OXFORD UNIVERSITY PRESS | Briefings in Functional Genomics

A Comprehensive Survey of Dimensionality Reduction and Clustering Methods for Single-cell and Spatial Transcriptomics Data

Journal:	*Briefings in Functional Genomics*
Manuscript ID	BFGP-23-0224.R1
Manuscript Type:	Review Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Sun, Yidi; Hainan University, Kong, Lingling; Hainan University Huang, Jiayi; Hainan University Deng, Hongyan; Hainan University, Bian, Xinling; Hainan University Li, Xingfeng; Hainan University Cui, Feifei; Hainan University, Dou, Lijun; Cleveland Clinic Lerner Research Institute Cao, Chen; Nanjing Medical University, Zou, Quan; School of Computer Science and Technology, Tianjin University, Zhang, Zilong; Hainan University, School of Computer Science and Technology
Keywords:	ScRNA-seq, Spatial transcriptomics, Dimensionality reduction, Clustering

A Comprehensive Survey of Dimensionality Reduction and Clustering Methods for Single-cell and Spatial Transcriptomics Data

Yidi Sun[1], Lingling Kong[1], Jiayi Huang[1], Hongyan Deng[1], Xinling Bian[1], Xingfeng Li[1], Feifei Cui[1], Lijun Dou[2], Chen Cao[3], Quan Zou[4, 5, *], Zilong Zhang[1, *]

[1] School of Computer Science and Technology, Hainan University, Haikou 570228, China

[2] Genomic Medicine Institute, Lerner Research Institute, Cleveland, OH 44106, USA

[3] School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 210029, China

[4] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China

[5] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, China

Abstract

In recent years, the application of single-cell transcriptomics and spatial transcriptomics analysis techniques has become increasingly widespread. Whether dealing with single-cell transcriptomic or spatial transcriptomic data, dimensionality reduction and clustering are indispensable. Both single-cell and spatial transcriptomic data are often high-dimensional, making the analysis and visualization of such data challenging. Through dimensionality reduction, it becomes possible to visualize the data in a lower-dimensional space, allowing for the observation of relationships and differences between cell subpopulations. Clustering enables the grouping of similar cells into the same cluster, aiding in the identification of distinct cell subpopulations and revealing cellular diversity, providing guidance for downstream analyses. In this review, we systematically summarized the most widely recognized algorithms employed for the dimensionality reduction and clustering analysis of single-cell transcriptomic and spatial transcriptomic data. This endeavor provides valuable insights and ideas that can contribute to the development of novel tools in this rapidly evolving field.

Key Words

ScRNA-seq; Spatial transcriptomics; Dimensionality reduction; Clustering

Key Points

- We summarized the principles of dimensionality reduction and clustering methods, with a focus on their applications in the analysis of single-cell and spatial transcriptomics data.
- We provided a comprehensive comparison of various methods, highlighting their strengths, weaknesses, and applicability across different data types and research contexts.
- We offered best practices for selecting and applying these methods, providing practical guidance for researchers.
- We provided insights into future directions for the development of dimensionality reduction and clustering algorithms for single-cell and spatial transcriptomics data.

Introduction

In recent years, single-cell transcriptomics and spatial transcriptomics have emerged as focal points of research interest. In practice, a single scRNA-seq experiment captures gene expression data from tens of thousands of individual cells [1]. However, dealing with such high-dimensional data poses challenges for subsequent analysis [2, 3]. Consequently, the adoption of dimensionality reduction techniques and clustering methods has become an indispensable strategy in transcriptomic analysis for noise reduction and enhancing data visualization[4, 5]. Dimensionality reduction and clustering help eliminate noise and redundancy in the original data [6, 7], and it aids in mitigating the impact of "dropout" events [8, 9], thereby facilitating subsequent steps in data visualization and clustering [10]. Furthermore, clustering plays a crucial role in uncovering cellular heterogeneity and exploring underlying biological mechanisms, holding significant biological implications [11-15]. Based on the aforementioned necessity, numerous methods for dimensionality reduction and clustering have been developed, tailored to different transcriptomic datasets.

It is noteworthy that each algorithm has its own strengths and limitations [16], leading to distinct characteristics and performance features among different algorithms [17, 18]. Due to the typically elevated levels of noise in spatial transcriptomics data [19, 20], many dimensionality reduction and clustering methods initially designed for single-cell RNA sequencing (scRNA-seq) data may not be directly applicable to the field of spatial transcriptomics [21, 22]. Consequently, researchers have recently begun developing specialized dimensionality reduction and clustering techniques tailored specifically to address the unique challenges posed by spatial transcriptomics.

This review aims to comprehensively examine the existing dimensionality reduction and clustering techniques applicable to both scRNA-seq and spatial transcriptomics datasets. It encompasses a comprehensive overview of dimensionality reduction and clustering methodologies for both single-cell and spatial transcriptomics, along with insights into their respective contexts of applicability. Furthermore, it offers valuable recommendations to guide researchers in selecting the most suitable dimensionality reduction and clustering methodologies tailored to the unique demands of scRNA-seq and spatial transcriptomics data. Our review aims to serve as a practical resource, equipping researchers with the necessary knowledge and references to make informed choices when working with scRNA-seq and spatial transcriptomics datasets, catering to their specific research requirements.

Overall Structure of the Article

This review is structured to provide an overview of dimensionality reduction methods for single-cell transcriptomics and spatial transcriptomics, followed by clustering methods for single-cell transcriptomics and spatial transcriptomics. The dimensionality reduction and clustering methods for single-cell and spatial transcriptomics data throughout the entire manuscript are broadly outlined in Figure 1, following a general analysis workflow.

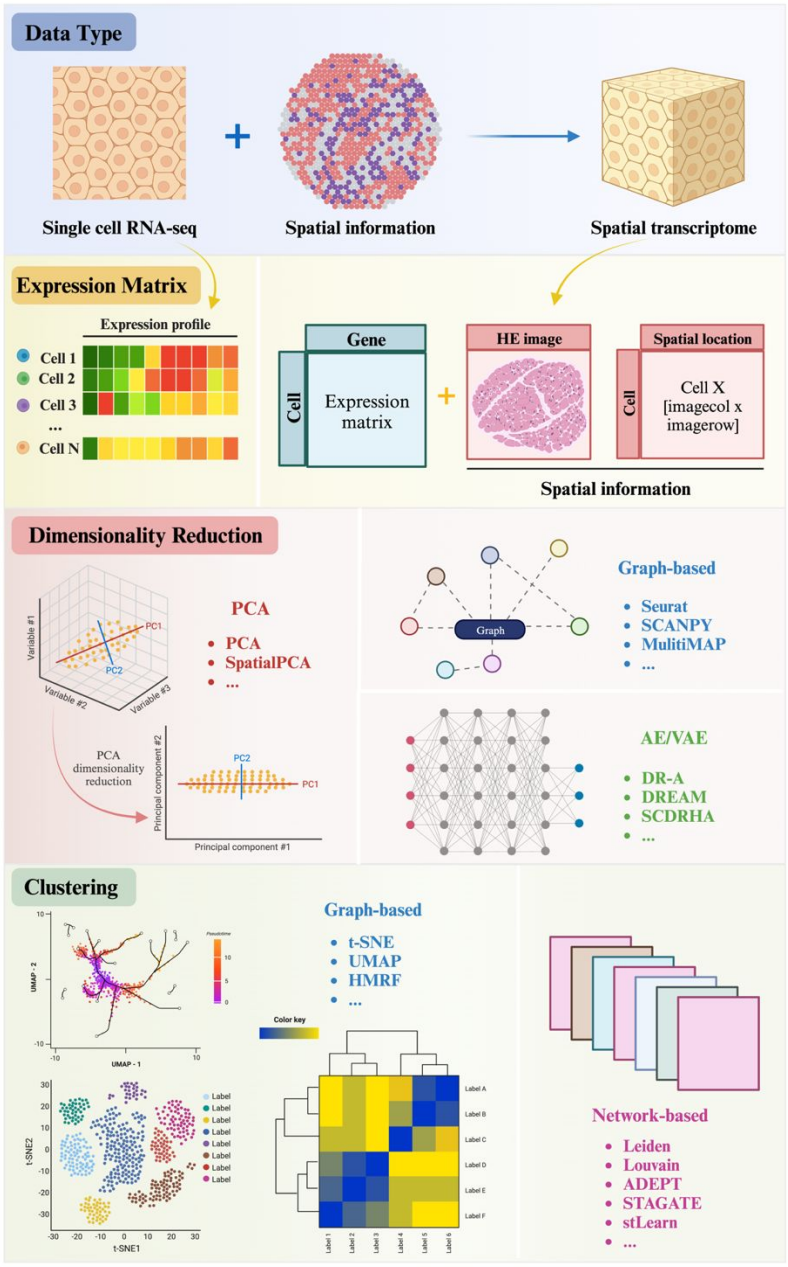


Figure 1. Dimensionality reduction and clustering methods for single-cell transcriptomic and spatial transcriptomic data.

