Published online 5 February 2021

Nucleic Acids Research, 2021, Vol. 49, No. 9 ***e50***
doi: 10.1093/nar/gkab043

SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes

Marc Elosua-Bayes[1], Paula Nieto[1], Elisabetta Mereu[1], Ivo Gut[1,2] and Holger Heyn[1,2,*]

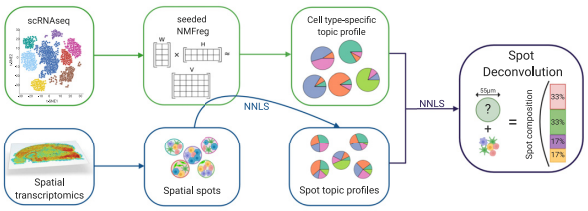
[1]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain and [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain

Received November 18, 2020; Revised January 04, 2021; Editorial Decision January 09, 2021; Accepted January 15, 2021

ABSTRACT

Spatially resolved gene expression profiles are key to understand tissue organization and function. However, spatial transcriptomics (ST) profiling techniques lack single-cell resolution and require a combination with single-cell RNA sequencing (scRNA-seq) information to deconvolute the spatially indexed datasets. Leveraging the strengths of both data types, we developed SPOTlight, a computational tool that enables the integration of ST with scRNA-seq data to infer the location of cell types and states within a complex tissue. SPOTlight is centered around a seeded non-negative matrix factorization (NMF) regression, initialized using cell-type marker genes and non-negative least squares (NNLS) to subsequently deconvolute ST capture locations (spots). Simulating varying reference quantities and qualities, we confirmed high prediction accuracy also with shallowly sequenced or small-sized scRNA-seq reference datasets. SPOTlight deconvolution of the mouse brain correctly mapped subtle neuronal cell states of the cortical layers and the defined architecture of the hippocampus. In human pancreatic cancer, we successfully segmented patient sections and further fine-mapped normal and neoplastic cell states. Trained on an external single-cell pancreatic tumor references, we further charted the localization of clinical-relevant and tumor-specific immune cell states, an illustrative example of its flexible application spectrum and future potential in digital pathology.

GRAPHICAL ABSTRACT



INTRODUCTION

Spatially resolved transcriptomics is key in advancing our understanding of tissue architectures. Unveiling the spatial disposition of cells enables researchers to determine cell-cell interactions and tissue reconstruction for a better knowledge of homeostasis and disease mechanisms. Array-based spatial transcriptomics (ST) is an unbiased and high-throughput approach to map genes within their spatial context. ST has been applied to chart the organizational landscape of tissues and diseases, such as prostate and pancreatic cancer (1,2), melanoma (3), amyotrophic lateral sclerosis (4) or the developmental human heart (5). Furthermore, recent studies successfully implemented ST to define the spatial topography of the human dorsolateral prefrontal cortex and its association with schizophrenia and autism (6).

Several technologies enable the spatial indexing of transcripts and the subsequent mapping of gene expression profiles, their main trade-off being a loss of single-cell resolution. Here, transcripts detected at capture locations (spots) are generally sampled from a mixture of cells which may be homo or heterogeneous. While widely used microarray-based ST techniques utilize 50–100 um spot diameters (10–20 cells) (7,8), bead array-based methods further minimized spot sizes to capture cell locations more precisely (2–10 um) (9,10). On the other hand, single-cell RNA sequencing (scRNA-seq) enables the profiling of thousands of single-cell transcriptomes without preserving the spatial context and potentially introducing recovery biases of cell compo-

*To whom correspondence should be addressed. Tel: +34 934020286; Email: holger.heyn@cnag.crg.eu

© The Author(s) 2021. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

sition. Successful integration of both data modalities could enable an in-depth study of tissue and organ architecture, elucidate cellular cross-talk, spatially track dynamic cell trajectories, and identify disease-specific interaction networks (e.g. between tumors and their microenvironment). Intersecting cell-type-specific genes from scRNA-seq with ST capture sites previously identified local enrichments, sufficient to segment tumor sections into normal and cancerous areas (2). However, while such analysis allowed predicting the presence or absence of cell types, it lacked the resolution to quantitatively infer cellular compositions at each capture site.

Here, we present SPOTlight, a deconvolution algorithm that builds upon a non-negative matrix factorization (NMF) regression algorithm which was previously applied to ST data (10). Importantly, SPOTlight adds prior information to the model, initializing both the basis and coefficient matrices with cell type marker genes, thereby greatly improving sensitivity and robustness. SPOTlight also relies on non-negative least squares (NNLS) to populate the coefficient matrix of capture locations as well as to determine a spot's composition. The latter is carried out by defining cell type-specific topic profiles, the distribution of genes defining a cell type or state, and by identifying the weights needed to reconstruct a spot profile. A unit-variance normalization step enables both paired and unmatched ST and scRNA-seq raw count matrices as input. We confirmed the sensitivity and accuracy of SPOTlight predictions on synthetic mixtures, testing scRNA-seq references of varying qualities (protocols, sequencing depth, cell numbers). SPOTlight showed excellent classification metrics even with low cell and molecule inputs. The possibility to integrate unpaired ST and scRNA-seq data enabled an automated, data-driven interpretation using large reference single-cell atlases, exemplified here using an adult mouse brain atlas (11). The automated interpretation of ST from patient sections has the potential to digitize pathology and improve patient stratification. As a proof-of-concept, we applied SPOTlight on pancreatic adenocarcinoma (PDAC) data and determined the spatial organization of clinical-relevant immune cell states in the tumor microenvironment.

MATERIALS AND METHODS

Implementation

Non-negative matrix factorization regression. The following annotations will be used when describing the model:

- N – Set of all cells from scRNAseq.
- M – Set of all capture locations from spatial data.
- G – Set of selected genes from scRNAseq, cell type marker genes + 3000 highly variable genes.
- G' – Set of all genes from spatial data.
- Gi – G∩G', intersection between G and G'.
- C – Number of cell types in the scRNAseq dataset
- K – Number of topics to use to reduce the dimensionalities, equal to C.
- V – matrix of dimensions Gi × N containing data from scRNAseq
- W – matrix of dimension Gi × K containing the gene distribution for each topic, basis between V and H.
- H – matrix of dimensions K × N containing the topic distribution for each cell.
- V' – matrix of dimensions Gi × M containing spatial data.
- H' – matrix of dimensions K × M containing the topic distributions for each capture location.
- Q – matrix of dimension K × C containing the topic distributions for each cell type.
- P – matrix of dimension C × M containing the cell type weights for each capture location.

