Genome Biology

METHODOLOGY

Open Access

Benchmarking component choices for unpaired single cell RNA and epigenomic integration

Fnu Naqing[1,2], Qiuyue Yuan[2,3] and Zhana Duren[1*]

*Correspondence:
zduren@iu.edu

[1] Center for Computational Biology and Bioinformatics, Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA
[2] Institute for Human Genetics, Department of Genetics and Biochemistry, Clemson University, Greenwood, SC 29646, USA
[3] Department of Human Cell Biology and Genetics, School of Medicine, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China

Abstract

Background: Single-cell multi-omics sequencing technologies enable profiling various cellular aspects, offering valuable biological insights. Integrating unpaired multi-omics data, which involves profiling different modalities from distinct cells within the same overall population, remains challenging yet crucial for a comprehensive understanding of cell states and molecular dynamics. Although numerous computational methods for integrating such unpaired data exist, a systematic evaluation of the choices at each step is lacking. The recent emergence of technologies simultaneously profiling multiple modalities within the same cell provides paired datasets, allows for the systematic evaluation of unpaired pipelines.

Results: We leverage paired scRNA-seq with scATAC-seq and histone modifications (ChIP-seq) to systematically evaluate methods for unpaired scRNA and peak-based epigenomic (ATAC-seq and ChIP-seq) data integration to establish a robust, general pipeline. We benchmark individual steps, including feature linking, dimension reduction, and clustering, and evaluate their combinatorial effects on pipeline performance by testing various choices at each stage. Our findings reveal that while gene activity scores show limited correlation with gene expression, they effectively preserve cellular neighborhoods for clustering. Dimension reduction emerges as the most critical step, with non-linear methods generally offering better performance and linear methods providing robustness. Optimal transport (OT)-based label transfer consistently outperforms other strategies across various embeddings.

Conclusions: This benchmark of unpaired integration provides valuable insights for developing methods suited for increasingly complex multi-omics study designs.

Keywords: Single cell multi-omics, Benchmarking, Unpaired integration

Background

Advancements in sequencing technologies have enabled the profiling of various molecular characteristics at the single-cell level. For instance, RNA-seq [1] can measure gene expression, while ATAC-seq [2] assesses chromatin accessibility. To gain a more

comprehensive understanding of the biological state of cells, it is essential to integrate data obtained from these distinct technologies [3–6]. Recently, multi-omics technologies, such as the simultaneous profiling of RNA-seq and ATAC-seq data within the same cells, have been developed, enabling studying multiple molecular features from a single cell [7–13]. However, integrating additional layers of cellular information—such as gene expression, chromatin accessibility, DNA methylation, 3D chromatin contacts, protein expression, and histone modifications—within the same cell at large scale remains technically challenging and costly. Consequently, future computational methods must be designed to handle complex scenarios involving different modalities measured from either matched or unmatched cells. Robust integration of such data is essential to fully support the diverse analytical needs of biological research.

To address the integration of unpaired multi-omics data, numerous computational methods have been developed, and the number continues to grow. Benchmarking studies have also been conducted to help guide researchers in selecting appropriate methods [14–17]. However, these benchmarks typically evaluate each method as a complete package, assessing overall performance across various aspects and grouping methods into broad categories. They generally do not systematically examine individual steps—or combinations of steps—within the integration process.

In our study, we view integration as a sequence of three key tasks: feature linking (Fig. 1a, column 1) dimensionality reduction (Fig. 1a, column 2), and clustering (Fig. 1a, column 3) —each of which can be achieved using multiple alternative approaches. For example, Seurat [18] integrates multiple modalities by leveraging Signac to calculate gene activity score (GAS) from ATAC-seq data (feature linking), employs Canonical Correlation Analysis (CCA) for joint dimensionality reduction, and transfers RNA cell labels to ATAC cells using an anchor-based label transfer technique (Fig. 1b-e, blue line); GLUE [19] integrates gene expression and chromatin accessibility through a peak-gene feature graph, learns cell embeddings with variational autoencoders, and performs joint clustering using the Leiden algorithm [20] (Fig. 1b-e, dark grey line). However, many other potential combinations of these steps remain unexplored (Fig. 1b-e, gray dash lines). Combining different approaches at different steps leads to a large number of potential method combinations. To address this, we systematically benchmark all combinations of these core steps to identify effective and robust pipelines for unpaired multimodal data integration.

In this study, we utilized nine RNA + ATAC paired multi-omics data [21–31] as well as three RNA + histone modifications (H3K4me1, H3K4me3, and H3K27ac) paired multi-omics data [32] as ground truth thereby benchmarking integration between transcriptomic and peak-based epigenomic modalities. We evaluated the integration pipeline at each step and benchmarked 180 combinations, including those implemented in existing packages (Fig. 1e) and novel combinations that had not been tested previously.

Results

A motivating example: standard integration may mislabel cells and distort downstream inference

To illustrate why systematic benchmarking is necessary, we did a case study using an E18 mouse brain multiome dataset (10 × Genomics) in which scRNA-seq and scATAC-seq were jointly measured in the same cells [23]. To mimic the common real-world setting

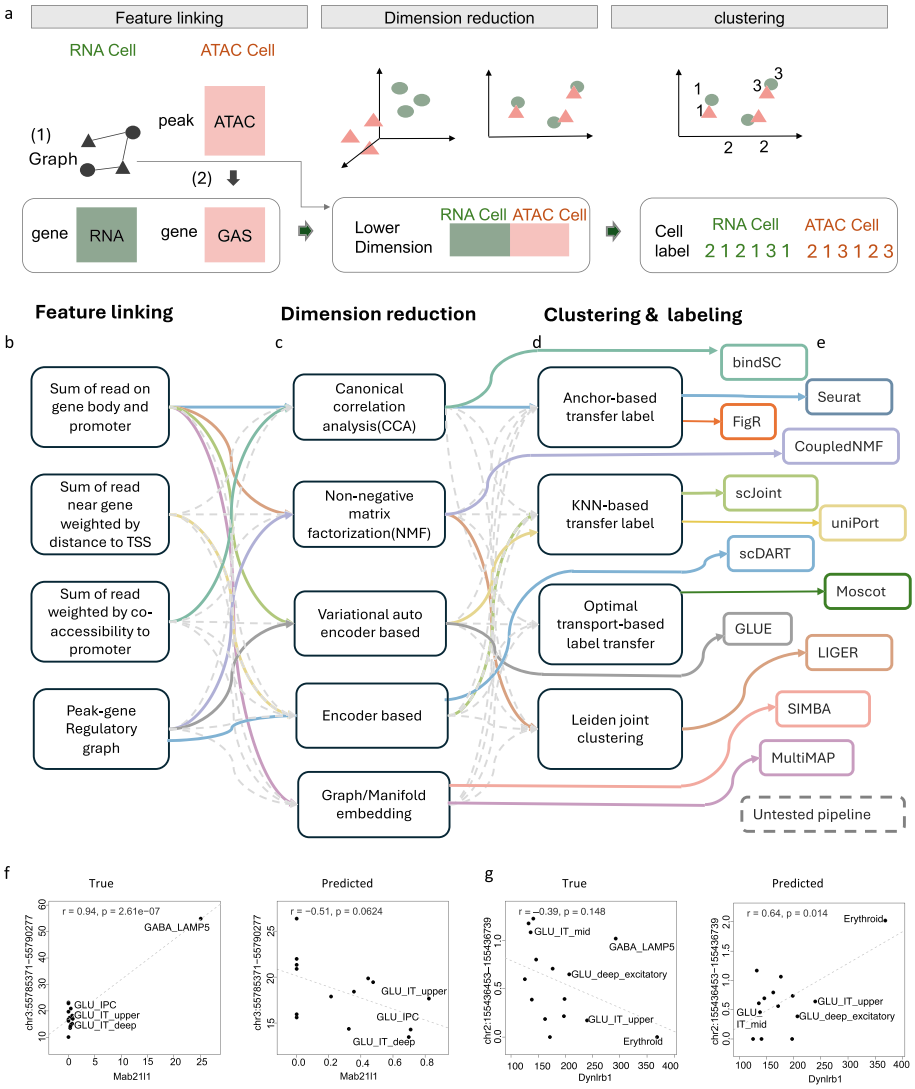

Fig. 1 Schematic demonstration of the study. **a** The three main steps of unpaired transcriptomic and peak-based epigenomic data integration (RNA + ATAC as example). First, features from transcriptomic and peak-based epigenomic data are harmonized by computing gene activity scores (GAS) from peaks or by constructing a peak-gene regulatory graph. Second, gene-level features from both modalities are jointly projected into a shared low-dimensional embedding. Third, cells are clustered in this joint space, two modalities jointly clustered and labeled or one modality cluster labels are transferred to the other. **b** Feature-linking strategies included in the benchmark. **c** Dimensionality-reduction strategies included in the benchmark. **d** Clustering and label-transfer strategies included in the benchmark. **e** Representative integration packages. Lines linking boxes: Colored solid lines denote module combinations of existing integration packages, whereas gray dashed lines denote potential combinations that are not implemented in these tools. Benchmarking was performed both for individual methods within each step and for pipelines formed by different combinations of adjacent steps. **f**, **g** Examples of peak-gene correlations across cell types in pseudo-bulk expression and accessibility. The pseudo-bulk accessibility is the aggregated single cell accessibility based on true (left) or predicted ATAC label (right). In comparison to the true label-based peak-gene correlation, the positive correlation of *Mab2111*-chr3:55,785,371–55,790,277 becomes negative in predicted (**f**), whereas the *Dynlrb1*- chr2:155,436,453–155,436,739 pair becomes a spurious positive correlation (**g**). Pearson correlation coefficients (r) and *p*-values are shown on scatter plot

where modalities are generated from different cells, we treated the two modalities as if they were unpaired and applied a widely used standard Seurat workflow for cross-modality label transfer. We first clustered and annotated RNA cells by marker genes (Additional file 1: Figs. S1-3). Then we embedded RNA and ATAC cells with Canonical Correlation Analysis (CCA) and transferred RNA-derived labels to ATAC cells using anchor-based label transfer, we refer to these transferred labels as Seurat ATAC label. Because the dataset is paired, each ATAC cell has a matched RNA profile with a corresponding "true" cell type label. This pairing allows us to directly quantify label-transfer accuracy by comparing each Seurat ATAC label to the label of its barcode-matched RNA cell. This analysis revealed substantial mislabeling, rare populations were entirely missed (e.g., 14:GABA_LAMP5), many cells were swapped among closely related excitatory subtypes, and others were incorrectly merged into unrelated groups. Overall, 45.07% of cells were misassigned (Additional file 1: Fig. S4a–c).

We further observed that such integration errors can substantially distort downstream biological analyses. As a representative example, we performed cell-type-specific cis-regulatory inference via peak–gene correlations using pseudo-bulk profiles aggregated by cell clusters. Specifically, we compared correlations computed using pseudo-bulk ATAC profiles aggregated by the true ATAC labels (referred as "True" in Fig. 1f, g) with those aggregated by the Seurat-transferred ATAC labels (referred as "Predicted" in Fig. 1f, g), while RNA aggregation was kept fixed. In this setting, mislabeling led to both the positive correlation misled as negative, such as the *Mab21l1*–chr3:55785371–55790277 pair that is specific to the unidentified 14:GABA_LAMP5 population and the introduction of spurious associations, exemplified by the *Dynlrb1*–chr2:155436453–155436739 pair, whose apparent correlation arises solely from cell-type misassignment. At a global scale, Seurat-transferred labels identified 33,397 peak–gene pairs as significant; however, 19,776 of these were not significant under the ground-truth labels, yielding 59.2% of reported significant pairs were false positives (Table S1).

Together, this test highlighted clear opportunities to further improve existing integration workflows and motivated our decision to systematically benchmark how choices at each step of the integration pipeline influence performance. A detailed account of this case study analysis, including method, results and the effect of improved pipeline designs, is provided in Additional file 1 [18, 23, 33–47].

Overview of the benchmarking process

The integration process can be broadly divided into three steps: feature linking, dimension reduction, and clustering/labeling, each with multiple available options. Combining different methods at each step can create many possible integration strategies. Some of these strategies have already been tested and are available as software packages (lines with color linking Fig. 1b-e), while others remain unexplored (grey dash lines between Fig. 1b-e). Among these untested combinations, there may be more effective approaches that improve integration accuracy.

The first step, feature linking, addresses the challenge of different data modalities using different feature spaces. This step connects features across modalities into a unified graph. There are two main approaches: (1) Directly utilizing the unified feature graph [19, 48] or (2) harmonizing features. The harmonization process converts

regulatory element (RE) features in ATAC-seq data into gene-level features by computing gene activity score (GAS) [49, 50] (Fig. 1a, column 1).

The second step, dimension reduction, reduces the complexity of the data and noise while keeping the important biological patterns and embedding different modalities into a common latent space (Fig. 1a, column 2). This step can be done using linear or non-linear methods. Linear methods include CCA [18] and Non-negative Matrix Factorization (NMF) [48, 51]. Non-linear methods use deep neural network approaches, such as encoders [52] and variational autoencoders (VAE) [19, 53] or graph/manifold-based methods [28, 29]. Additionally, methods based on Optimal transport (OT) have also been developed to enhance integration by aligning distributions across modalities [53, 54] (Fig. 1c).

The third step, clustering/labeling, is used to group similar cells within each modality and assign identities to both modalities (Fig. 1a, column 3). There are two main approaches: label transfer and joint clustering. Label transfer first assigns labels to one type of data, usually RNA-seq, and then transfers the labels to the other type of data. Joint clustering, on the other hand, groups cells from both data types simultaneously to create a shared clustering label (Fig. 1d).

