Frontiers in Cell and Developmental Biology

TYPE Original Research
PUBLISHED 09 March 2023
DOI 10.3389/fcell.2023.1065586

OPEN ACCESS

EDITED BY
Xin Wang,
The Chinese University of Hong Kong,
China

REVIEWED BY
Haoyun Lei,
Carnegie Mellon University, United States
Zheng Xia,
Oregon Health and Science University,
United States
Canping Chen,
Oregon Health and Science University
Portland, United States, in collaboration
with reviewer ZX

*CORRESPONDENCE
Saumyadipta Pyne,
spyne@ucsb.edu
Irina Dinu,
idinu@ualberta.ca

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

RECEIVED 09 October 2022
ACCEPTED 22 February 2023
PUBLISHED 09 March 2023

CITATION
Amini P, Hajihosseini M, Pyne S and Dinu I (2023), Geographically weighted linear combination test for gene-set analysis of a continuous spatial phenotype as applied to intratumor heterogeneity. *Front. Cell Dev. Biol.* 11:1065586.
doi: 10.3389/fcell.2023.1065586

Geographically weighted linear combination test for gene-set analysis of a continuous spatial phenotype as applied to intratumor heterogeneity

Payam Amini[1,2], Morteza Hajihosseini[3,4], Saumyadipta Pyne[5,6*†] and Irina Dinu[3*†]

[1]Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran, [2]School of Medicine, Keele University, Keele, Staffordshire, United Kingdom, [3]School of Public Health, University of Alberta, Edmonton, AB, Canada, [4]Stanford Department of Urology, Center for Academic Medicine, Palo Alto, CA, United States, [5]Health Analytics Network, Pittsburgh, PA, United States, [6]University of California, Santa Barbara, Santa Barbara, CA, United States

Background: The impact of gene-sets on a spatial phenotype is not necessarily uniform across different locations of cancer tissue. This study introduces a computational platform, GWLCT, for combining gene set analysis with spatial data modeling to provide a new statistical test for location-specific association of phenotypes and molecular pathways in spatial single-cell RNA-seq data collected from an input tumor sample.

Methods: The main advantage of GWLCT consists of an analysis beyond global significance, allowing the association between the gene-set and the phenotype to vary across the tumor space. At each location, the most significant linear combination is found using a geographically weighted shrunken covariance matrix and kernel function. Whether a fixed or adaptive bandwidth is determined based on a cross-validation cross procedure. Our proposed method is compared to the global version of linear combination test (LCT), bulk and random-forest based gene-set enrichment analyses using data created by the Visium Spatial Gene Expression technique on an invasive breast cancer tissue sample, as well as 144 different simulation scenarios.

Results: In an illustrative example, the new geographically weighted linear combination test, GWLCT, identifies the cancer hallmark gene-sets that are significantly associated at each location with the five spatially continuous phenotypic contexts in the tumors defined by different well-known markers of cancer-associated fibroblasts. Scan statistics revealed clustering in the number of significant gene-sets. A spatial heatmap of combined significance over all selected gene-sets is also produced. Extensive simulation studies demonstrate that our proposed approach outperforms other methods in the considered scenarios, especially when the spatial association increases.

Conclusion: Our proposed approach considers the spatial covariance of gene expression to detect the most significant gene-sets affecting a continuous phenotype. It reveals spatially detailed information in tissue space and can thus play a key role in understanding the contextual heterogeneity of cancer cells.

KEYWORDS

intratumor heterogeneity, gene-set analysis, geographically weighted regression, linear combination test, Spatial single cell analysis, cancer-associated fibroblast

Introduction

Globally, there were approximately 2.3 million new cases of and 685,000 deaths due to breast cancer (BC) in 2020 (Lei et al., 2021). BC is also the leading cause of cancer-related deaths among women (Beiki et al., 2012; World Health Organization, 2018) and the second leading cause of cancer deaths globally, worldwide (Beiki et al., 2012). The tumorigenesis involves uncontrolled growth of cells in breast tissues which can be either benign or malignant (Liu et al., 2013). Several studies on breast cancer patients have revealed the different anti- and pro-tumorigenic roles of the CAFs involved (Chang et al., 2012; Brechbuhl et al., 2017; Su et al., 2018).

Extensive studies over the past few decades have uncovered a variety of cell populations in tumors, thus leading to the active research area of intratumor heterogeneity (ITH) (Marusyk et al., 2020). 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. ITH is attributed to genetic, epigenetic, and microenvironmental factors (McGranahan and Swanton, 2017; Marusyk et al., 2020) and associated with poor prognosis, therapeutic resistance and treatment failure leading to poor overall survival in cancer patients (Landau et al., 2013; Patel et al., 2014; Zhang et al., 2014; Jamal-Hanjani et al., 2015; Jamal-Hanjani et al., 2017). Indeed, the persistence of some of the drug-tolerant intratumor cell populations could be attributed to their high phenotypic plasticity (Flavahan et al., 2017).

Interestingly, hierarchies of differentiation also exist among normal cells in healthy tissues, but the populations of tumor cells display far greater cell-to-cell variability and the resulting phenotypic instability (Landau et al., 2014; Jenkinson et al., 2017). Such ITH could be attributed to genetic causes ranging from aneuploidy to other factors such as complex contextual signals in the highly aberrant tumor microenvironments, or even global alterations in cancer cell epigenomes (Senovilla et al., 2012). ITH also involves immune cell infiltration, which is important to immunotherapies. Tumor antigen diversity could be determined by the T Cell clonality in the different regions of the same tumor (Senovilla et al., 2012). Studies have shown spatially complex interactions between tumor microenvironments and the patient's immune system (Hanahan and Weinberg, 2011; Vogelstein and Kinzler, 2015).

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) (Pietras and Östman, 2010; Cortez et al., 2014; Kalluri, 2016). Although the origin and activation mechanism of CAFs remains an area of active research (Anderberg and Pietras, 2009; Shiga et al., 2015; LeBleu and Kalluri, 2018; Chen and Song, 2019), studies have attributed the processes of formation and derivation of CAFs to various precursor cells (Anderberg and Pietras, 2009; Shiga et al., 2015; LeBleu and Kalluri, 2018; Chen and Song, 2019), which may be the source of the well-known heterogeneity among the CAFs (Du and Che, 2017; Öhlund et al., 2017; Costa et al., 2018; Raz et al., 2018; Lee et al., 2020). Indeed, in certain tumors, such as in the breast, in which the prevalence of CAFs could be as high as 80%, they can play both anti-as well as pro-tumorigenic roles (Chang et al., 2012; Brechbuhl et al., 2017; Su et al., 2018). 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 (Cuiffo and Karnoub, 2012; Junttila and De Sauvage, 2013). CAF-targeted treatments can have dual effects depending on the target and the tissue under consideration (Özdemir et al., 2014; Koliaraki et al., 2015; Wagner, 2016). For instance, spatial proximity to CAFs has been shown to impact molecular features and therapeutic sensitivity of breast cancer cells influencing clinical outcomes (Marusyk et al., 2016).

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 (Kalisky et al., 2018; Sun et al., 2021). Unlike previous transcriptomic studies that assayed a "bulk" sample, scRNA-seq data can provide a detailed characterization of each tumor. Indeed, the Human Tumor Atlas Network [https://humantumoratlas.org] is increasingly enriched with data on human cancers based on scRNA-seq assays. The high-resolution transcriptomic platform has led to several scRNA-seq studies of the composition of CAFs in different stages of cancer (Bernardo and Fibbe, 2013; Li et al., 2017; Puram et al., 2017; Lambrechts et al., 2018; Elyada et al., 2019; Hosein et al., 2019; Davidson et al., 2020; Dominguez et al., 2020; Friedman et al., 2020). For focused understanding of the heterogeneous expressions of genes, different sites of the same tumor were analyzed with multiregional RNA sequencing for different cancers (Gerlinger et al., 2012; Zhang et al., 2014; Yates et al., 2015; Thrane et al., 2018).

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. For a tumor sample, the presence of spatial contexts might play a major role which could be combined with scRNA-seq data with the explicit aim to capture microenvironmental heterogeneity. Spatial cell-to-cell communication in a given tissue image can be recovered from a spatial scRNA sequencing data *via* computational spatial re-mapping (Teves and Won, 2020). Alternatively, integration of high-resolution gene expression data with spatial coordinates can resolve such experimental shortcomings (Eng et al., 2019). 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 (Zhang et al., 2014; Eng et al., 2019). Towards this, many recent efforts have developed methods to analyze spatial information in single-cell studies (Lee et al., 2014; McKenna et al., 2016; Shah et al., 2016; Frieda et al., 2017; Alemany et al., 2018; Codeluppi et al., 2018; Raj et al., 2018; Spanjaard et al., 2018; Wang et al., 2018).

High-throughput transcriptomic data are useful not only for identifying genes that are differentially expressed, but also to test for co-regulation of multiple genes, i.e., a gene-set, based on existing empirical knowledge of biological pathways and gene signatures, e.g., the well-known hallmarks of cancer. In this direction, several methods for gene-set analysis (GSA) were introduced by (Goeman et al., 2004; Mansmann and Meister, 2005; Subramanian et al., 2005; Kong et al., 2006; Dinu et al., 2007; Efron and Tibshirani, 2007). Since the genes within such gene-sets share a common biological function, considering the correlations within each set is a key aspect of a useful GSA method. However, it was shown by (Tsai and Chen, 2009) that the above GSA methods were affected by large type II errors.

An important limitation of many GSA methods is that they can only accommodate binary outcomes, such as disease *versus* control. Our method, Linear Combination Test (LCT) is a GSA method that was designed to address these limitations by taking into account correlations across genes and outcomes, and dealing with binary, univariate or multivariate continuous outcomes, measured either at a single point in time or at multiple time points, and therefore, allow us to analyze a wider range of studies involving complex study designs (Wang et al., 2014). Studies have shown that LCT can overcome difficulties such as small sample size, large gene-sets, and can accommodate correlations across gene-sets, time points, and multiple correlated continuous phenotypes (Dinu et al., 2013). Thus, while a specific gene may not show consistent expression across individual cells, LCT is more likely than traditional approaches to detect the regulation of a functional process or biological pathway associated with the intercellular diversity of outcomes in a single cell level experiment.

