Vasconcelos *et al. Genome Biology* (2026) 27:21
https://doi.org/10.1186/s13059-025-03867-1

Genome Biology

METHODOLOGY Open Access

Differential expression analysis for spatially correlated data using smiDE

Ana Gabriela Vasconcelos[1†], Daniel McGuire[2†], Noah Simon[1], Patrick Danaher[2*] and Ali Shojaie[1*]

†Ana Gabriela Vasconcelos and Daniel McGuire contributed equally to this work.

*Correspondence:
Patrick Danaher
Patrick.Danaher@bruker.com
Ali Shojaie
ashojaie@uw.edu
[1]Department of Biostatistics, University of Washington, Seattle, USA
[2]Bruker Spatial Biology, Seattle, USA

Abstract
Differential expression is a key application of imaging spatial transcriptomics, moving analysis beyond cell type localization to examining cell state responses to microenvironments. However, spatial data poses new challenges to differential expression: segmentation errors cause bias in fold-change estimates, and correlation among neighboring cells leads standard models to inflate statistical significance. We find that ignoring these issues can result in considerable false discoveries that greatly outnumber true findings. We present a suite of solutions to these fundamental challenges, and implement them in the R package *smiDE*.

Keywords Spatial transcriptomics, Differential expression, Segmentation error mitigation, Spatial correlation, Spatial random effects model

Background

Imaging-based spatial transcriptomics (ST) technologies detect tissues' RNA molecules in situ, producing single cell expression data while preserving cell's spatial locations [1, 2]. This pairing of gene expression data with spatial context creates an opportunity to survey how cells respond to their environment or to interactions with other cell types. Consider, for example, a tumor containing cancer cells, T-cells, and other cell types (Fig. 1a) falling in two distinct regions: the tumor bed and the stroma (Fig. 1b). Spatial transcriptomics allows us to answer questions such as: how do one or more cell types modulate their gene expression differently across functional regions of the tissue (Fig. 1c)? Or, how does it change gene expression when co-localized with another cell type (Fig. 1d)? Or how does the cell type change expression in treated vs. untreated tumors?

These questions can be answered through differential expression (DE) analysis. Here we define DE as identifying changes in the expression level of genes associated with another factor, such as tissue region, treatment condition, or co-localization with another cell type. DE analysis has been a mainstay of gene expression studies, and becomes even more powerful in the context of single cell spatial transcriptomics. Many methods for DE analysis are available, such as DEseq2 [3], limma [4], and MAST [5], which were originally developed for traditional bulk or single cell gene expression studies. However,

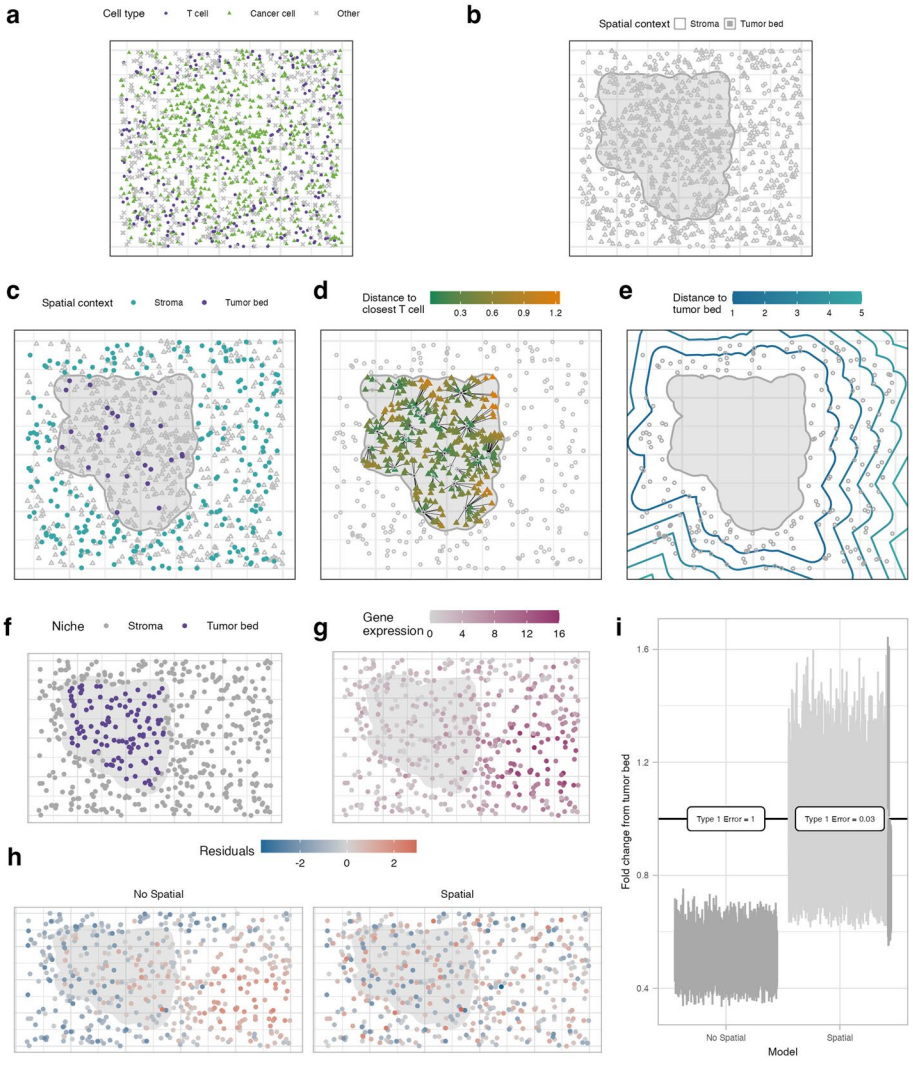

Fig. 1 Examples of spatially informed DE analyses and illustration of spatial confounding. We illustrate several ways DE can be formulated in spatial transcriptomics data. Assume we are interested in T-cells and cancer cells (**a**), which are located either in a tumor bed or stroma (**b**). Possible examples of DE analysis are: contrasting gene expression between T-cells in tumor bed vs stroma (**c**); analyzing how expression of genes of cancer cells in tumor bed differentiates based on distance to nearest T-cell (**d**); or determining how expression of genes of T-cells in stroma depends on distance to tumor bed (**e**). Panels **f**–**i** show a toy example illustrating how spatial correlation can induce spurious DE if ignored. We simulate data in two spatial contexts, tumor bed and stroma (**f**), with gene expression generated independently of region but following a spatially autocorrelated pattern (**g**). Panel **i** shows fold-change estimates of stroma vs. tumor bed: the independent model yields biased estimates with confidence intervals that fail to cover the true value of 1, falsely suggesting DE, whereas the spatial model properly accounts for spatial correlation and controls type I error

spatial transcriptomics data presents unique challenges that these methods were not designed for. First, imperfect segmentation of cell boundaries means some RNA molecules are incorrectly assigned to neighboring cells, inducing spatially-dependent bias in expression profiles. Second, cells that neighbor each other are likely to have correlated expression profiles, reflecting dependence that arises from shared microenvironments or tissue structure. If unaccounted for, this correlation can distort differential expression results (Fig. 1f-i). Finally, sample sizes can be enormous, with $> 10^6$ cells assayed per sample. While this problem is not unique to spatial transcriptomics, it makes it computationally expensive to model the spatial correlation between cells.

Despite their importance, the main challenges of DE analysis in spatial transcriptomics have not been fully addressed. A commonly used approach, CSIDE [6], is designed for spot-level data where each observation contains multiple cells, and it accounts for cell type composition through deconvolution. While CSIDE can also be applied to single-cell data, doing so requires an additional deconvolution step, making the application cumbersome. Moreover, like most existing DE approaches, CSIDE does not directly account for spatial correlation, which may be less problematic in spot-based data but can lead to misleading results in single-cell settings. Another recent approach is *niche-DE* [7], which performs DE to identify genes whose expression within a given cell type is up- or down-regulated depending on the presence of neighboring cell types (termed *spatial niches* or *effective niches*). While *niche-DE* leverages spatial information about local cell-cell composition, it does not include an explicit spatial random effect and therefore does not directly model residual spatial correlation. Importantly, it targets a single kind of DE question—association with a predefined effective-niche covariate—rather than allowing arbitrary contrasts (e.g., multi-region or one-vs-all comparisons, continuous spatial gradients, or interaction terms).

The importance of accounting for spatial correlation in DE analysis has been emphasized by Ospina et al. [8], who demonstrated that ignoring spatial dependence can lead to model misspecification and inflated false positives. Their framework uses Gaussian models and existing spatial methods, but it does not address the prevalence of zero inflation or overdispersion in count data, and the approaches they consider are computationally intensive.

A different class of methods, including SpatialDE [9], SPARK [10], Trendscreek [11] and BOOST-GP [12], focuses on identifying genes with *spatially variable expression* (SE). These methods perform variance component tests that decompose the variability between spatial and non-spatial sources. However, instead of finding genes with expressions that vary across space, many studies aim to interrogate how genes respond to specific, pre-defined variables, a problem not addressed by a simple search for variability [13–15]. This represents a distinct problem that existing SE-focused methods do not directly address.

Here, we introduce a comprehensive pipeline for spatial DE analysis that tackles these key challenges. Our framework accommodates different distributional assumptions, incorporates metrics to mitigate cell segmentation errors, and evaluates multiple strategies for modeling spatial correlation, including cluster-based approximations and stochastic partial differential equation (SPDE) representations, with an emphasis on computational efficiency. Through extensive simulations and real data analyses, we demonstrate that accounting for these factors is essential to reduce false discoveries and improve the reproducibility and generalizability of findings. An open-source implementation is provided in the R package *smiDE*, together with practical recommendations for best practices in spatial DE analysis.

Results

smiDE: a pipeline for differential expression in spatial transcriptomics

The *smiDE* pipeline aims to facilitate highly flexible regression modeling suitable for DE analysis with spatial transcriptomic data, where the scientific question involves one

or more covariates hypothesized to be associated with gene expression across physical space in a tissue.

The first step in the pipeline is to pose our question in a regression framework. Importantly, this definition of DE is intentionally broad: the factor of interest may represent a binary contrast (e.g., tumor bed vs. stroma), a categorical variable (e.g., multiple niches), or a continuous measure (e.g., distance to another cell type or anatomical landmark). By treating all of these as covariates in the design matrix, *smiDE* provides a unified framework for testing how gene expression varies with respect to spatially or biologically meaningful features. Once the contrast is defined, an appropriate distributional model can be selected. *smiDE* supports Poisson, Negative binomial, and Gaussian models, allowing users to balance computational efficiency and accuracy according to the characteristics of their data. The software automatically returns a rich set of results for each gene and covariate, and provides a unified syntax for fitting models across distributional models, as well as with or without sample level random effects.

To address technical challenges unique to spatial assays, *smiDE* incorporates two complementary procedures for mitigating bias from cell segmentation errors: filtering genes that are particularly prone to bias, and adjusting for the effects of segmentation errors within the modeling framework. In addition, the pipeline implements multiple strategies for handling spatial correlation, including Gaussian process-based models and scalable cluster-level approximations. These options balance statistical rigor with computational feasibility, making the framework adaptable to datasets of varying resolution and size.

Together, these components define a workflow tailored to the challenges of differential expression analysis in single-cell spatial transcriptomics, encompassing contrast specification, distributional modeling, and adjustments for contamination and spatial correlation. The pipeline is implemented in the open-source R package *smiDE* (available at https://github.com/Nanostring-Biostats/CosMx-Analysis-Scratch-Space/tree/Main/_code/smiDE). In the sections that follow, we describe each component in detail and provide practical guidelines for analysis, beginning with distributional modeling, then addressing contamination, and finally outlining and benchmarking strategies for incorporating spatial correlation.

Choosing parametric distribution for DE regression model

For scRNA-seq data, Negative binomial regression models are commonly used to model the observed expression of RNA transcripts across cells, often using random effects in multisubject studies to account for heterogeneity across independent biological samples [16].

Assuming the data are generated from a Negative binomial distribution is a reasonable a priori assumption, as count data are often overdispersed relative to a Poisson distribution due to both technical and biological sources of variation [17]. Due to the increasing scale of modern ST datasets, with potentially millions of cells across multiple biological samples, we compared performance tradeoffs of the principled choice of a reference Negative binomial regression model against approximations which may be more computationally tractable in large-scale studies.

A mixed Negative binomial model with random effects (RE) predicted for biological samples (Negative binomial RE) may be chosen from the first principles as the model that is expected to best describe the data. We may specify the regression model as

$$y_{ijg}|N_{ij}, \mu_{ijg} \sim \text{Negative binomial}(N_{ij}\mu_{ijg}, \phi_g), \tag{1}$$

where y_{ijg} are the observed counts for cell *i*, biological sample *j*, and gene *g*. N_{ij} are the total observed transcripts for cell *i* across all genes, and μ_{ijg} is the expected rate of expressed transcripts for cell *i*, gene *g*. The parameter ϕ_g is a size factor that correlates with the level of overdispersion in gene expression *g* relative to the Poisson distribution. The expected expression rate μ_{ijg} may be parameterized by covariates

$$\log(\mu_{ijg}) = \beta^T X_{ij} + b_j, \tag{2}$$

where X_{ij} and β represent the observed covariates of interest associated with cell *i* in sample *j* and their regression coefficients. The term b_j is a random effect shared by all cells belonging to sample *j*.

