



THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

COVER SHEET

Understanding the Immune Tumour Microenvironment by Integrating Single-Cell and Spatial Transcriptomics

Click or tap here to enter text.

Student Exam Number:

B242415

Submitted in partial fulfilment of the requirements for the
Degree of Master of Science in

Bioinformatics

of the University of Edinburgh 2023-2024

Name(s) of Supervisor(s):

Dr. Florent Petitprez

Word Count*

8207

*The word count should not include figure and table legends, text within tables, appendices or references.

1. Acknowledgements

I would like to thank my supervisor Dr. Florent Petitprez for his guidance, help and expertise throughout the entire project.

I would also like to thank my partner, who has been the most supportive and loving partner I could ever ask for.

2. Table of Contents

1. Acknowledgements.....	2
2. Table of Contents.....	3
3. Abstract	5
4. Introduction	6
4.1 <i>The Importance of the Spatial Architecture in the TME</i>	6
4.2 <i>Spatial Transcriptomics - Characterizing the Spatial Architecture of the TME</i>	8
4.3 <i>Spatial Deconvolution</i>	10
4.4 <i>Spatial Deconvolution in the TME</i>	11
4.5 <i>Aims</i>	11
5. Methodology.....	12
5.1 <i>Single Cell Reference Dataset</i>	13
5.2 <i>Modification of synthspot</i>	14
5.3 <i>Generating Synthetic Spot Data</i>	15
5.4 <i>Assigning Spatial Coordinates to Synthetic Spot Data</i>	16
5.5 <i>Spatial Deconvolution</i>	17
5.6 <i>Performance Statistics</i>	18
5.7 <i>benchdeconv</i>	19
5.8 <i>Compute Specifications</i>	20
6. Research Data Management.....	21
7. Results	21
7.1 <i>Synthetic ST Datasets</i>	21
7.2 <i>Performances Between Methods</i>	22
7.3 <i>Performances Between Datasets</i>	25
7.4 <i>Impact of Single Cell RNA Reference Dataset Size</i>	27
7.5 <i>Cell Type Specific Performance Biases</i>	28
7.6 <i>Spatial Performance Biases</i>	29
7.7 <i>Performances at Different Cell Densities of Immune Cell Types</i>	33
7.8 <i>Runtimes</i>	34
8. Discussion	35
8.1 <i>Validity of Cell Type Annotations in Source scRNA-seq Datasets</i>	35
8.2 <i>Distribution of the Number of Cells per Spot in the Synthetic Data</i>	36
8.3 <i>Validity of Synthetic Spatial Datasets</i>	37
8.4 <i>Heterogeneity of Cells in a Cell Type Complicates Deconvolution</i>	37
8.5 <i>Deteriorating SPOTlight Performance as scRNA-seq Size Increases</i>	37

8.6 Incomplete scRNA-seq Reference Data.....	38
8.7 More Tools and Datasets	39
9. Conclusion.....	40
10. References.....	42

3. Abstract

The spatial arrangement of cells within the tumour microenvironment is crucial for tumour development, subtyping, patient prognosis and tumour response to treatment. Characterizing the spatial architecture of the tumour microenvironment is primarily conducted through array based spatial transcriptomics. However, current widely available array based spatial transcriptomics do not reach single cell resolution. Each spot contains a mixture of gene expression from the multiple cells that reside in them. Spatial deconvolution was therefore developed to predict cell type fractions within the spatial transcriptomic data. Here we evaluate the performance of some of these spatial deconvolution methods within the context of the tumour microenvironment. We find that RCTD outperforms CARD and SPOTlight in deconvolution performance in almost every metric. We also find that though performance generally remains constant between different scRNA-seq datasets, the cell type specific performances vary.

4. Introduction

In 2022 alone, 20 million new incidents and 9.7 million deaths from cancer have been estimated to have occurred globally. These cancer statistics, as reported by the International Agency for Research on Cancer (IARC) in GLOBOCAN (Bray et al., 2024), make it clear the importance of investing further into solutions for research, prevention, detection, and treatment of cancers. It is crucial we continue our efforts to combat this disease, particularly as 35 million new cases are estimated to occur in 2050.

A key characteristic of cancers that is being studied is the heterogeneity of tumours. Not only does the tumour consist of the tumour cells, but also a diverse set of surrounding infiltrating and resident host cells, secreted factors, and extracellular matrix. These elements make up the tumour microenvironment (TME) (Anderson and Simon, 2020), which the interplay between elements of the TME greatly influences the tumour's development (de Visser and Joyce, 2023) and response to treatment (Hirata and Sahai, 2017, Petitprez et al., 2020b), therefore patient prognosis and clinical outcome (Baghban et al., 2020).

The complex signalling networks between the tumour cells and non-malignant cells allows the hijacking of the host cells to promote tumorigenesis and metastasis in all phases of development (Frisch et al., 2019). Characterization of the TME is therefore key in further treatment development. One key development in spatial technologies, 2D spatial transcriptomics (ST) (Williams et al., 2022), has uncovered the significance of the spatial architecture in determining the phenotypes of the TME.

4.1 The Importance of the Spatial Architecture in the TME

A key element the tumour cells interact with are the immunological elements, known as the tumour immune microenvironment (TiME) (Binnewies et al., 2018). These elements include innate and adaptive immune cells, extracellular immune factors, and cell surface molecules (Gajewski et al., 2013, Greten and Grivannikov, 2019). However, conflicting findings indicate that cell type

compositions of the TiME do not necessarily predict prognosis as claimed by Zemek et al. and disputed by Casasent et al. and Schürch et al. where Casasent et al. and Schürch et al. find patients with similar compositions in triple-negative breast cancer (TNBC) with different, unexplained prognoses (Casasent et al., 2018, Schürch et al., 2020, Zemek et al., 2019). Cell type composition of the TME therefore are not strong predictors of prognoses.

Tumour cell proliferation, angiogenesis, and immune cell infiltration occur during tumour expansion. Gathered immune cell infiltration spatial patterns have revealed different pro- and anti- tumour responses that are reliant on the TME spatial architecture. Immune cells during infiltration can be distributed diffusely or aggregated to create new cellular neighbourhoods where there are different types of local cellular compositions. Some of these cellular neighbourhoods are associated with enhanced immune cell function (Schürch et al., 2020). However, some types of cellular neighbourhoods also impede immune cell function (Keren et al., 2018). Since these spatial characteristics are strong predictors of cancer progression and response to treatment, attention has been shifted to characterizing the spatial architecture of the TiME.

Spatial structures within the TME exhibit different prognoses and responses to therapy. B-cell rich tertiary lymphoid structures in soft tissue sarcoma have been shown to be the strongest prognostic factor, even between different CD8⁺ T-cell and cytotoxic content levels (Petitprez et al., 2020a). This suggests that the spatial architecture may be used to inform treatment decisions, acting as progression and therapeutic biomarkers, thereby illustrating its wider use in clinical applications.

Tumour invasion occurs when the tumour cells expand into nearby environment. A key cellular process that facilitates tumour invasion is the epithelial-mesenchymal transition (EMT) (Pastushenko and Blanpain, 2019). Epithelial cells acquire mesenchymal features after EMT transition, leading to the loss of adherence junctions and gaining fibroblastic invasive phenotypes. In cancer, this allows solid tumours to increase their invasiveness and metastatic activity (Ribatti et al., 2020). Several spatial characteristics about EMTs have been identified: It has been shown that fibroblasts surround epithelial neoplastic areas proportional to the increase in EMT state within the TME, EMT states have also been shown to

cluster within epithelial pockets (Malagoli Tagliazucchi et al., 2023), and partial EMT cells have been shown to localize at the leading edge of primary tumours, as well as being located at the occurrence of invasion. Further studying these spatial characteristics may uncover additional cancer mechanisms and progression markers.

It is clear that the spatial architecture of the TME is crucial for tumour development and patient prognoses. Therefore, developing spatial techniques, such as spatial transcriptomics, is key in developing new biomarkers, identifying new spatial subtypes of cancers, predicting tumour response to treatment for precision medicine, and investigating spatial disease mechanisms.

4.2 Spatial Transcriptomics - Characterizing the Spatial Architecture of the TME

Advancements in ST provided the much-needed techniques for gathering information about the spatial architecture of the TME. ST techniques are a suite of techniques that aim to gather the transcriptome of cells in a spatially resolved manner. ST can broadly be categorized into two groups: Imaging ST and sequencing ST (Williams et al., 2022). Imaging ST consists of in situ hybridization (ISH), and in situ sequencing (ISS) while sequencing ST consists of array-based methods and microdissection. Microdissection is now largely defunct due to low throughput, resolution, and high labour and cost (Duan et al., 2023).

ISH involves hybridizing fluorescent probes to target transcripts *in situ*, where afterwards fluorescent microscopy is performed. This technique can reach sub-cellular resolution, enabling the ability to determine the transcript location within subcellular compartments. However, ISH is biased and thus only a panel of pre-selected genes can be assayed. This limits the scope of discovery where information on targets not included in the panel of genes is not available. ISS also involves hybridizing fluorescent probes and fluorescent microscopy but using a padlock probe and rolling circle amplification (Tang et al., 2023). Though not biased, it inherits the same disadvantages as ISH as well as only achieving a limited read length (Williams et al., 2022). However, imaging ST has a lower dropout rate than sequencing ST (Li et al., 2023), making it more suitable for

applications such as assessing the expression of a gene of interest from certain cell types.

Array based methods involve the capture of transcripts *in situ* by placing a 2D cross-section of the sample onto a slide, where the transcripts are released from the cells by permeabilizing the tissue. These transcripts are captured by probes of close spatial proximity. These probes are barcoded with a spatial and UMI (unique molecular identifier) tag which are printed onto the array slide. The spatial tag allows the identification of the spatial location of transcript, which can be mapped back thus preserving its spatial information (Ståhl et al., 2016). The captured transcripts are then sequenced. This allows the capture of the entire transcriptome of the cells in an unbiased manner which allows the identification certain activated cellular pathways, cellular subtypes, and cell states of certain cells at specific spatial locations that are not possible with imaging ST.

However, current widely available array ST platforms are not yet at single cell resolution. Multiple cells are able to reside in each spot (Shi et al., 2023). For illustration, the 10X Genomics Visium platform has a slide area of 6.5x6.5mm. Each slide contains 4992, 55µm diameter spots which 10X Genomics claims to generally house 1-10 cells (Genomics, 2024). However, observations by Dr. Florent Petitprez's lab suggests ranges closer to 1-43, with a median of 9 cells per spot. Therefore, only the average of the gene expression from the mixture of cells is measured. To identify the cell type compositions per spot, otherwise 'deconvolve' the cell type fractions, would therefore be an important step within the ST workflow for TME research.

10X Genomics Visium HD was released on March 26, 2024, where the theoretical resolution is 2µm (Oliveira et al., 2024). In addition, the 2x2µm "pixels", since they are square and no longer circular leading to less gaps, are smaller than almost all cell types. Therefore, multiple pixels need to be pooled to get more signal. These computational methods have not been developed yet. Due to the cost and availability of the new platform, as well as underdeveloped tools for Visium HD data analysis with early support tools and computation methods still only being in preprint (Polański et al., 2024) and claims not validated, spatial deconvolution will remain an important factor in ST analysis for the near future.

4.3 Spatial Deconvolution

Spatial deconvolution is the prediction of cell type identities and fractions within each spot of spatially resolved transcriptomic datasets (Ma and Zhou, 2022). Spatial deconvolution methods integrate annotated single cell RNA sequencing (scRNA-seq) datasets for training deconvolution protocols. The capture of the transcriptome of single cells in scRNA-seq allows for the precise characterization of gene expression profiles of different cell types by the deconvolution methods (Xie et al., 2021).

Spatial deconvolution can broadly be classified into two categories: signature matrix and model-based methods, shown in figure 1. Signature matrix-based methods utilize signature matrices, transcriptional fingerprints of each cell type that are learned from the reference scRNA-seq dataset. This is then used to deconvolve the cell fractions, often by expressing the vector of gene expression levels in the ST data as a matrix product of the signature matrix and cell fractions. This assumes each spot is the convolution, or sum, of the gene expression levels of each cell in the spot. Model based methods estimate deconvolution parameters that directly infer the cell fractions, whether via statistical modelling or neural networks.

