Supplemental Figure Legends

Supplemental Figure 1. Evaluation and benchmarking of the Capybara pipeline. Related to Figure 1 (A) Distributions of quadratic programming (QP) metrics: Error, Lagrangian multiplier, and Deviance. Black line: distribution of these metrics in the reference. Other lines: ideal distributions for discrete, hybrid, and unknown classifications, which are modeled based on the reference metrics. Assuming each cell in the reference has a discrete identity, the ideal distribution of deviance for discrete cells is modeled based on the reference identity scores. The ideal distribution of deviance for hybrid exhibits a two standard deviation shift left of the discrete distribution; the deviance distribution for unknown exhibits a two standard deviation shift left of the hybrid distribution. **(B)** Evaluation using five human pancreatic datasets and the Allen Mouse Brain (AMB) atlas against ten other classifiers, using an established benchmarking pipeline (Abdelaal et al., 2019). The classification performance was evaluated by the area under the receiver operating characteristic (AUROC) and the mean total time in seconds. The color of each square labels the assessment of each aspect. **(C)** Cross-validation using *Tabula Muris* processed with 10x droplet-based or Smart-seq2 technologies. In this evaluation, 90 cells were sampled from each cell type of each tissue to construct the high-resolution reference. Sample cells used for reference construction were used as training sets for scMap and SingleCellNet. The remaining cells were used as test samples. The performance was evaluated by AUROC scores. **(D)** Demonstration of inappropriate reference selection: We applied Capybara to a publicly available Mouse PBMC dataset (https://www.10xgenomics.com/resources/datasets) with two references from the Mouse Cell Atlas (MCA; Han et al., 2018): peripheral blood (Left) and brain (Right). 86.3% of the PBMC cells are classified as unknown using the brain reference. **(E)** Classification of the simulated dataset (described in Figure 1) using scMap with cell or cluster mapping. The color of each square labels the percentage of the actual annotation mapped to each scMap annotation. **(F)** Increased hybrid cell percentages with an increasing number of cell types in the reference, using different annotation complexity in the AMB Atlas. The AMB contains three levels of annotations: 1) Three cell types breaking down into GABAergic, Glutamatergic, and non-neuronal categories; 2) 20 cell types, placing cells into detailed cell categories, such as astrocytes and neural progenitor (NP) cells; 3) 108 cell types, breaking the individual cell types into their subtypes, such as astrocytes expressing *Aqp4* and NP cells expressing *Cpne7* or *Met*.

Supplemental Figure 2. Application of Capybara to classify hematopoietic cell identity. Related to Figure 2 and Table S2 (A) Comparison of log-normalized marker gene expression between classified cell types and other cells. The following genes are used as key markers for different populations: HSPCs (Hematopoietic Stem and Progenitor Cells): *Cd34*; Megakaryocytes: *Itga2b*; Neutrophils: *Cebpe*; Monocytes: *Csf1r*; Erythrocytes: *Car2*. We group the cells into two categories. One group represents all the cells classified as a specific cell type, and the other group represents the remaining cells that are not classified as the same identity. The comparison between the two groups was performed and tested with a two-sample Wilcoxon Rank Sum Test, using the wilcox.test function in R (****: P <=0.0001). **(B)** *Top:* Projection of hybrid cell populations on the PAGA-guided clustering, along with their corresponding discrete identities. *Bottom*: Comparison of pseudotime between hybrids and their discrete identity counterparts. A Wilcoxon test was used for significance testing. **(C)** PAGA-guided clustering of the Paul et al. dataset, consisting of a total of 2,730 myeloid progenitors enriched from mouse bone marrow, revealing a total of 24 clusters. As expected, PAGA connectivity links the clusters containing hybrid cells but does not pinpoint the hybrid state. **(D)** Correlation between log (PAGA connectivity scores) and log (transition scores).

Supplemental Figure 3. Evaluation of hybrid cells using ground-truth lineage tracing dataset. Related to Figure 3 and Table S3. **(A)** Manual annotation of the major differentiated cell populations identified from the Weinreb et al., 2020 hematopoiesis lineage tracing dataset, projected onto the SPRING embedding. **(B)** Comparison between Capybara classification and the manual annotation. We selected the major differentiated myeloid cell types, including basophils, eosinophils, mast cells, monocytes, and neutrophils, as a reference. Cell types not included in the reference are correctly identified as ‘Unknown.’ 95.1% of cells with unknown identities are labeled as undifferentiated in the original Weinreb et al. annotation. The color of each square denotes the percentage agreement between the manual and Capybara-based annotations. **(C)** Integrative analysis of monocyte-neutrophil hybrids, monocytes, neutrophils (identified by Capybara), with IG2 (Irf8-GMP2 bistable intermediates), and GMP (Granulocyte Monocyte Progenitor) cells (identified in Olsson et al., 2016). IG2 cells have the potential to differentiate into monocytes or granulocytes. To evaluate monocyte-neutrophil hybrids, we used Seurat V4 to integrate monocytes, neutrophils, and monocyte-neutrophil hybrids with the IG2 and GMP populations, showing the overlap between monocyte-neutrophil hybrids and the IG2 population. **(D)** The color of each square denotes the percentage of each population in each cluster. **(E)** Cosine similarity of percentage cluster representations across all populations, showing the highest similarity between monocyte-neutrophil hybrids and the IG2 population. **(F)** Comparison between RNA velocity and Capybara transition scores in cardiac reprogramming (Stone et al., 2019). *Left*: RNA velocity vectors projected onto the UMAP embedding (Square:

