Cell
Leading Edge

CellPress
OPEN ACCESS

Perspective

Mapping the transcriptome: Realizing the full potential of spatial data analysis

Eleftherios Zormpas,[1] Rachel Queen,[1,2] Alexis Comber,[3] and Simon J. Cockell[1,4,*]
[1]Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK
[2]Bioinformatics Support Unit, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK
[3]School of Geography and Leeds Institute for Data Analytics, University of Leeds, Leeds LS2 9NL, UK
[4]School of Biomedical, Nutritional and Sport Sciences, Faculty of Medical Sciences, Newcastle upon Tyne NE2 4HH, UK
*Correspondence: simon.cockell@newcastle.ac.uk
https://doi.org/10.1016/j.cell.2023.11.003

SUMMARY

RNA sequencing *in situ* allows for whole-transcriptome characterization at high resolution, while retaining spatial information. These data present an analytical challenge for bioinformatics—how to leverage spatial information effectively? Properties of data with a spatial dimension require special handling, which necessitate a different set of statistical and inferential considerations when compared to non-spatial data. The geographical sciences primarily use spatial data and have developed methods to analye them. Here we discuss the challenges associated with spatial analysis and examine how we can take advantage of practice from the geographical sciences to realize the full potential of spatial information in transcriptomic datasets.

INTRODUCTION

To understand the basis of tissue function, it is crucial to map out the spatial landscape and the molecular biology of the cells that comprise the tissue.[1] Since the 1960s, when the technique of *in situ* hybridization (ISH) was established, further methods for detecting RNA distribution and quantity in intact tissue have been developed. Such methods typically target a specific transcript or group of transcripts (a panel), relying on the binding of labelled complementary probes to a sequence of interest to detect its presence. Scaling these approaches from single transcripts to the whole transcriptome has long been a desirable goal.[2] Modern ISH technologies, such as single-molecule fluorescence *in situ* hybridization (smFISH), allow the visualization of sub-cellular transcript localization while retaining overall tissue structure information. The main limitation of these approaches has been in the number of transcripts that can be targeted in a single experiment, generally ranging from just one to a few thousand,[3] but not permitting the targeting of the whole transcriptome. This has made it impossible to measure spatial gene expression in an un-biased, hypothesis-free manner. However, recent developments with *in situ* hybridization techniques have improved the number of transcript targets available. For example, MERFISH[4] is an smFISH-based technology that currently offers panels of up to 500 targets, which could theoretically be expanded to 10,000 by extending the probe length.[3] STARmap[5] uses SNAIL (specific amplification of nucleic acids via intramolecular ligation) probes to target 160–1,020 genes and is capable of retaining three-dimensional positional information at a single-cell level. Another example of increased coverage is GeoMX, which can approach whole-transcriptome scale experiments, with the "Whole Transcriptome Atlas" offering 18,000 targets for human tissues.[6,7]

The phenotype of a cell is, in large part, determined by its transcriptome,[8] and a cell is the foundational unit of a tissue. It therefore follows that technologies to provide an unbiased per-cell transcriptome profile from a tissue are imperative to our understanding of tissue organization and functionality at the cellular level. Single-cell RNA sequencing (scRNA-seq) has proven invaluable as a method for generating relatively unbiased, hypothesis-free data at single-cell resolution, without relying on probe-based targeting of individual transcripts, unlike ISH-based techniques. A wide range of scRNA-seq techniques have been developed which take differing approaches to cell isolation and RNA amplification which in turn affect the sensitivity of a technique and the number of cells profiled.[9] Common to all scRNA-seq techniques, however, is the dissociation of cells from one another at the beginning of the experiment, leading to the loss of the spatial orientation and topological relationships of cells in the experiment to one another.

Until recently, researchers could assay gene expression either spatially (ISH-based techniques) or comprehensively (scRNA-seq), but not both. Following the methods of the first published spatial transcriptomics experiment,[10] an intact tissue section is placed on a specially prepared slide that has 100 μm spots of poly-dT probes—capable of binding to the polyadenylated tail of many RNA molecules. Each spot carries a unique spatial barcode. After the tissue is imaged, it is permeabilized and RNA

Box 1. Glossary

Aggregation problem: a variation in the results will be observed when we use alternative combinations of areal units, even on the same scale.

Bandwidth: a parameter describing how localized a geographically weighted regression model is.

Curse of dimensionality: the higher the number of dimensions used, the more equidistant the observations are.

Geographically weighted regression (GWR): a method that takes spatial heterogeneity into account because it allows for the relationships between the independent and the dependent variables to alter locally.

Global spatial autocorrelation measures: statistics assessing the degree of spatial autocorrelation for a variable in the whole dataset.

Hot and cold spots: aggregations in space of high (hot) or low (cold) values for a variable.

Interesting locations: a term used instead of "statistically significant" when considering pseudo-p-values in multivariate space.

Local indicators of multivariate spatial association (LIMSA): measure of the extent to which neighbors in multivariate space (i.e., spots with gene expression profiles similar to each other) are also neighbors in geographical space.

Local indicators of spatial association (LISA): see *local spatial autocorrelation measures*

Local spatial autocorrelation measures: statistics assessing the degree of spatial autocorrelation for a variable in every location compared to its neighbors.

Modifiable aerial unit problem (MAUP): the choice of spatial data aggregation scale will influence the statistical relationships and process understanding from analyses.

Multiscale GWR (MGWR): geographically weighted regression that operates in multiple scales by identifying the best bandwidth for each independent variable.

Pseudo-p-value: a p value calculated from a normal distribution generated by computational permutations that must be treated with caution.

Scale problem: different results will be obtained when we aggregate the same set of data on different scales.

Spatial autocorrelation (SA): nearby observations tend to be similar, violating one of the major assumptions of classical statistics—that of observation independence.

Negative SA: nearby values tend to be dissimilar.

Positive SA: nearby values tend to be similar.

Spatial heterogeneity: the factors associated with an outcome, and therefore the process, will vary in space. Also referred to as spatial non-stationarity.

from the permeabilized tissue hybridizes to the spot that is closest to it. Thus, it hybridizes to a spot that is more or less at its physical location in the tissue. Subsequent RNA sequencing can then be resolved to specific locations on the tissue image. Spatial transcriptomics thus combines tissue imaging and spatial resolution—effectively approaching the spatial resolution of ISH-based technologies—with the comprehensive transcriptome quantification abilities of RNA sequencing.[11]

Since the first spatial transcriptomics experiments in 2016, there has been rapid development of sequencing-based methods. For example, 10× Genomics Visium uses a smaller spot size of 55 μm (compared to the initial 100 μm), while Slide-Seq/V2[12,13] uses probes attached to 10 μm beads that randomly assemble into a monolayer on the tissue slide. Stereo-seq,[14] which uses a DNA nanoball patterned array and *in situ* RNA capture, offers improved resolution of 500 nm, meaning that sequencing-based spatial techniques can offer sub-cellular resolution.

Modern spatial transcriptomics methods generate three distinct but interrelated data types: (1) the image data, (2) the expression data, and (3) the spatial orientation and location of (2). A typical spatial transcriptomics analysis workflow (e.g., Orchestrating Spatially-Resolved Transcriptomics Analysis with Bioconductor) tends to treat individual observations as though they originate from single cell or mini-bulk RNA-seq experiments and deals with them accordingly, while ignoring the relative locations of the different observations. As we will discuss, expression data and their spatial properties are entangled to an extent that one cannot properly analye the former without taking account of the latter, and so treating individual observations independently risks missing vital information.

As spatial transcriptomics technologies become commercially available, the number of publications using the approach is steadily increasing. Current studies that utilize spatial transcriptomics can be broadly categorized into three groups. First are those that use the data for clustering, e.g., gene expression-based clustering of sarcoidosis granuloma skin samples to compare to single-cell clusters from the same tissue.[15] Second are those that consider specific cell/gene positioning, e.g., using marker gene expression to identify the locations of astrocytes in mouse brain samples.[16] Finally there are those that consider differential gene expression, e.g., finding changes in expression between spatial clusters found in mouse models of heart failure.[17] In many of these experiments, the spatial information is used to pre-determine particular locales of interest. The full potential of the spatial content, however, lies in using the coordinate information as an intrinsic part of the data to be analyed, and these aforementioned approaches do not attempt this. As we examine, the definition of these zones of interest can in itself give rise to statistical problems related to the aggregation of space.

By properly considering the issues of spatial data analysis, studied extensively in disciplines such as geography and ecology, and by treating space as a covariate, a new type of spatial transcriptomics data analysis becomes available. This analysis makes use of this additional dimension to further the understanding of how gene expression is organized in space. In the following sections, we will elucidate the ways in which spatial data are different from other molecular data, and look at the ways in which these differences can be managed and even exploited. A summary of the key terms we introduce here, particularly those from the geographic sciences, is given in the glossary (Box 1).

