A Zero-Inflated Hierarchical Generalized Transformation Model to Address Non-Normality in Spatially-Informed Cell-Type Deconvolution

Hunter J. Melton* Jonathan R. Bradley† Chong Wu ‡

June 24, 2024

Abstract

Oral squamous cell carcinomas (OSCC), the predominant head and neck cancer, pose significant challenges due to late-stage diagnoses and low five-year survival rates. Spatial transcriptomics offers a promising avenue to decipher the genetic intricacies of OSCC tumor microenvironments. In spatial transcriptomics, Cell-type deconvolution is a crucial inferential goal; however, current methods fail to consider the high zero-inflation present in OSCC data. To address this, we develop a novel zero-inflated version of the hierarchical generalized transformation model (ZI-HGT) and apply it to the Conditional AutoRegressive Deconvolution (CARD) for cell-type deconvolution. The ZI-HGT serves as an auxiliary Bayesian technique for CARD, reconciling the highly zero-inflated OSCC spatial transcriptomics data with CARD's normality assumption. The combined ZI-HGT + CARD framework achieves enhanced cell-type deconvolution accuracy and quantifies uncertainty in the estimated cell-type proportions. We demonstrate the superior performance through simulations and analysis of the OSCC data. Furthermore, our approach enables the determination of the locations of the diverse fibroblast population in the tumor microenvironment, critical for understanding tumor growth and immunosuppression in OSCC.

Keywords: Bayesian, Poisson distribution, Spatial transcriptomics, Uncertainty quantification

*Department of Statistics, Florida State University

†Department of Statistics, Florida State University. Corresponding author

‡Department of Biostatistics, University of Texas MD Anderson Cancer Center. Corresponding author

1 Introduction

Oral squamous cell carcinomas (OSCC) are the most common head and neck cancer, accounting for over 90% of oral cavity mucosal malignancies (Markopoulos, 2012). The insidious progression of OSCC, often undetectable until advanced stages due to the absence of early symptoms, leads to prognoses with an approximate 50% five-year survival rate (Markopoulos, 2012). Despite this, the advent of spatial transcriptomics (ST) technologies, which provide precise spatial localization of tissue-level gene expression (Moses and Pachter, 2022), has introduced new avenues for understanding the cellular and molecular landscapes within tumors. The recent OSCC ST dataset collected by Arora et al. (2023) has been particularly transformative, enabling the detailed characterization of the tumor microenvironment (TME) and illuminating the spatial distribution of cell types within the TME, thereby providing novel perspectives on tumorigenesis and cellular interactions (Moses and Pachter, 2022).

However, the application of ST in OSCC research encounters notable hurdles with spatial barcoding-based methods, such as the 10X Visium platform used to capture the OSCC ST data (Arora et al., 2023). Nearly all spatial barcoding-based methods collect expression measurements from tissue locations comprising a small to moderate number of single cells, which may represent diverse cell types (Ståhl et al., 2016; Rodriques et al., 2019; Stickels et al., 2021). Given that each assessed location contains a variety of cells, these techniques effectively quantify the average expression level of numerous cells at the location. Consequently, cell-type deconvolution, which probabilistically infers the proportion of cell types at each location, becomes a crucial analytical step for unraveling the spatial distribution of cell types and subsequent analyses (Liao et al., 2021; Rao et al., 2021).

Despite these advances, existing approaches often overlook the prevalent issue of zero

inflation in ST data. For example, CARD assumes a normal distribution for count-valued data, disregarding the zero-inflation and ties inherent in ST (Li et al., 2021; Zhao et al., 2022). The challenge is exemplified in the analysis of 12 OSCC samples (Arora et al., 2023), where the proportion of zero counts, or sparsity, of the samples ranges from 86%-91%, and the proportion of counts greater than 20 is less than 0.1%. This pattern of sparsity, echoed across numerous ST datasets (Moncada et al., 2020), poses analytical hurdles. While the high sparsity could be directly handled by the use of a zero-inflated Poisson model (Lambert, 1992) or zero-inflated Negative Binomial model (Greene, 1994), these solutions are both challenging and time-consuming, especially considering the size of the data—over 15 million data points for each sample. Consequently, few spatial transcriptomic analysis methods currently address zero-inflation (Zhao et al., 2022), and no cell-type deconvolution methods do so directly. Moreover, the process of uncertainty quantification (UQ) in cell-type deconvolution, a critical yet challenging task, often relies on model-based assumptions that may not align with the true distribution of the data. Fully Bayesian models present a method for deriving posterior credible intervals, thereby facilitating UQ, but their reliance on computationally intensive Markov chain Monte Carlo (MCMC) methods renders them impractical for the high dimensionality of ST datasets.

In classical statistics, when data diverge from the assumptions of a chosen model (often assumed normal), deterministic transformations, such as the Box-Cox (Box and Cox, 1964) or Yeo-Johnson power transformation (Yeo and Johnson, 2000), are frequently employed to align the data more closely with the assumptions preanalysis. However, these classical deterministic transformations have difficulty with data containing many ties (as is the case with the OSCC ST data), as the ties will remain present after applying the deterministic transformation, which results in data that are still non-normal. Contrastingly, the hier-

archical generalized transformation (HGT) (Bradley et al., 2020; Bradley, 2022; Bradley et al., 2023) introduces a probabilistic, rather than deterministic, transformation to data processing. The HGT is a new class of noisy transformations that breaks ties by adding a small amount of noise to the data, which is modeled via a Bayesian transformation. The HGT, however, has yet to be developed for zero-inflated data. Building on the HGT, we introduce the novel Zero-Inflated Hierarchical Generalized Transformation (ZI-HGT), an adaptation specifically designed to address the issues of zero inflation and ties in ST datasets. By integrating a type of zero-inflated Poisson model into the HGT framework, the ZI-HGT generates posterior replicates from the noisy transformation, effectively mitigating the impact of zeros and ties and ensuring the transformed data more closely aligns with the model assumption (such as the normality assumption in CARD). This novel combination of a zero-inflated model with HGT and CARD will produce a Bayesian hierarchical model (BHM) that can be implemented on a high-dimensional (over 15 million data points), multivariate (over the cell-types) spatial dataset, which is not possible with current fully Bayesian alternatives for zero-inflated multivariate spatial data (Musgrove et al., 2019).

The first step in implementing the ZI-HGT involves simulating posterior replicates of the transformation itself. The transformations are assumed to be unknown parameters in a hierarchical model that is known to overfit zero-inflated count data. Resultingly, the posterior replicates of the transformation are close to the original data and can reasonably be used in place of the original data. Then, the second step is to apply the chosen model (e.g., CARD) to each of the posterior replicates of the transformed data and calculate the average, standard deviation, and quantiles across the replicate draws to obtain the final results. The model can be applied in parallel across the multiple transformations produced in the first step, leading to an approach nearly as computationally efficient as the chosen model

(e.g., CARD). This two-step process efficiently addresses two crucial issues: accounting for uncertainty and model misspecification and enhancing deconvolution accuracy without compromising computational efficiency through parallel computing. Importantly, we can create pointwise credible intervals that allow for uncertainty quantification (UQ) without relying on untenable normal assumptions about the distribution of the ST data, which can make model-based UQ unreliable.

The rest of the manuscript is organized as follows: Section 2 delves into the motivating OSCC data. Preliminary concepts and the methodology are discussed in Sections 3 and 4, respectively. An empirical simulation study is presented in Section 5, which elucidates the viability of the model. Section 6 applies the model to OSCC data, presenting findings, and Section 7 provides conclusions and discussions. A schematic overview of the method can be found in Figure 1.

