doi:10.20944/preprints202209.0388.v1

*Article*

# Geostatistical Modeling and Heterogeneity Analysis of Tumor Molecular Landscape

**Morteza Hajihosseini[1], Payam Amini[2], Dan Voicu[3], Irina Dinu[1]*, Saumyadipta Pyne[4,5]***

[1] School of Public Health, University of Alberta, Edmonton, AB, Canada.
[2] Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran.
[3] Faculty of Engineering, McGill University, Montreal, QC, Canada.
[4] Health Analytics Network, Pittsburgh, PA, USA.
[5] Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA.
* Correspondence: idinu@ualberta.ca, spyne@ucsb.edu

**Simple Summary:**The present study introduces a new computational platform GATHER to conduct Geostatistical Analysis of Tumor Heterogeneity and Entropy in R. GATHER has several distinct advantages such as (a) novel use of single cell specific spatial information for kriging to synthesize high-resolution and continuous gene expression landscapes of a given tumor sample, (b) integration of such landscapes to identify and map the enriched regions of pathways of interest, (c) identification of genes that have spatial differential expression at locations representing specific phenotypic contexts, (d) computation of spatial entropy measures for quantification and objective characterization of intratumor heterogeneity, and (e) use of new tools for insightful visualization of spatial transcriptomic phenomena.

**Abstract:** Intratumor heterogeneity (ITH) is associated with therapeutic resistance and poor prognosis in cancer patients, and attributed to genetic, epigenetic, and microenvironmental factors. We developed a new computational platform, GATHER, for geostatistical modeling of single cell RNA-seq data to synthesize high-resolution and continuous gene expression landscapes of a given tumor sample. Such landscapes allow GATHER to map the enriched regions of pathways of interest in the tumor space and identify genes that have spatial differential expressions at locations representing specific phenotypic contexts using measures based on optimal transport. GATHER provides new applications of spatial entropy measures for quantification and objective characterization of ITH. It includes new tools for insightful visualization of spatial transcriptomic phenomena. We illustrate the capabilities of GATHER using real data from breast cancer tumor to study hallmarks of cancer in the phenotypic contexts defined by cancer associated fibroblasts.

**Keywords:** spatial single-cell analysis; intratumor heterogeneity; kriging; spatial entropy; Wasserstein distance; cancer; RNA-seq

doi:10.20944/preprints202209.0388.v1

## Introduction

In their well-known paper in 2010, Hanahan and Weinberg noted that tumors exhibit an additional dimension of complexity through their "tumor microenvironment" that contributes to the acquisition of the so-called hallmark traits of cancer. Indeed, extensive studies over the past decades have uncovered a great diversity of cell populations in tumors, thus leading to the active research area of intratumor heterogeneity (ITH) (1). It has been found to be associated with poor prognosis, therapeutic resistance and treatment failure leading to poor overall survival in cancer patients (2-6). ITH is attributed to genetic, epigenetic, and microenvironmental factors (1, 7). Tumors can develop a resistance to the treatment due to ITH by new genetic mutations, recovering functionality of previously inactivated genes, phenotypic changes, and variations in response to the microenvironment (8, 9).

The persistence of some of the drug-tolerant intratumor cell populations could be attributed to their high phenotypic plasticity. While hierarchies of differentiation also exist among normal cells in healthy tissues, the populations of tumor cells, in contrast, display far greater cell-to-cell variability and the resulting phenotypic instability (10, 11). Such ITH could be attributed to genetic causes ranging from aneuploidy to spontaneous cell fusions, say, between cancer and non-cancer cells, in addition to other factors such as complex contextual signals in the highly aberrant tumor microenvironments, or even global alterations in cancer cell epigenomes (12). ITH also involves immune cell infiltration, which is of obvious importance for immunotherapies. Tumor antigen diversity could be determined by the T cell clonality in the different regions of the same tumor (13). Studies have shown spatially complex interactions between tumor microenvironments and the patient's immune system (14, 15).

While heterogeneous cell types are prevalent within the tumor microenvironment some of which may account for cancer development and progression, it also contains different non-malignant components, including the cancer-associated fibroblasts (CAFs) (16-18). Although the origin and activation mechanism of CAFs remains an area of active research (19-22), studies have attributed the processes of formation and derivation of CAFs to various precursor cells (20, 23), which may be the source of the well-known heterogeneity among the CAFs (24-27). Indeed, certain tumors, such as in the breast, in which the prevalence of CAFs could as high as 80%, they can play both anti- as well as pro-tumorigenic roles (28-30). Importantly, CAFs can facilitate drug resistance dynamically by altering the cell-matrix interactions that control the outer layer of cells' sensitivity to apoptosis, producing proteins that control cell survival and proliferation, assisting with cell-cell communications, and activating epigenetic plasticity in neighboring cells (31, 32).

To understand the spatial heterogeneity of gene expressions, including drug targets, different sites of the same tumor were analyzed with multiregional RNA sequencing for different cancers (5, 33-35). It was observed, for instance, that if HER2+ breast tumors expressed HER2 uniformly across their cells, then the known HER2-targeted therapies were highly effective; and if not, then such patients were found to have shorter disease-free survival (36). In recent years, higher resolution, tissue-specific gene expression analysis is made possible by using new platforms such as single-cell RNA sequencing (scRNA-seq), which has rapidly evolved as a powerful and popular tool (37, 38). This has led to several scRNA-seq studies of the composition of CAFs in different stages of cancer (39-47). For instance, the Human Tumor Atlas Network [https://humantumoratlas.org] is increasingly enriched with data on human cancers based on scRNA-seq assays.

Despite the advancements and efficacy of scRNA-seq, the lack of spatial information in scRNA-seq analysis is a significant shortcoming for typical scRNA-seq methods to capture cellular heterogeneity. On the other hand, while oncologic pathologists have long studied cell signaling within tumors by manual scoring of discordance between individuals and variation between different batches using tissue immunostaining and microscopy, such techniques typically allow only a few selected markers to be observed per cell, and thus offer a limited reporting of the extent of potential heterogeneity. Combining high-resolution gene expression data with spatial coordinates can resolve these experimental shortcomings (48). For instance, spatial proximity to fibroblasts has been shown to impact molecular features and therapeutic sensitivity of breast cancer cells influencing clinical outcomes (49). While imaging the transcriptome *in situ* with high accuracy has been a major challenge in single-cell biology, development of high-throughput platforms for sequential fluorescence in situ hybridization such as RNA seqFISH+ and algorithms such as CELESTA can identify cell populations and their spatial organization in intact tissues (5, 50).

In order to approach the conceptualize the diversity in the spatial omic information, the concept of a habitat and its locations have been studied in association with genetic heterogeneity in a tumor (51, 52). In fact, it was noted that the spatial distribution of genetically distinct tumor cell populations may correlate with poor clinical outcomes (9). Landscape ecology is, therefore, a potentially effective modeling framework which – similar to the modeling of an ecosystem's behaviour in terms of the actions and interactions of individuals and groups of the different constituent species – could be adopted to study the spatio-temporally dynamic and heterogeneous system that is often represented by a tumor.

The present study introduces a new computational platform GATHER to conduct Geostatistical Analysis of Tumor Heterogeneity and Entropy in R. GATHER uses geostatistical modeling and spatial entropy measures for quantification and objective characterization of intratumor heterogeneity, and to identify different transcriptomic patterns in the molecular landscapes of a tissue sample. Geostatistical models provide a well-established theoretical framework for prediction and interpolation of spatial data. Kriging, for example, is a generalized least-square regression approach to predict spatial attributes at unobserved locations (53). GATHER applies kriging for estimating gene-specific, and thereby geneset-specific, expression values at every point of the given tumor space. By constructing such continuous molecular landscapes, it allows visualization and identification of local and regional transcriptomic variations. Further, GATHER provides quantitative characterization of ITH based on spatial entropy measures (54). Finally, GATHER applies a Wasserstein distance based 2-sample test, which is adapted specifically for use on single cell data (55), to provide 2 different approaches to identify genes that have spatial differential expression either (i) near a specific

doi:10.20944/preprints202209.0388.v1

location in the tumor space versus elsewhere, or (ii) across different regions in which a selected phenotypic context is present at different levels.

The concept of entropy in Information Theory, as defined on strings of symbols by Claude Shannon in 1948 (56), has been adapted and used in various contexts because of its ability to capture a broad set of notions such as information content, unexpectedness, uncertainty, diversity, and contagion (54). Indeed, it was shown that the cancer epigenome has higher entropy than its normal counterpart (10). In the present study, we are more specifically interested in the spatial entropy of a tumor's molecular information content. Despite early applications of Shannon's entropy ($H$) to evaluate spatial heterogeneity in geographical phenomena (57) and landscape ecology (58), researchers have noted that, while $H$ takes into account the number of symbols of each type in a string, it ignores the effect of their spatial arrangement (59). This has led to further development of different entropy measures that specifically include spatial information. In particular, the present study computed Batty's entropy to measure the spatial heterogeneity of diverse phenotypes in the tumor space, and Leibovici's co-occurrence based entropy for heterogeneity of a given geneset's enrichment in a particular phenotypic context.

