Title: Quantification and visualization of the tumor microenvironment heterogeneity from spatial transcriptomic experiments

Oscar E. Ospina¹, Christopher M. Wilson¹, Alex C. Soupir¹, Anders Berglund¹, Inna Smalley², Kenneth Y. Tsai^{2,3}, and Brooke L. Fridley^{1*}

¹ Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, USA

² Department of Tumor Biology, Moffitt Cancer Center, Tampa, FL, USA

² Department of Anatomic Pathology, Moffitt Cancer Center, Tampa, FL, USA

Running title: Tumor heterogeneity using spatial transcriptomics

Keywords: Spatial array, tumor microenvironment, spatial heterogeneity, RNA-Seq, spatial kriging, unsupervised clustering

Financial support: This work was supported by the NIH – H. Lee Moffitt Cancer Center T32 Integrated Program in Cancer and Data Science (T32CA233399), and the Biostatistics and Bioinformatics Shared Resource at the H. Lee Moffitt Cancer Center and Research Institute, an NCI designated Comprehensive Cancer Center (P30-CA076292). IS is supported by the NIH R00 CA226679. This content is solely the responsibility of the authors and does not necessarily represent the official view of the NIH.

*Correspondence:

Brooke L. Fridley, Ph.D.

12902 Magnolia Drive

Moffitt Cancer Center

Tampa, FL 33612

Phone: 813-745-1461

Fax: 813-745-6107

Email: brooke.fridley@moffitt.org

Conflicts of interests: None of the authors have conflict-of-interest or financial disclosures to declare.

Word count: 1692 (from Introduction to Discussion)

Tables and figures: five figures (two in main text, four in supplementary information); one supplementary table.

Abstract

Spatially-resolved transcriptomic technologies promise to increase our understanding of the tumor microenvironment, which will lead to better cancer prognosis and therapies. Several spatial transcriptomics technologies have been developed, as well as a few methods for data analysis. Nonetheless, analytical pipelines that take users from gene counts to association of tumor spatial heterogeneity with clinical data are not available. Here we present spatialGE, a software that provides visualizations and quantification of the tumor microenvironment heterogeneity using spatial transcriptomics. Our software includes: 1) generation of high-resolution gene expression surfaces via spatial interpolation; 2) quantification of spatial heterogeneity measures that can be compared against clinical information (e.g., patient survival); 3) cell deconvolution methods at the spot level; 4) spatially-informed clustering; and 5) a new data structure that allows storage and analysis of multiple ST samples simultaneously. We show the utility of spatialGE by studying the spatial heterogeneity of cutaneous malignant melanoma samples.

Significance: spatialGE enables the analysis of spatial transcriptomics samples and exploration of the tumor microenvironment, as well as the association of clinical outcomes with the spatial variation in gene expression.

Introduction

Spatially-resolved transcriptomic (ST) methods have allowed a better understanding of the tumor microenvironment (TME) at an unprecedented resolution (1). Characterizing heterogeneity within the TME is key to cancer prognosis and development of therapies (2,3). Advances in ST promise to unravel tumor heterogeneity at different scales, from gene expression differences among samples to clonal diversity within tissues (4). Importantly, the study of the TME is crucial to comprehend immune infiltration and the relationship with immunotherapy response (5).

Several ST technologies to study the TME have been developed (6-12). The most common approach involves mRNA probe hybridization in immobilized tissues on surfaces (7). These technologies can produce data for either a subset of genes (e.g., panel) or the whole transcriptome (7). Spatial transcriptomics preserves the localization of gene expression within the tissue architecture (13), as opposed to bulk RNA (RNA-seq) or in single-cell RNA sequencing (scRNAseq) where tissue homogenization or dissociation is necessary (14). Therefore, besides the high dimensionality of RNA-seq, the analysis needs to account for the spatial location of transcripts. Consequently, data structures are needed to accommodate both expression and spatial data, as well as related information (e.g., clinical information, demographics). Similarly, methods for the visualization and analysis of tissue spatial heterogeneity are needed, with existing tools varying in flexibility and approaches (15-20). Hence, we developed spatialGE, a pipeline for the processing and analysis of ST experiments. The pipeline borrows ideas from geostatistical techniques and applies them to the exploration of the TME. The novel features of spatialGE include: 1) generation of transcriptomic surfaces; 2) quantification of gene expression spatial heterogeneity and association with clinical information;

3) immune gene expression deconvolution methods; 4) spatially-informed clustering; and 5) a data structure to store gene counts, spatial, and sample metadata for multiple ST tissue sections.

