Spotless: a reproducible pipeline for benchmarking cell type deconvolution in spatial transcriptomics

Chananchida Sang-aram[1,2], Robin Browaeys[1,2], Ruth Seurinck[1,2], Yvan Saeys[1,2]

Abstract

Spatial transcriptomics (ST) is an emerging field that aims to profile the transcriptome of a cell while keeping its spatial context. Although the resolution of non-targeted ST technologies has been rapidly improving in recent years, most commercial methods do not yet operate at single-cell resolution. To tackle this issue, computational methods such as deconvolution can be used to infer cell type proportions in each spot by learning cell type-specific expression profiles from reference single-cell RNA-sequencing (scRNA-seq) data. Here, we benchmarked the performance of 11 deconvolution methods using 54 silver standards, 3 gold standards, and one in-depth case study on the liver. The silver standards were generated using our novel simulation engine *synthspot*, where we used six scRNA-seq datasets to create synthetic spots that followed one of nine different biological tissue patterns. The gold standards were generated using imaging-based ST technologies at single-cell resolution. We evaluated method performance based on the root-mean-squared error, area under the precision-recall curve, and Jensen-Shannon divergence. Our evaluation revealed that method performance significantly decreases in datasets with highly abundant or rare cell types. Moreover, we evaluated the stability of each method when using different reference datasets and found that having sufficient number of genes for each cell type is crucial for good performance. We conclude that while RCTD and cell2location are the top-performing methods, a simple off-the-shelf deconvolution method surprisingly outperforms almost half of the dedicated spatial deconvolution methods. Our freely available Nextflow pipeline allows users to generate synthetic data, run deconvolution methods and optionally benchmark them on their dataset (https://github.com/saeyslab/spotless-benchmark).

Introduction

Unlike single-cell sequencing, spatial transcriptome profiling technologies can uncover the location of cells, adding another dimension to the data that is essential for studying systems biology, e.g., cell-cell interactions and tissue architecture [1]. These approaches can be categorized into imaging- or sequencing-based methods, each of them offering complementary advantages [2]. As imaging-based methods use labeled hybridization probes to target specific genes, they offer subcellular resolution and high capture sensitivity, but are limited to a few hundreds or thousands of genes [3], [4]. On the other hand, sequencing- based methods offer an unbiased and transcriptome-wide coverage by using capture oligonucleotides with a polydT sequence [5], [6]. These oligos are printed in clusters, or *spots*, each with a location-specific barcode that allows identification of the originating location of a transcript. While the size of these spots initially started at 100µm in diameter, some recent sequencing technologies have managed to reduced their size to the subcellular level, thus closing the resolution gap with imaging technologies [7], [8]. However, these spots do not necessarily correspond to individual cells, and therefore, computational methods remain necessary to determine the cell-type composition of each spot.

Deconvolution and mapping are two types of cell type composition inference tools that can be used to disentangle populations from a mixed gene expression profile. In conjunction with the spatial

[1] Data mining and Modelling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium

[2] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

dataset, a reference scRNA-seq dataset from an atlas or matched sequencing experiment is typically required in order to build cell type-specific gene signatures. Deconvolution infers the proportions of cell types in a spot by utilizing a regression or probabilistic framework, and methods specifically designed for ST data often incorporate additional model parameters to account for spot-to-spot variability. On the other hand, mapping approaches score a spot for how strongly its expression profile corresponds to those of specific cell types [9]. As such, deconvolution returns the *proportion* of cell types per spot, and mapping returns the *probability* of cell types belonging to a spot. Unless otherwise specified, we use the term "deconvolution" to refer to both deconvolution and mapping algorithms in this study.

Although there have been previous benchmarking studies by Li et al. (2022) and Yan and Sun (2023), several questions remain unanswered [10], [11]. First, the added value of algorithms specifically developed for the deconvolution of ST data has not been evaluated by comparing them to a baseline or bulk deconvolution method. Second, it is unclear which algorithms are better equipped to handle challenging scenarios, such as the presence of a highly abundant cell type throughout the tissue or the detection of a rare cell type in a single region of interest. Finally, the stability of these methods to variations in the reference dataset arising from changes in technology or protocols has not been assessed.

In this study, we aim to address these gaps in knowledge and provide a comprehensive evaluation of 11 deconvolution methods in terms of performance, stability, and scalability (**Figure 1**). The tools include eight spatial deconvolution methods (cell2location[12], DestVI[13], DSTG[14], RCTD[15], SpatialDWLS[16], SPOTlight[17], stereoscope[18], and STRIDE[19]), one bulk deconvolution method (MuSiC [20]), and two mapping methods (Seurat[21] and Tangram[22]). For all methods compared, we discussed with the original authors, in order to ensure that their method was run in an appropriate setting and with good parameter values. We also compare method performance with two baselines, a “null distribution” based on random proportions from a Dirichlet distribution, and predictions from the non-negative least squares (NNLS) algorithm. We evaluated method performance on a total of 57 synthetic datasets (54 silver standards and 3 gold standards) and 1 application dataset. Our benchmarking pipeline is completely reproducible and accessible through Nextflow (github.com/saeyslab/spotless-benchmark). Furthermore, each method is implemented inside a Docker container, which enables users to run the tools without requiring prior installation.

Results

Synthspot allows simulation of artificial tissue patterns

Synthetic spatial datasets are commonly generated by developers of deconvolution methods as part of benchmarking their algorithms against others. However, these synthetic datasets typically have spots with random compositions that do not reflect the reality of tissue regions with distinct compositions, such as layers in the brain. To overcome this limitation, we developed our own simulator called *synthspot* that can generate synthetic tissue patterns with distinct regions, allowing for more realistic simulations (github.com/saeyslab/synthspot). We validated that our simulation procedure accurately models real data characteristics and that method performances are comparable between synthetic and real tissue patterns (**Supplementary Note 1**).

Within a synthetic dataset, *synthspot* creates artificial regions in which all spots belonging to the same region have the same prior frequencies. The prior frequencies correspond to the likelihood of sampling cells from a cell type, so spots within the same region will have similar compositions because they have the same prior frequencies. The priors are defined by the selected artificial tissue pattern, or

abundance pattern, which determine the uniformity, distinctness, and rarity of cell types within and across regions (**Figure 1a, Figure S1**). For example, the *uniform* characteristic will sample the same number of cells for all cell types in a spot, while *diverse* samples differing number of cells. The *distinct* characteristic constrains a cell type to only be present in one region, while *overlap* allows it to be present in multiple regions. Additionally, the *dominant cell type* characteristic randomly assigns a dominant cell type that is at least 5-15 times more abundant than others in each spot, while *rare cell type* does the opposite to create a cell type that is 5-15 times less abundant. The different characteristics can be combined in up to nine different abundance patterns, each representing a plausible biological scenario.

Cell2location and RCTD perform well in synthetic data

Our synthetic spatial datasets consist of 54 silver standards generated from synthspot and 3 gold standards generated from imaging data with single-cell resolution (**Supplementary Table 2a-b**). For the silver standards, we used 9 abundance patterns with 6 publicly available scRNA-seq datasets, four of which were from mouse brain tissues (cortex, hippocampus, and two from cerebellum), one from mouse kidney, and one from human skin cancer. We generated 10 replicates for each of the 54 combinations, with each replicate containing around 750 spots (**Figure S2**). Half of the cells from each scRNA-seq dataset were used to generate the synthetic spots and the other half used as the reference for deconvolution. For the gold standard, we used two seqFISH+ sections of mouse brain cortex and olfactory bulb (63 spots x 10,000 genes each) and one STARMap section of mouse primary visual cortex (108 spots x 996 genes). We summed up counts from cells within circles of 55-µm diameter, which are the size of spots in the 10x Visium commercial platform.

We evaluated method performance with the root-mean-square error (RMSE), area under the precision-recall curve (AUPR), and Jensen-Shannon divergence (JSD) (**Supplementary Note 2**). The RMSE measures how numerically accurate the predicted proportions are, the AUPR measures how well a method can detect whether a cell type is present or absent, and the JSD measure similarities between two probability distributions.