Dimensionality Reduction

Single-cell Transcriptomics

Principal component analysis (PCA) [23] is one of the most classical dimensionality reduction methods, seeking principal components to explain the maximum variance in the data. Non-negative matrix factorization (NMF) [24] methods achieve dimensionality reduction through learning non-negative constrained representations based on parts. However, the above two classical methods can only handle linear data. Given the high noise and complex structure in single-cell data, there have emerged some methods capable of nonlinear dimensionality reduction.

Ding et al. introduced the scvis dimensionality reduction method specifically tailored for single-cell transcriptome data [25]. scvis employs a probabilistic generative model to achieve dimensionality reduction by learning a parameter mapping from the high-dimensional space to the low-dimensional space. Scvis is based on a probabilistic model, yielding low-dimensional embeddings that are more interpretable while simultaneously quantifying the uncertainty associated with each mapping. Furthermore, it provides a measure of the likelihood of uncertainty for each mapping. Simulation results presented by Ding et al. showcase that scvis's dimensionality reduction outcomes effectively retain both global and local structures present in the original data. Even when working with smaller dataset sizes, scvis strives to preserve the underlying data structure to the greatest extent possible. This characteristic enhances the interpretability of scvis's dimensionality reduction results and provides more informative reference points for downstream analysis in the context of single-cell transcriptomics after dimensionality reduction.

The t-SNE algorithm, introduced by Maaten et al. in 2008 [26], has gained widespread acceptance and utilization. t-SNE, a variant of stochastic neighbor embedding (SNE), elegantly mitigates the Crowding Problem by replacing SNE's Gaussian distribution with Student’s t-distribution in low-dimensional embedding. It also tackles the optimization challenges of SNE's cost function by leveraging symmetric SNE. Key facets of t-SNE encompass: (1) Effective modeling of dissimilar data points with large pairwise distances. (2) Accurate modeling of similar data points with small pairwise distances. t-SNE exhibits remarkable proficiency in identifying local variations, endowing it with robust data structure representation capabilities within low-dimensional spaces. Nonetheless, the t-SNE algorithm may not consistently preserve global variations. To mitigate these limitations, researchers have introduced

several variant algorithms based on t-SNE in recent years. In 2023, Zhou et al. introduced the global t-SNE (g-SNE) [27], a technique that strikes a balance between preserving global and local variations through the introduction of a weight parameter λ. Compared to conventional t-SNE, g-SNE is adept at retaining a broader spectrum of global variations. The optimal selection of λ is contingent solely upon the intrinsic data structure itself. However, it's worth noting that the accuracy of the g-SNE algorithm was predominantly validated on smaller datasets, necessitating further verification of its dimensionality reduction capabilities when applied to larger-scale datasets. George et al. presented the FFT-accelerated Interpolation-based t-SNE (FIt-SNE) [28], an algorithm that expeditiously computes one- and two-dimensional t-SNE projections using polynomial interpolation, further accelerating the process through the utilization of the fast Fourier transform. FIt-SNE exhibits superior computational speed compared to traditional t-SNE, rendering it particularly well-suited for processing large datasets. Antti et al. developed qSNE [29], a high-speed t-SNE package that leverages a quasi-Newton optimizer to achieve quadratic convergence and automatic optimization of perplexity. Their findings demonstrate that these enhancements translate into qSNE being notably faster than conventional t-SNE packages. Additionally, it facilitates comprehensive analysis of extensive datasets, such as mass cytometry data, without the need for downsampling.

Compared to t-SNE, UMAP excels in preserving a broader range of global variations [30]. UMAP also boasts a faster computation speed when juxtaposed with the aforementioned scvis algorithm, and it is on par in terms of speed with the Flt-SNE algorithm. Becht et al. conducted an exhaustive and quantitative assessment, scrutinizing all these qualitative aspects. UMAP emerges as a robust and versatile choice for dimensionality reduction in single-cell techniques [31].

In a separate study, Archit et al. introduced the t-distributed gaussian process latent variable model (tGPLVM), a bayesian nonparametric model designed specifically for raw and unfiltered single-cell data [32]. tGPLVM possesses the capability to estimate informative manifolds from raw scRNA-seq data, utilizing unique molecular identifier (UMI) counts at the single-cell level. This attribute renders the resultant dimensionality reduction outcomes versatile for a range of downstream analyses in single-cell transcriptomic data, such as clustering and pseudotime analysis.

Eugene et al. introduced a novel dimensionality reduction method for scRNA-seq data in 2020, known as dimensionality reduction with adversarial autoencoder (DR-A)[33]. DR-A operates within the generative adversarial network (GAN) framework,

amalgamating the strengths of both adversarial autoencoders and variational autoencoders, thus harnessing the advantages of deep learning models. A significant enhancement in the DR-A framework compared to traditional GANs is the inclusion of an additional discriminator: The initial discriminator network, denoted as D1, is trained to accurately differentiate between samples originating from the sampled distribution and the latent distribution of the autoencoder. Subsequently, the secondary discriminator, labeled as D2, is trained to effectively classify whether the scRNA-seq data is genuine or synthetic. The efficacy of DR-A in dimensionality reduction exhibited further improvements with increasing dataset sizes. Additionally, Eugene et al. illustrated that the combination of DR-A for dimensionality reduction and the t-SNE algorithm for clustering yielded the most favorable clustering outcomes. DR-A can be employed as a complementary tool in conjunction with other algorithms to achieve even greater performance.

Zhao et al. introduced the single-cell dimensionality reduction using hierarchical autoencoder (SCDRHA) algorithm for scRNA-seq data [34]. The SCDRHA pipeline comprises two pivotal modules. The first module, known as the deep count autoencoder (DCA), serves the purpose of noise reduction within the data. The second module, the graph attention network (GAT), is responsible for projecting high-dimensional data into a lower-dimensional space. DCA, a deep learning approach rooted in unsupervised autoencoders, is constructed upon the zero-inflated negative binomial (ZINB) model. This framework adeptly captures the distinctive characteristics of single-cell transcriptomic data, especially in cases where dropout events are prevalent. Conversely, GAT represents a novel neural network framework designed to preserve the topological relationships within the latent space, with the goal of retaining cellular topology to the greatest extent possible during the dimensionality reduction process. With this combined structure, the SCDRHA algorithm significantly enhances the performance of data visualization and clustering.

The MultiMAP algorithm serves as a powerful tool for dimensionality reduction and the integration of single-cell transcriptomic datasets characterized by varying numbers and dimensions [35]. MultiMAP builds upon the principles of Riemannian geometry and algebraic topology to extend the UMAP framework, making it adept at handling diverse datasets with varying dimensions. It can seamlessly accommodate any number of datasets exhibiting different dimensions, effectively capturing geodesic distances across a unified latent manifold where data is uniformly distributed. Within MultiMAP, distances between data points within the same dataset are normalized using

dataset-specific neighborhood distances, while distances between data points originating from distinct datasets are normalized based on a shared feature space and its corresponding neighborhood parameter. Leveraging these distances, MultiMAP constructs a neighborhood graph known as the MultiGraph on the manifold. Ultimately, both the data and the manifold space are projected into a lower-dimensional embedding space through the minimization of the cross-entropy loss between the graph in the embedding space and the graph in the manifold space. Users retain the flexibility to fine-tune the weight of each dataset within the cross-entropy loss, thereby influencing the contribution of each dataset to the resulting layout. Furthermore, MultiMAP can streamline the integrated analysis of single-cell expression and chromatin accessibility profiles along a temporal trajectory, enabling researchers to delve into dynamic chromatin regulation in tandem with gene expression.

In a study conducted by Jing et al. [36], the authors introduced the DREAM algorithm tailored for dimensionality reduction and visualization of scRNA-seq data. The DREAM algorithm seamlessly combines variational autoencoders with gaussian mixture models, while introducing a zero-inflated layer to effectively address dropout events. The dimensionality reduction capabilities of the DREAM algorithm were meticulously validated across nine datasets, with comparisons made using various metrics. Remarkably, the DREAM algorithm consistently outperformed its counterparts in terms of average metrics. Furthermore, Jing et al. demonstrated the DREAM algorithm's remarkable ability to accurately capture the dynamic temporal sequence of human early embryonic development in vitro.