Materials and Methods

Spatial transcriptomics data input and transformation

In spatialGE, for each tissue section the user provides count data, with gene names following the HUGO nomenclature. Spot coordinates should also be provided per tissue section. Optionally, an additional, single table can be provided containing associated clinical variables (Fig. 1A). The example data set presented with the spatialGE software is the Thrane et al. study of cutaneous malignant melanoma (21). The spatialGE software processes raw counts with library size normalization (22) followed by voom transformation (23) to produce roughly Gaussian distributed gene expression values per gene (Supplementary Figure 1). Relative gene expression at each spot can then be visualized using “quilt plots” (Fig. 1B).

Spatial transcriptomic surfaces estimation via interpolation (“kriging”)

Visualization of spatial patterns in gene expression within the TME can be improved with the generation of transcriptomic surfaces. To produce smooth transcriptomic surfaces, spatialGE uses spatial interpolation (24) with either ordinary and universal kriging used to predict surfaces based on gene expression values or deconvoluted cell type scores (Fig. 1B-C). Ordinary kriging assumes a constant mean value across the sampled area (stationarity). Conversely, universal kriging assumes the mean gene expression varies across the tissue (24).

Spatial transcriptomics heterogeneity statistics

In addition to ST surfaces, spatialGE provides users with approaches to quantify spatial heterogeneity across and detection of “hot-spots”. A variety of measures are included to provide a summary of TME heterogeneity (Supplementary Table XXSP_STATS; Fig. 1B): Moran’s I (25) and Geary’s C (26) allow users to ascertain whether gene expression is uniform or concentrated in specific tissue regions. The Getis-Ord G_i^* statistic measures the tendency of a given gene to produce areas of concentrated high expression (hot-spots) or low expression (cold-spots) (27). The use of spatial statistics allows to quantify associations with other measurable variables from clinical outcomes. The statistics were implemented using the R package spdep (28).

Spatial transcriptomics gene expression deconvolution

For cell type deconvolution, xCell (29) has been implemented to infer cell compositions at each spot. Kriging can then be applied to these scores to obtain a spatial surface for a given cell type (Fig.1C). Additionally, tumor purity scores are generated using ESTIMATE (30), followed by classification of spots as primarily tumor or stroma using Gaussian model-based clustering, as implemented in the R package mclust (31).

Unsupervised spatially-informed clustering to infer TME expression subtypes

To understand the location and types of different TME compartments, unsupervised spatially-informed clustering can be performed (Fig. 1D). For clustering, principal component analysis (PCA) of gene counts is completed. Then, a spatially-informed distance matrix for clustering is computed based on two scaled distance matrices: 1) transcriptomic autocorrelation between spots using the principal components (D_1), and 2) spatial distances between spots (D_2).

Next, the autocorrelation matrix is “shrunk” toward the physical distance matrix by calculating a weighted average distance matrix:

$$D = [(1 - w) * D_1] + (w * D_2)$$

The user specifies the weight (w) to apply. Based on our experience a weight between 0.05 and 0.25 seems to best capture tissue heterogeneity, without too much “shrinkage”. Using the weighted matrix, D , hierarchical clustering with dynamic cut-points (32) can be completed to derive the TME compartments, without the need to specify the number of compartments.

Availability

spatialGE is available as an R package at <https://github.com/FridleyLab/spatialGE>. A tutorial is also provided showing users how to use the package and its various features. A script to reproduce the analyses in this manuscript is available in Supplementary Information.