2 Data and Motivation

We analyze the oral squamous cell carcinoma spatial transcriptomic (ST) data (Arora et al., 2023), which consists of 12 samples surgically resected from the oral cavity from 10 patients. Post-surgery, the tissues were embedded in optimal cutting temperature (OCT) compound and stored at -80°C until the original study. Ten μm sections were obtained using a cryostat, hematoxylin and eosin (H&E) stained, and utilized for spatial transcriptomics. These sections were then annotated by a pathologist into nine distinct tissue sections: squamous cell carcinoma (SCC), lymphocyte-positive stroma, lymphocyte-negative stroma, normal mucosa, glandular stroma, muscle, keratin, and artery/vein, and an artifact layer. To prepare the samples for ST, the optimal permeabilization time was determined through tissue optimization. ST was then conducted on the 12 samples using the 10x Genomics

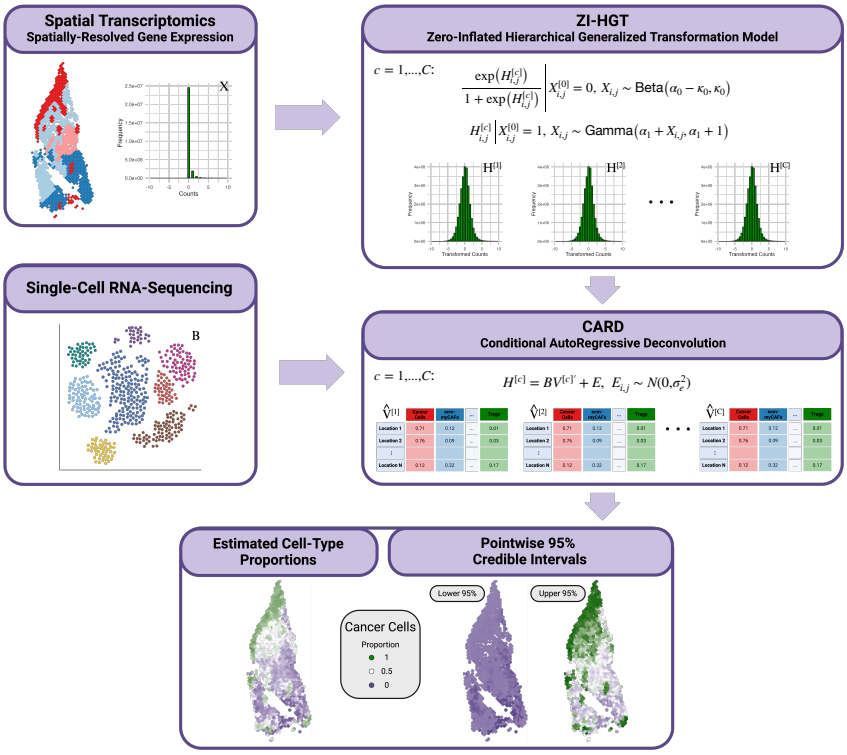


Figure 1: ZI-HGT + CARD Schematic Outline. The raw spatial transcriptomic data X is inputted into the ZI-HGT, generating C posterior replicates ($H^{[1]}$, ..., $H^{[C]}$), which are interpreted as transformations of the original data. These transformations better fit the assumptions of CARD, namely that the data are normally distributed. CARD is then applied to each transformed dataset and a corresponding single-cell RNA sequencing (scRNA-seq) dataset B, yielding cell-type proportion estimates ($\hat{V}^{[1]}$, ..., $\hat{V}^{[C]}$). These estimates are amalgamated into a single estimate $\hat{V}$.

Visium platform, which samples approximately 10 cells per location. Across all samples, reads were obtained for an average of $16,763$ unique transcripts over $2,073$ spatial locations. In addition to the ST data, our analysis incorporates a reference single-cell RNA-sequencing (scRNA-seq) dataset (Puram et al., 2017) (GSE103322), which is discussed in more detail in the Supplementary.

We consider OSCC ST Sample 1 as an example, but note that the patterns in the other samples are highly similar (see Supplementary Table 1). Sample 1 consists of read counts for 15,844 transcripts across 1,131 spatial locations. Supplementary Table 2 presents a frequency table of the gene expression read count values (see Supplementary Tables 3-13 for frequency tables of the other samples). We highlight two key features of the dataset. First, the extremely high zero inflation is noteworthy. Of the 17,919,564 elements in the spatial gene expression matrix for Sample 1, 16,384,736 (91.4%) are zeros. Second, the data naturally consists of counts, resulting in a significant number of ties. These characteristics underscore the limitations of current analytical approaches like CARD, which assumes that the ST data are normally distributed, an assumption rendered untenable by the zero-inflation and abundance of ties observed. As a result, estimated cell-type proportions will be less accurate than hoped for. In response to these challenges, we will introduce a novel zero-inflated HGT (a noisy transformation aimed at breaking the ties and addressing the zero-inflation) in the following section to address the non-normality of the data, rendering the normality assumption used in CARD more reasonable.

3 Preliminaries

3.1 Notation

Let X be the $G \times N$ gene expression matrix from one of the OSCC samples (Arora et al., 2023), containing transcript reads for G cell-type informative genes (discussed below) measured at N spatial locations, with the (i, j)-th element denoted $X_{i,j}$. Then, let $H(X)$, or H for brevity, be the noisy transformation of X, $X^{[0]}$ be a set of indicators for positive gene expression in X (i.e., $X^{[0]}_{i,j} = 1$ if $X_{i,j} > 0$), and $X^{(B)}$ be a data augmentation vari-

able that facilitates the implementation of a zero-inflated model without expensive Markov chain Monte Carlo (MCMC) sampling. We employ the corresponding single-cell RNA-sequencing data (Puram et al., 2017) as a deconvolution reference and select informative genes for distinguishing between cell types. Let B be the $G \times K$ single-cell gene expression matrix (often referred to as the reference basis matrix), where G is again the number of cell-type informative genes and K is the number of cell types. Lastly, let V be the $N \times K$ cell-type proportions matrix, where the rows of V contain the proportions of the K cell types for each spatial location. In the OSCC ST data, G and N may vary slightly from sample-to-sample, but $K = 14$.

Additionally, we introduce the following shorthand notation for probability distributions. "Point Mass" is a short-hand for the point mass distribution, and "TruncPoisson" is a short-hand for the zero-truncated Poisson distribution.

3.2 Review of CARD

CARD (Ma and Zhou, 2022) uses the non-negative matrix factorization model for X:

$$X = BV' + E, \tag{1}$$

where E is the $G \times N$ residual error matrix and $E_{i,j} \sim N(0, \sigma_e^2)$. B contains "cell-type informative" genes, which are classified as such if their mean expression level in any individual cell type has at least a 1.25 log-fold increase over the average expression level of every other included cell type. To consider the spatial colocalization of cell types, a conditional autoregressive (CAR) (Besag, 1974; Besag et al., 1991) modeling assumption is made for the columns of the cell-type proportions matrix V. For spatial location i and cell type k,

the full-conditional distribution for $V_{i,k}$ is given by

$$V_{i,k} = b_k + \phi \sum_{j \neq i} W_{i,j}(V_{j,k} - b_k) + \epsilon_{i,k}; \quad i = 1, ..., N, \; k = 1, ..., K, \tag{2}$$

where $V_{i,k}$ denotes the proportion of cells at location i belonging to cell type k, b_k is the cell-type intercept that represents the average cell-type composition for cell type k across all locations, and $W_{i,j}$ refers to the weight used for learning the cell-type composition at location i using the cell-type composition information at location j. The parameter ϕ represents the spatial autocorrelation, determining the strength of the spatial correlation in the cell-type composition, and $\epsilon_{i,k}$ represents the residual error such that $\epsilon_{i,k} \sim N(0, \sigma^2_{i,k})$. The CAR model represents $V_{i,k}$, the proportion of cells at location i belonging to cell type k, as a weighted sum of the proportions of cell type k at the other locations, allowing for the sharing of cell-type composition data across space. For a more detailed review, please refer to the Supplementary material.

Notably, Equation (1) assumes that the spatially resolved gene expression X is normally distributed, which does not align with the OSCC ST data due to the high level of sparsity and the prevalence of ties, as highlighted in Section 2. The minimum sparsity among the 12 OSCC ST samples is 86% (Supplementary Table 1), and the remaining data contains numerous ties (Supplementary Tables 2-13).

3.3 Review of HGT

Models like CARD assume the data is Gaussian, which is unrealistic considering the aforementioned presence of ties in the OSCC dataset. Adding noise to the data produces a new (transformed) dataset without ties, rendering the Gaussian assumption more reasonable. To this end, the HGT aims to add a small amount of noise to count data to break ties. The noise is added in a Bayesian way to fully account for the added uncertainty produced

via the noisy transformation. The HGT implements a Bayesian model that is known to overfit the original data (Bradley, 2022), and the posterior predictive replicates from the overfitted Bayesian model are interpreted as the noisy transformation. By "overfitting," we mean that the posterior replicates are close (in terms of their absolute difference) on average to a deterministic transformation of the data on some scale. The motivation for overfitting is that the noise added to the data should be "small" so as to avoid biasing results.

Recall that $H = H(X)$ denotes the noisy transformation of the original data X, so $H_{i,j}$ is the (i, j)-th element of the transformation H. The current use of the HGT would assume that $X_{i,j}\big|H_{i,j} \sim \text{Poisson}\,(H_{i,j})$ and $H_{i,j} \sim \text{Gamma}\,(\alpha, \kappa)$ so that the posterior distribution $H_{i,j}\big|X_{i,j} \sim \text{Gamma}\,(\alpha + X_{i,j}, \kappa + 1)$. When, for example, $\alpha \approx \kappa$, the posterior replicates of $H_{i,j}\big|X_{i,j}$ have mean $Z_{i,j}$ that overfits. CARD is then applied to the $G \times N$ matrix H in place of X. However, this current specification does not capitalize on the known zero-inflation present in the OSCC ST data, making it less than ideal. Simulations suggest that ignoring the zero-inflation in an HGT yields minimal improvement (Supplementary Figure 3).

