

Children's Mercy Kansas City

SHARE @ Children's Mercy

Manuscripts, Articles, Book Chapters and Other Papers

5-14-2024

Differential gene expression analysis of spatial transcriptomic experiments using spatial mixed models.

Oscar E. Ospina

Alex C. Soupir

Roberto Manjarres-Betancur

Guillermo Gonzalez-Calderon

Xiaoqing Yu

See next page for additional authors

Let us know how access to this publication benefits you

Follow this and additional works at: <https://scholarlyexchange.childrensmercy.org/papers>

Recommended Citation

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. Published 2024 May 14. doi:10.1038/s41598-024-61758-0

This Article is brought to you for free and open access by SHARE @ Children's Mercy. It has been accepted for inclusion in Manuscripts, Articles, Book Chapters and Other Papers by an authorized administrator of SHARE @ Children's Mercy. For more information, please contact hlsteel@cmh.edu.

Creator(s)

Oscar E. Ospina, Alex C. Soupir, Roberto Manjarres-Betancur, Guillermo Gonzalez-Calderon, Xiaoqing Yu, and Brooke L. Fridley



OPEN

Differential gene expression analysis of spatial transcriptomic experiments using spatial mixed models

Oscar E. Ospina¹, Alex C. Soupir¹, Roberto Manjarres-Betancur²,
Guillermo Gonzalez-Calderon², Xiaoqing Yu¹ & Brooke L. Fridley^{1,3}✉

Spatial transcriptomics (ST) assays represent a revolution in how the architecture of tissues is studied by allowing for the exploration of cells in their spatial context. A common element in the analysis is delineating tissue domains or “niches” followed by detecting differentially expressed genes to infer the biological identity of the tissue domains or cell types. However, many studies approach differential expression analysis by using statistical approaches often applied in the analysis of non-spatial scRNA data (e.g., two-sample t-tests, Wilcoxon’s rank sum test), hence neglecting the spatial dependency observed in ST data. In this study, we show that applying linear mixed models with spatial correlation structures using spatial random effects effectively accounts for the spatial autocorrelation and reduces inflation of type-I error rate observed in non-spatial based differential expression testing. We also show that spatial linear models with an exponential correlation structure provide a better fit to the ST data as compared to non-spatial models, particularly for spatially resolved technologies that quantify expression at finer scales (i.e., single-cell resolution).

The ability to measure gene expression within a spatial context, which is referred to as spatial transcriptomics (ST), includes a wide range of technologies, including assays based on the well-established in-situ fluorescent hybridization (FISH)^{1–3}, and groundbreaking in-situ spatial barcoding^{3–8}. Current ST techniques have the capacity for extensive multiplexing (i.e., hundreds to thousands of genes assayed in the same tissue) and generating an additional data modality representing the spatial position of the measured gene expression. The spatial information from ST experiments has allowed researchers to address questions about the tissue architecture of organs and diseases^{3,9–11}. Of particular importance has been the use of ST to assess tissue heterogeneity in many cancerous tissues^{6,12–21}, as well as infected tissues²². Spatial transcriptomics has also enabled a better understanding of cell-to-cell communication^{23–25} and identifying potential druggable targets^{18,26,27}.

One common step in ST analysis is the identification of genes that differentiate tissue domains within a sample (i.e., differentially expressed genes among tissue niches)^{28–30}. Although detecting spatially variable genes without a priori definition of tissue domains (i.e., clusters) is increasingly becoming a popular choice, many studies complete the identification of differentially expressed genes in ST data within domains in an analogous fashion as it is carried out among scRNA-seq cell clusters or cell populations. In those studies, once tissue niches have been identified in the ST samples via Louvain clustering, for example, researchers often proceed with non-parametric tests such as Wilcoxon’s rank sum test^{31–33} to identify differentially expressed (DE) genes among the niches. Although this approach may be appropriate for cases where transcriptomic differences between the compared domains are substantial (e.g., tumor vs. stroma), it does not account for the spatial dependency, which results in gene expression of neighboring sampling units (e.g., cell or spots) to be more similar than distant sampling units³⁴. Because the spatial dependency in ST data is a driving factor of the gene expression patterns observed in tissues^{35,36}, more sophisticated statistical methods could be used to account for the spatial dependency between sampling units^{37–39}. Common approaches in many novel methods include identifying genes with spatial patterns, such as gene expression “hot spots”, or testing for genes showing high expression on each tissue domain (i.e., cluster) detected in a sample^{35,38–44}. Benchmarking to compare the performance of these approaches has

¹Department of Biostatistics & Bioinformatics, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA. ²Biostatistics and Bioinformatics Shared Resource, Moffitt Cancer Center, Tampa, FL, USA. ³Biostatistics and Epidemiology Core, Division of Health Services & Outcomes Research, Children’s Mercy, Kansas City, MO, USA. ✉email: blfridley@cmh.edu

also been done⁴⁵, which is crucial to aid in method selection. However, despite the wide availability of methods to detect spatially variable genes, less effort has been directed to quantify the impact of disregarding spatial dependency in ST data analysis.

Quantifying the impact of non-spatial approaches for detecting differentially expressed genes is an important endeavor, given that failure to account for the spatial autocorrelation in ST experiments may result in inflation of the type I error rate^{40,46–48}. An increased type I error rate leads to more genes erroneously being identified as differentially expressed due to inaccuracy in the *p*-values (i.e., *p*-values too small). The impact of inflated type I error rates is increased due to unreliable estimation of gene expression variation, as the variation estimates do not consider the spatial correlation among the neighboring and distant sampling units. Even in non-spatial scRNA-seq, traditional differential expression methods fail to account for type I error inflation⁴⁶, which led us to believe that considering the spatial correlation in ST experiments can alleviate this phenomenon.

