Supplementary Materials for

A human cell atlas of fetal gene expression

Authors: Junyue Cao[1], Diana R. O'Day[2], Hannah A. Pliner[3], Paul D. Kingsley[4], Mei Deng[2], Riza M. Daza[1], Michael A. Zager[3,6], Kimberly A. Aldinger[2,5], Ronnie Blecher[1], Fan Zhang[7], Malte Spielmann[8,9], James Palis[4], Dan Doherty[2,3,5] , Frank J. Steemers[7], Ian A. Glass[2,3,5], Cole Trapnell[1,3,10,#], Jay Shendure[1,3,10,11,#]

Affiliations:
[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA
[2]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA
[3]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA
[4]Department of Pediatrics, University of Rochester Medical Center, Rochester, NY, USA
[5]Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA, USA
[6]Center for Data Visualization, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[7]Illumina Inc., San Diego, CA, USA
[8]Human Molecular Genomics Group, Max Planck Institute for Molecular Genetics, Berlin, Germany
[9]Institute of Human Genetics, University of Lübeck, Lübeck, Germany
[10]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA
[11]Howard Hughes Medical Institute, Seattle, WA, USA

Correspondence to: coletrap@uw.edu (CT) & shendure@uw.edu (JS)

This PDF file includes:

Materials and Methods

Cell culture

All mammalian cells were cultured at 37°C with 5% CO_2, and were maintained in high glucose DMEM (Gibco cat. no. 11965) supplemented with 10% FBS and 1X Pen/Strep (Gibco cat. no. 15140122; 100U/ml penicillin, 100 µg/ml streptomycin). Cells were trypsinized with 0.25% trypsin-EDTA (Gibco cat. no. 25200-056) and split 1:10 three times per week.

Nuclei isolation and fixation from cell lines

All cell lines were trypsinized, spun down at 300x**g** for 5 min (4°C) and washed once in 1X ice-cold PBS. 5M cells were combined and lysed using 1 mL ice-cold cell lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 0.1% IGEPAL CA-630 from (*192*), modified to also include 1% SUPERase In RNase inhibitor and 1% BSA). The filtered nuclei were then transferred to a new 15 ml tube (Falcon) and pelleted by centrifuge at 500x**g** for 5 min at 4°C and washed once with 1 ml ice-cold cell lysis buffer. The nuclei were fixed in 4 ml ice-cold 4% paraformaldehyde (EMS) for 15 min on ice. After fixation, the nuclei were washed twice in 1 ml nuclei wash buffer (cell lysis buffer without IGEPAL), and re-suspended in 500 ul nuclei wash buffer. The samples were split to 5 tubes with 100 ul in each tube and flash frozen in liquid nitrogen.

Tissue procurement and storage

Human fetal tissues were obtained by the UW Birth Defects Research Laboratory (BDRL) under a protocol approved by the University of Washington Institutional Review Board. Gestational age, reported as the number of weeks post-fertilization, was estimated from fetal foot length. Tissues of interest were isolated and rinsed in 1X HBSS (with Ca2+ and Mg2+) then blotted dry on a semi-damp gauze. Dried tissue was placed on a heavy-duty foil or in cryotube, snap frozen in liquid nitrogen, and then stored at -80°C.

Nuclei isolation and fixation of frozen fetal tissues

On the day of pulverization, we pre-cooled pre-labeled tubes and a hammer on dry ice with a cloth towel between the dry ice and metal. We created a "padding" by taking an 18" x 18" heavy duty foil, folded in half twice creating a rectangle and then folded twice to create a square. The frozen tissue was placed inside the foil "padding" then inside a pre-chilled 4 mm plastic bag to prevent tissue from falling out onto the dry ice in case the foil ruptured. We chilled the tissue packet between two slabs of dry ice. Using the pre-chilled hammer, we manually pulverized the tissue inside the packet with 3 to 5 impacts, avoiding a grinding motion. When necessary, we re-chilled the sample to avoid thawing and repeated the procedure until we generated small, uniform fragments. We then aliquoted the pulverized tissue into the pre-labeled and pre-chilled 1.5 ml LoBind and nuclease-free snap cap 1.5 ml tubes (Eppendorf cat. no. 022431021) and stored at -80°C until further processing. A subset of these aliquots were used for sci-RNA-seq3, and others

for sci-ATAC-seq3, as described in the companion paper.

On the day of nuclei isolation, aliquots of tissue powder (0.1-1g) were first incubated with 1 mL ice-cold cell lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 0.1% IGEPAL CA-630 from (*192*), modified to also include 1% SUPERase In (Thermo Fisher Scientific, AM2696) and 1% BSA (NEB, B9000S)) and then transferred to the top of a 40 µm cell strainer (VWR, 10199-654). Tissues were homogenized through the strainer with the rubber tip of a syringe plunger (VWR, BD309646) in 4 ml cell lysis buffer. The filtered nuclei were then transferred to a new 15 ml tube (VWR, 21008-936) and pelleted by centrifuge at 500x**g** for 5 min and washed once with 1 ml cell lysis buffer. The nuclei were fixed in 5 ml ice-cold 4% paraformaldehyde (EMS, 15-4-100) for 15 min on ice. After fixation, the nuclei were washed twice in 1 ml nuclei wash buffer (cell lysis buffer without IGEPAL), and re-suspended in 500 µl nuclei wash buffer. The samples were split into two tubes with 250 µl in each tube and flash frozen in liquid nitrogen. For human cell extraction in renal and digestive organs (kidney, pancreas, intestine, and stomach) and paraformaldehyde fixation, we followed the procedure described in (*13*).

Immunohistochemistry

Fetal tissues were fixed in formalin and embedded in paraffin. Sections of 4-5 µm thickness were cut and placed on Superfrost Plus slides (12-550-17, FisherBrand). For Immunohistochemistry, sections were subjected to heat mediated antigen retrieval (pH6.0) followed by blocking with normal serum. Primary antibodies were incubated overnight at 4°C. The primary antibody we used: GYPA (R&D, MAB1228, 1:250), CD34 (R&D, AF7227, 1:250), CD34 (Novus, NBP2-32933, 1:250), ANXA1 (R&D, AF3770, 1:500), TNFRS10C (R&D, MAB6301, 1:500), AFP (Novus, NBP1-76275, 1:400), ALB (R&D, MAB1455, 1:10K), AHSG (R&D, AF1184, 1:400), and APOA1 (R&D, MAB36641, 1:250). Species and subtype-appropriate fluorescent dye-labelled secondary antibodies were used (Alexa Fluor 488 and 594, 1:400, Jackson ImmunoResearch Lab) or biotinylated secondary antibody were used followed by ABC Elite Systems (PK-6100, Vector Lab) for DAB chromogen staining.

sci-RNA-seq3 library construction and sequencing

The paraformaldehyde fixed nuclei were processed similarly to the published sci-RNA-seq3 protocol (*11*) with slight modifications. A detailed version of the full sci-RNA-seq3 workflow including estimated hands-on time per step and established stopping points is available at protocols.io (http://dx.doi.org/10.17504/protocols.io.9yih7ue). Briefly, thawed nuclei were permeabilized with 0.2% TritonX-100 (in nuclei wash buffer) for 3 min on ice, and briefly sonicated (Diagenode, 12 sec on low power mode) to reduce nuclei clumping. The nuclei were then washed once with nuclei wash buffer and filtered through 1 ml Flowmi cell strainer (VWR, 10204-924). Filtered nuclei were spun down at 500x**g** for 5 min and resuspended in nuclei wash buffer. Nuclei from each sample were then distributed into several individual wells in four 96-well

plates. The links between well id and tissue id were recorded for downstream data processing. For each well, 80,000 nuclei (16 µL) were mixed with 8 µl of 25 µM anchored oligo-dT primer (5′-/5Phos/CAGAGCNNNNNNNN[10bp barcode]TTTTTTTTTTTTTTTTTTTTTTTTTTTT-3′, where "N" is any base; IDT) and 2 µL 10 mM dNTP mix (Thermo Fisher Scientific, R0192), denatured at 55°C for 5 min and immediately placed on ice. 14 µL of first-strand reaction mix, containing 8 µL 5X Superscript IV First-Strand Buffer (Invitrogen, 18090200), 2 µl 100 mM DTT (Invitrogen, 18090200), 2 µl SuperScript IV reverse transcriptase (200 U/µl, Invitrogen, 18090200), 2 µL RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen, 10777019), was then added to each well. Reverse transcription was carried out by incubating plates by gradient temperature (4°C 2 minutes, 10°C 2 minutes, 20°C 2 minutes, 30°C 2 minutes, 40°C 2 minutes, 50°C 2 minutes and 55°C 10 minutes).

After reverse transcription reaction, 60 µL nuclei dilution buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 1% BSA) was added into each well. Nuclei from all wells were pooled together and spun down at 500x**g** for 10 min. Nuclei were then resuspended in nuclei wash buffer and redistributed into another four 96-well plates with each well including 20 µL Quick ligase buffer (NEB, M2200L), 2 µL Quick DNA ligase (NEB, M2200L), 10 µL nuclei in nuclei wash buffer, 8µL barcoded ligation adaptor (100 uM, 5'- GCTCTG[9 bp or 10 bp barcode A]/dideoxyU/ACGACGCTCTTCCGATCT[reverse complement of barcode A]-3'). The ligation reaction was conducted at 25°C for 10min. After the ligation reaction, 60 µL nuclei dilution buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 1% BSA) was added into each well. Nuclei from all wells were pooled together and spun down at 600x**g** for 10min.

Nuclei were washed once with nuclei wash buffer, filtered with 1 ml Flowmi cell strainer (VWR, 10204-924) once, counted and redistributed into eight 96-well plates with each well including 2,500 nuclei in 5 µL nuclei wash buffer and 3 µL elution buffer (Qiagen, 19086). 1.33 µl mRNA Second Strand Synthesis buffer (NEB, E7550S) and 0.66 µl mRNA Second Strand Synthesis enzyme (NEB, E7550S) were then added to each well, and second strand synthesis was carried out at 16°C for 180 min.

For tagmentation, each well was mixed with 11 µL Nextera™ TD buffer (Illumina) and 1 µL i7 only TDE1 enzyme (62.5 nM, Illumina, diluted in Nextera™ TD buffer (Illumina)), and then incubated at 55°C for 5 min to carry out tagmentation. The reaction was then stopped by adding 24 µL DNA binding buffer (Zymo Research, D4004-1-L) per well and incubating at room temperature for 5 min. Each well was then purified using 1.5x AMPure XP beads (Beckman Coulter, A63882). In the elution step, each well was added with 8 µL nuclease free water, 1 µL 10X USER buffer (NEB, M5505L), 1 µL USER enzyme (NEB, M5505L) and incubated at 37°C for 15 min. Another 6.5 µL elution buffer was added into each well. The AMPure XP beads were removed by magnetic stand and the elution product (16 µL) was transferred into a new 96-well plate.

For PCR amplification, each well (16 μL product) was mixed with 2 μL of 10 μM indexed P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′; IDT), 2 μL of 10 μM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′, IDT), and 20 μL NEBNext High-Fidelity 2X PCR Master Mix (NEB, M0541L). Amplification was carried out using the following program: 72°C for 5 min, 98°C for 30 sec, 12-16 cycles of (98°C for 10 sec, 66°C for 30 sec, 72°C for 1 min) and a final 72°C for 5 min.

After PCR, samples were pooled and purified using 0.8 volumes of AMPure XP beads. Library concentrations were determined by Qubit (Invitrogen, Q32854) and the libraries were visualized by electrophoresis on a 6% TBE-PAGE gel (Invitrogen, EC6265BOX). All libraries were sequenced on an Illumina NovaSeq™ 6000 sequencer (Read 1: 34 cycles, Read 2: 100 cycles, Index 1: 10 cycles, Index 2: 10 cycles).

For paraformaldehyde fixed cells, they were processed similarly to the fixed nuclei with slight modifications: frozen fixed cells were thawed on 37°C water bath, spun down at 500x**g** for 5 min, and incubated with 500ul PBSI (1 x PBS, pH 7.4, 1% BSA, 1% SuperRnaseIn) including 0.2% Triton X-100 for 3min on ice. Cells were pelleted and resuspended in 500ul nuclease free water including 1% SuperRnaseIn. 3ml 0.1N HCl were added into the cells for 5min incubation on ice (*17*). 3.5ml Tris-HCl (pH = 8.0) and 35ul 10% Triton X-100 were added into cells to neutralize HCl. Cells were pelleted and washed with 1ml PBSR. Cells were pelleted and resuspended in 100ul PBSI. The following steps were similar with the above sci-RNA-seq3 protocol (with paraformaldehyde fixed nuclei) with slight modifications: (1) We distributed 20,000 fixed cells (instead of 80,000 nuclei) per well for reverse transcription. (2) We replaced all nuclei wash buffer in following steps with PBSI. (3) All nuclei dilution buffer were replaced with PBS + 1% BSA.

Processing of sequencing reads

Read alignment and gene count matrix generation for the single cell RNA-seq was performed using the pipeline that we developed for sci-RNA-seq3 (*11*) with minor modifications: base calls were converted to fastq format using Illumina's bcl2fastq/v2.16 and demultiplexed based on PCR i5 and i7 barcodes using maximum likelihood demultiplexing package deML (*193*) with default settings. Downstream sequence processing and single cell digital expression matrix generation were similar to sci-RNA-seq (*18*) except that RT index was combined with hairpin adaptor index, and thus the mapped reads were split into constituent cellular indices by demultiplexing reads using both the RT index and ligation index (ED < 2, including insertions and deletions). Briefly, demultiplexed reads were filtered based on RT index and ligation index (ED < 2, including insertions and deletions) and adaptor clipped using trim_galore/v0.4.1 with default settings. Trimmed reads were mapped to the human reference genome (hg19) for human fetal nuclei, or a

chimeric reference genome of human hg19 and mouse mm10 for HEK293T and NIH/3T3 mixed nuclei, using STAR/v 2.5.2b (*194*) with default settings and gene annotations (GENCODE V19 for human; GENCODE VM11 for mouse). Uniquely mapping reads were extracted, and duplicates were removed using the unique molecular identifier (UMI) sequence (ED < 2, including insertions and deletions), reverse transcription (RT) index, hairpin ligation adaptor index and read 2 end-coordinate (*i.e.* reads with UMI sequence less than 2 edit distance, RT index, ligation adaptor index and tagmentation site were considered duplicates). Finally, mapped reads were split into constituent cellular indices by further demultiplexing reads using the RT index and ligation hairpin (ED < 2, including insertions and deletions). For mixed-species experiment, the percentage of uniquely mapping reads for genomes of each species was calculated. Cells with over 85% of UMIs assigned to one species were regarded as species-specific, with the remaining cells classified as mixed cells or "collisions". To generate digital expression matrices, we calculated the number of strand-specific UMIs for each cell mapping to the exonic and intronic regions of each gene with python/v2.7.13 HTseq package (*195*). For multi-mapped reads, reads were assigned to the closest gene, except in cases where another intersected gene fell within 100 bp to the end of the closest gene, in which case the read was discarded. For most analyses we included both expected-strand intronic and exonic UMIs in per-gene single-cell expression matrices.

After the single cell gene count matrix was generated, cells with fewer than 250 UMIs were filtered out. Each cell was assigned to its original human fetal sample based on the RT barcode. Reads mapping to each fetus individual were aggregated to generate "pseudobulk RNA-seq" datasets. For sex assignments, we counted reads mapping to female-specific non-coding RNA (*TSIX* and *XIST*) or chrY genes (except genes *TBL1Y, RP11-424G14.1, NLGN4Y, AC010084.1, CD24P4, PCDH11Y, and TTTY14*, which are detected in both males and females). Fetuses were readily separated into females (more reads mapping to *TSIX* and *XIST* than chrY genes) and males (more reads mapping to chrY genes than *TSIX* and *XIST*).

Clustering analysis of pseudobulk transcriptomes was done with Monocle 3/alpha (*11*). Briefly, an aggregated gene expression matrix was constructed as described above for human fetal organs from each individual. Samples with over 5,000 total UMIs were selected. The dimensionality of the data was reduced by PCA (10 components), first on the top 500 most highly dispersed genes and then with UMAP (max_components = 2, n_neighbors = 10, min_dist = 0.5, metric = 'cosine').

Cell filtering, clustering and marker gene identification

For the detection of potential doublet cells, we first split the dataset into subsets for each organ and individual, and then applied the scrublet/v0.1 pipeline (*189*) to each subset with parameters (min_count = 3, min_cells = 3, vscore_percentile = 85, n_pc = 30, expected_doublet_rate = 0.06, sim_doublet_ratio = 2, n_neighbors = 30, scaling_method = 'log') for doublet score calculation. Cells with doublet score over 0.2 were annotated as detected doublets. We detected 6.4% potential doublet cells in the whole data set, which corresponds to an overall estimated doublet rate of 12.6%

(including both within- and between-cluster doublets).

For detection of doublet-derived subclusters for cells from each organ, we used an iterative clustering strategy as shown before (*11*). Briefly, gene count mapping to sex chromosomes were removed before clustering and dimensionality reduction. Preprocessing steps were similar to the approach used by ref (*196*). Briefly, genes with no count were filtered out and each cell was normalized by the total UMI count per cell. The top 1,000 genes with the highest variance were selected and the digital gene expression matrix was renormalized after gene filtering. The data was log transformed after adding a pseudocount, and scaled to unit variance and zero mean. The dimensionality of the data was reduced by PCA (30 components) first and then with UMAP, followed by Louvain clustering performed on the 30 principal components with default parameters. For Louvain clustering, we first fitted the top 30 PCs to compute a neighborhood graph of observations with local neighborhood number of 50 by scanpy.api.pp.neighbors function in scanpy/v1.0 (*197*). We then cluster the cells into sub-groups using the Louvain algorithm implemented as scanpy.api.tl.louvain function (*197*). For UMAP visualization, we directly fit the PCA matrix into scanpy.api.tl.umap function(*197*) with min_distance of 0.1. For subcluster identification, we selected cells in each major cell type and applied PCA, UMAP, Louvain clustering similarly to the major cluster analysis. Subclusters with a detected doublet ratio (by Scrublet) over 15% were annotated as doublet-derived subclusters.

For data visualization, cells labeled as doublets (by Scrublet) or from doublet-derived subclusters were filtered out. For each cell, we only retain protein-coding genes, lincRNA genes and pseudogenes. Genes expressed in less than 10 cells and cells expressing less than 100 genes were further filtered out. The downstream dimension reduction and clustering analysis were done by Monocle 3/alpha (*11*). The dimensionality of the data was reduced by PCA (50 components) first on the top 5,000 most highly dispersed genes and then with UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'). Cell clusters were identified using the Louvain algorithm implemented in Monocle 3 (louvain_res = 1e-04). Clusters were assigned to known cell types based on cell type-specific markers (Table S3). We found the above Scrublet and iterative clustering based approach is limited in marking cell doublets between abundant cell clusters and rare cell clusters (e.g. less than 1% of total cell population). To further remove such doublet cells, we took the cell clusters identified by Monocle 3 and first computed differentially expressed genes across cell clusters (within-organ) with the differentialGeneTest() function of Monocle 3. We then selected a gene set combining the top ten gene markers for each cell cluster (ordered by q-value and fold expression difference between first and second ranked cell cluster). Cells from each main cell cluster were selected for dimension reduction by PCA (10 components) first on the selected gene set of top cluster specific gene markers, and then by UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'), followed by clustering identification using the density peak clustering algorithm implemented in Monocle 3 (rho_thresh = 5, delta_thresh = 0.2 for most clustering analysis). Subclusters showing low expression of target cell cluster specific

markers and enriched expression of non-target cell cluster specific markers were annotated as doublets derived subclusters and filtered out in visualization and downstream analysis. Differentially expressed genes across cell types (within-organ) were re-computed with the differentialGeneTest() function of Monocle 3 after removing all doublets or cells from doublet-derived subclusters.

Adjudication of the 15 initially unannotated cell types

As noted in the main text, our first round of annotation was performed on a tissue-by-tissue basis by comparing observed cell types to those expected from prior knowledge of the same tissue. In general, we recovered all or nearly all main cell types identified by previous atlasing efforts directed at the same organs, despite differences with respect to species, stage of development and/or technology. In addition, we identified 15 cell types that we did not at least initially expect to observe in a given tissue. We labeled these based on the top enriched differentially expressed gene markers within that tissue, *e.g.* CSH1_CSH2 positive cells. Unsurprisingly, the initially unannotated cell types were rarer than annotated cell types (median 0.5% vs. median 2.2%, respectively, of all profiled cells within an organ), and a few exhibited low specificity scores upon intra-dataset cross-validation with an SVM classifier.

