Chapter 7
Comparative Analysis of Packages and Algorithms for the Analysis of Spatially Resolved Transcriptomics Data

Natalie Charitakis, Mirana Ramialison, and Hieu T. Nim

7.1 Introduction

Despite the natural stochasticity that can disrupt biological processes such as organ development, biological systems consistently produce the same gene expression pattern with sufficient robustness such that the embryo forms correctly (nearly) every time. Furthermore, the genes typically work together in networks, requiring a systems-wide transcriptomic approach to fully understand the spatial expression patterns. Many of these create well-defined regions of cells within developing tissues that can be easily reproduced, demonstrating how the spatial location of the gene regulatory networks is critical for the proper formation of tissues (Exelby et al. 2021). Determining these networks is an active study area in the emerging field of 'spatial biology', and calls for specialised computational techniques, many of which have been developed very recently.

The merits and limitations of single-cell RNA Sequencing (scRNA-Seq) have been well established (Hwang et al. 2018; Chen et al. 2019) and the method

Co-senior authors: Mirana Ramialison and Hieu T. Nim.

N. Charitakis
Murdoch Children's Research Institute, Parkville, VIC, Australia

Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia
e-mail: natalie.charitakis@mcri.edu.au

M. Ramialison (✉) · H. T. Nim (✉)
Murdoch Children's Research Institute, Parkville, VIC, Australia

Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia

Australian Regenerative Medicine Institute and Systems Biology Institute Australia, Monash University, Clayton, VIC, Australia
e-mail: mirana.ramialison@mcri.edu.au; hieu.nim@mcri.edu.au

G. A. Passos (ed.), *Transcriptomics in Health and Disease*,
https://doi.org/10.1007/978-3-030-87821-4_7

successfully applied across varying organs and conditions (Karaayvaz et al. 2018; Regev et al. 2017; Dong et al. 2018; He et al. 2020; Ximerakis et al. 2019; Tiklová et al. 2019; Zhou et al. 2021). scRNA-Seq is capable of identifying rare cell populations, including in disease states and developmental stages; however, the method yields noisy, variable data with lots of technical variation (Chen et al. 2019). Despite scRNA-Seq allowing for the study of cellular heterogeneity and cell type hierarchy, the loss of spatial information prevents the systematic study of physiological structure/function relationships in various tissues and organs. This was part of the drive in the development of spatial transcriptomics (ST) (Marx 2021) (now commercialised by 10x Genomics under the name Visium) and other spatially resolved transcriptomics (SRT) methods. The spatially resolved gene expression pattern within the context of a tissue is critical to achieving a full understanding of disease states and tissue development and function and the ability to investigate this is achievable using SRT (Ståhl et al. 2016).

Spatial transcriptomics is an area that is becoming more widely used and will continue to expand in the upcoming years (Marx 2021). Having been featured as Nature's 'Method of the Year' in 2020, the technology and the analytical opportunities it provides are going to keep growing rapidly (Marx 2021). As demonstrated in Fig. 7.1, the number of papers published on spatial transcriptomics has greatly increased since 2016, when the first technology named 'spatial transcriptomics' was published (Ståhl et al. 2016). Offering unprecedented spatial context to transcriptomic data presents an invaluable tool for studying tissues and their cellular

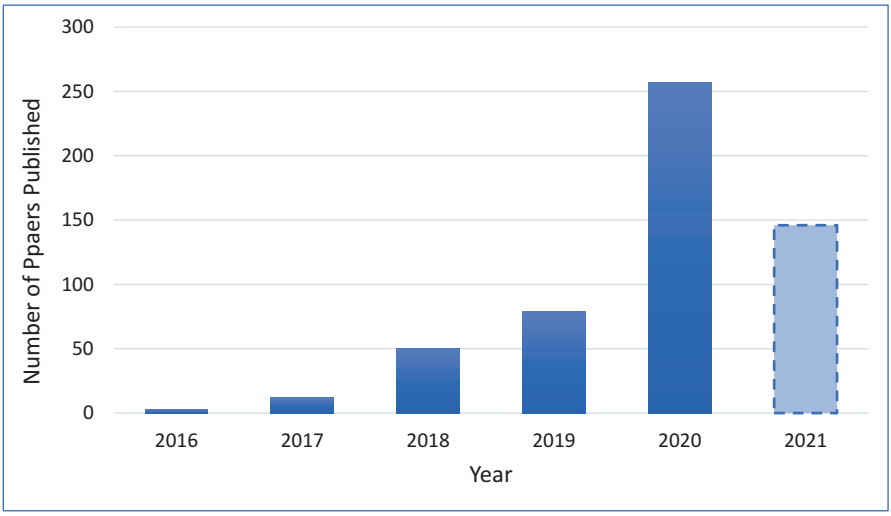

Fig. 7.1 Number of papers returned when a search was performed using the keywords 'Spatial Transcriptomics' using the software 'Publish or Perish'(Harzing 2016) to search PubMed and to manually search bioRvix, with the additional parameter of papers published from 01/01/2016 to 16/04/2021. Papers identified by searching both databases were consolidated; note that this is not a comprehensive view of all papers published on the topic since 2016. Bars in light blue with a dotted outline indicate that not all papers for the calendar year have been included

composition. As early as 2017, the merits of applying SRT to the discovery of spatial organisation of gene expression to improve transcriptional classification of cell types and localisation within a tissue had been discussed and even put to the test (Lein et al. 2017; Shah et al. 2016). The potential applications of this technology are continuously improving and expanding, as demonstrated by the integration of different methods to improve the resolution of current SRT methods (Moncada et al. 2020). The different techniques available to generate SRT data and their merits have been discussed (Lein et al. 2017; Crosetto et al. 2015; Asp et al. 2020; Waylen et al. 2020), but a review of data analysis tools is as of yet lacking. With an emphasis on obtaining spatially resolved data sets with single-cell resolution (Marx 2021), the method, aims and approaches to integrate and analyse the data generated are still in flux, with a clear 'gold standard' yet to distinguish itself. This chapter discusses some of the current packages and pipelines available to perform this analysis (Table 7.1).

7.2 Methods for Downstream Analysis of Spatially Resolved Transcriptomics Data

As identifying the spatial expression patterns of genes and how they vary across a tissue is a critical aim of spatial transcriptomics, many purpose-built tools for analysis of this data aim to identify spatially variable genes (SVGs) (Box 7.1) (Exelby et al. 2021). Building on the concept of highly variable genes in scRNA-Seq analysis, SVGs have a pattern of expression that depends on their location in the tissue and can give insight into biological function (Svensson et al. 2018). A complication of analysing these spatial transcriptomics data sets is accurately accounting for the spatial correlation across samples (Li et al. 2021), and different methods can be employed to tackle this problem. Various packages have been developed in primarily R or Python and are currently available to identify SVGs in spatial transcriptomic data sets.