Recently, we have extended LCT beyond any other "bulk" GSA method for application to single cell experiments (Dinu et al., 2021). However, GSA is considerably more complicated in the presence of spatial information since the analyzed gene-sets need not have a uniform impact over the entire area of a spatially continuous phenotype. In fact, the significance of association between a selected gene-set and a particular phenotypic context at various microenvironmental neighborhoods could be different. Yet, variable as they may be, since spatial effects are generally continuous in nature, proximity may determine more correlated associations than those across distant locations within the same tumor space. Notably, this alludes to Tobler's First Law of Geography, which states that "everything is related to everything else, but near things are more related than distant things." Traditional testing of such relationships involves global or "aspatial" regression, with the implicit assumption that the impact of the genes in a gene-set (covariates) on the phenotype (spatial outcome) is constant across the tumor space (study area). In the presence of ITH, such stationarity assumption is unlikely to be valid. Geographically weighted regression (GWR) is a well-known method (Brunsdon et al., 1996) that avoids this problem by performing the regression within local windows and each observation is weighted according to its proximity to the center of the window. Adaptive kernel bandwidths allow for heterogeneity among densities of gene expression over the windows in different parts of the study region. Local regression coefficients and associated statistics are mapped to visualize how the explanatory power of a gene-set on the associated phenotypes changes spatially.

In the present study, we combined gene-set analysis of LCT with the local spatial modeling of GWR with the aim to develop geographically weighted LCT (GWLCT) as a statistical test. We demonstrated it on spatial scRNA-seq data from a real breast tumor sample and obtained key insights into its molecular heterogeneity across different spatially continuous phenotypic contexts defined by five well-known markers of CAFs. We note that GWLCT has several distinct advantages. While the popular GSA methods are aspatial and use only bulk gene expression data, GWLCT is developed for spatial single cell gene expression data. The geographical weighting scheme allows nearby neighborhoods to contribute more to each local model, and the regions with significant association of a selected gene-set and a corresponding phenotype are detected using scan statistics on the local test scores and illustrated as maps. At each location, the combined significance of such associations for the selected gene-sets is computed and visualized with a spatial heatmap. We also present new 3D interactive tools for insightful visualization of the tumor space. In the next section, we describe the data and methods, followed by the results of real tumor data analysis and simulations of different association scenarios using GWLCT, and end with discussion, including future work.

Materials and methods

Data

Data for spatial transcriptomics were downloaded from the 10x Genomics website (https://www.10xgenomics.com/). In brief, the data were created using the Visium Spatial Gene Expression technique on an invasive breast cancer tissue sample that is expressing the Estrogen Receptor (ER), Progesterone Receptor (PR), and Human Epidermal Growth Factor Receptor (HER) negative. Illumina NovaSeq 6000 was used to generate the RNA sequencing data, which had a sequencing depth of 72,436 mean reads per cell. The downloaded dataset was filtered for average gene expression values greater than 1, and the resulting data matrix had 1,981 rows (genes) and 4,325 columns (single cells). The zero counts were substituted as part of the RNAseq data preparation with a relatively small random jitter about zero that would have the least impact on the remaining gene expression values. Using the bestNormalize package in the R programming language, we used a 10-fold cross-validation based data transformation strategy to normalize each gene's expression across samples (Peterson and Peterson, 2020).

Gene-sets

We downloaded from the Molecular Signatures Database (MSigDB) candidate gene-sets that represent commonly known "hallmarks" of cancer (Liberzon et al., 2011). To ensure their relevance and non-redundancy, we selected 8 hallmark gene-sets with at least 25% overlap with the expressed genes (see above text on

TABLE 1 An evaluation of cancer hallmark gene-sets associated with five continuous phenotypes C3, COL11A1, CXCL12, FBLN1, and S100A4 using the aspatial methods including GSEA, LCT, and RF-GSEA on a single cell breast cancer study.

Method	Phenotype	Gene-set name	Gene-set size	*p*-value	Q-value
GSEA (No phenotype specified)	Not Applicable	EMT	81	0.504	0.648
		ANGIOGENESIS	12	0.227	0.486
		DNA_Rep	41	0.425	0.763
		PI3K	28	0.059	0.155
		FAM	40	0.623	0.648
		P53	50	0.009	0.010
		ERE	64	0.178	0.486
		ERL	62	0.297	0.486
Global LCT	C3, COL11A1, CXCL12, FBLN1, S100A4	EMT	81	<0.001	<0.001
		ANGIOGENESIS	12	<0.001	<0.001
		DNA_Rep	41	<0.001	<0.001
		PI3K	28	<0.001	<0.001
		FAM	40	<0.001	<0.001
		P53	50	<0.001	<0.001
		ERE	64	<0.001	<0.001
		ERL	62	<0.001	<0.001
RF-GSEA	C3, COL11A1, CXCL12, FBLN1, S100A4	EMT	81	<0.001	<0.001
		ANGIOGENESIS	12	<0.001	<0.001
		DNA_Rep	41	<0.001	<0.001
		PI3K	28	<0.001	<0.001
		FAM	40	<0.001	<0.001
		P53	50	<0.001	<0.001
		ERE	64	<0.001	<0.001
		ERL	62	<0.001	<0.001

preprocessing) but mutual gene-set overlap of less than 10%. The selected hallmark gene-sets are: Epithelial Mesenchymal Transition (EMT, size = 81) (Sun et al., 2020), Angiogenesis (size = 12) (Madu et al., 2020), DNA Repair (DNA_Rep, size = 42) (Paluch-Shimon and Evron, 2019), Pi3k AKT MTOR Signaling (Pi3k, size = 28) (Dong et al., 2021), Fatty Acid Metabolism (FAM, size = 41) (Xu et al., 2021), P53 Pathway (P53, size = 50) (Gasco et al., 2002), Estrogen Response Early (ERE, size = 63) (Oshi et al., 2020), and Estrogen Response Late (ERL, size = 62) (Takeshita et al., 2021).

CAF markers

A selected set of five 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 (Vázquez-Villa et al., 2015). For details on the CAF markers, see reviews, e.g., (Gascard and Tlsty, 2016; Lee et al., 2020).

Statistical analysis

Similar to our previously developed GSA methods, GWLCT is motivated by a research gap, more precisely, the need for a statistical method taking into account spatial correlations across genes. The main goal of GSA methods is to efficiently screen large catalogues of *a priori* defined sets of genes sharing common biological functions, easily accessible to GSA users. GSA methods are testing for associations of such sets with a phenotype. To the best of our knowledge, there are no such methods developed for situations where gene measurements at spatial proximity could exhibit higher correlations. What is popularly known as Tobler’s “first law of geography” states that “everything is related to everything else, but near things are more related than distant things.” Based on this fundamental concept, which we borrowed from spatial data analysis, we provide below statistical derivations of an extension of LCT to geographically weighted spatial omic data.

Consider gene expression data of “g” gene variables $(X_1, X_2, X_3, \ldots, X_g)$, “$L$” cells (points) and “$K$” sets of genes at locations given by 2-dimensional Cartesian coordinates. The LCT approach assumes a null hypothesis in which there is no association between a linear combination of $X_1, X_2, X_3, \ldots, X_g$ with the phenotype (Dinu et al., 2013). For a local point (u_l, v_l), we can define a univariate regression as:

$$Y_{(u_l,v_l)} = \alpha_{0\,(u_i,v_i)} + \alpha_{1\,(u_l,v_l)}\beta_1 X_{1\,(u_l,v_l)} + \ldots + +\alpha_{g\,(u_l,v_l)}\beta_g X_{g\,(u_l,v_l)} + \varepsilon_{(u_l,v_l)}$$

where

$$Z(\beta)_{(u_l,v_l)} = \alpha_{0\,(u_l,v_l)} + \alpha_{1\,(u_l,v_l)}\beta_1 X_{1\,(u_l,v_l)} + \ldots + +\alpha_{g\,(u_l,v_l)}\beta_g X_{g\,(u_l,v_l)}$$

is the linear combination of $X_1, X_2, X_3, \ldots, X_g$ and $\varepsilon_{(u_l,v_l)} \sim N(0, \sigma^2)$. For each location in the dataset, we can find the most significant linear combination as follows:

$$\beta^*_{(u_l,v_l)} = Max\left[\rho\left(Y_{(u_l,v_l)}, Z_{(u_l,v_l)}\right)\right] = Max\left[\frac{\beta^T_{(u_l,v_l)} SCov(Y,X)_{(u_l,v_l)} \beta^T_{(u_l,v_l)}}{\beta^T_{(u_l,v_l)} SCov(X,X)_{(u_l,v_l)} \beta^T_{(u_l,v_l)}}\right]$$

where $SCov$ represents the weighted shrunken covariance matrix for each calibration location. The weights are generated using a bisquare kernel function, based on the Euclidean distance between two points l and l'

$$d_{ll'} = \sqrt{\left(Longitude_l - Longitude_{l'}\right)^2 + \left(Latitude_l - Latitude_{l'}\right)^2}$$

and bandwidth h_l, which determines the radius around the point l. Here, the optimal bandwidth is calculated using cross validation (CV) based on the sum of squared errors at each cell point and set of genes:

$$CV = \sum_{k=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{C}\left(Y_{(u_i.v_i).k} - \hat{Y}_{(u_{j\neq i}.v_{j\neq i}).k}\right)^2$$

The bandwidth with the least measure of CV is used for localization. Weighting functions of bisquare and tricube type kernels are used to take the weighted location at l against another location l' into account. The bisquare kernel weighting function is defined as:

$$w_{ll'}(u_i.v_i) = \left\{\left[1 - \left(\frac{d_{ll'}}{h_l}\right)^2\right]^2; d_{ll'} < h_l\, 0 \quad ; d_{ll'} \geq h_l\right.$$

and the tricube kernel weighting function as:

$$w_{ll'}(u_i.v_i) = \left\{\left[1 - \left(\frac{d_{ll'}}{h_l}\right)^3\right]^3; d_{ll'} < h_l\, 0 \quad ; d_{ll'} \geq h_l.\right.$$

For the weighting functions, the bandwidth can be determined either beforehand (fixed distance) or as the distance between the point l and its nearest neighbor (adaptive), which is predetermined as well.