As an alternative to the Negative binomial RE model, we also consider a Poisson model (Poisson RE) parametrized as

$$y_{ijg}|N_{ij}, \mu_{ijg} \sim \text{Poisson}(N_{ij}\mu_{ijg}), \tag{3}$$

with the same parameterization for μ_{ijg} as in (2). Finally, we also consider a simple Gaussian model (Gaussian RE). Due to the importance of the total transcript count N_{ij} as a cell-specific scaling term which accounts for technical variation in cells' measurement efficiency, and the desire to compare fold change estimates on the same scale as the log-linear Negative binomial and Poisson models, we consider the following normalization for regression with Gaussian family:

$$y^*_{ijg} = \frac{\bar{N} y_{ijg}}{N_{ij}} \tag{4}$$

$$y^*_{ijg}|\mu_{ijg} \sim \text{Normal}(\mu_{ijg}, \epsilon_{ijg}), \tag{5}$$

where $\bar{N}$ is the average total transcript count across all cells. For the Gaussian RE case, we avoid a direct log transformation of the normalized counts. Use of a log transformation would require taking the form $\log(y^*_{ijg} + c)$, where a constant '*c*' must be chosen to avoid infinite values when the raw count is zero. In this transformation, the choice of '*c*' is arbitrary and can potentially have a non-negligible impact on the regression model and its error distribution, particularly for low expressed genes where zero counts are frequent.

Using the CosMx NSCLC dataset [1] with 487,819 cells from 5 NSCLC tissues, we compared the above parametric models (Negative binomial RE, Poisson RE, and Gaussian RE) in terms of computation time, concordance in estimation of effect size, and significant gene calls. We ran DE regression models for 7 cell types (macrophage, T-reg, T-CD8, T-CD4, plasmablast, neutrophil, and tumor). A categorical covariate 'niche' annotates the functional region of tissue the cell resides in, and was treated as the primary covariate of interest in the regression models. This 'niche' covariate had 9 categories (tumor interior, storma, tumor-stroma boundary, myeloid-enriched stroma lymphoid structure, immune, plasmablast-enriched stroma, and macrophages). For each cell type, we assessed whether a gene was DE in a particular niche compared to the cell-weighted average of all other niches.

We found that concordance in fold change estimates between all three parametric families was high, with fold change estimates between Negative binomial and Poisson families virtually identical in macrophage cells (Fig. 2a), and across cell types (Additional file 1: Figure S1). Poisson RE models produced consistently larger test-statistics (in magnitude) and smaller *p*-values compared to Negative binomial and Gaussian RE models (Fig. 2b). Such inflated type I errors are expected from models that fail to account for count overdispersion, resulting in standard errors which are too small. Gaussian regression tended to call fewest significant genes of any family in macrophages (Fig. 2b-c) and across cell types (Additional file 1: Figure S2); however, a higher proportion of Gaussian RE significant gene-contrasts overlapped with Negative binomial RE (238/284 = 83%) compared to Poisson RE (320/408 = 78%) (Fig. 2c), a pattern which persisted across cell types (Additional file 1: Figure S3). We observed that the degree of overlap in significant gene sets between families tended to increase with the number of cells analyzed (Additional file 1: Figure S3).

In terms of computation efficiency, Gaussian RE models were fastest in elapsed time per single gene when the number of cells modeled were small, though the gap in modeling time became smaller as the number of cells increased. Model estimation time was similar overall at roughly 30-40k cells (Fig. 2d). Somewhat surprisingly, Negative binomial RE models were fastest at the largest sample size of DE for 227k tumor cells. Negative binomial and Poisson RE models were estimated using the NEBULA package, which implements a fast large sample analytical approximation to the model estimation, resulting in computation times which are orders of magnitude faster than other package implementations [16]. However, NEBULA does not support analysis of spatially

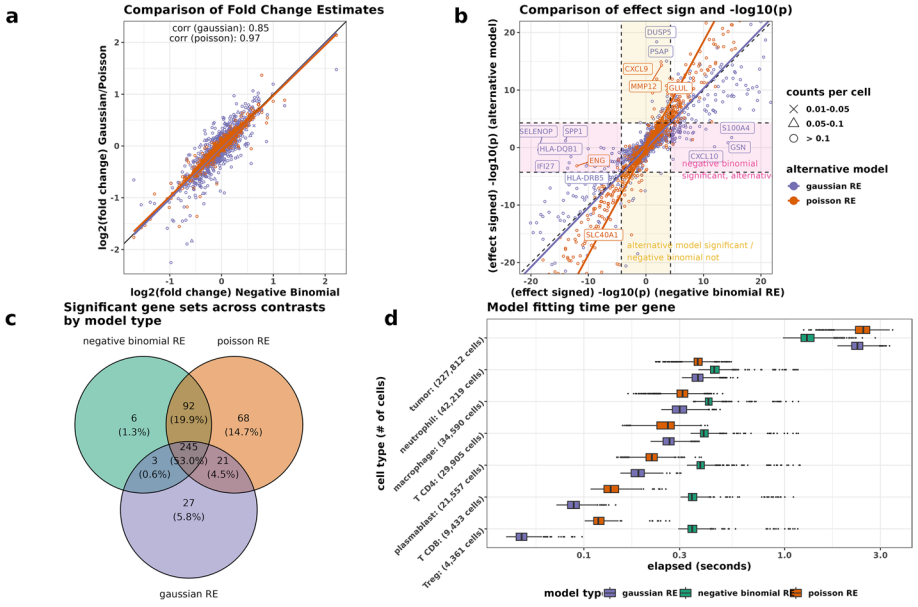

Fig. 2 Comparison between parametric families for cell type specific DE between spatial niches. **a** log2 fold changes for genes DE in macrophage cells between Negative binomial RE models (x-axis) compared to the corresponding log2 fold change estimates in gaussian or poisson RE models. **b** -log10(*p*-values) multiplied by the estimated direction of effect (+/-) compared between Negative binomial RE models (x-axis) with gaussian or poisson RE models (y-axis) for macrophage cells. Highlighted pink regions indicate genes where a spatial contrast and gene is significant using Negative binomial RE, and but not significant using alternative model at a bonferroni threshold of (0.05/(total # of genes in the panel)). **c** Venn diagram of overlapping significant gene contrasts between 3 model types. **d** Computation time comparisons between gaussian, Negative binomial, and poisson RE models across 7 cell types

correlated random effects, so the optimal models discussed in *Spatial Modeling* do not benefit from the speed described here.

Countermeasures for cell segmentation error

Often a key challenge in cell type specific analysis of ST data is dealing with the impact of erroneous transcript assignment due to imperfect or overlapping cell boundaries. While many methods have been developed for single cell segmentation of ST data [18, 19], even minor errors may potentially confound or distort cell type specific downstream analyses. As an illustrative example, we consider an analysis of macrophage cells which seeks to identify genes DE in tumor cell infiltrated regions (tumor-interior) of the tissue compared to the rest of the tissue. Identifying genes in macrophage cells that are DE in the tumor-interior may reveal important biological mechanisms by which macrophages promote immunosuppression or regulate cancer progression in the tumor [20–22]. Figure 3a shows a macrophage cell surrounded by cancer cells in a tumor bed. Due to imperfect definition of cell boundaries, a small number of KRT17 transcripts expressed by a neighboring cancer cell are falsely assigned to the macrophage cell. In the DE model, these errors are correlated with our scientific question in a way that confounds the DE analysis (Fig. 3b); the tumor-interior region is more likely to have KRT17-impacted macrophage cells than non-tumor regions of tissue due to the abundance of

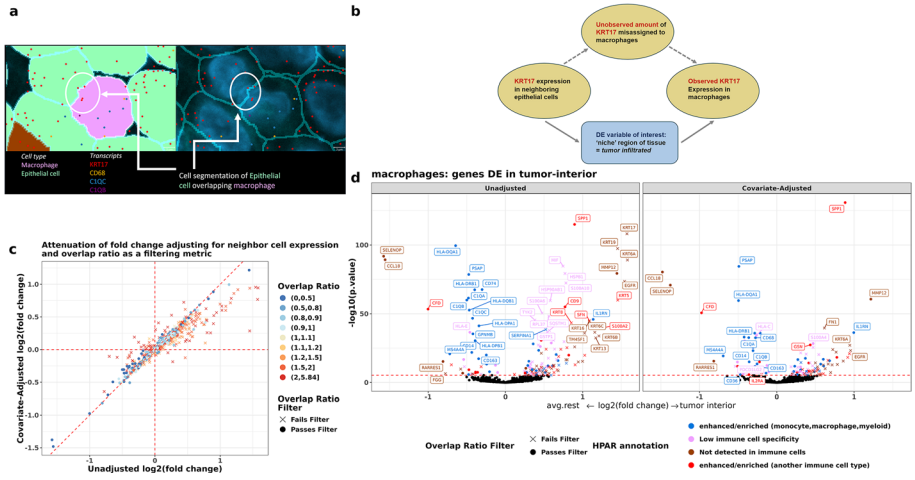

Fig. 3 Impact of segmentation error on DE analysis. **a** Illustrative example of a macrophage cell (pink) surrounded by cancerous cells in a tumor cell-infiltrated region of tissue. KRT17 transcripts (red dots) expressed by a neighboring epithelial cell breaches the estimated border of the macrophage cell. The cell borders are ambiguous; evident by overlapping nuclei of the two cells when viewed in two dimensions result in false positive counts of KRT17 assigned to the macrophage. **b** DAG showing causal path by which segmentation error correlates with both the response (KRT17 expression in macrophage) and estimand of interest (DE variable), acting as an unobserved confounder. Covariate adjustment via KRT17 expression in neighboring tumor cells can act as a proxy variable to reduce confounding bias. **c** Estimates of log2 fold change, with and without covariate adjustment for neighboring cell expression in macrophages. Genes which are more expressed in neighboring cell types (red, Overlap Ratio Metric >1), are attenuated towards zero after adjustment, while genes which are expressed more in the cell type of interest (blue, Overlap Ratio Metric < 1) have similar estimates with and without covariate adjustment. **d** A comparison between 'Unadjusted' and 'covariate-adjusted' volcano plots. Each gene is colored by a condensed description of the annotation provided for "Immune cell specificity" in the Human Protein Atlas (HPAR)[23, 24]. Based on prior biology, blue annotated genes are known to be expressed in the analyzed cell type (macrophage) and are the least controversial while brown genes ("Not detected in immune cells") are the most prone to be suspect. The majority of significant suspect genes become non-significant in the adjusted model. Some remaining suspect brown genes may be further disregarded in an analysis which applies a gene filtering metric

surrounding cancer cells—the abundance of cancer cells being a primary signature of the 'tumor-interior'.

Next, we introduce two countermeasures to the artifacts caused by segmentation errors. First, we propose a rule for flagging genes within a celltype most likely to be impacted by segmentation errors. Second, we expand our models to explicitly model gene expression in neighboring cells as a control variable, thereby weakening the effect of segmentation error as a confounder.

A segmentation overlap ratio metric for cell type-specific gene filtering

We motivate a metric for preemptively excluding genes likely to return spurious DE results due to imperfect or overlapping cell segmentation boundaries, which we call an "overlap ratio metric" (ORM). Intuitively, when analyzing a specific cell type, we expect that a gene may be prone to error if it has low expression in the cell type of interest and high expression in cells spatially adjacent to the cell type of interest (Fig. 3a).

Let $\bar{y}(t)_g$ denote the average expression of a particular gene g for the type t cells of interest. Further denote the average expression per cell of the gene in the *neighboring cells* of other cell types by $\bar{y}(t')_g$. Then the we define the ORM as

$$ORM_g = \frac{\bar{y}(t')_g}{\bar{y}(t)_g} \tag{6}$$

A more detailed formulation of the ORM is provided in Methods section. This metric quantifies susceptibility to confounding bias from overlapping or erroneous cell boundaries. To get a sense of how different choices of cutoff may reflect the number of analyzed genes and plausibility significant DE results, we used the CosMx NSCLC dataset (Additional file 1: Figure S4). The cumulative distribution plot in Figure S4a shows the number of accepted genes corresponding to different cutoffs for each of 7 cell types (macrophage, T-reg, T-CD8, T-CD4, plasmablast, neutrophil, and tumor). Thus number varies substantially by cell type; for example in tumor cells 67.2% (645/960) of all genes have ORM < 1, while only 13.5% (180/960) meet the same criteria for macrophage cells, and even fewer 5.6% (54/960) for plasmablast cells. This can reflect a cell type's heterogeneity of gene expression, accuracy of detection, and spatial density. For example, tumors from different patients display large heterogeneity in cellular composition, chromosomal structure, developmental trajectory [25] leading to highly heterogenous expression profiles. Plasmablast cells, on the other hand, are highly specific in function with a comparatively narrow expression profile.

