




Editor's Choice

Integration tools for scRNA-seq data and spatial transcriptomics sequencing data

Chaorui Yan, Yanxu Zhu, Miao Chen, Kainan Yang, Feifei Cui , Quan Zou  and Zilong Zhang 

Corresponding authors: Zilong Zhang, School of Computer Science and Technology, Hainan University, 570228, Haikou, China. Tel.: +8618306350568; E-mail: zhangzilong@hainanu.edu.cn, Quan Zou, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, 610054, Chengdu, China. Tel.: +8613656009020; E-mail: zouquan@nclab.net, Feifei Cui, School of Computer Science and Technology, Hainan University, 570228, Haikou, China. Tel.: +8615063589727; E-mail: feifeicui@hainanu.edu.cn

Abstract

Numerous methods have been developed to integrate spatial transcriptomics sequencing data with single-cell RNA sequencing (scRNA-seq) data. Continuous development and improvement of these methods offer multiple options for integrating and analyzing scRNA-seq and spatial transcriptomics data based on diverse research inquiries. However, each method has its own advantages, limitations and scope of application. Researchers need to select the most suitable method for their research purposes based on the actual situation. This review article presents a compilation of 19 integration methods sourced from a wide range of available approaches, serving as a comprehensive reference for researchers to select the suitable integration method for their specific research inquiries. By understanding the principles of these methods, we can identify their similarities and differences, comprehend their applicability and potential complementarity, and lay the foundation for future method development and understanding. This review article presents 19 methods that aim to integrate scRNA-seq data and spatial transcriptomics data. The methods are classified into two main groups and described accordingly. The article also emphasizes the incorporation of High Variance Genes in annotating various technologies, aiming to obtain biologically relevant information aligned with the intended purpose.

Keywords: integration; scRNA-seq data; spatial transcriptomics sequencing data; HVGs

INTRODUCTION

Cellular heterogeneity refers to the differences in protein expression levels between different cell types, caused by specific gene expression in different cellular compartments within the same tissue or cell population. Traditional sequencing technologies may mask cellular heterogeneity among different cells, detecting only differences between individuals or populations [1]. Single-cell RNA sequencing (scRNA-seq) technology surpasses the limitations of traditional sequencing by enabling gene expression analysis at the single-cell level, elucidating transcriptional activity in individual cells and uncovering distinct cell subpopulations [2–6]. There are many methods for scRNA-seq technology, with the most commonly used being 10x Genomics, Smart-seq [7], Drop-seq and InDrop [8]. Currently, scRNA-seq technology is constantly developing and playing a role in many fields such as life sciences and medicine [9–12].

Spatial transcriptomics is a technique that utilizes RNA sequencing to discern the spatial location and expression patterns of diverse cells or genes, while maintaining tissue structure integrity. It aids in uncovering their spatial distribution and intercellular communication. However, the sequencing depth of spatial transcriptomics is much lower than that of scRNA-seq [13]. Currently, spatial transcriptomics technology primarily falls into two categories: high-plex RNA imaging and spatial barcoding. Examples of high-plex RNA imaging technologies comprise MERFISH [14], seqFISH [15], STARmap [16], among others, while spatial barcoding technologies encompass Spatial transcriptomics [17], Slide-seq [18], ZipSeq [19] and others [2].

In scRNA-seq, tissue dissociation is a crucial step that results in the loss of spatial information pertaining to cells in the original tissue. Consequently, integrating spatial transcriptomics data with scRNA-seq data for combined analysis holds the potential to significantly enhance our comprehension of biological

Chaorui Yan is currently a master's student in computer science, Hainan University, Haikou, China. His research interests include bioinformatics, machine learning and deep learning.

Yanxu Zhu is currently an undergraduate student in computer science, Hainan University, Haikou, China. Her research interests include bioinformatics and AI.

Miao Chen is currently an undergraduate student in computer science, Hainan University, Haikou, China. Her research interests include bioinformatics and AI.

Kainan Yang is currently an undergraduate student in computer science, Hainan University, Haikou, China. Her research interests include big data and AI.

Feifei Cui is currently an associate professor in the School of Computer Science and Technology, Hainan University. She received the Ph.D. degree in bioinformatics from the University of Tokyo, Japan. Her research interests include bioinformatics, deep learning and biological data mining.

Quan Zou is currently a Professor in the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China. He received the PhD degrees in computer science from Harbin Institute of Technology, China. His research is in the areas of bioinformatics, machine learning and parallel computing.

Zilong Zhang is currently an associate professor in the School of Computer Science and Technology, Hainan University. He received the Ph.D. degree in bioinformatics from the University of Tokyo, Japan. His research interests include bioinformatics, machine learning and graph neural network.

Received: August 2, 2023. **Revised:** September 26, 2023. **Accepted:** January 3, 2024

© The Author(s) 2024. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

mechanisms in both spatial and temporal dimensions [20]. In practical applications, accurately mapping cell types to specific spatial locations enables the understanding of their distribution patterns within tissues. This approach not only reveals tissue structure and function but also provides insights into cell–cell interactions, differentiation and tissue development processes. Moreover, it enhances comprehension of drug mechanisms at the cellular and tissue levels, thereby promoting drug development and disease treatment strategies. For example, Marc Elosua-Bayes, Paula Nieto, Elisabetta Mereu and their research team utilized SPOTlight to investigate pancreatic ductal adenocarcinoma (PDAC) data and discern the spatial organization of clinically relevant immune cell states within the tumor microenvironment [21]. Similarly, Reuben Moncada, Dalia Barkley, Florian Wagner and their collaborators employed Multimodal Intersection Analysis (MIA) technology to identify the co-localization of inflammatory fibroblasts expressing stress response gene modules with cancer cells [13].

