Cell Reports Methods, Volume *3*

Supplemental information

BOMA, a machine-learning framework for comparative gene expression analysis across brains and organoids

Chenfeng He, Noah Cohen Kalafut, Soraya O. Sandoval, Ryan Risgaard, Carissa L. Sirois, Chen Yang, Saniya Khullar, Marin Suzuki, Xiang Huang, Qiang Chang, Xinyu Zhao, Andre M.M. Sousa, and Daifeng Wang

Supplemental Items

Supplemental Tables

Table S1. Summary of transcriptome datasets included in this study. Related to Figures 2,3,4,5 and STAR Methods

ID	Paper	Time Range*	Number of cells	Source	Sequencing technology	Use in this work
1	Li et al. 2018	8PCW – 40 Y	Bulk	brain	mRNA-seq	● Identify brain tissue-wise co-expression gene modules (development module) ● Bulk RNA-seq alignment (Figure 1)
2	Polioudaks, et al. 2019	17-18PCW	40,000	brain	scRNA-seq	Identify cell-type specific genes for brain development
3	Nowakowski, et al. 2017	6PCW – 32PCW	4,261	brain	scRNA-seq	Human brain single-cell dataset 1 for large dataset alignment (Figure 5)
4	Trevino et al. 2021	16PCW, 20PCW, 21PCW, 24PCW	57,868	brain	scRNA-seq	Human brain single-cell dataset 2 for large dataset alignment (Figure 5)
5	Bhaduri et al. 2020	14PCW, 18PCW, 22PCW	136,254	brain	scRNA-seq	Human brain single-cell dataset 3 for large dataset alignment (Figure 5)
6	Gordon et al. 2021	50D,75D,100D,150D,200D,250D,300D,350D,400D,600D	Bulk	organoid	mRNA-seq	Bulk RNA-seq alignment (Figure 1)
7	Kanton et al. 2019	0D, 4D, 10D, 15D, 1M, 2M,4M	73,358	organoid	scRNA-seq	● Cross-species single-cell alignment (Figure 3) ● Organoid single-cell alignment in Figure 4;
8	Birey et al. 2017	105D	11,838	organoid	scRNA-seq	Organoid single-cell dataset 1 for large dataset alignment (Figure 5)
9	Bhaduri et al. 2020	3W, 5W,8W,10W	189,346	organoid	scRNA-seq	Organoid single-cell dataset 2 for large dataset alignment (Figure 5)

*Y: year; M: Month; W: week; D: day; PCW: Postconceptional Week

Table S2. Number of Gene Modules identified by WGCNA for each brain region. Related to Figure 2 and STAR Methods.

Brain region	Name	Number of Gene Modules
A1C	primary auditory(A1) cortex	33
AMY	amygdala	92
CBC	cerebella cortex	114
DFC	dorsolateral prefrontal cortex	96
HIP	Hippocampus	53
IPC	posterior inferior parietal cortex	68
ITC	inferior temporal cortex	37
M1C	primary motor(M1) cortex	17
MD	mediodorsal nucleus of the thalamus	54
MFC	medial prefrontal cortex	126
OFC	orbital prefrontal cortex	38
S1C	primary somatosensory(S1) cortex	99
STC	superior temporal cortex	95
STR	Stratum	44
V1C	primary visual(V1) cortex	109
VFC	ventrolateral prefrontal cortex	116

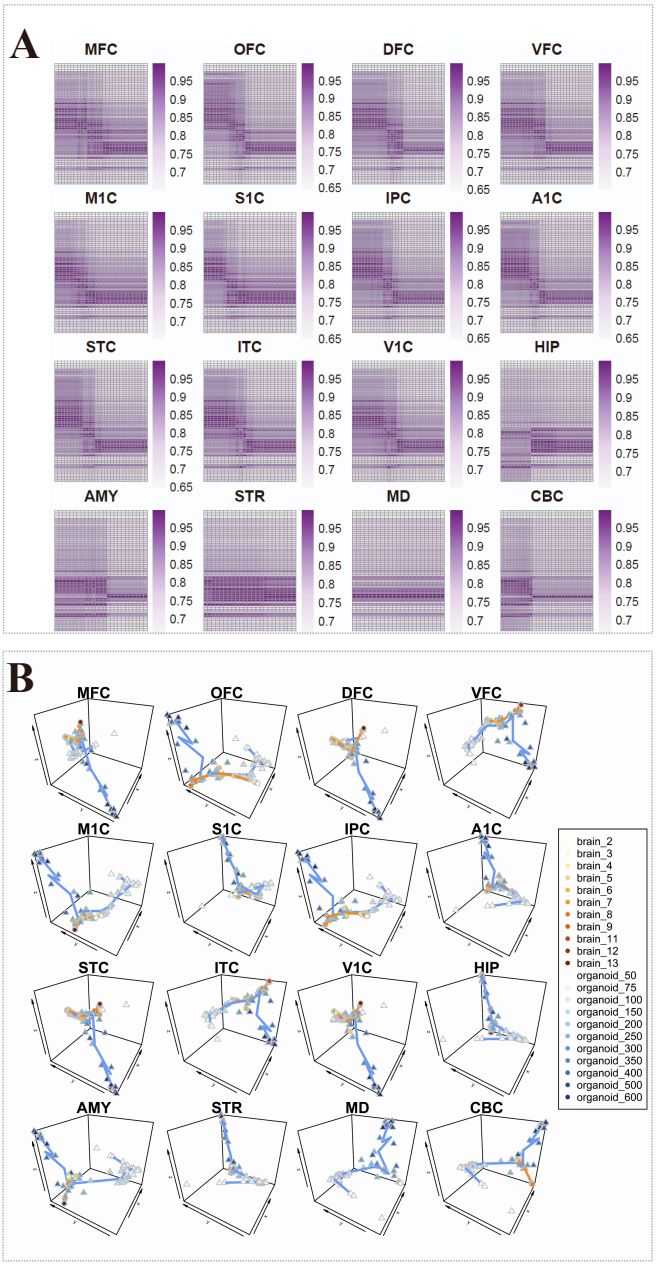
Table S3. Interpretation of cell-types annotated in Figure 4B. Related to Figure 4.