The shrunken covariance matrix of the gene expression data in the l^{th} cell and around the estimated bandwidth (h) can be written as:

$$\boldsymbol{SCov(X.X)}_{(u_l.v_l)}$$

$$= \begin{bmatrix} \frac{1}{L-1}\sum_{l'=1}^{L} w_{ll'}(x_{1ll'} - \bar{x}_1)(x_{1ll'} - \bar{x}_1) & \cdots & \frac{1}{L-1}\sum_{l'=1}^{L} w_{ll'}(x_{1ll'} - \bar{x}_1)\left(x_{gll'} - \bar{x}_g\right) \\ \vdots & \ddots & \vdots \\ \frac{1}{L-1}\sum_{l'=1}^{L} w_{ll'}\left(x_{gll'} - \bar{x}_g\right)(x_{1ll'} - \bar{x}_1) & \cdots & \frac{1}{L-1}\sum_{l'=1}^{L} w_{ll'}\left(x_{gll'} - \bar{x}_g\right)\left(x_{gll'} - \bar{x}_g\right) \end{bmatrix}$$

$$\boldsymbol{SCov(Y.X)}_{(u_l.v_l)}$$

$$= \begin{bmatrix} \left[\frac{1}{L-1}\sum_{l'=1}^{L} w_{ll'}(y_{ll'} - \bar{y})(x_{1ll'} - \bar{x}_1)\right]^2 & \cdots & \frac{1}{(L-1)^2}\sum_{l'=1}^{L} w_{ll'}(y_{ll'} - \bar{y})(x_{1ll'} - \bar{x}_1)\sum_{l'=1}^{L} w_{ll'}(y_{ll'} - \bar{y})\left(x_{gll'} - \bar{x}_g\right) \\ \vdots & \ddots & \vdots \\ \frac{1}{(L-1)^2}\sum_{l'=1}^{L} w_{ll'}(y_{ll'} - \bar{y})\left(x_{gll'} - \bar{x}_g\right)\sum_{l'=1}^{L} w_{ll'}(y_{ll'} - \bar{y})(x_{1ll'} - \bar{x}_1) & \cdots & \left[\frac{1}{L-1}\sum_{l'=1}^{L} w_{ll'}(y_{ll'} - \bar{y})\left(x_{gll'} - \bar{x}_g\right)\right]^2 \end{bmatrix}$$

Using the weighted shrunken covariance matrices, the most significant linear combination at location (u_l, v_l) can be determined as the maximum Eigenvector of $SCov(Y,X)_{(u_l,v_l)} SCov(X,X)_{(u_l,v_l)}{}^{-1}$

Combined significance mapping

We used the Fisher’s sum of log of independent p-values method to calculate the combined significance (CS) of association of the K gene-sets (e.g., 8 hallmarks in the present example) with a fixed phenotype at each location. The sum $X^2 = (-2) \times \sum_{k=1}^{K} \log p_k$ follows a chi-square distribution with $2K$ degrees of freedom, which yields a combined p-value CP corresponding to X^2 at a given location. Thus, the combined significance is computed as $CS = (-1) \times \log CP$ and plotted as a spatial heatmap.

Spatial cluster detection

Based on the count of significant gene-sets as determined by GWLCT at a given location, spatial clusters are detected and mapped for the user. For this purpose, assuming a Poisson distribution of such counts over a grid of points placed on the tumor space, scan statistics are computed with Openshaw’s Geographical Analysis Machine (GAM) (Openshaw et al., 1987) function as implemented in the R package DCluster.

Comparative analysis

A comparative analysis against popular aspatial GSA methods should help the reader understand the relevance of GWLCT extension to the GSA literature. In addition to GWLCT, the global LCT, GSEA, and a Random Forest based GSEA (RF-GSEA) (Chien et al., 2014) were also performed to identify the global gene-sets associated with the outcome. In the domain of regression, the random forest based technique is used when the outcome of concern is a continuous phenotype. The GSEA ignores the continuous phenotype and checks if the gene-sets show statistical difference between biological states. The RF-GSEA combines bootstrap and classification tree to find the proportion of explained variance of a continuous phenotype for a specific gene-set. The small sample size issue has been previously considered in these methods so that variable selection is conducted only from a

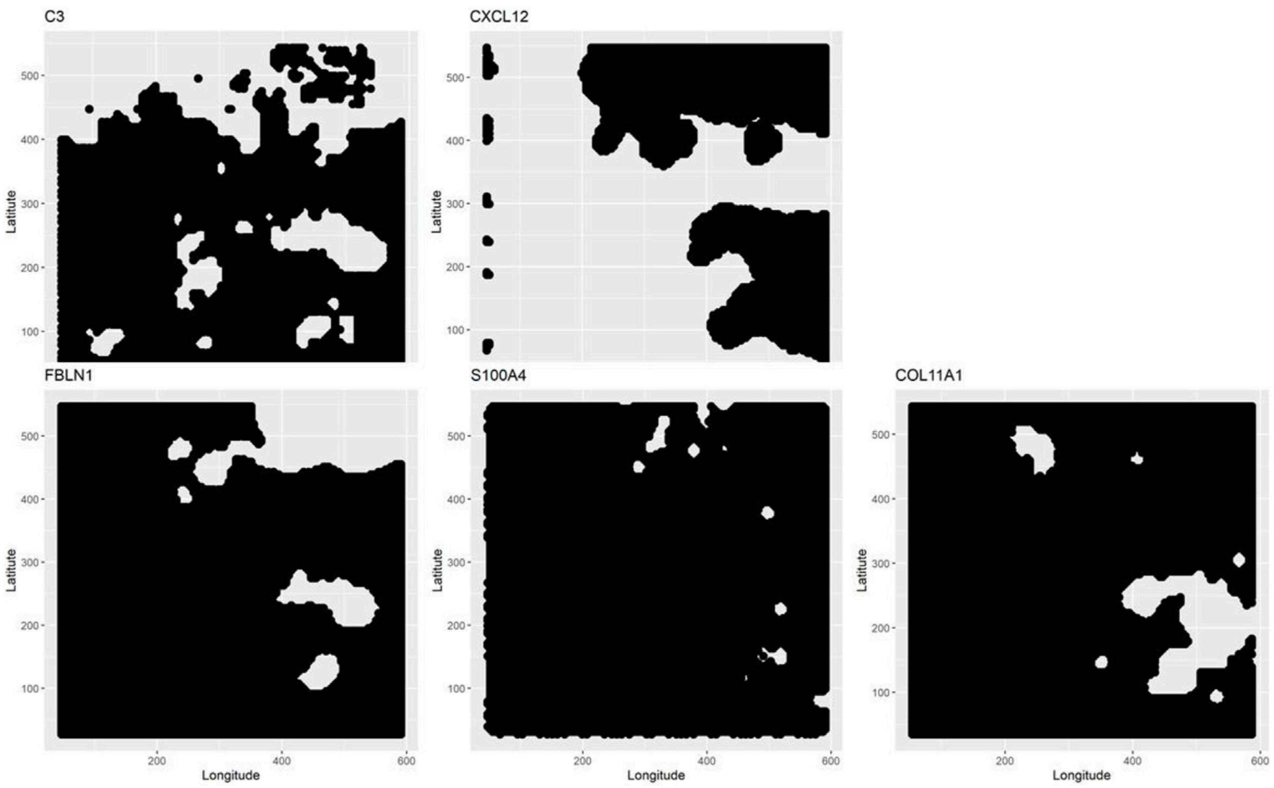

FIGURE 1
Scan statistics are used to detect and demarcate the putative clusters of cells which are enriched with selected gene-sets. The spatial clusters are defined on the tumor space by a scan statistic based on a Poisson distribution. At the locations where the scan statistics exceed a pre-specified threshold, the detected clusters for each phenotype are shown in white.

small random subset of the variables. Moreover, the RF-GSEA is able to accommodate a continuous phenotype when the associations between genesets and phenotypes are non-linear and contain complicated high-order interaction effects (Chien et al., 2014). Next, we give details of the simulation studies, including our approach to generate low and high spatial correlations among genes' expressions. This is a key aspect in observing and understanding advantages of GWLCT over aspatial methods in our study.

Simulation study

Several scenarios were designed to find the impact of simulation components such as bandwidth, number of coordinate points, number of genes, genes spatial association, phenotype-genes spatial association, and gene set probability on the statistical power of the methods. Similar to our previous simulations studies for LCT and its extensions (Wang et al., 2014), we generated a random subset of genes. The first step of the simulation was designed by making assumptions for the spatial distribution of the genes and phenotype. To do so, we used spatial and spatio-temporal geostatistical modeling, prediction and simulation function in R software called "gstat". This function creates an R object with the necessary fields for univariate or multivariate geostatistical prediction, its conditional or unconditional Gaussian, or indicator simulation equivalents (Pebesma, 2004). Values were set for the variogram model components as 10 for the partial sill, 3 for the range parameter, 10 for the nugget, and 30 for the number of nearest observations that are used for the kriging simulation. Moreover, a Gaussian model was assumed for the distribution of the gene expressions and phenotypes.

In the next step, the GWLCT components were defined. For the gene-set matrix, a binomial distribution was used to generate a membership indicator matrix in which the proportions of genes belonging to the gene-sets were characterized using the probability parameter (Low = 0.3, and High = 0.9). Three different values for the spatial covariance were considered to imply the spatial association among genes' expressions. The higher the spatial covariance, the lower the spatial association. Thus, a variance of 50 was considered as high which gives a corresponding low spatial association, a variance of 5 a moderate spatial association, and a variance of 0.1 a high spatial association.

As well, the spatial association between the continuous phenotype and the gene expression data was taken into account by a spatially and normally distributed phenotype generated from the gene expression data with the same parameters as for the spatial association among genes' expressions. High and low levels for the radius/bandwidth around each location for the local analyses were 20 and 6, respectively. The number of coordinate points was 10 by