At the core of our tool, we use non-negative matrix factorization (NMF) along with non-negative least squares (NNLS). NMF is used to factorize a matrix into two or more lower dimensionality matrices without negative elements. We first have an initial matrix V, which is factored into W and H. Unit variance normalization by gene is performed in V and V' in order to standardize discretized gene expression levels, '*counts-umi*' (10,12). Factorization is then carried out using the non-smooth NMF method (13), implemented in the R package *NMF* (14). This method is intended to return sparser results during the factorization in W and H, thus promoting cell-type-specific topic profile and reducing overfitting during training. Before running factorization, we initialize each topic, column, of W with the unique marker genes for each cell type with weights $1 - P$ value. The marker genes are obtained from Seurat's function *FindAllMarkers*. In turn, each topic of H in SPOTlight is initialized with the corresponding belongance of each cell for each topic, 1 or 0. This way, we seed the model with prior information, thus guiding it towards a biologically relevant result. This initialization also aims at reducing variability and improving the consistency between runs.

$$V \sim W * H$$

Second, NNLS regression is used to map each capture location's transcriptome in V' to H' using W as the basis. We obtain a topic profile distribution over each capture location which we can use to determine its composition.

$$V' \sim W * H'$$

Third, we obtain Q, cell-type specific topic profiles, from H. We select all cells from the same cell type and compute the median of each topic for a consensus cell-type-specific topic signature. We then use NNLS to find the weights of each cell type that best fit H' minimizing the residuals.

$$H' \sim Q * P$$

We use a minimum weight contribution to determine which cell types belong within a capture location. 0.09% is set by default, related to the expected number of cells at the capture locations (1–10 cells). In a scenario with 10 cells, we would detect all and also account for partially contributing cells.

By using NNLS, we are able to return a measure of error along with the predicted cell proportions. To do so, we calculate the total sum of squares (TSS) and the residual sum of squares (RSS) for each row. By dividing the RSS by the TSS we obtain the percentage of unexplained residuals for each spot. This measure can be used to assess the quality of

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

a predicted composition.

$$\text{TSS} = \sum_{i=1}^{C} (Y_i - 0)^2$$

$$\text{RSS} = \sum_{i=1}^{C} \left(Y_i - \hat{Y}_i\right)^2$$

$$\text{Unexplained residuals } (\%) = \text{RSS/TSS}$$

Parameters

Three important parameters can be adjusted and tuned in order to optimize the performance of this tool: (i) number of cells per cell-type, (ii) the supervised vs unsupervised approach along with marker gene sets and (iii) the minimum weight contribution threshold to include a cell as present. We benchmarked these parameters to assess their impact on the performance as follows:

- The number of cells per cell-type used to train the model. We identify the optimal number of cells maximizing performance along with computing time.
- The supervised versus unsupervised approach. For the former, we also tested marker genes together with different numbers of HVG. For the unsupervised approach, we used the 3000 HVG as in the original NMFreg (10).
- The minimum weight contribution. This refers to the NNLS weights from C that best fit H'. This weight contribution must be set. Due to the nature of NNLS, there may be cell types contributing a low amount just to residually minimize the squares, and therefore, adding noise to the prediction.

Synthetic mixtures

To be able to test the tool's performance, to benchmark parameters and to apply it on different data types, we generated synthetic mixtures of cells with defined composition. To generate these synthetic test mixtures, we selected between two and eight cells from the scRNA-seq datasets and combined their transcriptomic profiles. If the resulting mixture had >25 000 UMI counts we randomly downsampled it to 20 000 UMI counts in order to better simulate biological capture locations. Test mixtures can be generated using the SPOTlight function *test_spot_fun*.

Performance evaluation

To address how well the model performed, we assessed several parameters using synthetic and real datasets. From the predicted composition, we first evaluated if we were able to accurately predict when a cell type was correctly predicted within the mixture. Moreover, we also assessed if the predicted proportions were an accurate representation of the true composition. The former is a classification problem for which we used the following parameters; *sensitivity*, if a cell type correctly predicted to be present within the capture location; specificity, predicting its absence when its not present; *precision*, how good we are at identifying cell-types present; *accuracy*, percentage of correctly classified cell types; and *F1 score*, integrating sensitivity and precision. For the latter we used the *Jensen-Shannon Divergence* (JSD) distance metric used to measure the similarity between two probability distributions, P and Q, defining a probability space X. The JSD is a symmetric and smoothed version of the Kullback-Leibler divergence.

$$D_{kl}(P||Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

$$M = \frac{1}{2}(P + Q)$$

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

Benchmarking

For technology benchmarking, we assessed if the technology used to obtain scRNAseq data affected the performance of the model. We used data from peripheral blood mononuclear cells (PBMC) downsampled to the same sequencing depth (20 000 reads/cell). Specifications on how the data was generated and processed can be found elsewhere (15). For each technology, we trained the model and tested synthetic mixtures of 2–8 cells. We assessed: Cel-Seq2, ChromiumV2, Chromium V2 single-nucleus, C1HT-medium, C1HT-Small, ddSeq, Drop-Seq, gmcSCRB-Seq, ICELL8, inDrop, MARS-Seq, QUARTZ-Seq2 and SMART-Seq2. Data is publicly available through the Gene Expression Omnibus (GSE133549). To benchmark the effect of sequencing depth, we analyzed a Chromium V3 PBMC dataset (15) (GSE133549) downsampled to different depths using zUMI (16) (5000, 10 000, 15 000, 20 000, 50 000 reads/cell). Shared cell types between the different datasets were used excluding biases introduced by varying cell type numbers.

We benchmarked SPOTlight against other bulk and single-cell deconvolution tools: MuSiC (weighted and all-genes) (17), CIBERSORTx (18), DeconRNAseq (19), SCDC (20), RCTD (21) and the unsupervised NMFreg (10). Furthermore, we applied CoGAPS (22) to carry out the non-negative matrix factorization step in the SPOTlight workflow to assess if pre-existing single-cell specific tools showed an improvement on nNMF. To replace the nNMF approach by CoGAPS, we implemented the latest version of CoGAPS, CoGAPS 3 by Sherman *et al.*, using the Bioconductor R package version '3.6.0'. All tools were run with default settings as specified in their documentation and vignettes using the QUARTZ-Seq2 scRNA-seq dataset as the training set and the 1000 previously generated synthetic mixtures. Performances were assessed by the ability to correctly predict the presence/absence and the cell type proportions.