4 Methodology

4.1 Zero-inflated hierarchical generalized transformation (ZI-HGT)

When applying the ZI-HGT to CARD (Ma and Zhou, 2022), there are two features of the OSCC data we seek to address: the high zero inflation and the ties. To do so, we consider two observed datasets, X, the original OSCC ST data matrix, and $X^{[0]}$. Recall that the $G \times N$ matrix $X^{[0]}$ has (i, j)-th element $X^{[0]}_{i,j}$ representing an indicator that positive gene

expression was observed for gene i at location j, i.e., $X_{i,j}^{[0]} = 1$ for positive gene expression and $X_{i,j}^{[0]} = 0$ otherwise. Here, the superscript "B" stands for Bernoulli, and the introduction of $X_{i,j}^{(B)}$ will facilitate the implementation of a Bayesian zero-inflated model with a posterior distribution that is easily sampled from directly. We then assume the following hierarchical model:

$$X_{i,j} \big| X_{i,j}^{[0]} = 1, H_{i,j} \sim \text{TruncPoisson}\left(H_{i,j}\right), \tag{3}$$

$$X_{i,j} \big| X_{i,j}^{[0]} = 0, H_{i,j}, X_{i,j}^{(B)} \sim \text{Point Mass}\left(0\right), \tag{4}$$

$$H_{i,j} \big| X_{i,j}^{[0]} = 0, X_{i,j}^{(B)} \sim \text{Point Mass}\left(X_{i,j}^{(B)}\right), \tag{5}$$

$$H_{i,j} \big| X_{i,j}^{[0]} = 1 \propto \left(1 - \exp\left(-H_{i,j}\right)\right) \text{Gamma}(\alpha_1, \alpha_1), \tag{6}$$

$$\left(\frac{\exp\left(X_{i,j}^{(B)}\right)}{1 + \exp\left(X_{i,j}^{(B)}\right)}\right) \big| \alpha_0, \kappa_0 \sim \text{Beta}(\alpha_0, \kappa_0), \tag{7}$$

$$X_{i,j}^{[0]} \big| X_{i,j}^{(B)} \sim \text{Bernoulli}\left(\frac{\exp\left(X_{i,j}^{(B)}\right)}{1 + \exp\left(X_{i,j}^{(B)}\right)}\right); \quad i = 1, ..., N, \; j = 1, ..., G, \tag{8}$$

where the shorthand "$H_{i,j} \big| X_{i,j}^{[0]} = 1 \propto \left(1 - \exp\left(-H_{i,j}\right)\right) \text{Gamma}(\alpha_1, \alpha_1)$" represents the conditionally specified and unnormalized prior

$$f(H_{i,j} \big| X_{i,j}^{[0]} = 1) = \left(1 - \exp\left(-H_{i,j}\right)\right) \frac{\alpha_1^{\alpha_1}}{\Gamma(\alpha_1} H_{ij}^{\alpha_1 - 1} \exp(-\alpha_1 H_{ij}),$$

where a gamma probability density function is multiplied by $\left(1 - \exp\left(-H_{i,j}\right)\right)$ to offset the normalizing constant in the truncated Poisson likelihood.

In Equation (3), the non-zero count-valued OSCC ST data is modeled as truncated Poisson (to be ≥ 1) distributed, and in Equation (4) the zero-valued OSCC ST data is modeled as a point mass at zero. Zero-inflated models adopt a similar strategy but allow some zeros to be modeled in a Poisson mixture component. We instead use the truncated Poisson and the unnormalized prior for $H_{i,j}$, $\left(1 - \exp\left(-H_{i,j}\right)\right) \text{Gamma}(\alpha_1, \alpha_1)$, to achieve

a closed form posterior distribution that can be sampled from directly (see Equations (9) and (10) below). The prior distributions for the transformation $H_{i,j}$, given in Equations (6) and (7), are conditionally specified on the observed $X_{i,j}^{[0]}$. This choice is made to produce a posterior distribution for $H_{i,j}$ that exhibits the desired overfitting property (see Section 3.2 for the motivation for this choice).

The posterior transformation is known in closed form (see Supplementary Section Bayesian Hierarchical Model and Transformation Posterior Derivation) and is given by

$$H_{i,j}\big|X_{i,j}^{[0]} = 1, X_{i,j} \sim \text{Gamma}(\alpha_1 + X_{i,j}, \alpha_1 + 1); \quad i = 1, ..., N, \; j = 1, ..., G, \tag{9}$$

$$\frac{\exp(H_{i,j})}{1 + \exp(H_{i,j})}\Bigg|X_{i,j}^{[0]} = 0, X_{i,j} \sim \text{Beta}(\alpha_0, \kappa_0 + 1); \quad i = 1, ..., N, \; j = 1, ..., G, \tag{10}$$

which has mean $\frac{\alpha_1 + X_{i,j}}{\alpha_1 + 1}$ when $X_{i,j}^{[0]} = 1$ and mean $\frac{\alpha_0}{\alpha_0 + \kappa_0 + 1}$ when $X_{i,j}^{[0]} = X_{i,j} = 0$, and hence, overfits the data. A range for parameters α_0, κ_0, and α_1 is considered, and values are chosen to minimize the WAIC (Watanabe, 2010), a generalized version of the common AIC used to choose between models. The ranges of values for α_0 and κ_0 are determined to ensure that the transformed zero-data clusters near to zero, and the range of values for α_1 is determined to ensure that the transformed non-zero data does not shrink too close to zero.

Upon marginalizing $X_{i,j}^{[0]}$ in (4) and (3), we come to the expression

$$X_{i,j}\big|H_{i,j}, X_{i,j}^{(B)} \sim \pi_{i,j}^{*}\text{TruncPoisson}(H_{i,j}) + (1 - \pi_{i,j}^{*})\text{Point Mass}(0), \tag{11}$$

where $\pi_{i,j}^{*} = f\left(X_{i,j}^{[0]} = 1\big|H_{i,j}, X_{i,j}^{(B)}\right)$ (this expression is discussed further in the supplementary material), and the expression in Equation (11) is of the form of a zero-inflated model (Lambert, 1992). We recognize that it is more traditional to replace the truncated Poisson distribution with a Poisson distribution. However, we emphasize that the goal of the zero-inflated model in Equation (11) is different from traditional zero-inflated models.

In particular, our goal is to produce posterior replicates of a *a transformation* for use in a subsequent analysis, and not use posterior summaries $H_{i,j}$ directly. As such, the use of a truncated Poisson distribution is reasonable here, as this leads to (9) and (10), and hence, a noisy transformation $H_{i,j}$ that is close to the original unperturbed data $X_{i,j}$ on some scale.

4.2 Applying ZI-HGT to cell-type deconvolution

Denote the c-th replicate of $H = \{H_{i,j}\}$ from Equations (9) and (10) with $H^{[c]} = \{H^{[c]}_{i,j}\}$ for $c = 1, ..., C$. Then, following the HGT strategy, we assume

$$H^{[c]} = BV' + E, \quad c = 1, ..., C, \tag{12}$$

where $E = \{E_{i,j}\}$ is a set of independent normal random variables with constant variance, V is the cell-type proportions estimate and is restricted to be non-negative, and the columns of V are modeled using a conditional autoregressive (CAR) (Besag, 1974) model. The implementation of this model is given in Algorithm 1.

The ZI-HGT model is defined, via Bradley (2022), by Equations (3-8), (12) and (2). From Bradley (2022), the expression of the BHM requires some careful consideration of normalizing constants as Equations (3-8) condition on $H^{[c]}$, while Equation (12) includes H but not as a condition. The Bayesian hierarchical model for the ZI-HGT + CARD is given in the Supplementary material.