For the silver standards, RCTD and cell2location were the top two performers across all metrics, followed by SpatialDWLS, stereoscope, and MuSiC (**Figure 2b, Figure 3a**). All other methods ranked worse than NNLS in at least two metrics. We ranked the methods for each of the 54 scenarios based on the median value across 10 replicates and summed the ranks across all scenarios to determine the overall ranking. Method performances were more consistent between abundance patterns than between datasets (**Figure S4-S6**). Most methods had worse performance in the two abundance patterns with a dominant cell type, and there was considerable performance variability between replicates due to different dominant cell types being selected in each replicate. Only RCTD and cell2location consistently outperformed NNLS in all metrics for these patterns (**Figure S7**).

For the gold standards, cell2location, MuSiC, and RCTD are the top three performers as well as the only methods to outrank NNLS in all three metrics (**Figure 3b**). As each seqFISH+ dataset consisted of seven field of views (FOVs), we used the average across FOVs as the representative value to be ranked for that dataset. Several FOVs were dominated by one cell type (**Figure S3**), which was similar to the dominant cell type abundance pattern in our silver standard. Consistent with the silver standard results, half of the methods did not perform well in these FOVs. In particular, SPOTlight, DestVI, stereoscope, and Tangram tended to predict less variation between cell type abundances. DSTG, SpatialDWLS, and Seurat predicted the dominant cell type in some FOVs, but did not predict the remaining cell type compositions accurately. Most methods performed worst in the STARMap dataset except for SpatialDWLS, SPOTlight, and Tangram.

Detecting rare cell types remains challenging even for top-performing methods

Lowly abundant cell types often play an important role in development or disease progression, as in the case of stem cells and progenitor cells or circulating tumor cells [23]. As the occurrence of these cell types are often used to create prognostic models of patient outcomes, accurate detection of rare cell types is a key aspect of deconvolution tools [24], [25]. Here, we focus on the two *rare cell type* patterns in our silver standard (*rare cell type diverse* and *regional rare cell type diverse*), in which a rare cell type is 5-15 times less abundant than other cell types in either all or one synthetic region (**Figure S1g-h**). We evaluated methods based on the AUPR of the rare cell type only, in contrast to using two other metrics and considering all cell types together as done previously. This is because in prognostic models, the presence of rare cell types is often of more importance than the magnitude of the abundance itself. As such, it is more relevant that methods are able to correctly rank spots with and without the rare cell type.

We summed method rankings across the 12 combinations (6 scRNA-seq datasets x 2 rare cell type patterns), using the average AUPR of the replicates as the representative value for each combination. In line with our previous analysis, RCTD and cell2location were the best at predicting the presence of lowly abundant cell types (**Figure 2b, Figure 4a**). The AUPR is substantially lower if the rare cell type is only present in one region and not across the entire tissue, indicating that prevalence is also an important factor in detection. Nonetheless, the link between abundance and AUPR can be clearly observed, as seen by the PR curves of three cell types from the brain cortex dataset-regionally rare abundance pattern (**Figure 4b**). While most methods can detect cell types with moderate or high abundance, they usually have lower sensitivity for rare cell types, and hence lower AUPRs. To corroborate this, we visually inspected PR curves in each dataset-abundance pattern combination based on average cell type abundances and found similar patterns (**Figure S8**).

Technical variability between reference and spatial datasets can significantly impact predictions

Since deconvolution predictions exclusively rely on signatures learned from the scRNA-seq reference, it should come as no surprise that the choice of the reference dataset has been shown to have the largest impact on bulk deconvolution predictions [26]. Hence, deconvolution methods should ideally also account for platform effects, i.e., the capture biases that may occur from differing protocols and technologies being used to generate scRNA-seq and ST data.

To evaluate the stability of each method against reference datasets from different technological platforms, we devised the *inter*-dataset scenario, where we provided an alternative reference dataset to be used for deconvolution, in contrast to the *intra*-dataset analysis done previously, where the same reference dataset was used for both spot generation and deconvolution. We tested this on the brain cortex (SMART-seq) and two cerebellum (Drop-seq and 10x Chromium) silver standards. For the brain cortex, we used the 10x Chromium reference from the same atlas [27], and we simply swapped the references for the cerebellum datasets. To measure stability, we computed the JSD between the proportions predicted from the intra- and inter-dataset scenario.

Except for MuSiC, we see that methods with better performance – cell2location, RCTD, SpatialDWLS, and stereoscope – were also more stable against changing reference datasets (**Figure 2b, Figure 5**). Out of these four, only SpatialDWLS did not account for a platform effect in its model definition. Cell2location had the most stable predictions, ranking first in all three datasets, while both NNLS and MuSiC were consistently in the bottom three. For the rest of the methods, stability varied between datasets. DestVI performed well in the cerebellum datasets but not the brain cortex, and SPOTlight had the opposite problem. Stereoscope, RCTD, and SpatialDWLS always ranked in the top half, and

DSTG the bottom half. As expected, deconvolution performance generally worsens in the inter-dataset scenario (**Figure S10**).

Evaluation of methods on liver Visium datasets validate results on synthetic data

Although synthetic datasets are useful, it is also important to validate deconvolution methods on real sequencing-based ST data, as they may have different characteristics. For this purpose, we used four mouse liver Visium slides from the Guilliams et al. liver atlas [28]. The liver is a particularly interesting case study due to its tissue pattern, which is highly populated with hepatocytes (more than 60% abundant). This allows us to compare method performance with those of the *dominant cell type* abundance pattern from the silver standard, which was challenging for most methods. Furthermore, the liver atlas is also ideal for evaluating method stability, as it was generated using cells from three different experimental protocols: scRNA-seq following *ex vivo* digestion, scRNA-seq following *in vivo* liver perfusion, and single-nucleus RNA-seq (snRNA-seq) on frozen liver (**Figure S11**).

We assessed method performance using AUPR and JSD by leveraging prior knowledge of the *localization* and *composition* of cell types in the liver. Although the true composition of each spot is not known, we can assume the presence of certain cell types in certain locations in the tissue due to the zonated nature of the liver. Specifically, we calculated the AUPR of portal vein and central vein endothelial cells (ECs) under the assumption that they are present only in their respective locations. Next, we calculated the JSD of the expected and predicted cell type proportions in a liver sample. The expected cell type proportions were based on snRNA-seq data, which has been shown to best recapitulate the true cell compositions observed by confocal microscopy [28]. We ran each method on a combined reference containing all three experimental protocols (*ex vivo* scRNA-seq, *in vivo* scRNA-seq, and snRNA-seq), as well as on each protocol separately. To ensure consistency, we filtered each reference dataset to only include the nine cell types that were common to all three protocols.

We found that RCTD and cell2location were the best performers for both AUPR and JSD (**Figure 6**), achieving a JSD similar to the average pairwise JSD between four snRNA-seq samples. However, the JSD of Tangram, DSTG, SPOTlight, and stereoscope were worse than NNLS. Except for SPOTlight, these three methods were not able to accurately infer the overabundance of the dominant cell type, with stereoscope and Tangram predicting a uniform distribution of cell types (**Figure S12**). This corresponds to what we observed in the *dominant cell type* pattern of the silver standard and certain FOVs in the gold standard. However, there were some inconsistencies between the AUPR and JSD rankings, specifically with Seurat and SpatialDWLS ranking highly for JSD and lowly for AUPR. This is because the JSD is heavily influenced by the dominant cell type in the dataset, such that even when predicting only the dominant cell type per spot, Seurat still performed well in terms of JSD. However, it was unable to predict the presence of portal and central vein ECs in their respective regions (**Figure S13**). Therefore, both the AUPR and JSD must be considered when evaluating methods.

We observed that using the combination of the three experimental protocols as the reference dataset did not necessarily result in the best performance, and it was often possible to achieve better or similar results by using a reference dataset from a single experimental protocol. The best reference varied between methods, and most methods did not exhibit consistent performance across all references. Interestingly, Cell2location, MuSiC, and NNLS had much higher JSD when using the snRNA-seq data as the reference, while RCTD and Seurat had the lowest JSD using this reference. To further evaluate the stability of the methods, we calculated the JSD between proportions predicted with different reference datasets. RCTD and Seurat showed the lowest JSD, indicating higher stability (**Figure S14**). Finally, we examined the predicted proportions when using the entire atlas without filtering, which contains all three protocols and 23 cell types instead of the 9 common cell types as before (**Figure S12**).