Figure 1: High level overview of spatial deconvolution workflows

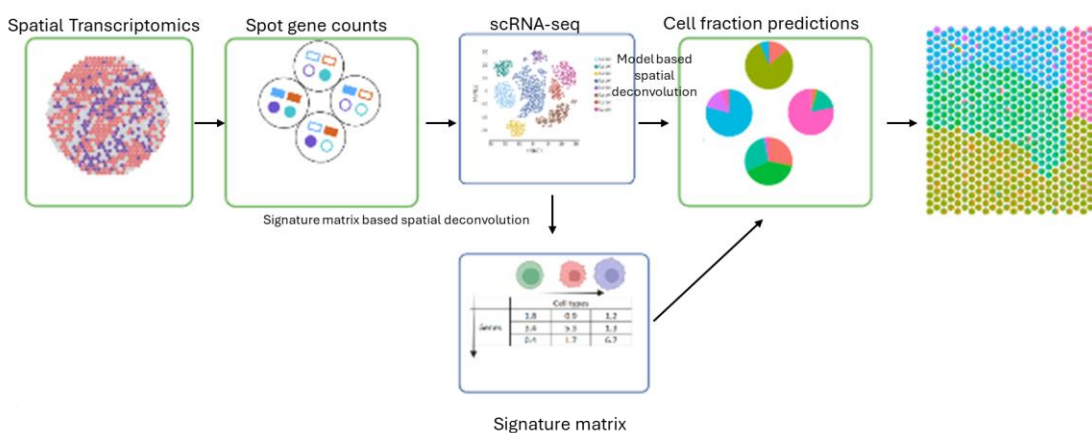


Figure 1 Legend: High level overview of spatial deconvolution workflows. Signature and model based deconvolution is pictured here. scRNA-seq is can be used directly for deconvolution (model based) or to create a signature matrix for deconvolution (signature matrix based). “Created with [BioRender.com](https://www.biorender.com/).”

4.4 Spatial Deconvolution in the TME

Though benchmarking studies have been conducted to evaluate spatial deconvolution performance (Sang-aram et al., 2024, Yan and Sun, 2022, Li et al., 2022), none have evaluated their performance within the context of the TME. It is unknown whether the deconvolution methods are able to adequately deconvolute TME datasets, particularly cell types that are known to be highly heterogeneous, such as cancer epithelial cells. The heterogeneity and variance of the transcriptome of these cells occurs not only between different tumours but also within tumours (Anderson and Simon, 2020)., further complicating spatial deconvolution. It is also unknown whether the type of cancer being deconvoluted affects performance. Though not specific to the TME, it is unknown whether the size of input reference scRNA-seq affects performance. Gathering this data would inform future ST experiments involving deconvolution, where extra capital can be saved by sequencing less cells in scRNA-seq while maintaining sufficient deconvolution performance.

4.5 Aims

Therefore, we aim to benchmark the performance of spatial deconvolution methods, with a focus on tumour microenvironment specific metrics. We will evaluate three deconvolution methods, RCTD (Cable et al., 2021)., CARD (Elmentaite et al., 2022). and SPOTlight (Elosua-Bayes et al., 2021) on three breast cancer subtype datasets, ER+, HER2+ and TNBC. Using the created benchmarking pipeline *benchdeconv*, we investigate the performance of each method:

1. with each breast cancer subtype
2. with varying sizes of reference scRNA-seq

3. in identifying TME cell types
4. at different densities of cell types per spot

5. Methodology

Figure 2 showcases the *benchdeconv* benchmarking pipeline. First, suitable and annotated tumour single cell transcriptomic datasets were identified and gathered. The samples in each dataset will be split into two groups. One set will be used to generate the synthetic spot data while the other will be used to train the deconvolution methods to deconvolute the synthetic spatial transcriptomic datasets. Each deconvolution method will then be evaluated, then compared with each other, with a report being automatically generated by the pipeline.

Figure 2: *benchdeconv* Benchmarking Pipeline Workflow Diagram

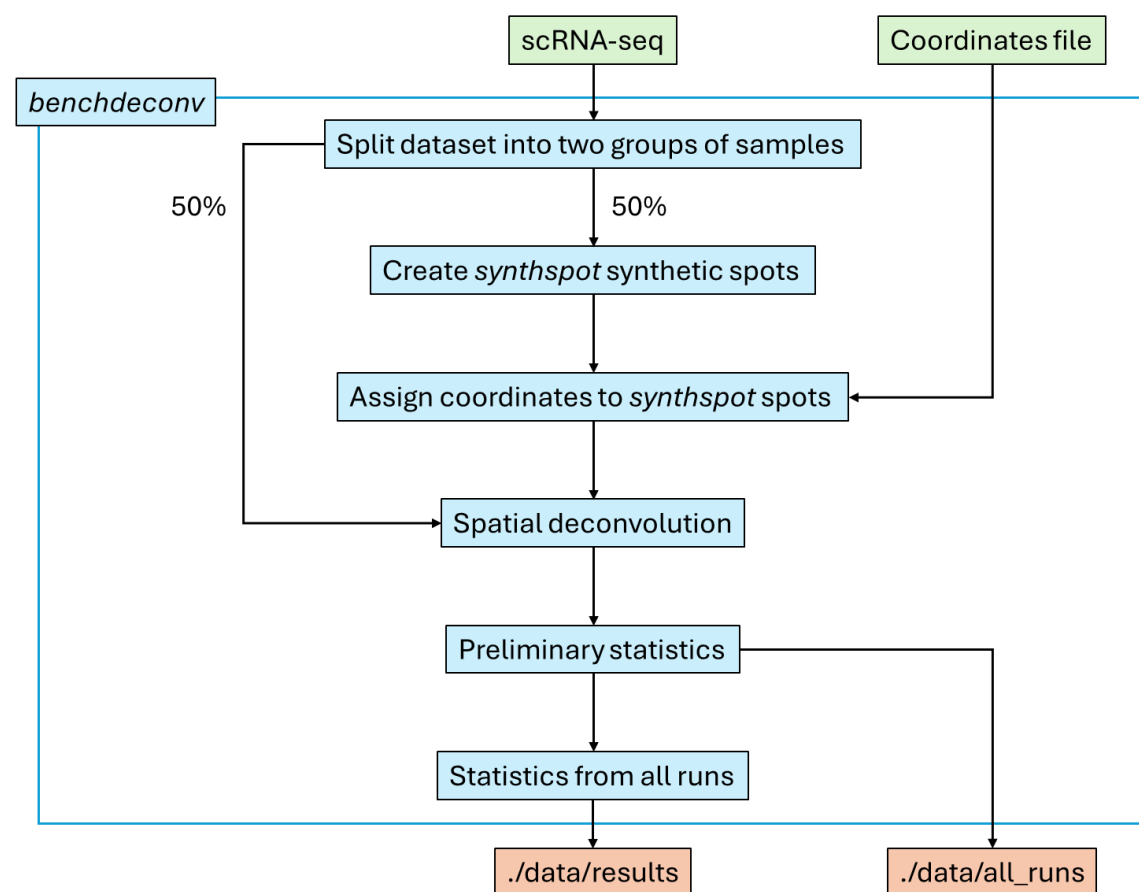


Figure 2 Legend: *benchdeconv* workflow diagram. Green are inputs, blue is *benchdeconv* and orange are output directories.

5.1 Single Cell Reference Dataset

For this benchmarking study, the deconvolution methods will be evaluated using annotated single cell transcriptomic datasets of tumours with a focus on breast cancers, which has been sourced from GEO series accession number GSE176078 (Wu et al., 2021). Here we focus on three subtypes of breast cancer: ER+, HER2+, and TNBC. The ER+ dataset contains 11 patients with a total of 38,241 cells. The HER2+ dataset contains 5 patients with a total of 19,311 cells. The TNBC dataset contains 10 patients with 42,512 cells. These datasets were gathered using single cell RNA-seq: 10X Genomics Chromium and Illumina NextSeq500 for library sequencing.

Figure 3: Table of cell types and their frequency in each dataset

Dataset	TNBC	ER+	HER2+
B-cells	956	292	304
CAFs	1343	1279	736
Cancer Epithelial	5457	5900	930
Endothelial	681	2629	493
Myeloid	3165	934	700
Normal Epithelial	490	1222	468
Plasmablasts	1324	311	108
PVL	613	1669	455
T-cells	7227	4885	5461

Figure 3 Legend: A table of all cell counts within each cell type of each dataset for the training scRNA-seq dataset.

For each subtype, the dataset was randomly split into two groups of samples. One subset of the dataset is used as a reference for generating the synthetic spot counts (the synth scRNA-seq dataset), while the other is used as a training dataset for the deconvolution methods (the training scRNA-seq dataset). Each subset will contain different samples from different tumours; different, non-overlapping groups of patients are in each subset. It is crucial separate the dataset to avoid data overlap in these two processes as it may lead to data leakage, where the deconvolution methods is more likely to interpret artifacts and prior information about the generation of the synthetic spots, which is not present in unseen datasets, which results in overfitting and overly optimistic performance predictions. Thus, to obtain more a more accurate evaluation

of the deconvolution method performances and their generalizability (Rosenblatt et al., 2024), the data inflows are separated.

To measure if there are any performance biases between each dataset source, we will evaluate the performance of the deconvolution methods with each subtype, ER+, HER2+, and TNBC. *benchdeconv* will run 10 times per dataset source to ensure reliable results. The datasets will be downsized to 5000 cells per cell type to reduce runtimes.

To measure the optimal size of the training dataset required for each deconvolution method, each method was run with different sizes of the training dataset where each cell type was downsized to have a specified number of cells in the training scRNA-seq dataset. Here we downsize to 500, 1000, 2500, 5000, and with all cells in the reference in each cell type. Performances were then compared between each run. Cell numbers in the synthspot synth scRNA-seq dataset were not downsized to best mimic real datasets. Each run with each downsized datasets will be repeated 10 times to ensure reliable results and deconvolution method stability. Due to project time constraints, only the TNBC dataset will be used as it contains the most cells in the reference dataset than ER+ and HER2+. Some cell types have less than 500 cells in the synth scRNA-seq. Therefore, if the subset dataset requires a higher number of cells than is in the reference for a cell type, it will use all the cells in the reference of that cell type.

5.2 Modification of synthspot

synthspot is an R package that generates synthetic Visiumlike spot data (Sangaram et al., 2024). It achieves this by assuming that each spot is a mixture of n cells and their gene expression. Therefore, it will sample n cells from the synth scRNA-seq data, where the probabilities of selection are defined a priori by the cell type frequencies in the synth scRNA-seq data. The summed expression data is then down sampled by a mean of 30,000 and standard deviation of 8,000, which are the default synthspot parameters, to better simulate the number of counts measured in real Visium data. Without down sampling, the dataset would report higher gene counts than can be captured by the 10X Visium platform.

However, some functionalities that are required for this benchmarking are not available. Therefore, *synthspot* has been adapted and modified (see Research Data Management for modified *synthspot* repository). First, *synthspot* was upgraded to take input scRNA-seq datasets using Seurat v5 objects while maintaining backwards compatibility with Seurat v3.

synthspot is able to generate spots that simulate tissue patterns into distinct 'regions'. These regions contain different prior probabilities for the cell types, thus mimicking tissues where tissue specific cell types tend to colocalize (Elmentaite et al., 2022). *synthspot* has 17 different 'dataset types' where the prior probabilities of cell types in each region can be somewhat controlled. Some of the dataset types are able to choose one cell type to make it the dominant, rare, or missing in the generated spots. However, *synthspot* chooses this cell type based on random sampling. Therefore, we modified *synthspot* so that the user is able to control which cell type is selected using the `select_celltype` parameter.

Finally, *synthspot* has been modified to produce spots with a specific cell density of a chosen celltype. This is done via the `min_cell_id_test` parameter to choose the density, and `select_celltype_min_id` parameter to select the cell type. After assigning the chosen number of cells of a cell type to a spot. The remaining cell type fractions will be downsized prior frequencies for the region the test spot inhabits so it is consistent with the surrounding spots, thus better mimicking real datasets.

5.3 Generating Synthetic Spot Data

The synth scRNA-seq dataset is then passed to the modified *synthspot* to generate the synthetic spot data. *synthspot* is able to generate spots with a user defined range of cell numbers per spot. Though 10X Genomics suggests that there are generally 1-10 cells per spot, Dr. Florent Petitprez's lab has observed upwards of 43 cells per spot in gathered cancer spatial transcriptomic datasets. Therefore, *n* has been set to 40 to better mimic real datasets.

