A comparative study of statistical methods for identifying differentially expressed genes in spatial transcriptomics

Yishan Wang[1,2], Chenxuan Zang[1], Ziyi Li[1], Charles C. Guo[3], Dejian Lai[2], Peng Wei[1✉]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.

[2]Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston (UTHealth), Houston, TX 77030, USA.

[3]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.

✉To whom correspondence should be addressed (Email: pwei2@mdanderson.org).

Abstract

Spatial transcriptomics (ST) provides unprecedented insights into gene expression patterns while retaining spatial context, making it a valuable tool for understanding complex tissue architectures, such as those found in cancers. Seurat, by far the most popular tool for analyzing ST data, uses the Wilcoxon rank-sum test by default for differential expression analysis. However, as a nonparametric method that disregards spatial correlations, the Wilcoxon test can lead to inflated false positive rates and misleading findings. This limitation highlights the need for a more robust statistical approach that effectively incorporates spatial correlations. To this end, we propose a Generalized Score Test (GST) in the Generalized Estimating Equations (GEEs) framework as a robust solution for differential gene expression analysis in ST. We conducted a comprehensive comparison of the GST with existing methods, including the Wilcoxon rank-sum test and the GEEs with the robust Wald test. By appropriately accounting for spatial correlations, extensive simulations showed that the GST demonstrated superior Type I error control and comparable power relative to other methods. Applications to ST datasets from breast and prostate cancer showed that the GST-identified differentially expressed genes were enriched in pathways directly implicated in cancer progression. In contrast, the Wilcoxon test-identified genes were enriched in non-cancer pathways and produced substantial false positives, highlighting its limitations for spatially structured data. Our findings suggest that the GST approach is well-suited for ST data, offering more accurate identification of biologically relevant gene expression changes. We have implemented the proposed method in R package "SpatialGEE", available on GitHub.

Keywords: differential expression; GEE; generalized score test; spatial transcriptomics; Wilcoxon rank-sum test; Type I error.

Author Summary

Spatial transcriptomics (ST) provides unprecedented insights into gene expression patterns while retaining spatial context, making it a valuable tool for studying complex tissue architectures and disease etiology. Seurat, a widely used software tool for analyzing ST data, relies on the Wilcoxon rank-sum test for differential expression analysis. However, this test ignores spatial correlations, leading to inflated false positive rates and misleading findings. This limitation highlights the need for a more robust statistical approach that effectively incorporates spatial correlations. To this end, we have proposed a Generalized Score Test (GST) in the Generalized Estimating Equations (GEEs) framework as a robust solution for differential gene expression analysis in ST. By appropriately accounting for spatial correlations, extensive simulations showed that the GST demonstrated superior false positive rate control and comparable power relative to other methods. Applications to ST datasets from breast and prostate cancer showed that GST identified cancer-related genes and pathways more accurately than the Wilcoxon test, which produced misleading results. We have implemented the proposed method in R package "SpatialGEE", available on GitHub.

1. Introduction

Spatial transcriptomics (ST) is an emerging high-throughput technology that can profile genome-wide gene expression across different regions of a tissue by retaining the spatial context, providing crucial information about cellular function and intercellular interactions (Ståhl et al., 2016; Svensson et al., 2018; Erickson et al., 2022; Tian et al., 2023; Jiang et al., 2024; Ma & Zhou, 2024; Shah et al., 2024). A common task in ST data analysis involves identifying differentially expressed (DE) genes across pathological grades (e.g., between ductal carcinoma in situ and invasive carcinoma for breast cancer), which is critical for understanding the complex organization of tissues, as well as both normal development and disease pathology. Among the most used statistical tools applied for this purpose, the Wilcoxon rank-sum test (Wilcoxon, 1945) is often the default choice due to its computational simplicity and ease of implementation from popular software suites such as Seurat (Butler et al., 2018; Stuart et al., 2019; Hao et al., 2024). spatialGE, a recently developed software package, further extends ST data analyses by integrating statistical and spatial models for differential expression analysis (Ospina et al., 2022). In addition to implementing the Wilcoxon rank-sum test, spatialGE employs a linear mixed model (LMM) with an exponential spatial covariance structure to account for spatial correlations in ST gene expression data. Both Wilcoxon test and LMM have been used to detect DE genes in ST data, but no comprehensive comparison has been conducted to evaluate Type I error control, power, numerical stability, or computational cost—highlighting a critical need for systematic evaluation.

Our preliminary analyses of real datasets and simulated data have revealed limitations of the Wilcoxon rank-sum test when applied to the spatially-correlated ST data. Specifically, we

observed that the Wilcoxon rank-sum test tends to have inflated Type I error rates in the presence of spatial correlations, resulting in an increased number of false positives. This issue raises concerns about the validity of power estimates and compromises the credibility of findings derived from spatial transcriptomic studies. Given that most transcriptomic data exhibit inherent spatial dependencies, Wilcoxon rank-sum test's assumption of independence between observations is frequently violated, emphasizing the need for alternative approaches that are more suitable for spatially structured data.

In this study, we considered two potential approaches: generalized linear mixed model (GLMM) and generalized estimating equations (GEE). The GLMM has been considered by many investigators as the gold standard for analysis of correlated data, since they are flexible in modeling complex dependency structures, thus allowing to account for both fixed and random effects (Breslow & Clayton, 1993). However, the GLMM can be computationally challenging in high-dimensional ST due to the large number of parameters involved and zero-inflated count data, which often leads to convergence issues and significant computational demands. To mitigate these computational challenges, we also investigate the GEE, a marginal modeling framework that offers a balance of robustness with computational efficiency (Liang & Zeger, 1986; Zeger et al., 1988). Unlike the GLMM, which models random effects explicitly, the GEEs use a "working" correlation matrix to effectively account for the spatial dependence between observations. More precisely, we first proposed and implemented the generalized score test (GST) (Rao, 1948; Liang & Zeger, 1986) within the GEE framework. We then compared two variations of the GEEs: the commonly used robust Wald test (White, 1980; Huber, 1967) and the GST. The robust Wald test requires fitting the model under the alternative hypothesis and uses a robust sandwich estimator for the standard error to account for correlations within the data. In

contrast, the GST requires only the fitting of the null model, which greatly enhances numerical stability and hence further reduces the computational burden, making it appealing for genome-wide scans.

Through extensive simulation studies, we demonstrated that both the Wilcoxon rank-sum test and the robust Wald test had inflated Type I error rates, whereas the GST exhibited superior Type I error control and comparable power. We excluded the GLMM in our comparison due to its computational intensity and convergence issues when applied to zero-inflated ST data. We then applied both the GST and the Wilcoxon rank-sum test to real datasets from breast and prostate cancer from 10x Genomics (Figure 1a and Figure 1b) to evaluate their effectiveness in detecting DE genes between tumor and normal tissues. The quantile-quantile (QQ) plots revealed significant false positive rate inflation by the Wilcoxon rank-sum test. The biological pathway enrichment analysis showed that major cancer pathways were enriched in the GST-identified gene sets, whereas the Wilcoxon rank-sum test frequently yielded misleading biological associations.

The rest of the paper is organized as follows. Section 2 reviews and introduces statistical methods under comparison for detecting DE genes in ST data, including Wilcoxon rank-sum test, GLMM, and robust Wald test and GST in the GEE framework. Section 3 presents a comprehensive simulation study to compare Type I error rates and power, followed by applications to a breast cancer and a prostate cancer ST dataset. We conclude with a discussion of our main findings and future directions.

