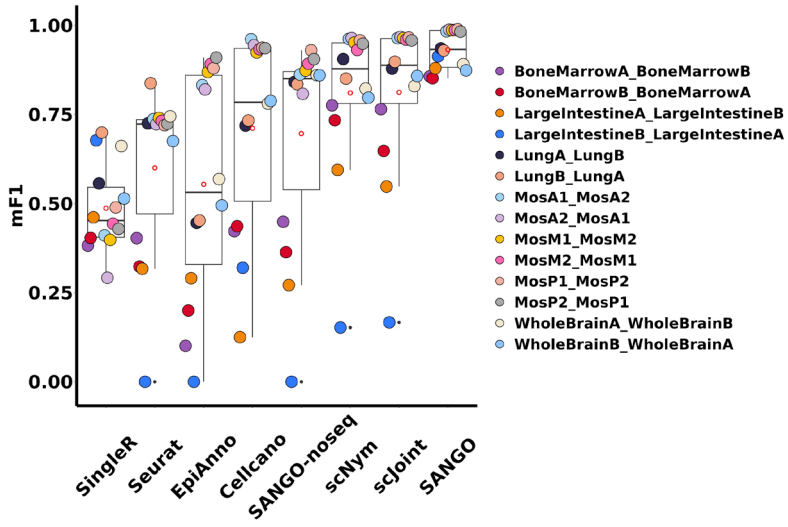
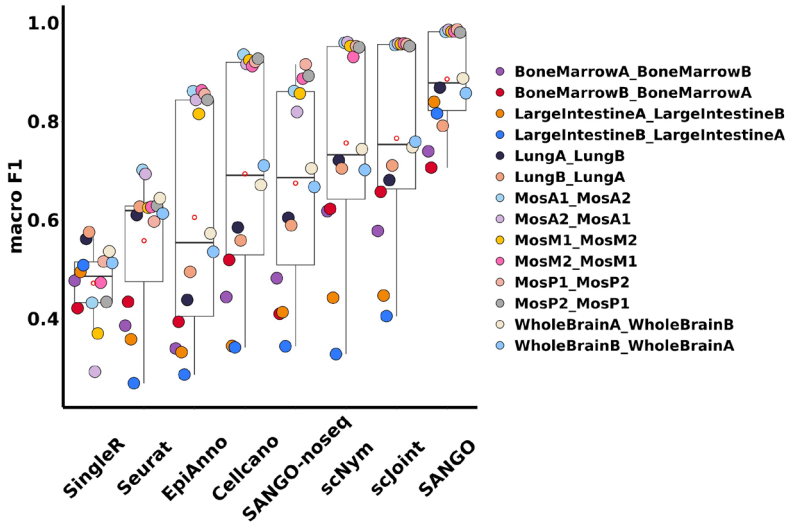
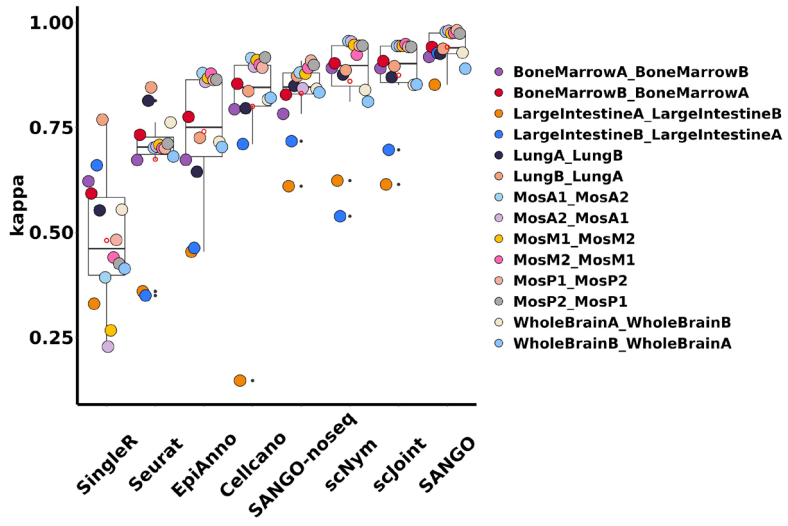
(a)


mF1
1.00
0.75
0.50
0.25
0.00
SingleR
Seurat
EpiAnno
Cellcano
SANGO-noseq
scNym
scJoint
SANGO
BoneMarrowA_BoneMarrowB
BoneMarrowB_BoneMarrowA
LargeIntestineA_LargeIntestineB
LargeIntestineB_LargeIntestineA
LungA_LungB
LungB_LungA
MosA1_MosA2
MosA2_MosA1
MosM1_MosM2
MosM2_MosM1
MosP1_MosP2
MosP2_MosP1
WholeBrainA_WholeBrainB
WholeBrainB_WholeBrainA

(b)

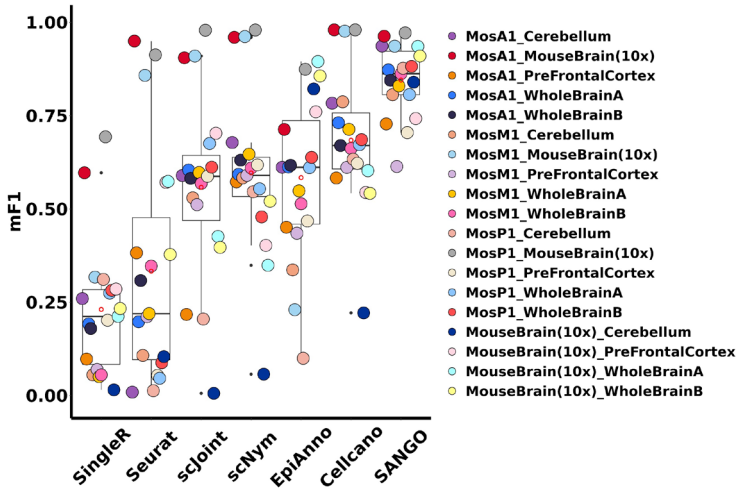

macro F1
1.0
0.8
0.6
0.4
SingleR
Seurat
EpiAnno
Cellcano
SANGO-noseq
scNym
scJoint
SANGO
BoneMarrowA_BoneMarrowB
BoneMarrowB_BoneMarrowA
LargeIntestineA_LargeIntestineB
LargeIntestineB_LargeIntestineA
LungA_LungB
LungB_LungA
MosA1_MosA2
MosA2_MosA1
MosM1_MosM2
MosM2_MosM1
MosP1_MosP2
MosP2_MosP1
WholeBrainA_WholeBrainB
WholeBrainB_WholeBrainA

(c)

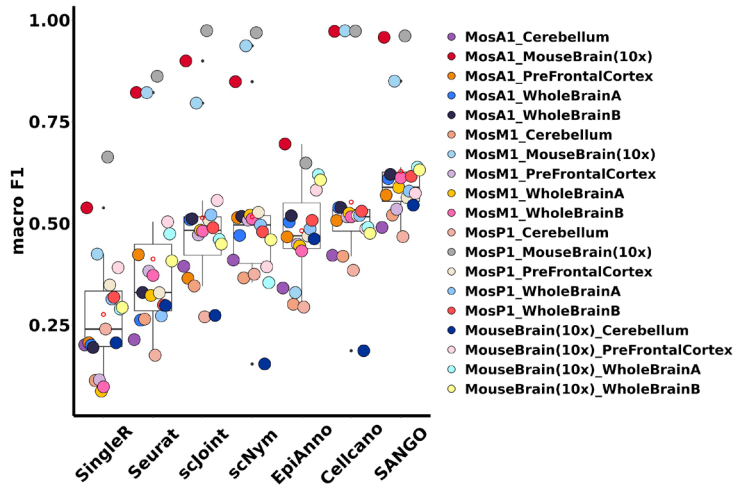

kappa
1.00
0.75
0.50
0.25
SingleR
Seurat
EpiAnno
Cellcano
SANGO-noseq
scNym
scJoint
SANGO
BoneMarrowA_BoneMarrowB
BoneMarrowB_BoneMarrowA
LargeIntestineA_LargeIntestineB
LargeIntestineB_LargeIntestineA
LungA_LungB
LungB_LungA
MosA1_MosA2
MosA2_MosA1
MosM1_MosM2
MosM2_MosM1
MosP1_MosP2
MosP2_MosP1
WholeBrainA_WholeBrainB
WholeBrainB_WholeBrainA

Supplementary figure 1. (a) Boxplots summarize the mF1 scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges) and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=14 biologically independent paired intra-datasets. (b) Boxplots summarize the macro F1 scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=14 biologically independent paired intra-datasets. (c) Boxplots summarize the Cohen's kappa scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=14 biologically independent paired intra-datasets. For all subfigures, the x-axis represents the various methods, while the y-axis denotes the measured values. The hollow red dot within the boxplot represents the average values, while black dots denote outliers.

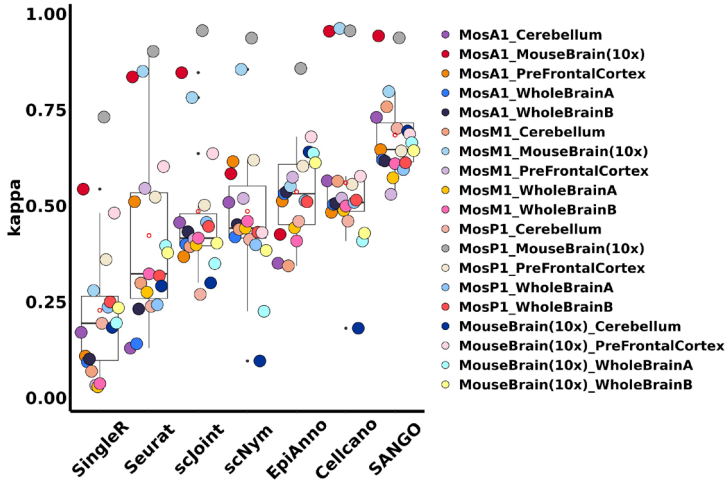
(a)



(b)

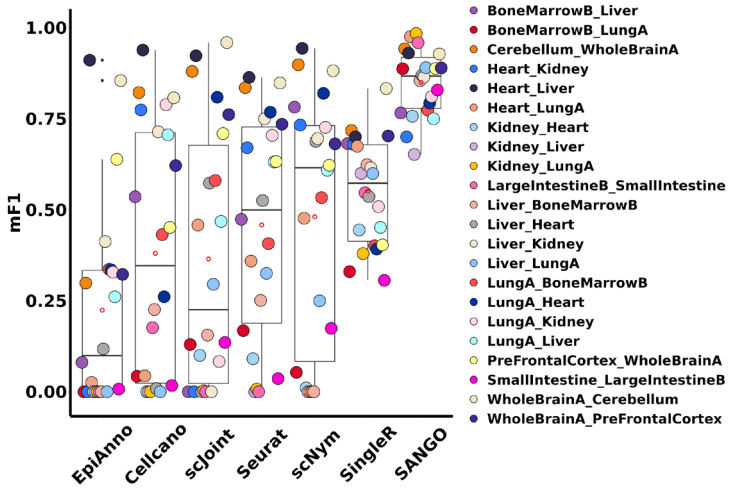


(c)

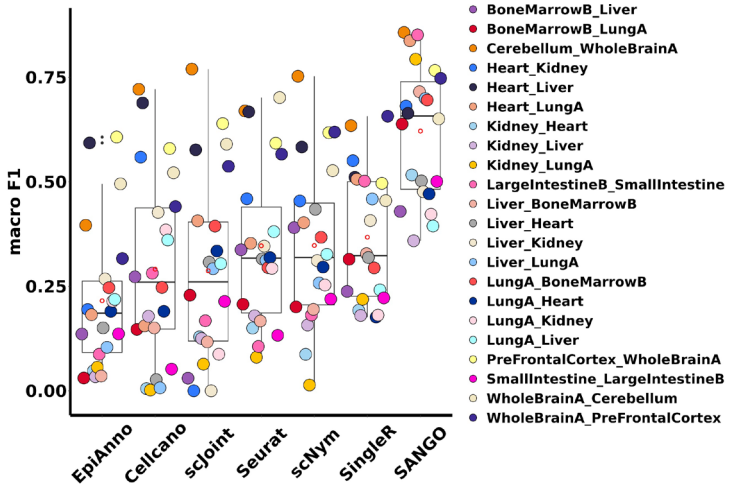


Supplementary figure 2. (a) Boxplots summarize the mF1 scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges) and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=19 biologically independent paired cross-platform datasets. (b) Boxplots summarize the macro F1 scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=19 biologically independent paired cross-platform datasets. (c) Boxplots summarize the Cohen's kappa scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=19 biologically independent paired cross-platform datasets. For all subfigures, the x-axis represents the various methods, while the y-axis denotes the measured values. The hollow red dot within the boxplot represents the average values, while black dots denote outliers.