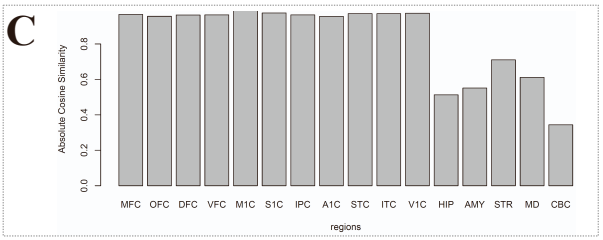
Cluster Name	Cluster Interpretation (Nowakowski, et al., 2017)
Astrocyte	Astocyte
Choroid	Choroid
Endothelial	Endothelial
EN-PFC	Early and Late Born Excitatory Neuron PFC
EN-V1	Early and Late Born Excitatory Neuron V1
Glyc	Glycolysis
IN-CTX-CGE	CGE/LGE-derived inhibitory neurons
IN-CTX-MGE	MGE-derived Ctx inhibitory neuron
IN-STR	Striatal neurons
IPC-div	Dividing Intermediate Progenitor Cells RG-like
IPC-Nen	Intermediate Progenitor Cells EN-like
MGE-div	dividing MGE Progenitors
MGE-IPC	MGE Progenitors
MGE-RG	MGE Radial Glia 1
Microglia	Micrgolia
Mural	Mural/Pericyte
nEN	Newborn Excitatory Neuron
nIN	MGE newborn neurons
OPC	Oligodendrocyte progenitor cell
RG	Radial Glia
U	Unknown early cells

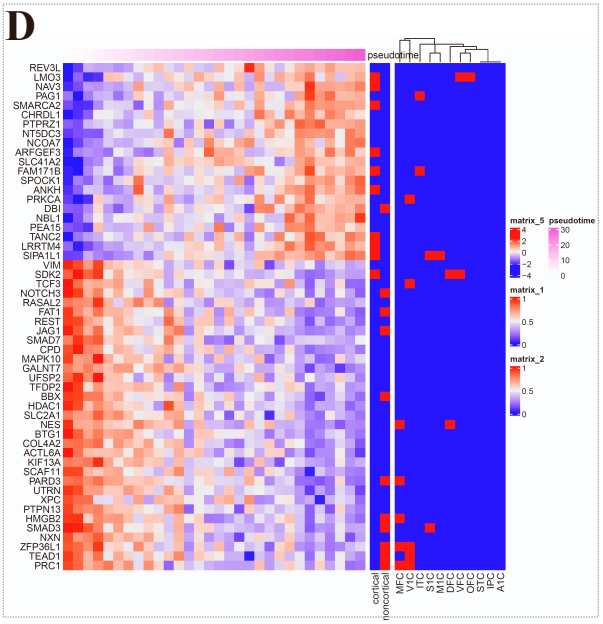
Table S4. Correspondence for sub-cell-types to be grouped into common major cell-types. Related to Figure 5.

