Benchmarking Computational Integration Methods for Spatial Transcriptomics Data

Yijun Li[1], Stefan Stanojevic[2], Bing He[2], Zheng Jing[3], Qianhui Huang[2], Jian Kang[1], Lana X. Garmire[1,2 *]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[3]Department of Applied Statistics, University of Michigan, Ann Arbor, MI, USA

[*]Corresponding author, email: lgarmire@med.umich.edu

Yijun Li, email : liyijun@umich.edu

Stefan Stanojevic, email: stanojes@med.umich.edu

Bing He, email: hbing@med.umich.edu

Zheng Jing, email: jingzhe@umich.edu

Qianhui Huang, email: qhhuang@umich.edu

Jian Kang, email: jiankang@umich.edu

Abstract

The increasing popularity of spatial transcriptomics has allowed researchers to analyze transcriptome data in its tissue sample's spatial context. Various methods have been developed for detecting SV (spatially variable) genes, with distinct spatial expression patterns. However, the accuracy of using such SV genes in clustering cell types has not been thoroughly studied. On the other hand, in single cell resolution sequencing data, clustering analysis is usually done on highly variable (HV) genes. Here we investigate if integrating SV genes and HV genes from spatial transcriptomics data can improve clustering performance beyond using SV genes alone. We evaluated six methods that integrate different features measured from the same samples including MOFA+, scVI, Seurat v4 , CIMLR, SNF, and the straightforward concatenation approach. We applied these methods on 19 real datasets from three different spatial transcriptomics technologies (merFISH, SeqFISH+, and Visium) as well as 20 simulated datasets of varying spatial expression conditions. Our evaluations show that the performances of these integration methods are largely dependent on spatial transcriptomics platforms. Despite the variations among the results, in general MOFA+ and simple concatenation have good performances across different types of spatial transcriptomics platforms. This work shows that integrating quantitative and spatial marker genes in the spatial transcriptomics data can improve clustering. It also provides practical guides on the choices of computational methods to accomplish this goal.

***Keywords*:** spatial transcriptomics, feature, clustering, dimension, spatial variation, single cell sequencing

---

Background

Spatial omics technologies are one of the breakthroughs in science in the last several years [1], [2–4]. Such technologies are able to measure transcriptome information systematically in the tissue space, thus preserving the spatial context of the tissue samples. The addition of spatial information allows researchers to further explore biological architecture and function and reveal more insights with respect to various disease mechanisms [5–7] [8] [9]. Various techniques for sequencing spatially resolved transcriptome data have been developed, including merFISH [10] [11] [12], SeqFISH [13], SeqFISH+ [14], Visium [15] [16,17], etc. Such technologies can be categorized into two general classes: fluorescence in situ hybridization (FISH)-based methods [14] [10], which directly extract transcriptome information at a molecular level and obtain the spatial locations of the cells by processing the FISH images; and array-based methods [16] [15], which attach probes with fixed physical locations to cryosections of tissues to obtain transcriptome information.

Since the function of cells is often related to their location in the tissue and their genetic information, spatial transcriptomics methods have great potential in helping us gain insights into biological mechanisms. Genes with a spatially distinct expression pattern, or spatially variable (SV) genes, are a novel type of markers for identifying cell types. Recently several computational methods, such as trendsceek [18], spatialDE [19], SPARK [20], Giotto [21], and MERINGUE [22], have been developed with the function to identify spatially variable genes (SV genes). However, clustering and cell-type identification are conventionally based on highly variable (HV) genes, in single cell RNA-Seq datasets which provide no spatial information

directly. In spatial transcriptomics data, the SV genes and HV genes are often quite distinct sets, despite the possibility of overlapping on some genes. We therefore asked if the integration of conventional HV genes and the novel SV genes can help to improve clustering performance, an area unexplored currently.

Towards this goal, we conducted a comprehensive benchmarking study to integrate HV genes and SV genes in spatial transcriptomics data. We adapted six different computational methods for this task, namely scVI, CIMLR, SNF, Seurat v4, MOFA+, and the straightforward concatenation approach. scVI is a neural-network based method originally designed for integrating single-cell transcriptomic data [23]. Seurat v4 is a Weighted Nearest Network method [24] and MOFA+ is a matrix factorization method [25], both of which were used to integrate multi-omics single cell data. We also adopted two additional methods originally developed to integrate bulk multi-omics data: CIMLR (Cancer Integration via Multikernel Learning) model is an integration method using Gaussian kernels [26] and SNF is a network fusion method based on patient-patient similarities [27]. While other methods align multiple types of genomics data from similar (but not the same) cells, such as LIGER [28], MATCHER [29] and Harmony [30], these methods are deemed unsuitable for the question in the study, since the SV genes and HV genes here are obtained from the same cells (or spots).

We applied these methods on a total of 19 real datasets across three spatial transcriptomics platforms, including the Mouse Somatosensory Cortex dataset by SeqFISH+[14], six datasets by 10X genomics' Visium platform [16,31–35] and twelve datasets by merFISH methods [10]. We also evaluated the clustering results on 20 simulated datasets. The clustering performance was measured by non-spatial clustering metrics using AMI (Adjusted Mutual Information), spatial clustering metrics via total Variation Norm, and by downstream effects on differential expression

analysis via the F-1 score, as well as practical running time and memory use. The benchmarking results are expected to provide practical guidelines for researchers conducting spatial omics clustering analysis, where the SV genes and HV genes present distinct properties of the data.

Results

Overview of Benchmark Comparisons

We conducted a comprehensive benchmarking study to evaluate the effect of integrating HV genes and SV genes in spatial transcriptomics data on 19 real datasets and 20 simulated datasets, using 5 different integration approaches. The characteristics of the integration methods used in the study are summarized in **Table 1**. We strictly selected only methods that integrate multi-omics data generated from the same cells and excluded methods that align multiple types of genomics data from similar (but not the same) cells, since the SV genes and HV genes here are obtained from the same cells (or spots). The processing pipeline for benchmarking is shown in **Fig. 1**. It starts with spatial transcriptomics data which has two components: a gene expression matrix, and spatial data which consists of the spatial coordinates of each spot or cell. After data preprocessing and normalization (see **Methods**), we extracted the HV genes using the gene expression matrix and the SV genes using both the gene expression matrix and spatial data. In order to remove any potential bias, we excluded the genes that were both highly variable and spatially variable. We then conducted dimensionality reduction to obtain the same number of extracted features, using either the internal steps within the methods (MOFA+, scVI, Seurat v4 and CIMLR) or using PCA (Principal Component Analysis) for the approach without this step (concatenation). Next, we applied clustering (default Leiden algorithm) for all methods except

for CIMLR, which has its own built-in clustering step. For a fair comparison of evaluation metrics, we repeatedly performed Leiden clustering until we obtained the same number of clusters as those from the ground truth. Finally, the clustering performance was assessed by non-spatial clustering metric AMI (Adjusted Mutual Information) and the spatial clustering metric VN (Variation Norm). We also performed differential expression analysis based on the clustering labels using the F-1 score, to assess the downstream quality of the clustering labels. We used the results obtained from using only SV genes as the baselines, in comparison to those obtained from integration methods.

Clustering Accuracy on Real Spatial Transcriptomics Datasets

We compared the clustering accuracy performance of paired multi-omics integration methods on 19 datasets with available ground truth information. Our real datasets span over diverse resolutions, tissue types and cell type compositions (**Table 2**). Six datasets are from Visium [15], a non-single-cell resolution technology, or its prototype technology. They include a Mouse Olfactory Bulb dataset [16], a Human Cerebellum dataset [31], a Mouse Kidney dataset [32], and three different Mouse Brain datasets [33–35]. We included the Mouse Somatosensory Cortex dataset from SeqFISH+ [14], a single-cell resolution dataset. We also selected twelve datasets on separate sections of the Mouse Hypothalamus by merFISH [10], another single-cell resolution technology.

For non-spatial clustering evaluation, we evaluated the methods' performance by computing AMI (Adjusted Mutual Information). We compared the clustering performances with those using solely SV genes as the baseline (**Fig. 2a-c, Supplementary Fig. 1a-b**). There is no single best

method for all datasets (**Fig. 2a-b, Supplementary Fig. 1a-b**). However on average, MOFA+ has the highest median AMI values across datasets (**Fig. 2c**), significantly better than the baseline (p-value = 0.0004). It is closely followed by concatenation and scVI, which also have significantly higher AMIs (p-values 0.0020 and 0.0446 respectively) than clustering solely based on SV genes. Oppositely, SNF, CIMLR and Seurat v4 are significantly worse (p-values 0.0020, 0.0012 and 0.0004 respectively) at integrating SV and HV genes, compared to the baseline of non-spatial clustering using SV genes only (**Fig.2c**). In particular, Seurat v4 has a much wider range of AMIs compared to other methods, prompting a closer look at the AMI values across different datasets. Indeed, Seurat v4 shows technology-dependent performance patterns (**Fig. 2a-b, Supplementary Fig. 1a-b**). It has some of the lowest AMIs in datasets of single cell resolutions, particularly in merFISH datasets; However, it is ranked better in the Visium platform. Conversely, scVI has higher AMI rankings across single-cell resolution datasets than those datasets that are not measured at single-cell resolutions.

Since spatial transcriptomics data measure gene expression *in situ,* we also evaluated the clustering performance of each integration method with respect to the spatial distribution of the clustering labels (**Fig. 2d-f, Supplementary Fig. 2a-b**). We derived spatial variation norm (VN) to measure the performance of spatial clustering (see **Methods**). Again, there is no single best method for all datasets (**Fig. 2d-e, Supplementary Fig. 2a-b**), however, some general trends do emerge (**Fig. 2f**). Concatenation has the highest average VN value, significantly better (p-value = 0.0407) than clustering solely based on SV genes (**Fig. 2f**). scVI and MOFA+ show comparable VN values to the baseline condition. Seurat v4 yields slightly worse average VN value and wider variations, Again, CIMLR (p-value = 0.0446) and SNF (p-value = 0.0005) significantly underperform the baseline clustering performance of using solely SV genes (**Fig.**

2f). We observe again technology-dependent performance patterns (**Fig. 2d-e, Supplementary Fig. 2a-b)**. Seurat v4 appears to yield better rankings of VN in non-single-cell resolution Visium datasets, but worse rankings in datasets of single cell resolutions. Overall, the evaluation of spatial clustering performance using VN is largely in concordance with non-spatial clustering performance based on AMI. This further confirms that both spatial and quantitative features are useful in cell type clustering, when the proper integration methods are used.