Mouse brain deconvolution

To assess the tool's performance on a biological dataset, we used mouse brain as model tissue. Despite its complexity

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

with multiple cell types and states, it presents well-defined structures with location-specific types. We used a mouse brain reference scRNA-seq dataset comprised of cells sampled from multiple cortical areas and the hippocampus, provided by the Allen Institute, with ∼76 000 cells and 47 annotated clusters sequenced using SMART-Seq2 (11,23) (GSE71585). The spatial transcriptomics data of an adult mouse brain (anterior and posterior sagittal slices) was obtained from 10X Genomics (24). Two replicates for each slice were available and used to confirm the predictions. To validate the predicted cell type spatial distribution within the brain structure, we used known cell-type gene markers along with reference *in situ hybridization* (ISH) image data at cellular-level resolution from the Allen Mouse Brain Atlas (25). The marker genes used for the hippocampal cell types represented in this study were: Cornu Ammonis 1 stratum pyramidale (CA1sp), *Fibcd1*; Cornu Ammonis 2 stratum pyramidale (CA2sp), *Ccdc3*; Cornu Ammonis 3 stratum pyramidale (CA3sp), *Pvrl3*; and Dentate gyrus (DG), *Prox1*, as reported in Cembrowski *et al*. (2016).

In situ hybridization images were obtained from the Allen Brain Atlas. Links to the images are the following:

- *Fibcd1*:mouse.brain-map.org/experiment/siv?id=69672462&imageId=69647545&initImage=ish&coordSystem=pixel&x=4464.5&y=3184.5&z=1
- *Ccdc3*:mouse.brain-map.org/experiment/siv?id=68844056&imageId=68705406&initImage=ish&coordSystem=pixel&x=4352.5&y=2880.5&z=1
- *Pvrl3*:mouse.brain-map.org/experiment/siv?id=69816733&imageId=69747543&initImage=ish&coordSystem=pixel&x=5744.5&y=3576.5&z=1
- *Prox1*:mouse.brain-map.org/experiment/siv?id=69289763&imageId=69177644&initImage=ish&coordSystem=pixel&x=5416.5&y=3720.5&z=1

Pancreatic ductal adenocarcinoma

We used pancreatic ductal adenocarcinoma (PDAC) ST data publicly available through the Gene Expression Omnibus (GSE111672) (2). Spatial data for this study was generated with the original spatial transcriptomics technology (9), while scRNAseq data was generated using inDrops. Further specifications on how the data was generated and processed can be found elsewhere (2). In total, 10 spatial slides from six tumor samples are available, two of which (PDAC-A and PDAC-B) have three biological replicates and paired scRNAseq data. For the purpose of this study, we used samples PDAC-A and PDAC-B and selected sections that harbored both normal and tumor areas (identified through the mapping of normal cell types and tumor clones). For PDAC-A, we used GSM3036911 (ST1 data) and GSM3036909, GSM3036910, GSM3405527, GSM3405528, GSM3405529, GSM3405530 (inDrops data). For PDAC-B, we used GSM4100723 (ST2 data) and GSM3405531, GSM3405532, GSM3405533. Filtering and data processing was carried out as specified in the original publication, keeping cells with ≥1000 UMIs, ≤20% mitochondrial transcripts, and ≤30% ribosomal transcripts (2). In PDAC-B, one cluster of ductal cells with low UMIs and high mitochondrial content was removed. A cell type annotation of the scRNA-seq datasets was provided by the authors of the original publication (2).

To generate a comprehensive immune cell type reference atlas for PDAC, we re-analyzed scRNA-seq data from Peng *et al*. (2019); Genome Sequence Archive: ID PRJCA001063). From this dataset only the tumoral pancreas samples were included. Cells with >20% of mitochondrial content and <100 UMIs were removed. We normalized, scaled, extracted the highly variable genes and performed PCA analysis on the remaining cells prior to clustering. Resulting clusters were annotated according to gene markers provided in the original manuscript. All the tumor and non-immune cells were identified and removed by marker gene analysis. For the detailed annotation of the immune cells, we used cell labels as defined in our Tumor Immune Cell Atlas (26). Briefly, we first used canonical markers to group cells into the major cell types (i.e. *CD79A*, *CD68* and *CD3E* for B-cells, myeloid cells, and T-cells, respectively). To further stratify the cells into cell states, we re-clustered and annotated each of them comparing the cluster markers to well-characterized single-cell gene sets of the tumor microenvironment (27–29) by computing the Jaccard similarity index using *matchSCore2* (15). We were able to identify all of the expected cell populations, including rare immune cell states.

When stratifying the tissue into tumoral and non-tumoral sections, tumoral spots contained >40% cancer-cell proportion. Cell type proportions within the spots were compared between regions and significance assessed using a non-parametric test (Mann–Whitney). To assess cell type enrichment between regions, we computed the proportion of spots containing each cell type. The significance between the proportions was assessed with a permutations test where the cell type specific statistic distribution was created randomly 10 000 times for each cell type. Moreover, we also assessed a third region, intermediate, between the tumoral and non-tumoral regions. Here, regions were defined as follows: tumoral, >40% cancer-cell proportion; intermediate, <40% cancer-cell and ductal-cell proportion; and non-tumoral, >40% ductal-cell type proportion. Again, cell type proportions within the spots were compared between regions and significance assessed with a Mann–Whitney test. Bonferroni adjusted *P*-values are reported for multiple comparisons. To calculate interaction networks, the edges between the cell-types represent the proportion of spots in which we detect co-localization.

Code versions and availability

This tool is developed to run with R versions ≥3.5; docker images with the appropriate environment are available at Docker hub: marcelosua/spotlight_env_rstudio and marcelosua/spotlight_env_r.

RESULTS AND DISCUSSION

At the core of SPOTlight, we identify cell type-specific topic profiles used to deconvolute ST spots (Figure 1). We set out to use NMF to obtain topic profiles due to its previous success in identifying biologically relevant gene expression programs (12), as well as its previous implementation in ST analysis (10). Its non-negative constraint allows it to