Results

The utility of spatialGE is demonstrated using the Thrane et al. ST tissue sections, with equidistant spots of 100μM diameter and 200μM apart (21). The eight tissue sections (representing four patients) had an average of 15,864 genes, and average number of spots of 293 (range = 212–383). Both quilt plots and transcriptomic surfaces of genes relevant to cutaneous malignant melanoma captured the heterogeneity of the tumor and stromal compartments (Fig. 2A, Supplementary Fig. 2), consistent with pathology images in the Thrane et al. study (21). With spatialGE, we could also reproduce their findings on the spatial patterns of gene expression. For example, in tissue section 2 (patient 1), the genes *PMEL* and *SPPI* were expressed throughout the tissue, except in a region classified as lymphoid tissue (Supplementary

Fig. 2A-B). Conversely, *CD74* and *IGLL5* exhibited the opposite pattern, with low expression in the melanoma region (Fig. 2A, Supplementary Fig. 2C-D). The expression of *SOX10* was high across the tissue section, whereas *SOX4* expression was low in spots classified as tumor by the ESTIMATE scores (Supplementary Fig. 2E-F). Future research could study the potential use of *SOX4* to identify melanoma regions using data from ST experiments.

The spatial autocorrelation statistics (Moran's I, Geary's C, and Getis-Ord G_i^*) captured gene expression heterogeneity across the tissue and showed moderately strong relationships with standard deviation ($0.49 < R^2 < 0.64$; Fig. 2B). A higher degree of spatial heterogeneity was found for *CD74* compared to *SOX10*, as indicated by Moran's I ($I=0.20$ and $I=0.02$, respectively). Nonetheless, *SOX10* showed higher Geary's C (Supplementary Fig. 2D-E), reflecting Geary's C higher sensitivity to smaller agglomerations of high expression. The Getis-Ord G_i^* statistic, a measure of "hot-spots", indicated that expression of *PMEL* and *SPPI* had a strong tendency to generate high expression regions (Supplementary Fig. 2A-B). Interestingly, Moran's I and Geary's C calculated for *CD74* showed an association between spatial heterogeneity in the biopsies and patient survival, with >120 months survival observed when *CD74* exhibited high spatial association (Fig. 2C).

After immune cell deconvolution for each spot, kriging of B cell scores showed a transition from high scores in stroma to low scores nearing the melanoma compartment. This finding was congruent with ESTIMATE binary tumor/stroma classifications (patient 1; Fig. 2A). Spatially-informed clustering with $w = 0.075$ seemed to capture the tissue heterogeneity of the tumor when compared to the pathology images in Thrane et al. (21) (Supplementary Fig. 3 and 4). The first and second clusters corresponded to melanoma and stromal areas. The third cluster

surrounded the melanoma area and included spots in the region of high B cell scores (Fig. 2D; Supplementary Fig. 3).

Discussion

Our pipeline, spatialGE, provides visualizations of spatially-resolved gene expression, as well as quantification of TME heterogeneity and exploration of its potential associations with clinical information. Gene expression visualizations at high resolution are offered by few tools (20), however spatialGE is unique in implementing kriging to achieve this task for multiple ST tissue sections simultaneously. Unlike other approaches, kriging gives users flexibility to choose visualization resolution with few modeling assumptions. Additionally, kriging is used to visualize tissue composition as inferred by deconvolution methods. Our pipeline also features unsupervised spatially-informed clustering, a simple yet informative method to detect TME compartments. Several ST clustering approaches are already available (16-18,20), however our method resembles tissue features by weighting transcriptomic similarities between spots by their distances to one another. Another innovation of spatialGE is the estimation of quantifiable metrics (spatial heterogeneity statistics) to capture transcriptomic complexity and the ability to compare them against patient clinical outcomes. A comparison between spatialGE and other software is available in the supplementary information (Supplementary Table 2).

Applying spatialGE to the melanoma ST data, we observed agreement between the transcriptomic surfaces and the tumor and stroma regions observed in pathology images (21). Our spatially-informed clustering approach provided TME compartments (i.e., clusters) that resembled the features annotated in pathology images (21), with an immune tumor border effect observed, in which the melanoma compartments were surrounded by regions with high B cell

scores. Finally, we observed an association between spatial heterogeneity and clinical variables providing support to the hypothesis that spatial heterogeneity is a predictor of patient outcomes (3). For example, higher levels of *CD74* (indicated by Moran's I and Geary's C) corresponded to longer survival.

In future spatialGE versions, we will include features to analyze a wider ST technology spectrum. We are currently developing gene expression visualizations and histological annotation of spots for technologies using non-gridded arrangements, as well as a web-based implementation of spatialGE. Another developing area is the integration of scRNA-seq data to provide ST cell-level insights. Finally, we are expanding our spatially-informed clustering method to identify genes that differ among clusters.

