# CITEgeist: Cellular Indexing of Transcriptomes and Epitopes for Guided Exploration of Intrinsic Spatial Trends

**Alexander Chih-Chieh Chang[1,2,⊕], Brent T. Schlegel[1,2,⊕], Neil Carleton[1,2], Priscilla F. McAuliffe[1,3], Steffi Oesterreich[1,4], Russell Schwartz[5,6], and Adrian V. Lee[1,4,7,*]**

[1] **Women's Cancer Research Center, UPMC Hillman Cancer Center, Magee-Womens Research Institute, Pittsburgh PA, USA**
[2] **University of Pittsburgh School of Medicine, Pittsburgh, PA, USA**
[3] **Department of Surgery, Division of Breast Surgical Oncology, University of Pittsburgh School of Medicine, Pittsburgh PA, USA**
[4] **Department of Pharmacology and Chemical Biology, University of Pittsburgh, Pittsburgh PA, USA**
[5] **Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA**
[6] **Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA**
[7] **Institute of Precision Medicine, Pittsburgh PA, USA**
[⊕] **This author contributed equally to this manuscript**
[*] **Corresponding Author**

## ABSTRACT

Spatial transcriptomics provides insights into tissue architecture by linking gene expression with spatial localization. Current deconvolution methods rely heavily on single-cell RNA sequencing (scRNA-seq) references, which are costly and often unavailable, mainly if the tissue under evaluation is limited, such as in a core biopsy specimen. We present a novel tool, CITEgeist, that deconvolutes spatial transcriptomics data using antibody capture from the same slide as the reference, directly leveraging cell surface protein measurements from the same tissue section. This approach circumvents the limitations of scRNA-seq as a reference, offering a cost-effective and biologically grounded alternative. Our method employs mathematical optimization to estimate cell type proportions and gene expression profiles, incorporating sparsity constraints for robustness and interpretability. Benchmarks against state-of-the-art deconvolution methods show improved accuracy in cell type resolution, particularly in dense tumor microenvironments, while maintaining computational efficiency. This antibody-based tool advances spatial transcriptomics by providing a scalable, accurate, and reference-independent solution for deconvolution in complex tissues. We validate this tool by using a combined approach of simulated data and clinical samples by applying CITEgeist to translational pre-treatment and post-treatment ER+ breast tumors from an ongoing clinical trial, emphasizing the applicability and robustness of CITEgeist.

Keywords: Spatial transcriptomics, Cancer, multi-omics, CITE, 10x Genomics, antibody capture, integer linear programming (ILP), expectation maximization (EM), breast cancer

## INTRODUCTION

In recent years, spatial transcriptomics has revolutionized our understanding of tissue architecture by providing spatially resolved gene expression data [1]. A standard method for analyzing Visium spatial sequencing data involves deconvolution using a single cell RNA sequencing (scRNA-seq) reference, which serves as a guide to estimate the proportions of cell types within each spatial spot [2, 3]. However, this approach has several inherent limitations. First, generating a single-cell reference requires additional, costly sequencing, doubling the experimental costs [4]. Furthermore, comprehensive single-cell atlases are unavailable for many cancer phenotypes and other specialized tissues, limiting the generalizability of this approach, such as post-treatment cancer samples [5].

A significant issue with the reliance on scRNA-seq references is that it often leads to tools optimized on standardized datasets, such as the mouse hippocampus demo data widely provided by 10X Genomics [6]. For example, all current tools were previously validated/published on this same mouse brain dataset [2, 7, 8, 9]. While these tools excel in resolving large and well-characterized tissue architectures, they do not reflect the complexity and heterogeneity of tissues such as cancer, where multiple cell types can be tightly packed within the same spatial region. Thus, the ability of these methods to handle intricate, dysplastic, mixed-cell environments, such as those found in tumors, is significantly limited. Additionally, there is a risk of artificially imposing single-cell data onto spatial maps, which can obscure unique biological signals in the tissue of interest.

A more biologically relevant and cost-effective alternative involves leveraging antibody capture data, which can be collected concurrently with RNA data on the same Visium slide. Unlike RNA data, which suffers from sparsity, antibody capture data offers direct measurements of cell surface proteins, providing a robust means to infer cell identity [10]. Despite the potential advantages of antibody capture, existing tools have not fully capitalized on its potential for spatial transcriptomic deconvolution.

We present a novel tool designed to deconvolute spatial transcriptomic data using antibody capture as the primary reference. This approach eliminates the need for costly scRNA-seq references while ensuring the estimated cell-type proportions are based on direct biological measurements from the same tissue section. By integrating antibody capture data into the deconvolution process, our tool offers a more accurate, cost-effective, and biologically meaningful method for spatial transcriptomic analysis. Furthermore, it was developed in response to a need identified in a translational trial with heterogeneous samples, which added to the robustness of the methods developed.

## MATERIALS AND METHODS

### Single-Cell RNA-seq Reference

To establish a baseline for gene expression profiles —– both for deconvolution tasks and the simulation of spatially resolved transcriptomic datasets —– we utilized a publicly available atlas of single-cell transcriptomic data from human breast cancers [5], explicitly focusing on estrogen receptor-positive (ER+) tumors. This dataset served as the reference for deconvolution methods like Seurat and cell2location, as well as the foundation for our simulations, enabling us to generate *in-silico* spatial datasets with a quantifiable ground truth of cell type proportions and gene expression levels.

We selected 12,000 cells from ER+ tumor samples across 11 ER+/HER2- patients in the Wu et al. dataset [5], maintaining cell type proportions through random downsampling from the primary atlas according to the annotations of 9 distinct cell types. Due to the computational constraints of the scCube autoencoder framework [11], this downsampled atlas served as our primary reference for spatial transcriptomic simulations.

For reference-based deconvolution methods, we used two references. We first retained the complete set of ER+ tumor cells from the 11 patients from the Wu et al. atlas. This ensured that the reference expression profiles for cell types were as comprehensive as possible, avoiding potential biases introduced by downsampling. We then re-ran the analysis with a reference downsampled from 30,000 to 8,000 cells to reflect both cancer heterogeneity and what using a single sample single cell reference would look like. By using both approaches, we ensured that the benchmarking fairly evaluated the deconvolution performance without artificial constraints on the reference profiles, while also testing what real-world use case would look like on our simulated tests.

### Spatial Transcriptomic Simulation

We utilized the scCube (v2.0.0) framework [11] to simulate spatial transcriptomic data from the downsampled ER+ scRNA-seq atlas. A Variational Autoencoder (VAE) with a hidden layer size 128, a learning rate of 0.0001, and 10,000 epochs generated single-cell gene expression profiles for nine major cell types. The model was trained on the input single-cell object, using the X attribute for expression and the 'celltype' key for cell types. Cells were generated based on the proportions of the annotated cell types, with a batch size of 512, executed on a CUDA-enabled A100 GPU. The trained model was saved for future use.

To capture the spatial distribution of cells, each sample was modeled on a 50 × 50 spatial unit grid, divided into hexagonal Visium-like spots, with an average of five cells per spot. This structure enabled the modeling of clustered and infiltrative patterns typical of the breast tumor microenvironment. Clusters of cells representing various types were assigned specific spatial parameters such as shape, twist, and scale, while infiltration patterns varied within and across clusters. Background cell types were also incorporated with proportions reflecting realistic tumor heterogeneity.

### Spatial Antibody Capture Simulation

Spatial antibody capture data was simulated for each replicate by first defining two simulated protein markers for each unique cell type in the dataset, resulting in a set of cell type-specific markers. For each simulated spatial spot, the proportion of each cell type was retrieved, and the expected expression for each marker was calculated by scaling the cell type proportion by a random factor drawn from a uniform distribution between 20 and 50. Expression values for each marker were then simulated using a negative binomial distribution with a mean proportional to the expected expression and a dispersion parameter of 0.5. A 5% dropout rate was applied to the simulated expression values, where a dropout event set the expression to zero. Low-level background expression was simulated using a negative binomial distribution with lower-scale parameters for spots where a given cell type was absent. In addition to the cell type-specific markers, nonspecific proteins were simulated across all spots using a negative binomial distribution with a constant mean expression of 10. The final output was a matrix of simulated protein expression values for each marker across all spatial spots.

To comprehensively evaluate deconvolution methods, we simulated mixed and highly segmented populations, acknowledging that actual cancer data spans a continuum between these two states. By showcasing performance across these diverse tissue architectures, we offer a thorough assessment encompassing the full spectrum of tissue heterogeneity. Our analysis emphasizes both population types: the mixed populations, which more closely resemble complex tumor microenvironments, and the highly segmented populations, which have been foundational in prior deconvolution method development. Benchmarking and simulation results for both datasets are detailed in the supplemental materials.

To account for spatial variability, we generated five distinct spatial replicates for each sample type (“high_seg” and “mixed”), each exhibiting similar yet unique spatial patterns. This approach yielded spatial gene expression matrices at single-cell and spot-level resolutions, where each spot’s gene expression values reflect the aggregated profiles of the cells within it. Additionally, we recorded the cell-type proportions per spot to provide a ground truth for evaluating spatial deconvolution accuracy. These datasets, featuring diverse spatial configurations and infiltration patterns, offer a realistic foundation for assessing different deconvolution methods on tumor-like spatial data.

### *Visium CytAssist v2 Library Preparation and Sequencing*

Before sectioning, samples were assessed for RNA quality, and all samples had a DV200 score above the minimum threshold ($\geq 30\%$). Spatial transcriptomic and proteomic libraries were generated with Visium CytAssist for the FFPE v2 gene expression kit (10x Genomics: 1000520) with the Human Immune Cell Profiling Panel (10x Genomics: PN-1000607).

Five-micron FFPE sections were placed on Schott Nexterion Hydrogel Coated Slides (Schott North America: 1800434) or Fisherbrand Superfrost Plus Slides (Fisher: 22-037-246) and processed through the 10x CytAssist for FFPE v2 protocol according to the manufacturer’s instructions. Library QC was completed with an Agilent TapeStation 4150. Libraries were normalized and pooled to 2 nM before loading on an Illumina NextSeq 2000 using a P3 100 flow cell. The pooled library was loaded at 650 pM, and sequencing was carried out with a 28/10/10/50 base pair read structure, targeting 125 million reads per sample for transcriptomic libraries, and 25 million reads per sample for proteomic libraries.

### *Total RNA Library Generation and Sequencing*

RNA was extracted from FFPE sections using the Purelink FFPE RNA isolation kit (Invitrogen: K156002). According to the manufacturer’s instructions, RNA-seq libraries were generated with the Takara SMART-Seq Stranded kit (Takara: 634447). RNA was normalized to 5 ng/$\mu$l in a total volume of 7 $\mu$l of input RNA. RNA fragmentation was not performed.

Ten cycles were used for PCR1, followed by depletion of ribosomal RNA using scZapR, and 12 cycles were completed for PCR2. Library quantification and evaluation were performed using a Qubit FLEX fluorometer and an Agilent TapeStation 4150. Libraries were normalized and pooled to 2 nM before sequencing on an Illumina NextSeq 2000 using a P4 200 flow cell. The pooled library was loaded at 750 pM. Sequencing was done with a 2×101 bp read structure, targeting 40 million reads per sample. A custom run chemistry was used to incorporate three dark cycles at the start of R2 to mask the three bases of the Takara adapter present in the read. Sequencing data was demultiplexed using the onboard Illumina DRAGEN FASTQ Generation software.

### *ddPCR for D538G Mutation*

Droplet Digital PCR (ddPCR) was performed using the QX200 Droplet Digital PCR System (Bio-Rad) to detect the D538G mutation. Reactions were prepared using the ddPCR Supermix for Probes (No dUTP) (Bio-Rad, Cat. No. 1863024) in a duplex assay with a FAM-labeled probe targeting the mutant allele and a HEX-labeled probe for the wild-type allele.

Custom PrimeTime® qPCR probes and primers were synthesized by Integrated DNA Technologies (IDT) with HPLC purification. The wild-type allele was detected using a FAM-labeled probe with the sequence: `/56-FAM/TC TAT GAC C/ZEN/T GCT GCT GGA GAT GCT /3IABkFQ/`, while the mutant allele was detected using a HEX-labeled probe with the sequence: `/5HEX/TC TAT GGC C/ZEN/T GCT GCT GGA GAT GCT /3IABkFQ/`.

Droplet generation was performed using the QX200 Droplet Generator (Bio-Rad). Following PCR, droplets were read using the QX200 Droplet Reader, and fluorescence signal intensities were analyzed using QX Manager Software (Standard Edition). The FAM to HEX fluorescence ratio was used to determine the proportion of mutant and wild-type alleles in each sample.

Reagents used in this assay included the ddPCR Supermix for Probes (No dUTP) (Bio-Rad, Cat. No. 1863024) and custom PrimeTime® qPCR probes synthesized by IDT.

### *Statistics for correlating Human Protein Atlas with CITEgeist deconvoluted signaling results*

Sender cell signals were extracted from a preprocessed AnnData object and assigned to the corresponding HPA gene expression values through a predefined dictionary that aligns cell types (e.g., ’CD8 T cells’ and ’CD4 T cells’ to ’T cells’). For each pathway—after removing irrelevant entries and extracting the ligand component for each pathway, we computed the mean sender signal for each mapped cell type. We obtained the corresponding average HPA nTPM value. Pathways with fewer than two matching cell types were excluded. For the remainder, Spearman rank correlations were calculated to assess the monotonic relationship between sender signal intensities and HPA expression. Finally, sender signal and HPA values were aggregated across pathways to compute a global Spearman correlation, visualized with a scatter plot overlaid by a linear regression line.

### *Phagocytosis Signature Validation*

To evaluate the enrichment of phagocytosis genes in surgical samples, we integrated human differential expression data (DESeq results) with a mouse phagocytosis signature from Gonzalez et al. [12]. Human genes with adjusted p-values below 0.05 were initially retained and merged with the mouse dataset based on overlapping homologous gene names. Pearson’s correlation was computed between the mouse phagocytosis $\log_2$fold change and the human surgical $\log_2$fold change, and a one-sample t-test was performed to assess whether the human fold changes significantly deviated from zero. Finally, the relationship between mouse and human fold changes was visualized using a scatter plot with an overlaid regression line and corresponding correlation statistics annotated.

### *ESR1 Mutant Signature Validation*

Gene lists corresponding to up- and down-regulated ESR1 mutant-associated genes were first extracted from the Estrogen 2.0 supplemental files [13] and filtered to retain only those in the cancer spatial dataset. For each spatial spot, an ESR1 signature score was computed as the difference between the mean expression of the upregulated genes and that of the down-regulated genes, with the resulting score appended to the metadata. Spots were then classified as ESR1 mutant or wild-type based on the “D538G Mutation” annotation, and the signature scores between these groups were compared using an unequal variance t-test, with 95% confidence intervals estimated via the t-distribution. Finally, the distribution of scores across groups was visualized using combined box and swarm plots.

### *External Tools*

COMMOT [14], GSEAPY [15], and PyDeSeq2 [16] were used as described in their respective vignettes. COMMOT was run with a 1% cutoff for signals screened. GSEAPY gene lists were derived from Scanpy Wilcoxon Rank Gene Group results with an adjusted p-value less than 0.05. PyDeSeq2 was run with size factors fit type set to ‘poscounts’ due to the sparse nature of spatial data.

### *Matched core biopsies and surgical specimens from patients on primary endocrine therapy*

As part of a prospective, pragmatic, hybrid de-centralized non-randomized clinical trial (NCT05914792) [17] designed to evaluate circulating tumor DNA (ctDNA) levels as a means to augment longitudinal monitoring of older patients with ER+ breast cancer receiving primary endocrine therapy (pET), tumor tissue samples were collected at baseline and at the time of surgical intervention [18]. In brief, patients aged 70 years and older with ER+/HER2- non-metastatic breast cancer signed informed consent (protocol approved by the University of Pittsburgh IRB under STUDY21100091 and conducted through the UPMC Hillman Cancer Center under protocol 22-088) and chose to forego upfront surgery in favor of pET. Because this trial focused on surgical de-escalation, most patients did not undergo surgery during the study follow-up. However, a subset of six patients, four responding and two progressing, based on imaging assessment and ctDNA results, underwent surgery after 3-50 months of pET (12 total specimens). We collected matched core biopsy and surgical specimens from these six patients (12 total specimens) for this analysis, providing a use case for the CITEgeist tool. To preserve patient anonymity, all samples are labelled HCC22-088-P#-S#, for HillmanCancerCenter trial 22-088 -Patient#-Sample#. Sample number is chronological; in all cases, the biopsy samples are one, and the surgical samples are 2.