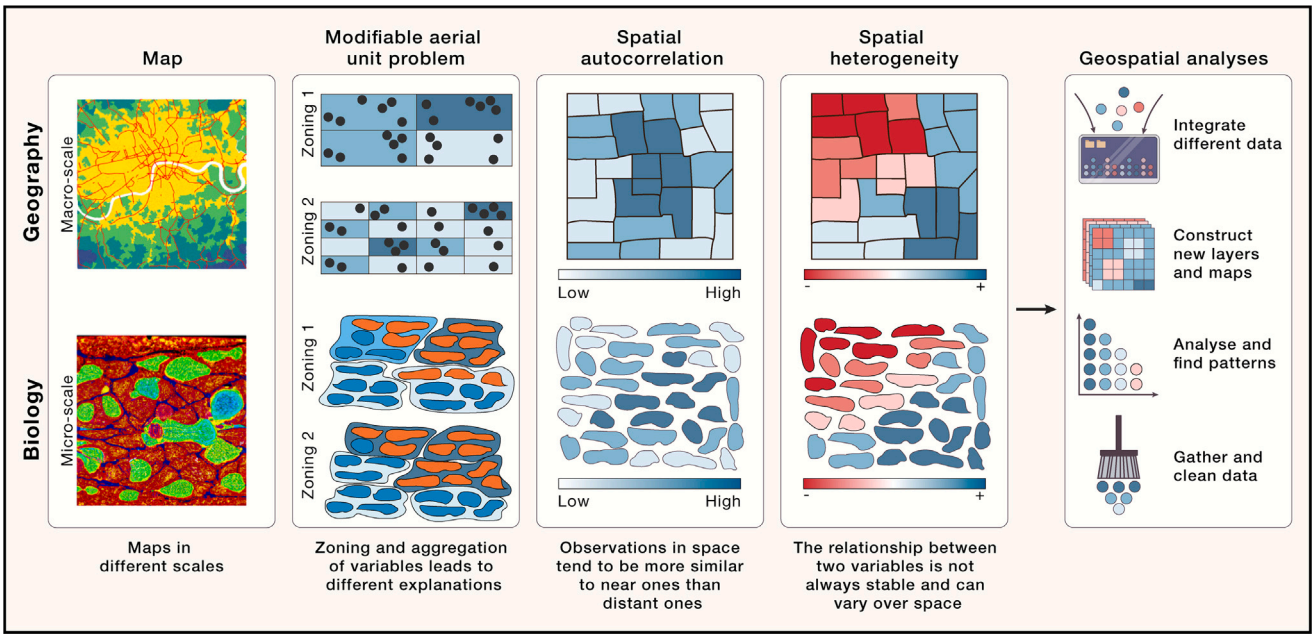


Figure 1. Conceptual illustration depicting the analogous problems encountered in the fields of geographical sciences and spatial transcriptomics

An illustrative example conceptualizing the similarities of the three main problems that underpin both geographical and biological spatial data. From left to right: geospatial and bio-spatial data can both be represented as maps on different scales. Geospatial data describe phenomena on a macro-scale, while bio-spatial data describe phenomena on a micro-scale. The modifiable aerial unit problem suggests caution when selecting analysis scales, since zoning and aggregation of variables can lead to differing results. Spatial autocorrelation advises that observations in space tend to be more similar to near ones than to distant ones. Finally, spatial heterogeneity demonstrates that the relationship between variables is not always stable and can vary over space.

KEY FEATURES OF SPATIAL DATA

Through decades of studying data that describe spatial problems, the geographic sciences have identified three main features governing spatially resolved datasets (Figure 1). First, the modifiable areal unit problem (MAUP, Box 1) arises from the choice of spatial data aggregation scale that influences the statistical relationships and process understanding from analyses (Figure 2). Second, spatial autocorrelation indicates that nearby observations tend to be similar, violating one of the major assumptions of classical statistics—that of observation independence (Figure 3). Finally, spatial heterogeneity or spatial non-stationarity states that the factors associated with an outcome, and therefore the process, will vary in space (Figure 4). Although the illustrative figures in this section use Visium data to demonstrate these phenomena, the concepts apply to all types of spatial transcriptomics data.

Modifiable areal unit problem

The MAUP (Box 1) is a concept in geography which states that the way we define and group geographic units can significantly impact our analysis and conclusions.[18] A spatial transcriptomics experiment results in a map of a biological tissue, along with associated information about the locations on that map (the expression of a range of genes). Typically, this information is analyed as *regions*, which could be anything from a single cell to a large multicellular section of tissue. The MAUP tells us that there is a *scale effect*[19] on the results of the analysis we undertake with these observations (Box 1). That means, if we change the size of the regions on our map, we can get different results. Common procedures for grouping regions in spatial transcriptomics data, for example clustering or expert annotation, produce zones on our map. The MAUP also describes a *zoning effect*,[20] where the boundaries that are set for each different region can influence the analysis and the results even if the regions themselves remain the same size. For instance, even slight variations in the input for clustering can produce different cluster layouts and these changing boundaries will affect downstream analysis. For example, changing the criteria by which highly variable genes are selected will necessarily influence the placement of cluster boundaries, which will in turn affect differential gene expression or marker gene selection. Figure 2 is a simple illustrative example of how the MAUP can influence the analysis of spatial transcriptomics data. This case shows how the aggregating of measurements at different scales or zones can affect the correlation of the expression of a pair of genes.

All analyses of spatial data are affected by the MAUP and the spatial scale of analysis. At its core the MAUP proposes that statistical distributions, relationships, and trends exhibit different properties when the same data are aggregated or combined over different reporting units, at different spatial scales. It describes the process of distortion in calculations and differences in outcomes caused by changes in statistical distributions, where variance reduces with aggregation. Thus, variations in statistical relationships are a result of the models generated from these data.

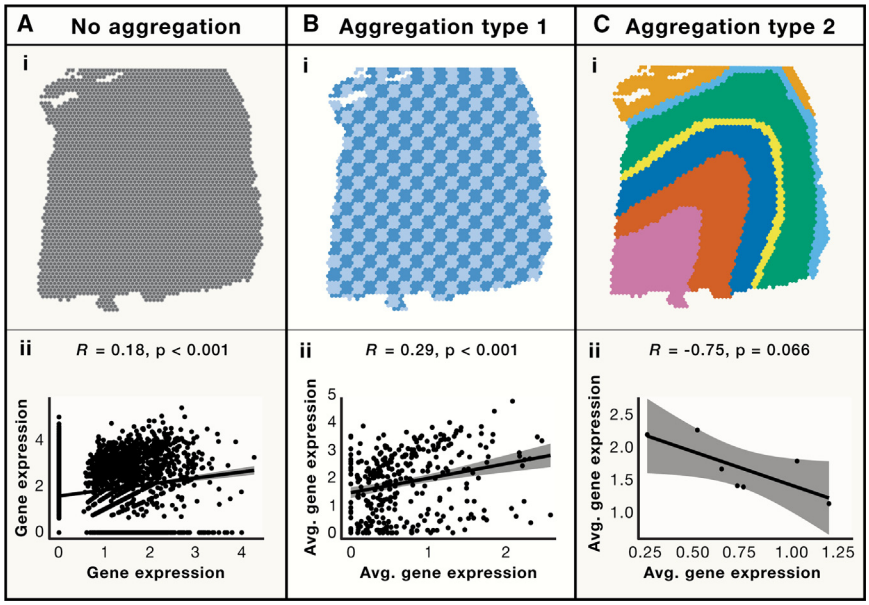


Figure 2. Demonstration of the effect of aggregation on correlation, as described by the modifiable areal unit problem (MAUP)

Different scales of aggregation change the correlation between two variables—each panel summarizes the same data points (3,610 observations of two genes).

(A) (i) There is no aggregation (gray tissue) and (ii) the expression of the two genes show positive correlation (R = 0.18).

(B) (i) The data points are aggregated to a square grid (light and dark blue) and the mean of each variable was calculated for each of the 325 grid cells and then the correlation was determined. (ii) This aggregation results in a positive correlation of 0.29.

(C) (i) A zonation based on the biology of the tissue was used to aggregate the data into 7 layers, and variable means were calculated. (ii) This aggregation scale resulted in a negative correlation of −0.75. In this example, different aggregation scales result in a changed relationship between the variables (from positive to negative correlation), but this does not always have to be the case as the effects of the MAUP are non-nested and non-hierarchical.

The MAUP also interacts with spatial sampling and spatial scale. The spatial scale includes the spatial support and the spatial extent of analysis.[21] Spatial support describes the area occupied by each observation—the space on which an observation is made. The MAUP arises from (1) the spatial support of spatial data and (2) the aggregation or interpolation of data to particular scales. The spatial support of observations affects statistically determined relationships between them, such as statistical models (e.g., regression or classification). In this way statistical analyses of spatial data over different spatial supports will result in different models, correlations, classifications, coefficient estimates, etc.[22–26]

Despite the MAUP being a core consideration, its effects are rarely tested even in geographic research. It is also frequently overlooked by the many disciplines now routinely using spatial data. However, there are suggestions and demonstrated approaches for determining the impact of the MAUP and for appropriate scales of analyses.[26] In brief, testing for the MAUP and an appropriate sampling or aggregation scale involves the identification of the spatial scales at which the investigated process is stationary (stable) with respect to their variance, covariance, and other moments. It is essential to be aware of these issues when working with spatial data and to consider different ways of aggregating or dividing data to better understand the potential impact of the MAUP.