Integration methods of scRNA-seq and spatial transcriptomics can be broadly classified into two main categories: Deconvolution and Mapping. Deconvolution methods involve the construction of mathematical or statistical inference models, wherein scRNA-seq data serves as the background data. These methods integrate the spatial transcriptomics sequencing data with the known cell types or expression patterns of marker genes derived from scRNA-seq data, enabling the inference of cell types for each spot. Conversely, mapping aims to establish correspondence between scRNA-seq data and spatial transcriptomics sequencing data, aligning and mapping them within the spatial domain. This approach enables spatial visualization and analysis of cell types or gene expression patterns. In contrast to the deconvolution approach, mapping typically doesn't necessitate the prior development of elaborate cell subtype models. Mapping is generally more flexible and better suited for situations demanding greater precision in cell source identification. Spatial transcriptomics sequencing data typically involve two sequencing methods: high-plex RNA imaging (HPRI) and Spatial barcoding technology. HPRI technology provides greater sequencing depth, resulting in higher RNA molecule abundance at each genomic location, while Spatial barcoding offers slightly higher resolution, allowing for better distinction and spatial precision among different locations.

DECONVOLUTION INTEGRATION METHODS FOR SINGLE-CELL RNA SEQUENCING AND SPATIAL TRANSCRIPTOMICS

Deconvolution methods for integrating scRNA-seq and spatial transcriptomics data can be classified into three distinct groups: enrichment score-based methods, regression model-based methods and probability model-based methods. Enrichment score-based methods estimate the relative abundance of cell types within each spot by calculating the enrichment score of a specific gene across spatial transcriptomics and scRNA-seq datasets. Regression model-based methods posit that the expression of each gene can be predicted using a linear or nonlinear regression model, which is trained using scRNA-seq data and subsequently applied to spatial transcriptomics data to forecast the gene expression in each spot. Probability model-based methods assume that the gene expression of different cell types across various locations can be characterized by a probability distribution. These methods utilize scRNA-seq data to build a probability model, which is then applied to spatial transcriptomics data, facilitating the estimation of the probability distribution for each cell type in every spot.

The integration of scRNA-seq data and spatial transcriptomics data often involves the utilization of highly variable genes (HVGs) to identify genes exhibiting substantial expression differences within the cell population. These differentially expressed genes may play important regulatory roles in terms of cell types, states and other aspects (Figure 1). HVGs are genes that display substantial expression level variations across individuals or in diverse conditions within a biological organization or a cell population [22]. The analysis of HVGs typically involves the utilization of tools like Seurat [23], scran [24], BASiCS [25] and other established academic protocols. By using HVGs, these differentially expressed genes are highlighted, providing insights into the functional diversity and biological differences within the cell population. In methods that do not rely on HVGs, the expression information of all genes is included in the integration analysis, which approach allows for a comprehensive consideration of the expression status of all genes, not just the highly variable ones. This comprehensive consideration may help discover genes that are overall stably expressed but exhibit differences in specific cell types or spatial locations. These genes might be overlooked in traditional differential analysis but play important roles in the characteristics of specific cell types or tissue structures, leading to the discovery of new key genes and novel biological insights (Summarized in Table 1). When working with large datasets, the computation of HVGs often necessitates additional computational resources [26]. In summary, classifying integration techniques into two subcategories based on HVG usage, as opposed to previous methods, better highlights the selection of integration techniques for achieving either a global or local perspective. This distinction is particularly valuable in scenarios involving large datasets, demanding substantial computational resources for HVG computation due to dataset scale.

SPOTlight [21] is a deconvolution integration method relies on non-negative matrix factorization regression. This method utilizes scRNA-seq data to perform non-negative matrix factorization and then deconvolves spatial transcriptomics sequencing data. The method is highly sensitive to scRNA-seq technology, with Quartz-Seq2, Smart-Seq2 and Chromium being the three best-performing scRNA-seq technologies. Additionally, experimental testing found that SPOTlight's performance severely deteriorates with low sequencing depth, and training with 100 cells per cell type can balance computational efficiency and performance.

Cell2location [27] is a probabilistic model-based deconvolution integration method that performs joint analysis of scRNA-seq data and spatial transcriptomics data using Bayesian statistical modeling. This model exhibits a strong capability to identify fine-grained cell types within complex tissues and demonstrates high sensitivity toward subtle cell type variations. Cell2location also handles different batches of data well.

SD² [28] is a deconvolution-based data integration method that leverages both dropout-based genes and spatial information. It uses graph convolutional networks (GCN) to perform cell-type deconvolution, where dropout genes are typically removed as obstructions during data analysis. Dropout-based genes constitute >80% of the scRNA-seq and spatial transcriptomics data. By utilizing this feature, SD² is suitable for data analysis with high dropout rates and generally outperforms cell2location [27], Deconvoluting Spatial Transcriptomics Data through Graph-based Convolutional Networks (DSTG) [29] and SPOTlight [21] in terms of performance.