## Benchmarking to State of the Art Reference-Based Deconvolution

We conducted a comprehensive comparison of CITEgeist with four prominent deconvolution methods—Seurat, RCTD, cell2location, and Tangram—chosen based on their recognition as “state-of-the-art” in recent benchmarking studies [19, 20].

Seurat, a prominent R-based framework for single-cell RNA sequencing analysis, employs an integration-based clustering approach that uses “anchor” features to transfer cell type labels probabilistically between spatial spots and reference single-cell RNAseq data [3]. While Seurat provides spot-level cell type proportions, it falls short of estimating gene expression profiles.

RCTD (Robust Cell Type Decomposition) is an R-based probabilistic framework that deconvolves spatial transcriptomics data by leveraging reference single-cell RNA-seq data. It models the spatial expression as a mixture of reference cell types and infers the cell type proportions at each spatial spot [9]. While RCTD provides accurate cell type proportions, it does not directly estimate gene expression profiles for individual cell types.

In contrast, cell2location utilizes a sophisticated Bayesian approach to deconvolution. It first estimates reference cell type gene expression profiles through negative binomial regression, then deconvolves the query expression matrix using these refined signatures [2]. This approach outputs cell type proportions and estimated cell type-specific gene expression signatures for each spot.

Similarly, Tangram is a deep learning-based method that leverages single-cell RNA-seq and spatial transcriptomics data for joint modeling. It employs a variational autoencoder (VAE) framework to map cell types to spatial locations. It estimates each spot's cell type proportions and cell type-specific gene expression profiles in a unified model [8]. Tangram's deep learning approach allows it to capture complex spatial patterns and heterogeneity in gene expression.

We deliberately selected these methods due to their established performance in deconvolving spatial transcriptomic data. Our comparative analysis employed multiple metrics to rigorously assess the accuracy of spot-level cell type deconvolution and gene expression profile inference. Each method was implemented following the standard pipeline procedures detailed in their respective methodological vignettes, and all associated scripts for each method are available via the CITEgeist GitHub repository.

### Implementation of CITEgeist

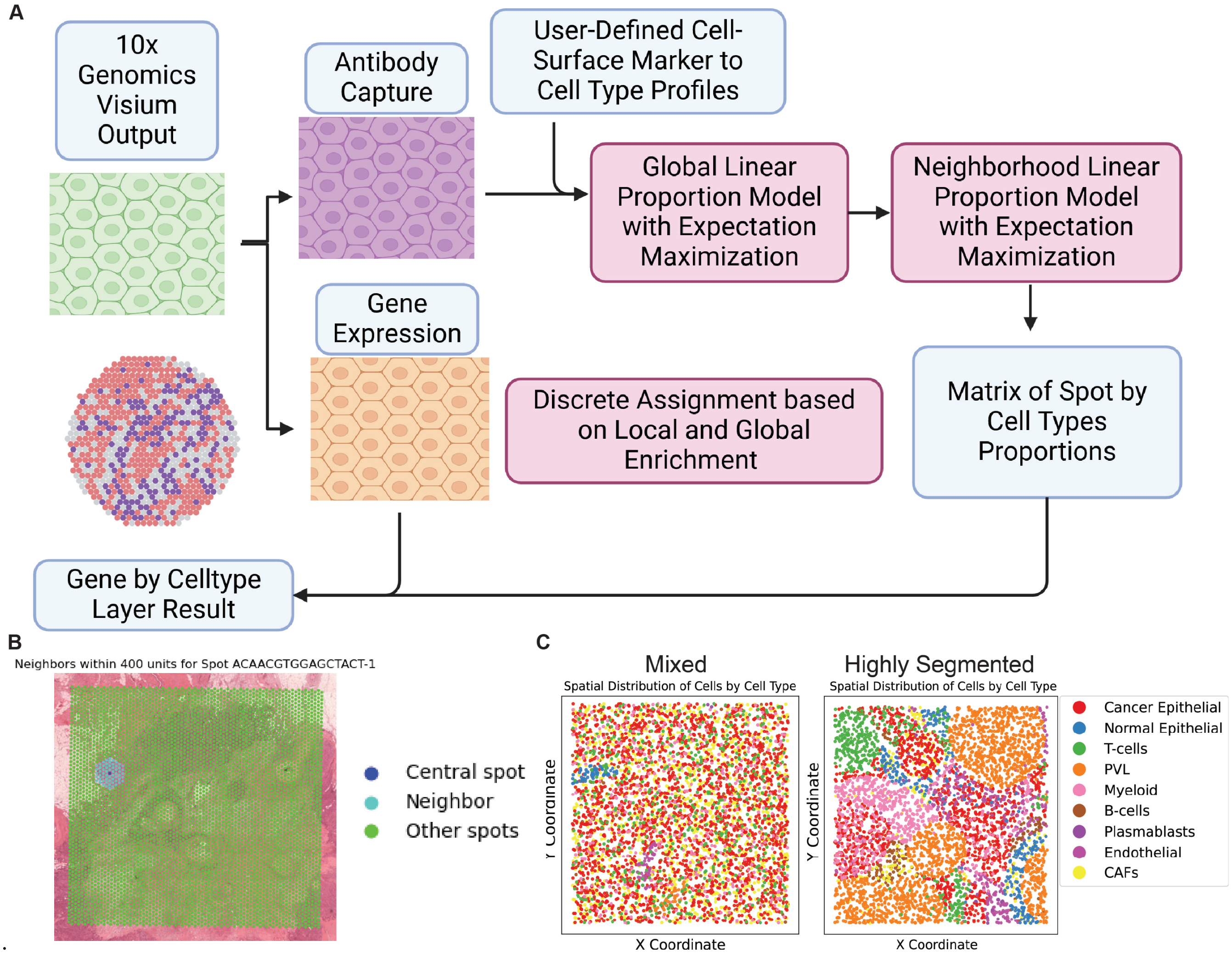

A
10x Genomics Visium Output
Antibody Capture
User-Defined Cell-Surface Marker to Cell Type Profiles
Global Linear Proportion Model with Expectation Maximization
Neighborhood Linear Proportion Model with Expectation Maximization
Gene Expression
Discrete Assignment based on Local and Global Enrichment
Matrix of Spot by Cell Types Proportions
Gene by Celltype Layer Result
B
Neighbors within 400 units for Spot ACAACGTGGAGCTACT-1
Central spot
Neighbor
Other spots
C
Mixed
Spatial Distribution of Cells by Cell Type
Highly Segmented
Spatial Distribution of Cells by Cell Type
Y Coordinate
X Coordinate
Y Coordinate
X Coordinate
Cancer Epithelial
Normal Epithelial
T-cells
PVL
Myeloid
B-cells
Plasmablasts
Endothelial
CAFs

**Figure 1.** **(A)** Graphical abstract describing the CITEgeist approach. **(B)** Example of "Neighborhood" definition in CITEgeist spatial analysis.**(C)** Examples of the spatial architecture of our "Mixed" and "Highly Segmented" test data sets.

### Pre-Processing

Before running the model, we perform the following preprocessing steps for the data. We filter genes down to those with a count $> 0$ in at least 1.0% of spots and mean expression $> 1.1$ in nonzero spots. We filter gene expression spots down to those with a minimum number of UMIs of 100 or greater. And then, we normalize counts to a target_sum of 10,000.

For the antibody capture data, we winsorize the antibody capture data to remove the top and bottom 5%, followed by global centered log ratio normalization.

### Protein Deconvolution with EM Algorithm

Our model aims to deconvolute spatial transcriptomics data by using integer linear programming (ILP) to infer the cell type proportions in each spatial spot based on antibody capture data, our model aims to deconvolute spatial transcriptomics data.Including an Expectation-Maximization (EM) algorithm enhances estimation by iteratively optimizing the cell type proportions and scaling factors.

To map antibody signals to cell types, the user defines a set of significant antibodies for each cell type based on known markers (**Figure 1A**). These antibodies are selected best to represent the cell type's known canonical surface markers. For each spot, the signals from the significant antibodies assigned to a particular cell type are averaged to generate a representative antibody signal for that cell type.

Let $N$ be the number of spatial spots, $T$ the number of cell types, and $S \in \mathbb{R}^{N\times T}$ the matrix of antibody capture data, where each entry $S_{ij}$ represents the antibody signal for spot $i$ and cell type $j$, scaled by a factor $\beta_j$. The scaling factor represents the unknown relationship between the amount of a surface marker and the recorded signal. We estimate the proportions $Y \in \mathbb{R}^{N\times T}$ of each cell type and the scaling factors $\beta \in \mathbb{R}^T$ iteratively and normalize beta values to a maximum value of 1. Elastic net regularization is applied to the proportions $Y$ to encourage sparsity and smooth informational landscapes.This ensures that only the most relevant cell types are assigned significant proportions in each spot.

The optimization at each EM step is defined as follows:

**E-step: Estimate $Y$ given $\beta$:**

$$\min_{Y} \sum_{i=1}^{N}\sum_{j=1}^{T}(S_{ij}-\beta_j Y_{ij})^2 + \lambda\left(\alpha\sum_{i=1}^{N}\sum_{j=1}^{T}|Y_{ij}| + (1-\alpha)\sum_{i=1}^{N}\sum_{j=1}^{T}Y_{ij}^2\right)$$

Subject to:

$$0 \le Y_{ij} \le 1, \quad 0.9 \le \sum_{j=1}^{T} Y_{ij} \le 1.2 \quad \text{for all } i.$$

Constraining $Y$ to a limited range of 0 to 1 limits the algorithm's search space range while allowing for interpretable results. Secondly, allowing the total of all $Y_{ij}$ to range between 0.9 and 1.2 allows for the accommodation of cell types not defined in the cell profile dictionary so that the method is robust even to variations in user accuracy. A small amount of flexibility also prevents the failure of the model to reach an optimal solution by adding flexibility in the objective function optimization.

**M-step: Update $\beta$ given $Y$:**

$$\beta_j = \frac{\sum_{i=1}^{N} S_{ij}Y_{ij}}{\sum_{i=1}^{N} Y_{ij}^2}, \quad \text{ensuring } \beta_j \ge 0 \text{ and normalized to the maximum } \beta.$$

The EM algorithm iterates until the convergence criteria are met, ensuring a robust estimation of $Y$ and $\beta$.

### Regional Refinement with Neighborhood Optimization

Although global optimization provides an initial estimate of cell-type proportions, it does not account for local spatial dependencies that can influence cell distributions. To refine these estimates, we incorporate a secondary optimization step that operates on regional neighborhoods (**Figure 1B**). This step aims to adjust the proportions of the cell types at each spatial point by incorporating information from its local vicinity, ensuring smoother transitions and greater contextual accuracy, especially in demarcated borders such as in the highly segmented dataset (**Figure 1C**).

For each spot $i$, we define a local neighborhood $\mathcal{N}(i)$ containing all spatially adjacent spots within a fixed radius $r$. The antibody signal matrix $S$ is then restricted to these local neighborhoods, and the Expectation-Maximization (EM) procedure is re-run within each neighborhood. The local optimization refines the cell type proportions $Y_{\mathcal{N}(i)}$ by solving the same model.

This formulation ensures that the inferred proportions remain biologically reasonable while allowing local adjustments. A key addition to this local optimization is the constraint in the following paragraph:

**Constraining Iterative Updates** To prevent instability in the optimization process, we introduce an explicit constraint on how much $Y$ is allowed to change between consecutive iterations during regional finetuning. This constraint ensures that each cell type proportion in a given spot does not fluctuate excessively from one iteration to the next. Specifically, we define:

$$\delta, \quad \text{where} \quad 0 \leq \delta \leq 1.$$

For each spatial spot $i$ and cell type $j$, the constraint enforces:

$$\max(0, Y_{ij}^{(t)} - \delta) \leq Y_{ij}^{(t+1)} \leq \min(1, Y_{ij}^{(t)} + \delta).$$

This guarantees that cell proportions remain biologically plausible while gradually refining their estimates. In our implementation, we set $\delta = 0.4$, allowing each $Y_{ij}$ to adjust by at most 40% per iteration while remaining in the range $[0, 1]$.

By enforcing gradual, bounded updates, this constraint prevents large oscillations in $Y$, promoting smooth convergence while maintaining the flexibility necessary for regional refinement.

The update rule for the scaling factor $\beta$ is also refined at the neighborhood level:

$$\beta_j^{(\text{new})} = \frac{\sum_{k \in \mathcal{N}(i)} S_{kj} Y_{kj}}{\sum_{k \in \mathcal{N}(i)} Y_{kj}^2}, \quad \text{ensuring } \beta_j \geq 0 \text{ and normalized to the maximum } \beta.$$

**Elastic Net Regularization for Global and Local Optimization** To enforce sparsity while maintaining numerical stability, we apply an elastic net regularization term for cell proportion during global and local optimization steps. At the global level, we set the elastic net mixing parameter to $\alpha = 0.5$, striking a balance between L1 and L2 penalties, and $\lambda = 1.0$.This ensures that cell type proportions remain sparse while allowing for smoothness across similar cell types.A grid search for finetuned local model deconvolution demonstrates that a regularization strength of $\lambda = 1.0$ and alpha value of $\alpha = 0.7$ in Mixed data (**Supplementary Figure S1**) and $\alpha = 0.9$ in Highly Segmented data (**Supplementary Figure S2**) is sufficient to enforce sparsity without over-penalizing the objective function. We will use the $\alpha = 0.9$ value on patient data for subsequent biological analysis.

**Iterative Local Refinement and Convergence** The local optimization is iteratively applied across all spatial spots, ensuring that each region is updated while maintaining consistency with its neighboring estimates. The algorithm terminates when the relative change in $Y$ and $\beta$ falls below a predefined threshold across all spots.

This method effectively finetunes the global estimates by integrating local spatial constraints while preserving biologically plausible transitions across spatial spots. The final refined estimates of $Y$ and $\beta$ represent a spatially coherent deconvolution of cell type proportions suitable for downstream analysis of cellular heterogeneity in the tissue microenvironment.

### *Spatial Gene Deconvolution with Variable-Sized Neighborhoods and Optional Priors*

Next, we use the cell-type proportion estimates from the antibody-guided model to deconvolute the observed gene expression data. Our approach integrates spatial context by assigning the same neighborhood of spots (within a user-specified radius) to each spot (**Figure 1B**), thereby accounting for local tissue structure and potential bleed-through of signals. Incorporating neighbors can improve the accuracy of allocating gene counts to distinct cell types for high-resolution assays where adjacent spots are closely spaced.

**Overview of the Approach.** Let $N$ denote the total number of spots on the array, and consider a specific spot $i$. For each spot $i$, we observe a vector of gene counts $\{X_{i,m}\}_{m=1}^{M}$, where $m$ indexes the genes ($1 \leq m \leq M$). We further define a neighborhood $\mathcal{N}_i$ as the set of spots located within a specified radius around spot $i$. Our goal is to partition the observed gene counts at spot $i$ among $T$ cell types, guided by (i) local information from $\mathcal{N}_i$, (ii) global enrichment patterns, and (iii) optional prior knowledge on expected gene expression levels.

### *Steps for Deconvolution at Spot i*

1. **Neighborhood Definition:** Identify $\mathcal{N}_i$ by selecting all spots $j$ such that the distance $d(i, j)$ is below a user-defined threshold (radius).

2. **Data Extraction:** From the neighborhood $\mathcal{N}_i$, gather:

   - **X**: The observed gene counts or expression levels for all genes, indexed as $(j, m)$ for $j \in \mathcal{N}_i$ and $m = 1, \ldots, M$.

   - **Y**: Cell-type proportions in each spot $j \in \mathcal{N}_i$, previously inferred by the EM-based antibody model.

3. **Expression-Aware Enrichment Calculation:**

   We compute an enrichment score for each cell type based on local and global expression patterns.For each gene $k \in \{1,\dots,M\}$:

   (a) *Gene Expression Vector:* Let
   $$\mathbf{g}_k = \text{expression\_data}[:,k],$$
   be the vector of gene $k$ expression across all spots in the dataset.