Sub cell-type	Major cell-type
Bhaduri et al. 2020 (organoid), Dataset 5	
Astrocyte	Astro
ExcitatoryNeuron	EN
InhibitoryNeuron	IN
IPC	IPC
Outlier,Unknown	Unknown
RadialGlia	RG
Nowakowski et al. 2017, Dataset 3	
Astrocyte	Astro
Choroid	Choriod
EN-PFC1,EN-PFC2,EN-PFC3,EN-V1-1,EN-V1-2,EN-V1-3,nEN-early1,nEN-early2,nEN-late	EN
IN-CTX-CGE1,IN-CTX-CGE2,IN-CTX-MGE1,IN-CTX-MGE2,IN-STR,nIN1,nIN2,nIN3,nIN4,nIN5	IN
IPC-div1,IPC-div2,IPC-nEN1,IPC-nEN2,IPC-nEN3,MGE-IPC1,MGE-IPC2,MGE-IPC3	IPC
MGE-RG1,MGE-RG2,oRG,RG-div1,RG-div2,RG-early,tRG,vRG	RG
Microglia	Microglia
Endothelial	Endothelial
Mural	Mural
OPC	OPC
Trevino et al. 2021, Dataset 4	
CGEIN,MGEIN	IN
earlyRG,lateRG,tRG	RG
EC	Endothelial
GluN1,GluN2,GluN3,GluN4,GluN5,GluN6,GluN7,GluN8	EN
MG	Microglia
OPC_Oligo	OPC
Peric	Mural
Bhaduri et al. 2020 (brain), Dataset 9	
Endothelial	Endothelial
ExcitatoryNeuron	EN
InhibitoryNeuron	IN
IPC	IPC
RadialGlia	RG
Outlier,Red blood cells	Unknown
Mural	Mural
Microglia	Microglia

Supplemental Figures







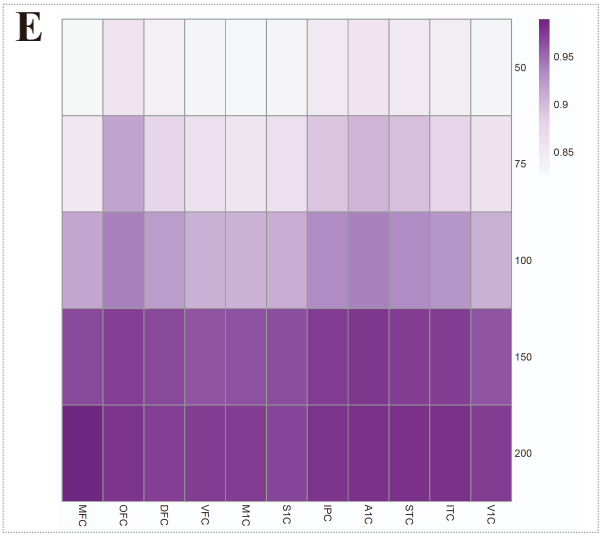


Figure S1. Alignment and scores between cortical organoids with all 16 brain regions across time. Related to Figure 2. A). Local alignment scores between organoids with 16 brain regions; B). 3D scatterplot of the alignment. Brain samples were colored red while organoids were colored blue. The color gradients represent the developing time, with the later time-point being darker colors. Blue/red curves link the averaged coordinates of each timepoint to represent the developing trajectory; C). Cosine similarity between trajectories of brains (stage 3-13) and organoids (day 75-300). D). Pseudotime-correlated genes potentially drive the alignment between brain regions and organoid in Figure 2. Organoids genes expression were correlated with the pseudotime (z-axis in Fig2A), genes with absolute Pearson correlation > 0.7 were shown in the heatmap. Brain cortical markers (N=51) and non-cortical markers (N=75) were identified by selecting genes significantly upregulated (log2FoldChange>0.7 and p-value<0.05) between stages 3-6. Similarly, individual cortical area markers (N=136) were identified by comparing each area with all the other cortical areas; E). Averaged local alignment scores between organoids versus the 11 cortical regions. S1C: primary somatosensory(S1) cortex; M1C: primary motor(M1) cortex; OFC: orbital prefrontal cortex; DFC: dorsolateral prefrontal cortex; MFC: medial prefrontal cortex; VFC: ventrolateral prefrontal cortex; STR: stratum; HIP: hippocampus; AMY: amygdala; IPC: posterior inferior parietal cortex; A1C: primary auditory(A1) cortex; V1C: primary visual(V1) cortex; STC: superior temporal cortex; ITC: inferior temporal cortex; MD: mediodorsal nucleus of the thalamus; CBC: cerebella cortex.