Subsequent to the initial round of annotation, we reexamined these 15 cell types based on their distribution in the global UMAP, whether they matched annotated cell types in mouse atlases, their distribution across tissues derived from different individuals, and their potential for maternal origin. Our updated interpretations are summarized below. We have grouped them into 8 to which we have assigned preliminary annotations based on these additional analyses, 4 that would be better characterized as subtypes of other cell types, and 3 that have high specificity scores but remain ambiguous. SS = within-tissue specificity score.

Annotated based on additional analyses and/or review

AFP_ALB positive cells (placenta: 0.29% of cells, SS 0.99; spleen: 0.71% of cells, SS 0.98). These cells are highly correlated and cluster with hepatoblasts (**Fig. 3A**; expressing high levels of serum albumin, alpha fetoprotein, and apolipoproteins), which are of course expected in the liver but unexpected in the placenta and spleen. Of note, although observed in at least two samples of each organ, their abundance was highly variable. Furthermore, at least in the placenta, similar hepatoblast-like, *AFP*+, *ALB*+ cells were observed in the mouse and were matched to these in cell type correlation analysis (**Fig. S13A**). Finally, followup immunostaining studies confirmed the presence of AFP-positive cells within human fetal splenic tissue (**Fig. 3D**). We believe that these cells correspond to hepatoblasts that are potentially circulating.

CSH1_CSH2 positive cells (adrenal: 0.024%, SS 0.87; lung: 0.0073%, SS 0.77): These cells express high levels of placental lactogen, chorionic gonadotropin, and aromatase, and in the global

analysis, are closely related to trophoblasts (**Fig. 3A**). We infer they correspond to trophoblasts that have entered fetal circulation and are present in sufficient numbers in at least the fetal adrenal gland and fetal lung, which were two of the most deeply sampled organs, to cluster independently of other cell types. Of note, although observed in at least two samples of each organ, their abundance was highly variable. Finally, followup immunostaining studies confirmed ANXA1-positive cells (these cells are also marked by high levels of *ANXA1* expression) in human fetal adrenal tissue, consistent with these cells corresponding to circulating trophoblasts (**Fig. 3C**).

IGFBP1_DKK1 positive cells (2.3% of cells in placenta, SS 1.0): In males, *IGFBP1*+, *DKK1*+ cells in the placenta expressed appreciable levels of *XIST* or *TSIX* (**Fig. 12B**). In the companion single cell atlas of chromatin accessibility, a corresponding placental cluster was identified, and shown by two methods to be dominated by maternally derived cells (*12*). Together with the fact that these cells are most strongly marked by expression of insulin-like growth factor binding protein-1 (*IGFBP1*), we conclude that these likely correspond to maternally-derived decidualized stromal cells (*35*).

PAEP_MECOM positive cells (0.96% of cells in placenta, SS 0.99): In males, *PAEP*+, *MECOM*+ cells in the placenta expressed appreciable levels of *XIST* or *TSIX* (**Fig. 12B**). In the companion single cell atlas of chromatin accessibility, a corresponding placental cluster was identified, and shown by two methods to be dominated by maternally derived cells (*12*). On cell type correlation analysis with the corresponding mouse tissue, they are strongly matched to labyrinth trophoblasts (**Fig. S13B**). Together with the fact that these cells are most strongly marked by expression of glycodelin (*PAEP*), we conclude that these likely correspond to maternally-derived endometrial epithelial cells (*35*).

PDE11A_FAM19A2 positive cells (1.6% of cells in eye, SS 0.75): These cells specifically express a phosphodiesterase, encoded by *PDE11A*, a gene marking the non-pigmented epithelial cells in the ciliary body of the eye (*198*). Although rare, these cells were observed in all eye tissues sampled. On cell type correlation analysis with mouse retinal tissue, they are strongly matched to a cell type annotated by MCA as *RIMS*+ amacrine cells (**Fig. S13C**).

MUC13_DMBT1 positive cells (6.5% of cells in stomach, SS 0.95): In global analyses, these cells are highly correlated with intestinal epithelial cells (**Fig. 3A**) and were observed in all stomach tissues sampled. On cell type correlation analysis with mouse stomach tissue, they are strongly matched to a cell type annotated by MCA as tuft cells (**Fig. S13D**).

CCL19_CCL21 positive cells (0.45% of cells in pancreas, SS 1.0): These cells are highly correlated with mesenchymal cells from other tissues (**Fig. 3A**). They specifically express *FDCSP,* which encodes follicular dendritic cell secreted protein, suggesting these cells may be follicular dendritic cells, which are thought to be of mesenchymal origin (*199*). These cells were observed

in all pancreatic tissues sampled.

STC2_TLX1 positive cells (34% of cells in spleen, SS 0.94): These cells are amongst the most common cell type observed in the spleen, and specifically express the glycoprotein *STC2*, as well as the transcription factors *TLX1* and *NKX2-3*, all associated with mesenchymal precursor or stem cells (*103–105*). These cells were observed in all splenic tissues sampled.

Low specificity scores

CLC_IL5RA positive cells (0.10% of cells in heart, SS 0.07): The cells are highly correlated with cardiomyocytes (thus their low specificity score) but express a distinct program that may reflect a specialized role, with specific expression of immune cell-related receptors, including interleukin 5 receptor Subunit Alpha (*IL5RA*) and hematopoietic-specific transmembrane protein 4 (*MS4A3*). These cells were observed in all hearts sampled. Of note, one of the top differentially expressed genes in this cell type encodes oncostatin M (*OSM*), which has reported to mediate cardiomyocyte dedifferentiation and remodeling during both acute and chronic cardiac diseases (*200*)

ELF3_AGBL2 positive cells (0.16% of cells in heart, SS 0.19): The cells are highly correlated with cardiomyocytes (thus their low specificity score) but express a distinct program that may reflect a specialized role, with specific expression of many genes associated with pulmonary alveolar surfactant secreting cells, including pulmonary secretory protein 1 (*SCGB3A2*), pulmonary surfactant-associated protein B (*SFTPB*) and pulmonary surfactant-associated protein C (*SFTPC*). These cells were observed in all hearts sampled.

PDE1C_ACSM3 positive cells (0.64% of cells in stomach, SS 0.75): The cells are highly correlated with goblet cells in the stomach (thus their modest specificity score), but express a distinct program that may reflect a specialized role, with specific expression of *NOX4*, *PDE1C*, and *ACSM3*. These cells were observed in all stomach tissues sampled.

SLC26A4_PAEP positive cells (0.018% of cells in adrenal, SS 0.0): In both the tissue-specific and global analyses (**Fig. 3A**), these cells are most closely related to adrenocortical cells, and their very low specificity score follows from that together with their rarity. As compared to other cells in the adrenal gland, these cells specifically expressed pendrin (*SLC26A4*) and a sodium-independent chloride-iodide exchanger, glycodelin (*PAEP*). Glycodelin produced by the maternal endometrium is a key regulator of fetomaternal tolerance during pregnancy (*201*). A fetal source of glycodelin is potentially interesting. These cells were observed in all adrenal glands sampled.

High specificity score but ambiguous

SLC24A4_PEX5L positive cells (1.8% of cells in cerebellum, SS 1.0): These cells express high

levels of GABAergic neuron markers GAD1 and GAD2 (*202*), as well as glutamate NMDA receptor subunit 3A (GluN3A, encoded by *GRIN3A*), a gene specifically expressed in a subtype of interneurons of the cerebellum (*203*). Notably, in the global analysis these cells cluster separately from all other neurons. These cells were observed in all cerebellar tissues sampled.

SATB2_LRRC7 positive cells (0.31% of cells in heart, SS 1.0): These cells overwhelmingly derive from a single heart sample and are strongly correlated with CNS excitatory neurons (**Fig. 3A**). Although this most likely reflects contamination of that sample with a fragment of brain during dissection, it is notable that other neuronal cell types were not identified within the same tissue sample.

SKOR2_NPSR1 positive cells (0.86% of cells in cerebrum, SS 1.0): In the global analysis, these cells are highly correlated with Purkinje neurons from the cerebellum, which are not to our knowledge expected in the cerebrum. These cells are not uniformly distributed across the 8 cerebral tissues sampled, but are clearly observed in 5 of these tissue samples. This could reflect contamination, but of note we did not observe other cerebellar cell types in those same tissue samples.

Clustering analysis of cells across organs

For clustering analysis of 77 main cell types across 15 organs, we sampled 5,000 cells from each cell type (or all cells for cell types with fewer than 5,000 cells in a given organ). The dimensionality of the data was reduced first by PCA (50 components) on the gene set combining top cell type-specific gene markers identified above (**Table S5**, qval = 0) and then with UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'). Differentially expressed genes across cell types were identified with the differentialGeneTest() function of Monocle 3. For annotating cell type-specific gene features, we intersected the cell type-specific genes identified above with the predicted secreted and membrane protein coding gene sets from the Human Protein Atlas (*190*), as well as the TF set annotated in the "motifAnnotations_hgnc" data from package RcisTarget/v1.2.1 (*191*).

For clustering analysis of blood cell across 15 organs, we extracted all blood cells corresponding to annotated clusters of myeloid cells, lymphoid cells, thymocytes, megakaryocytes, microglia, antigen presenting cells, erythroblasts, and hematopoietic stem/progenitor cells. The dimensionality of the data was reduced first by PCA (40 components) on the expression of a gene set combining the top 3,000 blood cell type-specific gene markers (**Table S5**, only genes specifically expressed in at least one blood cell type were selected (q-value < 0.05, fold expression difference between first and second ranked cell cluster > 2) and ordered by median qval across organs) and then with UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'). Cell clusters were identified using the Louvain algorithm implemented in Monocle 3 (louvain_res = 1e-04). Clusters were assigned to known cell types based on cell type-specific

markers. We then co-embedded the human fetal blood cells and a scRNA-seq atlas of blood cells from the fetal liver (*108*), using the Seurat v3 integration method (FindAnchors and IntegrateData) (*15*) with a chosen dimensionality of 30 on the top 3,000 highly variable genes with shared gene names in both datasets.

We then applied a similar analysis strategy as above for clustering analysis of endothelial or epithelial cells across organs. For endothelial cells, we first extracted cells corresponding to annotated clusters of vascular endothelial cells, lymphatic endothelial cells and endocardial cells across organs. The dimensionality of the data was reduced first by PCA (30 components) on the gene set combining top 1,000 endothelial cell type-specific gene markers identified above (Table S5, only genes specifically expressed in at least one endothelial cell type were selected (q-value < 0.05, fold expression difference between first and second ranked cell cluster > 2) and ordered by median qval across organs) and then with UMAP with the same parameters used for blood cells. Cell clusters were identified using the Louvain algorithm implemented in Monocle 3 (louvain_res = 1e-04), and then annotated based on the tissue origin of endothelial cells. For epithelial cells, we first extracted cells from the epithelial cell cluster in **Fig. S4B**, followed by dimension reduction first by PCA (50 components) first on the top 5,000 most highly dispersed genes and then with UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'). For validating the tissue specific endothelial cells, we then co-embedded the human fetal endothelial cells and a scRNA-seq atlas of endothelial cells from mouse adult tissues (*139*), using the Seurat v3 integration method (FindAnchors and IntegrateData) (*15*) with a chosen dimensionality of 30 on the top 3,000 highly variable genes with shared gene names in both datasets.

Intra-dataset cross-validation analysis

For cells from each organ, we randomly sampled up to 2,000 cells from each main cell type. We then followed the same process (*101*). Briefly, we combined all sampled cells from each organ and evaluated cell type specificity by applying a 5-fold cross-validation to the dataset, with a support vector machine (SVM) classifier (with linear kernel). Whole transcriptomes were used in cell type prediction. We then computed the cross-validation F-1 value as cell type specificity score. As control, we randomly permuted the cell type labels, followed by the same analysis pipeline. For cell type specificity analysis across all organs, we applied the same analysis strategy to the full dataset after sampling up to 2,000 cells of each main cell type.

Sub-clustering analysis

For each main cell type (with over 1,000 cells) in each organ, we applied Harmony/v1.0 for batch correction and dimension reduction (*102*). Briefly, the dimensionality of the data was reduced by PCA (30 components, or 10 components for cell types with less than 5,000 cells) first on the top 3,000 (or 1,000 for cell types with less than 5,000 cells) most highly variable genes, followed by batch correction on sample ID. Cell clusters were identified using the Louvain algorithm implemented in Seurat/v3.1.4 (*15*) (resolution = 0.5). We then applied the intra-dataset cross-

validation approach to evaluate the specificity of sub-clusters within each main cell type. For every sub-cluster pair, A and B, we computed the number of A cells mislabeled as B cells in cross-validation analysis with the true dataset (mislabeled cell number: n) or the permuted dataset (mislabeled cell number: m). A large n value suggests the two sub-clusters are not well separated by the full transcriptome. We thus iteratively merged similar sub-cluster pairs (n > m), and identified a total of 657 subtypes across 15 organs. The intra-dataset cross validation approach was applied to evaluating subtype specificity within each main cell type in each organ. To annotate the identity of subtypes, we applied the same cell type correlation analysis strategy described in (*11*) to compare cell subtypes from this study to cell types of the same organ from the Microwell-seq based Mouse Cell Atlas (MCA) (*16*). A similar comparison was performed for all subtypes from the brain against cell types annotated in a recent mouse brain atlas (MBCA) (*50*).

Validating erythropoiesis in the adrenal tissues from newborn mice

Adrenals and kidneys were harvested from CD1 Swiss albino mice (Charles River) on the day of birth (P0), and bone marrow cells were flushed from the femurs of the dams. Solid tissues were dissociated using collagenase and stained for imaging flow cytometry using the markers Ter119 (AF488), CD117 (PE-CF594), CD71 (PE), CD45 (EF450), and DRAQ5. Gating of maturing erythroblast populations was performed using published methods (*128*) and analyzed with IDEAS (Luminex) software.

Comparison of human and mouse developmental atlases

We first applied a slightly modified version of the strategy described in (*11*) to identify correlated cell types between this human fetal cell atlas and the mouse organogenesis cell atlas (MOCA) (*11*). We first aggregated cell type-specific UMI counts, normalized by the total count, multiplied by 100,000, and log-transformed after adding a pseudo-count. We then applied non-negative least squares (NNLS) regression to predict the gene expression of target cell type (T_a) in dataset A with the gene expression of all cell types (M_b) in dataset B:

$$T_a = \beta_{0a} + \beta_{1a} M_b$$

where T_a and M_b represent filtered gene expression for target cell type from data set A and all cell types from data set B, respectively. To improve accuracy and specificity, we selected cell type-specific genes for each target cell type by: 1) ranking genes based on the expression fold-change between the target cell type vs. the median expression across all cell types, and then selecting the top 200 genes. 2) ranking genes based on the expression fold-change between the target cell type vs. the cell type with maximum expression among all other cell types, and then selecting the top 200 genes. 3) merging the gene lists from step (1) and (2). β_{1a} is the correlation coefficient computed by NNLS regression.

Similarly, we then switch the order of datasets A and B, and predict the gene expression of target

cell type (T_b) in dataset B with the gene expression of all cell types (M_a) in dataset A:

$$T_b = \beta_{0b} + \beta_{1b} M_a$$

Thus, each cell type a in dataset A and each cell type b in dataset B are linked by two correlation coefficients from the above analysis: β_{ab} for predicting cell type a using b, and β_{ba} for predicting cell type b using a. We combine the two values by:

$$\beta = \beta_{ab} + \beta_{ba}$$

and find β reflects the matching of cell types between two data sets with high specificity. For each cell type in dataset A, all cell types in dataset B are ranked by β and the top cell type (with β > 0.06) is identified as the matched cell type. We compared all human cell types from this study to 10 main cell trajectories and 56 sub-trajectories from the mouse embryonic cell atlas (MOCA) (*11*).

As a different approach, we co-embedded the human fetal cell atlas and the mouse organogenesis cell atlas (MOCA) (*11*) using the Seurat v3 integration method (FindAnchors and IntegrateData) (*15*) with a chosen dimensionality of 30 on the top 3,000 highly variable genes with shared gene names in both human and mouse. We first integrated 65,000 human fetal cells (up to 1,000 cells randomly sampled from each of 77 cell types) and 100,000 mouse embryonic cells (randomly sampled from MOCA) with default parameters. We then applied the same integrative analysis strategy to extracted human and mouse cells from the hematopoietic, endothelial and epithelial trajectories.

For the co-embedded human and mouse hematopoietic cells, we annotated each mouse cell based on its k nearest neighbours of human cells. We chose a small k value (k = 3) such that rare cell types were also annotated. Differentially expressed genes across mouse hematopoietic cells were computed with the differentialGeneTest() function of Monocle 3/alpha.

Pseudotemporal ordering of mouse macrophage/microglia cells was done with Monocle 3/alpha with the reduction method of "DDRTree". Briefly, the top 3 principal components on the top 500 highly variable genes were used to construct the DDRTree pseudotime trajectory with UMI number per cell as a covariate (param.gamma = 120, norm_method = "log", residualModelFormulaStr = "~ sm.ns(Total_mRNAs, df = 3)"). The cells are separated into three branch trajectories in the DDRTree space. Differentially expressed genes across the three branches were computed with the differentialGeneTest() function of Monocle 3/alpha. We then clustered cells with k means clustering (k = 10) and computed the average development time for each cluster. The progenitor cell group was annotated based on the lowest average development time and appeared at the center of the three branches. Each cell was assigned a pseudotime value based on

its distance from the progenitor cells.

Using the Garnett models trained on this human cell atlas for cell type classification

The R package Garnett for Monocle 3 (version 0.2.9) was used to generate cell type classifiers for each of the 15 tissues. Marker genes for each cell type were assembled from literature searches by author HAP, and models were trained using train_cell_classifier using default parameters and the gene database org.Hs.eg.db (version 3.10.0). Models were trained on the entirety of each tissue dataset with the exception of cerebrum, where 100,000 cells were randomly sampled for training for computational efficiency. To compare cell type assignments to those obtained via manual annotation by author JC (*i.e.* the 77 main cell types), we applied the function classifiy_cells using the trained models with the following non-default parameters: cluster_extend = TRUE, cluster_extend_max_frac_incorrect = 0.25, cluster_extend_max_frac_unknown = 0.95. Garnett cell type assignments that matched the cell type assignment from manual annotation were considered 'correct' with the following exceptions: Garnett classification of "Chromaffin cells" was considered correct when manual annotation was "Sympathoblasts", Garnett classification of "B cells" or "T cells" was considered correct when manual annotation was "Lymphoid cells", Garnett classification of "Cap mesenchyme cells", "Collecting duct cells", "Distal tubule cells", "Loop of Henle cells", "Proximal tubule cells", and "Podocytes" were considered correct when manual annotation was "Metanephric cells", Garnett classification of "Ureter cells" and "Collecting duct cells" were considered correct when manual annotation was "Ureteric bud cells", Garnett classification of "Pancreatic Alpha cells", "Pancreatic Beta cells", and "Pancreatic Delta cells" was considered correct when manual annotation was "Islet endocrine cells", Garnett classification of "D cells" was considered correct with manual annotation of "Neuroendocrine cells".

To test the applicability of Garnett trained models to future data, we applied the pancreas model to human adult pancreas scRNA-seq data from reference (*100*). The model was applied using the function classify_cells with the same parameters as above. When compared cell type assignments to those provided by the authors, we considered the following cell types to be equivalent: acinar, Acinar cells; ductal, Ductal cells; endothelial, Endothelial cells; mast, Myeloid cells; macrophage, Myeloid cells; schwann, Glia; alpha, Pancreatic Alpha cells; beta, Pancreatic Beta cells; delta, Pancreatic Delta cells; activated_stellate, Pancreatic stellate cells; quiescent_stellate, Pancreatic stellate cells; t_cell, T cells.

Code availability

Scripts for processing sci-RNA-seq3 sequencing were written in python and R with code available at https://doi.org/10.5281/zenodo.4013713. Most trajectory analysis was done with Monocle3 with setup instructions and tutorial available at http://cole-trapnell-lab.github.io/monocle-release/monocle3.