SpatialDWLS [30] is a deconvolution method based on weighted least squares. Its implementation can be divided into two main

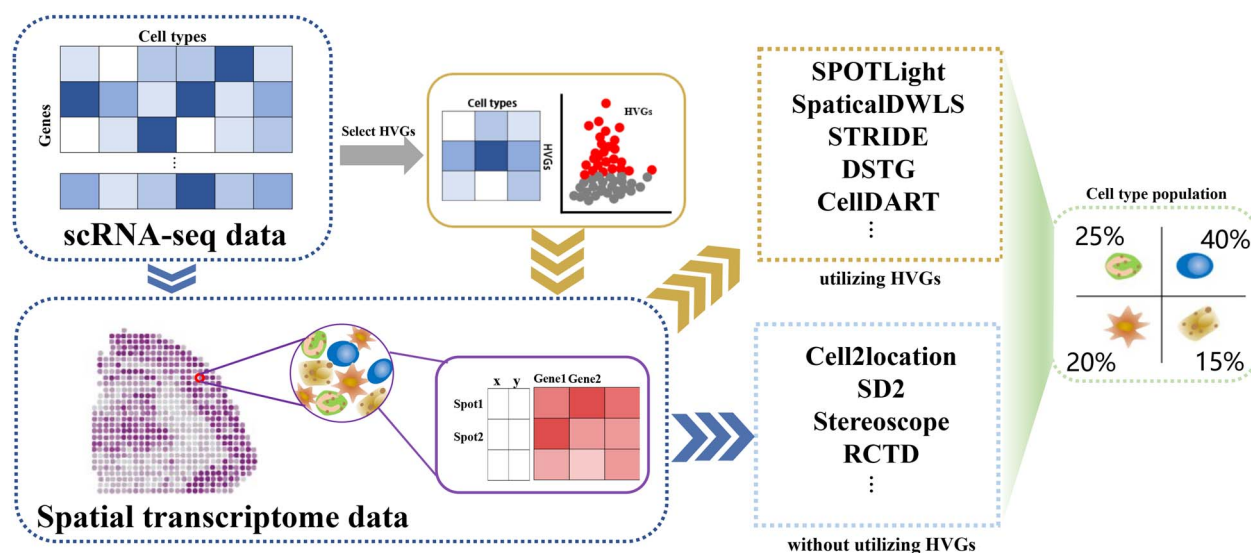


Figure 1: Deconvolution methods integrate single-cell sequencing data and spatial transcriptomics sequencing data, based on the utilization of HVGs.

Table 1. Summary of deconvolution methods

Methods	Key technique	GitHub address	Feature
SPOTLight(2021)	NNLS	https://github.com/MarcElosua/SPOTlight_deconvolution_analysis	Performs well even with small cell datasets
Cell2location(2022)	Bayesian model	https://github.com/BayraktarLab/cell2location	Allows joint analysis of multiple single-cell and spatial transcriptomic data with high accuracy, regardless of cell abundance
SD ² (2022)	GCN	https://github.com/leihouyeung/SD2	Including dropout data as valid input data can provide more convincing biological information
SpaticalDWLS (2021)	DWLS and enrichment analysis	https://github.com/RubD/Giotto	Possesses high computing efficiency
Stereoscope (2020)	Probabilistic model	https://github.com/almaan/stereoscope	Using similar organizations of scRNA-seq data and spatial transcriptomic sequencing data is possible, without requiring the pairing of these two types of data
RCTD (2022)	Probabilistic model	https://github.com/dmcable/spacexr	Considering platform effects, cross-platform learning can be achieved
STRIDE (2022)	topic-model-based	https://github.com/wanglabtongji/STRIDE	Can be used for integrating serial sectioned tissues and reconstructing the three-dimensional structure of tissues
DSTG (2021)	GCN	https://github.com/Su-informatics-lab/DSTG	Due to inherent algorithmic differences, spatial transcriptomics sequencing data and scRNA-seq data may not be compatible
CellDART (2022)	ADDA	https://github.com/mexchy1000/CellDART	Better than most existing integration methods in terms of spatial localization of multiple subtypes of excitatory neurons

steps. In the first step, cell type enrichment analysis is performed on each spatial transcriptomics sequencing data point using both scRNA-seq and spatial transcriptomics sequencing data. The second step involves the utilization of an enhanced dampened weighted least squares (DWLS) algorithm to estimate the cell type location for each data point. Ultimately, the cell type composition for each location is determined using the least squares method.

Specifically, the method examines differentially expressed genes in scRNA-seq data to identify cell type-specific marker genes. Subsequently, the identified marker genes are employed for deconvolution analysis of the spatial transcriptomics data, enabling the inference of cell type composition.

Stereoscope [31] is a probability model-based deconvolution method that considers each expression atlas to have its own inherent biological features. Therefore, Stereoscope utilizes

complete expression profile information rather than dependent on specific gene markers to deconvolve spatial transcriptomics data, and does not require normalization, gene selection and other preprocessing steps.

Initially, the method estimates the parameters that characterize the negative binomial distribution for each gene across different cell types by leveraging scRNA-seq data. These parameters capture the expression variations of each gene within specific cell types. Subsequently, the parameters of individual cells are combined using weighted averaging to generate an optimal interpretation of the spatial data. The weights are normalized to determine the relative abundance of each cell type, providing a comprehensive understanding of their distribution within the sample. Due to its unique characteristics, Stereoscope performs well in complex tissues with multiple similar cell types.