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

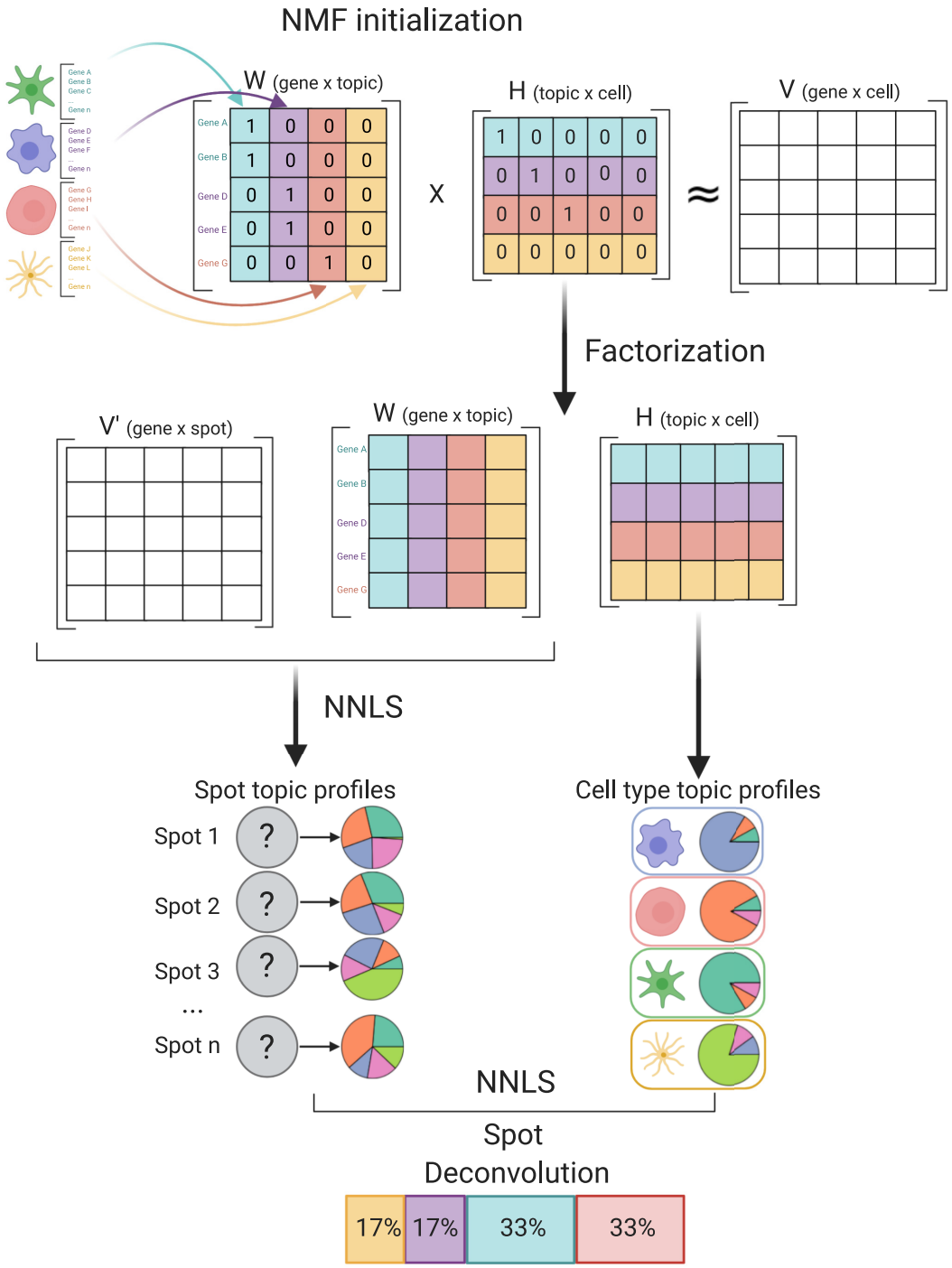


Figure 1. SPOTlight scheme. Step-by-step illustration of SPOTlight's algorithm. At the beginning of this process we have a count matrix, V, for scRNAseq data and a set of marker genes for the identified cell types. First, we use prior information to initialize the basis and coefficient matrices, W and H respectively. We assume the number of topics, k, to be equal to the number of cell types in the dataset. Each topic is then associated with a cell type; columns in W are initialized with marker genes for the associated cell type with that topic, while rows in H are initialized with the membership of each cell to its associated topic. Second, we proceed with the matrix factorization from which we obtain gene distributions for each topic in W, and topic profiles for each cell in H. Third, we use W to map the ST data, V', by means of non-negative least squares (NNLS) to obtain H'. Columns in H' represent the topic profile for each spot. Fourth, from the H matrix obtained from the scRNAseq data we consolidate all the cells from the same cell type to obtain cell type-specific topic profiles. Lastly, we use NNLS to find which combination of cell type-specific topics resembles each spot's topic profile.

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

model count data, which provides more interpretable results than standard matrix factorization. We seed the model with prior information, guiding it towards biologically relevant results and greatly improving the consistency between runs. Gene expression counts are used as input after a unit variance normalization (by gene) is performed to standardize discretized gene expression levels (10,12). Importantly, the NMF is initialized by the two main matrices: the basis matrix (W) with unique cell type marker genes and weights, and the coefficient matrix (H) in which each row is initialized, specifying the corresponding relationship of a cell to a topic (i.e. association with a cell type, Figure 1). Factorization is then carried out using non-smooth NMF (13,14). This step returns sparser results during the factorization, promoting cell type-specific topic profiles, while reducing overfitting during training. After factorization, we obtain cell type-specific topic profiles from the coefficient matrix and generate consensus topic signatures across all cells. Subsequently, NNLS regression is used to map each spot's transcriptome to a topic profile distribution using the unit-variance normalized ST count matrix and the basis matrix previously obtained. Lastly, NNLS is again applied to determine the weights for each cell type that best fit each spot's topic profile by minimizing the residuals. We use a minimum weight contribution threshold to determine which cell types are contributing to the profile of a given spot, also considering the possibility of partial contributions. NNLS also returns a measure of error along with the predicted cell proportions, allowing the user to estimate the reliability of predicted spot compositions.

Benchmarking SPOTlight performance

To evaluate the SPOTlight's performance, we benchmarked parameters and tested different scenarios with synthetically generated mixtures of cells of known cell type composition. To generate synthetic mixtures, we selected cells from peripheral blood mononuclear cell (PBMC) scRNA-seq datasets and combined their transcriptomic profiles to different proportions (Materials and Methods). PBMC scRNA-seq data have multiple well-characterized and discrete cell populations, providing an ideal input for benchmarking purposes. Synthetic mixtures then served as ground-truth to evaluate SPOTlight's performance to predict cell types and spot composition using the following parameters: *sensitivity* (correctly predicted cell type presence); *specificity* (correctly predicting absence); *precision* (performance when calling a cell type present); *accuracy* (percentage of correctly classified cell types); and *F1 score* (integrating recall/sensitivity and precision). To assess the similarity between the real and predicted proportions, we used the Jensen-Shannon Divergence (JSD), a distance metric that determines the similarity between two probability distributions. As JSD is a distance metric, values closer to 0 signify a higher similarity between both distributions.

When testing the performance on synthetic mixtures, we obtained a sensitivity of 0.911, an accuracy of 0.78, and an F1 score and specificity of 0.77 and 0.63, respectively. Median JSD values of 0.160 [CI:0.096–0.224] indicated a high accuracy of estimated cell type proportions. The benchmarking results are in line with results from subsequent applications of SPOTlight in different biological scenarios, such as brain tissue or PDAC patient samples; SPOTlight sensitively detected cell types and subtle cell states at their expected locations. A major challenge of NMF is its stochastic nature, requiring repeated iterations from different starting points in order to obtain valid results. To overcome this inherent variability, we initialized the basis and coefficient matrices by seeding them with prior information. Consequently, multiple iterations with seeded NMF regression obtained very similar results for synthetic cell type mixtures (JSD scores, Figure 2A–D). In line with these results, topic profiles from different cell types displayed consistent profiles in all iterations (Supplementary Figure S1) and single cells used to train the model presented comparable topic profiles (Supplementary Figure S2).