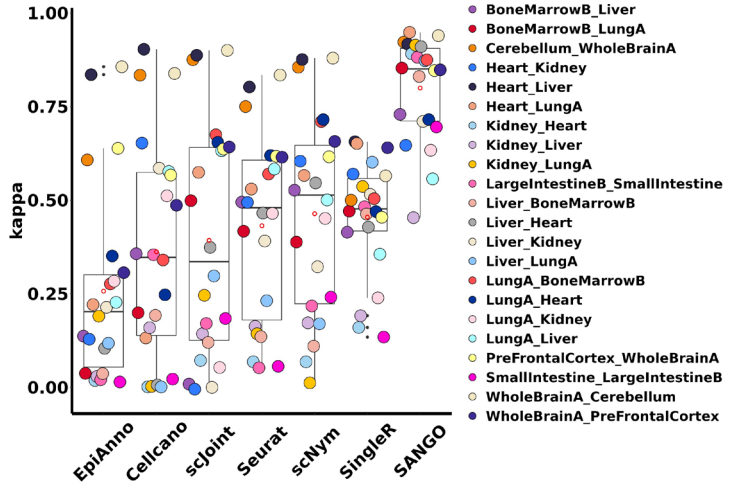
(a)



(b)



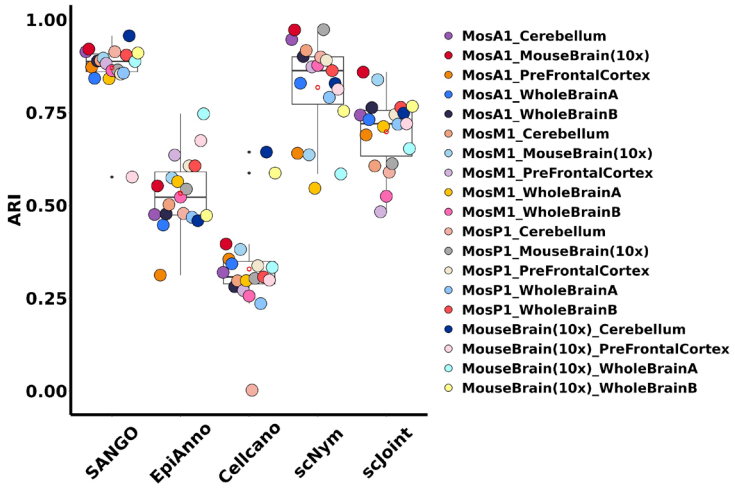
(c)



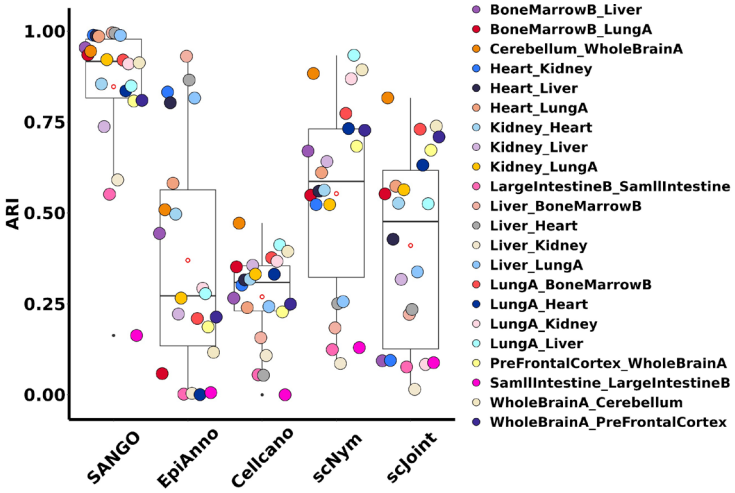
Supplementary figure 3. (a) Boxplots summarize the mF1 scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges) and 1.5 times the interquartile range

(whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=22 biologically independent paired cross-tissue datasets. (b) Boxplots summarize the macro F1 scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=22 biologically independent paired cross-tissue datasets. (c) Boxplots summarize the Cohen's kappa scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=22 biologically independent paired cross-tissue datasets. For all subfigures, the x-axis represents the various methods, while the y-axis denotes the measured values. The hollow red dot within the boxplot represents the average values, while black dots denote outliers.

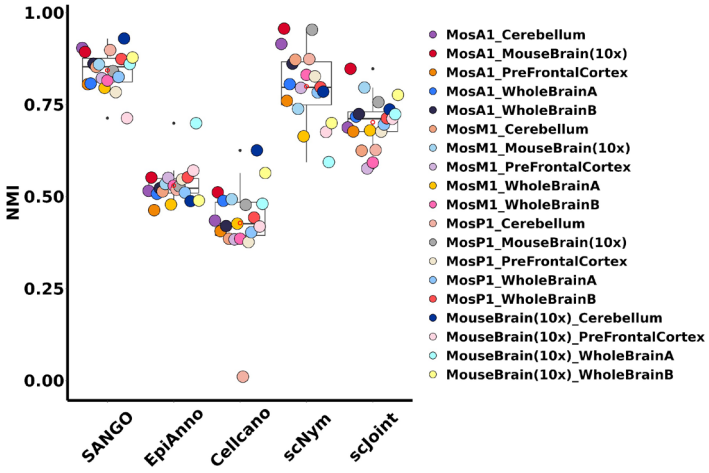
(a)



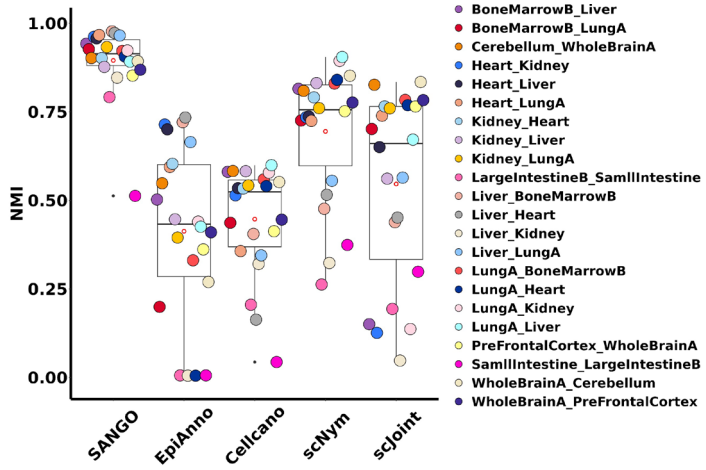
(b)



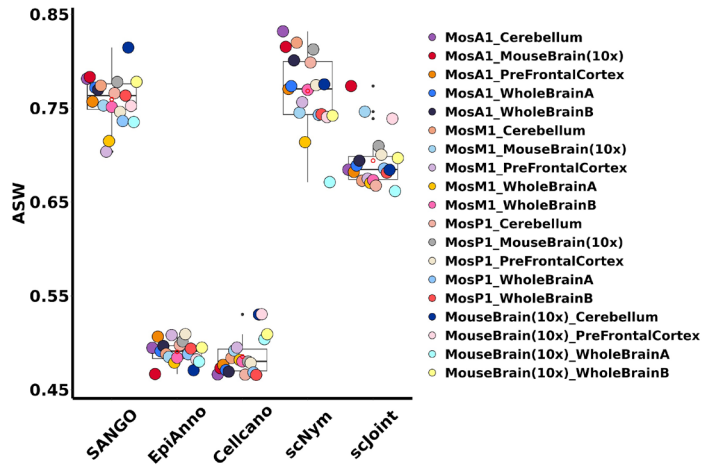
(c)


NMI
1.00
0.75
0.50
0.25
0.00
SANGO
EpiAnno
Cellcano
scNym
scJoint
MosA1_Cerebellum
MosA1_MouseBrain(10x)
MosA1_PreFrontalCortex
MosA1_WholeBrainA
MosA1_WholeBrainB
MosM1_Cerebellum
MosM1_MouseBrain(10x)
MosM1_PreFrontalCortex
MosM1_WholeBrainA
MosM1_WholeBrainB
MosP1_Cerebellum
MosP1_MouseBrain(10x)
MosP1_PreFrontalCortex
MosP1_WholeBrainA
MosP1_WholeBrainB
MouseBrain(10x)_Cerebellum
MouseBrain(10x)_PreFrontalCortex
MouseBrain(10x)_WholeBrainA
MouseBrain(10x)_WholeBrainB

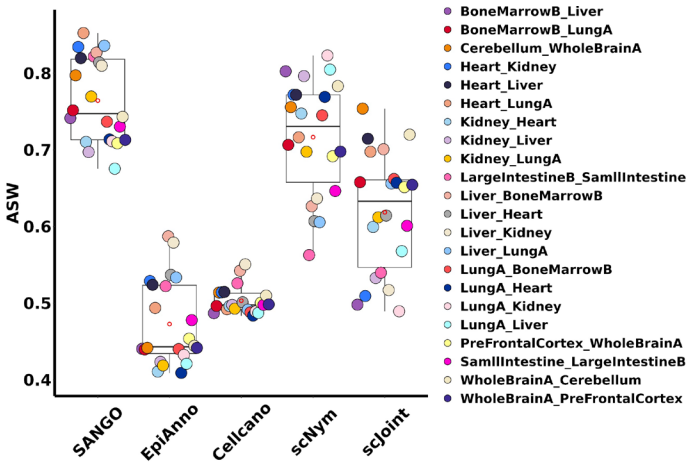
(d)


NMI
1.00
0.75
0.50
0.25
0.00
SANGO
EpiAnno
Cellcano
scNym
scJoint
BoneMarrowB_Liver
BoneMarrowB_LungA
Cerebellum_WholeBrainA
Heart_Kidney
Heart_Liver
Heart_LungA
Kidney_Heart
Kidney_Liver
Kidney_LungA
LargeIntestineB_SamllIntestine
Liver_BoneMarrowB
Liver_Heart
Liver_Kidney
Liver_LungA
LungA_BoneMarrowB
LungA_Heart
LungA_Kidney
LungA_Liver
PreFrontalCortex_WholeBrainA
SamllIntestine_LargeIntestineB
WholeBrainA_Cerebellum
WholeBrainA_PreFrontalCortex

(e)

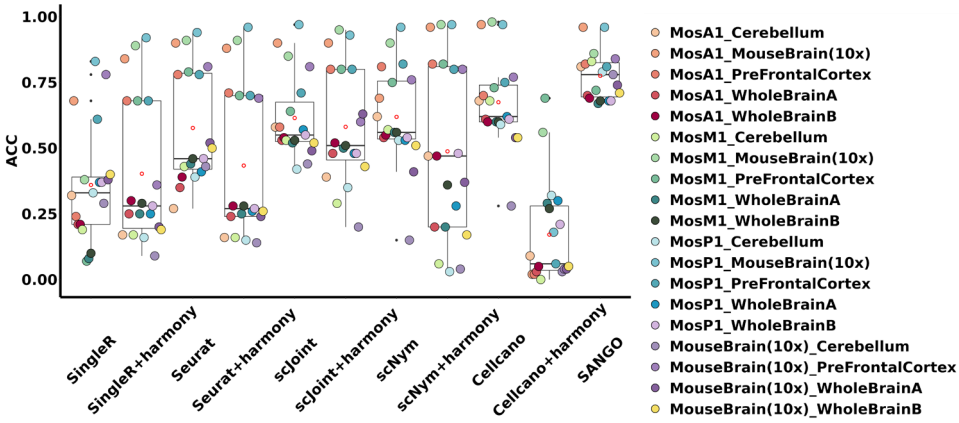

ASW
0.85
0.75
0.65
0.55
0.45
SANGO
EpiAnno
Cellcano
scNym
scJoint
MosA1_Cerebellum
MosA1_MouseBrain(10x)
MosA1_PreFrontalCortex
MosA1_WholeBrainA
MosA1_WholeBrainB
MosM1_Cerebellum
MosM1_MouseBrain(10x)
MosM1_PreFrontalCortex
MosM1_WholeBrainA
MosM1_WholeBrainB
MosP1_Cerebellum
MosP1_MouseBrain(10x)
MosP1_PreFrontalCortex
MosP1_WholeBrainA
MosP1_WholeBrainB
MouseBrain(10x)_Cerebellum
MouseBrain(10x)_PreFrontalCortex
MouseBrain(10x)_WholeBrainA
MouseBrain(10x)_WholeBrainB

(f)

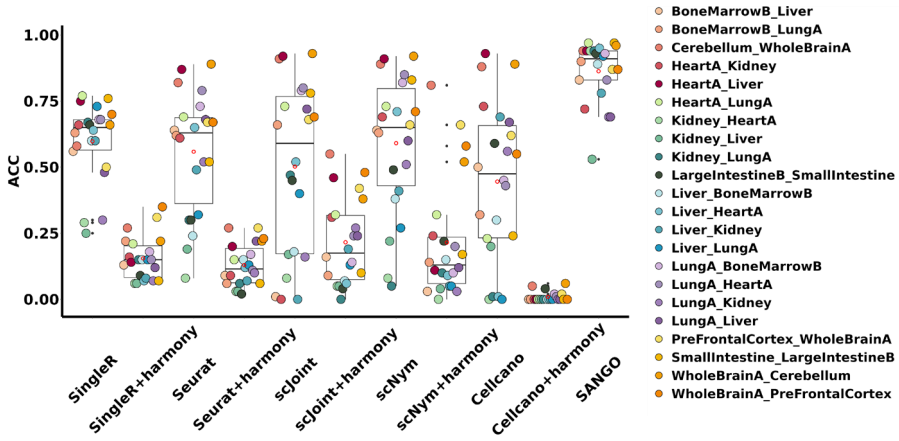


Supplementary figure 4. Quantitative evaluation of all methods with batch-effect correction is conducted using three assessment metrics: ARI, NMI, and ASW. (a) Boxplots summarize the ARI scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=19 biologically independent paired cross-platform datasets. (b) Boxplots summarize the ARI scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=22 biologically independent paired cross-tissue datasets. (c) Boxplots summarize the NMI scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=19 biologically independent paired cross-platform datasets. (d) Boxplots summarize the NMI scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=22 biologically independent paired cross-tissue datasets. (e) Boxplots summarize the ASW scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=19 biologically independent paired cross-platform datasets. (f) Boxplots summarize the ASW scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=22 biologically independent paired cross-tissue datasets. For all subfigures, the x-axis represents the various methods, while the y-axis denotes the measured values. The hollow red dot within the boxplot represents the average values, while black dots denote outliers.

