Comprehensive Benchmark on Spatial Transcriptomic Deconvolution Methods: SpatialDWLS, CARD, and Cell2Location

2137

CMML3 ICA2 mini-project Deconvolution, SpatialDWLS

Abstract. Spatial transcriptomics reveals transcript distributions but faces cell mixture challenges for many sequencing platforms. Deconvolution tools (Cell2location, CARD, SpatialDWLS) predict cell types from scRNA-seq references. This study benchmarks these tools across datasets and situations, focusing on accuracy and robustness assessment with both real data and simulated data, offering guidance for their application. The benchmark pipeline also provide a comprehensive framework for further evaluations.

Advancements in spatial transcriptomics improve the capacity to detect the spatial distribution of transcripts in tissue slices. Imaging-based methods like MERFISH and Xenium can achieve single-cellular resolution but are often limited in the total number of RNA transcripts detected and labor works (Zhang et al., 2023). Conversely, many widely used sequencing-based methods, such as 10X Visium and Slide-seq, offer near-cellular or even sub-cellular level resolution but contend with cell heterogeneity within each multi-cellular spot (Jonghe et al., 2024). This limitation reveals a challenge for resolving single-cell level transcriptomes with spatial context. To utilize the lower cost, higher throughput, and whole-genome scale of sequencing-based methods while addressing their resolution shortcomings, several bioinformatic tools have been developed. Typically, deconvolution tools use annotated single-cell transcriptomics (scRNA-seq) data as a reference and employ statistical or machine learning algorithms to estimate cellular distribution. Among these tools, Cell2location infers cell type proportions using a Bayesian statistical model based on gene expression signatures from scRNA-seq (Kleshchevnikov et al., 2022); CARD, a spatially informed regression based model, combines non-negative matrix factorization with a spatial statistical model (Ma & Zhou, 2023); and SpatialDWLS, derived from previous DWLS tool for bulk RNA-seq deconvolution (Tsoucas et al., 2019), uses a weighted-least-squares approach to calculate cell type composition (Dong et al., 2021).

The aim of this study is to conduct an impartial and meticulous benchmark of these differently designed deconvolution tools to evaluate their performance of accuracy and robustness (Figure 1a). To achieve this, we evaluated the performance of the three deconvolution methods using seven sequencing-based datasets and seven imaging-based datasets, covering five popular spatial transcriptomic sequencing platforms (details in Supplementary Tables S1, S2, and Supplementary Methods). For sequencing-based datasets, paired scRNA datasets were prepared. Imaging-based datasets were used to simulate datasets with known spot composition ground truth and to generate datasets with variable gene numbers or spot resolutions (Figure 1a). Performance was assessed using metrics like Root Mean Squared Error (RMSE), Jensen-Shannon Divergence (JSD), and Pearson Correlation Coefficient (PCC) for accuracy, alongside robustness tests (Li et al., 2022; Li et al., 2023).

In real datasets, Cell2location demonstrated the highest correlation score in the resolved mean marker gene PCC and generally identified more cell types with higher mixture levels (entropy), although these intrinsic metrics could vary depending on the target slices (Figure 1b). For a series of simulated MERFISH mouse brain slices at a resolution of 0.25 (where the number of binned spots was about 0.25 times the original cell count), Cell2location also performed better in overall accuracy when considering metrics evaluating the similarity between predicted cell proportion and ground truth. CARD appeared slightly better than SpatialDWLS in these simulations (Figure 1c). Interestingly, SpatialDWLS achieved the highest median marker gene PCC in these simulated data, suggesting that while it could accurately capture class-specific gene correlations, it had a comparatively worse ability in resolving cell type proportions (Supplementary Figure S1a). Overall, Cell2location showed the best deconvolution accuracy among the three tools, followed by CARD.