Supplementary Figures

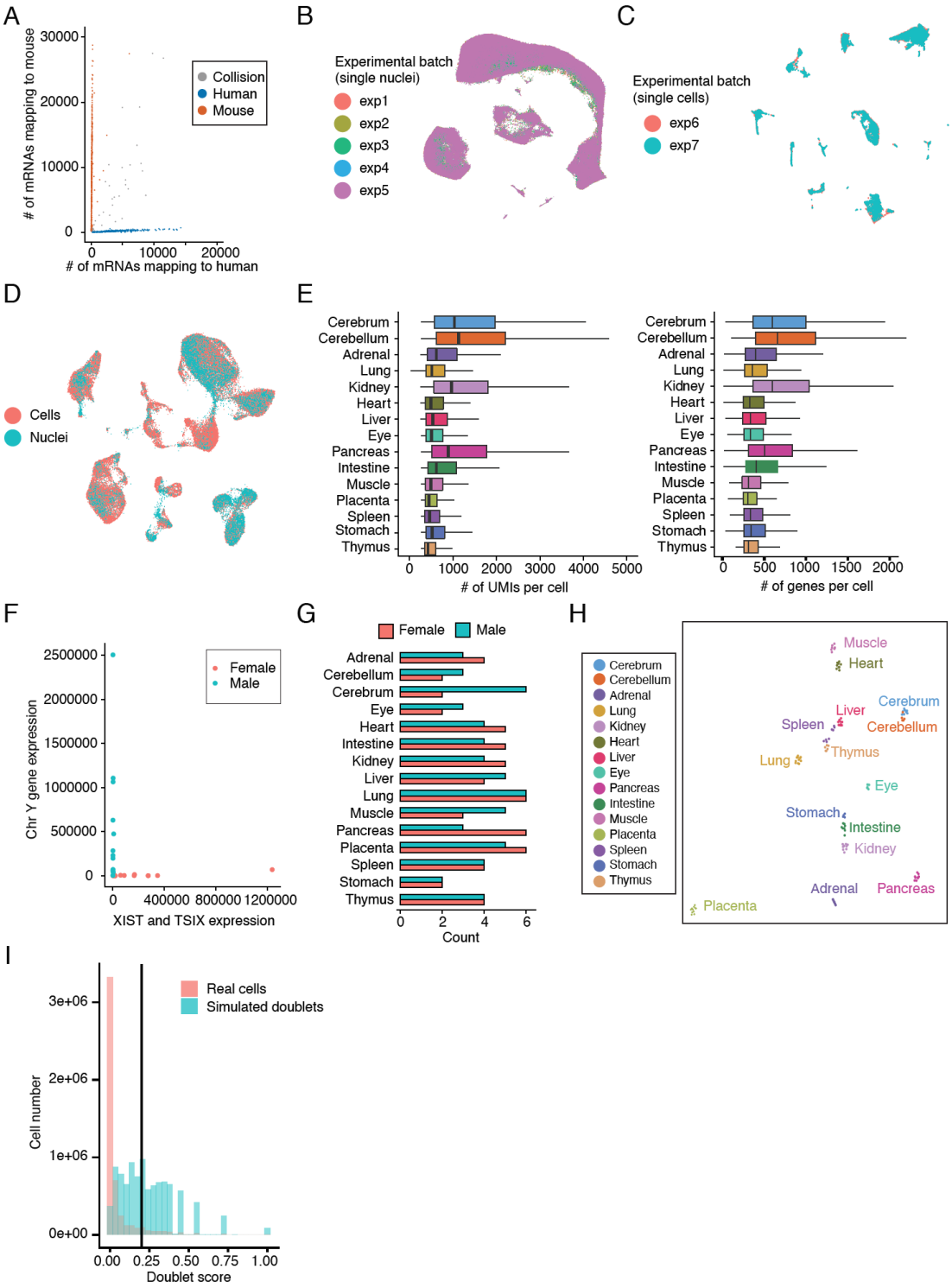


Fig. S1. Performance and quality control related analyses. (**A**) "Barnyard" scatter plot showing the number of mRNAs mapping to the human vs. mouse genome for single cell profiles derived

from a mixture of human HEK293T and mouse NIH/3T3 nuclei. (**B-C**) UMAP visualization of single cell profiles derived from "sentinel" tissues included as batch controls in each experiment, including 5 batches of nuclei (B) and 2 batches of cells (C). (**D**) UMAP visualization of both nuclei and cells sampled from the same tissue (pancreas), integrated with Seurat (*15*). (**E**) Boxplot showing the number of UMIs (left) and genes (right) recovered per cell for each organ. (**F**) Scatter plot showing the number of reads mapping to genes on the male-specific Y chromosome vs. female-specific *XIST* and *TSIX* transcripts per fetus. The sex of each fetus was inferred from this analysis. (**G**) Barplot showing the distribution of sexes for tissue samples corresponding to each organ. (**H**) UMAP visualization of the aggregated transcriptomes of single cells from each of 117 tissue samples, colored by organ. Four samples with fewer than 5,000 total UMIs were excluded from this analysis. (**I**) Histogram showing distribution of doublet scores for profiled cells (red) and simulated doublets (blue) using Scrublet (*20*). Black vertical line corresponds to a threshold of 0.20, above which cells were called as doublets and filtered out.

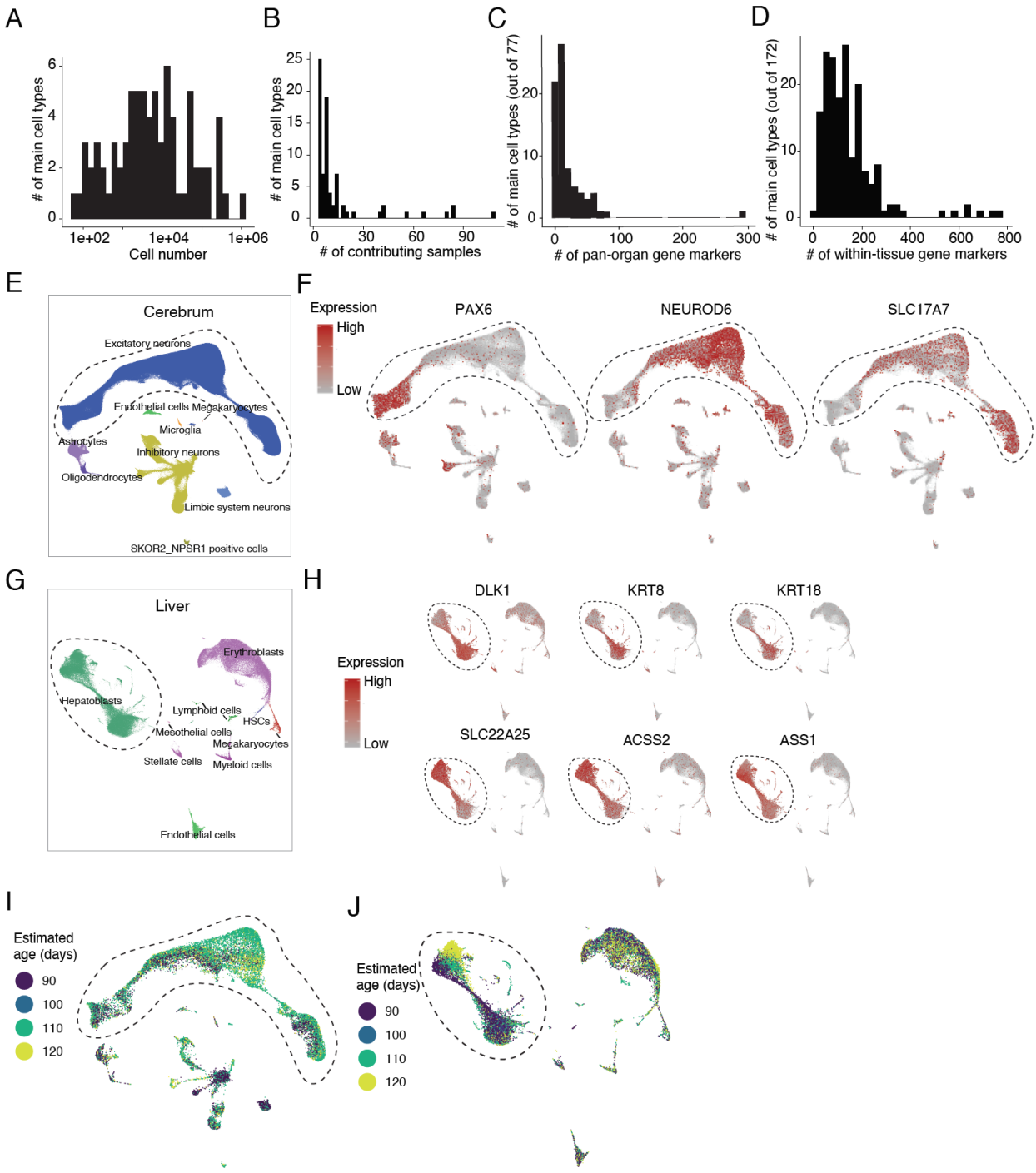
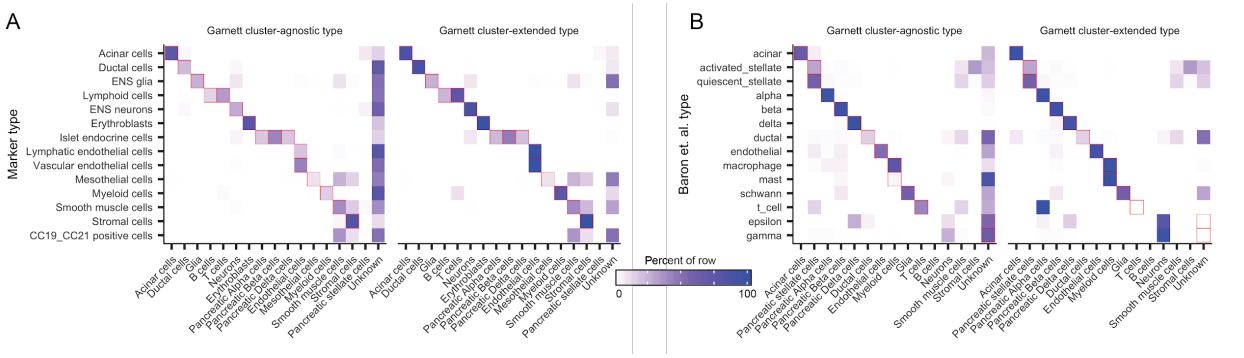


Fig. S2. Identifying main cell types and corresponding gene markers. (**A-D**) Histograms showing the distribution of the number of cells (A), contributing samples (B) and marker genes (C) for each of the 77 main cell types. For gene marker identification, we performed differential gene analysis to identify cell type-specific gene markers for each of the 77 main cell types (FDR of 5% and requiring at least a 5-fold expression difference between first and second ranked cell

type). (**D**) Histogram showing the distribution of the number of marker genes assessed on an organ-by-organ basis for 172 cell types (FDR of 5% and requiring at least a 5-fold expression difference between first and second ranked cell type within a given organ). (**E-H**) UMAP of profiled cerebrum (E, F) and liver (G, H) cells, colored by cell types as in corresponding panels of **Fig. 1C** (E, G) or normalized gene expression of selected genes (F, H). For gene expression, UMI counts for each gene were scaled for library size factor, log-transformed, and then mapped to Z-scores. (**I-J**) UMAP of profiled cerebrum (I) and liver (J) cells, colored by estimated post-conceptual age (rounded). We do not observe a clear correspondence between the differentiation trajectories and estimated post-conceptual age.



C	Cluster-agnostic			Cluster-extended		
	Correct	Incorrect	Unclassified	Correct	Incorrect	Unclassified
Adrenal	0.33	0.04	0.64	0.79	0.02	0.19
Cerebellum	0.44	0.07	0.49	0.89	0.03	0.08
Cerebrum	0.23	0.10	0.67	0.60	0.07	0.30
Eye	0.30	0.07	0.64	0.70	0.03	0.27
Heart	0.40	0.13	0.47	0.64	0.07	0.30
Intestine	0.37	0.05	0.58	0.90	0.02	0.09
Kidney	0.41	0.05	0.53	0.68	0.04	0.28
Liver	0.41	0.08	0.52	0.69	0.06	0.25
Lung	0.43	0.06	0.51	0.87	0.02	0.11
Muscle	0.32	0.03	0.65	0.79	0.02	0.19
Pancreas	0.58	0.07	0.35	0.88	0.05	0.07
Placenta	0.31	0.08	0.61	0.73	0.05	0.22
Spleen	0.32	0.35	0.33	0.50	0.30	0.20
Stomach	0.42	0.09	0.50	0.66	0.08	0.26
Thymus	0.26	0.08	0.67	0.79	0.04	0.17

Fig. S3. Semi-supervised classification with Garnett agrees with manual annotations and can be used for automatic classification of other datasets. Garnett classifiers were generated for each tissue using literature-derived markers (*99*). (**A**) A heatmap comparing manual classification (rows) to cluster-agnostic (left) and cluster-extended (right) cell type assignments by Garnett in pancreas. Color represents the percentage of cells with a certain manual annotation that were labeled as each type by Garnett. (**B**) Results of applying the classifier trained on the data in (A) to an independent adult human pancreas scRNA-seq data from (*100*). Similar to (A), but comparing cell type assignments from (*100*) (rows) to cell type assignments by Garnett (columns). (**C**) Table showing the fraction of Garnett organ-specific classifications that agree with manual classifications. Correct indicates that the strategies agree, incorrect that they do not agree, and unclassified that Garnett did not provide a classification.

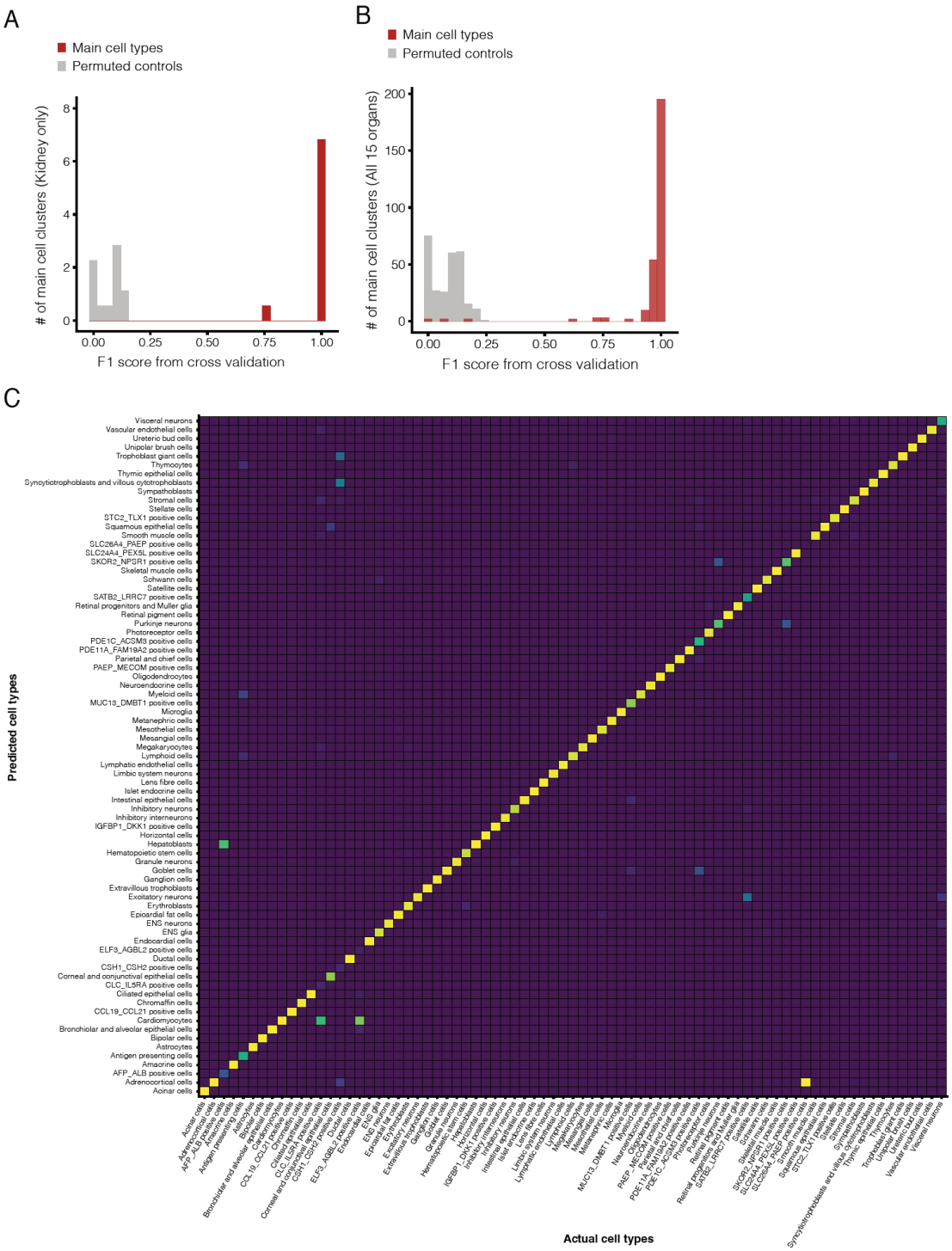


Fig. S4. Evaluation of main cell type specificity. (**A-B**) Histogram showing distribution of cluster specificity scores (F1 scores) of main cell types vs. permuted controls in kidney (A) or all 15 organs (B). (**C**) Confusion matrix for intra-dataset cell type cross-validation by SVM classifier for the 77 main cell types identified across all 15 organs.

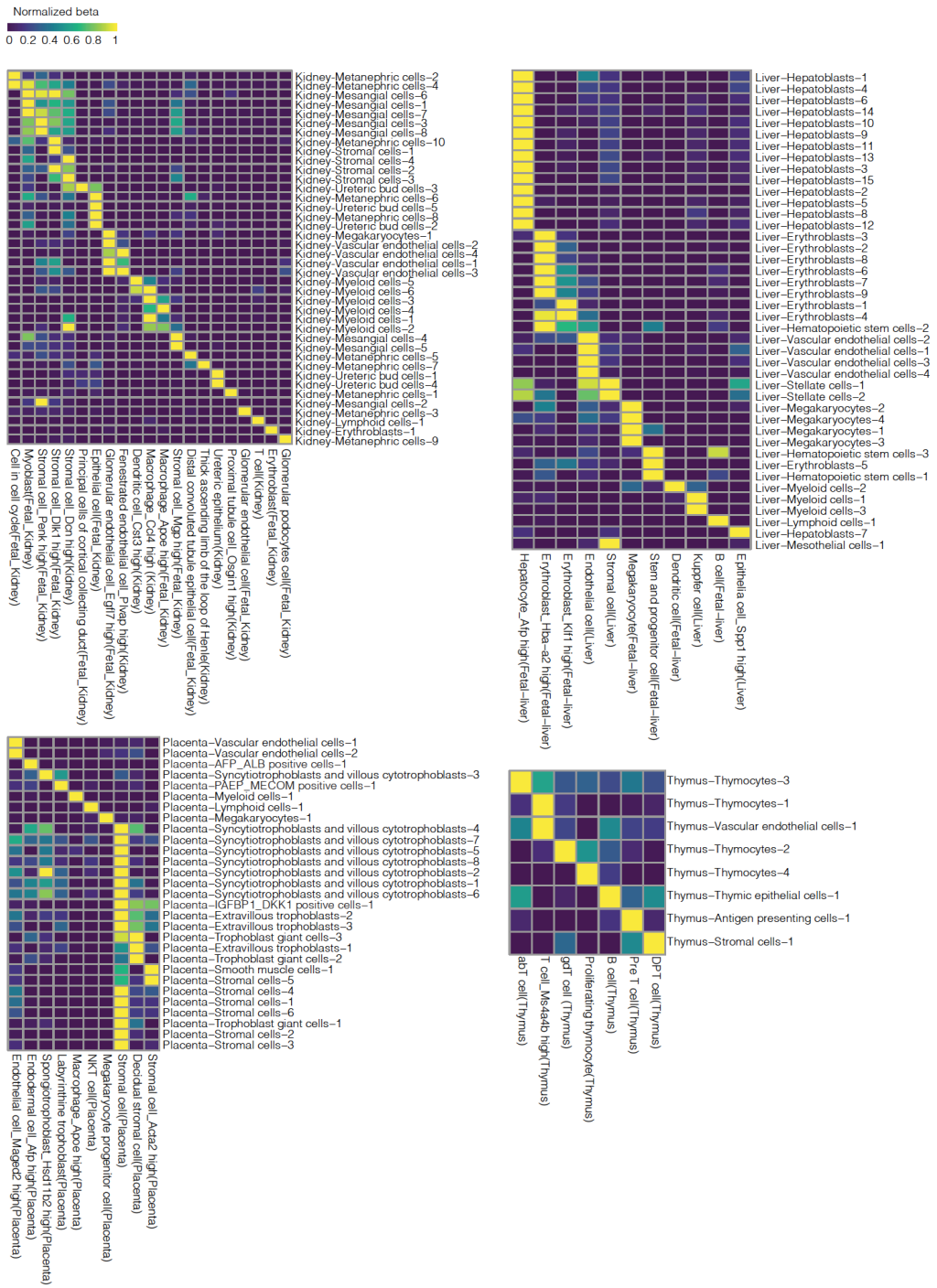


Fig. S5. Cell type correlation analysis with corresponding mouse tissues (kidney, liver, placenta, thymus). We compared putative human cell subtypes identified in our data (rows) against annotated mouse cell types from the corresponding tissues (*16*) (columns) by cell type correlation analysis. Colors correspond to beta values, normalized by the maximum beta value per row. All MCA cell types with a beta of a matched human cell type > 0.01, that is also the maximum beta for that human cell type, are shown for kidney, liver, placenta, and thymus.