Main Takeaways

In the overview of single-cell transcriptomic dimensionality reduction methods above, PCA and NMF are both based on linear data, scvis and SCDRHA are rooted in neural network models. Scvis, a probabilistic generative model, incorporates specific assumptions and inference processes. This probabilistic generative model enhances the interpretability of the acquired low-dimensional embeddings and facilitates the quantification of uncertainty for each mapping, thus augmenting the algorithm's adaptability. The SCDRHA algorithm seamlessly combines two modules, DCA and GAT. DCA, grounded in ZINB, proves particularly well-suited for data with a high prevalence of 'dropout' events, while GAT, designed within a neural network framework, effectively preserves cellular topological structures. The tGPLVM algorithm adeptly processes raw data from single-cell transcriptomics without any preprocessing steps. The DR-A algorithm amalgamates the framework of variational

autoencoders and adversarial autoencoders, with its dimensionality reduction capabilities scaling with larger datasets, rendering it suitable for handling extensive datasets. The MultiMAP algorithm exhibits remarkable versatility in handling datasets characterized by varying quantities and dimensions. The DREAM algorithm fuses variational autoencoders, Gaussian mixture models, and ZINB, and compared to other algorithms, it showcases the most stable dimensionality reduction performance. While the t-SNE algorithm excels in identifying local variables, the UMAP algorithm excels in preserving global variables and generating superior visualization results following dimensionality reduction.

Spatial Transcriptomics

The Seurat algorithm, developed by Satija et al., serves as a valuable tool for dimensionality reduction and clustering analysis within the domain of spatial transcriptomics data [37]. The input required for the Seurat algorithm comprises single-cell expression profiles and a spatial reference map detailing gene expression for a select set of specific genes. Users partition the tissue of interest into distinct spatial domains, often referred to as 'bins,' each characterized by its unique geometry and size. Across the entire map, landmark genes are classified as either 'on' or 'off' within each bin, and Seurat leverages the single-cell expression levels of these genes to infer the probable origin bins of individual cells. The input data for the Seurat algorithm comprises two components. Firstly, it includes dissociated single-cell RNA sequencing data, where spatial context information is typically lost during the dissociation process. Secondly, it encompasses in situ hybridization patterns for a series of genes. Seurat constructs a gene expression model for each landmark gene by harnessing the variability of other genes within the dataset, thereby reducing reliance on a single measurement and minimizing the impact of technical errors. Subsequently, Seurat establishes statistical models of gene expression for each bin, bridging the bimodal expression patterns derived from RNA-seq estimates with the binarized in situ data. Ultimately, Seurat leverages these models to infer the initial spatial position of cells, assigning a posterior probability of origin to each bin. Seurat possesses the capacity to either exclusively map to a single bin or assign probabilities to multiple bins, depending on specific scenarios. Furthermore, Seurat can be seamlessly combined with sequencing data from diverse instruments, aiding in the determination of cell developmental states or disease phenotypes. This integration allows for the amalgamation of Seurat with existing experimental results, facilitating a more comprehensive comprehension of cell fate.

SCANPY is an extensible toolkit developed in python for handling large-scale single-cell data[38], providing functionalities for data processing, visualization, clustering, pseudotime analysis, and trajectory inference. The strength of SCANPY lies in its highly modular design, making it easily extensible and maintainable by the community. Within the community, results obtained with different tools can be seamlessly transferred, as SCANPY's data storage formats and objects are language-independent and cross-platform.

Shang et al. introduced a dedicated dimensionality reduction technique tailored for spatial transcriptomics, known as SpatialPCA [21]. SpatialPCA excels in reducing the dimensionality of spatial transcriptomic data while simultaneously preserving critical biological signals and spatial structures. It effectively models the spatial correlations inherent in the underlying space throughout the dimensionality reduction process, thereby safeguarding the local similarities of the original data within the low-dimensional space. Consequently, the dimensionality-reduced outcomes generated by SpatialPCA are rich in spatial information, which can be seamlessly integrated with other scRNA-seq analysis methods, facilitating diverse downstream analyses within the realm of spatial transcriptomics. Table 1 illustrates all the mentioned dimensionality reduction methods.

Table 1. Summary of Dimensionality Reduction Methods.

No.	Category	Methods	Implementation language/ Platform	GitHub address/ Usage or download URL	Year of publication	Reference	Strengths	Limitations
1	Single cell	PCA	Jupyter/Python	https://github.com/erdogant/pca	1987	[23]	Simplified with significant dimensionality reduction effects	Only applicable to linear data
2	Single cell	NMF	R	https://github.com/renozao/NMF	1999	[24]	Utilizing non-negative decomposition is more applicable to biological data	Only applicable to linear data
3	Single cell	scvis	Python	https://bitbucket.org/jerry00/scvis-dev/src/master/	2018	[25]	Capable of preserving both global and local structures from the original data even when handling small-scale	——

							datasets	
4	Single cell	SCDRHA	Python	https://github.com/WHY-17/SCDRHA	2021	[34]	The composite structure is capable of simultaneously avoiding dropout events and preserving cellular topological structure	——
5	Single cell	tGPLVM	Python	https://github.com/architverma1/tGPLVM	2020	[32]	Can be used for raw, unfiltered data	——
6	Single cell	DR-A	Python	https://github.com/eugenelin1/DRA	2020	[33]	Utilizing deep learning models to enhance dimensionality reduction performance	The benefits for downstream analysis need further exploration
7	Single cell	DREAM	Python	https://github.com/Crystal-JJ/DREAM	2023	[36]	Strong stability in dimensionality reduction and suitable for pseudo-temporal analysis	——
8	Single cell	UMAP	JavaScript/Python	https://github.com/umap-project/umap	2018	[30], [31]	Capable of simultaneously preserving local and global variables	——
9	Single cell	MulitiMAP	Python	https://github.com/Teichlab/MultiMAP	2021	[35]	Able to integrate and analyze data of different quantities and dimensions	——
10	Spatial transcriptomics	Seurat	R	https://github.com/satijalab/seurat	2015	[37]	Applicable to different types of single-cell transcriptomic data	——
11	Spatial transcriptomics	SCANPY	Python	https://github.com/scverse/scanpy	2018	[38]	Suitable for handling large-scale single-cell data	——
12	Spatial transcriptomics	SpatialPCA	R	https://github.com/shangll123/SpatialPCA	2022	[21]	Able to preserve crucial biological signals and spatial structures while reducing dimensionality	Applicable only to transcriptomic data

Clustering

Single-cell Transcriptomics

The CellBIC algorithm, introduced by Kim et al., presents a top-down hierarchical clustering approach tailored for scRNA-seq data analysis [39]. CellBIC leverages the bimodal distribution frequently observed within scRNA-seq data when clustering cells. This bimodal distribution signifies varying expression levels of many genes across distinct cell types. More precisely, it implies that a gene may exhibit high expression levels in one cell type while displaying low expression levels in others. The CellBIC algorithm operates under the assumption that cells featuring a bimodal distribution across multiple genes share certain similarities, which can be harnessed for the clustering of diverse cell types using the multimodal distribution of these genes. This fundamental principle underpins the CellBIC algorithm. The authors rigorously applied the CellBIC algorithm to diverse datasets, including human pancreas, mouse cortex, and mouse lung data, all of which consistently exhibited exceptional clustering outcomes.

Zhu et al. introduced the hidden-markov random field (HMRF) algorithm, a groundbreaking approach that seamlessly integrates two distinct methodologies: scRNA-seq and single-molecule fluorescence in situ hybridization (smFISH) [40]. While scRNA-seq offers a comprehensive transcriptomic perspective, it lacks spatial information. On the other hand, smFISH allows for sensitive mRNA transcript detection while preserving limited spatial information for a select few hundred genes. By amalgamating HMRF with both methods, scRNA-seq acts as a guiding framework to determine cell types that correspond to cells described by smFISH. This unique approach systematically identifies varied spatial domain patterns from smFISH and leverages these patterns to interpret variations associated with environmental factors within the scRNA-seq dataset. This holistic approach has paved the way for the systematic dissection of the individual contributions of cell type and spatially dependent factors in driving cell-state variations, a milestone that has proven elusive in other research endeavors.

Ruth et al. introduced SAME-clustering, a hybrid clustering method for single-cell data, hinging on maximum likelihood estimation [41]. The authors rigorously assessed SAME-clustering's clustering capabilities across 15 datasets sourced from various origins. Across all datasets, SAME-clustering consistently outperformed at least three of the individual clustering methods, showcasing its robust clustering performance.