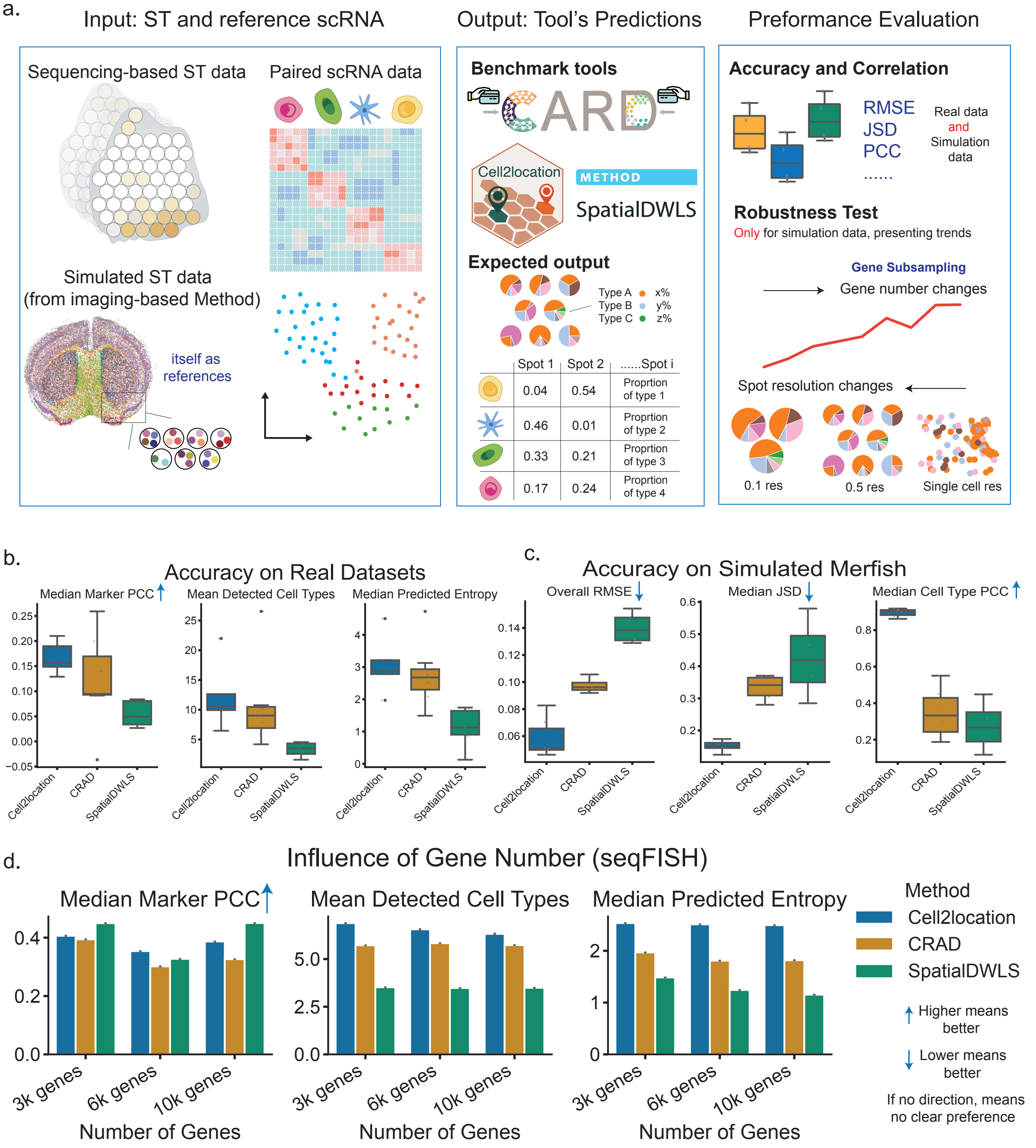


Figure 1. Benchmark workflow and performance evaluation. (a) Benchmark process overview. Inputs were ST and paired reference scRNA, containing real and simulated data; Expected outputs are cell type proportion matrix (b) Accuracy on real datasets (Median Marker PCC, Mean Detected Cell Types, Median Predicted Entropy). (c) Accuracy on simulated Merfish data (Overall RMSE, Median JSD, Median Cell Type PCC). (d) Influence of gene number on seqFISH+ data (Median Marker PCC, Mean Detected Cell Types, Median Predicted Entropy). The downward blue arrow means the higher value was better, and upward arrow means opposite.

Robustness evaluations were conducted across several aspects. In replicate tests using simulated seqFISH+ data, CARD and SpatialDWLS showed stability without any randomness. Cell2loca-tion exhibited slight differences between replicates, but this influence was considered negligible (Supplementary

Figure S1b). Regarding the effect of gene number, when decreased via subsampling, performance indicators fluctuated but remained relatively steady for datasets (Figure 1d). SpatialDWLS appeared more sensitive to changes in gene number compared to the other two. Effect of Spot resolution, a critical factor when comparing sequencing platforms, was also investigated. Using one MERFISH mouse brain slice, nine simulated slices with gradually varying resolutions (from approximately 800 to 4000 spots) were generated. After assess the results from simulated slices with ground truth, Cell2location maintained outstanding accuracy across all resolutions, with CARD also showing excellent and slightly lower scores (Figure 2a). SpatialDWLS, however, performed worst, and its performance showed degrading trend as resolution increased, suggesting it may not be robust to deal with near-cellular level sequencing data. In addition, the predicted cell proportions from all three tools showed similar patterns in regions with purer cell types but differed more in margin sections or areas with highly mixed cell environments (Figure 2b; more visual comparisons in Supplementary Figure S2).

Since some articles reported that SpatialDWLS had great performance in some cases (Li et al., 2024), an inner-tool evaluation was conducted to see if its default settings were suboptimal. Two MERFISH simulated datasets were processed with an clustering step before deconvolution as input as additional information. However, this did not significantly change performances, suggesting the lack of pre-clustering was not the primary reason for its lower scores in these tasks (Figure 2c).

