Article

Inferring Cell–Cell Communications from Spatially Resolved Transcriptomics Data Using a Bayesian Tweedie Model

Dongyuan Wu [1], Jeremy T. Gaskins [2], Michael Sekula [2] and Susmita Datta [1,*]

[1] Department of Biostatistics, University of Florida, Gainesville, FL 32603, USA; dongyuanwu@ufl.edu
[2] Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA; jeremy.gaskins@louisville.edu (J.T.G.); michael.sekula@louisville.edu (M.S.)
* Correspondence: susmita.datta@ufl.edu

Abstract: Cellular communication through biochemical signaling is fundamental to every biological activity. Investigating cell signaling diffusions across cell types can further help understand biological mechanisms. In recent years, this has become an important research topic as single-cell sequencing technologies have matured. However, cell signaling activities are spatially constrained, and single-cell data cannot provide spatial information for each cell. This issue may cause a high false discovery rate, and using spatially resolved transcriptomics data is necessary. On the other hand, as far as we know, most existing methods focus on providing an ad hoc measurement to estimate intercellular communication instead of relying on a statistical model. It is undeniable that descriptive statistics are straightforward and accessible, but a suitable statistical model can provide more accurate and reliable inference. In this way, we propose a generalized linear regression model to infer cellular communications from spatially resolved transcriptomics data, especially spot-based data. Our BAyesian Tweedie modeling of COMmunications (BATCOM) method estimates the communication scores between cell types with the consideration of their corresponding distances. Due to the properties of the regression model, BATCOM naturally provides the direction of the communication between cell types and the interaction of ligands and receptors that other approaches cannot offer. We conduct simulation studies to assess the performance under different scenarios. We also employ BATCOM in a real-data application and compare it with other existing algorithms. In summary, our innovative model can fill gaps in the inference of cell–cell communication and provide a robust and straightforward result.

Keywords: cellular communication; spatial transcriptomics; generalized linear regression model; Bayesian modeling; Tweedie distribution

check for updates

Citation: Wu, D.; Gaskins, J.T.; Sekula, M.; Datta, S. Inferring Cell–Cell Communications from Spatially Resolved Transcriptomics Data Using a Bayesian Tweedie Model. *Genes* **2023**, *14*, 1368. https://doi.org/10.3390/genes14071368

Academic Editors: Anelia D. Horvath and Liangjiang Wang

Received: 30 April 2023
Revised: 16 June 2023
Accepted: 26 June 2023
Published: 28 June 2023

1. Introduction

Different biochemical signalings from cellular communications control different activities of living organisms, which highlights the importance of understanding cell–cell communications (CCC) on biological processes and mechanisms [1,2]. In practice, we infer the CCC from some known ligand–receptor (LR) pairs because the interaction of LRs mediates communication. As single-cell RNA sequencing (scRNA-seq) technologies have matured, researchers have gradually gained opportunities to investigate CCC from scRNA-seq data since we know more about the ligand and receptor gene expression information and cell type annotation at the cellular level. Several approaches have been proposed for inferring the CCC from scRNA-seq data. For example, CellPhoneDB [3] calculates the mean of the average ligand expression level for one cell type and the average receptor expression level for another cell type and conducts a permutation test to determine the significance of this LR pair between two cell types. Instead of using the mean or product to measure the communication between two cell types, SingleCellSingleR [4] introduces a regularized product score for an LR pair and provides an ad hoc benchmark to decide an appropriate score threshold. In addition, CellChat [5] offers a more complicated measurement to reflect the interaction strength of an LR pair between cell types, but it also utilizes the permutation test to identify statistical significance. Even though these methods can infer cellular communications to some extent, one of the common limitations is that they do not consider the spatial information for each cell, which is crucial for cell signaling activities but lost in single-cell data. This restriction may lead to high false discovery rates in discovering intercellular communications [1].

Fortunately, in recent years, the development of various spatially resolved transcriptomics (SRT) technologies has made it possible to access cellular locations, opening up new opportunities to incorporate physical distances of cells into CCC analysis. Giotto [6] provides a similar communication measurement as CellPhoneDB [3], but it incorporates the spatial information. Its computations focus on proximal cells, and its permutation tests shuffle cell locations within the same cell type rather than mixing the cell type annotations. SpaOTsc [7] treats CCC analysis as an optimal transport problem, using a random forest model to estimate the spatial distance of a signaling pathway and adjusting the cost matrix of the optimal transport plan by incorporating inferred spatial constraints. COMMOT [8] shares a similar framework with SpaOTsc [7] but accounts for the competition between cells in the signaling analysis. SpaTalk [9] evaluates the communication scores using intercellular and intracellular scores. The intercellular score is based on the number of one-hop neighbor nodes of receiver cell types for each sender cell type, while the intracellular score is computed from their integrated LR transcription factor knowledge graph. More comprehensive introductions to the existing CCC methods can be found in [1,10].

One should note that the existing approaches for SRT data also have their own drawbacks. The existing spatial CCC approaches only consider the single-cell resolution data from technologies such as seqFISH+ [11] and STARmap [12]. However, some spot-based technologies, such as the widely used 10X Visium [13,14] and Slide-seqV2 [15], detect gene expression levels based on spots, which means that each pixel location may contain several cells. This challenge persists even with high-resolution technologies that can reach the size of mammalian cells, as cells may overlap with each other [16]. Thus, it is valuable to interpret the mixture of multiple cell types and their corresponding proportions for CCC analysis from the spot-based data. Some recent methods using cell type deconvolution, such as RCTD [16], SPOTlight [17], and STRIDE [18], have considered the mixture issue of the spot-based data, but few CCC approaches have been proposed to address it. In addition, most of the existing CCC methods focus on providing an ad hoc measurement to estimate intercellular communication instead of relying on a statistical model. While it is undeniable that descriptive statistics are straightforward and accessible to interpret, a suitable statistical model is needed to provide more accurate and reliable estimation and inference.

In this paper, we introduce a novel generalized linear regression model with compound Poisson–Gamma distributions, also known as Tweedie distribution with $p \in (1,2)$, to infer the communications between cell types. The model combines the physical locations of spots/cells and the proportions of cell types to estimate the signaling strength from one cell type to another. Its unique structure allows it to handle both spot-based SRT data and SRT data at the single-cell resolution, and it is able to consider the communication between different cell types simultaneously for a particular LR pair. For spot-based SRT data, our model uses a convolution strategy to integrate the possible interactions between the cell types at the sender spots and the cell types at the receiver spots to the average spot-to-spot communication scores. Furthermore, due to the properties of the regression model, our proposed method naturally provides the direction of the association between cell type communication and LR interaction that other approaches cannot offer. Since we utilize Bayesian inference for this model, we refer to our approach as BATCOM, which is shorthand for BAyesian Tweedie modeling of COMmunications.

The rest of this manuscript is organized as follows. In Section 2, we define the model structure and demonstrate how our model estimates the communication strength from one cell type to another based on the gene expression matrix and the spot locations from SRT data. In Section 3, both simulation studies and case studies are conducted to display the usability, reliability, and robustness of the proposed model. Finally, we summarize our conclusions and give a discussion in Section 4.

2. Materials and Methods

The workflow of our proposed approach is summarized in Figure 1. Firstly, we require the gene expression matrix from the SRT data with a list of known LR pairs to calculate the communication scores. We also need the spatial information (i.e., coordinate locations) of spots/cells to determine the distances between each pair of spots/cells. Next, we combine cell type annotations for spots/cells obtained from some upstream cell type deconvolution tools, such as RCTD [16], STRIDE [18], and SPOTlight [17]. Specifically, for spot-based SRT data, we should have information on the proportion of each cell type within each spot. On the other hand, for SRT data at the single-cell resolution, it is also easy to generate a matrix that identifies the cell type for each cell, as these data are a special case for our approach where the membership proportion is 1 for the corresponding cell type and 0 otherwise. For clarity, we will focus exclusively on spot-based SRT data in the following sections. After fitting a generalized linear regression model, the communication strength between different cell types is obtained from the regression coefficients. One can also display the communication strengths between cell types in detail using a heatmap and construct a network to visualize cell type communication.

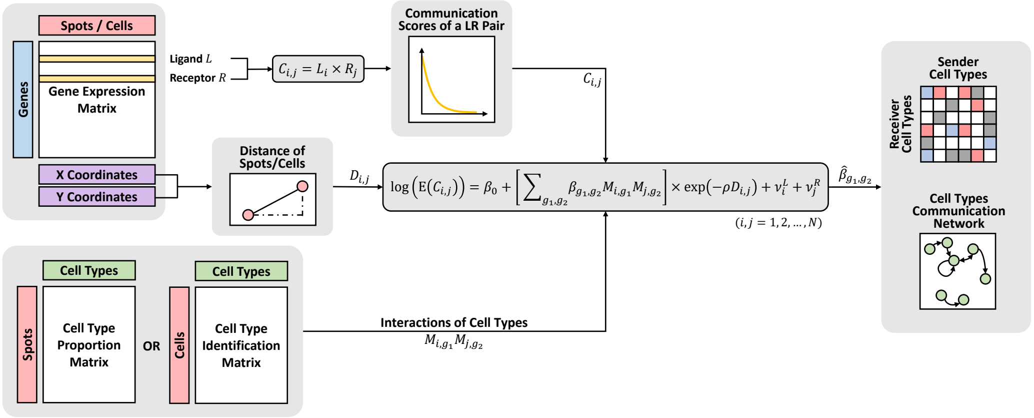
Spots / Cells
Genes
Gene Expression Matrix
X Coordinates
Y Coordinates
Ligand L
Receptor R
$C_{i,j} = L_i \times R_j$
Communication Scores of a LR Pair
$C_{i,j}$
Distance of Spots/Cells
$D_{i,j}$
$\log\left(\mathrm{E}(C_{i,j})\right) = \beta_0 + \left[\sum_{g_1,g_2} \beta_{g_1,g_2} M_{i,g_1} M_{j,g_2}\right] \times \exp(-\rho D_{i,j}) + \nu_i^L + \nu_j^R$
$(i,j = 1,2,\ldots,N)$
$\hat{\beta}_{g_1,g_2}$
Sender Cell Types
Receiver Cell Types
Cell Types Communication Network
Cell Types
Spots
Cell Type Proportion Matrix
OR
Cell Types
Cells
Cell Type Identification Matrix
Interactions of Cell Types
$M_{i,g_1} M_{j,g_2}$

Figure 1. Overview of the workflow. BATCOM requires three main inputs: (1) a gene expression matrix obtained from the SRT data with coordinates information, (2) a matrix that reflects the cell type annotations of spots/cells, and (3) a list of known LR pairs. After fitting the model, a heatmap and a network can be generated for visualization purposes.

2.1. Spot-to-Spot Communication Score

Since the spot-based SRT data only provide the gene expression level for each spot, we define the communication score $C_{i,j}^k$ $(i,j = 1,2,\ldots,N)$ from sender spot i to receiver spot j for the LR pair k as

$$C_{i,j}^k = L_i^k \times R_j^k, \tag{1}$$

where L_i^k and R_j^k are the expression level of ligand L at spot i and the expression level of receptor R at spot j, separately. Specifically, L_i^k and R_j^k can be estimated using the arithmetic mean if the ligand L or receptor R consists of more than one subunit. In other words,

$$L_i^k = \frac{1}{s_L} \sum_{s=1}^{s_L} y_i^{L_{k,s}},$$

$$R_j^k = \frac{1}{s_R} \sum_{s=1}^{s_R} y_j^{R_{k,s}},$$

where s_L and s_R denote the numbers of subunits of ligand L and receptor R, and y_i and y_j are the corresponding normalized gene expression levels of spot i and spot j in SRT data. Although the geometric mean of subunits can be an alternative option for estimation, this will introduce a significantly larger number of zeros present in $C_{i,j}^k$ compared to the arithmetic mean.

While public databases such as CellChatDB [5], CellPhoneDB [3], and CellTalkDB [19] offer an extensive list of LR pairs, it is crucial to note that not all pairs are relevant for the analysis since they may not exist in the data being studied. Therefore, we remove LR pairs whose spot-to-spot communication scores $C_{i,j}$ contain more than 98% zeros. Once we obtain the communication scores $C_{i,j}^k$ for each relevant LR pair k, we can calculate spot-to-spot communication scores for a specific signaling pathway or the entire system by summing up the corresponding LR pairs.

The $C_{i,j}^k$ communication scores in Equation (1) will be the outcome variables in the regression modeling yielding N^2 observations. In practice, this dimensionality can be reduced by only considering (i, j) pairs that are within a certain distance of each other.