In summary, spatialGE provides a suite of methods for the exploration of the TME via ST. By adapting geostatistical methods (e.g., kriging and autocorrelation statistics) and a novel spatially-informed clustering, we provided an innovative approach to the study processes such as immune infiltration and the effect of TME architecture on patient outcomes.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank the spatial research group at the KTH Royal Institute of Technology – SciLifeLab (Stockholm, Sweden) for making their melanoma data available.

References

1. Moor AE, Itzkovitz S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr Opin Biotechnol* **2017**;46:126-33
2. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* **2012**;12:323-34
3. Nederlof I, Horlings HM, Curtis C, Kok M. A High-Dimensional Window into the Micro-Environment of Triple Negative Breast Cancer. *Cancers (Basel)* **2021**;13
4. Saviano A, Henderson NC, Baumert TF. Single-cell genomics and spatial transcriptomics: Discovery of novel cell states and cellular interactions in liver physiology and disease biology. *J Hepatol* **2020**;73:1219-30
5. Tang H, Qiao J, Fu YX. Immunotherapy and tumor microenvironment. *Cancer Lett* **2016**;370:85-90
6. Stahl PL, Salmen F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **2016**;353:78-82
7. Zhang M, Sheffield T, Zhan X, Li Q, Yang DM, Wang Y, *et al.* Spatial molecular profiling: platforms, applications and analysis tools. *Brief Bioinform* **2021**;22
8. Xia C, Fan J, Emanuel G, Hao J, Zhuang X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc Natl Acad Sci U S A* **2019**;116:19490-9
9. Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, *et al.* Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat Biotechnol* **2020**;38:586-99

10. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **2018**;361
11. Vickovic S, Eraslan G, Salmen F, Klughammer J, Stenbeck L, Schapiro D, *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* **2019**;16:987-90
12. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **2019**;363:1463-7
13. Maniatis S, Petrescu J, Phatnani H. Spatially resolved transcriptomics and its applications in cancer. *Curr Opin Genet Dev* **2021**;66:70-7
14. Yu X, Abbas-Aghababazadeh F, Chen YA, Fridley BL. Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments. *Methods Mol Biol* **2021**;2194:143-75
15. Navarro JF, Sjostrand J, Salmen F, Lundeberg J, Stahl PL. ST Pipeline: an automated pipeline for spatial mapping of unique transcripts. *Bioinformatics* **2017**;33:2591-3
16. Bergenstrahle J, Larsson L, Lundeberg J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* **2020**;21:482
17. Tan X, Su A, Tran M, Nguyen Q. SpaCell: integrating tissue morphology and spatial gene expression to predict disease cells. *Bioinformatics* **2020**;36:2293-4
18. Dries R, Zhu Q, Dong R, Eng CL, Li H, Liu K, *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* **2021**;22:78
19. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat Methods* **2020**;17:193-200

265 20. Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, *et al.* Spatial
266 transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* **2021**

267 21. Thrane K, Eriksson H, Maaskola J, Hansson J, Lundeberg J. Spatially Resolved
268 Transcriptomics Enables Dissection of Genetic Heterogeneity in Stage III Cutaneous
269 Malignant Melanoma. *Cancer Res* **2018**;78:5970-9

270 22. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential
271 expression analysis of digital gene expression data. *Bioinformatics* **2010**;26:139-40

272 23. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model
273 analysis tools for RNA-seq read counts. *Genome Biol* **2014**;15:R29

274 24. Diggle PJ, Ribeiro PJ. Spatial prediction. Model-based Geostatistics. New York, NY:
275 Springer New York; 2007. p 134-56.

276 25. Moran PA. Notes on continuous stochastic phenomena. *Biometrika* **1950**;37:17-23

277 26. Geary RC. The contiguity ratio and statistical mapping. *The incorporated statistician*
278 **1954**;5:115-46

279 27. Getis A, Ord JK. The analysis of spatial association by use of distance statistics.
280 Perspectives on spatial data analysis: Springer; 2010. p 127-45.

281 28. Bivand R, Anselin L, Berke O, Bernat A, Carvalho M, Chun Y, *et al.* spdep: Spatial
282 dependence: weighting schemes, statistics and models: R package version 0.5-31, URL
283 <http://CRAN.R-project.org/package=spdep>; 2011.