In this study, we tested four different methods that calculates GAS: Signac [50], LIGER [51], Cicero [55], and MAESTRO [56]. Each has a different definition of the regulatory relation between genes and peaks (see below for detail). For dimension reduction, we evaluated 10 different methods, including four linear methods: CCA based method Seurat [18] and bindSC [57], NMF based method LIGER and Coupled-NMF [22], two encoder based methods scJoint [52] and scDART [58], two variational auto encoder methods: uniPort [53] and GLUE [19], and two graph/manifold based methods: SIMBA [59] and MultiMAP [60]. For clustering/labeling, we benchmarked 5 approaches including label transfer and joint clustering. Label transfer methods includes the two anchor-based methods, Seurat and FigR [61], an OT-based label transfer method implemented in MOSCOT [62], and a general K nearest neighbors (KNN)-based label transfer method. Leiden clustering was included as a representative joint clustering approach. In total, we tested 180 combinatorial pipelines (Additional file 2: Table S3).

Benchmarking was performed on a total of nine paired scRNA-seq and scATAC-seq datasets and three paired scRNA-seq and histone modification datasets, spanning multiple species and tissues. The scRNA-seq and scATAC-seq data from human peripheral blood mononuclear cells (PBMC) [21], human brain [22], fresh embryonic E18 mouse brain [23], mouse and human kidney cancer [24, 25], human small intestine [26] and human bone marrow mononuclear cells (BMMC) [27], and two SHARE-seq data from mouse brain [28, 29] and mouse skin [30, 31]. The scRNA-seq and histone modification datasets were derived from adult mouse frontal cortex and hippocampus [32] and include H3K4me1, H3K4me3, and H3K27ac profiles paired with gene expression (Table 1).

Notably, since existing multimodal data integration methods do not always follow the three-step process, some methods either combine steps or skip certain steps. Our study provides a comprehensive comparison of the available options for each step and helps to improve the overall integration process. We acknowledge that our evaluation

Table 1 Overview of benchmark datasets used in this study

Tissue/dataset	Species	Genome	Cells	Data modality
PBMC	Human	hg38	9,543	scRNA-seq + scATAC-seq
Human brain	Human	hg38	3,332	scRNA-seq + scATAC-seq
Mouse brain E18	Mouse	mm10	4,881	scRNA-seq + scATAC-seq
Bone marrow (NeurIPS 2021)	Human	hg38	69,249	scRNA-seq + scATAC-seq
Human kidney cancer	Human	hg38	22,772	scRNA-seq + scATAC-seq
Small intestine	Human	hg38	10,640	scRNA-seq + scATAC-seq
Mouse kidney cancer	Mouse	mm10	14,652	scRNA-seq + scATAC-seq
Mouse skin	Mouse	mm10	34,774	scRNA-seq + scATAC-seq
Mouse brain	Mouse	mm10	3,293	scRNA-seq + scATAC-seq
Mouse frontal cortex & hippocampus (H3K4me1)	Mouse	mm10	12,962	scRNA-seq + histone ChIP-seq
Mouse frontal cortex & hippocampus (H3K4me3)	Mouse	mm10	7,465	scRNA-seq + histone ChIP-seq
Mouse frontal cortex & hippocampus (H3K27ac)	Mouse	mm10	11,749	scRNA-seq + histone ChIP-seq

of individual steps does not fully represent the overall performance of a software package, as different tools focus on different aspects of multimodal data integration.

Benchmarking GAS calculation methods

GAS approximates the gene expression levels based on chromatin accessibility. So, a well-calculated GAS should be consistent with gene expression. We designed three matrices to evaluate the GAS: a) Across-cell Pearson correlation coefficient (PCC) measures GAS and gene expression correlation across cells; b) Local Neighborhood Consistency (LNC): Assesses whether cells with similar gene expression have similar GAS values; c) Cell-Type Average Silhouette Width (ASW): Evaluates the ability of GAS to cluster cells.

We benchmarked four widely used methods—Signac, LIGER, Cicero, and MAESTRO. Signac and LIGER determine gene activity by counting sequencing fragments overlapping promoter and gene body regions. Signac directly counts fragments spanning from 2000 base pairs upstream of the transcription start site (TSS) to the end of the gene body, while LIGER separately counts fragments in the promoter region (2000 base pairs upstream of the TSS to the TSS) and the gene body (TSS to the end of the gene body), summing them together. This results in double-counting for fragments spanning the TSS. In contrast, MAESTRO extends this approach by incorporating upstream and downstream peaks, applying a distance-dependent weight that decreases with increasing distance to the TSS. Cicero takes a different approach by constructing co-accessibility networks, scoring peaks based on their connectivity under the assumption that regions correlated with the gene's promoter are likely involved in regulating the gene. Another key difference is the data input requirements: Signac and LIGER use fragment files, whereas MAESTRO and Cicero rely solely on peak counts matrix, avoiding the need for fragment-level data.

Across-cells correlation (PCC)

We calculated the PCC between each gene's GAS and real expression. Figure 2b, column 1 presents the average PCC across all genes across all datasets for different methods. Signac perform best and LIGER close behind. In dataset-specific results in Additional file 1:

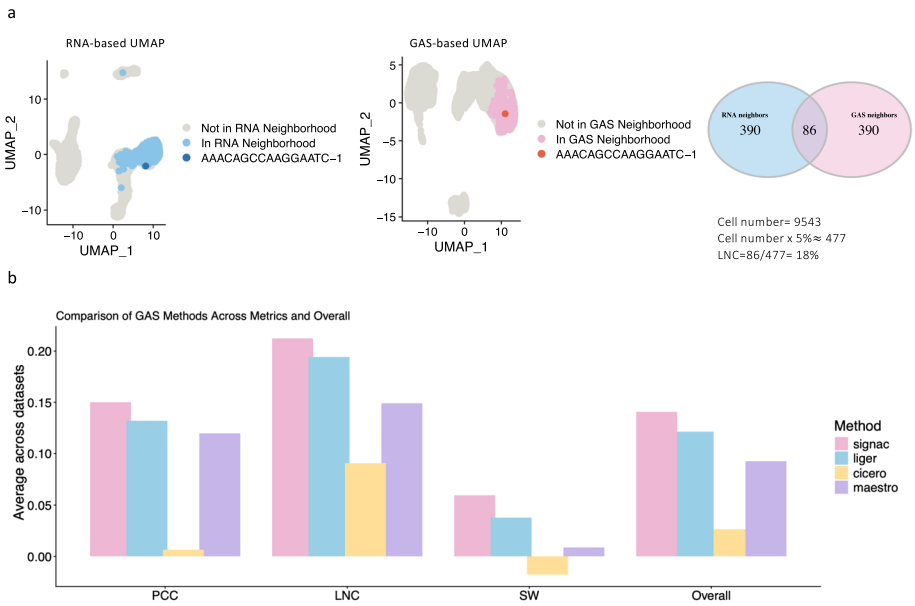

Fig. 2 Evaluation of gene activity score (GAS) methods. **a** Illustration of the Local Neighborhood Consistency (LNC) metric. For each cell, we identify the same number of nearest neighbors in the RNA-based UMAP (left) and the gene activity score (GAS)-based UMAP (middle), then compute LNC as the fraction of neighbors shared between the two spaces (right, Venn diagram; one example cell is highlighted). **b** Comparison of GAS methods (Signac, LIGER, Cicero, MAESTRO) across three metrics-Pearson correlation (PCC) with RNA expression LNC, and cell-type average silhouette width (ASW)-and an overall mean score. Metric values were computed per dataset and then averaged across datasets ($n = 8$ for PCC and LNC; $n = 3$ for ASW, which requires ground-truth cell-type labels)

Fig. S7a, we see Signac and LIGER are almost performing same except for some difference in human kidney cancer and human small intestine data. The overall correlation remains low, less than 0.2 for most data and methods. The best method Signac's mean PCC across datasets is 0.15 (Fig. 2b, column 1). This suggests that GAS does not strongly predict absolute gene expression levels. We also showed a cell type average PCC for 3 datasets that have cell-type labels in Additional file 1: Fig. S8, column 1. Where different cell types differ in PCC, this difference is consistent across methods. Overall, the cell type–specific results align with the global PCC trend.

Local neighborhood consistency (LNC)

To evaluate whether cells with similar gene expression have similar GAS values, we calculated LNC. First, we constructed similarity graphs among cells based on GAS and RNA separately (illustrated in Fig. 2a using UMAPs). We then quantified the overlapping between these two graphs. We define LNC for each cell as the percentage of overlapping neighbors in two graphs (illustrated in Fig. 2a using Venn plot). Figure 2b, column 2 shows the average LNC over all cells and across all datasets for different methods. Consistent with the PCC results, Signac ranks the highest and LIGER closely behind. On dataset-level results, these two methods perform almost identical only slightly different on human small intestine data (Additional file 1: Fig. S7b). The dataset-specific LNC is relatively higher than the PCC. In 5 out of 8 datasets, the best performing method achieves LNC greater than 0.2, and in 2 datasets it exceeds 0.3. This is unexpected, given

that the correlation of gene expression and GAS is low. This suggests that although GAS does not directly predict gene expression well, it captures meaningful population structure-helping explain why GAS-based methods are widely used for this integration. We also showed a cell type average LNC for 3 datasets with known cell-type labels in Additional file 1: Fig. S8, column 2, where some cell types have higher LNC and some cell types shows lower LNC, and the trend is consistent across methods. Overall, the cell type–specific LNC align with the global LNC trend.

Clustering performance (ASW)

To assess clustering performance of GASs, we did a principal component analysis (PCA) for the GASs of the data with known cell-type labels (PBMC, SHARE-seq datasets: mouse brain and mouse skin). Then we calculated the ASW on the first 15 principal components (PCs). Figure 2b, column 3 shows the mean ASW across 3 datasets for different methods. We can see the overall clustering performance directly on GAS's first 15 PCs is weak. However, from Additional file 1: Fig. S7c we can see this result is extremely dataset-specific, the super low average across datasets is because the mouse brain and mouse skin SHARE-seq dataset yield negative silhouette index, these two datasets also generate the worst LNC results, may indicating the cells structure of these two datasets is harder to capture. But for PBMC, the ASW index reaches 0.25. Additional file 1: Fig. S8, column 3 show ASW heatmaps of 3 datasets for different methods, with a column of ASW calculated on RNA-PC15 embedding is included as a reference. Overall, clustering performance using GAS is comparable to clustering based on RNA. In some cell types, GAS-based clustering outperforms RNA-based clustering (Dermal papilla, Macrophage DC in mouse skin; naïve CD8 T cells, myeloid DC in PBMC dataset), while RNA-based clustering is better in others. This indicates that chromatin accessibility and gene expression contribute differently to identifying cell types. Despite the variability, Signac and LIGER remain the best-performing methods in terms of ASW.

By benchmarking GAS methods across multiple datasets, we found that GAS does not provide a strong one-to-one prediction of gene expression levels (low PCC). Still, it does preserve cell neighborhood relationships (better LNC) and supports meaningful cell clustering (better LNC, better ASW). Importantly, our findings indicate that some cell types are more identifiable based on their gene expression patterns, whereas others are more clearly defined by chromatin accessibility. This highlights the complementary nature of RNA and ATAC data, emphasizing the importance of integrating both modalities to obtain a more comprehensive understanding of cellular identity. Signac and LIGER consistently performed the best across all metrics among most of the tested methods, indicates that including original fragment file in process of GAS calculation improve the accuracy. And because Signac and LIGER performed nearly identically, LIGER's double-counting of TSS-spanning peaks does not improve the results (Fig. 2b, column 4). Based on this, we excluded LIGER-derived GAS from the downstream GAS-based dimensional-reduction benchmarks to avoid redundancy.

Benchmarking of dimension reduction methods

Dimension reduction aims to reduce the high dimensionality of different modalities into low-dimensional joint embeddings. A good joint embedding should have two key

characteristics: (1) matching profiles from 2 modalities of the same cell should be similar to each other, and (2) different cells should remain distinguishable. To evaluate these aspects, we use three metrics: a) The percentage of mutual nearest neighbors (MNN%) assesses whether the RNA and ATAC profiles of the same cell are mutually within each other's nearest neighbors (top 1%); b) ASW, based on ground truth labels, measures how well cell types are separated in the embedding; c) Contrastive similarity, evaluates the distinguishability of different cells in the embedding, applicable even in absence of cell-type ground truth.

We benchmarked ten commonly used methods for dimension reduction: seven GAS-based methods including Seurat, LIGER, scJoint, uniPort, bindSC, MultiMAP and SIMBA and three graph-based methods scDART, GLUE and CoupledNMF. These methods can be categorized into 5 commonly used dimension reduction approach: CCA based, NMF based, encoder based, variational autoencoder based and graph manifold based (Fig. 1c). For GAS-based dimension reduction methods, we evaluated all possible combinations with different GAS inputs: Signac, Cicero, and MAESTRO. Additionally, for the Seurat and LIGER, we tested different embedding sizes (15, 30, 45, and 60) to assess the effect of embedding dimension size on performance. In total, this yield 42 distinct embeddings.

The percentage of mutual nearest neighbors (MNN%)

To evaluate if the paired RNA and ATAC profiles from the same cell are close on the joint embedding, we defined metric MNN%, which is the percentage of cells where each cell's RNA and ATAC embeddings are mutually found within top 1% nearest neighbors of the other modality's embedding space (see Methods for detail). Figure 3a, column 1 shows the average MNN% of 9 datasets on the 42 different embeddings with different feature linking methods, top 5 embeddings are labeled (see Additional file 1: Fig. S9 for dataset-level results). The GLUE performs the best, identifies mutual nearest neighbors ranges from 5 to 57% of the cells across all datasets, on average 25%, this means approximately one quarter of cells have RNA and ATAC embeddings that fall within each other's top 1% nearest-neighbor sets. Several GAS-based methods—such as Seurat, uniPort, SIMBA, and bindSC—also performed well, with average MNN% in 15–20% range, particularly when using Signac as the GAS input.