The additional 14 cell types made up around 20% of the ground truth proportions. While RCTD, Seurat, SpatialDWLS, and MuSiC retained the relative proportions of the 9 common cell types, the rest predicted substantially different cell compositions.

Runtime and scalability

Most methods were able to deconvolve the datasets in less than 30 minutes and had only slight variability in the runtime (**Figure 7a**). Model-based methods like DestVI, stereoscope, cell2location, and STRIDE had the advantage that their model can be reused for synthetic datasets that were generated from the same single-cell dataset. The model fitting step is then constant for all synthetic datasets, as each one contains roughly 1000 spots. Often these model-based methods can also make use of GPU acceleration (e.g. through a PyTorch implementation), which we applied when available. We do not recommend running them on a CPU, as the runtime can increase by more than tenfold.

Next, we investigated the scalability of the methods by varying the dimensions of the spatial dataset (**Figure 7b**). We tested 16 combinations of spots (100, 1k, 5k, 10k) and genes (5k, 10k, 20k, 30k), while the single-cell dataset was kept at 5k cells and 30k genes. Here, we considered both the model building and fitting step in the runtime (if applicable). Tangram, Seurat, and SPOTlight had a constant runtime across all dimensions, each deconvolving the largest dataset (3×10^8 elements in total) in less than ten minutes. Other methods had increased runtime both when spots and genes were increased, except for DestVI who was only affected by the number of spots, and STRIDE by the number of genes. This was because DestVI only used 2,000 highly variable genes for deconvolution. Similarly, the scalability of RCTD, SpatialDWLS, Tangram, and SPOTlight are because they only make use of differentially expressed or cell type-specific marker genes. Stereoscope was the least scalable by far, taking over 8 hours for the largest dataset, and 3.5 hours for the smallest one. Note that many methods allow for parameters that can greatly affect runtime, such as the number of training epochs and/or the number of genes to use (**Supplementary Table 1**). For instance, here we have used all genes to run stereoscope (default parameters), but the authors have suggested that it is possible to use only 5000 genes in the model and maintain the same performance.

Discussion

In this study, we performed a thorough investigation of various biologically relevant and previously unconsidered aspects related to deconvolution. We evaluated the performance of 11 deconvolution methods on 54 silver standards, 3 gold standards, and 1 Visium dataset using 3 complementary metrics. In addition, we incorporated two baselines, the NNLS algorithm and null distribution proportions. Our findings indicate that RCTD, cell2location, and SpatialDWLS were the highest ranked methods in terms of performance, consistent with previous studies [10], [11]. However, we also found that over half of the methods did not outperform the baseline (NNLS) and bulk deconvolution method (MuSiC). These results were consistent across the silver and gold standards, as well as the liver case study, demonstrating the generalizability and applicability of our simulation and benchmarking framework (**Figure 2**). We also found that the abundance pattern of the tissue and the reference dataset used had the most significant impact on method performance. Even top-performing methods struggled with challenging abundance patterns, such as when a rare cell type was present in only one region, or when a highly dominant cell type masks the signature of less abundant ones. Furthermore, different reference datasets could result in substantially different predicted proportions. Methods that accounted for technical variability in their models, such as cell2location and RCTD, were more stable to changes in the reference dataset than those that did not, such as SpatialDWLS.

Regarding the reference dataset, the number of genes per cell type (which is generally correlated to the sequencing depth) seems to have a significant impact on the performance of deconvolution

methods. We observed that methods were less accurate in datasets with fewer genes per cell type (**Figure S9**). For example, all methods performed better in the single-nucleus cerebellum silver standard dataset, which had the same number of cell types as the single-cell cerebellum dataset, but on average 1,000 more genes per cell type. The kidney dataset was the most challenging for all methods, with most of its 16 cell types having less than 1,000 features. This was evident from the RMSE and JSD scores that were relatively close to the null distribution (**Figure S4**). In contrast, the 18 cell types in the brain cortex dataset had an average of 3,000-9,000 features, leading to much better performance for most methods compared to the kidney dataset despite having more cell types. This trend was also observed in the STARMap gold standard dataset, which consisted of only 996 genes. Most methods performed worse in the STARMap dataset except for SpatialDWLS, SPOTlight, and Tangram. Since these three methods only use marker genes for deconvolution, this may explain why the small number of genes (most of which are already marker genes) did not affect them as much.

In addition to performance, the runtime and scalability of a method is also a factor to consider. Although most methods' runtimes were comparable on our silver standards, we have made use of GPU acceleration for Tangram, DestVI, stereoscope, and cell2location. As this might not be available for some users, using these methods on the CPU might require training on fewer epochs or selecting only a subset of genes. With all these factors in consideration, we recommend RCTD as a good starting point. In addition to being one of the best and fastest methods, it also allows CPU parallelization (i.e., using multiple cores) for users that may not have access to a GPU.

As a general guideline, we recommend comparing the result of multiple deconvolution methods, especially between cell2location and RCTD. If the predictions are highly contradictory, the reference dataset may not be of sufficiently good quality. We recommend obtaining scRNA-seq and spatial data from the same sample to reduce biological variability, as there will always be technical variability across platforms. This can also ensure that the same cell types will be present in both datasets [9]. If that is not possible, use a reference dataset that has sufficient sequencing depth (at least more than 1000 genes per cell type), preferably from a single platform. In addition to checking the dataset's sequencing depth, one can also evaluate the reference dataset quality using our simulation engine *synthspot.* As in our silver standards, users can select the abundance pattern most resembling the real tissue to generate the synthetic spatial dataset. For a good reference, both cell2location and RCTD should nearly have an AUPR of one and JSD of zero. We have integrated this entire workflow (simulating synthetic data, executing deconvolution methods, and evaluating the predictions) in our Nextflow pipeline as well.

As spatial omics is still an emerging field, the development of new deconvolution methods can be anticipated in the future. Our benchmark provides a reliable and reproducible evaluation framework for current and upcoming deconvolution tools, making it a valuable resource for researchers in the field.

Acknowledgements

We would like to thank all the authors of the methods who provided valuable feedback on how to optimally run their algorithms. Their input has been crucial in enabling us to compare the methods in a fair manner. This study was funded by BOF (Ghent University, C.S.) and The Research Foundation - Flanders (1181318N, R.B.).

Methods

Synthspot. We generated synthetic spatial datasets using the *synthspot* R package (github.com/saeyslab/synthspot), whose spot generation process is modeled after that of SPOTlight [17]. This simulator generates synthetic spot data by considering the gene expression of one spot as a mixture of the expression of n cells (n between 2 and 10, randomly sampled based on a uniform distribution). To generate one spot, the simulator samples n cells from the input scRNA-seq data, sums their counts gene by gene, and downsamples the counts. The target number of counts for downsampling is picked from a normal distribution with mean and standard deviation of 20000 ± 5000 by default, but these values can be changed by the user. To mimic biological tissue, synthspot generates artificial regions, or groups of spots with similar cell type compositions (**Figure S2**). The prior distribution of each cell type in each region is influenced by the selected *abundance pattern*, called *dataset type* in the package (**Supplementary Note 1**). We used 9 out of the 17 possible abundance patterns. For each single-cell dataset, we generated ten replicates of each abundance pattern. For each replicate, we ran *generate_synthetic_visium* with five regions (*n_regions*), with minimally 100 spots and maximally 200 spots per region (*n_spots_min*, *n_spots_max)*, and target counts of 20000 ± 5000 per spot (*visium_mean,* and *visium_sd*). There were on average 750 spots per replicate.

Methods execution. An overview of the 11 methods can be found in **Supplementary Table 1**. All methods require a reference scRNA-seq dataset with cell type annotations along with the spatial dataset as inputs. We ran the methods based on the tutorials provided, and tested different parameters whenever possible and used the best results (**Supplementary Note 3**). We contacted the authors for recommendations on how to run the methods. Cell2location, DestVI, stereoscope, and STRIDE were run in two steps, first training a model on the reference scRNA-seq data, and second predicting the cell type proportions using the model. Hence, a model trained on a scRNA-seq dataset can be reused for all abundance patterns of that dataset (9 patterns x 10 replicates).