   (b) *Expression Threshold:* Define
   $$\theta_k = \begin{cases} \text{median}(\mathbf{g}_k \mid \mathbf{g}_k > 0), & \text{if there exists at least one nonzero } g_{ki} \\ 0, & \text{otherwise.} \end{cases}$$

   (c) *High-Expression Spots:* Let
   $$H_k = \{\, i \mid g_{ki} \geq \theta_k \}.$$
   If $H_k$ is empty, then the enrichment vector is set uniformly:
   $$E_t = \frac{1}{T} \quad \forall t \in \{1,\dots,T\}.$$

   (d) *Normalized Cell-Type Proportions:* Otherwise, let $P_{it}$ be the proportion of cell type $t$ at spot $i$. Define
   $$\tilde{P}_{it} = \frac{P_{it}}{f_t + \varepsilon},$$
   where $f_t$ is the overall frequency of cell type $t$ across the entire dataset, and $\varepsilon = 10^{-10}$ is a small constant for numerical stability.

   (e) *High-Expression Mean vs. Global Mean:*
   $$\bar{P}_t^{(H)} = \frac{1}{|H_k|} \sum_{i \in H_k} \tilde{P}_{it}, \quad \bar{P}_t^{(\text{bg})} = \frac{1}{N} \sum_{i=1}^{N} \tilde{P}_{it}.$$

   (f) *Raw Enrichment Score:*
   $$E_t = \frac{\bar{P}_t^{(H)}}{\bar{P}_t^{(\text{bg})} + \varepsilon}.$$

   (g) *Smoothing and Normalization:* First apply a smoothing factor:
   $$E_t^{(\text{smooth})} = 0.8\, E_t + 0.2,$$
   Then normalize:
   $$E_t^{(\text{final})} = \frac{E_t^{(\text{smooth})}}{\sum_{t=1}^{T} E_t^{(\text{smooth})} + \varepsilon}.$$

   (h) *Local/Global Weighting:* For each gene $k$, define
   $$E_k = w_{\text{local}}\, E_k^{(\text{local})} + w_{\text{global}}\, E_k^{(\text{global})},$$
   with user-specified weights $w_{\text{local}}$ and $w_{\text{global}}$.

4. **Optional Prior Matrix:** If an external matrix $\mathbf{P}^{(\text{prior})} \in \mathbb{R}^{T \times M}$ is available, it provides a baseline expression strength for each cell type–gene pair $(j,m)$.A user-controlled parameter $\lambda_{\text{prior\_weight}}$ then governs how strongly the model penalizes assignments that deviate from $\mathbf{P}^{(\text{prior})}$.

5. **Integer Linear Programming (ILP) Formulation:** For each gene $m$ and cell type $j$, define an integer decision variable

$$X_{j,m} \geq 0,$$

representing the number of gene $m$ counts allocated to cell type $j$ at spot $i$. Enforce the constraint that all counts at spot $i$ must be allocated:

$$\sum_{j=1}^{T} X_{j,m} = \text{center\_counts}(m),$$

where center_counts$(m)$ denotes the total observed integer count of gene $m$ in spot $i$.

6. **Objective Function (Frequency Normalization and Tie-Breaking):** We *maximize* a score that rewards assigning counts of gene $m$ to cell types with high local/global enrichment. For each pair $(j,m)$,

$$\text{term}_{j,m} = \left(\text{gene\_specific\_enrichment}_{m,j}\right) \times \left(\text{cell-type\_preference}_j\right) \times \left(\text{normalized\_weights}_j\right) \times . \left(\text{tie-break factor}\right) \times X_{j,m},$$

where the tie break factor is a small multiplicative random deviation (e.g., drawn from $[0.9, 1.1]$) to avoid ties. If a prior matrix $\mathbf{P}^{(\text{prior})}$ is used, we add a penalty term

$$-\lambda_{\text{prior\_weight}} \left(1 - P_{j,m}^{(\text{prior})}\right) \times X_{j,m},$$

The overall objective is to maximize the sum of these terms to discourage assignments that conflict with prior knowledge.

7. **Solution via ILP Solver:** We solve the above integer optimization problem using a standard ILP solver (e.g., Gurobi). The solution yields an integer allocation $\{X_{j,m}\}$ that balances:

   - Enrichment signals and local/global frequencies,
   - Tie-breaking randomness,
   - (Optional) prior-based penalties.

   The solver returns an optimal gene-count assignment across all cell types at spot $i$, satisfying the integer-count constraints while maximizing the chosen objective function.

### *Advantages of the Method*

- **Flexible Neighborhoods:** The radius-based approach adapts to diverse spatial transcriptomics technologies, from coarser to finer resolutions.
- **Local and Global Insight:** By combining local neighborhood enrichment with global signals, the method balances tissue-scale patterns and immediate spot context.
- **Optional Priors:** The framework naturally accommodates external knowledge or pre-computed profiles, penalizing assignments that deviate from a trusted baseline. To demonstrate the efficacy of CITEgeist unsupervised, we do not use it in this subsequent analysis but leave it as an option for future users or other analytical extensions in future work.
- **Integer Constraints:** Ensuring that each gene's counts are allotted exactly preserves interpretability and respects the discrete nature of observed data.

This spatial deconvolution framework provides a principled way to redistribute gene-level counts among cell types by leveraging local spatial context and global expression trends. By introducing optional priors and configurable neighborhood sizes, the method remains robust and flexible across a range of spatial transcriptomics platforms.

**Advantages of Neighborhood Context:** Incorporating adjacent spots accounts for bleed-through for higher-resolution data, enhancing deconvolution accuracy. The adjustable neighborhood radius allows adaptation to varying spatial resolutions and tissue heterogeneity.

This approach enables scalable and localized optimization, providing robust gene expression estimates while leveraging spatial information.

### Code Implementation

All code for CITEgeist modeling and deconvolution is implemented in Python 3.10. ILPs are solved using Gurobi version 11.0.2. Benchmarking is implemented in Python (3.10.15) and R (4.4.0) using custom conda environments for reproducibility, for which specific details are provided in the GitHub repository. All code is run on the University of Pittsburgh High-throughput Computing core on a 2-core, 64GB memory SLURM-executed job, with cell2location also utilizing an A100 NVIDIA GPU.

### Discussion of the Framework

The model leverages antibody capture data as a ground truth for estimating cell type proportions and gene expression profiles. Incorporating the EM algorithm improves estimation accuracy by iteratively updating scaling factors $\beta$ and cell type proportions $Y$. We hypothesize that including this scaling factor helps capture variability in cell-marker expression and marker-to-signal capture ratio. Variable-sized neighborhoods enhance spatial resolution by capturing local context, which is particularly valuable for heterogeneous tissues and allows flexibility for future cases. In our simulated framework, we found that neighborhoods, as long as they contained as many spots as the number of searched cell types, were robust to a wide range of radii, making them reliable and user-friendly. Maximum and discrete constraints on the gene expression model also lead to easily interpretable outputs with accurate assignment without additional L1 or L2 regularization

This framework is computationally efficient due to parallelization across spatial spots and neighborhoods, enabling scalability to large datasets (See **Figure S3** for specific runtime comparisons).The EM algorithm and neighborhood-based deconvolution provide a novel, adaptable approach for spatial transcriptomics analysis.While we do take more time compared to other CPU-based methods, we are comparable to Cell2Location **Figure S3** in time, and we scale rapidly with a 16-core machine taking approximately 4 minutes a sample in the simulated data set.

## Evaluation and Benchmarking

We calculated several complementary metrics to evaluate the performance of the gene count prediction and cell type proportion models. These metrics provide a comprehensive assessment of the model's ability to accurately predict gene counts and cell type compositions.

### Evaluation of Prediction of Gene Counts

For the gene count predictions, we computed the root mean squared error (RMSE), normalized root mean squared error (NRMSE), and mean absolute error (MAE) between the ground truth and predicted gene counts.Raw gene count matrices were $\log(1+x)$ normalized before metric calculation. The RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where $n$ is the number of spots, $y_i$ is the ground truth gene expression value, and $\hat{y}_i$ is the predicted value. The NRMSE normalizes the RMSE to the range or mean of the ground truth counts, allowing for comparison across cell types with different gene expression magnitudes. The MAE is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

We calculated these metrics and reported the average and median RMSE, NRMSE, and MAE across all cell types.

### Evaluation of Prediction of Cell Type Proportions

For the cell type proportion predictions, we calculated the root mean squared error (RMSE) between the true and predicted cell type proportions, as well as the Jensen-Shannon Divergence (JSD), mean absolute error (MAE), and Pearson correlation coefficient.

The RMSE for the cell type proportions is defined in the same way as for the gene count predictions, where $n$ is the number of spots, $y_i$ is the ground truth proportion, and $\hat{y}_i$ is the predicted proportion.

We also measured the JSD between the true and predicted cell-type composition vectors for each spatial spot. The JSD provides a symmetric, bounded measure of the distance between two probability distributions $P$ and $Q$ and is defined as:

$$\text{JSD}(P||Q) = \frac{1}{2}D_{\text{KL}}(P||M) + \frac{1}{2}D_{\text{KL}}(Q||M)$$

where $D_{\mathrm{KL}}$ is the Kullback-Leibler divergence and $M = \frac{1}{2}(P+Q)$. We reported the median JSD for all spots as the primary metric.

Additionally, we calculated the Pearson correlation coefficient between the true and predicted cell type proportions to evaluate the linear relationship between the two.

### *Significance Testing*

To evaluate differences in performance across various deconvolution methods (namely Seurat, RCTD, cell2location, Tangram, and CITEgeist) for spot-level cell type deconvolution, we performed a one-way analysis of variance (ANOVA). The method was the independent variable, while the selected metric was the dependent variable. Post-hoc pairwise comparisons were conducted using Tukey's Honest Significant Difference (HSD) test to identify specific group differences. Similarly, we utilized one-way ANOVA and post-hoc Tukey's HSD to determine significant differences in metrics between cell2location, Tangram, and CITEgeist for the GEX layer deconvolution task. Statistical significance was determined at $p < 0.05$, with adjusted $p$-values reported for multiple comparisons.

# RESULTS

To demonstrate the Validity of our method, we created two separate test datasets (**Figure 1C**): a set of mixed samples ($n = 5$) that emphasizes the mixed nature of tumor microenvironment populations and a set of highly segmented samples ($n = 5$) that more closely approximates the mouse brain dataset used to validate other State of The Art (SOTA) tools such as Cell2Location [2].

## Simulation

| **Cell Type** | **Mixed Simulation** | | | | | | **Highly Segmented Simulation** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S1 | S2 | S3 | S4 | S5 | **AVG** | S1 | S2 | S3 | S4 | S5 | **AVG** |
| Cancer Epithelial. | 2175 | 2165 | 2209 | 2184 | 2199 | **2186.4** | 1208 | 295. | 527. | 843. | 641. | **702.8**. |
| CAFs | 1485 | 1509 | 1513 | 1528 | 1527 | **1512.4** | 247. | 193. | 385. | 401. | 143. | **273.8**. |
| T-cells | 491. | 473. | 498. | 467. | 478. | **481.4**. | 413. | 369. | 345. | 416. | 821. | **472.8**. |
| Myeloid | 456. | 464. | 456. | 443. | 486. | **461** | 617. | 518. | 337. | 228. | 357. | **411.4**. |
| B-cells | 246. | 220. | 249. | 225. | 238. | **235.6**. | 194. | 272. | 458. | 214. | 262. | **280.0**. |
| Normal Epithelial. | 52. | 45. | 0 | 61. | 0 | **31.6**. | 285. | 602. | 505. | 858. | 236. | **497.2**. |
| Plasmablasts | 34. | 35. | 27. | 2 | 38. | **27.2**. | 21. | 233. | 113. | 137. | 280. | **156.8**. |
| PVL | 23. | 41. | 0 | 38. | 34. | **27.2**. | 1650 | 1774 | 1848 | 1340 | 1524 | **1627.2** |
| Endothelial | 38. | 48. | 48. | 52. | 0 | **37.2**. | 365. | 744. | 482. | 563. | 736. | **578.0**. |
| **TOTAL** . | **5000** | **5000** | **5000** | **5000** | **5000** | **-** | **5000** | **5000** | **5000** | **5000** | **5000** | **-** |

**Table 1.** Summary of cell type counts across Mixed and Highly Segmented Visium replicates simulated via the scCube framework.

Using the scCube framework, we generated a comprehensive simulated dataset comprising five highly heterogeneous ("mixed") spatial transcriptomic samples and five highly segmented samples, each containing 5,000 cells distributed across nine distinct cell types. The simulated tissue architecture, specifically the Mixed population, was aimed at reflecting the typical cellular composition of breast cancer, with cancer epithelial cells representing the dominant population (mean = 2,187 cells), followed by cancer-associated fibroblasts (CAF) (mean = 1,513 cells). Immune populations were represented at physiologically relevant frequencies, with T-cells, Myeloid cells, and B-cells at intermediate levels (482, 461, and 236 cells on average, respectively). The remaining stromal and epithelial components—Normal Epithelial cells, Plasmablasts, PVL (Perivascular-like), and Endothelial cells—were present at lower frequencies (ranging from 28 to 33 cells on average), consistent with typical breast tumor composition (**Table 1**).

Notably, the simulation maintained consistent total cell counts (5,000 cells per sample) while introducing biologically relevant variation in cell type proportions across samples. This variation was particularly evident in less abundant populations, such as Plasmablasts (ranging from 2 to 38 cells) and PVL cells (0 to 41 cells), mimicking the heterogeneity observed in actual tumor samples. The simulation framework successfully generated spatial patterns characteristic of breast cancer tissue architecture. Cancer Epithelial cells and CAFs formed the bulk of the tissue mass while maintaining realistic infiltration patterns of immune and stromal populations.

Each sample was structured on a 50 × 50 spatial unit grid with Visium-like spot resolution, incorporating both clustered and infiltrative patterns typical of the tumor microenvironment (**Figure 1C**). This spatial organization and the cell type-specific gene expression profiles generated through VAE training provided a robust foundation for evaluating deconvolution method performance under realistic conditions.

## Benchmarking

### *Benchmarking Proportion Deconvolution*

Our benchmarking analysis assessed CITEgeist's ability to estimate cell type proportions relative to leading computational methods, including Seurat, RCTD, Tangram, and cell2location. Using simulated heterogeneous datasets designed to challenge spatial deconvolution methods, CITEgeist consistently outperformed or performed on par with most competing methods, except RCTD in specific scenarios.

CITEgeist demonstrated high accuracy in predicting cell type proportions across both highly segmented and mixed replicates (**Figure 2A**). Across both spatial patterns, CITEgeist achieved strong correlation scores (Corr of 0.869 in mixed and 0.948 in highly segmented conditions), outperforming most methods except RCTD (which achieved correlations of 0.964 and 0.973, respectively). For error-based metrics, CITEgeist exhibited low error rates, with MAE of 0.051 and RMSE of 0.090 in mixed conditions, performing substantially better than Cell2Location (MAE: 0.089, RMSE: 0.168) and Seurat (MAE: 0.084, RMSE: 0.134). In the highly segmented condition, CITEgeist maintained competitive performance (MAE: 0.034, RMSE: 0.084) comparable to Cell2Location (MAE: 0.026, RMSE: 0.083). In contrast, Tangram showed notably higher error rates (mixed condition MAE: 0.106, RMSE: 0.136; high segmentation MAE: 0.113, RMSE: 0.180) and low correlation (0.596 in mixed, 0.585 in high segmentation), regardless of the underlying pattern in the replicates.

Notably, CITEgeist demonstrated moderate JSD values of 0.322 in mixed and 0.154 in highly segmented conditions, indicating good similarity between predicted and ground-truth distributions. The performance gap was particularly evident between CITEgeist and methods like Seurat (JSD: 0.448 mixed, 0.259 high segmentation) and cell2location (JSD: 0.460 mixed, 0.056 high segmentation), which exhibited larger divergence in the Mixed condition.