Average silhouette width (ASW)

To evaluate the cell type separation on each embedding, we calculated the ASW for the datasets with ground truth cell labels: PBMC, Mouse brain (SHARE-seq) and Mouse skin data and BMMC data (Fig. 3a, columns 2–3). Top 5 embedding are labeled. Overall, most methods achieve higher ASW in the RNA modality than in ATAC. In RNA, GLUE performed the best, with CoupledNMF, Seurat, scJoint, uniPort combined with Signac are closely behind. In ATAC, although separation was generally weaker across methods. GLUE again produced the highest ATAC ASW, followed by Signac-bindSC, SIMBA and uniPort. LIGER have the most consistent performance for both modalities. Dataset-level results are in Additional file 1: Fig. S10, from which we can see the two datasets mouse brain (SHARE-seq) and mouse skin (SHARE-seq) yield ASW values close to zero for many methods, and even the best methods (e.g. GLUE, uniPort, Seurat)

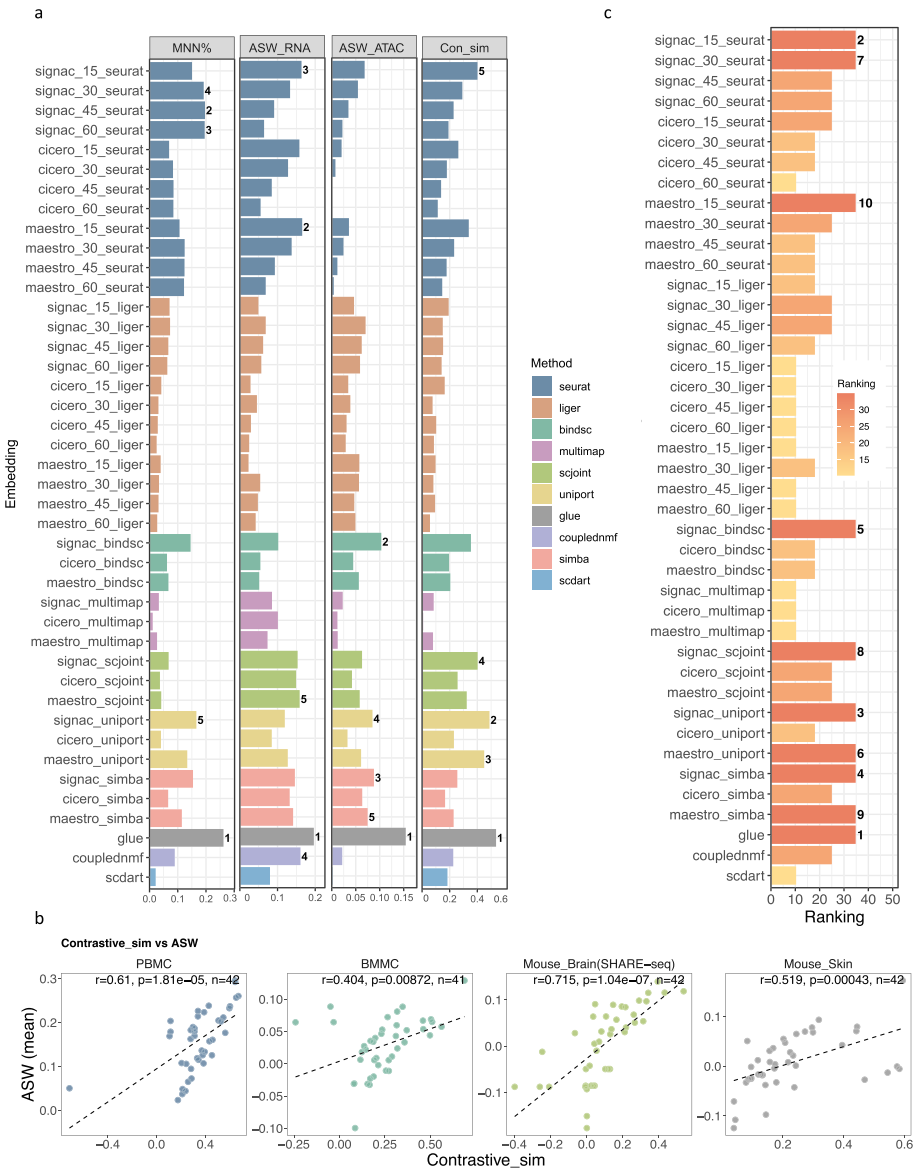

Fig. 3 Dimension reduction evaluation. **a** Metrics results. Column 1: Mutual nearest neighbors' percentage (MNN%) across embeddings. The x-axis shows the 42 distinct embeddings, and the y-axis shows the mean MNN% across all datasets. Colors indicate the dimension-reduction method. Column 2–3 Average silhouette width (ASW) for RNA (column2) and ATAC (column3) embeddings, calculated on datasets with ground-truth cell-type labels. Column 4: Dataset-averaged contrastive similarity. **b** Pearson correlation between contrastive similarity and ASW across the 42 embeddings, shown separately for PBMC, BMMC, Mouse brain (SHARE-seq), and Mouse skin datasets. Each point represents one embedding; ASW is the mean of RNA and ATAC ASW. Pearson correlation coefficients (r), *p*-values(p), and sample sizes (n) are indicated, demonstrating a consistent positive correlation between the two metrics. **c** Overall ranking of embeddings based on the average rank across all evaluation metrics: MNN%, ASW(RNA), ASW(ATAC), and contrastive similarity. Bars indicate the embeddings ranking tier. The better the embedding's overall performance the higher and deeper the bar. Top 10 embedding are labeled

only reach modest separation. This result is consistent with the result in GAS evaluation (PCA-reduced GAS embedding silhouette width metric in Fig. 2b, column 3, Additional file 1: Fig. S7c) and likely reflects intrinsic properties of these two datasets and/or their annotations label—such as more continuous trajectories, closely related or overlapping

cell types, and noisier or sparser accessibility profiles. Nevertheless, better methods still achieve consistently higher ASW than weaker ones on these data, indicating that the metric remains informative even when absolute values are low. Overall, non-linear methods (like uniPort, scJoint, SIMBA) are able to reach higher ASW, but are more sensitive to dataset quality while CCA based method, Seurat and bindSC, especially when paired with Signac GAS, are robust to dataset quality, while GLUE shows both high and robust to datasets performance.

Contrastive similarity

As the ASW requires ground truth cell labels, it can only be applied to certain datasets, limiting its applicability across all datasets. To evaluate the embeddings in terms of cell-separation on all datasets including those have not provided with cell labels, we introduced a new metric: contrastive similarity. This metric compares the similarity between paired RNA and ATAC profiles for the same cell (positive pair) to the average similarity of the same RNA (or ATAC) cell to all other cells within or across modalities (negative pairs). To assess whether this label-free metric effectively evaluate the cell-distinguishing ability similar to ASW, we calculated the correlation between the ASW and contrastive similarity across various embeddings on the datasets with ground truth cell labels. For each dataset, we used all available embeddings (42 per dataset, 41 for BMMC). For each embedding, we defined mean ASW as the average of its RNA and ATAC ASW values and then quantified its association with contrastive similarity across embeddings using PCC (r). Statistical significance was assessed using a two-sided Pearson correlation test (H_0: $r = 0$), and we report the resulting p-value (Fig. 3b). The scatter plots show a consistent positive association across all datasets (PBMC: $r = 0.61$, $p = 1.8 \times 10^{-5}$; BMMC: $r = 0.404$, $p = 8.7 \times 10^{-3}$; Mouse brain: $r = 0.715$, $p = 1.0 \times 10^{-7}$; Mouse skin: $r = 0.519$, $p = 4.3 \times 10^{-4}$). This result support contrastive similarity as a practical, label free surrogate for ASW. Under this metric, GLUE again achieves the best performance, followed other neural network–based methods such as uniPort and scJoint occupy the 2nd-4th places. Signac + CCA based method: Seurat and bindSC also shows good performance (Fig. 3a, column 4, dataset-level results are in Additional file 1: Fig. S11). These results also indicate that GAS quality strongly influence integration performance.

Embedding size effect on Seurat and LIGER

For Seurat and LIGER, we also examined how the dimensionality of the joint embedding affects performance for Seurat and LIGER. For Seurat CCA, increasing the embedding size from 15 to 60 dimensions increases MNN% (for example, with Signac GAS, average MNN% rises from ~ 0.15 at 15 dimensions to ~ 0.20 at 45–60 dimensions) but decreases ASW and contrastive similarity (with Signac, contrastive similarity falls from ~ 0.40 at 15 dimensions to ~ 0.19 at 60 dimensions). This indicates a trade-off: higher-dimensional Seurat embeddings align RNA and ATAC more tightly but blur cell-type boundaries, whereas lower-dimensional embeddings preserve biological separation at the cost of slightly weaker cross-modal alignment (Additional file 1: Fig. S12a). For LIGER NMF, the effect of embedding size is less pronounced. With Signac GAS, MNN% peaks around 30 dimensions but only changes modestly between 15 and 60 dimensions, and contrastive similarity is highest at 15 dimensions and gradually declines at larger sizes (Additional

file 1: Fig. S12b). For both Seurat and LIGER, Signac variants are always the best performing one, indicating the GAS effect is consistent through all embedding size.

To summarize overall embedding performance, we ranked each embedding for every evaluation metric, assigning scores from 42 (best) to 1 (worst), and averaged these ranks to obtain a final overall score. Based on this score, we group methods into four performance tiers and label only the top 10. The tiers correspond to ranks 1–10, 11–20, 21–30, and 31–42. We chose this tier-based visualization because many methods have very similar overall scores, and fine-grained differences among mid-ranked methods (e.g., rank 11 vs. 19) are difficult to interpret and may not be practically meaningful given dataset-to-dataset variability. The tiered view therefore emphasizes the main qualitative conclusions—identifying consistently top-performing methods versus moderate and weaker ones—while keeping the figure readable (Fig. 3c). From this ranking we see GLUE ranked the highest overall, reflecting its consistently strong performance in both cross-modal alignment and cell-separation metrics, in addition to it, among the top10 tier, we see Signac + CCA combinations: Signac_Seurat and Signac_bindSC as well as neural network–based methods such as uniPort (both Signac and MAESTRO GAS variation), scJoint (Signac based) and SIMBA(both Signac and MAESTRO GAS variation).

In conclusion, neural network–based methods (GLUE, uniPort, SIMBA, and scJoint) generally perform well, particularly when paired with Signac-derived GAS, and more effectively separate cell populations. In contrast, Signac + CCA methods exhibit stable performance across datasets and evaluation metrics, resulting in consistently high overall rankings. Together, these results highlight the central role of GAS quality in GAS-based integration methods' performance, as well as the advantage of complex neural network in multimodality integration, with GLUE combining high accuracy and strong robustness across diverse datasets.

Benchmarking of clustering

As the final step of integration, we expect the clustering methods to successfully align the clustering structures of both modalities while recovering the pairing relation of the two different profiles of the same cell. To evaluate the performance of clustering, we use two evaluation metrics. Metric1: Normalized Mutual Information (NMI), which is a widely used measure of clustering consistency. However, NMI is invariant to label permutations, meaning that if a group of cells is clustered together in both modalities but assigned different labels (e.g., cluster 1 in RNA and cluster 2 in ATAC), the NMI score can still be high. To directly quantify the agreement of cluster assignments, we introduce metric 2: Average label consistency. This metric is computed by averaging two percentages: for cells in each RNA cluster, calculate the percentage of corresponding ATAC cells assigned by the same label, and take the mean percentage. In reverse, calculate the same mean percentage for ATAC clusters. The final average label consistency score is obtained by taking the mean of these two percentages (see Methods for detail).

We applied five clustering/labeling strategies to the 42 embeddings generated in Step 2. Because Seurat anchor-based label transfer requires Seurat-compatible objects, in our implementation, it was applied only to the 12 Seurat embeddings. For all embeddings we implemented four additional strategies: Leiden joint clustering, KNN-based label transfer, OT–based label transfer, and FigR anchor-based label transfer. In total, this yields

180 embedding–clustering pipelines ($42 \times 4 + 12$). For label-transferring strategies, we used RNA as the reference modality. RNA labels were obtained using Seurat's standard single-modality scRNA-seq preprocessing and clustering workflow [18] and RNA label were transferred to ATAC cells using each label transfer method. We evaluated all pipelines across nine datasets using NMI and average label consistency. Dataset-level results for all embeddings and clustering/labeling strategies are provided in Additional file 1: Figs. S13–14.

Figure 4a–b summarizes performances averaged across datasets and grouped by the dimension reduction methods that generate the embeddings. For Seurat embeddings, Seurat anchor-based label transfer method performs the best overall, indicating that the Seurat CCA combined with its anchor framework forms a strong pipeline. Across all methods, GLUE stands out as the most consistently high-performing embedding: all clustering approaches applied to GLUE yield high NMI and high average label consistency. More broadly, when the embedding is fixed, NMI tends to show relatively small differences between clustering strategies whereas average label consistency separates strategies more clearly. This highlights that NMI alone can miss meaningful differences in cross-modality aligning performance: pipelines may recover similar within-modality population structure (similar NMI) while still differing substantially in how well RNA and ATAC are aligned in the shared space (captured by average label consistency).

