Title: spatialGE: Quantification and visualization of the tumor microenvironment heterogeneity using spatial transcriptomics

Oscar E. Ospina
Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL 33612 USA

Christopher M. Wilson
Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL 33612, USA

Alex C. Soupir
Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL 33612, USA

Anders Berglund
Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL 33612, USA

Inna Smalley
Department of Tumor Biology, Moffitt Cancer Center, Tampa, FL 33612, USA

Kenneth Y. Tsai
Department of Anatomic Pathology, Moffitt Cancer Center, Tampa, FL 33612, USA

Brooke L. Fridley*
Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL 33612, USA
***Contact:** Brooke.Fridley@Moffitt.org

Abstract

Summary: Spatially-resolved transcriptomics promises to increase our understanding of the tumor microenvironment and improve cancer prognosis and therapies. Nonetheless, analytical methods to explore associations between the spatial heterogeneity of the tumor and clinical data are not available. Hence, we have developed spatialGE, a software that provides visualizations and quantification of the tumor microenvironment heterogeneity through gene expression surfaces, spatial heterogeneity statistics (SThet) that can be compared against clinical information, spot-level cell deconvolution, and spatially-informed clustering (STclust), all using a new data object to store data and resulting analyses simultaneously.

Availability and implementation: The R package and tutorial/vignette are available at https://github.com/FridleyLab/spatialGE. A script to reproduce the analyses in this manuscript is available in Supplementary information.

Contact: Fridley.Lab@Moffitt.org or Brooke.Fridley@Moffitt.org

Supplementary information: Available at Bioinformatics online.

1 Introduction

Spatially-resolved transcriptomics (ST) has allowed a better understanding of the tumor microenvironment (TME), immune infiltration, and its relationship with immunotherapy response (Tang, et al., 2016), as well as promising to improve the development of cancer prognoses and therapies (Nederlof, et al., 2021). Many ST methods involve mRNA probe hybridization in immobilized tissues on surfaces (Zhang, et al., 2021). Hence, gene expression localization within the tissue architecture is preserved as opposed to other transcriptomic approaches, such as bulk RNA (RNA-seq) or single-cell RNA sequencing (scRNAseq)

(Maniatis, et al., 2021; Yu, et al., 2021). Given the high-dimensional nature of ST experiments, data structures are needed to accommodate the gene expression abundance, spatial locations, and clinical information for samples taken from multiple subjects. Similarly, methods for ST visualization and analysis are needed, with existing tools varying in flexibility and approaches (Bergenstrahle, et al., 2020; Dries, et al., 2021; Hao, et al., 2021; Navarro, et al., 2017; Sun, et al., 2020; Tan, et al., 2020; Zhao, et al., 2021).

To meet these unmet computational needs, we have developed a pipeline for the processing and analysis of probe-based NGS ST experiments and exploration of the TME, which we call spatialGE. Novel features of the software include the generation of transcriptomic surfaces; quantification of spatial heterogeneity and the association with clinical information; gene expression deconvolution; spatially-informed clustering (STclust); and a new data structure to store multi-sample spatial data, subject-level metadata, and analytical results.

2 Implementation

Our new software takes gene expression and spatial location data as input, and optionally, associated metadata (i.e., clinical information), and stores it in a new R object class referred to as 'STList'. Other existing data structures can store metadata, however the STList takes a single table with just one row of information for each sample. This approach makes inputting data easier for users and allows spatialGE to complete downstream analyses with few lines of code. Users also have the possibility to provide Visium outputs directly from space ranger or Seurat objects, as well as .dcc/.pkc files from GeoMx experiments. Additionally, results from spatialGE analyses are saved in the STList making downstream statistical analysis and visualization of results seamless. The software allows users to perform gene- or spot-wise filtering and generate

73 quality control visualizations, followed by library size normalization and logarithmic or voom
74 transformation (Law, et al., 2014). The relative gene expression at each spot can be visualized
75 using "quilt plots" (Fig. 1A). If the STList is created from space ranger outputs or Seurat objects,
76 histological images are imported automatically, allowing to plot gene expression and these
77 images side by side for comparative analysis. Higher resolution plots are achieved with
78 transcriptomic surfaces produced by spatial interpolation ("kriging"; Fig. 1B) (Diggle and
79 Ribeiro, 2007). Surfaces can also be generated for deconvoluted cell type scores (Fig. 1C), with
80 cell type compositions at each spot inferred via gene expression deconvolution in xCell (Aran, et
81 al., 2017). Lastly, tumor purity scores are generated using ESTIMATE (Yoshihara, et al., 2013),
82 followed by tumor/stroma classification using model-based clustering (Fig 1A-C) (Scrucca, et
83 al., 2016). See Supplementary methods for additional details.