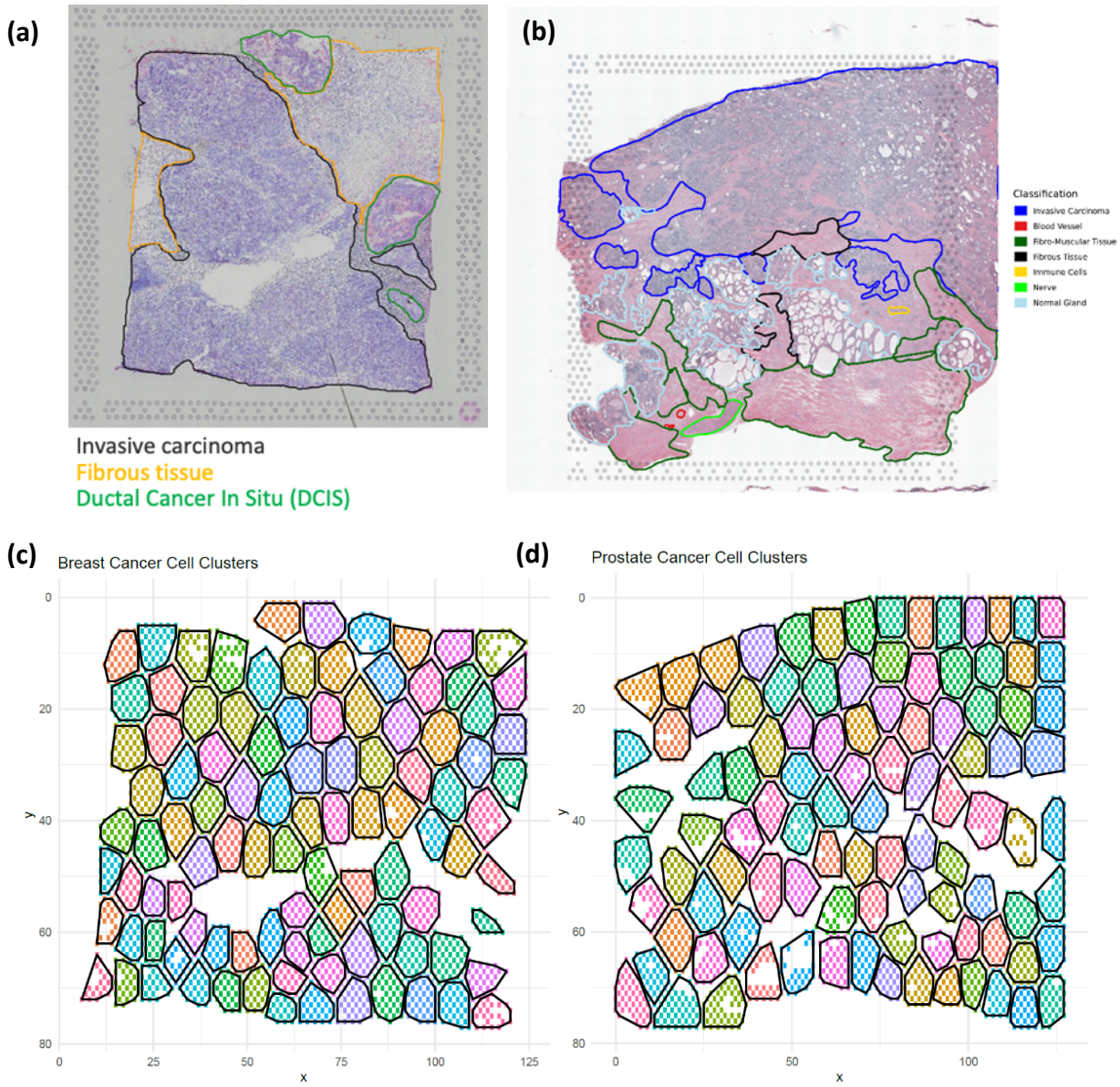

Fig. 1. H&E-stained images and spatial clusters for breast and prostate cancer. **(a)** H&E-stained image with pathology labels for breast cancer (10x Genomics). **(b)** H&E-stained image with pathology labels for prostate cancer (10x Genomics; Zang et al., 2024). **(c)** Spatial clusters (n = 100) for breast cancer. **(d)** Spatial clusters (n = 100) for prostate cancer.

2. Materials and Methods

2.1. Wilcoxon Rank-Sum Test

We begin with the Wilcoxon rank-sum test due to its use as the default method in popular ST data analysis software suites like Seurat. Its simplicity and computational efficiency make it the most common method for detecting DE genes across pathological grades in ST data.

Consider a tissue containing two pathology grades: Grade A and Grade B. Each spatial location i $(i = 1, \dots, n)$ is represented by a spot that includes multiple cells, along with associated gene expression data. Let X and Y represent the expression levels of a particular gene in Grades A and B, respectively. The number of spatial locations in Grade A is denoted by n_X, and the number of spatial locations in Grade B is denoted by n_Y.

The Wilcoxon rank-sum test assesses differential expression between the two grades by ranking all observations, calculating the sum of ranks for each grade, and computing the test statistic based on these sums. The rank of the ith observation when combining and ordering all observations from both grades is represented by R_i. The sum of ranks for Grade A is $W_X = \sum R_{i,X}$, where $R_{i,X}$ is the rank of the ith observation in Grade A. Similarly, the sum of ranks for Grade B is $W_Y = \sum R_{i,Y}$.

To evaluate the difference between the distributions of X and Y, the test statistic $W = min(W_X, W_Y)$ is used. When the sample sizes (n_X and n_Y) are sufficiently large, the distribution of W under the null hypothesis H_0: the distributions of X and $Y$are identical, can be approximated by a normal distribution with mean $\mu_W = \frac{n_X(n_X+n_Y+1)}{2}$ and variance $\sigma_W^2 = \frac{n_X n_Y(n_X+n_Y+1)}{12}$. The standardized test statistic is then calculated as $Z = \frac{W-\mu_W}{\sigma_W}$, and the two-sided p-value is obtained as $p - value = 2 \times \left(1 - \Phi(|Z|)\right)$, where $\Phi(|Z|)$ represents the cumulative distribution function (CDF) of the standard normal distribution for the absolute value of Z.

Despite its common use in practice, the Wilcoxon test assumes independence between observations, an assumption often violated in spatially correlated data. This can lead to inflated Type I error rates, highlighting the need to explore alternative methods that account for spatial dependencies. To control the false discovery rate (FDR), the Benjamini-Hochberg procedure can be applied for multiple testing in genome-wide scans (Benjamini & Hochberg, 1995).

2.2. Generalized Linear Mixed Model (GLMM)

Following the Wilcoxon rank-sum test, we explored the GLMM because it is considered the "gold standard" for analyzing spatially correlated data. Its flexibility in modeling fixed and random effects allows it to account for spatial dependencies, making it initially promising for identifying DE genes across pathology grades.

Let Y_{ij} $(i = 1, \dots, n;\ j = 1, \dots, p)$ represent the gene expression count for gene j at spatial location i. Define X_i as a binary dummy variable for pathology grade at location i (0 for Grade A, 1 for Grade B). The spatial coordinates are denoted by $s_i = (s_{i1}, s_{i2})$ for spatial location i. The model is specified as:

$$log(\mu_{ij}) = \boldsymbol{X_i^T \beta} + \varepsilon_{ij},$$

where μ_{ij} is the expected count for gene j at location i; $\boldsymbol{X_i^T} = [1, X_i]$ is the design matrix for fixed effects; $\boldsymbol{\beta} = [\beta_{j0}, \beta_{j1}]^T$ is the vector of fixed effect coefficients, where β_{j0} is the intercept and β_{j1} is the effect of Grade B compared to Grade A. The random effect ε_{ij} is assumed to follow a normal distribution, $\varepsilon_{ij} \sim \mathcal{N}\left(0, V\left(\sigma_j^2, \kappa_j, \tau\right)\right)$, representing the random effect for gene j at the ith spatial location.

For a given gene j, the spatial covariance matrix $\boldsymbol{V}(\boldsymbol{\sigma}_j^2, \boldsymbol{\kappa}_j, \boldsymbol{\tau})$ is defined based on the distances between pairs of spatial locations, and $\boldsymbol{\sigma}_j^2$ represents a vector of variance components (Li et al., 2009). The (i, i')th element of $\boldsymbol{V}(\boldsymbol{\sigma}_j^2, \boldsymbol{\kappa}_j, \boldsymbol{\tau})$ is given by $V_{ii'}(\sigma_j^2, \kappa_j, \tau) = \sigma_j^2 R(\tau, \kappa_j) = \sigma_j^2 exp(-\tau_{ii'}/\kappa_j)$, where $\tau_{ii'} = ||s_i - s_{i'}||$ denotes the Euclidean distance between two spatial locations i and i'. Here, $\kappa_j > 0$ is a parameter that determines the rate of decay in correlation with distance, with larger values of κ_j indicating stronger correlations and smaller semi-variances $(1 - exp(-\tau/\kappa))$. The exponential spatial structure used here is a specific case of the Matérn correlation structure $R(\tau) = \left(\frac{2\tau\sqrt{v}}{k}\right)^v K_v\left(\frac{2\tau\sqrt{v}}{k}\right)/(\Gamma(v)2^{v-1})$ when $v = 0.5$.