area containing a small number of moving vectors; Circle: area containing a large number of moving vectors). *Middle*: Transition scores projected onto the UMAP embedding (Square: area corresponding to the RNA velocity plot showing low transition scores; Circle: area corresponding to the RNA velocity plot showing high transition scores). *Right*: correlation between log (velocity vector counts) and log (transition scores).

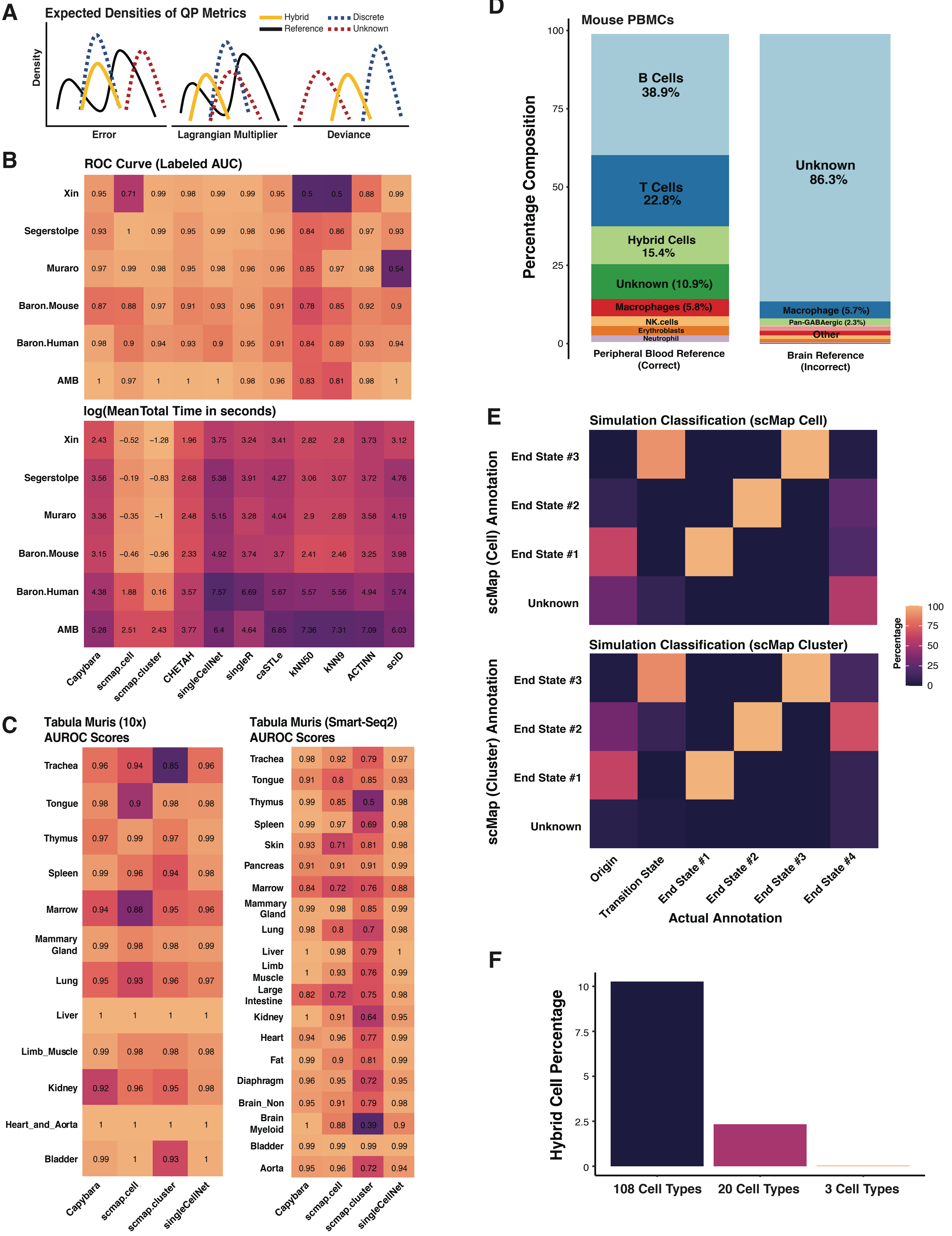
Supplemental Figure 4. Capybara analysis of direct cardiac reprogramming. Related to Figure 4 and Table S4. (A) Initial tissue-level classification of the Stone et al., 2019 dataset reveals four major tissues, which the high-resolution atlas restricts to two major tissues with the higher-resolution reference. **(B)** Normalized gene expression of cardiac markers labeling atrial vs. ventricular cardiomyocytes (****: P <=0.0001, Wilcoxon test). **(C)** Breakdown of cell types listed as "Other" in Figure 4. **(D)** Detailed hybrid cell-type breakdown for each time point of the Stone et al., cardiac reprogramming time course. **(E)** Integration of 10x dataset generated in this study with days 7 and 14 from the Stone et al., 2019 dataset, Cosine similarity = 0.804 between the independent studies. *Bottom right:* Projection of the two major cardiomyocyte populations, atrial and ventricular cardiomyocytes, onto the integrated UMAP. **(F)** RNA FISH of cells expressing only *Myh6* or *Myh7*. Scale bars = 50μm **(G)** Negative staining controls (MEFs) for RNA FISH and immunostaining. Scale bars = 10μm **(H)** Normalized gene expression of *Actc1* and *Tnnc1* across atrial, ventricular, and AV hybrids in the scRNA-seq data**. (I)** Immunofluorescence for MYL7 (atrial) and MYL2 (ventricular) proteins showing cells expressing a single protein (discrete) or co-expressing both proteins (hybrids) Scale bars = 10μm. **(J)** Quantification of discrete cells identified from immunostaining and scRNA-seq. MYL7-expressing cells are enriched relative to MYL2-expressing cells, confirming the atrial/ventricular bias observed from scRNA-seq data.