While RCTD performed well in both mixed and highly segmented simulated data, achieving the best results (mixed condition: JSD: 0.131, MAE: 0.025, RMSE: 0.049, Corr: 0.964; high segmentation: JSD: 0.145, MAE: 0.022, RMSE: 0.059, Corr: 0.973), its performance degraded significantly when the reference was downsampled (mixed condition: JSD: 0.710 [+0.579], MAE: 0.149 [+0.124], RMSE: 0.181 [+0.132], Corr: 0.144 [-0.821]; high segmentation: JSD: 0.782 [+0.637], MAE: 0.166 [+0.144], RMSE: 0.244 [+0.185], Corr: 0.253 [-0.720]).

Beyond numerical accuracy, CITEgeist (**Figure 2B**) demonstrated superior performance in reconstructing both cell type proportions and spatial organization. It accurately captured infiltrative patterns, resolved boundaries between clusters, and identified smaller populations such as Plasmablasts and Endothelial cells, which other methods like Seurat and Tangram struggled to detect. While RCTD exhibited reasonable spatial localization—improving upon Seurat and Tangram by 20–50%, depending on the metric—it remains fundamentally limited by its inability to infer gene expression layers.

CITEgeist demonstrated superior performance in reconstructing spatially resolved gene expression profiles compared to Tangram and Cell2Location. In the mixed dataset, which contains spots composed of multiple cell types, CITEgeist achieved significantly lower RMSE and MAE than competing methods across all nine reference cell types (**Figure 3**). In particular, CITEgeist maintained a low mean absolute error ($<0.1$), whereas Cell2Location and Tangram exhibited much higher error rates, ranging from 0.2 to 0.75. RCTD and Seurat were excluded from this analysis as they do not infer gene expression profiles.

A more detailed statistical comparison of Cell2Location, CITEgeist, and Tangram further supports CITEgeist's superior accuracy in the mixed dataset. On average, Cell2Location exhibited an RMSE of approximately 0.610 with a standard deviation of 0.095, whereas CITEgeist demonstrated a significantly lower RMSE of 0.272 with a standard deviation of 0.201. Tangram, in contrast, showed the highest RMSE values, highlighting its reduced accuracy in reconstructing gene expression in mixed spots. The differences in RMSE among the three methods were statistically significant, as confirmed by a one-way ANOVA ($p = 1.529655 \times 10^{-41}$, $F = 213.9206$). Tukey's post-hoc test identified substantial differences between Cell2Location and CITEgeist ($p < 0.001$) as well as between Tangram and both Cell2Location and CITEgeist ($p < 0.001$), indicating that CITEgeist significantly outperformed the other two methods in RMSE.

In contrast, the MAE values in the mixed dataset were more comparable between Cell2Location and CITEgeist. Cell2Location averaged an MAE of 0.066 (standard deviation: 0.038), whereas CITEgeist had a similar MAE of 0.064 (standard deviation: 0.080), and Tangram exhibited a substantially higher MAE of 0.312 (standard deviation: 0.129).

In high-segmentation (High_Seg) datasets, where a single cell type dominates each spot, CITEgeist continued to perform well but showed more comparable results to Cell2Location. While RMSE remained significantly lower for CITEgeist than Cell2Location ($p = 2.21 \times 10^{-15}$), indicating better overall reconstruction accuracy, the MAE differences were not statistically significant ($p = 0.273$). While CITEgeist excels in resolving complex mixed-cell environments, it remains robust in high-segmentation settings, achieving similar absolute errors to existing deconvolution methods.

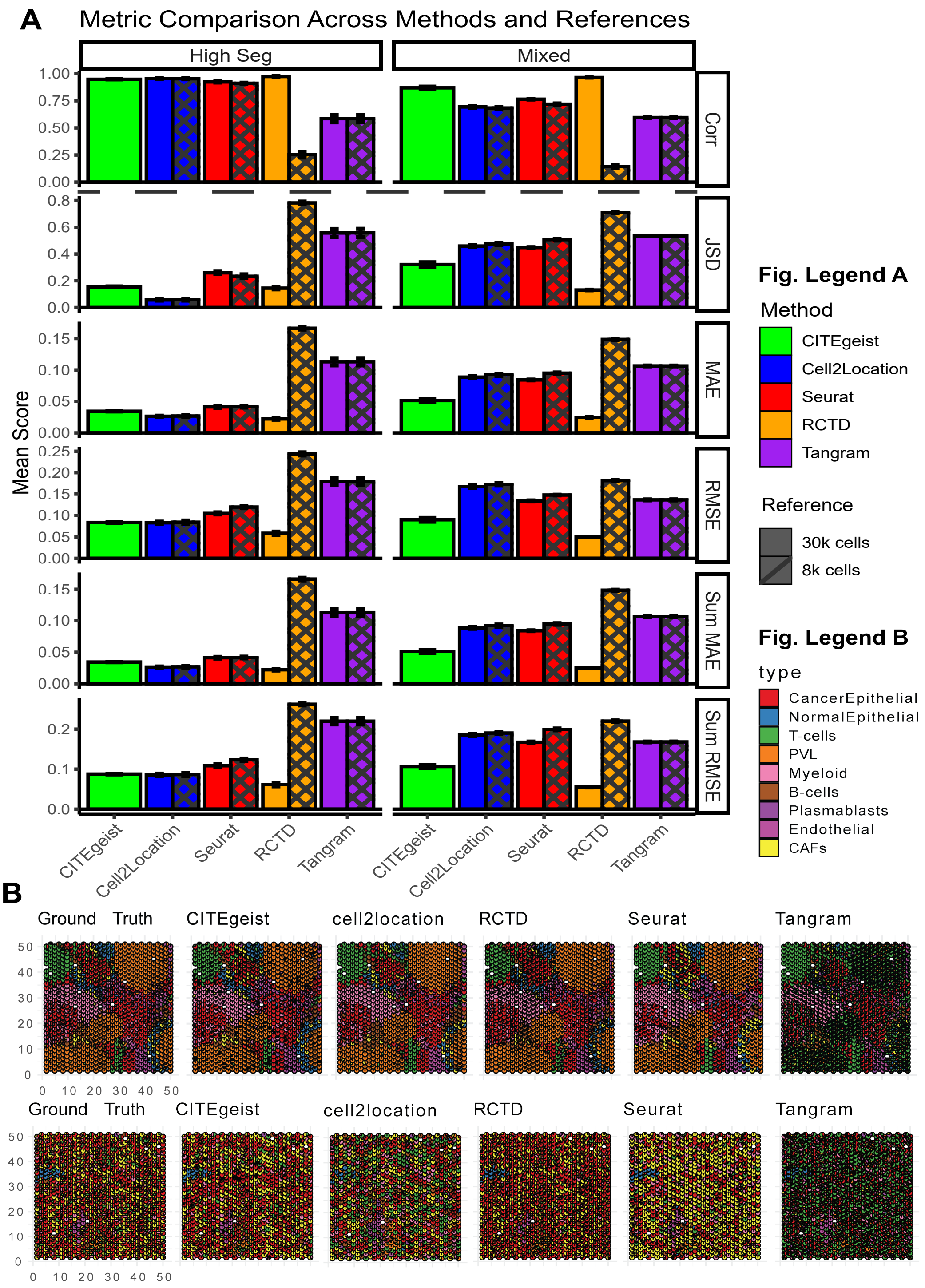


**Figure 2. Performance evaluation of CITEgeist and reference-based cell type proportion deconvolution methods on simulated Visium datasets.**. **(A)** Quantitative comparison of deconvolution accuracy across different spatial patterns ("Highly Segmented" and "Mixed") and reference atlas sizes (30,000 or 8,000 cells). Metrics include Pearson's Correlation (Corr), Jensen-Shannon Divergence (**JSD**), Mean Absolute Error (**MAE**), Root Mean Square Error (**RMSE**), mean MAE across spots (**Sum MAE**), and mean RMSE across spots (**Sum RMSE**). Error bars represent standard error (n = 5 independent replicates per condition) **(B)** Representative deconvolution results from benchmarked methods for Highly Segmented (**top**) and Mixed (**bottom**) samples. Each panel displays estimated cell-type proportions across spatial spots. Representative images were selected from n = 5 independent simulations per condition.

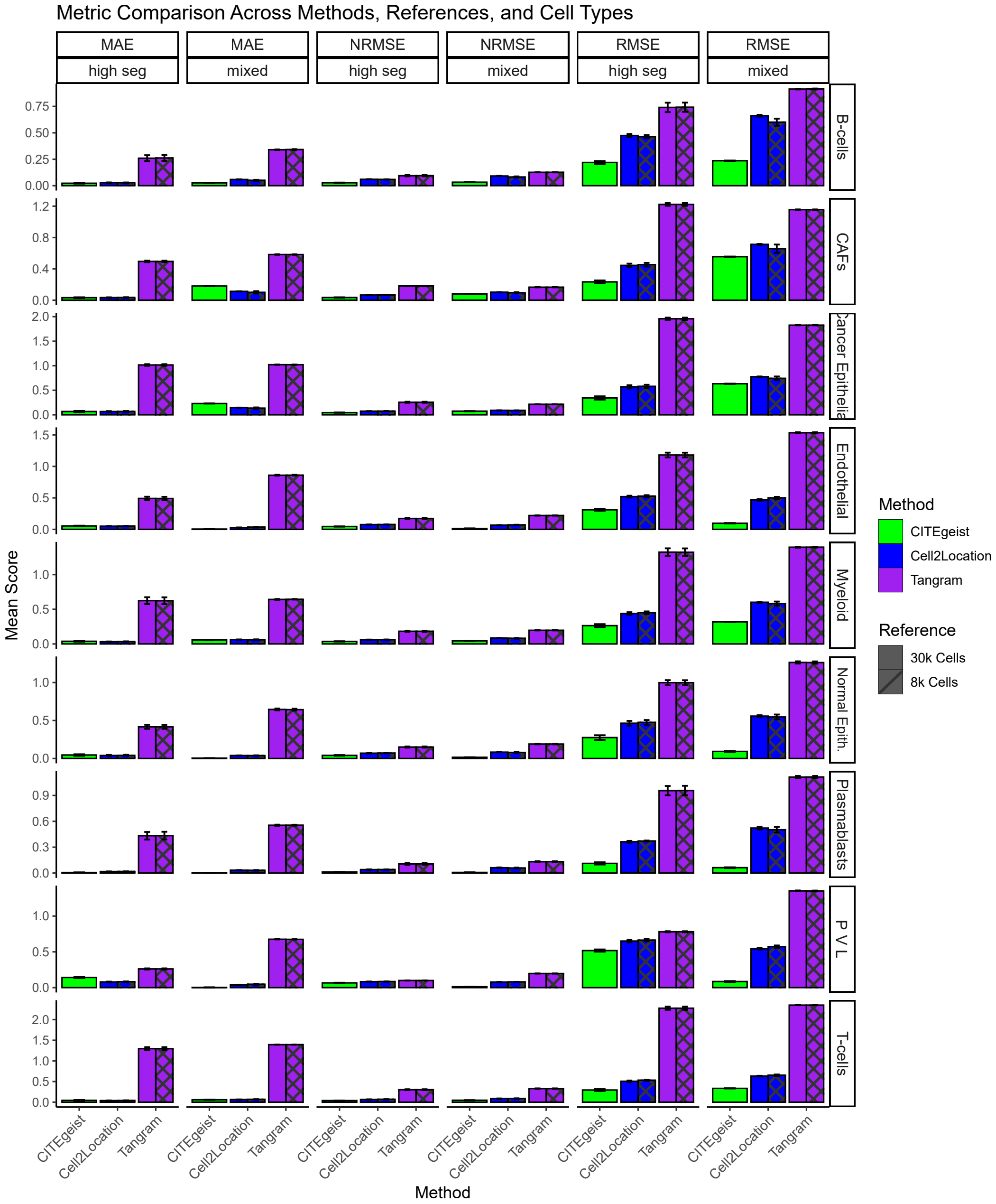


**Figure 3.** Benchmarking Gene Expression (GEX) Deconvolution Performance. This figure evaluates the performance of CITEgeist in deconvoluting cell-type Gene Expression (GEX) layers, comparing it to a reference-based deconvolution method. The performance is assessed using three metrics: **RMSE** (Root Mean Square Error), **NRMSE** (Range-Normalized Root Mean Square Error), and **MAE** (Mean Absolute Error). Higher values for these metrics indicate poorer deconvolution performance. The reference levels indicate the number of cells in the scRNA-seq reference dataset, which were benchmarked using reference sizes of either 30k or 8k cells for the reference-based methods. Metrics were calculated for each cell type's GEX profile for each replicate, with **N = 5** for each data type (high-seg or mixed). Error bars represent the standard error across replicates.

These findings underscore the robustness of CITEgeist in reconstructing spatial gene expression patterns without relying on single-cell reference data. Unlike competing methods, CITEgeist preserves fine-scale spatial heterogeneity and accurately captures tumor microenvironment complexity regardless of tissue architecture.

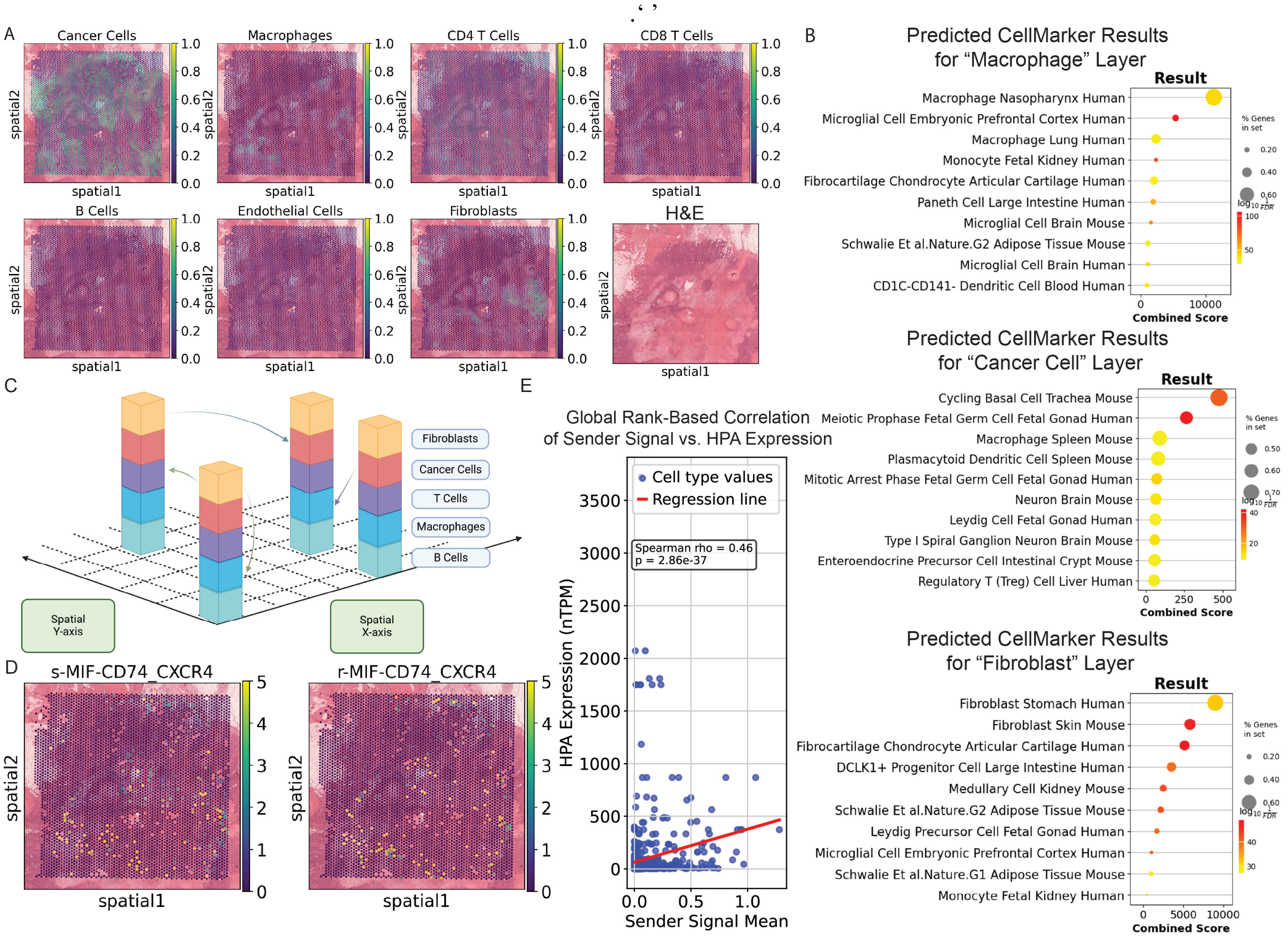


**Figure 4.** **(A)** Spatial plots showing the predicted proportions from CITEgeist output of the celltypes in sample HCC22-088-P1-S2 **(B)** CellMarker 2024 GSEApy calls, showing the top predicted celltypes for their respective 'assigned' gene expression layer. **(C)** Graphical abstract demonstrating how AnnData objects permit stacking of cell' compartments' in the same adata spatial spot, and how COMMOT calculates using this information **(D)** Spatial plot showing an example of the detected sender and receiver signal for pathway MIF-CD74_CD44. **(E)**.Global correlation regression plot showing the correlation between cell type assigned sender signal and said ligands expression in the Human Protein Atlas.