Spots in 5 different regions were generated with *synthspot* to examine the deconvolution methods performance in identifying the edges of these distinct regions, particularly since some methods consider the composition of the surrounding spots in space when deconvoluting.

The “artificial_diverse_overlap” mode was chosen for *synthspot* since no region metadata was given from GSE176078. Regional cell type prior frequencies are therefore artificial, with each region containing overlapping cell type contributions. This mode was chosen as it generates distinct regions where performance in the edges of regions can be evaluated.

By using the newly added `min_cell_id_test` and `select_celltype_min_id` parameters we added onto *synthspot*, we generate immune cell types B-cells, plasmablasts, T-cells, and myeloid cells at densities 0.05, 0.10, 0.20, 0.50, and 0.80 within each test spot. This is to examine the performance of the deconvolution tools at different cell densities of different immune cell types. Due to the long runtimes when using the full training scRNA-seq dataset, it will be downsized to 5000. Every run of each cell type at each density will be repeated 5 times to ensure reliable results. Due to project time constraints, only the TNBC dataset will be used as it contains the most cells in the reference dataset than ER+ and HER2+.

5.4 Assigning Spatial Coordinates to Synthetic Spot Data

The spots generated by *synthspot* were then assigned spatial coordinates, with region locations shown in figure 4. The region templates were generated randomly and can be customized by the user. It is important to note that due to the staggered nature of the spots on the Visium plates, it is highly recommended to export the barcodes from a tool such as the 10x Genomics Loupe Browser v8.0.0. For runs where the spots of specific cell densities are generated, a different region template will be used where the location of these spots will be included.

Figure 4: Region template for generating synthetic spatial datasets

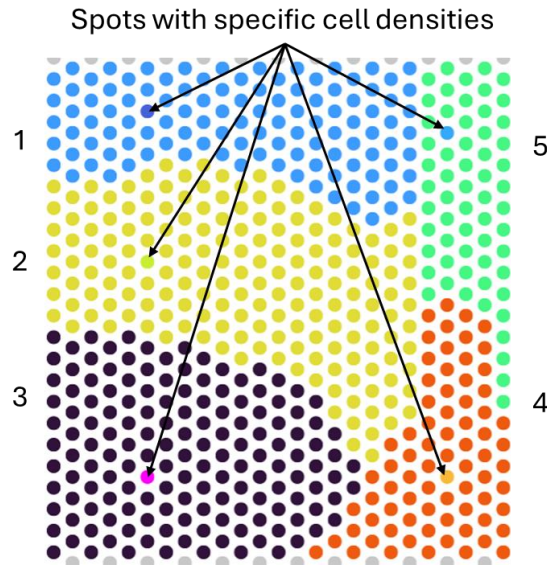


Figure 4 Legend: Region template for assigning coordinates to *synthspot* spots. Regions are numbered 1-5. Each color represents a different region with different cell type probabilities. The spots that are used for testing performances at different cell densities of certain cell types are also pointed out.

5.5 Spatial Deconvolution

The *benchdeconv* benchmarking pipeline currently will benchmark the deconvolution methods shown in figure 2. *spacedeconv* (Zackl, 2024), an R interface to the spatial deconvolution tools will be used to access the deconvolution methods. All methods were ran using their default parameters.

RCTD (Robust Cell Type Decomposition) is a model based deconvolution method where it assumes that the number of reads mapped to each gene in each spot follows a Poisson distribution, given the underlying expression levels of the gene in the celltypes. It takes the training scRNA-seq dataset to estimate the mean expression profiles of each cell, where the Poisson distribution parameters are then estimated. Maximum likelihood estimation (MLE) is then used to deconvolute the cell type fractions (Cable et al., 2021).

SPOTlight is a signature matrix based deconvolution method that utilizes both non-negative matrix factorization (NMF) and non-negative least squares

(NNLS) regression to estimate cell fractions. A signature matrix, a matrix that describes the transcriptional profiles of each cell type, is built from the input training scRNA-seq data. NMF is then used to identify representative factors composing each cell type for deconvolution (Elosua-Bayes et al., 2021).

CARD (Conditional AutoRegressive-based Deconvolution) is a signature matrix based deconvolution method which builds a signature matrix and utilizes NMF to deconvolute the cell fractions. It is able to borrow cell type compositions of neighbouring spots to estimate cell type composition correlations (Ma and Zhou, 2022), effectively modelling the observation that tissue specific cell types tend to colocalize (Elmentaite et al., 2022).

5.6 Performance Statistics

To measure the performance of each deconvolution method, RMSD (Root Mean Square Deviation) and JSD (Jensen-Shannon Divergence) were calculated. The Kruskal Wallis and Dunn tests were then performed to identify any significant differences in performances.

RMSD was calculated as follows:

$$\text{per Celltype RMSD} = \sqrt{\frac{\sum_{i=1}^N (x_i - \hat{x}_i)^2}{N}}$$

Where: N = the total number of spots, x_i = ground truth fraction of a single cell type and \hat{x}_i = predicted fraction of a single cell type. This returns the RMSD of each cell type, which is then averaged to give a representative RMSD value for each deconvolution attempt.

JSD was calculated as follows:

$$JSD(p_n \parallel q_n) = \frac{1}{2} D_{KL} \left(p_n \parallel \frac{p_n + q_n}{2} \right) + \frac{1}{2} D_{KL} \left(q_n \parallel \frac{p_n + q_n}{2} \right),$$

$$D_{KL}(a_n \parallel b_n) = \sum_{m=1}^M \left(a_{nm} \times \log \frac{a_{nm}}{b_{nm}} \right),$$

Where: p_n = ground truth fraction of a single cell type q_n = predicted fraction of a single cell type. As JSD is a symmetric and smoothed Kullback Leibler (KL) divergence, KL is also given where b_n = ground truth probability of a single cell type a_n = predicted probability of a single cell type. As with RMSD, this returns the JSD of each cell type, which is then averaged to give a representative JSD value for each deconvolution attempt.

The Kruskal Wallis and Dunn tests were then performed to determine if there were any significant performance differences:

1. between deconvolution methods
2. between scRNA-seq dataset sources ie. Different breast cancer subtypes
3. between each cell type in each method

The Benjamini–Hochberg multiple testing correction is applied to the Dunn tests to gather adjusted p values. This is to address the multiple testing problem. Benjamini-Hochberg was chosen over other more conservative multiple testing corrections, such as Bonferroni. This is because methods like Bonferroni would be too strict especially since controlling, rather than eliminating, the false discovery rate is wanted here.

Runtimes for *benchdeconv* and each deconvolution method will also be measured.

5.7 benchdeconv

Here we present *benchdeconv*, a benchmarking pipeline for spatial deconvolution methods. It takes single cell RNA datasets to produce synthetic spatial datasets for deconvolution and evaluation. *benchdeconv* can be installed via a GitHub repository (see Research Data Management), which includes the *benchdeconv.yaml* file to rebuild the conda environment, as well as the *R_installation.R* script for installing R dependencies. The coordinates for each region is stored in

data/spot_coords/out1.csv and out1_mintest.csv for the normal and minimum cell density for identification region coordinates respectively.

benchdeconv scripts are contained in *benchdeconv/scripts* where:

- | | |
|---------------|---|
| run.R | - Main <i>benchdeconv</i> script for importing data, generating synthetic datasets, deconvolution, and generating preliminary statistics. |
| run_hpc.sh | - Submission script for Eddie, the University of Edinburgh's Research Compute Cluster |
| run_stats.R | - R script to collect and generate statistics from Eddie runs |
| benchdeconv.R | - Module file for run.R containing run.R function definitions |

The --help flag for run.R and run_stats.R can be used to display the options and flags that need to be passed. Results are all reported in the "data/results" directory, if not specified by the --outdir option.

Addition of more deconvolution methods has been simplified, with the command for running more deconvolution methods can be input into the ##DECONVOLUTE section of run.R. The method name can be added to the method_names variable, and a data frame of deconvolution results for each method can be added to the method list variable to run the generation of relevant statistics.

5.8 Compute Specifications

All runs were conducted on Eddie, the University of Edinburgh's Linux compute cluster. It consists of 7000 Intel Xeon cores with 3TB member per compute node. A table of runs as well as the RMSD and JSD values of all runs are available in the supplementary materials, available upon request.

6. Research Data Management

The *benchdeconv* benchmarking pipeline can be found in:

<https://github.com/jaychowcl/benchdeconv>

The modified *synthspot* repository can be found in:

<https://github.com/jaychowcl/synthspot>

Breast cancer scRNA-seq dataset:

GEO assession: GSE176078

7. Results

7.1 Synthetic ST Datasets

Spatial scatter pies of the cell fractions of the generated ST datasets were generated. Run 1 plots are shown here in figure 1 as representative performances. RCTD outperforms the other methods by predicting correctly the presence of cell types within each spot, as well as their fractions, shown in figure 5 where the most similar to the ground truth is RCTD.

Figure 5: Ground truth and predicted spot compositions

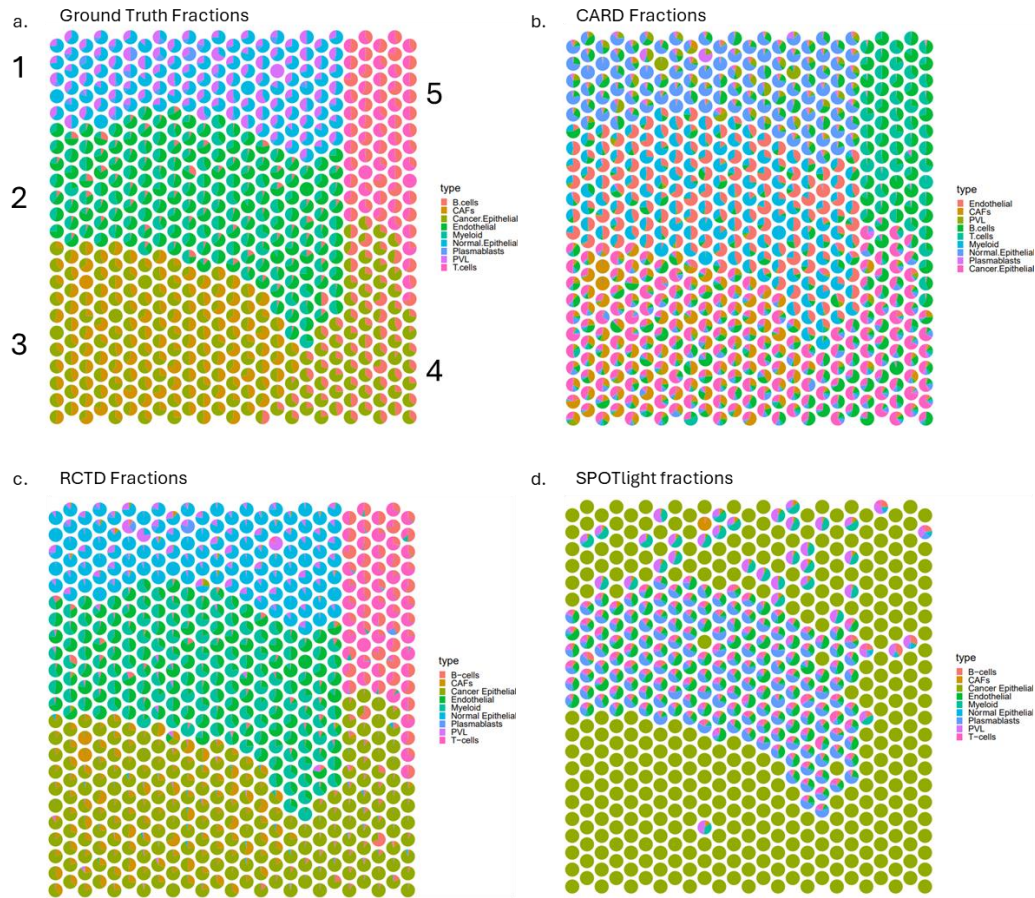


Figure 5 Legend: Spatial scatter pie visualization of ground truth and predicted spot composition of a single run. Run1 is plotted here with no downsizing in the TNBC dataset. Each spot is a pie chart showing the cell type compositions. Each colour represents a different cell type identity.

7.2 Performances Between Methods

The performance of each deconvolution method was then evaluated against each other. For each dataset, a boxplot of RMSD and JSD values of all runs in each dataset was generated, and statistical tests Kruskal-Wallis and Dunn were conducted to determine if there are any significant differences in medians of performances. RMSD and JSD values for each cell type from the 10 runs in each dataset were averaged to give a representative RMSD and JSD value for each cell type, which were then added to the plots.