To build intuition for how a cutoff may correlate with an analyst's tolerance for false discoveries, we summarized the descriptions provided for "Immune cell specificity" in the Human Protein Atlas (HPAR)[23, 24] and looked at their distribution by cell type and by cutoff Figures S4b-d. For example, for macrophage cell type (Additional file 1: Figure S4c) we see that of the 180 genes with ORM < 1, 11.1% are labeled as "Not detected in immune cells" according to HPAR, while 37.2% are noted to be enhanced or enriched with any of a relevant monocyte, macrophage, myeloid, or dendritic (DC) descriptor, and an additional 26.1% marked as having low immune cell specificity Figure S4c. Including genes for analysis beyond this threshold shows a sharp increase in the rate of genes "Not detected in immune cells", as well as genes that are "enhanced/enriched" in other cell types. Increasing the threshold to include more genes in a DE

analysis increases the risk that we may find a significant DE result due merely to segmentation errors. However, setting the threshold too low runs the risk of missing some real and interesting DE gene discoveries.

Genes where $ORM_g > 1$ are higher expressed, on average, in neighboring cells of other types than that of the cell type of interest; hence the observed number of transcripts in analyzed cells may contain unacceptable levels of error when the ORM is much greater than 1. While values other than 1 might be considered as cutoffs for determining whether to analyze a particular gene for a given cell type, we use the threshold of 1 as an intuitive cutoff in our analysis (Fig. 3c-d). The optimal cutoff will vary with the accuracy of cell segmentation and with the analyst's tolerance for false discoveries. Regardless of the cutoff, this statistic provides a measure of the skepticism each gene deserves.

A covariate adjustment approach using gene expression in neighboring cells

An alternative or complementary approach to removing a gene from analysis via ORM, is to adjust for propensity of transcript assignment error within the regression model. In practice, the exact amount of erroneous transcripts in individual cells is unknown, but is correlated with the expression of the analyzed gene in spatially neighboring cells, making 'total neighbor expression' a useful proxy covariate to include in the regression model (Fig. 3b).

Consider a gene and a set of cells of the same type. To capture the gene's propensity for error within the analyzed cells, we measure its total expression in neighboring cells of other cell types (we expect that error from neighbors of the same cell type will have negligible impact on analyses where neighboring cells have highly similar values of predictor variables). Denote this propensity for error in transcript assignment with a spatially lagged expression vector l_g, where $l_g = W(r)y(t')_g$; here, $W(r)$ is a neighborhood indicator matrix indicating the neighboring cells of other cell types within a distance r of the cells to be analyzed, and $y(t')$ is the expression of the gene in cell types *other than* type t. If we assume that errors in transcript assignment increase the average observed expression by a factor proportional to l_g, we can update our regression model in Eq. (2) to include this factor as a proxy covariate for unobserved transcript assignment error. The updated model is specified as

$$\log(\mu_{ijg}) = \beta^T x_{ij} + b_j + \delta l_{ijg}. \tag{7}$$

While we describe the expression in neighboring cells as having an additive effect, the inclusion of this covariate in log-linear regression models with a multiplicative effect can be a convenient and effective approximation.

Given that extremely close neighboring cells may be more likely to impact a cell of interest compared to farther neighboring cells, we may also choose to weight the neighboring expression by the distance of the neighboring cell, such that the elements of the neighborhood matrix $W(r)$ follow

$$w_{ij} = \begin{cases} 1/d_{ij} & d_{ij} \leq r \\ 0 & d_{ij} > r, \end{cases}$$

where d_{ij} is the distance between a cell i to a neighbor cell j. To stabilize model estimation with this covariate (which is typically heavily right-skewed) we apply a rank-based inverse normal transformation to l_g, i.e., $\tilde{l}_g = \Phi^{-1}(l_g)$, where Φ denotes the cumulative

distribution function for a standard normal distribution. In practice, l_g may be a noisy approximation of actual transcript assignment error, but a useful proxy variable, which may greatly reduce false positives while limiting the number of genes that must be disregarded via a gene filtering metric [25].

Covariate adjustment and gene filtering effectively mitigate false positive associations arising from segmentation error

We compared the behavior of the gene flagging metric and covariate adjustment method across DE analyses of macrophage, T-CD8, T-CD4, and T-reg cell types across 5 patients in the CosMx NSCLC dataset [1]. For each cell type, we fit Negative binomial RE regression models and looked at genes DE in any of the 9 spatial niche regions annotated (Fig. 3c-d). For each cell type, we considered both a naïve model without covariate adjustment for neighboring gene expression, as well as a covariate adjusted version, and considered the significant genes before and after applying a ORM filter (Fig. 3d).

To quantify the impact of segmentation error, we labeled genes according to their Human Protein Atlas Reference annotation for "Immune Cell Specificity" [23, 24]. Significant DE genes labeled as "Not detected in immune cells" are likely false positives that arise from segmentation errors, while genes labeled as "Enhanced or enriched" in the corresponding cell type may be considered plausible. Figure 3d compares volcano plots for macrophage cell DE genes in the tumor interior, where the naïve unadjusted model identifies a number of cancer cell enriched keratin genes, which are over-expressed in the tumor-interior due to errors in segmentation. Fold changes and p-values for false positive genes were largely attenuated by covariate adjustment, and almost entirely eliminated when removing genes according to the ORM (Fig. 3c-d). For example, likely false positive genes were removed for 112/138, 68/72, 27/27, and 14/16 significant genes labeled as "Not Detected in Immune cells" in macrophage, T-CD4, T-CD8, and T-reg cell types, respectively, when using a combined approach of covariate adjustment and gene filtering (Fig. 3d, Additional file 1: Figure S5).

In practice, we recommend a combined approach of filtering as well as covariate adjustment. Filtering is an effective and easy way to remove a majority of problematic genes, and covariate adjustment further refines estimated effect sizes and inference.

Accounting for spatial correlation

Spatial transcriptomics techniques offer the advantage of preserving the cell organization within tissue samples. A key consequence is that neighboring cells often exhibit more similar gene expression profiles compared to cells located at greater distances. Ignoring this spatial correlation can bias inference, as illustrated by the simulated example in Fig. 1f-i.

Here, we consider whether a gene is differentially expressed in T cells located in the tumor bed versus the stroma (Fig. 1f). Two modeling approaches are considered: one that accounts for spatial correlation (Gaussian Process explained further in Methods section) and one that assumes independence of cells ("Naive"). In this example, gene expression in some cells is higher due to a spatially smooth factor unrelated to niche (Fig. 1g), such as a localized inflammation for example. These factors generate spatial dependence in gene expression. Examining the residuals of the two models (Fig. 1h) reveals a spatial pattern under the Naive model, indicating poor model fit. This lack of

fit is mitigated when the spatial correlation is modeled. By not accounting for this correlation when testing for differential expression between T cells in tumor bed or in the stroma, we would incorrectly conclude that the gene is downregulated in T cells inside the tumor (Fig. 1i), when in reality the difference in expression is related to other unrelated factors. This simple example highlights the importance of accounting for spatial correlation in order to avoid the spatial bias and correctly understand the true underlying signal.

Modeling strategies

The toy example above illustrates how ignoring spatial correlation can lead to false discoveries in differential expression testing. To address this issue, we extend the baseline Negative binomial mixed model in Eq. (2) to include spatially structured random effects. Formaly, let y_{ig} denote the transcript counts of the gth gene ($g = 1, \ldots, p$) in the ith cell ($i = 1, \ldots, n$). For each gene g we model the conditional mean μ_g of $y_g = (y_{1g}, \ldots, y_{ng})$ as

$$g(\mu_g) = Z\alpha + X\beta + s, \tag{8}$$

where g is a link function, either the identity for Gaussian distribution, or log function for Negative binomial. The vector X represents the covariate of interest, with β the DE effect sizes associated with it, and Z represents the covariates to be adjusted for, for example the spatially lagged expression vectors $l_1, \ldots, l_p$, with α the coefficients associated with it. The random component s is assumed to have mean zero and covariance matrix Σ, which describes the correlation between cells. Different specifications of Σ lead to alternative strategies for modeling spatial dependence. In *smiDE*, we consider four main approaches: (i) Gaussian Process (GP) models, where correlation is determined by a distance-based kernel; (ii) cluster-based models, which capture dependence within local groups of cells; (iii) SPDE-based approximations, which provide scalable representations of GP models using sparse precision matrices; and (iv) the BYM2 model, which combines structured and unstructured spatial effects in a Bayesian framework. These strategies represent a spectrum of trade-offs between flexibility, interpretability, and computational efficiency, and are described in detail below.

Gaussian process models

GP models describe spatial correlation as a smooth function of pairwise (Euclidean) distances between cells. Typically a Matérn kernel is used, with the range parameter ρ determining the spatial scale of correlation and the variance parameter σ^2 capturing the overall magnitude of spatial variability (Methods section). This formulation is highly flexible and often considered the gold standard [8], but comes with significant computational cost. Spatial single-cell transcriptomics data often contain hundreds of thousands of cells, and modeling correlation at the single-cell level requires dealing with a covariance matrix of extremely large size. Since DE analysis is typically performed for thousands of genes, the computational challenge is further amplified. For these reasons, approximate strategies are necessary to make GP modeling feasible in practice.

Cluster-based models

A practical strategy to reduce computational burden is to model spatial correlation not between individual cells, but between clusters of cells obtained by clustering based on location (Fig. 4a). In this framework, random effects are defined at the cluster level, assuming cells within each cluster are similar to one another. We can either assume independence between clusters (capturing only within-cluster similarity) or model between-cluster correlation using BYM2 (adjacency of clusters) or a GP (distance between cluster centroids).

Modeling correlation at the cluster level introduces the challenge of choosing the number of clusters, which reflects a trade-off between computational efficiency and the granularity of spatial adjustment. Under the independence assumption, increasing the number of clusters approaches the case of no spatial correlation at all (Additional file 1: Figure S6). For spatial models, a larger number of clusters provides finer resolution and approximates single-cell modeling, but comes at the cost of increased computational burden. To make recommendations comparable across datasets with different number

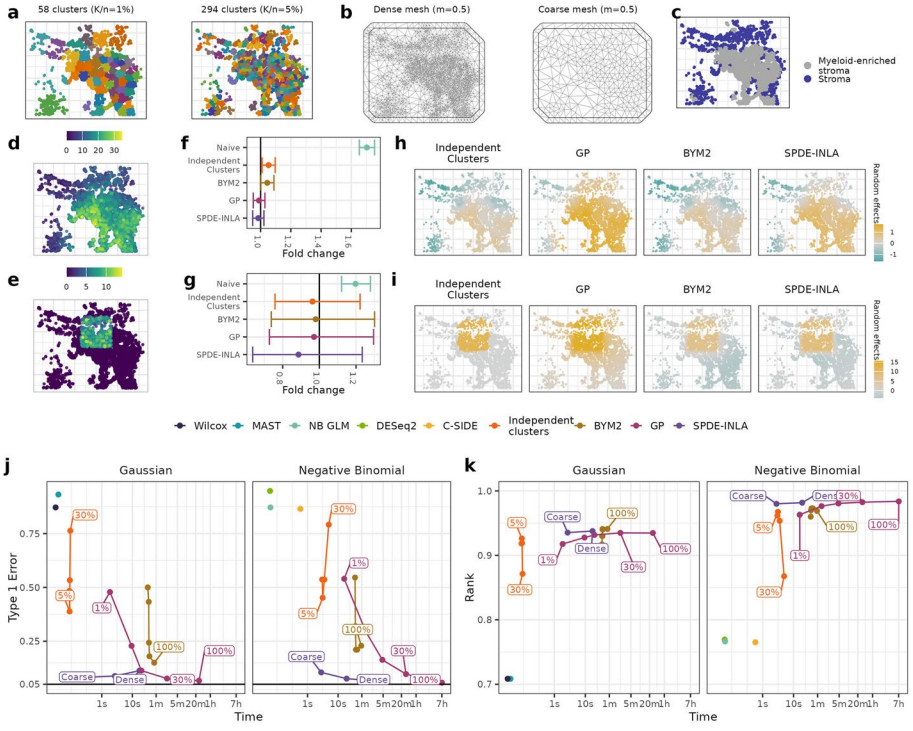

Fig. 4 Illustration of clustering, mesh construction, niche representation, and simulation results. **a** Example of clusters based on locations, with $K/n = 1\%$ and 5% showing how cluster size alters resolution. **b** Mesh representation with denser versus coarser grids. **c** Representation of Stroma and Myeloid-enriched stroma niches. **d**–**i** Two simulation scenarios for macrophage expression across niches: one with spatial variation generated by an LGCP (**d**, **f**, **h**) and one with sharp niche boundaries (**e**, **g**, **i**). Although both were generated under the null (no true DE), the naive model that ignores spatial correlation produce biased estimates and spurious DE. Figures **d**–**e** show transcript proportions per cell, **f**–**g** the estimated niche fold changes with intervals, and **h**–**i** the random effects. **j**–**k** Comparison of type I error and runtime across Gaussian and negative binomial models. Numbers in the plots indicate either the proportion of cells per cluster (*k*/*n* in %) or the mesh type (coarse vs dense) for SPDE models. Panel **j** shows that non-spatial models have inflated type I error, while spatial models achieve better calibration, with clear trade-offs between accuracy and runtime across cluster sizes and mesh resolutions. Panel **k** shows that incorporating spatial structure improves the ranking of true signals compared to non-spatial models, further emphasizing the advantage of spatial approaches

of cells, we parameterize by the cluster-to-cell ratio (K/n) rather than a fixed K, since a fixed number of clusters has different implications depending on n.