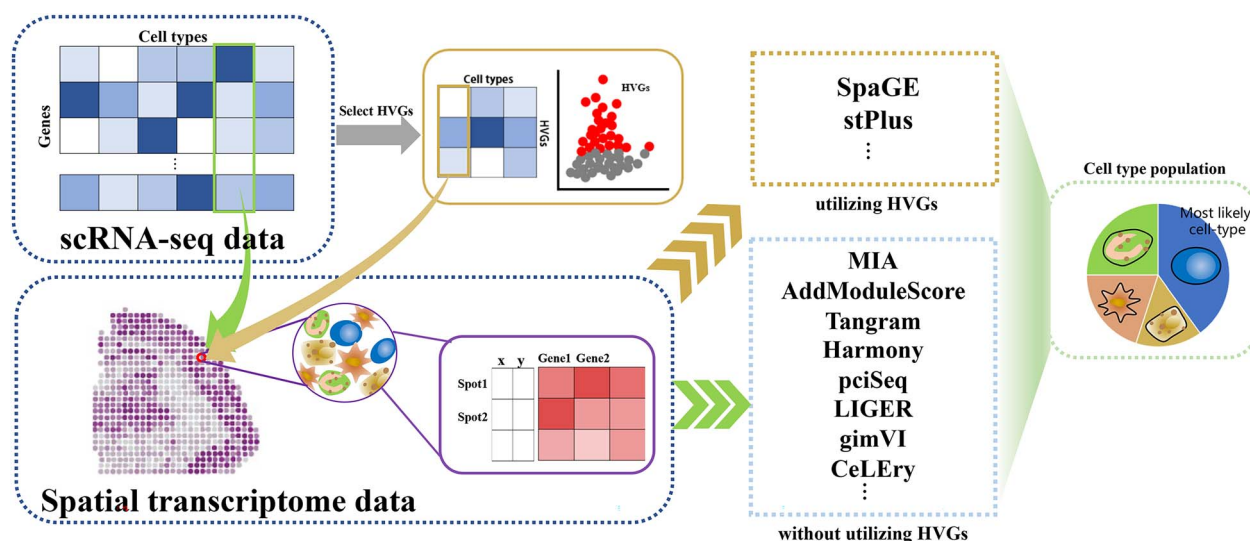


Figure 2: Mapping methods integrate single-cell sequencing data and spatial transcriptomics sequencing data, and classify them based on the utilization of HVGs.

Robust Cell Type Decomposition (RCTD) [32] is a supervised learning deconvolution method that utilizes a probabilistic model for robust cell type decomposition. By leveraging scRNA-seq data, RCTD acquires cell type expression profiles and employs them to deconvolve spatial transcriptomics data, providing the spatial distribution of cellular phenotypes for each spatial location. RCTD is capable of overcoming platform differences caused by various spatial transcriptomics technologies.

Spatial Transcriptomics Deconvolution by Topic Model (STRIDE) [33] can perform joint analysis of spatial transcriptomics sequencing and scRNA-seq data. By utilizing scRNA-seq data, this method deconvolves spatial expression data in tissue sections to identify and locate distinct cell types along with their transcriptional features. Moreover, it has the ability to construct 3D models for the same contiguous tissue. STRIDE employs scRNA-seq data to train a Latent Dirichlet Allocation model, acquiring gene-topic distributions that are subsequently utilized for deconvolving the cellular composition within spatial spots.

DSTG [29] is a graph-based convolutional neural network (GCN)-based joint analysis strategy. The DSTG algorithm generates pseudo-ST spot data from scRNA-seq data, constructs a graph using both real-ST spots data and the generated pseudo-ST spot data, and records and utilizes the real spatial information as features. Finally, DSTG employs GCN, a semi-supervised learning method, for the estimation of the cellular composition within real-ST spots.

CellDART [34] utilizes adversarial discriminative domain adaptation (ADDA) with neural networks to estimate the cell spatial distribution defined by single-cell level data integration. CellDART first constructs pseudo spatial spots using randomly selected cells from scRNA-seq data, with known cell components. Then, a neural network is built using these pseudo spots and real spots from spatial transcriptomic sequencing data. Meanwhile, a cell component classifier is also constructed. Through optimization, CellDART achieves estimation of cell proportions in real spots from spatial data.

MAPPING

The mapping method (Figure 2) is a commonly employed approach in the joint analysis of scRNA-seq and spatial

transcriptomics data. This method aims to establish a correspondence between gene expression patterns in ST data and cell type information in cell data, thereby associating gene expression information in ST data with specific cell types. This method allows us to enhance our comprehension of the distribution of various cell types and gene expression patterns within spatial transcriptomics data.

In particular, the mapping method can be divided into two steps. First, a set of gene features that can accurately distinguish different cell types needs to be determined in the cell data. Subsequently, the gene features are juxtaposed with the gene expression patterns present in the spatial transcriptome data, facilitating the identification of the cell type corresponding to each spatial location.

In practical applications, the mapping method typically requires a series of complex computational models and algorithms to be implemented. For example, some researchers use deep learning-based methods to train the mapping relationship between cell features and gene expression features, in order to achieve more accurate and refined analysis of spatial transcriptomics data. In addition, there are also some open-source tools and software available, such as scGNN [35], which can help us achieve joint analysis of cell data and ST data. Lastly, Table 2 provides a comprehensive summary of all the methods.