Figure 6: Boxplots of deconvolution method performances between different datasets

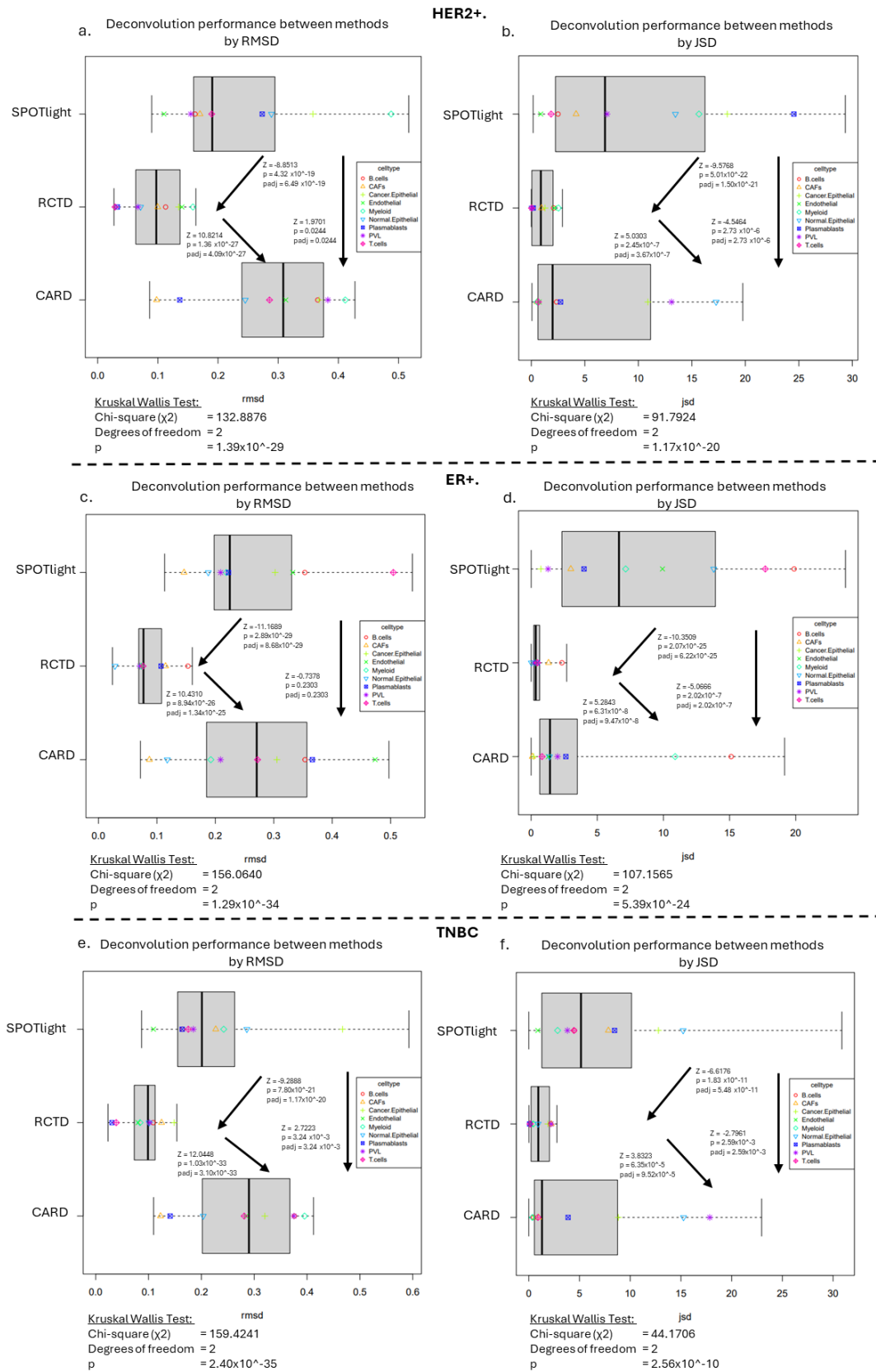


Figure 6 Legend: Boxplots of RCTD, SPOTlight and CARD performance in ER+, HER2+ and TNBC datasets by RMSD and JSD. The cell type performances have been averaged over 10 runs in each dataset and plotted, with each symbol representing a different cell type. Kruskal Wallis and Dunn tests are reported to test if there are any significant differences between means of each method.

As shown in figure 6, RCTD outperforms CARD and SPOTlight in each dataset. The ranking of deconvolution performance between the cancer subtype datasets ER+, HER2+ and TNBC remain consistent between datasets. However, there is a difference in performance between SPOTlight and CARD depending on whether the RMSD or JSD value is taken as an indicator of performance. In all datasets, the Kruskal-Wallis test returns a $p < 0.05$. We therefore accept the alternative hypothesis, median performance values are not equal and one method's performance method's median differs from another performance median, and reject the null hypothesis, that median performance values are equal. Further Dunn tests show that RCTD performs the best between all deconvolution methods with adjusted $p < 0.025$ (since two tailed $0.05/2 = 0.025$) in all datasets, showing that there is significant difference in the mean of RMSD and JSD between RCTD and all the other methods where the median of RCTD performance is higher, as shown by the lower RMSD and JSD values.

When considering RMSD, the next best performing method is SPOTlight, with adjusted $p < 0.025$ for the Dunn test when comparing against CARD in all datasets, showing that there is a significant difference in the median of RMSDs. However, when considering JSD, the opposite is true. Since RMSD is more sensitive to outliers, this may suggest that while SPOTlight is able to provide a more consistent performance, CARD is generally more accurate but may provide predictions with higher occurrences of outliers. This is further evidenced by the large positive skew in the CARD JSD boxplots. Interestingly the long tail may be due to the low prediction performance in certain celltypes where in both HER2+ and TNBC, CARD has lower performance when considering normal epithelial and PVL (perivascular-like) cells as shown in the cyan and purple points in figure 6. This is further explored when examining cell type specific performances.

7.3 Performances Between Datasets

The performance of the deconvolution methods between datasets were then evaluated and compared. For each dataset, a boxplot of RMSD and JSD values of all runs per dataset was generated. Statistical tests Kruskal-Wallis and Dunn were conducted to determine if there are any significant differences in medians of performances. RMSD and JSD values for each method from the 10 runs in each dataset were averaged to give a representative RMSD and JSD value for each method, which were then added to the plots. A heatmap of RCTD values for each cell type in each run was also generated.

Figure 7: Boxplots, Kruskal Wallis and Dunn Tests for Deconvolution Performance Between Datasets

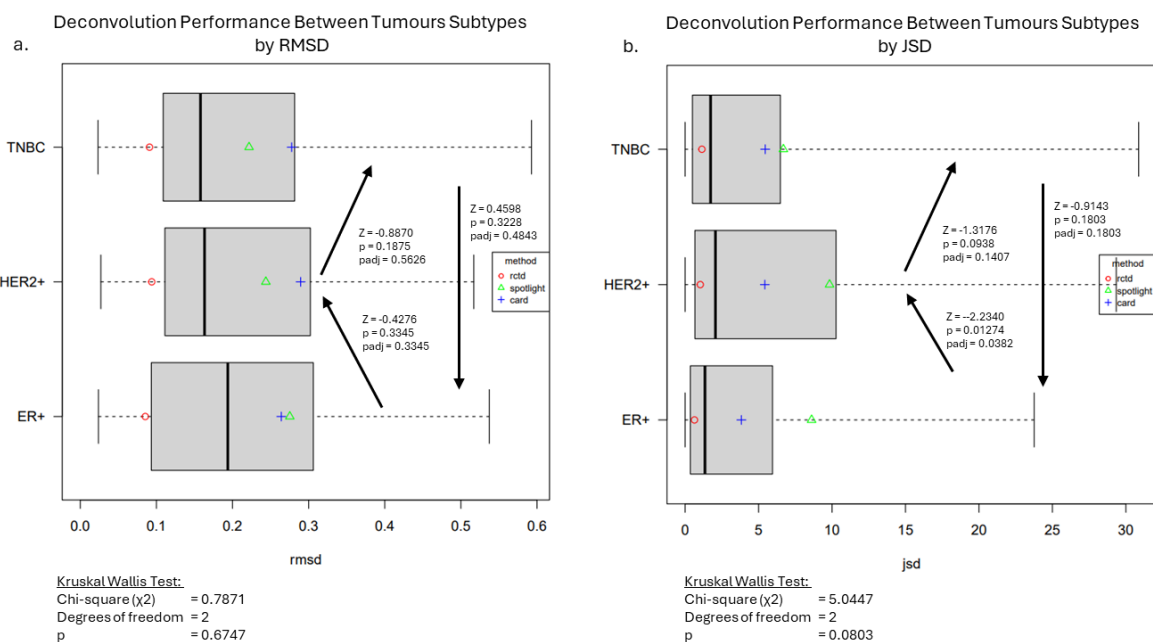


Figure 7 Legend: Boxplot of RCTD, CARD and SPOTlight performances within ER+, HER2+ and TNBC datasets by RMSD and JSD. The performance of each method has been averaged over 10 runs for each dataset and plotted. Each symbol represents a different deconvolution method. Kruskal Wallis and Dunn tests were performed to determine any significant changes in medians.