Given that no MCMC is required, the full method with CARD can be run efficiently by generating each of the posterior predictive replicates of the transformed data $H^{[c]}_{i,j}$ then computing the cell-type proportions estimates $V^{[c]} = \hat{V}\left(H^{[c]}\right)$ in parallel. The estimator $\hat{V}$ is the function that minimizes the least squares $E\left[||\hat{V}(H) - S(H)||_2^2 | X\right]$, where S is a generic real-valued estimator and $||\cdot||_2$ is the L_2 norm. The transformed OSCC data replicates H align with the normal assumption in CARD significantly better than the

Algorithm 1: Applying the ZI-HGT to CARD

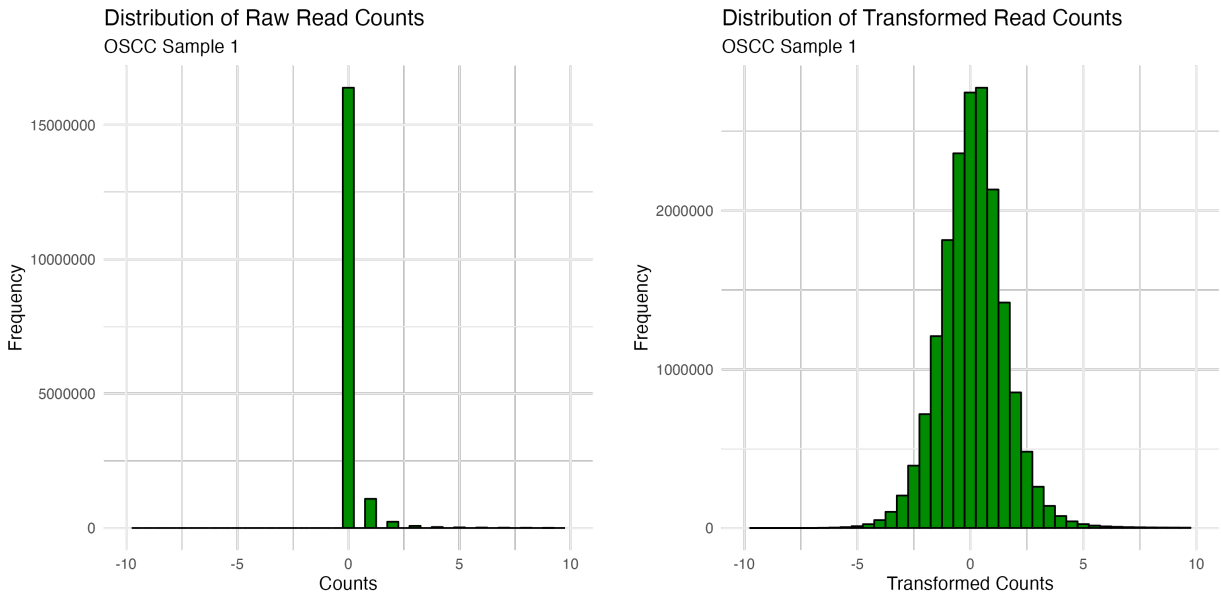
For $c \in \{1, ..., C\}$

(1) Sample posterior predictive transformed data $H_{i,j}^{[c]}$ from $H_{i,j}|X_{i,j}^{[0]}, X_{i,j}$ according to (10) and (9)

(2) Apply CARD to the transformed data: $V^{[c]} = \hat{V}\left(H^{[c]}\right)$

(3) Obtain overall cell-type proportions estimate $\hat{V} = \sum_{c=1}^{C} \hat{V}^{[c]}/C$

original data X (Figure 2). Specifically, H contains no zeros or ties, and the size of the large right tail of the original distribution is significantly reduced. As a result, the new ZI-HGT-generated cell-type proportion estimates are more accurate at no additional computational cost, as seen in the simulation study. We recommend setting $C = 100$, as done in this study. Additionally, uncertainty quantification (UQ) can be easily accomplished by using the individual estimates $\hat{V}^{[c]}$ to construct pointwise credible intervals for $\hat{V}$.



(a) Raw Read Counts

(b) Transformed Read Counts

Figure 2: Distribution of the gene expression read counts for OSCC Sample 1. a) Raw read counts. b) A single posterior replicate of the transformed read counts.

5 Simulations

We designed and conducted a rigorous empirical simulation study based on the OSCC data to evaluate the performance of the proposed approach, the ZI-HGT + CARD. We examined results in multiple settings, including distinct sparsity levels and differing numbers of cell types in the ST data. As a sanity check, we also compared our proposed ZI-HGT + CARD with several existing methods under the identical settings described in CARD (Ma and Zhou, 2022) to ensure comprehensive and unbiased evaluations.

5.1 Simulation Design

The simulation study primarily focuses on OSCC Sample 2, as we believe it to be fairly representative of the majority of the samples, though we ensure that the choice of sample has little impact (Supplementary Figure 2). Sample 2 includes data from $1,749$ locations and $15,624$ transcripts. Following convention, we begin by identifying the tissue layers and corresponding dominant cell types for each of these locations according to pathologist annotations (Arora et al., 2023) (see Supplementary Table 13 for a list of pathologist-annotated layers and corresponding dominant cell types). At this point, it is typical to use the matching scRNA-seq data (Puram et al., 2017) to build the simulated ST data; however, after doing so, we found that the median sparsity level of the simulated ST was approximately 52%, drastically different than the 90% sparsity for sample 2 and 87-91% across all 12 samples. To reconcile the simulated ST data with the real OSCC data, we chose to generate sparser scRNA-seq data with SPARSim (Baruzzo et al., 2020) to subsequently build sparser simulated ST data.

Generating scRNA-seq data: SPARSim employs a two-step Gamma-Multivariate Hypergeometric mixed model for generating simulated scRNA-seq data. In the first step,

the true scRNA-seq data W, an $N \times M$ matrix with rows corresponding to transcripts, or genes, and columns corresponding to individual cells, is simulated from a Gamma distribution. Consider a specific cell type ct, and let there be M_{ct} such cells in the dataset. Then, let $W_{i,j}$ be a random variable representing the expression level of gene i in cell j ($i = 1, ..., N;\ j = 1, ..., M_{ct}$). For any gene i,

$$W_{i,j} \sim Gamma\left(\text{shape} = \frac{1}{\Phi_{i,ct}},\ \text{scale} = Z_{i,ct} \cdot \Phi_{i,ct}\right), \tag{13}$$

where $Z_{i,ct}$ is the "average" intensity, or expression level of gene i in cell type ct, and $\Phi_{i,ct}$ describes the biological variability in the expression level of gene i in cell type ct. This procedure iterates through all genes i and then all cell types ct. Then, in the second step, the observed scRNA-seq data Y, also an $N \times M$ matrix with rows corresponding to transcripts and columns corresponding to individual cells, is simulated from a Multivariate Hypergeometric distribution. This step attempts to model the technical variability of observed scRNA-seq data, as no data capture process can perfectly record the true scRNA-seq data. For any cell j,

$$Y_j \sim \text{Multivariate Hypergeometric}(n = L_j, m = W_j), \tag{14}$$

where L_j is the library size, and W_j is the true scRNA-seq data to sample from. Note that the library size is the total number of gene expression reads included in the "observed" data Y_j for cell j.

The three SPARSim hyperparameters, $Z_{i,ct}$, $\Phi_{i,ct}$, and L_j, are originally estimated from the real scRNA-seq data (Puram et al., 2017). The intensity $Z_{i,ct}$ is the average value for the scran-normalized (Lun et al., 2016) expression read counts for gene i and cell type ct. The biological variability parameter $\Phi_{i,ct}$ is estimated using the "*estimateGLMTagwiseDisp*" function in the *edgeR* package (Robinson et al., 2010). The library size L_j is computed by

summing the raw counts for each cell. To alter the sparsity of the simulated scRNA-seq data, and therefore the sparsity of the downstream simulated ST data, we estimated the three hyperparameters with the real scRNA-seq data (Puram et al., 2017) and then re-scaled the biological variability $\Phi_{i,ct}$ and library size L_j. For each set of scaled hyperparameters, we generated 10 simulated scRNA-seq datasets.

Generating ST data: We construct the simulated ST data following the same scaffolding as CARD (Ma and Zhou, 2022), with a crucial modification: our simulations are based on both the real OSCC ST data (Arora et al., 2023) and the simulated scRNA-seq data. This modification was specifically designed to ensure closer alignment with the characteristics of the real-world datasets, such as the sparsity. We divided the OSCC samples into layers according to pathologist annotations and assigned each layer a dominant cell type: endothelial cells for veins and arteries, intermediate fibroblasts for glandular stroma, ecm-myCAFs for lymphocyte-negative stroma, cytotoxic CD8+ T cells for lymphocyte-positive stroma, myofibroblasts for muscles, macrophages for non-cancerous mucosa, and cancer cells for squamous cell carcinoma. The simulations proceeded as follows: for a desired OSCC sample (Sample 2 was used for most analyses), a simulated scRNA-seq dataset was selected, enabling control over the sparsity of the simulated ST data. The simulated scRNA-seq data was divided into two parts: one for constructing the simulated ST data and the other for use in the cell-type deconvolution process. For each tissue layer in the sample, the number of colocalizing cell types with the dominant cell type was sampled from a discrete uniform distribution with a minimum of 1 and a maximum of 13. Next, the cell type proportions for each location within that tissue layer were determined by simulating from a Dirichlet distribution with concentration set to one. The largest proportion was assigned to the dominant cell type, and the proportions of other colocalizing cell types

were randomly assigned. Ten cells were allocated to each location, with the counts for each cell type being ten times the sampled cell type proportions, rounded to the nearest whole number. Cells were sampled without replacement from the first split of the simulated scRNA-seq data, and gene expression levels were aggregated across the ten cells to determine the gene expression at each location. Finally, recall that ten simulated scRNA-seq datasets were generated for each set of SPARSim (Baruzzo et al., 2020) hyperparameters. From each of these datasets, ten sets of ST data were simulated, resulting in a total of 100 sets of simulated ST data for each set of hyperparameters.