Using linear mixed models offers a simple alternative for DE analysis in ST data. In bulk RNA-seq analysis, robust and well-established pipelines apply linear model fitting to test for differences in expression between two or more categories^{49,50}. However, their application to ST requires additional considerations, given the spatial nature of this modality. One such consideration, which takes advantage of the flexibility of linear mixed models, is the incorporation of spatial covariance structures and variogram analysis^{51,52}. To implement this approach as an alternative for the analysis of ST data, we performed differential gene expression analysis among groups of regions of interest (ROIs), spots, or cells in multiple ST experiments using a spatially aware implementation. The implementation tested for genes with significantly higher (or lower) expression in one group of ROIs, spots, or cells (e.g., cluster, tissue niche) to other clusters or tissue niches by fitting linear mixed models that explicitly account for the random spatial effects via spatial covariance structures. This implementation was tested on publicly available ST data sets generated with 10X Genomics' Visium platform and Nanostring's GeoMx and Spatial Molecular Imager (CosMx-SMI) platforms. We fitted corresponding non-spatial and spatial models to assess the impact of accounting for the spatial autocorrelation on the downstream DE analysis results.

Results

Comparison of non-spatial and spatial models

Models with or without spatial covariance structures were fitted for each gene to determine the most suitable alternative for capturing the expression differences among tissue domains. The tissue domain or cell type annotations for each ROI, spot, or cell were obtained from the studies that generated the data sets (Table 1; Supplementary Table S1). These studies generated the annotations using histopathology methods (Visium and GeoMx data sets) and cell phenotyping (CosMx data sets). Assessment of the models using the Akaike Information Criterion (AIC), an estimate of model fit, showed that spatial models with an exponential covariance structure provided a more accurate fit to Visium and SMI data than non-spatial models (Fig. 1). Among the four Visium samples, between 28 and 41% of the tests (i.e., gene expression in domain A vs gene expression in other domains) showed a better fit to the data when using a spatial model (i.e., lower AIC) compared to a non-spatial model. For the SMI datasets, the percentage of tests favoring the spatial models varied from 32 to 67%. In contrast, for the analysis of the GeoMx data sets, no more than 16% of the spatial models were favored over the non-spatial models (Fig. 1).

Technology	Sample	Annotations	Number of tested genes	Maximum tests	Completed tests
GeoMx	hu_brain_001	Layer I, Layer II–III, Layer IV, Layer V, White matter	5000	25,000	25,000
	hu_brain_004a			25,000	25,000
	hu_liver_001	Zone 1, Zone 2, Zone 3, Bile duct, Artery, Portal vein, Central vein, Liver macrophages, Fibroblasts ^a	5000	40,000	40,000
	hu_liver_002			45,000	45,000
Visium	151507	Layer I, Layer II, Layer III, Layer IV, Layer V, Layer VI, White matter	5000	35,000	34,968
	151673			35,000	35,000
	UKF243	Cellular, Infiltrative, Necrosis ^b , Necrotic edge ^b , Vascular hyper, None	5000	30,000	30,000
	UKF275			20,000	20,000
SMI	CancerousLiver (FOV 366)	Hep ^c , Hep 3 ^d , Hep 4 ^d , Hep 5 ^d , Hep 6 ^d , Antibody secreting B cells, CD3 Alpha Beta T cells, Stellate cells, Periportal LSECs, Cholangiocytes ^d , Inflammatory macrophages, Gamma Delta T cells 1 ^d , Non inflammatory macrophages, Mature B cells, Tumor 1 ^c , Central venous LSECs ^c	1000	10,000	10,000
	NormalLiver (FOV 174)			13,000	13,000
	Lung5Rep3 (FOV 28)	pDC, Plasmablast ^e , Neutrophil, Fibroblast, B cell ^e , T CD4 naive ^e , Epithelial, Macrophage, Mast ^e , T CD8 naive, Endothelial, Tumor 12, mDC, T CD4 memory, Monocyte, NK, Tumor 5 ^e , Tumor 6 ^e , T CD8 memory ^f , Treg ^f	960	16,320	16,320
	Lung6 (FOV 20)			14,400	14,400

Table 1. Summary of spatial transcriptomics samples used in the differential expression tests. The biological annotations present in each sample are also shown. The maximum number of tests performed corresponds to the combination of genes times the number of annotations. The completed tests column indicates the number of tests that reached convergence during the REML optimization in spaMM (using the exponential covariance structure). ^aNot present in sample hu_liver_00. ^bNot present in sample UKF275. ^cNot present in sample NormalLiver (FOV 174). ^dNot present in sample CancerousLiver (FOV 366). ^eNot present in sample Lung5Rep3 (FOV 28). ^fNot present in sample Lung6 (FOV 20).