Additionally, SAME-clustering proficiently estimated the number of clusters. When compared to individual clustering methods, SAME-clustering's integrated features notably bolstered its capacity to identify novel clusters.

Mori et al. introduced an innovative 3D reconstruction methodology, spatial reconstruction by stochastic self-organizing map (SPRESSO), which offers precise cell position assignments within tissues or organs based on gene expression data, even in scenarios devoid of spatial information [42]. SPRESSO leverages stochastic self-organizing map (stochastic-SOM) clustering to estimate the spatial domains of cells solely from the gene expression matrix. This algorithm strategically selects genes utilizing gene ontology (GO) criteria. SPRESSO yielded remarkable success rates, showcasing an outstanding capability to identify spatial discriminator genes pivotal in differentiation and tissue morphogenesis. However, it may not be readily applicable to entirely distinct or exclusive samples marked by entirely divergent gene expression patterns. Four years later, Mori et al. introduced an enhanced iteration of the SPRESSO algorithm named eSPRESSO [43]. In contrast to SPRESSO, eSPRESSO excels in facilitating three-dimensional reconstruction for more intricate tissues or organs, such as the heart and pancreas.

In 2019, Baran et al. introduced the 'metacell' method for clustering and cell type annotation in single-cell transcriptome data, marking a departure from conventional clustering approaches [44]. This innovative approach directly clusters cells based on highly variable genes within raw scRNA-seq data. Clustering outcomes are represented through a K-NN graph, which undergoes continuous splitting based on edge confidence until a group of cells with the strongest correlation emerges. The salient feature of Metacell lies in its capacity to elevate the ratio of data reliability to noise, consequently averting errors stemming from the introduction of incorrect models. In comparison to naive smoothing techniques, it exhibits reduced susceptibility to over-fitting and over-smoothing. Three years later, Ben-Kiki et al. introduced the Metacell-2 algorithm, a recursive divide-and-conquer algorithm that represents an enhanced iteration of the original Metacell algorithm [45]. Building upon the foundations of Metacell, Metacell-2 refines the algorithm for outlier cell detection and further augments the precision of rare cell type recognition during cell type annotation.

The Leiden algorithm[46], proposed by V. A. Traag et al. in 2019, is an extension of the Louvain algorithm. The Louvain algorithm achieves clustering through two functions: node movement and network aggregation[47]. This community-based algorithm is highly suitable for understanding large and complex network structures,

enabling clustering of single-cell data. The Leiden algorithm first performs an initial clustering and then iteratively improves the clustering accuracy by optimizing modularity. This addresses the issue of internal fragmentation observed in the Louvain algorithm, making it suitable for larger-scale single-cell data with faster execution speed.

stLearn, developed by Pham et al., stands as a comprehensive Python software package integrating three distinct algorithms for cell type identification, cell trajectory reconstruction, and the study of cell-cell interactions within morphologically intact tissue sections [48]. The primary workflow of stLearn encompasses several key steps:

1. Normalization: accomplished through neighborhood smoothing and morphological adjustment.

2. Clustering: conducted on normalized data using graph-based clustering to delineate spatially and transcriptionally defined clusters.

3. Spatial trajectory analysis: employing the potential spatial trajectory (PST) method to unveil local and global relationships among cell types.

4. Cell-cell interaction detection: alongside microenvironment identification.

5. Other common tasks: such as dimensionality reduction techniques.

Furthermore, stLearn offers a suite of visualization techniques that facilitate the incorporation of essential information onto tissue images, including gene expression, cluster labels, sub-cluster labels, microenvironments, and in vivo trajectories. A standout feature of stLearn clustering is its capacity to synergistically incorporate gene expression, spatial distance, and tissue morphology data. This comprehensive approach maximizes the utility of spatial transcriptomic data, thereby enabling the detection of rare cell types and enhancing overall clustering performance.

Wu et al. introduced an innovative approach for clustering single-cell transcriptomic data [49], termed jSRC, which effectively reframes the clustering problem as an optimization challenge. The jSRC algorithm seamlessly integrates dimensionality reduction and clustering within its comprehensive framework. It strategically selects features guided by cell clustering, thus significantly enhancing clustering accuracy. The jSRC algorithm encompasses three primary learning tasks: dimensionality reduction, sparse representation, and clustering. These tasks collaboratively operate to directly discern cell types and gene information based on the acquired features. Furthermore, jSRC demonstrates the capacity to automatically identify cell types and their respective quantities by proficiently learning the intrinsic features embedded within scRNA-seq data.

The scHFC algorithm, introduced by Wang et al., is an optimization of the fuzzy c-means (FCM) and gath-geva (GG) algorithms, tailored specifically for single-cell data clustering [50]. Initially, the scHFC algorithm employs PCA to effectively reduce the dimensionality of scRNA-seq data. Subsequently, the data undergoes clustering via an enhanced FCM algorithm, which cleverly incorporates both simulated annealing and genetic algorithms. This clustering process generates a membership matrix representing the initial clustering outcomes. The membership matrix then serves as input for the GG algorithm, culminating in the attainment of the definitive clustering results. The scHFC algorithm stands out for its exceptional performance in terms of clustering stability and robustness, thereby delivering highly valuable clustering outcomes for subsequent downstream analyses.

Main Takeaways

SAME-clustering is an ensemble algorithm that amalgamates multiple distinct scRNA-seq clustering algorithms to identify the most optimal clustering results. Conversely, jSRC redefines single-cell clustering as an optimization problem, perpetually enhancing accuracy and refining clustering outcomes through the dimensionality reduction process. jSRC can autonomously deduce the number of cell types via learning. scHFC demonstrates exceptional stability in clustering performance, delivering results that are notably advantageous for downstream analysis. CellBIC adopts a top-down hierarchical clustering approach, circumventing the challenges associated with simplistic distance measures often encountered in bottom-up algorithms dealing with multidimensional data. The SPRESSO algorithm excels in estimating precise cell spatial positions within tissues or organs, even when spatial information is absent, making it a valuable asset for clustering tasks devoid of spatial context. Its advanced iteration, eSPRESSO, faithfully replicates the three-dimensional structures of various tissues and organs. MetaCell and its upgraded counterpart, Metacell-2, introduce the innovative concept of 'metacell.' The Metacell algorithm excels in minimizing model errors and exhibits remarkable prowess in identifying rare cell types. The HMRF algorithm complements scRNA-seq and smFISH methods, with a particular focus on detecting spatial domain information. In contrast, the stLearn algorithm offers expanded functionality compared to its counterparts. It is ideally suited for data clustering, incorporating spatial distance and tissue morphology information alongside gene expression data. stLearn optimizes the utility of these combined data sources, enhancing the detection of rare cell types. The Leiden algorithm is more suitable for large-scale and more complex single-cell data.

Spatial Transcriptomics

Kats et al. introduced SpatialDE2 [51], a specialized tool designed specifically for delineating tissue regions within spatial transcriptomic data. SpatialDE2 comprises two pivotal modules: one dedicated to tissue region segmentation and the other focused on identifying genes exhibiting spatial variability. The tissue region segmentation module enhances user-friendliness and expedites processing compared to previous methods, notably featuring an adaptive count-based likelihood function that automatically determines the number of tissue regions. Conversely, the module for detecting spatially variable genes builds upon prior methodologies by introducing technical enhancements and computational optimizations, resulting in superior speed and accuracy. The integration of these two modules significantly elevates the precision and robustness of SpatialDE2.

Spatial graph convolutional network (SpaGCN), as introduced by Hu et al. [52], is a clustering method tailored for spatial transcriptomics analysis. SpaGCN capitalizes on graph construction to effectively capture the intricate relationships between spatial information and gene expression profiles across all points in the tissue. By integrating graph convolutional networks, SpaGCN adeptly aggregates gene expression data from neighboring points, and subsequently applies an unsupervised iterative clustering approach to the expression matrix. Each resulting cluster represents a distinct spatial domain, facilitating the identification of spatially variable genes that are enriched within specific domains through rigorous differential expression analysis. The authors conducted a comprehensive validation of SpaGCN's performance on seven distinct spatial transcriptomics datasets. In comparison to alternative methods, SpaGCN showcased heightened sensitivity in the detection of genes exhibiting enriched spatial expression patterns.

BayesSpace is a clustering method rooted in spatial transcriptomics models [53]. It leverages a t-distribution error model to effectively pinpoint spatial clusters that exhibit greater resilience to outliers arising from technical noise. BayesSpace harnesses markov chain monte carlo (MCMC) techniques for parameter estimation, thereby enhancing the exploration of spatial information. Through the utilization of a fixed precision matrix, BayesSpace enhances estimation accuracy without sacrificing clustering performance.