Box 7.1
One key aim of analysing RNA-Seq and scRNA-Seq datasets is to identify differentially expressed genes (DEGs) between two groups from within a group of highly variable genes (HVGs). DEGs are identified between two groups when a gene's expression is statistically significantly different between the two groups present (Exelby et al. 2021). While this approach has yielded many important findings, it removes organisational context from the groups in question, something that can be recovered using spatial transcriptomics (Marx 2021). This new technology has shifted the goalposts for transcriptomics analysis, resulting in many bioinformatics packages dedicated to discovering *spatially variable genes* (SVGs) (Svensson et al. 2018; Li et al. 2021; Sun et al. 2019; Edsgärd et al. 2018; Hao et al. 2021; Zhang et al. 2018). As the name suggests, these genes will have amplified expression in certain regions of the tissue or sample, often displaying an underlying pattern (Svensson et al. 2018; Hu et al. 2020). Determining the best method to achieve the most biologically accurate results and computational efficiency is challenging, and research in this area is ongoing.

Table 7.1 Comparison of computational packages for analysis of spatial transcriptomics data sets

Purpose	Package name	Main method	Implementation	Pros	Cons	GitHub
Identifying SVGs	SpatialDE	GP Regression	Python	Currently most popular package in this category	Labels genes as SVGs that have very low expression and two normalisation steps	https://github.com/Teichlab/SpatialDE
	SPARK	Generalised Linear Spatial Model	R	Does not require data to be normalised and controls for type I error	Accuracy not significant improvement on SpatialDE	https://github.com/xzhoulab/SPARK
	Trendsceek	Marked Point Process	R	Low false positives reported	Identifies very low number of SVGs and ineffective on larger data sets	https://github.com/edsgard/trendsceek
	BOOST-GP	Bayesian Hierarchical Model	R	Accuracy rate is better than other packages in data sets with many 0 counts	Accuracy rate still low in the presence of many 0 counts	https://github.com/Minzhe/BOOST-GP
	SOMDE	SOM	Python	Able to efficiently identify SVGs even in very large data sets	In low dropout rate datasets not as good as SpatialDE	https://github.com/WhirlFirst/somde
	scGCO	Graph Cut Algorithm	Python	Results were more reproducible than SpatialDE. Can be used on data sets with millions of cells	~35% of labelled SVGs not reproducible	https://github.com/WangPeng-Lab/scGCO

Identifying SVGs + other capabilities	GP counts	GP Regression	Python	Can determine temporal trajectories and perform pseudotime analysis	Efficiency on larger data sets unclear	https://github.com/ManchesterBioinference/GPcounts
	STUtility	Spatial Autocorrelation	R	Image processing and the ability to create a 3D model from multiple samples	Accuracy in identifying SVGs and defining tissue heterogeneity not comprehensively reviewed	https://github.com/jbergenstrahle/STUtility
Assigning lost transcripts	Sparcle	MLE	Python	Unique capability and can be used alongside other packages	Developed specifically for smFISH	https://github.com/sandhya212/Sparcle_for_spot_reassignments
Cell type identification	SpatialDWLS	DWLS	R	*A priori* knowledge can be incorporated	Performance not validated to other packages on real datasets	https://github.com/rdong08/spatialDWLS_dataset
	FICT	Generative Mixture Model	Python	Addresses problem of over-reliance on expression data	Performance drops on data sets with less cells	https://github.com/haotianteng/FICT
	RCTD	Supervised Learning	R	Can normalise across platform effects of scRNA-Seq and SRT data sets	Requires well-annotated scRNA-Seq data sets	https://github.com/dmcable/RCTD

(continued)

Table 7.1 (continued)

Purpose	Package name	Main method	Implementation	Pros	Cons	GitHub
Spot-to-spot clustering	SpatialCPie	Different Clustering Algorithms	R	Can perform clustering at different resolutions for different subtypes of tissue; cluster graph is a novel method of visualising cluster origin in SRT	Validation against other packages lacking	https://github.com/jbergenstrahle/SpatialCPie
Pipeline	Giotto	–	R	Choice of algorithm for identifying marker genes in cell types, dedicated pipelines for lower resolution SRT data	Validation against different biological tissues collected on different platforms lacking	https://github.com/RubD/Giotto
	Squidpy	–	Python	Modular, so can incorporate other packages in analysis	Cellular neighbourhoods not very reproducible	https://github.com/theislab/squidpy

7.2.1 Identifying Spatially Variable Genes

Among these, SpatialDE is a popular package based on Gaussian process (GP) regression, which can clearly identify localized gene expression patterns for data sets containing temporal and/or spatial annotations (Svensson et al. 2018). SpatialDE can recognise SVGs by creating a model with two different terms reflecting distinct variance present in the data set. The first term captures the non-spatial variance present within the data, while the second aims to capture the spatially related variance of gene expression within the data set, with the assumption that the covariance between a cell's gene expression profile decreases with an increase in distance between the cells (Svensson et al. 2018). A ratio calculated using these terms can then be used as a measure of the level of gene expression variance attributable to spatial location (Svensson et al. 2018). These are the key parameters used to fit the Gaussian model in a computationally efficient manner (Svensson et al. 2018). Testing to prove whether statistically significant SVGs are present is performed by comparing this model to a second one which lacks the spatial covariance parameter that represents a data set in which spatial localisation has no effect on gene expression patterns (Svensson et al. 2018). This process is repeated for each gene, and after correcting for multiple testing, the SVGs can be pulled out of the data set (Svensson et al. 2018). SpatialDE has the capability of taking this a step further by creating models with different covariance functions for SVGs and comparing them, this is in addition to the initial 10 Gaussian kernels it tests before selecting that with the lowest p-value. This creates the ability to determine whether each SVGs is most accurately expressed as a linear, periodic or general expression model (Svensson et al. 2018). However, for the data to fit certain underlying assumptions of the Gaussian model, two normalisation steps are performed, the first being a variance stabilising transformation (Svensson et al. 2018; Sun et al. 2019). It may affect the package's performance as the assumptions underlying the model and the necessary data transformations do not truly reflect the nature of the data (Li et al. 2021). A further functionality of SpatialDE is that it can implement an unsupervised learning technique built on the Gaussian Mixture Model to apply automatic expression histology (AEH), which can group together SVGs by their spatial expression pattern using hidden patterns learnt from the data (Svensson et al. 2018). The observation that SpatialDE may introduce false positives by labelling genes with low levels of expression as SVGs is an area which requires further investigation and can be improved upon in future releases of the package (Sun et al. 2019).