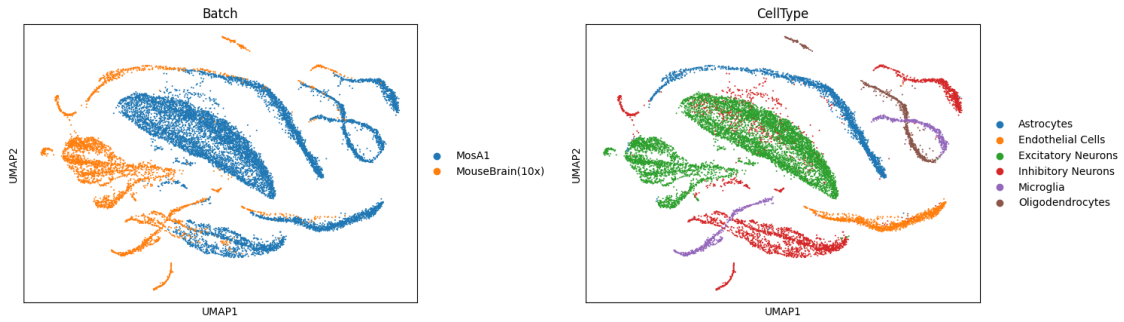
(a)


ACC
1.00
0.75
0.50
0.25
0.00
SingleR
SingleR+harmony
Seurat
Seurat+harmony
scJoint
scJoint+harmony
scNym
scNym+harmony
Cellcano
Cellcano+harmony
SANGO
MosA1_Cerebellum
MosA1_MouseBrain(10x)
MosA1_PreFrontalCortex
MosA1_WholeBrainA
MosA1_WholeBrainB
MosM1_Cerebellum
MosM1_MouseBrain(10x)
MosM1_PreFrontalCortex
MosM1_WholeBrainA
MosM1_WholeBrainB
MosP1_Cerebellum
MosP1_MouseBrain(10x)
MosP1_PreFrontalCortex
MosP1_WholeBrainA
MosP1_WholeBrainB
MouseBrain(10x)_Cerebellum
MouseBrain(10x)_PreFrontalCortex
MouseBrain(10x)_WholeBrainA
MouseBrain(10x)_WholeBrainB

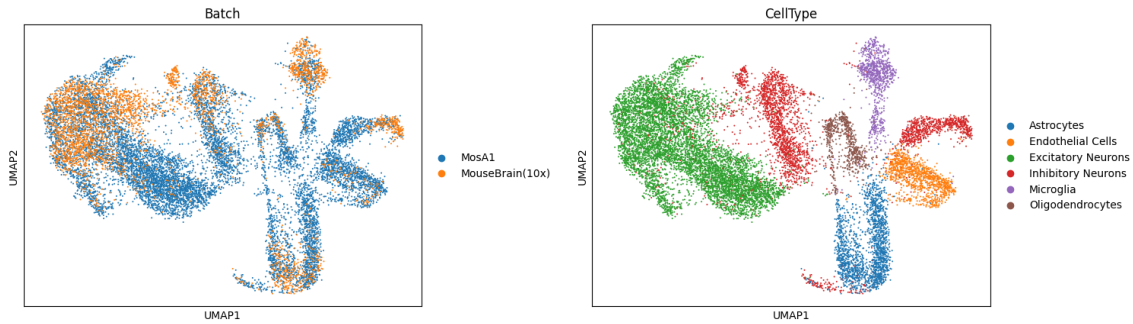
(b)


ACC
1.00
0.75
0.50
0.25
0.00
SingleR
SingleR+harmony
Seurat
Seurat+harmony
scJoint
scJoint+harmony
scNym
scNym+harmony
Cellcano
Cellcano+harmony
SANGO
BoneMarrowB_Liver
BoneMarrowB_LungA
Cerebellum_WholeBrainA
HeartA_Kidney
HeartA_Liver
HeartA_LungA
Kidney_HeartA
Kidney_Liver
Kidney_LungA
LargeIntestineB_SmallIntestine
Liver_BoneMarrowB
Liver_HeartA
Liver_Kidney
Liver_LungA
LungA_BoneMarrowB
LungA_HeartA
LungA_Kidney
LungA_Liver
PreFrontalCortex_WholeBrainA
SmallIntestine_LargeIntestineB
WholeBrainA_Cerebellum
WholeBrainA_PreFrontalCortex

(c)

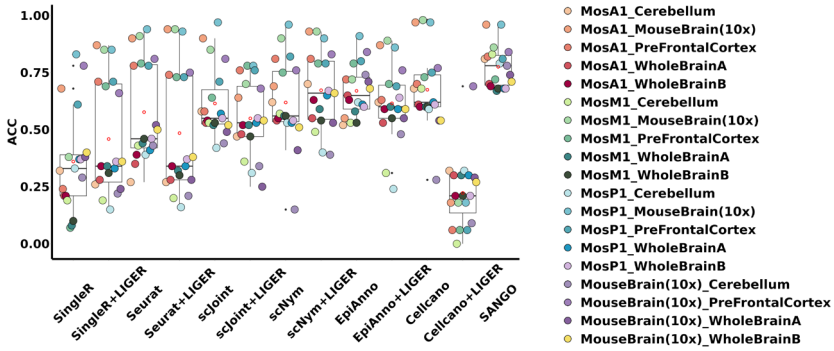

Batch
UMAP2
UMAP1
MosA1
MouseBrain(10x)
CellType
UMAP2
UMAP1
Astrocytes
Endothelial Cells
Excitatory Neurons
Inhibitory Neurons
Microglia
Oligodendrocytes

(d)

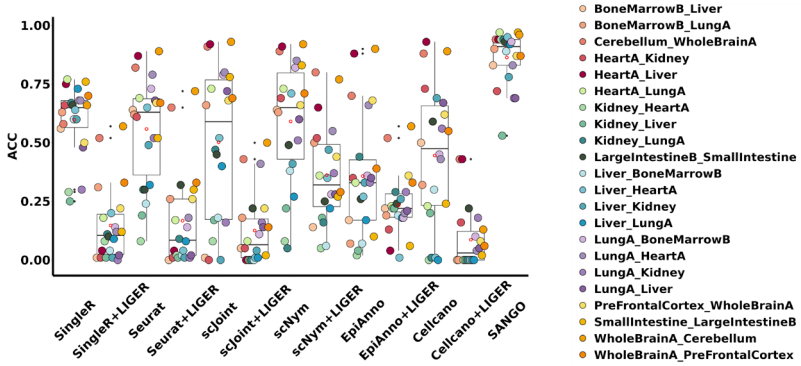

Batch
UMAP2
UMAP1
MosA1
MouseBrain(10x)
CellType
UMAP2
UMAP1
Astrocytes
Endothelial Cells
Excitatory Neurons
Inhibitory Neurons
Microglia
Oligodendrocytes

Supplementary figure 5. Quantitative evaluation of each method with batch effect method Harmony. (a) Boxplots summarize the ACC scores for each method with Harmony, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges) and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=19 biologically independent paired cross-platform datasets. (b) Boxplots summarize the ACC scores for each method with Harmony, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges) and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=22 biologically independent paired cross-tissue datasets. For all subfigures, the x-axis represents the various methods, while the y-axis denotes the measured values. The hollow red dot within the boxplot represents the average values, while black dots denote outliers. Visualization on batch effect removal showing one of the paired datasets MosA1_MouseBrain(10x). (c) shows the visualization before batch effect removal along with visualizations after batch effect removal conducted with (d) Harmony.

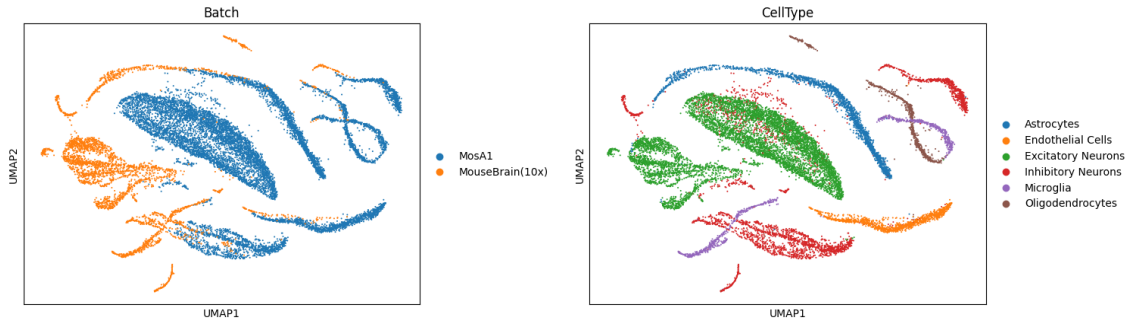
(a)



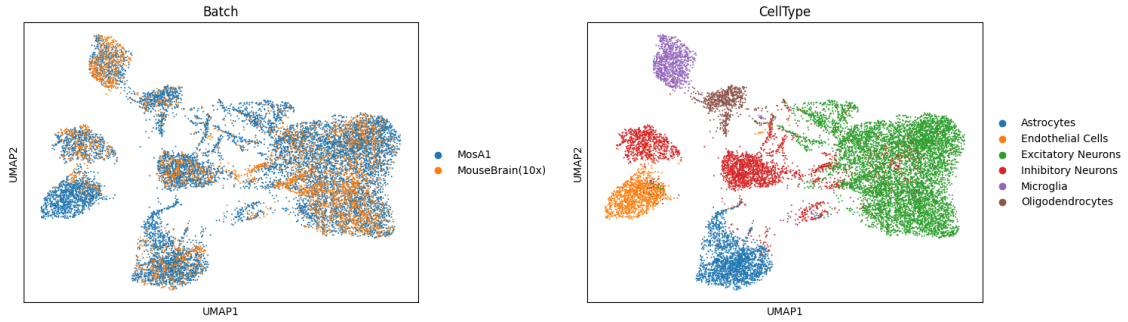
(b)



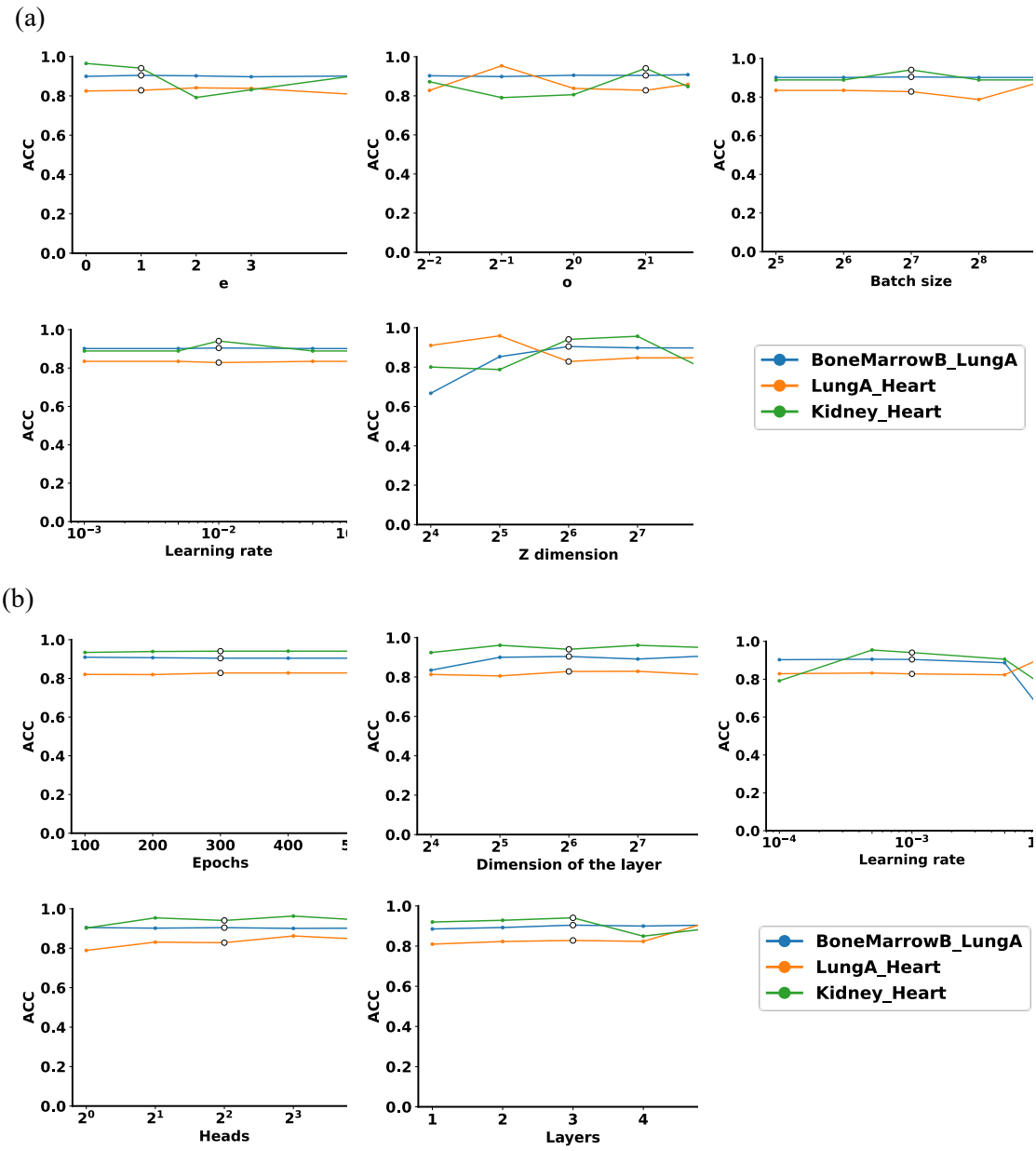
(c)