Clustering Accuracy on Simulated Spatial Transcriptomics Datasets

It is possible that the ground truth labels of real datasets are subject to biases from the genes and references that were used to obtain the annotations. Therefore, we also performed a simulation study where we could generate the dataset along with unambiguous ground truth labels. To simulate realistic transcriptomics data, we used the gene expression, spatial coordinates, and ground truth cell types of the Visium Mouse Olfactory Bulb dataset as our reference [16]. We aimed to explore how the level of spatial distinction impacted the integration and clustering performance of the methods. A gene's spatial expression is defined as its expression pattern in the context of the spatial distribution of the cells (or spots); the more distinct the spatial pattern, the more spatially expressed the gene is. We used the spatial pattern of cell types in the ground truth to closely approximate the actual spatial transcriptomics data. We varied the level of the spatial distinction profile of SV genes by simulating new spatial patterns for these genes (**Fig. 3**). The level of spatial distinction for the new spatial pattern is controlled by a probability parameter p (see **Methods**). We varied the values of probability p from 0.6 to 0.9 and created five random

repetitions for each value. We then evaluated how the probability parameter p affects the methods' performance using the same pipeline **(Fig. 1)** as the previous real datasets.

As with real datasets analysis, we evaluated the performances of clustering results from different integration methods, compared to those based solely on the SV genes (**Fig. 4**). With respect to non-spatial clustering metric, all methods except SNF, consistently yield significantly higher AMI values compared to the baseline condition of using SV genes solely **(Fig. 4a)**. Moreover, as the level of spatial distinction increases, the AMI values of all methods continue to improve, confirming that the level of spatial distinction certainty is informative of cell type clustering (**Fig.4a**). Under varying levels of spatial distinction, Seurat v4 has the best and the most stable non-spatial clustering performance (VN) among all the integration methods. Other methods such as concatenation and MOFA+ also do well. When the level of spatial distinction increases, the discrepancy between MOFA+, concatenation and the top performing method Seurat v4 become smaller. This suggests that MOFA+ and concatenation methods may require a certain level of spatial distinction to cluster accurately. Again SNF has relatively worse stability over different levels of spatial distinction.

For spatial clustering using VN, we observe an overall increasing accuracy trend as the level of spatial distinction increases for most integration methods (**Fig.4b**). Seurat v4 has the best and most stable spatial clustering performance among all the integration methods, followed by MOFA+. Both methods consistently outperform spatial clustering solely based on SV genes, over varying degrees of spatial distinction. For some other integration methods, the spatial clustering performance relative to the baseline is dependent on the level of spatial distinction. For example, the simple concatenation approach and scVI yield higher VNs than using SV genes, when the spatial probability parameter is above 0.7 and 0.8, respectively. Similar to the

real datasets, CIMLR and SNF have lower VN values compared to those using solely SV genes, and SNF has the worst stability over different levels of spatial distinction too.

Effect on Downstream Differential Expression

Clustering analysis is usually an important step for other downstream analyses. In order to further assess the downstream effect of clustering through integrating SV and HV genes, we performed differential expression analysis. For this we used the differential expression analysis results based on the ground truth labels as the references. We tested how many original markers can be recovered based on the integration cluster labels. We evaluated the downstream differential expression analysis performance using the F-1 score (see **Methods**), in comparison with that of the clustering based solely on SV genes.

For the 19 real datasets, concatenation had the highest average F-1 score and significantly better than the baseline performance based solely on SV genes (p-value = 0.0057). scVI, MOFA+ and CIMLR methods also achieve higher average F-1 scores than the baseline performance based solely on SV genes, but the advantage is not statistically significant. SNF and Seurat v4 yield significantly worse F-1 scores than the baseline, both with p-values less than 0.01 (**Fig. 5c**). The specific values and rankings of the F-1 scores again present technology dependent patterns, similar to those observed by AMI and VN. Seurat v4 obtains better rankings of F-1 scores in non-single-cell-resolution datasets than in single-cell resolution datasets (**Fig. 5a-b, Supplementary Fig. 3a-b**).

We performed similar downstream differential expression analysis for the simulated datasets (**Fig. 5d**), and observe the general increasing trend in average F-1 score for most integration methods. No method gives the best F-1 scores unanimously, over all levels of spatial distinction. The F-1 scores of these integration methods are dependent on the level of spatial distinction. Seurat v4 gives the highest F-1 rankings when spatial probability values are 0.7 and above. MOFA+ and concatenation also yield good F-1 rankings for differential expression analysis. Again all methods except SNF, show significantly better F-1 scores than the baseline condition. Also, the F-1 score of SNF is relatively unstable compared to other integration methods.

Runtime and Memory Usage

Runtime efficiency and memory usage are two important practical measures for computational methods. Therefore, we used real Human Cerebellum dataset by Visium platform for testing. We compared the runtime and memory usage of the six integration methods by subsetting the number of cells from 500, 1000, 1500, 3000, to 5000 and the number of features from 500, 1000, 3000, 5000, to 10000. All the integration methods were run on the Garmire Lab server, which operated on a Linux operating system (Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz, 219 GB RAM, 16 GB Nvidia GPU). MOFA+, scVI and SNF were run on GPU nodes and concatenation, CIMLR and Seurat v4 ran on regular nodes.

Increasing the number of cells has a more significant impact on both runtime efficiency and memory usage than increasing the number of features (**Fig. 6)**. For running time measurements, straightforward concatenation is the most time-efficient method among all, as expected (**Fig. 6 a, c**). The other iterative computational methods such as Seurat v4, CIMLR, SNF, MOFA+ and

scVI have an ascending order for computing time. Increasingly longer runtimes are needed, as the number of cells or the number of features increases (**Fig.6 a, c**). For memory usage, concatenation and SNF are the most memory-efficient methods over varying numbers of cells and features. On the other hand, CIMLR uses the most memory, which increases significantly more than the rest of the methods. This could be attributed to CIMLR's adaptation of parallel computation.

Discussion

The advent of spatial transcriptomics technologies gives us the potential to create a more comprehensive map of biological systems. Relative to single cell RNA-Seq technologies, the addition of spatial information has the potential to help discover novel SV markers. Previously, SV markers have been used for identifying clusters in spatial transcriptomics data [18–21]. However, these SV markers represent a different category from those HV gene markers detected by quantitative variabilities conventionally [36–38]. Questions remained: (1) if SV gene based clustering can be improved by integrating additional SV genes, which are normally used in single cell RNA-Seq analysis for clustering; (2) if integration of SV and HV genes can improve clustering results in spatial transcriptomics data, which computational method(s) to use. Here, we conducted a systematic benchmarking study and confirmed that combining these two types of markers can improve the clustering results, with appropriate integration methods.

Fig. 7 shows an overview of the non-spatial clustering accuracy (by AMI), spatial clustering accuracy (by VN), and downstream differential analysis performance (by F-1 score), runtime and memory usage for 19 real datasets across different technology platforms and 20 simulated

datasets across different spatial distinction levels. MOFA+ is recommended for general spatial transcriptomics datasets; straightforward concatenation is also a good approach especially considering the computational cost (running time and memory use); however, it is more sensitive to the level of spatial distinction in the dataset than MOFA+. Technology-specific patterns show that Seurat v4 performs better for integrating HV and SV genes of non-single-cell resolution datasets obtained from the Visium platform. SNF generally yields worse results than the baseline condition using solely on SV genes. Therefore, the current implementation of SNF is not suitable for integrating SV and HV genes in spatial transcriptomics data.

MOFA+ belongs to a class of matrix-factorization-based methods. Our evaluation shows that it may improve clustering results in integrating HV and SV genes, as well as subsequent downstream differential expression analyses. This class of models was previously shown quite effective in utilizing information from multiple sets of genomics features and extracting useful information for integration [28,39,40]. In particular, MOFA+ decomposes each component dataset as a product of a shared dimension-reduced latent space and a dataset-specific weight matrix. MOFA+ model can learn the shared dimensionally-reduced latent space by borrowing information from each component dataset, thus allowing for effective integration (see **Method**). Dimension reduction also appears to be an important factor in successful integration of features. Both Seurat v4 and SNF are based on building joint networks, but Seurat v4 has significantly better performance than SNF. Seurat v4 performs dimension reduction before building the networks whereas SNF attempts to directly integrate networks built on data not dimensionally reduced. Though slightly better than SNF, CIMLR generally performs worse than the other four methods. This could be attributed to the Gaussian kernel assumption that CIMLR makes (see

Method). Such an assumption might over-simplify the complexities of spatial transcriptomics data.

Conclusions

In general, we recommend MOFA+ and straightforward concatenation for integration of quantitatively variable and spatially variable genes in spatial transcriptomics data. We also stress the need for methods to be developed in this area which are further complicated by different spatial transcriptomics technology platforms.

Methods

Data Preprocessing

For all the computational methods including extracting HV and SV genes, we preprocessed the data by first filtering the raw gene expression dataset. The expression threshold was set to 1. The minimum number of cells a gene needs to be expressed in was set to 3 and the minimum number of expressed genes in a cell was set to at least 5% of the total number of genes. Such values were chosen to filter out unexpressed genes or cells without over-processing the data. Computational methods such as straightforward concatenation, MOFA+, SNF and CIMLR require normalized gene expression data. For these methods, the filtered raw gene expression data was normalized via log-normalization, and the cells and genes were scaled by Giotto's [21] default scale factor of 6000. We skipped normalization for Seurat v4 and scVI, as they require raw data.

Benchmark Analysis Parameters

We extracted the HV genes and the SV genes based on normalized expression data. Among the computational methods, MOFA+, scVI, Seurat v4 and CIMLR have inherent dimension reduction steps (see Supplementary Table 1 for details) which extract combined features on a lower dimension. For the rest of the methods that do not have inherent dimension reduction steps, we performed dimension reduction using PCA (Principal Component Analysis). The dimension of the extracted features was set to be the same across all methods for a fair comparison. The optimal number of features were determined empirically by comparing the general clustering performance across different values of reduced dimensions. We used Leiden clustering for all methods [41] except for CIMLR, which had its own built-in clustering step. For a fair comparison of evaluation metrics, we repeatedly performed Leiden clustering, tuning the resolution and number of nearest neighbors parameters for each computational method until we got the same number of clusters as there were ground truth clusters.