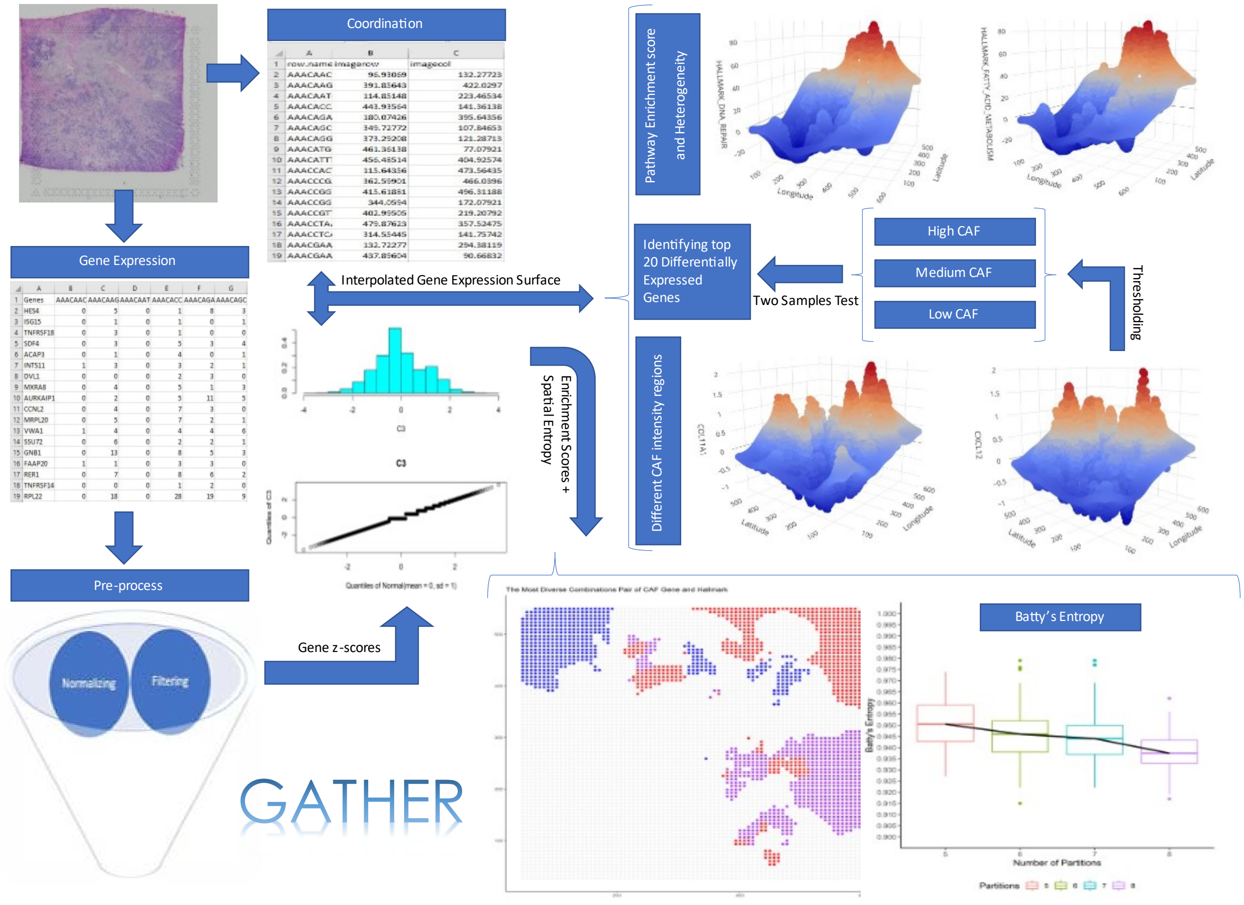
In 2020, a paper listed eleven grand challenges in single-cell data science, which included the challenge of "finding patterns in spatially resolved measurements" (60). Towards this, many recent efforts have produced computational methods to analyze spatial information in single-cell studies (61-69). The aim of the present study is to address the said challenge using a different – geostatistical modeling – approach in comparison to the existing ones. This gives GATHER several distinct advantages such as (a) use of single cell specific spatial information for kriging to synthesize high-resolution and continuous gene expression landscapes of a given tumor sample, (b) integration of such landscapes to identify and map the enriched regions of pathways of interest, (c) identification of genes that have spatial differential expression at locations representing specific phenotypic contexts, (d) computation of spatial entropy measures for quantification and objective characterization of ITH, and (e) use of new tools for insightful visualization of spatial transcriptomic phenomena. In the next section, we describe the data and methods, followed by the results of real tumor data analysis using GATHER, and end with discussion including future work.

## Data and Methods

### *Data*

The spatial transcriptomics data were downloaded from the 10x Genomics online resource [Available at: https://www.10xgenomics.com/resources/datasets/human-breast-cancer-whole-transcriptome-analysis-1-standard-1-2-0]. The data were generated using the Visium Spatial Gene Expression protocol run on an invasive breast cancer tissue sample that is Estrogen Receptor (ER) positive, Progesterone Receptor (PR) positive, and Human Epidermal Growth Factor Receptor 2 (HER2) negative. RNA sequencing data were generated with a paired-end, dual-indexed process using Illumina NovaSeq 6000, with a sequencing depth of 72,436 mean reads per cell. After filtering the downloaded dataset for average gene expression value

>1, the data matrix contained 1,981 rows (genes) and 4,325 columns (single cells). **Fig. 1** shows the steps of the GATHER workflow which includes spatial gene expression analysis. As part of our RNAseq data preprocessing, the zero counts were replaced with a small random jitter around zero that would minimally affect the remaining gene expression values. We normalized each gene's expression across samples with 10-fold cross-validation based data transformation method using *bestNormalize* package in R software (70).


Coordination
Gene Expression
Pre-process
Normalizing
Filtering
GATHER
Gene z-scores
Interpolated Gene Expression Surface
Enrichment Scores + Spatial Entropy
Pathway Enrichment score and Heterogeneity
HALLMARK_DNA_REPAIR
HALLMARK_FATTY_ACID_METABOLISM
Longitude
Latitude
Identifying top 20 Differentially Expressed Genes
Two Samples Test
High CAF
Medium CAF
Low CAF
Thresholding
Different CAF intensity regions
CCL11A1
CXCL12
C3
Quantiles of C3
Quantiles of Normal(mean = 0, sd = 1)
The Most Diverse Combinations Pair of CAF Gene and Hallmark
Batty's Entropy
Number of Partitions
Partitions 5 6 7 8

| | A | B | C |
|---|---|---|---|
| 1 | row.name | imagerow | imagecol |
| 2 | AAACAAC | 96.93069 | 132.27723 |
| 3 | AAACAAG | 391.85643 | 422.0297 |
| 4 | AAACAAT | 114.85148 | 223.46534 |
| 5 | AAACACC | 443.93564 | 141.36138 |
| 6 | AAACAGA | 180.07426 | 395.64356 |
| 7 | AAACAGC | 349.72772 | 107.84653 |
| 8 | AAACAGG | 373.29208 | 121.28713 |
| 9 | AAACATG | 461.36138 | 77.07921 |
| 10 | AAACATT | 456.48514 | 404.92574 |
| 11 | AAACCAC | 115.64356 | 473.56435 |
| 12 | AAACCCG | 362.59901 | 466.0396 |
| 13 | AAACCGG | 415.61881 | 496.31188 |
| 14 | AAACCGG | 344.0594 | 172.07921 |
| 15 | AAACCGT | 402.99505 | 219.20792 |
| 16 | AAACCTA | 479.87623 | 357.52475 |
| 17 | AAACCTC | 314.55445 | 141.75742 |
| 18 | AAACGAA | 132.72277 | 294.38119 |
| 19 | AAACGAA | 437.89604 | 90.66832 |

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Genes | AAACAAC | AAACAAG | AAACAAT | AAACACC | AAACAGA | AAACAGC |
| 2 | HES4 | 0 | 5 | 0 | 1 | 8 | 3 |
| 3 | ISG15 | 0 | 1 | 0 | 1 | 0 | 1 |
| 4 | TNFRSF18 | 0 | 3 | 0 | 1 | 0 | 0 |
| 5 | SDF4 | 0 | 3 | 0 | 5 | 3 | 4 |
| 6 | ACAP3 | 0 | 1 | 0 | 4 | 0 | 1 |
| 7 | INTS11 | 1 | 3 | 0 | 3 | 2 | 1 |
| 8 | DVL1 | 0 | 0 | 0 | 2 | 3 | 0 |
| 9 | MXRA8 | 0 | 4 | 0 | 5 | 1 | 3 |
| 10 | AURKAIP1 | 0 | 2 | 0 | 5 | 11 | 5 |
| 11 | CCNL2 | 0 | 4 | 0 | 7 | 3 | 0 |
| 12 | MRPL20 | 0 | 5 | 0 | 7 | 2 | 1 |
| 13 | VWA1 | 1 | 4 | 0 | 4 | 4 | 6 |
| 14 | SSU72 | 0 | 6 | 0 | 2 | 2 | 1 |
| 15 | GNB1 | 0 | 13 | 0 | 8 | 5 | 3 |
| 16 | FAAP20 | 1 | 1 | 0 | 3 | 3 | 0 |
| 17 | RER1 | 0 | 7 | 0 | 8 | 6 | 2 |
| 18 | TNFRSF14 | 0 | 0 | 0 | 1 | 2 | 0 |
| 19 | RPL22 | 0 | 18 | 0 | 28 | 19 | 9 |

**Fig. 1.** The GATHER workflow. It outlines the different analytical steps taken by GATHER starting from single cell omic data preparation including normalization and filtering to the generation of krging-predicted gene expression landscapes as well as iterative computation of spatial entropy measures. It also illustrates the interactive 3D visualization using GATHER of the computed gene- and geneset-specific landscapes defined over the input tissue space.

### *Constructing Gene Expression Landscape by Kriging*

Our dataset is defined on a 2-dimensional tissue space, with a specified coordinate system. We discretized this space using an evenly spaced grid of size $80 \times 80$, i.e., 6400 unique point locations over a rectangular area covering 50 units below (above) the minimum (maximum) values of x and y coordinates of the cells in our dataset.

