Supplementary Information for

Dissecting epigenetic heterogeneity in single-cell DNA methylomes with a unified framework

Songming Tang[1], Siyu Li[1], Guangxin Zhang[2], Aoran Lyu[3], Han Li[1,4] & Shengquan Chen[1,5,*]

[1] School of Mathematical Sciences and LPMC, Nankai University, Tianjin, China.

[2] SeekGene BioSciences, Beijing, China

[3] Department of Mechanical and Aerospace Engineering, The University of Manchester, Manchester, UK.

[4] Sheffield Institute for Translational Neuroscience, The University of Sheffield, Sheffield, UK.

[5] Academy for Advanced Interdisciplinary Studies, Nankai University, Tianjin, China.

[*] Corresponding author: chenshengquan@nankai.edu.cn

Supplementary Notes

Supplementary Note 1: Details of the evaluation metrics.

Let **T** denote the cell annotation labels, **P** denote the clustering labels, **N** denote the total number of cells, x_i denote the number of cells assigned to the ith cluster of **P**, y_j denote the number of cells that belong to the jth label of **T**, and n_{ij} denote the number of overlapping cells between the ith cluster and the jth label.

For ARI[1], suppose that C are the actual labels of the clusters and K are the clustering labels. In that case, we define a and b as follows: a, the number of pairs of elements in the same set in C and the same set in K. b, the number of pairs of elements in different sets in C and different sets in K. The following formula then gives the original Rand index:

$$RI = \frac{a+b}{C_2^{n_{\text{samples}}}}. \tag{1}$$

$C_2^{n_{\text{samples}}}$ is the total number of possible pairs in the datasets, and the following formula gives the ARI, which ensures random matching will get a value close to 0 and perfect matching gets a value close to 1:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}. \tag{2}$$

Mutual Information (MI) evaluates the degree of dependence between two random variables. AMI[2] adjusts MI by considering the expected value under random clustering, as shown in Equation (3), where H(·) is the entropy function and E(·) denotes the expectation function.

$$AMI = \frac{MI(\mathbf{P},\mathbf{T}) - E[MI(\mathbf{P},\mathbf{T})]}{avg[H(\mathbf{P}),H(\mathbf{T})] - E[MI(\mathbf{P},\mathbf{T})]}. \tag{3}$$

Moreover, NMI[3] scales MI to be between 0 and 1, and it is calculated as follows:

$$\mathrm{NMI} = \frac{\mathrm{MI}(\mathbf{P}, \mathbf{T})}{\sqrt{\mathrm{H}(\mathbf{P})\mathrm{H}(\mathbf{T})}}. \tag{4}$$

The homogeneity score[4] assesses whether the obtained clusters contain only cells of the same cell type, and it is equal to 1 if all the cells within the same cluster correspond to the same cell type. The homogeneity score is computed as follows:

$$\mathrm{Homogeneity} = 1 - \frac{H(\mathbf{T}|\mathbf{P})}{H(\mathbf{T})}, \tag{5}$$

where:

$$\begin{aligned} H(\mathbf{T}|\mathbf{P}) &= -\sum_{i=1}^{|\mathrm{P}|}\sum_{j=1}^{|\mathrm{T}|}\frac{n_{ji}}{N}\log\frac{n_{ji}}{\sum_{j=1}^{|T|} a_{ji}} \\ H(\mathbf{T}) &= -\sum_{j=1}^{|T|}\frac{\sum_{i=1}^{|P|} n_{ji}}{|T|}\log\frac{\sum_{i=1}^{|P|} n_{ji}}{|T|} \end{aligned} \tag{6}$$

Besides, the FMI[5] can be used as the geometric mean of pairwise precision and recall.

$$FMI = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}} \tag{7}$$

K-nearest neighbour batch effect test (kBET)[6] is a wrapper function of the implementation by Büttner, M et al.[7]. kBET measures the bias of a batch variable in the kNN graph. Specifically, kBET is quantified as the average rejection rate of Chi-squared tests of local vs global batch label distributions. This means that smaller values indicate better batch mixing.

Batch silhouette[8] measures the silhouette of a given batch. It assumes that a silhouette width close to 0 represents a perfect overlap of the batches; thus, the absolute value of the silhouette width is used to measure how well batches are mixed. The silhouette score evaluates the quality of clustering based on the distance between and within clusters. It is calculated using

the Equation (8), where x_i represents the i^{th} cell (totally N cells), and $a(x_i)$ and $b(x_i)$ represent the average distances between x_i and cells within and outside its cluster, respectively.

$$Silhouette\ score = \frac{1}{N}\sum_{i=0}^{N}\frac{b(x_i) - a(x_i)}{\max\left(a(x_i), b(x_i)\right)} \tag{8}$$

Graph iLISI is adapted from Korsunsky et al.[6] via shortest path-based distance computation on single-cell kNN graphs, and is noteworthy in that it does not require specifying cell type annotation labels. We calculated kBET, Batch silhouette, and Graph iLISI using scIB[8] and, by default, scaled between 0 and 1 so that larger scores are associated with better batch mixing.

The Macro-F1 Score is an evaluation metric for multi-class classification that computes the unweighted arithmetic mean of the per-class F1 scores, treating all classes equally regardless of their frequency. It is particularly useful for imbalanced datasets where minority class performance is critical. For each class i, compute:

$$\begin{aligned} \text{Precision}_i(P_i) &= \frac{TP_i}{TP_i + FP_i}, \\ \text{Recall}_i(R_i) &= \frac{TP_i}{TP_i + FN_i}, \\ F1_i &= 2 \times \frac{P_i \times R_i}{P_i + R_i}, \end{aligned} \tag{9}$$

where TP_i is the true positive for class i, FP_i is the false positive for class i, FN_i is the false negative for class i. And the formula of Macro-F1 can be written as:

$$\text{Macro-F1} = \frac{1}{C}\sum_{i=1}^{C} F1_i, \tag{10}$$

where C is the total number of classes.

Supplementary Note 2: Existing imputation strategies for scDNAm data

Existing imputation methods exhibit overall weaker performance, which can be largely attributed to their underlying design principles and implementation strategies. Specifically, EpiScanpy[9] adopts a feature-wise mean imputation approach, in which missing values for a given genomic region are replaced by the median (or mean) methylation level of that region across cells. Similarly, the pipeline proposed by Liu et al.[10] is effectively equivalent to imputing each missing entry with the genome-wide average methylation level of the corresponding cell. While these strategies are straightforward and computationally efficient, they do not account for cell-type-specific methylation patterns at individual genomic regions. In the presence of substantial sparsity and technical noise inherent to single-cell DNA methylation data, such mean-based imputations tend to further attenuate biologically informative signals, resulting in reduced performance.

MethSCAn[11] employs a different strategy based on iterative principal component analysis (PCA). In this approach, missing values are initialized to zero and then iteratively updated by repeatedly applying PCA until convergence, which can be interpreted as introducing a relatively strong noise component at the initialization stage. As a consequence, this method performs well for cell types with large numbers of cells and relatively stable methylation patterns, such as L6b and L5-IT neurons. However, for cell types with fewer cells or higher heterogeneity, the imputation becomes less stable, and performance degrades accordingly.

Fundamentally, these methods were not designed with the primary objective of reconstructing cell-type-specific methylation patterns. Rather, their goal is to provide

acceptable initial values for missing entries to enable downstream analyses, such as PCA or clustering, without errors caused by explicit missing values. As a result, they do not explicitly model the noise structure or cell-type-specific heterogeneity characteristic of scDNAm data. This difference in design objectives likely explains their limited ability to accurately recover cell-type-specific methylation signals and their comparatively lower performance.

Supplementary Note 3: scMethCraft correctly reconstructs cell-type-specific regions under extreme missingness

To further assess the robustness of scMethCraft under extreme missingness, we conducted an additional masking experiment focusing on cell-type-specific genomic regions. Specifically, in the ACC dataset, we selected regions overlapping well-established marker genes, including *STMN2*, *GAD2*, *MAP2*, and *SATB2*, and masked their methylation values by setting the corresponding entries to NA in the L2/3-IT and Lamp5 cell types. scMethCraft was then re-run to impute these fully masked regions.

As shown in Supplementary Figs. 21a,b and 22a,b, even when cell-type-specific regions exhibit extensive missingness, scMethCraft can recover methylation values that are consistent with known cell-type-specific patterns. This reconstruction is achieved by leveraging intercellular similarity learned from the remaining observed regions. To further illustrate this effect, we compared the methylation levels of the two cell types in the corresponding gene regions using violin plots before masking and after imputation (Supplementary Figs. 21c,d and 22c,d). Although the average methylation levels are not identical before masking and after imputation, scMethCraft clearly preserves the relative differences between cell types, reflecting the underlying cellular heterogeneity associated with these marker genes.

Together, these results demonstrate that scMethCraft can reliably reconstruct cell-type-specific methylation patterns even under extreme missingness, without collapsing biologically meaningful differences between cell populations.

Supplementary Note 4: Parameter selection in scMethCraft

Genomic bin size. By default, scMethCraft partitions the genome into fixed-length bins of 10 kbp, which is also used for all baseline methods to ensure fair comparison. Owing to its flexible network architecture, scMethCraft can be readily extended to alternative bin sizes. To assess the impact of bin size on model performance, we evaluated bin sizes of 5 kbp, 10 kbp, 20 kbp, 50 kbp, and 100 kbp on the ACC dataset. Bin sizes up to 100 kbp have minimal impact on overall visualization and clustering results. However, with larger bin sizes, certain closely related cell types, such as Vip, Sncg, Lamp5, and Lamp5-Lhx6 cells, become more difficult to distinguish, as highlighted by the purple boxes (Supplementary Fig. 23).

Region filtering. We next describe the rationale for applying an 80% missingness threshold to filter genomic regions. If a region lacks methylation measurements in all cells, including it in model training does not provide any informative signal for parameter updates. Examination of the region-wise distribution of missing (NA) rates in the ACC dataset shows that most regions exhibit moderate missingness, with NA rates concentrated around 0.1-0.4 (Supplementary Fig. 13). In contrast, a distinct peak near 1.0 indicates a non-negligible number of regions with extremely sparse or no coverage across cells. Excluding such regions from the model input is therefore a natural and data-driven choice. Furthermore, the density of regions with NA rates between 0.6 and 1.0 is very low, suggesting that the precise cutoff chosen within this range should not critically affect model behavior. Accordingly, scMethCraft is not expected to be sensitive to the exact threshold around 0.8. To empirically confirm this, we evaluated alternative filtering strategies by excluding regions missing in all cells, in more than 80% of cells, or in more than 60% of cells. The resulting analyses yielded highly consistent outcomes,

indicating that scMethCraft is robust to these alternative filtering choices (Supplementary Figs. 24, 25).

Embedding dimensionality. The default embedding dimensionality in scMethCraft is set to 50, which represents a moderate and commonly used choice in single-cell analysis. For example, Scanpy[12] maps cells into a 50-dimensional space by default using PCA. We initially selected this value empirically. To further examine the effect of embedding dimensionality, we evaluated scMethCraft on the ACC dataset using embedding dimensions ranging from 6 to 200, followed by visualization and clustering based on the resulting cell embeddings (Supplementary Figs. 26, 27). When the dimensionality is very low, the model is unable to sufficiently capture sequence-derived information. Once the dimensionality reaches approximately 20, the embeddings are generally adequate to encode relevant information, and further increases in dimensionality do not result in clear performance gains. Based on these observations, the embedding dimensionality is treated as an optional hyperparameter in scMethCraft, allowing users to adjust it according to specific data characteristics and analysis goals.

Supplementary Note 5: Selection of batch integration parameter.

First, we systematically evaluated the effect of β on batch integration performance across four datasets, which were used in the original manuscript, by varying β over the range [0, 1]. For the MTG-Pro dataset, we visualized the integration results using UMAP visualization under different values of β (Supplementary Fig. 28). These visualizations show that when β is close to 0, the two batches are poorly integrated. As β increases, batch mixing gradually improves and reaches an optimal level around $\beta \approx 0.2$. In addition to visualization, we evaluated integration quality using the graph iLISI metric, which does not rely on cell-type annotations and therefore provides a more direct assessment of batch integration (Supplementary Fig. 29). The results consistently show that performance increases from near zero at small values of β, reaches a plateau or optimum around $\beta \approx 0.2$, and then gradually decreases as β approaches 1. This trend is consistently observed across all four datasets. Based on these observations, we empirically selected β with 0.2 as the default value in our original experiments.

Second, the design of the batch integration strategy in scMethCraft is inspired by scCASE[13], which aims to balance intra-batch similarity and inter-batch similarity for the same cell types. Conceptually, the goal is that, after integration, similarities between cells of the same type should be comparable within and across batches. Motivated by this principle, we developed an adaptive strategy for selecting β based on batch-level statistics, including the average intra-batch similarity, average inter-batch similarity, and the number of cells in each batch:

We define a binary indicator matrix:

$$\mathbf{M} \in \{0,1\}^{N \times N}, \tag{11}$$

which encodes whether a pair of cells belongs to the same batch:

$$\mathbf{M} = \begin{cases} 1, & if\ B_i = B_j, \\ 0, & if\ B_i \neq B_j. \end{cases} \tag{12}$$

Using this matrix, we estimate an adaptive scaling factor β, which reflects the relative weighting between inter-batch similarity and intra-batch similarity:

$$\beta = \frac{\sum \mathbf{M}}{\sum (1 - \mathbf{M})} \cdot \frac{\mathrm{mean}\left(\mathbf{S_{norm}} \odot (1 - \mathbf{M})\right)}{\mathrm{mean}(\mathbf{S_{norm}} \odot \mathbf{M})} \tag{13}$$

Here, $\mathbf{S_{norm}}$ denotes the input cell-cell similarity matrix, and $\odot$ represents the Hadamard product. The term $\sum \mathbf{M}$ corresponds to the total number of within-batch cell pairs, while $\sum (1 - \mathbf{M})$ represents the total number of cross-batch cell pairs.

We applied this adaptive parameter selection strategy to the four datasets described above. In all cases, the adaptively selected β values were relatively small and consistently close to 0.2 (Supplementary Fig. 30a). Supplementary Figs. 30b-e and 31 show the resulting Graph cLISI scores and UMAP visualization obtained using the adaptive β values, demonstrating that the adaptive strategy achieves strong batch integration performance, indicating that the adaptive strategy recovers integration performance comparable to the empirically chosen default.

Finally, although scMethCraft is sensitive to the full range of β values between 0 and 1, this behavior is expected and reasonable, as the two extremes correspond to completely ignoring intra-batch information and treating intra-batch and inter-batch information equally, respectively, the latter effectively disabling batch integration. Importantly, the mathematical

interpretation of β enables principled, data-driven parameter selection, ensuring its effectiveness across different datasets. Moreover, the choice of β is decoupled from model training, which allows users to explore different values after training. In practice, users can test alternative integration strengths within minutes and select an appropriate value based on visual inspection, quantitative metrics, or specific application needs.

Supplementary Note 6: Single CpG site resolution scDNAm data analysis methods.