This benchmark provides a comparative analysis of Cell2location, CARD, and SpatialDWLS. Cell2location generally exhibited the highest accuracy and robustness across various conditions, making it a strong candidate for general use. CARD offered a good balance of performance. SpatialDWLS, while effective at capturing marker gene correlations, struggled with overall cell type proportion accuracy, and can be influenced at higher resolutions.

The choice of tool should depend on the specific dataset characteristics, resolution, and analytical goals, with an overall guidance summarized in Figure 2d. Cell2location requiring the highest computational resource, which caused lack of CUDA memory failure in some tests. SpatialDWLS also raised several errors without a clear fix solution, and do not have single clear guidance since it has been integrated into Giotto (Dries et al., 2021). All other aspects should be considered carefully.

However, small data size and non-specific preprocess could be a potential limitation when assessing the quality of current benchmarking, and it was not surprising that the relative ranking results may be different due to the choice of datasets (Li et al., 2023). But still, current principal and workflow of benchmark, could not only be applied to improve deconvolution tools evaluation in the future, can also be generalized into tools in other fields that also lacks of ground truth for evaluation. This study highlights the importance of both cross-tool and robustness evaluations for guiding researchers in selecting appropriate deconvolution methods for their spatial transcriptomics analyses.

References

1. Dong, R. and Yuan, G.-C. (2021) ‘SpatialDWLS: accurate deconvolution of spatial transcriptomic data’, *Genome Biology*, 22(1), p. 145.
2. Dries, R. et al. (2021) ‘Giotto: a toolbox for integrative analysis and visualization of spatial expression data’, *Genome Biology*, 22(1), p. 78.
3. Jonghe, J.D. et al. (2024) ‘scTrends: A living review of commercial single-cell and spatial ’omic technologies’, *Cell Genomics*, 4(12).
4. Kleshchevnikov, V. et al. (2022) ‘Cell2location maps fine-grained cell types in spatial transcriptomics’, *Nature Biotechnology*, 40(5), pp. 661–671.
5. Li, B. et al. (2022) ‘Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution’, *Nature Methods*, 19(6), pp. 662–670.
6. Li, H. et al. (2023) ‘A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics’, *Nature Communications*, 14(1), p. 1548.
7. Ma, Y. and Zhou, X. (2022) ‘Spatially informed cell-type deconvolution for spatial transcriptomics’, *Nature Biotechnology*, 40(9), pp. 1349–1359.
8. Tsoucas, D. et al. (2019) ‘Accurate estimation of cell-type composition from gene expression data’, *Nature Communications*, 10(1), p. 2975.
9. Zhang, M. et al. (2023) ‘Molecularly defined and spatially resolved cell atlas of the whole mouse brain’, *Nature*, 624(7991), pp. 343–354.