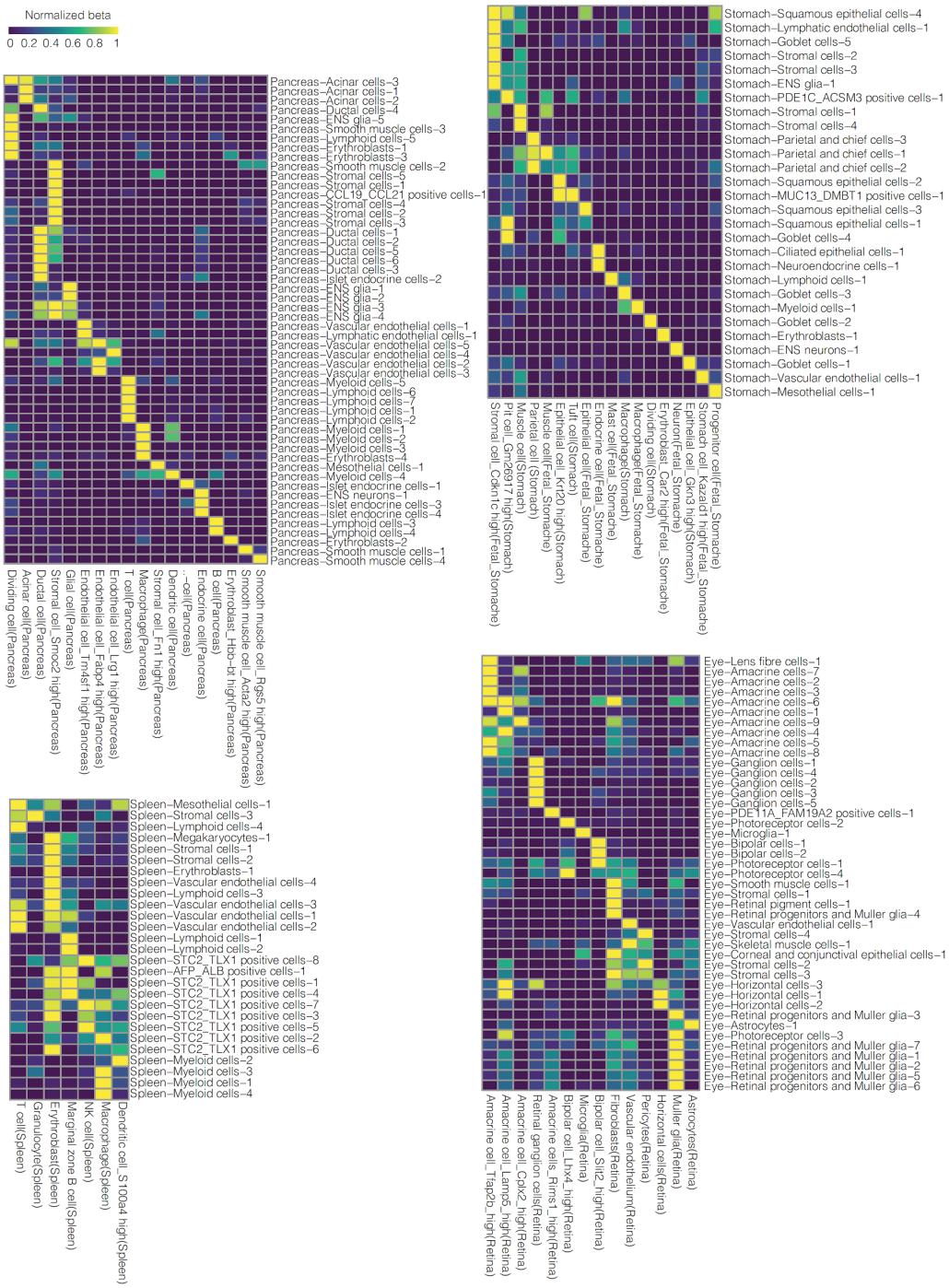


Fig. S6. Cell type correlation analysis with corresponding mouse tissues (pancreas, stomach, spleen, eye). We compared putative human cell subtypes identified in our data (rows) against annotated mouse cell types from the corresponding tissues (*16*) (columns) by cell type correlation analysis. Colors correspond to beta values, normalized by the maximum beta value per row. All MCA cell types with a beta of a matched human cell type > 0.01, that is also the maximum beta for that human cell type, are shown for pancreas, stomach, spleen and eye.

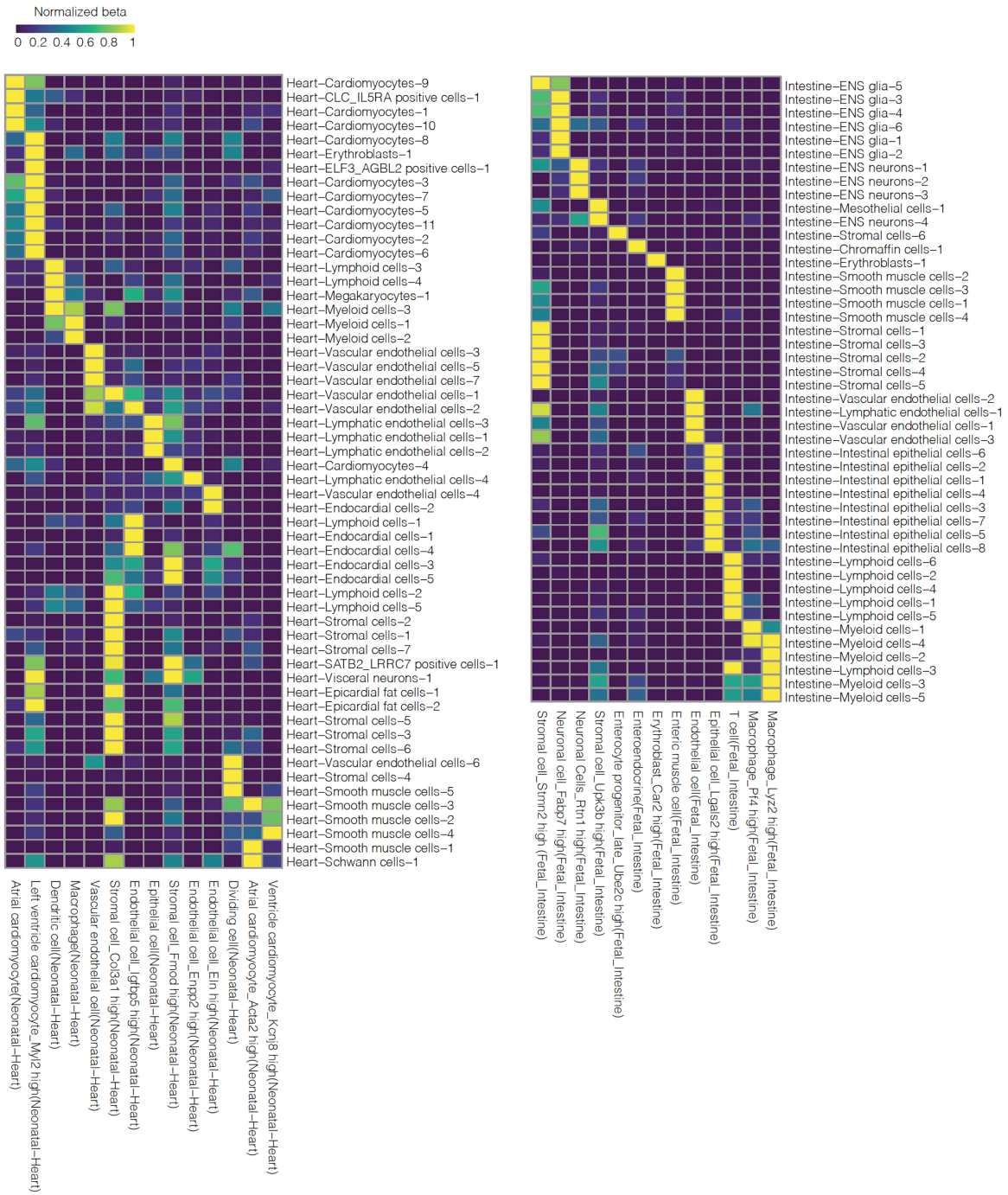


Fig. S7. Cell type correlation analysis with corresponding mouse tissues (heart, intestine). We compared putative human cell subtypes identified in our data (rows) against annotated mouse cell types from the corresponding tissues (*16*) (columns) by cell type correlation analysis. Colors correspond to beta values, normalized by the maximum beta value per row. All MCA cell types with a beta of a matched human cell type > 0.01, that is also the maximum beta for that human cell type, are shown for heart and intestine.

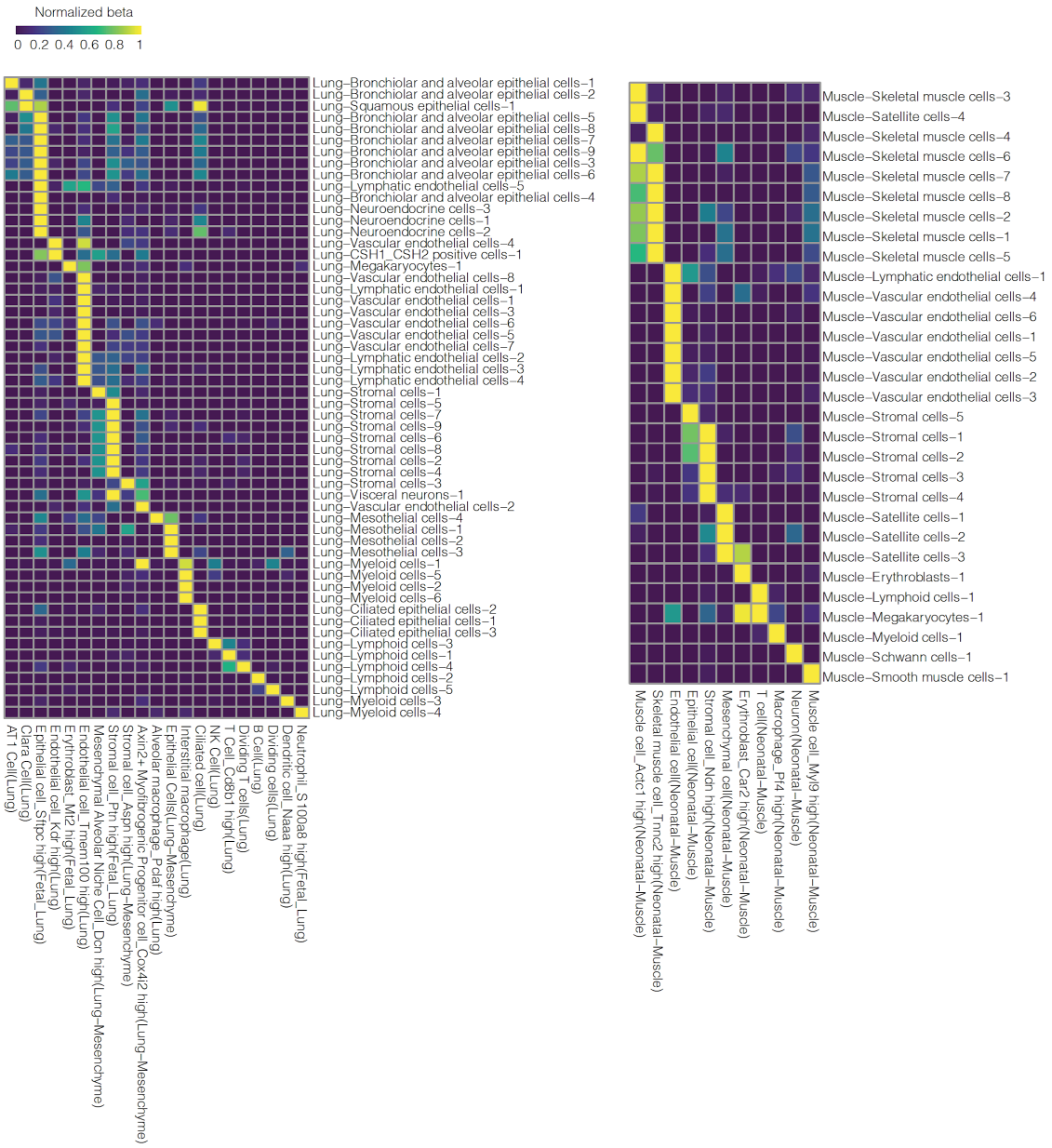


Fig. S8. Cell type correlation analysis with corresponding mouse tissues (lung, muscle). We compared putative human cell subtypes identified in our data (rows) against annotated mouse cell types from the corresponding tissues (*16*) (columns) by cell type correlation analysis. Colors correspond to beta values, normalized by the maximum beta value per row. All MCA cell types with a beta of a matched human cell type > 0.01, that is also the maximum beta for that human cell type, are shown for lung and muscle.

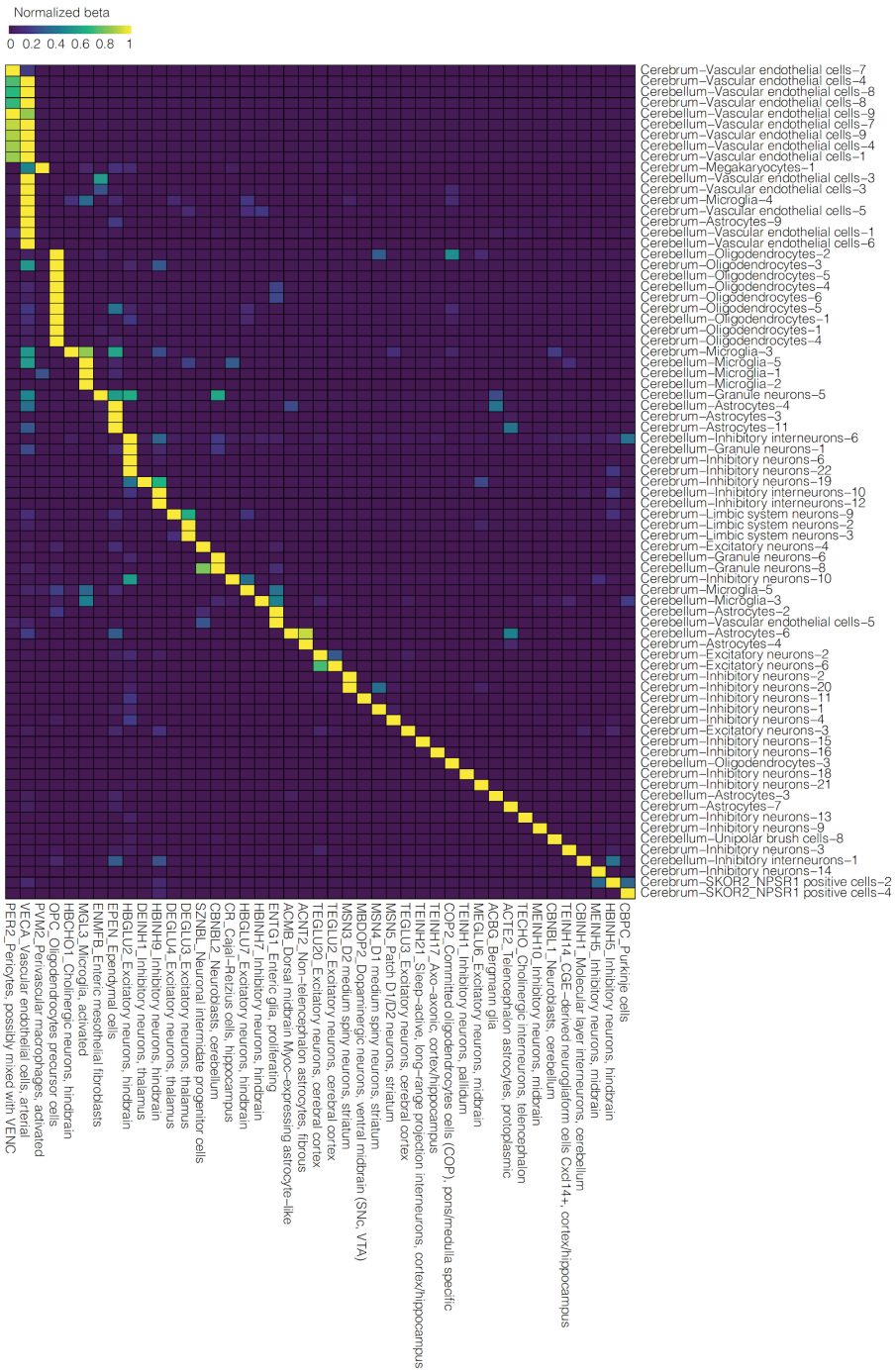


Fig. S9. Cell type correlation analysis with mouse brain cell atlas. We compared putative human cell subtypes identified in our data in the cerebrum or cerebellum (rows) against annotated mouse cell types from the Mouse Brain Cell Atlas (*50*) (columns) by cell type correlation analysis. Colors correspond to beta values, normalized by the maximum beta value per row. All MBCA cell types with a beta of a matched human cell type > 0.01, that is also the maximum beta for that human cell type, are shown together with matching human subtypes from cerebrum or cerebellum.