The MAUP and aggregation

The aggregation of observations into clusters (groups) based on attributes is commonly undertaken in order to reduce the complexity of the raw data and to provide convenient but coherent objects for subsequent analyses. In geography, groups may be driven by socioeconomic or environmental attributes, while in biology they may be related to cell types or gene expression profiles. The nature of the information loss associated with aggregation relates to the number of groups and their granularity. Furthermore, as individual observations are allocated to groups based on their attributes, aggregated group properties are summaries of sets of similar observations.

Aggregations over space are different. Here observations are grouped by their location and not by their attributes. They are spatially intersected with the aggregation areas and attributes are re-aggregated, including updating counts and the re-calculation of rates (proportions, percentages), over each aggregation area. Aggregation areas can be defined in a variety of ways, for example as a spatial grid (Figure 2B), or for spatial transcriptomics a biologically defined zone (analogous to, for example, an administrative area in the geographical sciences) (Figure 2C). These aggregation areas can be grouped together, based on the observations contained in each area. If the area changes for whatever reason, the observations contained will also change, and thus the grouping is likely to change too.

The scale (area, size, extent) of the aggregation units drives the specific manifestation of the MAUP. These manifestations are critically unpredictable because the effects of the MAUP can be non-nested, e.g., the boundaries of one aggregation unit will not always perfectly align or enclose the boundaries of another unit (compare Figures 2B and 2C). Furthermore, these effects can also be non-hierarchical, e.g., smaller units are not consistently aggregated into larger ones in a predictable manner.[26] The example shown in Figure 2 demonstrates that aspects of the MAUP apply to spatial transcriptomics data and show that it is worth considering the potential biological implications of this important concept.

Biological relevance of the MAUP

In geography, observations are usually aggregated over predefined areal units (such as census reporting areas) or defined by the spatial resolution of remote sensing instruments. The underlying processes that are captured by the observation are usually defined by the resolution of the aggregation, most evidently in remote sensing but also in socio-economic constructs; a classic example is the "neighborhood." In biology, processes

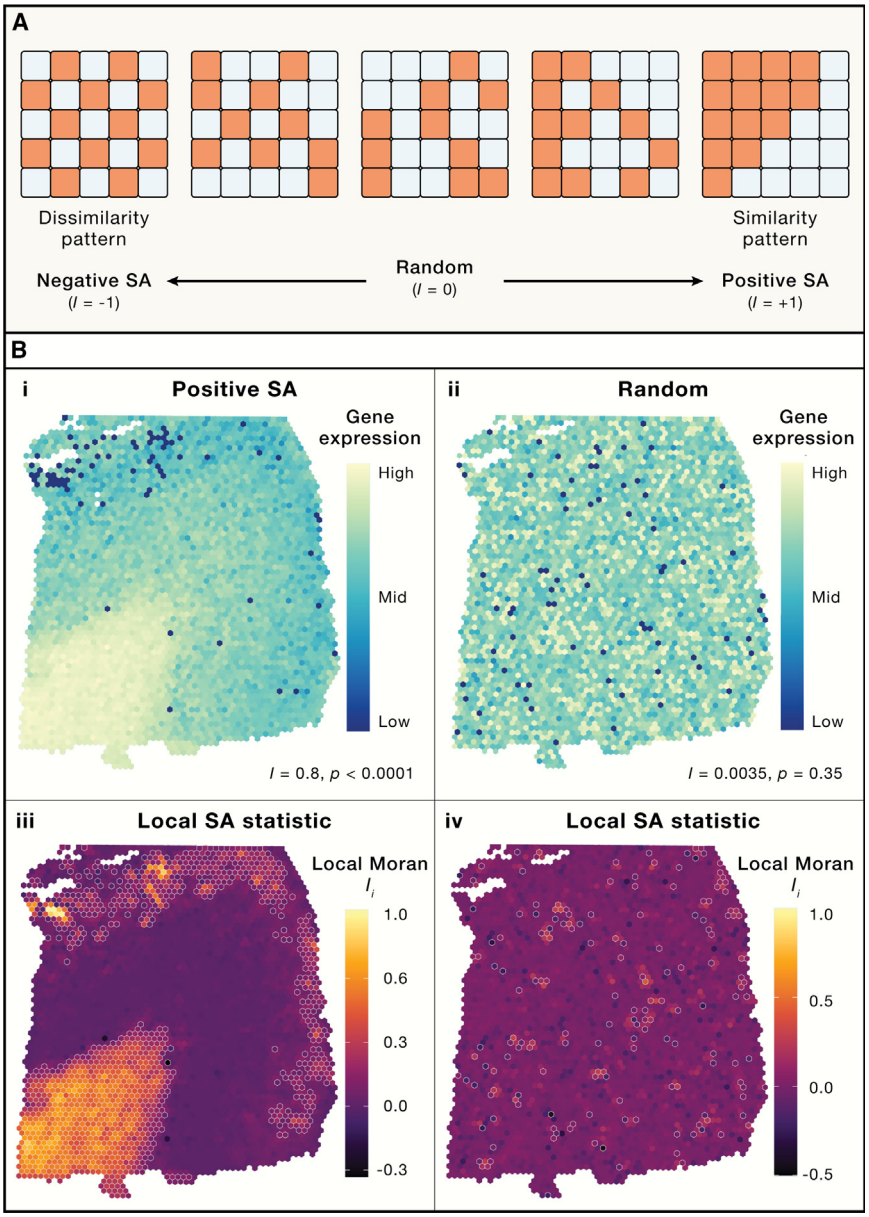


Figure 3. Spatial autocorrelation highlights regions of statistical similarity in gene-wise expression

(A) A simulated example of spatial autocorrelation (SA) for binary data associated with Moran's *I* statistics. The leftmost panel shows perfect dispersion or dissimilarity, corresponding to a Moran's *I* of −1. The center panel shows a random spatial distribution, corresponding to Moran's *I* of 0. The rightmost panel shows perfect spatial autocorrelation, clustering, or similarity, corresponding to a Moran's *I* of +1.

(B) Examples of SA from tissue measurements of gene expression. (i) Positive SA ($I = 0.8$): notice how high values cluster in one corner of the tissue while low values are highly localized. (ii) The same observations but randomly distributed over space ($I \approx 0$). (iii) The local Moran's *Ii* (LISA) for the variable in (i). Here, the local statistic provides a more detailed representation of SA over space, revealing locations with high SA. Significant locations ($p < 0.05$) are highlighted with a white border. (iv) The local significance of the local Moran's *Ii* (LISA) in (ii). Significant locations ($p < 0.05$) are highlighted with a white border.

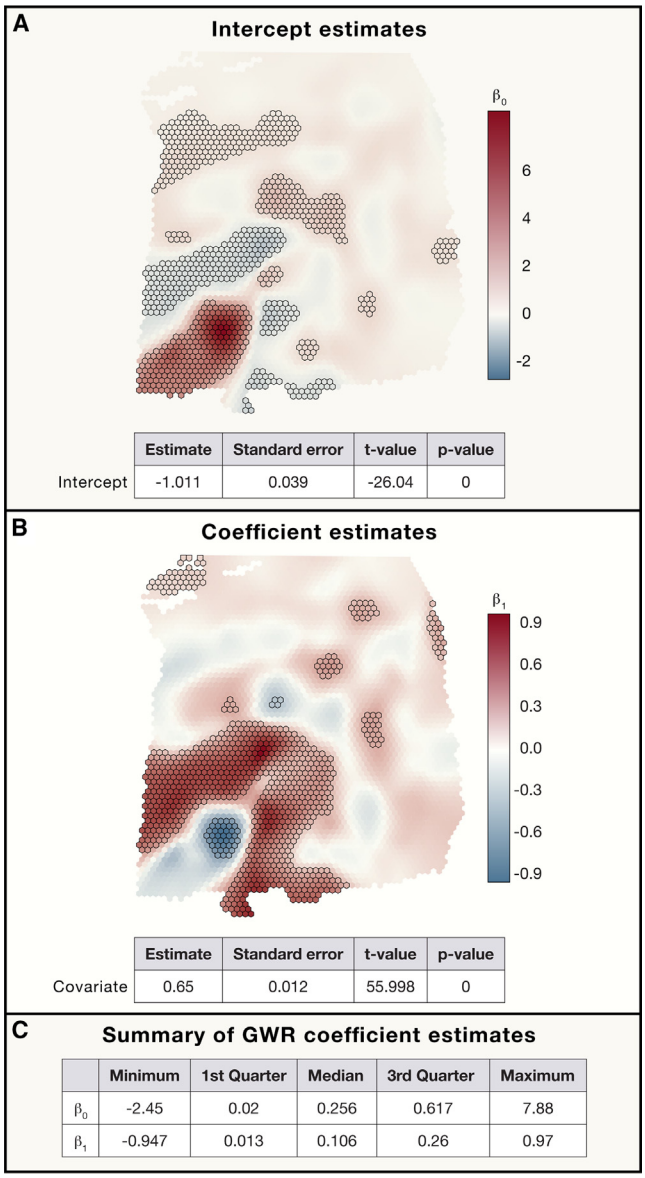
are more tightly defined and structured. A tissue is constructed of cells which for the purposes of partitioning and aggregation can be considered indivisible. Therefore, when working with spatially oriented biological data, it could be argued that the scales chosen for data analysis are somewhat pre-defined by the tissue structure itself, while recognizing that the understanding of the process and how they manifest themselves will be driven by this choice of analysis scale.