Singhal et al. introduced BANKSY [54]，a cutting-edge approach that seamlessly integrates molecular profiles with spatial locations, recognizing the comprehensive nature of cell representation within tissue when considering both its own transcriptome

and the transcriptome of its local microenvironment. BANKSY incorporates a spatial kernel to calculate the microenvironmental transcriptome, weighting contributions from neighboring cells. One prominent advantage of this approach in cell clustering is its ability to eliminate the assumption that cells of the same type or subtype must be physically close. Furthermore, BANKSY addresses the challenges of both cell type clustering and tissue domain segmentation within a unified framework of feature augmentation. Through this innovative strategy, BANKSY achieves precise labeling of spatially structured cell types and subtypes, even when subtle differences exist between subtypes in distinct microenvironments. Additionally, by fine-tuning a single hyperparameter, BANKSY accurately delineates tissue domains. The feature augmentation strategy employed by BANKSY seamlessly integrates with scalable graph clustering solutions capable of handling millions of cells. BANKSY can be seamlessly integrated with other spatial transcriptomic analysis methods, such as Seurat, SingleCellExperiment, and Scanpy, to unravel deeper biological insights.

The STAGATE algorithm, introduced by Dong et al., boasts an adaptive learning capability as its defining feature [55]. The STAGATE algorithm's workflow can be segmented into three distinct steps: First, it constructs a spatial neighbor network (SNN) based on the relative spatial positions of spots while concurrently performing pre-clustering of gene expression to identify regions containing different cell types. Subsequently, STAGATE employs a graph attention auto-encoder to learn low-dimensional latent embeddings of spatial information and gene expression. The unique advantage of STAGATE, setting it apart from other algorithms, is its use of an attention mechanism for adaptive learning in the intermediate layers of the encoder and decoder within the graph auto-encoder. Unlike other algorithms where the similarity of neighboring points is predefined before training the auto-encoder, STAGATE offers greater accuracy and flexibility in this regard. Lastly, UMAP is utilized for data visualization, and various clustering algorithms are applied to identify spatial domains. STAGATE's versatility extends to reconstructing 3D spatial transcriptomics and extracting 3D expression domains using multiple consecutive sections. It excels in downstream analyses, encompassing spatial domain identification, visualization, spatial trajectory inference, data denoising, and 3D expression domain extraction.

Stardust, an innovative extension of the widely adopted seurat clustering algorithm utilized in scRNA-seq data analysis [56], serves as a valuable complement. Seurat employs the Louvain algorithm, a network-based clustering approach, to represent each element of the dataset as a graph node. The connections between nodes are established

based on pairwise similarity, measured using the euclidean distance derived from transcriptional profiles. Subsequently, the algorithm conducts community detection to partition the dataset. Stardust enhances this framework by replacing the original distance matrix employed in seurat with a combined matrix that integrates both transcriptional information and spatial positioning of spots. To calculate the distance between pairs of nodes, it considers the distances of each node to all others. The transcriptional information matrix is generated through pairwise Euclidean distance computations across transcriptional profiles within PCA space, while the spatial position matrix reflects pairwise Euclidean distances between individual spots. What sets Stardust apart is its user-friendly approach, offering adaptability in determining the extent to which spatial information influences clustering dynamics across diverse tissue contexts. This method enhances overall stability and reliability, setting it apart from alternative clustering approaches.

To ensure that the low-dimensional embeddings obtained during dimension reduction remain relevant to the inferred class labels in the clustering step, Liu et al. introduced the dimension-reduction spatial-clustering (DR-SC) algorithm [57]. This innovative approach carries out dimension reduction and spatial clustering simultaneously within a unified framework. DR-SC is structured as a two-layer hierarchical model. In the first layer, DR establishes the connection between gene expression and latent low-dimensional embeddings, while the SC step aligns these latent low-dimensional embeddings with spatial coordinates and, when needed, clustering labels. Both layers work collaboratively to execute the clustering task. DR-SC is equipped with the ability to fine-tune the smoothness parameter within the spatial clustering component. By setting the smoothness parameter to zero, DR-SC can perform clustering on single-cell transcriptomic data devoid of spatial information. This integrated analysis methodology not only yields precise spatial clustering results but also ensures the efficient extraction of low-dimensional embedding information. During the dimension reduction process, DR-SC generates biologically relevant features compared to other dimension reduction methods and automatically determines the number of clusters using the modified minimum bayesian information criterion (MBIC), thereby enhancing clustering performance. Table 2 illustrates all the mentioned clustering methods.

Long et al. introduced GraphST [58], a novel graph-based self-supervised contrastive learning method that comprises three core modules: spatial clustering, multisample integration, and cell-type deconvolution. The distinctive feature of the

GraphST algorithm lies in its adept utilization of graph neural networks and contrastive learning techniques to capture both spatial transcriptomic gene expression information and spatial context. While other methods also employ neural networks to learn gene and spatial information, GraphST excels in preserving local spatial information through graph-based self-supervised contrastive learning. This enhancement results in more informative and discriminative learned representations, ultimately leading to superior clustering performance compared to these methods. Additionally, GraphST exhibits the versatility to integrate data from multiple samples and effectively identify cell types within spatial transcriptomic datasets. Long et al. rigorously assessed the performance of GraphST using diverse human and mouse datasets generated through various sequencing methods. GraphST has superior clustering performance and its remarkable ability to eliminate batch effects concurrently.

Hu et al. introduced ADEPT, a cutting-edge graph-based deep clustering algorithm tailored for spatial transcriptomic data [59]. The ADEPT algorithm encompasses five key steps: (1) It begins with input sequencing-based or image-based spatial transcriptomic data, treating each point in the gene expression profile as a node and considering the associated genes as feature vectors. A k-nearest neighbor (KNN) algorithm is employed to construct a graph structure based on the adjacency of nodes. (2) The resulting graph structure is passed through the generalized advantage estimator (GAE), which autonomously learns a low-dimensional representation of each node embedding. (3) Initial clustering is performed on the node embeddings. Multiple lists of differentially expressed genes (DEGs) are selected from these initial clusters based on the non-zero rates in the expression matrix. Once the reconstruction loss threshold is reached and the model has converged, a gaussian mixture model is applied to cluster the node embeddings. (4) The differential expression genes are amalgamated to form the final estimated matrix, subsequently utilized for the final clustering. (5) Downstream analysis is conducted based on the final clustering results. ADEPT exhibits robustness and superiority over other algorithms in diverse downstream analyses, highlighting its exceptional performance. Hu et al. conducted a validation of the ADEPT algorithm on a limited dataset, revealing potential scalability challenges that warrant further investigation and validation with larger sample sizes. It's worth noting that as the number of clusters increases, ADEPT's runtime also experiences an increment, suggesting opportunities for optimization in future iterations.

Table 2. Summary of Clustering Methods.

No.	Category	Methods	Implementation language/Platform	GitHub address/ Usage or download URL	Year of publication	Reference	Strengths	Limitations
1	Single cell	SAME-clustering	R/C++	https://github.com/yycunc/SAMEclustering	2019	[41]	Proficient in estimating the number of clusters	——
2	Single cell	jSRC	MATLAB/R	https://github.com/xkmaxidian/jSRC	2021	[49]	Integrated dimensionality reduction and clustering	——
3	Single cell	scHFC	Python	https://github.com/WJ319/scHFC	2022	[50]	High clustering stability	——
4	Single cell	CellBIC	MATLAB	https://github.com/neocaleb/CellBIC	2018	[39]	Top-down hierarchical	Only for scRNA-seq data analysis
5	Single cell	SPRESSO	R/Python/Shell	https://github.com/tmorikuicr/spresso	2019	[42]	Based on gene expression data (may lack spatial information)	Not applicable to entirely distinct or exclusive samples
6	Single cell	eSPRESSO	HTML/R	https://github.com/tmorikuicr/espresso	2023	[43]	Three-dimensional reconstruction	——
7	Single cell	MetaCell	R/C++	https://tanaylab.github.io/metacell/	2019	[44]	Can enhance the reliability of data by improving the signal-to-noise ratio	Not suitable for data lacking highly variable genes
8	Single cell	Metacell2	Python	https://github.com/tanaylab/metacells	2022	[45]	Achieves higher accuracy in identifying rare cell types	——