2.2. BATCOM Model Structure

To model the signaling strength for LR pair k across all combinations of cell types from the communication scores between two spots, we propose a generalized linear regression model

$$g\left(\mathrm{E}(C_{i,j}^k)\right) = \beta_0^k + \left[\sum_{g_1,g_2} \beta_{g_1,g_2}^k M_{i,g_1} M_{j,g_2}\right] \times \exp(-\rho D_{i,j}) + \nu_i^L + \nu_j^R,\ i,j = 1,2,\ldots,N, \quad (2)$$

where M_{i,g_1} and M_{j,g_2} are the proportions of cell types $g_1 (g_1 = 1,2,\ldots,G)$ at the sender spot i and $g_2 (g_2 = 1,2,\ldots,G)$ at the receiver spot j. Although the model typically includes G^2 interaction terms to account for the communications among G cell types, some of these product terms may be filtered out in practice due to minimal or non-existent observations, or they may be based on prior knowledge.

In addition, $\rho > 0$ is a communication constraint parameter, and $D_{i,j}$ is a suitably chosen distance metric between spots i and j, which we considered to be Euclidean in our applications. As spots i and j are spatially further apart ($D_{i,j}$ increasing), they have less ability to communicate. This effect is captured by the $\exp(-\rho D_{i,j})$ term in Equation (2), which down-weights the impact of the cell type memberships as distance increases. Larger values of ρ represent a faster spatial decay such that only adjacent spots may communicate, while smaller values of ρ allow communication across longer distances. However, the specific value of ρ should be chosen based on the scale of distances in the dataset. In this paper, we scaled D_{ij} so that adjacent spots had a distance of 1.

Based on Equation (1), $C_{i,j}^k$ $(i,j = 1,2,\ldots,N)$ are not independent of each other as they partially come from the same spot. For example, when $i = 1$, all $C_{1,j}^k$ $(j = 1,2,\ldots,N)$ should be correlated with each other because they all depend on the same sender spot expression L_1^k. Thus, Equation (2) includes two random effect parameters ν_i^L and ν_j^R to introduce correlation around the corresponding sender spot i for ligand L and receiver spot j for receptor R, respectively. Returning to our example, if spot 1 exhibits high expression levels for a specific ligand L_1^k, it will result in the corresponding communication scores $C_{1,j}^k$ being large or above average for all j. Neglecting to account for the shared structure across $C_{1,j}^k$ for $j = 1,2,\ldots,N$ may lead to an overestimation of the effect of the cell types most prevalent in spot 1. To address this issue, we introduce the inclusion of a large ν_1^L to capture the characteristics of spot 1, accounting for its high expression of ligand L and communication scores $C_{1,j}$. By incorporating this additional variability, the remaining variation in $C_{1,j}^k$ becomes associated with the primary target of interest: the cell type combinations.

The regression coefficients $\beta^k_{g_1,g_2}$ in Equation (2) are the parameters of interest in our model and thus reflect the communication strength of LR pair k from sender cell type g_1 to receiver cell type g_2. A positive $\beta^k_{g_1,g_2}$ indicates that as the memberships of cell type g_1 at the sender spot and cell type g_2 at the receiver spot jointly increase, the communication is predicted to increase. Conversely, negative coefficients suggest that larger cell type memberships will decrease the spot-to-spot communication of LR pair k. In this way, our approach is capable of deconvoluting the mean spot-to-spot communication scores into the interactions between the cell types at the sender spot and the cell types at the receiver spot.

It is important to note that the original gene expression matrix of SRT data is usually sparse. Moreover, if the ligand L or the receptor R is not expressed, the communication score $C^k_{i,j}$ will be zero according to Equation (1). As a result, $C^k_{i,j}(i, j = 1, 2, \ldots, N)$ is expected to be a sparse vector with continuous positive scores in the non-zero positions. Considering this property of the data, one common choice would be fitting a zero-inflated or hurdle model. However, both models require two sets of coefficients to account for the probability of zeros and the value of non-zeros separately [20–22], and it would be difficult to integrate these two different sets of coefficients together to reflect the strength of communication between cell types. To that end, we utilize the compound Poisson–Gamma distribution to model the communications scores $C^k_{i,j}$. This distribution effectively models the zero-inflated continuous values observed in the data while simplifying the modeling process and enhancing interpretability.

2.3. Compound Poisson–Gamma Distribution

For the compound Poisson–Gamma distribution CPG$(\lambda, \alpha, \gamma)$, the random variable C can be generated as follows:

$$C = \sum_{i=1}^{T} X_i, \quad T \sim \text{Poisson}(\lambda), \quad X_i \overset{iid}{\sim} \text{Gamma}(\alpha, \gamma), \quad T \perp\!\!\!\perp X_i, \tag{3}$$

where λ is the rate of the Poisson distribution, and α and γ are the shape and scale of the Gamma distribution, respectively. Based on the settings in Equation (3), we have

$$(C|T = t) = 0 \text{ if } t = 0,$$

$$(C|T = t) \sim \text{Gamma}(t\alpha, \gamma) \text{ if } t > 0,$$

which implies that the joint distribution of C and T is

$$\begin{aligned} p(c, t|\lambda, \alpha, \gamma) &= p(c|t, \alpha, \gamma)p(t|\lambda) \\ &= \begin{cases} \exp(-\lambda), & \text{if } t = 0, \\ \frac{c^{\alpha t-1}}{\gamma^{t\alpha}\Gamma(t\alpha)} \exp\left(-\frac{c}{\gamma}\right) \times \frac{\lambda^t}{t!} \exp(-\lambda), & \text{if } t > 0. \end{cases} \end{aligned} \tag{4}$$

Usually, one would integrate out T from Equation (4) to obtain a marginal distribution of C. However, the infinite summand $p(c|\lambda, \alpha, \gamma) = \sum_{t=0}^{\infty} p(c, t|\lambda, \alpha, \gamma)$ does not have a closed-form representation. We can only use approximation approaches, such as series expansion [23] or Fourier inversion [24], to approximate the infinite number of terms. Although several studies have conducted statistical estimation and inference for the marginal distribution of C based on the approximation [25,26], in this paper, our methodology uses the joint distribution of C and T. Our approach is related to the EM algorithm presented in [27], although we use a Bayesian data augmentation strategy.

The compound Poisson–Gamma distribution CPG$(\lambda, \alpha, \gamma)$ is equivalently known as the Tweedie distribution TW(μ, ϕ, p) when $1 < p < 2$. The Tweedie parametrization gradually shifts from a Poisson distribution to a Gamma distribution as p increases. Building a compound Poisson–Gamma generalized linear model in terms of the parameters of TW(μ, ϕ, p) is easier than the original CPG$(\lambda, \alpha, \gamma)$. Thus, it is critical to know the unique relationship between two sets of parameters (μ, ϕ, p) and $(\lambda, \alpha, \gamma)$ as follows:

$$\begin{cases} \lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \\ \alpha = \frac{2-p}{p-1}, \\ \gamma = \phi(p-1)\mu^{p-1}, \end{cases} \quad \Leftrightarrow \quad \begin{cases} \mu = \lambda\alpha\gamma, \\ \phi = \frac{\lambda^{1-p}(\alpha\gamma)^{2-p}}{2-p}, \\ p = \frac{2+\alpha}{1+\alpha}. \end{cases} \tag{5}$$

Smyth [28] provides a detailed description of the computational process.

Returning to the communication scores $C_{i,j}^k$, we will utilize the Tweedie parameterization for model specification. The mean $\mu_{i,j}^k$ is parameterized through Equation (2) using a log-link function for $g()$; hence, the parameters of interest β_{g_1,g_2} determine this mean of $C_{i,j}^k$. The other parameters ϕ and p are global, and their values are shared across all (i,j) pairs of spots.

2.4. Model Inference

2.4.1. Parameter Estimation

Considering the complexity and intractability of the proposed model, a Bayesian approach for inference will be a good choice. We use the Hamiltonian Monte Carlo (HMC) algorithm to make the sampling more efficient than the usual Gibbs sampling [29]. Because HMC requires the gradient of the log-posterior density function, we derived the closed-form solutions for it in Appendix A. The closed-form solutions are computationally less intensive and hence much faster than iterative methods.

As part of the Bayesian model specification, prior distributions for all parameters must be specified. Given the limited information available on the parameters, it is often preferable to choose weakly informative priors that offer both convenience and simplicity. Thus, the intercept term will have a disperse $\mathrm{N}(0, 100^2)$ prior, and the remaining coefficients β_{g_1,g_2} have $\mathrm{N}(0,1)$ priors. As for the other two Tweedie parameters ϕ $(\phi > 0)$ and p $(1 < p < 2)$, we transform them to $\log\phi$ and $\theta = \log\left(\frac{p-1}{2-p}\right)$ and assign a normal distribution and a logistic distribution as priors, respectively. Overall, the priors are

$$\begin{aligned} \beta_0 &\sim \mathrm{N}(0, 100^2), \\ \beta_{g_1,g_2} &\overset{iid}{\sim} \mathrm{N}(0,1), g_1 = 1,2,\ldots,G, g_2 = 1,2,\ldots,G, \\ \log\phi &\sim \mathrm{N}(0, 10^2), \\ \theta = \log\left(\frac{p-1}{2-p}\right) &\sim \mathrm{Logistic}(0,1). \end{aligned} \tag{6}$$

In addition to the parameters specified above, our methodology relies on the spatial tuning parameter ρ in Equation (2). This parameter is unknown; however, we can fit multiple models with varying values of ρ and select the most suitable model by considering model selection statistics.

Equation (2) reflects a mixed-effect model that accounts for both fixed and random effects. The random effects ν_i^L and ν_j^R are assumed to be independently and identically distributed according to the standard normal distribution $\mathrm{N}(0,1)$. In an initialization step, we fit this mixed-effects model using a Newton–Raphson algorithm on the posterior distribution. However, since the random effects are not our primary parameters of interest, we treat the estimated $\boldsymbol{\nu}^L$ and $\boldsymbol{\nu}^R$ as fixed parameters in the main Bayesian framework to reduce the computational burden.

As noted previously, we are using the joint distribution of (C, T) as the relevant likelihood function in our Bayesian model specification since it has a closed-form representation. Thus, every observation $C_{i,j}$ has a corresponding unknown latent variable $T_{i,j}$, and our HMC algorithm includes a data augmentation Gibbs step to sample values of T given the observed c and the current parameter values. We note that when $C = 0$, T must be equal to zero. Ts for the non-zero Cs are sampled from probabilities proportion to $p_{\text{Poisson}}(t'|\lambda) \times p_{\text{Gamma}}(c|t'\alpha, \gamma)$ for $t' = 1, 2, \ldots, T_{\max}$. $T_{\max}$ is an adaptive parameter in our algorithm that is increased and decreased depending on how large the sampled Ts are relative to this maximum threshold.

In general, our Markov chain Monte Carlo (MCMC) sampling algorithm consists of two steps. In one step, we apply HMC to jointly update the vector of model parameters $(\beta_0, \boldsymbol{\beta}_{g1,g2}, \phi, p)$ for the current values of T. The other step updates the augmentation parameters T for the non-zero communications, given the current parameter values. We note that there are often few changes to the T values, so it is more computationally efficient to run multiple steps of HMC for each update to the augmentation parameters. Typically, we consider 10 HMC steps for every single update of T.

During the implementation of the MCMC procedure, we first conduct a preliminary run of 6500 iterations to tune the maximum threshold $T_{\max}$ and the covariance matrix of the momentum variables of HMC, so that the acceptance rate of all parameters can be kept around 45% to 65%. We then run an additional 13,500 iterations and discard the first 3500 iterations as the burn-in period, at which point approximate convergence is achieved. This results in 10,000 retained samples, and the inference is made using this collection.

2.4.2. Hypothesis Testing

As previously mentioned, the regression coefficients β_{g_1,g_2} from Equation (1) are the main parameters of interest because they reflect the association between the cell-type communication and the LR interaction. Thus, for each coefficient, we test the null hypothesis $H_0 : \beta_{g_1,g_2} = 0$ against the alternative hypothesis $H_a : \beta_{g_1,g_2} \neq 0$. We calculate the mean and variance of the samples β_{g_1,g_2} from the inference period of HMC and then generate the statistic $W = \frac{\hat{\beta}^2_{g_1,g_2}}{\mathrm{var}(\beta_{g_1,g_2})}$ for a Wald test. As stated in [30], the standard Bayesian large sample theory implies that the test statistic W approximately follows an asymptotic $\chi^2(1)$ distribution under the null hypothesis, and we can easily obtain a pseudo-p-value by considering the tail probability beyond the W associated with the estimated coefficient $\hat{\beta}_{g_1,g_2}$. The pseudo-p-values across all interactions of cell types will be adjusted for multiple hypothesis testing using the false discovery rate (FDR) correction [31]. In this way, the inference will be considered in a frequentist framework for ease of interpretation. In this paper, we declare a CCC significant on an LR pair if the FDR adjusted p-value is less than 0.05.