This implies that observed and aggregated biological processes occur at different scales within a tissue and that the choice of scale influences the resolution at which such processes can be examined.[27,28] As an example, the activation state of an immune cell can be elucidated by studying its immediate neighborhood.[29] This implies that choosing the appropriate scale for downstream analysis is vital in understanding this activation state. Ideally, the chosen scale covers the immediate cellular neighborhood—as a finer or coarser scale would result in a different understanding of the state of the immune cell. High-resolution techniques such as Stereo-Seq[14] naturally invoke the MAUP, as the subcellular-level observations need to be aggregated at least at cell-level to facilitate analysis. Whether this aggregation is achieved via a grid-based system or an image-based approach such as cell segmentation will have considerable bearing on the results of the analysis.

Spatial autocorrelation

The concept of spatial autocorrelation (Box 1) has existed for many years. It is captured in what has become known as Tobler's "first law of geography" which states that *everything is related to everything else, but near things are more related than distant things*.[30] This reflects what we intuitively know about the world we live in—phenomena are clustered and not randomly distributed—and implies some form of distance decay in observation values and that measurements of a phenomenon will be correlated over space. This lack of observation independence has to be accounted for with statistical models.

Several established tests for spatial autocorrelation exist for measurements in a spatially resolved dataset which all have a similar underlying operation. Essentially, they generate measures of similarity by comparing each observation value with those in its neighborhood. The neighborhood may be defined by a distance or some form of adjacency (1st order to nth order),



	Estimate	Standard error	t-value	p-value
Intercept	-1.011	0.039	-26.04	0

	Estimate	Standard error	t-value	p-value
Covariate	0.65	0.012	55.998	0

	Minimum	1st Quarter	Median	3rd Quarter	Maximum
β_0	-2.45	0.02	0.256	0.617	7.88
β_1	-0.947	0.013	0.106	0.26	0.97

Figure 4. Spatial heterogeneity revealed by varying local regression coefficients across a spatial transcriptomics experiment

Geographically weighted regression (GWR) produces local (i.e., spatially varying) coefficient estimates, in contrast to the fixed global ones from a standard regression.

(A) The spatial variability of the estimate of the intercept in a GWR between the expression of two genes. The intercept (the estimated average expression of the response variable when the predictor variable is zero) varies across the tissue and takes both positive and negative values. The table at the bottom of the panel shows the intercept of the global linear model for the same comparison.

(B) The spatial variability in the coefficient of the same model as (A). Like the intercept, the coefficient (the average change in the response variable for a 1 unit increase in the predictor variable) varies considerably across the tissue and is rarely close to the global estimate (shown in the table at the foot of the panel). Notice how the fixed coefficient values from the global model (although statistically significant) mask the actual differences that were found over space.

(C) A summary of the range of values produced by the GWR model. The highlighted regions in (A) and (B) indicate the locations where significant local relationships were found.

and the spatial autocorrelation test may report local measures of spatial autocorrelation or a global (whole map) one (Box 1) which is essentially summarized from local measures (Box 1). For example, in spatial transcriptomics, a global measure will indicate whether a gene exhibits spatial autocorrelation or not (Figure 3Bi), while a local measure will reveal the exact pattern of said autocorrelation in space (Figure 3Biii). The advantages of the local measures are that they indicate the locales of spatial autocorrelation within the data, critically the local significance of the spatial autocorrelation test. In order to understand the local measures of spatial autocorrelation, we first need to consider their global counterparts.

Global measures of spatial autocorrelation

Moran's correlation coefficient,[31] known as Moran's *I*, is a univariate measure of spatial autocorrelation (Figure 3A). It is the most well-recognized and widely used spatial autocorrelation statistic. *I* can take values between −1 and +1, where values near +1 indicate strong positive spatial autocorrelation (Figure 3Bi) (Box 1), values near −1 indicate strong negative spatial autocorrelation (i.e., perfect dispersion) (Box 1), and values near zero suggest a randomly dispersed variable (Figure 3Bii).

Alternative measures to Moran's *I* include Geary's *C*[32] and Getis and Ord's *G* statistic.[33] Geary's *C* takes only positive values, where $C = 1$ means that there is a lack of spatial autocorrelation. Values of *C* approaching 0 show increasing positive spatial autocorrelation, and $C \gg 1$ is indicative of negative spatial autocorrelation which suggests large differences between a location and its neighbors. Thus, Geary's *C* is inversely related to Moran's *I*, though as the measures are calculated differently (Table 1), they are not directly related. Due to these differences in calculation, Moran's *I* is more a global measure and more sensitive to extreme values while Geary's *C* is sensitive to differences between values in neighboring areas. The *G* statistic is ideal for checking whether similar values co-locate as it only provides a measure of positive spatial autocorrelation. It is more useful as a local model of spatial autocorrelation, as discussed in the next section.

Local measures of spatial autocorrelation

The global tests for spatial autocorrelation generate a single statistic that evaluates spatial autocorrelation in the whole dataset and indicates the overall degree of spatial autocorrelation. However, some parts of the study may exhibit greater spatial autocorrelation than is found in others. Identifying local patterns of spatial autocorrelation can often be more informative. Each of the global indicators of spatial autocorrelation mentioned above has a local equivalent. Local indicators of spatial autocorrelation (LISA)[34] (Box 1) provide three kinds of information. First, the local spatial autocorrelation measure for each observation gives an indication of the degree of spatial clustering of values around that observation. Second, the local indicators of spatial autocorrelation provide a measure of local significance and third, provide a measure of the relative contribution made by each observation to the global spatial autocorrelation statistic (Figure 3B, iii and iv).

As with the global statistics, a local Moran's *Ii* is the most widely used measure. Other local statistics include local Geary's *Ci* which is based on the squared value difference between locations, and so fails to discriminate differences

Table 1. The mathematical formulas used for calculating spatial autocorrelation

SA Statistic	Formula
Global Moran's *I*	$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad i \neq j$ (Equation 1)
Global Geary's *C*	$C = \frac{(n-1)\sum_i\sum_j w_{ij}(x_i - x_j)^2}{2S_0\sum_i(x_i - \bar{x})^2}$ (Equation 2)
Global Getis and Ord's *G*	$G(d) = \frac{\sum_i\sum_j W_{ij}(d)x_i x_j}{\sum_i\sum_j x_i x_j} \quad i \neq j$ (Equation 3)
Local Moran's I_i	$I_i = z_i\sum_j w_{ij}z_j$ (Equation 4)
Local Geary's C_i	$C_i = \sum_j w_{ij}(z_i - z_j)^2$ (Equation 5)
Local Getis and Ord's G_i	$G(d) = \frac{\sum_j w_{ij}(d)x_j - W_i\mu}{\sigma\{[(n-1)S_{1i} - W_i^2]/(n-1)\}^2} \quad j \neq i$ (Equation 6)

The notation used is common throughout the table: $\boldsymbol{n}$ is the total number of observations for a variable, $\boldsymbol{(i, j)}$ are the locations of the observations, $\boldsymbol{W_{ij}}$ is the spatial weights matrix, $\boldsymbol{x}$ is the value of the variable on each location, $\bar{\boldsymbol{x}}$ is the mean of the variable, $\boldsymbol{S_0}$ is the sum of all weights $\left(\sum_i\sum_j \boldsymbol{w_{ij}}\right)$, $\boldsymbol{W_{ij}(d)}$ is the spatial weights matrix of one/zero with ones for all links that are within distance $\boldsymbol{d}$ from each other, $\boldsymbol{z_i}$ and $\boldsymbol{z_j}$ are the deviations from the mean, $\boldsymbol{w_{ij}}(\boldsymbol{d})$ is the weights matrix with a threshold distance $\boldsymbol{d}$ that defines how far away the neighborhood reaches, $\boldsymbol{W_i} = \sum_j \boldsymbol{w_{ij}}$, $\boldsymbol{S_{1i}} = \sum_j \boldsymbol{w_{ij}^2}$, and $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the classical sample mean and standard deviation, respectively, for sample size $\boldsymbol{(n-1)}$.

between high and low values,[35] but is useful for identifying nonlinear associations. Local Getis and Ord's *Gi* distinguishes between high- and low-value clustered regions since the *Z* score produced by the method informs whether a cluster of similar values is relatively high (hot spot, positive *Z* score) or low (cold spot, negative *Z* score)[36] (Box 1). The individual components of local *Gi* are not proportional to the global statistic, but it is a commonly used measure where it is important to distinguish between areas of clustered high and low values. To visualize this, let's imagine a gene exhibiting a pattern of high and low expression in different parts of the tissue which results in positive spatial autocorrelation. A local Moran's *Ii* and a local Geary's *Ci,* Getis and Ord's *Gi* can indicate not only where the positive spatial autocorrelation is located but also whether it is a result of high or low expression.