Selection of HV and SV genes

We used HV (highly variable) genes as features whose expression was informative of how the data clustered. We extracted HV genes using the R package *Giotto*'s *calculateHVG* function [21]. The function fits a LOESS regression model with each gene's log mean expression as the independent variable and the coefficient of variance as the dependent variable. Genes with a coefficient of variance that was higher than their predicted value past an arbitrary threshold were considered to be highly variable. We set the threshold value to the default: 0.1.

We used SV genes as features whose expression was informative of how clusters were distributed spatially. We extracted the SV genes using the R package *Giotto*'s *binSpect* function [21]. The function binarized each gene's expression value using the K-means method (number of centers set to 2). The function then constructed a spatial connectivity network based on the cells/spots' spatial coordinates using Delaunay's triangulation [42]. A contingency table was generated based on the connectivity of the spatial network and the binarized gene expression. The function performed the Fisher's test and generated a spatial expression score for each gene with higher values indicating a more distinct spatial expression pattern.

To avoid bias, we excluded genes that were both highly variable and spatially variable. Suppose there are g_1 unique HV genes and the optimal number of reduced dimensions is k. We denote the subset of normalized HV genes as E_1(g_1 genes $\times$ m cells). We performed PCA on E_1 and selected the top k principal components, represented by a $k \times m$ matrix denoted as W_{HV}. We used W_{HV} for further clustering analysis. Similarly, suppose there are g_2 unique SV genes. We denoted the normalized SV genes as E_2 (g_2 genes $\times m$ cells). We performed PCA on E_2 and selected the top k principal components, represented by a $k \times m$ matrix denoted as W_{SV}. We used W_{SV} for further clustering analysis.

Methods in comparison

Simple concatenation model: we concatenated the normalized HV genes E_1 and the normalized SV genes E_2, represented by a $(g_1 + g_2)\times m$ matrix denoted as E. We performed PCA on the concatenated dataset E and took the top k principal components for further clustering analysis.

CIMLR (Cancer Integration via Multikernel Learning) model [26]: we adapted CIMLR, which was originally developed to integrate multi-omics cancer data, to integrate normalized HV genes E_1 and normalized SV genes E_2. CILMR constructed a group of Gaussian kernels based on $E_1$and E_2 separately. CIMLR then computed the similarity matrix between the two datasets by combining the gaussian kernels. Given the number of clusters, dimension reduction and clustering via k-means clustering were then performed on the similarity matrix.

MOFA+ (Multi-Omics Factor Analysis v2) model [25]: MOFA+ is a statistical framework for integrating multi-omics single-cell data based on representing the original dataset in lower dimensions through matrix factorization infused with a Bayesian technique for enhancing model sparsity. We adapted MOFA+ to integrate HV genes and SV genes. We used the normalized HV genes E_1 and the normalized SV genes E_2 as input. MOFA+ aims to reconstruct E_i using lower dimensional matrices W_{MOFA+} $(m\times k)$ and U_i $(g_i\times k)$: (ϵ_i represents random noise) (Eq.2). W_{MOFA+} represents the low-dimensional factors common across all datasets and U_i represents each feature's association score with each factor. We used W_{MOFA+} for further clustering analysis.

$$E_i = W_{MOFA+}U_i^T + \epsilon_i,\ i = 1, 2 \quad (2)$$

scVI (Single-Cell Variational Inference) model [23]: scVI constructs latent space representations of cells using a variational autoencoder. The *encoder* neural network maps gene

expression values onto parameters of the probability distribution from which latent variables are sampled, while the *decoder* neural network takes as input such latent variables and aims to reproduce the original input, by producing the parameters of the ZINB (zero-inflated negative binomial) distribution modelling the gene expression values. The model also uses neural networks to model dropout patterns and adds scales to account for differences in gene expression between different cells. We concatenated the raw HV genes and the raw SV genes and fed them as input to the scVI model. We used scVI with a single hidden layer in an encoder and optimized the dimension of the hidden layer, and the dropout rate through a grid search, optimizing for the minimum reconstruction error.

SNF (Similarity Network Fusion) model [27]: we adapted the SNF algorithm to construct graph representations of normalized HV genes E_1 and normalized SV genes E_2 and merged them to construct a coherent picture. Each node on such graphs corresponds to different cells and the edge weights correspond to the degree of similarity (correlation) between them when separately considering highly variable and SV genes. The graphs were fused through iterative updates, inspired by diffusion processes, which gradually bring the graphs closer together. We performed clustering analysis on the integrated similarity graph.

Seurat v4 (Weighed Nearest Network) model [24]: we adapted Seurat v4 to compute cell-specific weights separately for the raw HV genes and the raw SV genes. Such weights reflected the "usefulness" of information from HV genes and SV genes. Seurat v4 integrated the HV genes and SV genes while leveraging their constructed weights so the integrated data reflected the information of its component datasets. Nearest neighbor graph construction and clustering were then performed on the integrated data.

Ground truths for real data sets

The quality of the ground truth labels are essential to the evaluation of methods' performance. We had complete control over the ground truth labels of the simulated datasets as they were arbitrarily generated. For the real datasets, we obtained ground truth labels from the original studies whenever possible. For the Mouse Olfactory Bulb dataset [16] generated by the technology that was later named Visium, we were able to recreate the cell type assignments based on the original supplementary materials. For the rest of the five Visium datasets, we used the clustering labels provided by 10X [31–35]. For the Mouse Somatosensory Cortex dataset by SeqFISH+ [14], we used the clustering labels generated by *Giotto* package [21]. For the merFISH datasets, we used the cell type annotations in the original report [10].

Simulation Study

Using the Mouse Olfactory Bulb dataset as reference, we used the original 3291 HV genes of the real dataset as the HV genes in this simulation study. We selected SV genes whose adjusted p-value is below 0.05. To ensure the simulated spatial expression patterns were meaningful with respect to identifying the ground truth cell types, we further selected SV genes whose spatial expression pattern matched the spatial pattern of a particular ground truth cell type by more than 60%. We found 335 such SV genes from the original Mouse Olfactory Bulb dataset.

We adapted *Giotto*'s [21] *simulateOneGenePatternGiottoObject* function to simulate spatial expression profiles for each individual selected spatially variable gene. We used the spatial patterns of cell types in the ground truth to guide customized simulations for spatial gene

expression patterns. Given a customized pattern that contains n cells, with probability p, the spots where the gene is highly expressed will fall into the customized pattern in the simulated dataset. When $p = 0.5$, the cells will fall into the customized pattern completely randomly; when $p = 1.0$, the cells will fall into the customized pattern with absolute certainty. We tested four different values for spatial probability p (0.6, 0.7, 0.8, 0.9) to represent different noise levels in clustering labels, and repeated the experiment 5 times for each value of p.

Hypothesis Tests

We performed non-parametric paired Wilcox tests to further investigate the efficacy of the performance of the integration methods in comparison to the baseline performance of solely SV genes. For each metric (AMI for non-spatial clustering, VN for spatial clustering, F-1 for downstream differential expression analysis), we performed the paired Wilcox test on each pair of SV genes metrics and integration method metrics. The null hypothesis is that the metric performance of solely SV genes comes from the same distribution as the performance of the integration method. The methods whose corresponding p-values are less than 0.05 were considered to perform statistically significantly different from the baseline SV genes.

Evaluation Metrics

Adjusted Mutual Information (AMI): To evaluate the clustering analysis' general performance, we computed the Adjusted Mutual Information (AMI) [43] of each set of clustering results compared with the ground truth. AMI is bounded between 0 and 1, with higher values indicating better clustering performance.

Variation Norm (VN): To evaluate the clustering results in the spatial context, we used a metric based on the total variation norm denoted as VN [44]. The larger VN is, the more accurate the spatial clustering is compared to the ground truth. We denote the set of clustering labels of a certain method as u* and the set of ground truth labels as u. For each set of labels, we computed the average pairwise Euclidean distance between the cells or spots within each cluster. We denote the density functions of such group-wise average distance as $q(s|u^*)$for the clustering labels u^* and $q(s|u)$ for the clustering labels u, respectively. We computed the total absolute error $IAE(u^*)$ as

$$IAE(u^*) = \int |q(s|u^*) - q(s|u)|ds$$

The accuracy measure VN is defined as

$$VN(u^*) = 1 - 0.5 * IAE(u^*)$$

F-1-score in differential expression: we performed traditional quantitative downstream differential expression analysis over the clustering results in a one-vs-all manner using the marker gene detection method "*Gini*" in Giotto package [21]. We used the marker gene detection results based on the ground truth clustering labels as reference. We compared the marker genes detected based on the integration method vs. those in the reference. Considering the marker gene detection as a binary classification problem, we counted the number of true positives (TP), false positives (FP) and false negatives (FN). We then computed the F-1 score. The larger the F-1 score is, the more accurate the prediction.

$$TP\text{: } true\ positives$$

$$FP: false\ positives$$

$$FN:\ false\ negatives$$

$$\text{F-1} = 2TP / (2TP + FN + FP)$$

Declarations

Ethics approval and consent to participate

All data utilized in this study has been previously published and is publicly available. The research performed in this study conformed to the principles of the Helsinki Declaration

Consent for publication

Not Applicable

Availability of data and materials

All data and code used in this benchmark study can be found in the following link

https://github.com/lanagarmire/ST_benchmark

Competing interests

The authors declare that they have no competing interests

Funding

L.X. Garmire was supported by NIH/NIGMS, R01 LM012373 and R01 LM012907 awarded by NLM, and R01 HD084633 awarded by NICHD. J.Kang's research was supported by NIH grants R01DA048993, R01 MH105561 and R01GM124061.