SPDE-based approximations

A further approach to scalability is to approximate GPs using stochastic partial differential equations (SPDE) [26]. The SPDE formulation discretizes the spatial domain into a triangular mesh, enabling the use of sparse precision matrices to represent correlation. The mesh resolution controls how finely spatial variation can be captured: coarser meshes improve speed at the expense of short-range fidelity, while denser meshes provide better accuracy but are more computationally demanding (Fig. 4b). Within *smiDE*, estimation is performed in a Bayesian framework using Integrated Nested Laplace Approximations (INLA) [27], which combine Laplace approximations with SPDE discretization. INLA achieves substantial computational gains over sampling-based methods, though it requires careful prior specification. We place penalized complexity (PC) priors on the correlation range and marginal variance to regularize the spatial field and avoid overfitting (see Methods section for details). These priors are specified relative to the spatial extent of the slide and typical cell-cell distances, balancing flexibility with stability.

BYM2 model

Finally, we consider the BYM2 model, a Bayesian framework originally developed for disease mapping [28]. BYM2 combines two random effects: a structured component, which smooths expression across neighboring units, and an unstructured component that captures independent variation. Instead of using pairwise distances, the structured component is defined by an adjacency matrix indicating which cells or clusters are considered neighbors. The BYM2 model assumes that the expression level of a given unit is primarily explained by its neighbors, such that once neighbor effects are accounted for, there is no additional dependence on more distant units. While this framework does not capture long-range smooth correlation in the same way as GP models, it provides a computationally efficient and interpretable alternative when tissues are partitioned into discrete domains such as clusters.

Evaluating type I error and gene ranking under spatial dependence

To benchmark methods, we conducted simulations designed to isolate the impact of spatial correlation on DE testing under controlled conditions. We anchored the simulations to a real FFPE non-small-cell lung cancer dataset by fixing the observed coordinates of 5,878 macrophages located in stroma and myeloid-enriched stroma niches (Fig. 4b) and testing for DE between these two niches.

As motivating examples, we first generated data under the null hypothesis of no niche effect. In one case, counts were simulated from a log-Gaussian Cox process (LGCP) [29], where the log-intensity includes a spatial GP random field with range 1. In another, we imposed an artificial spatial pattern by assigning higher expression values inside a square region (Fig. 4d). In both simulations (Fig. 4e-f), not accounting for spatial variation would identify the gene as differentially expressed between niches. When accounting for this correlation, regardless of the model, we achieve the correct conclusion that there is no difference between the niches. In Fig. 4g-h we can further investigate the spatial

random effects of each model. We see that the pattern obtained by the independent cluster, BYM2 and GP are similar, and all models are able to identify the patterns generated by the spatial effects.

We then conducted systematic evaluations under both Gaussian, with normalized counts, and Negative binomial likelihoods (Methods section). To provide a baseline, we included commonly used single-cell DE methods—Seurat's Wilcoxon rank sum test, Negative binomial regressions and MAST [30], DESeq2[3], and CSIDE [6] —alongside our spatial models. We assessed type I error, power, gene-ranking accuracy, and computational performance.

Naive non-spatial methods, including Seurat Wilcoxon, NB GLMs, MAST, and DESeq2, are the fastest(<1s per gene) but consistently exhibited severely inflated type I error, ranging from 0.86 (CSIDE) to 0.95 (DEseq2). By contrast, the cell-level GP achieved near-nominal calibration (type I error 0.057) but at the highest computational cost (7 hours per gene for NB, Fig. 4i). With the Gaussian likelihood, we observed a type I error of 0.067 with runtimes of ∼20 min per gene. Cluster-based GPs offered a trade-off between calibration and efficiency: at K/n=30%, type I error was near-nominal (0.077–0.097) with runtime reduced by more than 10fold. Independent-cluster models partially mitigated false positives at K/n around 5% ($\sim$ 0.38–0.45) but showed increased type I error as cluster size grew (up to 0.80). SPDE approximations using INLA provided the best balance of accuracy and scalability. Calibration depended on mesh resolution and likelihood: for NB, denser meshes improved type-I error control, whereas for Gaussian, denser meshes increased type-I error, making moderately coarse meshes preferable. Overall, SPDE models achieved near-nominal error rates ($\sim 0.05 - 0.07$) with runtimes of only 5–6 s per gene. The BYM2 model—based on adjacency rather than distance—showed elevated type I error (~ 0.25) under this design. Given that the data were generated from a continuous, distance-based GP with longer-range dependence, adjacency-based local smoothing mitigated some spatial confounding but did not fully capture the correlation structure.

Across non-null genes, models that explicitly model spatial correlation also delivered superior rankings (Fig. 4j). Spearman's correlation between estimated and true effects was ∼0.98 for NB and ∼0.93 for Gaussian fits with spatial terms, with broadly similar ranking across spatial models at comparable resolution. Even lightweight specifications—e.g., independent clusters at $K/n = 5\%$ or SPDE with a coarse mesh—achieved substantially better ranking at less time than with DESeq2 and NB-GLMs. Power analyses, restricted to models that at least partially controlled type I error (i.e., the spatial models), further highlight differences across distributions (Additional file 1: Figure S7). Negative binomial fits consistently achieved higher power than Gaussian fits, reflecting their better fit to overdispersed count data. For SPDE, power was largely robust to mesh resolution, with only minor differences between dense and coarse meshes, and slightly higher power for the coarse mesh in the Gaussian setting. For GP and BYM2, the number of clusters influenced power, with smaller clusters tending to reduce power at small effect sizes but higher power emerging at larger effects.

We also evaluated the influence of prior specification on the spatial parameters in the SPDE and BYM2 models (Additional file 1: Table S1). For the SPDE model, the true range parameter was set to 1, and we varied the prior by specifying the median range to 0.1, 1, 2, or 5. Under the negative binomial distribution, these choices had minimal

impact on type I error, whereas for the Gaussian distribution shorter priors produced higher type I error and longer priors produced lower type I error. For the variance of the spatial field, we set priors ranging from conservative, which shrink the spatial component, to weaker priors that allow more spatial variation (see Methods section). In the Gaussian setting, these choices led to only minor fluctuations in type I error (from 0.095 under the Moderate prior to 0.101 under the Permissive prior), while negative binomial models were essentially unaffected.

For the BYM2 model, we varied priors on ϕ, the parameter balancing spatially structured and unstructured effects, considering a balanced prior, one favoring i.i.d. structure, and one favoring spatial dependence. These choices had no effect on type I error. Priors on the precision parameter were varied analogously to the σ^2 priors in the SPDE model, and again resulted in only minimal differences.

Analysis of MERFISH whole brain dataset

We applied *smiDE* to the Allen Institute MERFISH whole-brain dataset (coronal section 51), which includes 98,934 cells profiled with a 500-gene panel [31, 32]. For our motivating example, we focused on excitatory projection neurons (class 01 IT-ET Glut) in the Isocortex, the largest cell population in the dataset, with 18,197 cells. We studied how these neurons modulate gene expression in response to contact with microglia ($n = 523$). Specifically, we ran DE models against a "microglia exposure" variable defined using a kernel-smoothed density of microglia cells centered at the neuron (Fig. 5a-b, Methods section).

To compare the tradeoffs of different DE analysis workflows and their impact on inference, we evaluated three analysis models built on the methodology above. (i) Naive: regression on the microglia exposure covariate without gene filtering for segmentation/contamination (no overlap-ratio filter), without neighbor-expression adjustment (Countermeasures for cell segmentation error section), and without spatial terms. (ii) Semi-naive: as in (i), but removing genes with ORM > 1 and adding the cross-type neighbor-expression covariate (Countermeasures for cell segmentation error section). (iii) Spatially aware: taking the approach of (ii) plus specifying an additional spatial random effect; we examined cell-level GP, and SPDE specifications (with multiple mesh resolutions), as well as cluster-based variants at cluster-to-cell ratios $K/n \in 5\%, 25\%, 50\%$. All models were fitted under both Negative binomial (counts) and Gaussian (normalized expression) likelihoods. For context, we also report results from *CSIDE* and *niche-DE*, both which controls the false discovery rate. Frequentist fits report both Bonferroni-adjusted p-values unless specified, whereas Bayesian fits report posterior means with Bonferroni-adjusted credible intervals at level $1 - 0.05/p$ (Methods section). All analyses were run using 5 CPU cores.

Detection of differentially expressed genes across workflows

We analyzed 426 genes with average expression above 0.1 across neurons. *CSIDE* detected 252 DE genes, comparable to *niche-DE*, which identified 269 DE genes under the negative binomial model and 241 under the Gaussian model. Both methods yielded results similar in scale to *smiDE* naive (178 with Bonferroni adjustment and 262 with FDR adjustment under the negative binomial, and 152 and 214 under the Gaussian) (Fig. 5c). However, many of these discoveries were not shared across methods, with each

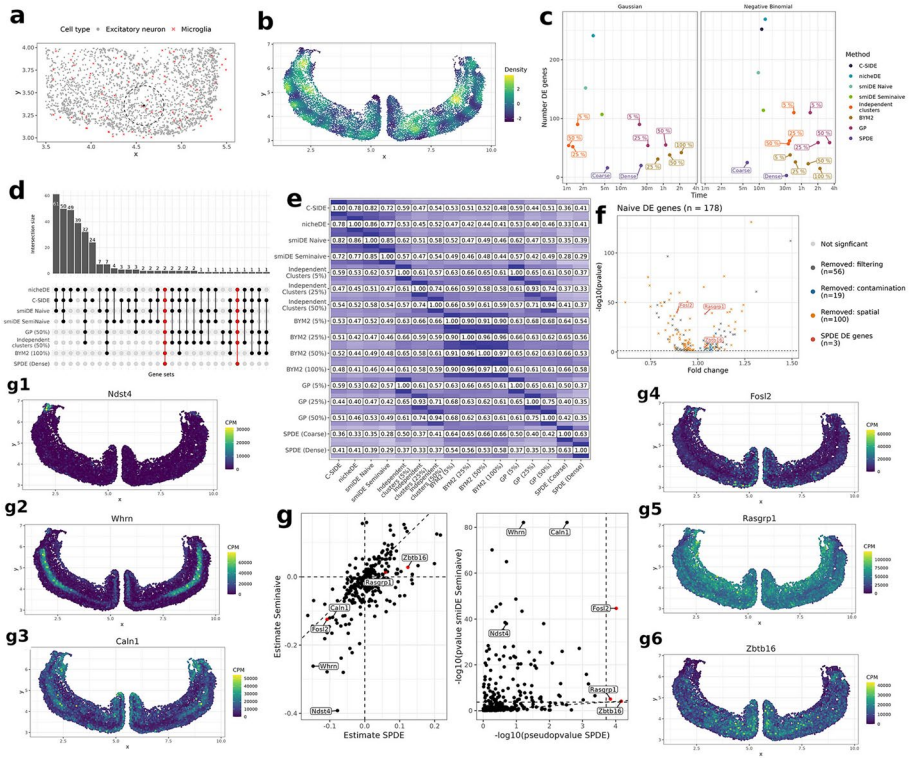

Fig. 5 Overview of DE analyses and comparisons across methods. **a** Spatial arrangement of excitatory neurons (points) and microglia (x), with circles indicating the kernel bandwidth used to compute local neighborhood density. **b** Kernel-density map of "microglia exposure," the DE covariate defined as the density of microglia around each neuron. Higher values indicate neurons located in regions with greater microglial presence. **c** Runtime (5 CPU cores) versus number of DE genes across methods for Gaussian and negative binomial distributions (labels show mesh coarseness for SPDE and cluster-to-cell ratio *K/n* for cluster-based methods; BYM2 at $K/n = 1$ with a balanced ϕ prior did not converge under Gaussian, so results are shown with the i.i.d.-favoring prior). This illustrates the trade-off between computational cost and number of discoveries: baseline methods identify many more DE genes, whereas spatially aware methods are more conservative, and clustering reduces runtime. **d** UpSet plot of DE overlaps, showing limited agreement among non-spatial methods. **e** Spearman correlations of (pseudo)-*p*-value rankings across methods and parameter settings, highlighting distinct groupings of non-spatial models, clustered models, and Bayesian spatial models. **f** smiDE naive-model DE results, with colors indicating whether each gene remains significant after contamination filtering, semi-naive adjustment, or spatial modeling. Nearly 98% of hits are eliminated after these corrections. **g** Comparison of smiDE naive versus spatially adjusted (SPDE dense mesh) statistics. Genes such as *CALN1*, *WHRN*, and *NDST4* are significant only under the naive model and do not persist after spatial adjustment. By contrast, *FOSL2*, *RASGRP1*, and *ZBTB16* remain significant, pointing to biologically meaningful associations with microglia that are robust to adjustment

approach identifying roughly 30 unique genes negative binomial (Fig. 5d). Applying the overlap ratio metric ($ORM < 1$) to reduce the impact of contamination, 258 genes remained for downstream analyses. Within this set, the semi-naive workflow uncovered 114 significant genes after FDR adjustment, including 11 additional DE genes not detected by the naive model (Fig. 5d). Due to additional internal computations for the contamination metric the computational cost increased from 9.61 to 11.8 minutes.