3. Results

We validated our modeling strategy through extensive simulation studies using data generated separately from two distinct models: our proposed compound Poisson–Gamma model and a pseudo-hurdle Gamma model. To accurately represent the spatial positions, we created a panel of 100 spots arranged in a 10 × 10 grid. Due to a lack of a comparable model structure among the existing CCC methods, we can only compare our model to variations to the structure of our BATCOM model. In the subsequent sections, we present the results of our simulation studies in Figures 2–5. For more detailed numerical results, please refer to Appendix B. Following that, we applied the proposed model to a real dataset and compared the results with other commonly used spatial CCC methods.

3.1. Simulation Study

3.1.1. Data Generated from the Compound Poisson-Gamma Model

To evaluate performance, we applied our method to simulated data generated from the proposed Tweedie model. Across simulations, we varied the number of cell groups G, the communication constraint parameter ρ, the sparsity rate of coefficients δ, and the two Tweedie parameters ϕ and p. For the proportions of cell types $M_{i,1}, \ldots, M_{i,G}$, we randomly generated each element from $\mathrm{U}(0,1)$ and then re-scaled each row so they sum to 1. Based on the regression structure, a dense vector of original coefficients $\boldsymbol{\beta}_O$ was independently sampled from $\frac{1}{2}\mathrm{U}(0.1, 0.5) + \frac{1}{2}\mathrm{U}(-0.5, -0.1)$, and the random effects $\boldsymbol{\nu}^L$ and $\boldsymbol{\nu}^R$ were sampled from $\mathrm{N}(0, 0.4)$ independently. Considering the sparsity rates of coefficients, we randomly picked $\delta \times G^2$ coefficients (excluding the intercept) in $\boldsymbol{\beta}_O$ and set them as zeros to get $\boldsymbol{\beta}_\delta$. To make different scenarios comparable, we fixed the $\boldsymbol{\beta}_\delta$, $\boldsymbol{\nu}_L$, and $\boldsymbol{\nu}_R$ across all corresponding scenarios. We also added -2 to the intercept to make the mean value $\boldsymbol{\mu} = \exp(\boldsymbol{X}\boldsymbol{\beta}_\delta + \boldsymbol{\nu}_L + \boldsymbol{\nu}_R)$ similar to real SRT data. For each combination of parameters, we generated 100 different simulated datasets to avoid uncertainty and detect variability.

For each dataset, the inference of BATCOM was performed by estimating and fixing the random effects, running the MCMC algorithm, and considering FDR-corrected Wald tests as discussed in Section 2.4. After obtaining these results, we generated a confusion matrix for each simulated dataset by comparing the results with the true values. A true positive (TP) was recorded when the estimate had the same sign as the true value and the adjusted p-value was less than 0.05. Similarly, a true negative (TN) was counted when the adjusted p-value was larger than 0.05, and the true value was zero. Conversely, a false negative (FN) was registered if the adjusted p-value was greater than 0.05, but the true value was not zero. Any estimate with an adjusted p-value less than 0.05 for a true value of zero or an estimate with a sign different from the true value was counted as a false positive (FP). Using the confusion matrix from the analysis of each generated dataset, we calculated the true positive rate (TPR), false positive rate (FPR), and observed FDR. We also plotted the receiver operating characteristic (ROC) curve by setting different cutoffs of the adjusted p-value to compute the area under the curve (AUC). For each scenario, we calculated the mean and standard deviations of these measurements across 100 simulated datasets.

Because we treat the communication constraint parameter ρ as a tuning parameter and do not estimate it during MCMC, it is critical to check the performance of the proposed model when the estimated value is close to and far away from the true value. Figure 2 presents the results of the proposed models fitted using $\hat{\rho} = 0.2, 0.5, 0.8$ under the scenarios with the true values of $\rho = 0.2, 0.4, 0.6, \text{and}\, 0.8$, separately. Additionally, we applied a standard Bayesian model selection strategy to choose the best value for this tuning parameter. To that end, we consider the widely applicable information criterion (WAIC$_2$) [29] according to the formula

$$\text{WAIC}_2^{\rho} = 2\sum_{i=1}^{n} \text{var}(\log p(c_i|\rho, \boldsymbol{\beta}^s, \phi^s, p^s)) - 2\sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(c_i|\rho, \boldsymbol{\beta}^s, \phi^s, p^s)\right),$$

where $i = 1, 2, \ldots, n$ reflects the observation and $s = 1, 2, \ldots, S$ represents the iteration of the inference period of HMC. For each dataset, we fit the model under $\hat{\rho} = 0.2, 0.5, 0.8$ and select the $\hat{\rho}$ that yields the lowest WAIC$_2$ as the BEST model choice.

In Figure 2, the results indicate that when the $\hat{\rho}$ used to fit the model is close to the true value, the performance is excellent, with high TPR and AUC, as well as low FPR and observed FDR. We also consider the value of $\hat{\rho} = 0.5$ as a useful default choice of the distance tuning parameter since it straddles the expected range of communication parameters $(0, 1)$. Empirically, this value performs well across the full range of ρ with only minor degradation in the more extreme cases of $\rho = 0.2, 0.8$. Thus, if running only one version of the model, we recommend using $\hat{\rho} = 0.5$ as the tuning parameter. For the remainder of this paper, we use the default $\hat{\rho} = 0.5$ choice. When computational resources permit, we suggest trying a small collection of ρ and selecting the best model based on WAIC$_2$.

We then compared our Bayesian framework (BATCOM) with a frequentist framework (referred to as TWGAM) using the *gam* function from R package mgcv [32]. TWGAM is a frequentist implementation of our model structure with the same distribution assumption and design matrix (scaled by distance) but without any random effect terms. Moreover, we compared our model structure concerning the proportions of cell types in each spot with other algorithm structures considering just one cell type. Many existing CCC methods treat each spot as containing only one cell type without considering the heterogeneity in each spot. To mimic this phenomenon, we constructed a corresponding zero-one matrix. For each spot (i.e., each row of the M matrix), we assigned a 1 to the cell type with the maximum proportion and a 0 to the rest of the cell types (named MAXPROP). MAXPROP uses all the other features of our proposed BATCOM except that the membership proportions $M_{i,g}$ $(i = 1, \ldots, N; g = 1, \ldots, G)$ are binary. For this MAXPROP, we employed our overall framework, including the random effect imputation, the MCMC sampling strategy for parameter estimation, and the Wald tests for inference. In addition, we compared our distribution assumption with the binomial logistic distribution of non-zero values (named LOGISTICS), which does not account for random effects.

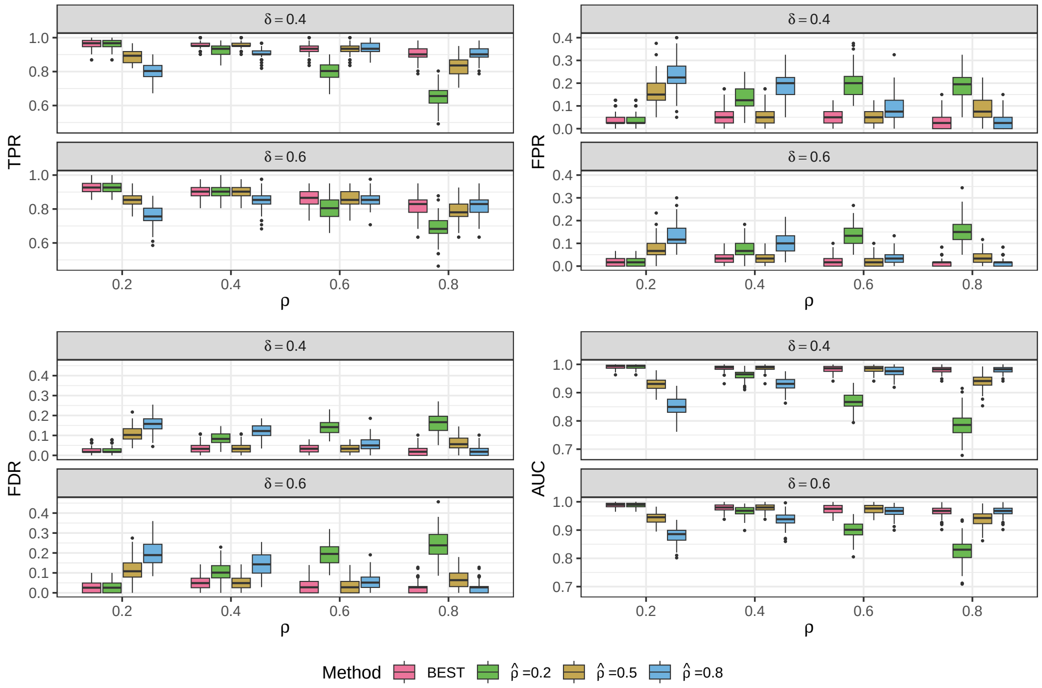

Figure 2. Results of BATCOM with different estimated ρ's based on the simulation data generated from the proposed compound Poisson–Gamma model. All scenarios were $G = 10$, $\phi = 3$, and $p = 1.5$. TPR: true positive rate; FPR: false positive rate; FDR: false discovery rate; AUC: area under the ROC curve.

Figure 3 displays the performance of different models using $\hat{\rho} = 0.5$ as the tuning parameter value under different scenarios, where the true communication constraint parameter ρ is either 0.4 or 0.6. It is easy to see that our proposed model (BATCOM) maintains a very robust performance across these different scenarios, consistently achieving high TPR and AUC when compared to three other models. Remarkably, BATCOM also effectively controls the FPR, and the observed FDR was around or below 0.05. In contrast, the frequentist framework (TWGAM), which shares the same distribution assumption as BATCOM, consistently performs worse, with much higher FPR and FDR. One possible explanation for this result is that TWGAM does not integrate prior knowledge about the parameters, and the failure to account for random effects could contribute to an increased FDR. The model that only considers one cell type for a spot (MAXPROP) is even worse than TWGAM in all aspects, including FPR, FDR, and AUC, with particularly high FPR and FDR. To some extent, these results suggest that existing CCC methods that ignore within-spot heterogeneity may produce numerous false discoveries. Additionally, although the model with the Bernoulli distribution assumption (LOGISTICS) generally maintains a small FPR and FDR, it has the lowest power (TPR), especially when we have a high value of the Tweedie parameter p. In other words, it tends to make too-conservative conclusions by inferring that most interactions among cell types are insignificant.

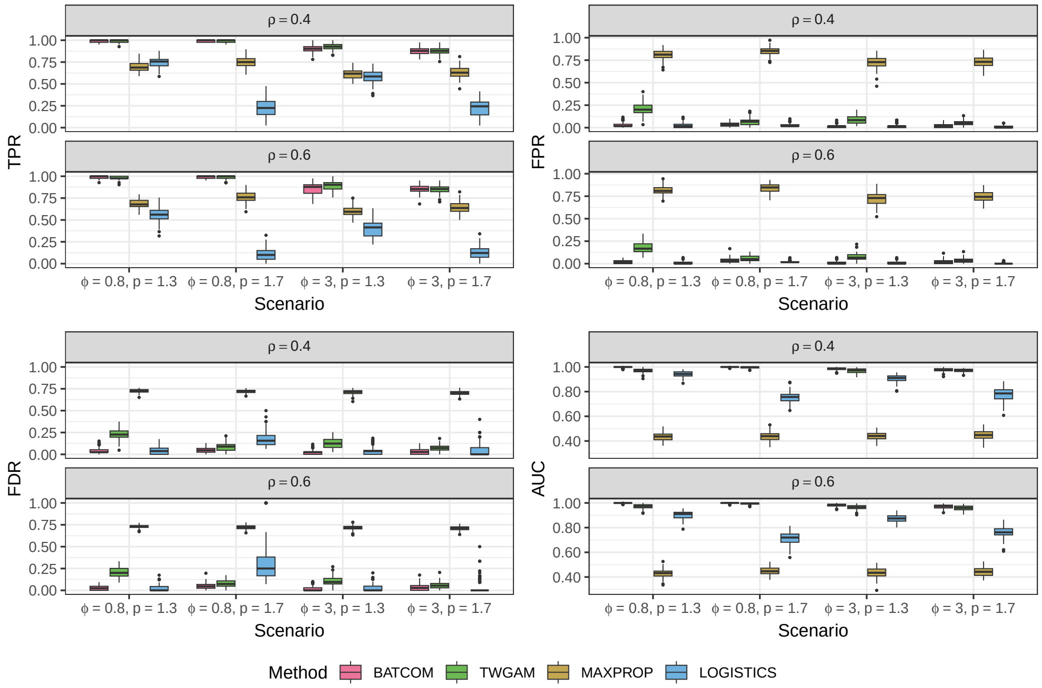

Figure 3. Results of different models based on the simulation data generated from the proposed compound Poisson–Gamma model. All scenarios were $G = 10$ and $\delta = 0.6$. All methods here used $\hat{\rho} = 0.5$. TPR: true positive rate; FPR: false positive rate; FDR: false discovery rate; AUC: area under the ROC curve.