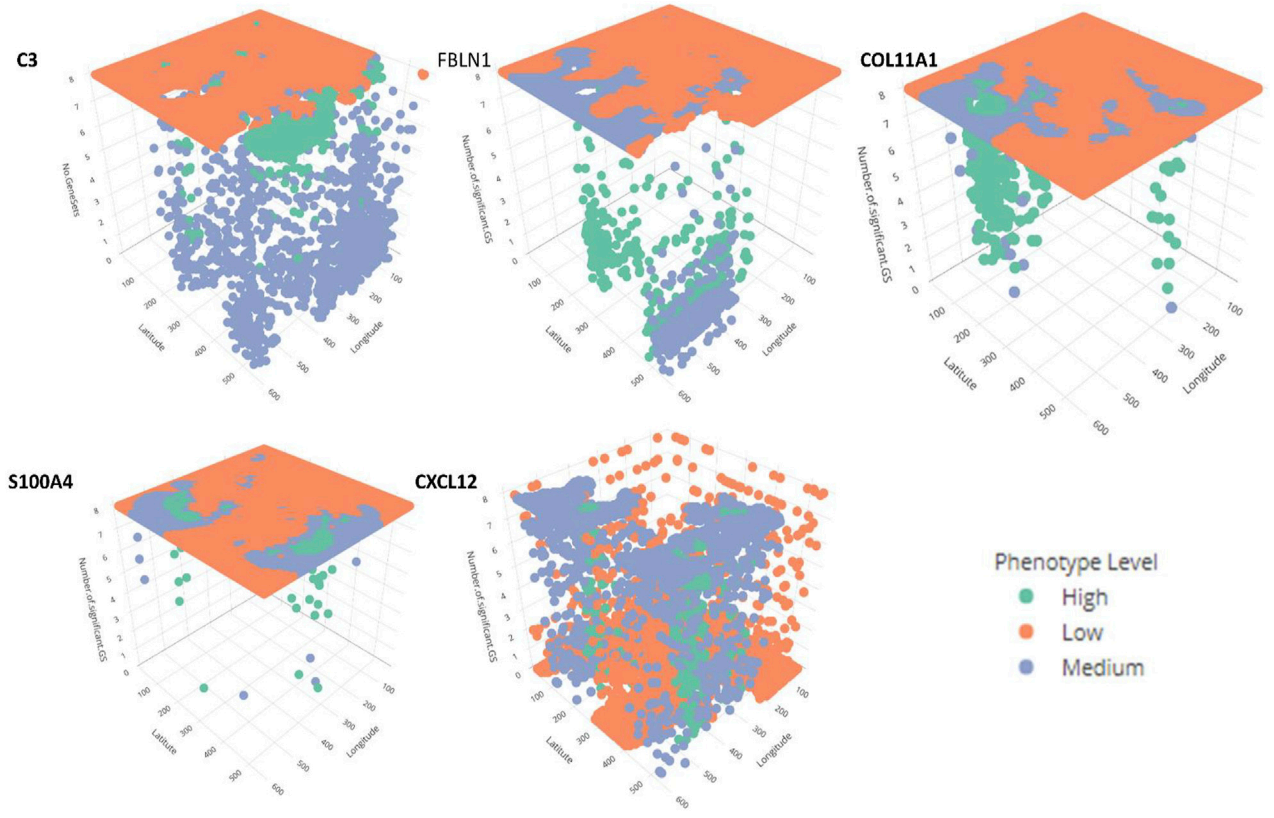

FIGURE 2
The interactive plots (3D) for the 5 phenotypes at Low, Medium and High CAF levels. Further information about the number and the name of significant gene-sets at each location can be obtained by rotating, zooming, and clicking each dot. The green, orange and blue dots reflect high, low, and medium CAF levels. (This figure is an illustrative static snapshot of the interactive 3D plots accessible online at https://mortezahaji.github.io/GWLCT-Project/.)

10 for low, and 100 by 100 for high. Finally, the two levels of low and high were defined for the total number of genes as 100 and 1000 respectively.

The above simulation components resulted in 144 scenarios in which an adaptive bandwidth kernel with the bisquare weighting function was used. The number of permutations was fixed and considered as 500, and the threshold of significance was assumed as 0.05. The methods were compared based on their statistical power. An average of statistical powers across the locations was computed and used as the performance measure for GWLCT.

R programming software version 4.1.1 is used for data analysis and packages such as corpcor (Schafer et al., 2017), qvcalc (Firth and Firth, 2020), stringr (Wickham, 2010), and plotly (Sievert, 2020).

Results

The three aspatial methods, LCT, RF-GSEA, and GSEA, were used to identify the “hallmark” cancer gene-sets that are significantly associated with five spatially continuous CAF phenotypes represented by their known markers C3, COL11A1, CXCL12, FBLN1, and S100A4 in the single cell breast cancer spatial transcriptomic data. The three aspatial methods, the phenotype, the size of each gene-set, *p*-value of the test, and the corresponding q-value are shown in Table 1. We note that such global *p*-values cannot be obtained from our proposed GWLCT, as this method is spatial in nature and assesses significance at every coordinate point rather than an overall significance measure. Using GSEA, the only significant gene-set was P53 (*p*-value = 0.009, q-value = 0.010). The results of LCT and RF-GSEA revealed that all the gene-sets are strongly associated with the five continuous phenotypes with *p*-value and q-value less than 0.001. This was expected since the candidate gene-sets represent commonly known “hallmarks” of cancer (Liberzon et al., 2011). The breast cancer data analysis interpretation resulting from the three aspatial methods considered is limited to the global significance values. This is an important limitation of aspatial methods. For the remainder of this section, we emphasize the advantages of GWLCT results interpretation in the context of spatial data analysis. Since the main advantage of GWLCT consists of an analysis beyond global significance, we study specifically at locations across the tumor space. Scan statistics provide a well-established computational method for detecting spatial clusters based on point count data. We computed scan statistics to detect the putative clusters of spatial regulation based on the number of cells with significantly enriched gene-sets occurring at locations where such counts exceeded what may be expected from an underlying Poisson distribution defined over the tumor space. The clusters are demarcated as white regions

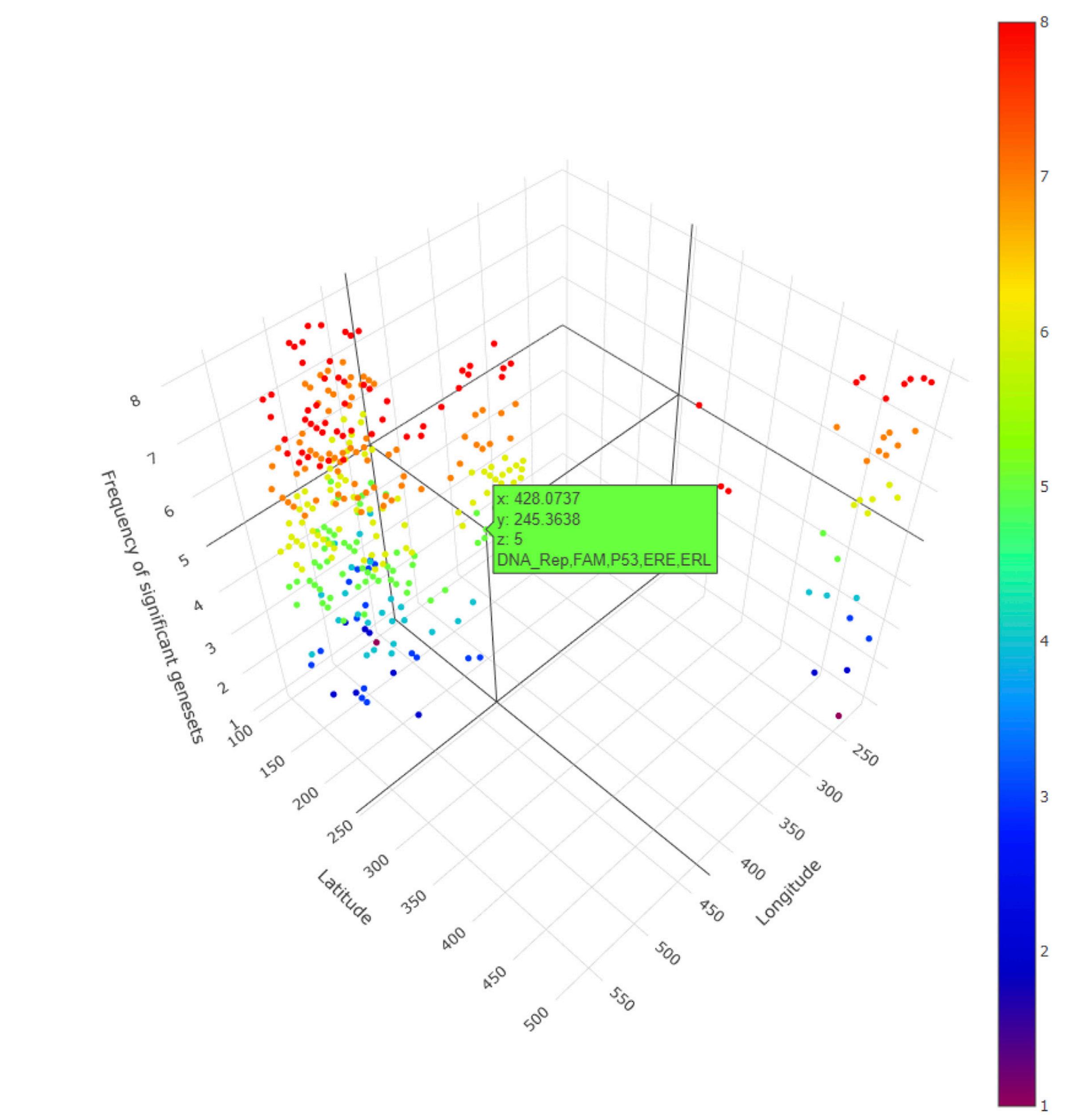

FIGURE 3
A snapshot of the 3D plot for COL11A1 at high CAF level. The number of significant gene-sets are shown in 8 colors (based on 8 gene-sets). The interactive plots can be accessed at: https://mortezahaji.github.io/GWLCT-Project/.

in Figure 1. Notably, the 3D plots, which are instrumental in showcasing the advantages of GWLCT over existing aspatial GSA methods, are presented here as well as at the following links on GitHub: https://mortezahaji.github.io/GWLCT-Project/. By clicking on any point in the plot, one can find the coordinate of the cell, corresponding number and name of significant gene-sets at three different levels: Low, Moderate, and High expressions of the selected CAF marker gene.

In addition, for local GWLCT, three different CAF categories were identified as Low (CAF gene expression less than 0.5), Moderate (CAF gene expression between 0.5 and 1), and High (CAF gene expression exceeding 1). Figure 2 demonstrates a snapshot of the 3D plots for the 5 phenotypes at three CAF levels. A snapshot of the 3D plot for the phenotype COL11A1 at high CAF level is also demonstrated in Figure 3. One is able to detect the frequency as well as the names of significant gene-sets at each location (based on the 8 gene-sets) by clicking on each dot, and rotating and zooming into the 3D interactive plot available at the above-mentioned website. The 3D plot in Figure 3 is divided into eight 2-dimension plots in Figure 4 so that one can evaluate the distribution of one to eight significant gene-sets across the regions with COL11A1 expressed at high CAF level.

Finally, Figure 5 shows the combined significance (CS) heatmap of the 8 hallmarks for COL11A1 phenotype at a high CAF level.