9	Single cell	HMRF	Python/C++ /Makefile	https://bitbucket.org/qzhudfci/smfishhmrf-py	2018	[40]	Identifies varied spatial domain patterns	——
10	Single cell	stlearn	Python	https://stlearn.readthedocs.io/	2020	[48]	Can collaboratively integrate gene expression, spatial distance, and tissue morphology data	——
11	Single cell	Leiden	Java/Python	https://github.com/CWTSLeiden/networkanalysis https://github.com/vtraag/leidenalg	2019	[46]	Suitable for larger-scale single-cell data	——
12	Spatial transcriptomics	Graph ST	Python	https://github.com/JinmiaoChenLab/GraphST	2023	[58]	Able to preserve local spatial information through graph self-supervised contrastive learning	——
13	Spatial transcriptomics	Spatial DE2	Python	https://github.com/PMBio/SpatialDE	2021	[51]	A tool specialized in delineating tissue regions in spatial transcriptomics data	——
14	Spatial transcriptomics	Banksy	Jupyter Notebook/Python	https://github.com/prabhakarlab/Banksy_py	2022	[54]	Eliminates the assumption that cells of the same type or subtype must be physically close to each other	——

15	Spatial transcriptomics	ADEPT	Python	https://github.com/maiziezhoulab/ADEPT	2023	[59]	Well-connected with downstream analyses	The runtime will increase with the growing volume of data
16	Spatial transcriptomics	STAGATE	Python	https://github.com/zhanglabtools/STAGATE	2022	[55]	Adding an attention mechanism to the graph autoencoder enables adaptive learning	——
17	Spatial transcriptomics	SpaGCN	Python/ Jupyter Notebook	https://github.com/jianhuupenn/SpaGCN	2021	[52]	Heightened sensitivity in the detection of genes exhibiting enriched spatial expression patterns	Applicable only to spatial transcriptomics clustering
18	Spatial transcriptomics	Stardust	R	https://github.com/InfOmics/stardust/	2022	[56]	More user-friendly	Incorporating limitations based on Louvain and other algorithms
19	Spatial transcriptomics	BayesSpace	R/C++	https://github.com/edward130603/BayesSpace	2021	[53]	Capable of improving estimation accuracy without sacrificing clustering performance	——
20	Spatial transcriptomics	DR-SC	R	https://github.com/feiyoung/DR-SC.Analysis	2022	[57]	Can perform clustering on data lacking spatial information	——

Main Takeaways

Among the spatial transcriptomic clustering methods mentioned above, four of them leverage graph neural network models. The GraphST algorithm integrates graph neural networks with contrastive learning, serving the dual purpose of clustering and batch effect removal in the samples. ADEPT builds a graph from the input sequencing or imaging data and employs graph autoencoders for clustering. STAGATE permits the definition of the similarity of neighboring points before training, enhancing the algorithm's adaptability for personalized analysis. SpaGCN introduces graph convolutional networks to aggregate gene expression information from neighboring points, facilitating unsupervised iterative clustering. SpatialDE2 takes raw gene count matrices as input data and relies on cell type count data obtained from the computational deconvolution step. The BayesSpace algorithm is versatile, suitable not only for spatial transcriptomic data but also for protein and multi-omics data analysis. BANKSY effectively characterizes cells within tissue by considering both their intrinsic transcriptomic data and the transcriptomic data of the local microenvironment, employing spatial kernels to calculate weighted averages of microenvironment transcriptomes from neighboring cells. Stardust minimizes the user-set parameters, with spatial weights being the sole required parameter. DR-SC allows for data-driven selection of the number of clusters, and the inferred cell types can be directly employed for subsequent differential gene expression analysis.

Conclusion and Outlook

In recent years, significant progress has been made in the fields of scRNA-seq and spatial transcriptomic analysis. Many researchers have developed a plethora of practical tools for dimensionality reduction and clustering in the context of single-cell transcriptomics and spatial transcriptomics.

It is evident that each algorithm has its unique characteristics, advantages, and provides valuable guidance for algorithm development. Different algorithms have specific requirements for input data. Spatial transcriptomics technology, when compared to single-cell transcriptomics, enriches the data with spatial information, necessitating spatial clustering for subsequent analysis. Moreover, many spatial transcriptomics clustering methods can be adapted for dimensionality reduction tasks in spatial transcriptomic data. Some of these algorithms are also suitable for downstream analysis after clustering.

In summary, computational methods related to dimensionality reduction and

clustering have significantly contributed to the advancement of scRNA-seq and spatial transcriptomics. Continuous innovations in these dimensionality reduction and clustering algorithms are poised to drive advancements in biological and clinical research, enhancing researchers' insights into the realm of bioinformatics. This review aims to guide researchers in selecting appropriate algorithms for their specific needs and inspire the development of related algorithms by summarizing dimensionality reduction and clustering algorithms for scRNA-seq and spatial transcriptomic data. All the methods mentioned in the document have been listed to https://github.com/YidiSun626/Dimensionality-Reduction-and-Clustering-Methods-for-Single-cell-and-Spatial-Transcriptomics-Data for researchers' reference.

Funding

The work was supported by the the National Natural Science Foundation of China (No. 62131004, No. 62102064, No. 62261018).

Conflict of Interest

None declared.

Acknowledgments

Figure 1 was created with BioRender.com.

References

1. Wagner, A., A. Regev, and N. Yosef, *Revealing the vectors of cellular identity with single-cell genomics.* Nature biotechnology, 2016. **34**(11): p. 1145-1160.
2. Xiang, R., et al., *A comparison for dimensionality reduction methods of single-cell RNA-seq data.* Frontiers in genetics, 2021. **12**: p. 646936.
3. Zhang, Z., et al., *Goals and approaches for each processing step for single-cell RNA sequencing data.* Briefings in Bioinformatics, 2021. **22**(4): p. bbaa314.
4. Wang, Y., et al. *SBSM-Pro: Support Bio-sequence Machine for Proteins*. 2023. arXiv:2308.10275 DOI: 10.48550/arXiv.2308.10275.
5. Zhang, Z., et al., *Single-cell RNA analysis reveals the potential risk of organ-specific cell types vulnerable to SARS-CoV-2 infections.* Comput Biol Med, 2021. **140**: p. 105092.
6. Sun, S., et al., *Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis.* Genome biology, 2019. **20**(1): p. 1-21.
7. Duan, H., et al., *Machine learning-based prediction model for distant metastasis of breast cancer.* Computers in Biology and Medicine, 2024. **169**: p. 107943.