Author Contributions

LG envisioned this project. YL developed the benchmark framework, simulation design, implemented the computational methods and evaluated the methods' performance. SS helped implement scVI and SNF, and double checked reproducibility. BH and QH helped select the real datasets. ZJ helped implement MOFA+. JK helped develop the spatial clustering evaluation metric. All authors have read and approved the final manuscript.

Acknowledgment

Not Applicable.

Figures legends

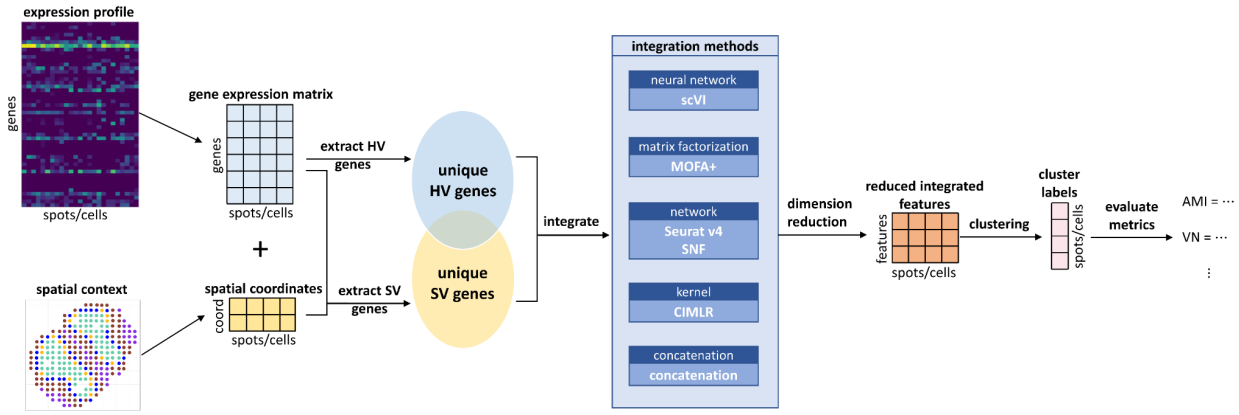
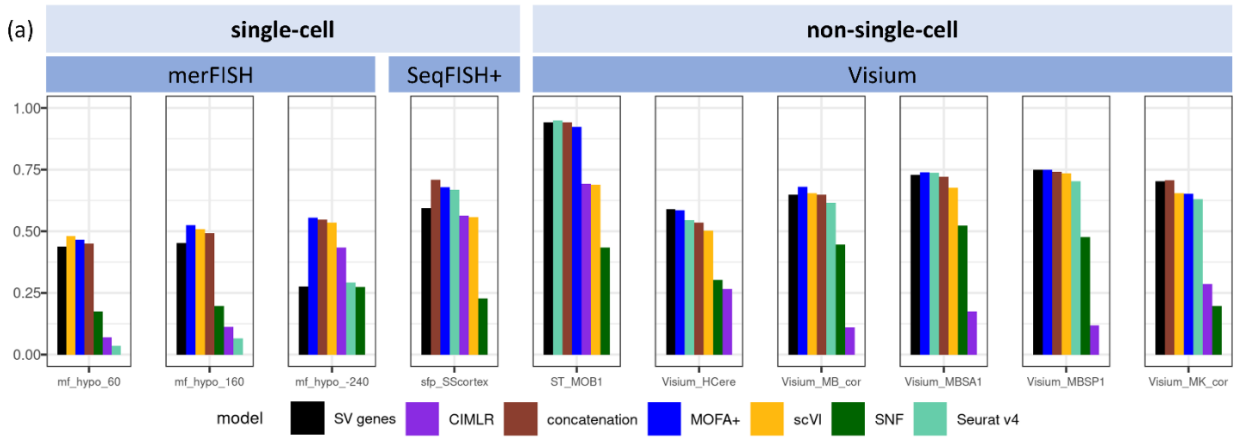
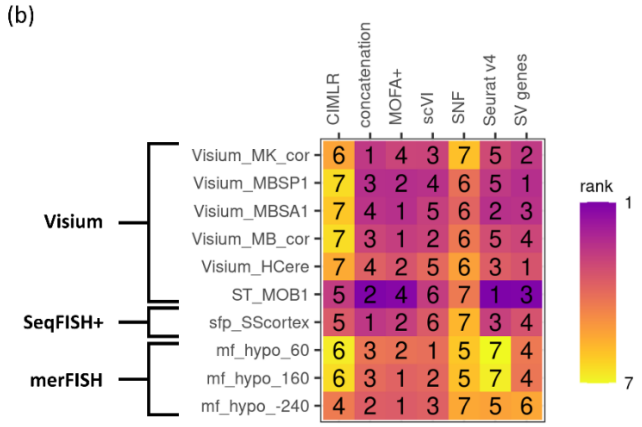
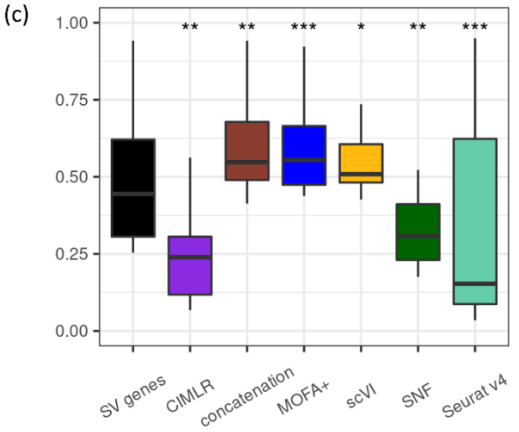
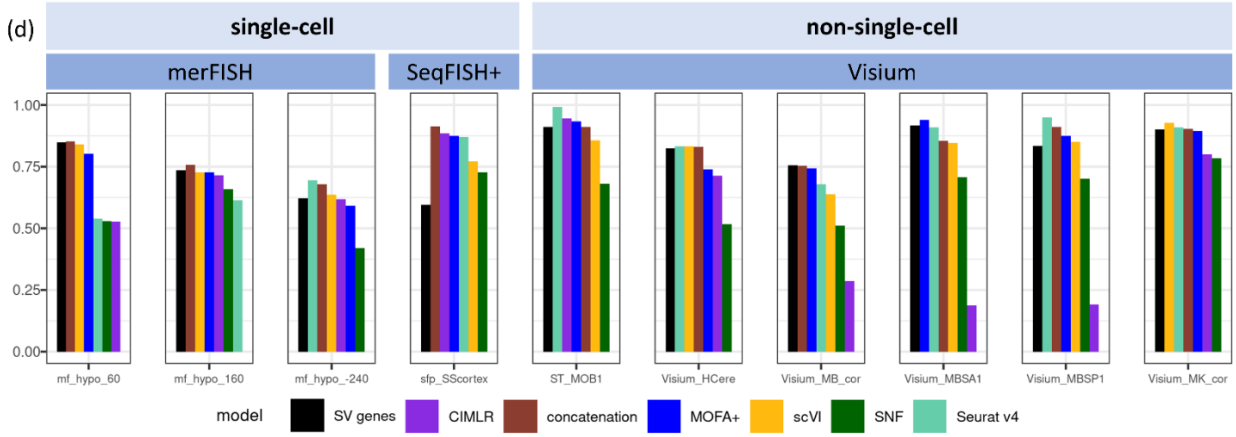


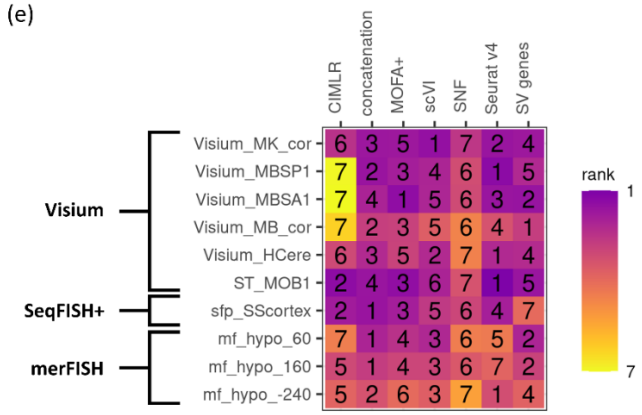
Figure 1. Benchmark workflow. The workflow is composed of three general steps. Step 1: extraction of the HV and SV genes. Step 2: integration and dimension reductions of the HV and SV genes. Step 3: clustering analysis and evaluations based on the integration results.


(a)
single-cell
non-single-cell
merFISH
SeqFISH+
Visium
1.00
0.75
0.50
0.25
0.00
mf_hypo_60
mf_hypo_160
mf_hypo_-240
sfp_SScortex
ST_MOB1
Visium_HCere
Visium_MB_cor
Visium_MBSA1
Visium_MBSP1
Visium_MK_cor
model
SV genes
CIMLR
concatenation
MOFA+
scVI
SNF
Seurat v4


(b)
CIMLR
concatenation
MOFA+
scVI
SNF
Seurat v4
SV genes
Visium
SeqFISH+
merFISH
Visium_MK_cor 6 1 4 3 7 5 2
Visium_MBSP1 7 3 2 4 6 5 1
Visium_MBSA1 7 4 1 5 6 2 3
Visium_MB_cor 7 3 1 2 6 5 4
Visium_HCere 7 4 2 5 6 3 1
ST_MOB1 5 2 4 6 7 1 3
sfp_SScortex 5 1 2 6 7 3 4
mf_hypo_60 6 3 2 1 5 7 4
mf_hypo_160 6 3 1 2 5 7 4
mf_hypo_-240 4 2 1 3 7 5 6
rank
1
7


(c)
1.00
0.75
0.50
0.25
0.00
**
**
***
*
**
***
SV genes
CIMLR
concatenation
MOFA+
scVI
SNF
Seurat v4


(d)
single-cell
non-single-cell
merFISH
SeqFISH+
Visium
1.00
0.75
0.50
0.25
0.00
mf_hypo_60
mf_hypo_160
mf_hypo_-240
sfp_SScortex
ST_MOB1
Visium_HCere
Visium_MB_cor
Visium_MBSA1
Visium_MBSP1
Visium_MK_cor
model
SV genes
CIMLR
concatenation
MOFA+
scVI
SNF
Seurat v4