Figure 8: Heatmap of RMSD values for each cell type for each run

	HER2+										ER+										TNBC										
Run	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	
B.cells	0.117	0.109	0.11	0.11	0.112	0.112	0.116	0.111	0.118	0.114	0.161	0.152	0.145	0.154	0.153	0.151	0.159	0.153	0.159	0.143	0.112	0.105	0.111	0.102	0.109	0.111	0.114	0.112	0.111	0.104	
CAFs	0.097	0.105	0.095	0.09	0.096	0.111	0.107	0.097	0.092	0.106	0.114	0.117	0.108	0.118	0.112	0.117	0.12	0.113	0.117	0.112	0.123	0.129	0.126	0.121	0.12	0.124	0.119	0.133	0.129	0.122	
Cancer.Epithelial	0.137	0.136	0.127	0.127	0.13	0.142	0.143	0.132	0.131	0.141	0.08	0.077	0.064	0.074	0.08	0.067	0.07	0.071	0.08	0.067	0.151	0.15	0.149	0.153	0.146	0.148	0.147	0.154	0.149	0.138	
Endothelial	0.14	0.142	0.14	0.137	0.143	0.139	0.135	0.142	0.145	0.137	0.072	0.077	0.07	0.079	0.077	0.066	0.065	0.076	0.073	0.069	0.083	0.08	0.072	0.085	0.079	0.075	0.081	0.088	0.087	0.073	
Myeloid	0.159	0.162	0.161	0.154	0.161	0.154	0.148	0.163	0.162	0.157	0.069	0.078	0.075	0.086	0.08	0.064	0.066	0.078	0.073	0.066	0.086	0.084	0.081	0.071	0.083	0.079	0.084	0.087	0.097	0.085	
Normal.Epithelial	0.074	0.069	0.076	0.065	0.075	0.072	0.08	0.063	0.072	0.066	0.026	0.033	0.031	0.026	0.024	0.03	0.029	0.027	0.033	0.024	0.105	0.094	0.104	0.098	0.102	0.105	0.098	0.102	0.102	0.101	
Plasmablasts	0.032	0.035	0.033	0.029	0.04	0.032	0.034	0.031	0.031	0.031	0.113	0.103	0.107	0.104	0.106	0.099	0.113	0.113	0.107	0.106	0.032	0.026	0.042	0.03	0.028	0.035	0.026	0.035	0.028	0.023	
PVL	0.065	0.067	0.072	0.057	0.075	0.068	0.081	0.056	0.064	0.063	0.077	0.07	0.067	0.07	0.07	0.073	0.07	0.064	0.081	0.064	0.103	0.094	0.099	0.102	0.103	0.113	0.099	0.1	0.104	0.103	
T.cells	0.027	0.029	0.03	0.031	0.028	0.029	0.03	0.028	0.029	0.028	0.081	0.073	0.075	0.072	0.075	0.076	0.078	0.08	0.075	0.075	0.038	0.04	0.041	0.034	0.034	0.039	0.047	0.033	0.043	0.037	
B.cells	0.36	0.366	0.36	0.375	0.37	0.36	0.371	0.353	0.36	0.38	0.357	0.364	0.334	0.353	0.396	0.338	0.318	0.362	0.351	0.362	0.412	0.362	0.367	0.412	0.371	0.375	0.368	0.352	0.39	0.35	
CAFs	0.096	0.092	0.092	0.096	0.107	0.098	0.1	0.106	0.088	0.113	0.093	0.084	0.086	0.104	0.087	0.104	0.08	0.072	0.08	0.08	0.118	0.128	0.128	0.126	0.111	0.126	0.111	0.136	0.133	0.11	
Cancer.Epithelial	0.355	0.362	0.379	0.356	0.37	0.369	0.375	0.387	0.368	0.359	0.304	0.303	0.312	0.306	0.324	0.294	0.3	0.305	0.308	0.297	0.331	0.292	0.304	0.313	0.308	0.308	0.326	0.362	0.301	0.357	
Endothelial	0.302	0.299	0.298	0.331	0.312	0.303	0.327	0.304	0.313	0.336	0.459	0.471	0.472	0.472	0.473	0.471	0.462	0.476	0.482	0.497	0.277	0.288	0.299	0.31	0.301	0.281	0.308	0.252	0.286	0.244	
Myeloid	0.417	0.428	0.416	0.396	0.407	0.413	0.406	0.423	0.42	0.389	0.225	0.188	0.199	0.196	0.163	0.202	0.192	0.185	0.191	0.18	0.399	0.396	0.395	0.378	0.4	0.387	0.386	0.403	0.405	0.403	
Normal.Epithelial	0.25	0.251	0.247	0.24	0.236	0.241	0.246	0.272	0.223	0.246	0.123	0.117	0.113	0.107	0.112	0.116	0.115	0.112	0.133	0.128	0.114	0.236	0.246	0.158	0.23	0.203	0.217	0.202	0.204	0.224	
Plasmablasts	0.138	0.138	0.143	0.137	0.136	0.137	0.139	0.129	0.136	0.131	0.371	0.357	0.384	0.352	0.362	0.365	0.345	0.37	0.373	0.38	0.138	0.133	0.125	0.134	0.145	0.146	0.127	0.146	0.157	0.159	
PVL	0.376	0.381	0.397	0.376	0.383	0.389	0.392	0.377	0.383	0.371	0.195	0.213	0.207	0.208	0.217	0.209	0.232	0.205	0.2	0.203	0.376	0.367	0.363	0.367	0.38	0.365	0.372	0.386	0.372	0.404	
T.cells	0.289	0.285	0.283	0.28	0.279	0.294	0.288	0.294	0.299	0.267	0.282	0.279	0.262	0.281	0.288	0.27	0.26	0.272	0.27	0.26	0.288	0.276	0.293	0.271	0.28	0.285	0.278	0.282	0.282	0.276	
B.cells	0.166	0.177	0.146	0.173	0.157	0.143	0.187	0.159	0.151	0.16	0.352	0.348	0.348	0.354	0.342	0.352	0.356	0.36	0.353	0.366	0.149	0.147	0.161	0.139	0.211	0.164	0.161	0.135	0.191	0.185	
CAFs	0.175	0.174	0.163	0.162	0.168	0.182	0.165	0.157	0.172	0.182	0.168	0.155	0.152	0.161	0.139	0.163	0.137	0.116	0.157	0.113	0.232	0.158	0.239	0.238	0.235	0.236	0.236	0.221	0.248	0.229	
Cancer.Epithelial	0.355	0.356	0.367	0.352	0.36	0.359	0.359	0.365	0.355	0.35	0.304	0.307	0.308	0.303	0.311	0.302	0.309	0.305	0.298	0.277	0.39	0.378	0.467	0.419	0.593	0.459	0.436	0.374	0.554	0.512	
Endothelial	0.101	0.09	0.12	0.124	0.114	0.113	0.09	0.107	0.11	0.131	0.316	0.323	0.324	0.317	0.34	0.331	0.32	0.33	0.318	0.408	0.123	0.113	0.114	0.104	0.117	0.107	0.108	0.087	0.112	0.107	
Myeloid	0.493	0.466	0.507	0.475	0.485	0.517	0.451	0.483	0.495	0.505	0.204	0.212	0.194	0.238	0.224	0.209	0.216	0.193	0.198	0.296	0.263	0.28	0.233	0.235	0.209	0.232	0.226	0.283	0.231	0.229	
Normal.Epithelial	0.286	0.288	0.292	0.288	0.284	0.295	0.284	0.291	0.287	0.292	0.183	0.185	0.185	0.2	0.184	0.19	0.183	0.203	0.196	0.17	0.283	0.299	0.3	0.293	0.293	0.281	0.294	0.274	0.277	0.278	
Plasmablasts	0.272	0.265	0.282	0.302	0.259	0.259	0.254	0.285	0.255	0.3	0.217	0.218	0.225	0.223	0.233	0.21	0.222	0.225	0.223	0.23	0.178	0.128	0.137	0.14	0.186	0.189	0.155	0.147	0.204	0.172	
PVL	0.14	0.187	0.135	0.133	0.174	0.135	0.207	0.154	0.158	0.123	0.214	0.216	0.21	0.196	0.224	0.213	0.207	0.208	0.22	0.181	0.218	0.182	0.172	0.205	0.169	0.201	0.203	0.175	0.168	0.154	
T.cells	0.189	0.186	0.196	0.192	0.186	0.191	0.186	0.193	0.19	0.187	0.537	0.512	0.518	0.513	0.467	0.535	0.496	0.475	0.523	0.469	0.155	0.289	0.162	0.142	0.174	0.154	0.144	0.172	0.181	0.18	

Figure 8 Legend: Heatmap of RMSD values for each cell type for a subset of runs using different datasets. The darker red colours represent higher RMSD values while lighter green colors represent lower RMSD values.

The heatmap in figure 8 shows that within each dataset, there is strong reproducibility between runs. It emphasizes that the main factor that modulates performance of the deconvolution tools is the cell type. Interestingly, it seems that the dataset impacts the deconvolution performance for certain cell types. For instance, RCTD performs well at deconvoluting normal epithelial cells in the spots for ER+ but performs worse at deconvoluting with normal epithelial cells in TNBC. This may be due to the different number of T cells between each dataset. However, as shown below, dataset size does not impact performance and may not explain this difference.

As shown in figure 7, there is no evidence that there is a significant difference in performance based on gathered observations between deconvoluting the different datasets from breast cancer subtypes. This can be shown in the $p > 0.05$ in the Kruskal-Wallis test. We therefore accept the null hypothesis, there is no significantly different performance differences between different datasets, and do not accept the alternative hypothesis, which is that there is a dataset with significantly different performances. This be further shown in the Dunn test, where $p > 0.025$ for all comparisons. This suggests that the performance of the deconvolution tools are not

impacted by the source of the reference dataset, as long as the reference scRNA-seq dataset comes from a suitable source which is representative of the source of ST data that is being deconvoluted.

The same observation that SPOTlight performs better than CARD when being evaluated with RMSD, and vice versa when considering JSD, can be made.

7.4 Impact of Single Cell RNA Reference Dataset Size

Here we assess the performance of the deconvolution tools when given differing sizes of input training scRNA-seq datasets. A scatter plot was generated to observe any trends.

Figure 9: Scatter plot of deconvolution performance with different sizes of scRNA-seq

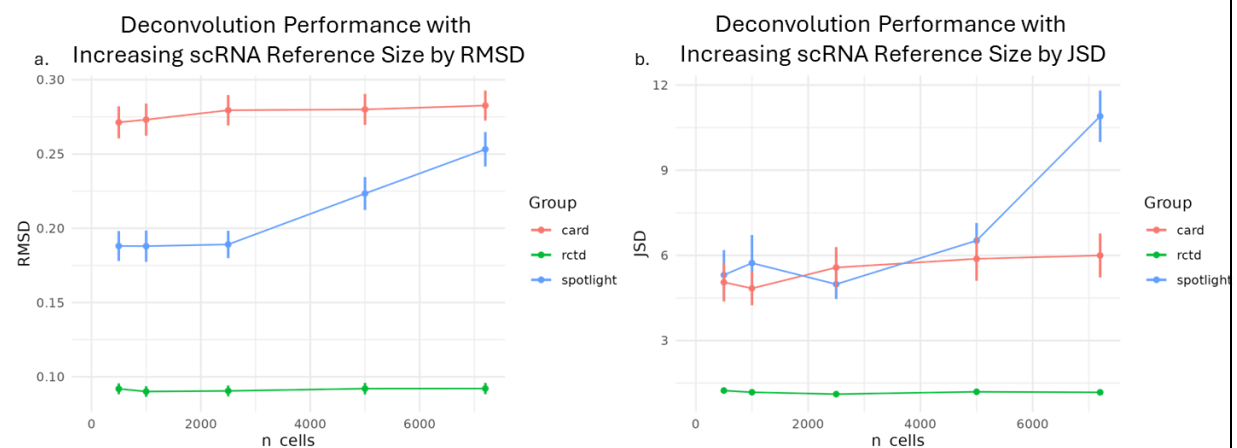


Figure 9 Legend: Scatter plot of deconvolution performance with varying sizes of reference scRNA-seq datasets by RMSD and JSD. Each point has been averaged over 10 runs, with 50 runs in total for the 5 different sizes of dataset.

RCTD and CARD performance remain stable at each size of reference scRNA-seq dataset. Both methods do not seem to be impacted by the number of cells in the reference dataset. This suggests that even with only 500 cells per cell type in the reference scRNA-seq dataset, it would be sufficient for adequate deconvolution. However, SPOTlight seems to deteriorate in performance as the number of cells in each cell type increases. This is unexpected as SPOTlight should theoretically have

more information to better determine the fingerprint of the cell types and therefore better deconvolution performance.

7.5 Cell Type Specific Performance Biases

Here we further explore the performance differences between different cell types. A boxplot of RMSD and JSD values for each cell type per method has been generated. Dunn test results are also reported.

Figure 10: Boxplots of the performance of different cell types in each method

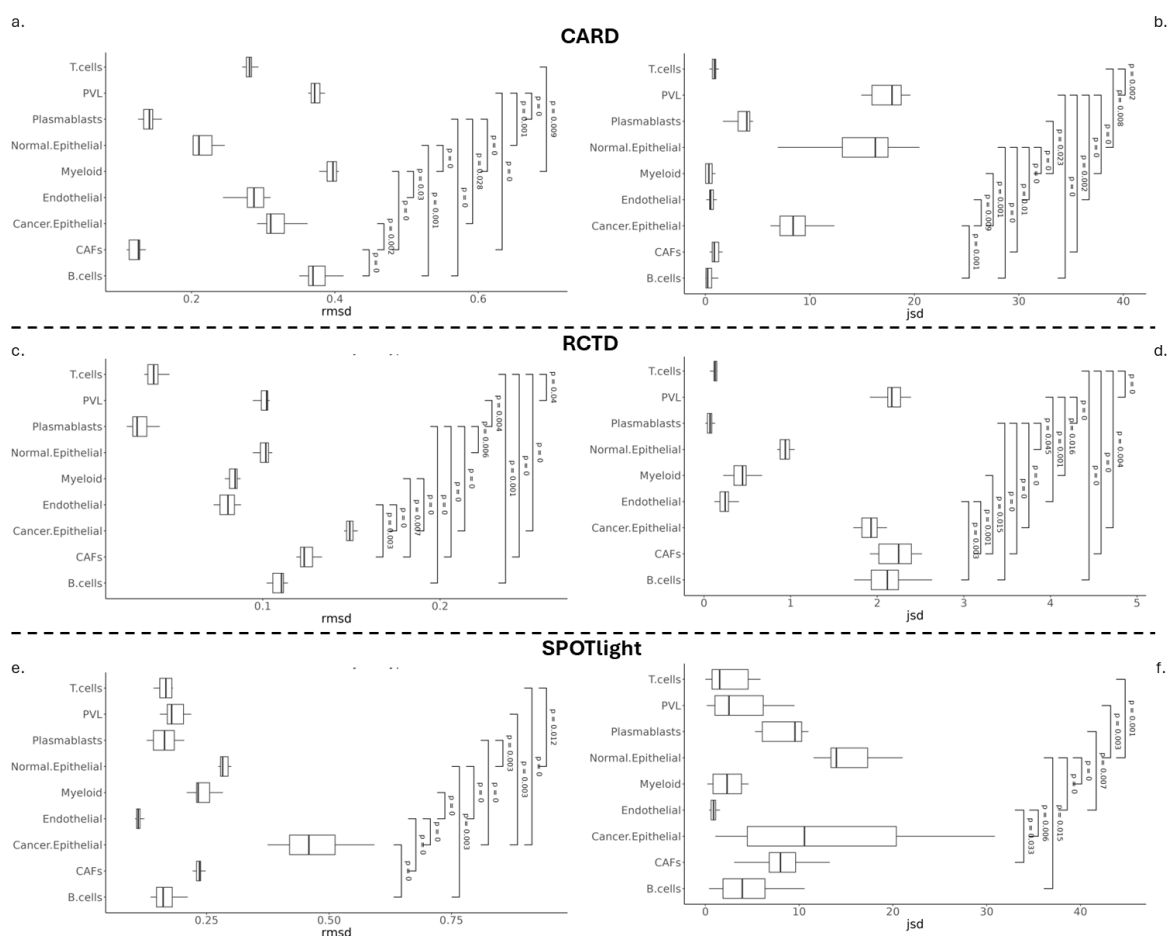


Figure 10 Legend: Boxplot of cell type specific performances by RMSD and JSD within the TNBC dataset by RCTD, CARD and SPOTlight. The performance of each method has been averaged over 10 runs for each dataset and plotted. Dunn tests were performed to determine any significant changes in medians.