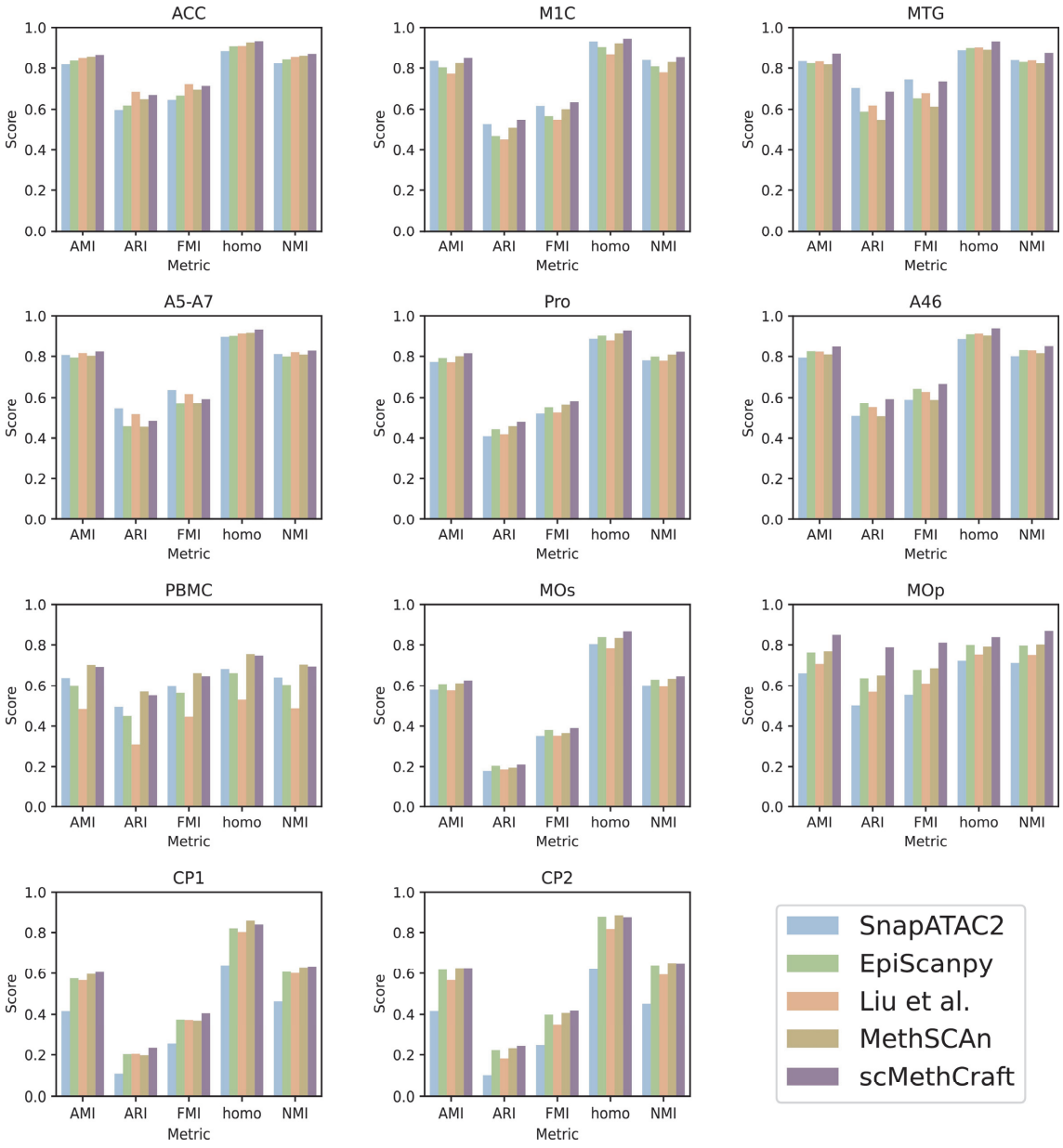
Several methods have been proposed for the analysis of single-cell DNA methylation data and have made important conceptual contributions to the field. For example, Epiclomal[14] is a probabilistic clustering approach based on hierarchical mixture models that jointly perform cell clustering and methylation state imputation on sparse scDNAm data. Melissa[15] is a Bayesian hierarchical model that exploits local methylation patterns to cluster cells and characterize epigenetic variability across the genome. scMET[16] extends hierarchical modeling to assess variability in methylation levels across genomic features. DeepCpG[17] and MambaCpG[18] are deep learning-based approaches that focus on predicting CpG level methylation states by integrating local DNA sequence context and information from neighboring CpG sites. Collectively, these methods have played an important role in advancing early single-cell methylation analysis and have substantially enriched the scDNAm methodological landscape.

Despite their methodological significance, these approaches share several limitations that restrict their applicability to modern large-scale scDNAm datasets. First, all of the above methods operate at single CpG site resolution, either for site-level prediction or for clustering based on CpG level signals. While such designs are effective for datasets with relatively small numbers of cells, they become increasingly impractical as the scale of scDNAm experiments continues to grow. With the rapid increase in throughput of single-cell DNA methylation technologies, region-based aggregation has become a more widely adopted analysis paradigm. In this setting, methylation signals are summarized across fixed genomic bins to produce a cell-by-region matrix that enables scalable downstream analysis.

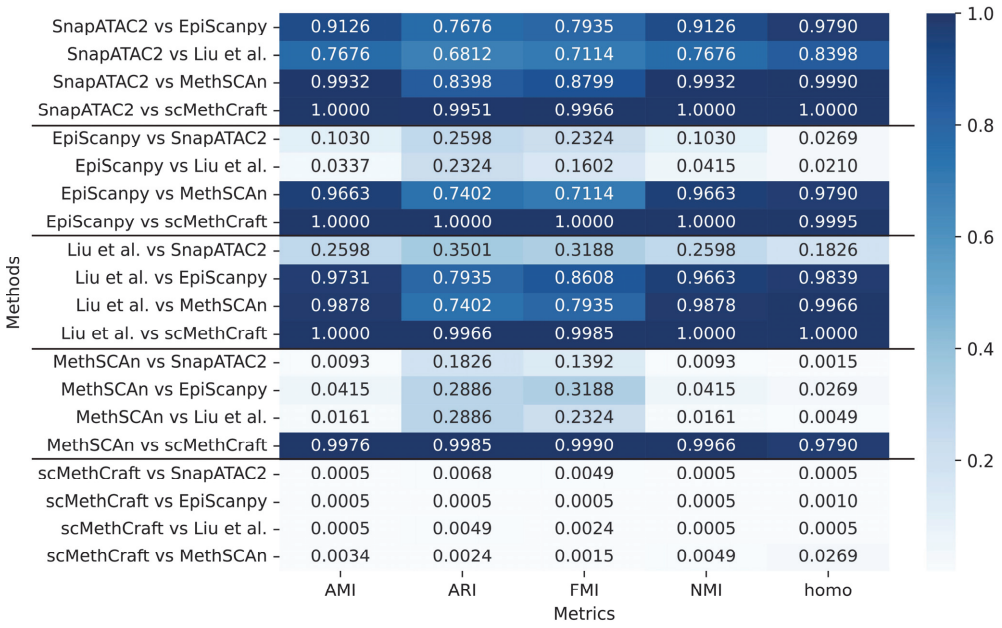
Second, the scalability of CpG level methods poses a major practical constraint. In the datasets analyzed in our study, which typically contain thousands of cells, we attempted to apply Epiclomal[14], Melissa[15], scMET[16], DeepCpG[17], and MambaCpG[18] under realistic computational settings. None of these methods could be completed within practical memory limits up to one terabyte or within reasonable runtime limits of a week. Similar scalability limitations have been explicitly documented in the peer review discussion of the MethSCAn[11], which provides a detailed analysis explaining why several earlier scDNAm methods are no longer suitable as baseline approaches for large-scale data analysis (https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-024-02347-x/MediaObjects/41592_2024_2347_MOESM2_ESM.pdf).

Third, these methods differ fundamentally from scMethCraft in terms of input representation, modeling objective, and application scenario. scMethCraft is specifically designed for binned scDNAm matrices and focuses on matrix-based downstream tasks such as imputation, dimensionality reduction, and data integration. As a result, direct performance comparisons between scMethCraft and CpG level models are not methodologically appropriate. To ensure completeness and fairness in benchmarking, we therefore focus our comparative analysis on methods that are compatible with region based scDNAm matrices and scalable to thousands of cells.

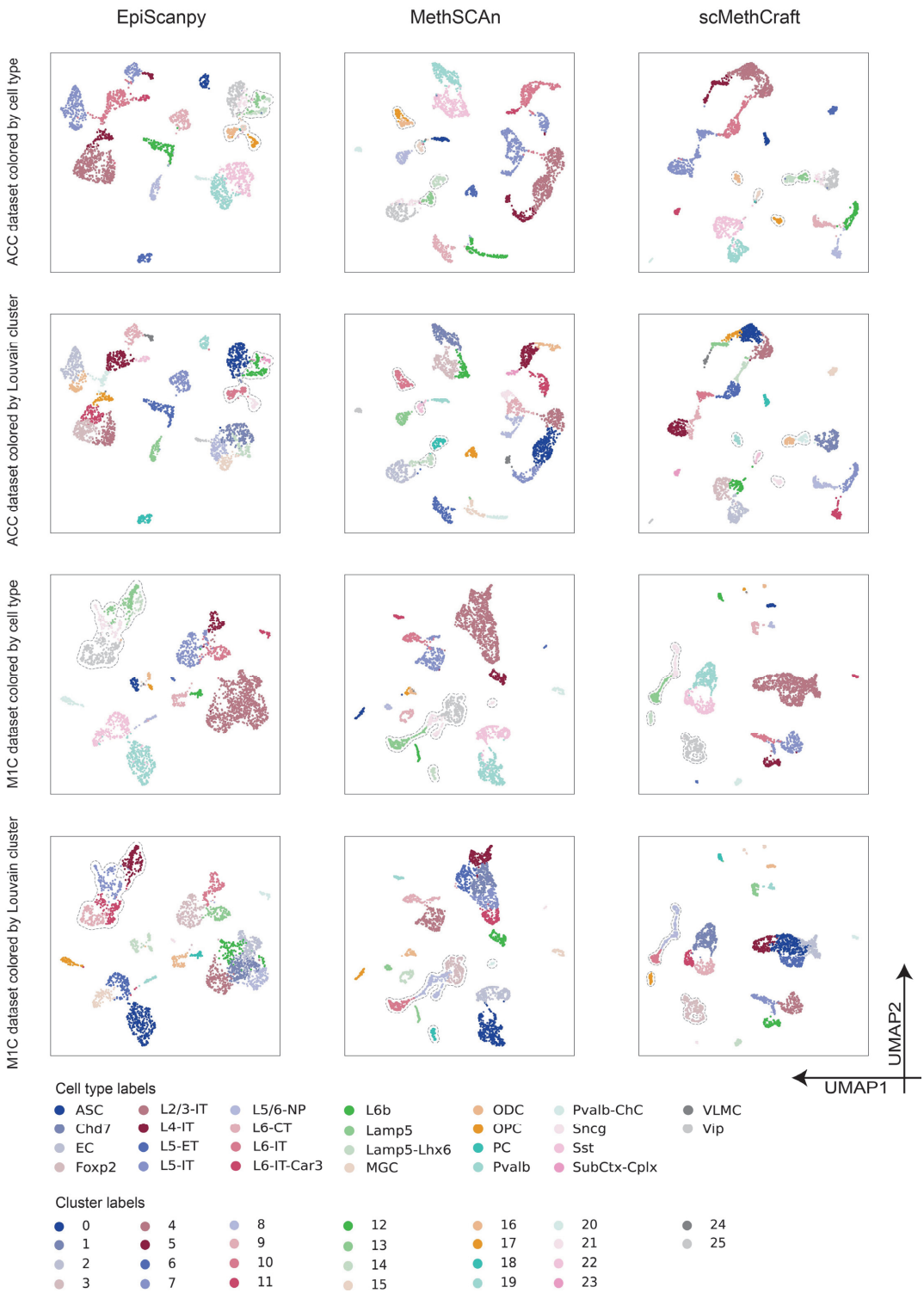
Supplementary Figures



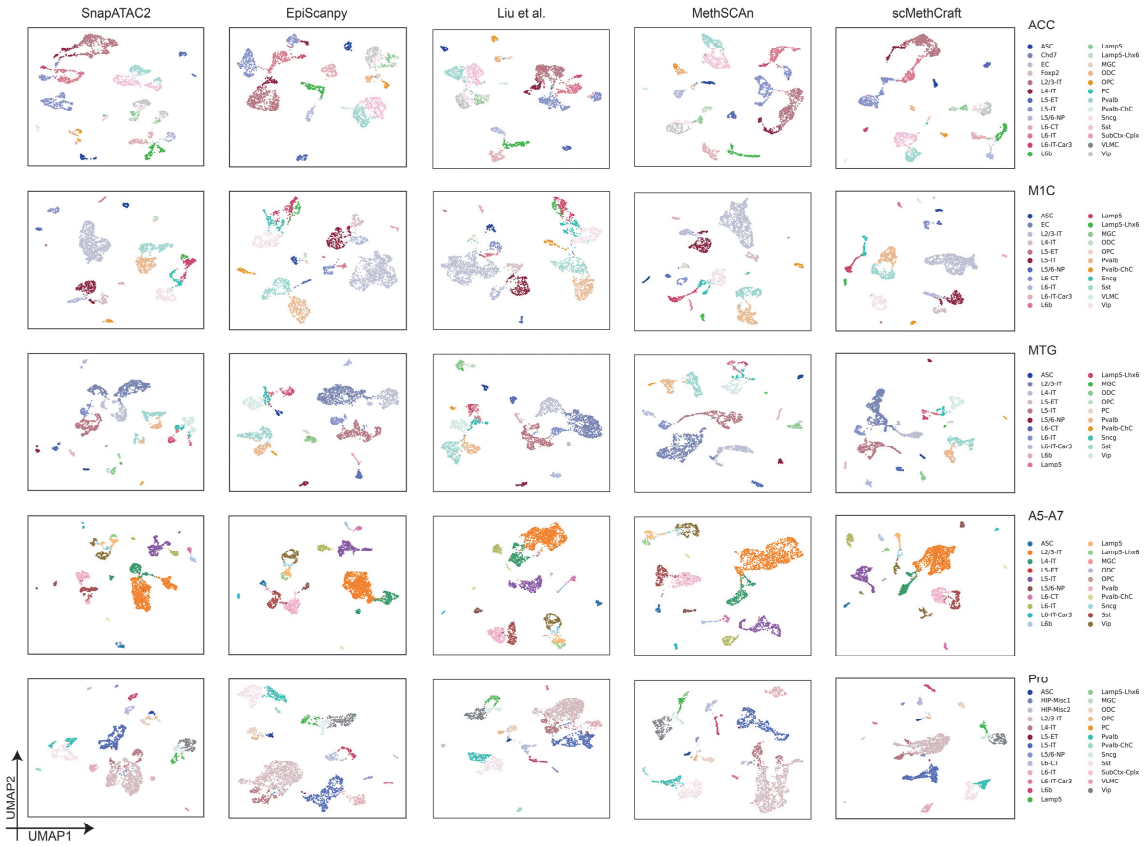
Supplementary Figure 1. The benchmarking of various dimensionality reduction methods. Cell clustering performance of embeddings generated by various methods was evaluated using five distinct metrics across 11 datasets.