(e)
CIMLR
concatenation
MOFA+
scVI
SNF
Seurat v4
SV genes
Visium
SeqFISH+
merFISH
Visium_MK_cor 6 3 5 1 7 2 4
Visium_MBSP1 7 2 3 4 6 1 5
Visium_MBSA1 7 4 1 5 6 3 2
Visium_MB_cor 7 2 3 5 6 4 1
Visium_HCere 6 3 5 2 7 1 4
ST_MOB1 2 4 3 6 7 1 5
sfp_SScortex 2 1 3 5 6 4 7
mf_hypo_60 7 1 4 3 6 5 2
mf_hypo_160 5 1 4 3 6 7 2
mf_hypo_-240 5 2 6 3 7 1 4
rank
1
7

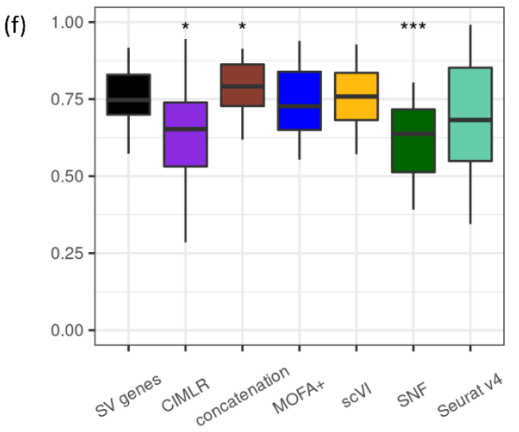

(f)
1.00
0.75
0.50
0.25
0.00
*
*
***
SV genes
CIMLR
concatenation
MOFA+
scVI
SNF
Seurat v4

Figure 2. Performance comparisons of different data integration methods on real spatial transcriptomics datasets, in three representative platforms including merFISH, SeqFISH+, and Visium. (a-c) Results on AMI, shown as (a) barplot of AMI values for three representative datasets by merFISH, the SeqFISH+ dataset and all six Visium datasets; (b) AMI ranking for three representative datasets by merFISH, the SeqFISH+ dataset and all six Visium datasets, and (c) AMI boxplot over all 19 real datasets, including two-sided paired Wilcoxon test results on the null hypothesis that the performance of the integration method is the same as the performance of solely SV genes. *: p-value < 0.05, ** : p-value < 0.01, ***: p-value < 0.001; (d-f) Results on VN, shown as (d) barplot of VN values for three representative datasets by merFISH, the SeqFISH+ dataset and all six Visium datasets; (e) VN ranking for three representative datasets by merFISH, the SeqFISH+ dataset and all six Visium datasets, and (f) VN boxplot over all 19 real datasets, including two-sided paired Wilcox test results on the null hypothesis that the performance of the integration method is the same as the performance of solely SV genes. *: p-value < 0.05, ** : p-value < 0.01, ***: p-value < 0.001.

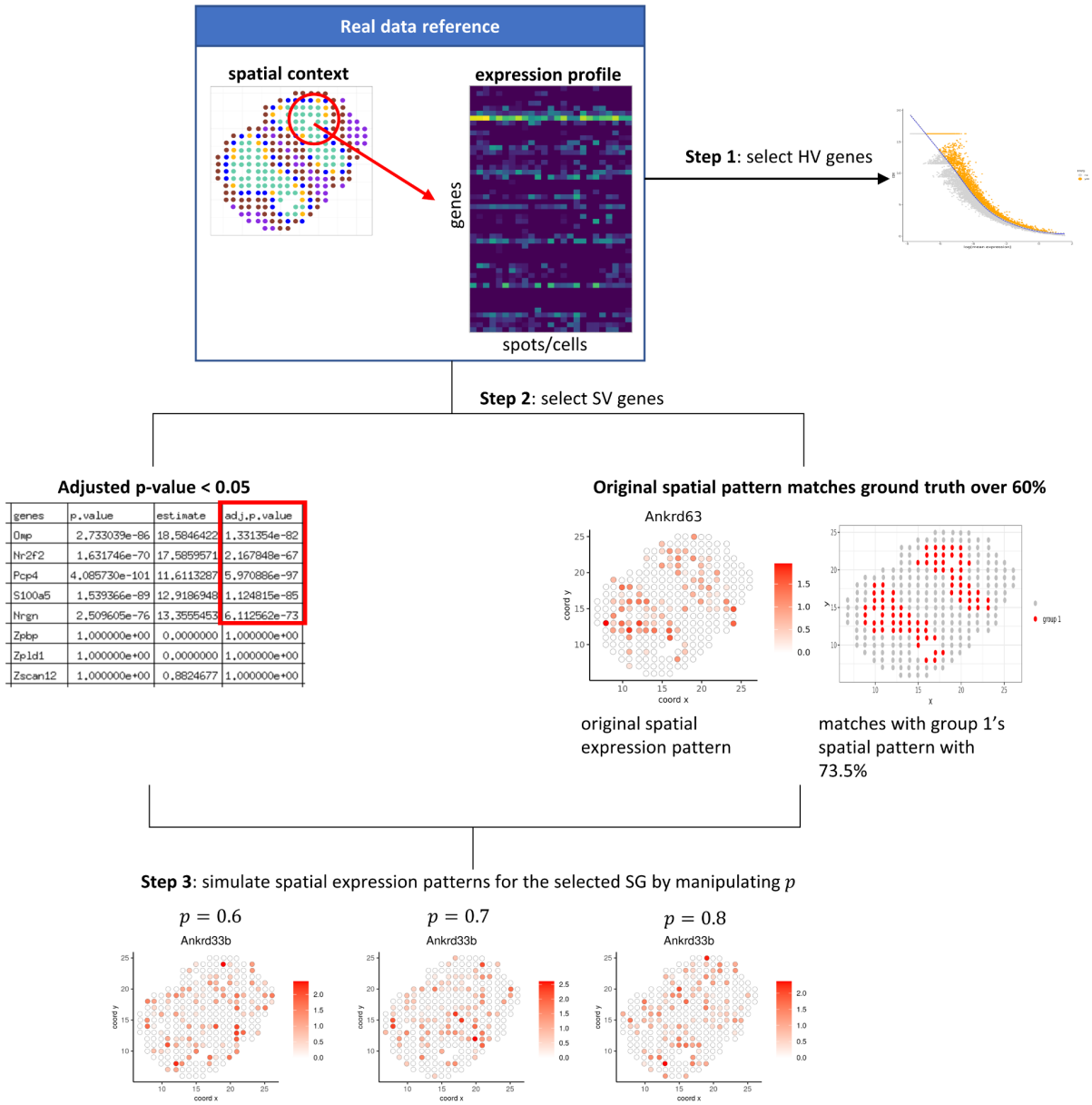


Figure 3. The design of simulation workflow based on the real dataset. The workflow consists of three major steps. Step 1: select HV genes. Step 2: select SV genes whose adjusted p-values are below 0.05 and whose original spatial expression pattern matches at least 60% of a ground truth cluster. Step 3: simulate new spatial expression patterns for the selected SV genes.

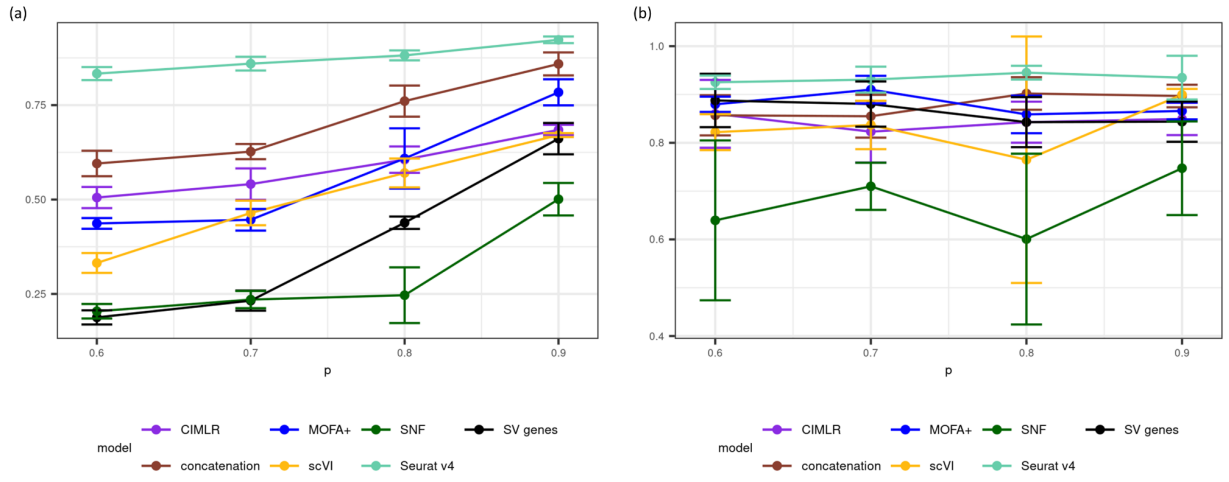


Figure 4. Clustering performance comparisons of different data integration methods on simulated spatial transcriptomics datasets, by varying probability (p) of spatial clustering certainty. Shown are metrics using (a) AMI and (b) VN. Five repetitions are run in each condition, and the average scores are displayed, with standard deviations.