A package with the same goal as SpatialDE is SPARK (Spatial Pattern Recognition via Kernels), which employs a generalised linear spatial model (GLSM) with different spatial kernels to identify SVGs (Sun et al. 2019). This model was built on previous work to take into consideration the effects of spatial correlation and covariate measurement error; it was built and tested on 2D data; however, it is capable of being expanded to 3D data sets (Sun et al. 2019). As in the case of SpatialDE, SPARK models gene expression for each gene across all spatial coordinates; however, this model operates under the assumption that the spatial data is

non-Gaussian (Sun et al. 2019). SPARK builds on other GLSMs by developing a hypothesis testing framework for the model (Sun et al. 2019). The power of this hypothesis testing is linked to how the spatial kernel function accurately represents the spatial pattern of the gene represented in the model; and as different gene expression patterns will most accurately be represented by different spatial kernel functions, SPARK considers 10 different kernels (similarly to SpatialDE) based on commonly observed biological patterns (Sun et al. 2019). Due to the heuristic nature of these kernels, this process could introduce biases that lead that package to choose more commonly observed biological patterns. SPARK can work with large data sets as it employs a penalised quasi-likelihood (PQL) algorithm for parameter estimation to circumvent the problem of the difficulty in solving GLSMs in short periods of time; this algorithm informs the parameters used in each of the spatial kernel functions. It further improves on the packages available at the time of publication, SpatialDE and Trendsceek, by not performing a normalisation step on the data, which decreases the power of the analysis (Sun et al. 2019). A drawback of SpatialDE that SPARK corrects for is to control for type 1 errors through the Cauchy combination rule, thus giving it additional power when identifying SVGs (Sun et al. 2019). The Cauchy combination rule groups the p-values generated from each spatial kernel function into a single p-value while still controlling for type 1 errors, which results in a single p-value per gene (Sun et al. 2019). The final steps involve controlling for FDR across all p-values and then determining which are SVGs (Sun et al. 2019). While SpatialDE and SPARK share the use of parametric test statistics, there are a few critical differences between the packages (Sun et al. 2019). As previously mentioned, SPARK does not model normalised data, while SpatialDE can only approximate p-values; SpatialDE first calculates an exact p-value per gene; and once it obtains the initial set of statistically significant genes, SpatialDE then performs additional analysis to determine their p-values (Sun et al. 2019). Furthermore, when validated against multiple data sets, it performed just as well or better than SpatialDE and Trendsceek (described in next paragraph) (Sun et al. 2019). When its ability to calculate true positives in two simulated data sets was tested across a total of six different spatial expression patterns with varying FDRs, SPARK outperformed Trendsceek and had better results than SpatialDE (Sun et al. 2019). While with certain simulated data sets, SPARK and Trendsceek performed similarly in computing well-calibrated p-values, but SpatialDE did not identify certain SVGs present (Sun et al. 2019). While the SPARK paper only tests the package's performance against SpatialDE and Trendsceek, it outperformed both in terms of the number of SVGs identified when validated against a spatial transcriptomics mouse olfactory bulb data set (Sun et al. 2019). However, not all genes identified by SpatialDE overlapped with those identified by SPARK (Sun et al. 2019). Despite this, the newly identified SVGs are in line with markers specific to the tissue they were annotated in, and GO enrichment analysis adds further confidence that the majority of these newly identified SVGs are biologically relevant (Sun et al. 2019). In terms of computational efficiency, when running with 10 parallel CPU threads, SPARK was more computationally efficient than the same analysis run on a single-threaded SpatialDE (although the difference in this instance is minimal) and

Trendsceek; its single-threaded performance is consistently less efficient than SpatialDE across 4 datasets of varying sizes (Sun et al. 2019).

Trendsceek is one of the earlier packages developed to identify SVGs using a non-parametric approach (Edsgärd et al. 2018). Trendsceek individually assesses each gene and normalises its expression through a log10 transformation (Edsgärd et al. 2018). It relies on a marked point process to model gene expression and cell location and later will test the null hypothesis by generating four non-parametric test statistics (Edsgärd et al. 2018). These four test statistics yield four p-values and a gene with a minimum of 1 p-value ≤ 0.05 after adjustment for multiple testing using the Benjamini-Hochberg method is determined to be an SVG (Edsgärd et al. 2018). A key difference that separates Trendsceek from SpatialDE and SPARK is its computing of non-parametric test statistics, meaning it lacks an underlying generative model. Trendsceek was tested against simulated data sets, and it demonstrated very low power to identify SVGs when they were present if less than 5% of cells in the data set had varying levels of expression (Edsgärd et al. 2018). This implies that as SRT datasets continue to increase in size, Trendsceek will not be able to distinguish SVGs present in a very small subset of cells within a tissue. When Trendsceek's performance in identifying SVGs across two spatial transcriptomics data sets is compared to SpatialDE and SPARK, it identified fewer SVGs, with numbers almost 10 times lower than the other packages (Sun et al. 2019). When compared to different packages in other studies, Trendsceek struggled to identify SVGs in real datasets, while other packages were able to (Sun et al. 2019).

Each new package developed aims to address the shortcomings of those already published; for example, BOOST-GP claims that many popular substitutes such as SpatialDE, SPARK and Trendsceek do not account for the substantial proportion of zero counts present in the data set and the effect the sparsity of the data can have analysis (Li et al. 2021). Therefore, BOOST-GP puts forth a new Bayesian hierarchical model aimed at accounting for the considerable number of zero counts present in spatial data sets, that other packages published up to this point had neglected (Li et al. 2021). A key difference to other packages is that BOOST-GP employs a negative binomial distribution when modelling count data, which should account for its observed over-dispersion (Li et al. 2021). This resembles the methods used by popular bulk RNA-Seq analysis packages rather than other spatial transcriptomics packages explored thus far (Li et al. 2021). BOOST-GP's performance was compared to that of SpatialDE's, SPARK and Trendsceek when there were false zeros present in the data, and BOOST-GP was clearly most adept at handling this complication, even if it still presented significant difficulties in retrieving a good Matthews correlation coefficient (used to determine the tool's accuracy) on a synthetic data set (Li et al. 2021). Furthermore, depending on the spatial pattern of the expression of the gene, the accuracy of BOOST-GP can differ slightly (Li et al. 2021). Alternatively, when the tool was tested on two real data sets, it was found that SPARK identified more SVGs than BOOST-GP; however, SpatialDE discovered the least (Li et al. 2021). In the analysis of human breast cancer data, despite identifying fewer SVGs than SPARK, BOOST-GP was able to identify novel, biologically relevant terms in the GO analysis, adding to its value in the analysis of SRT data (Li et al. 2021).

As larger datasets become increasingly common, packages must be created to efficiently analyse the vast amounts of data generated by SRT experiments. One of the newer packages is SOMDE (Hao et al. 2021). Built-in python, SOMDE aims to identify SVGs in large-scale datasets (Hao et al. 2021). By using a self-organising map (SOM) neural network and a Gaussian process to model the data, it can identify SVGs in large datasets much faster than SpatialDE, SPARK or Trendsceek (Hao et al. 2021). This is achieved as the data is organised into different nodes by the SOM neural network, the Gaussian process is used at the level of the nodes to identify the SVGs present in the data (Hao et al. 2021). The organisation of data into nodes minimises the sample space while preserving the original spatial organisation and expression data (Hao et al. 2021). The next stage which uses a Gaussian process identifies the SVGs from the reduced sample space (Hao et al. 2021). As seen in packages such as SpatialDE and BOOST-GP, the Gaussian process is a popular method for identifying SVGs (Svensson et al. 2018; Li et al. 2021). SOMDE also uses a log ratio test similar to that employed by SpatialDE to test the statistical significance of the spatial expression variability of each gene (He et al. 2020). When SOMDE was applied to discover the SVGs of five different data sets, it was able to do so without significant increase in computational time as the size of the data set increased, yielding results in under 5 min for the largest data set with over 20,000 data sites (Hao et al. 2021). It also demonstrated a faster running time compared to Giotto and SpatialDE on three differently sized data sets used for validation (Hao et al. 2021). Despite this, the package lacks validation on a data set of single-cell resolution (Hao et al. 2021). When its performance was compared to scGCO and SpatialDE on a simulated data set, SOMDE consistently outperformed scGCO but only had an improved performance compared to SpatialDE when a high dropout rate is incorporated into the data set (Hao et al. 2021). When its performance was compared to real data sets, most of the SVGs identified by SOMDE overlap with those identified by packages like scGCO, SPARK and SpatialDE (Hao et al. 2021).