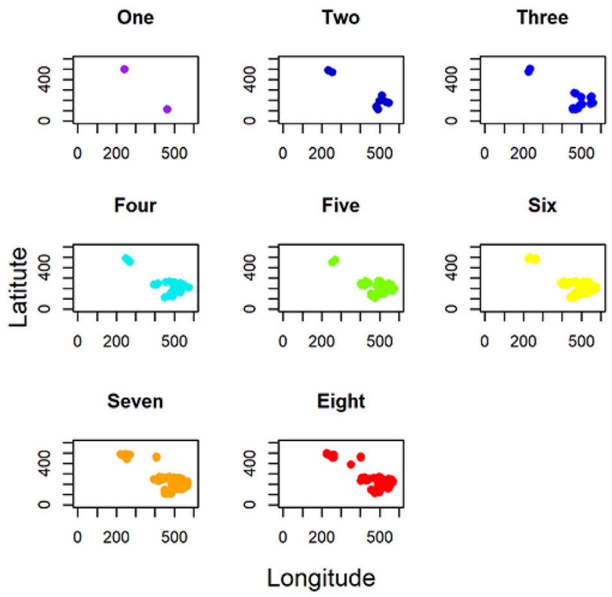

FIGURE 4
2D plots indicating the number of significant gene-sets across the tumor space for COL11A1 at high CAF level.

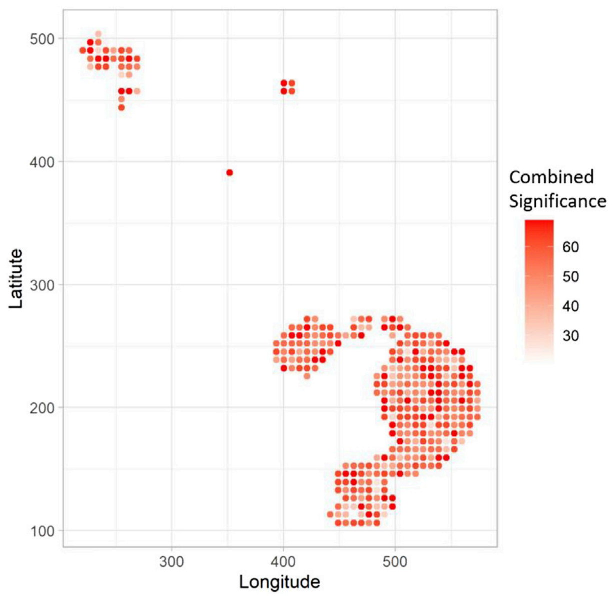

FIGURE 5
The spatial heatmap shows the Combined Significance (CS) of association of the 8 hallmarks with the COL11A1 phenotype at high CAF level. Bigger the value of CS at a location, the corresponding point is shown in darker red.

Higher values of *CS* in Figure 5 represent locations with a combined significant gene-set.

Simulation study results

Supplementary Table S1 and Figure 6 shows the estimated statistical power of the methods using the simulation study. GSEA has the least statistical power among the four methods at all the 144 scenarios. Using 500 iterations in the simulation study, it was revealed that regardless of any parameter in the simulation, GSEA and GWLCT always have the least and most statistical power among the methods, respectively. Overall, we note that bandwidth, number of coordinate points, gene-set size, or total number of genes in each simulation experiment, do not affect the statistical power performance of the methods, a desirable feature shared by sound gene-set analysis methods.

To be particular, by considering GWLCT as the reference category and using Dunnett's test, we compared the average power of GWLCT against Bulk GSEA, LCT, and RF-GSEA by varying from Low to Moderate to High levels of six different parameters, as shown in Figure 6A through F. (A) For the low bandwidth, the mean power of GWLCT was significantly higher than LCT and Bulk ($p < 0.001$). The GSEA and GWLCT were not statistically different in terms of statistical power ($p = 0.534$). For the high bandwidth level GWLCT was not different with the GSEA ($p = 0.709$), and LCT ($p = 0.126$) and different with Bulk ($p < 0.001$). (B) GWLCT performed better than Bulk ($p < 0.001$ for low and high level of number of coordinate points), and LCT ($p = 0.0.021$ for low level of number of coordinate points, and $p = 0.002$ for high level of number of coordinate points). No statistical difference was found between GWLCT and GSEA ($p = 0.541$ for low level of number of coordinate points, and $p = 0.715$ for a high level of number of coordinate points). (C) GWLCT performed better than Bulk ($p < 0.001$ for low and high levels of number of genes), and LCT ($p = 0.004$ for low level of number of genes, and $p = 0.019$ for high level of number of genes) and the same with GSEA ($p = 0.599$ for low level of number of genes, and $p = 0.590$ for high level of number of genes). (D) GWLCT performed better than Bulk and LCT ($p < 0.001$ for low and high levels of gene-set probability (size of gene-sets)). No statistical difference was found between GWLCT and GSEA ($p = 0.397$ for low level of gene-set probability, and $p = 0.820$ for high level of gene-set probability). (E) For the high level of spatial association between genes and continuous phenotype, GWLCT outperformed Bulk ($p < 0.001$) and LCT ($p = 0.005$). No statistical difference was found between GWLCT and GSEA ($p = 0.778$). For the moderate level of spatial association between genes and continuous phenotype, GWLCT outperformed Bulk ($p < 0.001$) and LCT ($p < 0.001$). No statistical difference was found between GWLCT and GSEA ($p = 0.399$). For the low level of spatial association between genes and continuous phenotype, GWLCT outperformed GSEA ($p < 0.001$), Bulk ($p < 0.001$), and LCT ($p < 0.001$). (F) For the high level of spatial association between genes, GWLCT outperformed Bulk ($p < 0.001$). No statistical difference was found between GWLCT with GSEA ($p = 0.849$) and with LCT ($p = 0.523$). For the moderate level of spatial association between genes, GWLCT outperformed Bulk ($p < 0.001$) and LCT ($p = 0.012$). No statistical difference was found between GWLCT and GSEA ($p = 0.768$). For the low level of spatial association between genes, GWLCT outperformed Bulk ($p < 0.001$), and LCT ($p < 0.001$). No statistical difference was found between GWLCT and GSEA ($p = 0.528$).

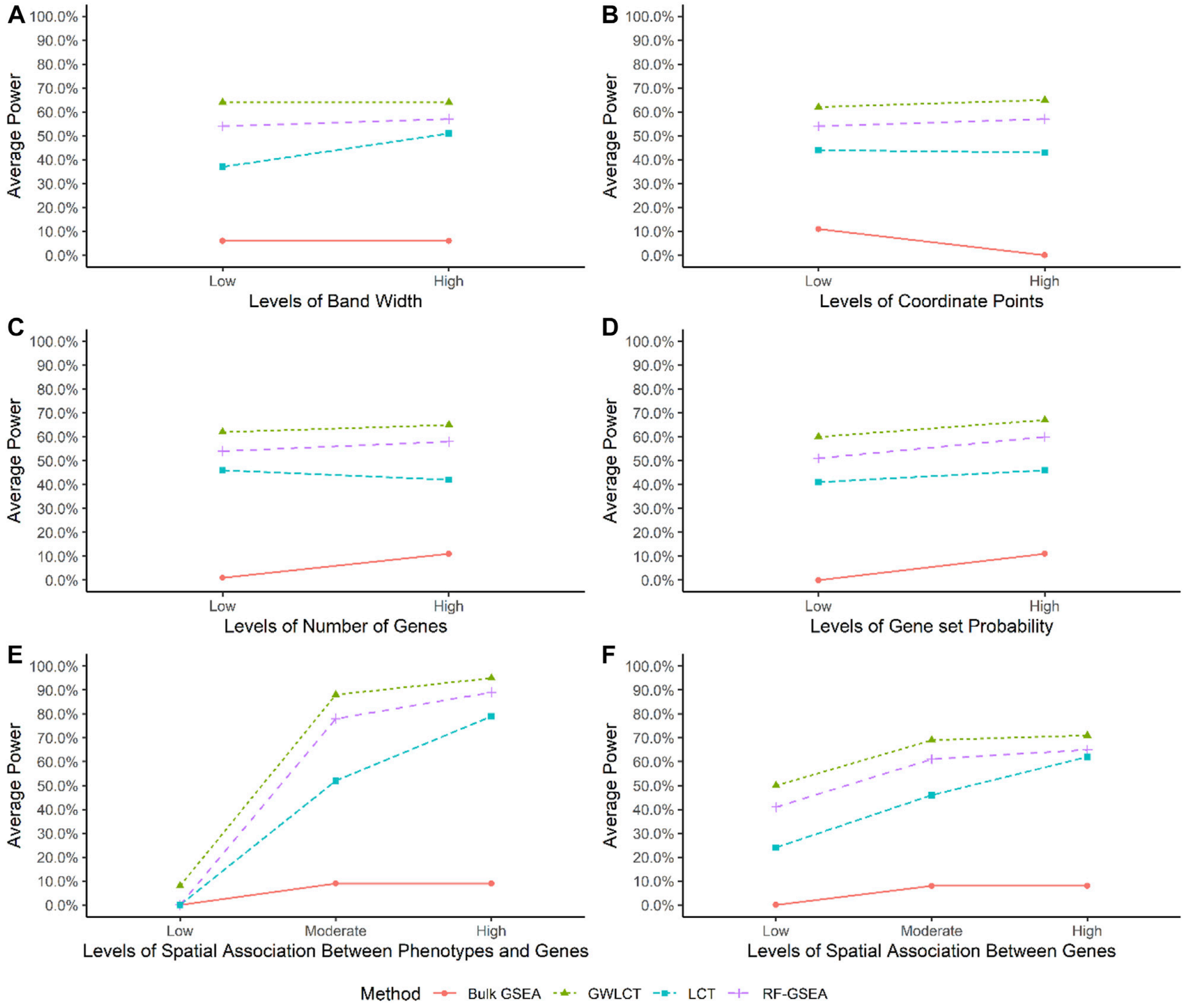

FIGURE 6
Comparing the average statistical power of each method across various levels of simulation parameters. Considering GWLCT as the reference category and using Dunnett's test, we compared the average power of GWLCT (green) against Bulk GSEA (red), LCT (blue) and RF-GSEA (purple) by varying from Low to Moderate to High levels of the following parameters: **(A)** the kernel bandwidth, **(B)** the number of coordinate points, **(C)** the number of genes, **(D)** gene-set probability (size of gene-sets), **(E)** the spatial association between genes and continuous phenotype, and **(F)** the spatial association between genes. For detailed interpretation of the results in terms of the testing significance given by the *p*-values, see text.