Simulation settings and evaluation criteria: In our simulation study, we explored various settings, encompassing different sparsity levels in the simulated spatial transcriptomics (ST) data (Figure 3), simulations based on diverse OSCC samples (Supplementary Figure 2), and varying numbers of cell types in the ST data (Supplementary Figure 1). Additionally, we compared the ZI-HGT + CARD to other cell-type deconvolution methods, including SPOTlight (Elosua-Bayes et al., 2021), SpatialDecon (Danaher et al., 2022), STdeconvolve (Miller et al., 2022), as well as CARD with a simple deterministic transformation and the HGT + CARD (without considering zero-inflation). That said, our primary focus is on comparing the ZI-HGT + CARD to CARD alone, as CARD is the current state-of-the-art method and performed best among all methods we evaluated. We primarily focused on OSCC Sample 2 because it is representative of the majority of the samples. Sparsity levels in the simulated ST data were controlled by multiplying the SPARSim hyperparameters by specific factors. Specifically, we generated the three SPARSim hyperparameters using the real scRNA-seq data (Puram et al., 2017), and through iterative adjustments, we found that multiplying the library size L_j by 0.05 and the biological variability $\Phi_{i,ct}$ by 50 yielded the most realistic data in terms of sparsity and structure.

We refer to these factors as the "library factor" and "Φ factor", respectively. Across 100 simulated ST datasets using these settings, sparsity levels ranged between 86% and 92%, closely aligning with the observed sparsity levels in the real OSCC data (Supplementary Table 1).

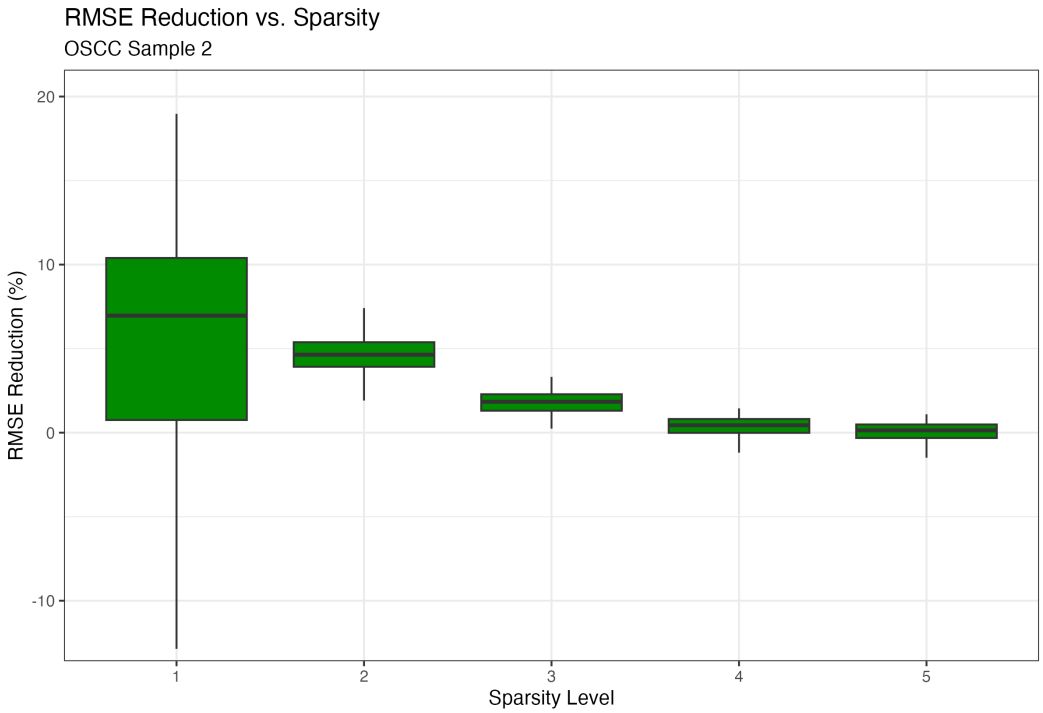
In the simulations, we set C, the number of ZI-HGT iterations, to 100. We let $\alpha_0 \in \{1.73, 2, 2.33\}$ and $\alpha_1 \in \{0.1, 0.3, 0.5\}$, then chose a single pair of hyperparameters by minimizing the WAIC (Watanabe, 2010). To assess accuracy, we measured the root mean square error (RMSE) for the estimated cell-type proportions, given by

$$\text{RMSE} = \sqrt{\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(V_{i,k} - \hat{V}_{i,k}\right)^2}, \tag{15}$$

where N is the number of spatial locations and K is the number of cell types. We then compared the ZI-HGT + CARD to the original CARD method by determining the percent RMSE improvement over CARD, calculated by $100\% \times \left(1 - \text{RMSE}_{\text{ZI-HGT+CARD}}/\text{RMSE}_{\text{CARD}}\right)$.

5.2 Simulation Results

We first considered the impact of the sparsity of the ST data on the performance of the ZI-HGT + CARD, depicted in Figure 3. Higher sparsity in the simulated ST data resulted in stronger performance of the ZI-HGT + CARD compared to CARD alone. At sparsity level 1 (library factor 0.05 and Φ factor 50), for which the median sparsity was 89.8% (i.e., the realistic simulated ST data), the ZI-HGT + CARD improved the RMSE by a median of 6.6% beyond CARD. At sparsity level 2, for which the median sparsity was 77.9%, the ZI-HGT + CARD improved the RMSE by a median 4.6%. This trend was expected because higher sparsity, or zero-inflation, leads to ST data that is less normally distributed, thus deviating further from CARD's assumptions. The ZI-HGT addresses zero-inflation; therefore, its impact increases as the ST data becomes "less normal."



Sparsity Level	Φ Factor	Library Factor	Minimum	Median	Maximum
1	50	0.05	86.8%	89.8%	91.0%
2	10	0.1	73.2%	77.9%	80.1%
3	4	0.2	58.4%	63.7%	67.5%
4	4	0.4	48.8%	54.4%	58.3%
5	4	0.6	43.9%	49.7%	54.0%

Figure 3: **RMSE reduction vs. sparsity level of simulated ST data.** RMSE reduction for the ZI-HGT + CARD across 100 simulated ST datasets and various sparsity levels. The sparsity levels in the boxplot have corresponding SPARSim hyperparameter factors and minimum, median, and maximum sparsity levels for all simulated ST datasets displayed in the table.

Next, we investigated the influence of the number of distinct cell types in the dataset on performance (Supplementary Figure 1). While the median RMSE reduction appears consistent across varying numbers of cell types, a noticeable trend emerges, showing a

reduction in variance as the number of cell types decreases. This outcome aligns with expectations because a smaller number of cell types leads to more compact and simplified simulated ST data.

To ensure the robustness of our findings, particularly in the context of our focus on OSCC Sample 2, we examined the potential impact of choosing a specific sample for simulating ST data (Supplementary Figure 2). We generated simulated ST data from SPARSim-generated simulated scRNA-seq data, adopting "realistic" settings for the library factor (0.05) and Φ factor (50). Applying our methods to these simulations, we observed reasonably consistent median RMSE reduction percentages, with a median of 6.6%. This reinforces our confidence in focusing on OSCC Sample 2 for the simulations, supporting it as a fair and reasonable choice.

Additionally, we compared the selection of hyperparameters for the ZI-HGT using the WAIC vs. the oracle hyperparameters, which were chosen to minimize the RMSE in each simulated dataset (Supplementary Figure 2). Fortunately, minimizing the WAIC consistently led to reasonable hyperparameter choices. The difference in the median RMSE reduction between the WAIC-chosen hyperparameter model and the oracle model is consistently small, nearly always within a 1% RMSE reduction.