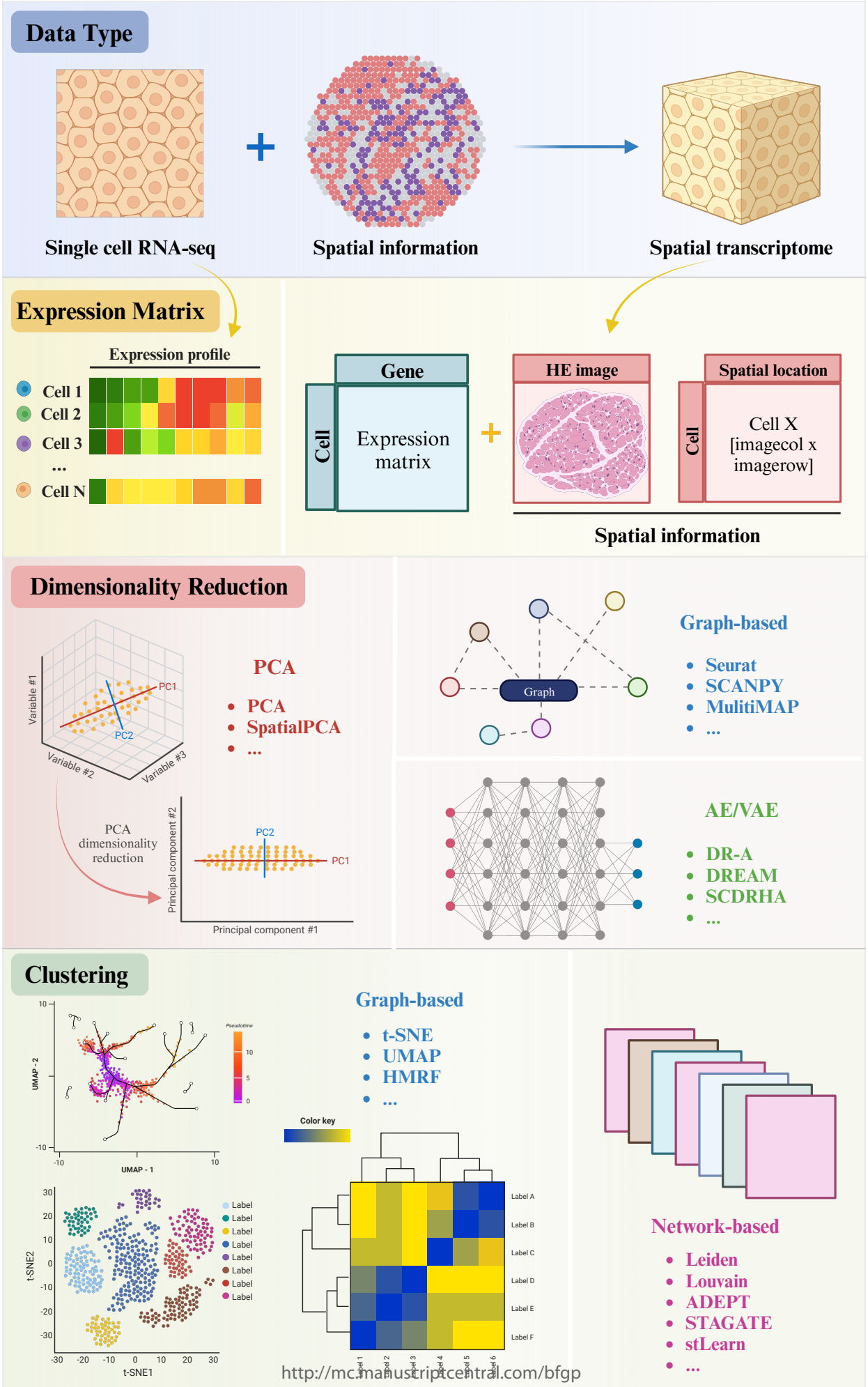
8. Gong, W., et al., *DrImpute: imputing dropout events in single cell RNA sequencing data.* BMC bioinformatics, 2018. **19**: p. 1-10.
9. Kharchenko, P.V., L. Silberstein, and D.T. Scadden, *Bayesian approach to single-cell differential expression analysis.* Nature methods, 2014. **11**(7): p. 740-742.
10. Kiselev, V.Y., et al., *SC3: consensus clustering of single-cell RNA-seq data.* Nature methods, 2017. **14**(5): p. 483-486.
11. Qi, R., et al., *Clustering and classification methods for single-cell RNA-sequencing data.* Briefings in bioinformatics, 2020. **21**(4): p. 1196-1208.
12. Zhang, Z., et al., *Critical downstream analysis steps for single-cell RNA sequencing data.* Briefings in Bioinformatics, 2021. **22**(5): p. bbab105.
13. Qi, R., et al., *A spectral clustering with self-weighted multiple kernel learning method for single-cell RNA-seq data.* Briefings in Bioinformatics, 2021. **22**(4): p. bbaa216.
14. Zhang, Z., et al., *Single-cell RNA sequencing analysis identifies key genes in brain metastasis from lung adenocarcinoma.* Current Gene Therapy, 2021. **21**(4): p. 338-348.
15. Wang, Y., et al., *SBSM-Pro: support bio-sequence machine for proteins.* arXiv preprint arXiv:2308.10275, 2023.
16. Stegle, O., S.A. Teichmann, and J.C. Marioni, *Computational and analytical challenges in single-cell transcriptomics.* Nature Reviews Genetics, 2015. **16**(3): p. 133-145.
17. Ziegenhain, C., et al., *Comparative analysis of single-cell RNA sequencing methods.* Molecular cell, 2017. **65**(4): p. 631-643. e4.
18. Wang, J., Q. Zou, and C. Lin, *A comparison of deep learning-based pre-processing and clustering approaches for single-cell RNA sequencing data.* Briefings in Bioinformatics, 2022. **23**(1): p. bbab345.
19. Rao, A., et al., *Exploring tissue architecture using spatial transcriptomics.* Nature, 2021. **596**(7871): p. 211-220.
20. Su, Y., et al. *Human-Spa: An Online Platform Based on Spatial Transcriptome Data for Diseases of Human Systems*. in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2023.
21. Shang, L. and X. Zhou, *Spatially aware dimension reduction for spatial transcriptomics.* Nature Communications, 2022. **13**(1): p. 7203.
22. Zhang, Z., et al., *webSCST: an interactive web application for single-cell RNA-sequencing data and spatial transcriptomic data integration.* Bioinformatics, 2022. **38**(13): p. 3488-3489.
23. Wold, S., K. Esbensen, and P. Geladi, *Principal component analysis.* Chemometrics and Intelligent Laboratory Systems, 1987. **2**(1): p. 37-52.
24. Lee, D.D. and H.S. Seung, *Learning the parts of objects by non-negative matrix factorization.* Nature, 1999. **401**(6755): p. 788-791.
25. Ding, J., A. Condon, and S.P. Shah, *Interpretable dimensionality reduction of single cell transcriptome data with deep generative models.* Nature Communications, 2018. **9**(1): p. 2002.

26. Laurens van der Maaten, G.H., *Visualizing Data using t-SNE.* Journal of Machine Learning Research, 2008. **9**(86): p. 2579−2605.
27. Zhou, Y. and T.O. Sharpee, *Using Global t-SNE to Preserve Intercluster Data Structure.* Neural Computation, 2022. **34**(8): p. 1637-1651.
28. Linderman, G.C., et al., *Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data.* Nature Methods, 2019. **16**(3): p. 243-245.
29. Häkkinen, A., et al., *qSNE: quadratic rate t-SNE optimizer with automatic parameter tuning for large datasets.* Bioinformatics, 2020. **36**(20): p. 5086-5092.
30. Mcinnes L , H.J., *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* The Journal of Open Source Software, 2018. **3**(29): p. 861.
31. Becht, E., et al., *Dimensionality reduction for visualizing single-cell data using UMAP.* Nature Biotechnology, 2019. **37**(1): p. 38-44.
32. Verma, A. and B.E. Engelhardt, *A robust nonlinear low-dimensional manifold for single cell RNA-seq data.* BMC Bioinformatics, 2020. **21**(1): p. 324.
33. Lin, E., S. Mukherjee, and S. Kannan, *A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis.* BMC Bioinformatics, 2020. **21**(1): p. 64.
34. Zhao, J., et al., *SCDRHA: A scRNA-Seq Data Dimensionality Reduction Algorithm Based on Hierarchical Autoencoder.* Front Genet, 2021. **12**: p. 733906.
35. Jain, M.S., et al., *MultiMAP: dimensionality reduction and integration of multimodal data.* Genome Biology, 2021. **22**(1): p. 346.
36. Jiang, J., et al., *Dimensionality reduction and visualization of single-cell RNA-seq data with an improved deep variational autoencoder.* Briefings in Bioinformatics, 2023. **24**(3).
37. Satija, R., et al., *Spatial reconstruction of single-cell gene expression data.* Nature Biotechnology, 2015. **33**(5): p. 495-502.
38. Wolf, F.A., P. Angerer, and F.J. Theis, *SCANPY: large-scale single-cell gene expression data analysis.* Genome Biology, 2018. **19**(1): p. 15.
39. Kim, J., D.E. Stanescu, and Kyoung J. Won, *CellBIC: bimodality-based top-down clustering of single-cell RNA sequencing data reveals hierarchical structure of the cell type.* Nucleic Acids Research, 2018. **46**(21): p. e124-e124.
40. Zhu, Q., et al., *Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data.* Nature Biotechnology, 2018. **36**(12): p. 1183-1190.
41. Huh, R., et al., *SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble.* Nucleic Acids Research, 2019. **48**(1): p. 86-95.
42. Mori, T., et al., *Novel computational model of gastrula morphogenesis to identify spatial discriminator genes by self-organizing map (SOM) clustering.* Scientific Reports, 2019. **9**(1): p. 12597.
43. Mori, T., et al., *eSPRESSO: topological clustering of single-cell*

transcriptomics data to reveal informative genes for spatio–temporal architectures of cells. BMC Bioinformatics, 2023. **24**(1): p. 252.