In this study, the geostatistical method of Ordinary Kriging (OK) was used for interpolating the expression value of each gene $g$ at each grid-point

$p$ based on the best linear unbiased prediction (blup) using a weighted average expression of $g$ in the cells that lie in a given neighborhood of $p$. The basic model for the OK predictor (Waller and Gotway 2004) of the expression $Z(g,s_0)$ of $g$ at a location $s_0$ in the given tissue space is computed as

$$\hat{Z}(g,s_0) = \sum_{i=1}^{N_g} \lambda_{g,i} Z(g,s_i)$$

where is the measured expression value at the location $s_i$ of cell $i$, $\lambda_{g,i}$ is the weight attributed to the measured expression of $g$ at location $s_i$, and $N_g$ is the number of available single cell measurements of the expression of $g$. For OK, we assume stationary $Z(g,\cdot)$ and a known semivariogram (of $g$). The kriging weights that determine the contributions of the measurements are defined by an empirical semivariogram function that describes the spatial dependence among the single cell expression values of $g$ in terms of intercellular distance (53). Typically, such contribution to the kriged expression value at $s_0$ decrease for a cell $s_i$ as it gets farther from $s_0$. GATHER also computes the kriging standard error (71) at the same location $s_0$ which gives a measure of the uncertainty of the prediction of $Z(g,s_0)$. Thus, GATHER constructs gene-specific, continuous transcriptomic landscapes, along with the maps of the corresponding standard errors, which could be visualized for each gene separately (or as spatially combined for a given geneset) an example of which is shown in Fig. 1.

### *Test of Spatial Differential Expression of Genes*

Our platform allows us to identify a spatial phenotype in terms of differential expression one or more "marker" genes that is known to characterize the phenotype. This allows us to demarcate and map the regions in the tissue space where the phenotype is significant. To map the co-occurrence of more than one phenotype, distinct colors were used. Further, the presence of these spatial phenotypes could serve as specific *contexts* within which certain genes of interest may show differential expression. Indeed, we developed methods for identifying such genes as well as measuring contextual enrichment of genesets and curated molecular pathways.

Differentially expressed genes were detected using the semi-parametric 2-Wasserstein distance test for single cell data (55). In this study, the test was applied in a spatially contextualized manner using two different approaches. In our first approach, we identified and mapped the significance of local expression of a given gene at any point of the tissue space, which is systematically discretized by a well-defined grid (see above). For this purpose, we begin by grouping the cells that are local to a given location and distinguish them from the group of nonlocal cells that are distant from this location. At each point $p$ of the grid, we defined a neighborhood $Nbd(p,r)$ centered at $p$ based on a circle of radius $r$. The value of $r$ is chosen to be the 25th percentile of all pairwise distances between the cells, thus ensuring proximity among the cells that lie in $Nbd(p,r)$. The cells that lie in $Nbd(p,r)$ are called "local", and the rest are termed "non-local".

In our second approach, the entire tissue space was partitioned into regions according to different levels of enrichment of a phenotype of interest, e.g., characterized by expression of markers of cancer associated fibroblasts (CAF). The regions of the landscape are thus marked by a pre-determined $l$

(=3) discrete levels of the selected phenotype: high (CAF $z > 1$), mid (CAF $0.5 < z \leq 1$) and low (CAF $z \leq 0.5$). These regions provide the graded spatial contexts in which certain genes may express. Thus, we used the 2-Wasserstein distance method to compare the single-cell level expression of each gene across successive levels of the phenotype, i.e., across (a) the high and the medium regions; and (b) the medium and the low regions. For a selected phenotype, the significantly differentially expressed genes are identified by permutation testing (with 100 repetitions) at a pre-determined FDR adjusted q-value level (say, 0.2) (72).

### *Spatial Analysis of Hallmark Genesets of Cancer*

Geneset enrichment landscape construction:

Let $L$ be the list of genes whose expressions are measured (and thus available as spatial z-scores) in the present study.

For a geneset $S$, we computed the spatial enrichment z-score, $SEZ(S,p)$, at each grid point $p$ using the Stouffer's sum of the spatial z-scores, $SZ(g,p)$, of expressions of the genes in $S$ and $L$ at $p$ as follows:

$SEZ(S,p) = \sum_{g \epsilon S \cap L} SZ(g,p) / \surd|S \cap L|$.

This allows us to construct a geneset enrichment landscape, which extends the idea of single gene expression landscape.

Cancer Hallmark Geneset Enrichment:

We downloaded from the Molecular Signatures Database (MSigDB) genesets (Table S1) that represent commonly known "hallmarks" of cancer (73). To ensure their relevance as well as non-redundancy, we selected 8 of those hallmark genesets that have at least 25% overlap with the expressed genes (see above text on preprocessing) but mutual geneset overlap of less than 10%.

### *Spatial Entropy of a Tumor Sample*

Calculating phenotypic diversity:

Given our interest to characterize the heterogeneity of a given tumor sample in terms of the spatial phenotypes therein, Batty's entropy measure was computed to evaluate the distribution of a candidate phenotype over the given tissue area by allowing for partitioning the same into subareas of different sizes and shapes. Let a tissue area of size $A$ partitioned into G subareas of size $A_g, g = 1, \dots G$. If a phenotype of interest $F$ occurs in $A$, and in $A_g$ with probability $p_g$, then $\sum_{g=1}^{G} p_g = 1$. The phenotype intensity in $A_g$is given by $\lambda_g = p_g / A_g$.

Batty's spatial entropy for phenotype $F$ occurring over a tissue area $A$ that is randomly partitioned into $G$ subareas is defined as:

$H_B(F) = \sum_{g=1}^{G} p_g \log(1/\lambda_g)$.

The maximum value of spatial entropy is $\log(A)$ when $F$ occurs with equal intensity ($\lambda_g = 1/A$) over all $G$ subareas partitioning the tissue area of size $A$. The spatial entropy attains a minimum value of $\log(A_{g*})$ when the entire $F$ is concentrated in the smallest subarea of size $A_{g*}$. Since the location and size of such subareas are unknown for the occurrence of an arbitrary phenotype, we randomly partitioned the landscape of the tumor for computing Batty's entropy over different values of $G$ ($G = 2,3 \dots 12$), and repeated the partitioning process ($N = 100$ times for each value of $G$) to output the median $H_B(F)$ as the final measure of spatial heterogeneity of $F$ over $A$.

Heterogeneity of geneset enrichment in a phenotypic context:

Batty's spatial entropy of a variable $X$ can be extended to a co-occurrence based entropy measure defined using a new categorical variable Z that takes values in the form of ordered pairs $(x_i, x_j)$ of $X$ that is considered co-occurrent if their distance is less than or equal to a pre-determined threshold $d$. Given $I$ categories of $X$, there are $I^2$categories of $Z$. As noted by Altieri et al. (2018) (74), an entropy measure based on Z is useful when the variable of interest has multiple categories and the aim is to understand how a spatial context (e.g., a local phenotype's enrichment in a tumor) may influence its neighborhood outcomes (say, a selected molecular pathway's expression). The discretized levels of a given (phenotype, geneset) pair (a realization of $Z$) at the observed locations could be viewed as multicategorical point data and their co-occurrence based Leibovici's spatial entropy (Leibovici et al. 2009) (75) is defined as follows:

$$H_L(Z|d) = \sum_{r=1}^{I^2} p(z_r|d) \log(1/p(z_r|d)).$$

*R libraries*

All statistical analyses were performed in R version 4.0.4. We used the *Seurat* package (76) for data preparation, *bestNormalize* (77) for normalization; *automap* (78) for making a standard grid and applying Ordinary Kriging; *waddR (79)* to detect differentially expressed genes based on the 2-Wasserstein distance test and *SpatEntropy* package (74) for Batty's and Leibovici's spatial entropy calculations. The 3-dimensional and interactive plots were generated with *plot3D* and *plotly* packages (80, 81).

## Results

The present study yields a new computational tool, GATHER, for geostatistical modeling and heterogeneity analysis of molecular landscapes in tumors and tissue samples. The different modules of GATHER are outlined in **Fig. 1**. These include (1) gene-specific expression landscape construction via kriging based geostatistical prediction, (2) estimating a measure of uncertainty associated with the kriging predictions, (3) computing the spatial geneset enrichment score, (4) identifying genes with spatial differential expression at selected phenotypic contexts, (5) identifying genes with spatial differential expression at selected locations, (6) computing Batty's spatial entropy to measure phenotypic heterogeneity, and (7) computing Leibovici's co-occurrence based entropy to quantify the heterogeneity of a selected geneset's enrichment in a given phenotypic context. Furthermore, GATHER provides tools for insightful visualization such as 2D and 3D normalized gene expression landscapes and the corresponding maps of standard errors, spatial enrichment surface of a geneset as well as spatial entropy associated diagrams.

We begin with an illustration of the gene-specific expression landscape construction via kriging based geostatistical prediction. In a past study of 100 breast tumors to understand the complexity of intratumor genetic heterogeneity, driver mutations were observed in several cancer genes (82). For instance, *TBX3*, which encodes for the transcription factor T-box 3 (TBX3), was found to be overexpressed in different types of carcinomas, including breast cancer. TBX3, a mostly cytoplasmic protein in both normal and breast cancer tissues, is significantly overexpressed in the latter, and thus, could serve as a potential diagnostic marker of breast cancer cells (83). Yet, TBX3 localizes differently depending on its role and the cell cycle phase (84). To gain insights into the possible spatial distribution, we used GATHER to construct the expression landscape of *TBX3*, which, along with the corresponding standard errors of the local kriging predictions, are mapped and shown in **Fig. 2**.