Word count Abstract: 63; Main: 997

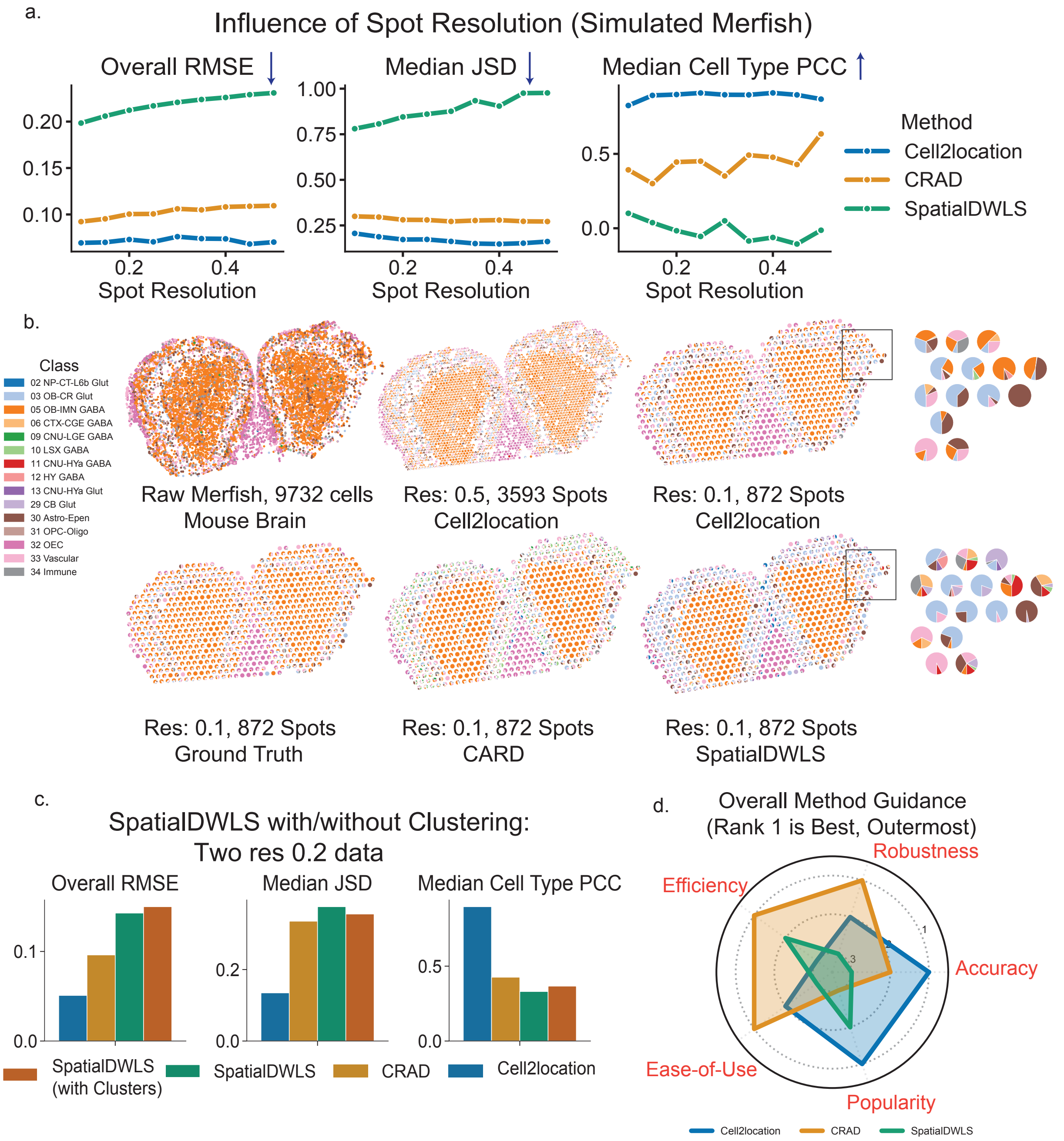


Figure 2. Robustness evaluation and general method guidance. (a) Impact of spot resolution on Overall RMSE, Median JSD, and Median Cell Type PCC. (b) Visual comparison of deconvolution on mouse brain slices at different resolutions, especially resolution 0.1 and 0.5 (c) SpatialDWLS performance with/without clustering (Overall RMSE, Median JSD, Median Cell Type PCC). (d) Overall method guidance radar chart (Robustness, Accuracy, Popularity, Ease-of-Use, Efficiency). Robustness and Accuracy were based on current evaluation. Popularity was based on citation and GitHub Star. Ease-of-use was based on subjective usage smoothness (could be improved if more user could feedback). Efficiency was based on computational resource requirement.

Supplementary Material

Methods

Implementation We benchmarked selected deconvolution tools: Cell2location (v0.1.4; Kleshchevnikov et al., 2022), CARD (v1.1; Ma $ Zhou, 2023), and SpatialDWLS (implementation with Giotto v4.2.1; Dong et al., 2021). Tools were performed using Python (v3.9.21) for Cell2location and R (v4.4.3) for CARD and Giotto, respectively, with recommended environment configuration (Supplementary Figure S3).

Data Preparation Seven sequencing-based datasets and seven imaging-based datasets were collected, covering five popular spatial transcriptomic sequencing platforms, including 10X Visium, Slide-seq, and Stereo-seq (Li et al., 2023; Lopez et al., 2022; details in Supplementary Table S1 and S2). Real datasets were obtained as raw gene expression matrices and underwent initial quality control (QC), filtering genes expressed in <3 cells/spots and cells/spots with <3 expressed genes. Subsequent pre-processing varied by tool as per their recommendations: For SpatialDWLS (via Giotto), both single-cell reference and spatial data were library-size normalized before signature matrix creation and deconvolution. For CARD, QC-filtered raw count matrices were input directly. Cell2location required unnormalized raw counts. All analyses used the intersection of genes common to reference and spatial datasets.

Simulated datasets with known ground truth cell type proportions were generated from imaging-based datasets (MERFISH, seqFISH+). MERFISH mouse brain slices (Zhang et al., 2023) were used to simulate varying spot resolutions by controlling the ratio of spots to real cells (e.g., 0.25 resolution means the number of binned spots is 0.25 times the original cell count). Simulated seqFISH+ data were obtained from Li et al., 2024.

Evaluation Metrics Deconvolution performance was evaluated using the predicted cell type proportion matrix (P) and the ground truth proportion matrix (T). Both are $N \times M$ matrices (N spots, M cell types). Comparisons use commonly aligned spots and cell types. Proportions P_{ij} and T_{ij} are row-normalized (sum to 1 per spot). All these metrics were learned or modified from Li et al., 2022 and Li et al., 2023.