To directly compare joint clustering versus label transfer, we focused on Leiden clustering (joint clustering) and the overall best performing label transfer method: OT-based label transfer. For each dimension reduction method, we selected the best-performing setting identified in Step 2 (e.g., embedding size and GAS). The scatter plot in Fig. 4c compared the average label consistency of the Leiden clustering and the OT-label transfer, across different dimension reduction methods and datasets. Most dataset–method combinations lie above the $y = x$ line, indicating that OT-based label transfer overall improves average label consistency relative to Leiden clustering.

The gain from switching from Leiden clustering to OT-based label transfer indicates that some embeddings retain a residual modality effect (i.e. RNA and ATAC are not fully mixed) but are still structurally alignable: paired cells may receive different cluster assignments under joint Leiden clustering because the two modalities are not perfectly co-embedded, but their overall geometry remains similar enough for OT to recover a reliable cross-modality mapping and improve average label consistency. We summarized this improvement across different dimension reduction methods using a boxplot (Fig. 4d). SIMBA showed the largest improvement, followed by MultiMAP, Coupled-NMF, and scJoint, indicating that these methods preserve within-modality structure while leaving residual modality differences that reduce paired-cell label agreement under joint clustering; OT-based transfer can partially correct this mismatch. In contrast, uniPort, GLUE, bindSC, and Seurat showed little to no change, consistent with the fact that these methods already enforce cross-modal alignment during embedding construction, leaving limited room for OT to further improve (e.g., CCA-based methods optimizes cross-modal correlation; uniPort incorporates OT during embedding; GLUE uses an adversarial objective to promote a shared space).

In the two-metrics summary (Fig. 4e), we plot each pipeline using its mean-NMI and mean-average label consistency across all datasets. Pipelines in the upper-right

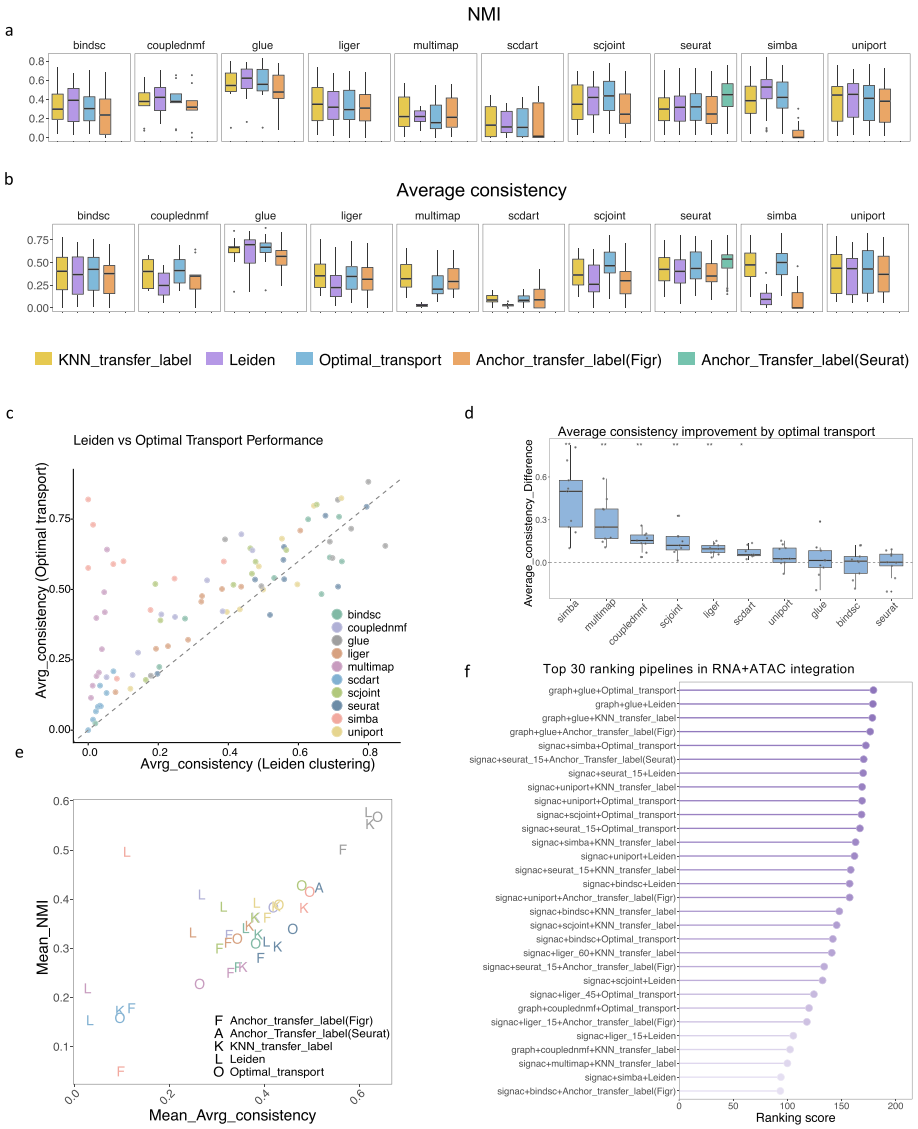

Fig. 4 Clustering and label-transfer evaluation for RNA + ATAC integration. **a**, **b** Boxplots of Normalized Mutual Information (NMI) (**a**) and Average label consistency (**b**) across nine datasets for each combination of dimension-reduction method (columns) and clustering strategy (colors). **c** Comparison of Leiden versus OT-based label transfer on metric average label consistency. Each point is one dataset; colors indicate the embedding method. The dashed line denotes equality ($y = x$). **d** Method-specific change in average label consistency (OT—Leiden) aggregated across datasets. Significance is from paired tests across datasets; stars denote $p < 0.05$ (*) and $p < 0.005$ (**); no star indicates $p \geq 0.05$. **e** Overall pipeline performance summarized by the dataset-mean value of the average label consistency (x-axis) and dataset-mean NMI (y-axis). Each point represents one pipeline; colors indicate embedding method and point shapes indicate clustering strategy. **f** Top 30 pipelines ranked by the mean of metric-specific ranks (NMI and average label consistency). Higher ranking scores indicate better overall performance; the best pipeline approaches the maximum score (~ 180)

therefore achieve both strong clustering agreement (high NMI) and strong cross-modality agreement (high average label consistency). The scatter plot shows a clear pattern in which pipelines cluster primarily by dimensionality reduction method. For most dimensionality reduction methods, changing the clustering approach has only

a modest effect on either metric, whereas the major differences in mean NMI and mean average label consistency are driven by the choice of dimensionality reduction rather than the clustering strategy. Because NMI is invariant to label permutations whereas average label consistency is not, pipelines in the upper-left (high NMI but low label consistency) may appear problematic: they preserve similar clustering structure within each modality yet assign different cluster IDs to paired cells across modalities. Interestingly, across our datasets this mismatch is largely correctable by OT-based label transfer. The four methods most clearly exhibiting the high-NMI/low-consistency pattern—SIMBA, MultiMAP, CoupledNMF, and scJoint—are also the top methods that gain the most in average label consistency after OT-based label transfer (as discussed above, Fig. 4d). This indicates that, although their embeddings do not fully align cluster identities across modalities, they still capture structurally compatible organization that OT can exploit, so these embeddings remain valuable as inputs to downstream label transfer. We next summarized these results with an overall ranking and show the top 30 pipelines (Fig. 4f). We ranked pipelines by averaging their rank based on dataset-mean-NMI and dataset-mean-average label consistency. Consistent with the scatter plot, the ranking in Fig. 4f is dominated by GLUE-based pipelines, with Seurat anchor-based label transfer on Seurat embeddings also rank among the top performers; uniPort, scJoint and SIMBA combined with Signac + OT appear in the top 10. A complete ranking table for three-step evaluation for all pipelines, including feature linking, dimension reduction, clustering choices, and their performance in individual evaluation metrics, is provided in Additional file 2: Table S3.

Conclusion: Dimension reduction is the main driver of integration performance. When embeddings are strong, downstream clustering choices have a smaller effect and most pipelines perform similarly well. Overall, GLUE combined with OT-label transfer achieves the best performance for RNA + ATAC integration, while Seurat CCA embeddings with Seurat anchor-based label transfer provide a robust and consistently high-performing alternative. For embeddings with weaker cross-modal alignment, OT-label transfer can substantially improve cross-modality agreement. In contrast, Leiden joint clustering often captures within-modality structure (high NMI) but tends to yield less consistent labels across modalities (lower average label consistency) for some embeddings; in many cases, this mismatch can be largely corrected by applying OT-based label transfer downstream.

Evaluation of pipelines in RNA + histone modifications data

To evaluate whether the best-performing pipeline generalizes to other multi-omics settings, we selected the dimension reduction methods that ranked in the top 10 in the RNA–ATAC benchmark (Fig. 3c) and applied them to three additional paired datasets in which RNA was profiled alongside a histone modification (H3K4me1, H3K4me3, or H3K27ac) [32]. We then applied all clustering methods to the resulting embeddings to identify the optimal integration pipeline for RNA–histone data. Consistent with our evaluation of RNA and ATAC data, we treated paired RNA and histone modification data as unpaired to assess the performance of different integration approaches. Given the absence of fragment-level information for the histone modification data, the second-best performing GAS: MAESTRO is used for all GAS-based method. For Seurat, we

used embedding size 15, the size that had optimal performance in our RNA and ATAC-seq data evaluations.

Dimension reduction

Figure 5a, b shows that on the H3K4me1 and H3K27ac datasets, GLUE have the highest MNN% and scJoint have the highest contrastive similarity. These two datasets show highly similar patterns across methods. In contrast, all methods do not perform well on the H3K4me3 dataset, with MNN% values approximately half of those observed for other histone modifications. This may be due to the complex role of H3K4me3 in transcription regulation, which makes direct prediction from histone modification to RNA expression particularly difficult. A similar pattern is observed in contrastive similarity. Specifically, methods perform robust on H3K4me1 and H3K27ac datasets, while exhibiting lower contrastive similarity on H3K4me3, only uniPort shows much higher contrastive similarity than others, probably the OT design in uniPort balanced the in modality and cross modality similarity while jointly reducing the dimension. Neural network–based methods again yield the highest contrastive similarity, while the linear methods Seurat and bindSC show relatively stable performance across datasets.

Clustering

Consistent with observations from RNA + ATAC datasets, clustering performance is strongly influenced by the quality of the dimension reduction (Fig. 5c-d). For the H3K4me3 dataset, clustering results are poor, while the H3K4me1 and H3K27ac datasets show more consistent and robust performance. Among dimension reduction methods, Seurat embeddings paired with anchor-based label transfer outperform all clustering strategies. OT performs second-best for Seurat embedding and achieves the best results for embeddings generated by other methods. KNN-based label transfer and Leiden clustering also perform well, with minimal performance differences between them.

Notably, scJoint embeddings produce high contrastive similarity on H3K4me1 and H3K27ac (Fig. 5b). However, for H3K4me3, scJoint yields extremely low MNN% and contrastive similarity (Fig. 5a, b), leading to poor clustering results when using Leiden or KNN label transfer. Interestingly, applying OT to the same embedding substantially improves clustering performance (Fig. 5c, d),

To explain this discrepancy, we hypothesized that although the scJoint embedding of H3K4me3 may be globally reasonable, it contains cells that are disproportionately close to many cross-modality cells, leading to failures in neighbor-based matching methods. To test this, we assessed the inequality of cross-modality nearest neighbor contributions using the Gini coefficient, which quantify how concentrated the distribution of cross-modality nearest neighbor is (Fig. 5e). Specifically, we computed how often each RNA or histone cell was selected as a nearest neighbor by cells from the other modality.