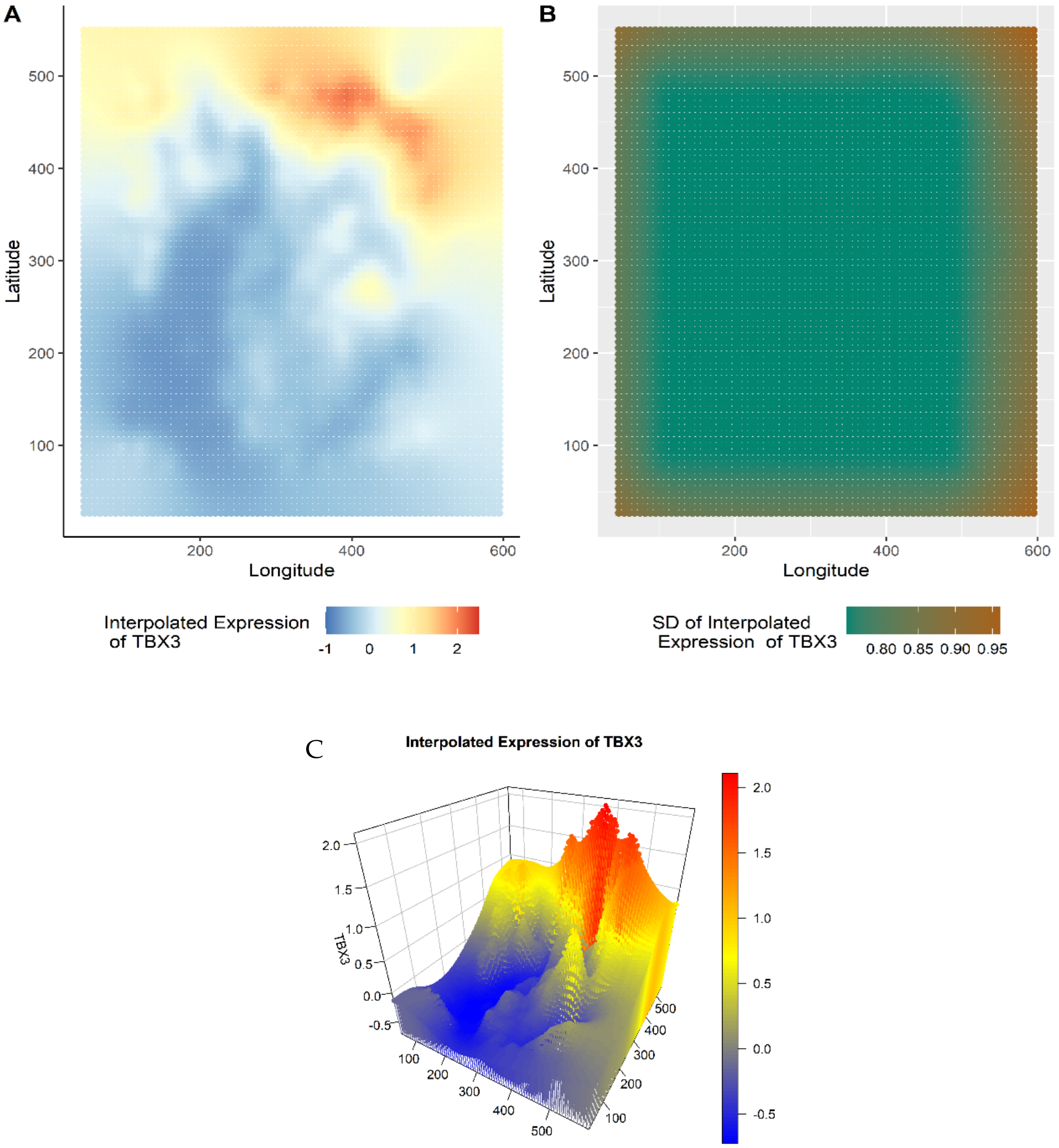


**Fig. 2.** Gene expression landscape generated by geostatistical modeling. Taking the gene *TBX3* as an example, plots (A) and (B) show kriging predicted value $Z$ of gene expression at each point of the tissue space and the associated standard error respectively. Plot (C) is a snapshot of the interactive 3D visualization of the plot (A). The x- and y-dimensions define the tissue space while the z-dimension in plot (C) represents the kriging predicted expression.

Notably, the geostatistical modeling based transcriptomic landscapes could also be viewed using the 3D interactive visualization tool of GATHER (Fig. 2C). Using a grid of evenly spaced points defined on the input tissue space, the $z$-dimension depicts the level of predicted gene expression at each point (x,y) of the synthesized landscape. The interactive 3D visualization tool could be useful for operations such as zooming in to identify and localize regions of phenotypic interest (say, to molecular oncologic pathologists), alignment of the landscapes of different genes for comparing their spatial expression signatures, demarcate those areas that reveal gene expression above (or below) a certain level for focused molecular analysis (e.g., test for specific hallmarks of cancer), and characterize overall intratumor diversity. The interactive version of all the 3D plots are available at the Landscape-Project GitHub webpage [https://mortezahaji.github.io/Landscape-Project/].

GATHER analyzes spatial differential gene expression in single cell transcriptomic data using 2 different approaches. An illustrative example is provided using a selected set of 5 CAF phenotypes, which were represented by the expression of the corresponding marker genes (the respective phenotypes are noted in parentheses): *CXCL12* (CAF-S1), *FBLN1* (mCAFs), *C3* (inflammatory CAFs), *S100A4* (sCAFs), and *COL11A1*, which is a fibroblast-specific "remarkable biomarker" that codes for collagen 11-α1 and shows expression gain in CAFs (85). For details on the CAF markers, see reviews, e.g., (86, 87).

In the first approach, at each point $p$ of the tumor space, GATHER computes the differential expression of each of the above CAF genes between 2 sets of samples drawn from spatial neighborhoods that are (i) near to $p$ versus (ii) distant from $p$ using a semi-parametric 2-sample test for single cell data based on the 2-Wasserstein distance (55). It outputs p-values obtained from the test, which are then adjusted for False Discovery Rate (FDR) by the Benjamini-Hochberg method. This allows GATHER to map the locally significant CAF phenotypes **Fig. 3** shows a 3D snapshot of the differentially expressed CAF genes at each point. For the list of all differentially expressed genes based on the above approach, see **Table 1**.

**Table 1:** Differentially expressed genes in five different CAF phenotypic contexts and their spatial entropy.

| **CAF Marker** | **High CAF**<br>Z≥1<br>N | **Medium CAF**<br>0.5<Z<1<br>N | **Low CAF**<br>Z≤0.5<br>N | **The top 20 most common expressed genes in 100-times permutation at q<0.2 (N=50 random samples for all groups)** | | **Median of Batty's Spatial Entropy** |
|---|---|---|---|---|---|---|
| | | | | **High CAF Vs. Medium CAF** | **Medium CAF Vs. Low CAF** | |
| **COL11A1** | 190 | 3,600 | 535 | MMP11, COL1A2, FN1, DCN, S100A6, CTSK, COL3A1, COL1A1, TIMP3, LUM, SDC1, B2M, S100A4, COL10A1, LGALS1, COL5A2, SERPINF1, SPARC, HLA.A, CTSD | COL1A2, ASPN, DCN, SDC1, LGALS1, COL1A1, SPARC, TAGLN, HTRA3, POSTN, COL5A1, PRSS23, AEBP1, CALD1, ACTA2, COL5A2, PTMS, FN1, COL6A2, FSTL1 | 0.983 |
| **S100A4** | 223 | 3,600 | 502 | LGALS1, S100A6, COL3A1, ACTB, HTRA1, S100A10, TAGLN, COL6A3, CD74, CRABP2, POSTN, TMSB10, HLA.DRB1, PALLD, CLU, SPARC, COL1A1, PTMS, COL6A1, SDC1 | FSTL1, SERPING1, COL3A1, COL6A2, FTL, ISLR, LGALS1, S100A6, SPARC, TAGLN, C1S, CILP, COL1A1, COL6A1, DCN, FLNA, HLA.DPA1, HLA.DPB1, PCOLCE, PTMS | 0.982 |
| **CXCL12** | 141 | 3,553 | 631 | COL6A2, DCN, MMP2, HSPG2, NBL1, SERPING1, SERPINF1, COL6A1, ISLR, | ACTB, ASPN, BGN, CALD1, CILP, COL1A1, COL3A1, | 0.983 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | AEBP1, ASPN, SPARC, LUM, COL5A2, THY1, LRP1, COL1A1, MMP11, COL3A1, RARRES2 | COL5A1, COL6A2, DCN, FLNA, FN1, FSTL1, HTRA3, LGALS1, LUM, S100A6, SDC1, SPARC, TAGLN | |
| **C3** | 206 | 3,501 | 618 | HLA.DRA, FTL, CYBA, HLA.DPB1, APOE, HLA.DPA1, CD74, A2M, RPL13, IFI27, LAPTM5, TYROBP, CTSB, VIM, ACTB, HLA.E, SERPING1, HLA.DRB1, PSAP, TMSB10 | APOE, COL5A1, FSTL1, SPARC, BGN, COL5A2, GPRC5A, PRCP, AP2M1, EDF1, HLA.DPA1, PITX1, ARHGAP1, COL6A1, COL6A2, CYB561, ATP5IF1, CD81, COL1A1, COL1A2 | 0.983 |
| **FBLN1** | 288 | 3,449 | 588 | LUM, COL3A1, COL6A2, FTL, C3, IFI27, COL1A1, COL1A2, MMP2, SERPING1, COL6A1, LRP1, SERPINF1, COL6A3, LGALS1, SPARC, FN1, ACTB, HTRA1, IFITM3 | COL3A1, DCN, SPARC, CILP, COL5A1, FN1, LGALS1, MYL9, ACTB, ASPN, CALD1, COL1A1, COL6A2, MMP11, POSTN, S100A6, TAGLN, TPM4, COL1A2, COL6A1 | 0.982 |