Independent clustering and GP gave results similar to the semi-naive workflow at a cluster-to-cell ratio of $K/n = 5\%$, suggesting that, unlike in the simulation setting, this cluster size does not capture spatial correlation. Both methods produced nearly identical results, likely because the data are very dense, so modeling correlation within clusters while ignoring correlation between them is sufficient. At $K/n = 25\%$, the number of DE genes dropped to around 50 for both Gaussian and negative binomial models, with Gaussian being faster (about 1 minute for independent clusters and $\sim$30 minutes

for GP). Results at $K/n = 50\%$ were similar, and this stability across cluster ratios demonstrates the robustness of the clustering approach: the number of DE genes and overall inference remained consistent across cluster settings, aligning with the simulation results. The BYM2 model had some convergence issues depending on the number of clusters and priors, particularly for Gaussian. When it converged, results were broadly consistent across cluster sizes, though runtime increased substantially. For Gaussian models, BYM2 failed at $K/n = 5\%$, but at $K/n = 25\%$ it detected 31 DE genes in 47 minutes and at $K/n = 100\%$ it detected 42 DE genes in 2.1 hours. For negative binomial models, BYM2 detected 38 DE genes at $K/n = 5\%$ in 37 minutes, but only 15 DE genes at $K/n = 100\%$ in 2.2 hours. This number of genes was comparable to those obtained with SPDE at a coarser mesh. Using the SPDE approximation with INLA, mesh density strongly influenced both runtime and discovery. Under the negative binomial model, a coarse mesh identified 25 DE genes in 5.9 minutes, whereas a dense mesh identified only 3 genes but required 30 minutes. Importantly, two of these genes were consistently identified by all other workflows, while one was not detected by BYM2, highlighting both the robustness of core signals and the occasional variability across spatial models.

To assess consistency across workflows, we compared rankings of *p*-values and pseudo-*p*-values across the 258 genes that passed contamination filtering, using Spearman correlation (Fig. 5e). Non-spatial methods clustered together: *CSIDE*, *niche-DE*, and *smiDE* naive/semi-naive were highly concordant ($r_s = 0.72 - 0.86$), but each showed weak agreement with SPDE (dense or coarse; $r_s \approx 0.25 - 0.41$). Independent clusters and GP produced very similar rankings to each other, and were moderately aligned with the non-spatial methods ($r_s \approx 0.44 - 0.62$ to *smiDE* naive). Importantly, both GP and BYM2 showed high internal consistency across different cluster-to-cell ratios, with BYM2 reaching $r_s \approx 0.9$, indicating that the use of clusters provides stable rankings even as resolution changes. Within the spatial models, BYM2 correlated more strongly with SPDE than did the non-spatial approaches ($r_s = 0.54 - 0.58$ to SPDE dense; $r_s = 0.64 - 0.66$ to SPDE coarse). SPDE itself was sensitive to mesh resolution (dense vs coarse, $r_s = 0.63$) and aligned more with other spatially aware approaches at lower cluster ratios. For practical ranking, lower-resolution clustering provides reasonable concordance, while finer SPDE meshes improve fidelity to spatially adjusted rankings.

Impact of contamination and spatial adjustment on biological interpretation

The cumulative effect of sequential adjustments is summarized in Fig. 5f. Starting from the naive *smiDE* model, which identified 178 DE genes, 56 (31.5%) were excluded by the contamination filter, and a further 19 (10.7%) were removed by the semi-naive adjustment—many of these had appeared up-regulated in neurons proximal to microglia. Among the filtered genes were examples predominantly associated with non-IT-ET-Glut cell classes, including oligodendrocyte markers (e.g., *MOG*, *OPALIN*, *LPAR1*, *DOCK5*), genes expressed mostly in inhibitory neurons (e.g., *CXCL14*, *VIP*, *NR2F2*, *CALB2*, *TOX3*), and genes showing low or near-zero expression across major cortical classes in the reference dataset (e.g., *GPX3*, *GPX4*, *DMRT2*, *ACTA2*, *HTR3A*) [33](Additional file 1: Figure S8).

Accounting for spatial correlation excluded 100 (56.2%) additional genes, underscoring spatial confounding as a major source of spurious associations. Among these, *NDST4* appeared strongly associated in the naive model but is expressed almost exclusively

outside the isocortex (Fig. 5g1), marking the signal as an artifact. Similarly, *WHRN* and *CALN1* showed apparent associations in the naive model but their expression is clearly layer-specific, not related to microglia proximity, and they lost significance once spatial correlation was modeled (Fig. 5g2-g3). By contrast, some genes remained robust to all adjustments (Fig. 5g). *FOSL2*, a member of the AP-1 transcription factor family, was consistently downregulated in excitatory neurons in the presence of microglia, consistent with prior evidence that AP-1/Fos proteins regulate neuron–microglia interactions and inflammatory activation [34]. In contrast, *RASGRP1*, a Ras guanine exchange factor broadly expressed in excitatory neurons, was upregulated near microglia, aligning with observations of altered neuronal *RASGRP1* expression in human brain disorders with immune involvement [35]. *ZBTB16*, a transcriptional regulator enriched in motor cortex, was also upregulated in neurons in microglia-rich regions, consistent with knockout studies showing that neuronal *Zbtb16* influences microglial density and synaptic remodeling [36]. Despite distinct regional expression patterns (*FOSL2* enriched in motor and outer layers, *RASGRP1* broadly across isocortex, *ZBTB16* enriched in motor cortex, Fig. 5g4-g6)), their associations with microglia persisted, suggesting that these these changes may represent neuron-intrinsic transcriptional adaptations to nearby immune cells.

In summary, of 426 genes analyzed, 269 were called significant by *niche-DE* naive, *CSIDE* and 178 by *smiDE* naive but only 3 of these findings remained after employing countermeasures to segmentation errors and modeling spatial correlation. In other words, less than 2% of hits from naive models were fully explainable by factors other than the design variable.

Benchmarking runtime and memory

We benchmarked runtime and memory usage across workflows (Additional file 1: Figure S9). In terms of memory, non-spatial *smiDE* naive and semi-naive were the most efficient, using only ∼22 MiB. Among existing baselines, *niche-DE* required ∼150 MiB and *CSIDE* nearly 1.5 GiB. Most spatially aware approaches were intermediate, requiring ∼ 100 MiB, with SPDE under a dense mesh reaching 113 MiB. For BYM2, however, memory scaled with the number of clusters because of the size of the adjacency matrix: while $K/n = 25\%$ required 248 MiB, the largest setting ($K/n = 100\%$) peaked near 1.5 GiB. Reducing the number of clusters substantially decreased memory requirements, making BYM2 feasible at coarser resolutions. Overall, despite their added complexity, most spatial models required less memory than *nicheDE* or *CSIDE*.

Runtime varied more widely. Some spatial models were faster than baseline methods: independent clusters with a Gaussian distribution completed in < 2 minutes, and SPDE with a coarse mesh required ∼5 minutes, both faster than naive workflows and *niche-DE* (∼13 minutes) or *CSIDE* (∼22 minutes).Other settings were slower: independent clusters with a negative binomial distribution took ∼30 minutes, GP models extended up to 1–4 hours depending on resolution, and BYM2 ranged from ∼30 minutes at $K/n = 5\%$ (negative binomial) to over 1 hour at $K/n = 50\%$. Therefore, spatially-aware models are not inherently more expensive than baseline models, many run in comparable or even shorter time than common baseline, while also requiring less memory. Moreover, clustering-based or SPDE approximations make fully spatial models computationally feasible for large datasets.

Priors specification

We evaluated the sensitivity of spatial models to prior specification on the Matérn range and variance parameters for SPDE-INLA (Additional file 1: Table S2 and Figures S10–S15) and on the ϕ and precision parameters for BYM2 (Additional file 1: Table S3 and Figures S16–S21).

For SPDE (dense mesh), negative binomial results were stable across priors, consistently detecting 2–3 DE genes. Gaussian models were more sensitive: short or weak range priors increased the number of DE genes (up to 24) and shifted p-values relative to the standard, with the largest effects observed for lowly expressed genes. Variance priors had smaller influence, though conservative priors tended to yield smaller p-values and permissive priors larger ones. Overall, SPDE was most sensitive to the range prior, specially when considering a Gaussian distribution.

For BYM2, convergence issues were more frequent, especially for Gaussian models and smaller cluster sizes. When models converged, differences across ϕ priors (balanced, i.i.d., spatial) were small, generally 1–2 DE genes, with localized shifts for genes of intermediate expression. Precision priors behaved similarly to the variance priors in SPDE: conservative settings produced smaller p-values, while flatter priors reduced significance. In Gaussian models, very conservative priors frequently failed to converge, while in negative binomial models the impact was minor and limited to lowly expressed genes. Thus, BYM2 exhibited modest prior effects but greater convergence instability than SPDE.

Analysis of FFPE colon cancer tissue

Using a publicly released CosMx 6,000-plex colon cancer dataset [37] with several tertiary lymphoid structures, we explored how B cells responded to the number of immune cells within a 15 micron radius (Fig. 6a). Among other functions in tumors, B cells form tertiary lymphoid structures (TLS), consisting of B cell dense germinal centers surrounded by T cells. TLS's are often associated with positive cancer outcomes [38]. The number of immune cells (including other B cells) in a B cell's immediate neighborhood is a continuous covariate which effectively serves as a measurement of how close that B cell is to the center of a TLS, and may provide insight into what genes are differentially regulated in these clinically relevant functional regions.

Using the three DE workflows introduced above (naive, semi-naive, and spatially aware), we analyzed B cells with the local immune-cell count within 15 μm as a continuous covariate indexing proximity to TLS centers. Results are summarized in volcano plots (Fig. 6b-d). In the spatially aware workflow, we included a spatial random effect using the GP-INLA specification (Accounting for spatial correlation section). All models were fit using Negative binomial regression.

The naive model discovered 141 significant genes based on a bonferroni p-value threshold of $p < 0.05/5914$, composed of 35 down-regulated genes (for which the expression increases when a B cell is surrounded by fewer immune cell neighbors), and 106 up-regulated genes (for which expression increases when surrounded by a larger number of immune cells). Unsurprisingly, the naive analysis highlighted a number of genes which a biologist may deem implausible, particularly among down-regulated genes not predicted to be expressed in immune cells (for example, ADCY1, COL3A1) [23, 24], whose expression would challenge the cell type identity of "B cell", and whose