(d)

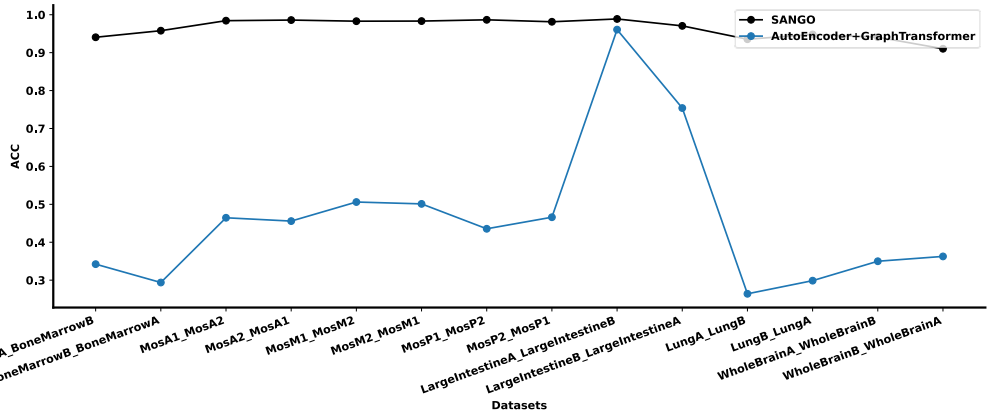


Supplementary figure 6. Quantitative evaluation of each method with batch effect method LIGER. (a) Boxplots summarize the ACC scores for each method with LIGER, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges) and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=19 biologically independent paired cross-platform datasets. (b) Boxplots summarize the ACC scores for each method with LIGER, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges) and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=22 biologically independent paired cross-tissue datasets. For all subfigures, the x-axis represents the various methods, while the y-axis denotes the measured values. The hollow red dot within the boxplot represents the average values, while black dots denote outliers. Visualization on batch effect removal showing one of the paired datasets MosA1_MouseBrain(10x). (c) shows the visualization before batch effect removal along with visualizations after batch effect removal conducted with (d) LIGER.

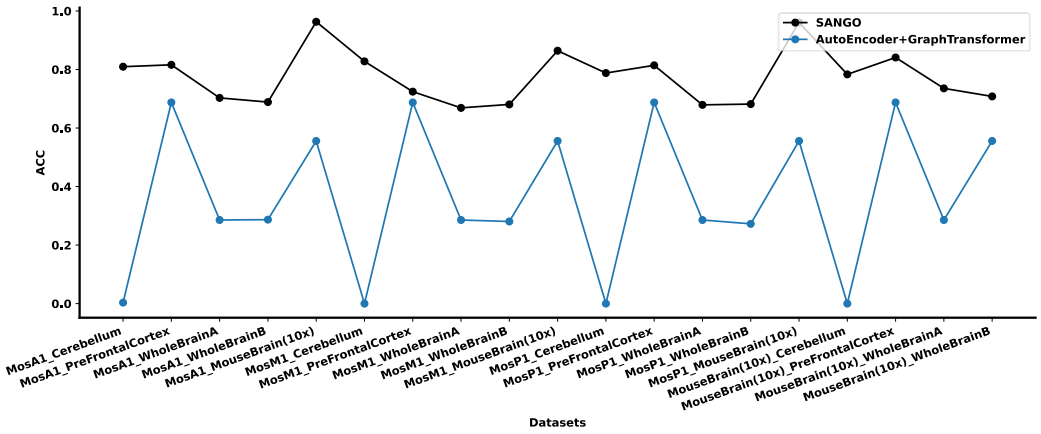


Supplementary figure 7 illustrates the robust evaluation of SANGO. (a) shows how the accuracy of SANGO changes when adjusting the hyperparameters of CACNN. (b) displays the accuracy variations of SANGO when modifying the hyperparameters of the graph transformer. Hollow circles represent the default parameter values used in SANGO.

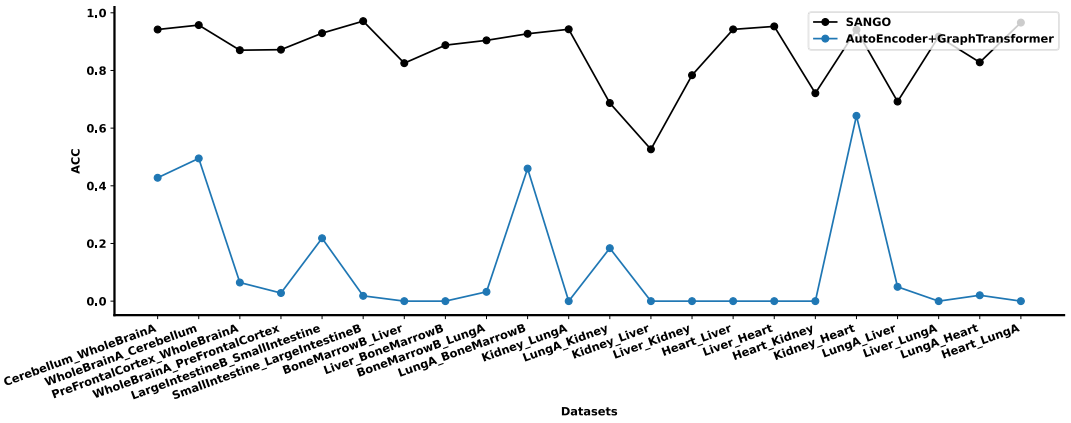
(a)



(b)

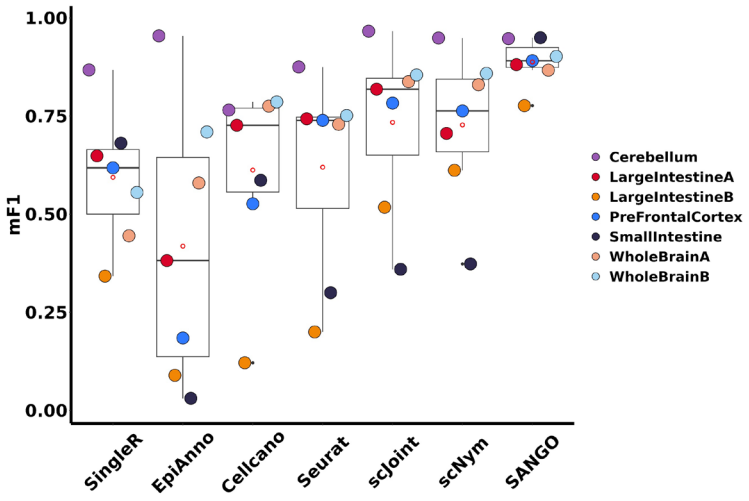


(c)

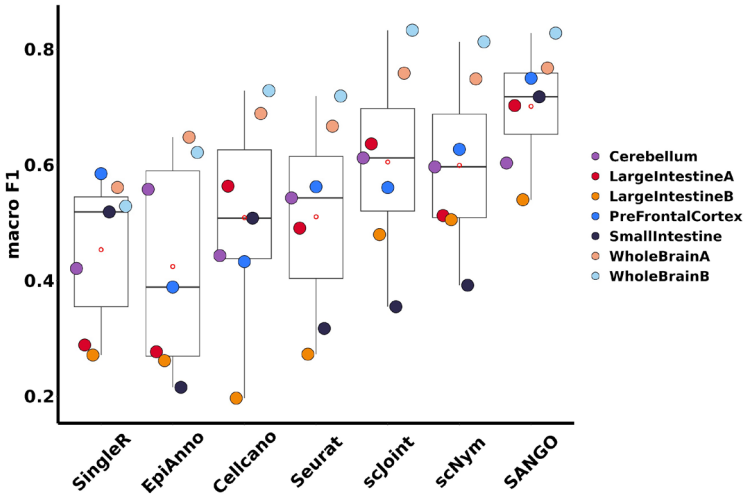


Supplementary figure 8. The performance comparison between SANGO (CACNN + GraphTransofomer) and (AE + GraphTransofomer) on the (a)intra-datasets, (b)cross-platform datasets, and (c)cross-tissue datasets.

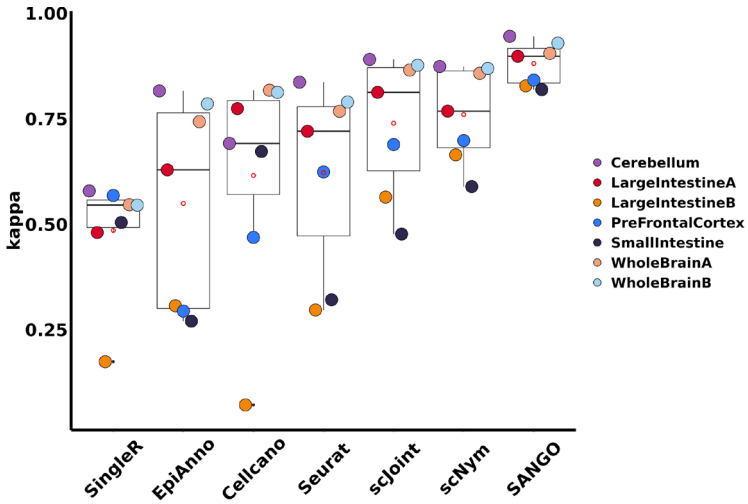
(a)


mF1
1.00
0.75
0.50
0.25
0.00
SingleR
EpiAnno
Cellcano
Seurat
scJoint
scNym
SANGO
Cerebellum
LargeIntestineA
LargeIntestineB
PreFrontalCortex
SmallIntestine
WholeBrainA
WholeBrainB

(b)

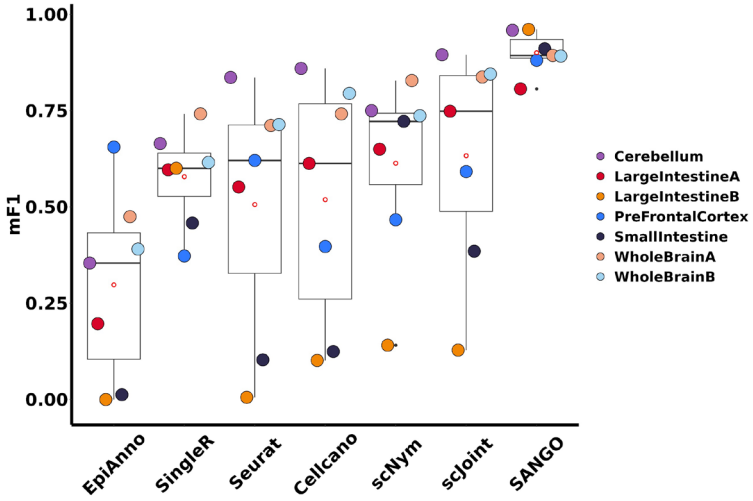

macro F1
0.8
0.6
0.4
0.2
SingleR
EpiAnno
Cellcano
Seurat
scJoint
scNym
SANGO
Cerebellum
LargeIntestineA
LargeIntestineB
PreFrontalCortex
SmallIntestine
WholeBrainA
WholeBrainB

(c)

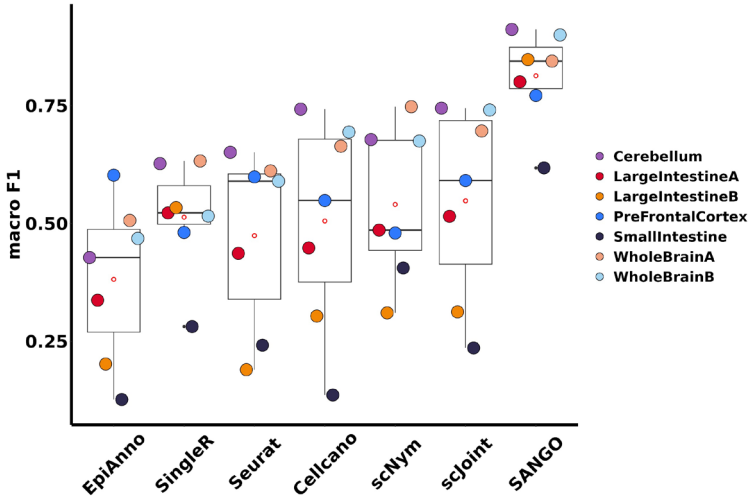

kappa
1.00
0.75
0.50
0.25
SingleR
EpiAnno
Cellcano
Seurat
scJoint
scNym
SANGO
Cerebellum
LargeIntestineA
LargeIntestineB
PreFrontalCortex
SmallIntestine
WholeBrainA
WholeBrainB

Supplementary figure 9. Performance on the single query data when using the multi-source data as the reference. The evaluation of each method on the two tissues including the mouse brain (consisting of four datasets) and Intestine (consisting of three datasets) by mF1, macro F1, and Cohen's kappa metrics, respectively. For each tissue, we iteratively left one as the query data and the rest as the multi-source reference data, resulting in 7 paired datasets. Each dataset in the legend represents the query data. (a) Boxplots summarize the mF1 scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges) and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=7 biologically independent paired datasets. (b) Boxplots summarize the macro F1 scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=7 biologically independent paired datasets. (c) Boxplots summarize the Cohen's kappa scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=7 biologically independent paired datasets. For all subfigures, the x-axis represents the various methods, while the y-axis denotes the measured values. The hollow red dot within the boxplot represents the average values, while black dots denote outliers.