In our previous simulation results, we considered all spot pairs for analysis. However, in real-world applications, we can effectively reduce the number of observations in the model by focusing only on spot pairs (i, j) that are within a specific distance threshold of each other, as mentioned earlier. To investigate the trade-off between efficiency and accuracy resulting from this reduction, we conducted an additional simulation experiment by varying the distance threshold and evaluating its impact on estimation accuracy. Specifically, we set the threshold values to 10 (i.e., including all spot pairs), 7, 5, and 3, as illustrated in Figure 4. It is evident that reducing the threshold to include fewer (i, j) spot pairs in the model inevitably affects estimation accuracy; however, the performance did not deviate significantly. The TPR and AUC showed a significant reduction when the threshold was set to 3, but the change remained below 10%. In contrast, the FPR and FDR exhibited no significant changes with varying thresholds. This observation suggests that reducing the number of spot pairs during model fitting leads to a more conservative decision-making process in our algorithm.

It is worth noting that the choice of the threshold is closely related to the selection of ρ. If ρ is large, only the closest spots have a substantial contribution, making a small threshold appropriate. However, when ρ is small, more spots contribute, and a larger threshold may be necessary to avoid excluding relevant spots. In Figure 4, we used $\hat{\rho} = 0.5$, resulting in a minimal weight for the exponential term of distance in Equation (2) when the distance between two spots exceeds 5.

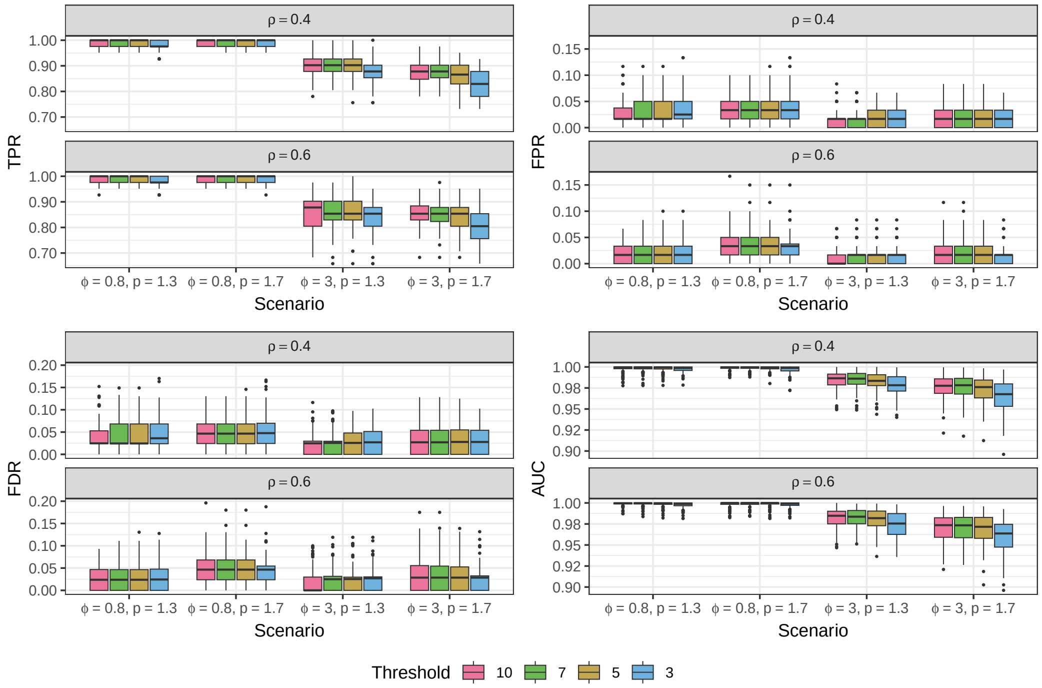

Figure 4. Results of BATCOM using different thresholds of distances of spots on the simulation data generated from the proposed compound Poisson–Gamma model. All scenarios were $G = 10$ and $\delta = 0.6$. All methods here used $\hat{\rho} = 0.5$. TPR: true positive rate; FPR: false positive rate; FDR: false discovery rate; AUC: area under the ROC curve.

3.1.2. Data Generated from the Pseudo-Hurdle Gamma Model

To complete our comprehensive evaluation of performance, we also simulated data from a structure that differs from our methodology while maintaining a comparable set of model parameters describing the relationship between cell type memberships and communication scores. We generated this additional simulated data from a pseudo-hurdle Gamma model. In the traditional hurdle model, one needs to have two different sets of coefficients to determine the probability of zeros and the distribution of non-zero values separately. Here, we generated the new data using the model

$$\Pr(C_{\text{new}} = 0) = 1/(1+\mu),$$

$$C_{\text{new}} | C_{\text{new}} > 0 \sim \text{Gamma}(\alpha_{\text{new}}, \frac{1+\mu}{\alpha_{\text{new}}}),$$

where μ depended on the same βs through Equation (2), the shape parameter α_{new} was randomly selected from a uniform distribution $\text{U}(0.5, 5)$, and the scale parameter $\gamma_{\text{new}} = \frac{1+\mu}{\alpha_{\text{new}}}$. In this way, the simulated data from this pseudo-hurdle Gamma model have the same mean μ as our proposed compound Poisson-Gamma distribution, ensuring that the parameter interpretation of β is comparable.

After switching to the pseudo-hurdle Gamma model as the data generator for our simulation scheme, we found that the results remained consistent with our previous findings using the compound Poisson–Gamma distribution. As shown in Figure 5, our proposed model outperforms other models, particularly in terms of its high TPR and low observed FDR in all scenarios. Even when faced with higher sparsity rates of coefficients and a larger number of cell types, our proposed model strikes a balance between identifying new discoveries and minimizing errors.

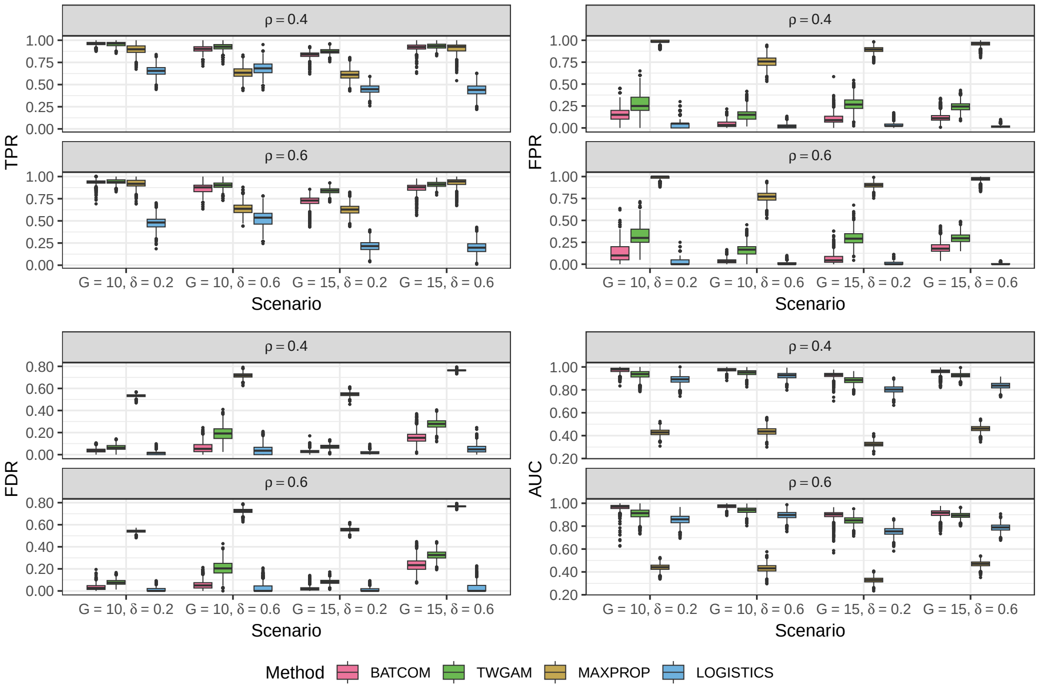

Figure 5. Results of different models based on the simulation data generated from the pseudo-hurdle Gamma model. All methods here used $\hat{\rho} = 0.5$. TPR: true positive rate; FPR: false positive rate; FDR: false discovery rate; AUC: area under the ROC curve.

3.2. Case Study

We applied our methods to the Visium spatial transcriptomics data of cutaneous squamous cell carcinoma (cSCC) [33], in which each spot contains multiple cells. Ji et al. [33] performed scRNA-seq on both tumors and normal skin and profiled SRT data on tumors simultaneously. As an example, we focused on the SRT data from replicate 2 of patient 2, which had a greater sequencing depth than other samples. This sample contains 1932 spots (after excluding spots with less than 100 genes detected) and 10,703 genes (after filtering out genes not expressed in at least 97.5% of the spots). To perform the upstream cell type deconvolution analysis, we utilized scRNA-seq data from the same patient as a reference and obtained the cell type proportion matrix using the full mode of RCTD [16].

Subsequently, we conducted a comparative analysis between our proposed method, employing $\hat{\rho} = 0.5$ as the tuning parameter, and other CCC algorithms tailored to SRT data, including Giotto [6], COMMOT [8], and SpaTalk [9]. To ensure a fair comparison, we utilized the same cell type proportion matrix from RCTD [16] ($G = 24$ cell types) and the same list of known LR pairs from CellTalkDB [19] (3398 pairs). However, Giotto [6] and COMMOT [8] required a single cell type to be specified for each spot, which posed a challenge for comparison. To address this, we assigned to each spot the cell type that had the highest proportion in the matrix, which introduced 12 cell types into the algorithms. Furthermore, we fit our model using the MAXPROP version, which is expected to be suboptimal for spot-based SRT data.

After filtering out LR pairs based on each algorithm's default rules, we found that our BATCOM and MAXPROP methods considered 712 LR pairs, whereas SpaTalk considered 515 LR pairs, Giotto considered 983 LR pairs, and COMMOT focused on 664 LR pairs. Regarding cell-type interactions, while there were a total of 576 interactions possible among the 24 cell types, we filtered out some interactions due to minimal or non-existent observations; this resulted in 364 interactions (out of 576) for BATCOM and 96 interactions (out of 144) for MAXPROP. SpaTalk does not consider interactions between the same cell type, leading to 552 interactions under its consideration. Meanwhile, Giotto and COMMOT dealt with 144 interactions due to the presence of 12 cell types.

Figure 6 illustrates the UpSet plot for the number of significant CCC on LR pairs identified by different methods. As we can see, our proposed method, BATCOM, and MAXPROP detected the third and second highest number of significant communications, respectively, following SpaTalk, which identified the most significant results. It is not surprising that our proposed model (BATCOM) identified fewer significant CCC results than SpaTalk, as SpaTalk does not correct the *p*-values for multiple comparisons. In contrast, Giotto and COMMOT found the fewest communication pairs. Notably, MAXPROP identified many more significant CCCs than BATCOM, which is consistent with the higher FDR observed in the preceding simulations. Compared to Giotto and COMMOT, which also utilize the modal cell type, MAXPROP found more significant communications. Specifically, one-third to one-half of the results from Giotto and COMMOT overlapped with MAXPROP but disagreed with BATCOM, which is the version of our model that uses all available information about the spot's cell type makeup.