For H3K4me1 and H3K27ac, the Gini coefficients for RNA were 0.53 and 0.53, and for histone were 0.49 and 0.48, respectively, indicating moderate concentration in cross-modality neighbor assignment. In contrast, the H3K4me3 embedding showed extreme concentration, with Gini coefficients of 0.87 (RNA) and 0.98 (histone), suggesting that only a few cells dominate neighbor assignments. This severe imbalance likely underlies the poor MNN% and contrastive similarity observed in H3K4me3. The performance

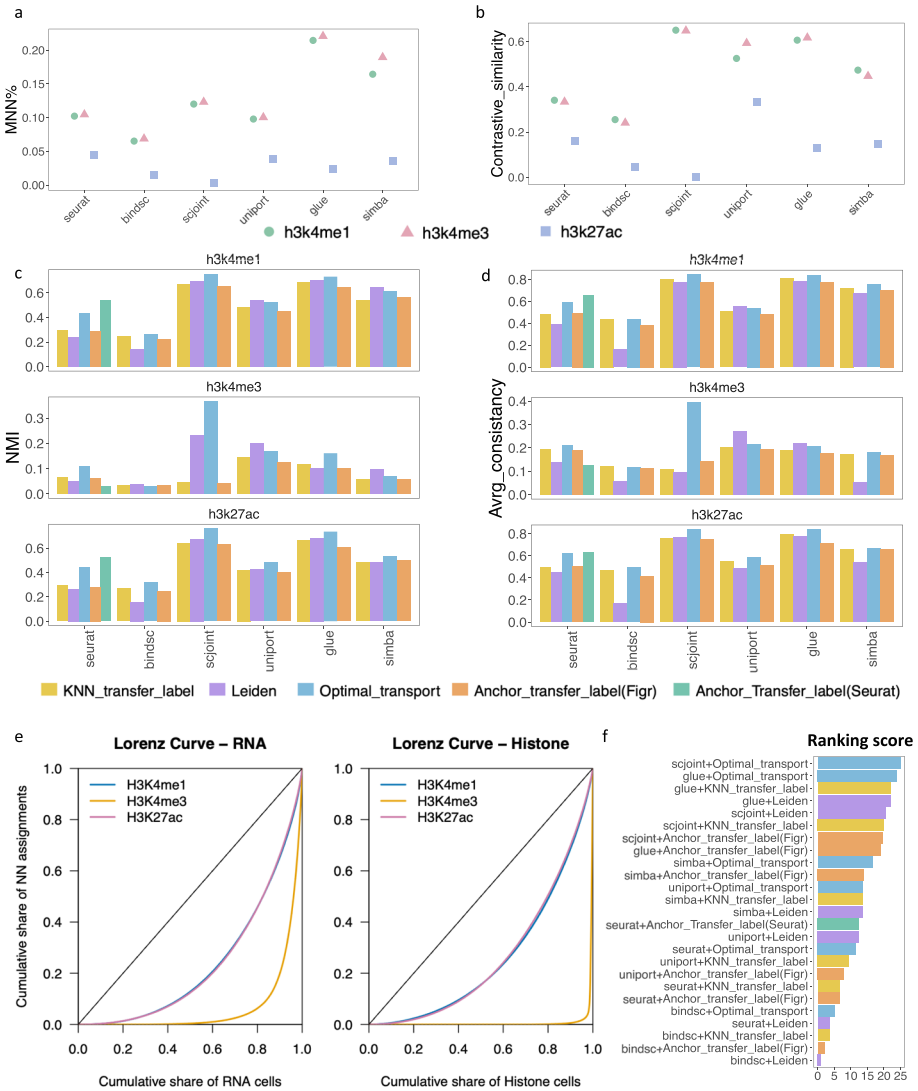

Fig. 5 RNA + Histone modification data integration analysis. **a**, **b** Dimension reduction method evaluation: Mutual nearest neighbor percentage (MNN%) (**a**) and Contrastive similarity (**b**) of embeddings from different dimension-reduction methods. Each point represents one dataset, with shapes indicating different datasets. Dimension reduction methods appear in the top 10 in RNA + ATAC benchmarking was benchmarked here. **c**, **d** Clustering method evaluation. Normalized mutual information (NMI) (**c**) and average label consistency (**d**). The x-axis shows embeddings from different dimension-reduction methods, bars with different colors correspond to different clustering strategies. **e** Lorenz curves illustrating inequality in cross-modality nearest-neighbor contributions for scJoint embeddings on histone modification datasets. Curves show the cumulative proportion of cells (x-axis) versus the cumulative proportion of nearest-neighbor assignments (y-axis). Left: RNA cells; right: Histone cells. Curves closer to the diagonal indicate more even neighbor contributions. **f** Overall ranking of integration pipelines combining dimension-reduction and clustering methods for RNA + Histone modification data, based on the mean value of 2 clustering metrics' dataset average performance metrics

boost from OT-based label transfer likely stems from its ability to consider global structure, rather than relying solely on local neighbors.

Similar to our RNA + ATAC benchmark, we enumerated distinct combinations of dimension reduction and clustering pipelines, computed dataset-averaged clustering metrics, and ranked pipelines by averaging their ranks based on dataset-mean NMI

and dataset-mean average label consistency (Fig. 5f). In the RNA + histone modification integration task, neural network–based methods—including scJoint, uniPort, and SIMBA—ranked more prominently among the top-performing pipelines compared with RNA + ATAC integration. We hypothesize that this shift reflects the increased complexity of histone modification signals, particularly H3K4me3, whose relationship with gene expression is mediated by multiple regulatory mechanisms and contextual factors. We identify scJoint combined with OT-based label transfer as the best-performing pipeline for RNA + Histone modification data integration, followed by GLUE with OT-based label transfer and other GLUE-based strategies. Notably, GLUE combined with OT-based label transfer remains among the top pipelines, highlighting the robustness of this approach across diverse multi-omics integration tasks.

Robustness of integration methods to unbalanced modality sizes

In real single-cell multi-omics experiments, the number of cells profiled in each modality is often unbalanced. Although paired datasets enable evaluation based on true RNA + ATAC correspondences, robustness to modality imbalance is an important practical consideration. To assess this, we performed one-sided subsampling experiments that preserve ground-truth RNA + ATAC pairings while introducing controlled imbalance.

Using the BMMC dataset, one of the largest paired datasets in our benchmark, we fixed the RNA modality and subsampled the ATAC modality to 10%–90% of its original size. At each subsampling level, we applied all dimensional-reduction methods and evaluated the resulting embeddings using the metrics: MNN% and contrastive similarity.

Robustness of dimensional reduction under imbalance

Across nine subsampling levels, most dimensional-reduction methods showed strong robustness to unbalanced modality sizes, with performance curves closely resembling those observed in the balanced setting (Additional file 1: Fig. S15a). Methods that performed well on the full dataset—such as GLUE, Seurat, scJoint, and uniPort—remained stable even when the ATAC modality was substantially reduced. Methods with lower performance exhibited both reduced scores and visibly larger fluctuations across perturbation levels. For scJoint and CoupledNMF, we observed a decline in performance as the retained ATAC fraction increased; this unexpected trend likely reflects increased computational difficulty at larger input sizes (e.g., longer runtimes or incomplete convergence), which can lead to suboptimal solutions in practice. Notably, these results also demonstrate that pairing-based evaluation metrics (MNN% and contrastive similarity) remain reliable even when one modality is much smaller.

Downstream robustness of clustering and label transfer

To examine whether robustness at the embedding level propagates to downstream analyses, we evaluated one joint clustering method (Leiden) and one label-transfer method (KNN-based transfer transfer) across the same subsampling settings. We evaluate the clustering results by taking the mean value of the 2-clustering metrics: NMI and average label consistency (Additional file 1: Fig. S15b). For Leiden clustering, the performance closely tracked embedding quality. Strong embeddings, like GLUE, Seurat also produced high performance when cluster with Leiden; scJoint and CoupledNMF's reproduced the

decline in performance as the retained ATAC fraction increased. Notably, for SIMBA embedding, Leiden clustering performance improve when the retaining fraction increases, suggesting that under unbalanced modality size it becomes harder for SIMBA to fully mix the two modalities, and residual modality effects become more pronounced. In contrast, KNN-based label transferring shows stable performance across subsampling fractions, including SIMBA.

Overall, these analyses show that most of the dimensional-reduction methods benchmarked in our study are robust to unbalanced modality sizes, and that our pairing-based metrics provide consistent assessments even when one modality is much smaller. A subset of methods changes as ATAC retention increases, likely reflecting increased computational difficulty at larger data sizes. On the other hand, more balanced RNA and ATAC cell numbers tend to improve global cross-modality mixing, leading to better clustering agreement for some embeddings.

Conclusion and discussions

We benchmarked 180 distinct combinations of unpaired integration pipelines utilizing paired RNA and ATAC-seq data, as well as RNA and histone modification datasets. We summarize our primary conclusions in the following points:

1. While GAS exhibits a weak correlation with gene expression levels, limiting their direct predictive power, they can preserve reasonable cell neighborhood relationships and support meaningful cell clustering, suggesting their utility lies in capturing cellular context rather than precise expression quantification. Using original fragment file improve the GAS accuracy.
2. Dimension reduction is the most critical step influencing the success of multi-omics integration, making it the primary target for future method development. GAS quality highly affects the GAS-based methods' performance. Neural network-based methods better integration performance, especially on cell type separation, while CCA-based methods demonstrate greater robustness to variations in data quality. The optimal clustering method depends on the type of embedding. A high-quality embedding can enable most clustering approaches to perform well.
3. In terms of clustering strategy, optimal transport-based clustering consistently demonstrates superior performance across various embeddings when compared to joint clustering. For embeddings with weaker cross-modal alignment, OT-label transfer can substantially improve cross-modality agreement.
4. Generally, a pipeline employing graph-based feature linking, GLUE for joint dimension reduction, and OT for label transfer represents a robust strategy that performs well across all tested multi-omic datasets.
5. For gene expression and histone modification data integration, employing GAS for feature linking, scJoint for dimension reduction, and OT for label transfer represents the optimal approach.

We acknowledge several recent benchmarking efforts for single-cell multi-omics integration, including the scRNA–scATAC benchmark by Xiao et al. [16]. The SCMMIB project on broad multimodal integration settings [63], the Nature Methods study

assessing prediction and integration across technologies by hu et al. [64], and the scIB benchmark for single-omics batch integration that helped establish many commonly used evaluation principles [14]. Our findings are broadly consistent with these studies, particularly the strong performance of deep generative and graph-based approaches and the need to balance modality/batch mixing with preservation of biological structure. At the same time, our work adds three distinct contributions: (i) we focus on unpaired "diagonal" integration of scRNA with peak-based epigenomic modalities (rather than primarily paired integration or batch correction), (ii) we adopt a modular three-step design (feature linking → dimensionality reduction → clustering/label transfer), enabling systematic benchmarking of component combinations rather than only end-to-end tools, and (iii) we introduce label neighborhood consistency and contrastive similarity as paired-ground-truth–aware metrics that directly leverage barcode pairing to quantify cross-modal alignment.

Next, we discuss the limitations in this study. First, the computational resources and time constraints inherent in a comprehensive benchmarking study meant that we did not evaluate every existing method for unpaired multi-omics integration. Second, the ground truth cell label that provided along with subset of our benchmark data PBMC, BMMC, Mouse-brain (SHARE-seq) and Mouse skin dataset, may itself be generated from computational tool (Seurat or scanpy) thus could introduce method-dependent bias and may affect label-based metrics: cell-type ASW (cell type average silhouette width). Finally, our three-step framework, while providing a structured approach for comparison, may not fully encompass the architecture and functionality of all available integration methods. Certain algorithms may employ unique strategies that do not readily align with our defined stages of feature linking, dimension reduction, and clustering/labelling.

We agree that technological development will enable more multimodalities profiling on the same cell and totally unpaired data will not be the most common experimental design in the future. Since measuring all modalities on the same cell is not feasible in near future, it is still needed new methods for integration of a mixed scenarios of matched and unmatched modalities. Since the unpaired integration is the most challenging integration, we believe the benchmarking results for unpaired data from this study will provide insights for future method development for more complicated cases.

Methods and materials

Datasets

To benchmark the integration process across three major stages, we used 9 sets of publicly available paired single-cell RNA and ATAC multiome datasets. Additionally, to assess the robustness of the optimal integration pipeline, we included three datasets where RNA was simultaneously measured alongside histone modification signals.

10 × Genomics scRNA and ATAC multiome datasets

The scRNA-seq and scATAC-seq data from human peripheral blood mononuclear cells (PBMC) [21], human brain [22], fresh embryonic E18 mouse brain [23], and mouse kidney cancer [24], human kidney cancer [25], human small intestine [26] were obtained from the 10 × Genomics Multiome ATAC + Gene Expression dataset portal. (https://

support.10xgenomics.com/single-cell-multiome-atac-gex/datasets). Among these datasets, PBMC10K includes a surrogate ground truth for 9,543 cells.

BMMC scRNA and ATAC multiome dataset from NeurIPS 2021 competition

The human bone marrow mononuclear cell (BMMC) dataset was downloaded as a preprocessed.h5ad file from Gene Expression Omnibus repository (GEO), accession number: GSE194122 [27]. This dataset is the training data of the NeurIPS 2021 Multimodal Single-Cell Data Integration Challenge [65]. The dataset contains paired RNA and ATAC profiles, along with cell-type labels and precomputed Signac gene activity score (GAS).

SHARE-seq datasets

Mouse brain and mouse skin SHARE-seq datasets were downloaded from GEO database. Both datasets provide paired single-cell RNA-seq and chromatin accessibility profiles with ground truth cell-type annotations. Mouse brain accession number RNA: GSM4156610 [28], ATAC: GSM4156599 [29]; Mouse skin accession number RNA: GSM4156608 [30], ATAC: GSM4156597 [31]. These data were originally generated and described by Ma et al. [11].

RNA and histone modification multiome data

We download RNA and histone modification (ChIP-seq) data obtained from adult mouse frontal cortex and hippocampus using the Droplet Paired-Tag microfluidic cell barcoding-based method. It comprises 12,962 cells with H3K4me1 and RNA, 7,465 cells with H3K4me3 and RNA, and 11,749 cells with H3K27ac and RNA. The histone modification data is available at GEO under accession number GSE152020 [32]. As the fragment files were not provided for the histone modification data, for analyses requiring GAS, we used tool MAESTRO convert DNA count matrix to GAS matrix.

The GAS methods benchmarking

Signac

GAS calculation using Signac (version 1.10.0) was performed within the Seurat single cell RNA and ATAC integration framework, following the tutorial: https://satijalab.org/seurat/articles/seurat5_atacseq_integration_vignette. The peak count matrix was made incorporated with the original fragment file and corresponding genomic annotation. Genome annotations were loaded using Ensembl annotation databases. For human datasets, we used EnsDb.Hsapiens.v86, and for mouse datasets, we used EnsDb.Mmusculus.v79. GAS was calculated by function: GeneActivity. Only peaks detected in at least 10 cells were retained. The promoter region is defined was the default: ± 2000 base pair upstream from the transcription start site (TSS).

LIGER

GAS calculation using LIGER (v2.1.0) followed the official tutorial https://welch-lab.github.io/liger/articles/Integrating_scRNA_and_scATAC_data.html. The original fragment files were sorted and split into promoter and gene-body fragments using a promoter definition of ± 2000 base pair from the TSS. GENCODE annotations (GRCh38 for human; GRCm38 for mouse) were converted into BED files for promoter and gene-body

regions. Promoter and gene-body fragments were overlapped with the corresponding BED regions to generate promoter and gene-body fragment count matrixes, which were then combined by LIGER to derive GAS.