The GSEA statistical power is affected by all the simulation variables, however this is mostly due to the zero power of some scenarios. The performance of GSEA is affected by larger number of coordinate points, higher number of genes, higher spatial association among gene expressions, as well as between the phenotype and the gene expressions, and higher number of genes at the gene-sets leads to higher statistical power. Obviously, GSEA is in a lower class of statistical power compared to the other three approaches. Regardless of the scenarios, the results of one-way analysis of variance shows that there are significant differences in the statistical power of LCT, GWLCT, and RF-GSEA (F = 7.842, $p < 0.001$). Tukey's multiple comparison revealed that the difference is due to lower statistical power of LCT compared to the other two methods. Considering a fixed effect for other variables in the simulation, one can find out that LCT has a lower statistical power compared to the other two GWLCT and RF-GSEA when the band width is low. For a study with a low number of coordinate points, the power of GWLCT is significantly higher in comparison to LCT. The almost flat trend of power for the methods also reveals that the performance of the methods is robust against this parameter. As the number of genes increases, the power of LCT reduces significantly compared to GWLCT and RF-GSEA. Moreover, as the amount of spatial association among genes decreases, the power of LCT is significantly lower compared to GWLCT and RF-GSEA. At high levels of genes spatial association, LCT performs reasonably well compared to GWLCT and RF-GSEA, as it is designed to accommodate correlations across genes in a set or biological pathway, *via* a shrinkage correlation matrix. However, at lower levels of gene spatial correlation, GWLCT and RF-GSEA outperform LCT. Moreover, as the amount of spatial association between phenotype and genes increases, the power of GWLCT is significantly higher compared to LCT and RF-GSEA. In addition, the statistical power is robust against the probability of genes belonging to the gene-sets, which is directly related to gene-set size. Although the power increases slightly for all the methods when the size of gene-sets are larger, the increase is not statistically significant between Low and High levels of binomial probability parameter.

Therefore, the most important variables influencing statistical power are the spatial association features. There is a dose response trend of improvement in statistical power for each method, as spatial association among genes increases. The GWLCT outperforms all other methods in all levels of low, medium, and high spatial association among genes. The performance gap narrows down as we move from Low to High levels, indicating higher magnitudes of correlations are easier to be picked up. Interestingly, GWLCT picks up even on subtle spatial associations across genes, exhibiting the largest improvement in statistical power over other methods at Low spatial association levels. The spatial association between phenotype and gene expressions also plays a key role in method performance. When there is a low spatial association between phenotype and genes, GWLCT is still able to detect true significant associations between gene-sets and phenotype, while all other methods have a flat zero statistical power. Similar to the spatial associations across genes, GWLCT picks up on subtle signals for low levels of phenotype-gene expression levels of spatial associations. GWLCT outperforms all other methods at Moderate and High spatial association between phenotype and genes, with the highest statistical power of 0.95, across all simulation scenarios. A significant increase happens when the spatial association between phenotype and genes increases. The highest power for the GWLCT and RF-GSEA can be achieved when both spatial association variables are high. In contrast to GWLCT, the RF-GSEA loses its statistical power when the spatial association among the genes is at Low levels. GWLCT can identify more subtle signals of spatial association, which is an attractive property of the proposed method. More details on the statistical power of the methods in different circumstances can be found in the appendix Supplementary Table S2.

Discussion

Observations of heterogeneity of cell subpopulations in a tumor and the complex interplay of functions involved in the diverse morphological and phenotypic profiles of cancers have a long history. Even in the 19th century, pleomorphism of cancer cells within tumors was observed by the “father of modern pathology”, Rudolph Virchow. More recently, in the 1970s, G.H. Heppner, I.J. Fidler and others showed the existence of distinct subpopulations of cancer cells in tumors, which differed in terms of their tumorigenicity, resistance to treatment, and potential to metastasize. Heppner reviewed the concept of tumor heterogeneity in 1984, and recognized cancers as being composed of multiple subpopulations (Heppner and Miller, 1983), which leads to heterogeneity of cellular morphology, gene expression, metabolism, motility, proliferation, etc (Marusyk et al., 2020). Importantly, ITH has been shown to be associated with poor outcome and decreased response to cancer treatment multiple human cancer types implying a general role in therapeutic resistance (Landau et al., 2013; Patel et al., 2014; Zhang et al., 2014).

The past decade has revealed the immense potential of immunotherapy in cancer. Therapies that promote anti-tumor immune responses have resulted in marked and durable responses in subsets of patients in several cancers (Egen et al., 2020). For instance, abundance of tumor-infiltrating lymphocytes (TILs) and absence of lymphovascular invasion were found to be useful prognostic factors for disease-free survival in patients with HR-/HER2+ breast cancer who were treated using adjuvant trastuzumab (Lee et al., 2015). Spatial transcriptomic approach (Ståhl et al., 2016) was used to identify a type I interferon response overlapping with regions of T Cell and macrophage subset co-localization in HER2+ breast tumors (Andersson et al., 2021). To address the complex interplay between different molecular backgrounds that can characterize ITH with spatial precision, we used GWLCT at a given location in the tissue space to test for the association between a phenotype of interest and different selected gene-sets. The *CS* score to summarize the overall significance of such associations is computed and visualized with a spatial heatmap.

Gene-set analysis (GSA) is a well-established methodological approach in bioinformatics to test for significant regulation of a selected collection of genes across given samples that represent distinct outcomes. At the level of single cells, GSA could be extended to samples that are individual cells which admit to different phenotypes of interest (Dinu et al., 2021). Furthermore, for spatial single cell analysis, such phenotypes would ideally have a spatially correlated and continuous representation. The gene-sets used in GSA are typically curated based on existing experimentally obtained knowledge of genes and their involvements in molecular pathways. In the present study, for illustrative purposes, we selected a collection of 8 gene-sets that represent certain distinctive hallmarks of cancer (Hanahan, 2022). To test for their enrichment in relevant intratumor contexts, we selected 5 different CAF phenotypes of interest since CAFs are well-known for their contribution to heterogeneity and plasticity in the tumor microenvironment (Ping et al., 2021).

The usual methods for GSA involve one of the two major approaches: (a) competitive, which examines if the correlation of a gene-set with the phenotype is the same as the other gene-sets, and (b) self-contained hypothesis, which investigates if the expression of a gene-set changes by the experimental condition. Our LCT method belongs to the former approach which is more likely than traditional methods to detect the regulation of a functional process or biological pathway that is significantly associated with the gene expression results of a given SCA experiment (Dinu et al., 2021). Interestingly, LCT also extends to longitudinal (Khodayari Moez et al., 2019), multivariate and continuous outcomes (Wang et al., 2014), which are capabilities that we built upon here for providing more accurate representation of single cell level stochasticity of the transcriptomic behavior than that of the univariate and discrete class labels typically used in traditional bulk sample studies.

Simulations, along with real omic data analysis, have served as a powerful and effective tool for establishing the performance of new GSA methods. Past studies have thus used simulation for comparative analysis of different criteria of performance of LCT and other major GSA methods. It was found that LCT has type I error and power that are comparable to MANOVA-GSA (Wang et al., 2014) and superior to SAM-GS (Dinu et al., 2007), particularly at higher magnitudes of correlation values across gene-sets (as is commonly noted during GSA). In terms of computational efficiency, LCT outperformed both methods. In another simulation study, LCT also outperformed GSEA (Moez et al., 2018). Along this direction, therefore, in the present study, we conducted a large number of simulations to compare the performance of GWLCT against multiple known GSA techniques based on a variety of well-defined criteria under different experimental assumptions (or scenarios). Interestingly, statistical power did not change with a variation in set size, number of coordination points or bandwidth, or total number of genes in the simulation dataset, for any of the methods considered, which represent desirable properties for sound GSA methods. Larger spatial correlation across gene expression measurements and between genes and phenotype are key aspects of improved statistical power across our simulation experiments.

In the present study, we introduced GWLCT as a new computational platform that presents a fusion of ideas from spatial data analysis (GWR) and bioinformatics (GSA). We understand that the dual modes—both spatial and single-cell—in which GWLCT provides a joint extension to other GSA approaches places it in a unique category thus making it difficult to compare with the existing methods. Yet, we conducted extensive simulation studies which revealed better performance of GWLCT based on several criteria as compared to many known GSA methods that work either on bulk transcriptomics for different scenarios or aspatial version of single-cell transcriptomics. In particular, the use of multiple different kernels and flexible (adaptive) choice of corresponding bandwidths for geographical weighting allows the linear combination test to test for local associations between selected gene-sets and phenotypic contexts within a tumor sample. Thus, GWLCT provides a novel spatial version of gene-set analysis using high-resolution spatial scRNA-seq data. It does have some limitations that will be addressed in our future work. For instance, as it is difficult to determine *a priori* the precise spatial scale at which a gene-set or pathway may be regulated in a given phenotypic context, new multi-scale geographical weighting techniques (Fotheringham et al., 2017) may prove to be useful. We will also extend GWLCT to other omic data as we have previously demonstrated with LCT (Khodayari Moez et al., 2019). As spatial single cell omic platforms become increasingly popular, GWLCT will enrich the ongoing efforts in this rapidly emerging area of research (Zhang et al., 2014; Eng et al., 2019; Hajihosseini et al., 2022). Clinical verification of such new analytical methods will require follow-up studies that must be systematically designed for that specific purpose.

Data availability statement

The de-identified “Spatial Gene Expression Dataset by Space Ranger 1.2.0” used in the present study was obtained from the 10X Genomics website at https://www.10xgenomics.com/resources/datasets/human-breast-cancer-whole-transcriptome-analysis-1-standard-1-2-0 Genomics obtained fresh frozen human Invasive Lobular Carcinoma breast tissue from BioIVT Asterand. It was AJCC/UICC Stage Group I, ER positive, PR positive, HER2 negative. For further details, see the above website.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

PA: Software coding, Data analysis, drafting the manuscript, reviewing the results and approving the final version of the manuscript. ID: study conception and design, interpreting the results, reviewing the results and approving the final version of the manuscript. SP: study conception and design, interpreting the results, reviewing the results and approving the final version of the manuscript. MH: Prepared the data, contributed data analysis, reviewing the results and approving the final version of the manuscript.

Funding

This research was supported by a Fellowship from Mathematics of Information Technology and Complex Systems Accelerate program (grant number: RES0047324, recipient: ID).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2023.1065586/full#supplementary-material

References

Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J., and Van Oudenaarden, A. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature* 556 (7699), 108–112. doi:10.1038/nature25969

Anderberg, C., and Pietras, K. (2009). *On the origin of cancer-associated fibroblasts*. Oxfordshire, United Kingdom: Taylor and Francis.