Spatial autocorrelation and statistical inference

Hypothesis testing for spatial autocorrelation statistics can be achieved using either a *Z* score calculation or a Monte-Carlo method of permutation simulation. Both of these cases involve the estimation of a pseudo p value (Box 1) and should be treated with caution, since this estimate is unlikely to be a true reflection of the potential for type I error.[34,35] This analysis therefore should be considered strictly exploratory and subjected to rigorous false discovery correction. This has direct relevance in cases where global or local measures of spatial autocorrelation are employed to guide subsequent analyses. For instance, spatially variable genes can be chosen independently or in conjunction with highly variable genes to perform dimensionality reduction and subsequent analysis. Failure to take into account the aforementioned statistical considerations can impact the quality of the selected gene group, consequently influencing the downstream results.

Multivariate spatial autocorrelation

The statistics described above are univariate, and thus would need to be considered for every measurement in a spatial dataset. Local indicators of multivariate spatial association (LIMSA) (Box 1) would also have utility in biological systems, where the observed phenotype is often the product of multiple co-regulated variables. In LIMSA, the core concept is to measure the extent to which neighbors in a multivariate feature space are also neighbors in geographical space.[35] LIMSA statistics are adversely affected by dimensionality, because the number of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially with the number of input variables (i.e., dimensionality) of the function (see Box 1, "curse of dimensionality").[37] If the number of features (i.e., genes) is bigger than the number of observations, there is the risk of overfitting the model. On the other hand, if there are too many dimensions, each observation is equidistant from the rest, resulting in no meaningful clusters.

The dimensionality problem, coupled with the pseudo p value estimation approach inherited from univariate measures of spatial autocorrelation, means that care should be taken with the interpretation of "statistically significant" locations in LIMSA. Pseudo p values can only provide an indication of interesting locations (Box 1). It is also important to note that significantly clustered locations in the univariate case do not necessarily translate to significantly clustered ones in multivariate space.[35]

Biological relevance of spatial autocorrelation

The behavior of cells *in situ* is influenced by their surrounding environment, which includes intra-cellular signals and direct contact with neighboring cells.[28] Consequently this implies that cell variables (e.g., expressed genes), when considered in their spatial context, are dependent on their surroundings as opposed to being independent variables. This suggests that well-established statistical methods of RNA-seq analysis are not perfectly suited to infer answers from spatially oriented spatial transcriptomics experiments as they typically assume value independence.[38–40]

The canonical example of precise spatial patterns that arise as a direct effect of spatial signalling is the embryonic development process.[41] Finely spatially organized cell blocks such as the blastoderm or the gastrula exhibit positive spatial autocorrelation due to the unique transcriptomic profile that each block presents.[42,43]

Similar examples of spatially structured cell neighborhoods that could exhibit spatial autocorrelation can be found in the liver's zonation system[44] where marker genes for different liver zones (e.g., central and portal vein proximity) are shown to exhibit significant spatial autocorrelation. The concept of zonation also applies to the bone marrow,[45] which contains various niches that are distinct in their cell-type composition. In the lung,[46] spatial profiling allowed the segregation of the lobar airways into four distinct zones, with differing cell compositions and consequently gene expression. A spatially ordered neighborhood can also be found in the tumor microenvironment,[47,48] which can exhibit coherent communities of different microenvironment subtypes that are spatially organized. As a result, a proper investigation of the tissue cannot be undertaken without taking spatial autocorrelation into consideration.

Spatial heterogeneity

Spatial heterogeneity (Box 1) describes the variation of a process across different locations within a defined area. It is a concept often used in geography, environmental science, and other fields to describe how things change from place to place.[49,50] Spatial heterogeneity is important when trying to understand the underlying spatial pattern of a process.[51] Conceptually, a single variable like the expression of a single gene can change in different parts of a tissue, in which case, this individual gene is said to be displaying spatial heterogeneity. Intriguingly, spatial heterogeneity can also apply to relationships between variables, which may not be constant across space. For example, as demonstrated in Figure 4, the global relationship between a pair of genes (in this case modelled by simple linear regression) may not accurately summarize the local-level variation seen in this relationship. In this example, few locations in the tissue have a local relationship between the given pair of genes that resembles the global regression estimates—and the local relationships are varying between being strongly positive or strongly negative depending on the location within the tissue.

Taking spatial heterogeneity into account in the analysis allows for the relationships between the independent and dependent variables to vary locally.[51] Geographically weighted regression (GWR) (Box 1) is one such method widely used in geographical sciences. Essentially, GWR undertakes a series of local regressions using a moving window or kernel. In contrast to standard linear regression, GWR includes a local spatial weights matrix to represent the spatial relationships of the data used in each local model. The output of GWR provides an estimate of spatial heterogeneity by generating a regression estimate per location, giving the ability to map the regression coefficients over space to visualize the relationships (Figure 4). The coefficients produced by GWR are the same as those derived from simple linear regression. β_0 is the intercept, the expected expression of gene Y when gene X expression is zero; β_1 is the coefficient estimate, describing how much the expression of gene Y changes for a single unit change in the expression of gene X (i.e., the slope of the regression line).

A key consideration in GWR is in the determination of the *bandwidth* (Box 1) for the analysis. The bandwidth defines the kernel size and thus the distance decay in the spatial weights matrix for data falling under the kernel. It is akin to a smoothing function whereby a large value will reduce the observation of local effects. Standard GWR identifies a single optimal bandwidth for the entire space, thereby assuming that all relationships between the dependent and independent variables operate at the same spatial scale.[52] However, in biology, this is rarely the case,[53] and as such a standard GWR identifies a *best on average* scale for examining spatial heterogeneity.[53] Alternatively, multi-scale GWR (Box 1) defines a bandwidth for each covariate, thereby allowing the scale of analysis to vary for each dependent to independent variable relationship.[54,55] Thus, in most cases multi-scale GWR is likely the most appropriate approach since it avoids the compromise of selecting a single scale for the whole map.[53]

Biological relevance of spatial heterogeneity

In the biosciences, spatial heterogeneity is most commonly used to describe an area of tissue that was expected to be homogeneous (for example with respect to the cell types in a given tissue) but is found to be not homogeneous. Taking into consideration the variation in space is common when studying phenomena like variation exhibited by natural ecosystems. This is analogous to the tumor microenvironment, which is often accompanied by strong regional variation.[56,57] For example, tumor-associated macrophages (TAMs) exhibit multi-dimensional heterogeneity in different cancer types like renal carcinoma, lung adenocarcinoma, and gastric cancer.[58–60] In gastric cancer, TAMs were also found to be polarized according to their location within the tumor microenvironment and display heterogeneity in their phenotype depending on this location (core, edge, and margin).[60] This example highlights the importance of considering spatial heterogeneity in the analysis of spatially resolved data, since without doing so these subtleties in the organization of systems like the tumor microenvironment would be lost.

SPATIALLY AWARE SPATIAL TRANSCRIPTOMICS ANALYSIS

The common features of geospatial data, discussed in the previous section, appear to be routinely exhibited by spatial transcriptomics data. This implies that those features should be considered when developing approaches to analyze these data. Several bioinformatics analysis packages exist that incorporate

space, for example by calculating a distance-related covariance matrix,[61] by building a neural network that uses locational information to construct itself,[62] or by accounting for spatial correlation of observations in a Bayesian model prior to clustering.[63]

Today, approximately one-third of spatial transcriptomics tools incorporate space in one way or another, and more of these tools are constantly being developed. Following, we discuss common spatial applications that are addressed by such approaches.

Spatial clustering

Clustering for data exploration and inference is a primary analysis output for spatial transcriptomics data. Standard clustering analysis doesn't consider the spatial distribution of values but only their similarity. As such, spatial autocorrelation is of particular concern with respect to clustering. Spatial autocorrelation underpins spatial patterns of gene expression resulting in hot and/or cold spots that can subsequently affect clustering. A distinction should be drawn here between *spatial clustering*, which seeks to define a geographically co-located group of similar observations, and *geographically weighted clustering*, which looks to account for spatial features of data (such as spatial autocorrelation) when producing clusters.

An example of spatial clustering is the method SpaGCN, which integrates gene expression, spatial location, and histology data to construct a network and detect spatially variable genes enriched in a specific spatial domain in order to generate clusters.[64] Geographically weighted clustering can be found in MERINGUE, which considers spatial positioning in addition to transcriptional profiles by weighting the edges of a K-nearest neighbor graph according to spatial adjacency, allowing spatially distinct cellular subpopulations to be resolved.[65]

Mapping of spatially variable genes

All transcriptomics experiments, whether bulk, single-cell, or spatial, are assaying gene expression. Spatial transcriptomics presents the opportunity to map this expression to specific locations on the tissue, which may also be associated with the presence of specific processes or cells. Generating activity maps of gene expression can reveal process-level insights into the functioning of these genes. More importantly, by accounting for spatial autocorrelation and examining spatial heterogeneity at the proper scale (or set of scales), it is possible to explore the underlying stationarity—or non-stationarity—of a process over space.