44. Baran, Y., et al., *MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions.* Genome Biology, 2019. **20**(1): p. 206.
45. Ben-Kiki, O., et al., *Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis.* Genome Biology, 2022. **23**(1): p. 100.
46. Traag, V.A., L. Waltman, and N.J. van Eck, *From Louvain to Leiden: guaranteeing well-connected communities.* Scientific Reports, 2019. **9**(1): p. 5233.
47. Blondel, V., et al., *Fast Unfolding of Communities in Large Networks.* Journal of Statistical Mechanics Theory and Experiment, 2008. **2008**.
48. Duy, P., et al., *stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues.* bioRxiv, 2020: p. 2020.05.31.125658.
49. Wu, W., Z. Liu, and X. Ma, *jSRC: a flexible and accurate joint learning algorithm for clustering of single-cell RNA-sequencing data.* Briefings in Bioinformatics, 2021. **22**(5).
50. Wang, J., et al., *scHFC: a hybrid fuzzy clustering method for single-cell RNA-seq data optimized by natural computation.* Briefings in Bioinformatics, 2022. **23**(2).
51. Ilia, K., V.-T. Roser, and S. Oliver, *SpatialDE2: Fast and localized variance component analysis of spatial transcriptomics.* bioRxiv, 2021: p. 2021.10.27.466045.
52. Hu, J., et al., *SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network.* Nature Methods, 2021. **18**(11): p. 1342-1351.
53. Zhao, E., et al., *Spatial transcriptomics at subspot resolution with BayesSpace.* Nature Biotechnology, 2021. **39**(11): p. 1375-1384.
54. Vipul, S., et al., *BANKSY: A Spatial Omics Algorithm that Unifies Cell Type Clustering and Tissue Domain Segmentation.* bioRxiv, 2022: p. 2022.04.14.488259.
55. Dong, K. and S. Zhang, *Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder.* Nature Communications, 2022. **13**(1): p. 1739.
56. Avesani, S., et al., *Stardust: improving spatial transcriptomics data analysis through space-aware modularity optimization-based clustering.* GigaScience, 2022. **11**.
57. Liu, W., et al., *Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data.* Nucleic Acids Research, 2022. **50**(12): p. e72-e72.
58. Long, Y., et al., *Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST.* Nature Communications, 2023. **14**(1): p. 1155.
59. Hu, Y., et al., *ADEPT: Autoencoder with differentially expressed genes and*

imputation for robust spatial transcriptomics clustering. iScience, 2023. **26**(6): p. 106792.


Data Type
Single cell RNA-seq
Spatial information
Spatial transcriptome
Expression Matrix
Expression profile
Cell 1
Cell 2
Cell 3
...
Cell N
Gene
Cell
Expression matrix
HE image
Spatial location
Cell
Cell X [imagecol x imagerow]
Spatial information
Dimensionality Reduction
Variable #1
Variable #2
Variable #3
PC1
PC2
PCA
PCA
SpatialPCA
...
PCA dimensionality reduction
Principal component #2
Principal component #1
Graph
Graph-based
Seurat
SCANPY
MulitiMAP
...
AE/VAE
DR-A
DREAM
SCDRHA
...
Clustering
UMAP - 1
UMAP - 2
Pseudotime
t-SNE1
t-SNE2
Label
Graph-based
t-SNE
UMAP
HMRF
...
Color key
Label A
Label B
Label C
Label D
Label E
Label F
Network-based
Leiden
Louvain
ADEPT
STAGATE
stLearn
...

No.	Category	Methods	Implementation language/Platform	GitHub address/Usage or download URL	Year of publication	Reference	Strengths	Limitations
1	Single cell	PCA	Jupyter/Python	https://github.com/erdogant/pca	1987	[23]	Simplified with significant dimensionality reduction effects	Only applicable to linear data
2	Single cell	NMF	R	https://github.com/renozao/NMF	1999	[24]	Utilizing non-negative decomposition is more applicable to biological data	Only applicable to linear data
3	Single cell	scvis	Python	https://bitbucket.org/jerry00/scvis-dev/src/master/	2018	[25]	Capable of preserving both global and local structures from the original data even when handling small-scale datasets	____
4	Single cell	SCDRHA	Python	https://github.com/WHY-17/SCDRHA	2021	[34]	The composite structure is capable of simultaneously avoiding dropout events and preserving cellular topological structure	____
5	Single cell	tGPLVM	Python	https://github.com/architverma1/tGPLVM	2020	[32]	Can be used for raw, unfiltered data	____
6	Single cell	DR-A	Python	https://github.com/eugenelin1/DRA	2020	[33]	Utilizing deep learning models to enhance dimensionality reduction performance	The benefits for downstream analysis need further exploration

7	Single cell	DREAM	Python	https://github.com/Crystal-JJ/DREAM	2023	[36]	Strong stability in dimensionality reduction and suitable for pseudo-temporal analysis	——
8	Single cell	UMAP	JavaScript/Python	https://github.com/umap-project/umap	2018	[30], [31]	Capable of simultaneously preserving local and global variables	——
9	Single cell	MulitiMAP	Python	https://github.com/Teichlab/MultiMAP	2021	[35]	Able to integrate and analyze data of different quantities and dimensions	——
10	Spatial transcriptomics	Seurat	R	https://github.com/satijalab/seurat	2015	[37]	Applicable to different types of single-cell transcriptomic data	——
11	Spatial transcriptomics	SCANPY	Python	https://github.com/scverse/scanpy	2018	[38]	Suitable for handling large-scale single-cell data	——
12	Spatial transcriptomics	SpatialPCA	R	https://github.com/shangll123/SpatialPCA	2022	[21]	Able to preserve crucial biological signals and spatial structures while reducing dimensionality	Applicable only to transcriptomic data

No.	Category	Methods	Implementation language/Platform	GitHub address/Usage or download URL	Year of publication	Reference	Strengths	Limitations
1	Single cell	SAME-clustering	R/C++	https://github.com/yycunc/SAMEclustering	2019	[41]	Proficient in estimating the number of clusters	——
2	Single cell	jSRC	MATLAB/R	https://github.com/xkmaxidian/jSRC	2021	[49]	Integrated dimensionality reduction and clustering	——
3	Single cell	scHFC	Python	https://github.com/WJ319/scHFC	2022	[50]	High clustering stability	——
4	Single cell	CellBIC	MATLAB	https://github.com/neocaleb/CellBIC	2018	[39]	Top-down hierarchical	Only for scRNA-seq data analysis
5	Single cell	SPRESSO	R/Python/Shell	https://github.com/tmorikuicr/spresso	2019	[42]	Based on gene expression data (may lack spatial information)	Not applicable to entirely distinct or exclusive samples
6	Single cell	eSPRESSO	HTML/R	https://github.com/tmorikuicr/espresso	2023	[43]	Three-dimensional reconstruction	——
7	Single cell	MetaCell	R/C++	https://tanaylab.github.io/metacell/	2019	[44]	Can enhance the reliability of data by improving the signal-to-noise ratio	Not suitable for data lacking highly variable genes
8	Single cell	Metacell2	Python	https://github.com/tanaylab/metacells	2022	[45]	Achieves higher accuracy in identifying rare cell types	——
9	Single cell	HMRF	Python/C++/Makefile	https://bitbucket.org/qzhudfci/smfishhmrf-py	2018	[40]	Identifies varied spatial domain patterns	——

10	Single cell	stlearn	Python	https://stlearn.readthedocs.io/	2020	[48]	Can collaboratively integrate gene expression, spatial distance, and tissue morphology data	——
11	Single cell	Leiden	Java/Python	https://github.com/CWTSLeiden/networkanalysis https://github.com/vtraag/lei	2019	[46]	Suitable for larger-scale single-cell data	——
12	Spatial transcriptomics	GraphST	Python	https://github.com/JinmiaoChenLab/GraphST	2023	[58]	Able to preserve local spatial information through graph self-supervised contrastive learning	——
13	Spatial transcriptomics	SpatialDE2	Python	https://github.com/PMBio/SpatialDE	2021	[51]	A tool specialized in delineating tissue regions in spatial transcriptomics data	——
14	Spatial transcriptomics	Banksy	Jupyter Notebook/Python	https://github.com/prabhakarlab/Banksy_py	2022	[54]	Eliminates the assumption that cells of the same type or subtype must be physically close to each other	——
15	Spatial transcriptomics	ADEPT	Python	https://github.com/maiziezhoulab/ADEPT	2023	[59]	Well-connected with downstream analyses	The runtime will increase with the growing volume of data

16	Spatial transcriptomics	STAGATE	Python	https://github.com/zhanglabtools/STAGATE	2022	[55]	Adding an attention mechanism to the graph autoencoder enables adaptive learning	——
17	Spatial transcriptomics	SpaGCN	Python/ Jupyter Notebook	https://github.com/jianhuupenn/SpaGCN	2021	[52]	Heightened sensitivity in the detection of genes exhibiting enriched spatial expression patterns	Applicable only to spatial transcriptomics clustering
18	Spatial transcriptomics	Stardust	R	https://github.com/InfOmics/stardust/	2022	[56]	More user-friendly	Incorporating limitations based on Louvain and other algorithms
19	Spatial transcriptomics	BayesSpace	R/C++	https://github.com/edward130603/BayesSpace	2021	[53]	Capable of improving estimation accuracy without sacrificing clustering performance	——
20	Spatial transcriptomics	DR-SC	R	https://github.com/feiyoung/DR-SC.Analysis	2022	[57]	Can perform clustering on data lacking spatial information	——