Again, it is clear that RCTD outperforms CARD and SPOTlight, as shown by the RMSD and JSD values of almost all cell type performances by RCTD since

they are lower than all the RMSD and JSD values in other methods. There is also a clear difference in performance between deconvoluting different cell types in each method. The cell type that each method consistently had trouble deconvoluting were cancer epithelial cells which is shown in the high JSD and RMSD values. Interestingly, other than cancer epithelial cells, each method seems to prefer deconvoluting different cell types. For instance, RCTD has trouble with CAFs (cancer associated fibroblasts) while CARD performs quite well. However, as shown in the table in figure 8, the performance in each cell type may differ between datasets. Despite this, the gathered performance data can be used as a guide to determine deconvolution performance for the cell types in the TNBC dataset.

7.6 Spatial Performance Biases

Here we gather error heatmaps for the deconvolution results to assess any spatial performance biases.

Figure 11: Error heatmaps for each deconvolution method

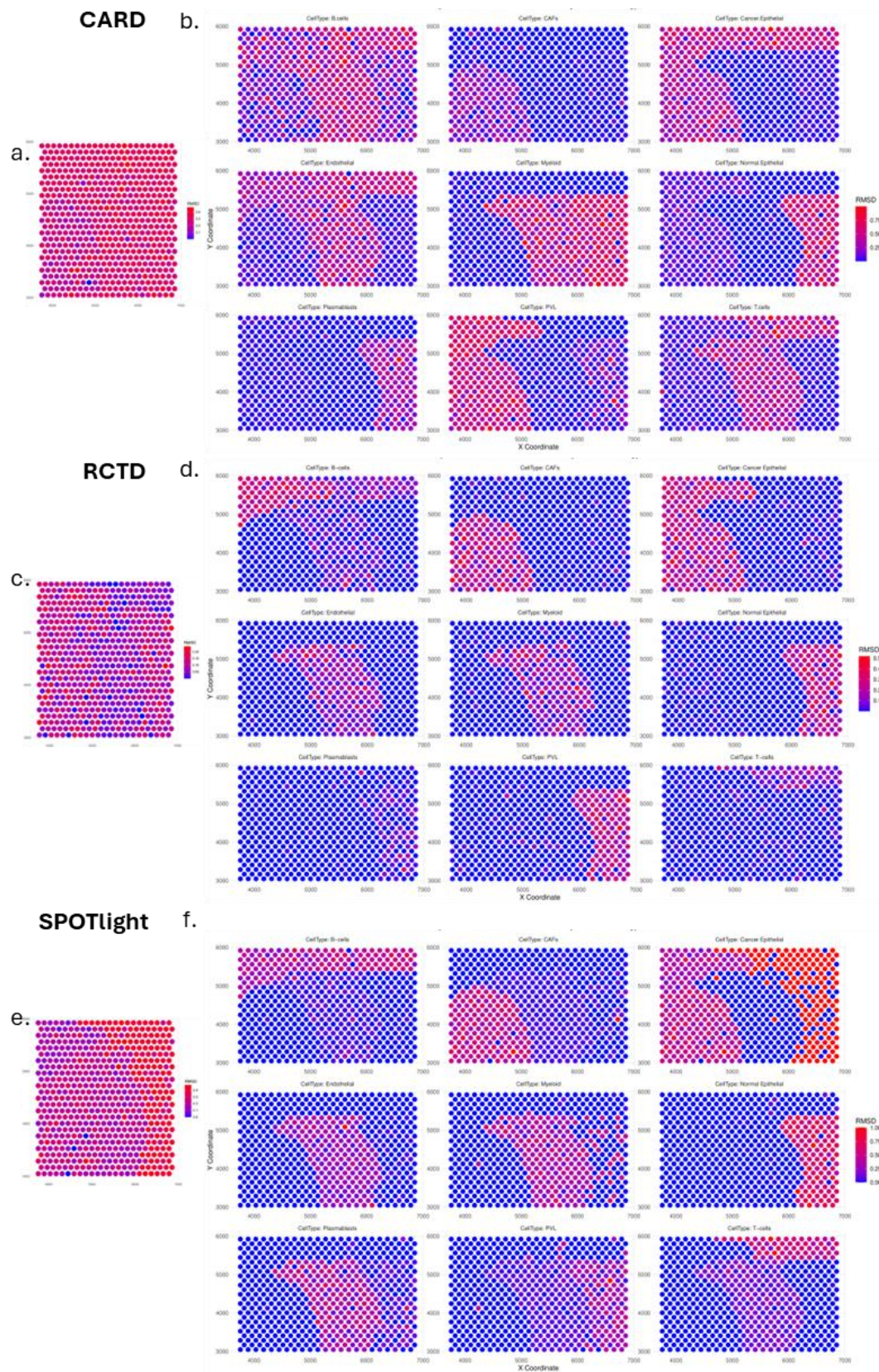


Figure 11 Legend: Error heatmaps for each deconvolution method where a. c. and e. represent total error heatmaps while b., d. and f. showcase cell type specific error heatmaps.

When considering the performance for all cell types there seem to be no spatial biases, as shown in figure 11.a, 11.c and 11.e where there are no clear patterns of error. However, when considering each cell type, there are clear patterns of error that generally follow the boundaries and areas of each region. Though CARD is the only method that borrows the composition information from surrounding spots for deconvolution, every method displays these patterns of error. This suggests that the errors do not come from any spatial biases but originate from the method's ability to determine the presence or absence of a certain cell type within a region, as well as determining the specific fraction of a cell type that is present.

RCTD is able to accurately determine the presence and absence of a cell type within each spot. This can be seen in the low error of spots in each cell type in figure 11.d, where the blue regions are all region that do not contain that specific cell type. The regions of higher error in all cell types correspond to regions where the cell type is present. This suggests that though RCTD is able to accurately determine the presence/absence of a cell type very well.

CARD displays the same problems with RCTD where the region of higher errors primarily arises from the error between predicted and actual cell type fractions of cell types that are present and has been detected to be present. However, CARD displays lower performance than RCTD when determining the presence and absence of a cell type. For example, CARD predicts that B-cells are in all regions even when it isn't, as shown in figure 11.b, leading to the undefined error borders that is present in other error heatmaps where not only is it unable to detect the absence of B-cells in some region 2 spots, but also drastically overestimates the fraction of B-cells in region 2. However, this may be due to the very low counts of B cells per spot within this region, leading to this reduced performance. The issue of detecting cell types can also be further be seen in figure 11.b where though PVL cell fractions were accurately predicted in region 1 where they are present, CARD predicts their presence in regions 3 and 4 where they are not present.

SPOTlight is the worst performing method where it is worse than the other methods at detecting the presence and absence of cell types, as well as predicting their cell fractions. Though figure 11 only shows one run, we find that these types of errors are widespread. Not only was SPOTlight unable to accurately determine the presence and absence of almost all cell types, where it predicts that most of the spots contain cancer epithelial cells, it is also unable to accurately determine the cell fractions, as seen in Figure 11.d where it predicts that most cells are 100% cancer epithelial. In region 2, the only region predicts the absence of cancer epithelial cells, it predicts that the spots are composed of a mixture of all cells barring cancer epithelial cells when it only is composed of endothelial and myeloid cells with a small fraction of B-cells. This highlights the poor performance of SPOTlight when determining the presence and absence of a cell type in a spot, as well as the fractions of present cell types.

7.7 Performances at Different Cell Densities of Immune Cell Types

Figure 12: Scatterplot of deconvolution performance at different densities of cell types

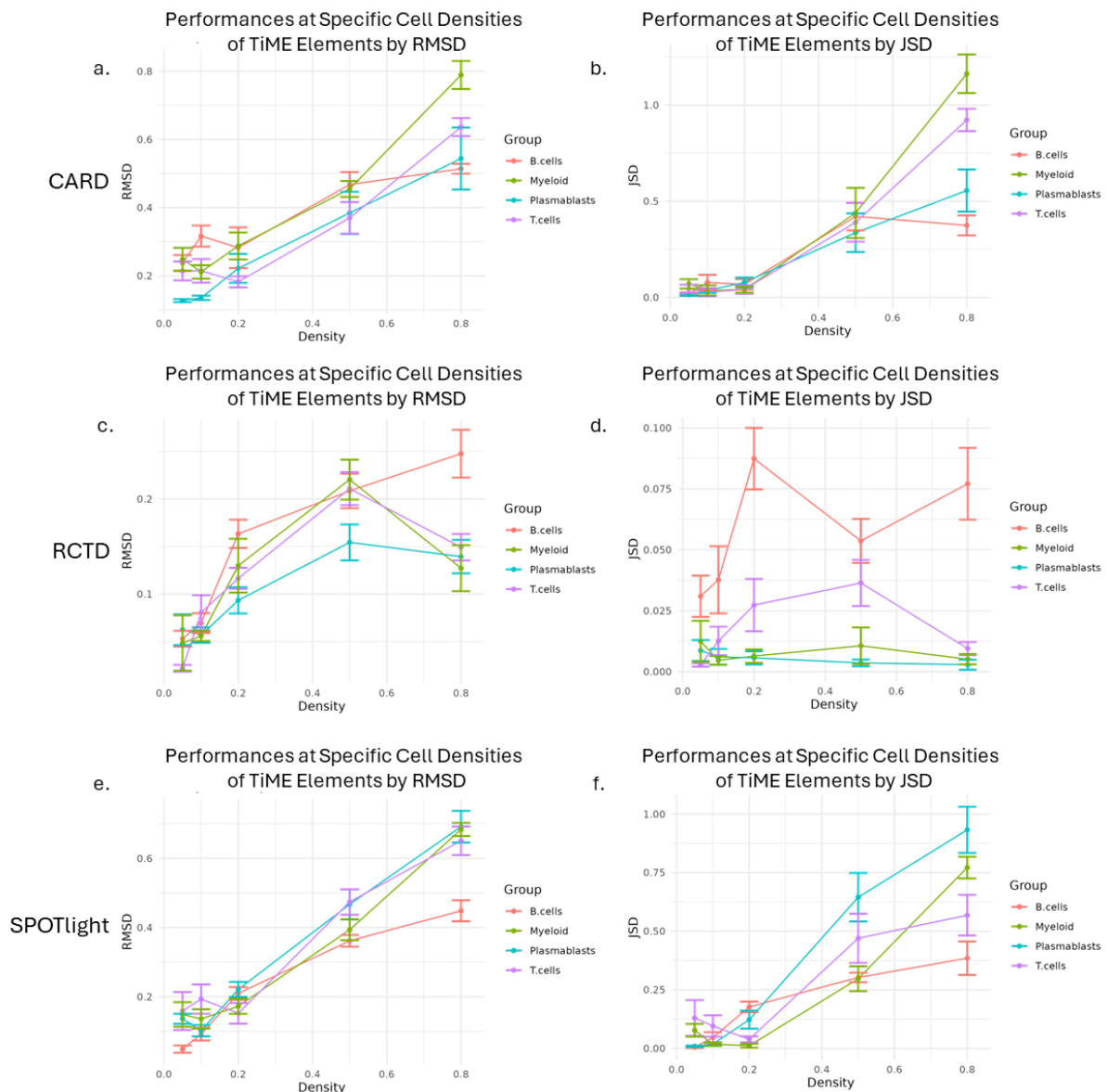


Figure 12 Legend: Scatterplot of performance by RMSD and JSD of RCTD, CARD and SPOTlight at different cell densities of B cells, myeloid, plasmablasts and T cells. Each point is an average of 5 runs, with a total of 100 runs for all points and all cell types.