Cicero

GAS calculation using Cicero (version 1.3.9) was performed following the tutorial: https://cole-trapnell-lab.github.io/cicero-release/docs_m3/#cicero-gene-activity-scores. The peak count matrix was binarized and converted into a cell_data_set object, and peaks were assigned to genes based on the GENCODE annotations (GRCh38 for human; GRCm38 for mouse). Chromatin co-accessibility links were inferred using run_cicero, and the gene activity matrix was computed with function build_gene_activity_matrix, followed by standard normalization.

MAESTRO

GAS calculation using MAESTRO (version 1.5.1) was performed using a Python-based implementation within R. The multi-omics feature matrix was read in, the peak count matrix was input to the ATACCalculateGenescore function from the MAESTRO package. The organism reference was set to "GRCh38" for human data and "GRCm38" for mouse data. The R function used for GAS calculation can be found on the GitHub page. (https://github.com/liulab-dfci/MAESTRO/tree/master/R).

Graph-based feature linking and graph construction

Non-GAS-based dimension reduction methods in our benchmark (GLUE, scDART, and CoupledNMF) rely on graph-based feature linking to encode prior regulatory relationships between RNA genes and ATAC peaks. In all cases, RNA genes and ATAC peaks are treated as nodes in a bipartite graph, with edges representing genomic proximity or overlap. Gene genomic coordinates were obtained from annotation GTF files corresponding to each dataset (GENCODE GRCh38 for human datasets and GENCODE GRCm38 for mouse datasets). Gene regions were defined as the union of the gene body and a 2000 base pair upstream promoter window around the TSS. Peak genomic coordinates were parsed from ATAC feature names of the form chr:start–end. Gene–peak edges were defined based on genomic overlap or proximity, as described below for each method.

For GLUE, gene annotations were added to the RNA object using scglue.data.get_gene_annotation, and genes without valid chromosome information were removed. Peak coordinates were parsed into chromosome, start, and end positions. The guidance graph was constructed using scglue.genomics.rna_anchored_guidance_graph with default settings, which links peaks to nearby genes overlapping the gene body $\pm$2000 base pair region and assigns distance-based edge weights. The resulting weighted bipartite graph was validated using scglue.graph.check_graph and used as the guidance graph input for GLUE.

For CoupledNMF, gene–peak coupling was constructed following the original MATLAB implementation, log2-transformed counts were filtered to retain highly detected genes (top ~5,000 by detection rate) and informative peaks, defined by excluding the top 5% most accessible peaks and retaining peaks with high background-normalized

accessibility. Gene and peak genomic coordinates were obtained using the read_ATAC_GEX function. Each gene was linked to nearby peaks within 1 million base pair on the same chromosome, with edge weights defined as an exponential decay of genomic distance, resulting in a sparse weighted gene–peak coupling matrix used to couple RNA and ATAC factorizations.

For scDART, a region-to-gene mapping file (region2gene.txt) was explicitly generated. Protein-coding gene regions (gene body + 2000 base pair upstream from TSS) were derived from the appropriate GTF file (GRCh38 or GRCm38) and intersected with dataset-specific ATAC peak regions using BEDTools. Overlapping gene–peak pairs were recorded and converted by scDART into a bipartite adjacency matrix, which was used as a graph constraint to encourage linked genes and peaks to share similar latent representations.

Dimension reduction methods

Seurat

Dimension reduction using Seurat (v4.3.0) was performed following the official tutorial: https://satijalab.org/seurat/articles/seurat5_atacseq_integration_vignette. Seurat objects were created for RNA and ATAC modalities. Standard preprocessing was applied to each object. The GAS matrix was added to the ATAC object, followed by normalization and scaling. Canonical Correlation Analysis (CCA) dimension reduction was performed using function: FindTransferAnchors, which internally conducts anchor finding and computes embeddings. To assess the effect of embedding size, the dims parameter was tested at values 15, 30, 45, and 60.

bindSC

Dimension reduction using bindSC (v1.0.0) was performed following the tutorial on: https://htmlpreview.github.io/?https://github.com/KChenlab/bindSC/blob/master/vignettes/mouse_retina/retina.html. RNA and ATAC were loaded as Seurat objects. RNA counts were log normalized. GAS matrix was added as an ACTIVITY assay into ATAC. We selected 5,000 highly variable genes (HVGs) from both the RNA and GAS, and restricted RNA and GAS matrices to their overlapping HVGs, yielding the bindSC inputs X (RNA), and Z0 (GAS), followed by a SVD joint dimension reduction for X and Z0 to generate x and z0. ATAC peak counts were TF–IDF normalized and reduced with latent semantic indexing (LSI) to 50 dimensions to form matrix Y. Modality-specific Leiden clustering on RNA's embedding from principal component analysis (PCA) and ATAC's LSI embedding provided cluster priors. Integration used bindSC's BiCCA model with default settings ($\alpha = 0.5$, $\lambda = 0.5$, $K = 15$, 50 iterations). The resulting U (RNA) and R (ATAC) embeddings were used for downstream benchmarking analyses.

CoupledNMF

CoupledNMF was run using the authors' original MATLAB implementation (downloaded from the Wong lab website: https://web.stanford.edu/group/wonglab/zduren/CoupledNMF/index.html). For each dataset, the input directory contained the 10x-formatted matrixes (matrix.mtx, features.tsv, barcodes.tsv) were read using the provided read_ATAC_GEX function. All parameters were kept at their default values (including $\lambda = 1$, $\mu = 0.001$, and $K = 100$).

LIGER

Dimension reduction using LIGER (v2.1.0) was performed following the official tutorial: https://welch-lab.github.io/liger/articles/Integrating_scRNA_and_scATAC_data.html. RNA counts matrix and GAS matrix were read in R, combined to create a LIGER object. Preprocessing followed LIGER's standard preprocessing: normalize, selectGenes and scaleNotCenter, all used default settings. The dimension reduction was then performed using the optimizeALS function, with the number of latent dimensions (k) set to 15, 30, 45, or 60 to evaluate the effect of embedding size. The object was then quantile-normalized using quantile_norm, and the resulting H_norm embeddings were used as the embedding.

scJoint

To implement dimension reduction using scJoint, we followed the tutorial: https://github.com/sydneybiox/scJoint/blob/main/tutorial/Analysis%20of%2010xGenomics%20data%20using%20scJoint.ipynb. RNA counts and GAS matrixes were read into R as Seurat objects, and converted to SingleCellExperiment objects, filtered RNA and GAS to retain only shared genes, and saved as.h5 files via function: write_h5_scJoint. RNA label is generated by Seurat RNA single modality analysis saved as csv file. These files were used as input for training the model. Training followed default parameters from the tutorial: batch size = 256, learning rate = 0.01 for both stage 1 and 3, and embedding size = 64. For E18 mouse brain data, stage 3 learning rate was reduced to 0.005 to ensure convergence.

scDART

We performed dimension reduction using scDART following the official scDART package demonstration: https://github.com/PeterZZQ/scDART. As preparation, we constructed region-to-gene graph as described in the Feature Linking section. RNA and ATAC were read into python. RNA counts were normalized, log-transformed, and 1000 HVGs were selected using Scanpy. The region-to-gene mapping was filtered to retain only links involving HVGs, and RNA and ATAC count matrixes were subset to genes and peaks present in the filtered graph. scDART was trained using RNA and ATAC counts, and a binary region-to-gene adjacency matrix. Hyperparameters were chosen following recommendations from the scDART paper: latent dimensionality was set to 4 or 8, the graph regularization weight (reg_g) was fixed at 1, the MMD regularization weight (reg_mmd) was tested at 1 or 10, the number of training epochs was set to 500, and the time-scale parameter ts was tested at {30, 50, 70}. For each dataset, all the parameter combinations were run and the one combination yielding the lowest final MMD loss was selected to be include in benchmarking. Final RNA and ATAC embeddings were extracted from the trained model.

uniPort (v1.2.2)

We run uniPort following official tutorial: https://uniport.readthedocs.io/en/latest/. GAS and RNA count matrixes were loaded as AnnData objects. Each modality was filtered separately by requiring at least three detected features per cell and retaining only features detected in at least 200 cells. Each modality was normalized for sequencing

depth (library-size normalization) and applied a log transformation, followed by selection of the top 2,000 HVGs within each modality. RNA and GAS were integrated using uniPort by supplying the two modality-specific datasets together with a combined (concatenated) matrix to guide cross-modal alignment, with the shared-structure regularization strength set to $\lambda = 1.0$.

GLUE (v0.3.2)

We followed the tutorial: https://scglue.readthedocs.io/en/latest/tutorials.html. RNA and ATAC matrixes were filtered to remove non-expressed features, and the top 2,000 HVGs were selected. RNA and ATAC were normalized and scaled separately; RNA underwent PCA to 100 components, and ATAC was reduced using LSI to 100 dimensions. Gene annotations were retrieved using *scglue.data.get_gene_annotation* with corresponding GTF file to the dataset. Peak positions were parsed from peak names. A guidance graph was constructed using scglue.genomics.rna_anchored_guidance_graph with default settings (detail described in the Feature Linking section).

MultiMAP

Integration using Multimap was performed following the official GitHub tutorial: https://github.com/Teichlab/MultiMAP. The scRNA-seq, scATAC-seq and GAS data were loaded as AnnData objects. Prior to integration, PCA was applied to RNA data, and TF–IDF normalization followed by latent semantic indexing (LSI) was applied to ATAC data following the MultiMAP workflow. LSI coordinates derived from ATAC modality were transferred to the corresponding GAS object to ensure consistent peak representations. Integration was performed using MultiMAP.Integration, providing RNA PCA and ATAC LSI embeddings as input modalities. Modality strength parameters were set to 0.8 for RNA and 0.2 for ATAC, consistent with the MultiMAP paper for RNA–ATAC multiome data. All other parameters were set to default values.

SIMBA (v 1.2)

Integration was performed following the official SIMBA tutorial: https://simba-bio.readthedocs.io/en/latest/multiome_10xpmbc10k_integration.html. RNA, ATAC and GAS matrixes were loaded as AnnData objects. Each modality was filtered separately by retaining only features detected in at least 3 cells. RNA data were library-size normalized, log-transformed, and 4000 HVGs were selected. ATAC data were further filtered to only keeping peaks associated with top 50 PCs. GAS matrix used the same normalization and transformation steps as RNA data and was restricted to the RNA-selected HVGs. In the original SIMBA workflow, GAS is computed internally using a MAESTRO-like approach. In this study, internal gene scoring was replaced with externally computed GAS matrix for benchmarking purposes. To ensure consistency with SIMBA's internal logic, MAESTRO-based GAS was computed using a decay distance parameter of 5,000 bp. Final RNA and ATAC embeddings were extracted from the trained SIMBA model for downstream benchmarking analyses.

Clustering methods

Seurat anchor–based label transfer

For Seurat Anchor-based transfer label framework, we directly pass the Seurat generated CCA embeddings to function FindTransferAnchors, Cell type labels defined in the RNA dataset were then transferred to ATAC cells using TransferData, parameters were left at the Seurat defaults. The resulting predicted labels for ATAC cells were used in downstream benchmarking analyses.

FigR Anchor–based label transfer

For FigR Anchor-based transfer label framework, we followed the online tutorial (https://buenrostrolab.github.io/FigR/articles/FigR_stim.html). We downloaded the relevant R source files (cellPairing.R, utils.R, FigR.R) from the FigR GitHub repository (https://github.com/buenrostrolab/FigR) and used only the pairCells functionality. Low-dimensional embeddings of RNA and ATAC cells were supplied as the RNA and ATAC inputs to pairCells.

Three key parameters—search_range, max_multimatch, and min_subgraph_size—were selected automatically based on dataset size and further adjusted in response to solver diagnostics. The search_range parameter determines the size of the geodesic K nearest neighbors (KNN) window (search_range × total cells = KNN size). For datasets with ≤ 5,000 cells we used search_range = 0.2(default); for medium sized datasets (~ 5,000–30,000 cells) we used search_range = 0.05; and for larger datasets such as mouse skin (~ 37 k cells) or BMMC (~ 70 k cells) we reduced this to search_range = 0.005 to limit problem size. If the matching routine reported that no feasible pairing could be found under the current constraints (e.g. errors indicating that matches could not be found and suggesting relaxing constraints), we interpreted this as an overly narrow search window and doubled search_range. Conversely, if errors indicated that the matching problem or KNN search was too large or memory intensive, we halved search_range. For each dataset we allowed up to three such adjustment attempts; if all three failed or the resulting subgraph was severely dominated by one modality, we consider FigR failed for that embedding. The max_multimatch parameter, which bounds how many cells in the larger modality can match to a given cell in the smaller modality, was tuned stepwise: we tried max_multimatch = 10, then 30, then 50 and stopped if all three failed. The minimum subgraph size used for geodesic pairing (min_subgraph_size) was also scaled with dataset size, using 50 cells (default) per modality for ≤ 5,000 cells, 100 for ~ 5,000–30,000 cells, and 200 for the larger mouse skin and BMMC datasets.

For MultiMap embeddings, we observed that the original FigR function: umap_knn_graph internally re-embeds the data with UMAP,which will produced errors when applied to the 2-dimensional MultiMap-embedding. To avoid this and also because MultiMap already provides a umap-like embedding and a second UMAP step is redundant and unstable in this case, we modified umap_knn_graph to accept the MultiMap embedding as the final coordinate space. Geodesic distance computation and bipartite matching in core function: cell_pairing were then performed on MultiMap based KNN graph using the same pipeline and parameter logic as for the other embeddings.