Finally, we compared the ZI-HGT + CARD to the use of a simple deterministic transformation of the ST data before running CARD, an HGT + CARD (i.e., the ZI-HGT without addressing the zero-inflation), and other common methods for cell-type deconvolution: SPOTlight (Elosua-Bayes et al., 2021), SpatialDecon (Danaher et al., 2022), and STdeconvolve (Miller et al., 2022) (Supplementary Figure 3). We chose a $log\,(1 + \epsilon + X_{i,j})$ transformation for the deterministic transformation to mimic the ZI-HGT; the log mitigates the right-skewness of the ST data, and the ϵ addition addresses the zero-inflation.

This deterministic transformation improves the RMSE by 2% beyond CARD alone, though it is unable to approach the approximate 7% RMSE decrease of the ZI-HGT + CARD. This result is expected; a deterministic transformation is not appropriate for the OSCC ST data because it cannot address the abundance of ties in the dataset, nor does it allow for uncertainty quantification. Adding a non-zero-inflated hierarchical generalized transformation model to CARD improves the RMSE by approximately 1% beyond CARD alone, emphasizing the need to address zero-inflation. Each of the other common cell-type deconvolution methods performs notably worse than CARD alone; the median RMSE of the best performing method of the three (SPOTlight) was approximately 15% worse than CARD.

To demonstrate the viability of our method in an unbiased manner, we also applied the ZI-HGT + CARD to CARD's exact simulations without any modifications. These simulations were constructed very similarly to ours, though they are based on two freely available mouse olfactory bulb (MOB) datasets, a scRNA-seq dataset (Zeisel et al., 2018) and an ST dataset (Ståhl et al., 2016). We note that the simulated ST data is much less sparse than the real ST datasets; the median sparsity level was 68.6% in these simulated datasets and 88.5% in the real OSCC ST datasets. The scRNA-seq reference comprises gene expression read counts for 20,515 cells across over six cell types: neurons, astrocytes, oligodendrocytes, vascular cells, immune cells, and ependymal cells. The scRNA-seq data is split into two pieces: one to construct the simulated ST data and the other to conduct cell-type deconvolution. The ST dataset contains 260 locations and three anatomic regions: the granule cell layer, the mitral cell layer, and the nerve layer. Each anatomic region was assigned a dominant cell type: oligodendrocytes for the granule cell layer, neurons for the mitral cell layer, and astrocytes for the nerve cell layer (Svensson et al., 2018). For additional robustness, Ma and Zhou (2022) considered simulations with added noise. Let p_n

be the percentage of the 260 spatial locations that are "noisy". Data at noisy locations are generated identically to non-noisy locations, with the exception that all cell-type proportions are randomly assigned, so the dominant cell type no longer automatically receives the largest sampled proportion of cells. Noisiness percentages $p_n \in \{0\%, 20\%, 40\%, 60\%\}$ are considered.

Five distinct simulation scenarios are considered: 1) The scRNA-seq data is applied as the reference basis matrix for cell-type deconvolution, exactly as usual. 2) The scRNA-seq data is applied as the reference basis matrix for cell-type deconvolution, but one cell-type is removed from the reference. 3) The scRNA-seq data is applied as the reference basis matrix for cell-type deconvolution, but one cell-type (blood cells) is added to the reference. 4) The scRNA-seq data is applied as the reference basis matrix for cell-type deconvolution, but two cell-types are merged in the reference. 5) A mismatched scRNA-seq reference dataset from a different platform, drop-seq, is used. Beyond CARD's simulation construction, we also consider the impact of altering the sequencing depth of the simulated spatial transcriptomic data. The sequencing depth of a platform is defined as the total number of read counts across the entire tissue. As comparators for applying the ZI-HGT to CARD, we also apply CARD to simulated datasets with total sequencing depth -10% and $+10\%$ as compared to the original simulated datasets. We achieve increased and decreased sequencing depth by decreasing and increasing the number of cells per location in the simulation to nine and eleven, respectively.

In Figure 4, we display the results of applying the ZI-HGT + CARD across 100 simulated ST datasets for all five CARD simulation scenarios with $p_n = 0\%$ (see Supplementary Figures 4 - 6 for $p_n \in \{20\%, 40\%, 60\%\}$). While applying the ZI-HGT + CARD makes less of an impact here than in the more realistic OSCC simulations (only a 2.1% median

RMSE reduction for scenario 1), this outcome is to be expected. The median sparsity of the simulated ST data for scenario 1 ($p_n = 0\%$) for the CARD simulations is 68.6%, which is significantly lower than is realistic for many ST datasets, including the OSCC data. As shown in Figure 3, the ZI-HGT is more impactful in sparser ST datasets because the zero-inflation is more drastic. In all, the ZI-HGT + CARD is more accurate than CARD alone for almost every noisiness percentage p_n and scenario. An exception is observed in scenario 5, where a mismatched scRNA-seq reference was used. As the ZI-HGT transforms the data by adding noise, the noise from the mismatched scRNA-seq reference is greatly amplified, leading to slightly diminished accuracy. Nonetheless, scenario 5 represents a relatively uncommon and suboptimal approach in practice, as using a mismatched scRNA-seq reference dataset is generally avoided when possible. In situations where the only option is a mismatched reference, the sole use of CARD is advised. It is crucial for users to exercise caution when interpreting and applying cell deconvolution results under these conditions due to the heightened risk of inaccuracies. As a sanity check, our analysis confirms that enhancing sequencing depth directly contributes to improved deconvolution accuracy. Comparatively, our novel method achieves a marginal reduction in accuracy enhancement compared to a 10% increase in sequencing depth under the CARD simulation settings.

6 Results

We now return to the OSCC data introduced in the Data Section (Arora et al., 2023) and report results from applying the ZI-HGT + CARD. Across the 12 analyzed samples, we set C, the number of ZI-HGT iterations, to 100. We used the previously discussed HNSCC scRNA-seq data from Puram et al. (2017) (GSE103322) as the deconvolution reference. Additionally, we applied CARD alone to serve as a comparison to the ZI-HGT + CARD.

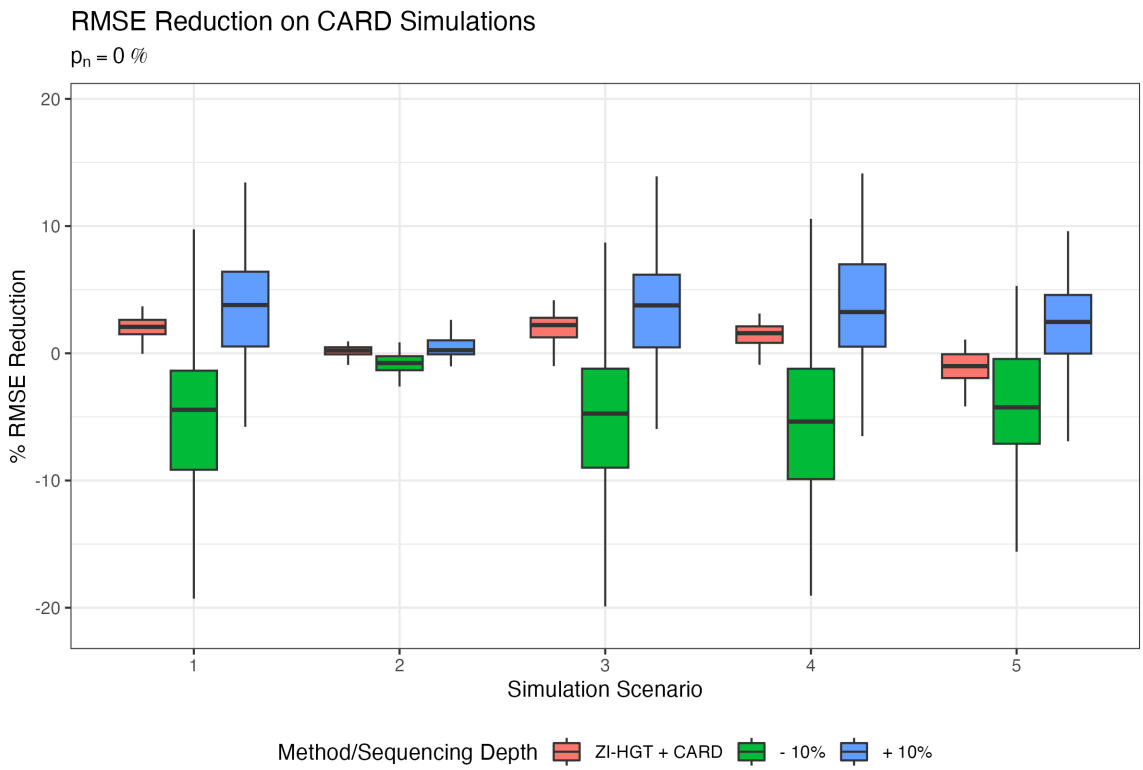


Figure 4: **RMSE reduction on CARD simulations.** RMSE reduction for the ZI-HGT + CARD across 100 simulated MOB ST datasets and various simulation scenarios following the CARD simulation setup with $p_n = 0\%$. The -10% and +10% sequencing depths refer to applying CARD to simulated ST data with accordingly decreased or increased sequencing depth.