Other methods have been developed to identify SVGs that differ from those presented thus far. One of these methods has been implemented in a python package called scGCO, which employs graph cut algorithms to identify SVGs (Zhang et al. 2018). scGCO first produces a graph by performing a Delaunay triangulation in which only true cell neighbours are connected by edges, allowing an accurate representation of cellular interactions in a sparse graph which is not memory intensive (Zhang et al. 2018). Subsequently, Voronoi diagrams are created which have previously been used to model cells (Zhang et al. 2018). Using a Markov random field (MRF) model and adapting methods traditionally used in object identification in images, scGCO can classify cells into two categories which provide efficient, low polynomial time computing and a result which is globally optimal (Zhang et al. 2018). Much like SpatialDE, scGCO employs Gaussian Mixture modelling but uses it to classify each gene's expression to ensure more accurate classification of cell types based on their gene expression (Svensson et al. 2018; Zhang et al. 2018). The performance of SVG was tested against a spatial transcriptomics data set obtained from a mouse olfactory bulb and compared to results obtained from the same data by SpatialDE (Zhang et al. 2018). A more comprehensive review of scGCO against

different packages would be beneficial to obtain a holistic understanding of its improved performance in SVG detection. scGCO successfully identified over 1,000 additional SVGs compared to SpatialDE, and at an FDR cut-off of 0.01 rather than 0.05 (Zhang et al. 2018). The majority of SVGs identified by scGCO were also identified by SpatialDE, and each formed its own spatial pattern. These results were consistent when validation was repeated across replicate mouse olfactory bulb data (Zhang et al. 2018). However, while scGCO yielded a smaller number of unreproducible SVGs across the different replicate data sets than SpatialDE, ~35% of identified SVGs were still unreproducible (an 11% reduction from SpatialDE) (Zhang et al. 2018). If replicate data sets are available for studies, then this is something that should be investigated further across all packages, resulting in the exclusion of non-reproducible SVGs for a more accurate final subset of SVGs. Additionally, when comparing between regions of the mouse olfactory bulb, scGCO was more adept at identifying SVGs than SpatialDE, while neither method entirely recovered all marker genes reported in the study which published the data set (Zhang et al. 2018). Additional validation was performed using data from breast cancer biopsies, with scGCO having a similar improved performed compared to SpatialDE when employed on the mouse olfactory bulb data set. Furthermore, the SVGs identified by SpatialDE within the breast cancer data set did not maintain consistent clustering pattern (Zhang et al. 2018). scGCO's performance on other spatial transcriptomics data sets was equally as robust (Zhang et al. 2018). scGCO also performed better in terms of computational time and memory required than SpatialDE and Trendsceek when used to analyse a simulated data set with up to a million cells.

7.2.2 Identifying Spatially Variable Genes and More

As evidenced by the packages reviewed so far, GPs are a popular method for analysing spatial transcriptomics data as they can model its spatial dependence. To this end, as new packages are developed, many are built on alternative GP regression models, such as GPcounts (BinTayyash et al. 2020). GPcounts can be used to model either spatial or temporal large-scale scRNA-Seq data through modelling count data using a negative binomial (NB) likelihood (BinTayyash et al. 2020). The NB likelihood model should more accurately capture the distribution of gene expression data compared to Gaussian likelihood model as it accounts for possible heteroscedastic noise and the presence of many zero-counts but requires UMI normalisation to be applied (BinTayyash et al. 2020). Furthermore, GPcounts evaluates its performance across different simulated data sets when it implements different underlying likelihood models to determine under which conditions each yields the best results (BinTayyash et al. 2020). Subsequently, it can be observed that employing an NB likelihood was effective in producing accurately identified SVGs in the package BOOST-GP (Li et al. 2021). However, GPcounts's primary aim is not to identify SVGs, it is also able to identify differentially expressed genes (DEGs), perform pseudotime inference and then identify branching genes and discover temporal

trajectories, widening its scope compared to most packages (BinTayyash et al. 2020). The GP model is stochastic and non-parametric, and there is a choice of kernel to find one that most accurately models the data, similarly to the step employed by SpatialDE (Svensson et al. 2018), and this is determined by the Bayesian Inference Criterion (BinTayyash et al. 2020). Using SpatialDE as a benchmark, GPcounts builds on and alters many of the steps implemented by SpatialDE (BinTayyash et al. 2020). This applies from the testing procedures used to determine SVGs and DEGs p-values to the type of normalisation applied to the data (BinTayyash et al. 2020). GPcounts has also implemented the additional step of a built-in check during its kernel function hyperparameter estimation to minimise the problems of getting stuck in a local optimum by restarting the optimisation as this is suspected (BinTayyash et al. 2020). This is so far one of the only optimisation-based methods that has implemented this kind of self-check and could give GPcounts a distinct advantage in the accurate identification of SVGs. An improved assessment of GPcounts performance when detecting DEGs would be to evaluate the package on published data sets in addition to the simulated data (BinTayyash et al. 2020). When evaluated for its identification of SVGs, GPcounts did use a real mouse olfactory bulb data set and compared its performance to SpatialDE, SPARK and Trendsceek (BinTayyash et al. 2020). GPcounts identifies the most SVGs out of any of the packages, with the vast majority of identified SVGs at a 5% FDR overlapping with those identified by SpatialDE and SPARK (BinTayyash et al. 2020). The unique SVGs identified by GPcounts have spatial patterns that match those depicted in the Allen Brain Atlas, indicating a high confidence in these findings (BinTayyash et al. 2020). GPcounts also identified 90% of the biologically important marker genes expressed in the dataset, although SPARK had a similar performance as it identified 80% (BinTayyash et al. 2020), while SpatialDE identified only 30% of the marker genes (BinTayyash et al. 2020).