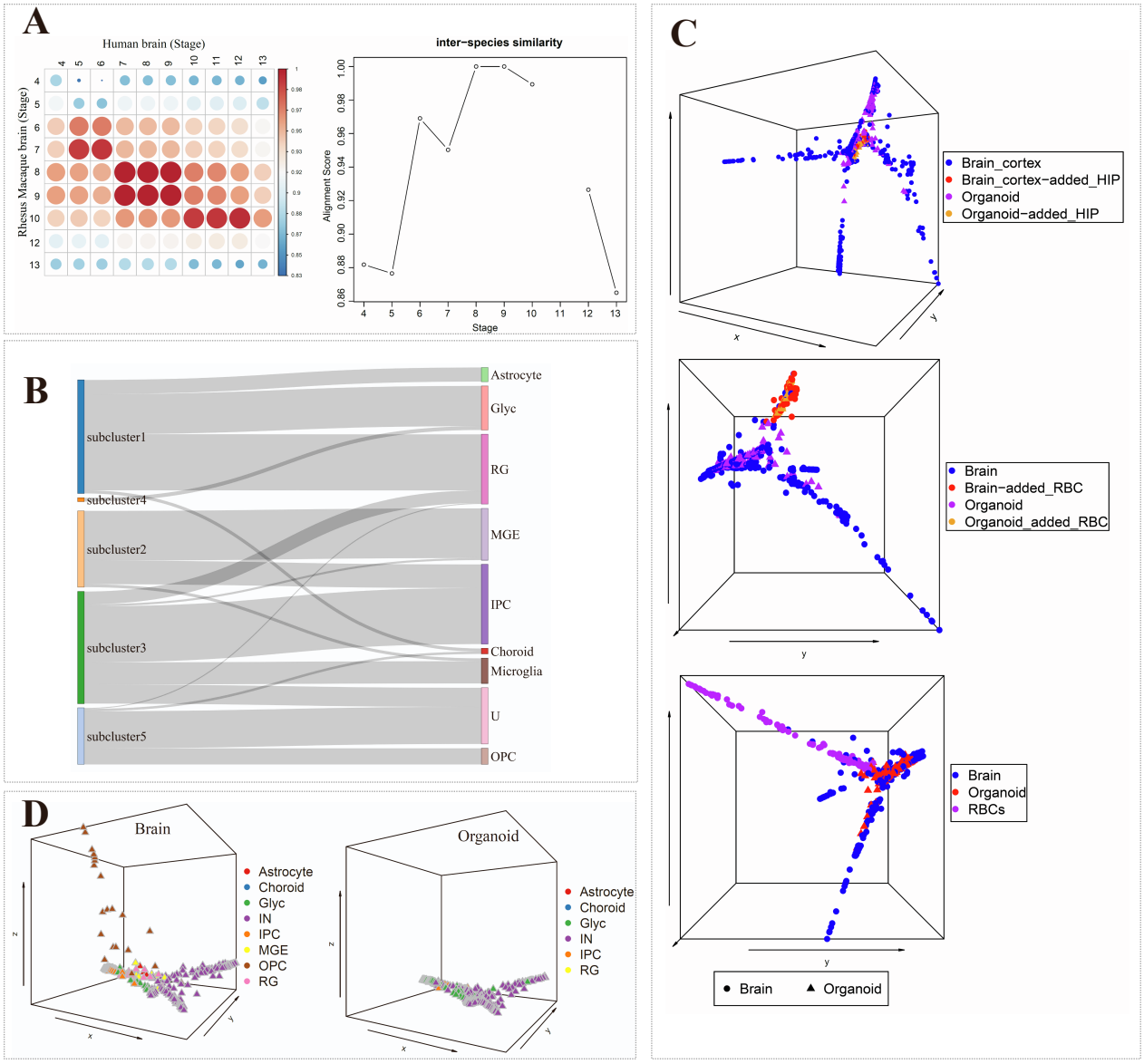
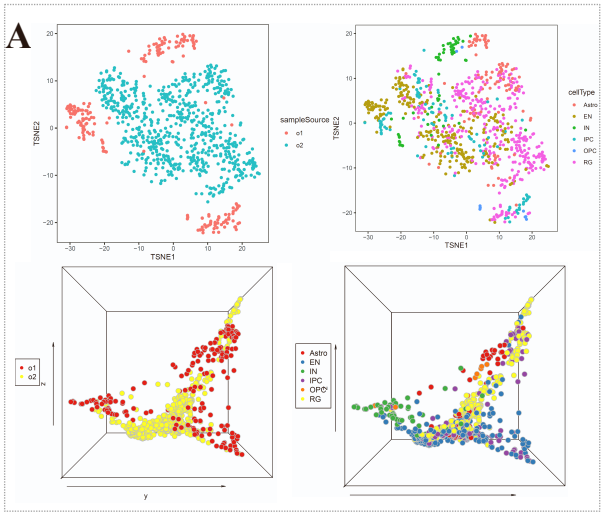
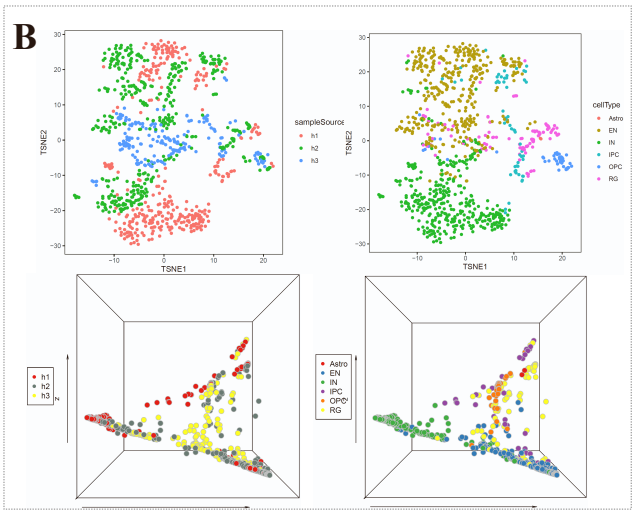
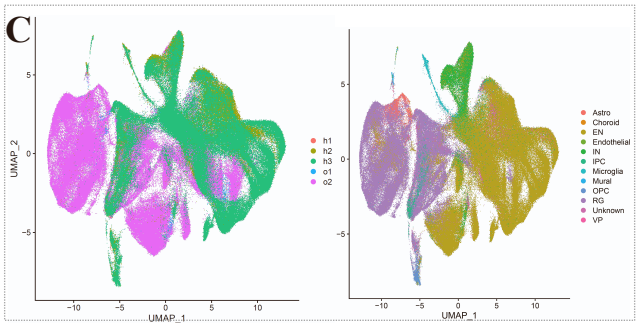
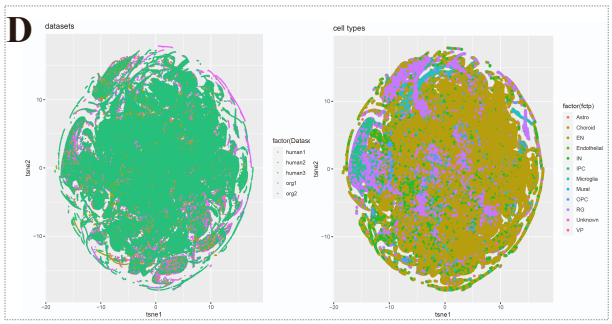