For reproducibility, each method was installed inside a Docker container and can be run either using Docker or Singularity (Apptainer). We implemented a Nextflow pipeline to run the methods concurrently and compute their performance [29]. Our pipeline can be run in three modes 1) *run_standard* which is used to reproduce our analysis, 2) *run_dataset* which users can run the pipeline on their own scRNA-seq and spatial datasets, and 3) *generate_and_run* which takes a scRNA-seq dataset, generates synthetic datasets from it, and runs the pipeline. Furthermore, our pipeline can accept the single-cell and spatial inputs as a Seurat (.rds) or AnnData (.h5ad) object in *any* combination, and the proportions are output as a tab-separated file. Parameter tuning is also implemented. All analysis scripts and parameter configurations can be found at https://github.com/saeyslab/spotless-benchmark.

All methods were run on a high-performance computing cluster (HPC) with a SLURM scheduler. We made use of the GPU cluster (NVIDIA Volta V100) whenever possible. We started with 1 core and 8GB memory, dynamically increasing the memory by 8GB if the process would fail.

Datasets. We used 3 and 54 datasets for the gold and silver standards, respectively. They can be downloaded on Zenodo (https://zenodo.org/record/5763377).

Gold. The seqFISH+ dataset consisted of two mouse brain tissue slices (cortex and olfactory bulb) with 7 field of views (FOVs) per slice. Ten thousand genes were profiled for all FOVs. Each FOV had a dimension of 206 µm x 206 µm and a resolution of 103 nm per pixel. We simulated Visium spots of 55 µm diameter and disregarded the spot-to-spot distance, resulting in 9 spots per FOV and 126 spots for the entire dataset. The FOVs had varying number of cells and cell types per simulated spot (**Supplementary Table 2A**). We created a reference dataset for each tissue slice using the combined single-cell expression profiles from all seven FOVs. There were 17 cell types in the cortex tissue section and 9 cell types in the olfactory bulb tissue section.

The STARMap dataset of mouse primary visual cortex (VISp) is 1.4 mm x 0.3 mm and contains 1,020 genes and 973 cells. We generated Visium-like spots as above, assuming that 1 pixel = 810 nm, as this information was not provided. We removed any spot with hippocampal cell types, and removed unannotated and Reln cells from spots, resulting in 108 spots comprising 569 cells and 12 cell types in total (**Supplementary Table 2A**). We used the VISp scRNA-seq dataset from the Allen Brain Atlas as the reference [30]. After filtering for common genes, 996 genes remained.

Silver. We used five scRNA-seq datasets and one snRNA-seq dataset for the generation of silver standards (**Supplementary Table 2B; Figure S9**). All datasets are publicly available and already contained cell type annotations. We downsized each dataset by filtering out cells with an ambiguous cell type label and cell types with less than 25 cells, and keeping only highly variable genes (HVGs) that are expressed in at least 10% of cells from one cell type. We set this threshold at 25% for the brain cortex dataset. We further downsampled the kidney dataset to 15,000 cells. For the two cerebellum datasets, we integrated them following the method of [21], performed gene filtering on the integrated dataset, and only retained common cell types.

Liver. We downloaded scRNA-seq and spatial datasets of healthy mice from the liver cell atlas [28] (**Supplementary Table 2C**). The scRNA-seq data has 185,894 cells and 31,053 genes from three digestion protocols: *ex vivo* digestion, *in vivo* perfusion, and nuclei isolation. We used the finer annotation of $CD45^-$ cells, which among others, subclustered endothelial cells into portal vein, central vein, lymphatic, and liver sinusoidal endothelial cells. We only selected cell types that are at least 50 cells present in all three digests, resulting in nine cell types. The spatial data consisted of four Visium slides, containing on average 1,440 spots. Based on hepatocyte zonation markers, the zonation trajectory of each spot was calculated and then each spot was annotated as part of central, mid, periportal, or portal zone.

We ran deconvolution using five (sub)sets of the scRNA-seq reference: three digestion protocols that were filtered to nine common cell types, the combined dataset of these three, and the entire atlas. This entire atlas was not used to compute evaluation metrics but only to visualize method stability (**Figure S12**). Due to the large size of the references, we ran some methods differently. We ran stereoscope with 5000 HVGs and subsampled each cell type to a maximum of 250 cells (*-sub*). We gave STRIDE a number of topics to test (*--ntopics 23 33 43 53 63*) instead of the default range of [#cell types, #cell types × 3]. For Seurat, we ran *FindTransferAnchors* with the "rpca" (reciprocal PCA) option instead of "pcaproject". MuSiC and SpatialDWLS use dense matrices in their code (in contrast to sparse matrices in other methods), resulting in a size limit of 2^{31} elements in a matrix. Whenever this is exceeded, we downsampled the reference dataset by keeping maximally 10,000 cells per cell type, and keeping 3000 HVGs along with any genes that are at least 10% expressed in a cell type. Finally, for the combined reference, we provided the digest information to cell2location and Seurat.

To compute the AUPR, we considered only spots annotated as being in the central or portal zone. We considered the following ground truth for portal vein and central vein ECs:

	Portal vein EC	Central Vein EC
Portal spot	1	0
Central spot	0	1

For the JSD, we first calculated the average abundance of each cell type for each slide separately. Then, we selected four samples from the snRNA-seq digest (ABU11, ABU13, ABU17, ABU20), as they were the only ones that contained all 9 cell types in the sample, and also averaged the abundance of each cell type. Then, we calculated the pairwise JSD between each of the four slides and four snRNA-seq samples and reported the average JSD. The reference value is obtained by averaging the pairwise JSD between the four snRNA-seq samples.

Scalability. We generated a synthetic dataset of 10,000 spots and ~31,000 genes from the liver atlas data using only the Nuc-seq part. We then used both the scRNA-seq data from the atlas as the reference dataset (both ex vivo and in vivo digestions). The synthetic dataset uses the prior distribution of cell types from the given data, so all spots have approximately the same composition (*generate_synthetic_data_lite* function from *synthspot*). The genes of the synthetic data were downsampled randomly, while the genes of the scRNA-seq data was downsampled based on the marker genes and HVGs. The spots/cells of both datasets were downsampled randomly (in a stratified manner for the reference dataset). We used the same parameters as we did for running the silver standard.

Benchmark metrics. The root-mean-squared error (RMSE) between the known and predicted proportions (p) of a spot s for cell type z, in a reference dataset with Z cell types in total, is calculated as

$$RMSE\ (s_{known}, s_{predicted}) = \sqrt{\frac{1}{Z}\sum_{z}^{Z}(p_{sz,known} - p_{sz,predicted})^2} \quad . \tag{1}$$

The Jensen-Shannon divergence (JSD) is a smoothed and symmetric version of the Kullback-Leibler divergence (KL). It is calculated as

$$JSD\left(s_{known} \,||\, s_{predicted}\right) = \frac{1}{2}KL(p_{s,known} \,||\, M) + \frac{1}{2}KL(p_{s,predicted} \,||\, M),$$

$$\text{where } M = \frac{1}{2}(p_{s,known} + p_{s,predicted})$$

The RMSE and JSD values across all spots are averaged and used as the representative value for a replicate k for each dataset-abundance pattern combination. Note that the RMSE calculation takes the number of cell types into account and hence should not be compared between datasets.

We used the area under the precision-recall curve (AUPR) instead of area under the receiver-operating curve (ROC) because in our silver standard there are more cell types absent than present per spot, and such imbalanced datasets are better evaluated with the AUPR [31]. For the AUPR, we get one value for the entire replicate. First, we binarized the known proportions by considering a cell type to be present in a spot if its proportion is greater than zero, and absent if it is equal to zero. Then, we compute the micro-averaged AUPR by aggregating all cell types together. We calculated the JSD and AUPR using the R packages *philentropy* and *precrec*, respectively [32], [33].

To provide a reference value for both metrics, we used (1) random proportions based on probabilities drawn from a Dirichlet distribution and (2) NNLS. For the former, we generated a reference value for all 56 combinations using the average value of 100 iterations. The dimension of the $\boldsymbol{\alpha}$ vector was equal to the number of cell types in the corresponding dataset, and all concentration values were set to 1.

For the NNLS baseline, we used the Lawson-Hanson NNLS implementation found in the *nnls* R package. We computed the single-cell reference profile by averaging the mean expression per gene for each cell type, and then used the function *nnls* to calculate the parameter estimates, which is then divided by the summed total to obtain the proportions. That is, we solve for β in the equation $\boldsymbol{Y} = \boldsymbol{X}\beta$, with $\boldsymbol{Y}$ the spatial expression matrix and $\boldsymbol{X}$ the average gene expression profile per cell type from the scRNA-seq reference.