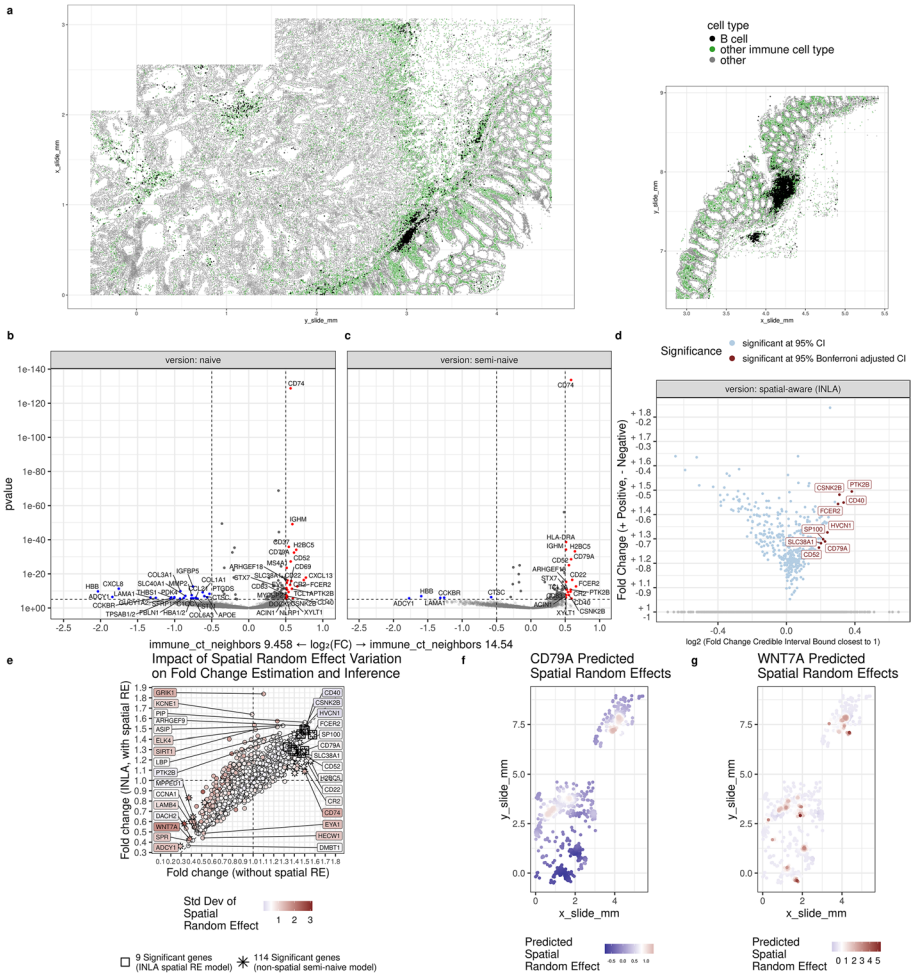

Fig. 6 **a** Colon tissue colored by cell type category; B cells (black), other immune cell types (green) include macrophage, mast, mDC, monocyte, neutrophil, NK, pDC, plasmablast, and T cell. 'Other' category includes epithelial and stromal cell types. Volcano plots for (**b**) naive model without gene filtering or covariate adjustment, (**c**) semi-naive model, with gene filtering, covariate adjustment, but without spatially correlated random effect (SRE), and (**d**) spatial-aware model results which include gene filtering, covariate adjustment, and SREs. For the bayesian GP-INLA model shown in (**d**), the x-axis coordinate represents the fold change (FC) credible interval (CI) bound closest to the null hypothesis (1), on log2-scale. FC distance is plotted on the y-axis for all genes with 95% CI's which excluded 1. Genes with α=(1-0.025/2922)% FC CI's which excluded 1 are highlighted in red. **e** FC estimates from semi-naive model without SREs (x-axis) plotted against FC estimates from spatial-aware SRE model (y-axis), colored by estimated standard deviation of SREs. Genes with higher variability in SRE (red color) are more likely to lose significance in spatial-aware model. Predicted SRE for (**f**) CD79A and (**g**) WNT7A are examples of low and high variability in SRE. WNT7A becomes not-significant in the spatial-aware model, while CD79A remains significant in presence of SRE

significance indicates likely confounding between the primary covariate (number of immune cell neighbors) and misassigned transcripts arising from segmentation error.

The semi-naive model automatically removed 3,214 genes from analysis by filtering genes with ORM > 1, and discovered 114 significant genes based on a bonferroni p-value threshold $p < 0.05/2700$, 98 of which were up-regulated DE genes. All genes which were significant in the semi-naive model were also significant in the original naive model.

Finally, the spatially aware model, which included a spatially varying random effect, discovered 9 significant genes, all of which are up-regulated in the presence of a larger number of immune cell neighbors. For the bayesian spatial-aware model

INLA, which does not compute p-values, we determined genes as signficant if their $\alpha = 1 - 0.025/2700$ level credible interval for fold change excluded the value of 1. This smaller set of genes represents a high-confidence gene set composed primarily of genes which can be defended based on existing literature. For example, FCER2 is known to initiate IgE-dependent antigen uptake and presentation to T cells [24, 39], which is plausible given the proximity of B cells to T cells in high-immune-cell TLS regions. Another DE gene, SP100, is a tumor suppressor gene and thought to play a roles in tumorigenesis, immunity, and gene regulation [24, 40]. In summary, of the 141 genes found significant by the naive analysis, 27 can be explained as resulting from errors in cell segmentation, and 105 could be explained by spatial variability unrelated to the biological question. The methods proposed here thus identified 93.6 percent of the naive method's findings as potentially false discoveries.

To give insight into the impact of the spatial random effect on inference, we plotted the fold change estimates for each gene from the semi-naive model against the fold change estimates from the spatially aware model (Fig. 6e), and colored the points by the estimated standard deviation of the spatial random effects from the spatially aware model. Genes with higher variation in spatial random effects tend to be less likely to be significant in a spatially aware model, as variability in the primary covariate can be increasingly explained by a spatially smoothed random effect. To further aid interpretation, we plotted the predicted random effects for CD79A (Fig. 6f) and WNT7A (Fig. 6g). WNT7A had high spatial variation, and was significant in the semi-naive model, but not in the spatially aware model. We can see based on the predicted random effects that there are apparent hyper-specific hot-spots of WNT7A expression that explain away the variation in expression we might otherwise attribute to our primary DE covariate. CD79A remained significant in the spatially aware model, and has comparably lower variation attributable to the spatial random effect, with the variation being visually smoother compared to the random effects for WNT7A. For this analysis of 2,922 B cells, average model estimation time per gene was 0.11 seconds for the naive model, 0.13 seconds for the semi-naive model, and 21.6 seconds for the spatial-aware model.

Discussion

Single-cell spatial transcriptomics has opened the door to new types of biological questions, such as how the expression of a gene varies with the presence of neighboring cell types, how gene programs of a specific cell type differ across tissue regions, and how expression shifts along spatial gradients such as anatomical boundaries. Answering these questions requires methods tailored to the unique challenges of spatial data, including sparse and overdispersed counts, contamination from misassigned transcripts, and spatial correlation. To address these issues, we developed a comprehensive pipeline for differential expression analysis that integrates strategies for contamination adjustment and spatial modeling, with an emphasis on computational efficiency for large datasets. All methods, including our new GP cluster approach, are implemented in the R package *smiDE*.

Our findings highlight several methodological insights. First, while the negative binomial distribution is ideally suited for overdispersed count data, Gaussian regression models can serve as viable alternatives with appropriate normalization. The Gaussian distribution provides accurate approximations and shorter runtimes, particularly when

accounting for spatial correlation, which is especially valuable in large datasets. Second, we demonstrate that addressing misassigned transcripts due to imperfect segmentation is essential for reliable DE inference. To this end, we introduce a data-driven metric to flag and remove genes likely driven by imperfect segmentation, as well as a simple proxy covariate that can be incorporated into DE regression models. Both strategies reduced false positives and improved power by filtering out genes driven by contamination. These strategies mitigate the impact of segmentation errors; fully correcting them involves a different class of specialized methods beyond the scope of this work. Third, accounting for spatial correlation is critical to control type I error, consistent with prior work [8], but our systematic evaluation reveals important tradeoffs across modeling strategies.

Across spatial modeling approaches, we observed tradeoffs between accuracy and computational efficiency. The GP provided the most accurate results but was computationally expensive (up to seven hours per gene with ∼6,000 cells). Cluster-based GP models substantially reduced runtime, with sufficient clusters achieving comparable error control and even small numbers producing reliable rankings. To achieve even faster computation, clusters can also be modeled independently; while this approach does not fully control type I error, it yields similar rankings of top genes, making it useful for exploratory analyses. BYM2 models were computationally efficient but struggled to fully capture spatial correlation and often showed convergence issues, although they may be advantageous in settings where graph-based adjacency structures are preferable to Euclidean distances, such as tissues with complex architectures. The SPDE approach offered a strong balance of speed and accuracy, completing analysis of ∼18,000 cells and 300 genes in roughly 30 minutes. This method required specification of a mesh and prior distributions, and while prior choices modestly influenced results—particularly for Gaussian models and lowly expressed genes—standard defaults were generally sufficient. Taken together, both simulations and real-data analyses demonstrate that spatially aware methods provide markedly better control of type I error than non-spatial models.

Our evaluation also shows how the proposed framework compares with existing methods. Non-spatial DE approaches, such as Seurat[30] or DESeq2[3], failed to adequately control error rates in the presence of spatial correlation. While well-suited for single-cell RNA-seq, they do not translate directly to spatial data, highlighting the need for caution when applied beyond their original scope. Among methods developed specifically for ST, C-SIDE [6] is a general DE framework but requires prior deconvolution, and niche-DE [7] incorporates spatial information through effective-niche covariates but is restricted to a single DE formulation. Neither method explicitly models spatial correlation, which explains the inflated type-1 error we observed in single-cell ST. They are more commonly applied to spot-based ST, where correlation is typically weaker, and their performance may differ in that setting. The method SpatialGE [8] is equivalent to a GP model without clustering, relying solely on Gaussian likelihoods that are computationally intensive and poorly suited for data with higher overdispersion, particularly in smaller sample sizes. SpatialGEE[41] models clusters independently using generalized estimating equations, essentially equivalent to our independent cluster specification and sharing its limitations. In contrast, our framework is more general, accommodating both Gaussian and negative binomial likelihoods, flexible priors, and multiple spatial modeling strategies, representing one of the first systematic evaluations of spatially aware DE across model classes.

These findings have direct consequences for interpreting spatial transcriptomics data and highlight their significance for both basic and translational research. Failing to adjust for misassigned transcripts can produce DE hits that simply reflect the composition of neighboring cell types, rather than cell type-specific biology, Meanwhile, ignoring spatial correlation inflates false positives, amplifies noise, and undermines reproducibility. By addressing both issues, our framework yields results that are more biologically meaningful and consistent across datasets. In basic research, this enables more reliable hypothesis generation, trustworthy pathway enrichment, and downstream analyses, as well as valid comparisons across conditions without artifacts from spatial bias. In translational contexts, these improvements enhance biomarker discovery by avoiding false positives that waste validation resources. From a practical standpoint, scalable approaches such as cluster-based GP and SPDE make it possible to analyze tens of thousands of cells in hours rather than days, enabling spatially aware inference as part of routine workflows.

A key limitation of our framework, and of spatial analyses more broadly, is spatial confounding, a well-recognized challenge in the field [42–45]. Spatial confounding occurs when the covariate of interest has strong spatial structure—such as tissue niches or brain regions—so that its effect is difficult to disentangle from underlying spatial variation. The severity of confounding depends on the degree of overlap between covariates and spatial patterns: when covariates are highly structured in space, the risk is greatest, whereas when they are more variable across tissue, the model can more easily separate their contribution from spatial effects. In these settings, adjusting for spatial effects can attenuate or mask the true covariate signal, while ignoring them inflates significance and increases false discoveries. The result is both bias in effect estimates and reduced power to detect true associations when covariates and spatial random effects capture overlapping sources of variation. If a covariate fully explains the variability in expression, no power is lost (Additional file 1: Figure S7); however, when covariates are only partially informative, their collinearity with spatial effects can compromise inference. Despite extensive discussion, the field has not reached consensus on the optimal way to address spatial confounding [46], and it remains a limitation of most spatial analyses.

Conclusions

This work advances differential expression analysis for single-cell spatial transcriptomics by explicitly modeling both contamination and spatial correlation—two major sources of bias that limit current methods. By systematically benchmarking multiple spatial models, we demonstrate how different formulations balance accuracy, computational feasibility, and interpretability, providing practical guidance for selecting methods under varying data scales and study goals.

While our framework substantially improves inference and reproducibility, important challenges remain. Spatial confounding continues to complicate interpretation when covariates overlap strongly with spatial structure, and further methodological advances are needed to mitigate its effects. Computational scalability also remains an active area of development. Emerging approaches such as NEBULA [16] offer promising strategies to accelerate mixed-model inference, and future extensions may incorporate spatial correlation directly, reducing tradeoffs between speed and accuracy.

Future work should also extend spatial modeling beyond single-sample analyses. Many biological and clinical applications require comparisons across patients or conditions,

where consistent modeling of spatial structure is essential. For now, we recommend performing differential expression separately within each sample and combining results using meta-analytic strategies such as inverse variance weighting [29], which captures both average and sample-specific effects. Developing frameworks that jointly model multiple samples while preserving spatial dependencies represents an important next step for the field.

Overall, our results highlight that appropriately accounting for spatial correlation and contamination is critical for extracting biologically meaningful signals from spatial transcriptomics data. By making spatially aware inference both accessible and computationally tractable through the *smiDE* package, this work lays the foundation for more reliable and reproducible analyses in both basic and translational research.