For the Results Section, we focus on OSCC Sample 1, though equivalent plots exist for the remaining 11 samples in the supplementary material(Supplementary Figures 7-41).

First, we recall that OSCC Sample 1 consists of spatially resolved gene expression read counts measured for 15,844 transcripts across 1,131 locations. We analyze the estimated proportions of each of the 14 cell types at each location in Figure 5 (ZI-HGT + CARD) and Supplementary Figure 7 (CARD). We display both the estimates of cell-type proportions and binary plots, with coloring applied only when the estimated cell-type proportion exceeds the threshold of 0.05. We selected this threshold as we expect approximately 10

cells per location for 10x Visium ST data. Therefore, if the cell-type proportion exceeds 0.05, it is reasonable to assume the presence of at least one cell of that type at the location.

Importantly, it is evident that the ZI-HGT + CARD demonstrates a marked improvement in identifying both cancer-associated fibroblasts (ecm-myCAFs and detox-iCAFs) and normal fibroblasts (myofibroblasts and intermediate fibroblasts) across the tumor microenvironment (TME) compared to CARD alone. Normal fibroblasts are known to exert suppressive functions against tumor growth, although they have the potential to transform into cancer-associated fibroblasts during tumorigenesis (Alkasalias et al., 2018). Cancer-associated fibroblasts play a pivotal role within the TME, supporting both immunosuppression and tumor promotion (Arina et al., 2016). Both forms of fibroblasts are expected to be distributed throughout the TME, though cancer-associated fibroblasts naturally colocalize more closely with tumor cells (Maia et al., 2021)—a pattern that is distinctly illustrated through the application of ZI-HGT + CARD (Figure 5a) but is less evident with CARD alone (Supplementary Figure 7a). Cancer-associated fibroblasts have been a primary target for treatment in recent years (Maia et al., 2021); therefore, better determining their locations in the TME could influence the development of cancer treatments and advance precision medicine.

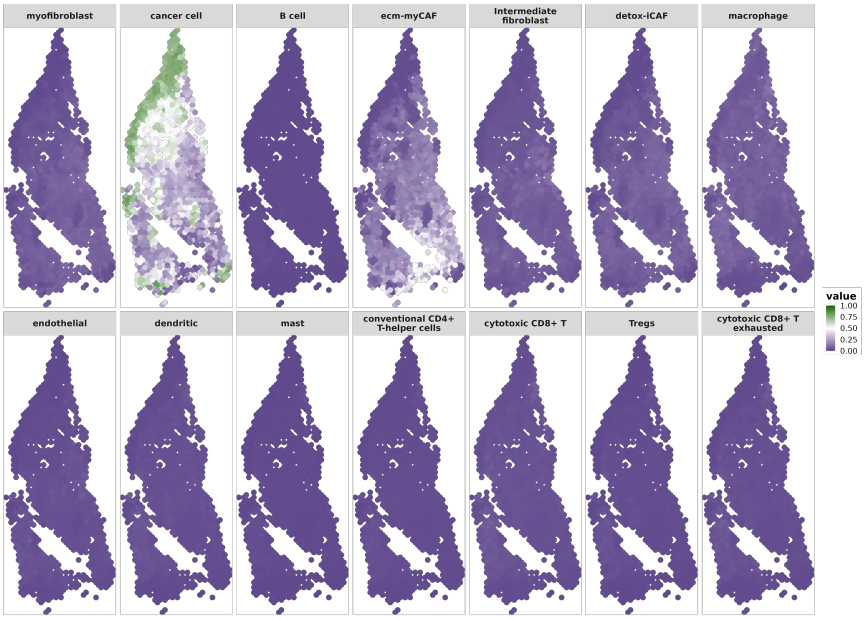
Assessing the precision of deconvoluted cell-type proportions in real datasets is challenging; however, a common metric evaluates Pearson's correlation between the average estimated cell-type proportions across the tissue and the proportions of the cell types in the corresponding scRNA-seq data (Ma and Zhou, 2022; Li et al., 2023). For OSCC Sample 1, the correlation coefficient was 0.93 for the ZI-HGT + CARD and 0.85 for CARD alone. Across all 12 samples, the ZI-HGT + CARD-estimated cell-type proportions show stronger correlation with the scRNA-seq proportions for 10 samples, with notable differ-

ences observed in the more complex samples, namely Samples 1, 3, 4, and 5. Additionally, 42.1% of the cells in the scRNA-seq dataset are cancer cells (Puram et al., 2017). However, the average cancer proportion across all locations and samples for CARD alone is 90.0%, which is a significant overestimation. The ZI-HGT + CARD also overestimates the overall proportion of cancer cells, but at a notably better 79.5%. Evidently, using the ZI-HGT helps reduce the substantial overestimation of the dominant cell type by CARD.

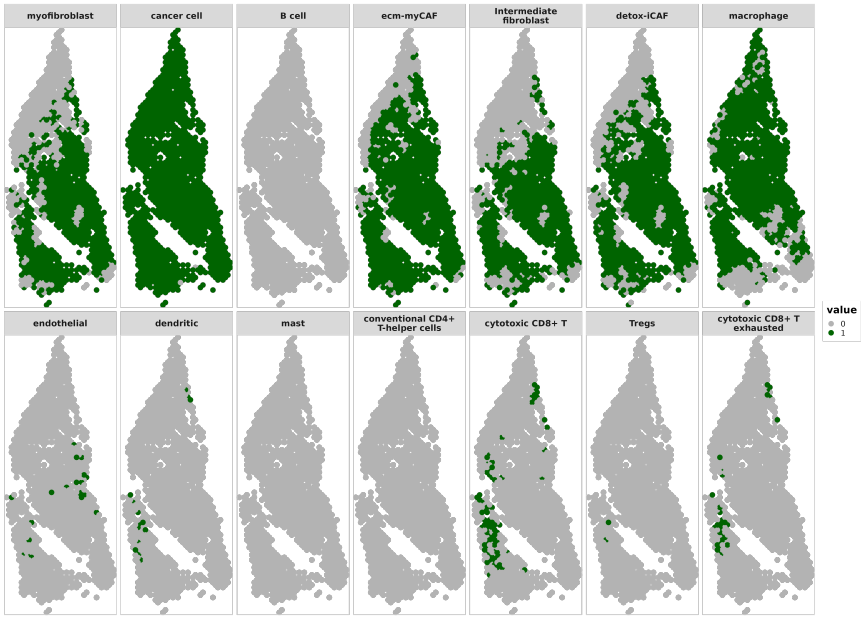
Beyond the capabilities of CARD alone, the ZI-HGT naturally quantifies uncertainty in estimates of the cell-type proportions by constructing pointwise Bayesian credible intervals (Supplementary Figure 8). There are $C = 100$ estimates of the cell-type proportions from the ZI-HGT + CARD. For each cell type and each location, we determine the lower and upper bounds of the interval by identifying the corresponding estimate of the cell-type proportions. For example, consider cancer cells at location L. We will refer to the proportion of cancer cells at L as $V_{\mathrm{L,cancer}}$. To construct a 95% pointwise credible interval, order the posterior predictions of the proportions of cancer cells at location L, $V^{[h]}_{\mathrm{L,cancer}}$, from smallest to largest. The lower bound of the interval will be the 2.5th percentile of the proportions, and the upper bound the 97.5th percentile. This is a very exciting direction to explore with the ZI-HGT, and we plan to do so in greater depth in future work.

7 Conclusions

By integrating hierarchical generalized transformation models (Bradley et al., 2020; Bradley, 2022; Bradley et al., 2023) with Bayesian methods for zero-inflation (Greene, 1994) and cell-type deconvolution with CARD (Ma and Zhou, 2022), we address CARD's assumptions (i.e., that the ST data are normally distributed) and account for uncertainty in the analysis of oral squamous cell carcinoma (OSCC) spatial transcriptomic data (Arora et al., 2023).



(a) Cell-Type Proportions Estimates



(b) Cell-Type Proportions > 0.05

Figure 5: OSCC Sample 1 ZI-HGT + CARD Cell-Type Proportions Estimates. a) Plots of cell-type proportions for all 14 cell types. b) Binary plots of cell-type proportions > 0.05 for all 14 cell types.