To test for DE genes across the two pathology grades, we test the null hypothesis $H_0: \beta_{j1} = 0$ against the alternative hypothesis $H_a: \beta_{j1} \neq 0$. This tests whether the expected count of gene expression significantly differs between the pathology grades after accounting for spatial correlation. Statistical inference is made using likelihood ratio tests based on the maximum likelihood estimation of the GLMM. For example, to test $H_0: \beta_{j1} = 0$, i.e., gene j is not differentially expressed across Grade A and Grade B, we compare the full model with the null model $log(\mu_{ij}) = \beta_0 + \varepsilon_{ij}$. DE genes are then identified by applying the Benjamini-Hochberg procedure to the p-values to adjust for multiple testing based on the FDR.

2.3. Generalized Estimating Equations (GEE)

Instead of explicitly modeling the spatial correlation structure using random effects, the GEE use a "working" correlation matrix to account for the spatial dependence between observations. We adopted the GEEs model with an independent working correlation structure by dividing the whole ST tissue into m spatial clusters (Figure 1c and Figure 1d) using K-means clustering

(MacQueen, 1967). The mean of Y_{ij}, denoted by μ_{ij}, is linked to the covariates through a log link function, and the model is specified as:

$$log(\mu_{ij}) = \boldsymbol{X_i^T \beta},$$

where $\boldsymbol{X_i^T} = [1, X_i]$ is the design matrix; $\boldsymbol{\beta} = [\beta_{j0}, \beta_{j1}]^T$ is the vector of fixed effect coefficients. The parameters $\boldsymbol{\beta}$ are estimated by solving the GEE:

$$\sum_{i=1}^{m} \boldsymbol{D_i^T V_i^{-1}(Y_i - \mu_i)} = 0,$$

where $\boldsymbol{D_i}$ is the derivative of the mean response with respect to $\boldsymbol{\beta}$ in the cluster i $\left(\boldsymbol{D_i} = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)$; $\boldsymbol{V_i}$ is the variance-covariance matrix of responses in the cluster i, which is a function of the working correlation matrix $\boldsymbol{R_i}$ ($\boldsymbol{V_i} = \boldsymbol{C_i^{1/2} R_i C_i^{1/2}}$); $\boldsymbol{C_i}$ is a diagonal matrix that includes the variances of the individual observations within the cluster i; $\boldsymbol{Y_i}$ is the response vector in cluster i; $\boldsymbol{\mu_i}$ is the mean vector in the cluster i. The robust variance estimates are used to make inferences for the estimated coefficients from the GEE. The robust variance estimate for the estimated coefficients is calculated by the "sandwich" estimator:

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{D_i^T V_i^{-1} D_i}\right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{D_i^T V_i^{-1}(Y_i - \mu_i)(Y_i - \mu_i)^T V_i^{-1} D_i}\right) \left(\sum_{i=1}^{m} \boldsymbol{D_i^T V_i^{-1} D_i}\right)^{-1}.$$

To provide a more efficient model fitting, the GEE uses an independent working correlation matrix $\boldsymbol{R}$ (an $n \times n$ matrix where n is the number of spatial locations). The advantage of the GEE framework is that it produces consistent and asymptotically normal estimates of the parameters even if the working correlation structure is incorrectly specified (Liang & Zeger, 1986).

A special case of the GEE with the robust Wald test is to treat each spatial location as a unique cluster, assuming independence among all spatial locations. This model is referred to as

Independent GEE and is conceptually similar to a two-sample z-test due to the assumption of independent observations.

2.3.1. Robust Wald Test

The robust Wald test requires fitting the full GEEs model under the alternative hypotheses to identify DE genes across the two pathology grades. This approach is the default test implemented in all currently available R packages associated with the GEE framework. To test $H_0: \beta_{j1} = 0$ against $H_a: \beta_{j1} \neq 0$, the robust Wald test statistic W is computed as:

$$W = \frac{\hat{\beta}_{j1}}{\widehat{se}(\hat{\beta}_{j1})},$$

where $\hat{\beta}_{j1}$ is the estimated coefficient and $\widehat{se}(\hat{\beta}_{j1})$ is the robust standard error, which is the square root of the "sandwich" variance estimator. W approximately follows a standard normal distribution under the null hypothesis.

2.3.2. Generalized Score Test (GST)

We propose the generalized score test as an alternative to the robust Wald test, sharing the same asymptotic properties but requiring only model fitting under the null hypothesis, making it more numerically stable. The asymptomatic equivalence between the GST statistic and the robust Wald statistic holds only as the number of spatial clusters is large (or approaches infinity in theory). However, with a finite number of clusters, as observed in real ST data, these tests can yield different Type I error and power results (Boos & Stefanski, 2013), supporting the need for more numerically robust methods in practical settings.

The score function $U(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})$ is the derivative of the quasi-loglikelihood with respect to the parameters $\boldsymbol{\beta}$, evaluated at the estimated parameters under the null hypothesis. Specifically, the score function $U(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})$ is given by:

$$U(\widehat{\boldsymbol{\beta}}_{\mathbf{0}}) = \sum_{i=1}^{m} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i).$$

The robust variance estimates of the score statistic, $U(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})$, is also estimated using the "sandwich" estimator $\widehat{Var}\left(U(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})\right)$. To test $H_0: \beta_{j1} = 0$ against $H_a: \beta_{j1} \neq 0$, the generalized score test (GST) statistic S is computed as:

$$S = \frac{U(\hat{\beta}_{j1})}{\widehat{se}\left(U(\hat{\beta}_{j1})\right)},$$

where $U(\hat{\beta}_{j1})$ is the score function evaluated at the estimated parameter $\hat{\beta}_{j1}$ under the null hypothesis, and $\widehat{se}(U(\hat{\beta}_{j1}))$ is the robust standard error of the score function. The test statistic S approximately follows a standard normal distribution under the null hypothesis.

3. Results

3.1. Simulation Studies

We conducted extensive simulation studies that closely resembled the conditions found in real data, including spatial structure and zero-inflated characteristics, for rigorously comparing the performance of different statistical methods in the context of ST. First, we fitted the GLMM described in Section 2.2 to a breast cancer ST dataset from 10x Genomics. This gave an estimation of the two key spatial parameters: σ^2 and κ, where σ^2 is the spatial variance that captures variability in gene expression across different spatial locations and κ is the spatial correlation that controls the rate of decay in correlation with distance between spatial locations.

These estimates were calculated from the fitted results of the GLMM for genes that had converged. The estimated parameters were subsequently used to generate simulated data that preserved the realistic spatial structure found in biological samples, closely resembling real ST datasets. Specifically, we employed the same GLMM structure by applying the estimated σ^2 and κ to ensure spatial correlation, while introducing zero-inflation through a small intercept (β_0) in the model. This approach preserved the spatial heterogeneity observed in real ST data and captured the intrinsic sparsity characteristic of biological samples. To better capture the range of spatial variability present in real tissues, we took set of three levels: 25th percentile, median, and 75th percentile in variation of the spatial parameters, thus representing the spectrum of variance in spatial gene expression and spatial correlation strengths from weak to strong. Regarding this, we considered three scenarios: (1) the 25th percentile of σ^2 and the 25th percentile of κ, (2) the median of σ^2 and the median of κ, and (3) the 75th percentile of σ^2 and the 75th percentile of κ.

These simulations target two aspects of model performance: Type I error rate control and statistical power. Type I error rate control is important to assess the reliability of the results, especially in ensuring that false positives are controlled at the nominal level. For each statistical method, we produced 10,000 replicates to evaluate how well the different approaches could control the rate of false positives at different levels of significance. The cut-offs ranged from the standard $\alpha = 0.05$ to the more stringent $\alpha = 0.0001$. It is in genome-wide analyses that such stringent levels of α are required, due to the large number of tests being conducted simultaneously. On the other hand, owing to the computational intensive nature of these simulations, we did not explore even lower levels of α.