Certain frameworks have been developed with a particular SRT technology in mind, in combination with addressing an area of data analysis the developers deem lacking. One of these is the STUtility workflow created in R and based and built on the Seurat analysis tool (Bergenstråhle et al. 2020a). Aiming to develop a package that allows the user to visualise multiple experiments in conjunction to create a 3D view of tissue, STUtility builds on well-established methods of analysis (moulded by those established for scRNA-Seq analysis) to focus on novel data visualization (Bergenstråhle et al. 2020a). Highlighting the importance of data normalisation and transformation to deconvolute technical noise from meaningful biological insight, the package uses a regularized negative binomial regression model successfully implemented in Seurat for normalisation (Bergenstråhle et al. 2020a). The image processing capabilities of STUtility focus on the alignment, automatic or manual, of multiple samples in addition to the removal of background noise (Bergenstråhle et al. 2020a). The removal of background noise – called masking in the study – is an integral part of image processing and allows the inside and outside of the tissue to be defined as well as decreasing the images' storage requirements (Bergenstråhle et al. 2020a). To automatically align multiple samples, the package identifies a reference image, then uses an iterative closest point (ICP) algorithm to align the

remaining samples to the reference, which can then be reconstructed into a 3D tissue model (Bergenstråhle et al. 2020a). While this method of creating a 3D model is not one which yields the most precise cell segmentation, this trade-off yields greater computational efficiency and still gives a faithful reconstruction of tissue morphology (Bergenstråhle et al. 2020a). Implementation of k-means clustering algorithms allows the package to clearly define the boundaries of the tissue (Bergenstråhle et al. 2020a). For the sequencing data, STUtility leans heavily on the functions created by the package Seurat (Bergenstråhle et al. 2020a). A decomposition of the normalised gene data called non-negative matrix factorization (NMF) is used to choose gene drivers and create a low dimensional representation of the data to be used in defining clusters and nearest neighbours (Bergenstråhle et al. 2020a). To obtain genes whose expression demonstrates spatial patterns, a connection network is created for each spot which allows the package to calculate the spatial-lag of each gene across spots. This is one of the inputs – the other being the normalised counts – used to calculate spatial correlation across the sample (Bergenstråhle et al. 2020a). Its ability to visualise spatial distinct features is clearly demonstrated in determining the spatial relation of gene expression to tissue areas (e.g., a tumour). STUtility is also able to identify SVGs using neighbourhood networks, but its accuracy in performing this function is not compared to other packages (Bergenstråhle et al. 2020a). Other capabilities were tested on a variety of human and mouse tissues (Bergenstråhle et al. 2020a). For both mouse brain and human breast cancer tissue samples, spatial gene expression patterns can be clearly identified (Bergenstråhle et al. 2020a). STUtility allows for the manual alignment of multiple images; however, a comparison as to the accuracy of this method compared to the automatic alignment is not offered and depending on the expertise of the user may vary significantly (Bergenstråhle et al. 2020a). Furthermore, while its implementation of neighbourhood networks offers a promising method to define subsections within a tissue and the heterogeneity within, as would be beneficial during the study of tumours, to see how well this correlates to the heterogeneity of the actual tissues of the sample is not reported (Bergenstråhle et al. 2020a; Palla et al. 2021).

7.2.3 Assigning Lost Transcripts

Other packages have been developed with the aim of addressing gaps in analysis that have not been adequately accounted for; one such package is Sparcle (Prabhakaran et al. 2021). When attempting to obtain an accurate gene counts matrix from image-based spatial transcriptomics techniques, often many transcripts are not assigned to cells after segmentation is performed, leading to a loss of data (Prabhakaran et al. 2021). Sparcle aims to recapture the data from these ‘dangling’ transcripts (Prabhakaran et al. 2021). Developed to be used in conjunction with data from any smFISH technology, Sparcle can build a probabilistic model which allows assignment of these dangling transcripts to the appropriate neighbouring cells using a maximum likelihood estimation (MLE). The MLE considers the dangling mRNA’s

distance to other transcripts, nearby cells and genes' covariance when calculating which nearby cell the transcript should most accurately be assigned to (Prabhakaran et al. 2021). Similar to other packages, Sparcle assumes that the most accurate representation of gene expression can be modelled using a multivariate Gaussian distribution (Svensson et al. 2018; Prabhakaran et al. 2021). Sparcle can employ two clustering methods when it first groups the cells in the chosen field of vision (FOV) by cell type based on a global count matrix: DPMM and Phenograph. Phenograph is an algorithm developed to cluster cell phenotypes in high-dimensions single-cell data and was originally applied to data from acute myeloid leukemia (Levine et al. 2015). Dirichlet process mixture model (DPMM) is a stochastic process which can feature all the individual Gaussian distributions for the expression of each gene and allows Sparcle to model all these distributions (Neal 2000). While having the additional flexibility to employ either algorithm at the clustering step, during its validation, Sparcle reports data based on the Phenograph algorithm but not on the performance when using DPMM, nor does it specify in which instance one method should be favoured over another (Prabhakaran et al. 2021). When used to assign dangling transcripts to a MERFISH data set, Sparcle was able to assign 68% or almost 2 million missed transcripts, and validation with scRNA-Seq data confirmed that the proportion of cell types assigned post use of Sparcle more closely matched the scRNA-Seq data (Prabhakaran et al. 2021). Validation against other neuronal data sets returned similarly desirable results. Despite this, there are limitations to the use of Sparcle. For example, when the programme draws an area around each dangling transcript that should mimic the size of a cell, the size of this area is optimised to the size of an average neuronal cell, meaning the package might not be well suited to non-neuronal data (Prabhakaran et al. 2021). Sparcle can run on approximately 80 cells in under 10 min with impressive mRNA recovery over three iterations; however, additional data on how this would scale with larger data sets is lacking, potentially causing computational bottlenecks in bigger data sets (Prabhakaran et al. 2021). It claims to improve on packages that remove the cell segmentation step entirely, such as Baysor and SSAM, by removing the need for *a priori* knowledge of the data set and not assuming that the cellular mRNA can be modelled by a uniform distribution (Prabhakaran et al. 2021). However, some further improvements could be made to enhance the performance, such as staining cellular membranes to better understand the size of neighbouring cells rather than estimating based on an area around the nucleus and calculating an estimate of the prior distribution of a gene's localised transcripts (Prabhakaran et al. 2021).

7.2.4 Estimation of Cell Type Composition

Identifying SVGs was the primary focus of the initial packages developed, but it is important to note that packages with alternative aims are increasingly being published. For example, SpatialDWLS was created to improve the identification of different cell types at locations in the data sets which do not have single-cell resolution