References

[1] A. Wagner, A. Regev, and N. Yosef, “Revealing the vectors of cellular identity with single-cell genomics,” *Nat. Biotechnol.*, vol. 34, no. 11, pp. 1145–1160, Nov. 2016, doi: 10.1038/nbt.3711.

[2] M. Asp, J. Bergenstråhle, and J. Lundeberg, “Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration,” *BioEssays*, p. 1900221, May 2020, doi: 10.1002/bies.201900221.

[3] C.-H. L. Eng *et al.*, “Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+,” *Nature*, vol. 568, no. 7751, pp. 235–239, 2019.

[4] C. Xia, J. Fan, G. Emanuel, J. Hao, and X. Zhuang, “Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression,” *Proc. Natl. Acad. Sci.*, vol. 116, no. 39, pp. 19490–19499, Sep. 2019, doi: 10.1073/pnas.1912459116.

[5] S. G. Rodriques *et al.*, “Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution,” *Science*, vol. 363, no. 6434, pp. 1463–1467, Mar. 2019, doi: 10.1126/science.aaw1219.

[6] P. L. Ståhl *et al.*, “Visualization and analysis of gene expression in tissue sections by spatial transcriptomics,” *Science*, vol. 353, no. 6294, pp. 78–82, Jul. 2016, doi: 10.1126/science.aaf2403.

[7] A. Chen *et al.*, “Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays,” *Cell*, vol. 185, no. 10, pp. 1777-1792.e21, May 2022, doi: 10.1016/j.cell.2022.04.003.

[8] C.-S. Cho *et al.*, “Microscopic examination of spatial transcriptome using Seq-Scope,” *Cell*, vol. 184, no. 13, pp. 3559-3572.e22, Jun. 2021, doi: 10.1016/j.cell.2021.05.010.

[9] S. K. Longo, M. G. Guo, A. L. Ji, and P. A. Khavari, “Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics,” *Nat. Rev. Genet.*, vol. 22, no. 10, pp. 627–644, Oct. 2021, doi: 10.1038/s41576-021-00370-8.

[10] B. Li *et al.*, “Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution,” *Nat. Methods*, vol. 19, no. 6, Art. no. 6, Jun. 2022, doi: 10.1038/s41592-022-01480-9.

[11] L. Yan and X. Sun, “Benchmarking and integration of methods for deconvoluting spatial transcriptomic data,” *Bioinformatics*, vol. 39, no. 1, p. btac805, Jan. 2023, doi: 10.1093/bioinformatics/btac805.

[12] V. Kleshchevnikov *et al.*, “Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics,” *bioRxiv*, p. 2020.11.15.378125, Jan. 2020, doi: 10.1101/2020.11.15.378125.

[13] R. Lopez *et al.*, “DestVI identifies continuums of cell types in spatial transcriptomics data,” *Nat. Biotechnol.*, Apr. 2022, doi: 10.1038/s41587-022-01272-8.

[14] Q. Song and J. Su, “DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence,” *Brief. Bioinform.*, 2021, doi: 10.1093/bib/bbaa414.

[15] D. M. Cable *et al.*, “Robust decomposition of cell type mixtures in spatial transcriptomics,” *Nat. Biotechnol.*, p. 2020.05.07.082750, Feb. 2021, doi: 10.1038/s41587-021-00830-w.

[16] R. Dong and G.-C. Yuan, “SpatialDWLS: accurate deconvolution of spatial transcriptomic data,” *Genome Biol.*, vol. 22, no. 1, p. 145, Dec. 2021, doi: 10.1186/s13059-021-02362-7.

[17] M. Elosua-Bayes, P. Nieto, E. Mereu, I. Gut, and H. Heyn, “SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes,” *Nucleic Acids Res.*, p. 2020.06.03.131334, Feb. 2021, doi: 10.1093/nar/gkab043.

[18] A. Andersson *et al.*, “Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography,” *Commun. Biol.*, vol. 3, no. 1, pp. 1–8, 2020, doi: 10.1038/s42003-020-01247-y.

[19] D. Sun, Z. Liu, T. Li, Q. Wu, and C. Wang, “STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing,” *Nucleic Acids Res.*, vol. 50, no. 7, pp. e42–e42, Apr. 2022, doi: 10.1093/nar/gkac150.
[20] X. Wang, J. Park, K. Susztak, N. R. Zhang, and M. Li, “Bulk tissue cell type deconvolution with multi-subject single-cell expression reference,” *Nat. Commun.*, vol. 10, no. 1, p. 380, Dec. 2019, doi: 10.1038/s41467-018-08023-x.
[21] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data,” *Cell*, vol. 177, no. 7, pp. 1888-1902.e21, Jun. 2019, doi: 10.1016/j.cell.2019.05.031.
[22] T. Biancalani *et al.*, “Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram,” *Nat. Methods*, vol. 18, no. 11, pp. 1352–1362, Nov. 2021, doi: 10.1038/s41592-021-01264-7.
[23] A. Jindal, P. Gupta, Jayadeva, and D. Sengupta, “Discovery of rare cells from voluminous single cell expression data,” *Nat. Commun.*, vol. 9, no. 1, Art. no. 1, Nov. 2018, doi: 10.1038/s41467-018-07234-6.
[24] H. R. Ali, L. Chlon, P. D. P. Pharoah, F. Markowetz, and C. Caldas, “Patterns of Immune Infiltration in Breast Cancer and Their Clinical Implications: A Gene-Expression-Based Retrospective Study,” *PLoS Med.*, vol. 13, no. 12, pp. 1–24, 2016, doi: 10.1371/journal.pmed.1002194.
[25] E. Sato *et al.*, “Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer,” *Proc. Natl. Acad. Sci.*, vol. 102, no. 51, pp. 18538–18543, Dec. 2005, doi: 10.1073/pnas.0509182102.
[26] F. Vallania *et al.*, “Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases,” *Nat. Commun.*, vol. 9, no. 1, p. 4735, Dec. 2018, doi: 10.1038/s41467-018-07242-6.
[27] Z. Yao *et al.*, “A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation,” *Cell*, vol. 184, no. 12, pp. 3222-3241.e26, Jun. 2021, doi: 10.1016/j.cell.2021.04.021.
[28] M. Guilliams *et al.*, “Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches,” *Cell*, vol. 185, no. 2, pp. 379-396.e38, Jan. 2022, doi: 10.1016/j.cell.2021.12.018.
[29] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, “Nextflow enables reproducible computational workflows,” *Nat. Biotechnol.*, vol. 35, no. 4, pp. 316–319, Apr. 2017, doi: 10.1038/nbt.3820.
[30] B. Tasic *et al.*, “Adult mouse cortical cell taxonomy revealed by single cell transcriptomics,” *Nat. Neurosci.*, vol. 19, no. 2, pp. 335–346, Feb. 2016, doi: 10.1038/nn.4216.
[31] J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.
[32] T. Saito and M. Rehmsmeier, “Precrec: fast and accurate precision-recall and ROC curve calculations in R,” *Bioinformatics*, vol. 33 (1), pp. 145–147, 2017, doi: 10.1093/bioinformatics/btw570.
[33] H.-G. Drost, “Philentropy: Information Theory and Distance Quantification with R,” *J. Open Source Softw.*, vol. 3, no. 26, p. 765, Jun. 2018, doi: 10.21105/joss.00765.