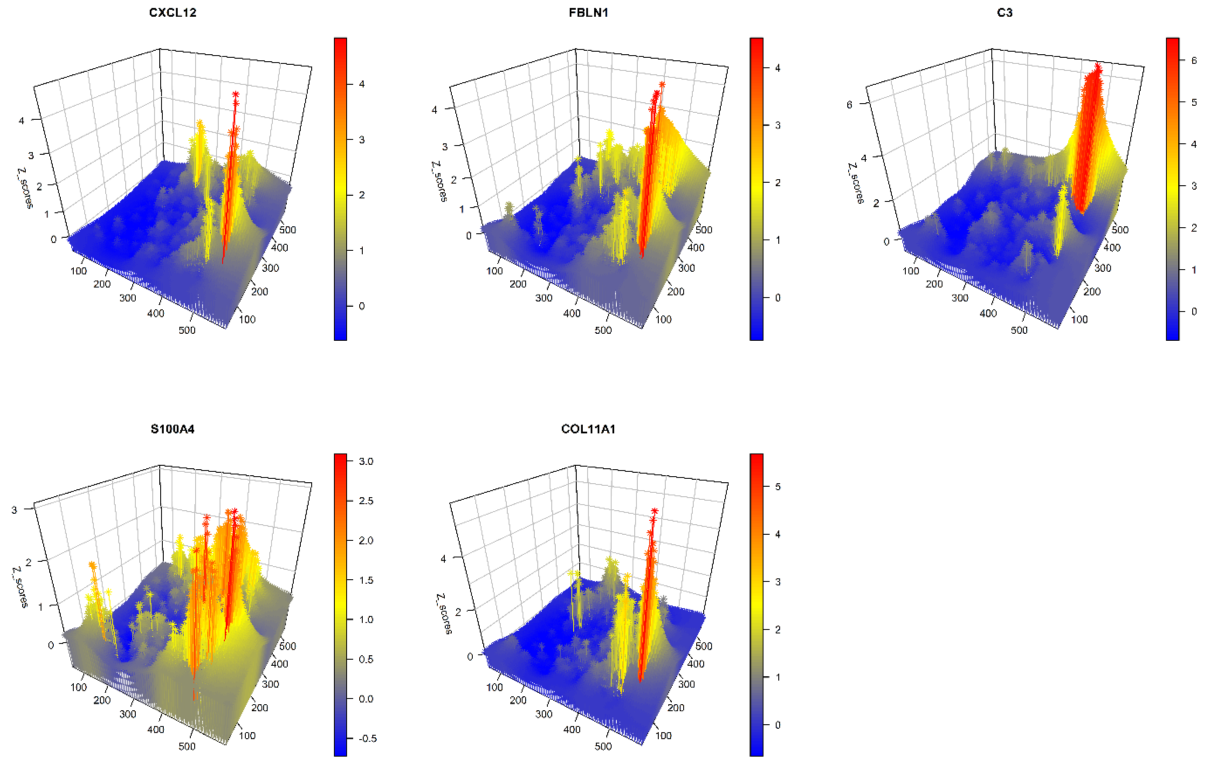


**Fig. 3** A 3D plot of gene-specific continuous transcriptomic landscapes of marker genes of different CAF phenotypes. The name of each CAF gene appears over its plot. The x- and y-dimensions define the tissue space while the z-dimension represents the kriging predicted expression value ($Z$) at each point of the tissue space.

In the second approach, GATHER partitions the tissue space into regions according to different levels of enrichment of a phenotype of interest, and identifies all genes that are expressed differentially across these regions. The regions of the landscape are characterized by $l = 3$ discrete levels of each CAF phenotype: high (CAF $z > 1$), mid (CAF $0.5 < z \leq 1$) and low (CAF $z \leq 0.5$). The levels provide graded spatial contexts in which certain genes

doi:10.20944/preprints202209.0388.v1

may express differentially. Again, we used the 2-Wasserstein distance method to identify the differentially expressed genes across (a) the high CAF versus the medium CAF regions; and (b) the medium CAF versus the low CAF regions. **Table 2** lists the genes that were thus found to be significantly differentially expressed across the spatial levels of CAF phenotypes**.**

**Table 2:** The hallmark genesets of cancer selected for the study.

| | Gene sets | Number of genes in geneset | Overlap with the gene list of the study (%) | Overlap among the 8 hallmark gene sets | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION | 201 | 81 (40%) | - | 2% | <1% | <1% | 0 | 0 | <1% | <1% |
| 2 | HALLMARK_ANGIOGENESIS | 37 | 12 (32%) | 2% | - | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | HALLMARK_ESTROGEN_RESPONSE_EARLY | 201 | 64 (32%) | <1% | 0 | - | 8% | <1% | 0 | <1% | <1% |
| 4 | HALLMARK_ESTROGEN_RESPONSE_LATE | 201 | 62 (31%) | 0 | 0 | 8% | - | 0 | 0 | <1% | 1% |
| 5 | HALLMARK_DNA_REPAIR | 151 | 42 (28%) | 0 | 0 | <1% | 0 | - | 0 | 0 | <1% |
| 6 | HALLMARK_PI3K_AKT_MTOR_SIGNALING | 106 | 28 (26%) | 0 | 0 | 0 | 0 | 0 | - | 0 | <1% |
| 7 | HALLMARK_FATTY_ACID_METABOLISM | 159 | 41 (26%) | <1% | 0 | <1% | <1% | 0 | 0 | - | <1% |
| 8 | HALLMARK_P53_PATHWAY | 201 | 50 (25%) | <1% | 0 | <1% | 1% | <1% | <1% | <1% | - |

Furthermore, we used GATHER to compute the spatial enrichment z-scores $\{SEZ(S,p)|S\epsilon C\}$ for a collection $C$ of hallmark genesets of cancer as shown in **Table 2**. For the given tumor, the landscapes defined by the enrichment scores of each selected hallmark of cancer are described in 3D in **Fig. 4**. In addition, the pointwise dominant hallmark, i.e., the geneset in $C$ having the highest spatial enrichment z-score at any given point, was determined and their distribution is depicted in 3D in **Fig. 5**.

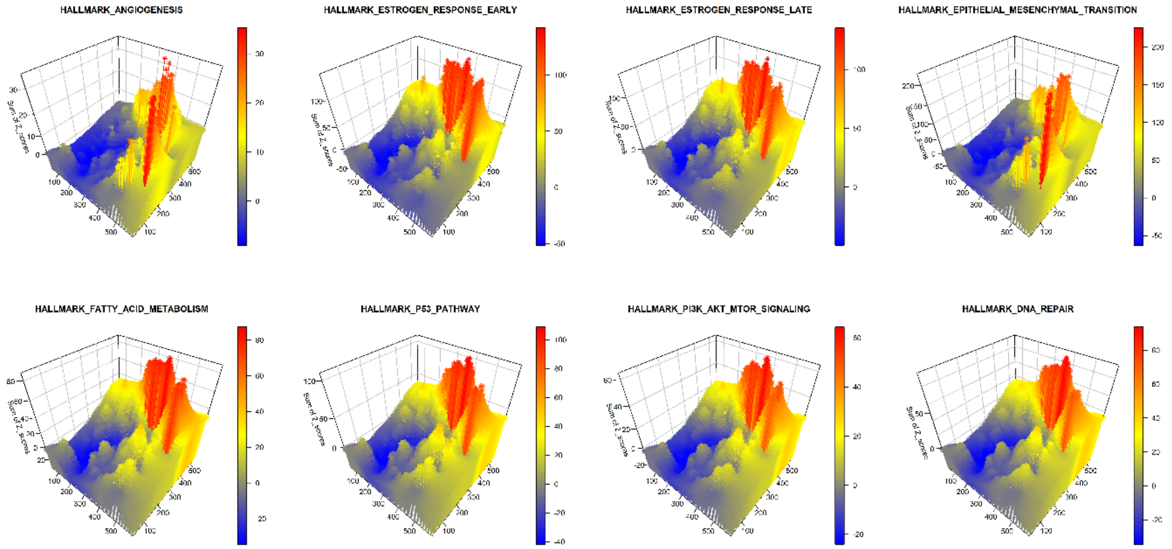


**Fig. 4.** A 3D snapshot of the spatial enrichment z-scores for different hallmark genesets of cancer. The x- and y-dimensions define the tissue space while the z-dimension represents the spatial enrichment z-score ($SEZ$) at a given point. The name of each hallmark geneset appears over its plot.

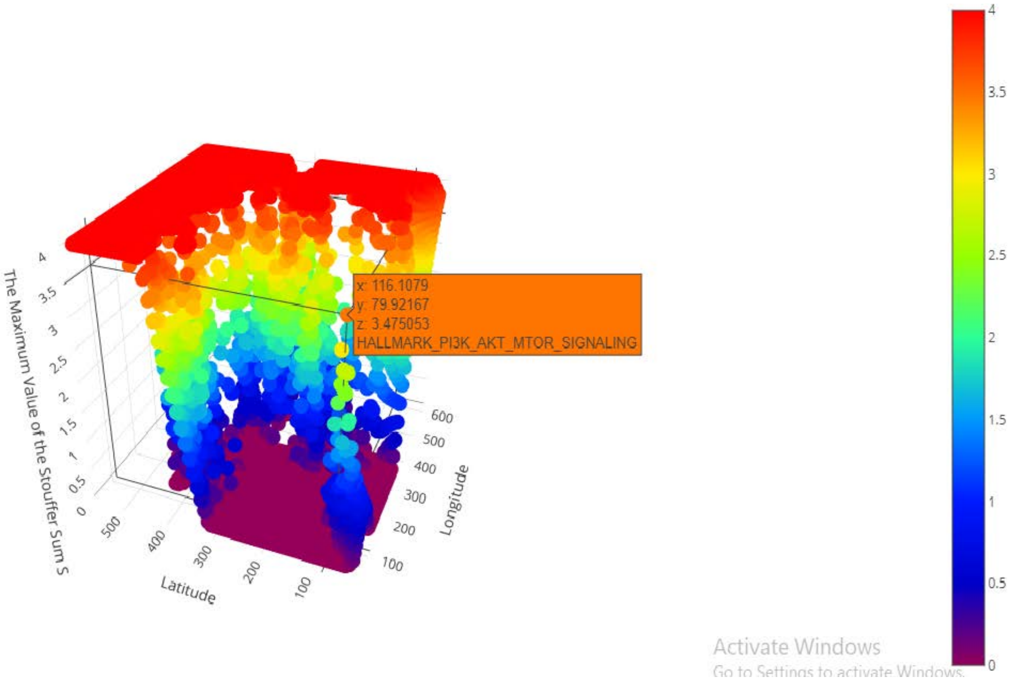


**Fig. 5.** A 3D snapshot of the pointwise dominant hallmark genesets of cancer. The x- and y-dimensions define the tissue space while the z-dimension represents the maximum spatial enrichment z-score ($SEZ$) at a given point among the selected hallmarks. One such point where PI3K_AKT_MTOR hallmark is dominant is shown as an example.