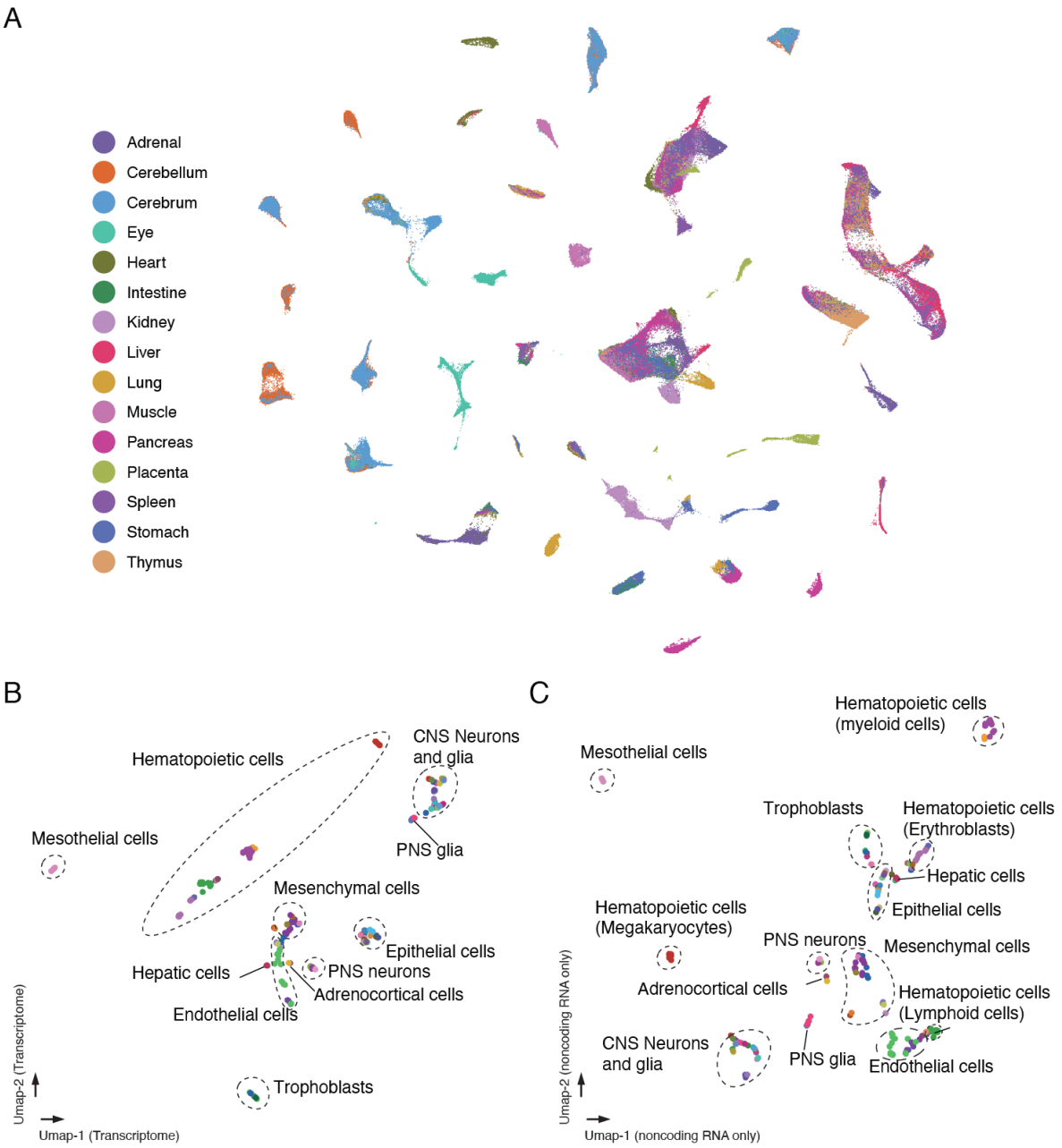


Fig. S10. Clustering analysis of 77 main cell types. (**A**) From each organ, we sampled 5,000 cells from each cell type (or all cells for cell types with fewer than 5,000 cells in a given organ). These were subjected to UMAP visualization based on the top differentially expressed genes across cell types within each organ. Here they are colored by tissue-of-origin. In **Fig. 3A**, the same UMAP visualization is colored by cell type labels. (**B-C**) We aggregated the transcriptomes for each of the 172 annotated clusters observed across 15 organs, *i.e.* prior to collapsing common labels. These pseudobulk profiles were subjected to UMAP visualization, either by top differentially expressed genes across cell types within each organ (B), or by differentially

expressed noncoding RNAs (C). Colors are the same as those used in **Fig. 1C**.

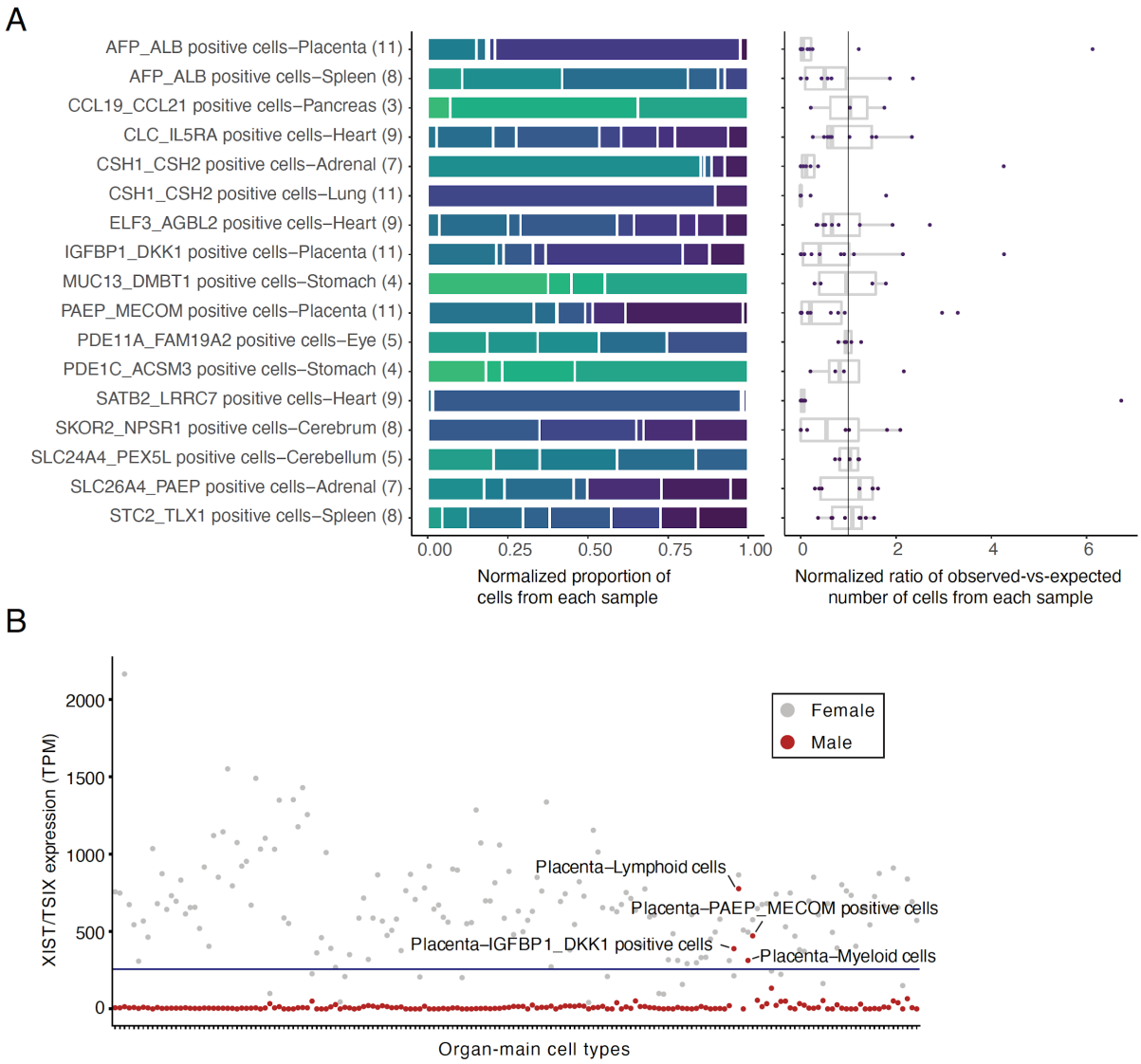


Fig. S11. **Distribution across individuals and potential for maternal origin of initially unannotated cell types across individuals.** (**A**) Bar plot (left) showing how initially unannotated cell types are distributed across all samples that were processed for the organ in which the unannotated cell type was identified (normalized by the total number of cells sampled in each sample). The number of samples represented for each organ is listed in parentheses. Point and box plot (right) showing the ratio of observed vs. expected numbers of cells for initially unannotated cell types in each sample of the organ in which they were identified ("expected" assumes that the overall number of cells of that type should be distributed equally across all tissue samples of the organ). (**B**) Scatter plot showing the summed expression of maternal specific genes (*XIST* and *TSIX*; TPM, transcripts per million) in all main cell types separated by male (colored in red) vs. female (colored in grey). The horizontal line represents 3 standard deviations above the mean gene

expression value of all cell types in male-derivied tissues.

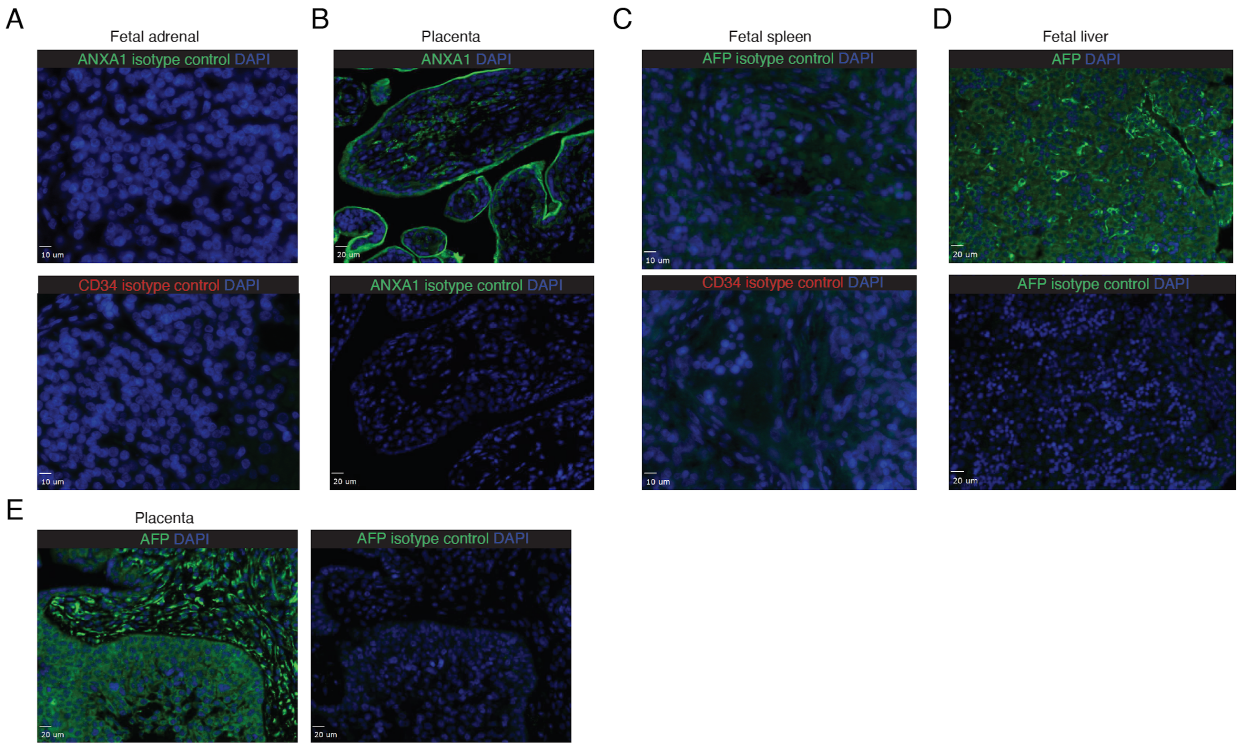


Fig. S12. Immunostaining with isotype controls and antibody validation in control tissues. (**A, C**) Representative fluorescence microscopy of human fetal adrenal (A) or spleen (C) tissue, staining for isotype controls for antibodies used in **Fig. 3CD**. (**B, D**) Representative fluorescence microscopy of control organs for ANXA1 (staining trophoblasts from placenta) or AFP (staining hepatocytes from liver) in corresponding control organs. (**E**) Representative fluorescence microscopy of placenta for *AFP*+, *ALB*+ cells, staining for AFP+. For controls, the concentration of isotype IgG used is the same as the concentration of primary antibody. Images were captured by the same exposure time and renormalized in the same range. Scale bars are shown in each image.

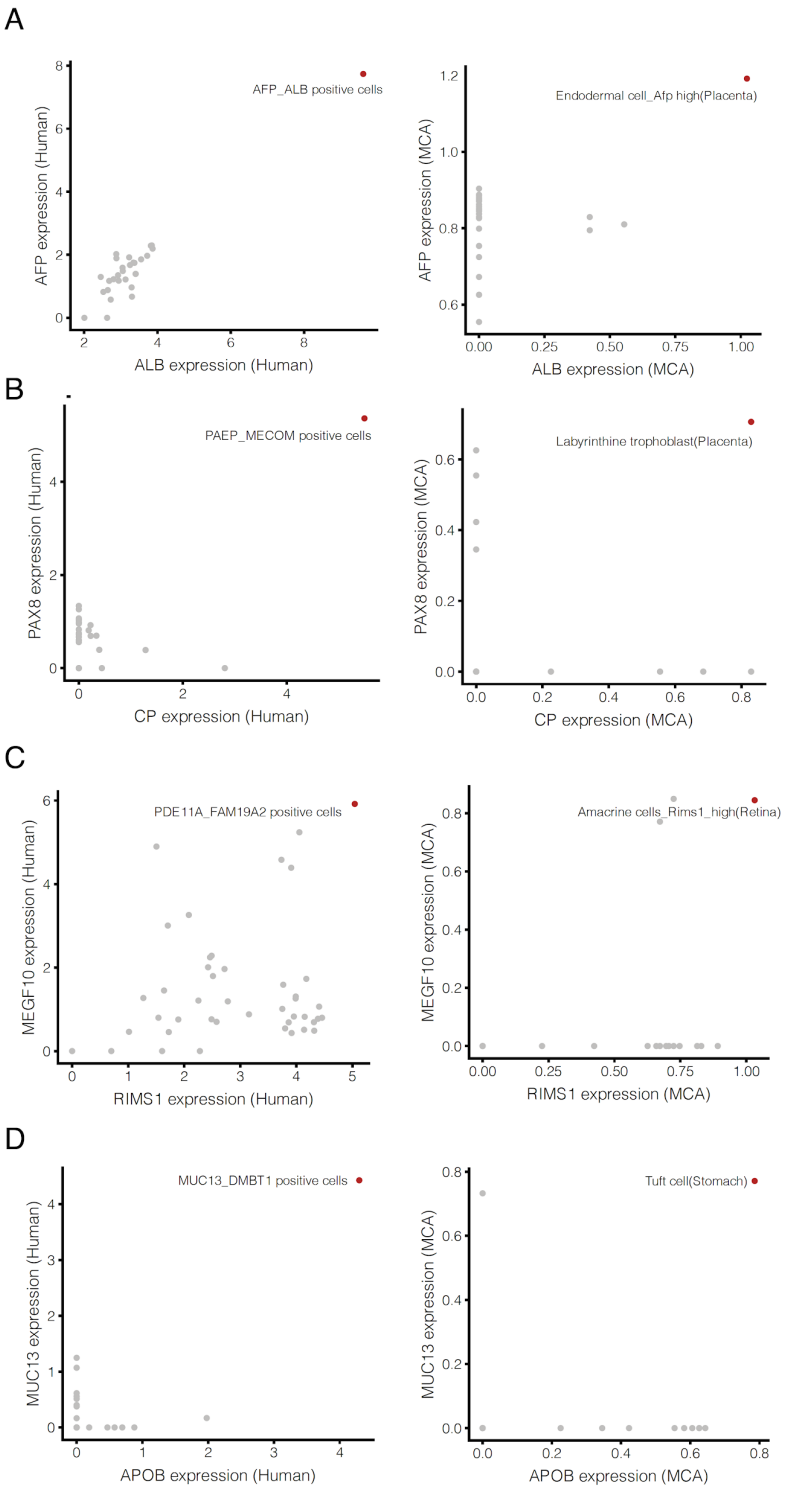


Fig. S13. Cell type correlation analysis informs annotation of some initially annotated cell types. Scatter plots showing the co-expression of top gene markers for four initially unannotated cell types in human subtypes (left) or matched mouse cell types from the Mouse Cell Atlas (MCA) (*16*) (right plots) from the same organ. The uncharacterized human cell type and matched mouse

cell types are colored in red, with other cell types colored in grey.

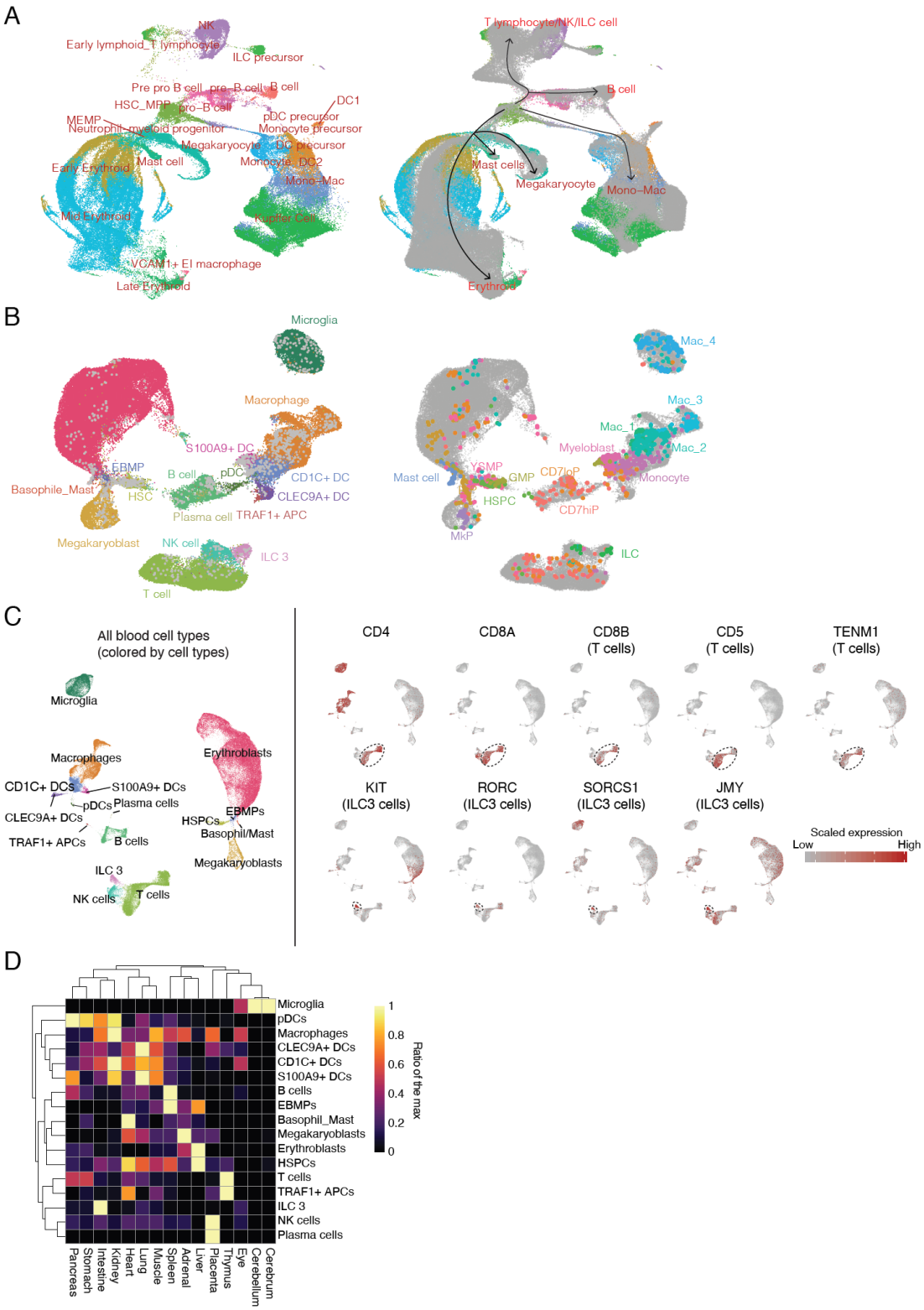


Fig. S14. Integration analysis and Identification of gene markers and cell fractions of blood cell types. (**A**) The same UMAP of **Fig. 4C** is shown on the left and the right, but cells from our study are not shown (left) or colored in dark grey (right). Cells from the scRNA-seq atlas of human

fetal liver blood cells (*108*) are colored based on the annotations from the original publication. Labels on the left derived from annotations in (*108*), and labels on the right are main blood cell categories. (**B**) UMAP visualization of blood cells, integrating across all profiled organs of this study (103,766 cells) and an scRNA-seq atlas of blood cells from human embryos (1,231 cells) (*109*). Cells are colored by blood cell types from our study (left) or the published dataset (right), with cells from the other study colored in grey. (**C**) UMAP visualization of all blood cells, colored as in **Fig. 4B** (reproduced at left) or normalized expression of cell type-specific markers (right). Colors indicate UMI counts for each gene that have been scaled for library size, log-transformed, and then mapped to Z-scores. The enriched cell types for each gene marker are labeled. (**D**) Heatmap showing the blood cell fraction of all blood cells derived from each organ, normalized by the max value of each cell type.