Supplemental Figure 5. Capybara analysis of motor neuron differentiation and programming. Related to Figure 5 and Table S5. (A) Capybara classification steps: with prior knowledge that these protocols aim to generate spinal motor neurons, we selected a single-cell spinal cord development atlas (Delile et al., 2019) as the high-resolution reference, omitting the general tissue selection step. We identified the major embryonic development stages corresponding to each protocol from Briggs et al., 2017. **(B)** UMAP embedding of this dataset, divided by protocol (Direct Programming: DP; Directed Differentiation: DD). *Top*: Projection of time points onto the UMAP embedding. *Bottom*: Projection of major discrete cell types, as identified via Capybara analysis, onto the UMAP embedding. **(C)** Transition scores for each protocol across experimental time points (****: P <=0.0001, n.s. = not significant, Wilcoxon test).

(D-E) Major hybrid populations in the direct programming and differentiation protocols. For each time point, we show the percentage of each hybrid type. **(F)** UMAP plot of the integrated datasets generated in this study, including four samples with different treatment groups. Expression of the motor neuron (MN) marker, *Mnx1* (left panel), and dorsal neuron marker, *Pou4f1* (right panel), are shown. **(G)** The major Capybara classification is shown for each cluster. **(H)** All discrete and hybrid cell type compositions for each treatment group.

Supplemental Figure 6. Capybara analysis of fibroblast to induced Endoderm Progenitor (iEP) Reprogramming. Related to Figure 6 and Table S6. (A) Tissue-level classification of iEP reprogramming (Biddy et al., 2018) reveals seven major tissues. With the high-resolution reference, three relevant tissues are selected. **(B)** Cell-type classification of long-term cultured iEPs using an expanded reference with embryonic populations. We selected a foregut organogenesis atlas (Han et al., 2020) and a gut tube development atlas (Nowotschin et al., 2018). We combined the two references and performed Capybara analysis. 99.9% of cells were classified as ‘unknown.’ We then combined this endoderm development atlas with the regenerative liver atlas and MCA: the iEPs primarily classified as injured BECs. **(C)** Hybrid populations in day 28 and long-term iEPs. **(D)** 3D-rendering of a microscopy z-stack for 3D-cultured iEPs stained for DAPI, EpCAM, and CK19, demonstrating branching. Scale bars = 50-100 μm. **(E)** Expression of other BEC markers on the UMAP embedding, including *Krt19, Sox9,* and *Cftr*. **(F)** Module scores comparing identified injured BECs, MEF/Stromal cells (LT-iEP Dataset), and primary BECs (Pepe-Mooney et al., 2019). BEC markers used for module scores are identified from Verhulst et al., 2019 and are listed in **Table S6**.