GATHER computes Batty's spatial entropy index with variable partitioning of the tissue space to output a quantitative measure of ITH. The tissue space is randomly partitioned into a fixed number of ($G$) polygons multiple ($N = 100$) times. For each iteration, the spatial entropy is computed, which results in a barplot for each choice of $G$. This is shown in **Fig. 6** for the spatial entropy of the expression of the gene $TBX3$ in the given tumor sample. While the median spatial entropy tends to decrease as the heterogeneity is likely to reduce within smaller polygons generated by higher values of $G$, we select the first value of $G$ for which the median entropy appears to flatten out as the optimal number of partitions. For the present example, the partition into $G^* = 7$ polygons is selected, and thus, GATHER outputs Batty's spatial entropy measure $H_B(TBX3)$ as 0.942.

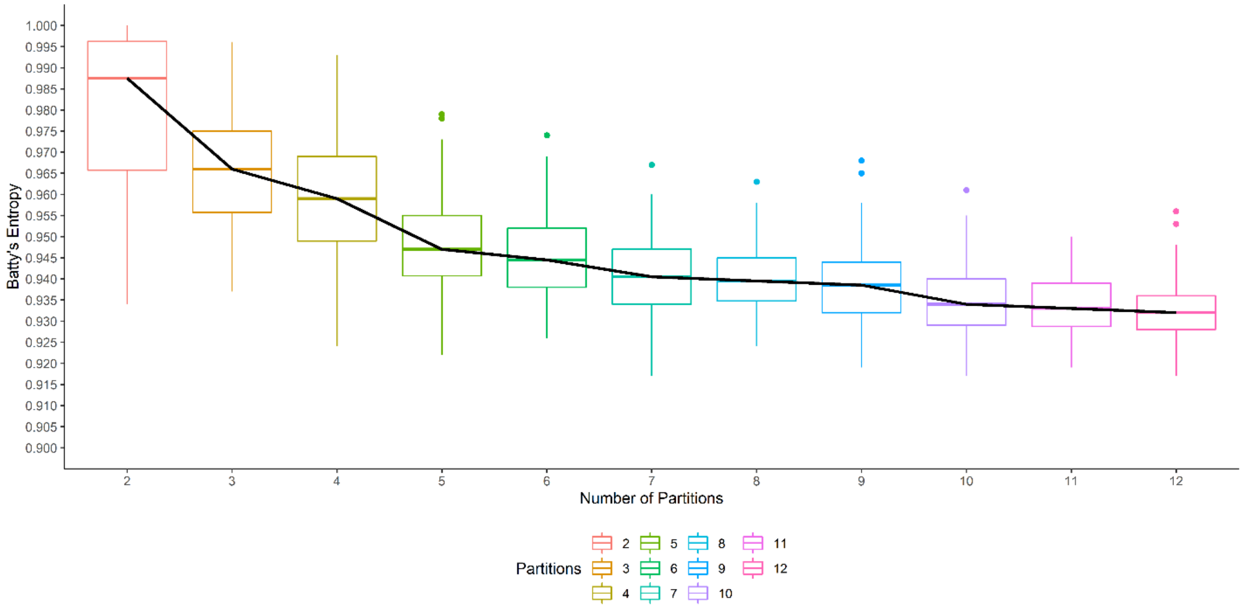


**Fig. 6.** Batty's spatial entropy as a measure of intratumor heterogeneity of gene (*TBX3*) expression. For different number of partitions (x-axis) of the tissue space, $N = 100$ spatial entropy values are calculated (y-axis) and shown with a boxplot. The trend of the median entropy values is shown with a black line.

Importantly, spatial heterogeneity of molecular signatures may be more insightful in the presence of a particular phenotypic context in a given tumor. To capture this with a quantitative measure, GATHER also computes spatial co-occurrence based Leibovici's entropy measure. It allows the user to define phenotypic contexts within which selected genes or genesets may express significant expression. We illustrate this using 6 contexts as defined by 5 CAF phenotypes (as described above) and a 6th context (namely, "None") where none of those phenotypes occur significantly. We test their co-occurrence with the enrichment of the selected hallmarks of cancer.

At each point $p$ of the tissue space, the thresholds for the expression $Z(C, p)$ of the dominant CAF phenotype $C$ as well as $SEZ(G, p)$ of the hallmark geneset $G$ were set at 0.5. Taking combinations of the different CAF marker genes and cancer hallmark genesets, the spatial heterogeneity of their co-occurrence is mapped. At each point, the combination with the most dominant phenotype is depicted. The map in **Fig. 7** uses the following colors to represent the (CAF, hallmark) pairs: red (*FBlN1*, PI3K_AKT_MTOR), blue (*C3*, Angiogenesis), purple (*COL11A1*, PI3K_AKT_MTOR), and grey (no significant CAF phenotype). The regional diversity of co-occurrence is clearly visible, which could be further analyzed by selecting other combinations with the platform's interactive 3D visualization tool.

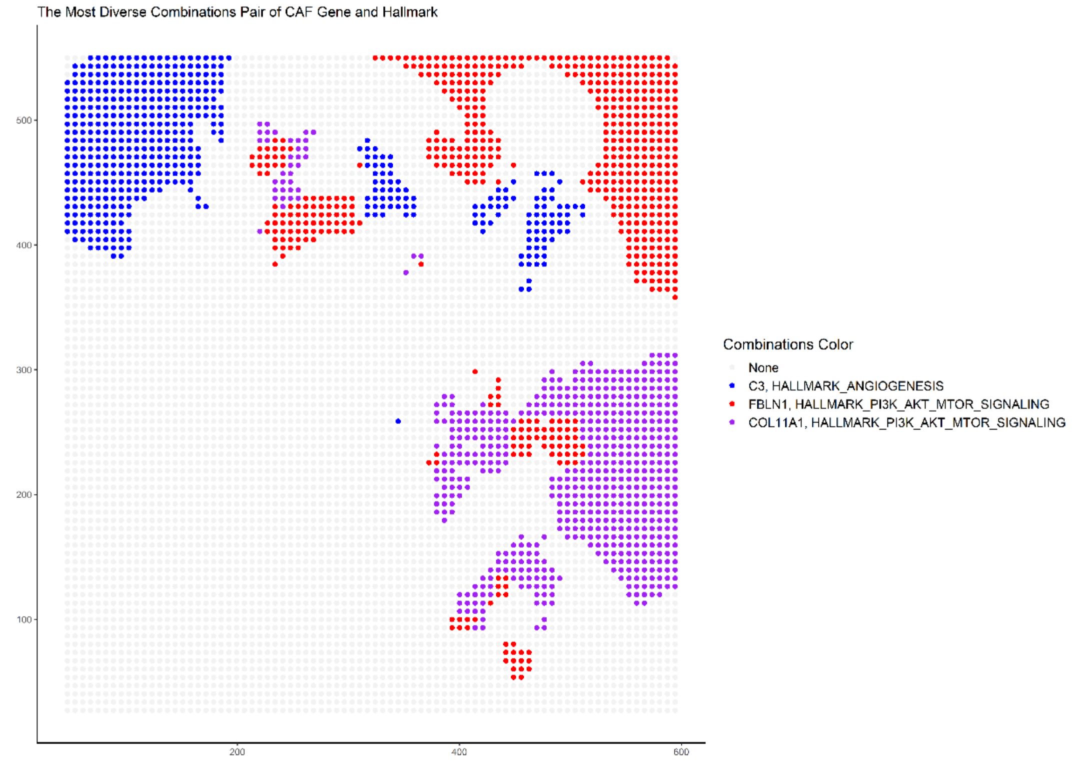


**Fig 7:** The co-occurrence based Leibovici's spatial entropy index. Taking combinations of the different CAF phenotypic contexts and cancer hallmark genesets, the spatial heterogeneity of their co-occurrence is described. At each point, the combination with the most dominant phenotype is depicted. The colors used to represent the (CAF marker, Hallmark geneset) pairs are red (FBlN1, PI3K_AKT_MTOR), blue (C3, Angiogenesis), purple (COL11A1, PI3K_AKT_MTOR), and grey (no significant CAF phenotype).

## Discussion

In the 19th century, Rudolph Virchow, the "father of modern pathology", had observed pleomorphism of cancer cells within tumors. In the 1970s, G.H. Heppner, I.J. Fidler and others showed the existence of distinct subpopulations of cancer cells within tumors, which differed in terms of their tumorigenicity, their resistance to treatment, and their ability to metastasize. ITH has been shown to be associated with poor outcome and decreased response to cancer treatment multiple human cancer types implying a universal role in therapeutic resistance (88, 89). Yet, quantitative assessment of cell-to-cell variation in the expression of a therapeutic target at the protein level is still a challenging task, which partly explains why explicit measures of ITH are not yet commonly used for guiding clinical decisions.

As noted above, GATHER has many practical advantages. The kriging estimates are based on a geostatistical model that allows GATHER to predict the expression value of a gene at any point of the transcriptomic landscape, which allows it to be represented as a surface that is both high-resolution and continuous. Thus, such landscapes can be visualized by an isopleth or contour map. Importantly, GATHER also computes and maps the standard errors of the gene-specific kriging estimates. Further, as the gene-specific landscapes are synthesized over a common grid, they can be aligned easily and systematically combined to produce surfaces that can depict spatial enrichment of genesets or pathways of interest. Moreover, a quantitative measure of error associated with the kriging predictions is available as a spatial measure of quality – and mappable at every location – of the transcriptomic landscape. As yet another advantage, since the kriging prediction at any point is based on every available observation in any given neighborhood, the synthesis of a gene's expression landscape by GATHER is not affected in general by the missing value problem that commonly afflicts single cell RNAseq data.

