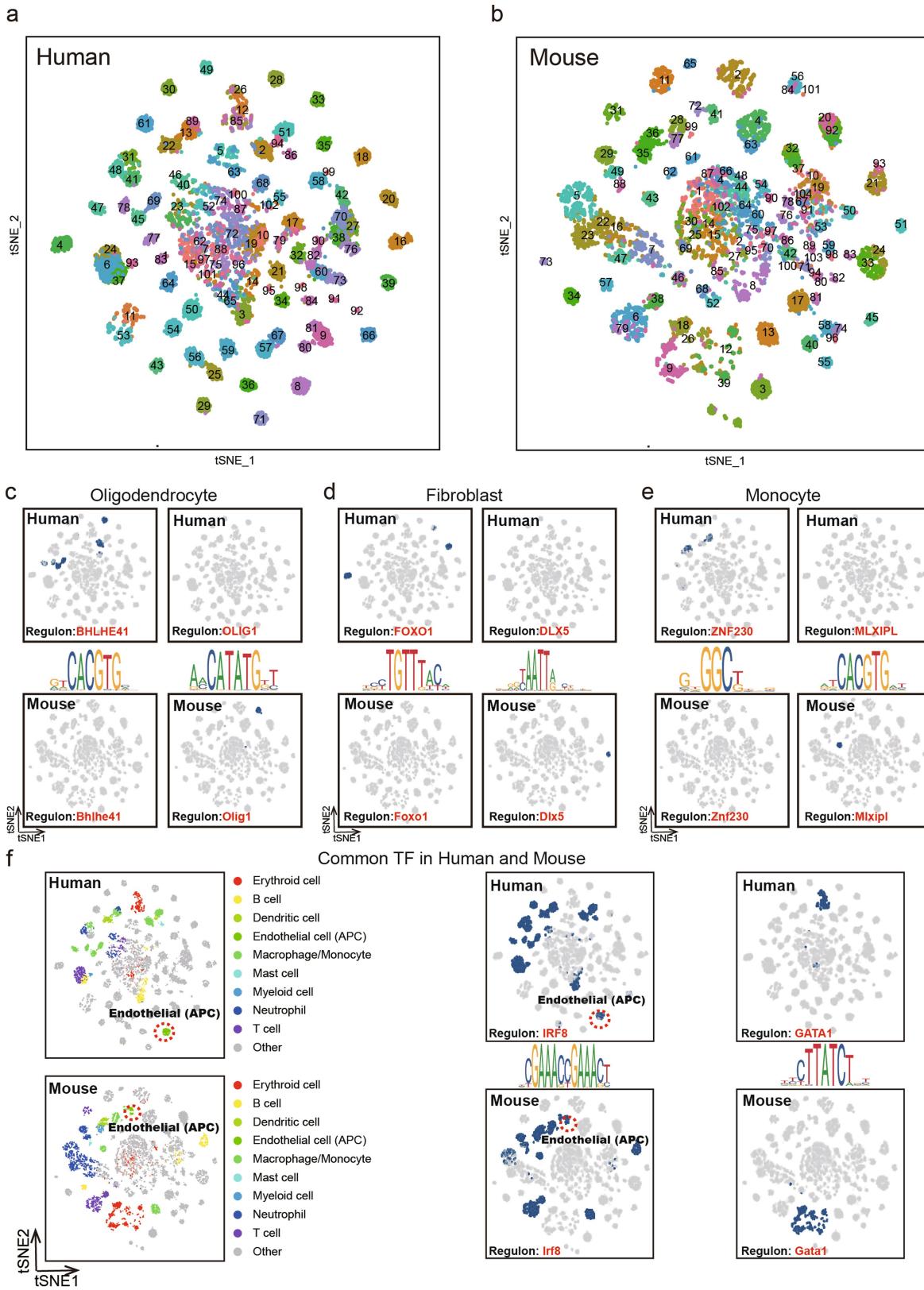
column represents a single cell in the customer data set. **d**, scHCL results for cerebral organoid data ( $n=508$  cells), as in **c**. **e**, tSNE map of single-cell data for day 20 embryoid bodies differentiated from human iPS cells ( $n=9,140$  cells). Cells are coloured according to cell-type cluster. **f**, scHCL results obtained using the data set of day 20 embryoid bodies differentiated from human iPS cells ( $n=9,140$  cells), as in **c**. **g**, Cell–cell correlation network for embryoid body. Each node represents a pseudo-cell in each cell type.