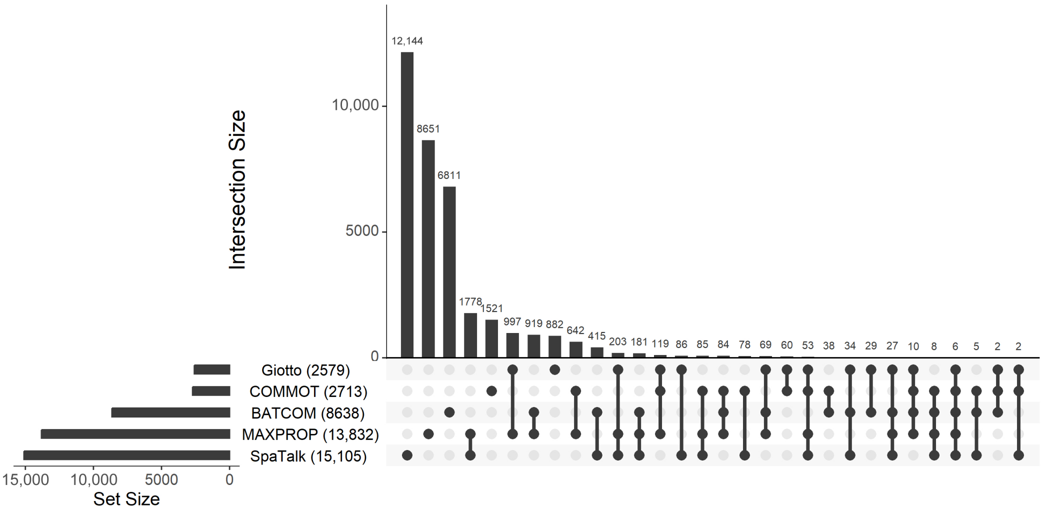

Figure 6. UpSet plot of significant CCC on LR pairs determined by different methods for the cutaneous squamous cell carcinoma data.

All methods share a common finding of six significant CCC pairs (Table 1). The results suggest that several LR interactions between different cell types may play a critical role in tumor-specific cellular crosstalk. Specifically, the ligand *SERPINE1*, which binds to the receptor *ITGB5*, has been found to promote tumor growth and angiogenesis in several types of cancer, including skin cancer [34]. Similarly, *THBS1*, *SDC4*, and *TLN1* have been linked to the development of metastasis and chemoresistance in skin cancer [35–37]. Notably, previous research has established *PLAU* and *ITGA5* as critical biomarkers for various types of squamous cell carcinoma [38–41]. Furthermore, the study by Fang et al. [39] suggests that *PLAU* affects the formation of inflammatory cancer-associated fibroblasts, which is consistent with the findings of our CCC analysis. These results emphasize the crucial role of specific LR interactions in cancer progression and highlight potential targets for therapeutic interventions.

Table 1. Significant CCC pairs shared in all methods.

Ligand	Receptor	Sender Cell Type	Receiver Cell Type
SERPINE1	*ITGB5*	Fibroblast	TSK
SERPINE1	*ITGB5*	TSK	Fibroblast
THBS1	*SDC4*	Fibroblast	TSK
PLAU	*ITGB5*	TSK	Fibroblast
TLN1	*ITGB5*	TSK	Tumor KC Diff
PLAU	*MRC2*	TSK	Fibroblast

TSK: tumor-specific keratinocytes; Tumor KC Diff: tumor-differentiating keratinocyte.

In addition to inferring CCC for a specific LR pair, exploring the overall cellular communication or communication within a specific signaling pathway based on SRT data can provide valuable insights. To demonstrate this, we aggregated the communication scores across all LR pairs and fit BATCOM using $C_{i,j} = \sum_k C_{i,j}^k$ as the outcome variable in Equation (2).

Figure 7 depicts the overall significant CCCs, indicating that cancer-related cells communicate closely with each other. Specifically, normal-differentiating keratinocytes (KC) exhibit positive communication with tumor-differentiating KC and plasmacytoid dendritic cells (PDC) while showing negative communication with fibroblasts. Notably, Ji et al. [33] found that the subpopulations of tumor KCs (basal, cycling, and differentiating) closely resemble the normal KC subpopulations, and they identified a fourth major tumor KC subpopulation, called tumor-specific keratinocytes (TSK), that exclusively exists in tumor skin and distinguishes itself from other tumor cells. Furthermore, Ji et al. [33] discovered that TSK and tumor-basal KC are both present in the leading edge of the tumor. Our results are consistent with these findings, as Figure 7 shows that tumor-basal KC frequently communicates with TSK.

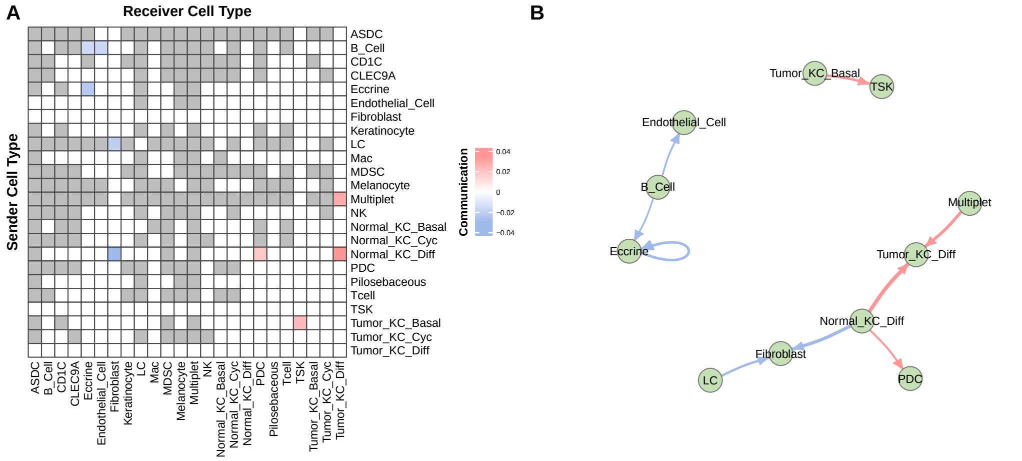

Figure 7. Overall CCC in the cutaneous squamous cell carcinoma data estimated by BATCOM. (**A**) Heatmap of CCC between the sender cell types and the receiver cell type. The gray blocks are the interactions of cell types that have been filtered out before fitting the model. The white blocks represent insignificant CCC. The colored blocks represent the significant CCC. (**B**) Network of CCC. The edge width reflects the strength of communication. The edge color shows the direction of the association.

4. Discussion

In this paper, we present a generalized linear regression model for inferring CCC based on LR interactions. Our model offers high flexibility in fitting both the single-cell resolution SRT data and spot-based SRT data. A significant challenge in spot-based SRT data is the presence of cell type mixtures in each spot, which we address by assuming that the mean spot-to-spot communication score is a convolution of possible interactions between cell types at the sender spot and the receiver spot. Our proposed model takes advantage of the regression model's properties to naturally handle communication between different cell types simultaneously, while also directly providing the direction of the association between CCC and LR interaction. Furthermore, our approach explicitly models the decreasing ability to communicate as the distance between cells or spots increases, differing from other algorithms that employ an arbitrary threshold to restrict communication.

Due to the limited information available on the parameters, our detailed Bayesian algorithm assumes the prior distributions defined in Equation (6). It is crucial to recognize that the choice of different prior distributions can lead to diverse model performances. To explore this, we also examined alternative priors such as $\mathrm{N}(0, 0.1^2)$ and $\mathrm{N}(0, 0.01^2)$ for the regression coefficients β_{g_1,g_2}, and we observed that the model's inference regarding significant connections remained robust relative to using the default $\mathrm{N}(0,1)$ prior. However, if necessary, the prior standard deviation can be easily adjusted in practice.

When comparing Bayesian and frequentist inference with the same distributional assumptions, we found that Bayesian inference provides more accurate estimation with lower FDR (Figure 3). However, MCMC algorithms can be time-consuming. For instance, in the cSCC case study, BATCOM and MAXPROP had an average running time of 67.38 and 19.80 min, respectively, per LR pair. In contrast, COMMOT and Giotto had average running times of 2.16 and 0.18 min, respectively, for each LR pair. SpaTalk had a different approach, identifying significant LR pairs for each cell type interaction, with an average running time of 4.38 min per interaction. To reduce the computational burden, one potential solution is to employ a threshold to control the number of included (i,j) spot or cell pairs. Although this approach inevitably impacts estimation accuracy, our simulation results (Figure 4) demonstrate that the resulting influence on performance will be minimal if using a moderate or large $\hat{\rho}$, such as 0.5. Alternatively, further exploration could focus on developing a more precise frequentist inference framework for the compound Poisson-Gamma distribution.

Currently, in this paper, we have defined the communication score as a product of the arithmetic mean of the ligand expressions at the sender spot and the mean receptor expression of the receiver spot. This simple approach implicitly assumes equal importance across the subunits of the LR pair, which may not always be the case. Certain subunits could have varying weights or specific distributions, necessitating a more sophisticated strategy in the future to accurately account for their expression. Moreover, we only considered the simple multiplication of ligands and receptors as the communication score between cells/spots (Equation (1)), while some other algorithms, such as CellChat [5], consider more complex relationships between ligands and receptors, including agonists and antagonists. It is certainly possible to design more intricate communication scores between cells/spots by accounting for these relationships. Given the versatility of our model, we can apply our approach directly to communication scores that have numerous zeros and positive continuous data, regardless of their complexity. This flexibility enables us to adapt our method to various scenarios and extend its applicability in future studies.

As previously mentioned, the tuning parameter ρ in Equation (2) is responsible for controlling the rate of decay of spot-to-spot communication as the distance between two spots increases. The appropriate value of ρ depends on the distance unit and potential communication assumption used in a specific tissue. While the parameter is not estimated during MCMC, we recommend using $\hat{\rho} = 0.5$ as a default value based on our simulation study results (Figure 2). For those with more computational resources, we suggest experimenting with different values of ρ and selecting the best one based on WAIC$_2$. For this manuscript, the utilization of Euclidean distance in our proposed model to account for the spatial proximity of cells or spots is specifically due to the current SRT data being derived from tissue slices relying on Cartesian coordinates. If future advancements in SRT technology enable the measurement of tissue shapes beyond the current capabilities, it will become imperative to explore alternative distance measurements that are better suited for such scenarios.

Furthermore, in the proposed model, we assume a decreasing trend in the communication probability when the distance between cells/spots increases. However, long-distance signaling is also essential in biological activities [42]. Therefore, a more comprehensive consideration of the relationship between communication and distance should be a focus of future research.

The results presented in Figure 6 indicate that various algorithms yield highly divergent results, with each method exhibiting a substantial number of distinct significant CCC. This observation aligns with the prior work by Li et al. [43]. It should be noted that the inferior performance of Giotto and COMMOT in our study may be partially attributed to directly assigning the cell type with the highest proportion to each spot. However, it is important to highlight that these CCC algorithms only allow for one cell type per spot. To ensure a fair comparison, we implemented the MAXPROP version of our model structure, aligning with the basic design of these algorithms. While we are confident that our methodology is statistically rigorous and reliable, the significant disparities across the different methods make it difficult to determine the most appropriate method at the biological level. Adding to this challenge is the lack of ground truth in this research domain. Moreover, although we conducted simulation studies with two different distributions to evaluate our model's performance, these assessments still rely on the underlying structure of our algorithm. Thus, it is imperative to undertake further experimental investigations and validations of CCC analysis to determine the most appropriate method for this area of inquiry.

As the field of CCC analysis continues to grow with the generation of more SRT data, we believe that our proposed model will serve as a valuable approach for inferring cellular communication in a flexible and accurate way. Our innovative approach can bridge gaps in current CCC inference methods and provide a straightforward outcome. Future studies could explore the applicability of our model to other more complicated structures and further validate its effectiveness in real-world scenarios.

Author Contributions: Conceptualization, D.W. and S.D.; methodology, D.W., J.T.G., M.S. and S.D.; software, D.W.; formal analysis, D.W.; investigation, D.W., J.T.G., M.S. and S.D.; data curation, D.W.; writing—original draft preparation, D.W.; writing—review and editing, D.W., J.T.G., M.S. and S.D.; visualization, D.W.; supervision, J.T.G. and S.D.; project administration, S.D.; funding acquisition, S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by NIH grant 1UL1TR000064 to S.D. from the Center for Scientific Review.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The cutaneous squamous cell carcinoma data analyzed in this study are openly available in the NCBI GEO database repository (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE144240 (accessed on 3 October 2022). The code scripts utilized in this study can be accessed on https://github.com/dongyuanwu/BATCOM.

Acknowledgments: We thank the editor and the reviewers for providing critical comments on an earlier version of the manuscript. We thank Indranil Mukhopadyay for the helpful suggestion. We appreciate the staff at the University of Florida Research Computing Center for providing computational resources and adequate technological support for carrying out the bioinformatics analysis.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The Closed Form of the Gradient of the Log-Posterior Density Function

The log-posterior density function can be viewed as the sum of the log-likelihood function and the log-prior density function. We first focus on the log-likelihood function. According to Equations (4) and (5), we define $\boldsymbol{\eta} = \log \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ as the link function in the proposed regression model, where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_N)^T$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)^T$, and $\boldsymbol{X}$ is an $N \times (q+1)$ design matrix. Then, the log-likelihood function of the ith ($i = 1, 2, \ldots, N$) observation is below.