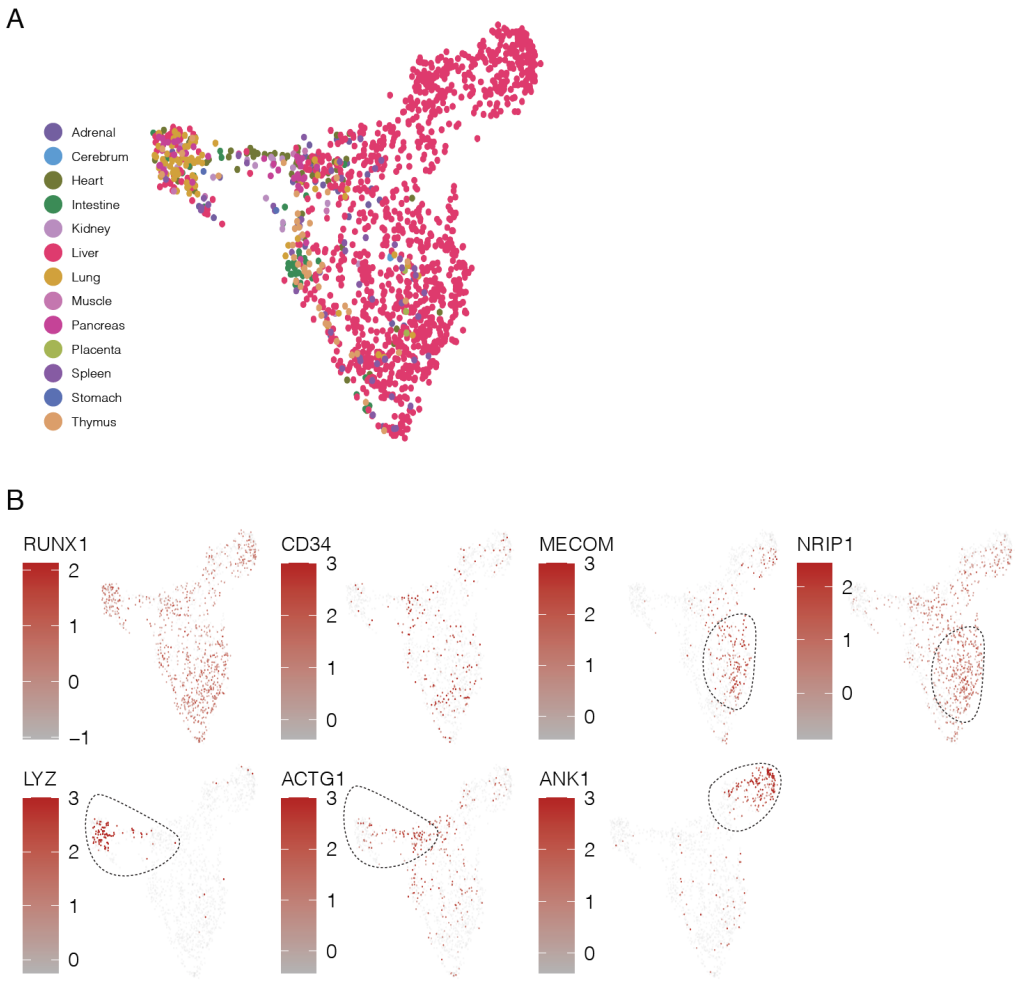


Fig. S15. UMAP visualization of hematopoietic stem/progenitor cells (HSPCs). (**A**) 1,608 HSPCs were selected for reanalysis and UMAP visualization, colored by tissue-of-origin. (**B**) Colors correspond to UMI counts for each gene that have been scaled for library size, log-transformed, and then mapped to Z-scores. The plotted genes include HSPCs markers (*RUNX1* and *CD34*), quiescent HSPCs markers (*MECOM* and *NRIP1*), and HSPCs differentiation markers (*LYZ, ACTG1, ANK1*).

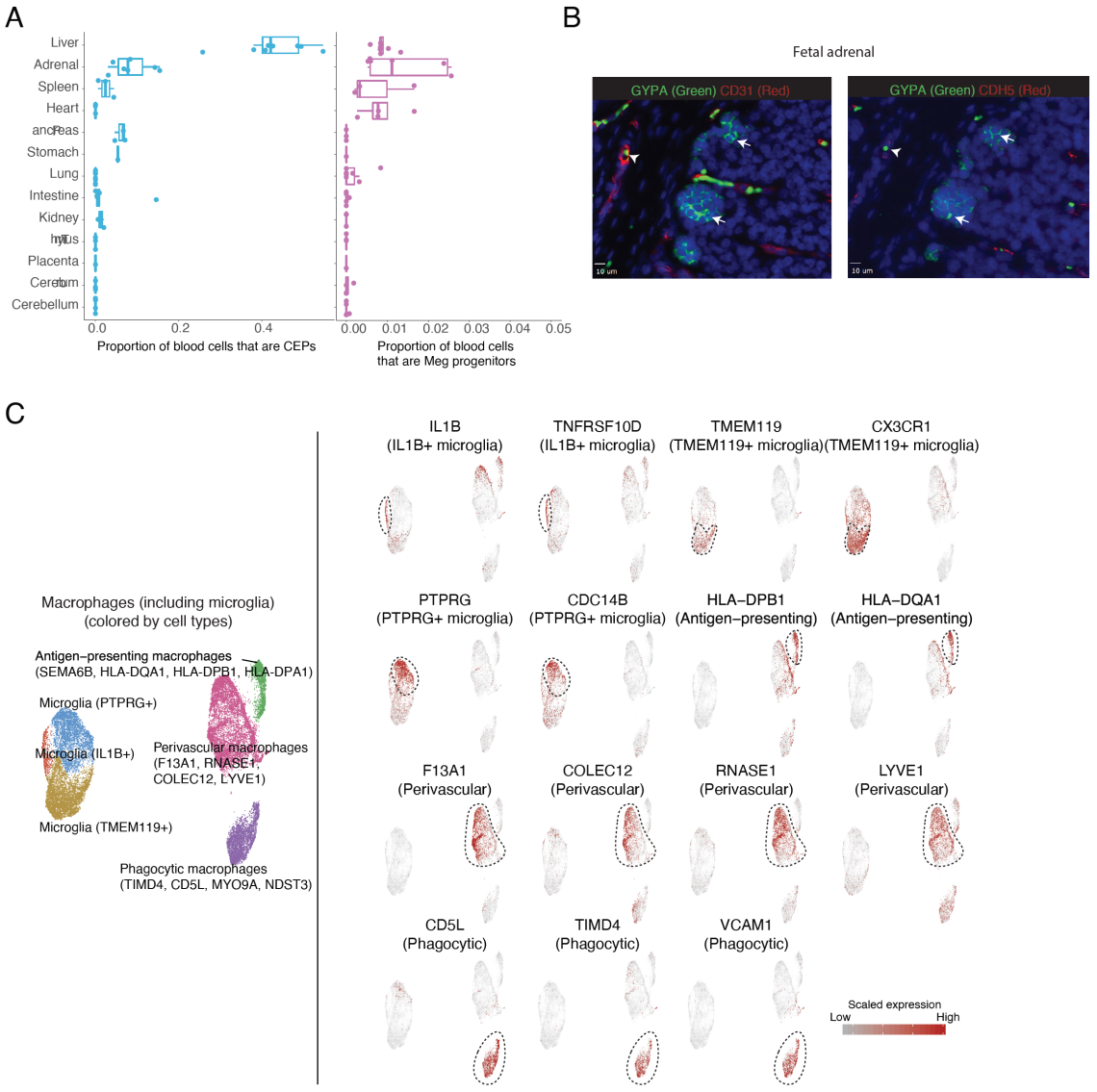


Fig. S16. Identification of gene markers and cell fractions of blood cell types. (**A**) Point and box plot showing the proportion of blood cells that are CEPs (left) or megakaryocyte progenitors (right) for individual samples of different organs. (**B**) Representative fluorescence microscopy of human fetal adrenal tissue staining for endothelium (CD31+, CDH5+) and erythroblasts (nucleated and GYPA+); nuclei stained with blue DAPI. The arrow indicates an GYPA+ erythroblast outside a CD31+, CDH5+ blood vessel. The arrow head indicates an GYPA+ erythroblast inside the CD31+, CDH5+ blood vessel. Scale bars, 10 μm. (**C**) UMAP visualization of all macrophages, colored by macrophage cell type as in **Fig. 5G** (reproduced at left) or normalized expression of cell type-specific markers (right). Colors indicate UMI counts for each gene that have been scaled for library size, log-transformed, and then mapped to Z-scores. The enriched cell types for each gene marker are labeled.

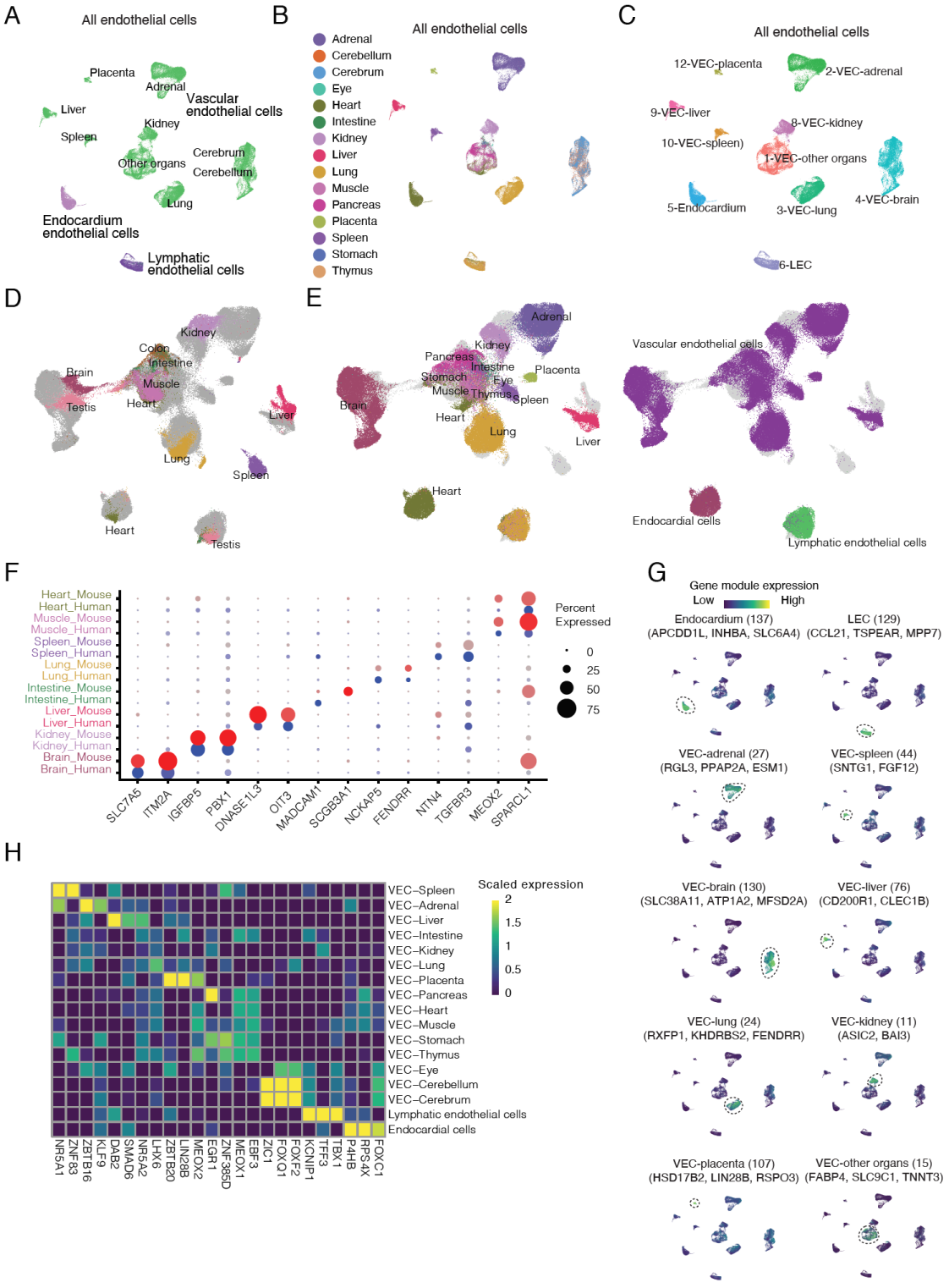


Fig. S17. Characterization and comparison of endothelial cell subtypes. (**A-C**) UMAP visualization of endothelial cell types colored by main cell types with tissue-of-origin of vascular

endothelial cells annotated (A) tissue-of-origin (B) or sub-cluster (C). VEC: vascular endothelial cells. LEC: Lymphatic endothelial cells. (**D**) UMAP visualization of single cells integrating human fetal ECs from our study with mouse adult ECs from (*139*). Mouse ECs from the published dataset are colored by tissue source (*139*). Human cells are colored in dark grey. (**E**) Similar to (D), but mouse cells are colored in light grey, while human cells are colored by the tissue source (left) or annotations (right). (**F**) Dot plot showing the expression of conserved gene markers for organ-specific endothelial cells (except intestinal endothelial cells, where conserved gene markers are not found). (**G**) UMAP visualization of human endothelial cells as in (A), colored by the normalized expression of cell type-specific genes (FDR of 0.05 and over 2-fold expression difference between first and second ranked cell type), with the number of cell type-specific genes and top gene names listed. UMI counts for these genes are scaled for library size, log-transformed, aggregated and then mapped to Z scores. VEC: Vascular endothelial cells. LEC: Lymphatic endothelial cells. (**H**) Heatmap showing cell type-specific TF expression for different endothelial cell types across organs. Aggregated UMI counts of each cluster for these TFs are scaled for library size, log-transformed, aggregated, and mapped to Z scores (capped to [0-2]).

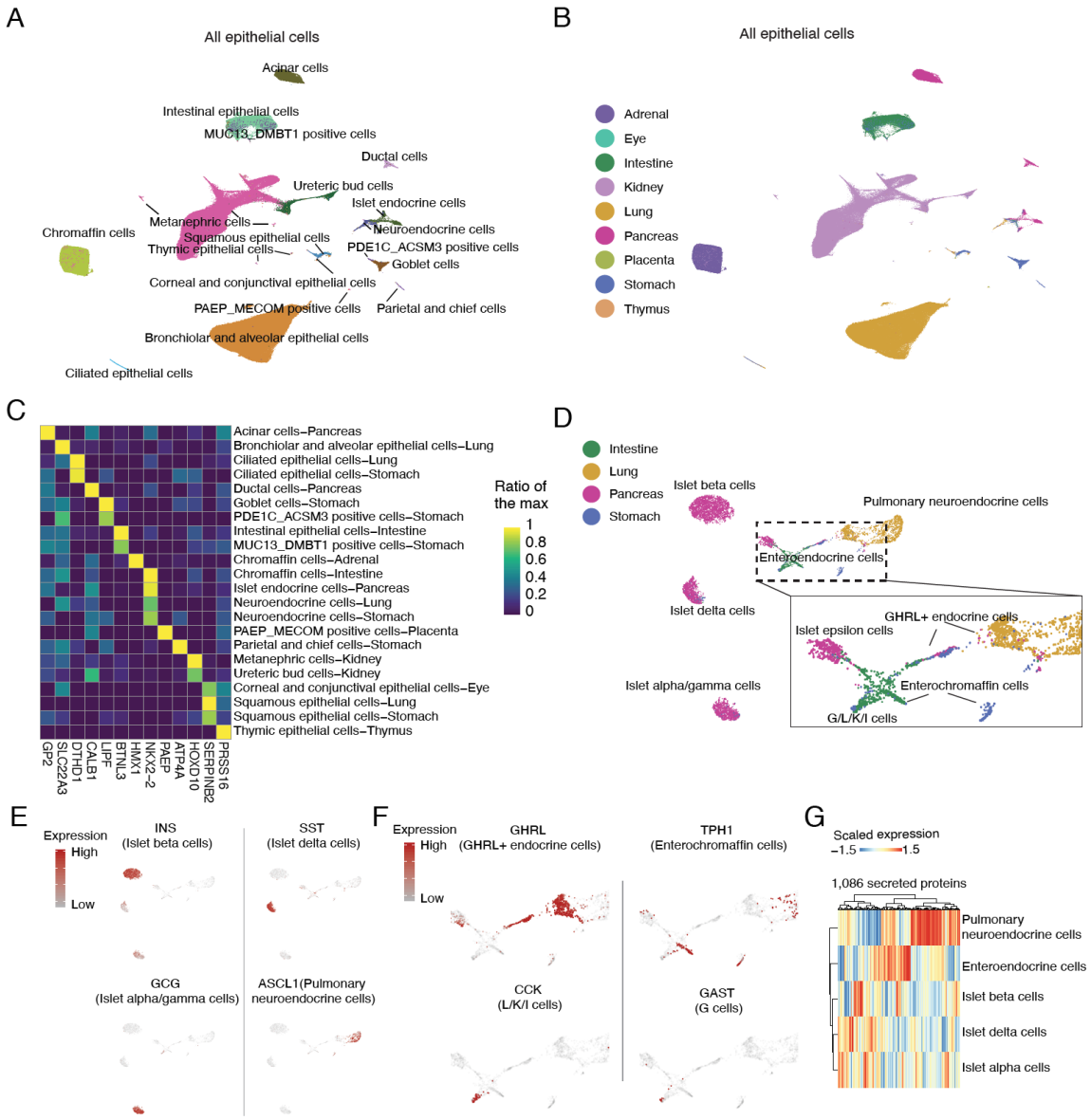


Fig. S18. Characterization and comparison of epithelial cell subtypes. (**A-B**) UMAP visualization of epithelial cells colored by main cell types (A) tissue-of-origin (B). (**C**) Heatmap showing a cell type-specific marker for different epithelial cell types across different organs. (**D-F**) UMAP visualization of neuroendocrine cells from digestive and respiratory organs colored by tissue-of-origin (D) or marker gene expression (E, F), with zoomed view of the enteroendocrine cells (F). Colors indicate UMI counts for each gene that have been scaled for library size, log-transformed, and then mapped to Z-scores. (**G**) Heatmap showing the relative expression of 1,086 secreted protein coding genes in consensus transcriptomes for each neuroendocrine cell cluster. The raw expression data (UMI count matrix) is log-transformed, column centered and scaled (using the R function scale).

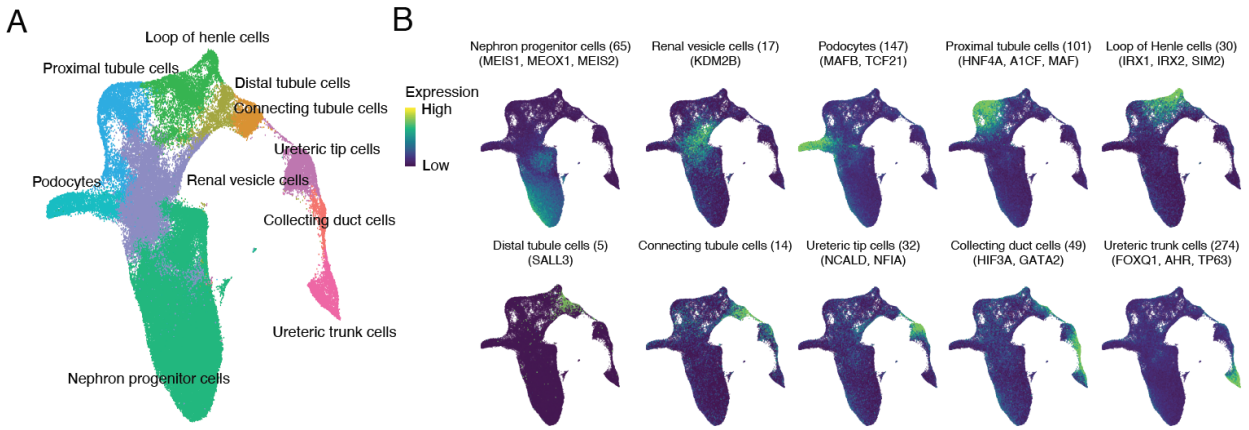


Fig. S19. Characterization and comparison of renal epithelial cell subtypes. (**A**) UMAP visualization of renal epithelial cells, colored by main progenitor and terminal cell types. (**B**) Plot similar with (A), colored by the normalized expression of cell type-specific genes (FDR of 0.05 and over 2-fold expression difference between first and second ranked cell type), with the number of cell type-specific genes and top TFs listed. UMI counts for these genes are scaled for library size, log-transformed, aggregated and then mapped to Z scores.