In addition to Type I error control, we investigated the statistical power of each method, which is indicative of the number of DE genes correctly identified. We generated 1,000 replicates

to compare the power of different statistical approaches in the detection of true signals while ensuring Type I error rates remain acceptable. Because GEE rely on asymptotic behavior and thus require large numbers of clusters to make proper inferences, under each scenario we set the number of spatial clusters to either 25 or 100. More precisely, we investigated the impact of the number of spatial clusters on the performance of the GEEs. When the model was configured to treat each spatial location as an independent cluster, it reduced to a special case we termed Independent GEE. This setup is conceptually similar to a two-sample z-test, as it treats all observations as independent.

Of note, although theoretically relevant for modeling correlated data, we excluded the GLMM from the final simulations and comparisons. We made this decision due to its inability to converge for many genes (about 43%) when applied to real ST data. The zero-inflated count nature of the real data resulted in weak spatial correlations, leading to convergence issues for the GLMM when estimating spatial parameters in the full likelihood framework. Failure to converge and yield stable results from the GLMM highlighted its limitations in handling the computational and numerical challenges in the context of ST data. Consequently, our comparisons focused on four methods: the Wilcoxon rank-sum test, the GEE with the robust Wald test, the GEE with GST, and the Independent GEE.

The simulation results are presented in Table 1 for Type I error rate comparison and in Figure 2 for power comparison. We found that when the number of spatial clusters was sufficiently large – k =100 clusters in our case, the GST was the most robust method within the GEE framework. The GST not only achieved superior control of Type I error but also demonstrated comparable statistical power to that of other methods (Figure 2 (b), (d), and (f)). When k =25 clusters, the GST had lower power (Figure 2 (a), (c), and (e)) and conservative Type

I error. In contrast, the robust Wald test in the GEE framework showed inflated Type I error rates across all scenarios, indicating its inability to fully account for the spatial dependencies in the data. Our findings of conservative Type I error control with the GST when the number of clusters (k) was small and inflated Type I error with the robust Wald test are consistent with Boos & Stefanski (2013), where similar patterns were observed. On the other hand, the Independent GEE, which does not account for spatial correlations, showed deflated Type I error rate under certain scenarios (e.g., when $\alpha = 0.0001$) and reduced power. While the Wilcoxon rank-sum test is used very frequently due to its simplicity and ease of implementation, it also led to inflated Type I error rates in some scenarios, especially at lower α levels and moderate/strong spatial correlations, making it unreliable for ST applications where stringent control of false positives is crucial. Therefore, our simulation studies suggest that the GST is the most reliable method for detecting DE genes in ST data.

		Significance Levels			
Clusters	**Methods**	**0.05**	**0.01**	**0.001**	**0.0001**
Weak Spatial Correlation and Spatial Variance (Scenario 1)					
25	GEE with GST	0.0554	0.0097	0.0003	**0.0001**
25	GEE with robust Wald test	0.0737	0.0251	0.0065	0.0019
25	Independent GEE with robust Wald test	0.0463	0.0088	0.0011	**0.0001**
25	Wilcoxon rank-sum test	0.0460	0.0079	**0.0010**	**0.0001**
100	GEE with GST	0.0508	0.0109	0.0008	**0.0001**
100	GEE with robust Wald test	0.0533	0.0121	0.0016	0.0003
100	Independent GEE with robust Wald test	0.0463	0.0088	0.0011	**0.0001**
100	Wilcoxon rank-sum test	0.0460	0.0079	**0.0010**	**0.0001**
Moderate Spatial Correlation and Spatial Variance (Scenario 2)					
25	GEE with GST	0.0615	0.0087	0.0004	0.0000
25	GEE with robust Wald test	0.0803	0.0256	0.0057	0.0017
25	Independent GEE with robust Wald test	0.0525	0.0110	0.0014	**0.0001**
25	Wilcoxon rank-sum test	0.0491	0.0104	0.0011	0.0002
100	GEE with GST	0.0563	0.0120	0.0012	**0.0001**
100	GEE with robust Wald test	0.0589	0.0142	0.0020	0.0004
100	Independent GEE with robust Wald test	0.0525	0.0110	0.0014	**0.0001**
100	Wilcoxon rank-sum test	0.0491	0.0104	0.0011	0.0002
Strong Spatial Correlation and Spatial Variance (Scenario 3)					
25	GEE with GST	0.0709	0.0107	0.0002	0.0000
25	GEE with robust Wald test	0.0856	0.0280	0.0068	0.0020
25	Independent GEE with robust Wald test	0.0587	0.0117	0.0014	0.0000
25	Wilcoxon rank-sum test	0.0483	0.0098	0.0007	0.0002
100	GEE with GST	0.0750	0.0185	0.0020	**0.0001**
100	GEE with robust Wald test	0.0690	0.0179	0.0022	0.0002
100	Independent GEE with robust Wald test	0.0587	0.0117	0.0014	0.0000
100	Wilcoxon rank-sum test	0.0483	0.0098	0.0007	0.0002

Table 1. Type I error comparison of GEE with robust Wald test, GEE with GST, Independent GEE with robust Wald test and Wilcoxon rank-sum test across different spatial scenarios and spatial cluster numbers. Type I error was evaluated at significance levels of $\alpha = 0.05, 0.01, 0.001, 0.0001$. Well controlled Type I error rates are in boldface.