**Extended Data Fig. 10 | Comparison of human and mouse tissues.** **a**, t-SNE analysis of 333,778 single cells from the MCA data, with 104 main cell-type clusters labelled. **b**, t-SNE analysis of 333,778 single cells from the MCA data, with tissue types labelled. **c**, t-SNE map of pseudo-cell 20 data for mouse ( $n=16,740$  pseudo-cells). Pseudo-cells are coloured according to cell-type cluster. **d**, Heat map showing the conserved cell types in human and mouse

adult kidney. AUROC scores were calculated from the Spearman correlation between human and mouse adult kidney pseudo-cells ( $n=1,197$  pseudo-cells). **e**, Heat map showing conserved cell types in human and mouse fetal kidney. AUROC scores were calculated from the Spearman correlation between human and mouse fetal kidney pseudo-cells ( $n=1,459$  pseudo-cells).

# Article



Extended Data Fig. 11 | See next page for caption.

**Extended Data Fig. 11 | Comparison of human and mouse regulons.**

**a, b**, Binary regulon activity t-SNE maps for human and mouse based on 259 human regulons (**a**) and 248 mouse regulons (**b**), created with R package SCENIC. Each dot represents a pseudo-cell of 20 in the HCL or MCA cell clusters. The t-SNE maps were created using binary regulon activity matrices from 17,028 human pseudo-cells and 16,740 mouse pseudo-cells. **c**, Binary RASs for human special regulon BHLHE41 and mouse special regulon Olig1 in the regulon activity t-SNE maps ( $n=17,028$  for human;  $n=16,740$  for mouse). **d**, Binary RASs for human special regulon FOXO1 and mouse special regulon

Dlx5 in the regulon activity t-SNE maps ( $n=17,028$  for human;  $n=16,740$  for mouse). **e**, Binary RASs for human special regulon ZNF230 and mouse special regulon Mlxip1 in the regulon activity t-SNE maps ( $n=17,028$  for human;  $n=16,740$  for mouse). **f**, Binary RASs for regulons IRF8/Irf8 and GATA1/Gata1 in the regulon activity t-SNE maps ( $n=17,028$  for human;  $n=16,740$  for mouse). Note that in the regulation of antigen-presenting endothelial cells, the IRF8/Irf8 regulon is conserved. In the regulation of erythroid cells, the GATA1/Gata1 regulon is conserved.

Corresponding author(s): Guoji Guo

Last updated by author(s): Mar 6, 2020

## Reporting Summary

Nature Research 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 Research policies, see [Authors & Referees](#) 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 HCL data were aligned to the *Homo sapiens* GRCh38 genome 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 2.4.3 in R3.5.0. Scanpy 1.4.3 was used for single cell gene expression analysis, such as lineage trajectory analysis. MetaNeighbor (available at <https://github.com/maggiecrow/MetaNeighbor>) was used to measure the similarity of cell types. Circlize 0.4.4 was used to make circos plot. SCENIC (available at <https://github.com/aertslab/SCENIC>) was used to infer gene regulatory networks. Cytoscape 3.5.0 was used for network visualization. Orthofinder 2.2.6 was used to infer orthologs. Software igraph was used to perform network analysis. SLICE 0.99.0 was used for single cell entropy analysis. Velocity 0.17.13 was used to calculate RNA velocity. Cellphonedb1.0.0 was used to make ligand-receptor analysis. Detailed codes for figures are provided in github (<https://github.com/ggjlab/HCL/>). An online R package is available for scHCL (<https://github.com/ggjlab/scHCL/>).

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [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 list of figures that have associated raw data
- A description of any restrictions on data availability

The CNGB Nucleotide Sequence Archive accession number is CNP0000325.

The GEO accession number is GSE134355.

The human DGE data are available at [https://figshare.com/articles/HCL\\_DGE\\_Data/7235471](https://figshare.com/articles/HCL_DGE_Data/7235471).

The mouse DGE data are available at [https://figshare.com/articles/MCA\\_DGE\\_Data/5435866](https://figshare.com/articles/MCA_DGE_Data/5435866).

Source data for Figs. 2 and 4 and Extended Data Figs. 2, 5, 7, 8, 9, 10, 11 are provided with the paper.

HCL Data can also be accessed on the website: (<http://bis.zju.edu.cn/HCL/>) and a mirror website for international users (<https://db.cngb.org/HCL/>).