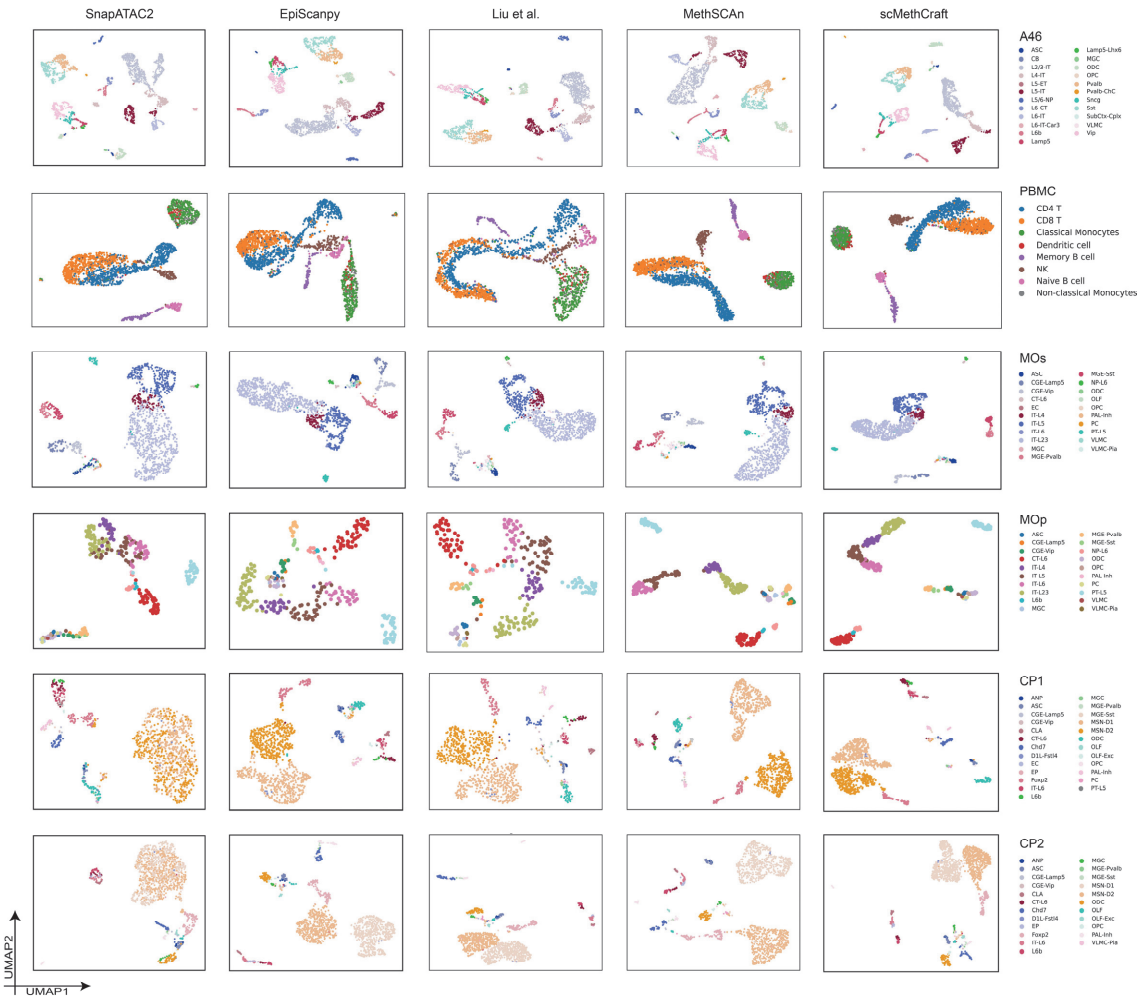
Supplementary Figure 2. Performance comparison of scMethCraft and other methods. P-values were obtained from one-sided paired Wilcoxon signed-rank tests to determine whether one method achieves significantly higher values of five metrics on the 11 datasets than another method (n = 11).



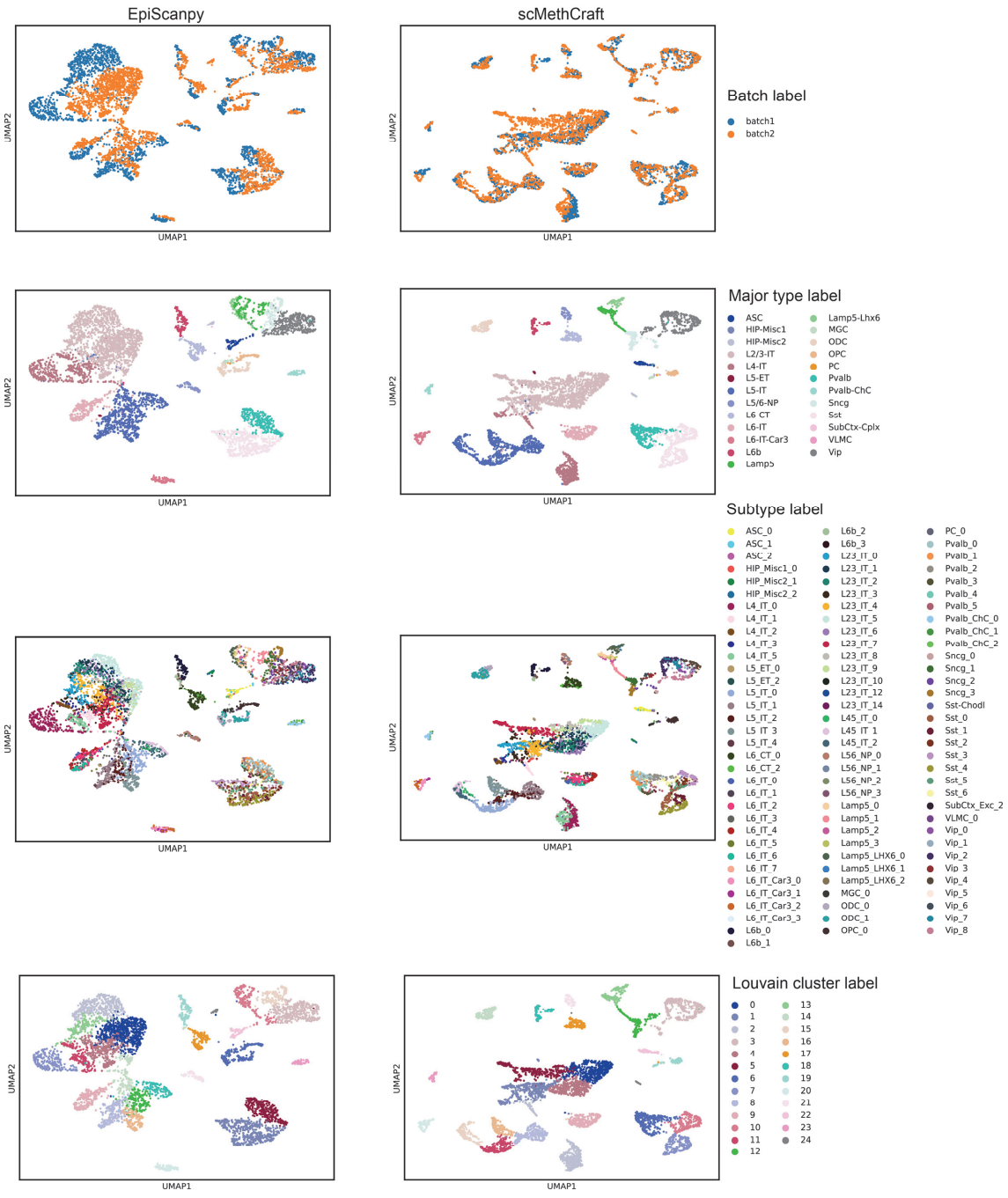
Supplementary Figure 3. UMAP visualization of cell embeddings provided by different methods on the ACC and M1C datasets. Cells are colored by cell types in the and by Louvain clusters.



Supplementary Figure 4. UMAP visualization of the cell embedding created by different methods. The UMAP visualization of ACC, M1C, MTG, A5-A7, and Pro datasets.

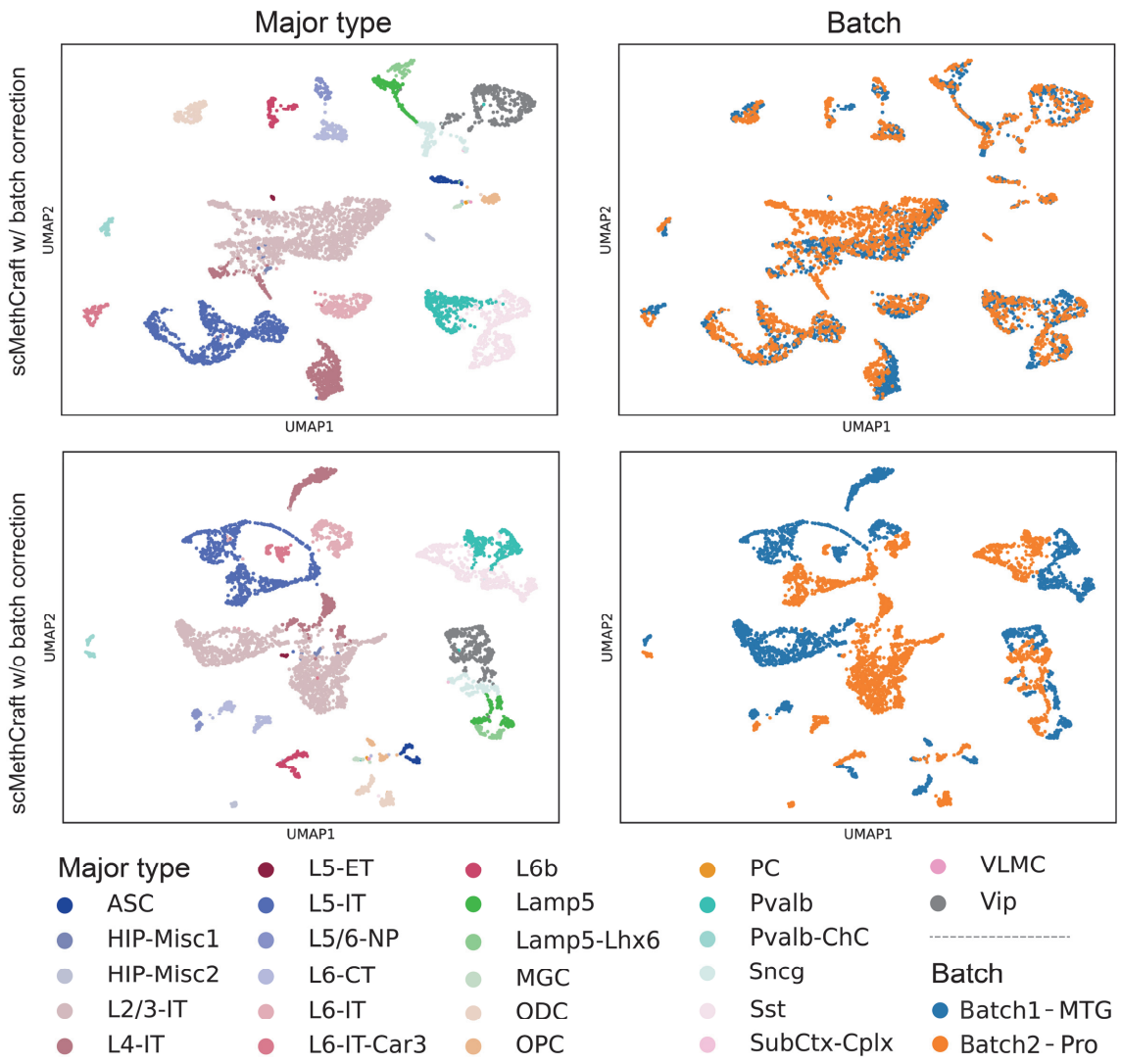


Supplementary Figure 5. UMAP visualization of the cell embedding created by different methods. The UMAP visualization of A46, PBMC, MOs, MOp, CP1, and CP2 datasets.