A
Human brain (Stage)
Rhesus Macaque brain (Stage)
inter−species similarity
Alignment Score
Stage
B
subcluster1
subcluster4
subcluster2
subcluster3
subcluster5
Astrocyte
Glyc
RG
MGE
IPC
Choroid
Microglia
U
OPC
C
Brain_cortex
Brain_cortex−added_HIP
Organoid
Organoid−added_HIP
Brain
Brain−added_RBC
Organoid
Organoid_added_RBC
Brain
Organoid
RBCs
Brain
Organoid
D
Brain
Organoid
Astrocyte
Choroid
Glyc
IN
IPC
MGE
OPC
RG

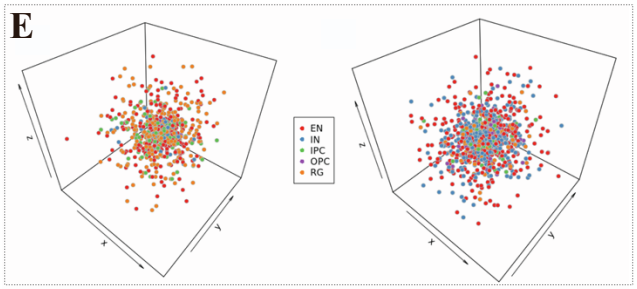
Figure S2. Evaluate and benchmark the robustness of BOMA. Related to Figure 3 and 4. A). Human VERSUS Macaque brain. The predicted developmental stages (Zhu Y. et.al., 2018) for Macaque brains was used for the comparative analysis; B).Subclusters of Cluster 1 in Figure 4. Each subcluster is specifically enriched with distinct cell-types. For example, subcluster2 is enriched with MGE and IPC; C). BOMA alignment with intentionally inserted mismatched region or cell-type in Figure 4. Top: BOMA alignment with intentionally added 6,676 Hippocampus (HIP) cells. HIP cells were equally split and inserted into brain cortical and organoid data. After rerunning the data processing and BOMA alignment, organoid cells were aligned with cortical cells but not hippocampal cells. Middle: BOMA alignment with intentionally added 2,170 red blood cells (RBCs). RBCs were equally split and inserted into brain cortical and organoid data in Figure 4. After rerunning the data processing and BOMA alignment, brain and organoid were aligned, while RBCs form a separate branch. Bottom: Alignment with RBCs only inserted to brain data. RBCs also form a separate developing branch in the manifold space. The Hippocampal cells and RBCs were obtained from (Bhaduri et al., 2020)[18]; D). BOMA alignment after intentionally leaving Excitatory Neurons (ExN) out from both Brain and Organoids in Figure 4. The same cell-types from brain and organoid can be well aligned.