1. **Overall Root Mean Squared Error (RMSE)**: Measures the difference between predicted (P_{ij}) and true (T_{ij}) proportions.

$$\mathrm{RMSE}_{\mathrm{overall}} = \sqrt{\frac{1}{N_c M_c}\sum_{i=1}^{N_c}\sum_{j=1}^{M_c}(P_{ij} - T_{ij})^2} \quad (1)$$

   where N_c, M_c are the numbers of common spots and cell types.
2. **Median Jensen-Shannon Divergence (JSD)**: Measures similarity between predicted (P_i) and true (T_i) proportion distributions for each spot i. Lower values (0 to 1) are better. The JSD for spot i is:

$$\mathrm{JSD}(P_i||T_i) = \frac{1}{2}D_{KL}(P_i||M_i) + \frac{1}{2}D_{KL}(T_i||M_i) \quad (2)$$

   where

$$M_i = \frac{1}{2}(P_i + T_i)$$

   , and

$$D_{KL}(A||B) = \sum_k A_k \log_2(A_k/B_k)$$

   is the Kullback-Leibler divergence. The median JSD across spots is reported.
3. **Median Cell Type Pearson Correlation Coefficient (PCC)**: Assesses correlation between predicted and true proportions for each cell type, across spots. The median of these per-cell-type PCCs should be reported.

Intrinsic prediction metrics (calculated from P only) included:

4. **Median Predicted Entropy**: The median Shannon entropy ($H(P_i) = -\sum_j P_{ij}\log_2 P_{ij}$) across spots. Lower values suggest more focused predictions.
5. **Mean Detected Cell Types**: The average number of cell types per spot with $P_{ij} > 0.01$.

A Marker Gene-based Metric was also used:

6. **Median Marker PCC**: For each cell type, the average PCC between its predicted proportions across spots and the spatial expression of its marker genes. Higher values indicate better consistency.

Code and Data Availability The datasets used and analyzed during the current study are available from the respective publications and public repositories as detailed in Supplementary Tables S1 and S2. The benchmark related results were stored in `https://github.com/AnonymityICAuser/CMML3_ICA2_data`. The code used for this mini-project is available in `https://github.com/AnonymityICAuser/CMML3_ICA2_code`.

Reflection

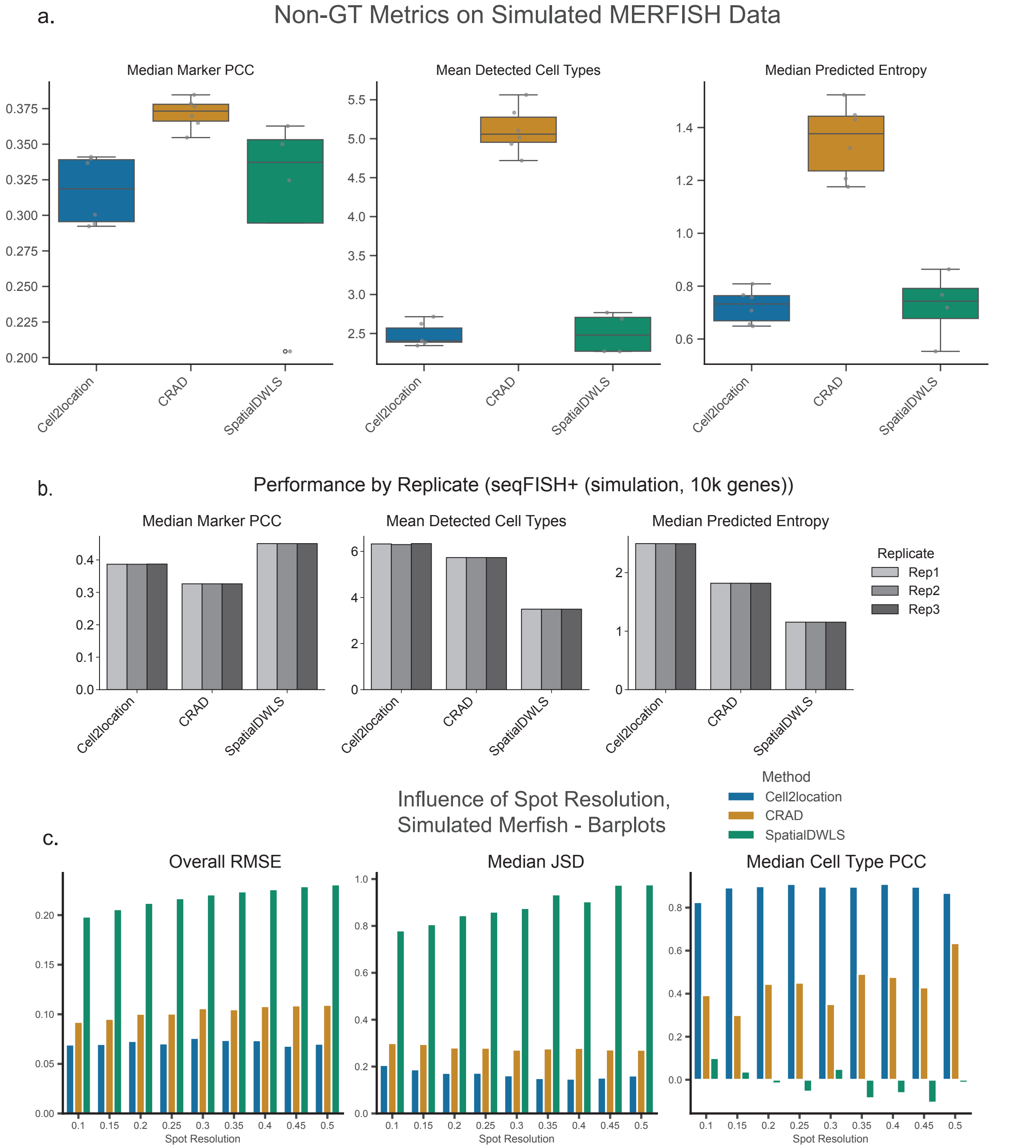
This mini-project provided valuable training in employing advanced deconvolution algorithms relevant to the CMML3 curriculum, covering the background of spatial transcriptomics deconvolution and the methodologies behind these tools. Insights from former benchmark papers with comprehensive comparative strategies and evaluation matrices were instrumental. Previous learning also equipped us with solid Python and R abilities and benchmarking principles. During this project, familiarity with ST and scRNA dataset processing was enhanced, along with experience in handling data from multiple sequencing platforms. The practical application of statistical evaluation metrics and visualization techniques learned in workshops was particularly helpful.