Approaches to mapping spatially variable genes (SVGs) include the modeling of gene expression with a generalized linear spatial model (GLSM)—the underlying stationary spatial process can be captured, and genes can be clustered based on the patterns that emerge.[66] An alternative approach uses a self-organizing map to construct a condensed map of neighboring cells clustered to nodes. Node-level spatial gene expression is used to identify SVGs using a Gaussian process.[62]

Spatial annotation

In a spatial transcriptomics experiment, the hematoxylin and eosin image that accompanies the expression data can be used for knowledge-driven segmentation. Once the tissue image is annotated, gene expression across tissue domains can be explicitly examined, and the coincidence between annotated regions and statistically defined clusters can be explored. Spatial transcriptomics studies often employ expert annotation to provide context,[63,67,68] though there is presently a lack of tools to enable this annotation to be produced systematically. SpatialLIBD is a method that offers functionalities that serve some of these needs. It enables the interactive visualization of spatial transcriptomics data and does provide for manual annotation, spot-by-spot.[69]

OTHER METHODS FOR SPATIAL TRANSCRIPTOMICS ANALYSIS

Spatial transcriptomics as a technique has a short history which also means that there is an ongoing proliferation of tools for many aspects of spatial transcriptomics data analysis. In many cases, these tools are leveraging techniques developed for bulk or, more commonly, single-cell analysis and applying them directly to spatial transcriptomics data. The plurality of analysis approaches for spatial transcriptomics data cover many use cases including clustering, deconvolution, image segmentation, 3D reconstruction, cell-cell interactions, and data integration, among others—a brief overview of the some of these aspects is provided below. Tools that fall outside of the category of spatially aware spatial transcriptomics analysis go beyond the scope of this work and have been well reviewed elsewhere.[27,70,71]

Clustering is a helpful tool because it structures and orders the data, allowing useful insights to be gained from complex, multivariate datasets and, subsequently, allowing researchers to use those insights to classify the observed data or to generate hypotheses. For example, in a study of inflammation in gingival tissue, k-means clustering with *t*-distributed stochastic neighbor embeddings (*t*-SNE) to group tissue locations identified three distinct region types.[72] Another method is Louvain clustering. It is implemented in some of the most popular single-cell analysis packages such as Seurat[73] and Scran[74] and has been used for clustering in a number of spatial transcriptomics studies. For example, Louvain clustering revealed the zonation patterns in healthy and diseased liver tissue using Visium data.[75]

Spatial transcriptomics technologies with resolutions greater than single cell, such as Visium, often require gene expression deconvolution to help understand the cell type composition of the captured regions. The most common approach to deconvolution relies on a related scRNA-seq dataset that can be used as a reference to identify the proportion of each cell type in the spatial transcriptomics data aggregates. Many computational approaches are taken to leverage this reference data to deconvolve the spatial transcriptomics data, including deep learning and AI,[76,77] Bayesian models,[78] and other statistical methods like maximum-likelihood estimation.[79]

In contrast to deconvolution, high-resolution techniques often require aggregation of data points at cell level. This can be achieved using image segmentation approaches to identify cell boundaries in the high-resolution microscope images of the tissue. Segmentation approaches can use the distribution of detected transcripts e.g., Baysor,[80] or can combine histology image data with the spatial transcriptomics output.[81]

Spatial transcriptomics data rarely stand alone, and it is often desirable to combine them with other related data types to enrich the interpretation of the individual datasets. CellTrek[82] is a computational method that uses a multivariate machine learning model to combine scRNA-seq and spatial transcriptomics datasets to achieve single-cell spatial mapping. Increasingly, multi-modal integration will be required, as tissues are probed with multiple technologies that produce different types and resolutions of output. Platforms for this type of integration are beginning to appear, such as Single-Cell Spatial Explorer[83] and Tangram.[76]

Many spatial transcriptomics analysis tools can undertake more than one analysis task or are designed with the whole spatial transcriptomics analysis pipeline in mind. These toolkits often provide a framework for the visualization of spatial transcriptomics data, as well as a platform for exploratory and statistical analysis. More than a dozen toolkits have been published to date and they incorporate a wide range of common features, including data preprocessing,[73,84] visualization,[85–87] data integration,[73,88] clustering,[73,87] and differential gene expression.[89,90] Although toolkits often share features with one another, these feature sets are not completely overlapping as each toolkit is designed with a specific analysis pipeline in mind and will often include bespoke methodologies.

Conclusions

As we have discussed, the considerations that arise from spatial data—the modifiable areal unit problem, spatial autocorrelation, and spatial heterogeneity—clearly manifest in spatial transcriptomics data. The existence of these factors, which could be seen as confounding to traditional analysis, is sufficient to suggest that spatially aware analysis of spatial transcriptomics data should be the default. There are instances of tools developed for spatial transcriptomics analysis that are essentially re-deriving statistics that have been in common use in the geographical sciences for decades. Notably, no specific aspects of spatial transcriptomics data prevent the direct use of already existing spatial models.

Around a third of the methods developed for spatial transcriptomics analysis to date incorporate space in some way, mostly in the identification of spatially variable genes and the spatial mapping of genes of interest. A minority of the tools that consider space are making active use of methodologies developed and proven in the geographical sciences. There are several tools that identify spatially "interesting" genes via spatial autocorrelation statistics such as Moran's *I* or Geary's *C*, but often the logical extension of this principle—that the presence of spatial autocorrelation undermines classical statistical models—is not taken. It is common to identify differentially expressed genes between tissue areas, defined either by expert annotation or clustering. Spatial autocorrelation informs us that these regions will vary in their gene expression simply because they are varying in space. Not accounting for this underlying variation will inevitably lead to a host of false-positive observations, alongside those genes that are genuinely changing expression.

To compound the effect of spatial autocorrelation, some of the typical analysis decisions taken, such as the aggregation of spots into zones of characterized biological function, introduce new challenges in terms of the choice of aggregation scale which are not currently considered in the downstream analysis. Recent advances in geography provide methods for determining the impact of the MAUP for the chosen aggregation scale and for selecting a more appropriate scale.[26] That being said, often the most rational way of dissecting a tissue region is by some known biology, because the results obtained can best be understood in light of that biology. However, the relationship between gene expression, space, and phenotype is complicated and multivariate. To rely on a one-to-one relationship between gene expression and observed biology is to make a big assumption. A data-driven approach to aggregation, while not disregarding the biology, could avoid these assumptions and consequently lead to more interpretable results.

One of the features that can be observed in Figure 4 is that even within biological "layers," genes do not behave consistently with respect to one another. This indicates that clusters are not defined by immovable boundaries and that gene expression within a cluster is not always homogeneous. It is safe to assume that the same principle applies at the edges of zones (whether layers or clusters or some other aggregation of locales). In ecology, these boundary regions have long been referred to as *ecotones*[91] and it is well recognized that borders are often not sharp but exhibit a gradient from one phenotype to another creating transition zones. This phenomenon is potentially amplified by the relatively low resolution of spatial transcriptomics platforms like 10X Visium, where the observation-level data are already an aggregate of around 10 cells that might span across different biological regions. Caution is therefore required when deriving conclusions based on the classification of spatial transcriptomics observations that incorporate hard boundaries.

The most commonly used tools for spatial transcriptomics analysis are not necessarily those that take a spatial approach. While this may change with time and with more user-friendly tools, it does not suggest that a "spatial by default" approach is taken for analyzing these data at present. Spatial transcriptomics data are bound to space. By properly considering the features of spatial data that separate it from other 'omics data, it is possible to improve the reliability of analysis output. Knowledge gained from the field of geospatial data analysis indicates that to fully explore the power offered by spatially resolved data, we need to investigate biology locally rather than dataset-wide. This reveals the existing heterogeneity over space that needs to be accounted for in analyses.

The geographical sciences have been working with spatial data for many decades and consequently have developed countless statistical techniques that account for or actively make use of space to understand a wide scope of processes. Methods such as geographically weighted regression,[55,92] geographically weighted principal component analysis,[93] spatially weighted approaches to clustering,[94] and generalized additive models[52] along with many more have all been extensively explored and applied to a range of data. The analysis of spatially resolved biomolecular data can be readily enhanced by the incorporation of these methods and collaboration with the scientists who have developed them. This inter-disciplinary approach will help with realizing the full potential of spatial transcriptomics data analysis.

DATA AND CODE AVAILABILITY

The data used to produce the figures come from the human dorsolateral prefrontal cortex (DLPFC) 10X Visium dataset[65] and specifically tissue section 151673. The code and processed data that produced the figures is available via Zenodo at: https://doi.org/10.5281/zenodo.8333525.