284 29. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity
285 landscape. *Genome Biol* **2017**;18:220

286 30. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, *et al.*
287 Inferring tumour purity and stromal and immune cell admixture from expression data.
288 Nat Commun **2013**;4:2612

289 31. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: Clustering, Classification and
290 Density Estimation Using Gaussian Finite Mixture Models. R Journal **2016**;8:289-317

291 32. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the
292 Dynamic Tree Cut package for R. Bioinformatics **2008**;24:719-20
293

Figure captions

Figure 1. Inputs (A) and outputs (B-D) from the analysis of spatially resolved transcriptomic data using spatialGE. **A.** Gene count matrices (M) with features (g) as rows and spots (s) as columns, as well as spot coordinates (C) are used to create a Spatial Transcriptomics List (STList). Optionally, a table with data on clinical variables (V) can be provided. **B.** Gene counts are voom-transformed, and “quilt” plots can be generated showing expression of a given gene at each spot. The transformed expression values can also be used to generate transcriptomic surfaces via “kriging”. Spatial autocorrelation statistics are also calculated (e.g., Moran’s I), which can be used to explore their relationship with clinical information (e.g., survival). **C.** Gene expression deconvolution via xCell and ESTIMATE is performed and immune cell type surfaces can be plotted, as well as classifications of spots into tumor or stroma compartments. **D.** spatialGE implements hierarchical clustering of the spots by weighting the transcriptomic differences in gene expression with the spatial distances between spots.

Figure 2. Results from spatialGE analysis of a melanoma stage IIIc using spatial transcriptomics (patient 1, slice 2 (21)). **A.** Quilt plot showing voom-transformed expression at each spot for immune gene *CD74*, with squares and crosses indicating tumor and stromal spots respectively, as inferred by ESTIMATE (A–top left). The tumor and stromal assignments can also be observed also in the accompanying plot (A–top right). A transcriptomic surface of *CD74* expression allows for a clearer visualization of the tissue heterogeneity (A–bottom left), with gray squares indicating the ESTIMATE tumor assignments. The inferred abundance of B cells in the tissue can be visualized by the predicted surface of xCell scores (A–bottom right). **B.** Relationship between the spatial autocorrelation statistics and standard deviation of gene expression. The red

dot represents *CD74*, and asterisk (*) indicates a significant relationship between the spatial statistic and the standard deviation. **C.** Scatter plots showing the relationships between spatial statistics calculated for *CD74* and the survival of patients in months (21). The colored arrows provide a guide for the interpretation of the spatial statistics. The dot with black border indicates the tissue section from patient 1 featured in panels the rest of the panels. **D.** ESTIMATE tumor/stroma assignments and spatially-informed clusters (using a spatial weight of 0.1) are shown. The yellow (1) and blue (2) clusters roughly represent tumor and stroma spots respectively. The red cluster (3) likely correspond to spots showing immune activity given the topological match with B cell scores (see **A**).

Supplementary Figure captions

Supplementary Fig. 1. Raw and voom-transformed counts for a set of genes within the melanoma ST array 1 from patient 1 (21): **A.** *PMEL*; **B.** *SPPI*; **C.** *SOX10*; **D.** *CD74*.

Supplementary Fig. 2. Transcriptomic surfaces resulting from spatial interpolation (“kriging”) of *PMEL* (**A**), *SPPI* (**B**), *CD74* (**C**), *IGLL5* (**D**), *SOX10* (**E**), and *SOX4* (**F**). The same genes are shown for all the tissue slices in Thrane et al. (21).

Supplementary Fig. 3. Cluster assignment of the spots as generated by unsupervised spatially-informed clustering and dynamic tree cuts. Spatial weights (w) from 0.0 to 0.5 were evaluated. **A.** $w=0$; **B.** $w=0.025$; **C.** $w=0.05$; **D.** $w=0.075$; **E.** $w=0.1$; **F.** $w=0.15$; **G.** $w=0.2$; **H.** $w=0.5$.

Supplementary Fig. 4. Cluster assignment of the spots as generated by unsupervised spatially-informed clustering and different values of k from 2 to 5. **A.** $k=2$; **B.** $k=3$; **C.** $k=4$; **D.** $k=5$. Each page in this file shows clusters at different spatial weights from $w=0.0$ to $w=0.5$.