Due to time limitations, only three tools were tested. However, the current benchmarking framework and pipeline can be adapted for a larger number of deconvolution tools. Future work could enhance simulation studies by incorporating different levels of expression noise, altered sequencing depths, and other modifications to test robustness. Furthermore, exploring the integration of these deconvolution results with downstream analyses, such as cell-cell communication or differential expression within predicted cell types, would be a valuable extension, applying advanced machine learning concepts from the course. This project also provides a foundation for conducting benchmark tasks in other bioinformatics fields using similar principles.

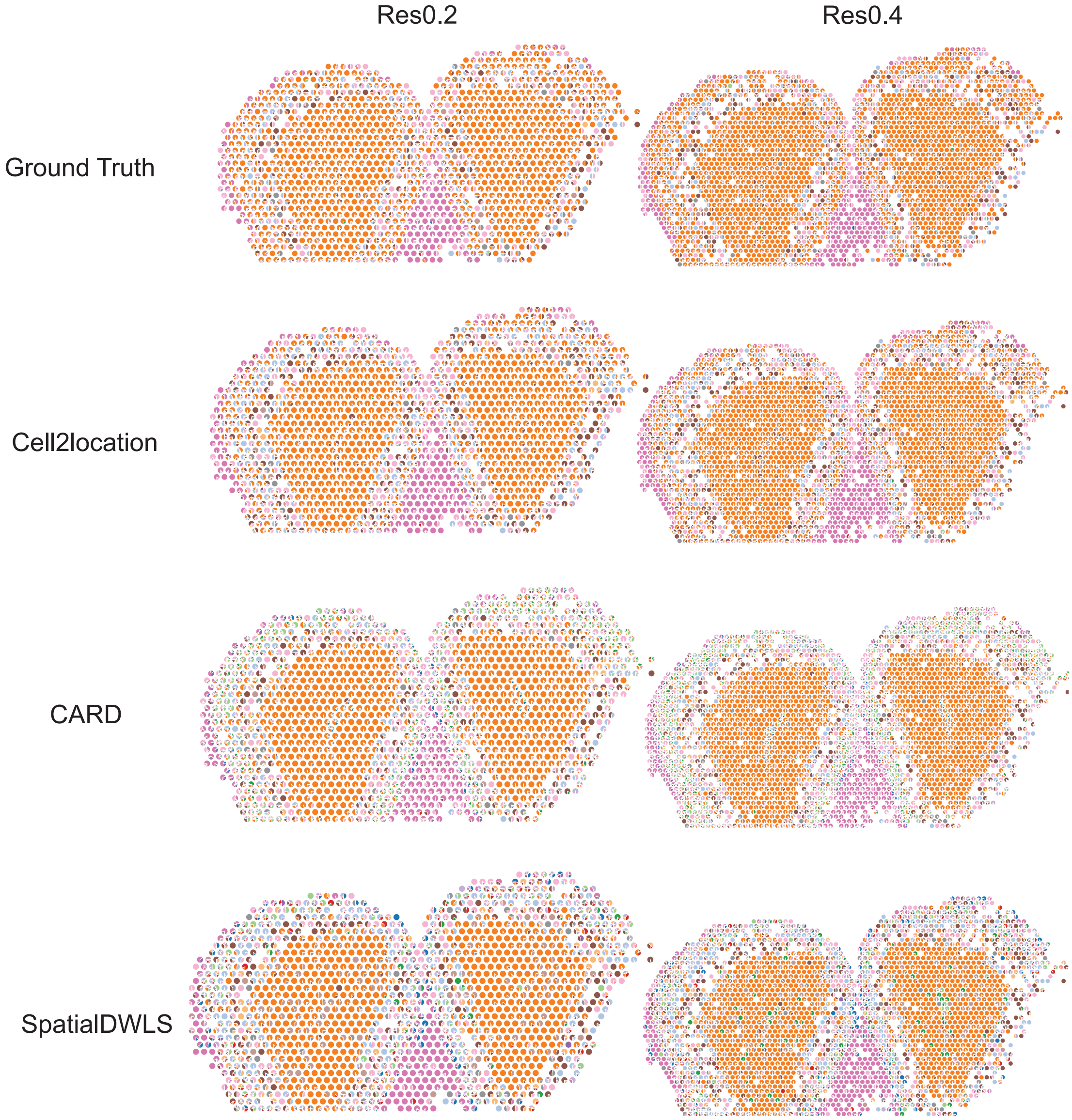
References

1. Dong, R. et al. (2021) 'SpatialDWLS: accurate deconvolution of spatial transcriptomic data', *Genome Biology*, 22(1), p. 145.