MIA [13] is a mapping approach akin to the principles of enrichment analysis. It integrates scRNA-seq data and ST sequencing data to deduce the degree of enrichment for cell subpopulations within the tissue space by assessing the correlation of specific genes across both data types. Specifically, MIA first identifies the gene set significantly upregulated in each cell type relative to other cells using scRNA-seq data, which is defined as cell type-specific gene set. Then, similar to the processing of scRNA-seq data, MIA identifies the gene set significantly upregulated in each spatial region relative to other regions using ST data, which is defined as tissue region-specific gene set. Next, MIA analyzes the intersection between the gene specific to each cell type set and the gene specific to each tissue region set using the MIA method to determine whether the overlap between them is higher or lower than random expectation, in order to determine the correlation degree between them and ultimately obtain the association between cell types and tissue regions. MIA has

Table 2. Summary of mapping methods

Methods	Key technique	GitHub address	Feature
MIA (2020)	Enrichment analysis	https://github.com/reubenmoncada/Multimodal-intersection-analysis-MIA	Good performance in the localization of cell types in cross-tissue regions
AddModuleScore (2020)	Enrichment analysis	https://github.com/WalterMuskovic/AddModuleScore	Focusing on scoring selected genes
Tangram (2021)	Deep Learning	https://github.com/broadinstitute/Tangram	Detect conserved cell type patterns across species
Harmony (2019)	Nonlinear dimensionality reduction and batch correction techniques	https://github.com/WalterMuskovic/AddModuleScore	Requires less computing resources
pciSeq (2020)	Probabilistic model	https://github.com/acycliq/pciSeq	Requires only low-magnification imaging
LIGER (2019)	integrative non-negative matrix factorization (iNMF)	https://github.com/welch-lab/liger	Suitable for large-scale datasets, with consideration of platform effects
SpaGE (2020)	PRECISE and kNN	https://github.com/tabdelaal/SpaGE	Adapts to large-scale datasets, considering platform effects. It has interpretability, high predictive accuracy for HVGs and does not require high computer resources
gimVI (2019)	deep generative model	https://github.com/scverse/scvi-tools	Efficient gene expression data imputation and integration capability
stPlus (2021)	reference-based method	https://github.com/xy-chen16/stPlus	Can be applied to large-scale datasets
CeLEry (2023)	feedforward neural network	https://github.com/QijhuangZhang/CeLEry	Consider the sample size challenge inherent in spatial transcriptomics datasets and provide a dedicated method for generating test set samples to enhance performance

shown promising results for localizing cell types across tissue regions.

The AddModuleScore function in the Seurat package [36] is a method used for the integrated analysis of ST and scRNA-seq transcriptomics. It scores gene sets based on scRNA-seq data to appraise the enrichment of cell subtypes. This score is then mapped onto the ST data metric for assessing cell-type enrichment in each spot of the spatial sample.

Tangram [37], a mapping approach built upon a deep learning framework, enables the integrated analysis of diverse ST data and single-cell or scRNA-seq data. It effectively assigns the most probable cell type to each spot within the spatial data. Initially, Tangram randomly assigns expression patterns derived from sc/snRNA-seq data to spatial data, simulating the correlation between sc/snRNA-seq data and gene expression at the spatial level using an objective function. Then, it reorganizes the expression data from sc/snRNA-seq to obtain a probability mapping matrix that describes the likelihood of each cell type at every spatial location.

Harmony [38] is a popular tool for integrating scRNA-seq data. Harmony utilizes nonlinear dimensionality reduction and batch correction techniques to integrate scRNA-seq datasets from different experimental conditions into a continuous expression matrix. Harmony can also be used to integrate scRNA-seq data and ST data. To do so, the scRNA-seq and spatial transcriptomics data need to be transformed into the same expression matrix format. Then, Harmony is used to perform batch correction on these two datasets to eliminate batch effects from different experimental conditions. Ultimately, the corrected data can be utilized for subsequent analyses, including clustering, cell type identification and gene expression analysis. For integrating scRNA-seq and ST data, Harmony performs a shared gene selection on the two datasets to obtain a gene list, which is used as an index to transform the scRNA-seq data and spatial transcriptomics data into two gene-cell matrices. The two

gene-cell matrices are subsequently combined, facilitating the integration of scRNA-seq and ST data into a unified expression matrix. The expression matrix undergoes batch correction to remove batch effects, and nonlinear dimensionality reduction is employed to aid in subsequent analyses, including clustering, cell type identification and gene expression analysis.

Probabilistic cell typing by in situ sequencing (pciSeq) [39] is a probabilistic model-based algorithm that utilizes probabilistic methods to integrate single-cell and ST data, enabling the identification and classification of cell types in spatial context. By combining in situ sequencing (ISS) technology with scRNA-seq, pciSeq enables spatial positioning of RNA molecules in tissues and measurement of RNA sequences for each cell, followed by the use of scRNA-seq data to establish an RNA sequence database to aid in identifying and labeling cell types. Finally, pciSeq combines the labeled cell type information with cell position information from ISS images to derive spatial maps depicting the distribution patterns of different cell types.

Specifically, the workflow of pciSeq involves clustering of scRNA-seq data to categorize cells into distinct cell types, selection of marker genes for padlock probes based on different cell types in scRNA-seq, generation of barcodes using synthesized probes for spatial amplification of specific RNA molecules in tissue slices, followed by sequencing. Next, a probabilistic model is established based on gene expression profiles of various cell types in scRNA-seq and the gene expression map obtained by location sequencing, to assign gene reads to individual cells and allocate individual cells to their corresponding cell types. Finally, spatial distribution maps of cell types are generated.