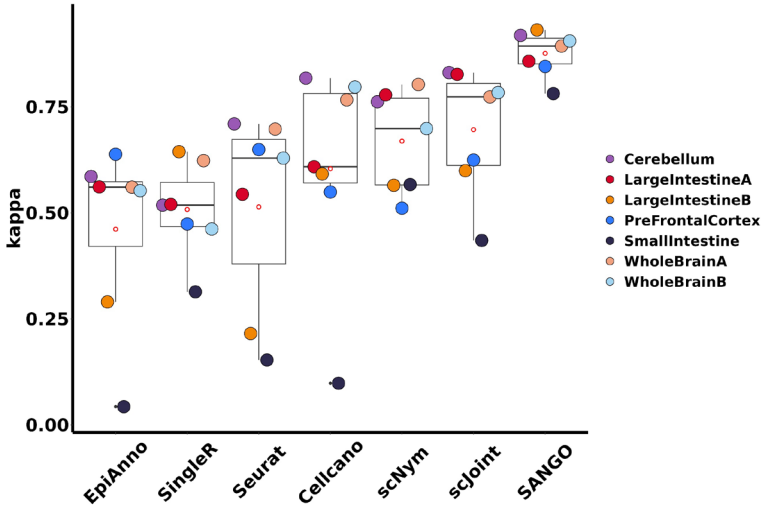
(a)



(b)

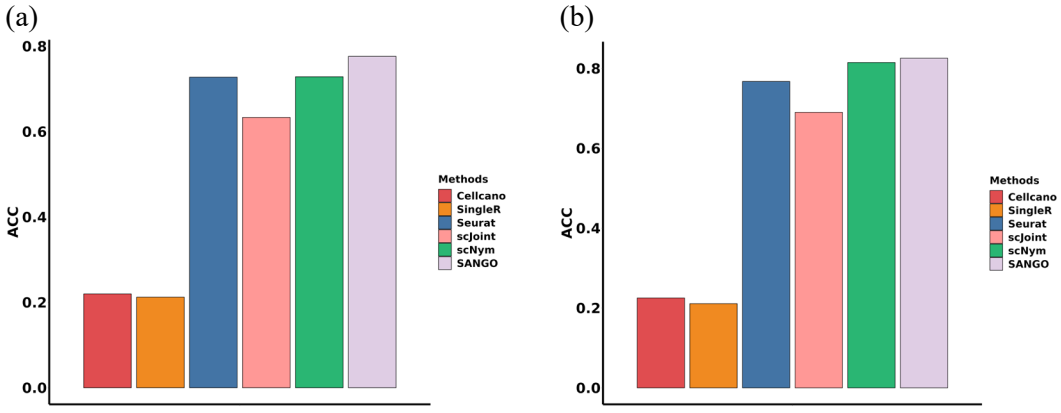


(c)



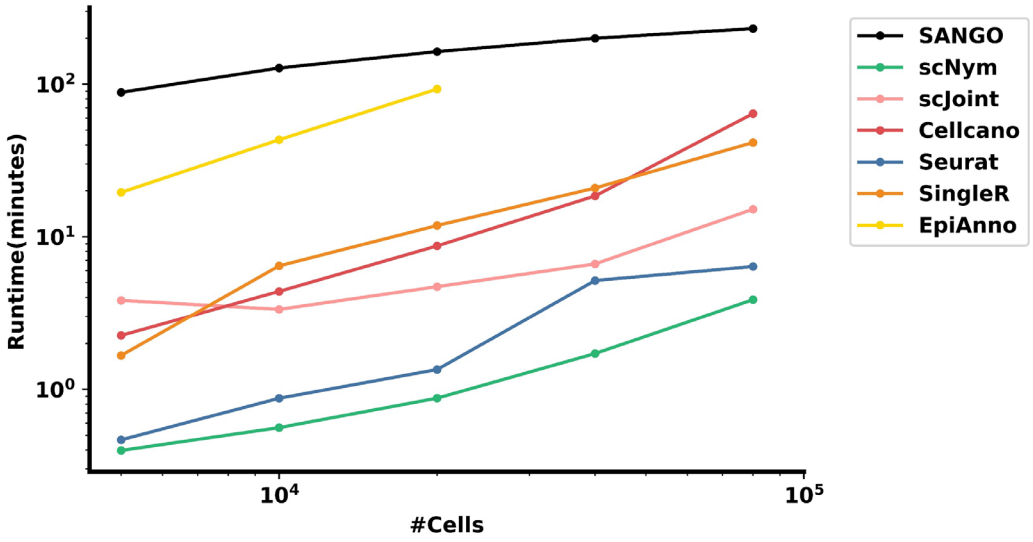
Supplementary figure 10. Performance on the combined query data when using the single reference data. The evaluation of each method on the two tissues including the mouse brain (consisting of four datasets) and Intestine (consisting of three datasets) by mF1, macro F1, and Cohen's kappa metrics, respectively. For each tissue, we iteratively left one as the reference data and the rest as the combined query data, resulting in 7 paired datasets. (a) Boxplots summarize the mF1 scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges) and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=7 biologically independent paired datasets. (b) Boxplots summarize the macro F1 scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=7 biologically independent paired datasets. (c) Boxplots summarize the Cohen's kappa scores for

each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. This analysis includes n=7 biologically independent paired datasets. For all subfigures, the x-axis represents the various methods, while the y-axis denotes the measured values. The hollow red dot within the boxplot represents the average values, while black dots denote outliers.

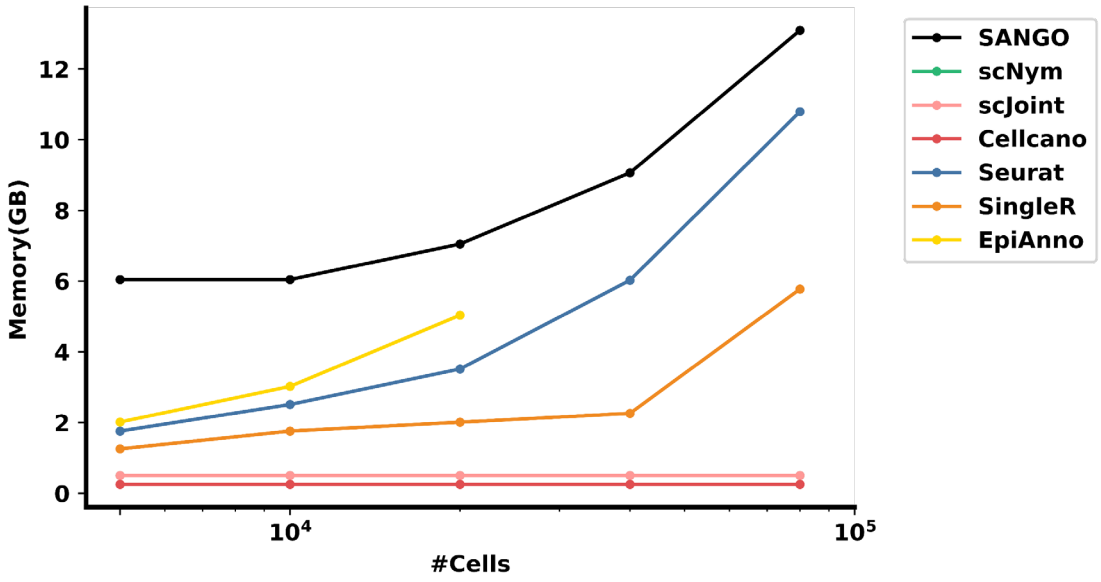


Supplementary figure 11. The performance of each method when using reference atlas data consisting of multiple anatomical entities. We selected the WholeBrain tissue as the query data and other tissues as the reference. The WholeBrain tissue has 9 cell types overlapped with the Cerebellum tissue. (a) In the first scenario, we kept 9 cell types only in the Cerebellum, and removed the cell types from other tissues. (b) In the 2nd scenario, we removed cell types so that five cell types appeared only in the Cerebellum, and the other four cell types appeared only in other tissues.

(a)

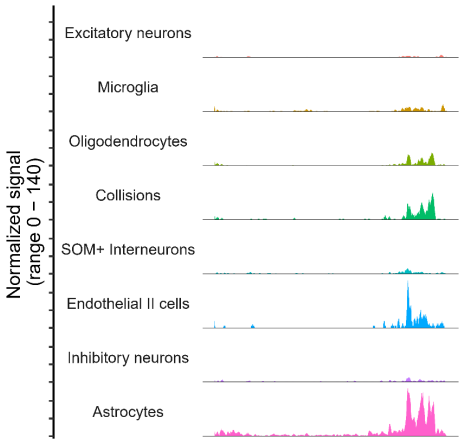


(b)



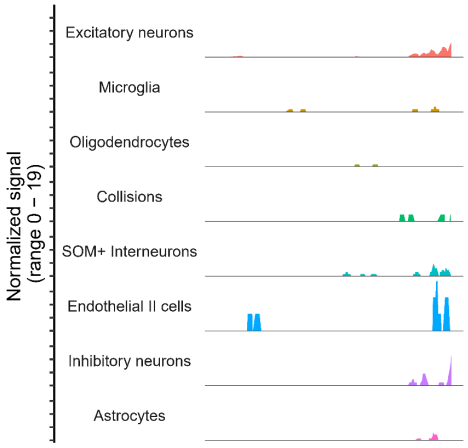
Supplementary figure 12. Evaluation of time consumption (a) and memory usage (b) for each approach across datasets with diverse cell counts.

(a)



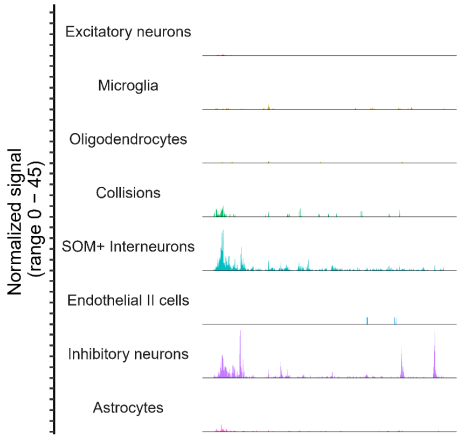
Gene: *Ndrg2*
Region: chr14-51903000-51915546

(b)



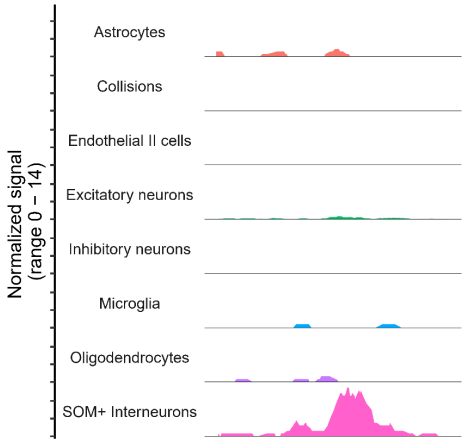
Gene: *Itm2a*
Region: chrX-107398771-107403376

(c)



Gene: *Gad2*
Region: chr2-22620000-22683000

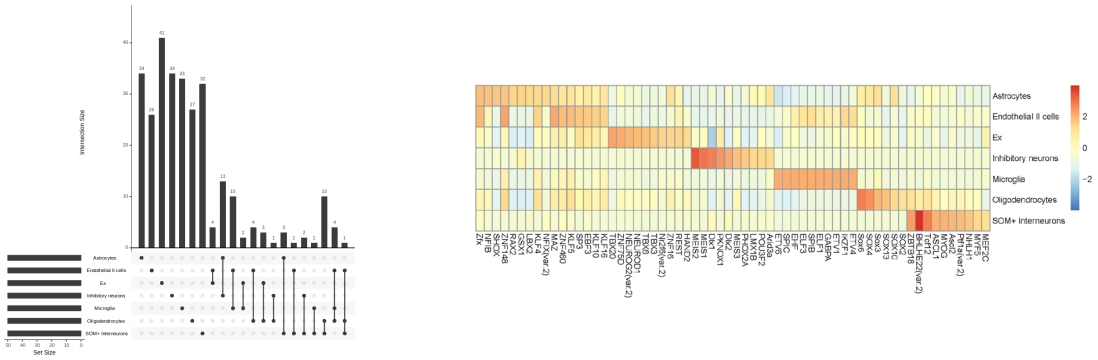
(d)



Gene: *Sst*
Region: chr16-23889920-23891500

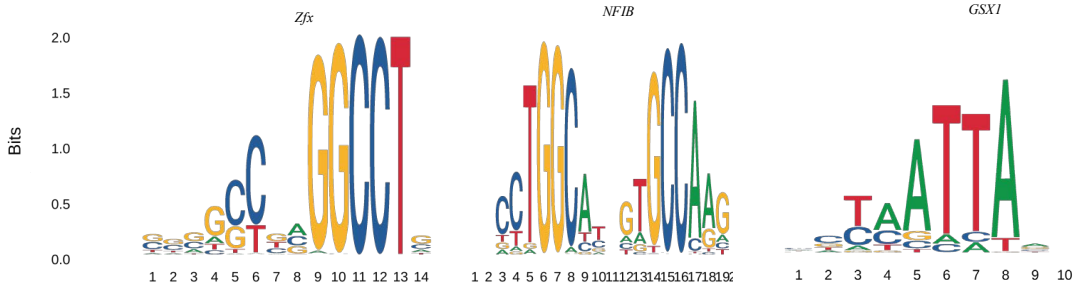
Supplementary figure 13. Coverage plots of chromatin accessibility in normal cortex data from the mouse brain over signature genes across predicted all cells: (a) *Ndrg2* for Astrocytes, (b) *Itm2a* for Endothelial II cells, (c) *Gad2* for Inhibitory neurons, (d) *Sst* for SOM+ interneurons. The term “Region” in each subgraph represents a genomic region of the chromosome.