Supplemental Figure 1



ROC Curve (Labeled AUC)

	Capybara	scmap.cell	scmap.cluster	CHETAH	singleCellNet	singleR	caSTLe	kNN50	kNN9	ACTINN	scID
Xin	0.95	0.71	0.99	0.98	0.99	0.99	0.95	0.5	0.5	0.88	0.99
Segerstolpe	0.93	1	0.99	0.95	0.99	0.98	0.96	0.84	0.86	0.97	0.93
Muraro	0.97	0.99	0.98	0.95	0.98	0.96	0.96	0.85	0.97	0.98	0.54
Baron.Mouse	0.87	0.88	0.97	0.91	0.93	0.96	0.91	0.78	0.85	0.92	0.9
Baron.Human	0.98	0.9	0.94	0.93	0.9	0.95	0.91	0.84	0.89	0.93	0.94
AMB	1	0.97	1	1	1	0.98	0.96	0.83	0.81	0.98	1

log(MeanTotal Time in seconds)

	Capybara	scmap.cell	scmap.cluster	CHETAH	singleCellNet	singleR	caSTLe	kNN50	kNN9	ACTINN	scID
Xin	2.43	−0.52	−1.28	1.96	3.75	3.24	3.41	2.82	2.8	3.73	3.12
Segerstolpe	3.56	−0.19	−0.83	2.68	5.38	3.91	4.27	3.06	3.07	3.72	4.76
Muraro	3.36	−0.35	−1	2.48	5.15	3.28	4.04	2.9	2.89	3.58	4.19
Baron.Mouse	3.15	−0.46	−0.96	2.33	4.92	3.74	3.7	2.41	2.46	3.25	3.98
Baron.Human	4.38	1.88	0.16	3.57	7.57	6.69	5.67	5.57	5.56	4.94	5.74
AMB	5.28	2.51	2.43	3.77	6.4	4.64	6.85	7.36	7.31	7.09	6.03