To convert the resulting cell pairs into ATAC labels, we implemented a custom function, transfer_rna_labels_to_atac. First, each ATAC cell that has directly paired to an RNA cell inherited that RNA cell's label. Remaining not paired ATAC cells were then assigned labels by k-nearest-neighbors voting in the ATAC embedding space, with neighbors drawn from already labeled ATAC cells. The final label set for all ATAC cells was used for downstream evaluation.

Leiden clustering

Performed using Scanpy in Python. RNA and ATAC embeddings were first loaded and concatenated to form a joint embedding matrix. This matrix was used to create an AnnData object. A nearest neighbor graph was then constructed using cosine distance. For all datasets, the number of neighbors was set to 20, except for the BMMC dataset, where 200 neighbors were used to account for the larger data size. Leiden clustering was applied with a resolution parameter set to 1.0 using sc.tl.leiden.

KNN label transfer

To transfer RNA-derived cell type labels to ATAC cells using K-nearest neighbors, we read in the RNA and ATAC embeddings and implement the *KNeighborsClassifier* from the scikit-learn Python package, transfer the RNA label to the ATAC data with setting: Euclidean distance and 5 neighbors.

Optimal transport–based label transfer

We used the Moscot package (v0.4.0) to perform optimal transport (OT)–based label transfer, following the official tutorial: https://moscot.readthedocs.io/en/latest/notebooks/tutorials/600_tutorial_translation.html#. RNA and ATAC data were provided in.h5ad format, with features not expressed in any cell excluded. For the OT process, the algorithm requires RNA- and ATAC -specific low-dimensional embeddings, as well as a joint embedding. Instead of computing PCA and LSI embeddings respectively for RNA and ATAC data, we used the corresponding modality-specific embeddings generated during the dimension reduction step for both separate and joint inputs. This approach yielded better results and more accurate label transfer.

Evaluation metrics for GAS

Pearson correlation coefficient (PCC)

To assess the correlation between the predicted GAS and actual gene expression levels, we calculate the PCC for each gene by comparing its GAS and RNA expression across all cells. Since different methods detect varying numbers of genes, we perform the evaluation on a common set of genes to ensure a fair comparison.

Average local neighborhood consistency (LNC)

To evaluate whether cells with similar gene expression profiles also receive similar GAS, we measure the local LNC between RNA and GAS modalities. A high LNC score indicates that GAS effectively captures the population structure of the cells, preserving the underlying biological relationships within the dataset. To compute LNC, we first construct two graphs: one using RNA expression data and another using GAS values, both

based on cosine similarity. For each dataset, we define the neighborhood size (K) as 5% of all cells. Then, for each cell, we determine the number of overlapping neighbors between the two graphs and divide this by the neighborhood size K (Fig. 2a).

Cell type average silhouette width (ASW)

To evaluate how well GAS preserve biological structure within the ATAC modality, we calculate the ASW. This is a commonly used metric for assessing the quality of reduced-dimensional embeddings. To compute ASW, we followed the standard Seurat single-modality RNA preprocessing and PCA workflow. Specifically, genes with zero variance across cells were first removed. The remaining GAS matrix was used to construct a Seurat object, followed by library-size normalization, selection of 2,000 highly variable features, data scaling, and PCA. PCA was performed using the HVGs identified by Seurat, and the first 15 principal components (PCs) were used as the embedding for downstream evaluation. Each GAS method was processed independently using this same pipeline before calculating ASW. Since ASW requires true cell-type labels, we apply this evaluation only to the PBMC10k, SHARE-seq mouse brain and mouse skin datasets.

Evaluation metrics for dimension reductions

The percentage of mutual nearest neighbors (MNN%)

To evaluate whether paired RNA and ATAC profiles are closely aligned in the joint embedding produced by dimension reduction methods, we define the Percentage of Mutual Nearest Neighbors (MNN%). This metric is based on the cosine similarity between RNA and ATAC cell embeddings in the joint space. For each RNA cell, we identify its K nearest ATAC neighbors based on similarity, and for each ATAC cell, we determine its K nearest RNA neighbors. A paired RNA and ATAC cell is considered a mutual nearest neighbor if they both appear in each other's K nearest neighbors' lists. The K is defined as the 5% of the cell number of the dataset. The final MNN% is calculated as the proportion of such mutually listed pairs relative to the total number of cells. A higher MNN% indicates that the embedding effectively captures the similarity between a cell's RNA and ATAC profiles, demonstrating better cross-modality alignment. This metric is in 0–1 range.

Cell type average silhouette width (ASW)

To evaluate how well the reduced-dimensional embedding distinguishes different cell types within each modality, we calculate the silhouette width separately for RNA and ATAC embeddings using true cell-type labels. We then compute the average silhouette width (ASW) per cell type. This evaluation metric is applied to the PBMC10k, Mouse brain SHARE-seq, Mouse skin SHARE-seq data and BMMC datasets.

Contrastive similarity of paired cells

To evaluate how well the embeddings capture similarities while distinguishing differences in the joint space without relying on true labels, we introduce a contrastive similarity metric based on InfoNCE [66]. This metric quantifies how similar the paired RNA and ATAC cells are on the joint embedding compared to their similarity with all non-paired cells. In practice, Specifically, we compute cosine similarities N pairs of cells

across and within modalities using a 2N by 2N similarity matrix. The top-left quadrant captures within-modality similarities among RNA cells, and the bottom-right quadrant represents within-modality similarities among ATAC cells. The off-diagonal blocks contain cross-modality similarities between RNA and ATAC cells. The diagonal of these cross-modality blocks corresponds to the paired RNA–ATAC cells. We define the numerator as the average cosine similarity across all paired RNA–ATAC cells, and the denominator as the average similarity across all non-paired cell pairs:

$$\text{contrastive similarity} = \log\left(\frac{exp\left(Sim_{i=j}/\tau\right)}{exp\left(Sim_{i\neq j}/\tau\right)}\right)$$

Here, $sim_{i=j}$ represents the average similarity between all the paired RNA and ATAC cells (positive pair), while $sim_{i\neq j}$ represents the average similarity of all non-paired cells (negative pairs), including those from the same and alternative modalities. To normalize the scale, both the numerator and denominator are divided by a temperature hyperparameter (τ), (set to 1 in our paper), followed by an exponential transformation. A positive value of this metric indicates that paired RNA and ATAC cells are more similar to each other than to unrelated cells, reflecting successful integration. A negative value suggests that paired cells are less similar than unrelated cells, indicating poor alignment in the embedding space.

Evaluation metrics for clustering and labeling approaches

NMI between RNA and ATAC label

To evaluate the performance of clustering methods, we measure the clustering agreement of two modalities cells by calculating the NMI, with values ranging from 0 (no agreement) to 1 (perfect match). For joint clustering methods, we computed NMI between the clustering labels obtained from the two modalities. For label transfer methods, we measured the agreement between the source labels (RNA clustering labels obtained using Seurat in our study) and the predicted labels for the ATAC modality.

Average consistency of RNA label and ATAC label

To evaluate the label consistency between the two modalities at the cell type level, we designed a metric based on bidirectional label agreement. Specifically, for each RNA cluster, we calculated the percentage of corresponding ATAC cells that were assigned the same label and then averaged these percentages across all RNA clusters. Conversely, for each ATAC cluster, we computed the percentage of matched RNA cells with the same label and took the average across all ATAC clusters. The final metric is the mean of these two averages. This consistency score ranges from 0 to 1, with higher values indicating better agreement. Notably, this metric is sensitive to misclassification of small populations.

Benchmark implementation

Ground-truth labels and where they are used

Ground-truth cell-type labels were provided by the original data authors and were available only for a subset of datasets (PBMC, BMMC, and SHARE-seq mouse brain and mouse skin). We used these labels only for evaluations that require known cell identities, including (i) ASW for GAS evaluation (computed on PCA-reduced GAS) and (ii)

ASW-based evaluation of dimensional-reduction embeddings. We also used them for cell type–level summaries of GAS results on these labeled datasets.

GAS evaluation and gene-set harmonization

Within each dataset, different GAS methods can output different gene sets. To ensure fair comparisons, GAS metrics were computed using the intersection of genes shared across all GAS methods within that dataset (and the intersection of cells when method-specific filtering led to differences in cell sets). The BMMC dataset was downloaded as a preprocessed `.h5ad` object and did not include fragment files; therefore, BMMC was excluded from fragment-based GAS evaluations.

Failure definition, timeouts, and handling missing results

Runs were treated as failures for a dataset if they did not finish within 72 h, or if they failed after more than three parameter-adjustment attempts and still did not produce results. Failed runs were excluded from downstream analyses for that dataset. We recorded failures for both dimension-reduction methods (Additional file 2: Table S4) and dimension-reduction + clustering combinations (Additional file 2: Table S5) across the nine RNA + ATAC datasets. When computing mean performance across the nine RNA + ATAC datasets, failed dataset–method entries were assigned a value of 0 for the corresponding evaluation metrics.

Metric truncation for visualization

For presentation purposes in the dimension-reduction benchmarks, negative ASW and contrastive similarity values were truncated to zero (Fig. 3a). For correlation analyses between these two metrics (Fig. 3b), we used the original, untruncated values.

Unless otherwise stated, all methods were run using tutorial-recommended default parameters.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04071-5.

Additional file 1. Supplementary Note: Case study—How scRNA-seq and scATAC-seq integration accuracy impacts cell-type–specific peak–gene regulatory inference in E18 mouse brain multiome data. Contains Table S1. Confusion matrix for Seurat label transfer in detecting significant peak–gene pairs. Table S2. Confusion matrix for GLUE + optimal transport label transfer. Fig. S1. UMAP of the E18 mouse brain RNA data. Fig. S2. Markers genes in E18 mouse brain data. Fig. S3. Annotation of E18 mouse brain data. Fig. S4. Comparison of ground truth ATAC cell labels and Seurat transferred ATAC cell labels. Fig. S5. Examples of peak–gene correlations under different label transfer methods. Fig. S6. Overall comparison of Seurat vs. GLUE + OT. Fig. S7. Dataset-specific evaluation of gene activity score methods. Fig. S8. Cell type–specific evaluation of gene activity score (GAS) methods. Fig. S9. Dataset-specific MNN. Fig. S10. Dataset-specific average silhouette width (ASW). Fig. S11. Dataset-specific contrastive similarity. Fig. S12. Effect of embedding size on Seurat and LIGER embedding performance. Fig. S13. Dataset-specific NMI across embeddings and clustering algorithms. Fig. S14. Dataset-specific average consistency across embeddings and clustering algorithms. Fig. S15. Robustness of dimension reduction methods and downstream clustering under unbalanced RNA–ATAC cell numbers.

Additional file 2: Table S3. Summary of the three-step evaluation of 180 RNA + ATAC integration pipelines. Table S4. Completion status of dimension reduction run across dataset. Table S5. Completion status of distinct embedding + clustering run across dataset.

Acknowledgements
We acknowledge the use of computational resources provided by Clemson University and the Institute for Human Genetics Bioinformatics Core.

Peer review information
Sergei Mangul, Andrew Cosgrove and Claudia Feng were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
F.N., Q.Y. and Z.D. designed the analytical approach. F.N. wrote the code and performed the data analysis. F.N. and Z.D. wrote, revised, and contributed to the final manuscript. The authors read and approved the final manuscript.

Funding
This work was partially supported by NIH grants R35GM150513 and R21DA060503.

Data availability
All datasets used in this study are publicly available. 10 × Genomics datasets, including PBMC [21], human brain [22], embryonic mouse brain [23], human and mouse kidney cancer [24, 25], and human small intestine [26], were obtained from https://www.10xgenomics.com/datasets. The human bone marrow mononuclear cell (BMMC) dataset is available in the Gene Expression Omnibus (GEO) under accession GSE194122 [27]. SHARE-seq datasets for mouse skin (RNA: GSM4156608 [30]; ATAC: GSM4156597 [31]) and mouse brain (RNA: GSM4156610 [28]; ATAC: GSM4156599 [29]) were downloaded from GEO. RNA and histone modification (H3K4me1, H3K4me3, H3K27ac) data from adult mouse frontal cortex are available under GEO accession GSE152020 [32]. All code used in this study is available at Github: https://github.com/Durenlab/UnPairBench [67] and archived in Zenode: https://doi.org/10.5281/zenodo.19188899 [68] and is distributed under the MIT License.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
ZD is an Editorial Board Member for Genome Biology but was not involved in the editorial process of this manuscript.

Received: 16 May 2025 Accepted: 3 April 2026
Published online: 15 April 2026

References

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. Mrna-seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377–82. https://doi.org/10.1038/nmeth.1315.
2. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523:486–90. https://doi.org/10.1038/nature14590.
3. Yuan Q, Duren Z. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. Nat Biotechnol. 2025;43:247–57. https://doi.org/10.1038/s41587-024-02182-7.
4. Duren Z, Lu WS, Arthur JG, Shah P, Xin J, Meschi F, et al. Sc-compreg enables the comparison of gene regulatory networks between conditions using single-cell data. Nat Commun. 2021;12:4763. https://doi.org/10.1038/s41467-021-25089-2.
5. Duren Z, Chang F, Naqing F, Xin J, Liu Q, Wong WH. Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with scREG. Genome Biol. 2022;23:114. https://doi.org/10.1186/s13059-022-02682-2.
6. Baul S, Tanvir Ahmed K, Jiang Q, Wang G, Li Q, Yong J, et al. Integrating spatial transcriptomics and bulk RNA-seq: predicting gene expression with enhanced resolution through graph attention networks. Brief Bioinform. 2024;25:bbae316. https://doi.org/10.1093/bib/bbae316.
7. Mimitou EP, Lareau CA, Chen KY, Zorzetto-Fernandes AL, Hao Y, Takeshima Y, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat Biotechnol. 2021. https://doi.org/10.1038/s41587-021-00927-2.
8. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;361:1380–5. https://doi.org/10.1126/science.aau0730.
9. Zhu C, Yu M, Huang H, Juric I, Abnousi A, Hu R, et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. Nat Struct Mol Biol. 2019;26:1063–70. https://doi.org/10.1038/s41594-019-0323-x.
10. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol. 2019;37:1452–7. https://doi.org/10.1038/s41587-019-0290-0.
11. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell. 2020;183:1103-1116.e20. https://doi.org/10.1016/j.cell.2020.09.056.