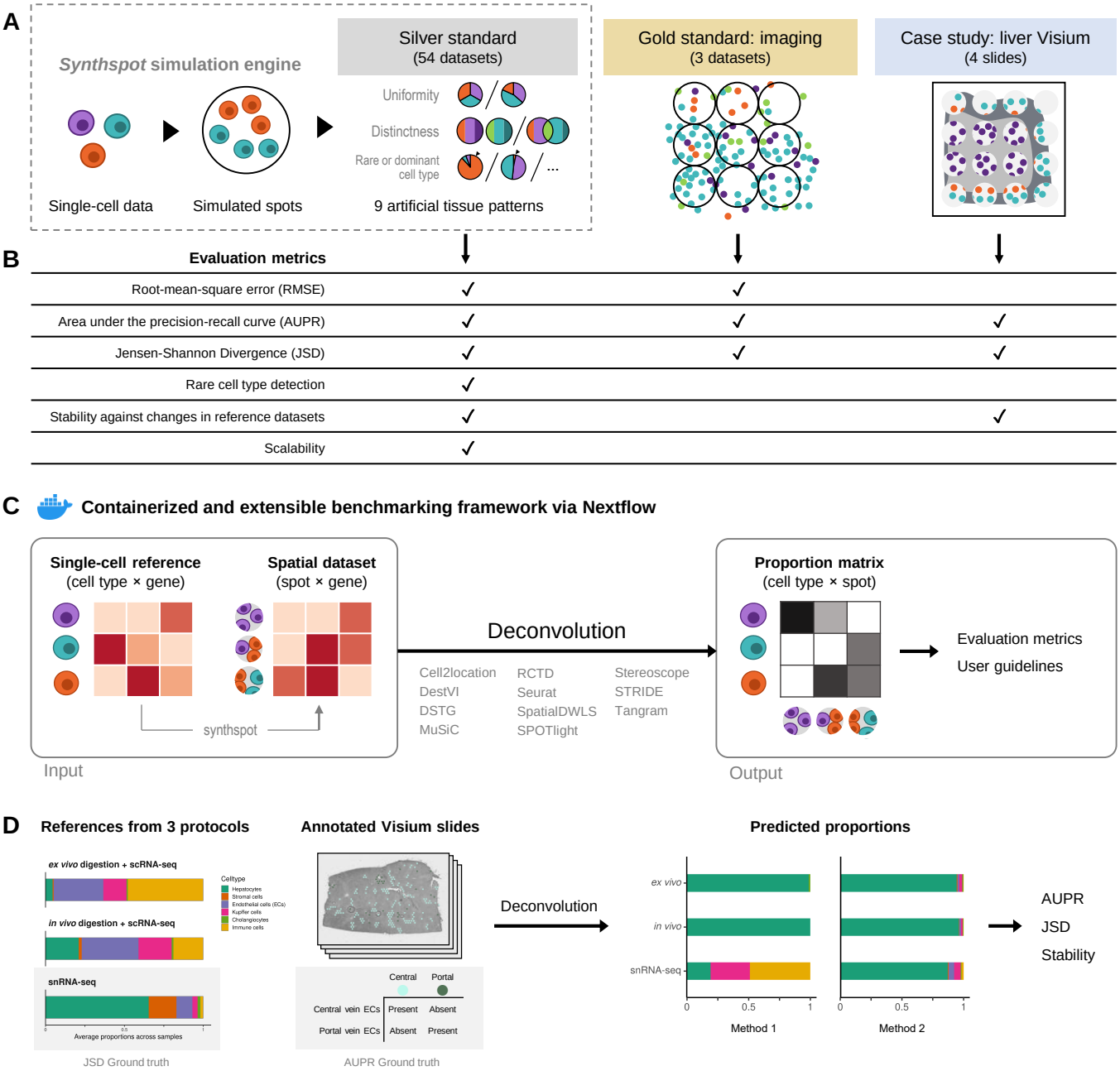

Figure 1. Overview of the benchmark. **(A)** The datasets used consist of silver standards generated from single-cell RNA-seq data, gold standards from imaging-based data, and a case study from liver data. Our simulation engine synthspot enables creation of artificial tissue patterns. **(B)** We evaluated deconvolution methods on three overall performance metrics (RMSE, AUPR, and JSD), and further checked specific aspects of performance, i.e., how well methods detect rare cell types and handle reference datasets from different sequencing technologies. **(C)** Our benchmarking pipeline is entirely accessible and reproducible through the use of Docker containers and Nextflow. **(D)** To evaluate performance on the liver case study, we leveraged prior knowledge of the localization and composition of cell types in the liver to calculate the AUPR and JSD. We also investigated method performance on three different sequencing protocols.

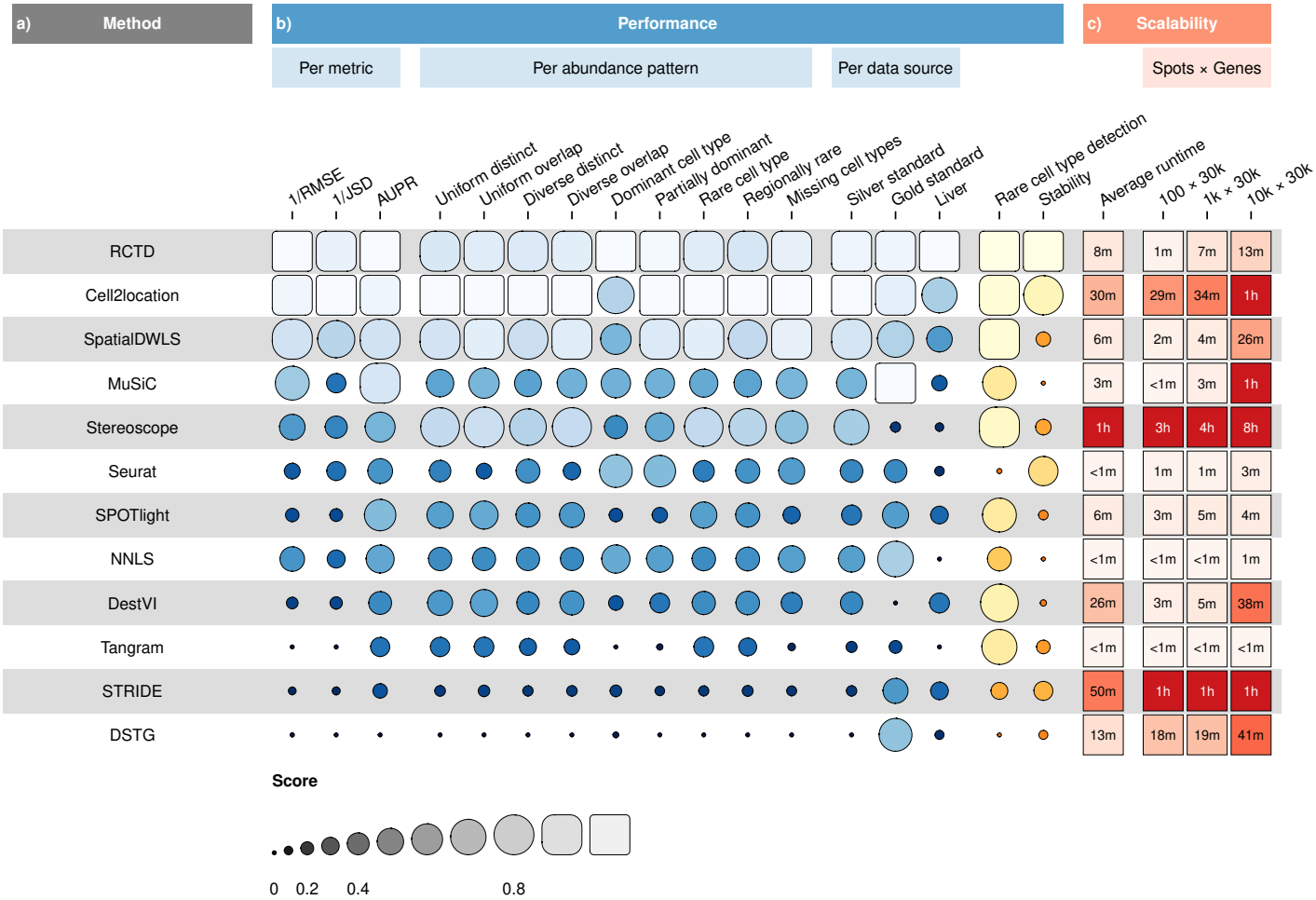

Figure 2. (a) Methods ordered according to the aggregated rankings of performance and scalability. **(b)** Performance of each method across metrics, artificial abundance patterns in the silver standard, and data sources. The ability to detect rare cell types and robustness against different reference datasets are also included. **(c)** Average runtime across silver standards and scalability towards more spots.