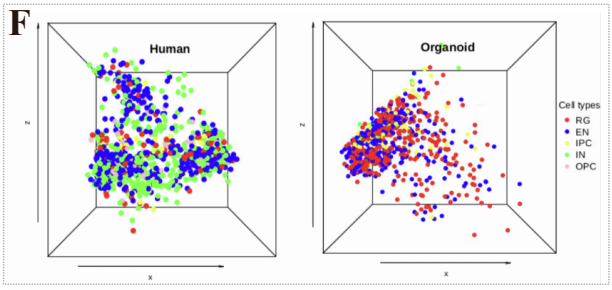

A
TSNE1
TSNE2
sampleSource
o1
o2
cellType
Astro
EN
IN
IPC
OPC
RG

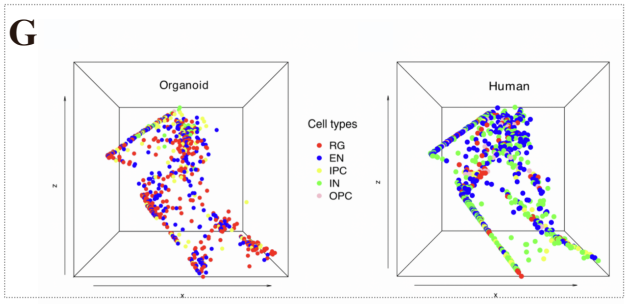

B
TSNE1
TSNE2
sampleSource
h1
h2
h3
cellType
Astro
EN
IN
IPC
OPC
RG


C
UMAP_1
UMAP_2
h1
h2
h3
o1
o2
Astro
Choroid
EN
Endothelial
IN
IPC
Microglia
Mural
OPC
RG
Unknown
VP


D
datasets
cell types
tsne1
tsne2


E
EN
IN
IPC
OPC
RG


F
Human
Organoid
Cell types
RG
EN
IPC
IN
OPC


G
Organoid
Human
Cell types
RG
EN
IPC
IN
OPC

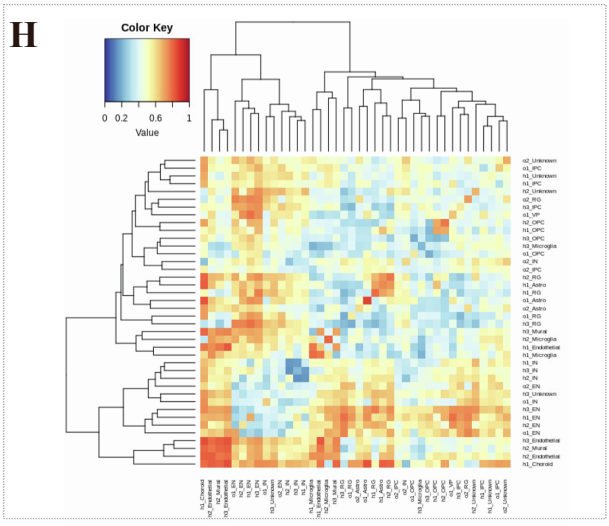

H
Color Key
0 0.2 0.6 1
Value

Figure S3. Benchmark BOMA using brain and organoid datasets integrated from multiple studies. Related to Figure 5. A,B). Minimum batch effects across merged organoid samples (A) and brain samples (B) were observed by tSNE and BOMA. Top left: tSNE analysis of pseudo-cells colored by the corresponding sample sources dataset. Top right: tSNE analysis of pseudo-cells colored by the corresponding cell-types. Bottom left: 3D scatterplot of pseudo-cells colored by sample sources after BOMA; Bottom right: 3D scatterplot of pseudo-cells colored by cell-types after BOMA; C). Human brain and organoid single-cells integrated by Seurat(C) and Liger(D). Left: UMAP colored by datasets. Right: UMAP colored by cell-types; E,F,G). Human brain and organoid pseudo-cells aligned by MMD-MA (E), SCOT (F) and Unioncom (G). Left: 3D scatter plots show the alignment results for 766 organoid pseudo-cells. Right: 3D scatter plots show the alignment results for 1,016 human brain pseudo-cells. The dots are colored by the cell-types. RG cells are marked in red, EN cells are marked in blue, IPC cells are marked in yellow, IN cells are marked in green and OPC cells are marked in pink; H). Cell-types similarity across scRNA-seq datsets quantified by MetaNeighbor. Heatmap generated by applying MetaNeighbor on 10% randomly sampled single-cells from integrated scRNA-seq dataset in Figure 5. The datasets are h1(human1, Nowakowski 2017), h2(human2, Trevino et al. 2021), h3(human3, Bhaduri et al. 2020), o1(organoid1, Birey et al. 2017) and o2(organoid2, Bhaduri et al. 2020).