We reason that different input qualities (transcriptome complexity), quantities (cell numbers) and proportion (extended discussion) would critically impact the performance of SPOTlight. Therefore we opted to test different input scenarios, including scRNA-seq protocols, sequencing depth, cell numbers and other tunable parameters to simulate variable experimental designs and to identify ideal inputs and limitations of the tools.

We previously benchmarked scRNA-seq protocols for their performance in producing complex sequencing libraries and their suitability to generate reference cell atlases (15). First, we assessed if the scRNA-seq technologies used to generate data affected the performance of SPOTlight. Different protocols produced vastly variable data qualities and we expected this to impact downstream applications, such as deconvolution algorithms. We used downsampled scRNA-seq datasets (20,000 reads per cell) and trained the SPOTlight model on synthetic mixtures for each protocol (Figure 2A). The best performance was achieved with Quartz-Seq2, Smart-Seq2, and Chromium protocols that also showed excellent benchmarking performance. It is worth highlighting the performance of single-nucleus (sn) sequencing in this context (Chromium sn), which resulted in deconvolution metrics that were comparable to scRNA-seq despite the sampling from a reduced transcriptome pool. In general, scRNA-seq with defined clusters and cell type-specific markers are ideal for optimal performance of SPOTlight. However, other commonly used sc/snRNA-seq protocols also return accurate predictions.

Second, we benchmarked the impact of reduced sequencing depth to identify the performance peak for a cost-effective reference atlas generation. An increased sequencing depth enables the detection of more molecules and genes, including lowly expressed transcripts. When testing SPOTlight on step-wise downsampled datasets (5000–50 000 reads per cell), we observed a critical drop in performance at lower sequencing depth (Figure 2B). While accuracy and specificity were comparable to deeply sequenced datasets, the sensitivity and accuracy of estimated cell type proportions (JSD index) was reduced at lower depths. Nevertheless, despite the lower sensitivity, shallowly-sequenced data such as large atlas projects (30,31) are also suitable inputs for accurate localization of cell types in space. We detected a peak in performance ∼20 000 reads per cell; this sequencing depth was also identified to be most cost-efficient for high-throughput scRNA-seq protocols (32).

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

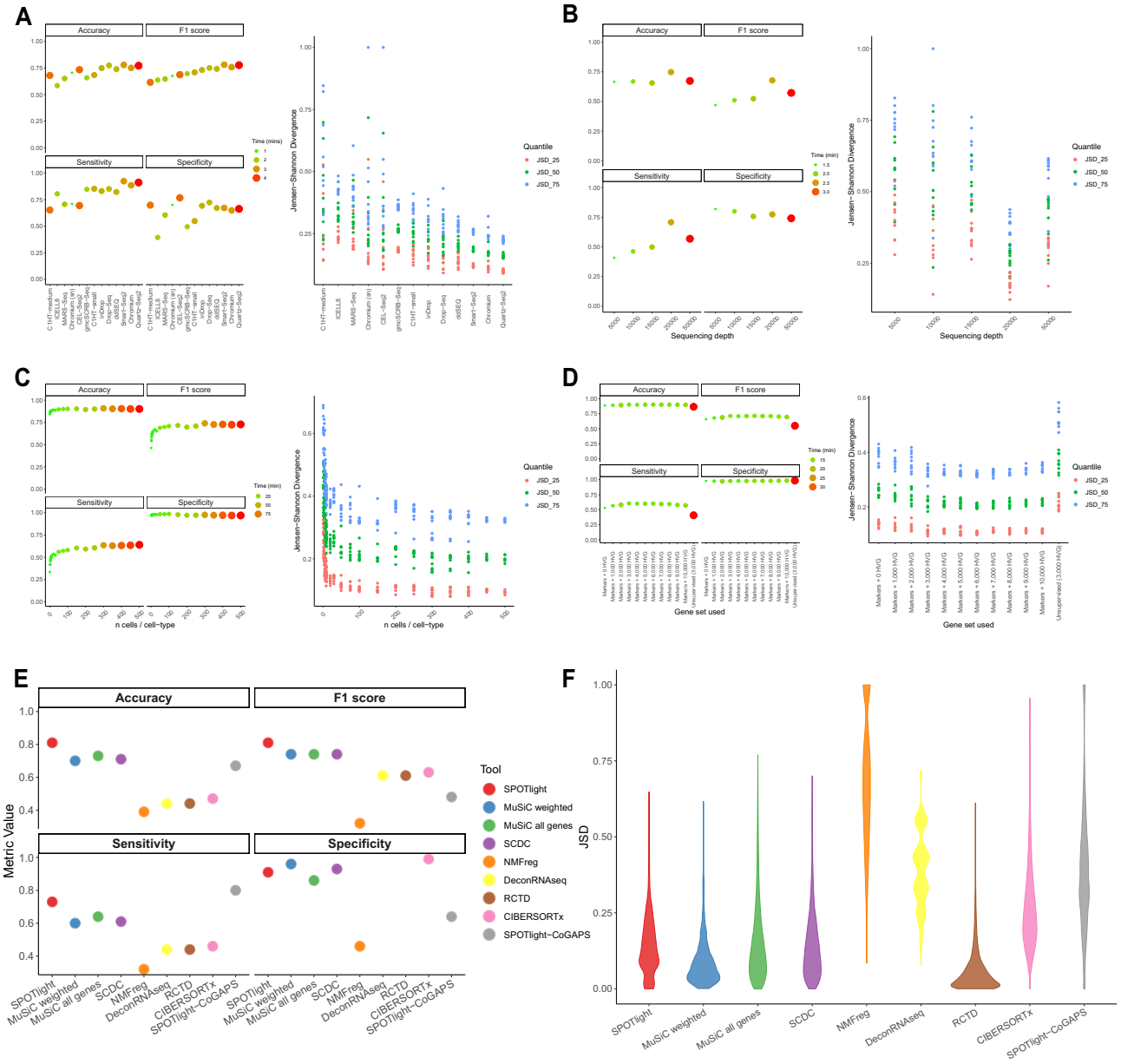


Figure 2. Benchmarking SPOTlight under different technical conditions and parameter optimization. In the classification metrics, the mean of 10 iterations is shown; when assessing JSD we show the results for each iteration. (**A**) The performance of SPOTlight was assessed with different scRNA-seq protocols to identify optimal input protocols (20 cells per cell type). (**B**) Benchmarking SPOTlight on the same dataset downsampled to different sequencing depths (20 cells per cell type). Performance improved at higher depths, but no sharp decrease at shallow sequencing depths was observed. (**C**) Optimizing the number of cells per cell type used to train the SPOTlight model. Peak performance was obtained using 100 cells per cell type, less cells decreased performance while more cells increased computational time without improving performance. (**D**) Optimizing the gene sets used to train the SPOTlight model. Optimal performance was obtained when using the union gene set between marker genes and 3000 highly variable genes (HVG). The unsupervised approach using only the 3000 HVG performed the worst. (**E**) Benchmarking classification performance of bulk and single-cell deconvolution tools on 1000 synthetic mixtures. SPOTlight proofed to be the most accurate and with the highest F1 score. (**F**) Proportion prediction performance of the different deconvolution tools on 1000 synthetic mixtures.