## CITEgeist accurately deconvolutes tumor microenvironment cell-type-specific gene expression and cell-type-specific signalling in an unsupervised manner

To validate CITEgeist's ability to deconvolute gene expression layers that have demonstrable downstream analytic utility, we used a set of tissue specimens collected from clinical trial NCT05914792 [17]. This is an ongoing clinical trial determining ctDNA as a means to augment monitoring of older women with ER+ breast cancer who receive pET and forgo surgery. To validate and demonstrate the utility of CITEgeist, we devised three hypotheses/analytical tasks. First, can CITEgeist accurately deconvolute gene markers and biological activity using only antibody information? Secondly, can samples deconvoluted by CITEgeist be integrated for downstream analysis? Lastly, can CITEgeist deconvoluted results uncover novel biology? For this first test, we used sample HCC22-088-P1-S2, a post-treatment surgical sample from a patient progressing on imaging after 50 months of endocrine therapy.

Our first test assessed CITEgeist's ability to deconvolute cell type proportions and gene expression layers, even without ground truth. To focus on spots with any given cell type, we filter each adata object layer to retain only those where a single cell type comprises at least 20% of the total composition, based on the heuristic that approximately five cells can fit in a single

Visium spot [2]. We extracted each filtered subset's corresponding cell-type-subset gene expression layer, creating sub-data objects that capture distinct cell-type profiles. These are concatenated into a unified object, appending the cell type as a suffix to maintain identity while allowing comparative analysis across layers.

We then analyzed the rank gene group results to identify significantly upregulated genes in each inferred cell type. As shown in **Figure 4A**, some cell types are rare but align well with expected population distributions, particularly dominant cell types in the tumor microenvironment such as macrophages, cancer cells, and fibroblasts [21]. This suggests that CITEgeist effectively reconstructs spatially relevant cell-type populations even without explicit ground truth.

We then identify the significantly upregulated genes in each predicted population, with a log fold change more significant than one and an adjusted p-value less than 0.05. These gene lists are then inputted into the GSEApy CellMarker 2024 gene set [15], and the predicted cell-type calls are shown in **Figure 4B**. These results demonstrate that using proportion-based deconvolution alone, CITEgeist can accurately identify marker genes and assign them to the correct layer despite having no prior information. Macrophages and Fibroblasts are immediately identified with a high degree of accuracy, and cycling basal cells are effectively equivalent to Cancer Cells due to the stem-like proliferative nature of cancer gene expression profiles, a known difficulty in standard cell type callers [22].

Additionally, CITEgeist is built to be fully functional within the Scanpy ecosystem [23]. Scanpy Anndata objects allow spots to have the exact spatial coordinates, and spatial plotting functions, by default, plot the highest value in a spot. A simple function provides for the expansion and re-merging of different cell-type compartments vertically within the same Scanpy object, as seen in **Figure 4C**. We then demonstrate that such a package is fully functional within externally developed tools such as the COMMOT cell-to-cell communication package [14], and call cross-celltype pathways of interest, such as MIF-CD74_CXCR4 in **Figure 4D**. To further validate the signals' accuracy and precise cell types assignment, such as chemokines to macrophages and fibroblasts. We analyze the concordance between predicted sender-signal strengths from spatial transcriptomics data and corresponding gene expression measurements from the Human Protein Atlas (HPA) [24]. For each pathway detected in the "Sender" list, the ligand is extracted from the pathway name and used to subset the HPA data. Sender cell types (extracted from the spatial data) are then mapped to their corresponding HPA cell types via a predefined dictionary. The mean sender signal and the average HPA nTPM are computed for each mapped cell type. Only pathways with at least two matching cell types are retained, and for these, a Spearman rank correlation is calculated between the sender signal means and the HPA expression values. Finally, the aggregated data across pathways generates a global scatter plot with a fitted regression line. The global Spearman correlation coefficient (rho = 0.46) and its p-value ($p = 2.86 \times 10^{-37}$) is annotated on the figure **Figure 4E**, thereby providing a rank-based assessment of how well the predicted signals correspond to established HPA expression profiles. This demonstrates that the secreted protein signals and their respective cell-type assignments from CITEgeist and COMMOT correlate well with external databases regarding which cell types commonly secrete these ligands.

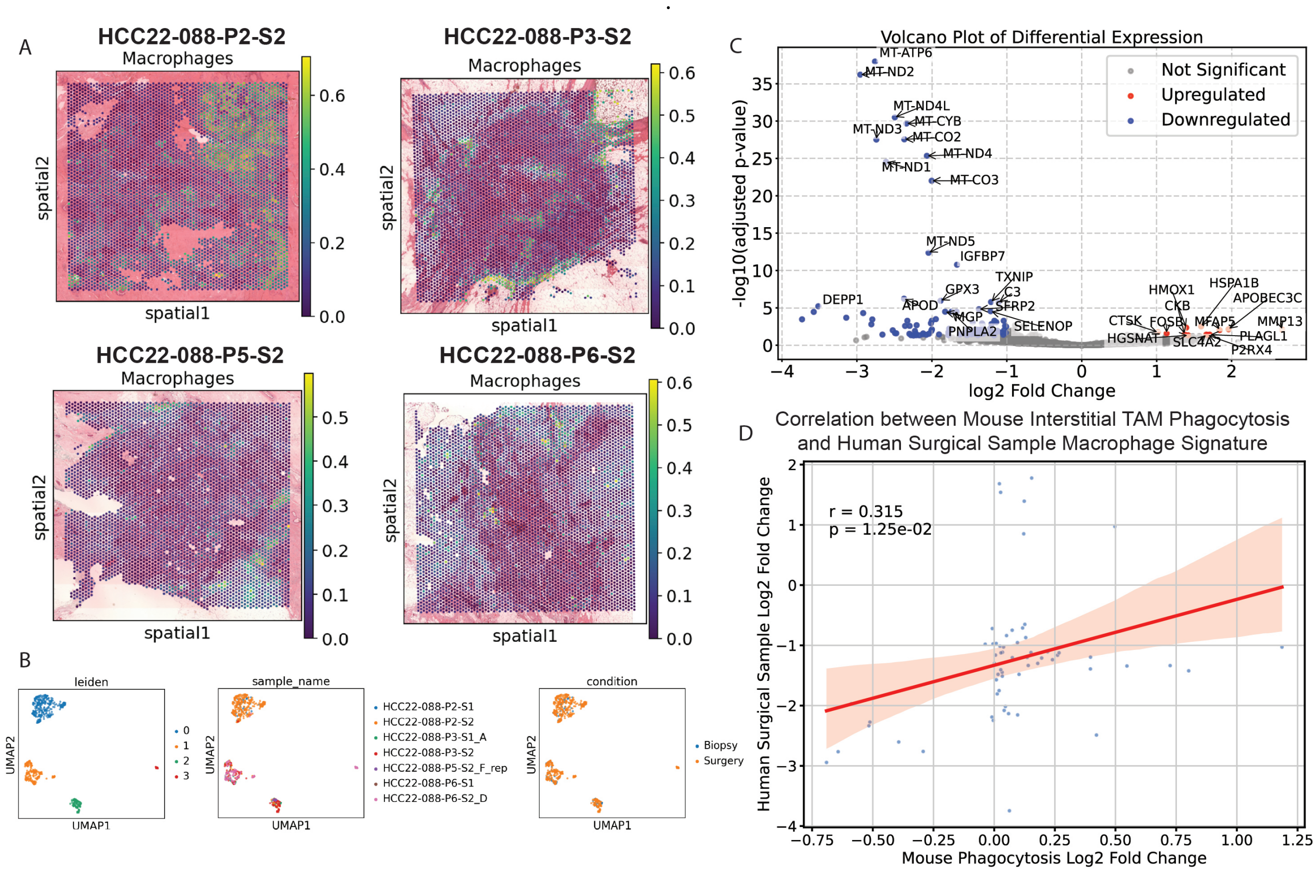


**Figure 5.** **(A)** Spatial plots showing CITEgeist Macrophage proportions in breast cancer surgical samples from the clinical trial in the four patients showing response by imaging to pET **(B)** UMAPs showing successful Harmony integration of Macrophage gene expression layers both across patients and across conditions. **(C)** PyDeSeq2 results show significantly up and downregulated genes between pre-treatment biopsy and post-pET-treatment surgical conditions. Genes expressed higher in the biopsies are on the left in blue, and genes expressed higher in the surgical samples are on the right in red. **(D)** Correlation plot confirming that major expression differences between pre-treatment biopsy and post-pET-treatment surgical macrophage gene expression correlate with differences found between tumor-associated interstitial macrophages performing phagocytosis based on the gene signature from Gonzalez et al.

## CITEgeist results can be integrated with Harmony, revealing interpretable underlying biological changes across samples.

Of critical concern for such a tool is its ability to integrate the outputs across multiple patients. Due to the discrete assignment component of CITEgeist, resultant outputs effectively resemble normalized count values and thus are usable for most tools. Here, we extract the Macrophage layer in eight samples from four patients (HCC22-088-P2-S2, HCC22-088-P3-S2, HCC22-088-P5-2, HCC22- 088-P6-S2) who had an imaging response to pET at the time of surgery, but elected removal of the tumor (**Figure 5A**). First, we demonstrate in **Figure 5B** that these layers are fully compatible with Harmony integration [25]. Secondly, to demonstrate biological utility, we hypothesize that the increase in macrophage infiltration in the responding surgical samples resulted from increased macrophage involvement in the clean-up of dead cancer cells or fibrotic tissue [26]. To validate this, we conducted DeSeq2 analysis via the PyDeSeq2 [16] package shown in **Figure 5C**. Then we utilized a tumor-associated interstitial macrophage-specific phagocytosis signature from Gonzalez et al. [12] and showed that the significantly upregulated and downregulated genes (Adjusted p-value $< 0.05$, $|logFC| > 1$) between the pre-treatment biopsy samples and post-pET-treatment surgical samples correlated with said signature, with a correlation coefficient of 0.315, p-value of 1.25e-02. Supporting that hypothesis and validating CITEgeist's utility in downstream analysis.

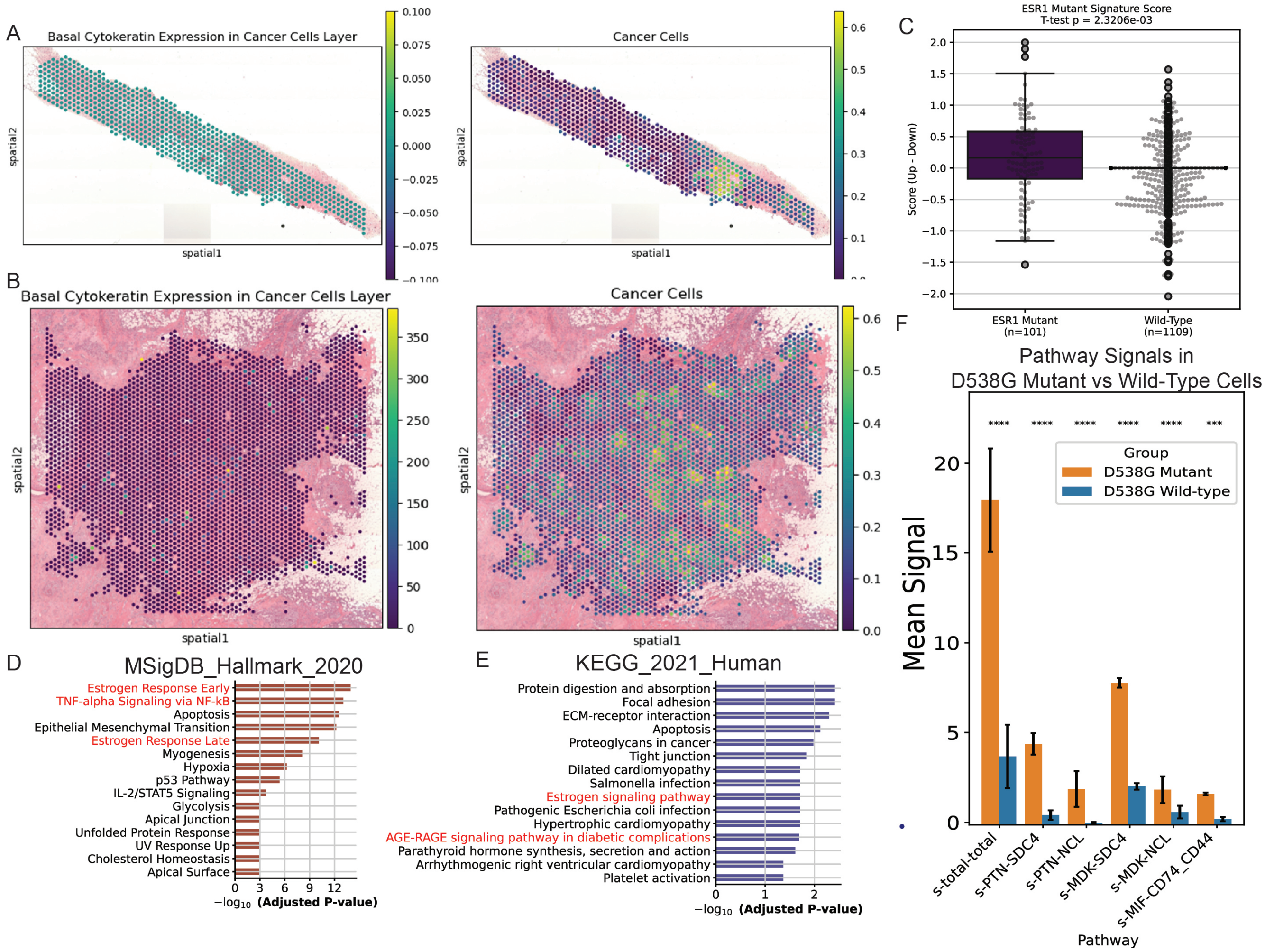


**Figure 6.** **(A)** Spatial plots showing the predicted cancer population in sample HCC22-088-P4-S1, and zero expression of basal cytokeratins in the Cancer Cells layer. **(B)** Spatial plots showing the heterogenous deconvoluted result of elevated basal cytokeratin expression in the Cancer Cells layer in the surgical sample from the same patient in sample HCC22-088-P4-S2, which has an ESR1 D538G mutation. **(C)** Combined box and scatter plot showing the average ESR1 Mutant Gene Signature Score between ESR1 D538G mutant cancer cells and WT cells. (T-test value 3.1229, p-value $2.3206e{-}03$ **(D)** Gene pathways upregulated in the D538G cancer spots from the MSigDB_Hallmark_2020 gene set. Red text pathways are previously confirmed upregulated activities in ESR1 mutant cells from Li et al. [27] **(E)** Gene pathways upregulated in the D538G cancer spots from the KEGG_2021_Human gene set. Red text pathways are previously confirmed upregulated activities from Li et al. [27] **(F)** Bar plot showing the statistically significant COMMOT sender signal differences between D538G cancer spots and WT cancer spots. A Mann-Whitney U test was performed to compare pathway-specific sender signals between cells with and without the D538G mutation. Error bars are SEM. The two-way test statistic and p-value were computed, and pathways were ranked by significance. The mean difference in signal between mutation-positive and mutation-negative cells was calculated. Benjamini-Hochberg FDR correction was applied to control for multiple comparisons, and pathways with FDR $< 0.05$ were identified as significantly different.