84 spatialGE also provides users with spatial statistics to quantify TME heterogeneity
85 (Supplementary Table 1; Supplementary Fig. 1). The Moran's I (Moran, 1950) and Geary's C
86 (Geary, 1954) allow users to ascertain whether gene expression is uniform throughout the tissue.
87 The Getis-Ord G_i statistic measures the tendency of a gene to produce expression hot- or cold-
88 spots (Getis and Ord, 2010). Users can associate per-sample spatial heterogeneity statistics with
89 other measurable clinical outcomes present in the sample metadata (SThet; Fig 1D). Finally,
90 spatialGE includes a computationally efficient spatially-informed unsupervised clustering
91 method, referred to as STclust, to detect TME compartments or "niches" (Fig. 1E). First, the
92 approach computes principal components (PCs) from gene counts. Then, a distance matrix is
93 computed based on two scaled distance matrices: 1) transcriptomic autocorrelation between spots
94 using the PCs (D_1), and 2) spatial distances between spots (D_2). Next, the autocorrelation matrix
95 (D_1) is "shrunk" toward the spatial distances matrix (D_2) by calculating its weighted average as

$D = [(1 - w) * D_1] + (w * D_2)$. The user specifies the weight (w) to apply. Based on our experience, a weight between 0.025 and 0.25 seems to best capture tissue heterogeneity (Supplementary Figs. 2 and 3).

3 Discussion

The spatialGE package is a comprehensive analytical R package for the simultaneous analysis of multiple tissues assayed with probe-based ST technologies (i.e., Visium, GeoMx) using the new STList R object class. spatialGE is unique in implementing kriging to generate gene expression surfaces at high resolution for multiple ST tissue sections simultaneously (for an alternative method see (Zhao, et al., 2021)). Several ST clustering approaches are already available (Supplementary Table 2) (Bergenstrahle, et al., 2020; Dries, et al., 2021; Tan, et al., 2020; Zhao, et al., 2021), however our spatially-informed clustering approach, STclust, is computationally efficient (Supplementary Fig. 4) and resembles tissue features by only weighting transcriptomic similarities by their distances between spots. A comparison of Louvain clustering as implemented in the widely used software Seurat to STclust, showed a Rand similarity index of 0.8 for k=4 (Supplementary Fig. 5), indicating good agreement of spot assignments between the two methods. Another innovation of spatialGE is the estimation of quantifiable metrics (spatial heterogeneity statistics, SThet) to capture transcriptomic complexity and the ability to compare them against clinical outcomes.

Applying spatialGE to the Thrane et al. melanoma ST data, we observed agreement between the transcriptomic surfaces (Supplementary Fig. 6) and tumor and stroma regions observed in pathology images (see (Thrane, et al., 2018)). STclust provided TME compartments (i.e., clusters; Supplementary Figs. 2 and 3) that resembled the features annotated in pathology

images (Thrane, et al., 2018). Finally, we observed an association between spatial heterogeneity and patient survival, providing support to the hypothesis that heterogeneity is a predictor of patient outcomes (Nederlof, et al., 2021).

In future versions, spatialGE will include analysis tools for different ST technologies (e.g., Slide-seq). We are currently developing additional analytical methods and visualizations that use histological annotation tools for non-gridded technologies (e.g, GeoMx), as well as a web-based implementation of spatialGE. Another developing area is the integration of scRNA-seq data with ST for spot-level cell phenotyping.

Acknowledgments

Funding: This work was supported by the National Institutes of Health (NIH) [T32-CA233399, P30-CA076292, R00-CA226679]. This content is solely the responsibility of the authors and does not necessarily represent the official view of the NIH. We thank the spatial research group at the KTH Royal Institute of Technology – SciLifeLab (Stockholm, Sweden) for making their melanoma data available.

Conflicts of interests: none declared.

References

Aran, D., Hu, Z. and Butte, A.J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* 2017;18(1):220.

Bergenstrahle, J., Larsson, L. and Lundeberg, J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* 2020;21(1):482.