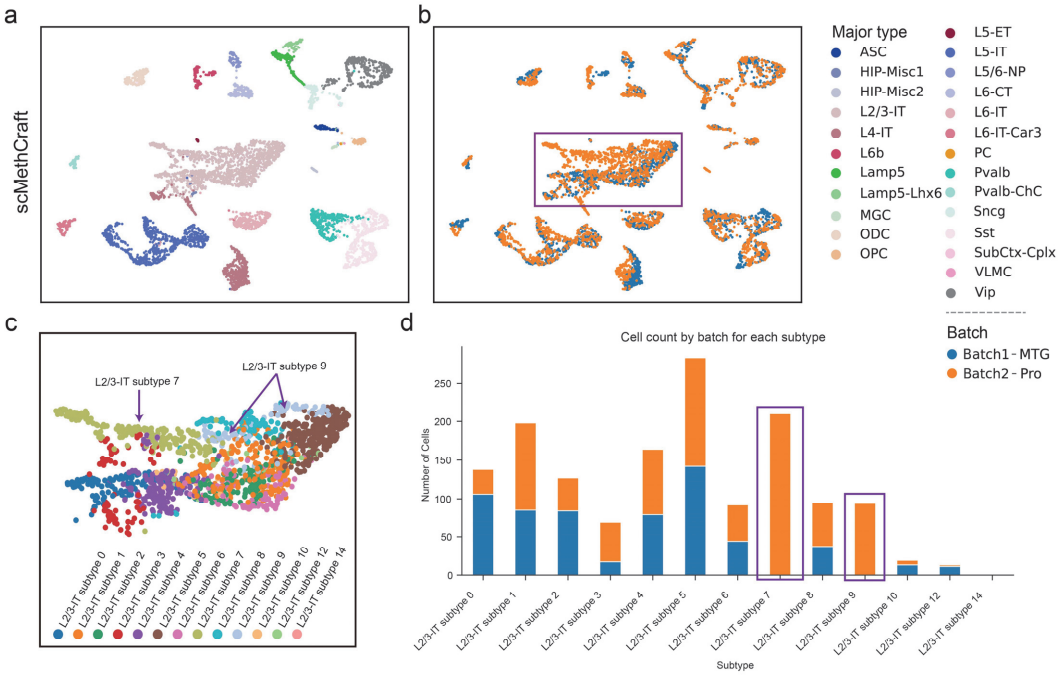


261

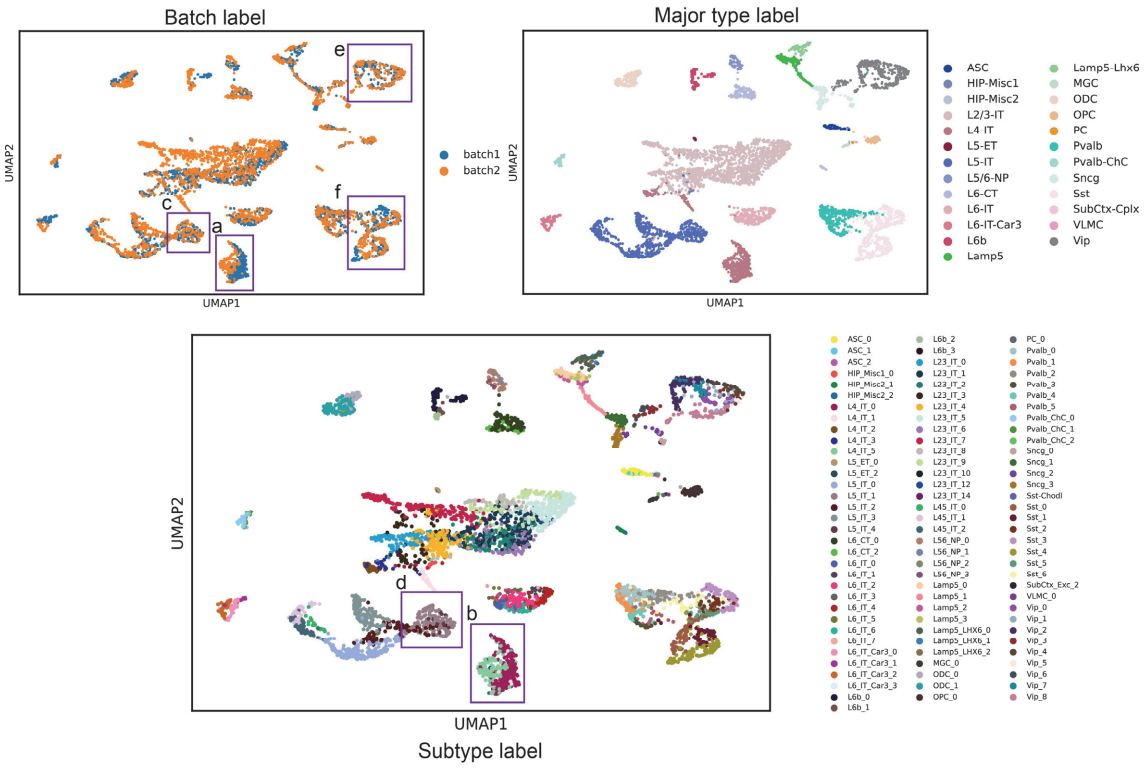
262 **Supplementary Figure 6. The visualization of different methods in batch integration.**
263 UMAP visualization of the integrated embeddings created by EpiScanpy and scMethCraft on
264 the MTG-Pro dataset. Batch labels, major type labels, subtype labels, and Louvain cluster
265 labels are projected onto the visualizations.



Supplementary Figure 7. The visualization of scMethCraft with batch correction and scMethCraft without batch correction on the MTG-Pro dataset. UMAP visualization of the integrated embeddings created by scMethCraft with batch correction and scMethCraft w/o batch correction on the MTG-Pro dataset. Batch labels and major type labels are projected onto the visualizations.



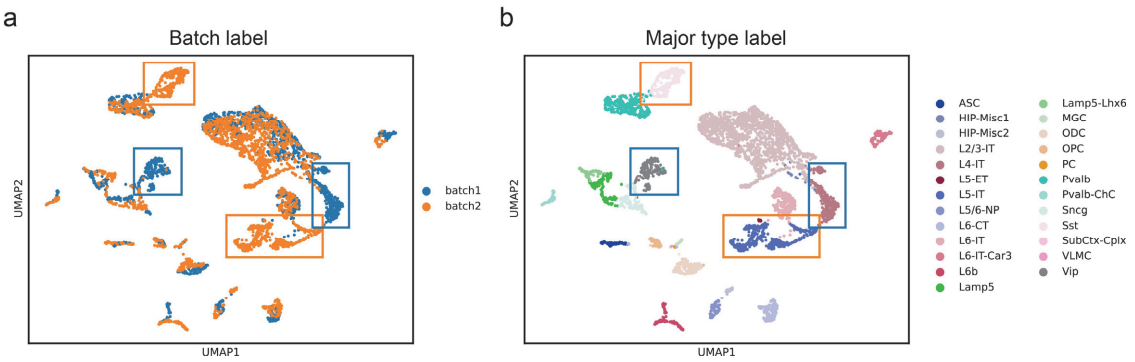
Supplementary Figure 8. scMethCraft preserves region-specific L2/3-IT subtypes during batch integration. a-b, UMAP visualization of the integrated embeddings provided by scMethCraft on the MTG-Pro dataset, colored by **(a)** Major type and **(b)** batch. The purple box indicates L2/3-IT cells. **c,** The L2/3-IT cells of scMethCraft's integrated embedding, colored by subtype. **d,** Subtype-specific batch composition of L2/3-IT cells.



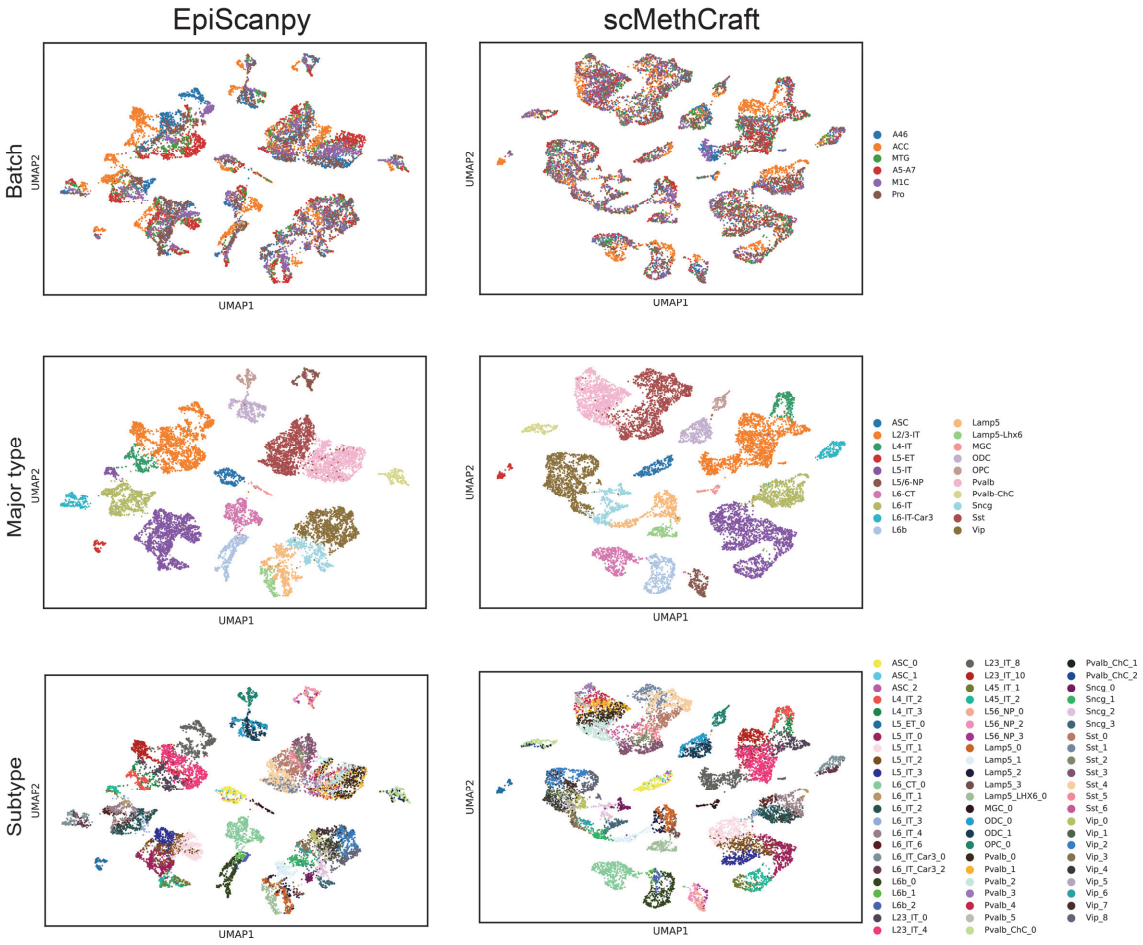
280

281 **Supplementary Figure 9. scMethCraft preserves biological heterogeneity while**
282 **integrating cell populations across batches.** scMethCraft preserves biological heterogeneity
283 while integrating cell populations across batches. UMAP visualization of the integrated
284 embeddings provided by scMethCraft on the MTG-Pro dataset, colored by batch, major type,
285 and subtype.

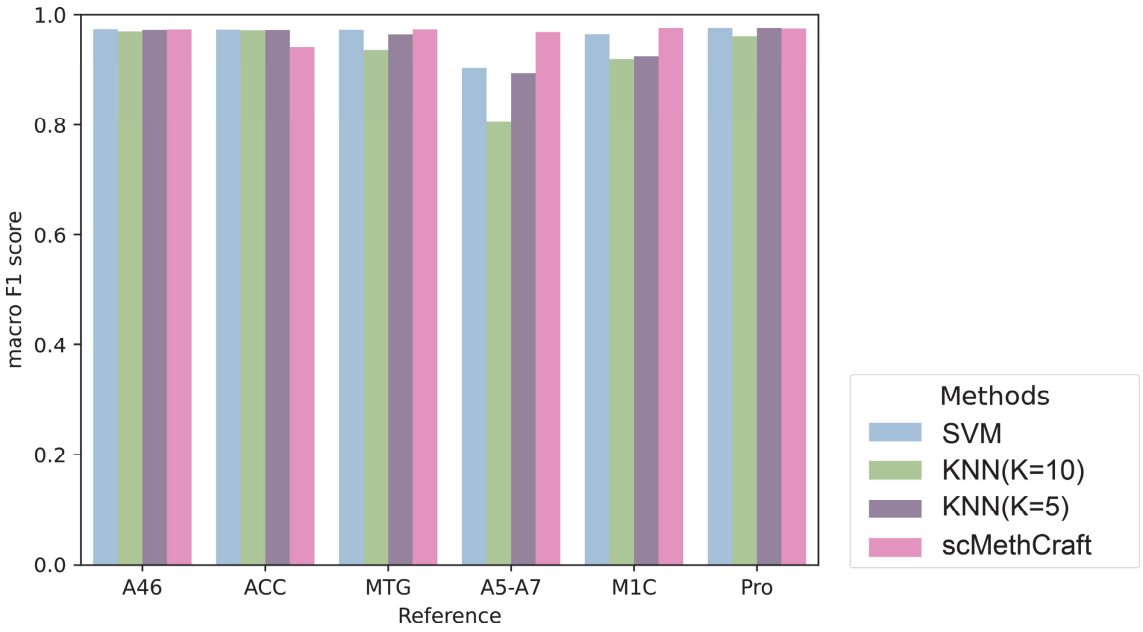
286



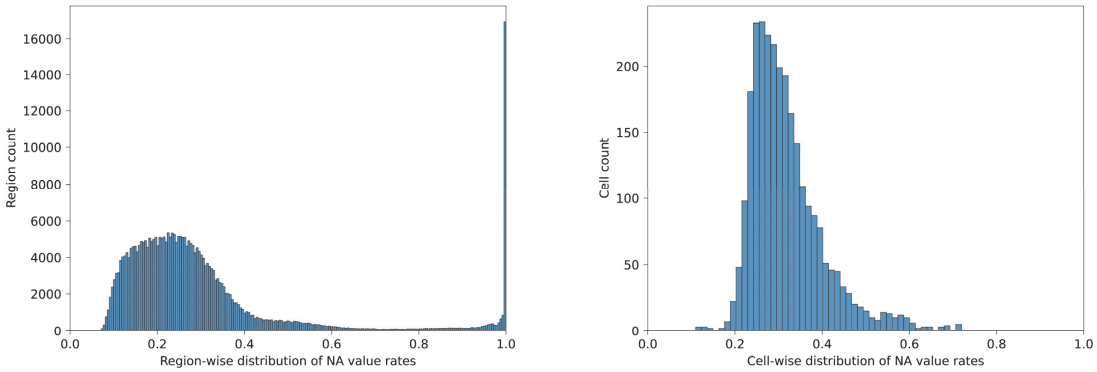
Supplementary Figure 10. Batch integration behavior of scMethCraft under partially overlapping cell type compositions. A partially overlapping scenario was generated by excluding L5-IT and SST cells from batch 1 and L4-IT and VIP cells from batch 2, producing a mixture of shared and batch-specific cell types for evaluating integration performance**. a,** UMAP visualization colored by batch labels. **b,** UMAP visualization colored by cell-type annotations. Blue boxes indicate cell types present exclusively in batch 1, while orange boxes indicate cell types present exclusively in batch 2.



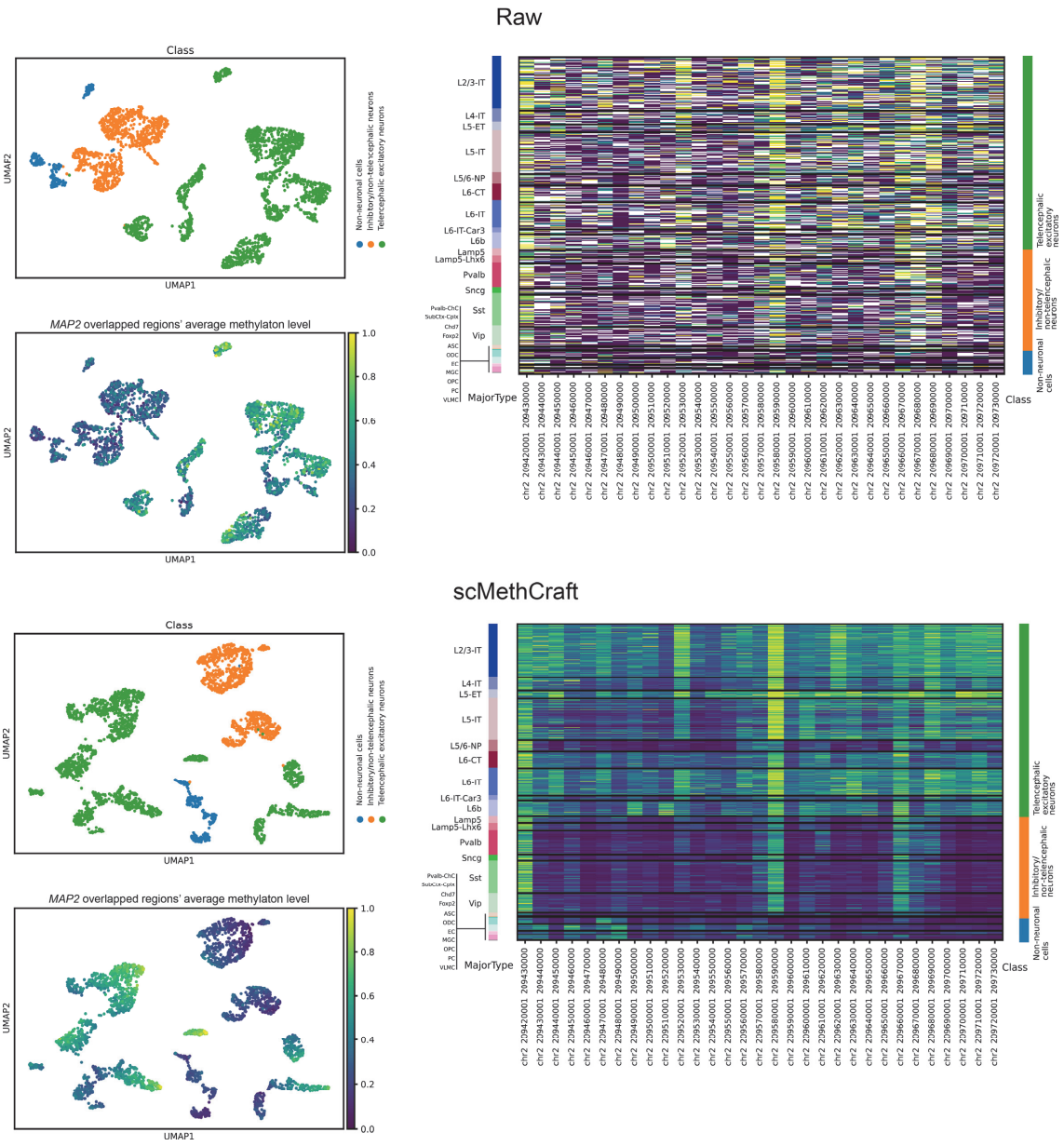
Supplementary Figure 11. The visualization of different methods in multi-batch integration. UMAP visualization of the integrated embeddings created by EpiScanpy and scMethCraft on the Mixed dataset. Batch labels, major type labels, and subtype labels are projected onto the visualizations.