LIGER [40] is a popular tool for integrating scRNA-seq data across individuals and species to identify shared cell types and dataset-specific characteristics, facilitating a cohesive examination of diverse single-cell datasets. In addition to its application in integrating single-cell transcriptomics and spatial transcriptomics data, LIGER can be utilized for integrating

other multi-modal datasets. The core approach employed by LIGER for data integration is integrative non-negative matrix factorization (iNMF). By performing iNMF, we can obtain two factor matrices and a coefficient matrix, where one of the factor matrices captures dataset-specific factors, while the other represents shared factors. Subsequently, the two factor matrices and the coefficient matrix are merged to form a low-dimensional embedding matrix. Each cell within this matrix is characterized by a linear combination of both its dataset-specific and shared factors. Utilizing this low-dimensional embedding matrix, various downstream analyses can be performed, including clustering, visualization and differential expression analysis.

Spatial Gene Enhancement (SpaGE) [41] predicts the expression of unmeasured genes in spatial transcriptomics data by matching tissue scRNA-seq data. SpaGE embeds spatial transcriptomics and scRNA-seq data using a joint embedding approach to establish a cohesive representation. SpaGE consists of two steps: first, PRECISE is used to compare the input ST data set and the scRNA-seq data set, and then the k-nearest-neighbor algorithm predicts the expression of unmeasured genes, leveraging the comparison results. Specifically, ST data and scRNA-seq data are input as query and reference matrices, respectively. Then, the PRECISE algorithm identifies the genes that are shared by these two datasets as the new dataset and performs principal component analysis on the three datasets, resulting in two independent sets of principal components (PCs). SpaGE employs cosine similarity matrix computation and SVD decomposition to compare the PCs, followed by utilizing the k-nearest-neighbor algorithm to estimate the expression of unmeasured genes in the spatial data based on the comparison results.

Gene imputation with Variational Inference (gimVI) [42] is a linear deep generative model based on deep learning. When integrating scRNA-seq data and ST data, gimVI treats this two datasets as two different perspectives of observations and models them using a joint variational autoencoder (VAE). Specifically, gimVI considers scRNA-seq data as the gene expression view and spatial transcriptomics data as the spatial position view. It models each view using a VAE and represents the distribution of the data using a Gaussian distribution. In addition, gimVI introduces a shared variable between the two views to facilitate information exchange between the two datasets. Through training the deep model, missing values can be imputed and ST data and scRNA-seq data can be integrated. In this way, gimVI can perform both imputation and integration of ST data and scRNA-seq data simultaneously.

stPlus [43] is a data integration method based on multivariate statistics and machine learning, aiming to capture a more comprehensive image of tissue architecture and cell types by combining scRNA-seq data with spatial transcriptomics data. The three main steps of the stPlus technique are data processing, joint embedding and prediction of gene expression in spatial data for genes that are not directly measured, utilizing cell embedding and reference scRNA-seq data. Specifically, data processing selects the top 2000 variable genes from the reference scRNA-seq data, and both spatial transcriptomics sequencing data and scRNA-seq data are input as gene-cell matrices, with the gene order of spatial transcriptomics sequencing data consistent with the processed scRNA-seq data, and the missing values filled with 0. The joint embedding step is pivotal in the stPlus method, involving the utilization of an autoencoder and a loss function to perform cell embedding for both spatial transcriptomics data and reference scRNA-seq data. Lastly, the stPlus method employs the weighted KNN algorithm to predict gene expression in spatial transcriptomics data where measurements are lacking.

CeLery [44] employs a mapping-based data integration methodology, utilizing a feedforward neural network. Initially, it extracts the association between gene expression and spatial coordinates from spatial transcriptomics sequencing data, subsequently applying this acquired model to scRNA-seq data, thereby deconvolving spatial location data from scRNA-seq data. Additionally, this approach addresses situations involving limited training sample sizes in the spatial transcriptomics (ST) dataset by offering users an optional data augmentation procedure grounded in VAEs. This process generates replicates of the ST data that are well-suited for training neural networks.

CONCLUSION AND OUTLOOK

From a comprehensive perspective, the methods for integrating single-cell and spatial transcriptome data are still being continuously developed and improved. Each method has its own advantages, disadvantages and applicable scenarios. Specific choices need to be considered based on specific circumstances. We have compiled 19 integration methods and established a GitHub repository based on categorization: <https://github.com/ChaoruiYan019/Integration-Tool-for-scRNA-seq-Data-and-Spatial-Transcriptomics-Sequencing-Data>. In practical applications, it may be necessary to select suitable methods for data integration and analysis based on experimental purposes and data characteristics.

At the same time, it should be noted that integrating single-cell and spatial transcriptome data also faces some challenges and limitations, such as large data volume, data noise, batch effects, spatial resolution, etc. It is necessary to adopt appropriate data preprocessing and quality control methods, as well as develop more accurate algorithms and tools for analysis and interpretation.

Concurrently, it is imperative to recognize the numerous challenges and limitations associated with integrating single-cell and spatial transcriptomic data. Successful integration mandates data standardization and calibration across different batches and platforms to ensure compatibility for merging and comparison. Crucially, as the dimensions of the generated data expand, enhancing the applicability of integration algorithms to large-scale datasets becomes paramount. Furthermore, the interpretation of the final integrated data is essential to unveil the spatial distribution patterns of cell types and gene expressions within tissues. This endeavor demands a profound comprehension of the data's significance and its integration with existing biological knowledge.