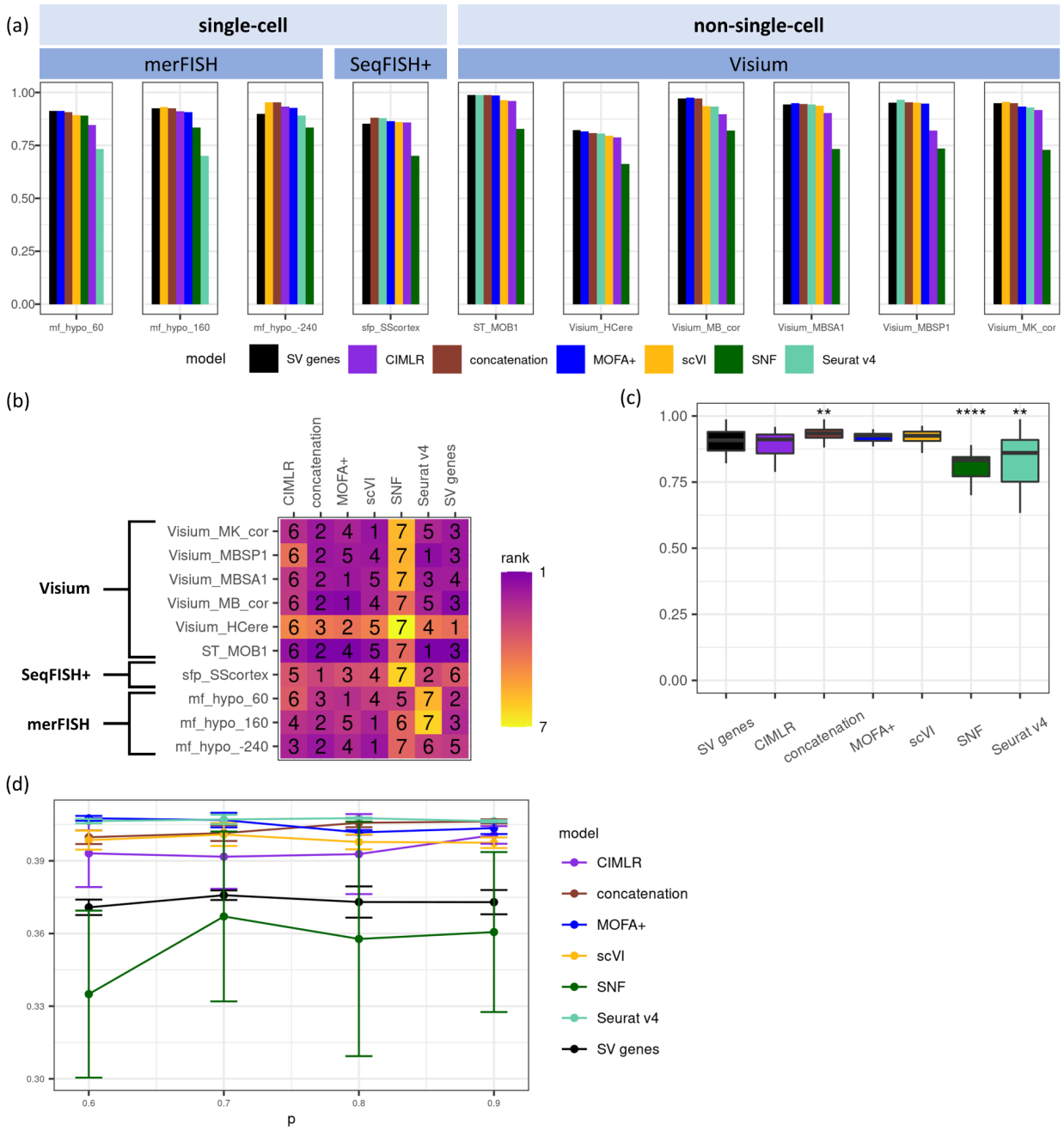


Figure 5. Effect of integration methods on downstream differential expression results, measured by F-1 scores. (a-c) Results on real datasets, in three representative platforms, including merFISH, SeqFISH+, and Visium, shown as (a) barplot of F-1 score values for three representative datasets by merFISH, the SeqFISH+ dataset and all six Visium datasets; (b) F-1 score ranking for three representative datasets by merFISH, the SeqFISH+ dataset and all six Visium datasets; (c) F-1 score boxplot over all 19 real

datasets, including two-sided paired Wilcox test results on the null hypothesis that the performance of the integration method is the same as the performance of solely SV genes. *: p-value < 0.05, ** : p-value < 0.01, ***: p-value < 0.001, ****: p-value < 0.0001; and (d) Results on the same simulation datasets as in Figure 4.

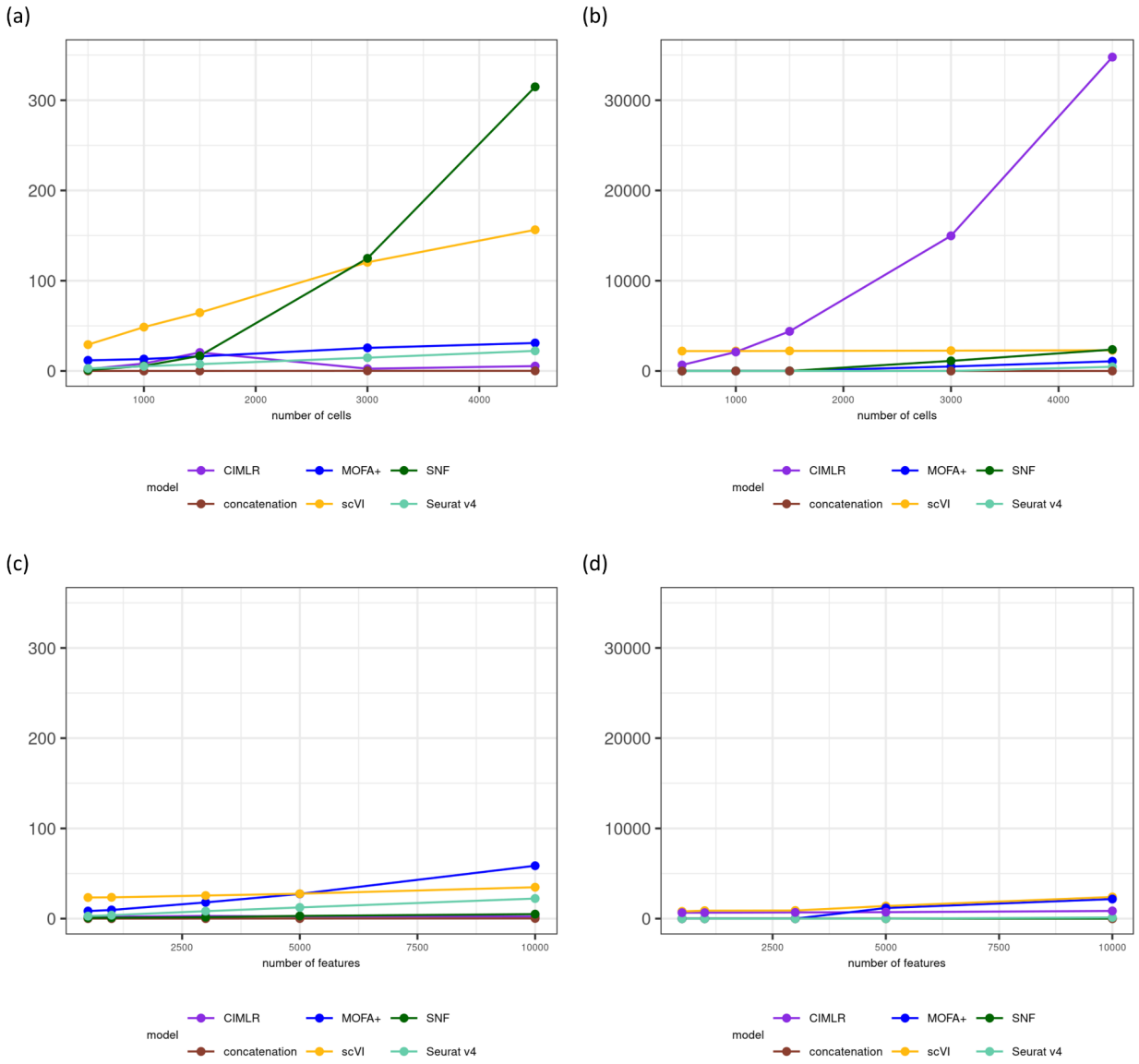


Figure 6. Comparison of running time and memory usage of integration methods. (a) The practical performance relative to the number of cells, measured by (a) running time, and (b) memory usage. (a) The practical performance relative to the number of features , measured by (a) running time, and (b) memory usage.