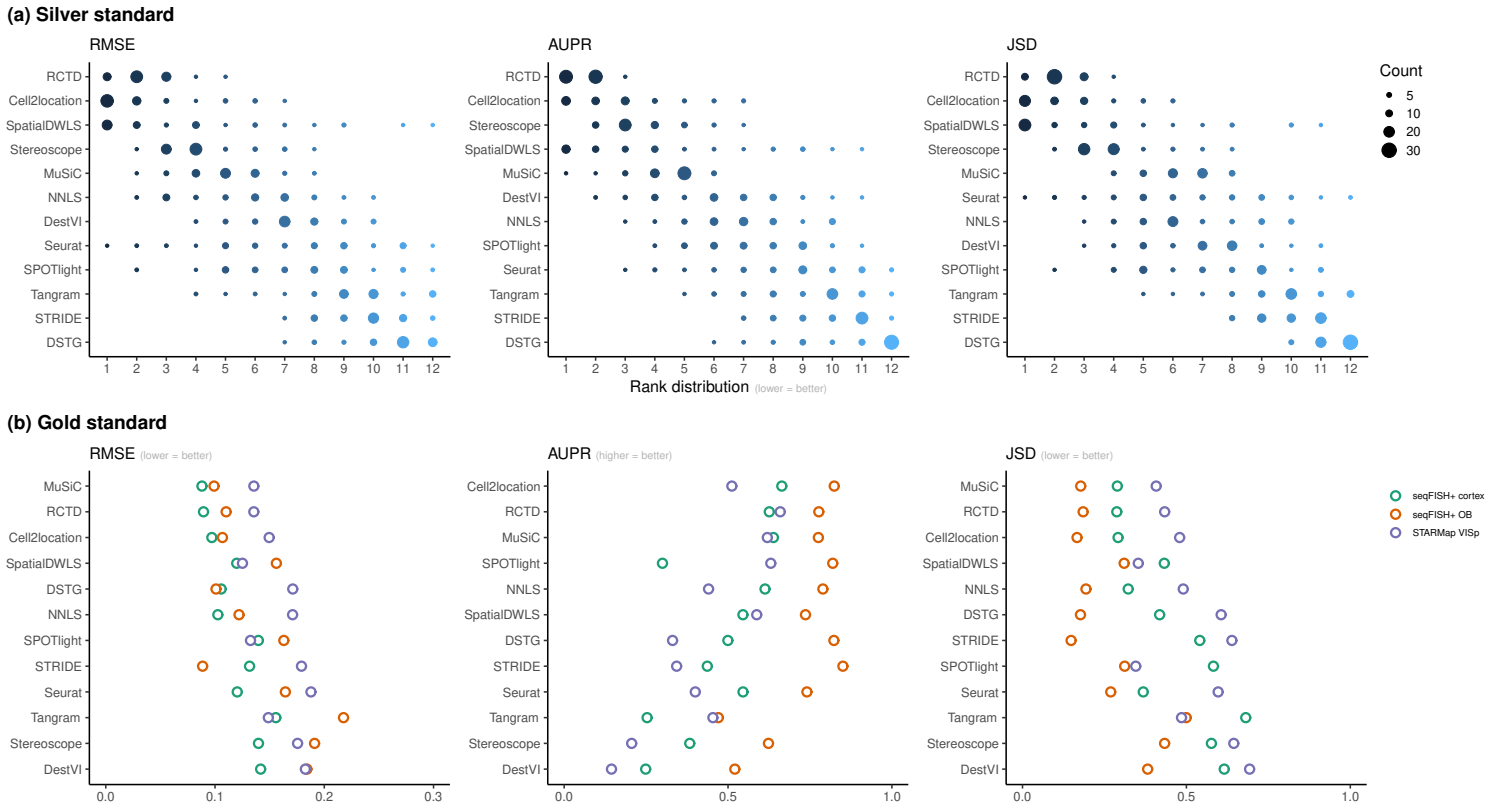

Figure 3. Method performance on synthetic datasets, evaluated using root-mean-squared error (RMSE), area under the precision-recall curve (AUPR), and Jensen-Shannon divergence (JSD). NNLS is the baseline algorithm (shaded). Methods are ordered from best to worst summed ranks. **(a)** The rank distribution of each method across all 54 silver standards, based on the best median value across ten replicates for that standard. **(b)** Gold standards of seqFISH+ datasets (10,000 genes) and STARMap (400 genes). We took the average over seven field of views for the seqFISH+ dataset.

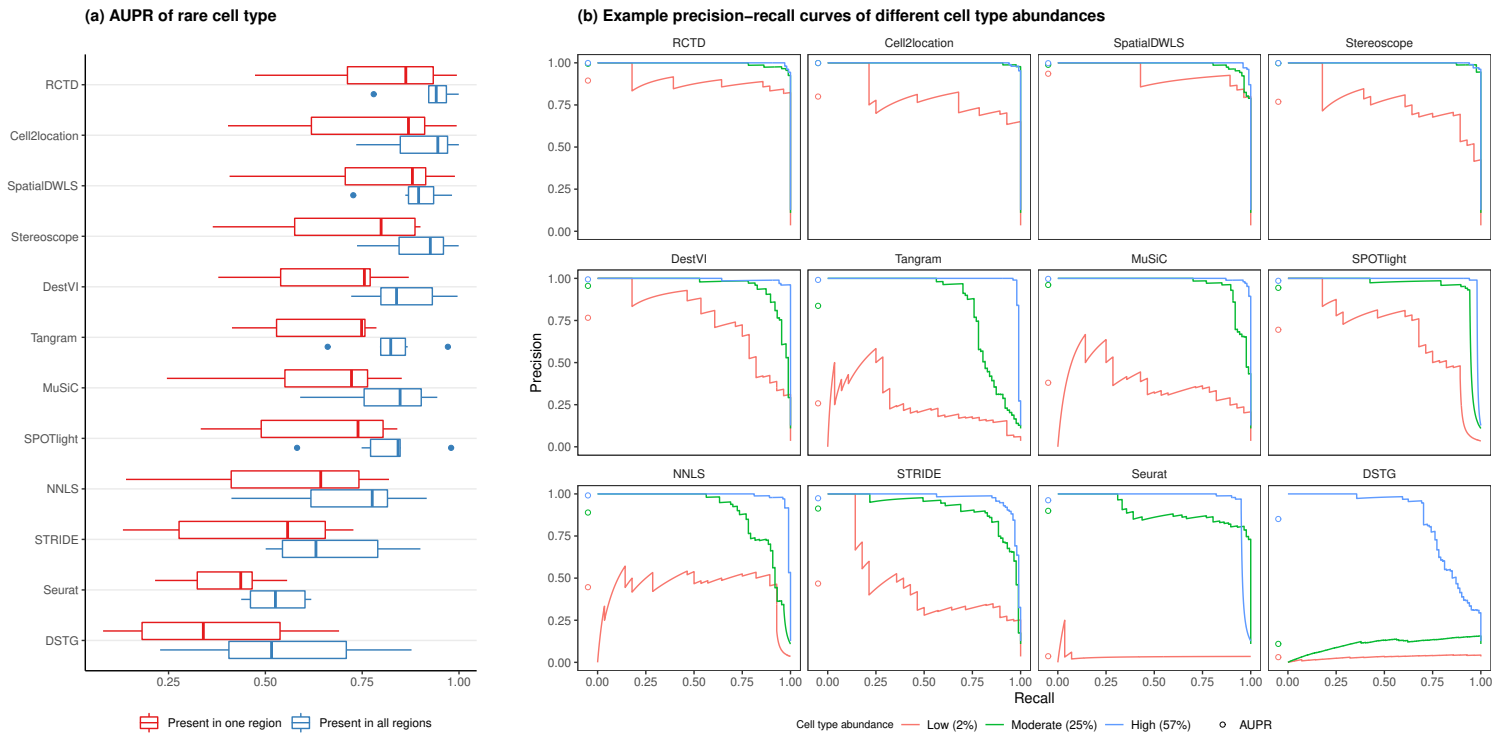

Figure 4. Detection of the rare cell type in the two *rare cell type* abundance patterns. **(a)** Area under the precision-recall curve in six datasets, averaged over ten replicates. Methods generally have better AUPR if the rare cell type is present in all regions compared to just one region. **(b)** Most methods can detect moderately and highly abundant cells, but their performance drops for lowly abundant cells.