(Dong and Yuan 2021). This is termed cell type deconvolution (Dong and Yuan 2021). Other published packages have been developed for this aim, but SpatialDWLS claims to improve on the results of these packages (Dong and Yuan 2021). How SpatialDWLS performs cell type deconvolution can be summarised in two steps: the first uses a cell type enrichment analysis method to identify which kinds of cells have a high probability of being at each location, and the second uses an extension of the dampened weighted least squares (DWLS) method to pinpoint the precise composition of cell types at the specified location (Dong and Yuan 2021). Firstly, signature genes can either be supplied by the user to be identified by differential expression analysis (Dong and Yuan 2021). Building on the previously developed DWLS method for scRNA-Seq data, this was extended to SRT data by incorporating the signature genes step (Dong and Yuan 2021). Furthermore, SpatialDWLS builds on clustering and gene marker identification used in Giotto (Dong and Yuan 2021; Dries et al. 2019). This would imply that any shortcoming with Giotto's performance in these areas would be transferred to SpatialDWLS. When evaluated on a simulated spatial transcriptomics dataset, SpatialDWLS outperformed RCTD and stereoscope in terms of having a lower Root Mean Square Error (RMSE) and in terms of computational time (Dong and Yuan 2021). However, when its performance was tested against a real mouse brain Visium data set, SpatialDWLS's performance was not benchmarked against the other three packages, thus making its performance on real data unclear (Dong and Yuan 2021). Despite this, the authors reported that the spatial location of the cell types assigned by SpatialDWLS was consistent with those reported in the Allen Mouse Brain Atlas (Dong and Yuan 2021). An interesting application of this package was to identify the change of cell type organisation in a spatial-temporal context throughout embryonic heart development (Dong and Yuan 2021). In addition to quantifying an increase in ventricular cardiomyocytes and smooth muscle cells as time went on, by calculating the assortativity coefficient (here used as a measure of whether neighbouring cells were of the same type) the study was able to determine that spatial organisation of the developing heart becomes increasingly defined in terms of neighbourhoods of cell types during development (Dong and Yuan 2021).

Assigning cell types to a spatial transcriptomics dataset can be approached more than one way. By incorporating *a priori* knowledge to a probabilistic likelihood function, FICT (FISH Iterative Cell Type assignment) can blend expression and spatial information to assign cell type to spatial transcriptomics data sets (Teng et al. 2021). This is achieved by creating a generative mixture model using a reduced dimensions representation of expression levels through a denoising autoencoder and assigning each cell as cell type defined by its neighbourhood (represented in an undirected graph); the parameters of this model can be learnt by an expectation maximization approach, which is an iterative process (Teng et al. 2021). Finally, the cell can be classified by a posterior distribution of the model (Teng et al. 2021). During this process, the problem of over-reliance on expression data needs to be addressed, which occurs because in a dataset it is likely that there are more genes being expressed than cell types present (Teng et al. 2021). To circumvent this problem, a named power factor acts as a weight term to balance the

dimensionally reduced expression component with the spatial component (Teng et al. 2021). The package was validated using three simulated and real data sets and compared to the results of GMM, scanpy, Seurat and smfishHmrf (Teng et al. 2021). Across all three simulated data sets, FICT has the highest median accuracy, reaching a high of approximately 0.89 in one of the simulated data sets (Teng et al. 2021). When evaluated on a real MERFISH mouse hypothalamus data set, the ground truth of the location of different cell types is unavailable, so clustering results obtained from different animals are compared using the Adjusted Rand Index. When comparing across this metric, FICT is more consistent in applying clusters to the majority of the paired animals, indicating its superior performance in assigning cell type clusters (Teng et al. 2021). FICT has the potential to identify novel subclusters within the data set (Teng et al. 2021). However, FICT's performance drops when applied to data sets with smaller numbers of cells, although this is observed across all packages validated (Teng et al. 2021). Furthermore, its decreased performance was still in line with packages with similar functions, and as spatial transcriptomics data sets become larger, this should not interfere with FICT being applied in future (Moncada et al. 2020). However, despite its greater accuracy when applied to larger datasets, FICT's runtime in these instances could still be improved (Moncada et al. 2020).

RCTD is another package created with the final aim of identifying cell types in a spatial transcriptomics data set (Cable et al. 2020). While identifying SVGs is extremely informative, it is important to understand how the role of underlying cell types contributes to a gene's spatially variable expression patterns (Cable et al. 2020). Robust Cell Type Decomposition (RCTD) makes use of annotated scRNA-Seq data to create cell type profiles for expected cell populations in the data, then labels spatial transcriptomics pixels with cell types using a supervised learning method (Cable et al. 2020). As one of the major hurdles in this analysis is the fact that the current spatial transcriptomics data sets can contain multiple cell types within a single pixel, RCTD can also fit a statistical model to determine multiple cell types present within a pixel and normalise across platform effects between the scRNA-Seq and SRT datasets (Cable et al. 2020). To achieve this, RCTD first creates a spatial map of cell types and estimates the number of different cell types in each pixel where the gene counts are assumed to have a Poisson distribution (Cable et al. 2020). This should circumvent the problem introduced by the current unsupervised learning methods that overlook clustering cells that co-localise transcriptionally as well as spatially (Cable et al. 2020). Using this approach, RCTD was able to classify cells across platforms with almost 90% accuracy. However, as with any supervised learning approach, the cell types one can detect using this tool are limited to how accurately and fully the reference data set is annotated, which may present difficulties. Also, while the study tested RCTD using references and data sets generated by many different kinds of scRNA-Seq and SRT technology, the effects that specific platforms may have on cell type assignment is still undetermined.

7.2.5 *Spot-by-Spot Clustering*

A common step in the analysis of many kinds of omics data sets is to perform clustering, and this is prevalent when analysing SRT data. This section will discuss techniques that cluster spots on an SRT array, which may contain multiple cell types, based on the overall gene expression profile of the spot (Bergenstråhle et al. 2020b). Despite being common, this is not a straightforward step. Understanding the results after different iterations can prove difficult, as does choosing the correct hyperparameters (Bergenstråhle et al. 2020b). This is further confounded as each barcode is associated with multiple cells (Bergenstråhle et al. 2020b). To address these issues, an R package called SpatialCPie was developed which focuses on clustering spots on the array based on the gene expression profile to allow annotation of regions of the tissue (Bergenstråhle et al. 2020b). SpatialCPie allows the user to choose which algorithm to implement and clusters the data at different resolutions from the start (Bergenstråhle et al. 2020b). The user is then free to choose which conformations of clusters created at which resolution most accurately represent their data. By creating a cluster graph and an array plot, SpatialCPie gives the user varied insight into how different resolutions affect the clustering outcomes (Bergenstråhle et al. 2020b). The cluster graph displays how the different clusters relate to one another across different resolutions, and conveys the origins of new clusters as they emerge at higher resolutions (Bergenstråhle et al. 2020b). The edges of the graph link the percentage of spots in new clusters that descend from different lower resolution clusters (Bergenstråhle et al. 2020b). The second visualisation method is the array plot, which represents the SRT array, but each spot is depicted as a pie cart that shows how similar the gene expression is between cluster centroids and spatial regions (Bergenstråhle et al. 2020b). SpatialCPie offers the novel, to the best of the authors' knowledge, option to choose a particular region of the dataset for further sub-clustering which may be appropriate depending on the tissue of interest (Bergenstråhle et al. 2020b). While SpatialCPie only compares itself to ST viewer – in a limited capacity – its overall performance is promising (Bergenstråhle et al. 2020b). However, additional validation of its performance compared to other similar packages such as ST viewer would be beneficial to understand its accuracy.