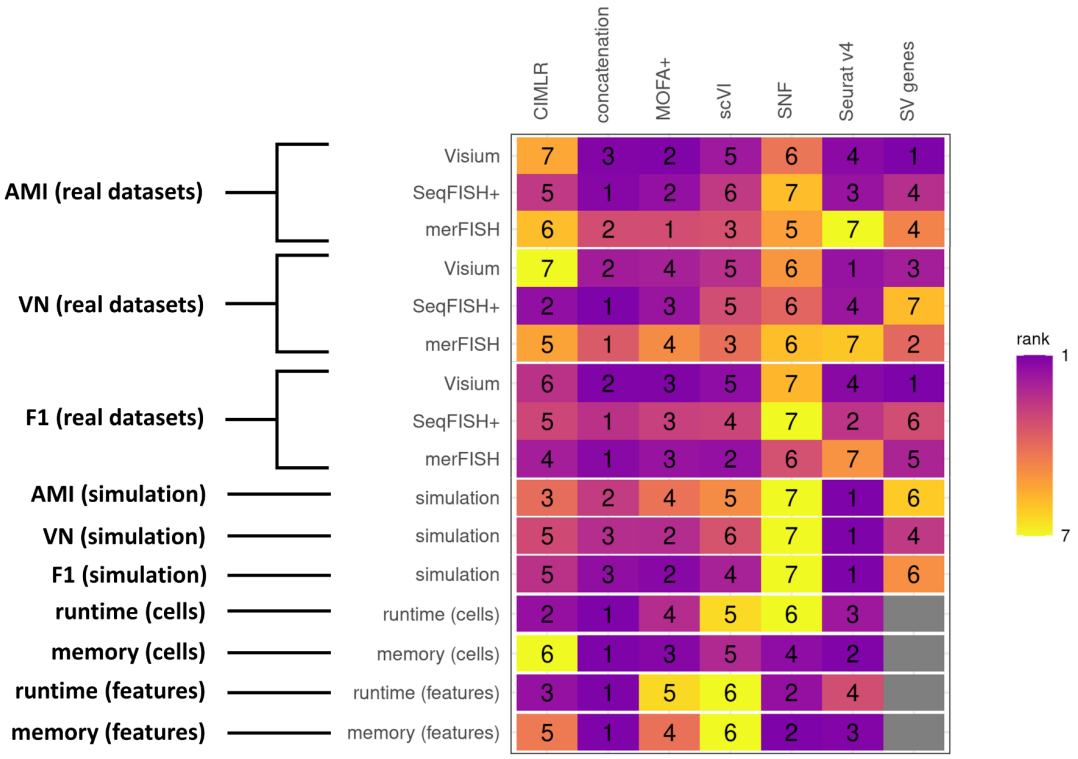
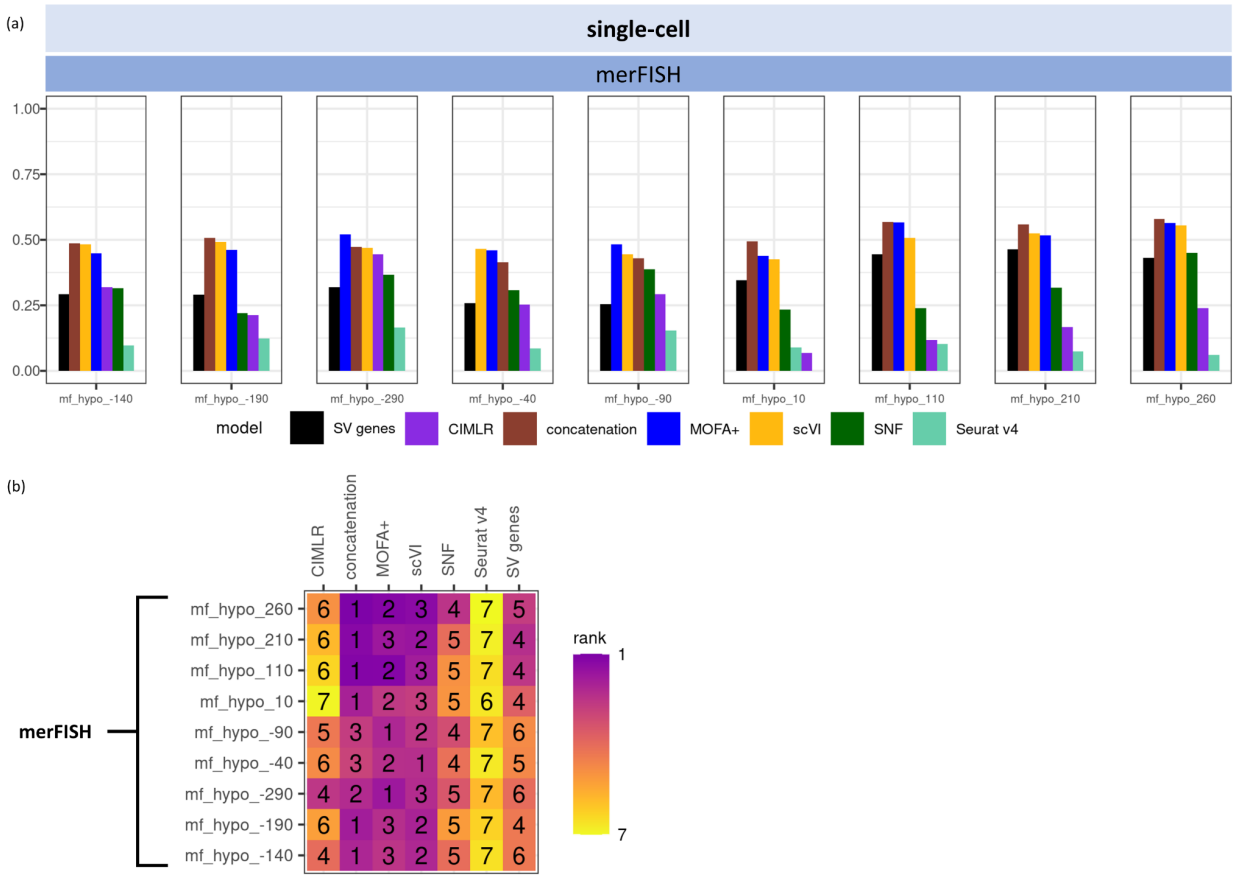
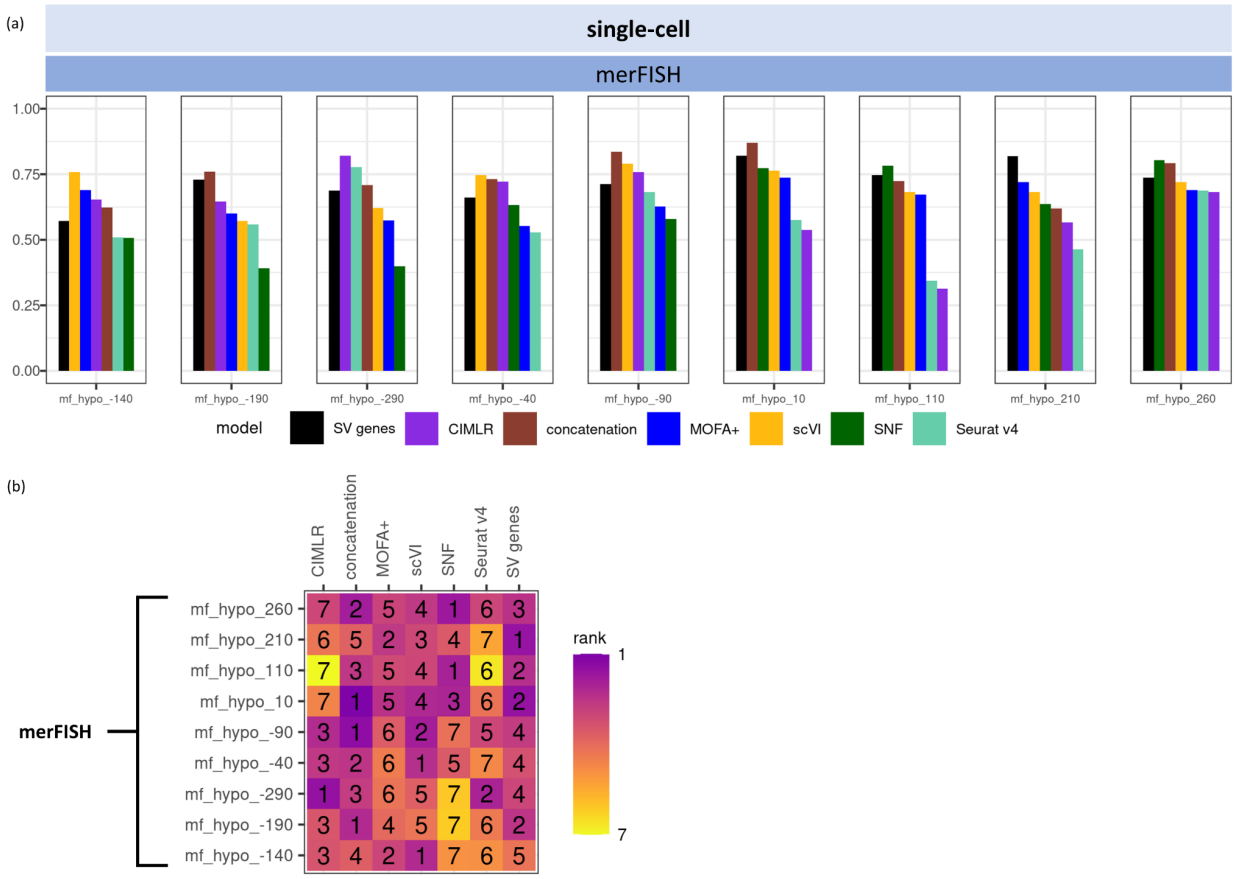


Figure 7. Heatmap summary of metrics ranking of integration methods over real and simulated datasets, by AMI, VN, F-1 statistics of downstream differential expression analysis, runtime and memory usage.

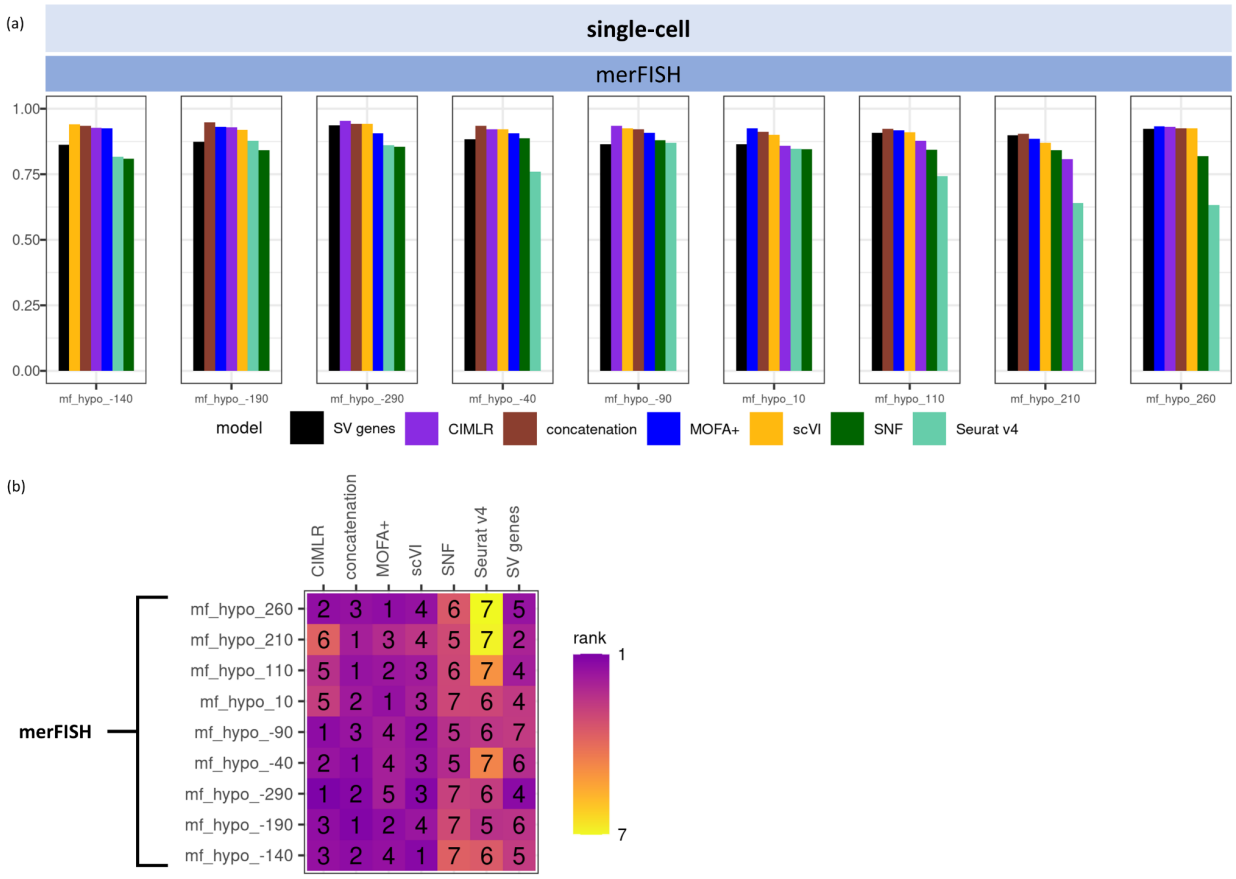
Supplementary materials



Supplementary Figure 1. (a) barplot of AMI values for the remaining nine datasets by merFISH; (b) AMI rankings for the remaining nine datasets by merFISH.