141 Diggle, P.J. and Ribeiro, P.J. Spatial prediction. In, *Model-based Geostatistics*. New York, NY:
142 Springer New York; 2007. p. 134-156.
143 Dries, R.*, et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression
144 data. *Genome Biol* 2021;22(1):78.
145 Geary, R.C. The contiguity ratio and statistical mapping. *The incorporated statistician*
146 1954;5(3):115-146.
147 Getis, A. and Ord, J.K. The analysis of spatial association by use of distance statistics. In,
148 *Perspectives on spatial data analysis*. Springer; 2010. p. 127-145.
149 Hao, Y.*, et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;184(13):3573-3587
150 e3529.
151 Law, C.W.*, et al.* voom: Precision weights unlock linear model analysis tools for RNA-seq read
152 counts. *Genome Biol* 2014;15(2):R29.
153 Maniatis, S., Petrescu, J. and Phatnani, H. Spatially resolved transcriptomics and its applications
154 in cancer. *Curr Opin Genet Dev* 2021;66:70-77.
155 Moran, P.A. Notes on continuous stochastic phenomena. *Biometrika* 1950;37(1/2):17-23.
156 Navarro, J.F.*, et al.* ST Pipeline: an automated pipeline for spatial mapping of unique transcripts.
157 *Bioinformatics* 2017;33(16):2591-2593.
158 Nederlof, I.*, et al.* A High-Dimensional Window into the Micro-Environment of Triple Negative
159 Breast Cancer. *Cancers (Basel)* 2021;13(2).
160 Scrucca, L.*, et al.* mclust 5: Clustering, Classification and Density Estimation Using Gaussian
161 Finite Mixture Models. *R Journal* 2016;8(1):289-317.
162 Sun, S., Zhu, J. and Zhou, X. Statistical analysis of spatial expression patterns for spatially
163 resolved transcriptomic studies. *Nat Methods* 2020;17(2):193-200.

164 Tan, X., *et al.* SpaCell: integrating tissue morphology and spatial gene expression to predict
165 disease cells. *Bioinformatics* 2020;36(7):2293-2294.
166 Tang, H., Qiao, J. and Fu, Y.X. Immunotherapy and tumor microenvironment. *Cancer Lett*
167 2016;370(1):85-90.
168 Thrane, K., *et al.* Spatially Resolved Transcriptomics Enables Dissection of Genetic
169 Heterogeneity in Stage III Cutaneous Malignant Melanoma. *Cancer Res* 2018;78(20):5970-5979.
170 Yoshihara, K., *et al.* Inferring tumour purity and stromal and immune cell admixture from
171 expression data. *Nat Commun* 2013;4:2612.
172 Yu, X., *et al.* Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA
173 Sequencing Experiments. *Methods Mol Biol* 2021;2194:143-175.
174 Zhang, M., *et al.* Spatial molecular profiling: platforms, applications and analysis tools. *Brief*
175 *Bioinform* 2021;22(3).
176 Zhao, E., *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol*
177 2021.
178

179 **Figure captions**

180 **Graphical abstract.** Overview of spatialGE features. **A.** The STList data object from spatialGE
181 can be creared from several sources, including comma- or tab-separated files containing gene
182 counts and spatial coordinates. The object can also be created directly from Visium outputs,
183 Seurat objects, or GeoMx outputs. **B.** Users can optionally provide a metadata file, containing
184 information associated with each sample (one row per sample, or per ROI if GeoMx data). **C.**
185 Methods for quality control of data are provided by spatialGE, including visualizations of counts
186 and genes per spot, as well as filtering of spots or genes within user-determined thresholds. **D.** A
187 novel method (STclust) performs spatially informed clustering of spots and tissue domain
188 identification. **E.** spatialGE provides different types of data visualization, including gene
189 expression at each spot ("quilt plots"), as well as adaptation of spatial interpolation ("kriging") to
190 spatial transcriptomics data (transcriptomic surface). **F.** spatialGE also leverages spatial statistics
191 (Moran's I, Geary's C, Getis-Ord Gi) to quantitatively describe heterogeneity within the tumor
192 microenvironment and to explore associations between spatial heterogeneity and clinical
193 oucomes. **G.** Gene expression deconvolution can also be applied to each spot to detect immune
194 cell types (xCell) and classification of spots as tumor or stroma (ESTIMATE).