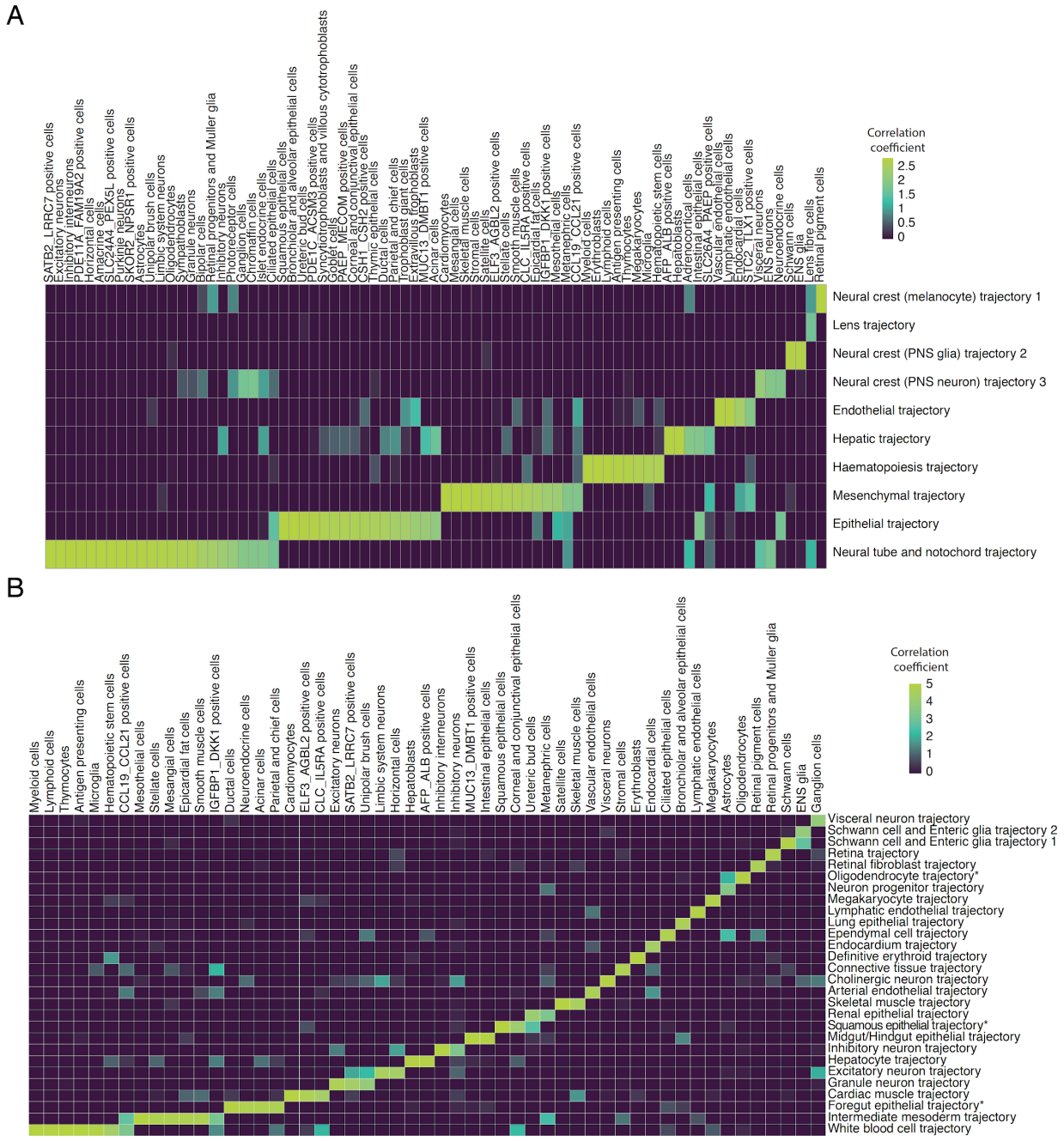


Fig. S20. Cell type correlation analysis between human and mouse cell atlases. (**A**) We compared 77 human fetal cell types (columns) against 10 mouse embryonic cell trajectories (MOCA) (*11*) (rows) with a cell type cross-matching method (*11*). Colors correspond to beta values, scaled by row and capped to [0, 6]. All main human cell types are shown (n = 77), as are MOCA major trajectories that are top matches for 1+ human cell types (n = 10). One unexpected match is between human retinal pigment cells and the mouse melanocyte trajectory. However, the marker genes for these cell types are very similar, skin was not represented in the human data, and

a subcluster of the mouse melanocyte trajectory does in fact appear to correspond to retinal pigment cells. (**B**) We compared the 77 human fetal cell types (columns) against 56 mouse embryonic cell sub-trajectories (MOCA) (*11*) (rows) with the same method. Colors correspond to beta values, scaled by column and capped to [0, 6]. All human cell types where the summed NNLS regression coefficient was > 0.6 are shown (53 columns), as are all MOCA sub-trajectories that are top matches for 1+ human cell types (28 rows). Reannotated mouse trajectories are marked with an asterisk. Upon comparison to organ-resolved human cell types and further consideration of markers, *e.g. Krt5* (*204*), *Dsp (205)*, *Grhl3 (206)*, we conclude that MOCA "pericardium" was incorrectly annotated in (*11*) and in fact corresponds to "squamous epithelium" (renamed in figure). "Stomach epithelial trajectory" has been renamed "foregut epithelial trajectory" as it mapped to epithelial cell types in organs derived from foregut (stomach and pancreas), rather than stomach alone. The previous uncharacterized "*Pdgfra* positive glial trajectory" has been renamed "oligodendrocyte trajectory", validated by specific expression of *Olig1, Olig2,* and *Brinp3 (166, 167)*.

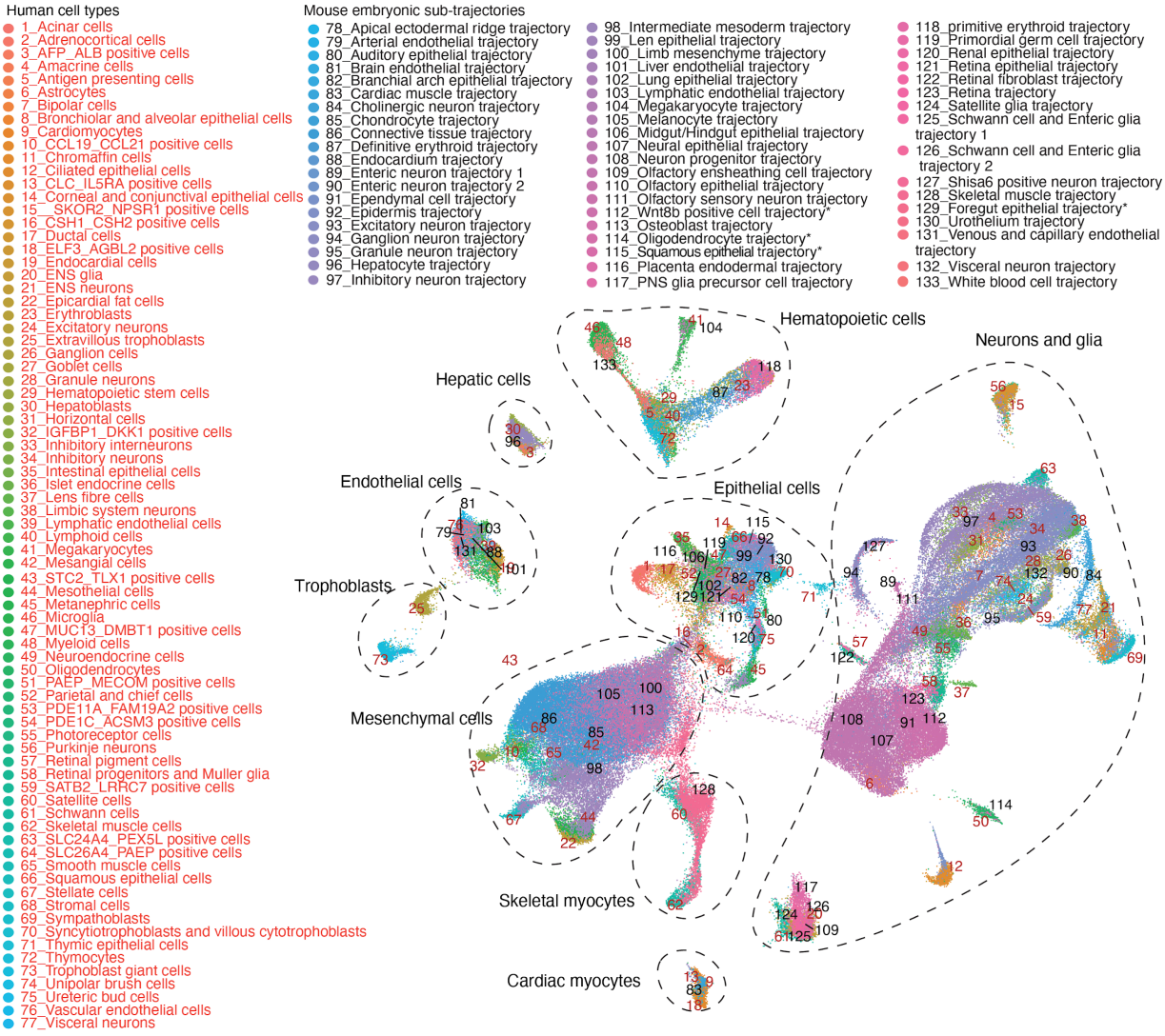


Fig. S21. Additional views of integration of human fetal and mouse embryonic cell atlases. After downsampling as described in the text, we applied Seurat (*15*) to jointly analyze human fetal and mouse embryonic cells (*11*). The same UMAP as in **Fig. 6A-C** is also shown here. Mouse cells are colored by the identity of mouse sub-trajectory (*11*). Cells are colored by human cell type (red labels) or mouse sub-trajectory (black labels). Note that several of the mouse sub-trajectory labels (*11*) have been modified pursuant to this work, indicated by asterisks (112: "oligodendrocyte precursor" → "*Wnt8b*+ cell"; 114: "*Pdgfra*+ glia" → "oligodendrocyte"; 115: "pericardium" → "squamous epithelial"; 129: "stomach epithelial" → "foregut epithelial"; see main text for justifications). Dotted circles indicate major lineages.

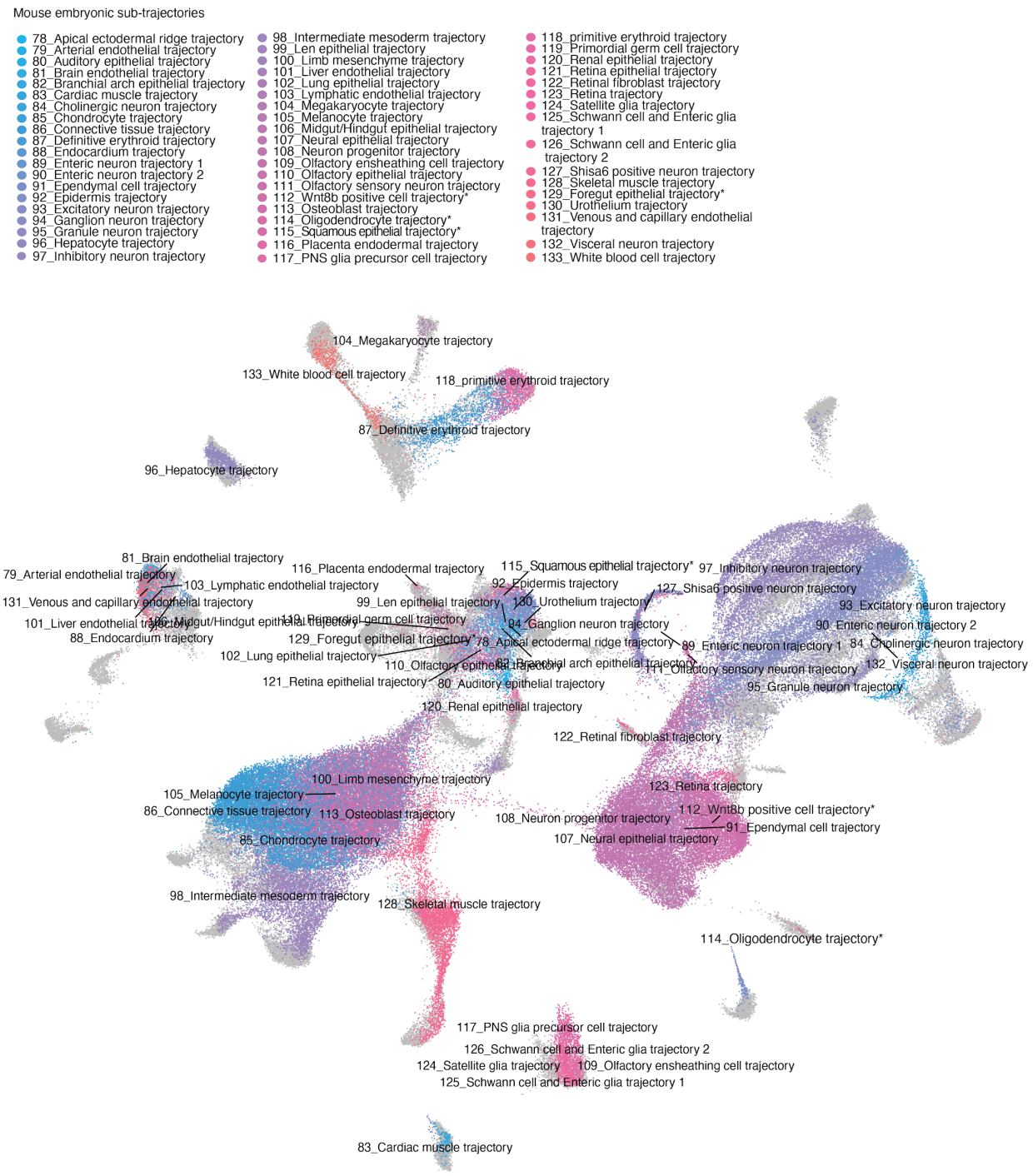


Fig. S22. Additional views of integration of human fetal and mouse embryonic cell atlases. After downsampling as described in the text, we applied Seurat (*15*) to jointly analyze human fetal and mouse embryonic cells (*11*). The same UMAP as in **Fig. 6A-C** and **Fig. S21** is also shown here. Mouse cells are colored by the identity of mouse sub-trajectory (*11*). Human cells are colored in grey.

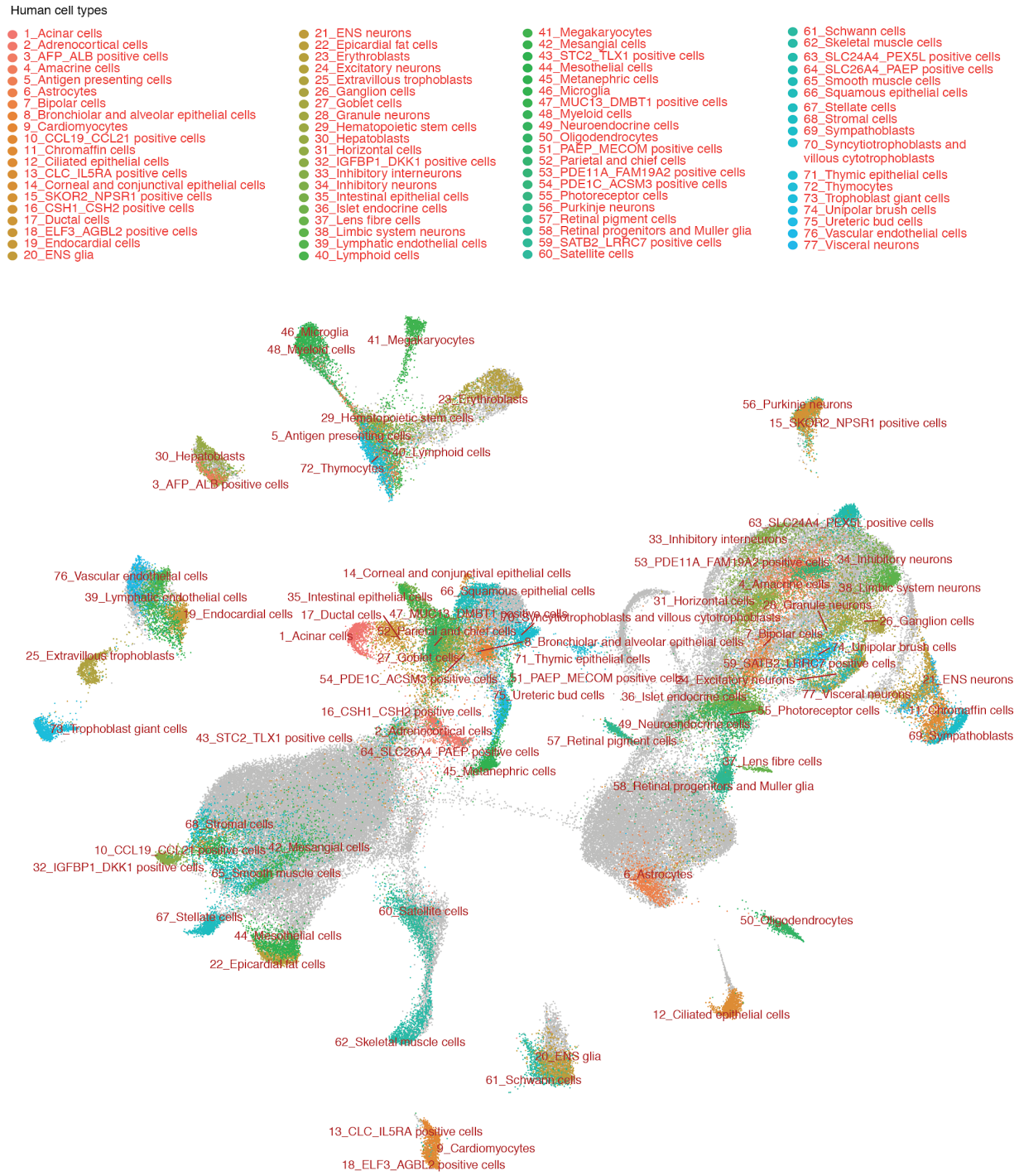


Fig. S23. Additional views of integration of human fetal and mouse embryonic cell atlases. After downsampling, we applied Seurat (*15*) to jointly analyze human fetal and mouse embryonic cells (*11*). The same UMAP as in **Fig. 6A-C** and **Fig. S21** is also shown here. Human cells are colored according to the 77 main cell types annotated here. Mouse cells are colored in grey.

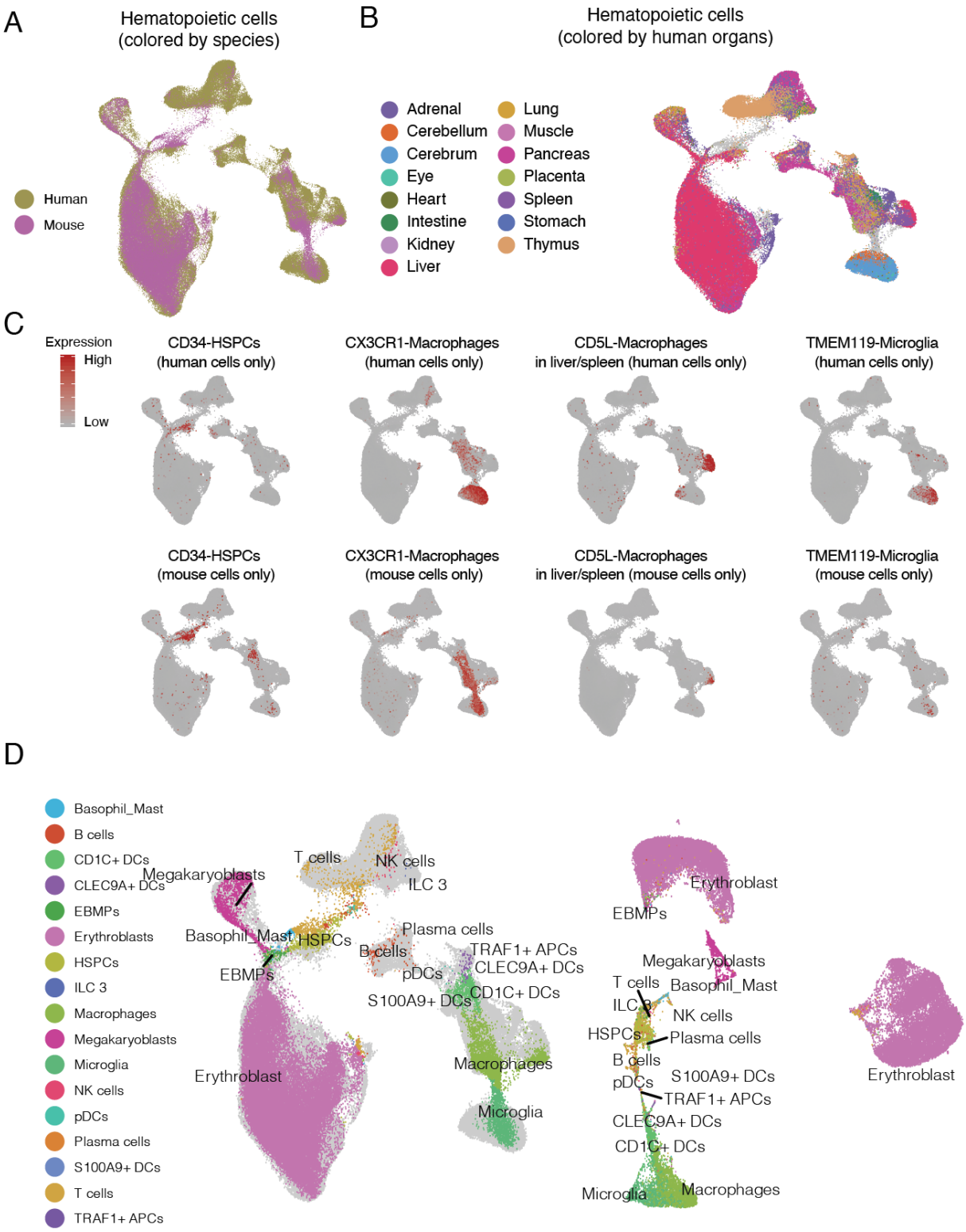


Fig. S24. Integration of all hematopoietic cells from human fetal and mouse embryonic cell atlases. We applied Seurat (*15*) to jointly analyze 103,766 human and 40,606 mouse hematopoietic

cells. (**A**) The UMAP visualization is the same with **Fig. 6D**. Cells are colored by source species. (**B**) The UMAP visualization is the same with **Fig. 6D**. Human cells are colored according to their tissue-of-origin. Mouse cells are colored in grey. (**C**) Similar UMAP visualization as above, colored by normalized gene expression in human cells only (top) or mouse cells only (bottom). Colors indicate UMI counts for each gene that have been scaled for library size, log-transformed, and then mapped to Z-scores. (**D**) Left: similar UMAP visualization as above, colored by annotated mouse blood cell types (based on the votes of k = 3 nearest human cells) or grey (human cells). Right: UMAP visualization of mouse blood cells only, colored by the cell type annotations shown in the left figure.