Invasive and metastatic tumors often have thorough tissue disorganization leading to a microenvironment defined by cellular and paracrine interactions that allow for selection and diversification of certain phenotypes that are not observed otherwise. For instance, blood and lymphatic vasculature in tumors are disorganized with significant functional, spatial, and temporal heterogeneity (90, 91). The resulting variability in nutrients, oxygenation, growth factors, and pH (92) can lead to various abnormal contextual signals that are absent in healthy normal tissues. While spatial phenotypic contexts have been challenging to capture precisely with traditional approaches, high-resolution landscapes constructed by GATHER allow easy demarcation of such regions with the expression levels – above a selected cutoff – of the often well-characterized markers of these contexts.

In the present study, we used different CAF phenotypes as illustrative examples. The significance of such phenotypes could be understood from several experimental models of breast cancer and human tumors that reveal spatial separation of the CAF subtypes attributable to different origins, including the peri-vascular niche, the mammary fat pad and the transformed epithelium. Indeed, not only do the cancer cells and CAFs share location-specific signaling pathways, the gene expression profiles for each CAF subtype indicate distinctive functional programs and hold independent prognostic capability in clinical cohorts by association to metastatic disease. GATHER is able to effectively identify at single cell level the genes with significant differential expression across the diverse spatial contexts as defined by the complex phenotypes that occur in heterogeneous tumor microenvironments.

Notably, an innovative quantitative feature of GATHER is its use of spatial entropy measures to evaluate ITH in a given tumor sample. It computed Batty's entropy to evaluate the distribution of a particular phenotype, as determined by the expression of the corresponding markers, over a given tissue area. Furthermore, as a tissue area could provide the locations for more than one phenotypes or the expressions of multiple markers of a complex phenotype, GATHER also computes a co-occurrence based spatial entropy measure due to Leibovici. The randomization over the tissue space allows the resulting spatial entropy to yield a robust measure of ITH.

Next-generation genetically engineered mouse models can more accurately mimic human cancers (93), new multiplex immunostaining techniques, digital pathology, and specialized computational platforms are able to provide more accurate quantitative assessment of ITH. New approaches such as MIBI (94) and cycIF (95) conduct assays on intact tissue samples thereby maintaining tumor topology and cellular contexts. New computational approaches have also been developed to use next-generation sequencing data to assess ITH and infer clonal evolution of a tumor. Although such techniques could be used for identifying the different subpopulations of cells in a tumor's microenvironment using a "parts-list" approach, it is much harder to clearly dissect the complex phenotypes of tumor cells in terms of their corresponding spatial contexts. Towards this, GATHER provides an efficient solution by substituting the approach of clustering discrete cells each with their stochastically variable gene expressions with constructing continuous transcriptomic landscapes via a long-established geostatistical modeling approach.

We note that our present work has some limitations. The assumption of stationary mean by Ordinary Kriging may not always hold in real data, although the method is known to yield relatively unbiased estimates despite non-stationarity (96). Alternatively, other approaches such as Universal Kriging may be implemented in future work. GATHER does not explicitly group the cell subtypes as clusters like some of the other scRNAseq analysis tools, although the expression landscapes of known markers for different cell subtypes could be used to demarcate the regions that are enriched above a certain threshold and thus yield the cells therein. In our earlier papers, we have developed a Linear Combination Test (LCT) that can rigorously test for enrichment of expression of genes in a pathway against multivariate, continuous phenotypes of samples as opposed to univariate, binary outcomes used for traditional geneset analysis (97-99). Recently, we extended LCT to conduct single cell geneset expression analysis but without using spatial phenotypes (100). In our future work, we will extend LCT to test the enrichment of pathways across complex spatial phenotypes based on the capability of GATHER to analyze tissue heterogeneity in a given sample.

## Acknowledgment

The authors declare no conflict of interest.

## IRB / Ethical Review

Not applicable to our study as it uses publicly available anonymized data for secondary analysis.

## Author Contribution

SP and ID conceived the study. SP, ID, PM, DV, and MH participated in the design and analysis and writing of the manuscript.

## Funding

This research received no funding

## References

1. Marusyk A, Janiszewska M, Polyak K. Intratumor heterogeneity: the rosetta stone of therapy resistance. Cancer cell. 2020;37(4):471-84.
2. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TB, Veeriah S, et al. Tracking the evolution of non–small-cell lung cancer. New England Journal of Medicine. 2017;376(22):2109-21.
3. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013;152(4):714-26.
4. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396-401.
5. Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science. 2014;346(6206):256-9.
6. Jamal-Hanjani M, Quezada SA, Larkin J, Swanton C. Translational implications of tumor heterogeneity. Clinical cancer research. 2015;21(6):1258-66.
7. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. Cell. 2017;168(4):613-28.
8. McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer cell. 2015;27(1):15-26.
9. Janiszewska M, Liu L, Almendro V, Kuang Y, Paweletz C, Sakr RA, et al. In situ single-cell analysis identifies heterogeneity for PIK3CA mutation and HER2 amplification in HER2-positive breast cancer. Nature genetics. 2015;47(10):1212-9.
10. Jenkinson G, Pujadas E, Goutsias J, Feinberg AP. Potential energy landscapes identify the information-theoretic nature of the epigenome. Nature genetics. 2017;49(5):719-29.
11. Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer cell. 2014;26(6):813-25.
12. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. Nature. 2007;447(7143):433-40.
13. Reuben A, Gittelman R, Gao J, Zhang J, Yusko EC, Wu C-J, et al. TCR Repertoire Intratumor Heterogeneity in Localized Lung Adenocarcinomas: An Association with Predicted Neoantigen Heterogeneity and Postsurgical RecurrenceTCR Intratumor Heterogeneity and Relapse in Lung Cancer. Cancer discovery. 2017;7(10):1088-97.
14. Balkwill FR, Capasso M, Hagemann T. The tumor microenvironment at a glance. Journal of cell science. 2012;125(23):5591-6.
15. Whiteside T. The tumor microenvironment and its role in promoting tumor growth. Oncogene. 2008;27(45):5904-12.
16. Kalluri R. The biology and function of fibroblasts in cancer. Nature Reviews Cancer. 2016;16(9):582-98.
17. Pietras K, Östman A. Hallmarks of cancer: interactions with the tumor stroma. Experimental cell research. 2010;316(8):1324-31.
18. Cortez E, Roswall P, Pietras K, editors. Functional subsets of mesenchymal cell types in the tumor microenvironment. Seminars in cancer biology; 2014: Elsevier.
19. Chen X, Song E. Turning foes to friends: targeting cancer-associated fibroblasts. Nature reviews Drug discovery. 2019;18(2):99-115.
20. LeBleu VS, Kalluri R. A peek into cancer-associated fibroblasts: origins, functions and translational impact. Disease models & mechanisms. 2018;11(4):dmm029447.
21. Anderberg C, Pietras K. On the origin of cancer-associated fibroblasts. Taylor & Francis; 2009.
22. Shiga K, Hara M, Nagasaki T, Sato T, Takahashi H, Takeyama H. Cancer-associated fibroblasts: their characteristics and their roles in tumor growth. Cancers. 2015;7(4):2443-58.
23. Sahai E, Astsaturov I, Cukierman E, DeNardo DG, Egeblad M, Evans RM, et al. A framework for advancing our understanding of cancer-associated fibroblasts. Nature Reviews Cancer. 2020;20(3):174-86.
24. Öhlund D, Handly-Santana A, Biffi G, Elyada E, Almeida AS, Ponz-Sarvise M, et al. Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer. Journal of Experimental Medicine. 2017;214(3):579-96.
25. Costa A, Kieffer Y, Scholer-Dahirel A, Pelon F, Bourachot B, Cardon M, et al. Fibroblast heterogeneity and immunosuppressive environment in human breast cancer. Cancer cell. 2018;33(3):463-79. e10.
26. Du H, Che G. Genetic alterations and epigenetic alterations of cancer-associated fibroblasts. Oncology letters. 2017;13(1):3-12.
27. Raz Y, Cohen N, Shani O, Bell RE, Novitskiy SV, Abramovitz L, et al. Bone marrow–derived fibroblasts are a functionally distinct stromal cell population in breast cancer. Journal of Experimental Medicine. 2018;215(12):3075-93.
28. Chang PH, Hwang-Verslues WW, Chang YC, Chen CC, Hsiao M, Jeng YM, et al. Activation of Robo1 signaling of breast cancer cells by Slit2 from stromal fibroblast restrains tumorigenesis via blocking PI3K/Akt/β-catenin pathway. Cancer Res. 2012;72(18):4652-61.
29. Su S, Chen J, Yao H, Liu J, Yu S, Lao L, et al. CD10+GPR77+ Cancer-Associated Fibroblasts Promote Cancer Formation and Chemoresistance by Sustaining Cancer Stemness. Cell. 2018;172(4):841-56.e16.
30. Brechbuhl HM, Finlay-Schultz J, Yamamoto TM, Gillen AE, Cittelly DM, Tan A-C, et al. Fibroblast Subtypes Regulate Responsiveness of Luminal Breast Cancer to Estrogen. Clinical Cancer Research. 2017;23(7):1710.
31. Cuiffo BG, Karnoub AE. Mesenchymal stem cells in tumor development: emerging roles and concepts. Cell adhesion & migration. 2012;6(3):220-30.