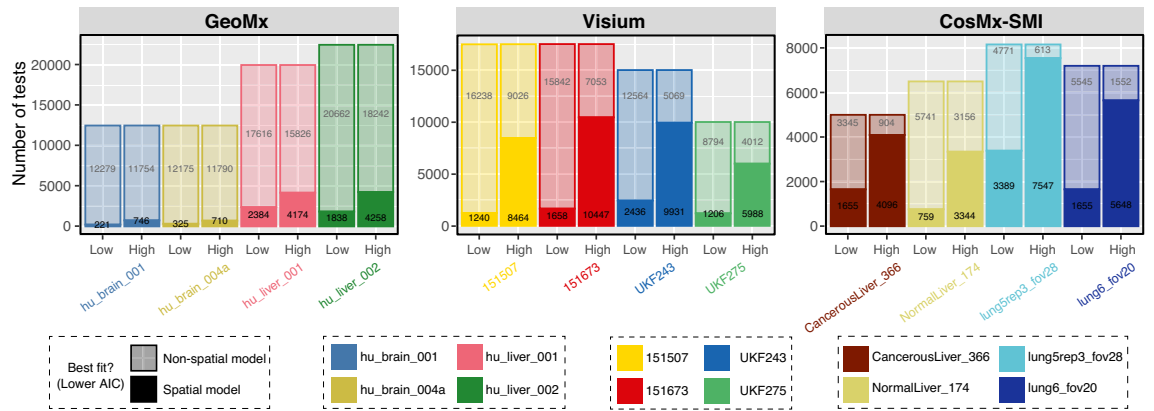


Figure 1. The results of model comparison between non-spatial models and spatial models with exponential covariance structure using AIC. For each gene x cluster test, the models with the lowest AIC were deemed to be a better fit to the data (solid color: spatial model with lower AIC, translucent color: non-spatial model with lower AIC). The tests were separated according to the average gene expression across all ROIs/spots/cells in the tissue sample (high vs low expression based on the median gene expression as threshold).

When considering only genes with high expression in the samples (above the median expression), the proportion of favored spatial models increased to 48–66% in Visium studies and 51–93% in SMI studies (Fig. 1).

Control of type I error by spatial models

The differential expression *p*-values tended to be smaller in the non-spatial models than the spatial models, possibly due to an increase in the type I error inflation. However, these patterns were dissimilar among the ST technologies (Fig. 2). In the Visium experiments, 65–71% of the *p*-values were larger in the spatial models compared to the non-spatial models. In SMI, 60–66% of the *p*-values from the spatial models were larger than those from the non-spatial models. In the GeoMx experiments, the *p*-values from the spatial models were larger in 40–54% of the tests compared to the non-spatial models. These modeling results suggest a potential slight inflation in the type I error rate for the non-spatial models, whereby *p*-values generated by non-spatial models are too small likely due to inaccurate estimation of the variance in test statistic. In other words, the variance estimation for the non-spatial models is too small, resulting in a larger test statistic and artificially smaller *p*-value.

In the tests, we grouped all the spots or cells that did not belong to the tissue niche or cell type in which differentially expressed genes were being detected. Hence, we also tested for pairwise differentially expressed

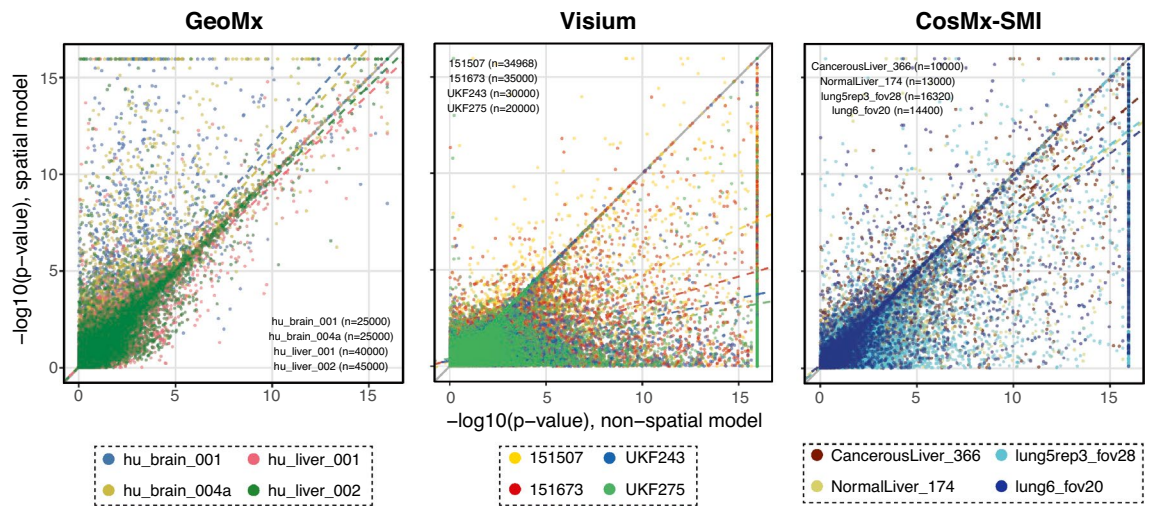


Figure 2. Comparison between non-spatial and spatial (exponential model) differential expression tests. Each point corresponds to the $-\log_{10}(p\text{-value})$ resulting from a (non-spatial or spatial) linear model fit between the expression of a gene and a binary variable indicating whether an ROI/spot/cell belongs to a biological annotation. The *p*-values indicate if the gene is differentially expressed (model coefficient different to zero) for a specific biological annotation compared to the rest of the ROIs/spots/cells. The solid line indicates a 1:1 correspondence (i.e., non-spatial and spatial models yield the same *p*-values). The colored dashed lines indicate the linear trend of the *p*-values for each sample. If a colored line lies below the solid line, *p*-values from the non-spatial model tend to be larger than those from the spatial model.

genes among three cell types in the two SMI data sets. Similar to the other tests pooling cell types, 44–64% of the p -values from the spatial models were larger than the non-spatial model p -values (Supplementary Fig. S1).