Moreover, the resolution of spatial transcriptome sequencing technology is steadily advancing. Initially, spatial transcriptome sequencing techniques heavily depended on methods like *in situ* hybridization or optical microscopy image analysis, yielding resolutions typically in the range of hundreds of micrometers to a few millimeters. Yet, the advent of innovative spatial transcriptome sequencing technologies, such as DBiT-seq [45] and Stereo-CITE-seq [46], has raised resolutions to subcellular or even single-cell levels. In the context of integrating scRNA-seq and spatial transcriptomic data, if spatial transcriptomic technology advances to the point of directly supplying sequencing data at scRNA-seq depth, this challenge will be effectively addressed. Nonetheless, these technologies currently cannot autonomously replace single-cell data. For example, despite DBiT-seq sequencing technology achieving a 10-micrometer pixel resolution, it remains unable to directly analyze individual cells. Therefore, in the near future, research on integrating single-cell and spatial transcriptomic data remains both valuable and essential.

Currently, artificial intelligence (AI) technologies have achieved notable success in the field of bioinformatics, and the integration of single-cell and spatial transcriptomic data has propelled advancements across various biological domains. For instance, the team led by Xianliang Hou, through the integration of single-cell and spatial transcriptomic data, has elucidated that endometrial carcinoma cells can shape the tumor microenvironment by suppressing immune cells and modulating it through MDK-NCL signal [47]. In another instance, Reuben Moncada's team, utilizing the MIA technique to integrate single-cell and spatial transcriptomic data, has unveiled the tissue architecture of PDAC, offering novel perspectives for cancer intervention [48]. Furthermore, Jianfei Zhu's team, through the concurrent integration of these two data types, has depicted the dynamic evolution from early stage tumors to invasive lung adenocarcinoma (LUAD), thereby advancing research in personalized treatment strategies for LUAD [49]. The development of AI technologies has not only driven progress in the field of bioinformatics but also provided robust solutions to the challenges encountered within the realm of bioinformatics.

From the research outlined in this review and previous studies, it is evident that among the 19 integration methods previously mentioned, the Tangram, MIA, Cell2location and Seurat integration algorithms have consistently been popular since their introduction [50]. Presently, several emerging technologies, like CelEry [44] and EnDecon [51], have recently been introduced. Although these technologies have not gained widespread adoption, their strong performance in testing indicates their future potential. Regarding the earlier integration techniques, they retain their reference value and can offer guidance to future researchers. For example, EnDecon enhances its strategy by amalgamating results from 14 individual integration methods to achieve the optimal outcome.

In summary, the integration of scRNA-seq data and spatial transcriptomic data remains an important research area with significant implications for understanding tissue and cellular functions, development and diseases. Deep learning algorithms have shown tremendous potential in integrating these two types of sequencing data and have made encouraging progress in the past few years. With ongoing model optimization and improvements, deep learning algorithms are expected to achieve higher accuracy and reliability in data integration. The performance of integrating scRNA-seq and spatial transcriptomic data using deep learning algorithms is expected to further improve. Furthermore, deep learning algorithms have the potential to integrate multi-modal data effectively, allowing for the handling of platform effects arising from different sequencing platforms. We anticipate that deep learning algorithms will have broader applications and achieve technological breakthroughs, contributing to the integration of single-cell and spatial transcriptomic data and providing new opportunities and challenges for life science research.

Key Points

- This review article presents 19 methods that aim to integrate single-cell RNA sequencing data and spatial transcriptomics data. The methods are classified into two main groups and described accordingly.
- Taking into account the influence of highly variable genes (HVGs) on the integration perspective, the incorporation of HVGs in these techniques is documented.

- Each integration method is explained in terms of its main ideas, techniques and characteristics. Additionally, relevant literature pertaining to each integration technique is compiled.

Funding

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

References

1. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015;**16**(3):133–45.
2. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat Rev Genet* 2021;**22**(10):627–44.
3. Zhang Z, Cui F, Wang C, et al. Goals and approaches for each processing step for single-cell RNA sequencing data. *Brief Bioinform* 2021;**22**(4):bbaa314.
4. Zhang Z, Cui F, Lin C, et al. Critical downstream analysis steps for single-cell RNA sequencing data. *Brief Bioinform* 2021;**22**(5):bbab105.
5. Gao Y, Cai GY, Fang W, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun* 2020;**11**(1):5033.
6. Wu Y, Zhang K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nat Rev Nephrol* 2020;**16**(7):408–21.
7. Ramsköld D, Luo S, Wang Y-C, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 2012;**30**(8):777–82.
8. Klein AM, Macosko E. InDrops and drop-seq technologies for single-cell sequencing. *Lab Chip* 2017;**17**(15):2540–1.
9. Zhang Z, Cui F, Cao C, 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**:105092.
10. Villani AC, Satija R, Reynolds G, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science (New York, NY)* 2017;**356**(6335):eaah4573.
11. Tirosh I, Venteicher AS, Hebert C, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendrogloma. *Nature* 2016;**539**(7628):309–13.
12. Zhang Z, Cui F, Zhou M, et al. Single-cell RNA sequencing analysis identifies key genes in brain metastasis from lung adenocarcinoma. *Curr Gene Ther* 2021;**21**(4):338–48.
13. Moncada R, Barkley D, Wagner F, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* 2020;**38**(3):333–42.
14. Moffitt JR, Zhuang X. RNA imaging with multiplexed error-robust fluorescence in situ hybridization (MERFISH). *Methods Enzymol* 2016;**572**:1–49.
15. Shah S, Lubeck E, Zhou W, Cai L. In situ transcription profiling of single cells reveals spatial Organization of Cells in the mouse hippocampus. *Neuron* 2016;**92**(2):342–57.
16. Wang X, Allen WE, Wright MA, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018;**361**(6400):eaat5691.
17. Rusk N. Spatial transcriptomics. *Nat Methods* 2016;**13**(9):710–10.