Importantly, the developed techniques are generally applicable. Our ZI-HGT could serve as an auxiliary technique for future cell-type deconvolution methods that share assumptions with CARD.

The integration of the ZI-HGT and CARD yielded compelling insights into the spatial distribution of cell types within the OSCC tumor microenvironment (TME). Notably, the ZI-HGT + CARD exhibited better accuracy than CARD alone, enabling the identification of the diverse fibroblast population, which is composed of critical players in the TME with implications for tumor growth and immunosuppression, within OSCC Sample 1. A notable strength of the ZI-HGT + CARD lies in its natural quantification of uncertainty. We constructed pointwise Bayesian credible intervals without making unrealistic assumptions regarding the distribution of the data.

Numerous methodological research directions remain open. New developments in spatial transcriptomic data capture techniques emerge annually, and adaptations of the ZI-HGT + CARD could be applied to 3-dimensional ST data or ST data with single-cell resolution. Approaches relying on cell-type proportion estimates, such as the cell-type-specific differential gene expression method C-SIDE (Cable et al., 2022), could be integrated into a deep model structure incorporating ZI-HGT and CARD. Furthermore, techniques beyond cell-type deconvolution may benefit from the ZI-HGT, such as the super-resolution gene expression prediction model iSTAR (Zhang et al., 2024). These promising ideas require substantial further development and are reserved for future research. Lastly, from a biology perspective, further implications may be drawn from the spatial patterns of the estimated cell-type proportions. Collaboration and consultation with disease-area experts are essential to fully realize the potential of these findings.

We conclude by discussing potential limitations of the ZI-HGT + CARD. Primarily,

the integration, or relative lack thereof, of the ZI-HGT and CARD may be considered a weakness. CARD itself could potentially be altered to better align with the ZI-HGT and the data. CARD assumes a conditional autoregressive (CAR) spatial structure in the data, which may be too simple. The CAR model with nearest neighbor adjacencies is not suitable for all forms of spatial data, such as non-stationary data. A more flexible spatial basis function expansion approach may be a better choice in this scenario, though further exploration of the spatial structure of the ST data would be necessary to answer such questions. However, the independence of the ZI-HGT and CARD can also be considered a strength. Alternative methods, such as the aforementioned iSTAR (Zhang et al., 2024), could replace CARD in the ZI-HGT + CARD model structure to better accomplish tasks beyond cell-type deconvolution. We posit that the ZI-HGT lays a groundwork for future method development, enabling researchers to design models that safely circumvent the zero-inflation and ties intrinsic to ST data and simply assume a normal distribution, providing a versatile foundation for novel approaches.

SUPPLEMENTARY MATERIAL

Section S.1 discusses the preparation of the scRNA-seq data and the different cell types considered in the deconvolution process. Section S.2 provides the Bayesian hierarchical model (BHM) for the ZI-HGT + CARD and derives the transformation posterior. Section S.3 demonstrates that the ZI-HGT can be written in the form of a zero-inflated Poisson. Section S.4 provides a method of constructing inverse transformations for the ZI-HGT. Section S.5 presents additional tables showing the sparsity for each OSCC sample and the distribution of the counts in each sample individually. Section S.6 gives additional results from both real and simulated studies.

References

Alkasalias, T., Moyano-Galceran, L., Arsenian-Henriksson, M., and Lehti, K. (2018). Fibroblasts in the tumor microenvironment: Shield or spear? *International Journal of Molecular Sciences*, 19:1532.

Arina, A., Idel, C., Hyjek, E. M., Alegre, M. L., Wang, Y., Bindokas, V. P., Weichselbaum, R. R., and Schreiber, H. (2016). Tumor-associated fibroblasts predominantly come from local and not circulating precursors. *Proceedings of the National Academy of Sciences of the United States of America*, 113:7551–7556.

Arora, R., Cao, C., Kumar, M., Sinha, S., Chanda, A., McNeil, R., Samuel, D., Arora, R. K., Matthews, T. W., Chandarana, S., Hart, R., Dort, J. C., Biernaskie, J., Neri, P., Hyrcza, M. D., and Bose, P. (2023). Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nature Communications*, 14:1–19.

Baruzzo, G., Patuzzi, I., and Camillo, B. D. (2020). Sparsim single cell: a count data simulator for scrna-seq data. *Bioinformatics*, 36:1468–1475.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems on jstor. *Journal of the Royal Statistical Society Series B*, pages 192–236.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B*, 26:211–252.

Bradley, J. (2022). Joint bayesian analysis of multiple response-types using the hierarchical generalized transformation model. *Bayesian Analysis*, 17(1):127–164.

Bradley, J., Wikle, C. K., and Holan, S. H. (2020). Hierarchical models for spatial data with errors that are correlated with the latent process. *Statistica Sinica*, 30(1):81–109.

Bradley, J., Zhou, S., and Liu, X. (2023). Deep hierarchical generalized transformation models for spatio-temporal data with discrepancy errors. *Spatial Statistics*.

Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., and Irizarry, R. A. (2022). Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4):517–526.

Danaher, P., Kim, Y., Nelson, B., Griswold, M., Yang, Z., Piazza, E., and Beechem, J. M. (2022). Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. *Nature Communications*, 13:1–13.

Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., and Heyn, H. (2021). Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49(9):e50.

Greene, J. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models. *NYU Working Paper No. EC-94-10*.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14.

Li, H., Zhou, J., Li, Z., Chen, S., Liao, X., Zhang, B., Zhang, R., Wang, Y., Sun, S., and Gao, X. (2023). A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nature Communications*, 14:1–10.

Li, Q., Zhang, M., Xie, Y., and Xiao, G. (2021). Bayesian modeling of spatial molecular profiling data via gaussian process. *Bioinformatics*, 37:4129–4136.

Liao, J., Lu, X., Shao, X., Zhu, L., and Fan, X. (2021). Uncovering an organ's molecular architecture at single-cell resolution by spatially resolved transcriptomics. *Trends in biotechnology*, 39(1):43–58.

Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome Biology*, 17:1–14.

Ma, Y. and Zhou, X. (2022). Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*, 40(9):1349–1359.

Maia, A., Wiemann, S., Reshkin, S. J., and Cardone, R. A. (2021). Cancer-associated fibroblasts: Implications for cancer therapy. *Cancers*, 13:3526.

Markopoulos, A. K. (2012). Current aspects on oral squamous cell carcinoma. *The open dentistry journal*, 6:126–130.

Miller, B. F., Huang, F., Atta, L., Sahoo, A., and Fan, J. (2022). Reference-free cell-type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nature Communications*, 13:1–13.

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., Hajdu, C. H., Simeone, D. M., and Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, 38:333–342.

Moses, L. and Pachter, L. (2022). Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546.

Musgrove, D., Young, D. S., Hughes, J., and Eberly, L. E. (2019). *A Sparse Areal Mixed Model for Multivariate Outcomes, with an Application to Zero-Inflated Census Data*, pages 51–74. Springer International Publishing, Cham.

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C. L., Mroz, E. A., Emerick, K. S., Deschler, D. G., Varvares, M. A., Mylvaganam, R., Rozenblatt-Rosen, O., Rocco, J. W., Faquin, W. C., Lin, D. T., Regev, A., and Bernstein, B. E. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171:1611–1624.

Rao, A., Barkley, D., França, G. S., and Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. *Nature*, 596:211–220.

Robinson, M., McCarthy, D., and Smyth., G. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467.

Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Di Bella, D. J., Arlotta, P., Macosko, E. Z., and Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., ke Borg, Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O.,

Lundeberg, J., and Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (New York, N.Y.)*, 353:78–82.

Svensson, V., Teichmann, S. A., and Stegle, O. (2018). Spatialde: identification of spatially variable genes. *Nature methods*, 15(5):343–346.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

Yeo, I. N. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87.

Zeisel, A., Hochgerner, H., Lannerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Haering, M., Braun, E., Borm, L. E., Manno, G. L., Codeluppi, S., Furlan, A., Lee, K., Skene, N., Harris, K. D., Hjerling-Leffler, J., Arenas, E., Ernfors, P., Marklund, U., and Linnarsson, S. (2018). Molecular architecture of the mouse nervous system. *Cell*, 174:999–1014.

Zhang, D., Schroeder, A., Yan, H., Yang, H., Hu, J., Lee, M. Y. Y., Cho, K. S., Susztak, K., Xu, G. X., Feldman, M. D., Lee, E. B., Furth, E. E., Wang, L., and Li, M. (2024). Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*.

Zhao, P., Zhu, J., Ma, Y., and Zhou, X. (2022). Modeling zero inflation is not necessary for spatial transcriptomics. *Genome Biology*, 23:1–19.