Discussion

Researchers often aim to detect differences in gene expression between cells or tissue niches, with many methods available for non-spatially informed assays, such as single-cell or “bulk” RNAseq^{49,50,53,54}. Although spatial statistics methods have existed in the literature for several decades⁵¹, only recently have spatial statistics been applied to detect spatially variable genes in biological tissues assayed with ST^{35,38–44}. In this study, we have shown that detecting differentially expressed genes in ST data benefits from statistical models that consider spatial autocorrelation. This leads to a more accurate estimate of the variance and thus produces more stable estimates of p -values. In other words, the spatial models account for the non-independence in the cells/spots, which is not addressed by traditional non-spatial linear models (i.e., two sample t -tests assuming independence between observations). Failure to consider this dependency between observations may cause the tests to underestimate the variance of the test statistic resulting in overly small p -values. Our results highlight the importance of considering the spatial dependency present in spatial-resolved transcriptomics data, which is often neglected in many studies conducting differential expression analyses. Notably, an excess of small p -values has also been noted in non-spatial scRNA-seq differential expression analysis⁴⁶.

Our results comparing the models with and without a spatial correlation structure indicated that for densely sampled ST data (e.g., Visium, SMI), spatial models present a better model fit. For non-densely sampled experiments (e.g., GeoMx using ROIs), there was a slight tendency for non-spatial models to fit the data better when compared to spatial models, probably due to less spatial correlation among ROIs that are often sampled distant from one another. Considering this finding, using non-spatial models, such as two-sample t -tests, may be appropriate to study differential gene expression in studies using GeoMx where the ROIs are more spatially distant. Nonetheless, the correlation among ROIs within a single slide and the technical variation among slides in the same study could be considered when testing for differentially expressed genes⁵⁵. Our results also indicate that for Visium and SMI, the spatial models performed better than non-spatial models in cases where the differential expression test involved a highly expressed gene. Nonetheless, the utility of spatial models in moderating the excess of small p -values might depend on the relative sample size of the groups being compared. If one of the groups is represented by a few cells, the non-spatial and spatial models produce similar p -values (Supplementary Fig. S1). In addition, genes with low expression are likely to show excessive zeroes (a characteristic of ST data^{56,57}), and hence, fitting spatial mixed models may become challenging. Novel application of Bayesian methods to detect spatially variable genes appears robust to excessive zeroes in ST data^{57,58}.

Our results were indicative that p -values obtained from the spatial model constituted a more biologically informative ranking metric for gene set enrichment analysis (GSEA). Using Benjamini-Hochberg (FDR) adjusted p -values from the non-spatial and spatial models as ranking metrics, we performed GSEA for the Hallmark gene sets with the R package *fgsea*^{59,60}. The GSEA was conducted individually for each histopathology-defined domain in the glioblastoma Visium data set⁶¹. We observed that across all the significantly enriched Hallmark gene sets, the results were more significant using the p -values from the spatial models as compared to the non-spatial models, with the exceptions of oxidative phosphorylation in the necrosis niche and KRAS signaling downregulation in the necrotic edge niche (Fig. 3). A lower score of the KRAS signaling is expected in the necrotic edge, assuming that the tumor cells in this niche are not actively proliferating⁶². Although the GSEA was conducted on a single Visium sample (UKF243), and comprehensive testing is required to evaluate the information p -values can provide for pre-ranked GSEA, our analysis suggests that p -values derived from spatial models can be more appropriate for gene set enrichment analysis when using ST data.

Testing for differential gene expression is time-consuming for modern single-cell or spatial applications, as hundreds to thousands of individual tests are performed (i.e., each combination of gene expression in domain A vs gene expression in other domains). In addition, each test often includes hundreds to thousands of cells or spots. When applying spatial models for differential expression, the advantages of accurate estimation come at the cost of longer computation times than the non-spatial models (Fig. 4). Previously, we performed these models using the long-supported R package *nlme*. However, the estimation of parameters was exceedingly time-consuming (data not shown). Hence, we switched to using the R package *spaMM* to fit the statistical models. Using a High-Performance Computing environment (HPC), differential expression of a single gene between two tissue domains can take anywhere from a few seconds to more than 2 h in Visium- or SMI-generated data. Each test was run using a single core and 8 GB of memory, resources not typically available in conventional laptop computers if run across thousands of tests simultaneously. After considering these results, we opted to implement differential gene expression analysis using *spaMM* (as opposed to *nlme*) in our R package for spatial transcriptomics analysis *spatialGE*⁶³, and we have named this approach STdiff. In the *spatialGE* R package, we made efforts to parallelize the analyses, but such efforts alone are not enough to achieve feasible computing times on personal computers and require the use of an HPC environment.

We also give a word of caution to researchers completing differential expression analysis on tissue domains or clusters defined on the same expression data, which leads to circularity and could result in overinterpretation of the function of the defined tissue domains. We propose that our approach and any other method that tests for differential expression on clusters defined with the same tested data should be only used to provide biological identity to the clusters (i.e., phenotyping). A non-circular application of these methods would require delineating tissue domains based on images by an expert pathologist, followed by differential expression analyses on the pathologist’s annotations. An example of this application is our testing on the glioblastoma Visium dataset⁶¹ included in this study.