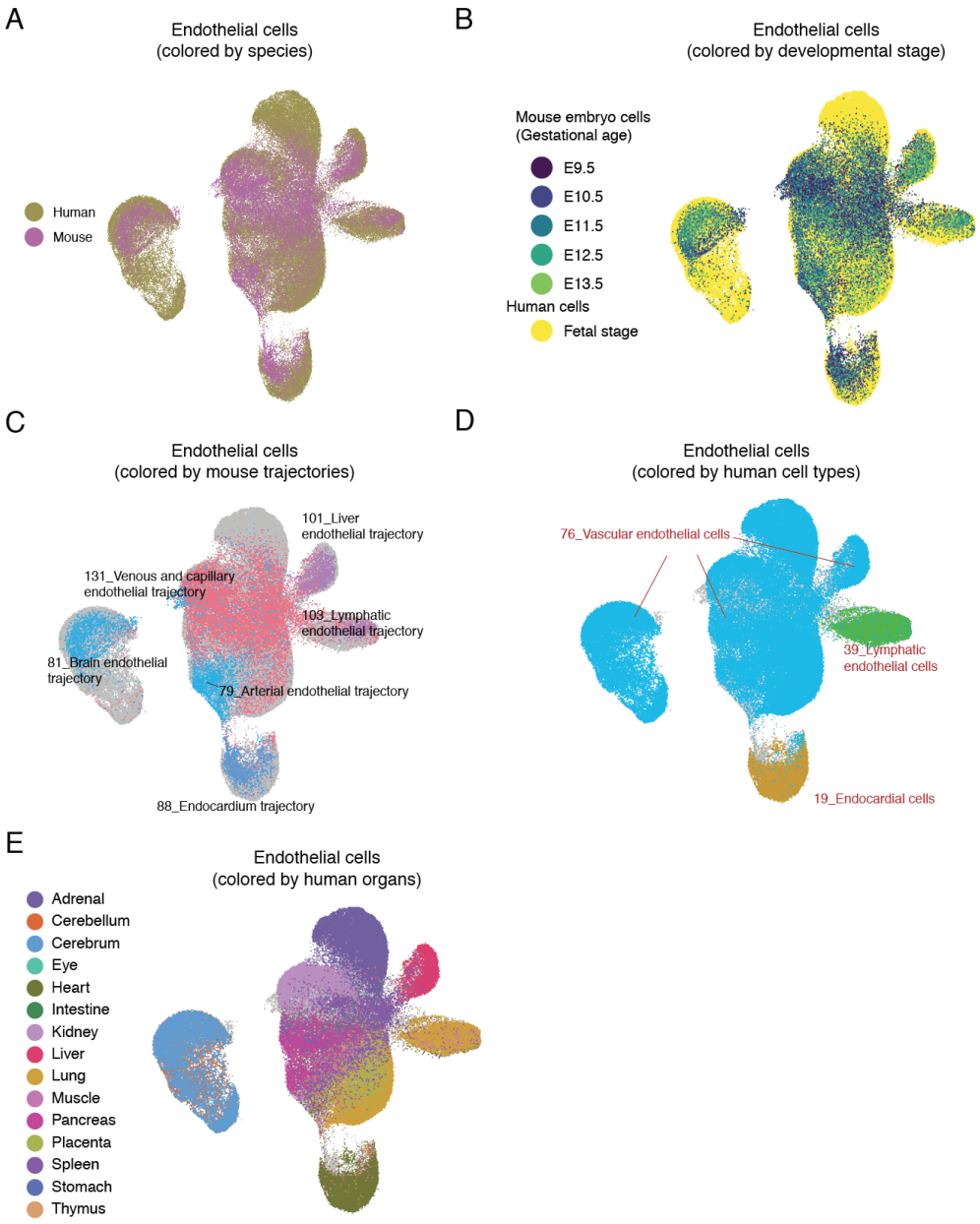


Fig. S25. Integration of all endothelial cells from human fetal and mouse embryonic cell atlases. We applied Seurat (*15*) to jointly analyze 89,291 human and 25,301 mouse endothelial cells. The same UMAP visualization is shown in all panels. (**A**) Cells are colored by source species. (**B**) Cells are colored by source/development stage. (**C**) Mouse cells are colored by the identity of mouse sub-trajectory (*11*). Human cells are colored in grey. (**D**) Human cells are colored according to the 77 main cell types annotated here. Mouse cells are colored in grey. (**E**) Human cells are colored according to their tissue-of-origin. Mouse cells are colored in grey.

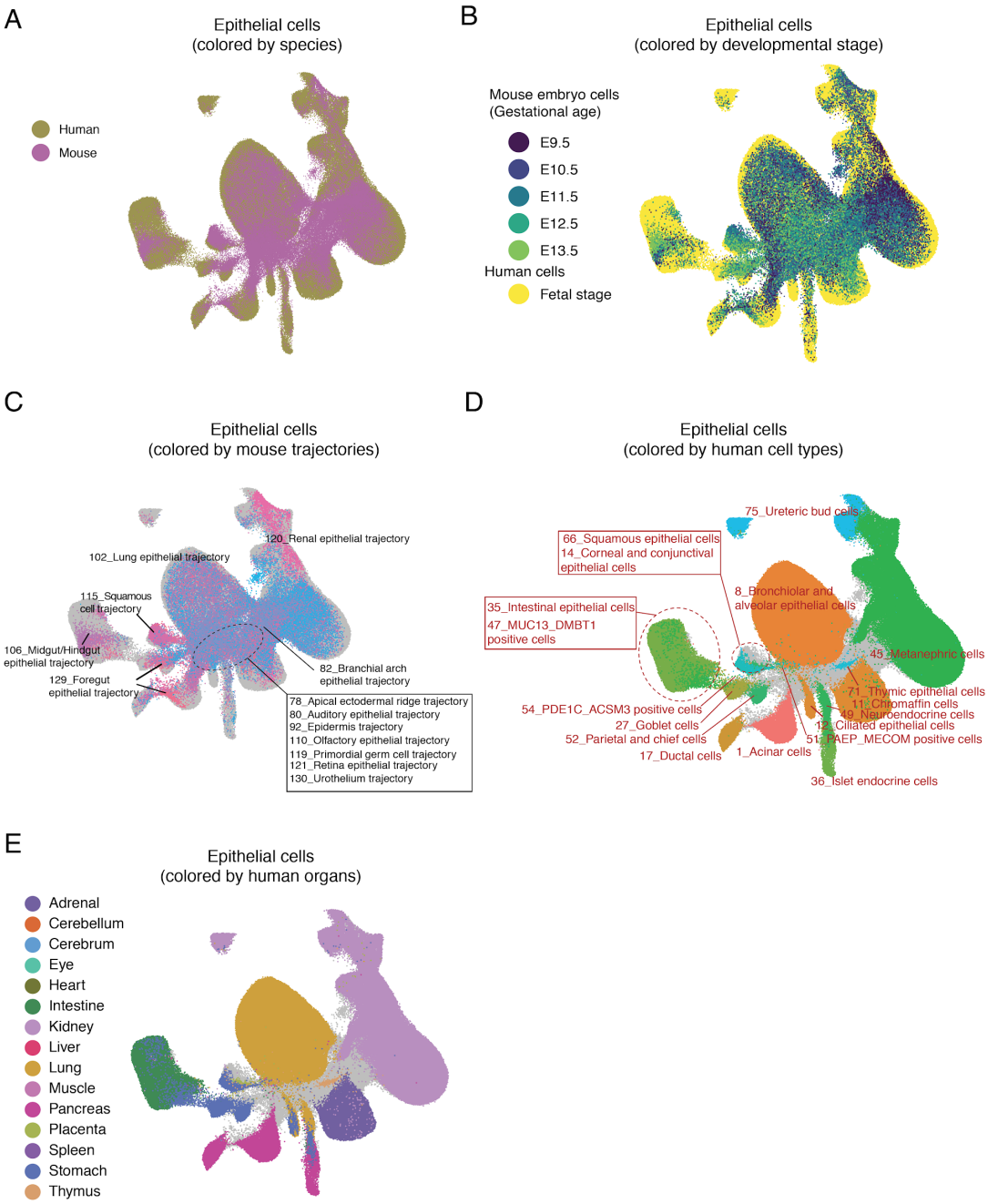


Fig. S26. Integration of all epithelial cells from human fetal and mouse embryonic cell atlases. We applied Seurat (*15*) to jointly analyze 282,262 human and 65,449 mouse epithelial cells. The same UMAP visualization is shown in all panels. (**A**) Cells are colored by source species. (**B**) Cells are colored by source/development stage. (**C**) Mouse cells are colored by the identity of mouse sub-trajectory (*11*). Human cells are colored in grey. (**D**) Human cells are colored according to the 77 main cell types annotated here. Mouse cells are colored in grey. (**E**) Human cells are colored according to their tissue-of-origin. Mouse cells are colored in grey.

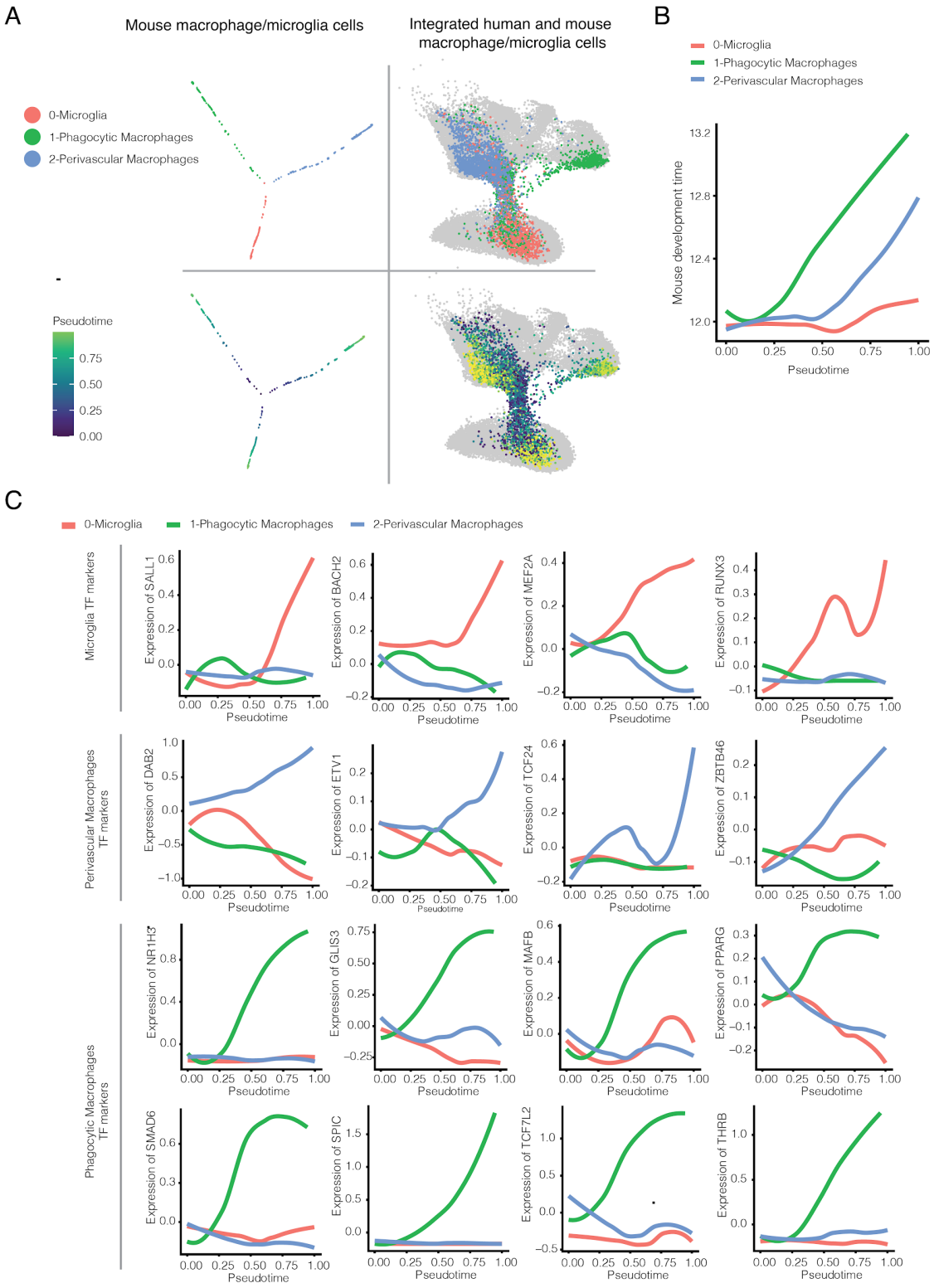


Fig. S27. Trajectory analysis of microglia and tissue resident macrophages. (**A**) Pseudotime trajectory (left) or UMAP coordinates from **Fig. 6D** (right) of all mouse microglia and macrophages, colored by differential branches (top) or pseudotime computed based on each cell's

euclidean distance to the intersection point in pseudotime trajectory (bottom). (**B-C**) Smoothed line plots showing dynamics of developmental time (B) or TF expression (C) as a function of pseudotime for cells in each of the three branches. For gene expression, UMI counts for each gene are scaled for library size, log-transformed, and then mapped to Z-scores.

Legends for Supplementary Tables

Table S1 | Metadata for 121 fetal tissue samples. Includes individual id, organ, estimated post-conceptual age, number of cells profiled, average and median mRNA/gene count, median exonic reads ratio, number of cells included in downstream analysis, and annotation for trisomy 18 samples.

Table S2 | Differential gene expression test results for the aggregated "pseudobulk" transcriptomes of different organs. For each gene, the "max_organ" is the organ with the highest expression by transcripts per million (TPM) ("max.expr"). The "second_organ" is the organ with the second highest expression by transcripts per million (TPM) ("second.expr"). The "fold.change" is the fold change between the max expression and second max expression. The "qval" is the false detection rate (one-sided likelihood ratio test with adjustment for multiple comparisons) for the differential gene expression test across the aggregated pseudobulk transcriptomes of different organs.

Table S3 | Metadata for main cell types annotated in each organ. We list the 172 cell types annotated during our organ-by-organ review, together with the gene markers supporting that annotation and citations to the literature supporting those gene markers. The table also includes basic statistics for each main cell type, including number of cells profiled, average and median mRNA/gene count, median exonic reads ratio, number of samples, and statistics from intra-data cross-validation analysis: cross-validation accuracy (CV_accuracy), recall value (CV_recall) and F1 score (CV_F1_score).

Table S4 | Differential gene expression test results for 77 main cell types. For each gene, the "max_cell_type" is the cell type with the highest expression by transcripts per million (TPM) ("max.expr"). The "second_cell_type" is the cell type with the second highest expression by transcripts per million (TPM) ("second.expr"). The "fold.change" is the fold change between the max expression and second max expression. The "qval" is the false detection rate (one-sided likelihood ratio test with adjustment for multiple comparisons) for the differential gene expression test across different cell types.

Table S5 | Differential gene expression test results for main cell types within each organ. Similar to Table S4 but on an organ-by-organ basis. For each gene, the "max_cell_type" is the cell type with the highest expression within the organ by transcripts per million (TPM) ("max.expr"). The "second_cell_type" is the cell type with the second highest expression within the organ by transcripts per million (TPM) ("second.expr"). The "fold.change" is the fold change between the max expression and second max expression. The "qval" is the false detection rate (one-sided likelihood ratio test with adjustment for multiple comparisons) for the differential gene expression test across different cell types within the organ.

Table S6 | Metadata for subtypes within each organ. Includes subtype name, top matched cell type in the mouse cell atlas (MCA) (*16*) and mouse brain cell atlas (MBCA) (*50*) and beta values in cell type correlation analysis. The table also includes other basic statistics for each subtype, including number of cells profiled, average and median mRNA/gene count, median exonic reads ratio, number of samples, and statistics from intra-data cross-validation analysis: cross-validation accuracy (CV_accuracy), recall value (CV_recall) and F1 score (CV_F1_score).

Table S7 | Differential gene expression test results for blood cell types. Same column annotation as Table S5.

Table S8 | Differential gene expression test results for cell clusters along erythropoiesis trajectory. Same column annotation as Table S5.

Table S9 | Differential gene expression test results for macrophage/microglia subtypes. Same column annotation as Table S5.

Table S10 | Differential gene expression test results for endothelial subtypes. Same column annotation as Table S5.

Table S11 | Differential gene expression test results for neuroendocrine cell subtypes. Same column annotation as Table S5.

Table S12 | Differential gene expression test results for renal epithelial cell subtypes. Same column annotation as Table S5.

Table S13 | Top matched MOCA (mouse organogenesis) trajectory for each human fetal cell type. Beta value represents the sum of regression coefficients from non-negative least squares (NNLS) regression. Added beta values over 0.6 are considered to be strong matches.

Table S14 | Top matched MOCA (mouse organogenesis) sub-trajectory for each human fetal cell type. Beta value represents the sum of regression coefficients from non-negative least squares (NNLS) regression. Summed beta values over 0.6 are considered to be strong matches.

Table S15 | Conserved blood cell type markers across human and mouse. For each gene, "cell_type_name" is the human and mouse cell type with the highest expression. The "qval_human" and "qval_mouse" are the false detection rates (one-sided likelihood ratio test with adjustment for multiple comparisons) for the differential gene expression test across different blood cell types in human and mouse, respectively.

Table S16 | Differential gene expression test results for mouse embryonic microglia and macrophage subtypes. Same column annotation as Table S5.

Legends for Supplementary Files

File S1 | Metadata of high-quality cells. Includes sample metadata and various per-cell QC stats, Louvain cluster id and cell type annotation, for each of the 4,062,965 high-quality cells used in the downstream analyses.

File S2 | Metadata of genes. Includes gene id, short name and gene type information for each gene.

File S3 | Gene count matrix of cells. Includes expression UMI values for each gene in each cell.

File S4 | Matrix of gene expression values across tissues. Includes normalized gene expression values (transcripts per million) for each tissue.

File S5 | Matrix of proportion of cells with each gene detected across tissues. Includes the proportion of cells in each tissue in which a given gene was detected (UMI > 0).

File S6 | Matrix of gene expression values across main cell types. Includes normalized gene expression values (transcripts per million) for each cell type in each tissue.

File S7 | Matrix of proportion of cells with each gene detected across cell types. Includes the proportion of cells in each major cell type in which a given gene was detected (UMI > 0).