If $t_i = 0$ (i.e., $y_i = 0$),

$$\log p(y_i, t_i|\cdot) = -\lambda_i = -\frac{\mu_i^{2-p}}{\phi(2-p)} = -\frac{1}{2-p}\exp[(2-p)\eta_i - \log\phi],$$

and if $t_i > 0$ (i.e., $y_i > 0$),

$$\begin{aligned}
\log p(y_i, t_i|\cdot) &= -\lambda_i + t_i \log \lambda_i - \log t_i! - \log \Gamma(t_i\alpha) - t_i\alpha \log \gamma_i + (\alpha t_i - 1)\log y_i - \frac{y_i}{\gamma_i} \\
&= -\frac{1}{2-p}\exp[(2-p)\eta_i - \log\phi] + t_i[(2-p)\eta_i - \log\phi - \log(2-p)] - \log t_i! \\
&\quad - \log\Gamma(t_i\alpha) - t_i\alpha[\log\phi + \log(p-1) + (p-1)\eta_i] + \alpha t_i \log y_i - \log y_i \\
&\quad - \frac{y_i}{p-1}\exp[(1-p)\eta_i - \log\phi] \\
&= -\frac{1}{2-p}\exp[(2-p)\eta_i - \log\phi] - t_i[\log\phi + \log(2-p)] - \log t_i! \\
&\quad - \log\Gamma(t_i\alpha) - t_i\alpha[\log\phi + \log(p-1)] + \alpha t_i \log y_i - \log y_i \\
&\quad - \frac{y_i}{p-1}\exp[(1-p)\eta_i - \log\phi]
\end{aligned}$$

Because the parameters sampled in HMC are $\boldsymbol{\beta}$, $\log\phi$, and $\theta = \log\left(\frac{p-1}{2-p}\right)$, we need their gradients. To simplify the calculation, we only discuss the gradient of the ith ($i = 1, 2, \ldots, N$) contribution to the log-likelihood. The gradient of the overall log-likelihood can be computed by summing all the N individual log-likelihood contributions.

When $t_i = 0$ (i.e., $y_i = 0$),

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}\log p(y_i, t_i|\cdot) &= \frac{\partial}{\partial\eta_i}\log p(y_i, t_i|\cdot)\cdot\frac{\partial\eta_i}{\partial\beta_j} \\
&= -\frac{1}{2-p}\exp[(2-p)\eta_i - \log\phi](2-p)\cdot x_{ij} \\
&= -\exp[(2-p)\eta_i - \log\phi]\cdot x_{ij},
\end{aligned}$$

where x_{ij} is the ith row and $(j+1)$th column of the design matrix $\boldsymbol{X}$, and $j = 0, 1, \ldots, q$. Moreover, it is easy to know that

$$\begin{cases} \phi = \exp(\log\phi), \\ p = \frac{2e^\theta + 1}{e^\theta + 1}, \end{cases} \quad \Rightarrow \quad \begin{cases} \frac{\mathrm{d}}{\mathrm{d}\log\phi}\phi = \exp(\log\phi), \\ \frac{\mathrm{d}}{\mathrm{d}\theta}p = \frac{e^\theta}{(e^\theta+1)^2} = (p-1)(2-p). \end{cases}$$

Thus,

$$\frac{\partial}{\partial\log\phi}\log p(y_i, t_i) = -\frac{1}{2-p}\exp[(2-p)\eta_i - \log\phi](-1) = \frac{1}{2-p}\exp[(2-p)\eta_i - \log\phi],$$

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log p(y_i, t_i|\cdot) &= \frac{\partial}{\partial p}\log p(y_i, t_i|\cdot)\cdot\frac{\mathrm{d}p}{\mathrm{d}\theta} \\
&= -\frac{1}{(2-p)^2}\{\exp[(2-p)\eta_i - \log\phi](-\eta_i)(2-p) + \exp[(2-p)\eta_i - \log\phi]\} \\
&\quad \times (p-1)(2-p) \\
&= \frac{p-1}{2-p}\{\exp[(2-p)\eta_i - \log\phi]\eta_i(2-p) - \exp[(2-p)\eta_i - \log\phi]\}.
\end{aligned}$$

On the other hand, when $t_i > 0$ (i.e., $y_i > 0$),

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}\log p(y_i, t_i|\cdot) &= \frac{\partial}{\partial\eta_i}\log p(y_i, t_i|\cdot)\cdot\frac{\partial\eta_i}{\partial\beta_j} \\
&= \left\{-\exp[(2-p)\eta_i - \log\phi] - \frac{1}{p-1}y_i\exp[(1-p)\eta_i - \log\phi](1-p)\right\}\cdot x_{ij} \\
&= \{y_i\exp[(1-p)\eta_i - \log\phi] - \exp[(2-p)\eta_i - \log\phi]\}\cdot x_{ij},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \log\phi}\log p(y_i,t_i|\cdot) &= \frac{1}{2-p}\exp[(2-p)\eta_i-\log\phi]-t_i-t_i\alpha+\frac{1}{p-1}y_i\exp[(1-p)\eta_i-\log\phi],\\
\frac{\partial}{\partial\theta}\log p(y_i,t_i|\cdot) &= \frac{\partial}{\partial p}\log p(y_i,t_i|\cdot)\cdot\frac{\mathrm{d}p}{\mathrm{d}\theta}\\
&= \frac{p-1}{2-p}\{\exp[(2-p)\eta_i-\log\phi]\eta_i(2-p)-\exp[(2-p)\eta_i-\log\phi]\}\\
&\quad+\frac{t_i}{2-p}\cdot\frac{\mathrm{d}p}{\mathrm{d}\theta}-\frac{\mathrm{d}}{\mathrm{d}(t_i\alpha)}\log\Gamma(t_i\alpha)\cdot t_i\cdot\frac{\mathrm{d}\alpha}{\mathrm{d}p}\cdot\frac{\mathrm{d}p}{\mathrm{d}\theta}-t_i\log\phi\cdot\frac{\mathrm{d}\alpha}{\mathrm{d}p}\cdot\frac{\mathrm{d}p}{\mathrm{d}\theta}\\
&\quad-t_i\log(p-1)\cdot\frac{\mathrm{d}\alpha}{\mathrm{d}p}\cdot\frac{\mathrm{d}p}{\mathrm{d}\theta}-\frac{t_i\alpha}{p-1}\cdot\frac{\mathrm{d}p}{\mathrm{d}\theta}+t_i\log y_i\cdot\frac{\mathrm{d}\alpha}{\mathrm{d}p}\cdot\frac{\mathrm{d}p}{\mathrm{d}\theta}\\
&\quad-\frac{1}{(p-1)^2}\{y_i\exp[(1-p)\eta_i-\log\phi](-\eta_i)(p-1)\\
&\qquad\qquad-y_i\exp[(1-p)\eta_i-\log\phi]\}\cdot\frac{\mathrm{d}p}{\mathrm{d}\theta}\\
&= \frac{p-1}{2-p}\{\exp[(2-p)\eta_i-\log\phi]\eta_i(2-p)-\exp[(2-p)\eta_i-\log\phi]\}\\
&\quad+t_i(p-1)+\frac{\mathrm{d}}{\mathrm{d}(t_i\alpha)}\log\Gamma(t_i\alpha)\cdot t_i\cdot\alpha+t_i\alpha\log\phi+t_i\alpha\log(p-1)-t_i\alpha(2-p)\\
&\quad-t_i\alpha\log y_i+\frac{2-p}{p-1}\{y_i\exp[(1-p)\eta_i-\log\phi]\eta_i(p-1)\\
&\qquad\qquad+y_i\exp[(1-p)\eta_i-\log\phi]\},
\end{aligned}$$

where $\alpha=\frac{2-p}{p-1}$ and $\frac{\mathrm{d}\alpha}{\mathrm{d}p}=-\frac{1}{(p-1)^2}$.

Next, we can solve the gradient of the log-prior functions according to Equation (6):

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}\log p(\boldsymbol{\beta}) &= -\frac{\beta_j}{\sigma_j^2},\ \text{where } \sigma_j^2 \text{ is the prior variance of } \beta_j, j=0,1,\ldots,q,\\
\frac{\mathrm{d}}{\mathrm{d}\log\phi}\log p(\log\phi) &= -\frac{\log\phi}{100},\\
\frac{\mathrm{d}}{\mathrm{d}\theta}\log p(\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\log\left[\frac{e^\theta}{(e^\theta+1)^2}\right]=\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\theta-2\log(e^\theta+1)\right]=\frac{1-e^\theta}{1+e^\theta}.
\end{aligned}$$

Additionally, due to the parameter transformation, the last puzzle of the gradient of the log-posterior density function is the gradient of the corresponding log-Jacobian terms, i.e.,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\log\phi}\log J_\phi &== \frac{\mathrm{d}}{\mathrm{d}\log\phi}\log\phi=1,\\
\frac{\mathrm{d}}{\mathrm{d}\theta}\log J_p &== \frac{\mathrm{d}}{\mathrm{d}\theta}\log[(p-1)(2-p)]=\frac{\mathrm{d}}{\mathrm{d}\theta}\log\frac{e^\theta}{(e^\theta+1)^2}=\frac{1-e^\theta}{1+e^\theta}.
\end{aligned}$$

Finally, the gradient of the log-posterior density function should be

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}\log p(\cdot|y_i,t_i) &= \frac{\partial}{\partial\beta_j}\log p(y_i,t_i|\cdot)+\frac{\partial}{\partial\beta_j}\log p(\boldsymbol{\beta}),\\
\frac{\partial}{\partial\log\phi}\log p(\cdot|y_i,t_i) &= \frac{\partial}{\partial\log\phi}\log p(y_i,t_i|\cdot)+\frac{\partial}{\partial\log\phi}\log p(\log\phi)+\frac{\mathrm{d}}{\mathrm{d}\log\phi}\log J_\phi,\\
\frac{\partial}{\partial\theta}\log p(\cdot|y_i,t_i) &= \frac{\partial}{\partial\theta}\log p(y_i,t_i|\cdot)+\frac{\partial}{\partial\theta}\log p(\theta)+\frac{\mathrm{d}}{\mathrm{d}\theta}\log J_p.
\end{aligned}$$

Appendix B. The Detailed Numerical Results of Simulation Studies

Table A1. Detailed results of BATCOM with different estimated ρ values based on the simulation data generated from the proposed compound Poisson–Gamma model.