(a) (b)

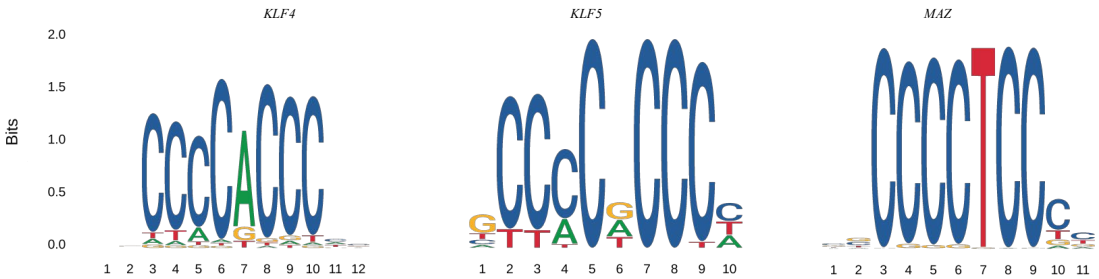


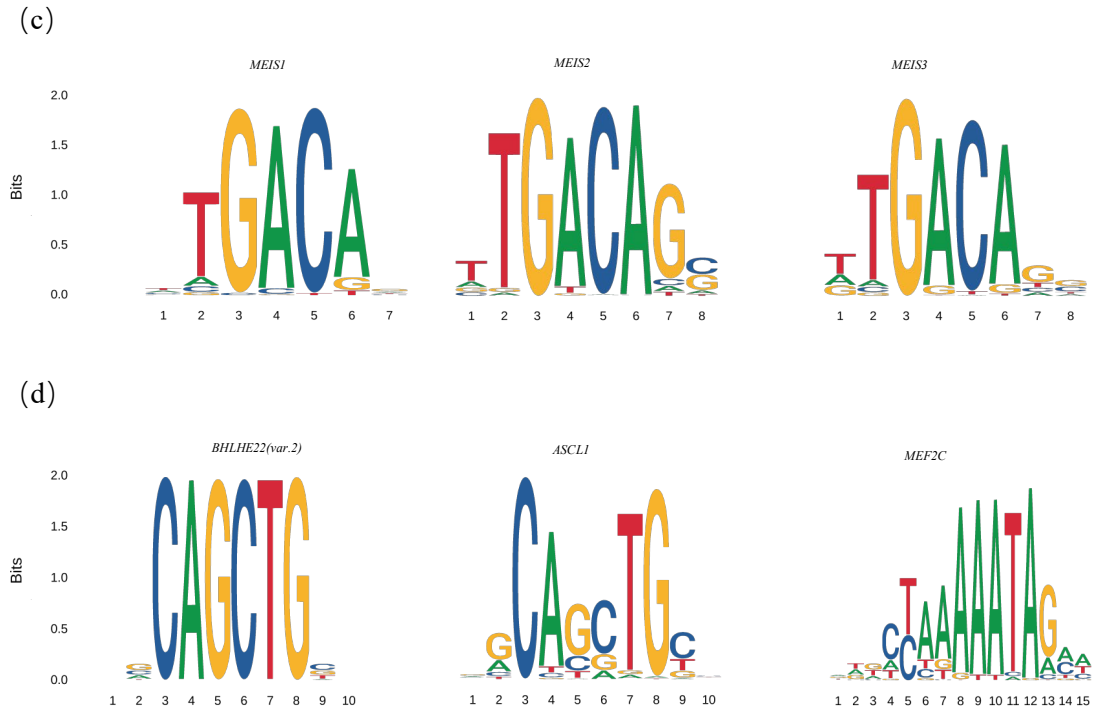
Supplementary figure 14. The motifs associated with each cell type as determined by our predictive method. (a) The UpSet diagram displays cell-specific motifs and the common motifs shared among different cell types. The horizontal bars represent the top 50 significant motifs for each cell type, while the vertical bars indicate cell type-specific and shared motifs. (b) The heatmap plot showcases the top 10 significant motifs for each cell type.

(a)

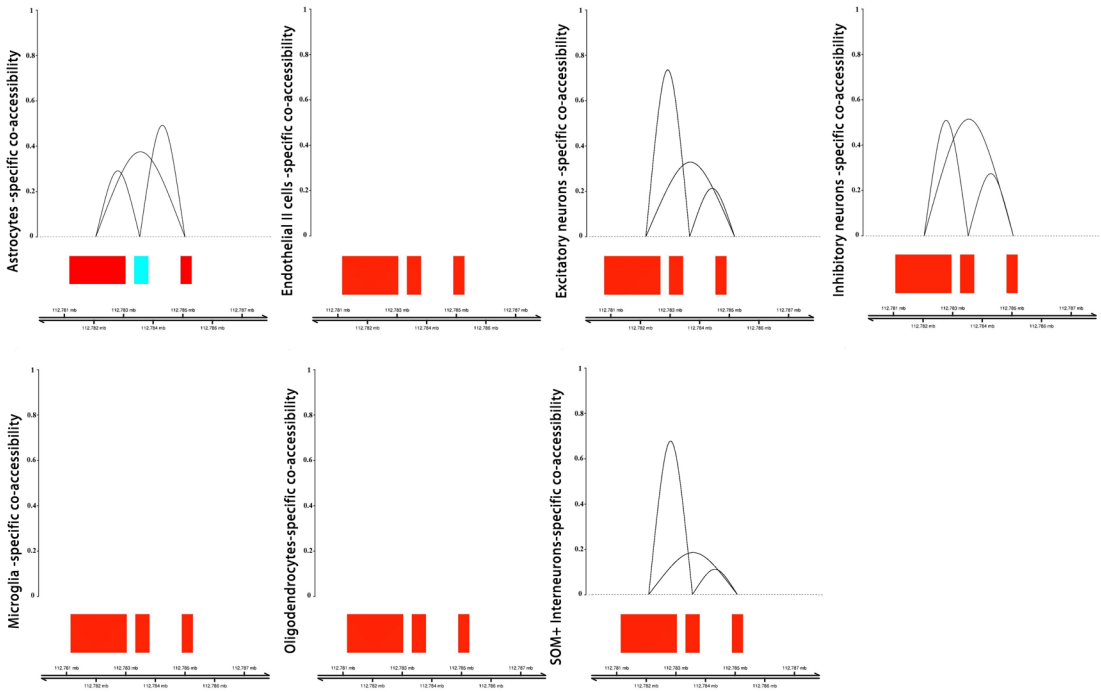


(b)

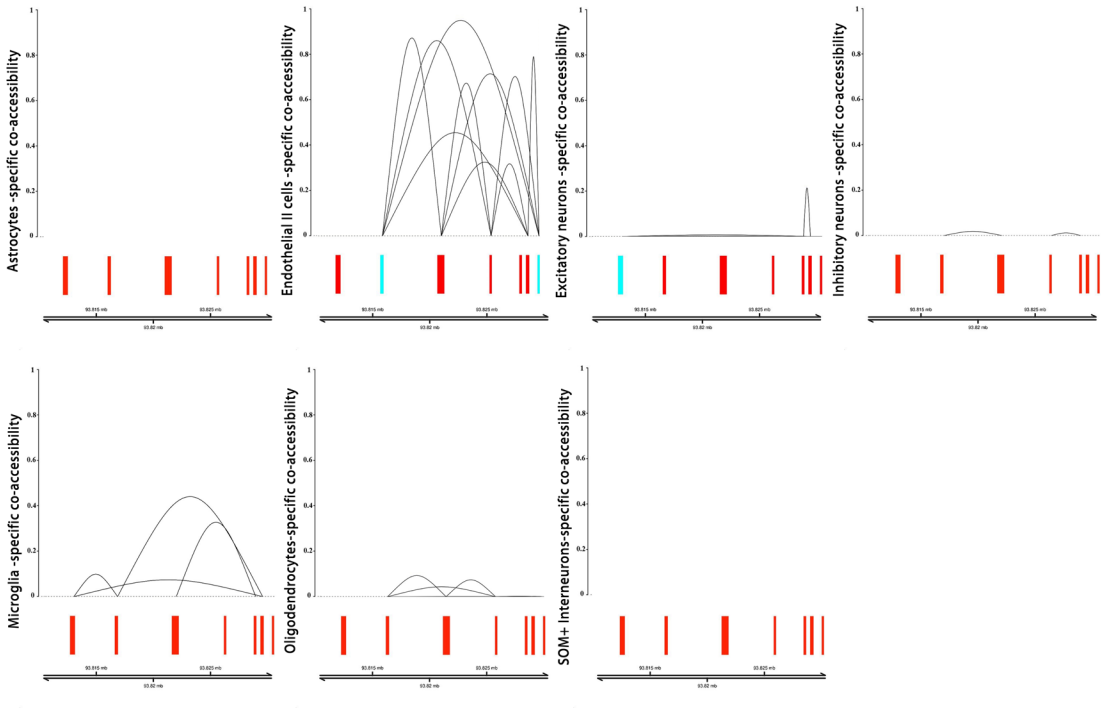




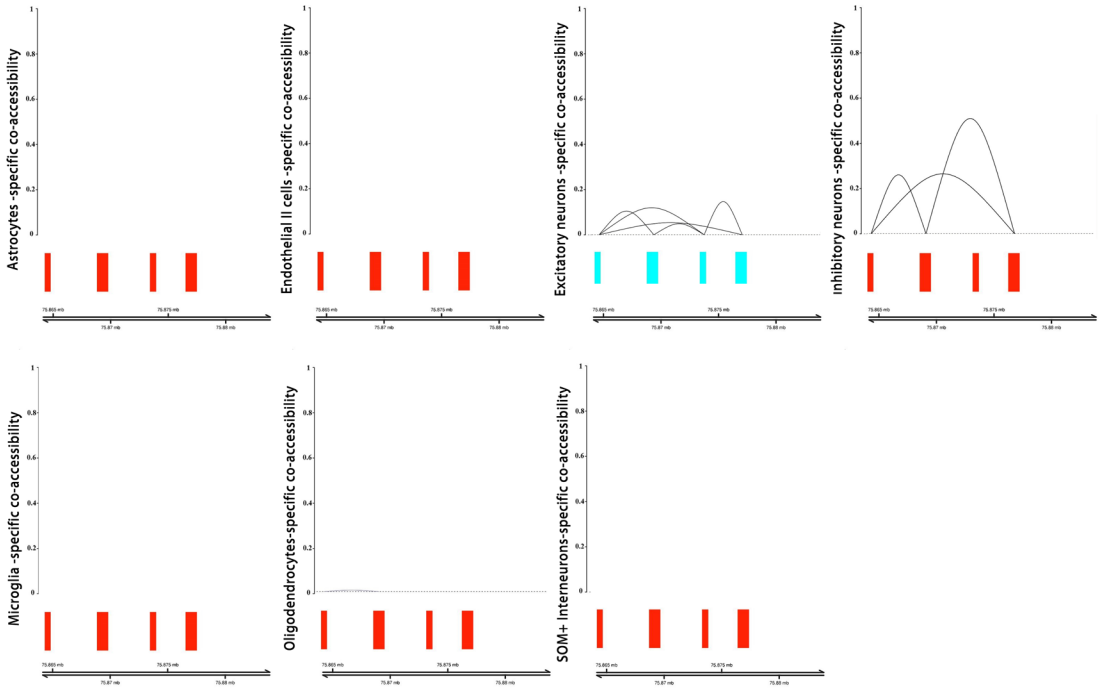
Supplementary figure 15. Overrepresented DNA motifs are identified through cell type-specific accessibility peaks in (a) Astrocytes, (b) Endothelial II cells, (c) Inhibitory neurons, (d) SOM+ interneurons.



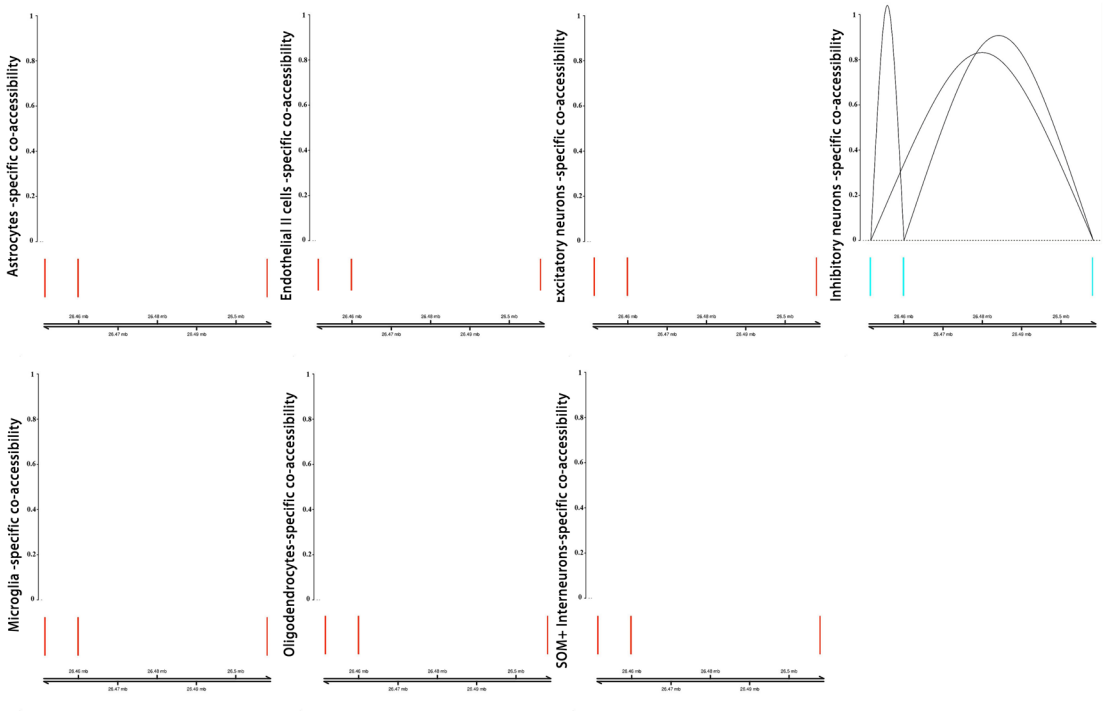
Supplementary figure 16. Cis-regulatory chromatin interactions predicted by Cicero with scATAC-seq data from Astrocytes. The cell type-specific peaks identified by SANGO are marked in cyan.