18. Rodriques SG, Stickels RR, Goeva A, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 2019;**363**(6434):1463–7.
19. Hu KH, Eichorst JP, McGinnis CS, et al. ZipSeq: barcoding for real-time mapping of single cell transcriptomes. *Nat Methods* 2020;**17**: 833–43.
20. Zhang Z, Cui F, Su W, et al. webSCST: an interactive web application for single-cell RNA-sequencing data and spatial transcriptomic data integration. *Bioinformatics (Oxford, England)* 2022;**38**(13):3488–9.
21. Elosua-Bayes M, Nieto P, Mereu E, et al. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res* 2021;**49**(9):e50.
22. Osorio D, Yu X, Zhong Y, et al. Single-cell expression variability implies cell function. *Cell* 2019;**9**(1):14.
23. Satija R, Farrell JA, Gennert D, et al. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**(5): 495–502.
24. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* 2016;**5**:2122.
25. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol* 2015;**11**(6): e1004333.
26. Yip SH, Sham PC, Wang J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform* 2019;**20**(4):1583–9.
27. Kleshchevnikov V, Shmatko A, Dann E, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* 2022;**40**(5):661–71.
28. Li H, Li H, Zhou J, Gao X. SD2: spatially resolved transcriptomics deconvolution through integration of dropout and spatial information. *Bioinformatics* 2022;**38**(21):4878–84.
29. Song Q, Su J. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief Bioinform* 2021;**22**(5):bbaa414.
30. Dong R, Yuan GC. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol* 2021;**22**(1):145.
31. Andersson A, Bergenstrahle J, Asp M, et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol* 2020;**3**(1):565.
32. Cable DM, Murray E, Zou LS, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* 2022;**40**(4):517–26.
33. Sun D, Liu Z, Li T, et al. STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. *Nucleic Acids Res* 2022;**50**(7):e42.
34. Bae S, Na KJ, Koh J, et al. CellDART: cell type inference by domain adaptation of single-cell and spatial transcriptomic data. *Nucleic Acids Res* 2022;**50**(10):e57.
35. Wang JX, Ma AJ, Chang YZ, et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat Commun* 2021;**12**(1):1882.
36. Ji AL, Rubin AJ, Thrane K, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* 2020;**182**(2):497–514.e22.
37. Biancalani T, Scalia G, Buffoni L, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with tango. *Nat Methods* 2021;**18**(11):1352–62.
38. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**(12):1289–96.
39. Qian X, Harris KD, Hauling T, et al. Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nat Methods* 2020;**17**(1):101–6.
40. Welch JD, Kozareva V, Ferreira A, et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 2019;**177**(7):1873–1887.e17.
41. Abdelaal T, Mourragui S, Mahfouz A, Reinders MJT. SpaGE: spatial gene enhancement using scRNA-seq. *Nucleic Acids Res* 2020;**48**(18):e107.
42. Lopez R, Nazaret A, Langevin M, Samaran J, Regier J, Jordan MI, Yosef N: A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. arXiv preprint arXiv:1905.02269, 2019. <https://doi.org/10.48550/arXiv.1905.02269>.
43. Shengquan C, Boheng Z, Xiaoyang C, et al. stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics* 2021;**37**:i299–307.
44. Zhang Q, Jiang S, Schroeder A, et al. Leveraging spatial transcriptomics data to recover cell locations in single-cell RNA-seq with CeLery. *Nat Commun* 2023;**14**(1):4050.
45. Liu Y, Yang M, Deng Y, et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* 2020;**183**(6):1665–1681.e18.
46. Qie J, Liu Y, Wang Y, et al. Integrated proteomic and transcriptomic landscape of macrophages in mouse tissues. *Nat Commun* 2022;**13**(1):7389.
47. Hou X, Yang Y, Li P, et al. Integrating spatial Transcriptomics and single-cell RNA-seq reveals the gene expression Profiling of the human embryonic liver. *Front Cell Dev Biol* 2021;**9**:652408.
48. Moncada R, Barkley D, Wagner F, et al. Author correction: integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* 2020;**38**(12):1476–6.
49. Zhu J, Fan Y, Xiong Y, et al. Delineating the dynamic evolution from preneoplasia to invasive lung adenocarcinoma by integrating single-cell RNA sequencing and spatial transcriptomics. *Exp Mol Med* 2022;**54**(11):2060–76.
50. Li B, Zhang W, Guo C, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat Methods* 2022;**19**(6): 662–70.
51. Tu JJ, Li HS, Yan H, Zhang XF. EnDecon: cell type deconvolution of spatially resolved transcriptomics data via ensemble learning. *Bioinformatics* 2023;**39**(1):btac825.