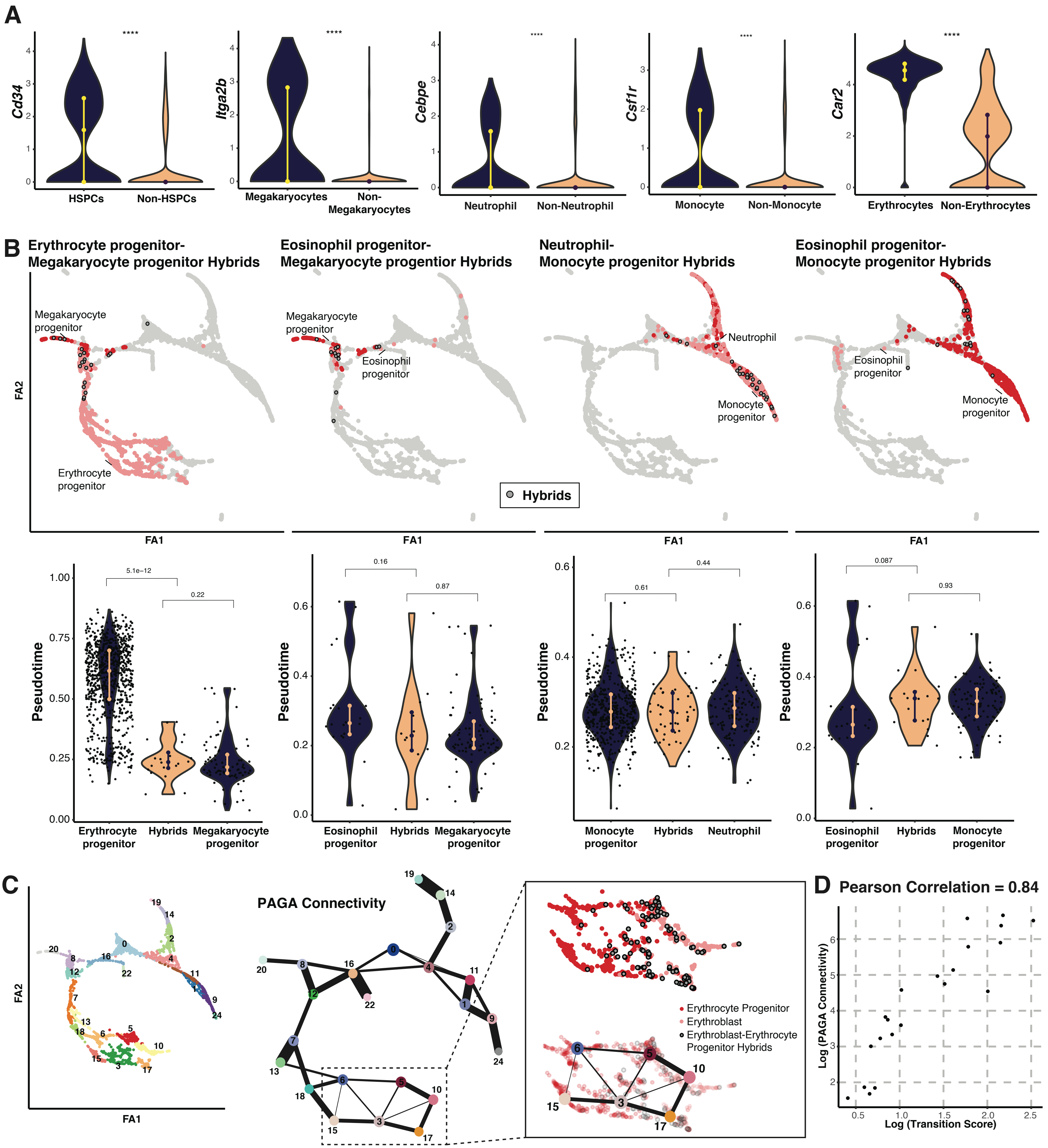
Tabula Muris (10x)
AUROC Scores

	Capybara	scmap.cell	scmap.cluster	singleCellNet
Trachea	0.96	0.94	0.85	0.96
Tongue	0.98	0.9	0.98	0.98
Thymus	0.97	0.99	0.97	0.99
Spleen	0.99	0.96	0.94	0.98
Marrow	0.94	0.88	0.95	0.96
Mammary Gland	0.99	0.98	0.98	0.99
Lung	0.95	0.93	0.96	0.97
Liver	1	1	1	1
Limb_Muscle	0.99	0.98	0.98	0.98
Kidney	0.92	0.96	0.95	0.98
Heart_and_Aorta	1	1	1	1
Bladder	0.99	1	0.93	1