32. Junttila MR, De Sauvage FJ. Influence of tumour micro-environment heterogeneity on therapeutic response. Nature. 2013;501(7467):346-54.
33. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl j Med. 2012;366:883-92.
34. Thrane K, Eriksson H, Maaskola J, Hansson J, Lundeberg J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. Cancer research. 2018;78(20):5970-9.
35. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nature medicine. 2015;21(7):751-9.
36. Rye IH, Trinh A, Sætersdal AB, Nebdal D, Lingjærde OC, Almendro V, et al. Intratumor heterogeneity defines treatment-resistant HER 2+ breast tumors. Molecular oncology. 2018;12(11):1838-55.
37. Kalisky T, Oriel S, Bar-Lev TH, Ben-Haim N, Trink A, Wineberg Y, et al. A brief review of single-cell transcriptomic technologies. Briefings in Functional Genomics. 2018;17(1):64-76.
38. Sun G, Li Z, Rong D, Zhang H, Shi X, Yang W, et al. Single-cell RNA sequencing in cancer: Applications, advances, and emerging challenges. Molecular Therapy-Oncolytics. 2021;21:183-206.
39. Bernardo ME, Fibbe WE. Mesenchymal stromal cells: sensors and switchers of inflammation. Cell stem cell. 2013;13(4):392-402.
40. Davidson S, Efremova M, Riedel A, Mahata B, Pramanik J, Huuhtanen J, et al. Single-cell RNA sequencing reveals a dynamic stromal niche that supports tumor growth. Cell reports. 2020;31(7):107628.
41. Dominguez CX, Müller S, Keerthivasan S, Koeppen H, Hung J, Gierke S, et al. Single-cell RNA sequencing reveals stromal evolution into LRRC15+ myofibroblasts as a determinant of patient response to cancer immunotherapy. Cancer discovery. 2020;10(2):232-53.
42. Elyada E, Bolisetty M, Laise P, Flynn WF, Courtois ET, Burkhart RA, et al. Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. Cancer Discov. 2019;9(8):1102-23.
43. Friedman G, Levi-Galibov O, David E, Bornstein C, Giladi A, Dadiani M, et al. Cancer-associated fibroblast compositions change with breast cancer progression linking the ratio of S100A4+ and PDPN+ CAFs to clinical outcome. Nature Cancer. 2020;1(7):692-708.
44. Hosein AN, Huang H, Wang Z, Parmar K, Du W, Huang J, et al. Cellular heterogeneity during mouse pancreatic ductal adenocarcinoma progression at single-cell resolution. JCI insight. 2019;5(16).
45. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med. 2018;24(8):1277-89.
46. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet. 2017;49(5):708-18.
47. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017;171(7):1611-24.e24.
48. Stuart T, Satija R. Integrative single-cell analysis. Nature Reviews Genetics. 2019;20(5):257-72.
49. Marusyk A, Tabassum DP, Janiszewska M, Place AE, Trinh A, Rozhok AI, et al. Spatial Proximity to Fibroblasts Impacts Molecular Features and Therapeutic Sensitivity of Breast Cancer Cells Influencing Clinical OutcomesStromal Fibroblasts and Therapy Resistance. Cancer research. 2016;76(22):6495-506.
50. Eng C-HL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature. 2019;568(7751):235-9.
51. Gillies RJ, Brown JS, Anderson AR, Gatenby RA. Eco-evolutionary causes and consequences of temporal changes in intratumoural blood flow. Nature Reviews Cancer. 2018;18(9):576-85.
52. Lloyd MC, Cunningham JJ, Bui MM, Gillies RJ, Brown JS, Gatenby RA. Darwinian Dynamics of Intratumoral Heterogeneity: Not Solely Random Mutations but Also Variable Environmental Selection ForcesDarwinian Dynamics of Intratumoral Heterogeneity. Cancer research. 2016;76(11):3136-44.
53. Goovaerts P. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology. 2000;228(1):113-29.
54. Altieri L, Cocchi D, Roli G. Spatial entropy for biodiversity and environmental data: The R-package SpatEntropy. Environmental Modelling & Software. 2021;144:105149.
55. Ramdas A, Trillos NG, Cuturi M. On Wasserstein two-sample testing and related families of nonparametric tests. Entropy. 2017;19(2):47.
56. Shannon, CE. (1948). A Mathematical Theory of Communication. Bell System Technical Journal. 27(4): 623–656.
57. Batty M. Entropy in spatial aggregation. Geographical Analysis. 1976;8(1):1-21.
58. MacArthur R. Fluctuations of animal populations and a measure of community stability. ecology. 1955;36(3):533-6.
59. Leibovici DG, Birkin MH. On geocomputational determinants of entropic variations for urban dynamics studies. Geographical Analysis. 2015;47(3):193-218.
60. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome biology. 2020;21(1):1-35.

61. Frieda KL, Linton JM, Hormoz S, Choi J, Chow K-HK, Singer ZS, et al. Synthetic recording and in situ readout of lineage information in single cells. Nature. 2017;541(7635):107-11.
62. McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J. Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science. 2016;353(6298).
63. Shah S, Lubeck E, Zhou W, Cai L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron. 2016;92(2):342-57.
64. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, et al. Highly multiplexed subcellular RNA sequencing in situ. Science. 2014;343(6177):1360-3.
65. 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(6400).
66. Raj B, Wagner DE, McKenna A, Pandey S, Klein AM, Shendure J, et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nature biotechnology. 2018;36(5):442-50.
67. Alemany A, Florescu M, Baron CS, Peterson-Maduro J, Van Oudenaarden A. Whole-organism clone tracing using single-cell sequencing. Nature. 2018;556(7699):108-12.
68. Spanjaard B, Hu B, Mitic N, Olivares-Chauvet P, Janjuha S, Ninov N, et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. Nature biotechnology. 2018;36(5):469-73.
69. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. Nature methods. 2018;15(11):932-5.
70. Peterson RA, Peterson MRA. Package ‘bestNormalize’. Normalizing transformation functions R package version. 2020;1.
71. Haining RP, Haining R. Spatial data analysis: theory and practice: Cambridge university press; 2003.
72. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289-300.
73. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739-40.
74. Altieri L, Cocchi D, Roli G. SpatEntropy: Spatial Entropy Measures in R. arXiv preprint arXiv:180405521. 2018.
75. Leibovici DG, editor Defining spatial entropy from multivariate distributions of co-occurrences. International Conference on Spatial Information Theory; 2009: Springer.
76. Gribov A, Sill M, Lück S, Rücker F, Döhner K, Bullinger L, et al. SEURAT: visual analytics for the integrated analysis of microarray data. BMC medical genomics. 2010;3(1):1-6.
77. Peterson RA, Peterson MRA. Package ‘bestNormalize’. Published online. 2020;27.
78. Hiemstra P, Hiemstra MP. Package ‘automap’. compare. 2013;105:10.
79. Schefzik R, Flesch J, Goncalves A. Fast identification of differential distributions in single-cell RNA-sequencing data with waddR. Bioinformatics. 2021;37(19):3204-11.
80. Soetaert K. plot3D: Tools for plotting 3-D and 2-D data. R package version. 2014:10-2.
81. Sievert C. Interactive web-based data visualization with R, plotly, and shiny: CRC Press; 2020.
82. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012;486(7403):400-4.
83. Aliwaini S, Lubbad AM, Shourfa A, Hamada HA, Ayesh B, Abu Tayem HEM, et al. Overexpression of TBX3 transcription factor as a potential diagnostic marker for breast cancer. Molecular and clinical oncology. 2019;10(1):105-12.
84. Willmer T, Cooper A, Peres J, Omar R, Prince S. The T-Box transcription factor 3 in development and cancer. Bioscience trends. 2017;11(3):254-66.
85. Vázquez-Villa F, García-Ocaña M, Galván JA, García-Martínez J, García-Pravia C, Menéndez-Rodríguez P, et al. COL11A1/(pro) collagen 11A1 expression is a remarkable biomarker of human invasive carcinoma-associated stromal cells and carcinoma progression. Tumor Biology. 2015;36(4):2213-22.
86. Gascard P, Tlsty TD. Carcinoma-associated fibroblasts: orchestrating the composition of malignancy. Genes & development. 2016;30(9):1002-19.
87. Lee YT, Tan YJ, Falasca M, Oon CE. Cancer-associated fibroblasts: epigenetic regulation and therapeutic intervention in breast cancer. Cancers. 2020;12(10):2949.
88. Almendro V, Cheng Y-K, Randles A, Itzkovitz S, Marusyk A, Ametller E, et al. Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. Cell reports. 2014;6(3):514-27.
89. Morris LG, Riaz N, Desrichard A, Şenbabaoğlu Y, Hakimi AA, Makarov V, et al. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. Oncotarget. 2016;7(9):10051.
90. Carmeliet P, Jain RK. Angiogenesis in cancer and other diseases. nature. 2000;407(6801):249-57.
91. Stacker SA, Williams SP, Karnezis T, Shayan R, Fox SB, Achen MG. Lymphangiogenesis and lymphatic vessel remodelling in cancer. Nature Reviews Cancer. 2014;14(3):159-72.
92. Korenchan DE, Flavell RR. Spatiotemporal pH heterogeneity as a promoter of cancer progression and therapeutic resistance. Cancers. 2019;11(7):1026.

doi:10.20944/preprints202209.0388.v1

93. Kersten K, de Visser KE, van Miltenburg MH, Jonkers J. Genetically engineered mouse models in oncology research and cancer medicine. EMBO molecular medicine. 2017;9(2):137-53.
94. Angelova M, Mlecnik B, Vasaturo A, Bindea G, Fredriksen T, Lafontaine L, et al. Evolution of metastases in space and time under immune selection. Cell. 2018;175(3):751-65. e16.
95. Lin J-R, Fallahi-Sichani M, Sorger PK. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. Nature communications. 2015;6(1):1-7.
96. Gotway CA, Wolfinger RD. Spatial prediction of counts and rates. Statistics in Medicine. 2003;22(9):1415-32.
97. Khodayari Moez E, Hajihosseini M, Andrews JL, Dinu I. Longitudinal linear combination test for gene set analysis. BMC Bioinformatics. 2019;20(1):650.
98. Vatanpour S, Pyne S, Leite AP, Dinu I. Gene set analysis and reduction for a continuous phenotype: Identifying markers of birth weight variation based on embryonic stem cells and immunologic signatures. Computers in Biology and Medicine. 2019;113:103389.
99. Wang X, Pyne S, Dinu I. Gene set enrichment analysis for multiple continuous phenotypes. BMC Bioinformatics. 2014;15(1):260.
100. Khodayari Moez E, Hajihosseini M, Andrews JL, Dinu I. Longitudinal linear combination test for gene set analysis. BMC bioinformatics. 2019;20(1):1-19.