Andersson, A., Larsson, L., Stenbeck, L., Salmén, F., Ehinger, A., Wu, S. Z., et al. (2021). Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat. Commun.* 12 (1), 6012. doi:10.1038/s41467-021-26271-2

Beiki, O., Hall, P., Ekbom, A., and Moradi, T. (2012). Breast cancer incidence and case fatality among 4.7 million women in relation to social and ethnic background: A population-based cohort study. *Breast Cancer Res.* 14 (1), R5–R13. doi:10.1186/bcr3086

Bernardo, M. E., and Fibbe, W. E. (2013). Mesenchymal stromal cells: Sensors and switchers of inflammation. *Cell stem Cell* 13 (4), 392–402. doi:10.1016/j.stem.2013.09.006

Brechbuhl, H. M., Finlay-Schultz, J., Yamamoto, T. M., Gillen, A. E., Cittelly, D. M., Tan, A-C., et al. (2017). Fibroblast subtypes regulate responsiveness of luminal breast cancer to estrogen. *Clin. Cancer Res.* 23 (7), 1710–1721. doi:10.1158/1078-0432.CCR-15-2851

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geogr. Anal.* 28 (4), 281–298. doi:10.1111/j.1538-4632.1996.tb00936.x

Chang, P. H., Hwang-Verslues, W. W., Chang, Y. C., Chen, C. C., Hsiao, M., Jeng, Y. M., et al. (2012). Activation of Robo1 signaling of breast cancer cells by Slit2 from stromal fibroblast restrains tumorigenesis via blocking PI3K/Akt/β-catenin pathway. *Cancer Res.* 72 (18), 4652–4661. doi:10.1158/0008-5472.CAN-12-0877

Chen, X., and Song, E. (2019). Turning foes to friends: Targeting cancer-associated fibroblasts. *Nat. Rev. Drug Discov.* 18 (2), 99–115. doi:10.1038/s41573-018-0004-1

Chien, C-Y., Chang, C-W., Tsai, C-A., and Chen, J. J. (2014). MAVTgsa: an R package for gene set (enrichment) analysis. *BioMed Res. Int.* 2014, 346074. doi:10.1155/2014/346074

Codeluppi, S., Borm, L. E., Zeisel, A., La Manno, G., van Lunteren, J. A., Svensson, C. I., et al. (2018). Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. methods* 15 (11), 932–935. doi:10.1038/s41592-018-0175-z

Cortez, E., Roswall, P., and Pietras, K. (2014). Functional subsets of mesenchymal cell types in the tumor microenvironment. *Seminars cancer Biol.* 25, 3–9. Elsevier. doi:10.1016/j.semcancer.2013.12.010

Costa, A., Kieffer, Y., Scholer-Dahirel, A., Pelon, F., Bourachot, B., Cardon, M., et al. (2018). Fibroblast heterogeneity and immunosuppressive environment in human breast cancer. *Cancer Cell* 33 (3), 463–479. doi:10.1016/j.ccell.2018.01.011

Cuiffo, B. G., and Karnoub, A. E. (2012). Mesenchymal stem cells in tumor development: Emerging roles and concepts. *Cell adhesion Migr.* 6 (3), 220–230. doi:10.4161/cam.20875

Davidson, S., Efremova, M., Riedel, A., Mahata, B., Pramanik, J., Huuhtanen, J., et al. (2020). Single-cell RNA sequencing reveals a dynamic stromal niche that supports tumor growth. *Cell Rep.* 31 (7), 107628. doi:10.1016/j.celrep.2020.107628

Dinu, I., Moez, E. K., Hajihosseini, M., Leite, A., and Pyne, S. (2021). Use of linear combination test to identify gene signatures of human embryonic development in single cell RNA-seq experiments. *Statistics Appl.* 19 (1), 431–442.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., et al. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinforma.* 8 (1), 242–313. doi:10.1186/1471-2105-8-242

Dinu, I., Wang, X., Kelemen, L. E., Vatanpour, S., and Pyne, S. (2013). Linear combination test for gene set analysis of a continuous phenotype. *BMC Bioinforma.* 14 (1), 212–219. doi:10.1186/1471-2105-14-212

Dominguez, C. X., Müller, S., Keerthivasan, S., Koeppen, H., Hung, J., Gierke, S., et al. (2020). Single-cell RNA sequencing reveals stromal evolution into LRRC15+ myofibroblasts as a determinant of patient response to cancer immunotherapy. *Cancer Discov.* 10 (2), 232–253. doi:10.1158/2159-8290.CD-19-0644

Dong, C., Wu, J., Chen, Y., Nie, J., and Chen, C. (2021). Activation of PI3K/AKT/mTOR pathway causes drug resistance in breast cancer. *Front. Pharmacol.* 12, 628690. doi:10.3389/fphar.2021.628690

Du, H., and Che, G. (2017). Genetic alterations and epigenetic alterations of cancer-associated fibroblasts. *Oncol. Lett.* 13 (1), 3–12. doi:10.3892/ol.2016.5451

Efron, B., and Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann. Appl. statistics* 1 (1), 107–129. doi:10.1214/07-aoas101

Egen, J. G., Ouyang, W., and Wu, L. C. (2020). Human anti-tumor immunity: Insights from immunotherapy clinical trials. *Immunity* 52 (1), 36–54. doi:10.1016/j.immuni.2019.12.010

Elyada, E., Bolisetty, M., Laise, P., Flynn, W. F., Courtois, E. T., Burkhart, R. A., et al. (2019). Cross-species single-cell analysis of pancreatic ductal adenocarcinoma reveals antigen-presenting cancer-associated fibroblasts. *Cancer Discov.* 9 (8), 1102–1123. doi:10.1158/2159-8290.CD-19-0094

Eng, C-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 568 (7751), 235–239. doi:10.1038/s41586-019-1049-y

Firth, D., and Firth, M. D. (2020). *Package 'qvcalc'*.

Flavahan, W. A., Gaskell, E., and Bernstein, B. E. (2017). Epigenetic plasticity and the hallmarks of cancer. *Science* 357 (6348), eaal2380. doi:10.1126/science.aal2380

Fotheringham, A. S., Yang, W., and Kang, W. (2017). Multiscale geographically weighted regression (MGWR). *Ann. Am. Assoc. Geogr.* 107 (6), 1247–1265. doi:10.1080/24694452.2017.1352480

Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K-H. K., Singer, Z. S., et al. (2017). Synthetic recording and *in situ* readout of lineage information in single cells. *Nature* 541 (7635), 107–111. doi:10.1038/nature20777

Friedman, G., Levi-Galibov, O., David, E., Bornstein, C., Giladi, A., Dadiani, M., et al. (2020). Cancer-associated fibroblast compositions change with breast cancer progression linking the ratio of S100A4+ and PDPN+ CAFs to clinical outcome. *Nat. Cancer* 1 (7), 692–708. doi:10.1038/s43018-020-0082-y

Gascard, P., and Tlsty, T. D. (2016). Carcinoma-associated fibroblasts: Orchestrating the composition of malignancy. *Genes and Dev.* 30 (9), 1002–1019. doi:10.1101/gad.279737.116

Gasco, M., Shami, S., and Crook, T. (2002). The p53 pathway in breast cancer. *Breast cancer Res.* 4 (2), 70–76. doi:10.1186/bcr426

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366, 883–892. doi:10.1056/NEJMoa1113205

Goeman, J. J., Van De Geer, S. A., De Kort, F., and Van Houwelingen, H. C. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics* 20 (1), 93–99. doi:10.1093/bioinformatics/btg382

Hajihosseini, M., Amini, P., Voicu, D., Dinu, I., and Pyne, S. (2022). Geostatistical modeling and heterogeneity analysis of tumor molecular landscape. *Cancers (Basel)* 14 (21), 5235. Preprints; 2022090388. doi:10.3390/cancers14215235

Hanahan, D. (2022). Hallmarks of cancer: New dimensions. *Cancer Discov.* 12 (1), 31–46. doi:10.1158/2159-8290.CD-21-1059

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013

Heppner, G. H., and Miller, B. E. (1983). Tumor heterogeneity: Biological implications and therapeutic consequences. *Cancer Metastasis Rev.* 2 (1), 5–23. doi:10.1007/BF00046903

Hosein, A. N., Huang, H., Wang, Z., Parmar, K., Du, W., Huang, J., et al. (2019). Cellular heterogeneity during mouse pancreatic ductal adenocarcinoma progression at single-cell resolution. *JCI insight* 5 (16), e129212. doi:10.1172/jci.insight.129212

Jamal-Hanjani, M., Quezada, S. A., Larkin, J., and Swanton, C. (2015). Translational implications of tumor heterogeneity. *Clin. cancer Res.* 21 (6), 1258–1266. doi:10.1158/1078-0432.CCR-14-1429

Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B., Veeriah, S., et al. (2017). Tracking the evolution of non–small-cell lung cancer. *N. Engl. J. Med.* 376 (22), 2109–2121. doi:10.1056/NEJMoa1616288

Jenkinson, G., Pujadas, E., Goutsias, J., and Feinberg, A. P. (2017). Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nat. Genet.* 49 (5), 719–729. doi:10.1038/ng.3811

Junttila, M. R., and De Sauvage, F. J. (2013). Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* 501 (7467), 346–354. doi:10.1038/nature12626

Kalisky, T., Oriel, S., Bar-Lev, T. H., Ben-Haim, N., Trink, A., Wineberg, Y., et al. (2018). A brief review of single-cell transcriptomic technologies. *Briefings Funct. Genomics* 17 (1), 64–76. doi:10.1093/bfgp/elx019

Kalluri, R. (2016). The biology and function of fibroblasts in cancer. *Nat. Rev. Cancer* 16 (9), 582–598. doi:10.1038/nrc.2016.73

Khodayari Moez, E., Hajihosseini, M., Andrews, J. L., and Dinu, I. (2019). Longitudinal linear combination test for gene set analysis. *BMC Bioinforma.* 20 (1), 650–719. doi:10.1186/s12859-019-3221-7

Koliaraki, V., Pasparakis, M., and Kollias, G. (2015). IKKβ in intestinal mesenchymal cells promotes initiation of colitis-associated cancer. *J. Exp. Med.* 212 (13), 2235–2251. doi:10.1084/jem.20150542

Kong, S. W., Pu, W. T., and Park, P. J. (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics* 22 (19), 2373–2380. doi:10.1093/bioinformatics/btl401

Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., et al. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* 24 (8), 1277–1289. doi:10.1038/s41591-018-0096-5

Landau, D. A., Carter, S. L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M. S., et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* 152 (4), 714–726. doi:10.1016/j.cell.2013.01.019

Landau, D. A., Clement, K., Ziller, M. J., Boyle, P., Fan, J., Gu, H., et al. (2014). Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* 26 (6), 813–825. doi:10.1016/j.ccell.2014.10.012