	$\delta = 0.4$				$\delta = 0.6$			
	TPR	**FPR**	**FDR**	**AUC**	**TPR**	**FPR**	**FDR**	**AUC**
$\phi = 3, p = 1.5, G = 10, \rho = 0.2$								
BEST	0.96 (0.03)	0.04 (0.03)	0.03 (0.02)	0.99 (0.01)	0.92 (0.03)	0.02 (0.02)	0.03 (0.03)	0.99 (0.01)
$\hat{\rho} = 0.2$	0.96 (0.03)	0.04 (0.03)	0.03 (0.02)	0.99 (0.01)	0.92 (0.03)	0.02 (0.02)	0.03 (0.03)	0.99 (0.01)
$\hat{\rho} = 0.5$	0.89 (0.04)	0.17 (0.06)	0.11 (0.04)	0.93 (0.02)	0.85 (0.05)	0.08 (0.04)	0.12 (0.05)	0.94 (0.02)
$\hat{\rho} = 0.8$	0.81 (0.05)	0.23 (0.07)	0.16 (0.04)	0.85 (0.03)	0.76 (0.06)	0.13 (0.06)	0.20 (0.07)	0.88 (0.03)
$\phi = 3, p = 1.5, G = 10, \rho = 0.4$								
BEST	0.96 (0.02)	0.06 (0.04)	0.04 (0.03)	0.99 (0.01)	0.89 (0.04)	0.03 (0.02)	0.05 (0.04)	0.98 (0.01)
$\hat{\rho} = 0.2$	0.93 (0.03)	0.13 (0.05)	0.08 (0.03)	0.96 (0.02)	0.90 (0.04)	0.08 (0.04)	0.11 (0.05)	0.97 (0.02)
$\hat{\rho} = 0.5$	0.96 (0.02)	0.06 (0.04)	0.04 (0.03)	0.99 (0.01)	0.89 (0.04)	0.03 (0.02)	0.05 (0.04)	0.98 (0.01)
$\hat{\rho} = 0.8$	0.91 (0.03)	0.20 (0.06)	0.12 (0.03)	0.93 (0.02)	0.85 (0.05)	0.10 (0.05)	0.14 (0.06)	0.94 (0.02)
$\phi = 3, p = 1.5, G = 10, \rho = 0.6$								
BEST	0.93 (0.04)	0.05 (0.03)	0.03 (0.02)	0.98 (0.01)	0.86 (0.05)	0.02 (0.02)	0.04 (0.03)	0.97 (0.02)
$\hat{\rho} = 0.2$	0.80 (0.06)	0.20 (0.06)	0.14 (0.03)	0.87 (0.03)	0.80 (0.06)	0.14 (0.05)	0.20 (0.06)	0.90 (0.03)
$\hat{\rho} = 0.5$	0.93 (0.04)	0.05 (0.03)	0.03 (0.02)	0.98 (0.01)	0.86 (0.05)	0.02 (0.02)	0.04 (0.04)	0.97 (0.02)
$\hat{\rho} = 0.8$	0.94 (0.03)	0.09 (0.05)	0.05 (0.03)	0.97 (0.02)	0.86 (0.05)	0.03 (0.03)	0.05 (0.04)	0.96 (0.02)
$\phi = 3, p = 1.5, G = 10, \rho = 0.8$								
BEST	0.91 (0.04)	0.03 (0.03)	0.02 (0.02)	0.98 (0.01)	0.81 (0.06)	0.02 (0.02)	0.03 (0.03)	0.97 (0.02)
$\hat{\rho} = 0.2$	0.65 (0.06)	0.19 (0.06)	0.16 (0.05)	0.79 (0.04)	0.69 (0.07)	0.15 (0.05)	0.24 (0.07)	0.83 (0.04)
$\hat{\rho} = 0.5$	0.83 (0.05)	0.09 (0.05)	0.06 (0.03)	0.94 (0.02)	0.78 (0.06)	0.04 (0.03)	0.07 (0.05)	0.94 (0.03)
$\hat{\rho} = 0.8$	0.91 (0.04)	0.03 (0.03)	0.02 (0.02)	0.98 (0.01)	0.81 (0.06)	0.02 (0.02)	0.03 (0.03)	0.97 (0.02)

TPR: true positive rate; FPR: false positive rate; FDR: false discovery rate; AUC: area under the ROC curve.

Table A2. Detailed results of different models based on the simulation data generated from the proposed compound Poisson–Gamma model. All methods here used $\hat{\rho} = 0.5$.

	$\rho = 0.4$				$\rho = 0.6$			
	TPR	**FPR**	**FDR**	**AUC**	**TPR**	**FPR**	**FDR**	**AUC**
$\phi = 0.8, p = 1.3, G = 10, \delta = 0.6$								
BATCOM	0.99 (0.01)	0.03 (0.03)	0.04 (0.04)	1.00 (0.00)	0.99 (0.02)	0.02 (0.02)	0.03 (0.02)	1.00 (0.00)
TWGAM	0.99 (0.02)	0.20 (0.06)	0.23 (0.06)	0.97 (0.02)	0.98 (0.02)	0.18 (0.06)	0.21 (0.06)	0.97 (0.02)
MAXPROP	0.69 (0.06)	0.81 (0.06)	0.73 (0.02)	0.43 (0.04)	0.68 (0.05)	0.81 (0.05)	0.73 (0.02)	0.43 (0.04)
LOGISTICS	0.75 (0.06)	0.03 (0.03)	0.05 (0.04)	0.94 (0.03)	0.56 (0.08)	0.01 (0.01)	0.03 (0.04)	0.90 (0.03)
$\phi = 0.8, p = 1.7, G = 10, \delta = 0.6$								
BATCOM	0.99 (0.01)	0.04 (0.02)	0.05 (0.03)	1.00 (0.00)	0.99 (0.01)	0.03 (0.03)	0.05 (0.03)	1.00 (0.00)
TWGAM	0.99 (0.01)	0.06 (0.04)	0.08 (0.05)	1.00 (0.01)	0.98 (0.02)	0.06 (0.03)	0.08 (0.04)	0.99 (0.01)
MAXPROP	0.75 (0.06)	0.85 (0.05)	0.72 (0.02)	0.44 (0.04)	0.76 (0.06)	0.84 (0.05)	0.72 (0.02)	0.45 (0.03)
LOGISTICS	0.22 (0.10)	0.03 (0.02)	0.17 (0.09)	0.75 (0.04)	0.10 (0.07)	0.02 (0.01)	0.34 (0.25)	0.72 (0.05)
$\phi = 3, p = 1.3, G = 10, \delta = 0.6$								
BATCOM	0.90 (0.05)	0.01 (0.02)	0.02 (0.03)	0.98 (0.01)	0.86 (0.06)	0.01 (0.02)	0.02 (0.03)	0.98 (0.01)
TWGAM	0.92 (0.04)	0.09 (0.05)	0.13 (0.06)	0.97 (0.02)	0.89 (0.05)	0.07 (0.04)	0.10 (0.05)	0.96 (0.02)
MAXPROP	0.61 (0.06)	0.72 (0.07)	0.71 (0.02)	0.44 (0.03)	0.60 (0.06)	0.73 (0.07)	0.72 (0.02)	0.43 (0.04)
LOGISTICS	0.59 (0.08)	0.02 (0.02)	0.04 (0.04)	0.91 (0.03)	0.40 (0.09)	0.01 (0.01)	0.02 (0.04)	0.87 (0.03)

Table A2. *Cont.*

	$\rho = 0.4$				$\rho = 0.6$			
	TPR	**FPR**	**FDR**	**AUC**	**TPR**	**FPR**	**FDR**	**AUC**
				$\phi = 3, p = 1.7, G = 10, \delta = 0.6$				
BATCOM	0.87 (0.04)	0.02 (0.02)	0.04 (0.03)	0.98 (0.01)	0.85 (0.05)	0.02 (0.02)	0.04 (0.03)	0.97 (0.02)
TWGAM	0.88 (0.04)	0.04 (0.03)	0.06 (0.04)	0.97 (0.01)	0.85 (0.05)	0.04 (0.03)	0.06 (0.04)	0.96 (0.02)
MAXPROP	0.63 (0.06)	0.73 (0.06)	0.70 (0.03)	0.45 (0.04)	0.64 (0.06)	0.75 (0.06)	0.71 (0.03)	0.44 (0.03)
LOGISTICS	0.23 (0.09)	0.01 (0.01)	0.05 (0.08)	0.78 (0.06)	0.12 (0.08)	0.00 (0.01)	0.04 (0.08)	0.76 (0.05)

TPR: true positive rate; FPR: false positive rate; FDR: false discovery rate; AUC: area under the ROC curve.

Table A3. Detailed results of BATCOM using different thresholds of distances of spots on the simulation data generated from the proposed compound Poisson–Gamma model. All methods here used $\hat{\rho} = 0.5$.

	$\rho = 0.4$				$\rho = 0.6$			
	TPR	**FPR**	**FDR**	**AUC**	**TPR**	**FPR**	**FDR**	**AUC**
				$\phi = 0.8, p = 1.3, G = 10, \delta = 0.6$				
$D \leq 10$	0.99 (0.01)	0.03 (0.03)	0.04 (0.04)	1.00 (0.00)	0.99 (0.02)	0.02 (0.02)	0.03 (0.02)	1.00 (0.00)
$D \leq 7$	0.99 (0.01)	0.03 (0.03)	0.04 (0.03)	1.00 (0.00)	0.99 (0.01)	0.02 (0.02)	0.03 (0.03)	1.00 (0.00)
$D \leq 5$	0.99 (0.02)	0.03 (0.03)	0.04 (0.03)	1.00 (0.00)	0.99 (0.02)	0.02 (0.02)	0.03 (0.03)	1.00 (0.00)
$D \leq 3$	0.98 (0.02)	0.03 (0.03)	0.04 (0.03)	1.00 (0.00)	0.98 (0.02)	0.02 (0.02)	0.03 (0.03)	1.00 (0.00)
				$\phi = 0.8, p = 1.7, G = 10, \delta = 0.6$				
$D \leq 10$	0.99 (0.01)	0.04 (0.02)	0.05 (0.03)	1.00 (0.00)	0.99 (0.01)	0.03 (0.03)	0.05 (0.03)	1.00 (0.00)
$D \leq 7$	0.99 (0.01)	0.04 (0.02)	0.05 (0.03)	1.00 (0.00)	0.99 (0.01)	0.03 (0.03)	0.05 (0.03)	1.00 (0.00)
$D \leq 5$	0.99 (0.01)	0.04 (0.02)	0.05 (0.03)	1.00 (0.00)	0.99 (0.01)	0.03 (0.03)	0.05 (0.03)	1.00 (0.00)
$D \leq 3$	0.99 (0.02)	0.04 (0.03)	0.05 (0.04)	1.00 (0.01)	0.99 (0.02)	0.03 (0.03)	0.04 (0.03)	1.00 (0.00)
				$\phi = 3, p = 1.3, G = 10, \delta = 0.6$				
$D \leq 10$	0.90 (0.05)	0.01 (0.02)	0.02 (0.03)	0.98 (0.01)	0.86 (0.06)	0.01 (0.02)	0.02 (0.03)	0.98 (0.01)
$D \leq 7$	0.90 (0.05)	0.02 (0.02)	0.02 (0.03)	0.98 (0.01)	0.86 (0.06)	0.01 (0.02)	0.02 (0.03)	0.98 (0.01)
$D \leq 5$	0.90 (0.05)	0.02 (0.02)	0.03 (0.03)	0.98 (0.01)	0.86 (0.06)	0.01 (0.02)	0.02 (0.03)	0.98 (0.01)
$D \leq 3$	0.88 (0.05)	0.02 (0.02)	0.03 (0.03)	0.98 (0.01)	0.84 (0.06)	0.01 (0.02)	0.02 (0.03)	0.97 (0.02)
				$\phi = 3, p = 1.7, G = 10, \delta = 0.6$				
$D \leq 10$	0.87 (0.04)	0.02 (0.02)	0.04 (0.03)	0.98 (0.01)	0.85 (0.05)	0.02 (0.02)	0.04 (0.03)	0.97 (0.02)
$D \leq 7$	0.87 (0.04)	0.02 (0.02)	0.04 (0.03)	0.98 (0.01)	0.85 (0.05)	0.02 (0.02)	0.04 (0.03)	0.97 (0.02)
$D \leq 5$	0.86 (0.04)	0.02 (0.02)	0.04 (0.04)	0.97 (0.02)	0.84 (0.05)	0.02 (0.02)	0.03 (0.03)	0.97 (0.02)
$D \leq 3$	0.83 (0.05)	0.02 (0.02)	0.03 (0.03)	0.97 (0.02)	0.81 (0.06)	0.02 (0.02)	0.03 (0.03)	0.96 (0.02)

TPR: true positive rate; FPR: false positive rate; FDR: false discovery rate; AUC: area under the ROC curve.

Table A4. Detailed results of different models based on the simulation data generated from the pseudo-hurdle Gamma model. All methods here used $\hat{\rho} = 0.5$.