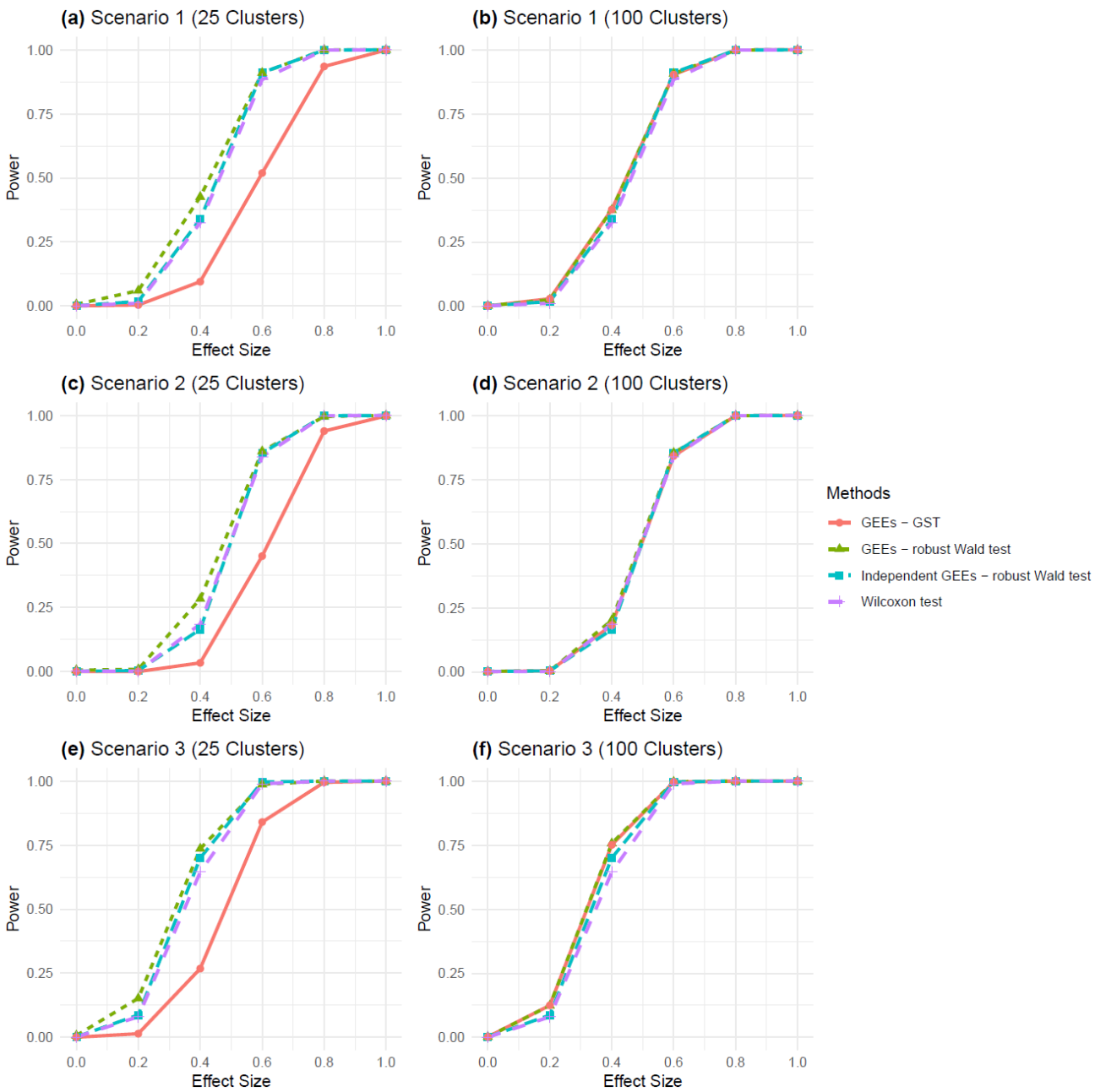

Fig. 2. Power comparison of GEE with robust Wald test, GEE with GST, Independent GEE with robust Wald test and Wilcoxon test across different spatial scenarios and spatial cluster numbers. Panels **(a)**–**(f)** represent power curves for three spatial scenarios with 25 and 100 clusters: **(a)** Scenario 1 (25 clusters), **(b)** Scenario 1 (100 clusters), **(c)** Scenario 2 (25 clusters), **(d)** Scenario 2 (100 clusters), **(e)** Scenario 3 (25 clusters), and **(f)** Scenario 3 (100 clusters). Scenarios represent varying spatial correlation and spatial variance: Scenario 1 (weak), Scenario 2 (moderate), and Scenario 3 (strong). Type I error and power were evaluated at a significance level of $\alpha = 0.001$.

3.2. Real Data Applications

In addition to simulation studies, we applied the statistical methods under comparison to two ST datasets for breast cancer and prostate cancer to evaluate their performance in real-world scenarios.

3.2.1. Breast Cancer ST dataset

We applied our methods to a real ST dataset of breast cancer sample provided by 10x Genomics Visium spatial platform (https://www.10xgenomics.com/datasets/human-breast-cancer-block-a-section-1-1-standard-1-1-0). The ST dataset comprises 3,798 spots and 24,923 genes. After retaining the 3,000 most variable genes, the percentage of zeros has a first quartile of 68.0%, a median of 95.5%, and a third quartile of 99.6%. The haematoxylin and eosin (H&E)-stained tissue image with pathology labels is shown in Figure 1a. The analysis used 100 spatial clusters generated via K-means clustering, as illustrated in Figure 1c. Here, the analysis was aimed at identifying the DE genes across pathological grades, specifically comparing ductal carcinoma in situ (DCIS) with fibrous tissue (FT) and invasive carcinoma (IC) with FT. By focusing on breast cancer, we aimed to explore the consistency of our simulation findings using a well-characterized dataset, therefore providing a comprehensive evaluation of the statistical methods under comparison.

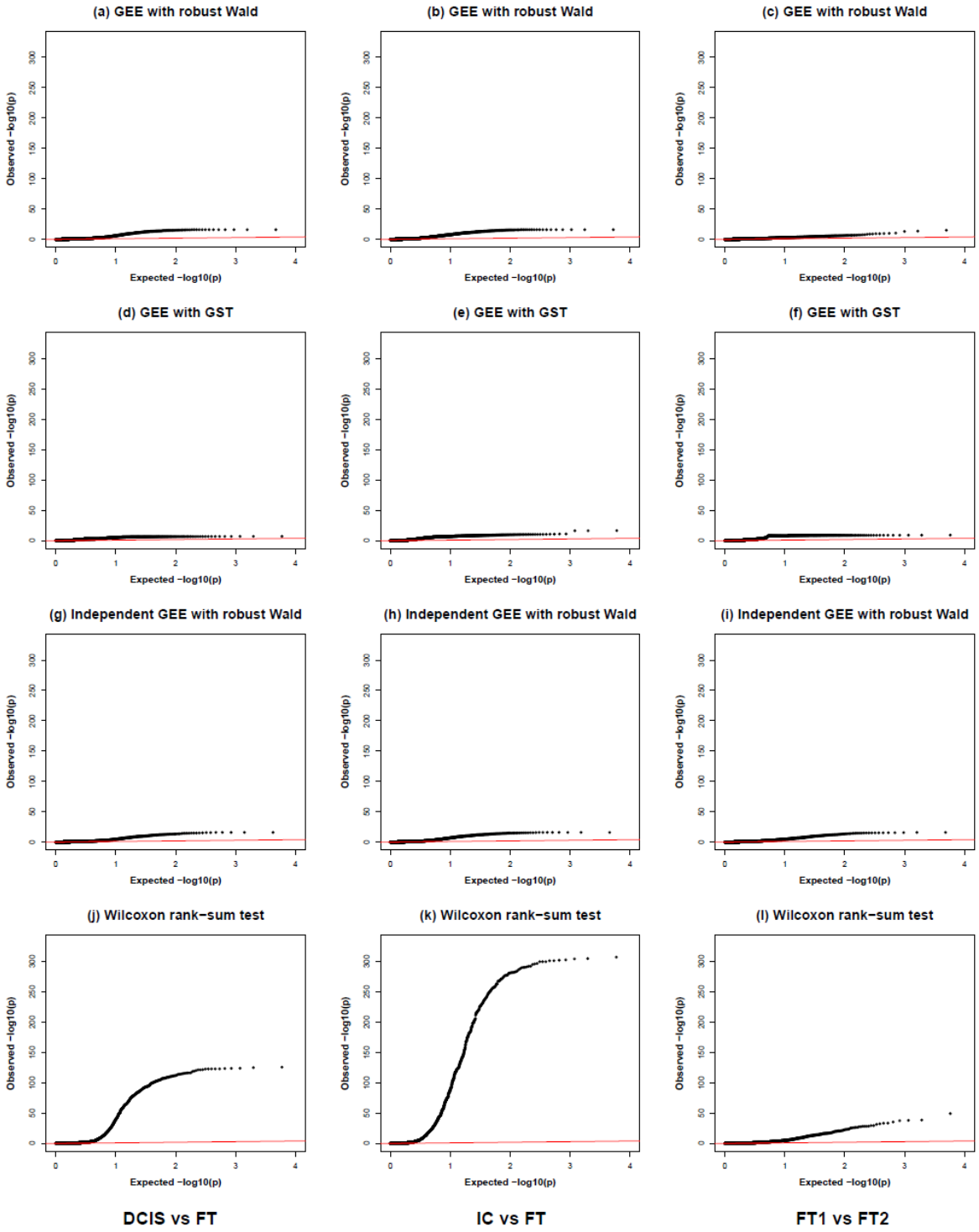

Fig. 3. QQ plots of GEE with robust Wald test, GEE with GST, Independent GEE with robust Wald test and Wilcoxon rank-sum test for breast cancer comparisons: tumor-controls (DCIS vs FT: **(a)**, **(d)**, **(g)**, **(j)**; IC vs FT: **(b)**, **(e)**, **(h)**, **(k)**) and control-control (FT1 vs FT2: **(c)**, **(f)**, **(i)**, **(l)**). The red line represents the expected distribution under the null hypothesis. DCIS: ductal carcinoma in situ; IC: invasive carcinoma; FT: fibrous tissue.

The raw dataset was then pre-processed for the analysis by quality control, normalization, and selection of highly variable genes for downstream analyses. Quality control included the filtration of spots based on low total gene count and the removal of genes expressed in very few spots to ensure a robust analysis. We then normalized the data by regulating the differences in sequencing depth across varying spots. Using the Seurat software suite, we selected 3,000 most variable genes. This has enabled us to focus on the most informative features of the data and therefore enhance the statistical power in subsequent analyses.