ACKNOWLEDGMENTS

E.Z. is supported by a studentship from the Medical Research Council Discovery Medicine North (DiMeN) Doctoral Training Partnership (grant number: MR/N013840/1).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Zhuang, X. (2021). Spatially resolved single-cell genomics and transcriptomics by imaging. Nat. Methods *18*, 18–22.
2. Avital, G., Hashimshony, T., and Yanai, I. (2014). Seeing is believing: New methods for in situ single-cell transcriptomics. Genome Biol. *15*, 110.
3. Xia, C., Fan, J., Emanuel, G., Hao, J., and Zhuang, X. (2019). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proc. Natl. Acad. Sci. USA *116*, 19490–19499.
4. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*, aaa6090.
5. Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., et al. (2018). Three-dimensional Intact-Tissue Sequencing of Single-Cell Transcriptional States. Science *361*, eaat5691.
6. Zollinger, D.R., Lingle, S.E., Sorg, K., Beechem, J.M., and Merritt, C.R. (2020). GeoMx™ RNA Assay: High Multiplex, Digital, Spatial Analysis of RNA in FFPE Tissue. Methods Mol. Biol. *2148*, 331–345.
7. Jerby-Arnon, L., Neftel, C., Shore, M.E., Weisman, H.R., Mathewson, N.D., McBride, M.J., Haas, B., Izar, B., Volorio, A., Boulay, G., et al. (2021). Opposing immune and genetic mechanisms shape oncogenic programs in synovial sarcoma. Nat. Med. *27*, 289–300.
8. Kim, J., and Eberwine, J. (2010). RNA: State memory and mediator of cellular phenotype. Trends Cell Biol. *20*, 311–318.
9. Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol. *18*, 35–45.
10. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82.
11. Rao, A., Barkley, D., França, G.S., and Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. Nature *596*, 211–220.
12. Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467.
13. Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., Di Bella, D.J., Arlotta, P., Macosko, E.Z., and Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat. Biotechnol. *39*, 313–319.
14. Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., et al. (2022). Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell *185*, 1777–1792.e21.
15. Krausgruber, T., Redl, A., Barreca, D., Doberer, K., Romanovskaia, D., Dobnikar, L., Guarini, M., Unterluggauer, L., Kleissl, L., Atzmüller, D., et al. (2023). Single-cell and spatial transcriptomics reveal aberrant lymphoid developmental programs driving granuloma formation. Immunity *56*, 289–306.e7.
16. Hasel, P., Rose, I.V.L., Sadick, J.S., Kim, R.D., and Liddelow, S.A. (2021). Neuroinflammatory astrocyte subtypes in the mouse brain. Nat. Neurosci. *24*, 1475–1487.
17. Ko, T., Nomura, S., Yamada, S., Fujita, K., Fujita, T., Satoh, M., Oka, C., Katoh, M., Ito, M., Katagiri, M., et al. (2022). Cardiac fibroblasts regulate the development of heart failure via Htra3-TGF-β-IGFBP7 axis. Nat. Commun. *13*, 3275.
18. Openshaw, S., and Taylor, P. (1979). A Million or So Correlated Coefficients: Three Experiments on the Modifiable Areal Unit Problem. In Statistical Applications in the Spatial Sciences, N. Wrigley, ed., pp. 127–144. London: Pion.
19. Dungan, J.L., Perry, J.N., Dale, M.R.T., Legendre, P., Citron-Pousty, S., Fortin, M.-J., Jakomulska, A., Miriti, M., and Rosenberg, M.S. (2002). A balanced view of scale in spatial statistical analysis. Ecography *25*, 626–640.
20. Wong, D.W.S. (2004). The Modifiable Areal Unit Problem (MAUP). In WorldMinds: Geographical Perspectives on 100 Problems: Commemorating the 100th Anniversary of the Association of American Geographers 1904–2004, D.G. Janelle, B. Warf, and K. Hansen, eds. (Springer Netherlands), pp. 571–575.
21. Atkinson, P.M., and Graham, A.J. (2006). Issues of Scale and Uncertainty in the Global Remote Sensing of Disease. In Advances in Parasitology Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications, S.I. Hay, A. Graham, and D.J. Rogers, eds. (Academic Press), pp. 79–118.
22. Brunsdon, C., and Comber, A. (2021). Opening practice: Supporting reproducibility and critical spatial data science. J. Geogr. Syst. *23*, 477–496.
23. Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. Environ. Plan. A *16*, 17–31.
24. Jones, K., Manley, D., Johnston, R., and Owen, D. (2018). Modelling residential segregation as unevenness and clustering: A multilevel modelling approach incorporating spatial dependence and tackling the. MAUP. Environment and Planning B: Urban Analytics and City Science *45*, 1122–1141.
25. Arbia, G., Benedetti, R., and Espa, G. (1996). Effects of the maup on image classification. Geogr. Syst., 123–141.
26. Comber, A., and Harris, P. (2022). The Importance of Scale and the MAUP for Robust Ecosystem Service Evaluations and Landscape Decisions. Land *11*, 399.
27. Palla, G., Fischer, D.S., Regev, A., and Theis, F.J. (2022). Spatial components of molecular tissue biology. Nat. Biotechnol. *40*, 308–318.
28. Zinner, M., Lukonin, I., and Liberali, P. (2020). Design principles of tissue organisation: How single cells coordinate across scales. Curr. Opin. Cell Biol. *67*, 37–45.
29. Codeluppi, S., Borm, L.E., Zeisel, A., La Manno, G., van Lunteren, J.A., Svensson, C.I., and Linnarsson, S. (2018). Spatial organization of the somatosensory cortex revealed by osmFISH. Nat. Methods *15*, 932–935.
30. Tobler, W.R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. Econ. Geogr. *46*, 234–240.
31. Moran, P.a.P. (1950). Notes on continuous stochastic phenomena. Biometrika *37*, 17–23.
32. Geary, R.C. (1954). The Contiguity Ratio and Statistical Mapping. Inc. Statistician *5*, 115–146.
33. Getis, A., and Ord, J.K. (1992). The analysis of spatial association by use of distance statistics. Geographical analysis *24*, 189–206.

34. Anselin, L. (1995). Local indicators of spatial association - lisa. Geogr. Anal. *27*, 93–115.

35. Anselin, L. (2019). A Local Indicator of Multivariate Spatial Association: Extending Geary's c. Geogr. Anal. *51*, 133–150.

36. Ord, J.K., and Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geogr. Anal. *27*, 286–306.

37. Keogh, E., and Mueen, A. (2017). Curse of dimensionality. In Encyclopedia of machine learning and data mining, C. Sammut and G.I. Webb, eds. (Springer US), pp. 314–315.

38. Wang, T., Li, B., Nelson, C.E., and Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinf. *20*, 40.

39. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

40. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

41. Shahbazi, M.N. (2020). Mechanisms of human embryo development: From cell fate to tissue shape and back. Development *147*, dev190629.

42. Briscoe, J., and Small, S. (2015). Morphogen rules: Design principles of gradient-mediated embryo patterning. Development *142*, 3996–4009.

43. Tyser, R.C.V., Mahammadov, E., Nakanoh, S., Vallier, L., Scialdone, A., and Srinivas, S. (2021). Single-cell transcriptomic characterization of a gastrulating human embryo. Nature *600*, 285–289.

44. Hildebrandt, F., Andersson, A., Saarenpää, S., Larsson, L., Van Hul, N., Kanatani, S., Masek, J., Ellis, E., Barragan, A., Mollbrink, A., et al. (2021). Spatial Transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver. Nat. Commun. *12*, 7046.

45. Baccin, C., Al-Sabah, J., Velten, L., Helbling, P.M., Grünschläger, F., Hernández-Malmierca, P., Nombela-Arrieta, C., Steinmetz, L.M., Trumpp, A., and Haas, S. (2020). Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nat. Cell Biol. *22*, 38–48.

46. Kadur Lakshminarasimha Murthy, P., Sontake, V., Tata, A., Kobayashi, Y., Macadlo, L., Okuda, K., Conchola, A.S., Nakano, S., Gregory, S., Miller, L.A., et al. (2022). Human distal lung maps and lineage hierarchies reveal a bipotent progenitor. Nature *604*, 111–119.

47. Carmona-Fontaine, C., Deforet, M., Akkari, L., Thompson, C.B., Joyce, J.A., and Xavier, J.B. (2017). Metabolic origins of spatial organization in the tumor microenvironment. Proc. Natl. Acad. Sci. USA *114*, 2934–2939.

48. Danenberg, E., Bardwell, H., Zanotelli, V.R.T., Provenzano, E., Chin, S.-F., Rueda, O.M., Green, A., Rakha, E., Aparicio, S., Ellis, I.O., et al. (2022). Breast tumor microenvironment structures are associated with genomic features and clinical outcome. Nat. Genet. *54*, 660–669.

49. Brunsdon, C., Fotheringham, A.S., and Charlton, M.E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geogr. Anal. *28*, 281–298.

50. Fotheringham, A.S., Charlton, M., and Brunsdon, C. (1997). Two techniques for exploring non-stationarity in geographical data. J. Geogr. Syst. *4*, 59–82.

51. Fotheringham, A.S., Brunsdon, C., and Charlton, M. (2003). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships (John Wiley & Sons).

52. Comber, A., Harris, P., Murakami, D., Tsutsumida, N., and Brunsdon, C. (2022). Geographically Varying Coefficient Regression: GWR-Exit and GAM-On? In 15th International Conference on Spatial Information Theory (COSIT 2022) Leibniz International Proceedings in Informatics (LIPIcs), T. Ishikawa, S.I. Fabrikant, and S. Winter, eds., pp. 13–21.