Methods

Choosing parametric distribution for DE regression model

Using the CosMx NSCLC dataset [1, 47] with 487,819 total cells from 5 NSCLC tissues, we compared three parametric regression models (Negative binomial RE, Poisson RE, or Gaussian RE) in terms of computation time, concordance in effect size estimation, and significant gene calls. Specifically, we ran DE regression models for 7 cell types (macrophage, T-reg, T-CD8, T-CD4, plasmablast, neutrophil, and tumor). We used the rank-normalized and inverse distance weighted total expression of neighboring cells of other cell types as described in Eq. (7) as an adjustment covariate in our regression model. A categorical covariate 'niche' annotates the functional region of tissue the cell resides in, and was treated as the primary covariate of interest in the regression models. This 'niche' covariate had 9 categories (tumor interior, storma, tumor-stroma boundary, myeloid-enriched stroma lymphoid structure, immune, plasmablast-enriched stroma, and macrophages). The `emmeans` package [48] was used to calculate 'one vs. the rest' contrasts for each niche, such that for each cell type, we assessed whether a gene was DE in a particular niche compared to the cell-weighted average of all other niches. This constitutes 9 contrasts per analyzed gene. We further subset the number of analyzed genes for each cell type using Eq. (8), removing genes with overlap ratio metric (ORM) > 1.

Countermeasures for cell segmentation error

We compared the behavior of the gene flagging metric and covariate adjustment method across DE analyses of macrophage, T-CD8, T-CD4, and T-reg cell types across 5 patients in the CosMx NSCLC dataset as described above. For each cell type, we fit regression models and tested whether genes were DE in any of the 9 'niches' compared to a cell-weighted average of other niches. For each cell type, we fit models both with and without a control variable. For models including the control variable, the covariate was defined as the rank normalized, inverse-distance weighted total expression of neighboring cells of other cell types as described in Eq. (7). To determine the effectiveness of the overlap ratio metric, we considered an unfiltered analysis of all genes, as well as removing those with ORM > 1. When assessing plausibility of significant genes based on their annotation for "Immune Cell Specificity" in Human Protein Atlas [23, 24], we considered them to be "Enhanced or Enriched" in T-reg,T-CD8,T-CD4 if the annotation contained "enhanced" or "enriched" in the description of Immune Cell Specificity, in addition to any of the phrases "T-cell", "T-reg", or "Treg". For the macrophage cell type, a

gene was considered to be "enhanced/enriched" if the description contained the phrase "enhanced", "enriched", in addition to "monocyte", "macrophage", "myeloid", or "DC". Genes were considered as "enhanced/enriched (another immune cell type)" if they contained "enhanced" or "enriched" in their description, but did not meet the cell-type specific criteria described above. Genes were labeled as "Low Immune Cell Specificity" or "Not Detected in Immune Cells" if this was the exact description for the gene provided in Human Protein Atlas.

An overlapping cells ratio metric for cell type specific gene filtering

We develop an Overlapping cells ratio metric for filtering out the most implausible genes in DE which are most likely to arise from segmentation errors.

In a DE model for a specific cell type, let *t* denote the cell type of interest and t' denote all other cell types. Intuitively, when analyzing a specific cell type, we expect that a gene may be prone to error from overlapping segmentation if it has higher expression in spatially close neighboring cells of other cell types t', than the cell type of interest, *t*.

With n_t denoting the number of cells of type *t* and $n_{t'}$ denoting the number of cells in all other cell types t', let $W(r)$ denote an $n_t \times n_{t'}$ spatial neighborhood matrix, where each element $w_{ij} = 1$ if cell *j* is within distance *r* of cell *i*, and 0 otherwise. Let $M(r)$ denote a $n_{t'} \times n_{t'}$ diagonal matrix with diagonal elements $1/m_{ii}$ indicating the number of neighbors to cell *i* within radius *r* in other cell types t', $m_{ii} = \sum_{j=1}^{n_{t'}} w_{ij}$. Further let $y(t)_g$ and $y(t')_g$ denote the n_t and $n_{t'}$ length vectors of expression for gene *g*. Then, the average expression of the gene in neighboring cells of other cell types is given by $\bar{y}(t')_g = {n_t}^{-1}[W(r)M(r)y(t')_g]$. Letting $\bar{y}(t)_g = {n_t}^{-1}y(t)_g$ denote the average expression of gene *g* for the type *t* cells of interest, the ratio

$$ORM_g = \frac{\bar{y}(t')_g}{\bar{y}(t)_g} \tag{9}$$

is a metric by which we quantify the susceptibility to confounding bias from overlapping or erroneous cell boundaries. Genes with $ORM_g > 1$ are higher expressed, on average, in neighboring cells of other types than the cell type of interest; hence the observed number of transcripts in analyzed cells may have an unacceptably high rate of false positives.

Accounting for spatial correlation

Generalized linear mixed models

Assume we have *n* cells and *p* genes, let Y_{ig} represent the expression of the *g*th gene in *i*th cell, with conditional expectation μ_{ig} modeled as

$$g(\mu_{ig}) = \beta_g^T X_i + s_g,$$

where *X* is the design matrix for fixed effects, with coefficients β, and the random effects *b* follow a Gaussian distribution with covariance matrix Σ that models the spatial correlation between cells. This correlation can be modeled in several ways described in the subsections below.

Gaussian process models

The Gaussian Process assumes an exponential covariance model

$$\Sigma_{ij} = \sigma^2 \exp(-\frac{d_{ij}}{\rho})$$

where d_{ij} is the Euclidean distance between the centroids of cells *i* and *j*th, σ^2 is the marginal variance of the spatial effect and $\rho > 0$ is the correlation range. The variance parameter σ^2 represents the marginal standard variability of the spatial random effect: when $\sigma^2 = 0$ there is no spatial component, while larger values of σ^2 indicate stronger spatial heterogeneity. The range parameters ρ controls the spatial scale of dependence. In spatial transcriptomics, nearby cells often share signals from the same environment. A small ρ means this influence is limited to immediate neighbors, while a larger ρ means the effect spreads across a broader tissue area. In the frequentist setting (β, σ^2, ρ) are estimated by maximizing the restricted marginal likelihood, as implementer in the R package `spaMM` [49].

BYM2 model

Alternatively, the BYM2 model [28], assumes a covariance structure

$$\Sigma = \sigma^2((1-\phi)I - \phi Q^{-1})^{-1}$$

where Q the precision matrix obtained based on the adjacency graph of cells, and $\phi \in [0, 1]$ is the proportion of variability explained by the spatially structured component. The adjacency matrix was defined using Delaunay triangulation, though any graph of neighbors could be considered, allowing flexible definitions of local spatial context.

This formulation is equivalent to setting

$$b = \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\phi}u + \sqrt{\phi}v\right)$$

where $\tau = 1/\sigma^2$ is the precision parameter, $u \sim N(0, I)$ represents an unstructured effect, and $v \sim \mathrm{ICAR}(Q)$ is a scaled structured effect following an intrinsic conditional autoregressive (ICAR) prior.

In practical terms, the parameter ϕ controls the balance between structured and unstructured variation. Values of ϕ close to 0 indicate that most expression variability is independent of spatial location, while values close to 1 indicate that expression is strongly shaped by the surrounding neighborhood. Thus, ϕ can be interpreted biologically as reflecting how much of the variability in gene expression is explained by local tissue context versus random cell-to-cell differences.

Parameters are estimated under a Bayesian framework using the Integrated Nested Laplace Approximation (INLA). We place Penalized Complexity (PC) priors [50] on both ϕ and the precision of the spatial effect τ. For ϕ we use as shrinkage prior $P(\phi < \phi_0) = 0.5$, with default $\phi_0 = 0.5$ with alternatives favoring unstructured $(\phi_0 = 0.3)$ or spatial $(\phi_0 = 0.7)$ variation. For the precision, we define PC priors corresponding to upper-tail probabilities $P(\tau > \tau_0) = 0.01$ with $\tau_0 = 0.3, 0.5, 0.75(\text{default}), 1.0, \text{ and } 2.0$ for Negative binomial models, which operate on the log scale of expression counts, and $\tau_0 = 0.5, 1.0, 1.5(\text{default}), 2.0, \text{ and } 3.0$ for Gaussian models of normalized expression. These specifications range from conservative to flat, and the distinction ensured that priors remained interpretable and biologically

meaningful across outcome types. Balanced and Standard priors for ϕ and τ are used, unless stated otherwise.

SPDE-based approximations

For scalability, we use the stochastic partial differential equation (SPDE) representation of a Matèrn field [26], yielding a sparse precision matrix which expedites computation. To solve SPDE numerically we discretize the tissue into a triangular mesh and random effects are defined on mesh vertices and values for cell locations are obtained by interpolation. The mesh resolution is determined by the *cutoff* parameter, which specifies the minimum distance between nodes value, preventing over-refinement where cells are very dense. Larger cutoff values produce a coarser mesh (fewer triangles, faster computation, but less ability to capture fine-scale variation); smaller values yield a denser mesh (better short-range fidelity, higher cost). We set *cutoff* as $m \times \text{extent}/\sqrt{n}$, where extent is the maximum span of the tissue in either x or y direction, *n* is the number of cells, and *m* is a coarseness factor (e.g, $m = 0.5$ dense;$m = 2$ coarse). Additionally, we add a buffer around the data domain (set to 5–10% of extent) to avoid boundary effects.

We use INLA for estimating parameters and place PC priors on the correlation range ρ and marginal standard deviation σ. For the range we set priors of the form $P(\rho < \rho_0) = 0.5$ so that ρ_0 corresponds to the prior median range, defined relative to the spatial extent of the slide or to the median nearest-neighbor distance (mnn) between cells. For the default we set $(\rho_0 = 10 \times \text{mnn})$, that is, 50% prior probability that correlation between cells decays within 10 times the median distance between two nearest neighbor cells. We also set a long-range prior $(\rho_0 = 20 \times \text{mnn})$, an alternative allowing shorter range $(\rho_0 = 5 \times \text{mnn})$ and a weakly informative choice $(\rho_0 = \text{extent}/3)$. For the variance prior, in Negative binomial models we use $P(\sigma > \sigma_0) = 0.01$ with $\sigma_0 \in \{0.5, 0.75, 1.0(\text{default}), 1.5, 3.0\}$ whereas in Gaussian we scale to gene-level dispersion by setting $\sigma_0 \in \{0.3, 0.5, 1.0(\text{default}), 1.5, 3.0\} \times \hat{sd}_g$, where $\hat{sd}_g$ is the empirical standard deviation of gene *g*. For both distributions, these specifications range from more conservative (shrinking the spatial field away from dominating the linear predictor) to less restrictive cases. Standard priors for ρ and σ are used, unless stated otherwise.

Cluster based

To reduce computational burden while retaining spatial structure, we aggregate cells into *K* spatial clusters using *k*-means on the (*x*, *y*) centroids, so that nearby cells are assigned to the same cluster. We consider cluster-level random effects and model the spatial dependence at the cluster scale assuming a $K \times K$ covariance matrix Σ_K. We consider three specifications. First an independence model with $\Sigma_K = \sigma^2 I_K$, which assumes similarities within clusters and no correlation between clusters. Second we assume a cluster-GP model with $\Sigma_K = \sigma^2 \exp\{-d_{kl}/\rho\}$, where d_{kl} is the Euclidean distance between cluster centroids *k* and *l*, providing distance-based smoothing at the cluster level. Third, a cluster-BYM2 model in which the structured component is defined on a cluster adjacency graph (constructed via Delaunay triangulation or *k*-nearest neighbors, $k = 2, 4$), inducing local smoothing across neighboring clusters; the unstructured component captures residual heterogeneity, as described above for BYM2.

The number of clusters K acts as a tuning parameter governing the resolution-efficiency trade-off. Larger K increases resolution and makes the GP and BYM2 model approximate single-cell smoothing, but also increases computational cost; in the independence case, letting K grow large approaches a non-spatial model (Additional file 1: Figure S6). Smaller K yields greater efficiency but risks oversmoothing and loss of longer-range dependence.

Inference and decision rules

For frequentist models, DE for each gene is tested with two-sided Wald statistics for the niche coefficient β, and p-values are adjusted across genes using Bonferroni correction, with significance threshold α/p, with $\alpha = 0.05$ and p the number of genes analyzed. For Bayesian models fit with INLA (SPDE and BYM2) we summarize the posterior of β by its mean, credible intervals and pseudo p-values. Pseudo-p-values are computed from the posterior probability that β is greater or less than zero, providing a continuous measure of evidence for ranking genes, while inference is based on credible intervals. The credible interval can correspond to an acceptance region for a bayesian hypothesis test under assumptions described in [51] and [52]. Specifically, we form Bonferroni-adjusted equal-tailed credible intervals and declare DE if its $1 - \alpha/p$ credible interval excludes zero. This choice provides a conservative, family-wise multiplicity adjustment that is directly comparable to frequentist Bonferroni control.

In settings with extremely large sample sizes (on the order of hundreds of thousands of cells), where even small effects can yield extremely small p-values, we emphasize interpreting results using effect sizes, ranked lists of genes, and visualizations such as volcano plots.