## CITEgeist results correlate well with known biological signals and have interpretable outputs for uncovering new biological mechanisms

In the NCT0591479 dataset, a D538G ESR1 mutation was identified in the surgical sample HCC22-088_P4_S2 on RNA sequencing that was previously not detected in the biopsy (**Supplemental Figure S4**). Previous research demonstrates that such mutations increase basal cytokeratin expression in cancer cells [27]. Thus, to validate the ability of CITEgeist to find such features, we assessed the ability of CITEgeist to deconvolute said increase in basal cytokeratin expression.

By analyzing the patient's pre-treatment biopsy (HCC22-088-P4-S1) and post-treatment surgical resection (HCC22-088-P4-S2) via CITEgeist, we demonstrate that the elevated basal cytokeratin expression expected from the D538G ESR1 mutation is accurately assigned to the Cancer Cell population in the surgical sample and is non-existent in the biopsy sample that does not have a D538G mutation. Furthermore, it demonstrated heterogeneous patterns within the tumor previously suspected but difficult to validate, in line with the heterogeneous nature of breast cancer lesions, as well as the allelic frequencies seen in ddPCR (**Supplemental Figure S5**).

Furthermore, to extend upon this finding, we took the ESR1 mutant gene signature from the EstroGene2.0 paper [13] and showed that it was in concurrence with the mutant cells detected in this sample. ESR1 mutant cells in this sample had an average score of 0.173 (95% CI $0.0384-0.3083$), while the WT cells had an average score of $-0.041$ (95% CI $-0.0595--0.0227$)—a statistically significant increase with a p-value of $2.32\times10^{-3}$ (**Figure 6C**). Furthermore, analysis of the pathways using rank gene groups and GSEApy in **Figure 6D** and **Figure 6E** showed that spots with ESR1 D538G mutant cancer cells were upregulated in multiple pathways, including inflammatory pathways, estrogen response pathways, and AGE-RAGE signaling, consistent with previous research that identified these same changes in ESR1 mutant breast cancer cells [27].

Additionally, we conducted COMMOT signaling analyses to compare the most prominently elevated sender signals in D538G mutant cancer spots relative to wild-type (WT) spots (**Figure 6F**). We observed a striking upregulation in the s-total-total pathway, which exhibited a mean value of $19.244\pm3.287$ SEM in mutant cells compared to $3.612\pm0.562$ SEM in WT (FDR-adjusted $p=5.57\times10^{-29}$). Likewise, s-PTN-SDC4 increased to $6.019\pm2.225$ SEM in D538G mutants, in contrast to $0.277\pm0.149$ SEM in WT (FDR-adjusted $p=8.09\times10^{-36}$), while s-PTN-NCL rose to $2.111\pm0.765$ SEM compared to $0.104\pm0.075$ SEM (FDR-adjusted $p=6.38\times10^{-26}$). We further noted that s-MDK-SDC4 reached $7.608\pm1.752$ SEM in D538G mutants versus $2.221\pm0.375$ SEM in WT (FDR-adjusted $p=9.33\times10^{-13}$), whereas s-MDK-NCL showed a mean of $1.876\pm0.611$ SEM relative to $0.708\pm0.180$ SEM (FDR-adjusted $p=1.41\times10^{-12}$). Finally, s-MIF-CD74_CD44 was also significantly elevated in D538G mutants ($1.631\pm0.731$ SEM) compared to WT ($0.301\pm0.101$ SEM; FDR-adjusted $p=8.51\times10^{-4}$). These findings highlight the markedly enhanced signaling activity in D538G mutants across multiple pathways.

# DISCUSSION

Here, we propose a novel algorithm for deconvoluting spatial sequencing data by leveraging proteomic information from the same slide. This approach is uniquely applied to clinical specimens, providing a direct and biologically grounded method for cell-type deconvolution. To our knowledge, this is the first method of its kind, and we demonstrate its superior performance compared to prevailing single-cell reference-based approaches. Notably, the clinical trial from which these specimens were obtained is currently in follow-up, further underscoring the translational relevance of our work.

## Advantages of using less-sparse protein signal information

The main advantage of our approach is that it is intuitive. RNA molecular information begins to degrade once the tissue is removed from the patient's body. Furthermore, the amount of any given canonical transcriptomic marker does not have a one-to-one correlation to its protein levels in any given cell. Protein markers, by contrast, are much more stable structures, have long-lasting signals even hours later, and are not as sensitive to input chemistries [28]. Furthermore, by using antibody information from the same specimen, we negate the need to develop a scRNAseq reference and avoid problems from heterogeneous cancer samples that add noise to the informational landscape.

## Advantages of a 'reference-free' deconvolution approach

We propose that our method is better than current single-reference-based methods for several reasons. Firstly, it is significantly less expensive. Single nuclei sequencing, at minimum, doubles the experiment cost, while antibody panels (as of November 2024) cost a fraction of the amount; this cost increases if a more extensive reference is required.

Additionally, there are concerns regarding the availability of tissue. A single nuclei extraction on the same block from which spatial transcriptomic data is obtained can typically require four 20um scrolls to reach the necessary numbers of intact nuclei [29]. Smaller tissue samples, such as diagnostic biopsies (which provide crucial pre-treatment molecular information), may not have enough.

Lastly, spatial and single-cell transcriptomics information landscapes will likely not be the same. Even in the ideal scenario, the single nuclei information is, by definition, from sections two to four cuts beneath the section used in spatial sequencing. This informational shift is especially true in cancer examples where regional changes in blood supply access and tissue density can have significant effects upon underlying biology [30]. Additionally, by transferring a single-cell informational landscape from single nuclei sequencing down to spatial sequencing, there is a risk of obfuscating accurate spatial information. This is because spatial sequencing is even sparser than single-cell sequencing datasets: by using a few markers cross-referencing anchors to 'paint' single-cell data onto spatial landscapes, there is a high risk of giving false spatial context to analysis that only is true in single-cell contexts, such as obfuscating real immune infiltration effects and variance in underlying spatial data vs. single-cell sequencing data (See **Figure 2B**, where Tangram over-calls T-cells due to the distinctiveness of their immune cell single-cell RNA signatures in comparison to other cell types, not due to their spatial presence in the dataset).

CITEgeist demonstrates superior performance compared to all other methods, even when evaluated with atlas-level references containing 30,000 single-cell profiles (**Figure 3A**), with the sole exception of RCTD. This exception can be attributed to an inherent characteristic of RCTD's algorithm, which identifies optimal combination of single-cell profiles to fit each spatial spot [9]. RCTD excels at assigning cells to their correct spatial locations when provided with a complete and comprehensive reference dataset. However, under more realistic conditions, including reduced reference datasets, tumor heterogeneity, and practical constraints such as cost and tissue availability with an 8,000-cell reference, RCTD's performance declines markedly. This is primarily because it relies on the presence of a complete and detailed reference landscape, which is often unavailable in real-world applications, particularly in tumor tissues. Additionally, it is crucial to emphasize that CITEgeist offers a distinct advantage: the ability to infer cell type and spot-specific gene expression layers — a feature RCTD does not currently support.

### Using a user-defined cell-protein marker library

While our model requires user expertise and design, we contend it is no different from when the single-cell reference requires user intervention to label the cell type classes being transferred onto spatial data. Further work is currently being conducted on using additional antibody signals, such as CD74 for M1-polarized macrophages, to further deconvolute down to sub-profiles of different cell types. Antibody-based approaches are easily extensible compared to single-cell reference-based approaches.

Furthermore, this provides key advantages. As long as the classes are defined to be broad enough to characterize most courses in the environment, there is little risk of a false positive result. The only major weakness of our approach is that it requires the antibody information to be collected at the time of sample submission. Thus, a large body of spatial sequencing data out there may not be usable under our new method. However, we contend that all computational techniques to analyze cancer should be considered in the framework of starting from the tissue. Thus, we argue that our method is optimal for researchers if it is less expensive and yields more accurate results from the beginning of a study, clinical trial, or analysis.

### Validity of the simulation framework

Given the lack of paired spatial gene expression and antibody capture datasets with ground-truth spot annotations—and the structured nature of tissues used for benchmarking deconvolution tools—we simulated spatial CITE-seq data to evaluate method accuracy in heterogeneous tumor tissues. Using the scCube framework and a gold-standard breast tumor atlas, we generated unbiased simulated datasets tailored for benchmarking.

Our simulation effectively captures biological variability through probabilistic distributions. Average protein expression for each cell type per spot is first sampled from a uniform distribution (20–50) to reflect baseline heterogeneity and scaled based on cell type proportions, followed by sampling from a negative binomial distribution (dispersion = 0.5) to account for overdispersion typical of gene and protein expression data. A 5% dropout rate introduces technical noise, replicating sparsity and measurement limitations in real datasets. These elements ensure the simulated data reflects spatial transcriptomics' variability and transcriptomics'stics.

While our framework does not model partial cells within spots—each containing an average of 5 cells without overlap—we addressed this limitation by adding low-level background expression for absent cell types and increasing antibody capture variability. Future iterations could model partial cell overlaps to enhance biological realism.

Although antibodies in our simulation were modeled to be highly specific to cell types, the accurate reconstruction of gene expression profiles demonstrates that high specificity and accuracy can still be achieved with one or two highly specific antibodies per cell type. This result highlights the natural correlation between gene-to-cell type and marker-protein-to-cell type expression. However, a potential limitation is the need for users to identify antibodies with reasonably high specificity for their target cell types. Provided this condition is met, CITEgeist can accurately deconvolute complex tissues, offering robust performance even with minimal antibody input.

### Aaccuracy of non-simulated patient sample expression layers from CITEgeist

The results presented in **Figure 4A** align with the expected demographic composition of the tissue, accurately reflecting the proportional distribution of cell types. **Figure 4B** further supports the robustness of our approach by demonstrating a strong correspondence between the global gene expression profile and multiple cell types across distinct organisms. Specifically, the observed high gene presence (60–70%) within the predefined gene sets indicates a reliable and biologically consistent assignment of cell type-specific gene markers unsupervised.

In **Figure 4E**, our results provide further validation, particularly in the context of "sender" signals, which correspond to the secretion of related ligands. The agreement between our inferred ligand activity and Human Protein Atlas (HPA) data underscores the biological relevance of our approach. COMMOT generates relative expression values, while HPA utilizes normalized transcripts per million (nTPM), so we employ Spearman correlation to assess concordance between the datasets. While moderate, the resulting correlation coefficient of 0.38 is consistent with prior observations that RNA-derived signaling estimates exhibit variable but significant correlation with proteomic data. This finding aligns with established reports indicating a broad concordance range of 30–85% between RNA and protein-level measurements [31]. Despite the absence of direct ground truth validation, these results provide confidence in the biological accuracy of the CITEgeist inference.

### Validation of a Phagocytic Signature in Post-Treatment Surgical Samples

This analysis highlights two key findings. First, the integration of macrophage-enriched spatial transcriptomic spots using Harmony demonstrates effective mixing across patient samples while preserving distinct clusters with shared transcriptional features (**Figure 5B**). Notably, the observed mixing of biopsy- and surgery-derived macrophages indicates that these populations are not entirely distinct but retain common transcriptional signatures, suggesting conserved functional states after deconvolution.

Second, differential expression analysis using PyDESeq2 (**Figure 5C**) robustly identifies genes associated with macrophage-mediated phagocytosis and fibrinolytic activity, such as MMP13 [32]. Moreover, pathway-level validation is provided through correlation with the interstitial tumor-associated macrophage signature defined by Gonzalez et al. [12]. The Pearson correlation coefficient of 0.353 is significant given the adaptation of the signature from a murine model to human homologs, further supported by a p-value of $4.95 \times 10^{-3}$. Given that phagocytic activity among macrophages is highly heterogeneous and context-dependent [33], the strong correlation observed with this specific signature—derived from a near-identical biological context—validates both our hypothesis and the capacity of CITEgeist to deconvolute functionally relevant cell states accurately.

### Increased Signaling in the D538G Mutant Population

Lastly, we validate the accuracy and consistency of D538G cancer cell deconvolution through three independent methods, demonstrating agreement at both the gene and pathway levels, consistent with findings reported by Li et al. [27].

A long-standing question in the study of ESR1 mutations in endocrine therapy-resistant breast cancer has been how mutations with relatively low allelic fractions—22% at the DNA level in this case (**Figure S5A**)—can drive significant phenotypic changes [34]. Our analysis identifies the most pronounced signaling differences between D538G-mutant and wild-type cancer cells in the upregulation of midkine (MDK) and pleiotrophin (PTN) (**Figure 6F**), two structurally related growth factors known to be induced by estrogen [35, 36]. Notably, MDK has been implicated in fibroblast-driven tumor progression in ovarian metastases of gastric cancer [37], as well as in the upregulation of age-related changes in $ER^+$ mammary tumors [38] and tumor microenvironment remodeling during aging-associated tumorigenesis [39].

These findings, enabled by CITEgeist, suggest that the constitutive, ligand-independent activation of ESR1 by the D538G mutation drives the upregulation of known estradiol-regulated growth factors, leading to known localized inflammatory signaling and tumor-promoting changes within the microenvironment. This mechanism, uncovered by CITEgeist, may explain how a mutation with a low allelic fraction exerts a broad influence over endocrine therapy resistance, contributing to the aggressive phenotype observed in these tumors.

### Study Limitations

While our method demonstrates comparable or superior performance and extended functionality relative to the current state of the art, several limitations warrant consideration in future work. Three specific challenges are immediately evident. Firstly, our method cannot distinguish between cells that share surface markers but exhibit distinct biological gene expression programs, such as normal and malignant epithelial cells. However, this limitation can be mitigated by incorporating clustering approaches to differentiate gene expression profiles. Secondly, antibody capture information is inherently limited and may not always serve as direct corollaries for cell types; for example, in our real-world datasets, $\alpha$-SMA—identified as the best available fibroblast marker in the Visium panel—is also expressed by cancer cells undergoing epithelial-mesenchymal transition (EMT) [40]. Refinement of multi-marker signatures will be necessary to enhance specificity. Thirdly, our approach does not provide the single-cell resolution afforded by single-cell sequencing, as spatial transcriptomic data remain sparse in UMI counts and signal density [41]. However, this limitation ensures that our spatial data represent accurate biological signals rather than artifacts introduced by single-cell transposition.

Additionally, while we demonstrate the applicability of our method using clinical trial specimens, it is essential to acknowledge that these tissues represent a relatively small cohort with short-term follow-up. As such, our biological findings should be interpreted cautiously, and future validation in more extensive, longitudinally tracked patient cohorts will be necessary to assess our approach's generalizability and clinical utility.

Based on the fundamental considerations we present regarding tissue availability and cost and the reliability of CITEgeist in three separate applications, we believe our method is a valuable tool in advancing the utility of spatial transcriptomics in cancer research.

## DATA AVAILABILITY

The Wu et al. breast cancer atlas can be accessed according to the instructions in their original publication; the processed scRNA-seq data utilized in this study can be downloaded through GEO (accession number: GSE176078, accessed 10/25/24). The down-sampled reference and simulated test datasets will be available via FigShare at publication (doi: 10.6084/m9.figshare.28385675).

### Visium Patient Datasets

Visium datasets used in this paper will be published at the time of publication and are currently embargoed. Reviewers can access the data via the link and private access token provided to the journal in the cover letter at time of submission.

## CODE AVAILABILITY

The CITEgeist package, including the tool itself, benchmarking scripts, and code for simulating test datasets, is available on GitHub at https://github.com/leeoesterreich/CITEgeist. The repository provides the complete implementation for generating the spatially-resolved CITE-seq data simulations and evaluating deconvolution accuracy and gene expression (GEX) profile inference performance. It includes workflows for benchmarking CITEgeist against state-of-the-art methods and instructions for replicating the analysis and customizing simulations for new datasets. Comprehensive documentation and example scripts are included to ensure usability and reproducibility. A persistent version of CITEgeist will be made available via FigShare at the time of publication (doi: 10.6084/m9.figshare.28385675)

## ACKNOWLEDGMENTS

This research was partly supported by the University of Pittsburgh Center for Research Computing (RRID: SCR 022735) through the resources provided. We utilized the HTC cluster, supported by NIH award S10OD028483, and the H2P cluster, supported by NSF award OAC-2117681. Additionally, this work was funded by US NIH Grant R01HG010589, NIH Grant 5F30CA264963-03, and a Developmental Pilot Award from the UPMC Hillman Cancer Center, supported through NIH grant P30CA047904.