2. Kleshchevnikov, V. et al. (2022) 'Cell2location maps fine-grained cell types in spatial transcriptomics', *Nature Biotechnology*, 40(5), pp. 661–671.
3. Li, B. et al. (2022) 'Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution', *Nature Methods*, 19(6), pp. 662–670.
4. Li, H. et al. (2023) 'A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics', *Nature Communications*, 14(1), p. 1548.
5. Lopez, R. et al. (2022) 'DestVI identifies continuums of cell types in spatial transcriptomics data', *Nature Biotechnology*, 40(9), pp. 1360–1369.
6. Ma, Y. et al. (2022) 'Spatially informed cell-type deconvolution for spatial transcriptomics', *Nature Biotechnology*, 40(9), pp. 1349–1359.
7. Zhang, M. et al. (2023) 'Molecularly defined and spatially resolved cell atlas of the whole mouse brain', *Nature*, 624(7991), pp. 343–354.

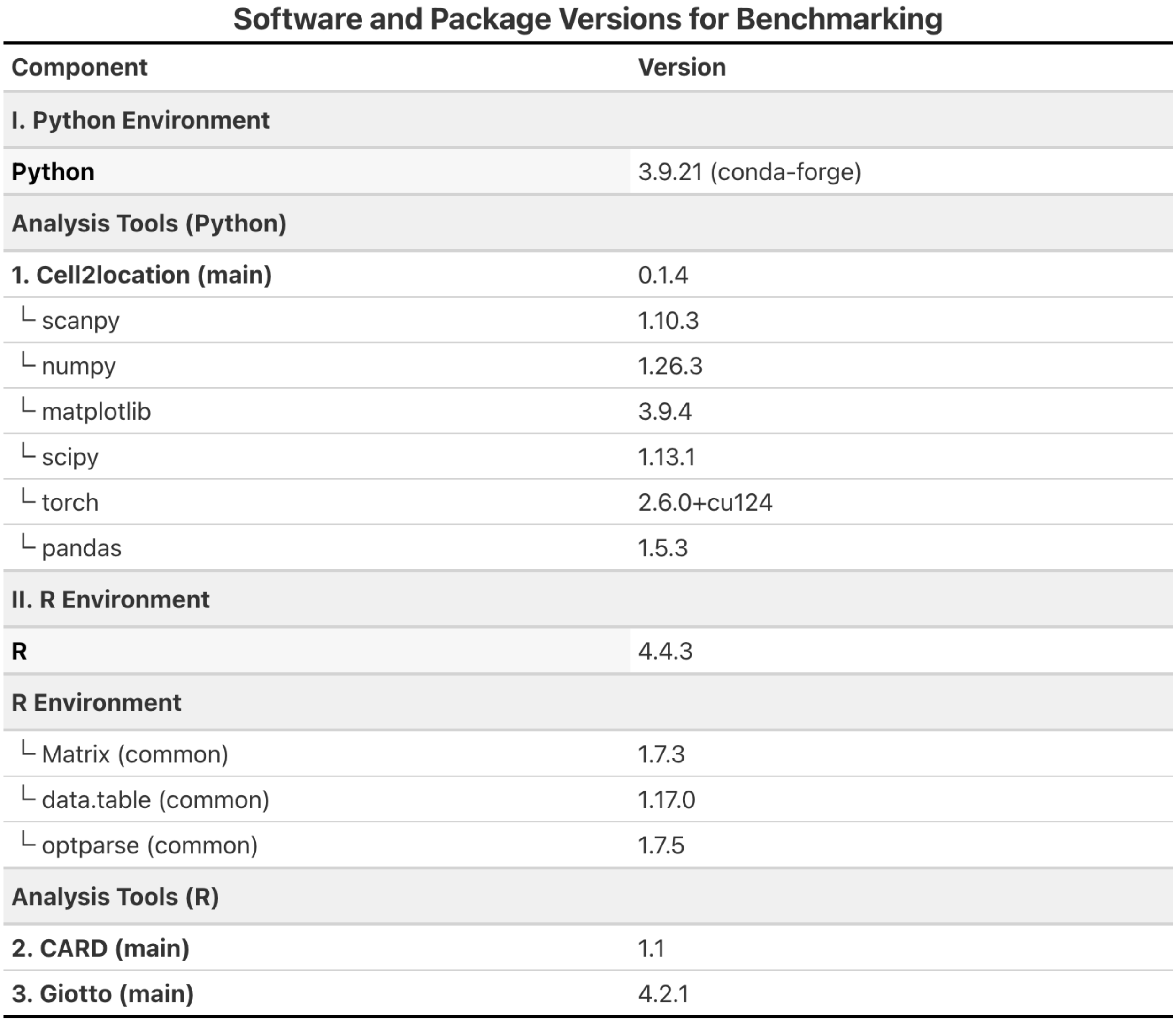
Word count Methods: 496; Reflection: 194