We first generated QQ plots of the genome-wide DE scan p-values (Figure 3) to provide a global assessment of the statistical methods. QQ plots are effective in illustrating discrepancies between observed and expected distributions, particularly in assessing how well each method controls false positives. Among these methods, the Wilcoxon rank-sum test showed heavy inflated in Type I error rates, which is consistent with our findings from the simulation studies (at $\alpha = 0.0001$). This inflation suggested that the Wilcoxon rank-sum test may not be appropriate for identifying DE genes in ST data where the spatial correlations are pervasive. Such false positive inflation can be of particular concern in high-throughput genomic studies since this would lead to spurious findings misinforming the biological interpretation.

Given that the ground truth of differential expression status of genes is unknown in real datasets, we adopted an internal negative control strategy by comparing the control versus control samples to identify possible false positive DE genes. This approach allows us to indirectly estimate the precision for each method individually without any predefined ground truth. Specifically, we randomly divided the fibrous tissue samples into two subsets (FT1 and FT2) and compared the two subsets against each other. The overlap between the gene sets identified between the comparison of IC with FT and the control-control comparisons (FT1

versus FT2) gave further insight into the reliability of each statistical approach. A method that has large numbers of overlapping genes suggests poor specificity and is more likely to generate false positives. We observed that the Wilcoxon rank-sum test had more overlapping genes than the GST, suggesting a higher false positive rate for the former. Additionally, the QQ plots for the control-control comparison (Figure 3 (c), (f), (i), and (l)) showed that although the Wilcoxon rank-sum test less deviated the most from the expected null distribution than in the tumor-control comparison (Figure 3 (j) and (k)), inflated Type I error rates remained apparent compared to the GEE-based tests, further supporting the findings of our simulation studies.

After Type I error control evaluation, we conducted biological pathway enrichment analysis on the gene sets identified as differentially expressed by the GST and the Wilcoxon rank-sum methods. We performed the pathway analysis to validate whether the gene sets uniquely identified by IC versus FT, excluding overlapping genes between IC versus FT and FT1 versus FT2, were predominantly enriched in cancer-related biological pathways, therefore further providing a biological context to the statistical findings. The enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways using g:Profiler (Kolberg et al., 2023) are presented in Figure 4a and Figure 4b using dot plots. The GST-identified DE genes were enriched in pathways related to cancer progression, including critical KEGG cancer pathways such as focal adhesion, PI3K-Akt signaling, ECM-receptor interaction, and pathways in cancer. Focal adhesion plays an important role in cell signaling and interaction with the extracellular matrix during tumor invasion and metastasis (Jin & Varner, 2004). PI3K-Akt signaling pathway commonly is activated in a variety of cancers and controls central cellular processes like survival, growth, and metabolism (Vivanco & Sawyers, 2002). ECM-receptor interaction is essential in cancer to mediate communication between tumorigenic cells and the surrounding

stromal environment for tumor progression (Lu et al., 2011). Pathways in cancer consist of broad categories of signaling mechanisms related to tumor development, survival, and metastasis. These are all well-established pathways in the literature to be implicated in the development and progression of cancer, and further supports the reliability of the GST method. In contrast, the Wilcoxon rank-sum test missed enrichment in proteoglycans in cancer pathway that plays a vital role in cancer biology in modulating cell proliferation, migration, and angiogenesis (Iozzo & Sanderson, 2011). It instead largely enriched non-cancer-related pathways, which included several pathways related to general metabolic processes and less specifically associated with the cancer biology. This discrepancy underlines even further the tendency of the Wilcoxon rank-sum test to yield spurious associations that may mislead interpretation of underlying biological mechanisms.

Moreover, we conducted an additional biological pathway enrichment analysis on the overlapping genes. This analysis aimed to determine whether the overlapping genes were enriched within known biological pathways relevant to noncancer. The GST method predominantly enriched the non-cancer-related KEGG pathways (Figure 4b), whereas the Wilcoxon rank-sum test did the opposite and largely enriched the cancer-related KEGG pathways (Figure 4a), further emphasizing its tendency of yielding false positives that give misleading biological associations.

3.2.2. Prostate Cancer ST dataset

We extended our comparison further by applying the statistical methods to a prostate cancer ST data, publicly available from 10x Genomics Visium spatial platform (https://www.10xgenomics.com/datasets/human-prostate-cancer-adenocarcinoma-with-invasive-carcinoma-ffpe-1-standard-1-3-0). The ST dataset comprises 4,371 spots and 16,907 genes. After

retaining the 3,000 most variable genes, the percentage of zeros has a first quartile of 50.0%, a median of 80.4%, and a third quartile of 97.4%. Figure 1b shows the H&E-stained tissue image with pathology labels. Figure 1d illustrates the 100 spatial clusters generated via K-means clustering. It was of interest to identify DE genes between the tissue types, comparing invasive carcinoma (IC) against adjacent non-cancerous fibro-muscular tissue (FT). The raw data was preprocessed, including quality control, normalization, and the selection of 3,000 most variable genes using the Seurat pipeline.

As in the breast cancer application, we first created QQ plots of genome-wide DE scans to compare each statistical method's performance in controlling false positives. As shown in Supplemental Materials Figure S1, the QQ plots showed extreme Type I error inflation in the Wilcoxon rank-sum test, confirming the findings obtained from the breast cancer analysis and the simulation studies. This further supports our cautionary note on the Wilcoxon rank-sum test for spatial transcriptomic data analysis where stringent control of false positives is essential.

We then resorted to a similar internal negative control strategy as in the analysis of the breast cancer dataset. We identified possible false positives for each approach by dividing the non-cancerous tissue into two subsets and comparing these subsets against one another. In fact, the Wilcoxon rank-sum test showed more overlapping genes between IC and FT than those identified by the GST, indicating higher false positives. Additionally, the QQ plots for the control-control comparison showed persistent Type I error inflation in the Wilcoxon rank-sum test, further supporting our findings from the breast cancer analysis and the simulation studies.

We next conducted a biological pathway enrichment analysis to assess the biological relevance of the genes uniquely identified as differentially expressed when using either the GST or the Wilcoxon rank-sum method. More precisely, we considered genes uniquely identified by

comparing IC with FT, discarding the overlap with control-control comparisons. Figure 4c and Figure 4d show the enriched KEGG pathways as dot plots. In general, the enriched pathways associated with tumor progression, such as focal adhesion, PI3K-Akt signaling pathway, ECM-receptor interaction, and MAPK signaling pathway, were dominant in the GST method. The MAPK signaling pathway controls cell proliferation, differentiation, and apoptosis and therefore is very important in the processes of oncogenesis and metastasis (Pearson et al., 2001). By contrast, the Wilcoxon rank-sum test did not enrich any of the cancer-related pathways identified by the GST, echoing our findings in the breast cancer ST data application.

Furthermore, we conducted an additional enrichment analysis on the overlapping genes: the Wilcoxon rank-sum test-identified DE genes were enriched more substantially for cancer-related pathways (Figure 4c) than did the GST method (Figure 4d), further underlining the former's failure by producing misleading false positive results.