We want to thank Jingyang Qian for their gracious guidance in utilizing the scCube method to simulate and benchmark spatially resolved Visium-like datasets from a single-cell reference. We also thank Jian Chen from the Lee-Oesterreich Lab for her help with RNA sequencing prep and ddPCR.

We acknowledge the assistance of the physicians and clinical support staff at the University of Pittsburgh Medical Center health system for their contributions to the clinical trial—special thanks to Natera Inc. for their support with the clinical trial NCT05914792.

Histology sectioning was performed in the Pitt Biospecimen Core (RRID: SCR 025229), and the services and instruments used in this project were supported, in part, by the University of Pittsburgh and the Office of the Senior Vice Chancellor for Health Sciences. Work conducted at the UPMC Hillman Cancer Center Tissue and Research Pathology Services (TARPS) Shared Resource Facility was also supported, in part, by the University of Pittsburgh and the National Cancer Institute of the NIH under Award Number P30CA047904.

Lastly, 10x Visium library generation, Takara Smart Stranded Total RNA library generation, and Illumina sequencing were performed in the Health Sciences Sequencing Core (RRID: SCR 023116) at UPMC Children's Hospitalh, Rangos Research Center. These services and instruments were generously supported, in part, by the University of Pittsburgh, the Office of the Senior Vice Chancellor for Health Sciences, the Department of Pediatrics, the Institute for Precision Medicine, and the Richard K Mellon Foundation for Pediatric Research.

## AUTHOR CONTRIBUTIONS

A.C.C. conceived the project, designed the computational framework, wrote the code, conducted the bioinformatic analyses, and prepared the initial manuscript draft. B.T.S. performed the benchmarking, simulation studies, and statistical analysis and contributed to discussing the benchmarking and simulation strategies. N.C. facilitated the collection of clinical samples and coordinated the clinical trial. P.F.A. served as the clinical trial's principal investigator and performed breast surgery collecting the samples. S.O. oversaw the overall project management and shaped the project's goals. R.S. contributed to manuscript revisions and provided expertise in mathematical modeling and computational methodologies. A.V.L. guided the project design and objectives. All authors reviewed and edited the manuscript.

## COMPETING INTERESTS

All authors have no competing interests to declare.

## MATERIALS & CORRESPONDENCE

All correspondence should be addressed to Adrian V. Lee at leeav@upmc.edu.

## REFERENCES

[1] Cameron G. Williams et al. "An Introduction to Spatial Transcriptomics for Biomedical Research". In: *Genome Medicine* 14.1 (June 27, 2022), p. 68. ISSN: 1756-994X. DOI: 10.1186/s13073-022-01075-1. URL: https://doi.org/10.1186/s13073-022-01075-1 (visited on 11/22/2024).

[2] Vadim Kleshchevnikov, Andrei Shmatko, Elliot Dann, et al. "Cell2location maps fine-grained cell types in spatial transcriptomics". In: *Nature Biotechnology* 40 (2022), pp. 661–671. DOI: 10.1038/s41587-021-01139-4.

[3] Tim Stuart et al. "Comprehensive integration of single-cell data". In: *Cell* 177.7 (2019), 1888–1902.e21. DOI: 10.1016/j.cell.2019.04.031.

[4] *Single Cell Core Pricing — Pitt Center for Advanced Genomics*. URL: https://www.advancedgenomics.pitt.edu/pricing/single-cell-core-pricing (visited on 11/22/2024).

[5] Sunny Z. Wu, Ghamdan Al-Eryani, Daniel L. Roden, et al. "A single-cell and spatially resolved atlas of human breast cancers". In: *Nature Genetics* 53.10 (2021), pp. 1334–1347. DOI: 10.1038/s41588-021-00911-1.

[6] *Visium v1 Mouse Brain Dataset Introduction - Official 10x Genomics Support*. 10x Genomics. URL: https://www.10xgenomics.com/support/software/loupe-browser/latest/tutorials/assay-analysis/visium-tutorial-introduction (visited on 02/02/2025).

[7] Xiaomeng Wan et al. "Integrating Spatial and Single-Cell Transcriptomics Data Using Deep Generative Models with SpatialScope". In: *Nature Communications* 14.1 (Nov. 29, 2023), p. 7848. ISSN: 2041-1723. DOI: 10.1038/s41467-023-43629-w. URL: https://www.nature.com/articles/s41467-023-43629-w (visited on 02/02/2025).

[8] T. Biancalani, G. Scalia, L. Buffoni, et al. "Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram". In: *Nature Methods* 18.11 (2021), pp. 1352–1362. DOI: 10.1038/s41592-021-01264-7.

[9] D.M. Cable, E. Murray, L.S. Zou, et al. "Robust decomposition of cell type mixtures in spatial transcriptomics". In: *Nature Biotechnology* 40.4 (2022), pp. 517–526. DOI: 10.1038/s41587-021-00830-w.

[10] *Using Seurat with Multimodal Data*. URL: https://satijalab.org/seurat/articles/multimodal_vignette#visualize-multiple-modalities-side-by-side (visited on 11/22/2024).

[11] Jingyang Qian et al. "Simulating multiple variability in spatially resolved transcriptomics with scCube". In: *Nature Communications* 15 (2024), p. 5021. DOI: 10.1038/s41467-024-49445-0.

[12] Michael A. Gonzalez et al. "Phagocytosis Increases an Oxidative Metabolic and Immune Suppressive Signature in Tumor Macrophages". In: *The Journal of Experimental Medicine* 220.6 (Mar. 30, 2023), e20221472. ISSN: 0022-1007. DOI: 10.1084/jem.20221472. pmid: 36995340. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10067971/ (visited on 02/03/2025).

[13] Zheqi Li et al. "EstroGene2.0: A multi-omic database of response to estrogens, ER-modulators, and resistance to endocrine therapies in breast cancer". en. In: *bioRxiv* (July 2024), p. 2024.06.28.601163. DOI: 10.1101/2024.06.28.601163. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11244912/ (visited on 02/06/2025).

[14] Zixuan Cang et al. "Screening Cell–Cell Communication in Spatial Transcriptomics via Collective Optimal Transport". In: *Nature Methods* 20.2 (Feb. 2023), pp. 218–228. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01728-4. URL: https://www.nature.com/articles/s41592-022-01728-4 (visited on 01/30/2025).

[15] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. "GSEApy: A Comprehensive Package for Performing Gene Set Enrichment Analysis in Python". In: *Bioinformatics* 39.1 (Jan. 1, 2023), btac757. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btac757. URL: https://doi.org/10.1093/bioinformatics/btac757 (visited on 02/02/2025).

[16] Boris Muzellec et al. "PyDESeq2: A Python Package for Bulk RNA-seq Differential Expression Analysis". In: *Bioinformatics* 39.9 (Sept. 5, 2023), btad547. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btad547. pmid: 37669147. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10502239/ (visited on 02/02/2025).

[17] Priscilla McAuliffe. *Longitudinal ctDNA Monitoring in Older Women With ER+ Breast Cancer Who Forego Upfront Surgery in Favor of Primary Endocrine Therapy*. Clinical trial registration NCT05914792. clinicaltrials.gov, July 7, 2024. URL: https://clinicaltrials.gov/study/NCT05914792 (visited on 02/02/2025).

[18] Neil Carleton et al. "Longitudinal Monitoring of ctDNA for Disease Surveillance in Older Women with ER+ Breast Cancer on Primary Endocrine Therapy to Facilitate Surgical De-Escalation: A Prospective, Pragmatic, Hybrid-Decentralized Trial with Correlative Analyses". In: *San Antonio Breast Cancer Symposium*. Abstract Number: SESS-1343, Presented at the San Antonio Breast Cancer Symposium 2024. 2024.

[19] Bin Li et al. "Benchmarking Spatial and Single-Cell Transcriptomics Integration Methods for Transcript Distribution Prediction and Cell Type Deconvolution". In: *Nature Methods* 19.6 (June 2022), pp. 662–670. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01480-9. URL: https://www.nature.com/articles/s41592-022-01480-9 (visited on 02/02/2025).

[20] Haoyang Li et al. "A Comprehensive Benchmarking with Practical Guidelines for Cellular Deconvolution of Spatial Transcriptomics". In: *Nature Communications* 14.1 (Mar. 21, 2023), p. 1548. ISSN: 2041-1723. DOI: 10.1038/s41467-023-37168-7. URL: https://www.nature.com/articles/s41467-023-37168-7 (visited on 02/02/2025).

[21] Shimrit Mayer et al. "The Tumor Microenvironment Shows a Hierarchy of Cell-Cell Interactions Dominated by Fibroblasts". In: *Nature Communications* 14.1 (Sept. 19, 2023), p. 5810. ISSN: 2041-1723. DOI: 10.1038/s41467-023-41518-w. URL: https://www.nature.com/articles/s41467-023-41518-w (visited on 02/02/2025).

[22] Jean Fan, Kamil Slowikowski, and Fan Zhang. "Single-Cell Transcriptomics in Cancer: Computational Challenges and Opportunities". In: *Experimental & Molecular Medicine* 52.9 (Sept. 2020), pp. 1452–1465. ISSN: 2092-6413. DOI: 10.1038/s12276-020-0422-0. URL: https://www.nature.com/articles/s12276-020-0422-0 (visited on 02/02/2025).

[23] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis". In: *Genome Biology* 19.1 (Feb. 6, 2018), p. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0. URL: https://doi.org/10.1186/s13059-017-1382-0 (visited on 02/02/2025).

[24] *The Human Protein Atlas*. URL: https://www.proteinatlas.org/ (visited on 02/05/2025).

[25] Ilya Korsunsky et al. "Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony". In: *Nature Methods* 16.12 (Dec. 2019), pp. 1289–1296. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0619-0. URL: https://www.nature.com/articles/s41592-019-0619-0 (visited on 02/02/2025).

[26] Isaure Vanmeerbeek et al. "The Interface of Tumour-Associated Macrophages with Dying Cancer Cells in Immuno-Oncology". In: *Cells* 11.23 (Dec. 2, 2022), p. 3890. ISSN: 2073-4409. DOI: 10.3390/cells11233890. pmid: 36497148. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9741298/ (visited on 02/02/2025).

[27] Zheqi Li et al. "ESR1 Mutant Breast Cancers Show Elevated Basal Cytokeratins and Immune Activation". In: *Nature Communications* 13.1 (Apr. 19, 2022), p. 2011. ISSN: 2041-1723. DOI: 10.1038/s41467-022-29498-9. URL: https://www.nature.com/articles/s41467-022-29498-9 (visited on 01/29/2025).

[28] Wenguang Shao et al. "Comparative Analysis of mRNA and Protein Degradation in Prostate Tissues Indicates High Stability of Proteins". In: *Nature Communications* 10.1 (June 7, 2019), p. 2524. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10513-5. URL: https://www.nature.com/articles/s41467-019-10513-5 (visited on 02/02/2025).

[29] Taopeng Wang et al. "snPATHO-seq, a Versatile FFPE Single-Nucleus RNA Sequencing Method to Unlock Pathology Archives". In: *Communications Biology* 7.1 (Oct. 16, 2024), pp. 1–12. ISSN: 2399-3642. DOI: 10.1038/s42003-024-07043-2. URL: https://www.nature.com/articles/s42003-024-07043-2 (visited on 11/22/2024).

[30] Chengheng Liao et al. "Tumor Hypoxia: From Basic Knowledge to Therapeutic Implications". In: *Seminars in cancer biology* 88 (Jan. 2, 2023), p. 172. DOI: 10.1016/j.semcancer.2022.12.011. pmid: 36603793. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9929926/ (visited on 11/22/2024).

[31] Raquel de Sousa Abreu et al. “Global Signatures of Protein and mRNA Expression Levels”. In: *Molecular bioSystems* 5.12 (Dec. 2009), pp. 1512–1526. ISSN: 1742-206X. DOI: 10.1039/b908315d. pmid: 20023718. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4089977/ (visited on 02/06/2025).

[32] Jonathan A. Fallowfield et al. “Scar-Associated Macrophages Are a Major Source of Hepatic Matrix Metalloproteinase-13 and Facilitate the Resolution of Murine Hepatic Fibrosis”. In: *Journal of Immunology (Baltimore, Md.: 1950)* 178.8 (Apr. 15, 2007), pp. 5288–5295. ISSN: 0022-1767. DOI: 10.4049/jimmunol.178.8.5288. pmid: 17404313.

[33] Christine Vogel and Edward M. Marcotte. “Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses”. In: *Nature Reviews Genetics* 13.4 (Apr. 2012), pp. 227–232. ISSN: 1471-0064. DOI: 10.1038/nrg3185. URL: https://www.nature.com/articles/nrg3185 (visited on 02/06/2025).

[34] Sung Gwe Ahn et al. “Primary Endocrine Resistance of ER+ Breast Cancer with ESR1 Mutations Interrogated by Droplet Digital PCR”. In: *NPJ Breast Cancer* 8 (May 2, 2022), p. 58. ISSN: 2374-4677. DOI: 10.1038/s41523-022-00424-y. pmid: 35501333. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9061813/ (visited on 02/07/2025).

[35] Gang Xi et al. “Estrogen Stimulation of Pleiotrophin Enhances Osteoblast Differentiation and Maintains Bone Mass in IGFBP-2 Null Mice”. In: *Endocrinology* 161.4 (Mar. 13, 2020), bqz007. ISSN: 0013-7227. DOI: 10.1210/endocr/bqz007. pmid: 32168373. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7069688/ (visited on 02/06/2025).

[36] Minakshi Saikia et al. “Role of Midkine in Cancer Drug Resistance: Regulators of Its Expression and Its Molecular Targeting”. In: *International Journal of Molecular Sciences* 24.10 (10 Jan. 2023), p. 8739. ISSN: 1422-0067. DOI: 10.3390/ijms24108739. URL: https://www.mdpi.com/1422-0067/24/10/8739 (visited on 02/06/2025).

[37] Simeng Hu et al. “The Estrogen Response in Fibroblasts Promotes Ovarian Metastases of Gastric Cancer”. In: *Nature Communications* 15.1 (1 Sept. 30, 2024), pp. 1–19. ISSN: 2041-1723. DOI: 10.1038/s41467-024-52615-9. URL: https://www.nature.com/articles/s41467-024-52615-9 (visited on 02/06/2025).

[38] Priscilla A. Furth et al. “Overexpression of Estrogen Receptor  in Mammary Glands of Aging Mice Is Associated with a Proliferative Risk Signature and Generation of Estrogen Receptor –Positive Mammary Adenocarcinomas”. In: *The American Journal of Pathology* 193.1 (Jan. 2023), pp. 103–120. ISSN: 0002-9440. DOI: 10.1016/j.ajpath.2022.09.008. pmid: 36464513. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9768686/ (visited on 02/06/2025).

[39] Pengze Yan et al. “Midkine as a Driver of Age-Related Changes and Increase in Mammary Tumorigenesis”. In: *Cancer Cell* 42.11 (Nov. 11, 2024), 1936–1954.e9. ISSN: 1878-3686. DOI: 10.1016/j.ccell.2024.09.002. pmid: 39366375.

[40] Dimitris Anastassiou et al. “Human Cancer Cells Express Slug-based Epithelial-Mesenchymal Transition Gene Expression Signature Obtained in Vivo”. In: *BMC Cancer* 11.1 (1 Dec. 2011), pp. 1–9. ISSN: 1471-2407. DOI: 10.1186/1471-2407-11-529. URL: https://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-11-529 (visited on 02/07/2025).

[41] Yue You et al. “Systematic Comparison of Sequencing-Based Spatial Transcriptomic Methods”. In: *Nature Methods* 21.9 (Sept. 2024), pp. 1743–1754. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02325-3. URL: https://www.nature.com/articles/s41592-024-02325-3 (visited on 02/07/2025).