Gene set enrichment scores from ranked p-values
Sample: UKF243

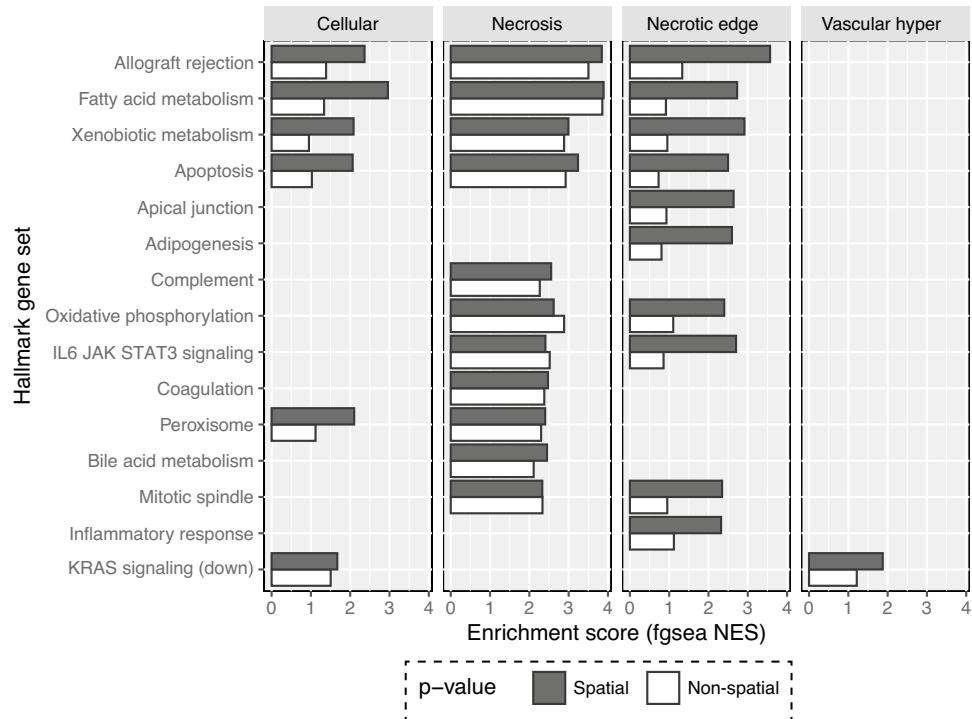


Figure 3. Scores resulting from the GSEA analysis calculated using the fgsea package. Genes were ranked using the p -values obtained by the non-spatial (white) or spatial models (dark grey). The spatial niches (cellular, necrosis, necrotic edge, vascular_hyper) were generated via histopathology examination by a previous study⁶¹. The gene sets depicted here represent Hallmark gene sets showing significant enrichment (adjusted p -value < 0.05). The “vascular_hyper” niche refers to tumor tissue with high vascularization.

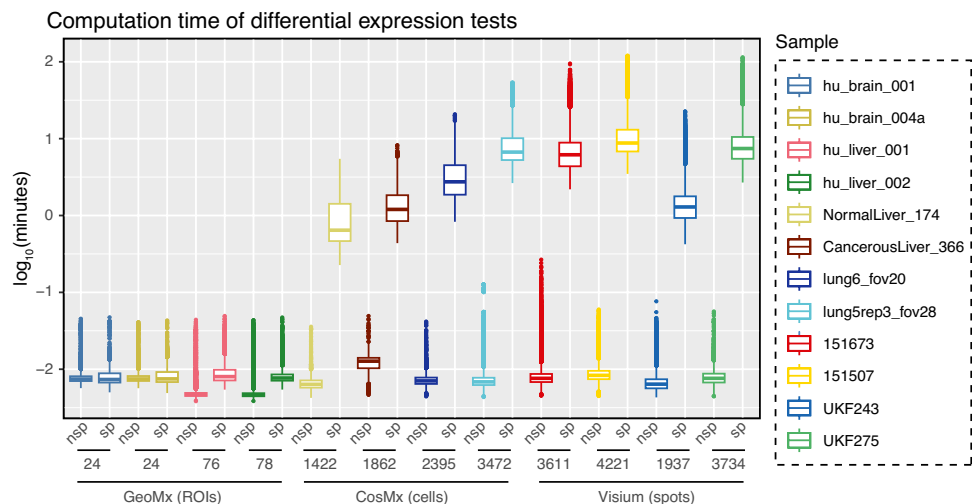


Figure 4. Time of execution in \log_{10} (minutes) for the non-spatial (nsp) and spatial (sp) test conducted. Each dot represents a single test involving a gene and a cell type or tissue domain. Execution time corresponds to the time a single CPU core took to run the “fitme” function from the spaMM R package. The number of ROIs, spots, or cells is shown on the x-axis.

In summary, considering spatial dependency is needed when conducting differential expression analysis in densely sampled spatially resolved transcriptomic experiments. In this study, we demonstrate that applying mixed models with spatial correlation structure effectively accounts for the correlation between spots or cells, thereby controlling for the inflated type I error rates observed in non-spatial models. Specifically, we show that spatial models with an exponential correlation structure provide a better fit to ST data than non-spatial models.

Material and methods

Spatial transcriptomic data sets

Spatial transcriptomics technologies are diverse, ranging in cellular and molecular resolution. Hence, we tested the utility of spatial linear mixed models for differential gene expression analysis using a series of data sets that reflected the spectrum of cellular and molecular resolution in ST technologies. We obtained publicly available ST data from spatial-barcoding technologies, including 10X Genomics' Visium and NanoString's GeoMx platforms, as well as the imaging technology produced from NanoString's CosMx Spatial Molecular Imager (SMI). The Visium data sets were generated by studies of the brain motor cortex⁶⁴ and glioblastoma⁶¹. The GeoMx and SMI data sets were obtained from NanoString's Spatial Organ Atlas repository⁶⁵. For each technology, we selected two tissue types with two samples for each tissue type (i.e., a total of 4 samples for each technology). More details of the selected samples and their access links are provided in the supplemental materials (Table 1; Supplementary Table S1). Using these data sets, we tested the utility of spatial models to detect DE genes. For this reason, a requisite for sample selection was that it contained biologically meaningful annotations (i.e., tissue domains, niches, or clusters) for each ROI/spot/cell. Preparation of expression and annotation data was carried out using the R statistical programming software version 4.1⁶⁶. Data was normalized using library size normalization and log-transformation in the package *spatialGE*⁶⁷.