7.2.6 *Pipelines*

As the area of SRT continues to expand, pipelines, rather than just analysis packages, will become more commonplace. One of the first available pipelines written in R is Giotto, which is a platform that can be used on both transcriptomics and proteomics data; it is divided into a data analysis and visualisation module (Dries et al. 2019). With a focus on being user-friendly and reproducible, Giotto does provide the opportunity for more complex spatial analysis using HMRF models (Dries et al. 2019). As a foundation, Giotto creates a neighbourhood network of cells and a

spatial grid for downstream analysis which includes ligand-receptor identification, gene expression pattern analysis and determining preferential cell neighbours (Dries et al. 2019). Giotto is tested on ten different data sets obtained with varying technologies and from varied tissues to examine its performance across a range of benchmarks (Dries et al. 2019). The initial steps in the analysis are similar to those performed in scRNA-Seq analysis, but Giotto does offer three different algorithms for identifying marker genes, one of which (Gini) was specifically developed for the pipeline, which differ in their strength in identifying particular kinds of marker genes (Dries et al. 2019). The Scran method evaluates the markers between two groups of cells by running t-test (default) and then determining marker genes (Lun et al. 2016). Mast identifies marker genes between two cell groups by employing a hurdle model (Finak et al. 2015). The Gini algorithms score marker genes within a cluster based on Gini coefficients, which were developed to identify rare cell types from an adapted model implemented in the social sciences (Jiang et al. 2016). All of these algorithms were developed to score marker genes between clusters in single-cell data sets. When evaluated, Gini discovered the most marker genes for the 12 cell types when compared to Mast and Scran; however, when identifying the top 20 markers using each method, Gini had the lowest sensitivity but highest specificity in both the endothelial and oligodendrocyte populations (Dries et al. 2019). The sensitivity and specificity of each algorithm vary slightly across the different cell populations they investigated when evaluated against a sequential fluorescence in situ hybridization (seqFISH+) somatosensory cortex dataset, and this is important when deciding which algorithm to employ; furthermore, this needs to be tested against data sets generated from different biological material and technologies to best understand the true limitations of each algorithm (Dries et al. 2019). Giotto also has analysis pipelines designed specifically for SRT data sets with lower resolution (Dries et al. 2019). By using one of three algorithms to provide an enrichment score between a location's expression pattern and a cell's gene signature, it is possible to assign a cell type to a location which contains more than one cell (Dries et al. 2019). Once again, the availability of multiple algorithms at this step which require different inputs allows Giotto to be flexibly implemented on a number of different datasets (Dries et al. 2019). These three enrichment algorithms were validated on a simulated dataset similar to one generated using seqFISH+ with the hypergeometric algorithms having the lowest AUC score (0.8) and both PAGE and RANK scoring similarly well when predicting cell type at a particular location (Dries et al. 2019). When applied to real data sets, the two best scoring algorithms RANK and PAGE performed well and should be used when employing the Giotto pipeline (Dries et al. 2019). To analyse spatial patterns of gene expression, Giotto creates a spatial network to represent the data using a Delaunay triangulation network, which is the same as the method employed by scGCO (Zhang et al. 2018; Dries et al. 2019). While the option is available to alternatively construct a spatial network with two different methods offering the user greater control on downstream parameters, the analysis results appear insensitive to these adjustments (Dries et al. 2019). To uncover SVGs, Giotto introduces two new methods, BinSpect-kmeans and BinSpect rank, as well as incorporated methods from SpatialDE, Trendsceek and SPARK

(Dries et al. 2019). When evaluated, each of the methods identified unique SVGs, with 103 genes being identified by all five methods (Dries et al. 2019).

As the field of SRT continues to expand, so will the analytical tools available. As an increasing number of downstream analysis packages are published for SVG identification amongst other analyses, pipelines and frameworks will become increasingly complex in the scope of their abilities. A new framework developed to combine and encompass all aspects of analysis for spatial-omics technology is Squidpy (Palla et al. 2021). While not built specifically for the analysis of SRT data, the Squidpy framework developed in Python brings common tools for analysis and visualisation to any spatial-omics data and takes advantage of the additional information available to improve exploration (Palla et al. 2021). Offering a broader and more modular approach than Giotto, Squidpy offers the opportunity for other packages to be easily integrated into its pre-existing framework to expand its capabilities (Palla et al. 2021). Squidpy will store the image data in an Image Container and create a neighbourhood graph of spatial coordinates so that it can be used on a wide array of technologies (Palla et al. 2021). A feature of Squidpy that adds additional analytical opportunity is its in-built image analysis tools (Palla et al. 2021). While the packages discussed so far require an image as part of the input for analysis, none extend so far as to allow the user to investigate the data contained in this image to the same extent as Squidpy, which is the capability that differentiates it most from Giotto (Palla et al. 2021). The first step in the investigation of cellular neighbourhoods and spatial patterns is the construction of a spatial graph (Palla et al. 2021). When compared to similar processes in Giotto, Squidpy had a more efficient run time when constructing both a spatial graph and calculating neighbourhood enrichment, although for data sets with a smaller number of observations the difference was not great (Palla et al. 2021). Despite offering an interesting perspective on the direction of spatial-omics analysis frameworks and pipeline and reporting limited but promising results with regards to its ability to reproduce results about cellular neighbourhoods, Squidpy does not report its performance in accurately discovering SVGs nor does it quantify how its results relate to those reported in the previous studies (Palla et al. 2021).

7.2.7 Discussion

Despite being a relatively novel technology, SRT – often alongside scRNA-Seq or other techniques – has already been successfully applied to identify gene expression changes in a variety of tissues and disease states. One example was its application in mouse brains to understand spatially DEGs involved in early-stage Alzheimer's disease (Navarro et al. 2020). Different SRT methods are best suited to studying different cell types within a tissue to distinguish differences between them in disease states, such as comparing the dopamine neurons from two regions in Parkinson's patients (Aguila et al. 2018). To further demonstrate how this technology can be applied to an array of conditions and diseases, Modlin and colleagues successfully

actioned it as part of an investigation into the organisation of cellular subtypes that contribute to the antimicrobial capabilities of human leprosy granulomas (Ma et al. 2020).

This clear increase in the popularity of SRT has prompted the recent development of many different packages and pipelines for the downstream data analysis of SRT data sets. While it seems that certain studies are still reliant on packages developed for scRNA-Seq data adapted to included SRT analysis such as Seurat (Ortiz et al. 2019), the variety of purpose-built available tools will likely replace these. A package for easily identifying SVGs seems to be the most popular aim, and even the pipelines developed so far have centred around this same purpose (Svensson et al. 2018; Li et al. 2021; Sun et al. 2019; Edsgärd et al. 2018; Hao et al. 2021; Zhang et al. 2018; Palla et al. 2021; Dries et al. 2019). However, the scope of developing packages continues to expand to further improve the capabilities of analysis, such as Sparcle, which was developed to be used in conjunction with other packages.