12. Unterman A, Sumida TS, Nouri N, Yan X, Zhao AY, Gasque V, et al. Single-cell multi-omics reveals dyssynchrony of the innate and adaptive immune system in progressive COVID-19. Nat Commun. 2022;13:440. https://doi.org/10.1038/s41467-021-27716-4.
13. Zhou Y, Geng P, Zhang S, Xiao F, Cai G, Chen L, et al. Multimodal functional deep learning for multiomics data. Brief Bioinform. 2024;25:bbae448. https://doi.org/10.1093/bib/bbae448.
14. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19:41–50. https://doi.org/10.1038/s41592-021-01336-8.
15. Lee MYY, Kaestner KH, Li M. Benchmarking algorithms for joint integration of unpaired and paired single-cell RNA-seq and ATAC-seq data. Genome Biol. 2023;24:244. https://doi.org/10.1186/s13059-023-03073-x.
16. Xiao C, Chen Y, Meng Q, Wei L, Zhang X. Benchmarking multi-omics integration algorithms across single-cell RNA and ATAC data. Brief Bioinform. 2024;25:bbae095. https://doi.org/10.1093/bib/bbae095.
17. Liu C, Ding S, Kim HJ, Long S, Xiao D, Ghazanfar S, et al. Multitask benchmarking of single-cell multimodal omics integration methods. Nat Methods. 2025;22:2449–60. https://doi.org/10.1038/s41592-025-02856-3.
18. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177:1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
19. Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat Biotechnol. 2022;40:1458–66. https://doi.org/10.1038/s41587-022-01284-4.
20. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9:5233. https://doi.org/10.1038/s41598-019-41695-z.
21. 10x Genomics. pbmc_granulocyte_sorted_10k - PBMC from a healthy donor - granulocytes removed through cell sorting (10k). Dataset. 2020. https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-1-0-0.
22. 10x Genomics. Human_brain_3k - Frozen human healthy brain tissue (3k). Dataset. 2020. https://www.10xgenomics.com/datasets/frozen-human-healthy-brain-tissue-3-k-1-standard-1-0-0.
23. 10x Genomics. E18_mouse_brain_fresh_5k - Fresh embryonic E18 mouse brain (5k). Dataset. 2020. https://www.10xgenomics.com/datasets/fresh-embryonic-e-18-mouse-brain-5-k-1-standard-1-0-0.
24. 10x Genomics. Mouse Kidney Nuclei Isolated with Chromium Nuclei Isolation Kit, SaltyEZ Protocol, and 10x Complex Tissue DP (CT Sorted and CT Unsorted).Dataset. 2023. https://www.10xgenomics.com/datasets/mouse-kidney-nuclei-isolated-with-chromium-nuclei-isolation-kit-saltyez-protocol-and-10x-complex-tissue-dp-ct-sorted-and-ct-unsorted-1-standard.
25. 10x Genomics. Human Kidney Cancer Nuclei isolated with Chromium Nuclei Isolation Kit, SaltyEZ Protocol, and 10x Complex Tissue DP (CT Sorted and CT Unsorted). Dataset. 2023. https://www.10xgenomics.com/datasets/human-kidney-cancer-nuclei-isolated-with-chromium-nuclei-isolation-kit-saltyez-protocol-and-10x-complex-tissue-dp-ct-sorted-and-ct-unsorted-1-standard.
26. 10x Genomics. Human Jejunum Nuclei Isolated with Chromium Nuclei Isolation Kit, SaltyEZ Protocol, and 10x Complex Tissue DP (CT Sorted and CT Unsorted). Dataset. 2023. https://www.10xgenomics.com/datasets/human-jejunum-nuclei-isolated-with-chromium-nuclei-isolation-kit-saltyez-protocol-and-10x-complex-tissue-dp-ct-sorted-and-ct-unsorted-1-standard.
27. Burkhardt DB, Lücken MD, Lance C, Cannoodt R, Pisco AO, Krishnaswamy S, et al. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. Dataset. Gene Expression Omnibus (GEO); 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122.
28. Ma S, Buenrostro J. Integrative single-cell chromatin and transcriptome profiling uncovers cell-type specific regulatory interactions-mouse brain (RNA-Seq). Dataset. Gene Expression Omnibus (GEO); 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4156610.
29. Ma S, Buenrostro J. Integrative single-cell chromatin and transcriptome profiling uncovers cell-type specific regulatory interactions-mouse brain (ATAC-Seq). Dataset. Gene Expression Omnibus (GEO); 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4156599.
30. Ma S, Buenrostro J. Integrative single-cell chromatin and transcriptome profiling uncovers cell-type specific regulatory interactions-mouse skin late anagen (RNA-Seq). Dataset. Gene Expression Omnibus (GEO); 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4156608.
31. Ma S, Buenrostro J. Integrative single-cell chromatin and transcriptome profiling uncovers cell-type specific regulatory interactions-mouse skin late anagen (ATAC-Seq). Dataset. Gene Expression Omnibus (GEO); 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4156597.
32. Zhu C, Ren B. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. Dataset. Gene Expression Omnibus (GEO); 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152020.
33. Chang T, Zhang M, Zhu J, Wang H, Li C-C, Wu K, et al. Simulated vestibular spatial disorientation mouse model under coupled rotation revealing potential involvement of Slc17a6. iScience. 2023;26:108498. https://doi.org/10.1016/j.isci.2023.108498.
34. Lattke M, Guillemot F. Understanding astrocyte differentiation: clinical relevance, technical challenges, and new opportunities in the omics era. WIREs Mech Dis. 2022;14:e1557. https://doi.org/10.1002/wsbm.1557.
35. Scholzen T, Gerdes J. The Ki-67 protein: from the known and the unknown. J Cell Physiol. 2000;182:311–22. https://doi.org/10.1002/(SICI)1097-4652(200003)182:3<311::AID-JCP1>3.0.CO;2-9.
36. Englund C, Fink A, Lau C, Pham D, Daza RAM, Bulfone A, et al. Pax6, Tbr2, and Tbr1 are expressed sequentially by radial glia, intermediate progenitor cells, and postmitotic neurons in developing neocortex. J Neurosci. 2005;25:247–51. https://doi.org/10.1523/JNEUROSCI.2899-04.2005.
37. Cubelos B, Sebastián-Serrano A, Beccari L, Calcagnotto ME, Cisneros E, Kim S, et al. Cux1 and Cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. Neuron. 2010;66:523–35. https://doi.org/10.1016/j.neuron.2010.04.038.
38. Jabaudon D, Shnider SJ, Tischfield DJ, Galazo MJ, Macklis JD. RORβ induces barrel-like neuronal clusters in the developing neocortex. Cereb Cortex N Y N 1991. 2012;22:996–1006. https://doi.org/10.1093/cercor/bhr182.

39. Yao Z, van Velthoven CTJ, Nguyen TN, Goldy J, Sedeno-Cortes AE, Baftizadeh F, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell. 2021;184:3222-3241.e26. https://doi.org/10.1016/j.cell.2021.04.021.
40. Tessier-Lavigne M, Goodman CS. The molecular biology of axon guidance. Science. 1996;274:1123–33. https://doi.org/10.1126/science.274.5290.1123.
41. Schwarz Q, Vieira JM, Howard B, Eickholt BJ, Ruhrberg C. Neuropilin 1 and 2 control cranial gangliogenesis and axon guidance through neural crest cells. Development. 2008;135:1605–13. https://doi.org/10.1242/dev.015412.
42. Fancy SPJ, Baranzini SE, Zhao C, Yuk D-I, Irvine K-A, Kaing S, et al. Dysregulation of the Wnt pathway inhibits timely myelination and remyelination in the mammalian CNS. Genes Dev. 2009;23:1571–85. https://doi.org/10.1101/gad.1806309.
43. Machold R, Rudy B. Genetic approaches to elucidating cortical and hippocampal GABAergic interneuron diversity. Front Cell Neurosci. 2024;18:1414955. https://doi.org/10.3389/fncel.2024.1414955.
44. Boldog E, Bakken TE, Hodge RD, Novotny M, Aevermann BD, Baka J, et al. Transcriptomic and morphophysiological evidence for a specialized human cortical GABAergic cell type. Nat Neurosci. 2018;21:1185–95. https://doi.org/10.1038/s41593-018-0205-2.
45. Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature. 2018;563:72–8. https://doi.org/10.1038/s41586-018-0654-5.
46. Li S, Nih LR, Bachman H, Fei P, Li Y, Nam E, et al. Hydrogels with precisely controlled integrin activation dictate vascular patterning and permeability. Nat Mater. 2017;16:953–61. https://doi.org/10.1038/nmat4954.
47. Vanlandewijck M, He L, Mäe MA, Andrae J, Ando K, Del Gaudio F, et al. A molecular atlas of cell types and zonation in the brain vasculature. Nature. 2018;554:475–80. https://doi.org/10.1038/nature25739.
48. Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. Proc Natl Acad Sci U S A. 2018;115:7723–8. https://doi.org/10.1073/pnas.1805681115.
49. Yuan Q, Duren Z. Integration of single-cell multi-omics data by regression analysis on unpaired observations. Genome Biol. 2022;23:160. https://doi.org/10.1186/s13059-022-02726-7.
50. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods. 2021;18:1333–41. https://doi.org/10.1038/s41592-021-01282-5.
51. Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. Jointly defining cell types from multiple single-cell datasets using LIGER. Nat Protoc. 2020;15:3632–62. https://doi.org/10.1038/s41596-020-0391-8.
52. Lin Y, Wu T-Y, Wan S, Yang JYH, Wong WH, Wang YXR. ScJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol. 2022;40:703–10. https://doi.org/10.1038/s41587-021-01161-6.
53. Cao K, Gong Q, Hong Y, Wan L. A unified computational framework for single-cell data integration with optimal transport. Nat Commun. 2022;13:7419. https://doi.org/10.1038/s41467-022-35094-8.
54. Alatkar SA, Wang D. CMOT: cross-modality optimal transport for multimodal inference. Genome Biol. 2023;24:163. https://doi.org/10.1186/s13059-023-02989-8.
55. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al. Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. Mol Cell. 2018;71:858-871.e8. https://doi.org/10.1016/j.molcel.2018.06.044.
56. Wang C, Sun D, Huang X, Wan C, Li Z, Han Y, et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. Genome Biol. 2020;21:198. https://doi.org/10.1186/s13059-020-02116-x.
57. Dou J, Liang S, Mohanty V, Miao Q, Huang Y, Liang Q, et al. Bi-order multimodal integration of single-cell data. Genome Biol. 2022;23:112. https://doi.org/10.1186/s13059-022-02679-x.
58. Zhang Z, Yang C, Zhang X. Scdart: integrating unmatched scrna-seq and scatac-seq data and learning cross-modality relationship simultaneously. Genome Biol. 2022;23:139. https://doi.org/10.1186/s13059-022-02706-x.
59. Chen H, Ryu J, Vinyard ME, Lerer A, Pinello L. SIMBA: single-cell embedding along with features. Nat Methods. 2024;21:1003–13. https://doi.org/10.1038/s41592-023-01899-8.
60. Jain MS, Polanski K, Conde CD, Chen X, Park J, Mamanova L, et al. MultiMAP: dimensionality reduction and integration of multimodal data. Genome Biol. 2021;22:346. https://doi.org/10.1186/s13059-021-02565-y.
61. Kartha VK, Duarte FM, Hu Y, Ma S, Chew JG, Lareau CA, et al. Functional inference of gene regulation using single-cell multi-omics. Cell Genom. 2022;2:100166. https://doi.org/10.1016/j.xgen.2022.100166.
62. Klein D, Palla G, Lange M, Klein M, Piran Z, Gander M, et al. Mapping cells through time and space with moscot. 2023. https://doi.org/10.1101/2023.05.11.540374.
63. Fu S, Wang S, Si D, Li G, Gao Y, Liu Q. Benchmarking single-cell multi-modal data integrations. Nat Methods. 2025;22:2437–48. https://doi.org/10.1038/s41592-025-02737-9.
64. Hu Y, Wan S, Luo Y, Li Y, Wu T, Deng W, et al. Benchmarking algorithms for single-cell multi-omics prediction and integration. Nat Methods. 2024;21:2182–94. https://doi.org/10.1038/s41592-024-02429-w.
65. Luecken M, Burkhardt D, Cannoodt R, Lance C, Agrawal A, Aliee H, et al. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In: Vanschoren J, Yeung S, editors. Proc Neural Inf Process Syst Track Datasets Benchmarks. 2021. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/158f3069a435b314a80bdcb024f8e422-Paper-round2.pdf.
66. van denOord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv. 2019.
67. Naqing FNU. Benchmarking component choices for unpaired single-cell RNA and epigenomic integration. Github; 2026. https://github.com/Durenlab/UnPairBench.
68. Naqing FNU. Benchmarking component choices for unpaired single-cell RNA and epigenomic integration. Zenodo; 2026. https://doi.org/10.5281/zenodo.19188899.

Publisher's Note