Model

In differential gene expression analysis, the goal is to identify genes for which the average expression in a group is significantly higher or lower than that in other groups. In the context of ST, the sampling units (cells, spots, ROIs) are grouped using either a clustering method or prior knowledge of the tissue (e.g., tissue domains or niches). Hence, the objective remains the same: To detect genes with significantly higher or lower expression in one group of cells, spots, or ROIs (i.e., spots or cells in a domain or tissue niche) compared to ROIs/spots/cells in another tissue domain or outside of the tissue domain of interest.

For the non-spatial case of our DE analysis proposal, the expression of a given gene (y_s) at a given sample unit location (s) can be modeled as:

$$y_s = \mu_k + \varepsilon_s$$

where μ_k is the mean expression of the gene in cluster k , and ε_s is the random error at location s , with $\varepsilon_s \sim N(0, \sigma^2)$. In order to extend this model to the spatial case, we add the effect of the spatial dependency as part of the random effects (U_s) term to account for the correlation among neighboring sampling units as:

$$y_s = \mu_k + U_s + \varepsilon_s$$

where U_s is defined as $U_s \sim MVN(0, V(\theta, d))$, where d represents the distance between two ROIs/spots/cells. Several types of covariance structures can define the spatial dependency. In this study, we have tested the use of the commonly used exponential covariance structure, which is a particular case of the Matérn covariance structure, $V(\theta, d) = \tau^2 \exp\left(-\frac{d}{\rho}\right)$. Other spatial covariance structures could be used. However, the *spaMM* R package includes support for the exponential structure. Other methods for detecting spatially variable genes also use exponential or Gaussian covariance structures (e.g., *nnSVG*⁴³, *SPARK-X*⁴⁰). The use of semiovariograms⁵¹ can be exploited in future studies that assess the fit of different covariance structures to spatial transcriptomics data.

Application of models on spatial transcriptomic data sets

The application of spatial models to densely sampled tissues can be computationally intensive, particularly as the number of ROIs/spots/cells increases. Spatial transcriptomics technologies such as Visium and SMI contain thousands of spots or cells, respectively, resulting in massive covariance matrices to manipulate thousands of genes. To test for the utility of spatial models over non-spatial linear models, we randomly chose 5000 genes in each sample of the GeoMx and Visium data sets. All genes were used in testing for the SMI data sets. Next, annotations for each ROI/spot/cell were used to indicate whether the ROI/spot/cell belonged to a biological cluster or tissue domain. For each combination of gene and ROI/spot/cell annotation, we fit non-spatial and spatial models with exponential covariance structure to test for differential expression between the ROI/spot/cells assigned to that biological annotation and the rest ROI/spot/cells (Table 1). Additionally, we assessed the utility of spatial models in pairwise comparisons between two given cell types of the SMI data sets. Specifically, we tested for differentially expressed genes among tumor cells, macrophages, and T cells in the non-small cell lung cancer (NSCLC) data set and among hepatocytes, stellate cells, and non-inflammatory macrophages of the liver data set. The models were fit using the *spaMM*⁶⁸ R package on a high-performance computing (HPC) environment with one core assigned to each test and 8 GB of memory per core. The Akaike Information Criterion (AIC) was used to compare the spatial and non-spatial models. The AIC is an estimate of model fit based on the log-likelihood penalized by the complexity of the model using the formula $AIC = 2k - 2\ln(\hat{L})$, where \hat{L} is the estimated maximum likelihood of the model given the data and k is the number of parameters in the model. Given a set of models, the best-fitting model out of the group is the one with the smallest AIC. All analyses were conducted in R (version 4.1)⁶⁶, and visualizations with the *ggplot2*⁶⁹ package.

Data availability

All data sets in this study are publicly available. Please refer to Supplementary Table S1 for more information and links to access the data sets.

Code availability

The code to conduct data pre-processing and running the models in an HPC environment can be found at https://fridleylab.github.io/diff_expression_spatial_linear_models/diff_expr_spatial_linear_models.html.