195

196 **Figure 1.** Results from spatialGE analysis of a melanoma stage IIIc using spatial transcriptomics
197 (patient 1, sample 2 (Thrane, et al., 2018)). **A.** Quilt plot showing log-expression at each spot for
198 immune gene *CD74*, with squares and crosses indicating tumor and stromal spot classifications
199 respectively. **B.** Transcriptomic surface of *CD74* expression allows for a clearer visualization of
200 the tissue heterogeneity, with gray squares indicating tumor compartment. **C.** The inferred
201 abundance of B cells in the tissue can be visualized by the predicted surface. **D.** Scatter plots

202 showing the relationships between spatial statistics calculated for *CD74* and the survival of
203 patients in months from (Thrane, et al., 2018). The colored arrows provide a guide for the
204 interpretation of the spatial statistics. The dot with black border indicates the tissue section from
205 patient 1 featured in the other panels in Fig 1. **E.** Tumor/stroma assignments and spatially-
206 informed clusters with STclust (using a spatial weight of 0.025) are shown. The yellow (1) and
207 blue (2) clusters roughly represent tumor and stroma regions of the tissue, respectively. The red
208 cluster (3) likely correspond to spots showing immune activity given the topological match with
209 *CD74* (Fig. 1B) and B cell scores (see Fig. 1**C**).
210
211

212 **Supplementary Figure captions**

213 **Supplementary Fig. 1.** Relationship between the spatial autocorrelation statistics and standard
214 deviation of gene expression. The red dot represents *CD74*, and asterisk (*) indicates a
215 significant relationship between the spatial statistic and the standard deviation.

216

217 **Supplementary Fig. 2.** Cluster assignment of the spots as generated with STclust and applying
218 automatic dynamic tree cutting. Spatial weights (*w*) from 0.0 to 0.5 were evaluated. **A.** *w*=0; **B.**
219 *w*=0.025; **C.** *w*=0.05; **D.** *w*=0.0.75; **E.** *w*=0.1; **F.** *w*=0.15; **G.** *w*=0.2; **H.** *w*=0.5.

220

221 **Supplementary Fig. 3.** Cluster assignment of the spots as generated with STClust using
222 different values of k, k=2 to 5. **A.** k=2; **B.** k=3; **C.** k=4; **D.** k=5. Each page in this file shows
223 clusters at different spatial weights from *w*=0.0 to *w*=0.5.

224

225 **Supplementary Fig. 4.** Execution time of unsupervised spatially-informed clustering with
226 STclust (k=4) in relation to the number of spots from each of the tissue slices in (Thrane, et al.,
227 2018). Labels accompanying points indicate the patient and replicate numbers.

228

229 **Supplementary Fig. 5.** Results Louvain clustering using Seurat and STclust with k=4 and
230 spatial weights *w*=0 (**A**); *w*=0.025 (**B**); *w*=0.05 (**C**); *w*=0.075 (**D**); *w*=0.1 (**E**); *w*=0.15 (**F**); *w*=0.2;
231 (**G**); and *w*=0.5 (**H**). Louvain clustering used a resolution parameter=0.8 (**I**). Clustering was
232 applied to tissue slice 2 from patient 1 (Thrane, et al., 2018). Rand indexes (R) are shown for the
233 comparison between STclust (**A-H**) and Louvain clustering (**I**). The red boxes indicate the two
234 most concordant clustering solutions (R=0.81).