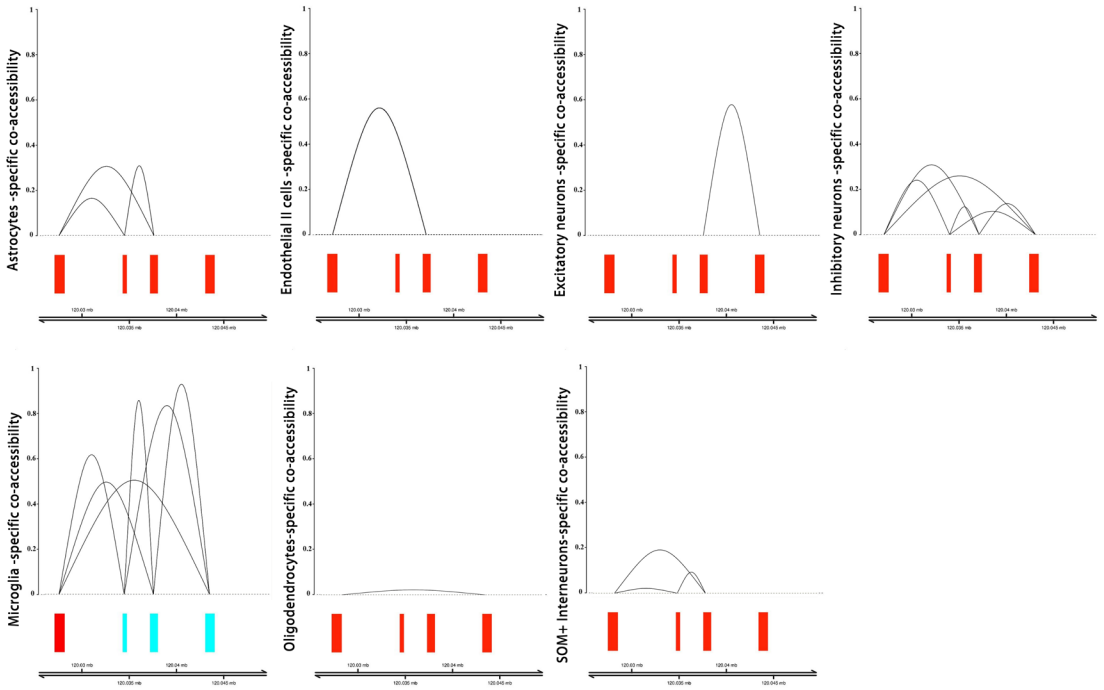
Supplementary figure 17. Cis-regulatory chromatin interactions predicted by Cicero with scATAC-seq data from Endothelial II cells. The cell type-specific peaks identified by SANGO are marked in cyan.



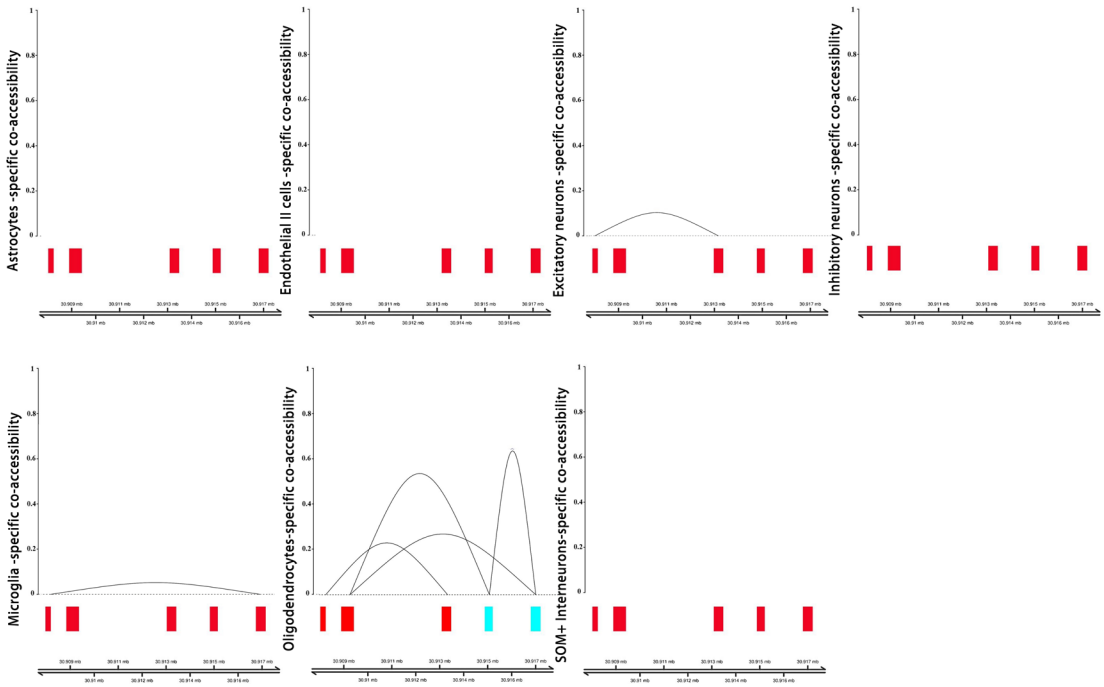
Supplementary figure 18. Cis-regulatory chromatin interactions predicted by Cicero with scATAC-seq data from Excitatory neurons. The cell type-specific peaks identified by SANGO are marked in cyan.



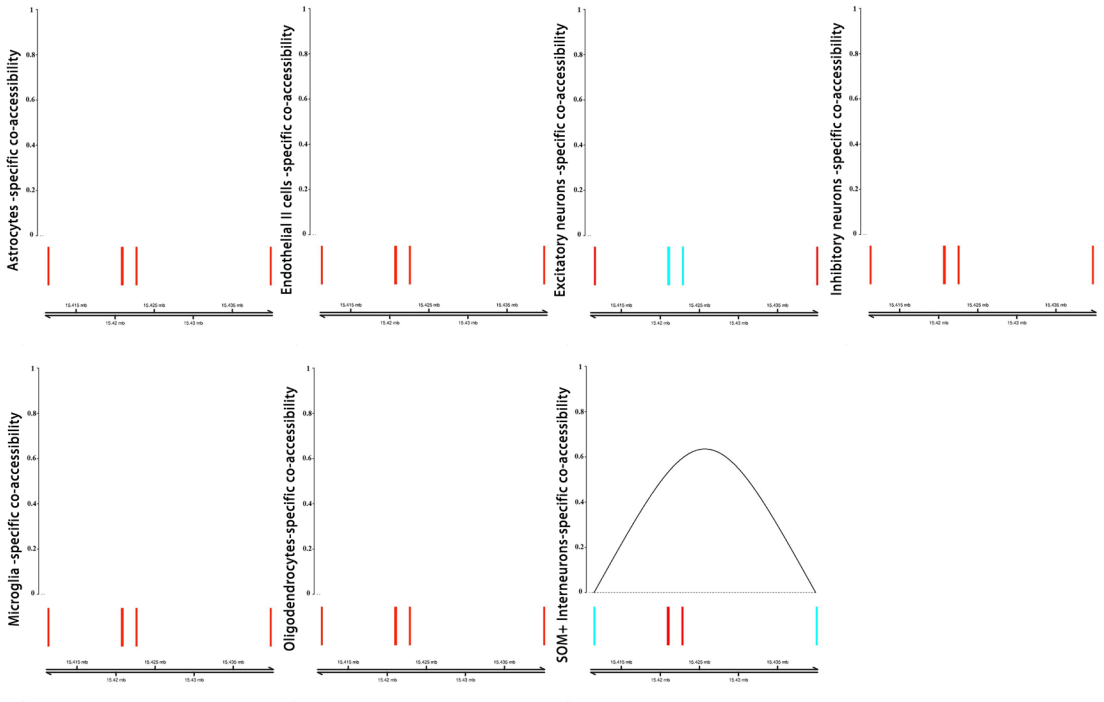
Supplementary figure 19. Cis-regulatory chromatin interactions predicted by Cicero with scATAC-seq data from Inhibitory neurons. The cell type-specific peaks identified by SANGO are marked in cyan.



Supplementary figure 20. Cis-regulatory chromatin interactions predicted by Cicero with scATAC-seq data from Microglia. The cell type-specific peaks identified by SANGO are marked in cyan.

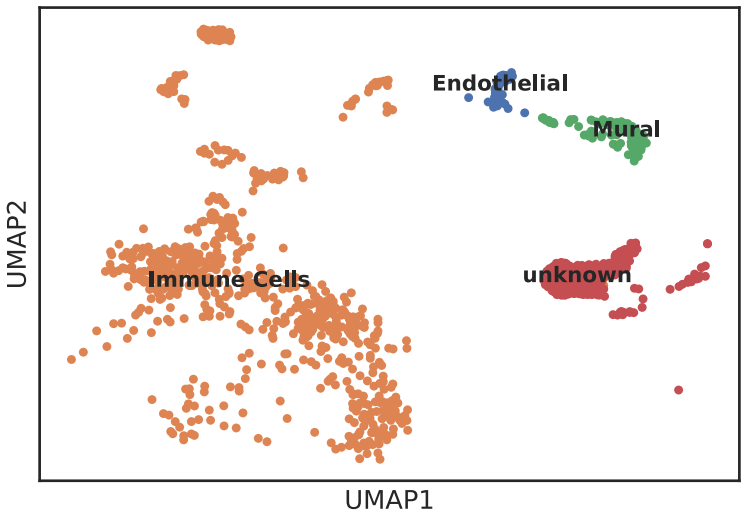


Supplementary figure 21. Cis-regulatory chromatin interactions predicted by Cicero with scATAC-seq data from Oligodendrocytes. The cell type-specific peaks identified by SANGO are marked in cyan.

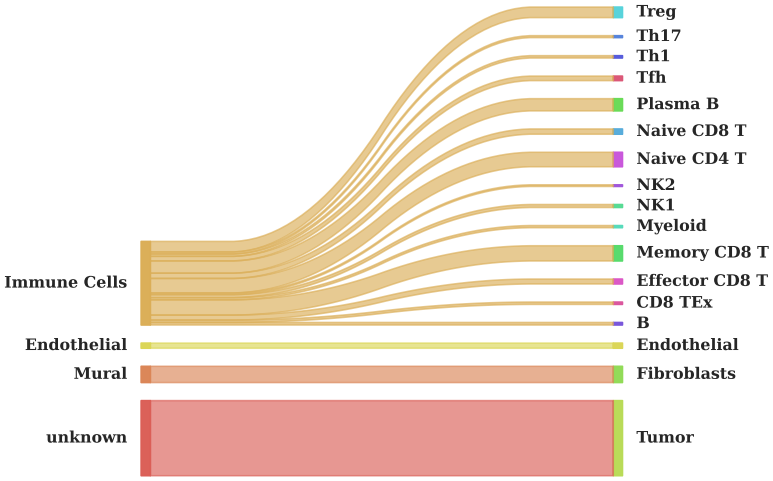


Supplementary figure 22. Cis-regulatory chromatin interactions predicted by Cicero with scATAC-seq data from SOM+ interneurons. The cell type-specific peaks identified by SANGO are marked in cyan.