Supplementary Figure S 1. Additional performance metrics. (a) Non-Ground Truth (intrinsic) metrics on simulated MERFISH data. (b) Performance by replicate for seqFISH+ simulated data (10k genes). (c) Bar plots showing influence of spot resolution on simulated Merfish data.



Supplementary Figure S 2. Visual comparison of deconvolution results from Cell2location, CARD, and SpatialDWLS against Ground Truth on simulated Merfish data at Res0.2 and Res0.4.



Software and Package Versions for Benchmarking

Component	Version
I. Python Environment	
Python	3.9.21 (conda-forge)
Analysis Tools (Python)	
1. Cell2location (main)	0.1.4
└ scanpy	1.10.3
└ numpy	1.26.3
└ matplotlib	3.9.4
└ scipy	1.13.1
└ torch	2.6.0+cu124
└ pandas	1.5.3
II. R Environment	
R	4.4.3
R Environment	
└ Matrix (common)	1.7.3
└ data.table (common)	1.17.0
└ optparse (common)	1.7.5
Analysis Tools (R)	
2. CARD (main)	1.1
3. Giotto (main)	4.2.1

Supplementary Figure S 3. Software and package versions used for benchmarking.

Table 1. Supplementary Table S1, Overview of Real Sequencing-Based Datasets used for Benchmarking.

Dataset ID	Species	Tissue	ST Technology	scRNA Tech.	Download Link / Source
MCA205 Tumor	Mouse	MCA205 Tumor	10x Visium	10X Chromium	DestVI Reprod. (GitHub)
MOp Visium	Mouse	Primary Motor Cortex (MOp)	10x Visium	10X Chromium	Tangram Demo (ST) & Google Cloud (scRNA)
MOp Slide-seq	Mouse	Primary Motor Cortex (MOp)	Slide-seq	10X Chromium	Tangram Demo (ST) & Google Cloud (scRNA)
Human Lymph Node	Human	Lymph Node	10x Visium	10X Chromium	Cell2location Tutorial
Zebrafish Embryo	Zebrafish	Embryo	Stereo-seq	Stereo-seq matched	Zenodo Rec. 7674290
Mouse Brain Stereo	Mouse	Brain	Stereo-seq	Matched scRNA	Zenodo Rec. 7674290
Mouse Brain Visium	Mouse	Brain	10x Visium	Matched scRNA	Zenodo Rec. 7674290

Table 2. Supplementary Table S2, Overview of Simulated and Imaging-Based Datasets used for Benchmarking.

Dataset Origin	Basis / Type	Species	Simulation Aspect	Download Link
MERFISH	Real Imaging Data (used for simulation)	Mouse	6 independent slices, 07-12	Allen Inst. ABC Atlas
MERFISH	Real Imaging Data (used for simulation)	Mouse	Multiple resolutions, 07 (spot sizes simulated)	Allen Inst. ABC Atlas
seqFISH+ (Sim.)	Simulated (imaging-inspired)	Mouse	Multiple gene numbers	Zenodo Rec. 7674290 (Sim. by Li et al., 2023)