Supplementary Figure 2. (a) barplot of VN values for the remaining nine datasets by merFISH; (b) VN rankings for the remaining nine datasets by merFISH.



Supplementary Figure 3. (a) barplot of F-1 score values for the remaining nine datasets by merFISH; (b) F-1 score rankings for the remaining nine datasets by merFISH.

Table 1: Summary of integration methods in comparison.

Method	**Working principle**	**Existence of Internal dimension reduction step**	**Required input data**
Concatenation	Straightforward concatenation	no	normalized

CIMLR	Kernel-based method	yes	normalized
MOFA+	Matrix Factorization	yes	normalized
scVI	Variational Autoencoder	yes	raw
SNF	Similarity network based fusion method	NA	normalized
Seurat v4	Weighted Nearest Neighbor	yes	raw

Table 2: Summary of datasets and their spatial transcriptomics platforms.

Study	**Genes utilized in ground truth labelling**	**Technology platform**	**Resolution**	**Number of cells (or spots)**	**Number of features**
Mouse Olfactory Bulb [18]	All	Visium	non-single-cell	265	16573
Mouse Somatosensory Cortex [14]	highly variable genes	SeqFISH+	single-cell	523	10000
Human Cerebellum [31]	All	Visium	non-single	4992	23808

			-cell		
Mouse Kidney Coronal Section [32]	All	Visium	non-single -cell	1438	20100
Mouse Brain Coronal Section [33]	All	Visium	non-single -cell	2702	21949
Mouse Brain Sagittal Anterior Section [34]	All	Visium	non-single -cell	2695	21334
Mouse Brain Sagittal Posterior Section [35]	All	Visium	non-single -cell	3355	21363
Mouse Hypothalamus (slice -240) [10]	All	merFISH	single-cell	5543	157
Mouse Hypothalamus (slice -190) [10]	All	merFISH	single-cell	5803	156
Mouse Hypothalamus (slice -140) [10]	All	merFISH	single-cell	5926	157
Mouse Hypothalamus (slice 60) [10]	All	merFISH	single-cell	5343	157

Mouse Hypothalamus (slice -40) [10]	All	merFISH	single-cell	5488	158
Mouse Hypothalamus (slice 110) [10]	All	merFISH	single-cell	5070	156
Mouse Hypothalamus (slice -90) [10]	All	merFISH	single-cell	5557	158
Mouse Hypothalamus (slice 10) [10]	All	merFISH	single-cell	5338	157
Mouse Hypothalamus (slice -290) [10]	All	merFISH	single-cell	5517	156
Mouse Hypothalamus (slice 260) [10]	All	merFISH	single-cell	4832	153
Mouse Hypothalamus (slice 210) [10]	All	merFISH	single-cell	4787	153
Mouse Hypothalamus (slice 160) [10]	All	merFISH	single-cell	5169	155

References

1. Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. Nature. 2021;596:211–20.

2. Liu Y, Yang M, Deng Y, Su G, Enninful A, Guo CC, et al. High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. Cell. 2020;183:1665–81.e18.

3. Deng Y, Bartosovic M, Ma S, Zhang D, Liu Y, Qin X. Spatial-ATAC-seq: spatially resolved chromatin accessibility profiling of tissues at genome scale and cellular level. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.06.06.447244v1.abstract

4. Deng Y, Zhang D, Liu Y, Su G, Enninful A, Bai Z, et al. Spatial epigenome sequencing at tissue scale and cellular level. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.03.11.434985v1.abstract

5. Lee J-K, Wang J, Sa JK, Ladewig E, Lee H-O, Lee I-H, et al. Spatiotemporal genomic architecture informs precision oncology in glioblastoma. Nat Genet. 2017;49:594–9.

6. Hunter MV, Moncada R, Weiss JM, Yanai I, White RM. Spatial transcriptomics reveals the architecture of the tumor/microenvironment interface. bioRxiv [Internet]. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.11.05.368753v1.abstract

7. Moncada R, Barkley D, Wagner F, Chiodin M, Devlin JC, Baron M, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat Biotechnol. Nature Publishing Group; 2020;38:333–42.

8. Takei Y, Yun J, Zheng S, Ollikainen N, Pierson N, White J, et al. Integrated spatial genomics reveals global architecture of single nuclei. Nature. 2021;590:344–50.

9. Ravasio A, Myaing MZ, Chia S, Arora A, Sathe A, Cao EY, et al. Single-cell analysis of EphA clustering phenotypes to probe cancer cell heterogeneity. Communications Biology. Nature Publishing Group; 2020;3:1–12.

10. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science [Internet]. 2018;362. Available from: http://dx.doi.org/10.1126/science.aau5324

11. Xia C, Fan J, Emanuel G, Hao J, Zhuang X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proc Natl Acad Sci U S A. 2019;116:19490–9.

12. Wang G, Ang CE, Fan J, Wang A, Moffitt JR, Zhuang X. Spatial organization of the transcriptome in individual neurons. bioRxiv [Internet]. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.12.07.414060v1.abstract

13. Zhu Q, Shah S, Dries R, Cai L, Yuan G-C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nat Biotechnol [Internet]. 2018; Available from: http://dx.doi.org/10.1038/nbt.4260

14. Eng C-HL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature. Nature Publishing Group; 2019;568:235–9.

15. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, et al. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell. 2020;182:1661–2.

16. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353:78–82.

17. 10x G. Spatial Transcriptomics [Internet]. Available from: https://www.10xgenomics.com/spatial-transcriptomics

18. Edsgärd D, Johnsson P, Sandberg R. Identification of spatial expression trends in single-cell gene expression data. Nat Methods. 2018;15:339–42.

19. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nat Methods. 2018;15:343–6.

20. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods. 2020;17:193–200.

21. Dries R, Zhu Q, Dong R, Eng C-HL, Li H, Liu K, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biol. 2021;22:78.

22. Miller BF, Bambah-Mukku D, Dulac C, Zhuang X, Fan J. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomics data with nonuniform cellular densities. Genome Res [Internet]. 2021; Available from: http://dx.doi.org/10.1101/gr.271288.120

23. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053–8.

24. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573–87.e29.

25. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21:111.

26. Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat Commun. 2018;9:4453.

27. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11:333–7.

28. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell. 2019;177:1873–87.e17.

29. Welch JD, Hartemink AJ, Prins JF. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol. 2017;18:138.

30. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.

31. 10x G. Human Cerebellum: Whole Transcriptome Analysis, Spatial Gene Expression Dataset by Space Ranger 1.2.0 [Internet]. 2020, October 27. Available from: https://support.10xgenomics.com/spatial-gene-expression/datasets/1.2.0/Parent_Visium_Human_Cerebellum

32. 10x G. Mouse Kidney Section (Coronal), Spatial Gene Expression Dataset by Space Ranger 1.1.0 [Internet]. 2020, June 23. Available from: https://support.10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Mouse_Kidney

33. 10x G. Mouse Brain Section (Coronal), Spatial Gene Expression Dataset by Space Ranger 1.1.0 [Internet]. 2020, June 23. Available from: https://support.10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Adult_Mouse_Brain

34. 10x G. Mouse Brain Serial Section 1 (Sagittal-Anterior), Spatial Gene Expression Dataset by Space Ranger 1.1.0 [Internet]. 2020, June 23. Available from: https://support.10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Mouse_Brain_Sagittal_Anterior

35. 10x G. Mouse Brain Serial Section 1 (Sagittal-Posterior), Spatial Gene Expression Dataset by Space Ranger 1.1.0 [Internet]. 2020, June 23. Available from: https://support.10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Mouse_Brain_Sagittal_Posterior

36. Zhu X, Wolfgruber T, Tasato A, Garmire DG, Garmire LX. Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists [Internet]. bioRxiv. 2017 [cited 2021 Aug 12]. p. 110759. Available from: https://www.biorxiv.org/content/10.1101/110759v4

37. Garmire D, Zhu X, Mantravadi A, Huang Q, Yunits B, Liu Y, et al. GranatumX: A community engaging, modularized and flexible software environment for single-cell analysis [Internet]. bioRxiv. 2021 [cited 2021 Aug 12]. p. 385591. Available from: https://www.biorxiv.org/content/10.1101/385591v3

38. Poirion OB, Zhu X, Ching T, Garmire L. Single-Cell Transcriptomics Bioinformatics and Computational Challenges. Front Genet. 2016;7:163.

39. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14:e8124.

40. Kriebel AR, Welch JD. Nonnegative matrix factorization integrates single-cell multi-omic datasets with partially overlapping features [Internet]. bioRxiv. 2021 [cited 2021 Aug 12]. p. 2021.04.09.439160. Available from: https://www.biorxiv.org/content/10.1101/2021.04.09.439160v1.full

41. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. Nature Publishing Group; 2019;9:1–12.

42. Delaunay B, Others. Sur la sphere vide. Izv Akad Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk. 1934;7:1–2.

43. Nguyen XV, Epps J, Bailey J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? [Internet]. 2009 [cited 2021 Aug 6]. Available from: https://openreview.net/forum?id=rJVTUsbObS

44. Wand MP, Ormerod JT, Padoan SA, Frühwirth R. Mean Field Variational Bayes for Elaborate Distributions. ba. International Society for Bayesian Analysis; 2011;6:847–900.