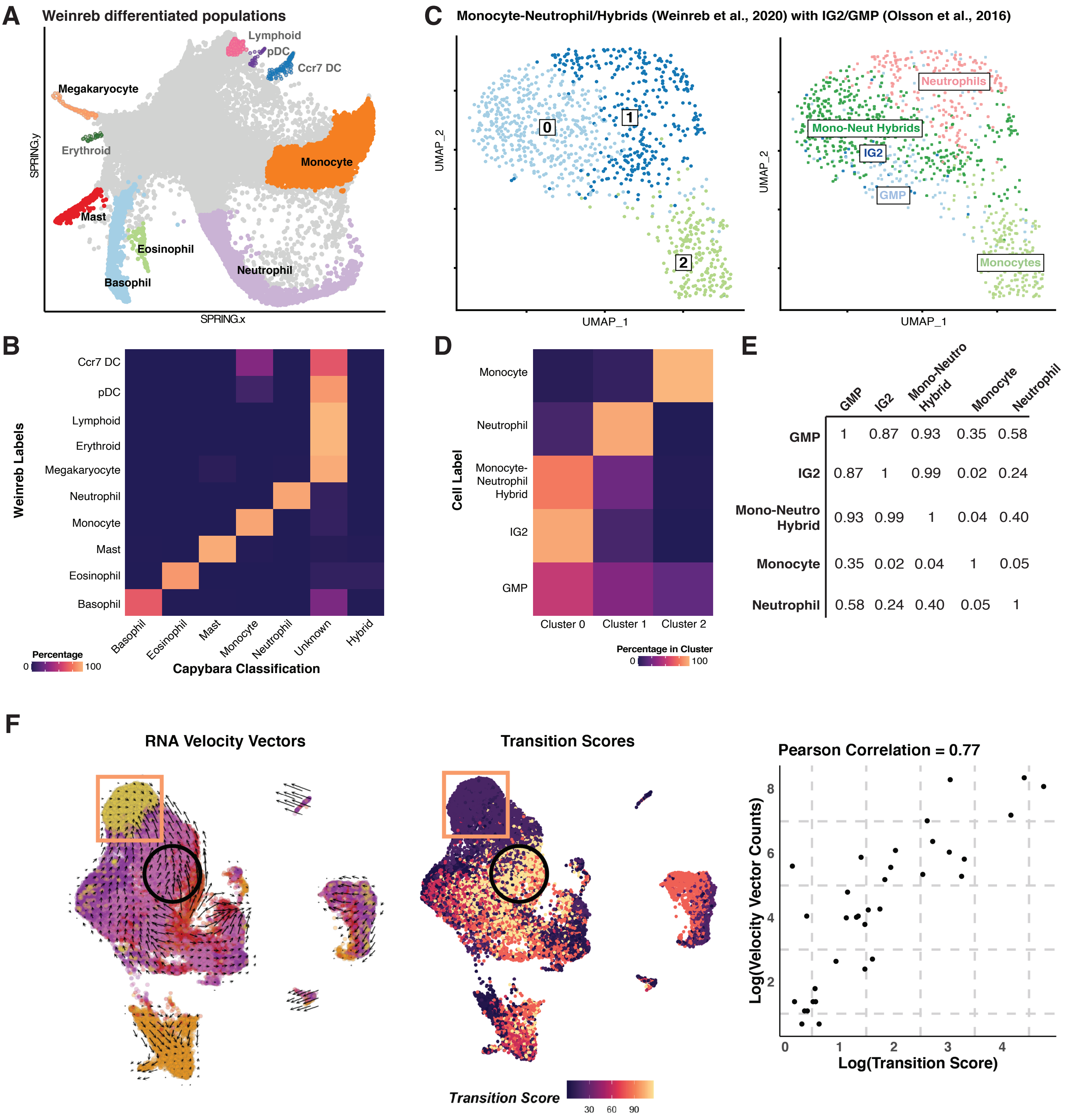
Tabula Muris (Smart-Seq2)
AUROC Scores

	Capybara	scmap.cell	scmap.cluster	singleCellNet
Trachea	0.98	0.92	0.79	0.97
Tongue	0.91	0.8	0.85	0.93
Thymus	0.99	0.85	0.5	0.98
Spleen	0.99	0.97	0.69	0.98
Skin	0.93	0.71	0.81	0.98
Pancreas	0.91	0.91	0.91	0.99
Marrow	0.84	0.72	0.76	0.88
Mammary Gland	0.99	0.98	0.85	0.99
Lung	0.98	0.8	0.7	0.98
Liver	1	0.98	0.79	1
Limb Muscle	1	0.93	0.76	0.99
Large Intestine	0.82	0.72	0.75	0.98
Kidney	1	0.91	0.64	0.95
Heart	0.94	0.96	0.77	0.99
Fat	0.99	0.9	0.81	0.99
Diaphragm	0.96	0.95	0.72	0.95
Brain_Non	0.95	0.91	0.79	0.98
Brain Myeloid	1	0.88	0.39	0.9
Bladder	0.99	0.99	0.99	0.99
Aorta	0.95	0.96	0.72	0.94