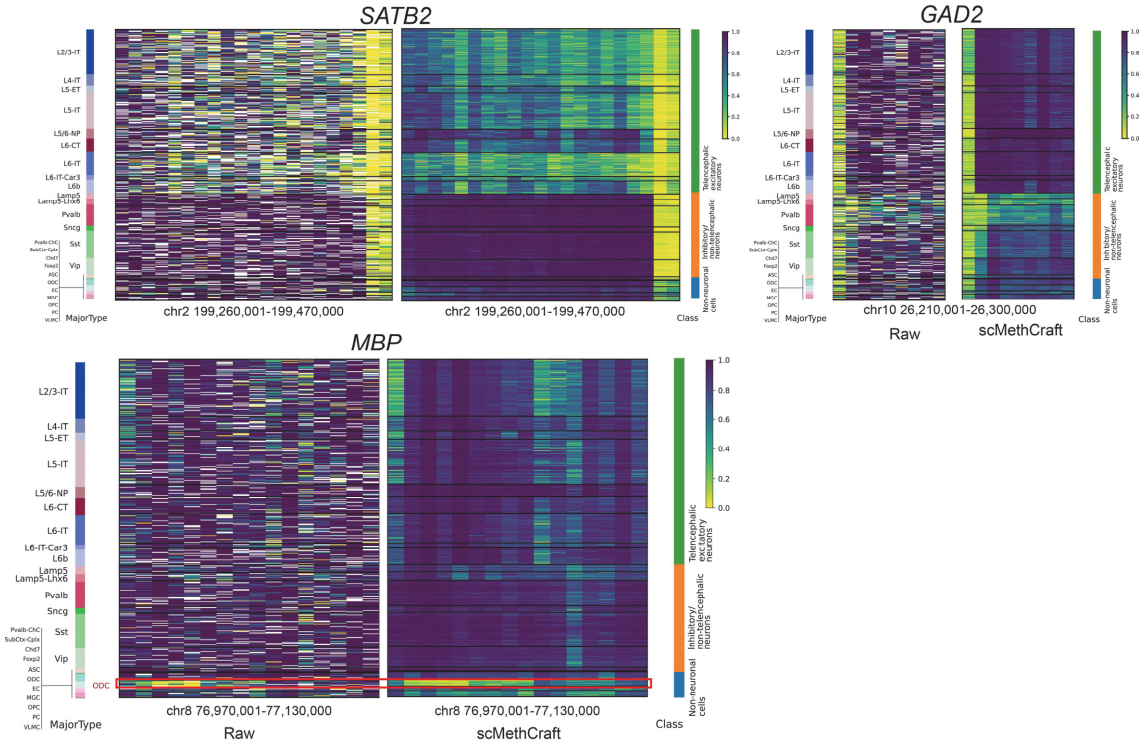
Supplementary Figure 12. Cell type annotation performance. Macro-F1 scores of scMethCraft and baseline methods in major type-level annotations.



Supplementary Figure 13. The proportion of explicit NA values of the ACC dataset. The histogram of the region-wise and cell-wise distribution of NA value rates. The x-axis represents the rate of NA value, and the y-axis represents the counts of regions or cells.

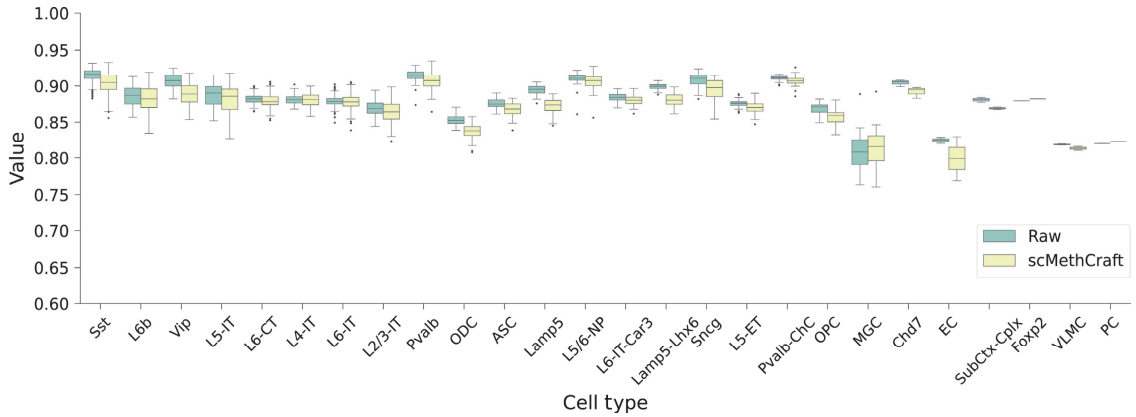


Supplementary Figure 14. The visualization and heatmap of the *MAP2* gene overlapped regions of the raw data and scMethCraft enhanced matrix. UMAP visualization of the raw and scMethCraft enhanced ACC dataset, class labels, and average methylation level of *MAP2* gene overlapping regions are projected onto the visualizations (left column). Heatmaps of the raw matrix and scMethCraft-enhanced matrix. The regions overlapping with the *MAP2* gene are shown (right column).

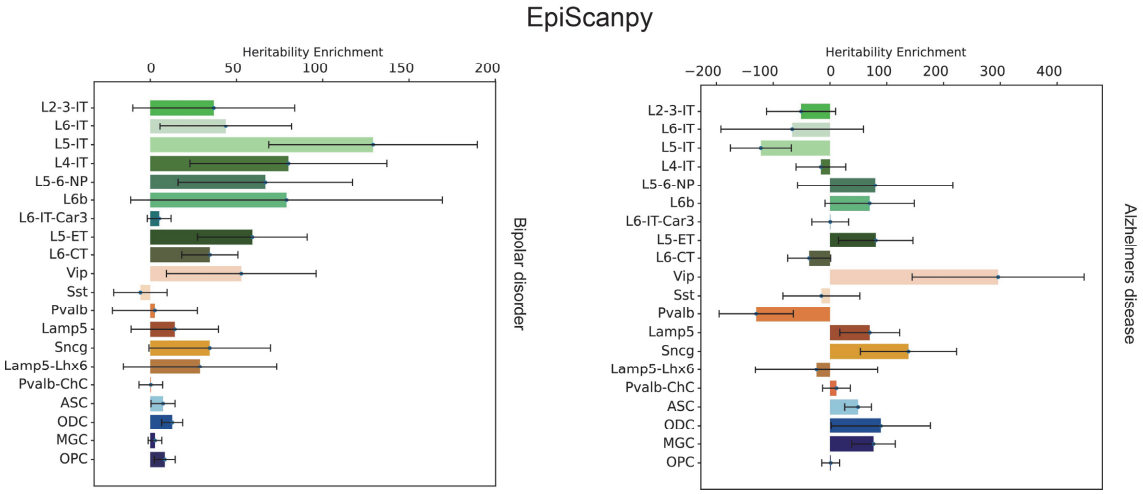


316

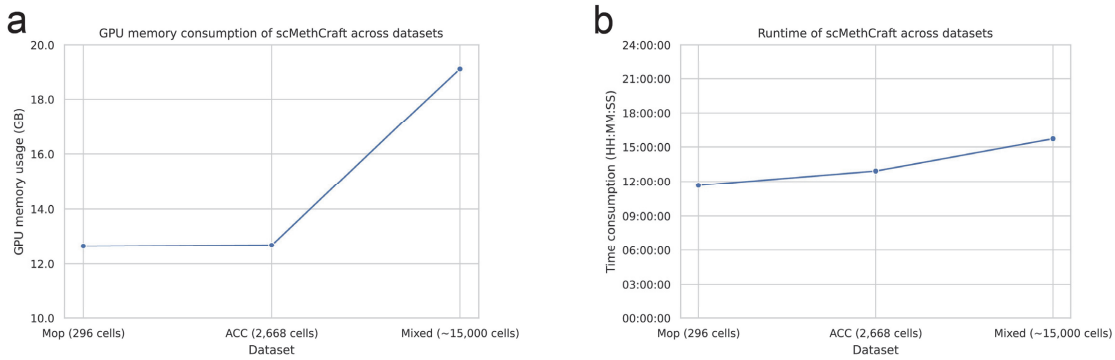
317 **Supplementary Figure 15. Heatmap of the *SATB2*, *MBP,* and *GAD2* genes overlapped**
318 **regions of raw data and scMethCraft enhanced matrix.** Heatmaps of the raw matrix and
319 scMethCraft-enhanced matrix. The regions overlapping with the *SATB2*, *MBP,* and *GAD2*
320 genes are shown.



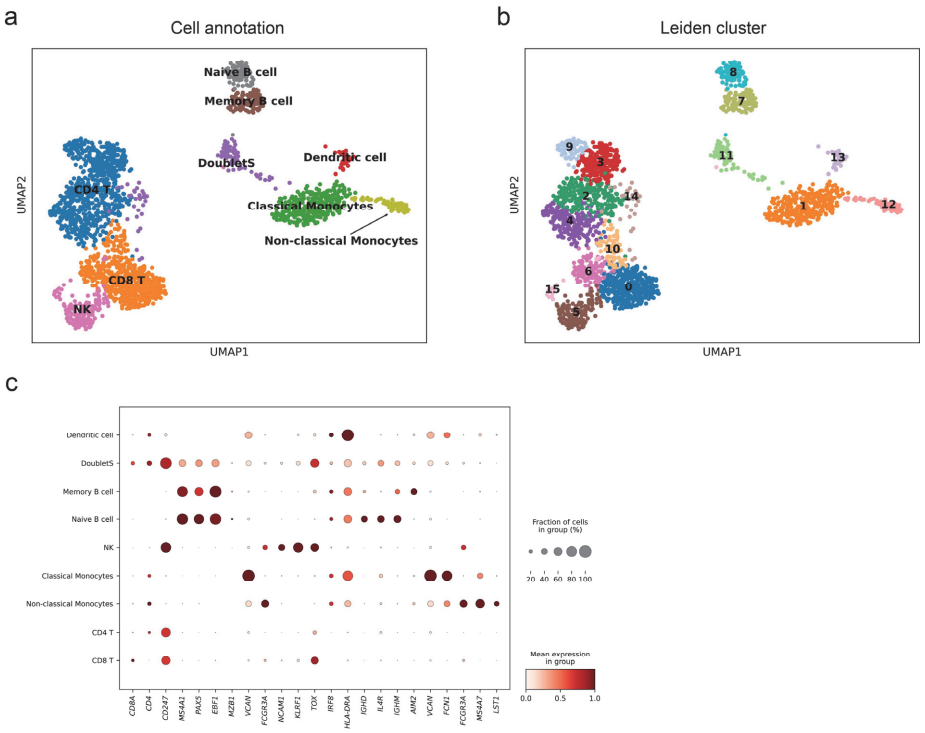
Supplementary Figure 16. Distribution of average DNA methylation levels within each cell type. The midline represents the median, the boxes represent the interquartile range, whiskers represent 1.5× interquartile and points represent outliers. The sample size varies across cell types and is provided in the Source Data.



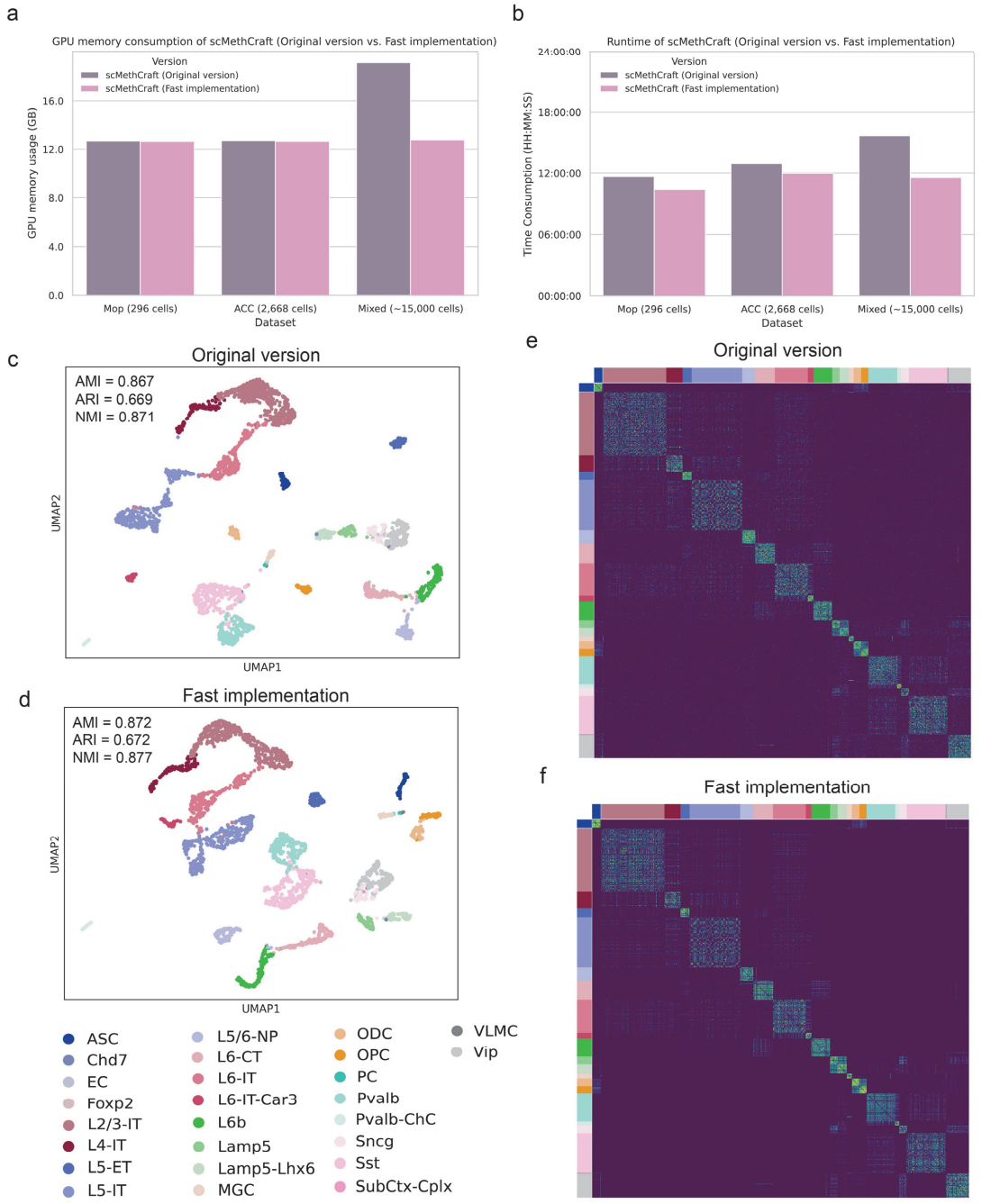
Supplementary Figure 17. Heritability enrichment estimated by stratified LDSC of the SNPs within the major type-specific hypo-DMRs identified by EpiScanpy, for two brain-related diseases. The error bars and centers of error bars represent the standard errors and average values of 200 equally sized jackknife blocks of adjacent SNPs about the estimates of enrichment.