## 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](https://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

702,968 single cells were analyzed in total for the first stage Human Cell landscape construction. A total of 67 human tissues and cell culture types were analyzed. In Extended Figure 1d, 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 702,968 single cells.

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

Different single cells were randomly captured before analysis. Human samples were not randomized due to practical constraints.

Blinding

We are blinded to analyzed cell types before single cell analyses.

## 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                                     |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Antibodies            |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology                    |
| <input type="checkbox"/>            | <input type="checkbox"/> Animals and other organisms      |
| <input type="checkbox"/>            | <input type="checkbox"/> Human research participants      |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data                    |

### Methods

- |                                     |   |
|-------------------------------------|---|
| n/a                                 | Involved in the study                           |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

## Antibodies

Antibodies used

Anti-HLA-DR antibody [EPR3692] (ab92511), abcam  
<https://www.abcam.com/hla-dr-antibody-epr3692-ab92511.html>  
CD31 Antibody Mouse Monoclonal| Catalog number: 66065-1-Ig |CloneNo.: 2A1E2, Proteintech  
<https://www.ptgblab.com/products/PECAM1,CD31-Antibody-66065-1-Ig.htm>  
Cytokeratin 17 Monoclonal Antibody (E3), MAS-13539, Thermo Fisher  
<https://www.thermofisher.com/antibody/product/Cytokeratin-17-Antibody-clone-E3-Monoclonal/MAS-13539>  
Anti-CXCL2 (AHP773, Bio-rad)  
<https://www.bio-rad-antibodies.com/polyclonal/human-gro-beta-antibody-ahp773.html?f=purified>  
The dilutions for different antibodies: [anti-HLA-DR (1:200), anti-CD31 (1:100), anti-Krt17 (1:50), and anti-CXCL2 (1:100)]

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)	H9 is from Wicell <a href="https://www.wicell.org/">https://www.wicell.org/</a>
Authentication	The H9 cell line is authenticated by multilineage differentiation experiments and mRNA-seq experiments. It shows that same self-renewal and differentiation abilities, as well as the same gene expression profiles as the other published H9 data.
Mycoplasma contamination	The cell line is negative for mycoplasma contamination.
Commonly misidentified lines (See <a href="#">ICLAC</a> 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. Mouse adult adrenal gland, omentum, pleura and stomach tissues were collected from 8–10 week-old female mice. Mouse fetal pancreas and stomach tissues were collected from E14.5 mouse embryos. 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.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve any samples from the field.
Ethics oversight	Mouse experiments in this study were approved by the Animal Ethics Committee of Zhejiang University; experiments conformed to the 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.

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Chinese Han population was analyzed in the study. The Han population is China's main ethnic group, with a population of about 1.3 billion, accounting for about 19% of the world's total population, and is distributed all around the world. Among 16 fetal tissue donors, 9 donors are male and 7 donors are female; all donors are between 10 weeks to 14 weeks except for 1 donor at 7 weeks and 1 donor at 26 weeks. Among 42 adult tissue donors, 22 donors are male and 20 donors are female; all donors are between 21 years to 66 years old except for one donor > 83 years and one donor < 1 year.
Recruitment	Tissue donors were recruited by collaborating doctors in the affiliated hospitals of Zhejiang University School of Medicine following local protocols. There is no potential self-selection in the volunteer tissue donation process.
Ethics oversight	The human sample collection and research conducted in this study were approved by the Research Ethics Committee of the Zhejiang University School of Medicine, Research Ethics Committee of the First Affiliated Hospital, Research Ethics Committee of the Second Affiliated Hospital and Research Ethics Committee of Women's Hospital at Zhejiang University (Approval Number: 20170029, 20180017, 20190034, 2018015, 2018507, 2018766 and 2018185). Informed consent for fetal tissue collection and research was obtained from each patient after her decision to legally terminate her pregnancy but before the abortive procedure was performed. Informed consent for collection and research of surgically removed adult tissues was obtained from each patient before the operation. Informed consent for the collection and research of tissues from deceased-organ donation was obtained from the donor family after the cardiac death of the donor. Details on donor information are provided in Supplementary Table 1. All the protocols used in this study were in strict compliance with the legal and ethical regulations of Zhejiang University School of Medicine and Affiliated Hospitals. All the protocols used in this study complied with the 'Interim Measures for the Administration of Human Genetic Resources' administered by The Ministry of Science and Technology and The Ministry of Public Health (Approval Number: 2020BAT0007).

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