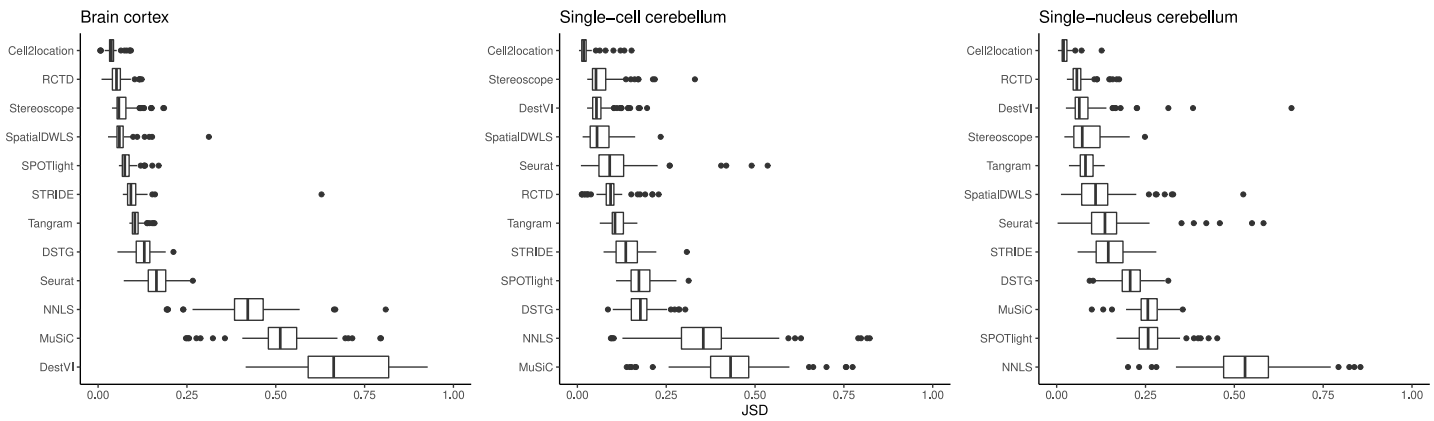

Figure 5. Stability against different reference datasets. For the same synthetic spatial datasets, two reference datasets were given and the Jensen-Shannon divergence (JSD) was computed between the predicted proportions. Methods are ordered based on stability, with a lower JSD indicating higher stability.

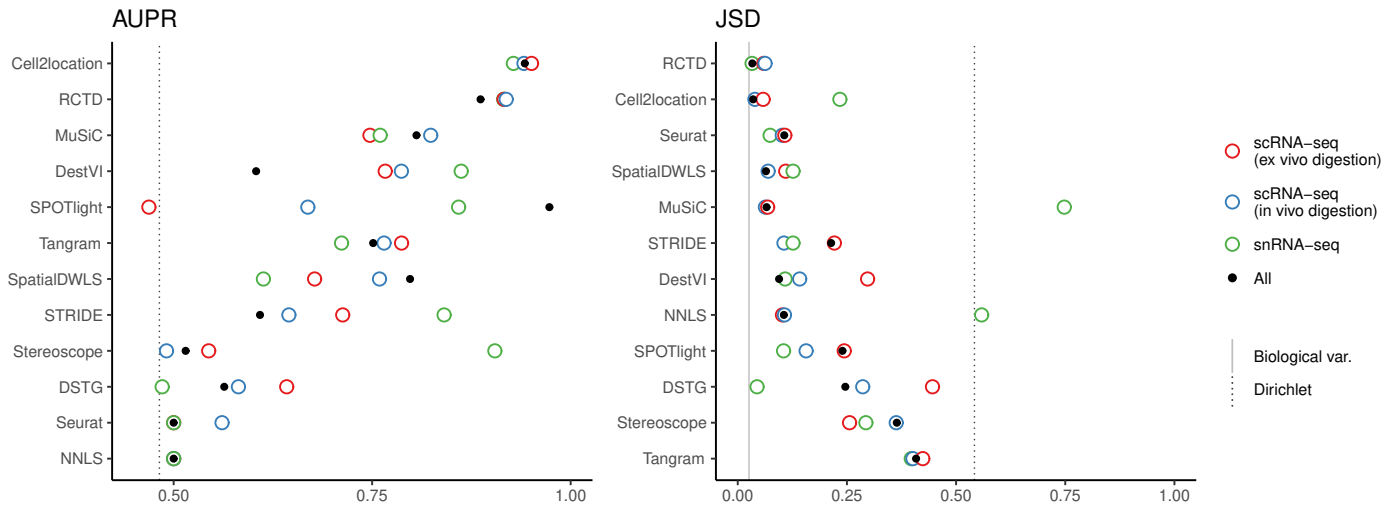

Figure 6. Method performance based on detecting portal/central vein endothelial cells in portal and central veins (AUPR), and by comparing distributions of all cell types (JSD). The biological variation is the average JSD between four snRNA-seq samples. All reference datasets here contain 9 cell types. Methods are ordered based on summed rank of all data points.

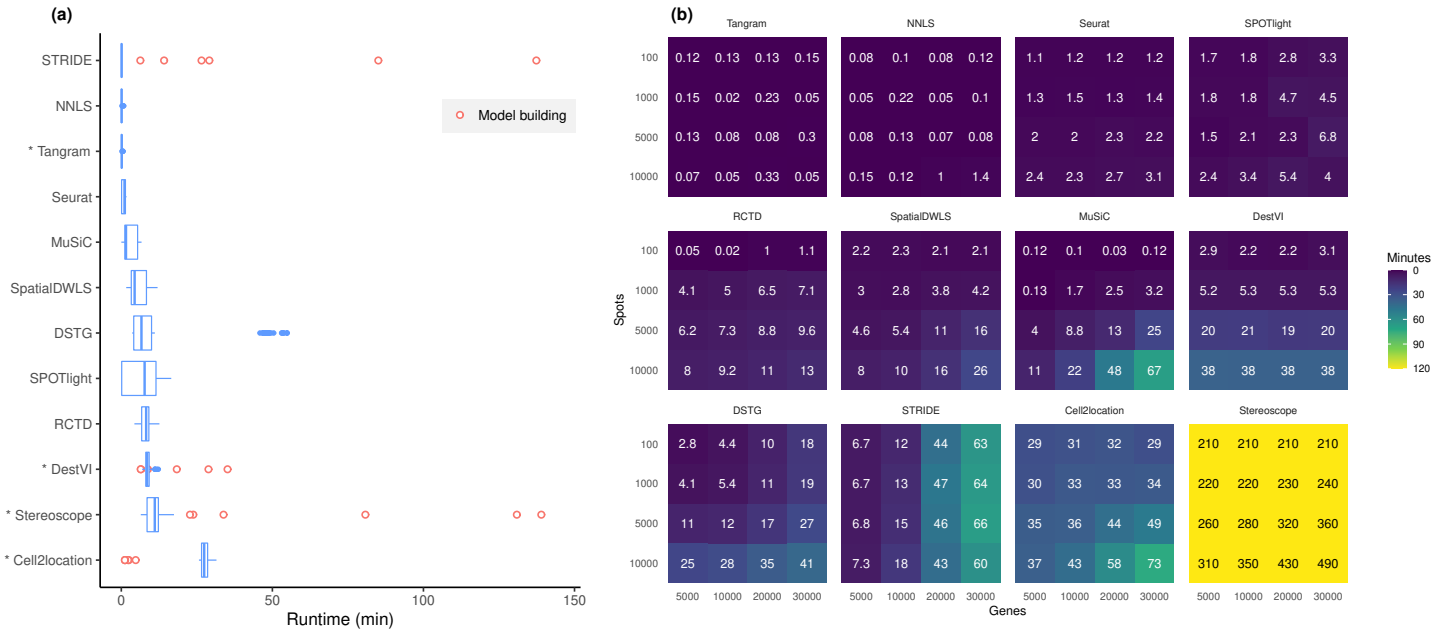

Figure 7. (a) Runtime over the 540 silver standards, ordered by median runtime. GPU acceleration is used whenever possible (asterisks). Cell2location, stereoscope, DestVI, and STRIDE first run with a model building step for each single-cell reference (red points), which can be reused for 90 synthetic datasets. **(b)** Scalability of the methods based on varying the dimensions of the spatial dataset. We added the model building and fitting time to model-based methods. Methods are ordered based on total time.