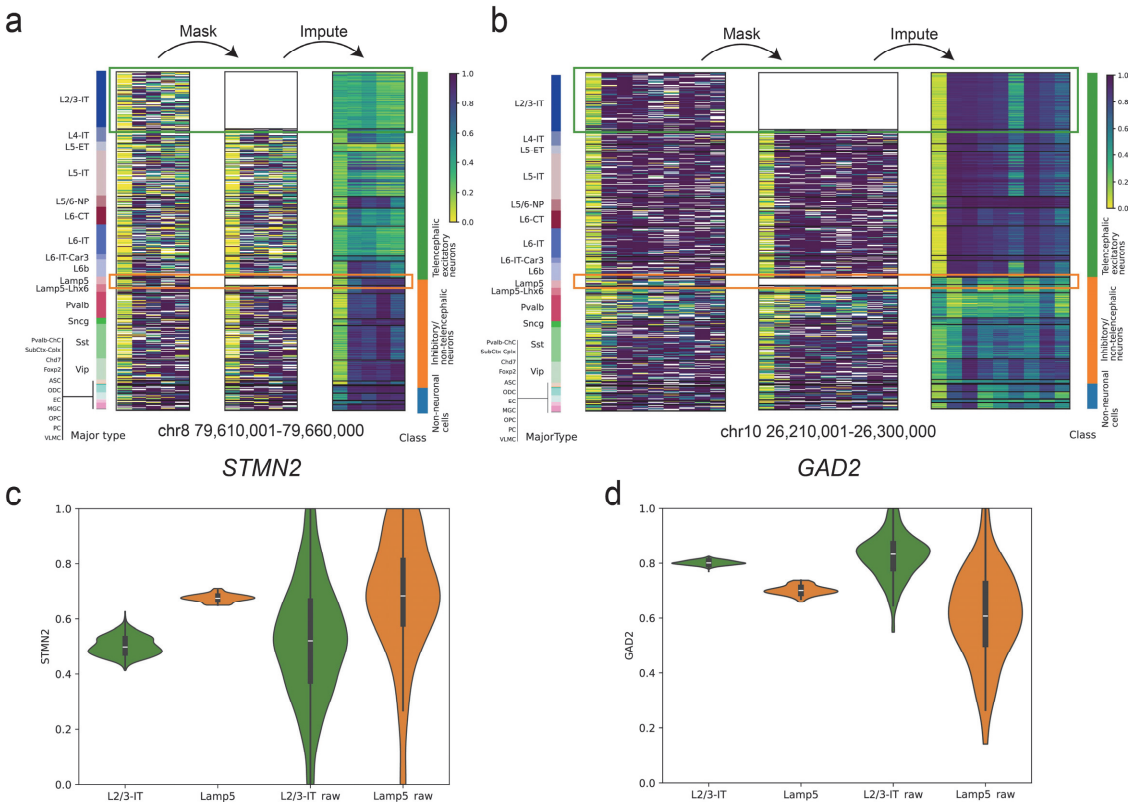
Supplementary Figure 18. GPU memory consumption and runtime of scMethCraft across datasets of increasing scale. a, Peak GPU memory usage during model training. **b,** Runtime of model training.



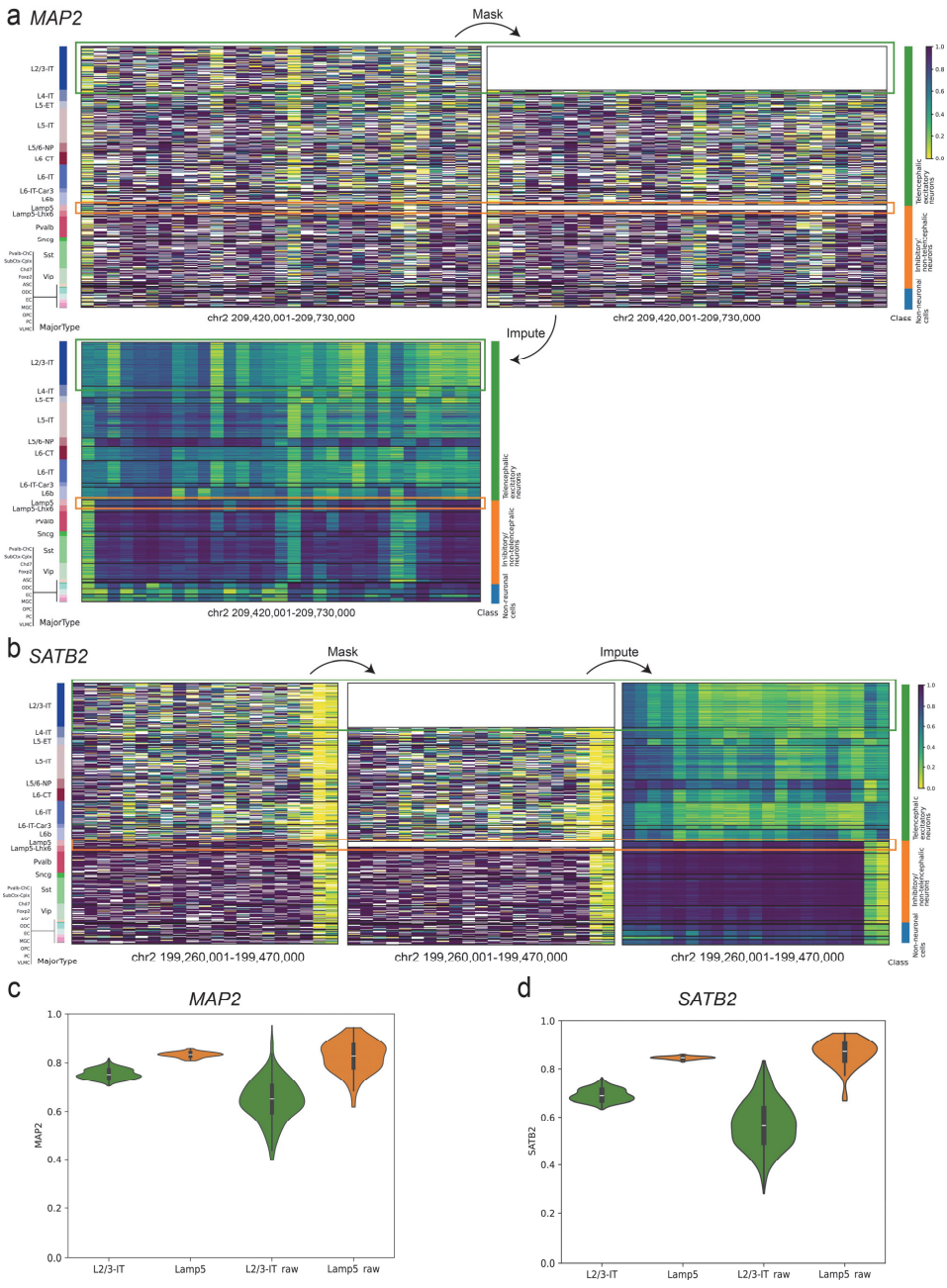
Supplementary Figure 19. Cell type annotation based on paired scRNA-seq data. a-b, UMAP visualization of scRNA-seq data colored by cell-type annotations and Leiden clusters, illustrating the correspondence between transcriptome-based clustering and manual cell-type annotation. **c,** Dot plot of the expression of canonical marker genes across annotated cell types based on paired scRNA-seq data.



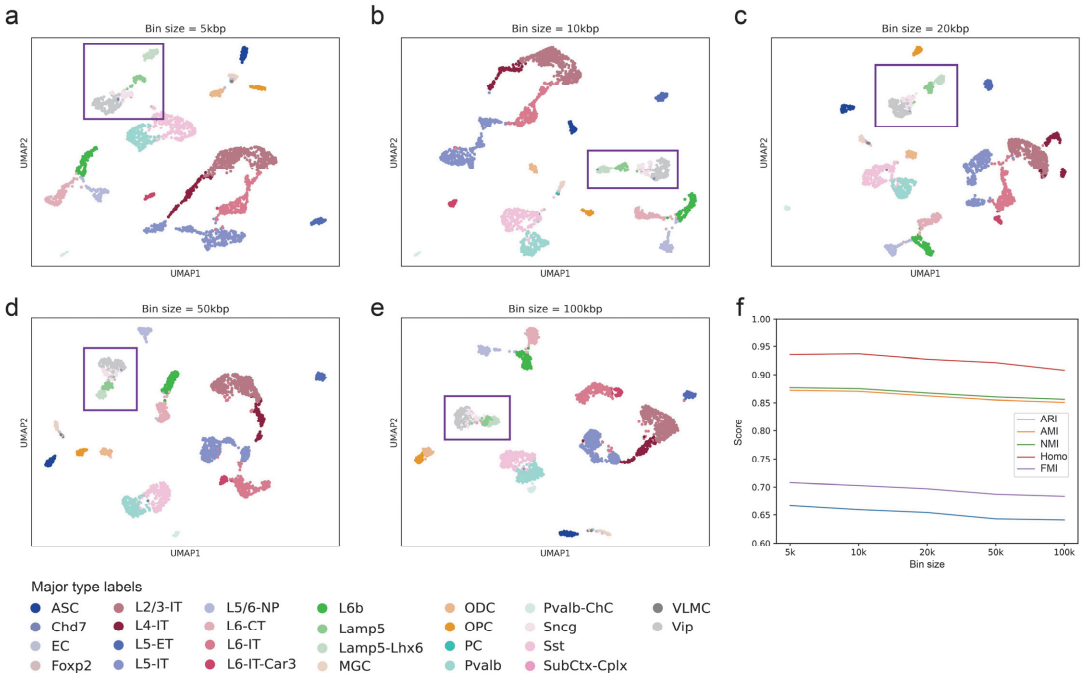
Supplementary Figure 20. Performance comparison of scMethCraft between the original version and the fast implementation. a, GPU memory consumption, and **b,** runtime of scMethCraft across datasets of increasing scale, contrasting the original version with the fast implementation. **c-d,** UMAP visualization and clustering performance of cell embeddings produced by **(c)** the original version and **(d)** the fast implementation. **e-f,** The heatmap of the cell-to-cell similarity matrix created by **(e)** the original version and **(f)** the fast implementation, where cells were sorted according to their major cell types. All similarity values are linearly scaled to the range 0-1.



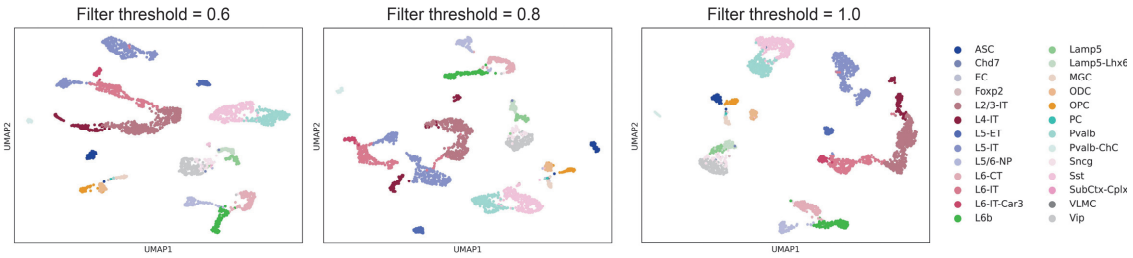
Supplementary Figure 21. scMethCraft effectively recovers masked cell-type-specific regions. a-b, Heatmaps of **(a)** STMN2 and **(b)** GAD2 gene overlapping regions. For each panel, the three heatmaps from left to right correspond to the raw data before masking, the raw data after masking, and the imputed results obtained by applying scMethCraft to the masked data. Green boxes indicate L2/3-IT cells, and orange boxes indicate Lamp5 cells. **c-d,** Violin plots showing the average methylation levels of L2/3-IT and Lamp5 cells in the **(c)** STMN2 and **(d)** GAD2 gene overlapping regions before masking and after imputation. The maximum width of each violin is identical, and kernel density estimates have been truncated to the data range.



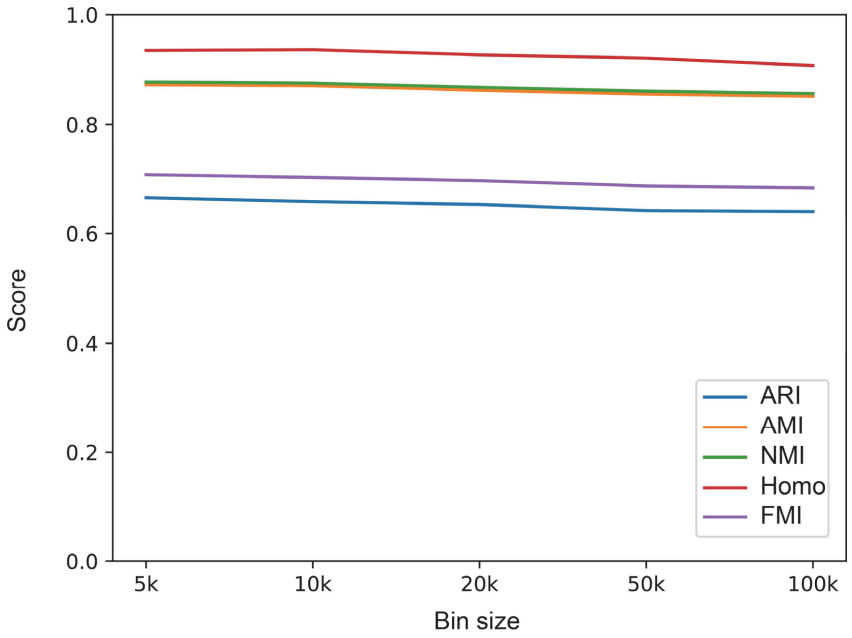
Supplementary Figure 22. scMethCraft effectively recovers masked cell-type-specific regions. a-b, Heatmaps of **(a)** MAP2 and **(b)** SATB2 gene overlapping regions. For each panel, the three heatmaps from left to right correspond to the raw data before masking, the raw data after masking, and the imputed results obtained by applying scMethCraft to the masked data. Green boxes indicate L2/3-IT cells, and orange boxes indicate Lamp5 cells. **c-d,** Violin plots showing the average methylation levels of L2/3-IT and Lamp5 cells in the **(c)** MAP2 and **(d)** SATB2 gene overlapping regions before masking and after imputation. The maximum width of each violin is identical, and kernel density estimates have been truncated to the data range.