Third, scRNA-seq protocols scale differently with droplet-based methods processing up to millions of cells, while plate-based protocols (e.g. Smart-seq2) generally generate datasets for a few thousand cells. We assessed the impact of input cell numbers on SPOTlight's performance, also considering computational time as an important factor using synthetic mixtures generated from a large and complex Smart-seq2 dataset (11,23). We found that the cell number per cell type to train the model was a key parameter (Figure 2C). The optimal value to strike a balance between deconvolution performance and computational time was around 100 cells. Selecting fewer cells would decrease computational time, but the performance has not plateaued. Selecting more cells would drastically increase computational time with marginal improvements on performance. As 100 cells per cell type are in the range of both

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

droplet- and plate-based methods, SPOTlight is suitable for the most commonly used formats of scRNA-seq data.

To train the model, different sources (gene selection) can be used as input. The selection of highly variable genes (HVG) has been shown to be critical for the clustering of scRNA-seq data and we reason that it could also be crucial for spot deconvolution. We further quantified the improvements due to the addition of cell type gene markers, a main difference to previous tools using NMF on ST data (10). We found that SPOTlight's performance was optimal when combining both HVG and specific cell types markers to seed the model (Figure 2D). Marker genes critically improved all metrics compared to an unsupervised approach using the 3000 HVG alone as proposed by the original NMF regression documentation (10). The number of HVG used had a marginal impact on the performance; however, optimal performance was observed using gene markers combined with the 3000 HVG.

Lastly, we benchmarked SPOTlight against published bulk and single-cell deconvolution methods, including MuSiC (weighted and all-genes) (17), CIBERSORTx (18), DeconRNAseq (19), SCDC (20), RCTD (21), and the unsupervised NMFreg (10). Predicting the presence/absence of cell types within a synthetic mixture, SPOTlight showed the highest accuracy, F1 score and sensitivity (Figure 2E). Assessing the performance of the predicted proportions, SPOTlight (median JSD 0.1276) obtained comparable high results as the best performing methods RCTD (0.0361) and MuSiC weighted (0.0674, (Figure 2F). We further benchmarked SPOTlight against CoGAPS3, another factorization methods specifically designed for single-cell data (22). CoGAPS3 uses Bayesian non-negative matrix factorization for single-cell analysis with asynchronous updates and sparse data structures as well as a Markov Chain Monte Carlo allowing it to escape local maxima. Using CoGAPS3, we observed a slightly improved sensitivity, however, it underperformed seeded nNMF in the other metrics assessed (Figure 2E, F).

Deconvoluting ST derived mouse brain tissue

To validate the SPOTlight performance on complex tissue architectures, we used mouse brain sections, a thoroughly cataloged tissue, presenting well-defined structures, and a plethora of cell types and states with specific molecular fingerprints. As a reference, we used scRNA-seq datasets (Smart-seq2) derived from multiple cortical areas as well as the hippocampus (11) (∼76 000 cells and 47 annotated cell types/states; Supplementary Table S1, Supplementary Figure S3). To anatomically match the sampling site, we analyzed ST data of the adult mouse brain obtained from anterior and posterior sagittal slices (24). Two biological replicates for each slice were analyzed to test the robustness of SPOTlight predictions. To validate the predicted spatial cell type distribution within brain areas, we used canonical cell type gene markers along with *in situ hybridization* (ISH) images with cell-level resolution (25).

SPOTlight spatial deconvolution of the mouse brain ST data accurately reconstructed the layered and segmented structure of brain anatomy (Figure 3A). The predicted localization of the 47 annotated clusters confirmed their enrichment in distinct layers (e.g. cortical areas) or specific regions (e.g. hippocampus) of the mouse brain (Supplementary Figure S4 and Extended Discussion). The joint analysis of brain cell types and states resulted in a high-level segmentation, but also provided more detailed information about heterogeneity (composition) of specific areas. A closer inspection confirmed the regional enrichment of specific cell types on their known structures, confirming the high accuracy and sensitivity of the SPOTlight predictions. The results on independent anterior and posterior sections also reflected robust predictions (Supplementary Figure S5).

Illustrative examples include the SPOTlight deconvolution to delineate the spatial organization of different cortical layers, L2/3 to L6, including layer-specific neuronal subtypes (Figure 3B). Consistent with the strictly layered structure of the cortex, subpopulations aligned along stretched areas descending towards the center (L2–L6, (Figure 3C–J). L6 contributed multiple neuronal subtypes that were all accurately predicted to the respective layer substructure (Figure 3H–J). The ability to differentiate between cortical neuronal subtypes underlines the tool's sensitivity when similar cell types and states are present in complex tissues.

The hippocampus architecture was first delineated using canonical markers: Cornu Ammonis 1 stratum pyramidale (CA1sp), *Fibcd1*; Cornu Ammonis 2 stratum pyramidale (CA2sp), *Ccdc3*; Cornu Ammonis 3 stratum pyramidale (CA3sp), *Pvrl3*; and Dentate gyrus (DG), *Prox1* (33). With SPOTlight, we could clearly discern between CA1sp, CA2sp, CA3sp and the DG, which was subsequently confirmed by ISH images (Supplementary Figure S6). Gene expression measurements of cell type markers from ST alone provided noisy signals (CA1sp, CA2sp, DG) or complete absence (CA3sp) related to the sparsity of ST data; highlighting the need for more sophisticated spatial annotation tools.

Charting spatial heterogeneity in human cancer

To further validate a broader application spectrum and to test its performance in complex human tissues, we applied SPOTlight on ST data from PDAC patient samples (2), generated with a different ST protocol version than the mouse brain data (9) (Extended Discussion). Sample-matched scRNA-seq data (inDrop) was analyzed to chart the tumor composition and subsequently used to train the SPOTlight model (Figure 4A). When integrating scRNA-seq and ST (PDAC-A) (Supplementary Table S2), we observed a discrete regional enrichment of normal pancreatic and neoplastic cell types (Figure 4B). In detail, normal cell types of the pancreas were mainly excluded from the tumor fraction and further split into acinar and ductal areas. Centroacinar ductal populations appeared in the duct epithelium, while terminal ductal populations were found in both duct epithelium as well as co-localizing in the cancerous part of the tissue (Supplementary Figure S7). In line with previous results (2), we detected the intermixing of two distinct tumor cell clones and the enrichment of a ductal population with a hypoxia gene signature in the cancerous region (Figure 4C).