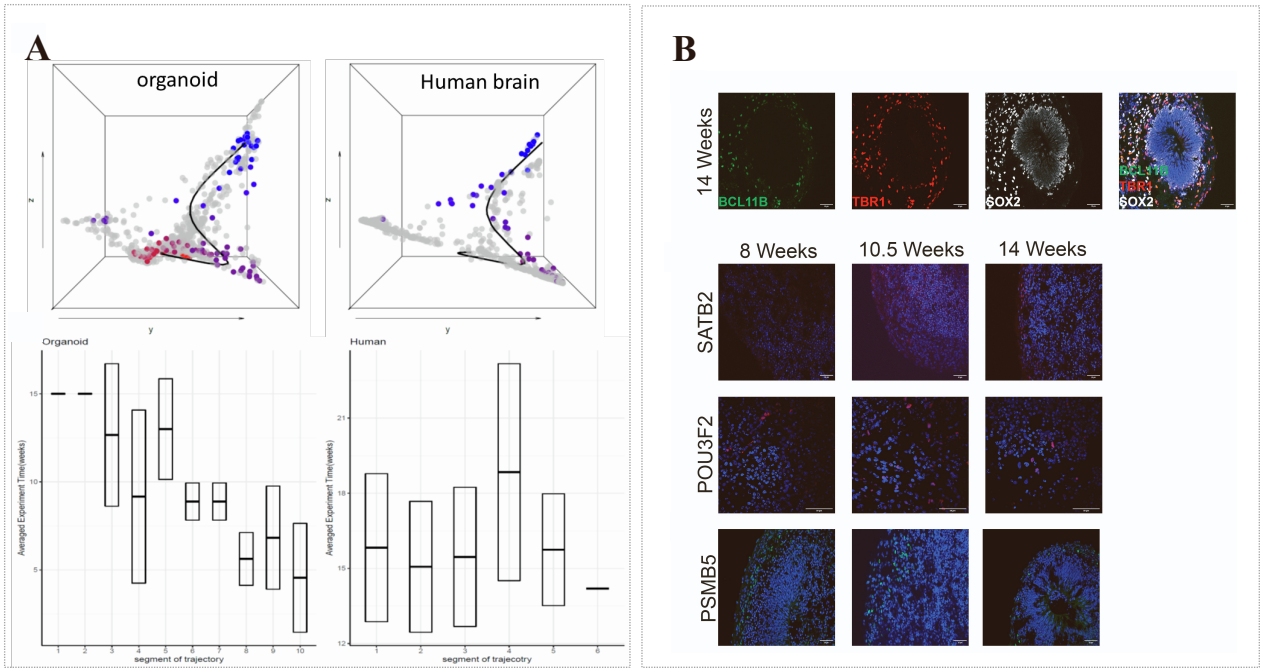
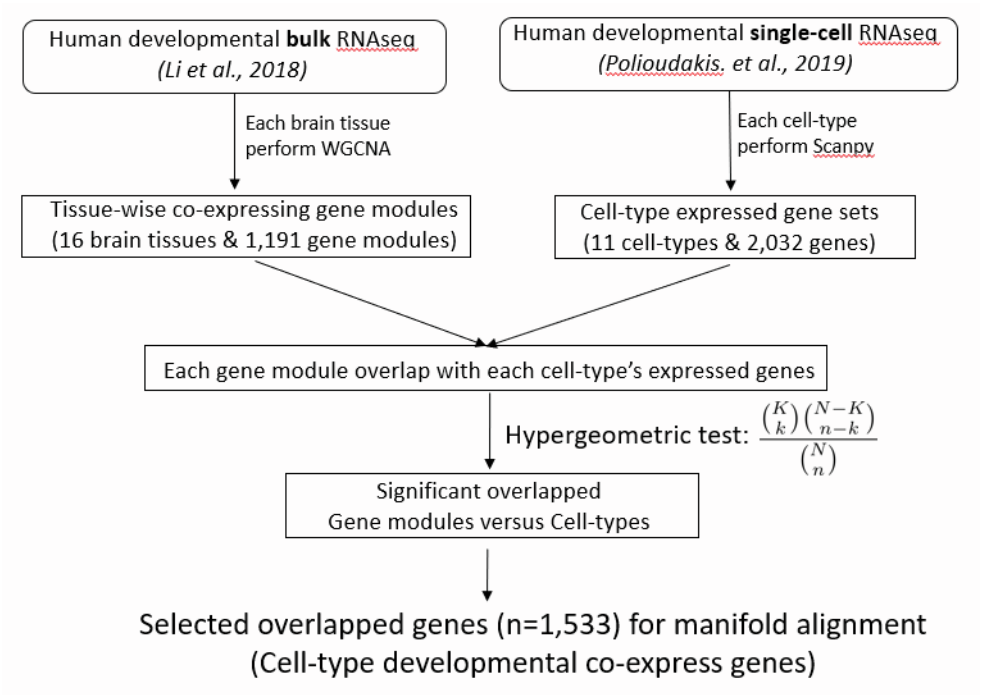
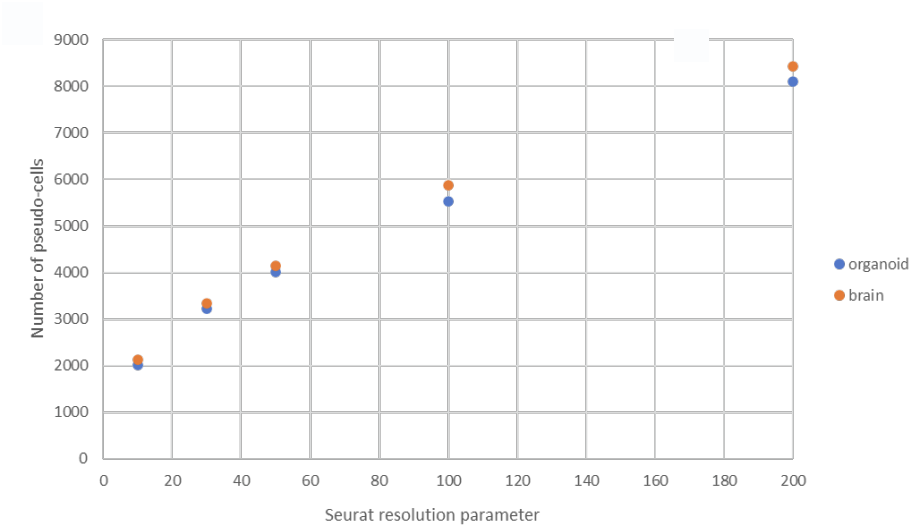