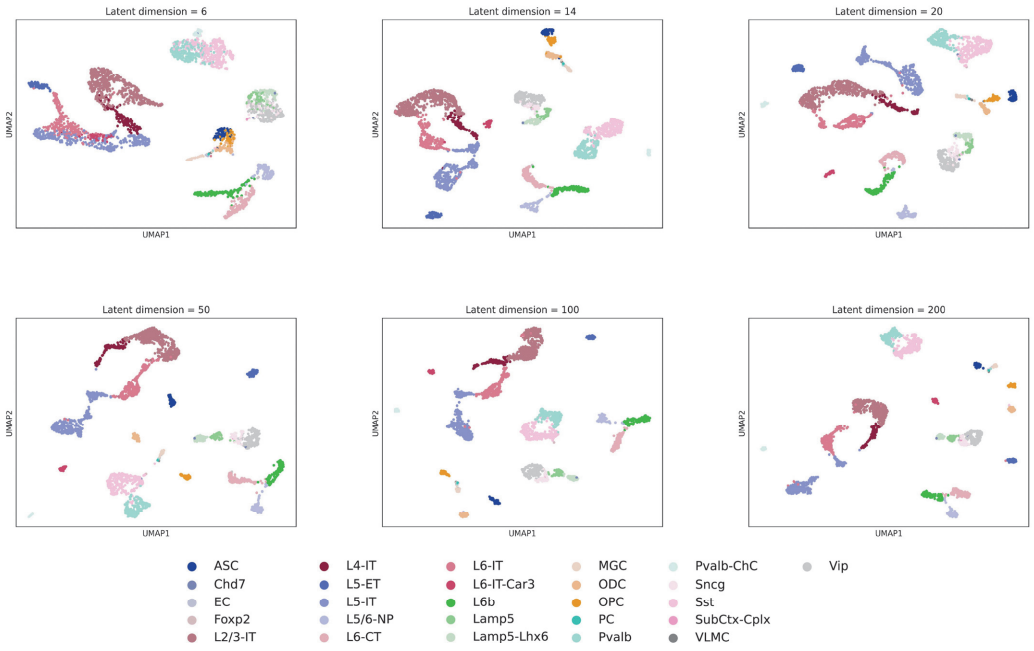
Supplementary Figure 23. Cell clustering performance of scMethCraft with varying bin sizes. a-e, UMAP visualization of the embedding created by scMethCraft with varying bin size. **f,** Cell clustering performance of various bin sizes.



Supplementary Figure 24. UMAP visualization of the embedding created by scMethCraft with varying filter thresholds. UMAP visualization of the embedding created by scMethCraft with varying filter thresholds from 0.6 to 1.0.



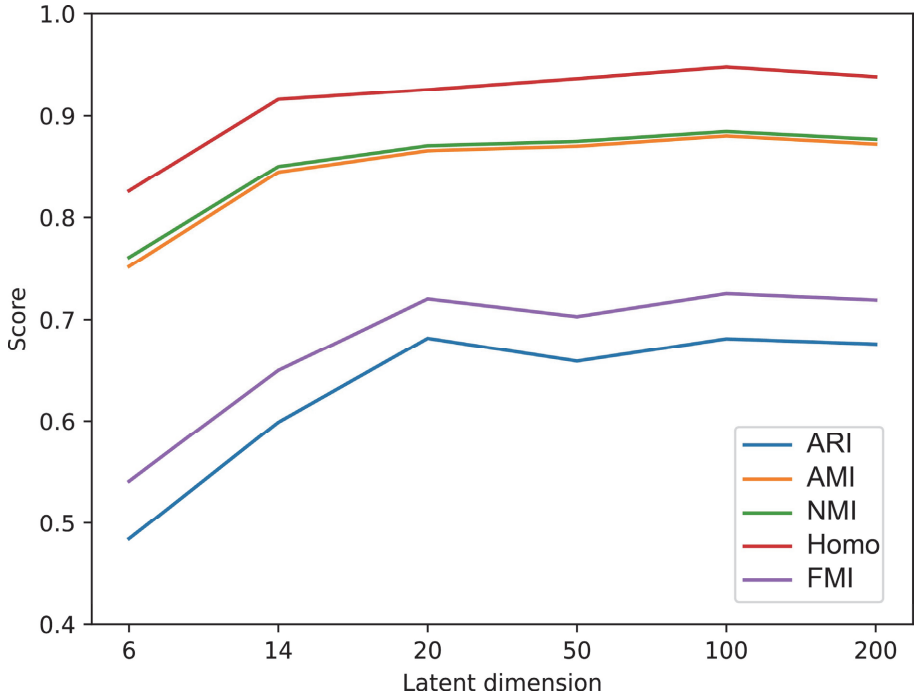
Supplementary Figure 25. Cell clustering performance of various bin sizes. Cell clustering performance of various bin sizes from 5k to 100k.



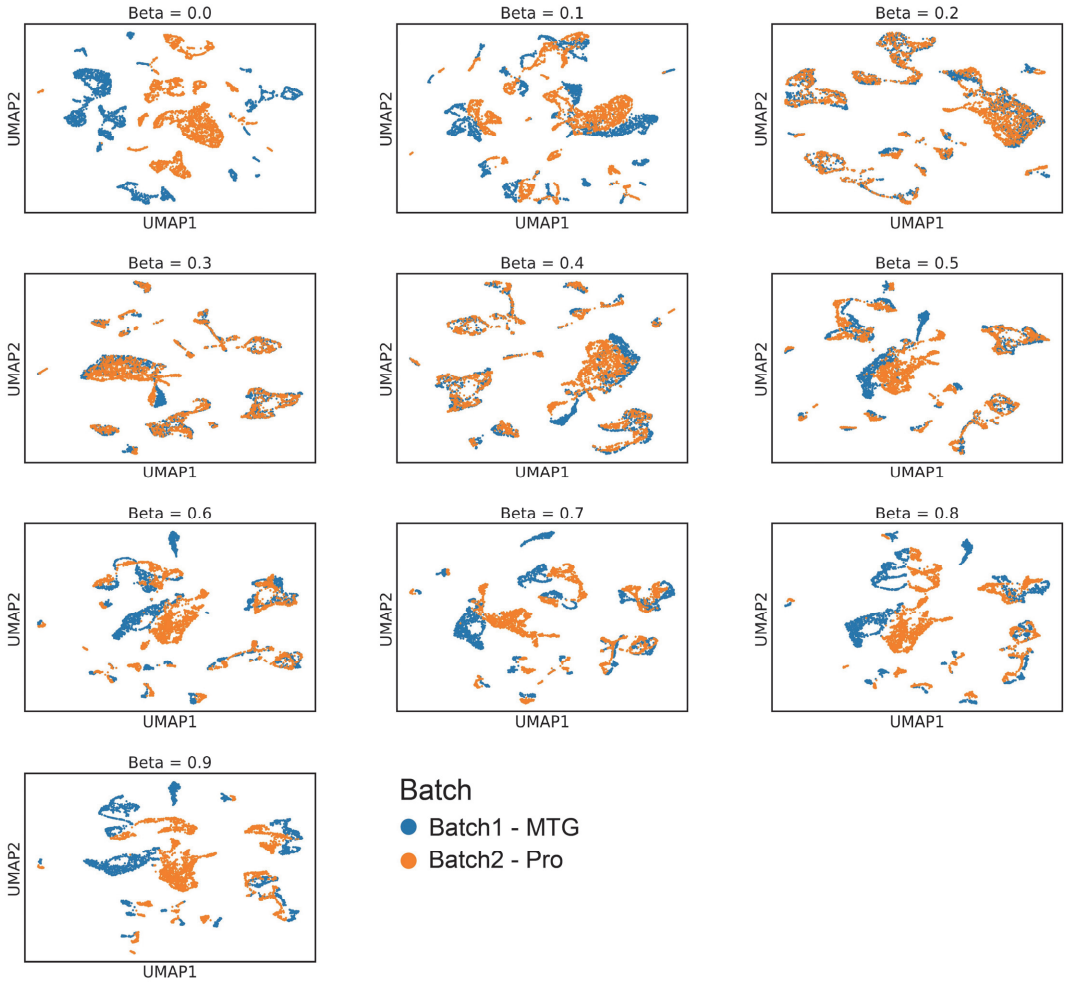
386

387 **Supplementary Figure 26. UMAP visualization of the embedding created by scMethCraft**
388 **with varying latent dimensions.** UMAP visualization of the embedding created by
389 scMethCraft with varying latent dimensions from 6 to 200.

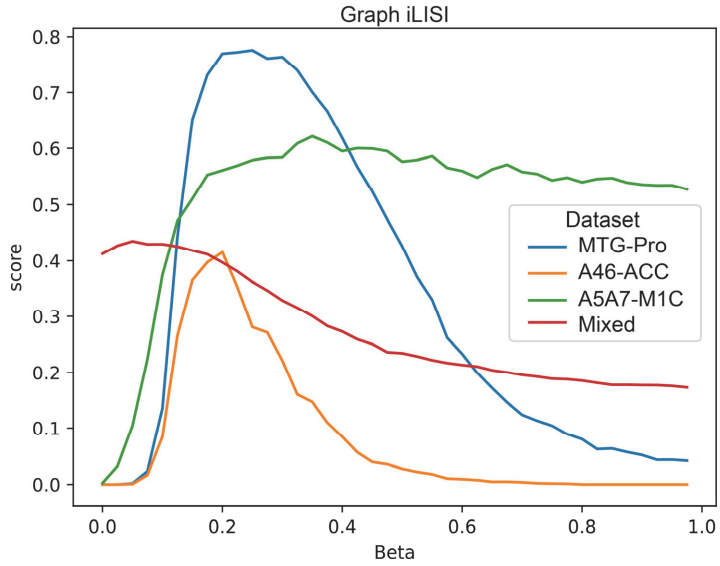
390



Supplementary Figure 27. Clustering performance of various latent dimensions. Clustering performance of various latent dimensions from 6 to 200.

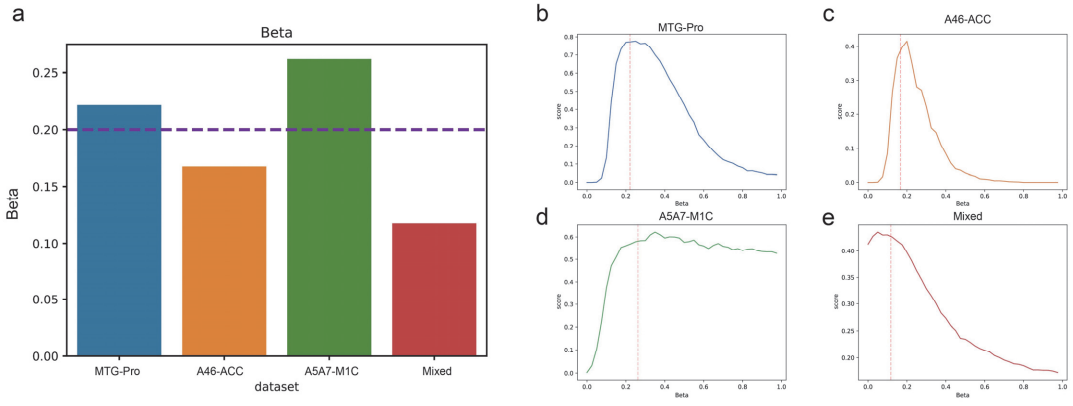


Supplementary Figure 28. UMAP visualization under different values of β. UMAP visualization under different values of β from 0 to 0.9.



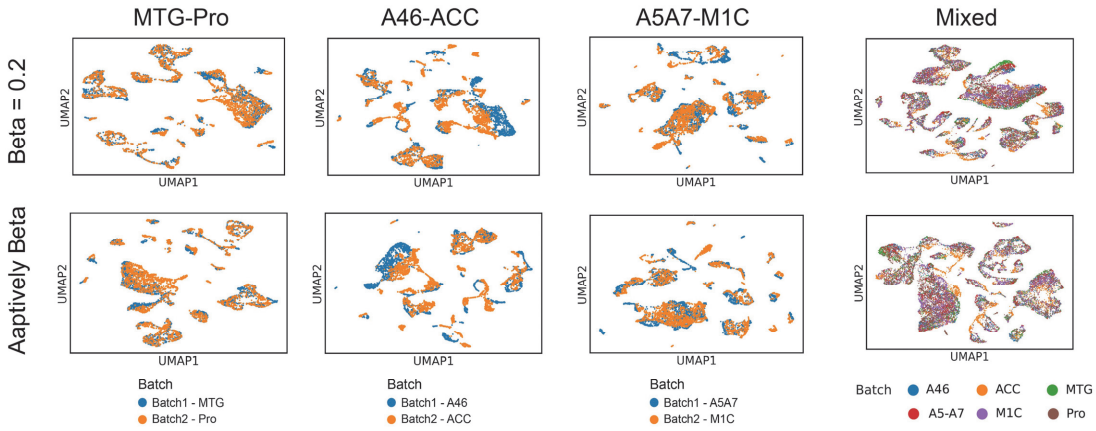
Supplementary Figure 29. Graph iLISI scores across different β values on four datasets.

Graph iLISI scores across different β values (from 0 to 1.0) on four datasets.



403

404 **Supplementary Figure 30. Adaptive parameter selection strategy β values across four**
405 **datasets. a,** Adaptively selected β values for each dataset. **b-e,** Graph cLISI scores evaluated
406 across different β values for each dataset. The dashed vertical line indicates the adaptively
407 selected β value.



408

409 **Supplementary Figure 31. UMAP visualization of adaptive parameter selection strategy**

410 ***β*** **values across four datasets.** UMAP visualization of adaptive parameter selection strategy

411 β values and fixed β values (0.2) across four datasets.

Supplementary Tables

Supplementary Table 1. Summary of datasets used in this study.