LeBleu, V. S., and Kalluri, R. (2018). A peek into cancer-associated fibroblasts: Origins, functions and translational impact. *Dis. models Mech.* 11 (4), dmm029447. doi:10.1242/dmm.029447

Lee, H. J., Kim, J. Y., Park, I. A., Song, I. H., Yu, J. H., Ahn, J-H., et al. (2015). Prognostic significance of tumor-infiltrating lymphocytes and the tertiary lymphoid structures in HER2-positive breast cancer treated with adjuvant trastuzumab. *Am. J. Clin. pathology* 144 (2), 278–288. doi:10.1309/AJCPIXUYDVZ0RZ3G

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Yang, J. L., Ferrante, T. C., et al. (2014). Highly multiplexed subcellular RNA sequencing *in situ*. *Science* 343 (6177), 1360–1363. doi:10.1126/science.1250212

Lee, Y. T., Tan, Y. J., Falasca, M., and Oon, C. E. (2020). Cancer-associated fibroblasts: Epigenetic regulation and therapeutic intervention in breast cancer. *Cancers* 12 (10), 2949. doi:10.3390/cancers12102949

Lei, S., Zheng, R., Zhang, S., Wang, S., Chen, R., Sun, K., et al. (2021). Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. *Cancer Commun.* 41 (11), 1183–1194. doi:10.1002/cac2.12207

Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* 49 (5), 708–718. doi:10.1038/ng.3818

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27 (12), 1739–1740. doi:10.1093/bioinformatics/btr260

Liu, X., Wang, L., Zhang, J., Yin, J., and Liu, H. (2013). Global and local structure preservation for feature selection. *IEEE Trans. neural Netw. Learn. Syst.* 25 (6), 1083–1095. doi:10.1109/TNNLS.2013.2287275

Madu, C. O., Wang, S., Madu, C. O., and Lu, Y. (2020). Angiogenesis in breast cancer progression, diagnosis, and treatment. *J. Cancer* 11 (15), 4474–4494. doi:10.7150/jca.44313

Mansmann, U., and Meister, R. (2005). Testing differential gene expression in functional groups. *Methods Inf. Med.* 44 (03), 449–453. doi:10.1055/s-0038-1633992

Marusyk, A., Janiszewska, M., and Polyak, K. (2020). Intratumor heterogeneity: The rosetta stone of therapy resistance. *Cancer Cell* 37 (4), 471–484. doi:10.1016/j.ccell.2020.03.007

Marusyk, A., Tabassum, D. P., Janiszewska, M., Place, A. E., Trinh, A., Rozhok, A. I., et al. (2016). Spatial proximity to fibroblasts impacts molecular features and therapeutic sensitivity of breast cancer cells influencing clinical outcomes. *Cancer Res.* 76 (22), 6495–6506. doi:10.1158/0008-5472.CAN-16-1457

McGranahan, N., and Swanton, C. (2017). Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell* 168 (4), 613–628. doi:10.1016/j.cell.2017.01.018

McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* 353 (6298), aaf7907. doi:10.1126/science.aaf7907

Moez, E. K., Pyne, S., and Dinu, I. (2018). Association between bivariate expression of key oncogenes and metabolic phenotypes of patients with prostate cancer. *Comput. Biol. Med.* 103, 55–63. doi:10.1016/j.compbiomed.2018.09.017

Öhlund, D., Handly-Santana, A., Biffi, G., Elyada, E., Almeida, A. S., Ponz-Sarvise, M., et al. (2017). Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer. *J. Exp. Med.* 214 (3), 579–596. doi:10.1084/jem.20162024

Openshaw, S., Charlton, M., Wymer, C., and Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *Int. J. Geogr. Inf. Syst.* 1 (4), 335–358. doi:10.1080/02693798708927821

Oshi, M., Tokumaru, Y., Angarita, F. A., Yan, L., Matsuyama, R., Endo, I., et al. (2020). Degree of early estrogen response predict survival after endocrine therapy in primary and metastatic ER-positive breast cancer. *Cancers* 12 (12), 3557. doi:10.3390/cancers12123557

Özdemir, B. C., Pentcheva-Hoang, T., Carstens, J. L., Zheng, X., Wu, C-C., Simpson, T. R., et al. (2014). Depletion of carcinoma-associated fibroblasts and fibrosis induces immunosuppression and accelerates pancreas cancer with reduced survival. *Cancer Cell* 25 (6), 719–734. doi:10.1016/j.ccr.2014.04.005

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344 (6190), 1396–1401. doi:10.1126/science.1254257

Paluch-Shimon, S., and Evron, E. (2019). Targeting DNA repair in breast cancer. *Breast* 47, 33–42. doi:10.1016/j.breast.2019.06.007

Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Comput. geosciences* 30 (7), 683–691. doi:10.1016/j.cageo.2004.03.012

Peterson, R. A., and Peterson, M. R. A. (2020). *Package 'bestNormalize'*. Normalizing transformation functions R package version.

Pietras, K., and Östman, A. (2010). Hallmarks of cancer: Interactions with the tumor stroma. *Exp. Cell Res.* 316 (8), 1324–1331. doi:10.1016/j.yexcr.2010.02.045

Ping, Q., Yan, R., Cheng, X., Wang, W., Zhong, Y., Hou, Z., et al. (2021). Cancer-associated fibroblasts: Overview, progress, challenges, and directions. *Cancer gene Ther.* 28 (9), 984–999. doi:10.1038/s41417-021-00318-4

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 171 (7), 1611–1624. doi:10.1016/j.cell.2017.10.044

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., et al. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* 36 (5), 442–450. doi:10.1038/nbt.4103

Raz, Y., Cohen, N., Shani, O., Bell, R. E., Novitskiy, S. V., Abramovitz, L., et al. (2018). Bone marrow–derived fibroblasts are a functionally distinct stromal cell population in breast cancer. *J. Exp. Med.* 215 (12), 3075–3093. doi:10.1084/jem.20180818

Schafer, J., Opgen-Rhein, R., Zuber, V., Ahdesmaki, M., Silva, A. P. D., Strimmer, K., et al. (2017). *Package 'corpcor'*.

Senovilla, L., Vitale, I., Martins, I., Tailler, M., Pailleret, C., Michaud, M., et al. (2012). An immunosurveillance mechanism controls cancer cell ploidy. *Science* 337 (6102), 1678–1684. doi:10.1126/science.1224922

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). *In situ* transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* 92 (2), 342–357. doi:10.1016/j.neuron.2016.10.001

Shiga, K., Hara, M., Nagasaki, T., Sato, T., Takahashi, H., and Takeyama, H. (2015). Cancer-associated fibroblasts: Their characteristics and their roles in tumor growth. *Cancers* 7 (4), 2443–2458. doi:10.3390/cancers7040902

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. Florida, United States: CRC Press.

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., et al. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* 36 (5), 469–473. doi:10.1038/nbt.4124

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353 (6294), 78–82. doi:10.1126/science.aaf2403

Su, S., Chen, J., Yao, H., Liu, J., Yu, S., Lao, L., et al. (2018). CD10+GPR77+ cancer-associated fibroblasts promote cancer formation and chemoresistance by sustaining cancer stemness. *Cell* 172 (4), 841–856. doi:10.1016/j.cell.2018.01.009

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Sun, G., Li, Z., Rong, D., Zhang, H., Shi, X., Yang, W., et al. (2021). Single-cell RNA sequencing in cancer: Applications, advances, and emerging challenges. *Mol. Therapy-Oncolytics* 21, 183–206. doi:10.1016/j.omto.2021.04.001

Sun, X., Wang, M., Wang, M., Yao, L., Li, X., Dong, H., et al. (2020). Exploring the metabolic vulnerabilities of epithelial–mesenchymal transition in breast cancer. *Front. Cell Dev. Biol.* 8, 655. doi:10.3389/fcell.2020.00655

Takeshita, T., Oshino, T., Tokumaru, Y., Oshi, M., Patel, A., Tian, W., et al. (2021). Clinical relevance of estrogen reactivity in the breast cancer microenvironment. *Front. Oncol.* 12, 865024. doi:10.3389/fonc.2022.865024

Teves, J. M., and Won, K. J. (2020). Mapping cellular coordinates through advances in spatial transcriptomics technology. *Mol. Cells* 43 (7), 591–599. doi:10.14348/molcells.2020.0020

Thrane, K., Eriksson, H., Maaskola, J., Hansson, J., and Lundeberg, J. (2018). Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res.* 78 (20), 5970–5979. doi:10.1158/0008-5472.CAN-18-0747

Tsai, C-A., and Chen, J. J. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics* 25 (7), 897–903. doi:10.1093/bioinformatics/btp098

Vázquez-Villa, F., García-Ocaña, M., Galván, J. A., García-Martínez, J., García-Pravia, C., Menéndez-Rodríguez, P., et al. (2015). COL11A1/(pro) collagen 11A1 expression is a remarkable biomarker of human invasive carcinoma-associated stromal cells and carcinoma progression. *Tumor Biol.* 36, 2213–2222. doi:10.1007/s13277-015-3295-4

Vogelstein, B., and Kinzler, K. W. (2015). The path to cancer—Three strikes and you're out. *N. Engl. J. Med.* 373 (20), 1895–1898. doi:10.1056/NEJMp1508811

Wagner, E. F. (2016). Cancer: Fibroblasts for all seasons. *Nature* 530 (7588), 42–43. doi:10.1038/530042a

Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 361 (6400), eaat5691. doi:10.1126/science.aat5691

Wang, X., Pyne, S., and Dinu, I. (2014). Gene set enrichment analysis for multiple continuous phenotypes. *BMC Bioinforma.* 15 (1), 260–269. doi:10.1186/1471-2105-15-260

Wickham, H. (2010). stringr: modern, consistent string processing. *R. J.* 2 (2), 38. doi:10.32614/rj-2010-012

World Health Organization (2018). *Breast cancer: Breast cancer and early diagnosis*. Georgia, United States: American Concer Society.

Xu, S., Chen, T., Dong, L., Li, T., Xue, H., Gao, B., et al. (2021). Fatty acid synthase promotes breast cancer metastasis by mediating changes in fatty acid metabolism. *Oncol. Lett.* 21 (1), 27. doi:10.3892/ol.2020.12288

Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., et al. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* 21 (7), 751–759. doi:10.1038/nm.3886

Zhang, J., Fujimoto, J., Zhang, J., Wedge, D. C., Song, X., Zhang, J., et al. (2014). Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* 346 (6206), 256–259. doi:10.1126/science.1256930