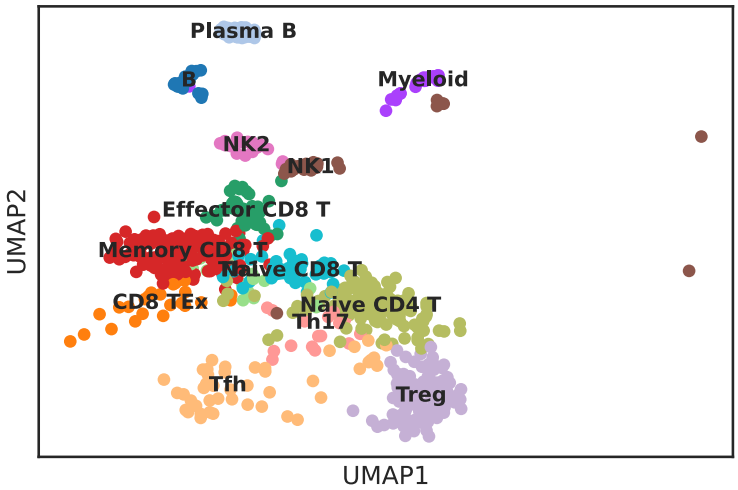
(a)


Endothelial
Mural
Immune Cells
unknown
UMAP2
UMAP1

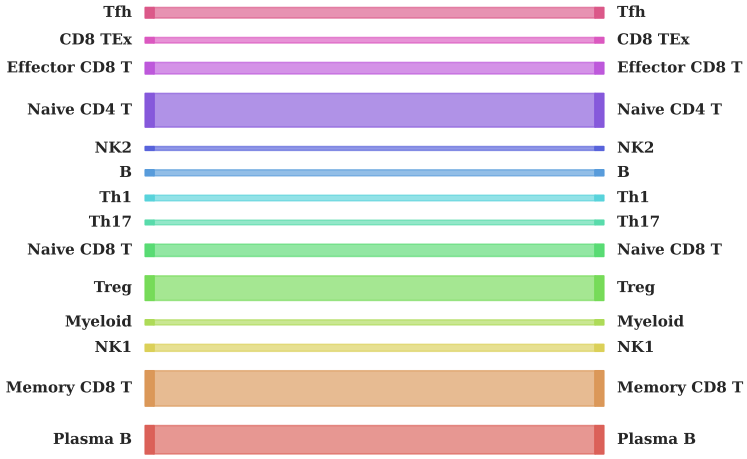
(b)


Immune Cells
Endothelial
Mural
unknown
Treg
Th17
Th1
Tfh
Plasma B
Naive CD8 T
Naive CD4 T
NK2
NK1
Myeloid
Memory CD8 T
Effector CD8 T
CD8 TEx
B
Endothelial
Fibroblasts
Tumor

(c)

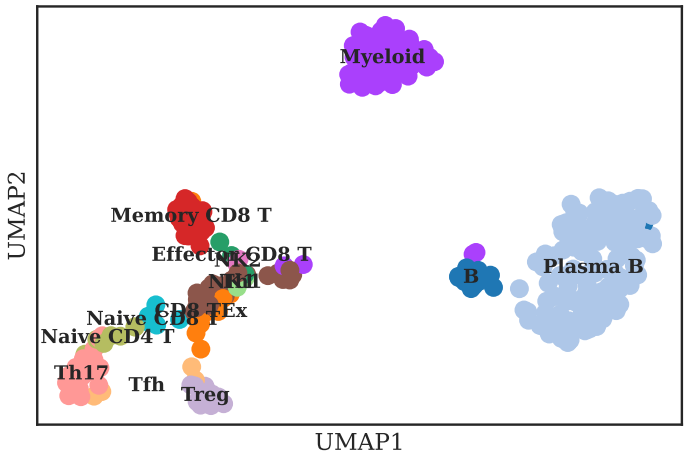

Plasma B
B
Myeloid
NK2
NK1
Effector CD8 T
Memory CD8 T
Naive CD8 T
CD8 TEx
Naive CD4 T
Th17
Tfh
Treg
UMAP2
UMAP1

(d)

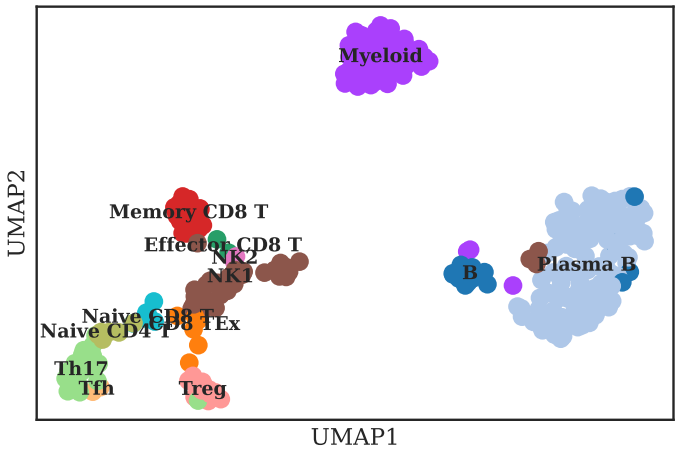


Supplementary figure 23. Visualization of the results of identifying multi-level cell types in basal cell carcinoma data through re-clustering labels. (a) The cells within query data are colored by re-clustering labels, and the cells with higher probability scores are recognized as unknown cell types. (b) River plot mapping the re-clustering coarse-grained cell types (left) to actual cell types (right). (c) The cells within query data are colored by re-clustering fine-grained immune cells. (d) River plot mapping the re-clustering immune cells (left) to actual cell labels (right).

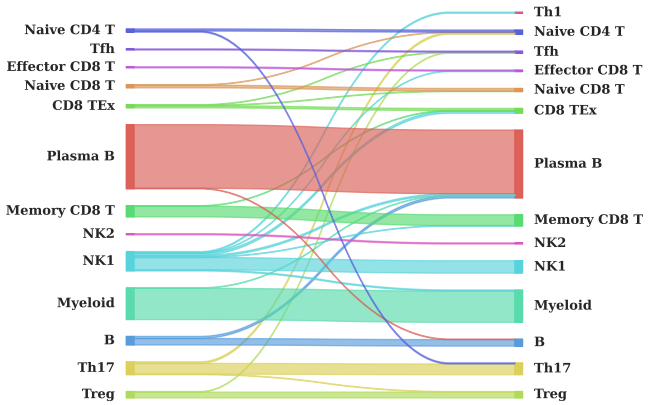
(a)



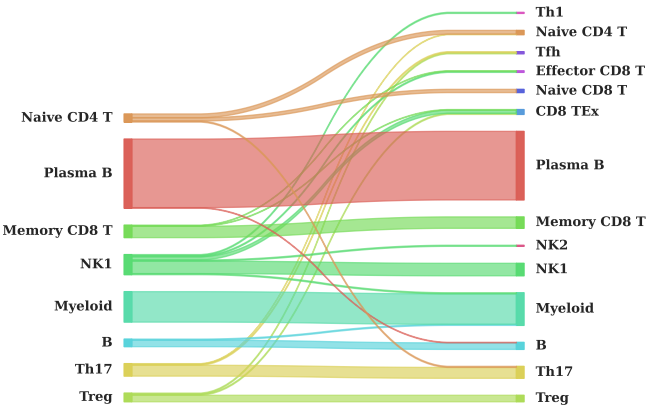
(b)



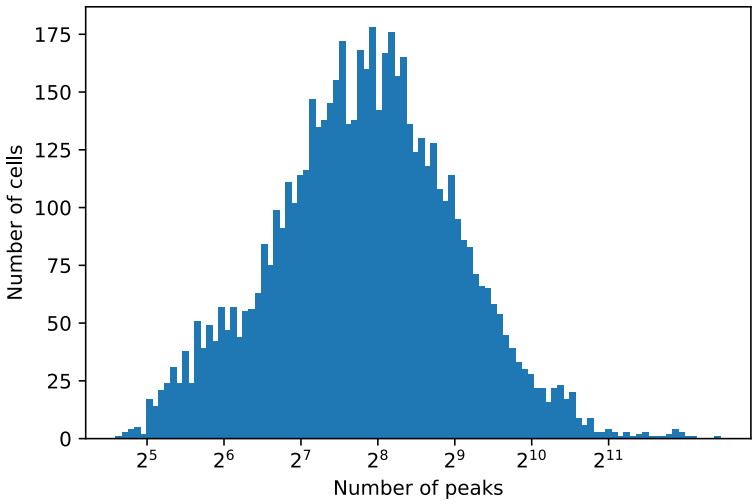
(c)



(d)



Supplementary figure 24. The performance of SANGO on the query data consisting of one pre-treatment SU008_Immune_Pre and one post-treatment sampleSU006_Total_Post in the BCC atlas. UMAP visualized cells of the query data colored by (a) truth cell types and (b) predicted cell types. (c) River plot mapping cell types annotated by SANGO (left) to actual cell labels (right). (d) River plot mapping re-clustering cell types (left) to actual cell labels (right).



Supplementary figure 25. The data distribution in the tissue dataset BoneMarrow. The x-axis represents the number of peaks, while the y-axis denotes the number of cells having peak number at a range.

Supplementary Note 1

For instance, excitatory neurons cells (Ex.neurons) were found to be enriched with the binding motifs of *TBX20*, *NEUROG2*, and *NEUROD1*[1-3]. The forced expression of *NEUROG2* and *NEUROG1* in human embryonic stem cells leads to the formation of a complex network comprising excitatory neurons [1]. Microglia were found to be enriched with the binding motifs of *ETV6*, *ELF3*, and *SPIB* [4-6]. Primary microglia overexpressing *ELF3*, tagged with *GFP*, ingest fluorescently labeled synaptosomes. Oligodendrocytes were found to be enriched with the binding motifs of *Sox6*, *Sox3*, and *SOX13* [7-9]. *SOX13* plays a functional role in complementing *Sox5* and *Sox6*, serving as crucial developmental modulators in mouse spinal cord oligodendrocytes. We also found that astrocytes were enriched with the binding motifs of *Zfx* [10], *NFIB*[11], and *GSX1* [12]. Endothelial II cells were enriched with the binding motifs of *KLF4*, *KLF5*, and *MAZ* [13-15]. Inhibitory neurons were enriched with the binding motifs of MEIS1, MEIS2, and *MEIS3* [16-18]. SOM+ Interneurons were enriched with the binding motifs of *BHLHE22*, *ASCL1*, and *MEF2C* [19-21].

[1] C. Lu, X. Shi, A. Allen, et al., "Overexpression of NEUROG2 and NEUROG1 in human embryonic stem cells produces a network of excitatory and inhibitory neurons," *The FASEB Journal,* vol. 33, no. 4, p. 5287, 2019.

[2] S. Tutukova, V. Tarabykinand L. R. Hernandez-Miranda, "The role of neurod genes in brain development, function, and disease," *Frontiers in Molecular Neuroscience,* vol. 14, p. 662774, 2021.

[3] M.-R. Song, R. Shirasaki, C.-L. Cai, et al., "T-Box transcription factor Tbx20 regulates a genetic

program for cranial motor neuron cell body migration," 2006.

[4] A. Grubman, X. Y. Choo, G. Chew, et al., "Transcriptional signature in microglia associated with Aβ plaque phagocytosis," *Nature communications,* vol. 12, no. 1, p. 3015, 2021.

[5] W. Gao, W. Kong, S. Wang, et al., "Biomarker Genes Discovery of Alzheimer's Disease by Multi-Omics-Based Gene Regulatory Network Construction of Microglia," *Brain Sciences,* vol. 12, no. 9, p. 1196, 2022.

[6] S. Xu, S. Mei, J. Lu, et al., "Transcriptome analysis of microglia reveals that the TLR2/IRF7 signaling Axis mediates Neuroinflammation after subarachnoid hemorrhage," *Frontiers in Aging Neuroscience,* vol. 13, p. 645649, 2021.

[7] T. Baroti, A. Schillinger, M. Wegner, et al., "Sox13 functionally complements the related Sox5 and Sox6 as important developmental modulators in mouse spinal cord oligodendrocytes," *Journal of neurochemistry,* vol. 136, no. 2, pp. 316-328, 2016.

[8] T. Turnescu, J. Arter, S. Reiprich, et al., "Sox8 and Sox10 jointly maintain myelin gene expression in oligodendrocytes," *Glia,* vol. 66, no. 2, pp. 279-294, 2018.

[9] S. A. Hoffmann, D. Hos, M. Küspert, et al., "Stem cell factor Sox2 and its close relative Sox3 have differentiation functions in oligodendrocytes," *Development,* vol. 141, no. 1, pp. 39-50, 2014.

[10] X. Fang, Z. Huang, W. Zhou, et al., "The zinc finger transcription factor ZFX is required for maintaining the tumorigenic potential of glioblastoma stem cells," *Stem cells,* vol. 32, no. 8, pp. 2033-2047, 2014.

[11] E. Matuzelski, J. Bunt, D. Harkins, et al., "Transcriptional regulation of Nfix by NFIB drives astrocytic maturation within the developing spinal cord," *Developmental biology,* vol. 432, no. 2, pp. 286-297, 2017.

[12] M. Patel, Y. Li, J. Anderson, et al., "Gsx1 promotes locomotor functional recovery after spinal cord injury," *Molecular therapy,* vol. 29, no. 8, pp. 2469-2482, 2021.

[13] X.-h. Wang, C.-y. Yanand J.-r. Liu, "Hyperinsulinemia-induced KLF5 mediates endothelial angiogenic dysfunction in diabetic endothelial cells," *Journal of Molecular Histology,* vol. 50, pp. 239-251, 2019.

[14] J.-R. Moonen, J. Chappell, M. Shi, et al., "KLF4 recruits SWI/SNF to increase chromatin accessibility and reprogram the endothelial enhancer landscape under laminar shear stress," *Nature communications,* vol. 13, no. 1, p. 4941, 2022.

[15] M. Smits, T. Wurdinger, B. van het Hof, et al., "Myc-associated zinc finger protein (MAZ) is regulated by miR-125b and mediates VEGF-induced angiogenesis in glioblastoma," *The FASEB Journal,* vol. 26, no. 6, pp. 2639-2647, 2012.

[16] S. Castagnola, J. Cazareth, K. Lebrigand, et al., "Agonist-induced functional analysis and cell sorting associated with single-cell transcriptomics characterizes cell subtypes in normal and pathological brain," *Genome Research,* vol. 30, no. 11, pp. 1633-1642, 2020.

[17] Y. M. Elkouby, H. Polevoy, Y. E. Gutkovich, et al., "A hindbrain-repressive Wnt3a/Meis3/Tsh1 circuit promotes neuronal differentiation and coordinates tissue maturation," *Development,* vol. 139, no. 8, pp. 1487-1497, 2012.

[18] A. Antic Nikolic, "Restless legs syndrome-associated variants contribute to MEIS1 regulation during neurodevelopment," Technische Universität München, 2021.

[19] D. J. Dennis, S. Hanand C. Schuurmans, "bHLH transcription factors in neural development, disease, and reprogramming," *Brain research,* vol. 1705, pp. 48-65, 2019.

[20] N. Kessaris, L. Magno, A. N. Rubin, et al., "Genetic programs controlling cortical interneuron fate," *Current opinion in neurobiology,* vol. 26, pp. 79-87, 2014.

[21] T. Riedemann, "Diversity and function of somatostatin-expressing interneurons in the cerebral cortex," *International journal of molecular sciences,* vol. 20, no. 12, p. 2952, 2019.