Dataset	Species	Tissue	No. of cells	No. of major types	No. of subtypes	Accession number
ACC	*Homo sapiens*	Brain – Cerebral cortex (Cx) – Anterior cingulate cortex – ACC	2,698	26	89	GSE168066
M1C	*Homo sapiens*	Brain – Cerebral cortex (Cx) – Precentral gyrus (PrCG) – Primary motor cortex – M1C	2,951	22	84	GSE168645
MTG	*Homo sapiens*	Brain – Cerebral cortex (Cx) – Middle Temporal Gyrus – MTG	2,496	21	86	GSE168734
A5-A7	*Homo sapiens*	Brain – Cerebral cortex (Cx) – Supraparietal lobule (SPL) – Posterosuperior (dorsal) parietal cortex – A5-A7	2,786	20	85	GSE179610
Pro	*Homo sapiens*	Brain – Cerebral cortex (Cx) – Cuneus, rostral part – Area Prostriata – Pro	2,730	25	95	GSE179971
A46	*Homo sapiens*	Brain – Cerebral cortex (Cx) – Middle frontal gyrus (MFG) – A46	2,518	23	84	GSE167577
PBMC	*Homo sapiens*	PBMC	2,094	8	0	PRJCA056651
MOs	*Mus musculus*	Brain – Secondary motor area	1,367	21	63	GSE130553
MOp	*Mus musculus*	Brain – Primary motor area	296	20	50	GSE131354
CP1	*Mus musculus*	Brain – Caudoputamen	1,106	26	50	GSE131393
CP2	*Mus musculus*	Brain – Caudoputamen	1,433	23	50	GSE131406

Supplementary Table 2. The description of the major types of six human cerebral cortex datasets, adapted from Tian et al.[19]

Major type	Full terminology	NO. of subtypes	Cell class
L2/3-IT	Intratelencephalic projection excitatory neurons in cortex layer 2/3	13	Telencephalic excitatory neurons
L6-IT	Intratelencephalic projection excitatory neurons in cortex layer 6	9	Telencephalic excitatory neurons
L5-IT	Intratelencephalic projection excitatory neurons in cortex layer 5	8	Telencephalic excitatory neurons
L4-IT	Intratelencephalic projection excitatory neurons in cortex layer 4	6	Telencephalic excitatory neurons
L5/6-NP	Near-projecting excitatory neurons in cortex layer 5/6	5	Telencephalic excitatory neurons
L6b	Excitatory neurons in cortex layer 6b	4	Telencephalic excitatory neurons
L6-IT-Car3	Excitatory neurons of both intratelencephalic projection neurons in cortex layer 6 and neurons from Cla	4	Telencephalic excitatory neurons
L5-ET	Extratelencephalic projection excitatory neurons in cortex layer 5	3	Telencephalic excitatory neurons
L6-CT	Corticothalamic projection excitatory neurons in cortex layer 6	2	Telencephalic excitatory neurons
HIP-Misc1	A rare cell type from hippocampus	1	Telencephalic excitatory neurons
HIP-Misc2	Another rare cell type from hippocampus	2	Telencephalic excitatory neurons
Vip	Cortical inhibitory neurons with genes Vip hypomethylated	9	Inhibitory/non-telencephalic neurons
Sst	Cortical inhibitory neurons with genes Sst hypomethylated	8	Inhibitory/non-telencephalic neurons
Pvalb	Cortical inhibitory neurons with genes Pvalb hypomethylated	6	Inhibitory/non-telencephalic neurons
Lamp5	Cortical inhibitory neurons with gene Lamp5 hypomethylated	4	Inhibitory/non-telencephalic neurons
Sncg	Cortical inhibitory neurons, homologous cell types of which in mouse have genes Sncg hypomethylated	4	Inhibitory/non-telencephalic neurons

Lamp5-Lhx6	Cortical inhibitory neurons with genes Lamp5 and Lhx6 hypomethylated	4	Inhibitory/non-telencephalic neurons
Pvalb-ChC	Chandelier cells	3	Inhibitory/non-telencephalic neurons
SubCtx-Cplx	Subcortical cell types with huge heterogeneity	3	Inhibitory/non-telencephalic neurons
Chd7	Neurons from BF/BN with gene CHD7 hypomethylated	2	Inhibitory/non-telencephalic neurons
Foxp2	Neurons from BF/BN with gene Foxp2 hypomethylated	1	Inhibitory/non-telencephalic neurons
CB	Granule cells from cerebellum	1	Inhibitory/non-telencephalic neurons
ASC	Astrocytes	3	Non-neuronal Cells
ODC	Oligodendrocytes	2	Non-neuronal Cells
MGC	Macro glia	1	Non-neuronal Cells
OPC	Oligodendrocyte progenitor cell	1	Non-neuronal Cells
EC	Endothelial cell	1	Non-neuronal Cells
VLMC	Vascular Leptomeningeal Cell	1	Non-neuronal Cells
PC	Perivascular Cell	1	Non-neuronal Cells

416

Supplementary Table 3. The description of the major types of four mouse brain datasets, adapted from Liu et al.[10]

Major type	Notes	NO. of subtypes	Cell class
ET-L5	Cortex ET Exc neuron in Layer 5	6	Excitatory neuron
NP-L6	Cortex NP Exc neuron in Layer 6	5	Excitatory neuron
IT-L6	Cortex IT Exc neuron in Layer 6	4	Excitatory neuron
CT-L6	Cortex CT Exc neuron in Layer 6	4	Excitatory neuron
IT-L23	Cortex IT Exc neuron in Layer 2-3	4	Excitatory neuron
IT-L5	Cortex IT Exc neuron in Layer 5	3	Excitatory neuron
CLA	Excitatroy neruon from Claustrum. Inferred by dissection regions and region specific marker genes, e.g., B3gat2 ISH from ABA https://mouse.brain-map.org/gene/show/93607	2	Excitatory neuron
IT-L4	Cortex IT Exc neuron in Layer 4	2	Excitatory neuron
L6b	Cortex L6b neuron	2	Excitatory neuron
EP	Excitatory neuron from Endopriform Nucleus. Inferred by dissection regions and region specific marker genes, e.g., Npsr1 ISH from ABA https://mouse.brain-map.org/gene/show/106174	2	Excitatory neuron
OLF-Exc	Excitatory neuron from olfactory area dissections including MOB, AON and PIR. Inferred by dissection regions	2	Excitatory neuron
MGE-Sst	MGE derived Sst+ Inh.	10	Inhibitory neuron
CGE-Vip	CGE derived Vip+ Inh.	7	Inhibitory neuron

MGE-Pvalb	MGE derived Pvalb+ Inh.	7	Inhibitory neuron
PAL-Inh	PAL Interneuron	7	Inhibitory neuron
CGE-Lamp5	CGE derived Lamp5+ Inh.	5	Inhibitory neuron
MSN-D1	Medium Spiny Neuron D1-type, hypomethylation in Drd1 gene	5	Inhibitory neuron
Foxp2	Inh cell marked by Foxp2 gene, mainly come from STR and PAL	4	Inhibitory neuron
MSN-D2	Medium Spiny Neuron D2-type, hypomethylation in Drd2 gene	4	Inhibitory neuron
Chd7	Inh cell shown hypomethylation in both Lhx6 and Chd7 gene	3	Inhibitory neuron
D1L-Fstl4	D1 Like Interneuron, marked by Fstl4 and Drd1	3	Inhibitory neuron
OLF	Granule layer of OLF and associated cell types	2	Inhibitory neuron
ASC	Astrocyte	3	Non-neuronal cells
VLMC	Vascular Leptomeningeal Cell	2	Non-neuronal cells
OPC	Oligodendrocyte Precursor Cell	2	Non-neuronal cells
ODC	Oligodendrocyte	2	Non-neuronal cells
VLMC-Pia	Vascular Leptomeningeal Cell mainly in Pia	1	Non-neuronal cells
MGC	Microglia	1	Non-neuronal cells
PC	Perivascular Cell	1	Non-neuronal cells
EC	Endothelial Cell	1	Non-neuronal cells
ANP	Adulut Neuron Precursor	1	Non-neuronal cells

418

Supplementary Table 4. ODC-relevant genes associated with ODC hypo-DMRs identified by scMethCraft that appeared in at least three instances.

Gene	Description	Reference
ZBTB20	This gene acts as a transcriptional repressor and plays a role in many processes including neurogenesis, glucose homeostasis, and postnatal growth.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=ZBTB20&keywords=ZBTB20
NFIB	Involved in brain development and regulation of DNA-templated transcription. Located in fibrillar center and nucleoplasm.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=NFIB&keywords=NFIB
PDE4B	The encoded protein regulates the cellular concentrations of cyclic nucleotides and thereby play a role in signal transduction. Altered activity of this protein has been associated with schizophrenia and bipolar affective disorder.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=PDE4B&keywords=PDE4B
HTRA1	By acting on TGF-beta signaling, may regulate many physiological processes, including retinal angiogenesis and neuronal survival and maturation during development.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=HTRA1&keywords=HTRA1
ADGRG1	Receptor involved in cell adhesion and probably in cell-cell interactions. Mediates cell matrix adhesion in developing neurons and hematopoietic stem cells. Receptor for collagen III/COL3A1 in the developing brain and involved in regulation of cortical development, specifically in maintenance of the pial basement membrane integrity and in cortical lamination.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=ADGRG1&keywords=%27ADGRG1%27
NFIA	Diseases associated with NFIA include Brain Malformations With Or Without Urinary Tract Defects and Bilateral Polymicrogyria.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=NFIA&keywords=%27NFIA%27
FYN	Plays a role in neural processes by phosphorylating DPYSL2, a multifunctional adapter protein within the central nervous system, ARHGAP32, a regulator for Rho family GTPases implicated in various neural functions, and SNCA, a small pre-synaptic protein.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=FYN&keywords=FYN
APBB2	The protein encoded by this gene interacts with the cytoplasmic domains of amyloid beta (A4) precursor protein and amyloid beta (A4) precursor-like protein 2. This protein contains two phosphotyrosine binding (PTB) domains, which are thought to function in signal transduction. Polymorphisms in this gene have been associated with Alzheimer's disease. Alternative splicing results in multiple transcript variants.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=APBB2&keywords=APBB2
LRRC8D	LRRC8D (Leucine Rich Repeat Containing 8 VRAC Subunit D) is a Protein Coding gene. Diseases associated with LRRC8D include Megalencephalic Leukoencephalopathy With Subcortical Cysts and Leukoencephalopathy With Vanishing White Matter 5.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=LRRC8D&keywords=LRRC8D
NTM	This gene encodes a member of the IgLON (LAMP, OBCAM, Ntm) family of immunoglobulin (Ig) domain-containing glycosylphosphatidylinositol (GPI)-anchored cell adhesion molecules. The encoded protein may promote neurite outgrowth and adhesion via a homophilic mechanism.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=NTM&keywords=NTM
PTPRZ1	Protein tyrosine phosphatase that negatively regulates oligodendrocyte precursor proliferation in the embryonic spinal cord. Required for normal differentiation of the precursor cells into mature, fully myelinating oligodendrocytes. May play a role in protecting oligondendrocytes against apoptosis. May play a role in the establishment of contextual memory, probably via the dephosphorylation of proteins that are part of important signaling cascades	https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTPRZ1&keywords=PTPRZ1

PTN	Binds cell-surface proteoglycan receptor via their chondroitin sulfate (CS) groups. Thereby regulates many processes like cell proliferation, cell survival, cell growth, cell differentiation and cell migration in several tissues namely neuron and bone.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTN&keywords=PTN
GPRC5B	Diseases associated with GPRC5B include Megalencephalic Leukoencephalopathy With Subcortical Cysts 3 and Megalencephalic Leukoencephalopathy With Subcortical Cysts.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=GPRC5B&keywords=GPRC5B
KANK1	Mutations in this gene cause cerebral palsy spastic quadriplegic type 2, a central nervous system development disorder.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=KANK1&keywords=KANK1
DIP2C	This gene encodes a member of the disco-interacting protein homolog 2 family. The protein shares strong similarity with a Drosophila protein which interacts with the transcription factor disco and is expressed in the nervous system.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=DIP2C&keywords=DIP2C
NOTCH1	During cerebellar development, functions as a receptor for neuronal DNER and is involved in the differentiation of Bergmann glia.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=NOTCH1&keywords=NOTCH1
PACRG	This gene encodes a protein that is conserved across metazoans. In vertebrates, this gene is linked in a head-to-head arrangement with the adjacent parkin gene, which is associated with autosomal recessive juvenile Parkinson's disease. These genes are co-regulated in various tissues and they share a bi-directional promoter. Both genes are associated with susceptibility to leprosy. The parkin co-regulated gene protein forms a large molecular complex with chaperones, including heat shock proteins 70 and 90, and chaperonin components. This protein is also a component of Lewy bodies in Parkinson's disease patients, and it suppresses unfolded Pael receptor-induced neuronal cell death. Multiple transcript variants encoding different isoforms have been found for this gene.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=PACRG&keywords=PACRG
SCD5	Involved in neuronal cell proliferation and differentiation through down-regulation of EGFR/AKT/MAPK and Wnt signaling pathways.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=SCD5&keywords=SCD5
PIP4K2A	Diseases associated with PIP4K2A include B-Lymphoblastic Leukemia/Lymphoma With Hyperdiploidy and Bipolar Disorder.	https://www.genecards.org/cgi-bin/carddisp.pl?gene=PIP4K2A&keywords=PIP4K2A

421

References

1. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
2. Vinh, N.X., Epps, J. & Bailey, J. in Proceedings of the 26th annual international conference on machine learning 1073–1080 (2009).
3. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
4. Rosenberg, A. & Hirschberg, J. in Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL) 410–420 (2007).
5. Fowlkes, E.B. & Mallows, C.L. A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* **78**, 553–569 (1983).
6. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
7. Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A. & Theis, F.J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
8. Luecken, M.D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
9. Danese, A. et al. EpiScanpy: integrated single-cell epigenomic analysis. *Nat. Commun.* **12**, 5228 (2021).
10. Liu, H. et al. DNA methylation atlas of the mouse brain at single-cell resolution. *Nature* **598**, 120–128 (2021).
11. Kremer, L.P. et al. Analyzing single-cell bisulfite sequencing data with MethSCAn. *Nat. Methods* **21**, 1616–1623 (2024).
12. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
13. Tang, S. et al. scCASE: accurate and interpretable enhancement for single-cell chromatin accessibility sequencing data. *Nat. Commun.* **15**, 1629 (2024).
14. PE de Souza, C. et al. Epiclomal: probabilistic clustering of sparse single-cell DNA methylation data. *PLoS Comput. Biol.* **16**, e1008270 (2020).
15. Kapourani, C.-A. & Sanguinetti, G. Melissa: Bayesian clustering and imputation of single-cell methylomes. *Genome Biol.* **20**, 61 (2019).
16. Kapourani, C.-A., Argelaguet, R., Sanguinetti, G. & Vallejos, C.A. scMET: Bayesian modeling of DNA methylation heterogeneity at single-cell resolution. *Genome Biol.* **22**, 114 (2021).
17. Angermueller, C., Lee, H.J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18**, 1–13 (2017).
18. Zhao, Q. et al. MambaCpG: an accurate model for single-cell DNA methylation status imputation using mamba. *Brief. Bioinform.* **26** (2025).
19. Tian, W. et al. Single-cell DNA methylation and 3D genome architecture in the human brain. *Science* **382**, eadf5357 (2023).