Of all the packages discussed, SpatialDE seems to be the most popular, followed by SPARK, Trendsceek and Giotto in terms of being used as benchmarks by which to validate new packages. SpatialDE indicated a tendency to label genes with very low expression as SVGs (Sun et al. 2019), and certain discrepancies in performance compared to other packages tested on real data sets. This alongside the potential introduction of false positives indicates an area of improvement for this popular package. A current limitation of the validation of package performance is that most commonly two data sets (Ståhl et al. 2016), obtained using the same Visium method, are used which will surely introduce inherent bias to the benchmarking process. It would be beneficial to understand the package's performance across datasets from different tissues (instead of exclusively olfactory bulb and breast) generated using a different technology.

To most comprehensively establish the relative performance of all packages, a review should be conducted which benchmarks all packages simultaneously against the same datasets, generated by different SRT methods in different tissues and a standard method for validation established. More packages that are modular and can be integrated alongside one another to expand the scope of analysis are critical and will help advance the field and uptake of this technology. Additionally, the further development of user-friendly pipelines will also make analysing SRT results more accessible. As the array of available tools for analysis of SRT data becomes greater, the results from studies employing the technology will improve and the scope of biological problems that can be addressed will simultaneously expand.

Declarations **Consent for Publication**: All authors provide their consent for publication.

Competing Interest: The authors declare no competing financial interest.
Authors' Contributions: NC drafted the manuscript with intellectual input contributions from HTN and MR. HTN and MR reviewed the manuscript. All authors approved the final manuscript.

References

Aguila J et al (2018) Spatial transcriptomics identifies novel markers of vulnerable and resistant midbrain dopamine neurons. bioRxiv. https://doi.org/10.1101/334417

Asp M, Bergenstråhle J, Lundeberg J (2020) Spatially resolved transcriptomes—next generation tools for tissue exploration. BioEssays 42:1900221

Bergenstråhle J, Larsson L, Lundeberg J (2020a) Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics 21:482

Bergenstråhle J, Bergenstråhle L, Lundeberg J (2020b) SpatialCPie: an R/Bioconductor package for spatial transcriptomics cluster evaluation. BMC Bioinformatics 21:161

BinTayyash N et al (2020) Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments. bioRxiv 2020.07.29.227207. https://doi.org/10.1101/2020.07.29.227207

Cable DM et al (2020) Robust decomposition of cell type mixtures in spatial transcriptomics. bioRxiv 2020.05.07.082750. https://doi.org/10.1101/2020.05.07.082750

Chen G, Ning B, Shi T (2019) Single-cell RNA-seq technologies and related computational data analysis. Front Genet 10:317

Crosetto N, Bienko M, Van Oudenaarden A (2015) Spatially resolved transcriptomics and beyond. Nat Rev Genet 16:57–66

Dong R, Yuan G-C (2021) SpatialDWLS: accurate deconvolution of spatial transcriptomic data. bioRxiv 2021.02.02.429429. https://doi.org/10.1101/2021.02.02.429429

Dong J et al (2018) Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. Genome Biol 19:31

Dries R et al (2019) Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. bioRxiv. https://doi.org/10.1101/701680

Edsgärd D, Johnsson P, Sandberg R (2018) Identification of spatial expression trends in single-cell gene expression data. Nat Methods 15:339–342

Exelby K et al (2021) Precision of tissue patterning is controlled by dynamical properties of gene regulatory networks. Development 148:dev.197566

Finak G et al (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 16:278

Hao M, Hua K, Zhang X (2021) SOMDE: a scalable method for identifying spatially variable genes with self-organizing map. bioRxiv. https://doi.org/10.1101/2020.12.10.419549

Harzing AW (2016) Publish or perish? Harzing.com https://harzing.com/resources/publish-or-perish/os-x. Accessed 26 Apr 2021

He S et al (2020) Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. Genome Biol 21:294

Hu J et al (2020) Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. bioRxiv 2020.11.30.405118. https://doi.org/10.21203/RS.3.RS-119776/V1

Hwang B, Lee JH, Bang D (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 50:96

Jiang L, Chen H, Pinello L, Yuan GC (2016) GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol 17:144

Karaayvaz M et al (2018) Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat Commun 9:1–10

Lein E, Borm LE, Linnarsson S (2017) The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. Science 358:64–69

Levine JH et al (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell 162:184–197

Li Q, Zhang M, Xie Y, Xiao G (2021) Bayesian modeling of spatial molecular profiling data via Gaussian process, Bioinformatics, 37(22):4129–4136, https://doi.org/10.1093/bioinformatics/btab455

Lun ATL, McCarthy DJ, Marioni JC (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. F1000Research 5:1–71

Ma F et al (2020) Single cell and spatial transcriptomics defines the cellular architecture of the antimicrobial response network in human leprosy granulomas. bioRxiv 12.01.406819. https://doi.org/10.1101/2020.12.01.406819

Marx V (2021) Method of the year: spatially resolved transcriptomics. Nat Methods 18:9–14

Moncada R et al (2020) Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat Biotechnol 38:333–342

Navarro JF et al (2020) Spatial transcriptomics reveals genes associated with dysregulated mitochondrial functions and stress signaling in Alzheimer disease. iScience 23:1–19

Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. J Comput Graph Stat 9. http://www.jstor.org/about/terms.html

Ortiz C et al (2019) Molecular atlas of the adult mouse brain. bioRxiv. https://doi.org/10.1101/784181

Palla G et al (2021) Squidpy: a scalable framework for spatial single cell analysis. bioRxiv 2021.02.19.431994. https://doi.org/10.1101/2021.02.19.431994

Prabhakaran S, Nawy T, Pe'er' D (2021) Sparcle: assigning transcripts to cells in multiplexed images. bioRxiv 2021.02.13.431099. https://doi.org/10.1101/2021.02.13.431099

Regev A et al (2017) The human cell atlas. elife 6:e27041

Shah S, Lubeck E, Zhou W, Cai L (2016) In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron 92:342–357

Ståhl PL et al (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 353:78–82

Sun S, Zhu J, Zhou X (2019) Statistical analysis of spatial expression pattern for spatially resolved transcriptomic studies. bioRxiv. https://doi.org/10.1101/810903

Svensson V, Teichmann SA, Stegle O (2018) SpatialDE: identification of spatially variable genes. Nat Methods 15:343–346

Teng H, Yuan Y, Bar-Joseph Z (2021) Cell type assignments for spatial transcriptomics data. bioRxiv 2021.02.25.432887. https://doi.org/10.1101/2021.02.25.432887

Tiklová K et al (2019) Single-cell RNA sequencing reveals midbrain dopamine neuron diversity emerging during mouse brain development. Nat Commun 10:1–12

Waylen LN, Nim HT, Martelotto LG, Ramialison M (2020) From whole-mount to single-cell spatial assessment of gene expression in 3D. Commun Biol 3:1–11

Ximerakis M et al (2019) Single-cell transcriptomic profiling of the aging mouse brain. Nat Neurosci 22:1696–1708

Zhang K, Feng W, Wang P (2018) Identification of spatially variable genes with graph cuts. bioRxiv 491472. https://doi.org/10.1101/491472

Zhou S et al (2021) Single-cell RNA-seq dissects the intratumoral heterogeneity of triple-negative breast cancer based on gene regulatory networks. Mol Ther Nucleic Acids 23:682–690