Figure S4. Trajectory analysis for Intermediate Progenitor Cells (IPCs) and experimental validation. Related to Figure 5 and 6. A). Trajectory analysis for IPCs. 3D scatter plot shows the inferred trajectory of IPC from the alignment. Organoid (top left) and brain (top right) cells are colored by pseudo-time along the trajectory. Blue means earlier while red means later. The trajectory was divided into 10 segments and correlated with experiment timepoints. The box plot (Bottom) shows the distribution of experimental time-points; B). Representative images of human cortical organoids immunostaining. Top: Image shows expressing progenitor cells positive for SOX2, deep layer cortical neurons positive for BCL11B and TBR1. Bottom: Immunostaining of different stages of differentiation in cortical organoids positive for SATB2, POU3F2 and PSMB5 cells. Quantification is shown in Figures 6C, 6F, and 6I. Scale bars, 50µm.

A


Human developmental **bulk** RNAseq
(Li et al., 2018)
Human developmental **single-cell** RNAseq
(Polioudakis. et al., 2019)
Each brain tissue
perform WGCNA
Each cell-type
perform Scanpy
Tissue-wise co-expressing gene modules
(16 brain tissues & 1,191 gene modules)
Cell-type expressed gene sets
(11 cell-types & 2,032 genes)
Each gene module overlap with each cell-type's expressed genes
Hypergeometric test: $\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$
Significant overlapped
Gene modules versus Cell-types
Selected overlapped genes (n=1,533) for manifold alignment
(Cell-type developmental co-express genes)

B


Number of pseudo-cells
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
120
140
160
180
200
Seurat resolution parameter
organoid
brain

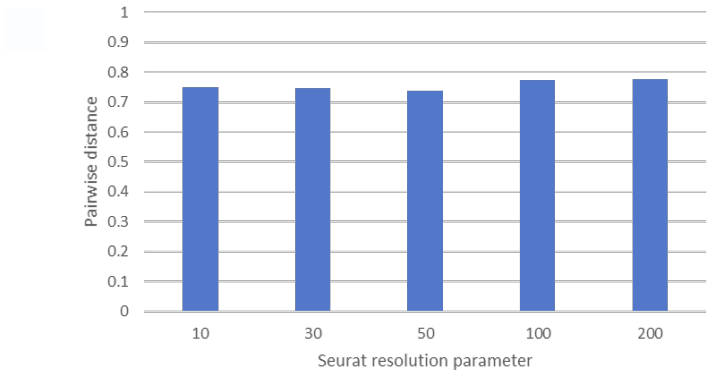

Pairwise distance
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
30
50
100
200
Seurat resolution parameter

Figure S5. Selecting human brain developmental genes and the scalability of BOMA to the pseudo-cell numbers. Related to STAR Methods. A). Flowchart of selecting human brain developmental genes. BH adjusted P-value <0.01 was used as statistical significance threshold. B). BOMA is scalable to the number of pseudo-cells. Top: Number of pseudo-cells generated at different Seurat resolution; Bottom: Averaged pairwise distances between pseudo-cells of the same cell-type.