Supplemental Figure 2

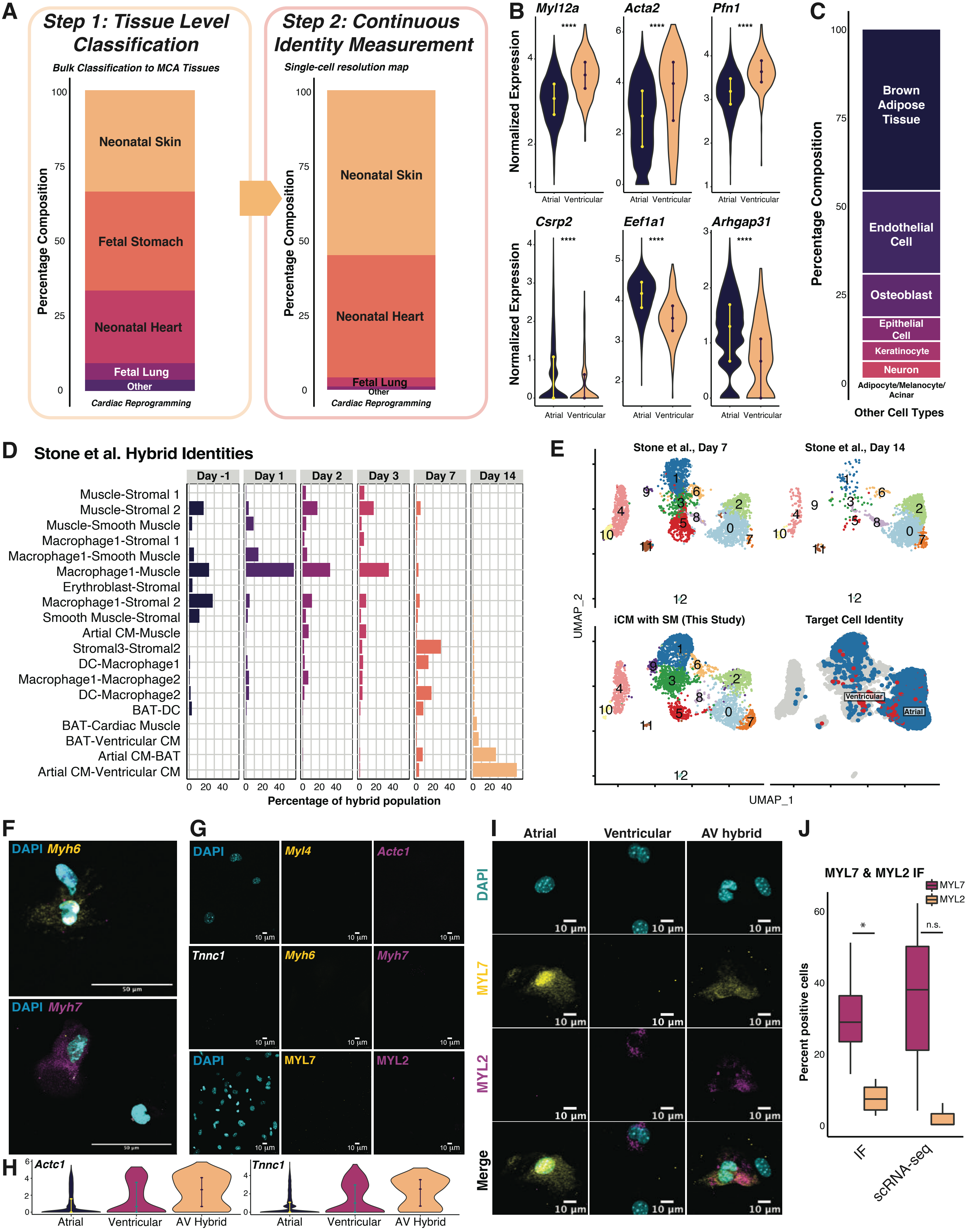


Supplemental Figure 3

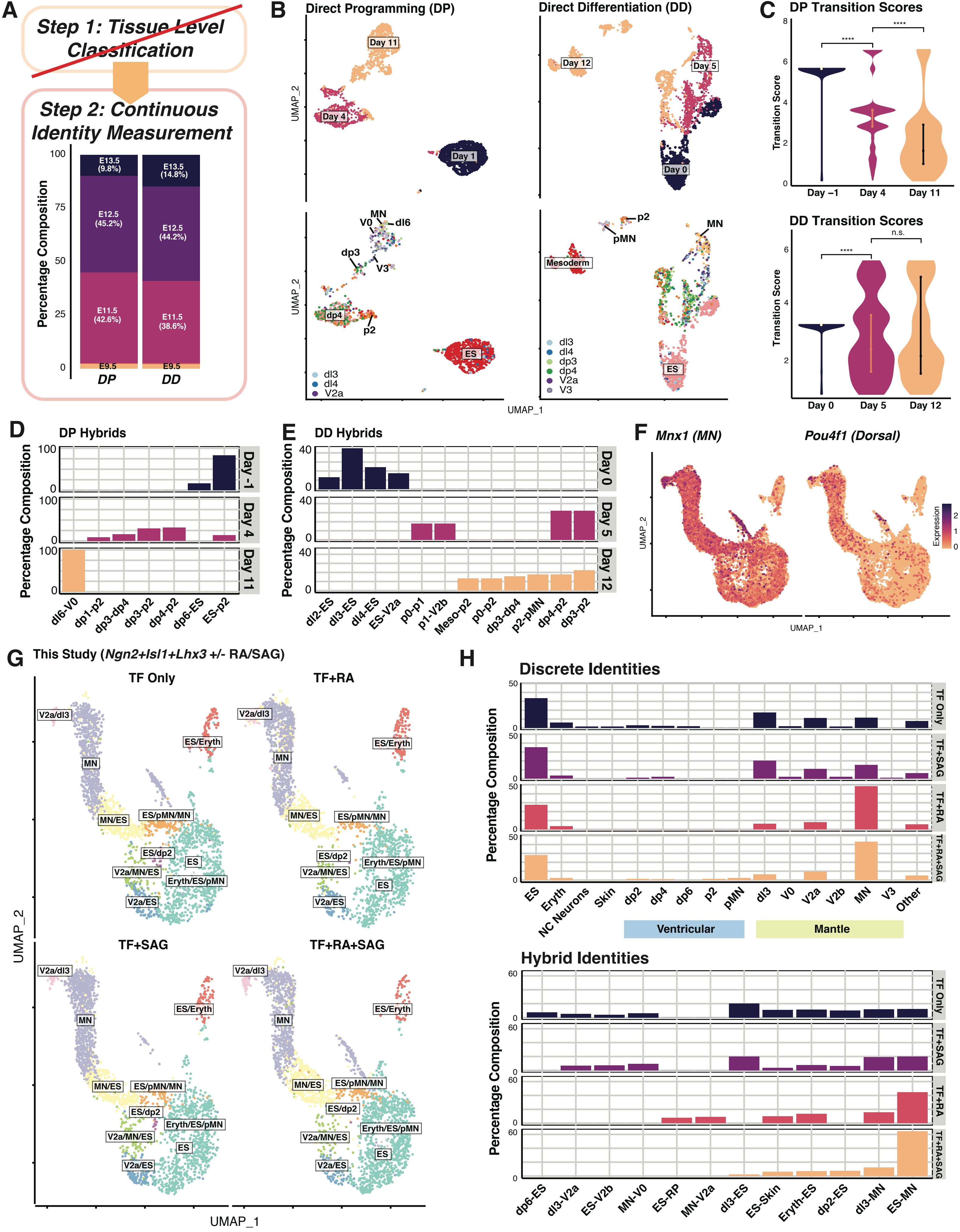


	GMP	IG2	Mono-Neutro Hybrid	Monocyte	Neutrophil
GMP	1	0.87	0.93	0.35	0.58
IG2	0.87	1	0.99	0.02	0.24
Mono-Neutro Hybrid	0.93	0.99	1	0.04	0.40
Monocyte	0.35	0.02	0.04	1	0.05
Neutrophil	0.58	0.24	0.40	0.05	1