To shed light on the distribution of immune cells in the tumor sections, we integrated, clustered, and annotated an ex-

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

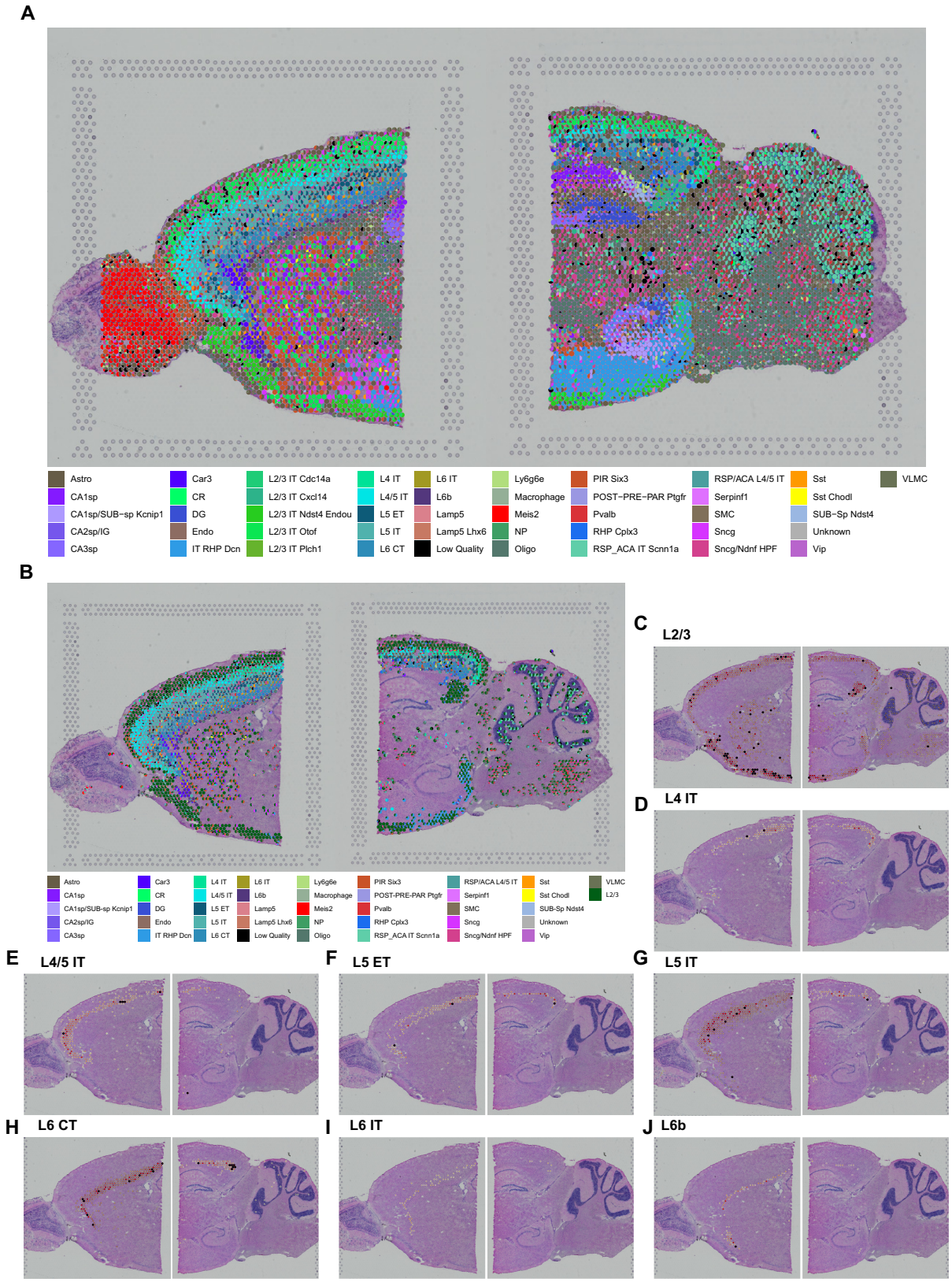


Figure 3. Cell type mapping on sagittal adult mouse brain anterior and posterior slices. (**A**) Spatial scatter pie plot representing the proportions of the cells from the reference atlas within capture locations in the adult mouse brain; we can observe the substructures of anatomical regions in the brain as defined by their specific cell types. (**B**) Proportions of the cortical cells from the reference atlas within capture locations; SPOTlight is able to capture the cortical structure being able to discern between highly similar neuronal cell types. (**C–J**) Proportion within each capture location of each specific cortical neuron type.