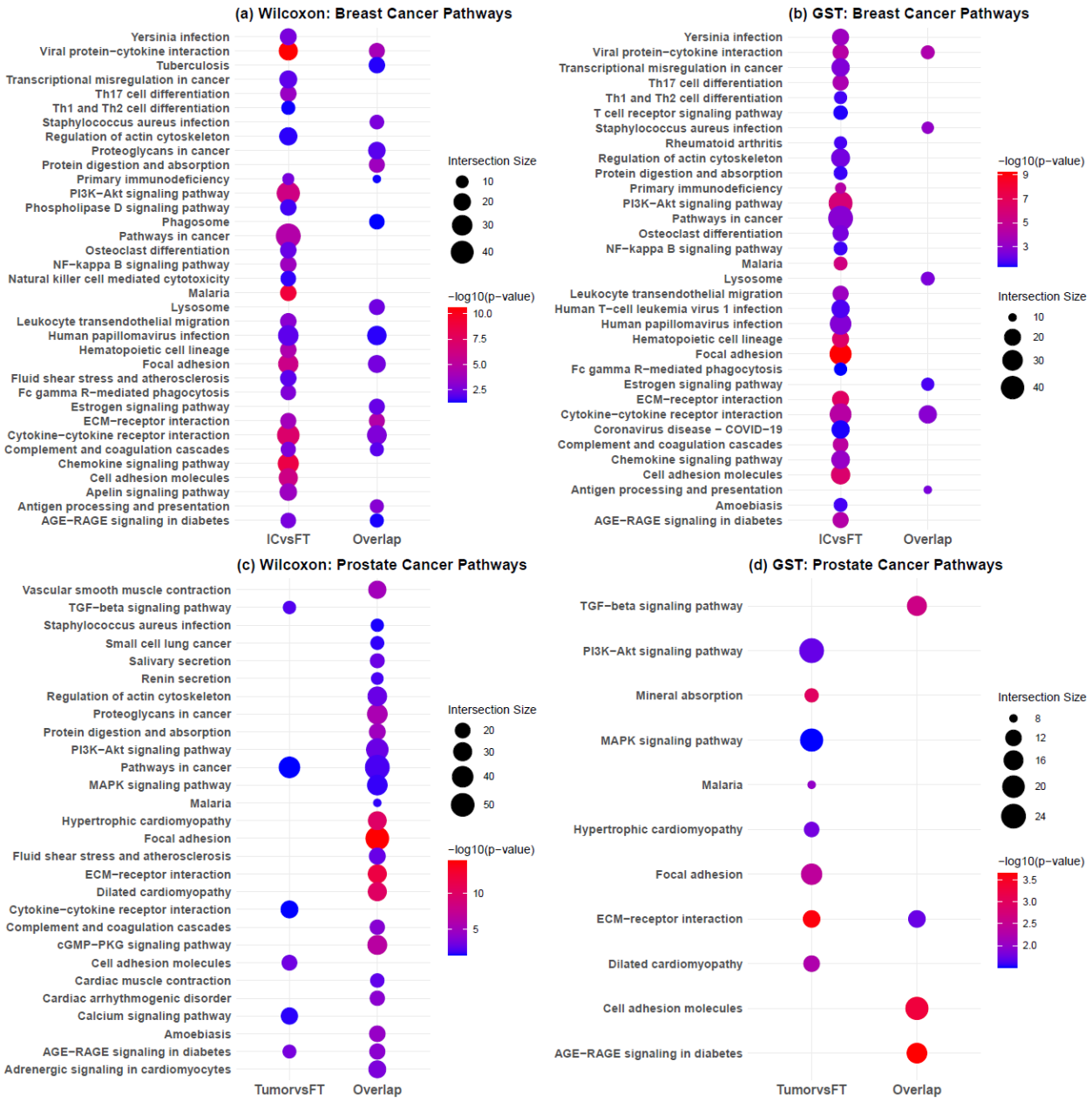

Fig. 4. Dot plots of enriched KEGG pathways identified in breast and prostate cancer. **(a)** Pathways identified by the Wilcoxon rank-sum test in breast cancer, **(b)** pathways identified by GST in breast cancer, **(c)** pathways identified by the Wilcoxon rank-sum test in prostate cancer, and **(d)** pathways identified by GST in prostate cancer. IC: invasive carcinoma; FT: fibrous tissue; Overlap: intersection of the differential expressed gene sets identified by IC/Tumor vs FT comparison and FT1 vs FT2 (control-vs-control) comparison.

4. Discussion

We have performed a comprehensive comparison of statistical methods for identified DE genes in ST data with evaluation of Type I error control, power, and identification of biologically relevant pathways. Through extensive simulations and applications to breast cancer and prostate cancer ST datasets, we compared the robustness of each method in handling the spatial correlations inherent in ST data. Our findings demonstrate that the GST within the GEE framework outperforms alternative approaches, including the Wilcoxon rank-sum test, the GEE with the robust Wald test and the Independent GEEs, in identifying DE genes in ST data.

A key finding of simulation is the superior Type I error control achieved by the GST, especially when the number of spatial clusters is large enough to support the asymptotic properties of the GEE. In contrast, the Wilcoxon rank-sum test, despite its popularity due to computational simplicity, consistently exhibited significant inflation of Type I error across simulations and real data analysis. This inflation was particularly apparent in the QQ plots, suggesting its over-detection of DE genes in spatially structured data.

Recent statistical genetics research also supports the stability of the GST with similar findings, especially when many single nucleotide polymorphisms (SNPs) are correlated and rare genetic variants produce sparse data patterns. In these cases, score tests have demonstrated greater numerical stability than the Wald test, which requires full model fitting under the alternative hypothesis and can be unstable or infeasible with sparse or correlated data (Pan et al., 2014). This limitation of the Wald test is consistent with our findings in ST, where the GST provides a more stable and reliable alternative in the presence of spatial correlations. Furthermore, extensions of the GST within the GEE framework in genetic association studies

have demonstrated that this approach is adaptive and robust for high-dimensional data with complex correlation structures (Wang et al., 2013; Zhang et al., 2014).

The real data analysis on breast cancer and prostate cancer further validated our simulation findings. We first focused on the gene sets uniquely identified in cancerous versus non-cancerous tissues by excluding overlapping genes with internal control comparisons. The GST consistently enriched pathways critical to cancer biology, such as PI3K-Akt signaling, ECM-receptor interaction, and focal adhesion – pathways essential for cell proliferation, apoptosis, and tumor growth. In contrast, the Wilcoxon rank-sum test failed to identify these key cancer-related pathways in prostate cancer data, instead highlighting nonspecific pathways, raising concerns about its reliability in ST. Next, we examined the overlapping genes, i.e., the intersection of the DE gene sets identified by tumor vs normal comparison and normal vs normal comparison, which further underlined the differences of these methodologies. The GST resulted in lower overlap among enriched gene sets, indicating high specificity, whereas the Wilcoxon rank-sum test displayed significant overlap, suggesting a higher false positive rate and reduced robustness. Additionally, the pathway enrichment analysis on the overlapping genes showed that the GST predominantly enriched non-cancer pathways, while the Wilcoxon rank-sum test had a strong tendency to enrich cancer-related pathways. It further emphasized the latter's susceptibility to producing spurious associations.

Our findings from extensive simulations and real data applications emphasize the importance of statistical methods selection in ST data analysis. Given the intrinsic spatial dependencies of such data, methods that do not consider these dependencies are at risk of producing unreliable results and ensuing misleading biological interpretation. The GST approach within the GEE framework provides a robust solution to these challenges by appropriately

modeling spatial correlations without the computational burdens associated with random effects modeling in the GLMM. These findings support the GST as a suitable method for DE analysis in ST data, providing both robust Type I error control and biologically meaningful insights.

Several ST platforms are available and generate data with different characteristics, such as 10x Visium ST, 10x Xenium, 10x Visium HD, Slide-Seq, MerFISH, and CosMX. Our proposed GST approach is in principle widely applicable to all these platforms, similar to the Wilcoxon rank-sum test implemented in Seurat. We specifically used two 10x Visium ST datasets in our real data applications, as 10x Visium is thus far the most widely adopted platform. Future research is warranted to study GST's robustness across other ST technologies. Of note, the GEE-GST is computationally feasible for genome-wide scans: it took 204 seconds to complete the analysis of 2212 spots and 2550 genes on a 10-core processor, with 2GB as the total peak memory usage.

Data and Code Availability

We have implemented our proposed methods in R package 'SpatialGEE' available at https://github.com/yishan03/SpatialGEE. The spatial transcriptomics datasets of breast cancer and prostate cancer analyzed here are available on 10x Genomics websites: https://www.10xgenomics.com/datasets/human-breast-cancer-block-a-section-1-1-standard-1-1-0 and https://www.10xgenomics.com/datasets/human-prostate-cancer-adenocarcinoma-with-invasive-carcinoma-ffpe-1-standard-1-3-0.

Supporting Information

The supplementary materials include Supplementary Figure S1.

Acknowledgments

This work was supported by Cancer Prevention and Research Institute of Texas (CPRIT) grant RP230166 (to PW). PW was partially supported by National Institutes of Health (NIH) grant P50CA217674. The authors declare no conflict of interest.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289-300.

Boos, D. D., & Stefanski, L. A. (2013). Essential statistical inference: Theory and methods. Springer.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421), 9-25.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5), 411-420.