As shown in figure 12, RMSD and JSD generally increases as cell density increases. This observed increase is counterintuitive since the theoretically the method will have more access to information about the identity of the cell type.

However, this is due to the low number of cells counts at the lower densities. For instance, at density 0.05, there are only 2 cells of a chosen cell type in spot when there are a maximum of 40 cells in a spot. This effectively caps the error since potential variability is limited, making it artificially low. This constrains the RMSD and JSD at lower densities, since even if the method did not detect the cell type the RMSD and JSD penalty would be limited due to the small difference in 0 and 0.05.

Interestingly, RCTD is the only method where performance increases after cell density reaches 0.5, with the exception of B-cells. This suggests that RCTD can better predict cell fractions when there are a higher number of cells in a spot, but only after a certain density.

7.8 Runtimes

Figure 13: Table of runtimes for *benchdeconv* and deconvolution methods

Method	runtimes (min)		
	no downsize	5000 cells per cell type	500 cells per cell type
benchdeconv	120.2	75.6	15.2
rctd	5.4	5.4	5.4
spotlight	109.4	64.2	4.9
card	0.5	0.4	0.2

Figure 13 Legend: The minutes each deconvolution method, as well *benchdeconv* itself, took for deconvolution. Runtimes of different sizes of scRNA-seq datasets are also shown.

Figure 13 shows runtimes of representative runs for each deconvolution method, as well as the runtime for the *benchdeconv* pipeline. CARD and RCTD takes a very short time when compared to SPOTlight. Interestingly, RCTD and CARD are very minimally affected by the size of the reference scRNA-seq dataset while SPOTlight takes an increasing amount of time as dataset size increases.

8. Discussion

Three deconvolution tools (RCTD, SPOTlight, and CARD) were evaluated for their performance for deconvoluting ST data within the context of various breast cancer subtype TMEs (TNBC, HER2+, ER+). A benchmarking pipeline was built for the purpose of evaluating and comparing spatial deconvolution performances between methods. We investigate the performance of each method, the impact of changing TME datasets on performance, TME cell type specific performances, spatial biases in deconvolution, and prediction performances at different cell densities of TME cell types. We find that RCTD outperforms SPOTlight and CARD by a large margin, being able to better distinguish the presence and absence of cell types in each spot as well as being better able to predict the cell fractions.

8.1 Validity of Cell Type Annotations in Source scRNA-seq Datasets

There are currently active discussions on the reliability of cell type annotations, particularly annotations gathered using dimensionality reduction techniques such as UMAP or t-SNE. It has been shown that after reducing to lower dimensions, distances between higher dimensional clusters are highly distorted leading to inaccurate cell type annotations if clusters are gathered directly from the 2 dimensional representation (Chari and Pachter, 2023). This is not an issue here as the GSE176078 annotations have been gathered using *scSubtype* (Wu et al., 2021). *scSubtype* is an intrinsic subtyping method where pseudo-bulk RNA-seq profiles of different tumour subtypes are generated and an adapted PAM50 centroid predictor is applied. Intrinsic subtypes are subtypes that are determined by their gene expression profiles, and the PAM50 centroid predictor is the most widely used intrinsic subtyping technique for breast cancer. It uses a set of 50 genes to train and find 5 centroids that represent different subtypes. The nearest centroid approach (Tibshirani et al., 2002) is then used to classify new cells. As this method does not apply dimension reduction, only a nearest centroid classification approach, distances and relationships between centroids and new data are preserved and not distorted. This method would therefore alleviate the

concerns of using dimension reduction for classification, leading to higher confidence than UMAP or t-SNE annotation methods.

8.2 Distribution of the Number of Cells per Spot in the Synthetic Data

In the current study, *benchdeconv* builds synthetic spots with *synthspot* with 1-40 cells per spot. This range can be set by the user, thus allowing future spatial deconvolution benchmarking with different overall cell densities in each spot. However, *synthspot* chooses the number of cells per spot from a uniform distribution, meaning it is equally likely to choose any number between 1-40. This is not representative of real ST spot data, where Dr. Florent Petitprez's lab observed numbers between 1-40, with a median of 9 cells per spot in typical cancer ST. This would lead to over inflated cell numbers in each cell, which may impact performance.

Higher cell numbers would make it harder for the deconvolution methods to identify identities of low-density cell types in each spot. For instance, figure 5 shows that while RCTD is able to identify a distinct cell type of low frequency in region 2, it misidentifies the cell identity as T-cells rather than B-cells, its actual identity. This may be due to the low number of the cell type in the spot, where the likelihood of gathering the marker genes that each method uses to fingerprint and identify the cell type is lower, thus reducing the ability of the methods accurately determine the cell type identity.

Therefore, to more accurately represent real ST spot cell densities, *synthspot* may be further adapted to sample the no. of cells per spot from a skewed distribution with high positive skew. However, further research must be conducted to determine the best distribution to use that reflects real ST spot densities as the current cell density distributions are unknown. It is also unknown whether these distributions vary depending on the source of the sample.

8.3 Validity of Synthetic Spatial Datasets

The current template for the regions has been created randomly and the prior frequencies of the cell types in each region, though informed by cell type frequencies in the reference scRNA-seq, have been created artificially. The region areas, arrangement and composition are therefore artificial and may not represent real ST data well. This may deteriorate CARD performance, as it is the only method evaluated here that considers the composition of surrounding spots. Therefore, *synthspot* performance in replicating TME ST data must be further validated.

8.4 Heterogeneity of Cells in a Cell Type Complicates Deconvolution

As seen in figure 10, there is strong evidence that suggests that deconvolution performance is related with how heterogenous a cell type is. For instance, cancer epithelial cells are highly heterogenous. This may lead to an increase in the variability in the representative fingerprints of each cell type, thus leading to a decrease in performance. This is consistent with B cell performances, another highly heterogenous cell type (Wei et al., 2021) with lower deconvolution performance.

8.5 Deteriorating SPOTlight Performance as scRNA-seq Size Increases

Figure 9 shows the deterioration of SPOTlight performance past 2000 cells per cell type. It is unknown why performance degrades with this trend since intuitively, the method would have more access to information that can identify the cell type identities with a bigger reference scRNA-seq.

This deterioration in performance may be due to the uneven number of cells per cell type in the reference. This can be shown in figure 3, where some cell types do not have enough cells to fill the minimum dataset size of 500 cells per cell type. This imbalance in cell types in the reference may introduce errors to each method when identifying the fingerprint of the cell type. For instance, a gene expressed in an overrepresented cell type in the reference may be assigned as a marker gene for a

specific cell type even when it is expressed in multiple underrepresented cell types, leading to a decrease in performance.

The deterioration in performance may also be due to the increased variability within a cell type when introducing more cells to the reference. This stems from the fact that there is a high level of heterogeneity even within the same cell type; the gene expression of cells of the same cell type may be very different depending on their condition and major or minor classifications. This is compounded by the fact that cancer cells are highly heterogeneous (Xu et al., 2024), suggesting that performance in deconvoluting cancer cells will be less effective than with other cell types. This phenomenon can be observed in figure 10 where the methods consistently have bad performance in deconvoluting cancer epithelial fractions. The methods therefore have more difficulty in identifying the fingerprint and marker genes that represent a cell type.

8.6 Incomplete scRNA-seq Reference Data

Spatial deconvolution requires a reference scRNA-seq dataset. All cell types that are present in ST must also therefore be present in the scRNA-seq for spatial deconvolution to proceed normally. Without reference scRNA-seq data on a cell type, a deconvolution method is unable to build a transcriptional fingerprint for the missing cell type for deconvolution.

Unfortunately, single cell sequencing of neutrophils and other granulocytes remain a challenge. The low RNA content, high RNases and inhibitory compounds within these cell types complicate sequencing. Neutrophils in particular are especially fragile. They are extremely sensitive to degradation and cannot be isolated using typical density gradient centrifugation (Xie et al., 2020). Therefore, neutrophils are typically not included in these scRNA-seq datasets even if specialized protocols exist due to the time, cost and additional complexity. This is an issue because neutrophils are a key element in cancer progression that promotes angiogenesis, immunosuppression and cancer metastasis (Wu et al., 2020). Without their inclusion in the scRNA-seq data, the deconvolution process will fail to accurately capture the cellular composition of the samples, potentially leading to incomplete or misleading interpretations of spatial transcriptomic data.

This limitation is particularly critical in studies of cancer and other inflammatory diseases, where neutrophils play a pivotal role in shaping the tumor microenvironment and influencing disease progression.

8.7 More Tools and Datasets

Only 3 deconvolution tools have been evaluated here: RCTD, CARD, and SPOTlight. However, many deconvolution methods have been developed, many of which have not yet been benchmarked in the context of the TME. Therefore, to attain a more comprehensive understanding of current spatial deconvolution capabilities, more tools need to be evaluated. The *benchdeconv* pipeline simplifies this by adding a section in `run.R`, `##DECONVOLUTE` where the command for deconvolution can be added. The results can be added to the list of results in the `methods` and `name` in the `method_names` variable where it will automatically be added into the pipeline.

For further spatial deconvolution performance evaluations, immunostaining information can be used to construct ground truth fractions before array ST. Though only feasible for some cell types, this allows the evaluation of the deconvolution methods with real ST data where the ground truth is known (for some cell types), thus better evaluating performance in real datasets.

The deconvolution methods can additionally be evaluated using synthetic datasets originating from imaging ST. Cell identities can be determined from imaging ST by using probes that bind to marker gene transcripts. After identifying cell identities, gene counts from cells within a 55µm diameter circle (to simulate Visium spots) can be summed to give synthetic ST spot counts with the ground truth cell fraction of the cell type in each 'spot'. Though not the complete transcriptome since imaging ST only assays a panel of transcripts, the true ground truth cell compositions can be known. However, imaging ST is panel based and do not measure the transcriptome. This is likely to strongly impact the performance, dependent on the statistical framework underlying the deconvolution method, and in non-predictable ways. This would therefore make the benchmarking unreliable.

This study mainly focuses on the TNBC dataset due to time constraints. However, as seen in figure 8, cell type specific performances vary between datasets. However, it is unknown whether these differences in performance are due to the differences in dataset source, or differences in individual datasets. Therefore, further benchmarking studies should include multiple datasets of the same cancer type. If cell type specific performances vary significantly between these datasets, it will suggest that the difference in cell type performance is not due to the dataset source, but rather the dataset itself regardless of its source. In addition, scRNA-seq datasets from other tumours may be evaluated to gather the cell type specific performances in each tumour type, and to confirm that different tumours impart different cell type specific performance when deconvoluting.

Future work within this space may also choose to focus on the ability of each deconvolution method to segregate cell types at different granularities. Granularities refers to the level of detail in which the cell types are categorized. For instance, T-cells can be further subtyped into T cells CD4+ and T cells CD8+. Here we only evaluate performances using the `celltype_major` annotations. In GSE176078 however, finer grain annotations are available, `celltype_minor` and `celltype_subset`. It is unknown whether these deconvolution methods will perform better or worse when facing finer grain annotations. It is unknown whether the deconvolution methods will perform better, perhaps due to each cell type having less variability thus making it easier for spatial deconvolution to determine cell types, or whether the deconvolution methods will perform worse since these sub-subtypes of cells will be very similar to each other. This may lead to the inability for the deconvolution methods to generate a transcriptional profile for a certain cell type that sufficiently differentiates it from other similar cell types.

9. Conclusion

Here we demonstrate the spatial deconvolution benchmarking pipeline *benchdeconv* in evaluating and comparing different deconvolution methods and their performances in the TME. We show that RCTD outperforms SPOTlight and CARD by a large margin. This holds true for all dataset types, highlighting RCTD

robustness. However, an issue every deconvolution method faced is the high degree of heterogeneity of certain cell types within the TME, such as cancer epithelial cells. Future work in this space should therefore aim to evaluate more deconvolution methods and datasets to further our understanding of spatial deconvolution in the TME. This will be crucial for informing further research into characterizing the spatial architecture of the TME on the best spatial deconvolution method to use and when.