Received: 1 February 2024; Accepted: 9 May 2024

Published online: 14 May 2024

References

- Lein, E., Borm, L. E. & Linnarsson, S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64–69. <https://doi.org/10.1126/science.aan6827> (2017).
- Burgess, D. J. Spatial transcriptomics coming of age. *Nat. Rev. Genet.* **20**, 317. <https://doi.org/10.1038/s41576-019-0129-z> (2019).
- Ospina, O., Soupir, A. & Fridley, B. L. in *Statistical Genomics* Vol. 2629 (eds B. L. Fridley & X. Wang) 115–140 (2023).
- Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090. <https://doi.org/10.1126/science.aaa6090> (2015).
- He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat. Biotechnol.* **40**, 1794–1806. <https://doi.org/10.1038/s41587-022-01483-z> (2022).
- Stahl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82. <https://doi.org/10.1126/science.aaf2403> (2016).
- Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319. <https://doi.org/10.1038/s41587-020-0739-1> (2021).
- Cho, C. S. *et al.* Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* **184**, 3559–3572. <https://doi.org/10.1016/j.cell.2021.05.010> (2021).
- Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546. <https://doi.org/10.1038/s41592-022-01409-2> (2022).
- Seferbekova, Z., Lomakin, A., Yates, L. R. & Gerstung, M. Spatial biology of cancer evolution. *Nat. Rev. Genet.* **24**, 295–313. <https://doi.org/10.1038/s41576-022-00553-x> (2023).
- Cheng, M. *et al.* Spatially resolved transcriptomics: A comprehensive review of their technological advances, applications, and challenges. *J. Genet. Genomics* **50**, 625–640. <https://doi.org/10.1016/j.jgg.2023.03.011> (2023).
- Theocharidis, G. *et al.* Single cell transcriptomic landscape of diabetic foot ulcers. *Nat. Commun.* **13**, 181. <https://doi.org/10.1038/s41467-021-27801-8> (2022).
- He, B. *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **4**, 827–834. <https://doi.org/10.1038/s41551-020-0578-x> (2020).
- Stur, E. *et al.* Spatially resolved transcriptomics of high-grade serous ovarian carcinoma. *iScience* **25**, 103923. <https://doi.org/10.1016/j.isci.2022.103923> (2022).
- Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9**, 2419. <https://doi.org/10.1038/s41467-018-04724-5> (2018).
- Backdahl, J. *et al.* Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin. *Cell Metab.* **33**, 1869–1882. <https://doi.org/10.1016/j.cmet.2021.07.018> (2021).
- Andersson, A. *et al.* Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat. Commun.* **12**, 6012. <https://doi.org/10.1038/s41467-021-26271-2> (2021).
- Tavares-Ferreira, D. *et al.* Spatial transcriptomics of dorsal root ganglia identifies molecular signatures of human nociceptors. *Sci. Transl. Med.* **14**, eabj8186. <https://doi.org/10.1126/scitranslmed.abj8186> (2022).
- Dhainaut, M. *et al.* Spatial CRISPR genomics identifies regulators of the tumor microenvironment. *Cell* **185**, 1223–1239. <https://doi.org/10.1016/j.cell.2022.02.015> (2022).
- Garcia-Alonso, L. *et al.* Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. *Nat. Genet.* **53**, 1698–1711. <https://doi.org/10.1038/s41588-021-00972-2> (2021).
- Chen, H. *et al.* Dissecting mammalian spermatogenesis using spatial transcriptomics. *Cell Rep.* **37**, 109915. <https://doi.org/10.1016/j.celrep.2021.109915> (2021).
- Delorey, T. M. *et al.* COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature* **595**, 107–113. <https://doi.org/10.1038/s41586-021-03570-8> (2021).
- Rao, A., Barkley, D., Franca, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220. <https://doi.org/10.1038/s41586-021-03634-9> (2021).
- Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* **22**, 627–644. <https://doi.org/10.1038/s41576-021-00370-8> (2021).
- Marshall, J. L. *et al.* High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways. *iScience* <https://doi.org/10.1016/j.isci.2022.104097> (2022).
- Joshi, N. *et al.* A spatially restricted fibrotic niche in pulmonary fibrosis is sustained by M-CSF/M-CSFR signalling in monocyte-derived alveolar macrophages. *Eur. Respir. J.* <https://doi.org/10.1183/13993003.00646-2019> (2020).
- Su, S. & Li, X. Dive into single, seek out multiple: Probing cancer metastases via single-cell sequencing and imaging techniques. *Cancers* <https://doi.org/10.3390/cancers13051067> (2021).
- Fang, S. *et al.* Computational approaches and challenges in spatial transcriptomics. *Genomics Proteomics Bioinf.* <https://doi.org/10.1016/j.gpb.2022.10.001> (2022).
- Ren, Y. *et al.* Spatial transcriptomics reveals niche-specific enrichment and vulnerabilities of radial glial stem-like cells in malignant gliomas. *Nat. Commun.* **14**, 1028. <https://doi.org/10.1038/s41467-023-36707-6> (2023).
- Lyubetskaya, A. *et al.* Assessment of spatial transcriptomics for oncology discovery. *Cell Rep. Methods* **2**, 100340. <https://doi.org/10.1016/j.crmeth.2022.100340> (2022).
- Zhu, J. *et al.* Delineating the dynamic evolution from preneoplasia to invasive lung adenocarcinoma by integrating single-cell RNA sequencing and spatial transcriptomics. *Exp. Mol. Med.* **54**, 2060–2076. <https://doi.org/10.1038/s12276-022-00896-9> (2022).
- Buzzi, R. M. *et al.* Spatial transcriptome analysis defines heme as a hemopexin-targetable inflammatoin in the brain. *Free Radic. Biol. Med.* **179**, 277–287. <https://doi.org/10.1016/j.freeradbiomed.2021.11.011> (2022).
- Luo, W. *et al.* Single-cell spatial transcriptomic analysis reveals common and divergent features of developing postnatal granule cerebellar cells and medulloblastoma. *BMC Biol.* **19**, 135. <https://doi.org/10.1186/s12915-021-01071-8> (2021).
- Qiu, Z. *et al.* Detection of differentially expressed genes in spatial transcriptomics data by spatial analysis of spatial transcriptomics: A novel method based on spatial statistics. *Front. Neurosci.* **16**, 1086168. <https://doi.org/10.3389/fnins.2022.1086168> (2022).

35. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: Identification of spatially variable genes. *Nat. Methods* **15**, 343–346. <https://doi.org/10.1038/nmeth.4636> (2018).
36. Fornito, A., Arnatkeviciute, A. & Fulcher, B. D. Bridging the gap between connectome and transcriptome. *Trends Cogn. Sci.* **23**, 34–50. <https://doi.org/10.1016/j.tics.2018.10.005> (2019).
37. Su, J. *et al.* Smoother: A unified and modular framework for incorporating structural dependency in spatial omics data. *Genome Biol.* **24**, 291. <https://doi.org/10.1186/s13059-023-03138-x> (2023).
38. Edsgard, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* **15**, 339–342. <https://doi.org/10.1038/nmeth.4634> (2018).
39. Hu, J. *et al.* SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351. <https://doi.org/10.1038/s41592-021-01255-8> (2021).
40. Zhu, J., Sun, S. & Zhou, X. SPARK-X: Non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biol.* **22**, 184. <https://doi.org/10.1186/s13059-021-02404-0> (2021).
41. Miller, B. F., Bambah-Mukku, D., Dulac, C., Zhuang, X. & Fan, J. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. *Genome Res.* **31**, 1843–1855. <https://doi.org/10.1101/gr.271288.120> (2021).
42. Dries, R. *et al.* Giotto: A toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78. <https://doi.org/10.1186/s13059-021-02286-2> (2021).
43. Weber, L. M., Saha, A., Datta, A., Hansen, K. D. & Hicks, S. C. nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes. *Nat. Commun.* **14**, 4059. <https://doi.org/10.1038/s41467-023-39748-z> (2023).
44. Deshpande, A. *et al.* Uncovering the spatial landscape of molecular interactions within the tumor microenvironment through latent spaces. *Cell Syst.* **14**, 285–301. <https://doi.org/10.1016/j.cels.2023.03.004> (2023).
45. Chen, C., Kim, H. J. & Yang, P. Evaluating spatially variable gene detection methods for spatial transcriptomics data. *Genome Biol.* **25**, 18. <https://doi.org/10.1186/s13059-023-03145-y> (2024).
46. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692. <https://doi.org/10.1038/s41467-021-25960-2> (2021).
47. Park, Y. P. & Kellis, M. CoCoA-diff: Counterfactual inference for single-cell gene expression analysis. *Genome Biol.* **22**, 228. <https://doi.org/10.1186/s13059-021-02438-4> (2021).
48. Sonesson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261. <https://doi.org/10.1038/nmeth.4612> (2018).
49. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140. <https://doi.org/10.1093/bioinformatics/btp616> (2010).
50. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. <https://doi.org/10.1186/s13059-014-0550-8> (2014).
51. Cressie, N. A. C. *Statistics for spatial data* Vol. 900 (Wiley & Sons, 1993).
52. Pinheiro, J. C. & Bates, D. M. *Mixed-effects models in S and S-PLUS* (Springer, 2000).
53. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296. <https://doi.org/10.1186/s13059-019-1874-1> (2019).
54. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122. <https://doi.org/10.12688/f1000research.9501.2> (2016).
55. Bergholtz, H. *et al.* Best practices for spatial profiling for breast cancer research with the GeoMx digital spatial profiler. *Cancers* <https://doi.org/10.3390/cancers13174456> (2021).
56. Zhao, P., Zhu, J., Ma, Y. & Zhou, X. Modeling zero inflation is not necessary for spatial transcriptomics. *Genome Biol.* **23**, 118. <https://doi.org/10.1186/s13059-022-02684-0> (2022).
57. Jiang, X., Xiao, G. & Li, Q. A Bayesian modified Ising model for identifying spatially variable genes from spatial transcriptomics data. *Stat. Med.* **41**, 4647–4665. <https://doi.org/10.1002/sim.9530> (2022).
58. Li, Q., Zhang, M., Xie, Y. & Xiao, G. Bayesian modeling of spatial molecular profiling data via Gaussian process. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btab455> (2021).
59. Fast gene set enrichment analysis v. 1.26 (Bioconductor, 2019).
60. Liberzon, A. *et al.* The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425. <https://doi.org/10.1016/j.cels.2015.12.004> (2015).
61. Ravi, V. M. *et al.* Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. *Cancer Cell* **40**, 639–655. <https://doi.org/10.1016/j.ccell.2022.05.009> (2022).
62. Zhu, G., Pei, L., Xia, H., Tang, Q. & Bi, F. Role of oncogenic KRAS in the prognosis, diagnosis and treatment of colorectal cancer. *Mol. Cancer* **20**, 143. <https://doi.org/10.1186/s12943-021-01441-4> (2021).
63. spatialGE: An R package for visualization and analysis of spatially-resolved gene expression v. 1.2 (GitHub, 2023).
64. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436. <https://doi.org/10.1038/s41593-020-00787-0> (2021).
65. Nanostring. *Spatial organ atlas*, <<https://nanostring.com/products/geomx-digital-spatial-profiler/spatial-organ-atlas/>> (2022).
66. R: A language and environment for statistical computing v. v4.1.2 (R Foundation for Statistical Computing, Vienna, Austria, 2021).
67. Ospina, O. E. *et al.* spatialGE: Quantification and visualization of the tumor microenvironment heterogeneity using spatial transcriptomics. *Bioinformatics* **38**, 2645–2647. <https://doi.org/10.1093/bioinformatics/btac145> (2022).
68. Rousset, F. & Ferdy, J.-B. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography* **37**, 781–790. <https://doi.org/10.1111/ecog.00566> (2014).
69. Wickham, H. *ggplot2: Elegant graphics for data analysis* (Springer, 2016).

Acknowledgements

This work has been supported in part by the National Institutes of Health (NIH) (U01 CA274489) and by the Biostatistics and Bioinformatics Shared Resource at the H. Lee Moffitt Cancer Center & Research Institute, an NCI-designated Comprehensive Cancer Center (P30 CA076292).

Author contributions

O.E.O., A.C.S., R.M., G.G., X.Y., and B.L.F. conceptualized and reviewed the study. B.L.F. and O.E.O. wrote the manuscript with contributions from all authors. O.E.O. collected the data sets and developed the code. O.E.O. and A.C.S. performed the linear model tests.

Funding

Funding was provided by National Institutes of Health (Grant number: U01 CA274489, U01 CA274489).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-61758-0>.

Correspondence and requests for materials should be addressed to B.L.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024