Simulations

To investigate the impact of spatial correlation on DE results under realistic scenarios, we simulated count data using a log-Gaussian Cox process (LGCP). For each cell i, we assume

$$Y_i \sim \text{Poisson}(\mu_i), \quad \log \mu_i = \log(N_i) + \beta_0 + \beta \cdot \text{niche}_i + s_i$$

where N_i is the observed total transcripts counts for the ith cell, β_0 is an intercept to guarantee that cell counts are on the scale similar to real data, $\text{niche} \in \{0, 1\}$ indicated the cell's niche (stroma vs. myeloid-enriched stroma), β is the DE effect, and $s = (s_1, \ldots, s_n))^T$ is a Gaussian-Process with exponential covariance matrix $\Sigma_{ij} = \sigma^2 \exp(-\frac{d_{ij}}{\rho})$. Distances d_{ij} were computed from the observed coordinates of $n = 5,878$ macrophages in the CosMx NSCLC dataset.

For each combination of parameters we generated 100 replicates of 21 genes: 6 null genes with $\beta = 0$ and 15 non-null genes with β spanning 0.1 to 1.5. Each dataset was analyzed with the spatial models in *smiDE* (GP, SPDE, BYM2, and cluster-level variants) and with non-spatial baselines (Seurat, DEseq2 and CSIDE). For Gaussian based models counts were normalized by observed total counts N_i. We report empirical type-I error using the null genes and power across non-null effect sizes. Ranking accuracy was summarized as Spearman's correlation between estimated and true β across genes and then averaged across replicates. Sensitivity analyses of prior specifications were performed

using the values described in the previous section; for the range parameter, priors were set in terms of $\rho_0 = 0.1, 1$, and 5.

Kernel-smoothed density

In the MERFISH whole-brain analysis, we used a "microglia exposure" variable as the DE covariate, defined as a Gaussian kernel–smoothed density of microglia centered at each neuron. This is equivalent to the "spatial (effective) niche" of Mason et al. [7]. For direct comparability, we computed this covariate using `nicheDE`'s `calculateEffectiveNiche()` with $\sigma = 1000$ (native coordinate units), without additional preprocessing. This DE variable is defined as

$$X_i = \sum_{j \in \mathcal{M}} K_h(\|s_i - s_j\|), \qquad K_h(r) = \exp\left(-\frac{r^2}{2h^2}\right),$$

where $\mathcal{M}$ indexes microglia, distances are Euclidean in the dataset's native units, and h is the kernel bandwidth (here $h = \sigma = 1000$). Intuitively, X_i is a distance-weighted count of nearby microglia: closer microglia contribute more, and larger h spreads the weighting over a broader neighborhood.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03867-1.

Additional File 1. Contains all supplementary figures and tables. These include further comparisons between Negative Binomial, Poisson, and Gaussian distributions; sensitivity analysis of the overlap-ratio threshold; number of DE genes under different contamination adjustments; cluster representations; power comparisons across model parameters; and sensitivity analyses of prior specifications in both simulationsand real data. Also included are an expression heatmap of genes by cell typeand benchmarking of memory usage

Acknowledgements
Not applicable.

Peer review information
George Inglis and Claudia Feng were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
A.S., P.D. and N.S conceived and supervised the study. A.G.V and D.M. analyzed the data, wrote the manuscript and implemented the methods. D.M. and A.G.V. wrote the software. All authors read and approved the final manuscript.

Funding
Not applicable.

Data availability
The smiDE package (spatial molecular imager Differential Expression) used to perform the analyses is publicly available at https://github.com/Nanostring-Biostats/CosMx-Analysis-Scratch-Space/tree/Main/_code/smiDE under a non-commercial use license from Bruker Spatial Biology, Inc. All analysis scripts and materials used to reproduce the results reported in this manuscript are freely available at https://github.com/dan11mcguire/smiDE_Analysis and have been archived in Zenodo https://doi.org/10.5281/zenodo.17179173 and https://doi.org/10.5281/zenodo.17519131.
The CosMx NSCLC FFPE dataset used for simulations is publicly available from NanoString Technologies, Inc. as part of the CosMx Spatial Molecular Imager demonstration dataset https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/nsclc-ffpe-dataset and corresponds to the dataset described in He et al. [1].
The MERFISH whole-mouse-brain dataset (Mouse whole-brain transcriptomic cell type atlas – MERSCOPE v1) is available from the Brain Image Library (Allen Institute for Brain Science, Zeng H. et al. Brain Image Library. https://doi.org/10.35077/g.610, 2023).

Declarations

Ethics approval and Consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
D.M. and P.D. are employees and shareholders of Bruker Corporation.

Received: 15 July 2024 / Accepted: 10 November 2025
Published online: 10 January 2026

References

1. He S, Bhatt R, Brown C, Brown EA, Buhr DL, Chantranuvatana K, et al. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. Nat Biotechnol. 2022;40(12):1794–806.
2. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348(6233):aaa6090.
3. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):1–21.
4. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47–e47.
5. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16(1):1–13.
6. Cable DM, Murray E, Shanmugam V, Zhang S, Zou LS, Diao M, et al. Cell type-specific inference of differential expression in spatial transcriptomics. Nat Methods. 2022;19(9):1076–87.
7. Mason K, Sathe A, Hess PR, Rong J, Wu CY, Furth E, et al. Niche-DE: niche-differential gene expression analysis in spatial transcriptomics data identifies context-dependent cell-cell interactions. Genome Biol. 2024;25(1):14.
8. Ospina OE, Soupir AC, Manjarres-Betancur R, Gonzalez-Calderon G, Yu X, Fridley BL. Differential gene expression analysis of spatial transcriptomic experiments using spatial mixed models. Sci Rep. 2024;14(1):10967.
9. Svensson V, Teichmann SA, Stegle O. Spatialde: identification of spatially variable genes. Nat Methods. 2018;15(5):343–6.
10. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods. 2020;17(2):193–200.
11. Edsgärd D, Johnsson P, Sandberg R. Identification of spatial expression trends in single-cell gene expression data. Nat Methods. 2018;15(5):339–42.
12. Li Q, Zhang M, Xie Y, Xiao G. Bayesian modeling of spatial molecular profiling data via Gaussian process. Bioinformatics. 2021;37(22):4129–36.
13. Liu S, Iorgulescu JB, Li S, Borji M, Barrera-Lopez IA, Shanmugam V, et al. Spatial maps of T cell receptors and transcriptomes reveal distinct immune niches and interactions in the adaptive immune response. Immunity. 2022;55(10):1940–52.
14. Danaher P, Hasle N, Nguyen E, Hayward K, Rosenwasser N, Alpers CE, et al. Single cell spatial transcriptomic profiling of childhood-onset lupus nephritis reveals complex interactions between kidney stroma and infiltrating immune cells. bioRxiv. 2023;2023–11. https://doi.org/10.1101/2023.11.09.566503.
15. Biermann J, Melms JC, Amin AD, Wang Y, Caprio LA, Karz A, et al. Dissecting the treatment-naive ecosystem of human melanoma brain metastasis. Cell. 2022;185(14):2591–608.
16. He L, Davila-Velderrain J, Sumida TS, Hafler DA, Kellis M, Kulminski AM. NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. Commun Biol. 2021;4(1):629.
17. Choudhary S, Satija R. Comparison and evaluation of statistical error models for scrna-seq. Genome Biol. 2022;23(1):27.
18. Petukhov V, Xu RJ, Soldatov RA, Cadinu P, Khodosevich K, Moffitt JR, et al. Cell segmentation in imaging-based spatial transcriptomics. Nat Biotechnol. 2022;40(3):345–54.
19. Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. Nat Methods. 2021;18(1):100–6.
20. Panni RZ, Linehan DC, DeNardo DG. Targeting tumor-infiltrating macrophages to combat cancer. Immunotherapy. 2013;5(10):1075–87.
21. Zhou L, Jiang Y, Liu X, Li L, Yang X, Dong C, et al. Promotion of tumor-associated macrophages infiltration by elevated neddylation pathway via NF-κB-CCL2 signaling in lung cancer. Oncogene. 2019;38(29):5792–804.
22. Christofides A, Strauss L, Yeo A, Cao C, Charest A, Boussiotis VA. The complex role of tumor-infiltrating macrophages. Nat Immunol. 2022;23(8):1148–56.
23. Karlsson M, Zhang C, Méar L, Zhong W, Digre A, Katona B, et al. A single-cell type transcriptomics map of human tissues. Sci Adv. 2021;7(31):eabh2169. https://doi.org/10.1126/sciadv.abh2169.
24. Human Protein Atlas. proteinatlas.org. Accessed Jan 2025.
25. Miao W, Geng Z, Tchetgen Tchetgen EJ. Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika. 2018;105(4):987–93.
26. Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2011;73(4):423–98.
27. Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and spatio-temporal models with R-INLA. Spat Spatio Temporal Epidemiol. 2013;7:39–55.
28. Riebler A, Sørbye SH, Simpson D, Rue H. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Stat Methods Med Res. 2016;25(4):1145–65.
29. Møller J, Syversveen AR, Waagepetersen RP. Log gaussian cox processes. Scand J Stat. 1998;25(3):451–82.
30. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.
31. Yao Z, Van Velthoven CT, Kunst M, Zhang M, McMillen D, Lee C, et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature. 2023;624(7991):317–32.
32. for Brain Science AI, Zeng H, Kunst M, Waters J, McMillen D. Mouse whole-brain transcriptomic cell type atlas – MERSCOPE v1. 2023. Dataset, Brain Image Library. https://doi.org/10.35077/g.610.

33. Bakken TE, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature. 2021;598(7879):111–9.
34. Nomaru H, Sakumi K, Katogi A, Ohnishi YN, Kajitani K, Tsuchimoto D, et al. Fosb gene products contribute to excitotoxic microglial activation by regulating the expression of complement C5a receptors in microglia. Glia. 2014;62(8):1284–98.
35. De Rosa A, Di Maio A, Torretta S, Garofalo M, Giorgelli V, Masellis R, et al. Abnormal RasGRP1 expression in the post-mortem brain and blood serum of schizophrenia patients. Biomolecules. 2022;12(2):328.
36. Usui N, Berto S, Konishi A, Kondo M, Konopka G, Matsuzaki H, et al. Zbtb16 regulates social cognitive behaviors and neocortical development. Transl Psychiatry. 2021;11(1):242.
37. Danaher P. CosMx 6000-plex colon cancer dataset. 2023. https://figshare.com/articles/dataset/CosMx_6000-plex_colon_cancer_dataset/24164388. https://doi.org/10.6084/m9.figshare.24164388.v1.
38. L ME, JL R, VC M, RA K. Tertiary lymphoid structures in cancer - considerations for patient prognosis. Cell Mol Immunol. 2020. https://doi.org/10.1038/s41423-020-0457-0.
39. Pirron U, Schlunck T, Prinz JC, Rieber EP. IgE-dependent antigen focusing by human B lymphocytes is mediated by the low-affinity receptor for IgE. Eur J Immunol. 1990. https://doi.org/10.1002/eji.1830200721.
40. Wasylyk C, Schlumberger SE, Criqui-Filipe P, Wasylyk B. Sp100 interacts with ETS-1 and stimulates its transcriptional activity. Mol Cell Biol. 2002. https://doi.org/10.1128/MCB.22.8.2687-2702.2002.
41. Wang Y, Zang C, Li Z, Guo CC, Lai D, Wei P. A comparative study of statistical methods for identifying differentially expressed genes in spatial transcriptomics. bioRxiv. 2025;2025-2. https://doi.org/10.1101/2025.02.17.638726.
42. Clayton DG, Bernardinelli L, Montomoli C. Spatial correlation in ecological analysis. Int J Epidemiol. 1993;22(6):1193–202.
43. Reich BJ, Hodges JS, Zadnik V. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics. 2006;62(4):1197–206.
44. Hodges JS, Reich BJ. Adding spatially-correlated errors can mess up the fixed effect you love. Am Stat. 2010;64(4):325–34.
45. Paciorek CJ. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. Stat Sci Rev J Inst Math Stat. 2010;25(1):107.
46. Khan K, Berrett C. Re-thinking spatial confounding in spatial linear mixed models. 2023;2023-1. arXiv preprint arXiv:2301.05743. https://doi.org/10.48550/arXiv.2301.05743.
47. NanoString Technologies, Inc . CosMx NSCLC FFPE dataset; n.d. Dataset, NanoString Technologies. https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/nsclc-ffpe-dataset. Accessed June 2023.
48. Lenth RV. emmeans: Estimated Marginal Means, aka Least-Squares Means. 2023. R package version 1.8.6. https://github.com/rvlenth/emmeans.
49. Rousset F, Ferdy JB. Testing environmental and genetic effects in the presence of spatial autocorrelation. Ecography. 2014;37(8):781–90. https://doi.org/10.1111/ecog.00566.
50. Fuglstad GA, Simpson D, Lindgren F, Rue H. Constructing priors that penalize the complexity of Gaussian random fields. J Am Stat Assoc. 2019;114(525):445–52.
51. Pereira CADB, Stern JM. Evidence and credibility: full Bayesian significance test for precise hypotheses. Entropy. 1999;1(4):99–110.
52. Thulin M. Decision-theoretic justifications for Bayesian hypothesis testing using credible sets. J Stat Plann Inference. 2014;146:133–8. https://doi.org/10.1016/j.jspi.2013.09.014.

Publisher's Note