10. References

- ANDERSON, N. M. & SIMON, M. C. 2020. The tumor microenvironment. *Curr Biol*, 30, R921-r925.
- BAGHBAN, R., ROSHANGAR, L., JAHANBAN-ESFAHLAN, R., SEIDI, K., EBRAHIMI-KALAN, A., JAYMAND, M., KOLAHIAN, S., JAVAHERI, T. & ZARE, P. 2020. Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Communication and Signaling*, 18, 59.
- BINNEWIES, M., ROBERTS, E. W., KERSTEN, K., CHAN, V., FEARON, D. F., MERAD, M., COUSSENS, L. M., GABRILOVICH, D. I., OSTRAND-ROSENBERG, S., HEDRICK, C. C., VONDERHEIDE, R. H., PITTET, M. J., JAIN, R. K., ZOU, W., HOWCROFT, T. K., WOODHOUSE, E. C., WEINBERG, R. A. & KRUMMEL, M. F. 2018. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat Med*, 24, 541-550.
- BRAY, F., LAVERSANNE, M., SUNG, H., FERLAY, J., SIEGEL, R. L., SOERJOMATARAM, I. & JEMAL, A. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74, 229-263.
- CABLE, D. M., MURRAY, E., ZOU, L. S., GOEVA, A., MACOSKO, E. Z., CHEN, F., IRIZARRY, R. A., CABLE, D. M., MURRAY, E., ZOU, L. S., GOEVA, A., MACOSKO, E. Z., CHEN, F. & IRIZARRY, R. A. 2021. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology* 2021 40:4, 40.
- CASASENT, A. K., SCHALCK, A., GAO, R., SEI, E., LONG, A., PANGBURN, W., CASASENT, T., MERIC-BERNSTAM, F., EDGERTON, M. E. & NAVIN, N. E. 2018. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell*, 172, 205-217 e12.
- CHARI, T. & PACHTER, L. 2023. The specious art of single-cell genomics. *PLoS Comput Biol*, 19, e1011288.

- DE VISSER, K. E. & JOYCE, J. A. 2023. The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. *Cancer Cell*, 41, 374-403.
- DUAN, H., CHENG, T. & CHENG, H. 2023. Spatially resolved transcriptomics: advances and applications. *Blood Sci*, 5, 1-14.
- ELMENTAITE, R., DOMÍNGUEZ CONDE, C., YANG, L. & TEICHMANN, S. A. 2022. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nature Reviews Genetics*, 23, 395-410.
- ELOSUA-BAYES, M., NIETO, P., MEREU, E., GUT, I. & HEYN, H. 2021. SPOTlight seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49.
- FRISCH, J., ANGENENDT, A., HOTH, M., PRATES ROMA, L. & LIS, A. 2019. STIM-Orai Channels and Reactive Oxygen Species in the Tumor Microenvironment. *Cancers (Basel)*, 11.
- GAJEWSKI, T. F., SCHREIBER, H. & FU, Y. X. 2013. Innate and adaptive immune cells in the tumor microenvironment. *Nat Immunol*, 14, 1014-22.
- GENOMICS, X. 2024. What is the spatial resolution and configuration of the capture area of the Visium v1 Gene Expression Slide?
- GRETEN, F. R. & GRIVENNIKOV, S. I. 2019. Inflammation and Cancer: Triggers, Mechanisms, and Consequences. *Immunity*, 51, 27-41.
- HIRATA, E. & SAHAI, E. 2017. Tumor Microenvironment and Differential Responses to Therapy. *Cold Spring Harb Perspect Med*, 7.
- IM, Y. & KIM, Y. 2023. A Comprehensive Overview of RNA Deconvolution Methods and Their Application. *Mol Cells*, 46, 99-105.
- KEREN, L., BOSSE, M., MARQUEZ, D., ANGOSHTARI, R., JAIN, S., VARMA, S., YANG, S. R., KURIAN, A., VAN VALEN, D., WEST, R., BENDALL, S. C. & ANGELO, M. 2018. A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell*, 174, 1373-1387 e19.

- LI, B., ZHANG, W., GUO, C., XU, H., LI, L., FANG, M. & ET AL. 2022. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods*, 19, 662-670.
- LI, H., ZHOU, J., LI, Z., CHEN, S., LIAO, X., ZHANG, B., ZHANG, R., WANG, Y., SUN, S. & GAO, X. 2023. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nature Communications*, 14, 1548.
- MA, Y. & ZHOU, X. 2022. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*, 40, 1349-1359.
- MALAGOLI TAGLIAZUCCHI, G., WIECEK, A. J., WITHNELL, E. & SECRIER, M. 2023. Genomic and microenvironmental heterogeneity shaping epithelial-to-mesenchymal trajectories in cancer. *Nature Communications*, 14, 789.
- OLIVEIRA, M. F., ROMERO, J. P., CHUNG, M., WILLIAMS, S., GOTTSCHO, A. D., GUPTA, A., PILIPAUSKAS, S. E., MOHABBAT, S., RAMAN, N., SUKOVICH, D., PATTERSON, D. & TAYLOR, S. E. B. 2024. Characterization of immune cell populations in the tumor microenvironment of colorectal cancer using high definition spatial profiling. *bioRxiv*, 2024.06.04.597233.
- PASTUSHENKO, I. & BLANPAIN, C. 2019. EMT Transition States during Tumor Progression and Metastasis. *Trends Cell Biol*, 29, 212-226.
- PETITPREZ, F., DE REYNIÈS, A., KEUNG, E. Z., CHEN, T. W.-W., SUN, C.-M., CALDERARO, J., JENG, Y.-M., HSIAO, L.-P., LACROIX, L., BOUGOÛIN, A., MOREIRA, M., LACROIX, G., NATARIO, I., ADAM, J., LUCCHESI, C., LAIZET, Y. H., TOULMONDE, M., BURGESS, M. A., BOLEJACK, V., REINKE, D., WANI, K. M., WANG, W.-L., LAZAR, A. J., ROLAND, C. L., WARGO, J. A., ITALIANO, A., SAUTÈS-FRIDMAN, C., TAWBI, H. A. & FRIDMAN, W. H. 2020a. B cells are associated with survival and immunotherapy response in sarcoma. *Nature*, 577, 556-560.

- PETITPREZ, F., MEYLAN, M., DE REYNIÈS, A., SAUTÈS-FRIDMAN, C. & FRIDMAN, W. H. 2020b. The Tumor Microenvironment in the Response to Immune Checkpoint Blockade Therapies. *Frontiers in Immunology*, 11.
- POLAŃSKI, K., BARTOLOMÉ-CASADO, R., SARROPOULOS, I., XU, C., ENGLAND, N., JAHNSEN, F. L., TEICHMANN, S. A. & YAYON, N. 2024. Bin2cell reconstructs cells from high resolution Visium HD data. *bioRxiv*.
- RIBATTI, D., TAMMA, R. & ANNESE, T. 2020. Epithelial-Mesenchymal Transition in Cancer: A Historical Overview. *Transl Oncol*, 13, 100773.
- ROSENBLATT, M., TEJAVIBULYA, L., JIANG, R., NOBLE, S. & SCHEINOST, D. 2024. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15, 1829.
- SANG-ARAM, C., BROWAEYS, R., SEURINCK, R. & SAEYS, Y. 2024. Spotless, a reproducible pipeline for benchmarking cell type deconvolution in spatial transcriptomics. *eLife*, 12, RP88431.
- SCHÜRCH, C. M., BHATE, S. S., BARLOW, G. L., PHILLIPS, D. J., NOTI, L., ZLOBEC, I., CHU, P., BLACK, S., DEMETER, J., MCILWAIN, D. R., KINOSHITA, S., SAMUSIK, N., GOLTSEV, Y. & NOLAN, G. P. 2020. Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell*, 182, 1341-1359 e19.
- SHI, J., PAN, Y., LIU, X., CAO, W., MU, Y. & ZHU, Q. 2023. Spatial Omics Sequencing Based on Microfluidic Array Chips. *Biosensors*, 13, 712.
- STÅHL, P. L., SALMÉN, F., VICKOVIC, S., LUNDMARK, A., NAVARRO, J. F., MAGNUSSON, J., GIACOMELLO, S., ASP, M., WESTHOLM, J. O., HUSS, M., MOLLBRINK, A., LINNARSSON, S., CODELUPPI, S., BORG, Å., PONTÉN, F., COSTEA, P. I., SAHLÉN, P., MULDER, J., BERGMANN, O., LUNDEBERG, J. & FRISÉN, J. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. United States: © 2016, American Association for the Advancement of Science.

- TANG, X., CHEN, J., ZHANG, X., LIU, X., XIE, Z., WEI, K., QIU, J., MA, W., LIN, C. & KE, R. 2023. Improved in situ sequencing for high-resolution targeted spatial transcriptomic analysis in tissue sections. *J Genet Genomics*. China: © 2023 Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, and Genetics Society of China. Published by Elsevier Ltd.
- TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. & CHU, G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99, 6567-72.
- WEI, Y., HUANG, C.-X., XIAO, X., CHEN, D.-P., SHAN, H., HE, H. & KUANG, D.-M. 2021. B cell heterogeneity, plasticity, and functional diversity in cancer microenvironments. *Oncogene*, 40, 4737-4745.
- WILLIAMS, C. G., LEE, H. J., ASATSUMA, T., VENTO-TORMO, R. & HAQUE, A. 2022. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14, 68.
- WU, L., SAXENA, S. & SINGH, R. K. 2020. Neutrophils in the Tumor Microenvironment. In: BIRBRAIR, A. (ed.) *Tumor Microenvironment: Hematopoietic Cells – Part A*. Cham: Springer International Publishing.
- WU, S. Z., AL-ERYANI, G., RODEN, D. L., JUNANKAR, S., HARVEY, K., ANDERSSON, A., THENNAVAN, A., WANG, C., TORPY, J. R., BARTONICEK, N., WANG, T., LARSSON, L., KACZOROWSKI, D., WEISENFELD, N. I., UYTINGCO, C. R., CHEW, J. G., BENT, Z. W., CHAN, C. L., GNANASAMBANDAPILLAI, V., DUTERTRE, C. A., GLUCH, L., HUI, M. N., BEITH, J., PARKER, A., ROBBINS, E., SEGARA, D., COOPER, C., MAK, C., CHAN, B., WARRIER, S., GINHOUX, F., MILLAR, E., POWELL, J. E., WILLIAMS, S. R., LIU, X. S., O'TOOLE, S., LIM, E., LUNDEBERG, J., PEROU, C. M. & SWARBRICK, A. 2021. A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet*, 53, 1334-1347.
- XIE, B., QIN, J., MORA, A. & LI, X. 2021. Automatic cell type identification methods for single-cell RNA sequencing. *Computational and Structural Biotechnology Journal*, 19, 5874-5887.

- XIE, X., SHI, Q., WU, P., ZHANG, X., KAMBARA, H., SU, J., YU, H., PARK, S. Y., GUO, R., REN, Q., ZHANG, S., XU, Y., SILBERSTEIN, L. E., CHENG, T., MA, F., LI, C. & LUO, H. R. 2020. Single-cell transcriptome profiling reveals neutrophil heterogeneity in homeostasis and infection. *Nat Immunol*, 21, 1119-1133.
- XU, L., SAUNDERS, K., HUANG, S. P., KNUTSDOTTIR, H., MARTINEZ-ALGARIN, K., TERRAZAS, I., CHEN, K., MCARTHUR, H. M., MAUÉS, J., HODGDON, C., REDDY, S. M., ROUSSOS TORRES, E. T. & CHAN, I. S. 2024. A comprehensive single-cell breast tumor atlas defines epithelial and immune heterogeneity and interactions predicting anti-PD-1 therapy response. *Cell Rep Med*, 5, 101511.
- YAN, L. & SUN, X. 2022. Benchmarking and integration of methods for deconvoluting spatial transcriptomic data. *Bioinformatics*, 39, btac805.
- ZACKL, C. 2024. spacedeconv: Unified Interface to spatial transcriptomics deconvolution tools.
- ZEMEK, R. M., DE JONG, E., CHIN, W. L., SCHUSTER, I. S., FEAR, V. S., CASEY, T. H., FORBES, C., DART, S. J., LESLIE, C., ZAITOUNY, A., SMALL, M., BOON, L., FORREST, A. R. R., MUIRI, D. O., DEGLI-ESPOSTI, M. A., MILLWARD, M. J., NOWAK, A. K., LASSMANN, T., BOSCO, A., LAKE, R. A. & LESTERHUIS, W. J. 2019. Sensitization to immune checkpoint blockade through activation of a STAT1/NK axis in the tumor microenvironment. *Sci Transl Med*. United States: © 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works.