## SUPPLEMENTAL FIGURES

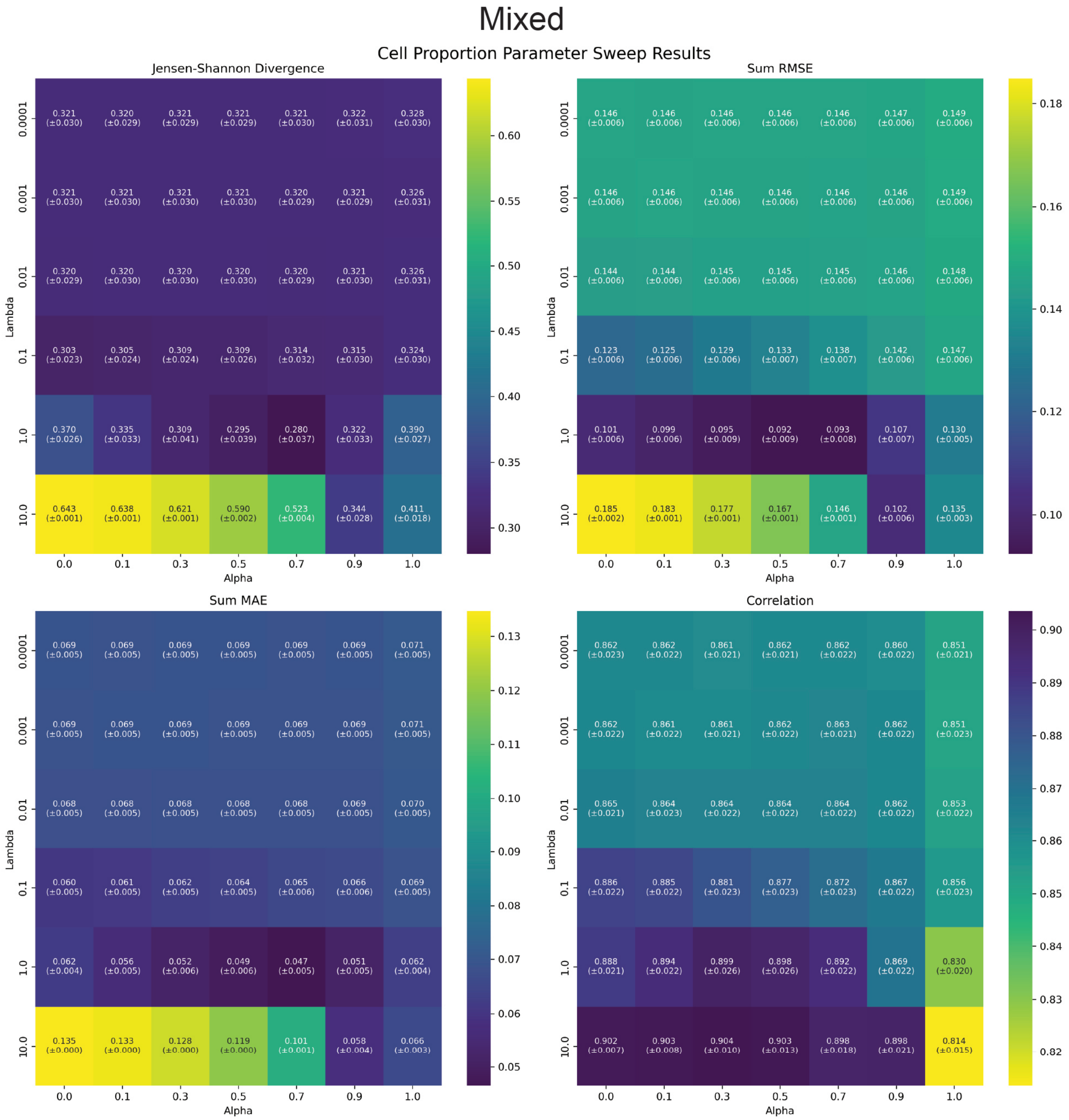


**Figure S1.** Grid search metrics results at varying alpha and lambda values for cell proportion finetuning Neighborhood model in the Mixed data set.

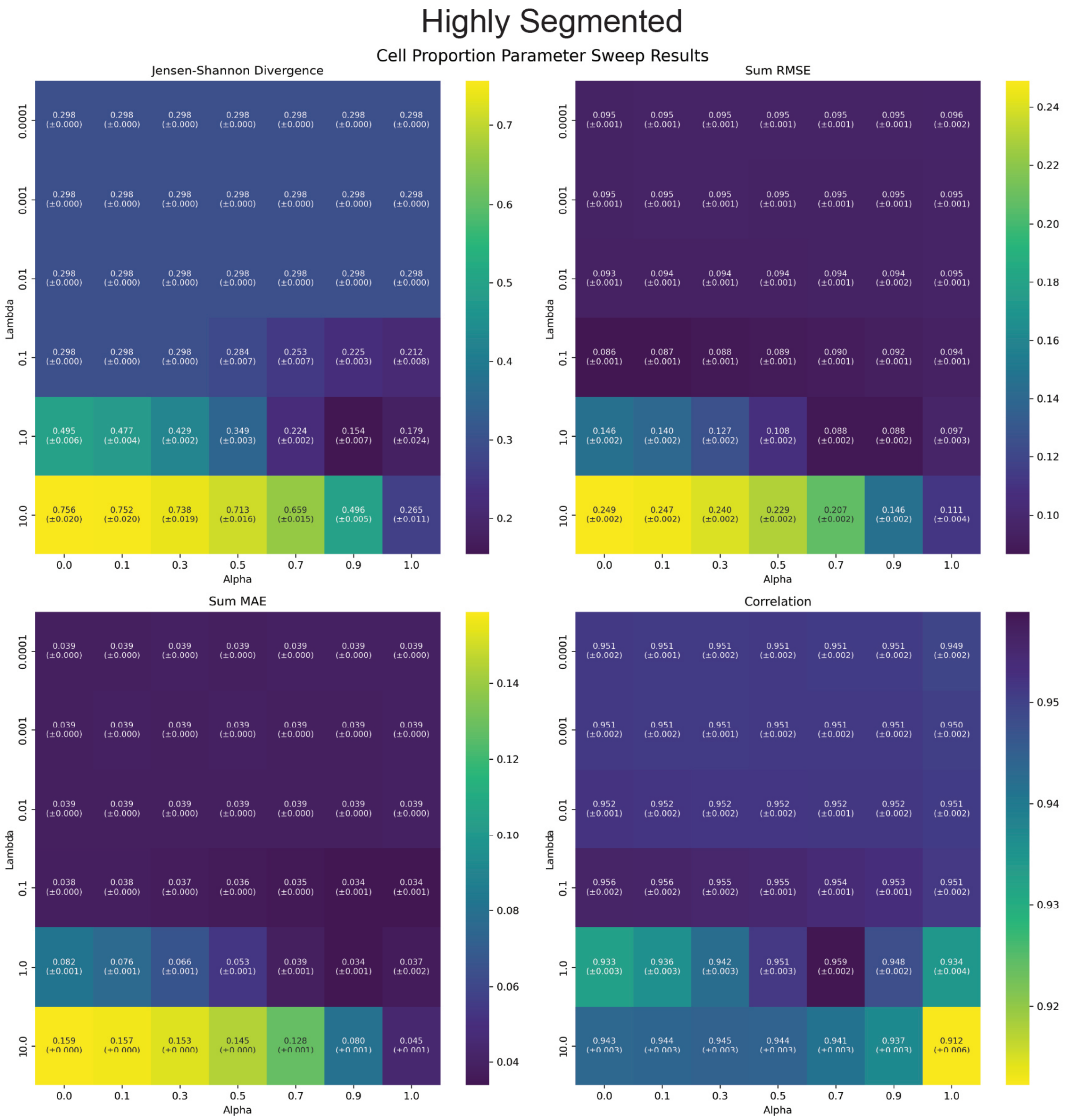
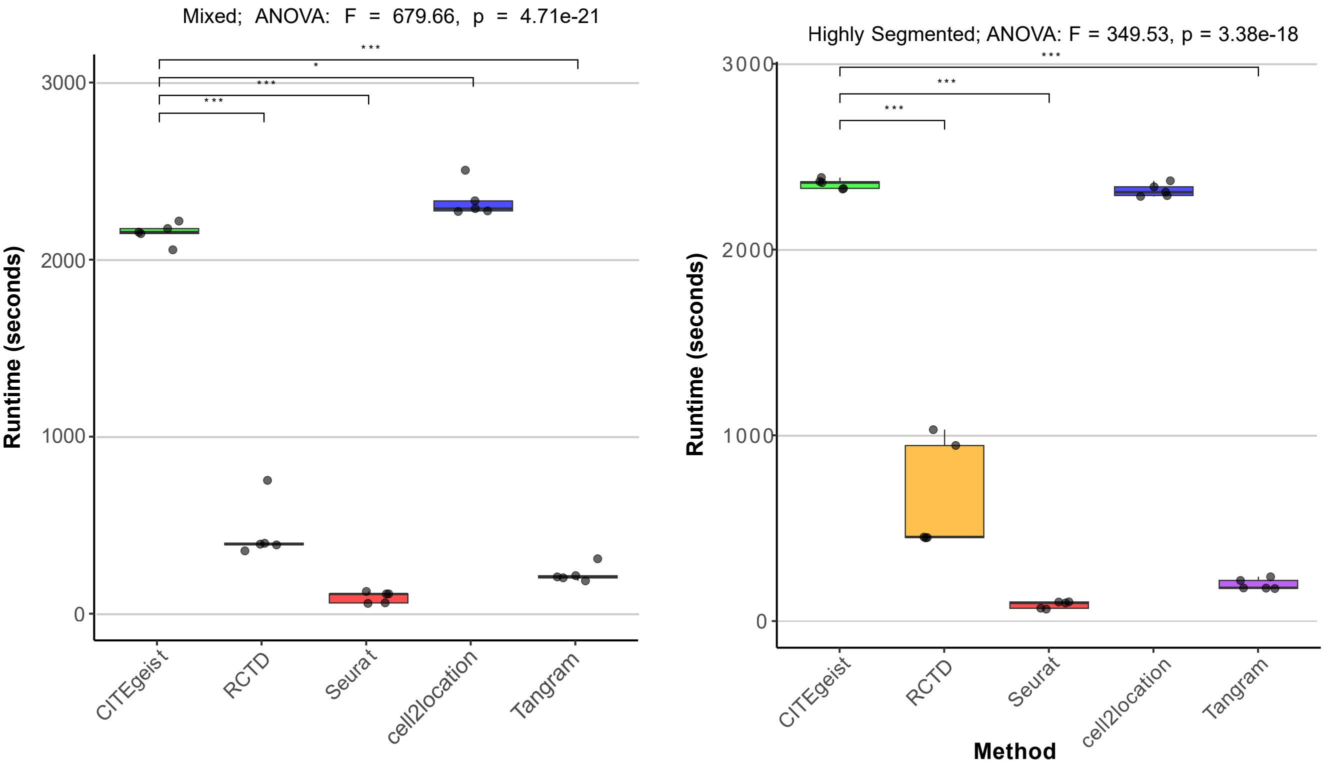


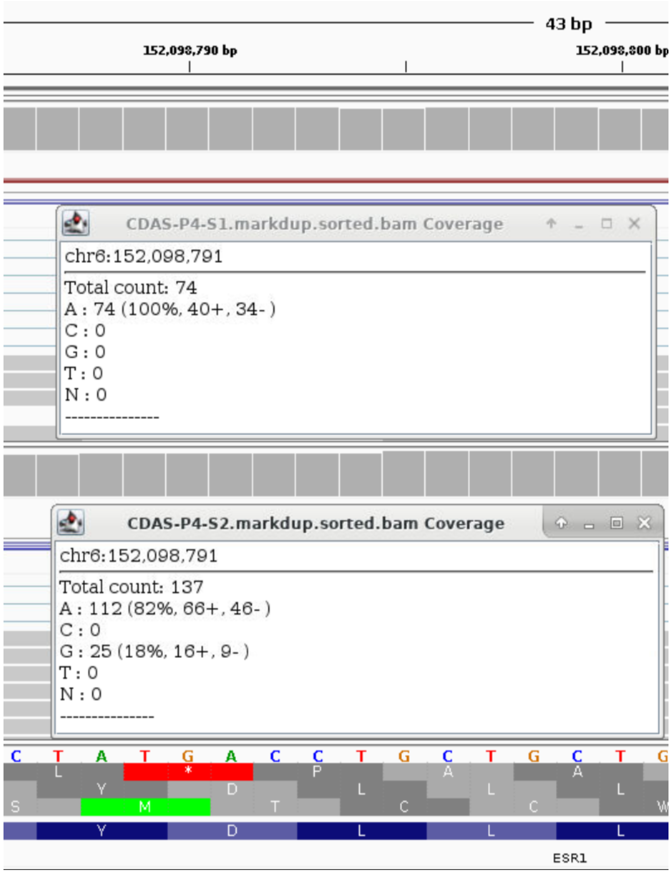
**Figure S2.** Grid search metrics results at varying alpha and lambda values for cell proportion fine-tuning Neighborhood model in the Highly Segmented data set.



**(a) Mixed replicates**

**(b) Highly segmented replicates**

**Figure S3.** Comparison of deconvolution method runtimes. Panels display execution times for benchmarked deconvolution methods under different segmentation conditions. Boxplots show computation times across deconvolution methods (seconds). Statistical significance was assessed using one-way Analysis of Variance (ANOVA), followed by post-hoc Tukey's Honest Significant Difference (HSD) test, with $p < 0.001$ denoted by ***. For clarity, only pairwise comparisons involving CITEgeist are shown.



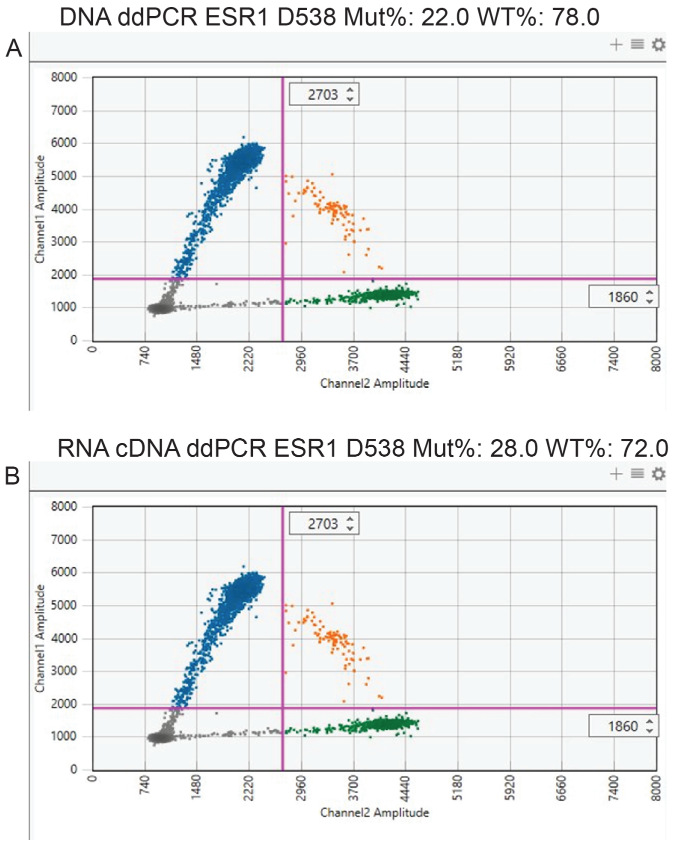
CDAS-P4-S1.markdup.sorted.bam Coverage

```
chr6:152,098,791
Total count: 74
A : 74 (100%, 40+, 34- )
C : 0
G : 0
T : 0
N : 0
---------------
```

CDAS-P4-S2.markdup.sorted.bam Coverage

```
chr6:152,098,791
Total count: 137
A : 112 (82%, 66+, 46- )
C : 0
G : 25 (18%, 16+, 9- )
T : 0
N : 0
---------------
```

**Figure S4.** Visualization of aligned RNAseq reads in IGV showing the ESR1 D538G mutation in the surgical sample (CDAS-P4-S2) but not the biopsy (CDAS-P4-S1)


DNA ddPCR ESR1 D538 Mut%: 22.0 WT%: 78.0

RNA cDNA ddPCR ESR1 D538 Mut%: 28.0 WT%: 72.0

**Figure S5.** ddPCR confirmation of D538G mutation in Sample CDAS-P4-S2 **(A)** DNA, **(B) RNA**