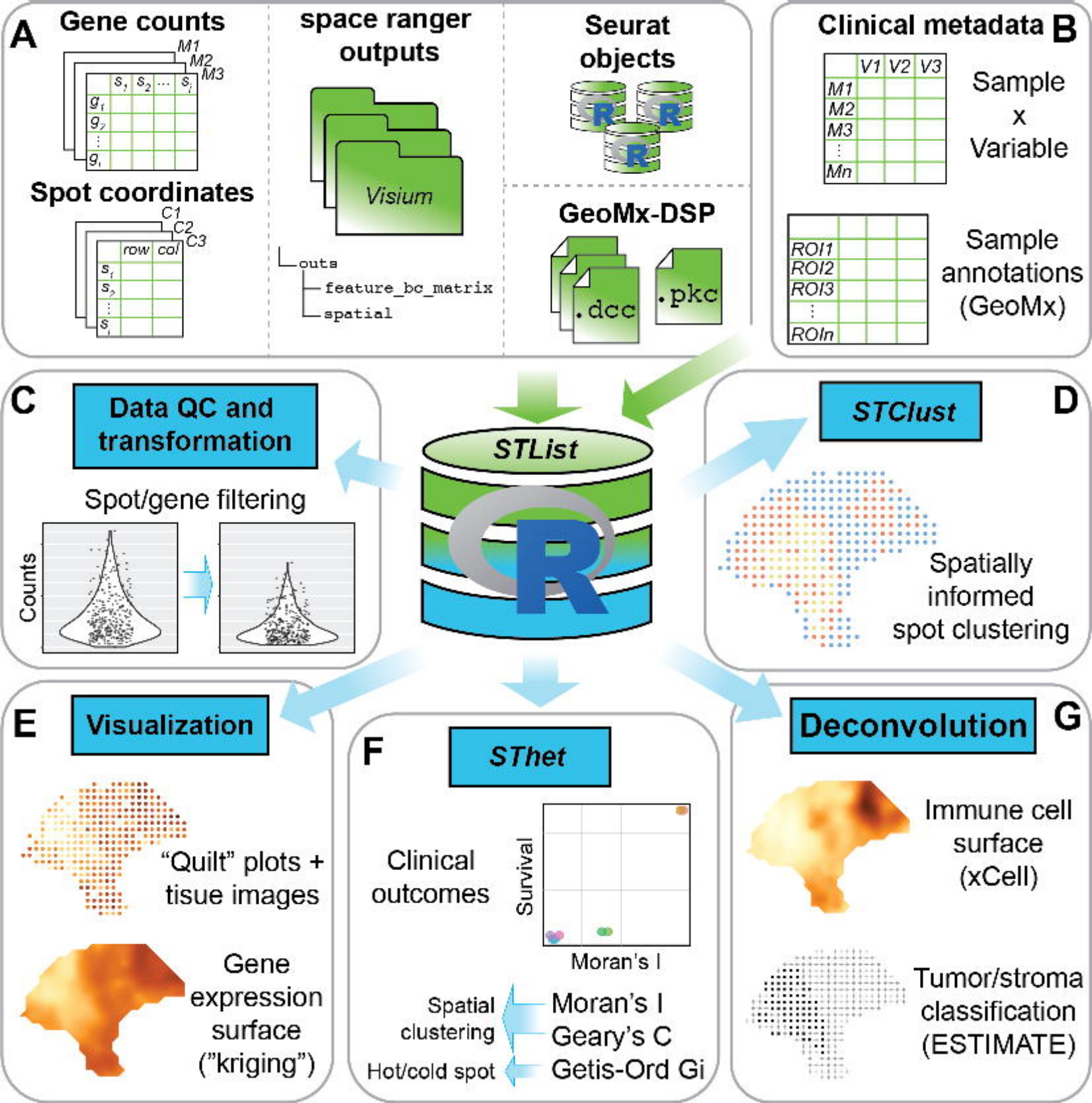
235

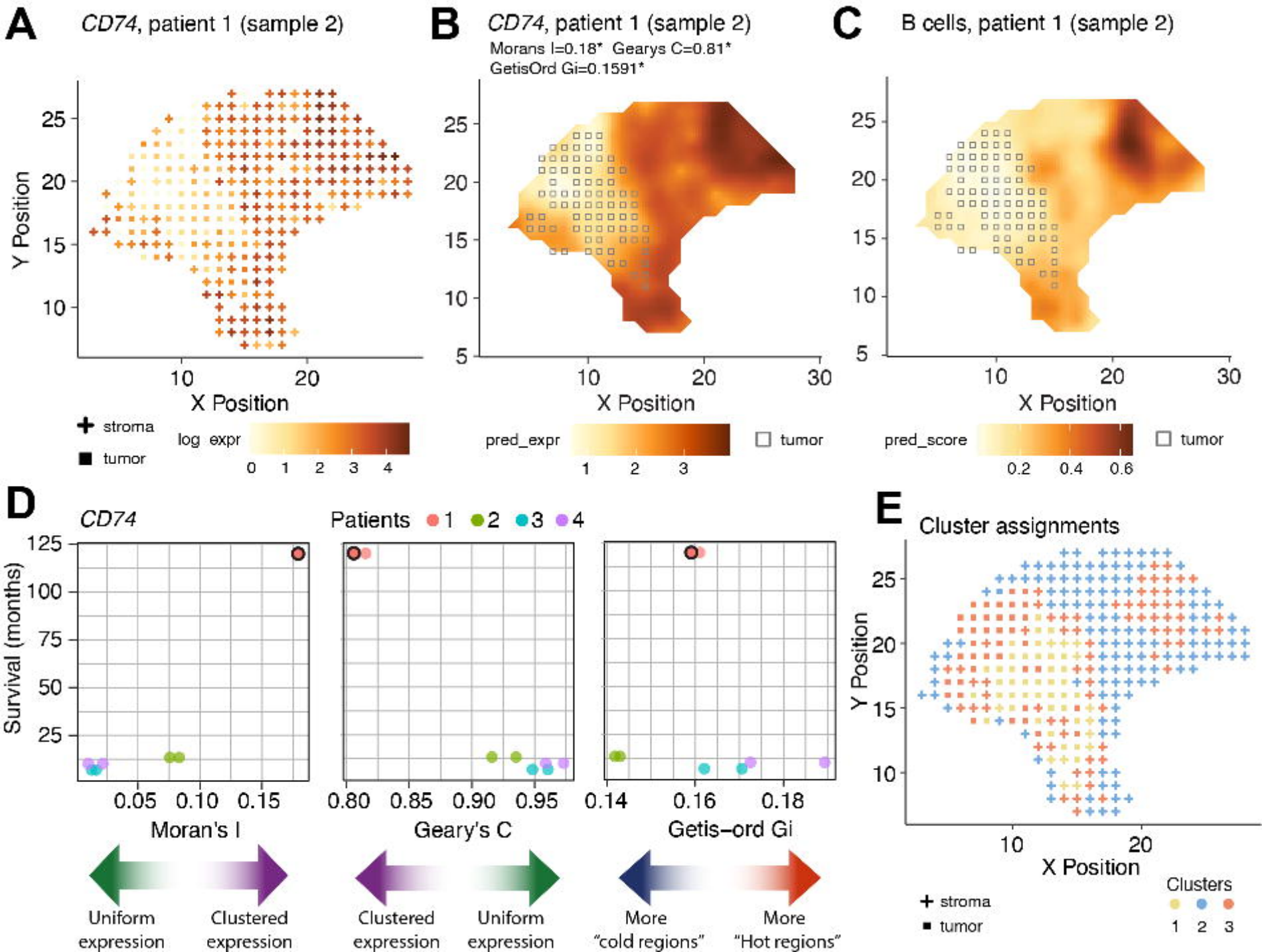
236 **Supplementary Fig. 6.** Transcriptomic surfaces resulting from spatial interpolation ("kriging")

237 of *PMEL* (**A**), *SPP1* (**B**), *CD74* (**C**), *IGLL5* (**D**), *SOX10* (**E**), and *SOX4* (**F**). The same genes are

238 shown for all the tissue slices in (Thrane, et al., 2018).

239


A
Gene counts
M1
M2
M3
Spot coordinates
C1
C2
C3
row
col
space ranger outputs
Visium
outs
feature_bc_matrix
spatial
Seurat objects
GeoMx-DSP
.dcc
.pkc
Clinical metadata
B
V1
V2
V3
M1
M2
M3
Mn
Sample x Variable
ROI1
ROI2
ROI3
ROIn
Sample annotations (GeoMx)
C
Data QC and transformation
Spot/gene filtering
Counts
STList
STClust
D
Spatially informed spot clustering
E
Visualization
"Quilt" plots + tisue images
Gene expression surface ("kriging")
F
SThet
Clinical outcomes
Survival
Moran's I
Spatial clustering
Moran's I
Geary's C
Hot/cold spot
Getis-Ord Gi
Deconvolution
G
Immune cell surface (xCell)
Tumor/stroma classification (ESTIMATE)


A
CD74, patient 1 (sample 2)
Y Position
X Position
stroma
tumor
log expr
0 1 2 3 4
B
CD74, patient 1 (sample 2)
Morans I=0.18* Gearys C=0.81*
GetisOrd Gi=0.1591*
X Position
pred_expr
1 2 3
tumor
C
B cells, patient 1 (sample 2)
X Position
pred_score
0.2 0.4 0.6
tumor
D
CD74
Patients 1 2 3 4
Survival (months)
Moran's I
Geary's C
Getis–ord Gi
Uniform expression
Clustered expression
Clustered expression
Uniform expression
More "cold regions"
More "Hot regions"
E
Cluster assignments
Y Position
X Position
stroma
tumor
Clusters
1 2 3