Erickson, A., He, M., Berglund, E., & Others. (2022). Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature*, 608, 360–367.

Hao, Y., Stuart, T., Kowalski, M. H., & Others. (2024). Dictionary learning for integrative, multimodal, and scalable single-cell analysis. *Nature Biotechnology*, 42, 293–304.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 221-233.

Iozzo, R. V., & Sanderson, R. D. (2011). Proteoglycans in cancer biology, tumour microenvironment and angiogenesis. *Journal of Cellular and Molecular Medicine*, 15(5), 1013-1031.

Jiang, X., Wang, S., Guo, L., Zhu, B., Wen, Z., Jia, L., Xu, L., Xiao, G., & Li, Q. (2024) iIMPACT: integrating image and molecular profiles for spatial transcriptomics analysis. *Genome Biology*, 2024, 25, 147.

Jin, Y., & Varner, J. (2004). Integrins: Roles in cancer development and as treatment targets. *British Journal of Cancer*, 90(3), 561-565.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22.

Li, Y., Tang, H., & Lin, X. (2009). Spatial linear mixed models with covariate measurement errors. *Statistica Sinica*, 19(3), 1077.

Lu, P., Takai, K., Weaver, V. M., & Werb, Z. (2012). Extracellular matrix degradation and remodeling in development and disease. *Cold Spring Harbor Perspectives in Biology*, 3(12), a005058.

Ma, Y., & Zhou, X. (2024). Accurate and efficient integrative reference-informed spatial domain detection for spatial transcriptomics. *Nature Methods*, 21(7), 1231–1244.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297.

McLachlan, G. J., & Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.

Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., & Peterson, H. (2023). g:Profiler—Interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*.

Ospina, O. E., Wilson, C. M., Soupir, A. C., Berglund, A., Smalley, I., Tsai, K. Y., & Fridley, B. L. (2022). spatialGE: quantification and visualization of the tumor microenvironment heterogeneity using spatial transcriptomics. *Bioinformatics*, 38(9), 2645–2647.

Pearson, G., Robinson, F., Beers Gibson, T., Xu, B. E., Karandikar, M., Berman, K., & Cobb, M. H. (2001). Mitogen-activated protein (MAP) kinase pathways: Regulation and physiological functions. *Endocrine Reviews*, 22(2), 153-183.

Pan, W., Kim, J., Zhang, Y., Shen, X., & Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics*, 197(4), 1081–1095.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44, 50-57.

Shah, K., Guo, B., Hicks, S.C. (2024) Addressing the mean-variance relationship in spatially resolved transcriptomics data with spoon. *bioRxiv*, 2024.11.04.621867. doi: 10.1101/2024.11.04.621867.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., ... & Frisen, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294), 78-82.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., ... & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, 177(7), 1888-1902.

Svensson, V., Teichmann, S. A., & Stegle, O. (2018). SpatialDE: Identification of spatially variable genes. *Nature Methods*, 15(5), 343-346.

Tian, L., Chen, F., & Macosko, E. Z. (2023). The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 41, 773–782.

Vivanco, I., & Sawyers, C. L. (2002). The phosphatidylinositol 3-Kinase AKT pathway in human cancer. *Nature Reviews Cancer*, 2(7), 489–501.

Wang, X., Lee, S., Zhu, X., Redline, S., & Lin, X. (2013). GEE-Based SNP set association test for continuous and discrete traits in family-based association studies. *Genetic Epidemiology*, 37(8), 778-786.

Wei, P., & Pan, W. (2008). Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, 24(3), 404-411.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817-838.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.

Zang, C., Guo, C. C., Wei, P., & Li, Z. (2024). TUSCAN: Tumor segmentation and classification analysis in spatial transcriptomics. *bioRxiv*.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44(4), 1049-1060.

Zhang, Y., Xu, Z., Shen, X., & Pan, W. (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage*, 96, 309-325.

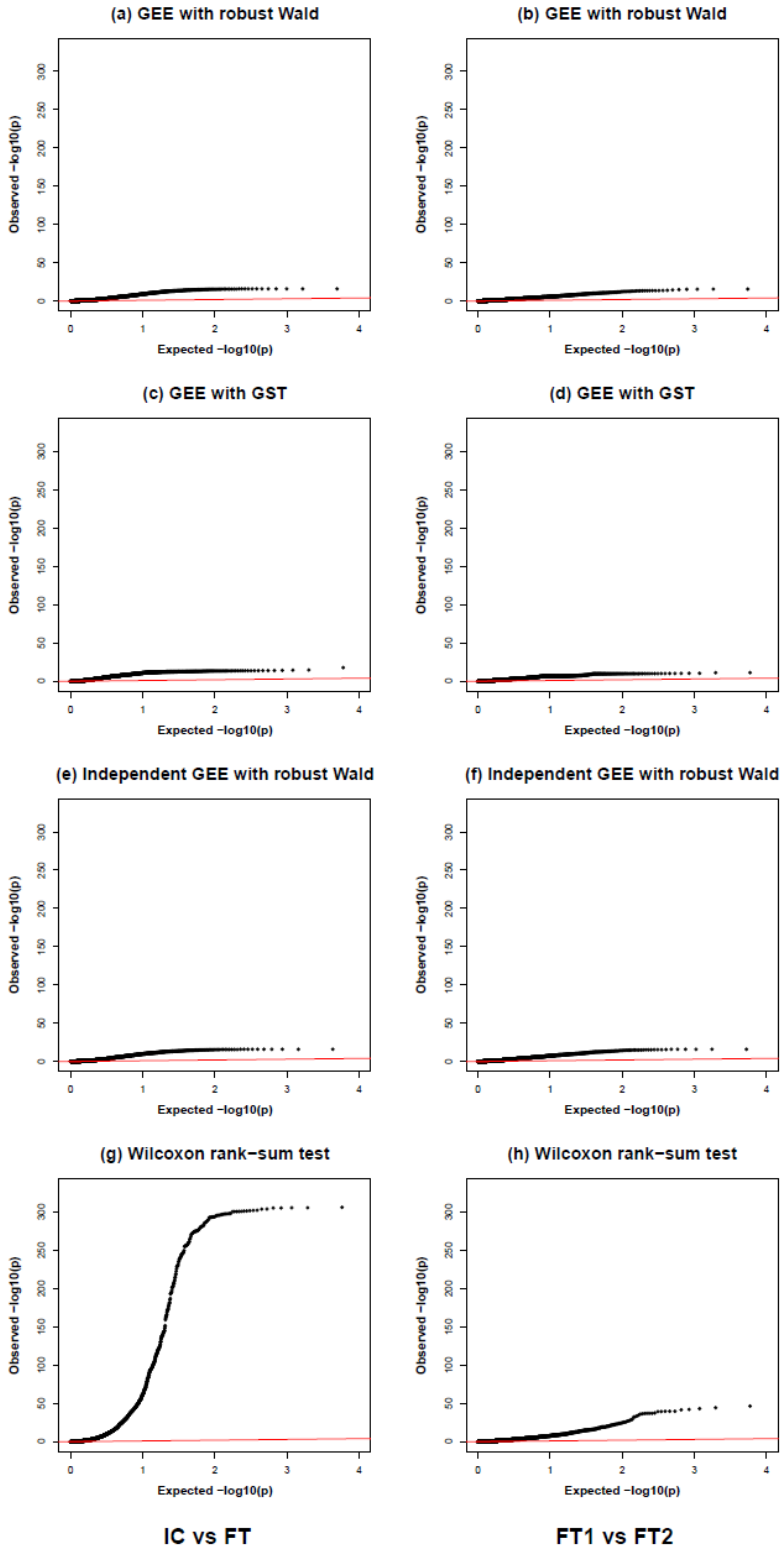

Fig. S1. QQ plots of GEE with robust Wald test, GEE with GST, Independent GEE with robust Wald test and Wilcoxon rank-sum test for prostate cancer comparisons: tumor-controls (IC vs FT: **(a)**, **(c)**, **(e)**, **(g)**) and control-control (FT1 vs FT2: **(b)**, **(d)**, **(f)**, **(h)**). The red line represents the expected distribution under the null hypothesis. IC: invasive carcinoma; FT: fibrous tissue.