53. Comber, A., Brunsdon, C., Charlton, M., Dong, G., Harris, R., Lu, B., Lü, Y., Murakami, D., Nakaya, T., Wang, Y., et al. A Route Map for Successful Applications of Geographically Weighted Regression. Geogr. Anal. 55, 155-178.

54. Fotheringham, A.S., Yang, W., and Kang, W. (2017). Multiscale Geographically Weighted Regression (MGWR). Annals of the American Association of Geographers *107*, 1247–1265.

55. Oshan, T., Li, Z., Kang, W., Wolf, L., and Fotheringham, A. (2019). Mgwr: A Python Implementation of Multiscale Geographically Weighted Regression for Investigating Process Spatial Heterogeneity and Scale. ISPRS Int. J. Geo-Inf. *8*, 269.

56. Junttila, M.R., and de Sauvage, F.J. (2013). Influence of tumour micro-environment heterogeneity on therapeutic response. Nature *501*, 346–354.

57. Yuan, Y. (2016). Spatial Heterogeneity in the Tumor Microenvironment. Cold Spring Harb. Perspect. Med. *6*, a026583.

58. Chevrier, S., Levine, J.H., Zanotelli, V.R.T., Silina, K., Schulz, D., Bacac, M., Ries, C.H., Ailles, L., Jewett, M.A.S., Moch, H., et al. (2017). An Immune Atlas of Clear Cell Renal Cell Carcinoma. Cell *169*, 736–749.e18.

59. Lavin, Y., Kobayashi, S., Leader, A., Amir, E.-A.D., Elefant, N., Bigenwald, C., Remark, R., Sweeney, R., Becker, C.D., Levine, J.H., et al. (2017). Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. Cell *169*, 750–765.e17.

60. Huang, Y.-K., Wang, M., Sun, Y., Di Costanzo, N., Mitchell, C., Achuthan, A., Hamilton, J.A., Busuttil, R.A., and Boussioutas, A. (2019). Macrophage spatial heterogeneity in gastric cancer defined by multiplex immunohistochemistry. Nat. Commun. *10*, 3928.

61. Zhu, J., Sun, S., and Zhou, X. (2021). SPARK-X: Non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. Genome Biol. *22*, 184.

62. Hao, M., Hua, K., and Zhang, X. (2021). SOMDE: A scalable method for identifying spatially variable genes with self-organizing map. Bioinformatics *37*, 4392–4398.

63. Allen, C., Chang, Y., Neelon, B., Chang, W., Kim, H.J., Li, Z., Ma, Q., and Chung, D. (2023). A Bayesian multivariate mixture model for high throughput spatial transcriptomics. Biometrics *79*, 1775–1787.

64. Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D.J., Lee, E.B., Shinohara, R.T., and Li, M. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat. Methods *18*, 1342–1351.

65. Miller, B.F., Bambah-Mukku, D., Dulac, C., Zhuang, X., and Fan, J. (2021). Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. Genome Res. *31*, 1843–1855.

66. Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat. Methods *17*, 193–200.

67. Yoosuf, N., Navarro, J.F., Salmén, F., Ståhl, P.L., and Daub, C.O. (2020). Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. Breast Cancer Res. *22*, 6.

68. Maynard, K.R., Collado-Torres, L., Weber, L.M., Uytingco, C., Barry, B.K., Williams, S.R., Catallini, J.L., Tran, M.N., Besich, Z., Tippani, M., et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat. Neurosci. *24*, 425–436.

69. Pardo, B., Spangler, A., Weber, L.M., Page, S.C., Hicks, S.C., Jaffe, A.E., Martinowich, K., Maynard, K.R., and Collado-Torres, L. (2022). spatialLIBD: An R/Bioconductor package to visualize spatially-resolved transcriptomics data. BMC Genom. *23*, 434.

70. Moses, L., and Pachter, L. (2022). Museum of spatial transcriptomics. Nat. Methods *19*, 534–546.

71. Du, J., Yang, Y.-C., An, Z.-J., Zhang, M.-H., Fu, X.-H., Huang, Z.-F., Yuan, Y., and Hou, J. (2023). Advances in spatial transcriptomics and related data analysis strategies. J. Transl. Med. *21*, 330.

72. Lundmark, A., Gerasimcik, N., Båge, T., Jemt, A., Mollbrink, A., Salmén, F., Lundeberg, J., and Yucel-Lindberg, T. (2018). Gene expression profiling of

periodontitis-affected gingival tissue by spatial transcriptomics. Sci. Rep. *8*, 9370.

73. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29.
74. Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A Step-by-step Workflow for Low-Level Analysis of Single-Cell RNA-Seq Data with Bioconductor. F1000Res *5*, 2122.
75. Guilliams, M., Bonnardel, J., Haest, B., Vanderborght, B., Wagner, C., Remmerie, A., Bujko, A., Martens, L., Thoné, T., Browaeys, R., et al. (2022). Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. Cell *185*, 379–396.e38.
76. Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg, C.R., Segerstolpe, Å., Zhang, M., et al. (2021). Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nat. Methods *18*, 1352–1362.
77. Song, Q., and Su, J. (2021). DSTG: Deconvoluting spatial transcriptomics data through graph-based artificial intelligence. Brief. Bioinform. *22*, bbaa414.
78. Chu, T., Wang, Z., Pe'er, D., and Danko, C.G. (2022). Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat. Cancer *3*, 505–517.
79. Cable, D.M., Murray, E., Shanmugam, V., Zhang, S., Zou, L.S., Diao, M., Chen, H., Macosko, E.Z., Irizarry, R.A., and Chen, F. (2022). Cell type-specific inference of differential expression in spatial transcriptomics. Nat. Methods *19*, 1076–1087.
80. Petukhov, V., Xu, R.J., Soldatov, R.A., Cadinu, P., Khodosevich, K., Moffitt, J.R., and Kharchenko, P.V. (2022). Cell segmentation in imaging-based spatial transcriptomics. Nat. Biotechnol. *40*, 345–354.
81. Chen, H., Li, D., and Bar-Joseph, Z. (2023). SCS: Cell segmentation for high-resolution spatial transcriptomics. Nat. Methods *20*, 1237–1243.
82. Wei, R., He, S., Bai, S., Sei, E., Hu, M., Thompson, A., Chen, K., Krishnamurthy, S., and Navin, N.E. (2022). Spatial charting of single-cell transcriptomes in tissues. Nat. Biotechnol. *40*, 1190–1199.
83. Pont, F., Cerapio, J.P., Gravelle, P., Ligat, L., Valle, C., Sarot, E., Perrier, M., Lopez, F., Laurent, C., Fournié, J.J., and Tosolini, M. (2023). Single-cell spatial explorer: Easy exploration of spatial and multimodal transcriptomics. BMC Bioinf. *24*, 30.
84. Sztanka-Toth, T.R., Jens, M., Karaiskos, N., and Rajewsky, N. (2022). Spacemake: Processing and analysis of large-scale spatial transcriptomics data. GigaScience *11*, giac064.
85. Queen, R., Cheung, K., Lisgo, S., Coxhead, J., and Cockell, S. (2019). Spaniel: Analysis and interactive sharing of Spatial Transcriptomics data. Preprint at bioRxiv.
86. Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A.C., Kuemmerle, L.B., Rybakov, S., Ibarra, I.L., Holmberg, O., Virshup, I., et al. (2022). Squidpy: A scalable framework for spatial omics analysis. Nat. Methods *19*, 171–178.
87. Ospina, O.E., Wilson, C.M., Soupir, A.C., Berglund, A., Smalley, I., Tsai, K.Y., and Fridley, B.L. (2022). spatialGE: Quantification and visualization of the tumor microenvironment heterogeneity using spatial transcriptomics. Bioinformatics *38*, 2645–2647.
88. Bergenstråhle, J., Larsson, L., and Lundeberg, J. (2020). Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genom. *21*, 482.
89. Dries, R., Zhu, Q., Dong, R., Eng, C.-H.L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., Bao, F., et al. (2021). Giotto: A toolbox for integrative analysis and visualization of spatial expression data. Genome Biol. *22*, 78.
90. Wang, L., Maletic-Savatic, M., and Liu, Z. (2022). Region-specific denoising identifies spatial co-expression patterns and intra-tissue heterogeneity in spatially resolved transcriptomics data. Nat. Commun. *13*, 6912.
91. Fortin, M.-J., Drapeau, P., and Fortin, M.J. (1995). Delineation of Ecological Boundaries: Comparison of Approaches and Significance Tests. Oikos *72*, 323–332.
92. Lu, B., Harris, P., Charlton, M., and Brunsdon, C. (2014). The GWmodel R package: Further topics for exploring spatial heterogeneity using geographically weighted models. Geo-Spatial Inf. Sci. *17*, 85–101.
93. Gollini, I., Lu, B., Charlton, M., Brunsdon, C., and Harris, P. (2015). GWmodel: An R package for exploring spatial heterogeneity using geographically weighted models. J. Stat. Softw. *63*, 1–50.
94. Wijayanto, A.W., and Purwarianti, A. (2014). Improvement design of fuzzy geo-demographic clustering using Artificial Bee Colony optimization. In 2014 International Conference on Cyber and IT Service Management (CITSM), pp. 69–74.