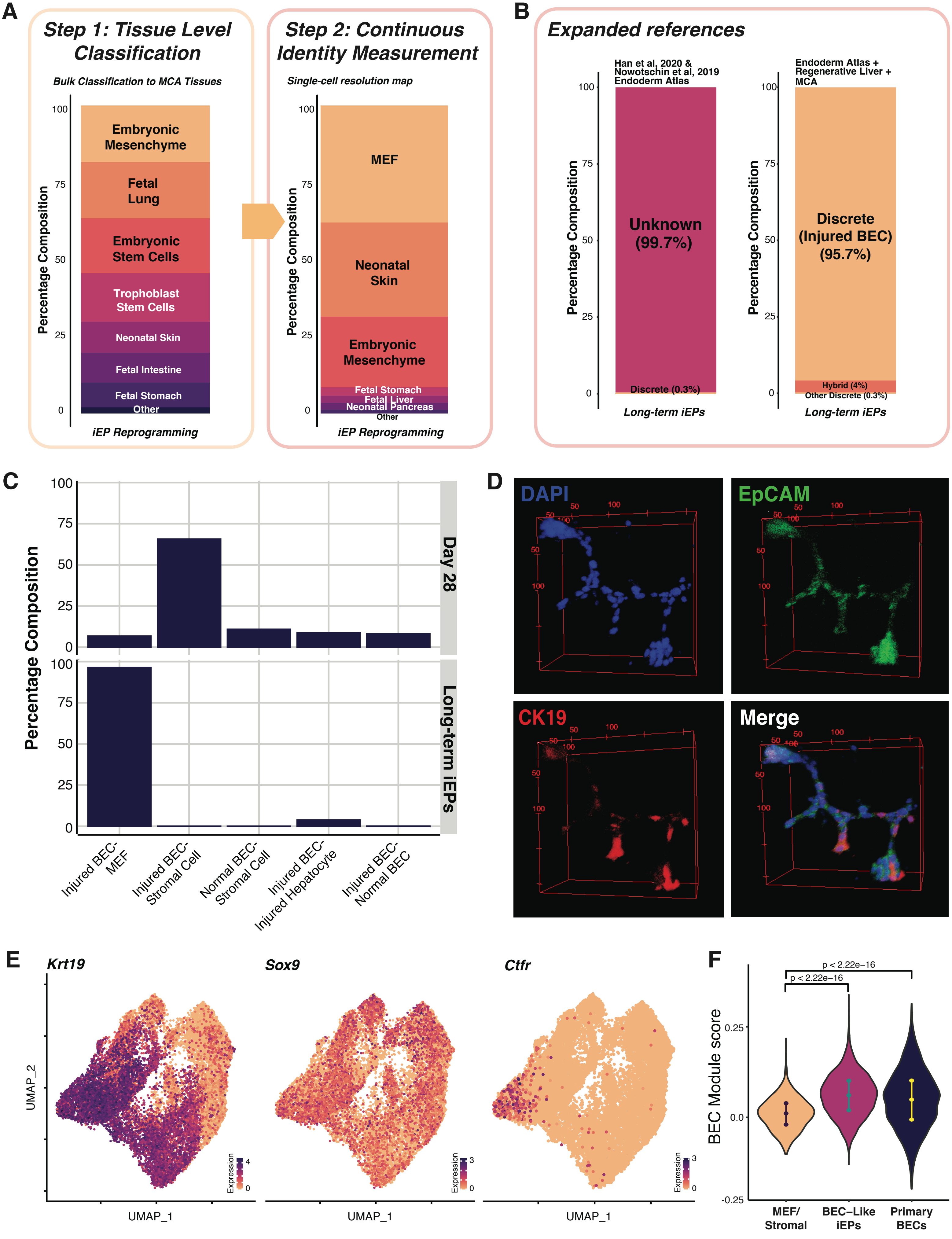
Supplemental Figure 4



Supplemental Figure 5



Supplemental Figure 6


A
Step 1: Tissue Level Classification
Bulk Classification to MCA Tissues
Percentage Composition
Embryonic Mesenchyme
Fetal Lung
Embryonic Stem Cells
Trophoblast Stem Cells
Neonatal Skin
Fetal Intestine
Fetal Stomach
Other
iEP Reprogramming
Step 2: Continuous Identity Measurement
Single-cell resolution map
Percentage Composition
MEF
Neonatal Skin
Embryonic Mesenchyme
Fetal Stomach
Fetal Liver
Neonatal Pancreas
Other
iEP Reprogramming
B
Expanded references
Han et al, 2020 & Nowotschin et al, 2019 Endoderm Atlas
Percentage Composition
Unknown (99.7%)
Discrete (0.3%)
Long-term iEPs
Endoderm Atlas + Regenerative Liver + MCA
Percentage Composition
Discrete (Injured BEC) (95.7%)
Hybrid (4%)
Other Discrete (0.3%)
Long-term iEPs
C
Percentage Composition
Day 28
Long-term iEPs
Injured BEC-MEF
Injured BEC-Stromal Cell
Normal BEC-Stromal Cell
Injured BEC-Injured Hepatocyte
Injured BEC-Normal BEC
D
DAPI
EpCAM
CK19
Merge
E
Krt19
Sox9
Ctfr
UMAP_2
UMAP_1
Expression
F
BEC Module score
p < 2.22e−16
p < 2.22e−16
MEF/Stromal
BEC−Like iEPs
Primary BECs