	$\rho = 0.4$				$\rho = 0.6$			
	TPR	**FPR**	**FDR**	**AUC**	**TPR**	**FPR**	**FDR**	**AUC**
				$G = 10, \delta = 0.2$				
BATCOM	0.96 (0.02)	0.16 (0.09)	0.04 (0.02)	0.97 (0.02)	0.94 (0.03)	0.13 (0.08)	0.03 (0.02)	0.96 (0.03)
TWGAM	0.96 (0.02)	0.27 (0.11)	0.07 (0.02)	0.93 (0.04)	0.94 (0.03)	0.32 (0.11)	0.08 (0.03)	0.91 (0.04)
MAXPROP	0.89 (0.05)	0.98 (0.02)	0.53 (0.01)	0.43 (0.03)	0.92 (0.05)	0.99 (0.02)	0.54 (0.01)	0.44 (0.03)
LOGISTICS	0.65 (0.06)	0.04 (0.05)	0.01 (0.02)	0.89 (0.03)	0.48 (0.07)	0.02 (0.03)	0.01 (0.01)	0.86 (0.04)
				$G = 10, \delta = 0.6$				
BATCOM	0.90 (0.04)	0.04 (0.03)	0.06 (0.04)	0.97 (0.02)	0.87 (0.05)	0.03 (0.03)	0.05 (0.04)	0.97 (0.02)
TWGAM	0.92 (0.04)	0.15 (0.06)	0.19 (0.06)	0.95 (0.02)	0.90 (0.04)	0.16 (0.06)	0.21 (0.06)	0.94 (0.03)
MAXPROP	0.63 (0.06)	0.76 (0.06)	0.72 (0.02)	0.44 (0.03)	0.64 (0.06)	0.77 (0.06)	0.72 (0.02)	0.43 (0.04)
LOGISTICS	0.69 (0.07)	0.02 (0.02)	0.04 (0.04)	0.92 (0.03)	0.53 (0.08)	0.01 (0.01)	0.03 (0.04)	0.90 (0.03)

Table A4. *Cont.*

	$\rho = 0.4$				$\rho = 0.6$			
	TPR	FPR	FDR	AUC	TPR	FPR	FDR	AUC
				$G = 15, \delta = 0.2$				
BATCOM	0.84 (0.03)	0.10 (0.05)	0.03 (0.01)	0.93 (0.02)	0.72 (0.05)	0.06 (0.05)	0.02 (0.02)	0.90 (0.04)
TWGAM	0.87 (0.03)	0.27 (0.08)	0.07 (0.02)	0.88 (0.03)	0.84 (0.03)	0.30 (0.08)	0.08 (0.02)	0.85 (0.03)
MAXPROP	0.61 (0.05)	0.89 (0.03)	0.55 (0.02)	0.33 (0.02)	0.63 (0.06)	0.90 (0.03)	0.56 (0.02)	0.33 (0.02)
LOGISTICS	0.45 (0.05)	0.03 (0.03)	0.02 (0.02)	0.80 (0.03)	0.21 (0.06)	0.01 (0.01)	0.01 (0.02)	0.75 (0.04)
				$G = 15, \delta = 0.6$				
BATCOM	0.92 (0.03)	0.12 (0.04)	0.16 (0.05)	0.96 (0.02)	0.87 (0.05)	0.18 (0.06)	0.24 (0.06)	0.91 (0.03)
TWGAM	0.93 (0.03)	0.24 (0.05)	0.28 (0.04)	0.93 (0.02)	0.91 (0.03)	0.30 (0.05)	0.32 (0.04)	0.89 (0.02)
MAXPROP	0.91 (0.06)	0.96 (0.03)	0.76 (0.01)	0.46 (0.03)	0.94 (0.05)	0.97 (0.02)	0.77 (0.01)	0.47 (0.02)
LOGISTICS	0.44 (0.06)	0.02 (0.01)	0.05 (0.04)	0.84 (0.03)	0.20 (0.06)	0.00 (0.01)	0.03 (0.04)	0.79 (0.03)

TPR: true positive rate; FPR: false positive rate; FDR: false discovery rate; AUC: area under the ROC curve.

References

1. Almet, A.A.; Cang, Z.; Jin, S.; Nie, Q. The landscape of cell–cell communication through single-cell transcriptomics. *Curr. Opin. Syst. Biol.* **2021**, *26*, 12–23. [CrossRef]
2. Armingol, E.; Officer, A.; Harismendy, O.; Lewis, N.E. Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* **2021**, *22*, 71–88. [CrossRef]
3. Efremova, M.; Vento-Tormo, M.; Teichmann, S.A.; Vento-Tormo, R. CellPhoneDB: Inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **2020**, *15*, 1484–1506. [CrossRef]
4. Cabello-Aguilar, S.; Alame, M.; Kon-Sun-Tack, F.; Fau, C.; Lacroix, M.; Colinge, J. SingleCellSignalR: Inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res.* **2020**, *48*, e55. [CrossRef]
5. Jin, S.; Guerrero-Juarez, C.F.; Zhang, L.; Chang, I.; Ramos, R.; Kuan, C.H.; Myung, P.; Plikus, M.V.; Nie, Q. Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **2021**, *12*, 1088. [CrossRef]
6. Dries, R.; Zhu, Q.; Dong, R.; Eng, C.H.L.; Li, H.; Liu, K.; Fu, Y.; Zhao, T.; Sarkar, A.; Bao, F.; et al. Giotto: A toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **2021**, *22*, 78. [CrossRef]
7. Cang, Z.; Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.* **2020**, *11*, 2084. [CrossRef] [PubMed]
8. Cang, Z.; Zhao, Y.; Almet, A.A.; Stabell, A.; Ramos, R.; Plikus, M.V.; Atwood, S.X.; Nie, Q. Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nat. Methods* **2023**, *20*, 218–228. [CrossRef] [PubMed]
9. Shao, X.; Li, C.; Yang, H.; Lu, X.; Liao, J.; Qian, J.; Wang, K.; Cheng, J.; Yang, P.; Chen, H.; et al. Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with SpaTalk. *Nat. Commun.* **2022**, *13*, 4429. [CrossRef] [PubMed]
10. Heydari, A.A.; Sindi, S.S. Deep learning in spatial transcriptomics: Learning from the next next-generation sequencing. *Biophys. Rev.* **2023**, *4*, 011306. [CrossRef]
11. Eng, C.H.L.; Lawson, M.; Zhu, Q.; Dries, R.; Koulena, N.; Takei, Y.; Yun, J.; Cronin, C.; Karp, C.; Yuan, G.C.; et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **2019**, *568*, 235–239. [CrossRef]
12. Wang, X.; Allen, W.E.; Wright, M.A.; Sylwestrak, E.L.; Samusik, N.; Vesuna, S.; Evans, K.; Liu, C.; Ramakrishnan, C.; Liu, J.; et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **2018**, *361*, eaat5691. [CrossRef] [PubMed]
13. Ståhl, P.L.; Salmén, F.; Vickovic, S.; Lundmark, A.; Navarro, J.F.; Magnusson, J.; Giacomello, S.; Asp, M.; Westholm, J.O.; Huss, M.; et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **2016**, *353*, 78–82. [CrossRef] [PubMed]
14. Salmén, F.; Ståhl, P.L.; Mollbrink, A.; Navarro, J.F.; Vickovic, S.; Frisen, J.; Lundeberg, J. Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. *Nat. Protoc.* **2018**, *13*, 2501–2534. [CrossRef]
15. Stickels, R.R.; Murray, E.; Kumar, P.; Li, J.; Marshall, J.L.; Di Bella, D.J.; Arlotta, P.; Macosko, E.Z.; Chen, F. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **2021**, *39*, 313–319. [CrossRef] [PubMed]
16. Cable, D.M.; Murray, E.; Zou, L.S.; Goeva, A.; Macosko, E.Z.; Chen, F.; Irizarry, R.A. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **2022**, *40*, 517–526. [CrossRef]
17. Elosua-Bayes, M.; Nieto, P.; Mereu, E.; Gut, I.; Heyn, H. SPOTlight: Seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* **2021**, *49*, e50. [CrossRef]
18. Sun, D.; Liu, Z.; Li, T.; Wu, Q.; Wang, C. STRIDE: Accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. *Nucleic Acids Res.* **2022**, *50*, e42. [CrossRef]
19. Shao, X.; Liao, J.; Li, C.; Lu, X.; Cheng, J.; Fan, X. CellTalkDB: A manually curated database of ligand–receptor interactions in humans and mice. *Briefings Bioinform.* **2021**, *22*, bbaa269. [CrossRef] [PubMed]

20. Finak, G.; McDavid, A.; Yajima, M.; Deng, J.; Gersuk, V.; Shalek, A.K.; Slichter, C.K.; Miller, H.W.; McElrath, M.J.; Prlic, M.; et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **2015**, *16*, 278. [CrossRef]
21. Sekula, M.; Gaskins, J.; Datta, S. Detection of differentially expressed genes in discrete single-cell RNA sequencing data using a hurdle model with correlated random effects. *Biometrics* **2019**, *75*, 1051–1062. [CrossRef] [PubMed]
22. Miao, Z.; Deng, K.; Wang, X.; Zhang, X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* **2018**, *34*, 3223–3224. [CrossRef] [PubMed]
23. Dunn, P.K.; Smyth, G.K. Series evaluation of Tweedie exponential dispersion model densities. *Stat. Comput.* **2005**, *15*, 267–280. [CrossRef]
24. Dunn, P.K.; Smyth, G.K. Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. *Stat. Comput.* **2008**, *18*, 73–86. [CrossRef]
25. Bonat, W.H.; Kokonendji, C.C. Flexible Tweedie regression models for continuous data. *J. Stat. Comput. Simul.* **2017**, *87*, 2138–2152. [CrossRef]
26. Mallick, H.; Chatterjee, S.; Chowdhury, S.; Chatterjee, S.; Rahnavard, A.; Hicks, S.C. Differential expression of single-cell RNA-seq data using Tweedie models. *Stat. Med.* **2022**, *41*, 3492–3510. [CrossRef]
27. Zhang, Y. Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models. *Stat. Comput.* **2013**, *23*, 743–757. [CrossRef]
28. Smyth, G.K. Regression analysis of quantity data with exact zeros. In Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management, Gold Coast, Australia, 17–19 July 1996; pp. 17–19.
29. Gelman, A.; Carlin, J.; Stern, H.; Dunson, D.; Vehtari, A.; Rubin, D. *Bayesian Data Analysis*, 3rd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2013. [CrossRef]
30. Matz, M.V.; Wright, R.M.; Scott, J.G. No control genes required: Bayesian analysis of qRT-PCR data. *PLoS ONE* **2013**, *8*, e71448. [CrossRef]
31. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
32. Wood, S.N. *Generalized Additive Models: An Introduction with R*; CRC Press: Boca Raton, FL, USA, 2017.
33. Ji, A.L.; Rubin, A.J.; Thrane, K.; Jiang, S.; Reynolds, D.L.; Meyers, R.M.; Guo, M.G.; George, B.M.; Mollbrink, A.; Bergenstråhle, J.; et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **2020**, *182*, 497–514. [CrossRef]
34. Klein, R.M.; Bernstein, D.; Higgins, S.P.; Higgins, C.E.; Higgins, P.J. SERPINE 1 expression discriminates site-specific metastasis in human melanoma. *Exp. Dermatol.* **2012**, *21*, 551–554. [CrossRef]
35. Jayachandran, A.; Anaka, M.; Prithviraj, P.; Hudson, C.; McKeown, S.J.; Lo, P.H.; Vella, L.J.; Goding, C.R.; Cebon, J.; Behren, A. Thrombospondin 1 promotes an aggressive phenotype through epithelial-to-mesenchymal transition in human melanoma. *Oncotarget* **2014**, *5*, 5782. [CrossRef] [PubMed]
36. Keller-Pinter, A.; Gyulai-Nagy, S.; Becsky, D.; Dux, L.; Rovo, L. Syndecan-4 in tumor cell motility. *Cancers* **2021**, *13*, 3322. [CrossRef] [PubMed]
37. Rezaie, Y.; Fattahi, F.; Mashinchi, B.; Kamyab Hesari, K.; Montazeri, S.; Kalantari, E.; Madjd, Z.; Saeednejad Zanjani, L. High expression of Talin-1 is associated with tumor progression and recurrence in melanoma skin cancer patients. *BMC Cancer* **2023**, *23*, 302. [CrossRef] [PubMed]
38. Chen, G.; Sun, J.; Xie, M.; Yu, S.; Tang, Q.; Chen, L. PLAU promotes cell proliferation and epithelial-mesenchymal transition in head and neck squamous cell carcinoma. *Front. Genet.* **2021**, *12*, 651882. [CrossRef]
39. Fang, L.; Che, Y.; Zhang, C.; Huang, J.; Lei, Y.; Lu, Z.; Sun, N.; He, J. PLAU directs conversion of fibroblasts to inflammatory cancer-associated fibroblasts, promoting esophageal squamous cell carcinoma progression via uPAR/Akt/NF-κB/IL8 pathway. *Cell Death Discov.* **2021**, *7*, 32. [CrossRef]
40. Zhou, C.; Shen, Y.; Wei, Z.; Shen, Z.; Tang, M.; Shen, Y.; Deng, H. ITGA5 is an independent prognostic biomarker and potential therapeutic target for laryngeal squamous cell carcinoma. *J. Clin. Lab. Anal.* **2022**, *36*, e24228. [CrossRef]
41. Fan, Q.C.; Tian, H.; Wang, Y.; Liu, X.B. Integrin-α5 promoted the progression of oral squamous cell carcinoma and modulated PI3K/AKT signaling pathway. *Arch. Oral Biol.* **2019**, *101*, 85–91. [CrossRef]
42. McMillen, P.; Oudin, M.J.; Levin, M.; Payne, S.L. Beyond neurons: Long distance communication in development and cancer. *Front. Cell Dev. Biol.* **2021**, *9*, 739024. [CrossRef]
43. Li, B.; Zhang, W.; Guo, C.; Xu, H.; Li, L.; Fang, M.; Hu, Y.; Zhang, X.; Yao, X.; Tang, M.; et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat. Methods* **2022**, *19*, 662–670. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.