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

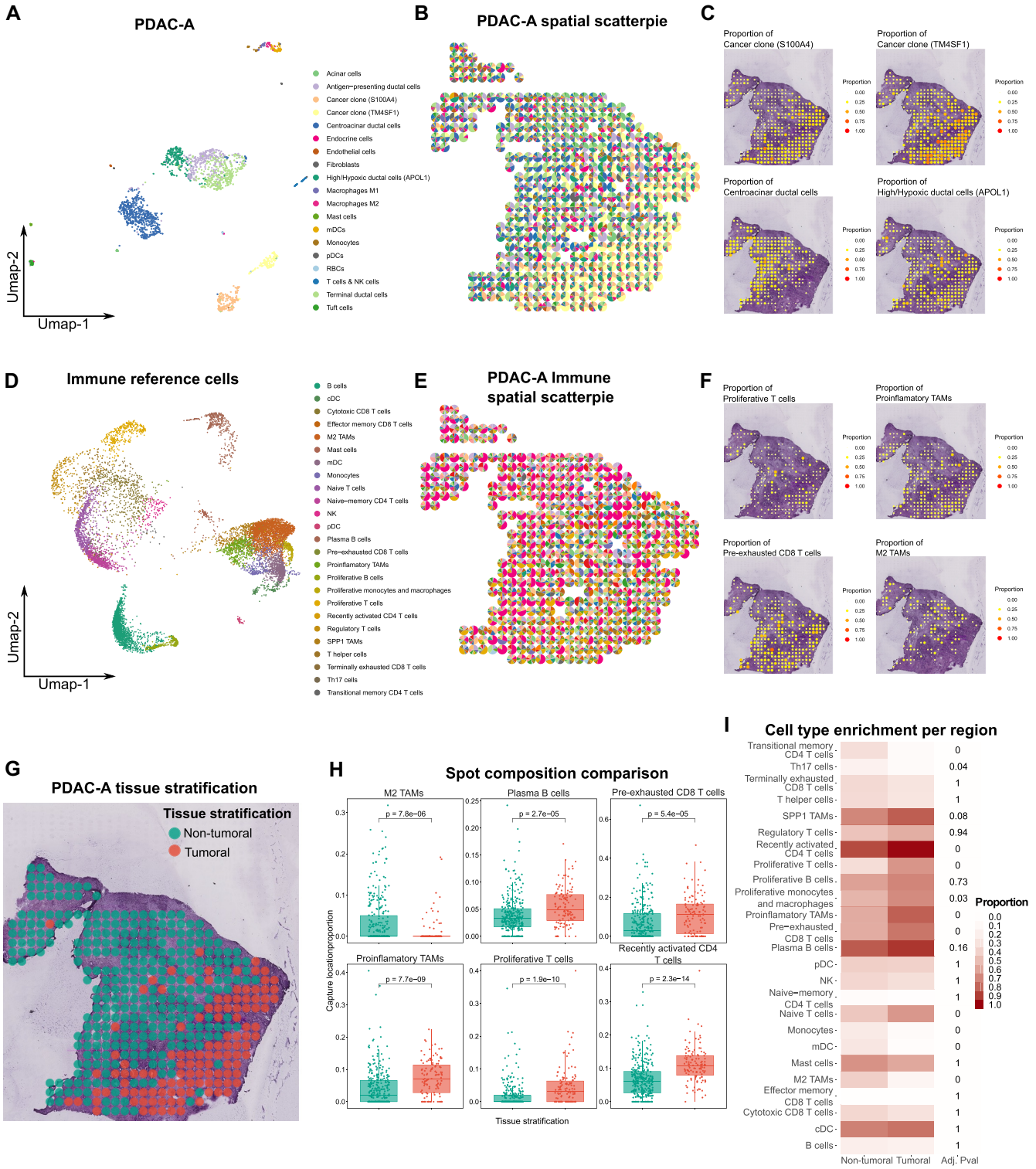


Figure 4. Mapping cell subpopulations across the tissue, charting tumoral and immune cell distribution on the tissue to identify differential immune microenvironments in tumoral versus non-tumoral regions. (**A**) UMAP projections of 1926 cells from PDAC-A, paired data from tissue slices. Cells are colored and labelled according to the cell type annotations from the original paper. (**B**) Spatial scatter pie plot representing the proportions of the cell types in the paired inDrop dataset within the capture locations. (**C**) Predicted proportion within each capture location for cancer clones S100A4 and TM4SF1 and centroacinar and hypoxic ductal cells. (**D**) UMAP projections of pancreatic immune reference cells mapped onto PDAC-A ST1. (**E**) Spatial scatter pie plot representing the proportions of the immune cells within the capture locations. (**F**) Predicted proportion within each capture spot for proliferative T-cells, pre-exhausted CD8 cells as well as proinflammatory and M2 TAMs. (**G**) Tissue stratification by tumoral - non-tumoral capture locations, stratification coincides with pathologist's annotation. (**H**) Cell type proportion comparison within each spot between tumoral and non-tumoral sections. (**I**) Proportion of capture locations containing each immune cell type within the tumoral and non-tumoral sections.

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

ternal single-cell PDAC dataset with a specific focus on the tumor immune microenvironment (34). Briefly, scRNA-seq data from 24 PDAC patients and 41 986 cells were merged to identify a total of 10 623 immune cells (Figure 4D). Clustering and curated annotation resulted in 22 immune subpopulations with 12 T-cell, 3 macrophage/monocyte, 2 B-cell, 4 dendritic and 1 MAST cell clusters. SPOTlight trained on PDAC immune cells and applied on the PDAC-A ST slides resulted in a remarkable local enrichment of tumor-specific cell states (Figure 4E, F and Supplementary Figure S8). In line with the regional distribution of normal and cancer cells, we identified a striking segmentation of immune cell states in the PDAC section (Figure 4G and Supplementary Figure S9). While anti-inflammatory M2 TAMs and transitional memory CD4 T-cells were enriched in the normal pancreas tissue, recently activated CD4 and pre-exhausted CD8 T-cells as well as proliferative CD8 cells and pro-inflammatory TAMs were significantly increased in the tumor ($P < 0.01$, Figure 4H, I and Supplementary Figure S9). In a second PDAC patient section (PDAC-B) (Supplementary Table S3), recently activated CD4 and pre-exhausted CD8 T-cells again co-localized with the tumor areas, while transitional memory CD4 T-cells and M2 TAMs were depleted from that area and mainly found together with endothelial and endocrine cells (Supplementary Figure S10). Most importantly, the enrichment of recently activated CD4 cells could not be detected through their presence alone. While the PDAC-B case showed an exclusive localization to the tumor area, recently activated CD4 cells were highly abundant in all areas, but to higher proportions in the tumor in PDAC-A. This finding strongly underlines the need to sensitively deconvolute spot composition to enable precise pathology assessments. The regional differences and local immune cell enrichments further allowed us to compute cell-cell interaction networks using the cell's co-localization in the PDAC sections (Supplementary Figure S11). Such visualization underlined the concerted interaction of tumor-resident immune cells and could provide further insight into the peculiarities of tumor microenvironments.

CONCLUSION

SPOTlight proved to be a robust, accurate, and sensitive tool to determine cell-type locations and a fine-grained composition of ST spots. We showed that scRNA-seq quality can impact its performance, obtaining the best results with deeply sequenced data from complex sequencing libraries. Nevertheless, SPOTlight also returns accurate predictions with shallowly sequenced references; an important feature when using large atlas projects as a reference. We further showed that as few as 100 cells per cell-type were sufficient to train the model without prolonged computation time. Benchmarking SPOTlight against other bulk and single-cell deconvolution tools confirmed its high accuracy for detecting cell types and for predicting the composition of ST spots. Applying SPOTlight on vastly different biological scenarios, different technology versions, and using matched and external references confirmed its broad and flexible application spectrum. This makes it a universal tool to combine both pillars of the single-cell genomics field (35) and to deduce cellular function and organization *in situ*. We are particularly excited about the potentially transformative impact on pathological assessments. Using an external immune reference to delineate the localization of immune cells in tumors could be implemented in automated digital pathology systems, where query ST patient samples are screened for immune cell composition and distribution. Importantly, both features have been related to patient prognosis and (immuno-) therapy response. Thus, we foresee spatial deconvolution using SPOTlight or similar tools to have a major impact on future cancer patient management and on precision oncology.

DATA AVAILABILITY

The SPOTlight code and the analysis notebooks to reproduce the aforementioned analysis are hosted at https://github.com/MarcElosua/SPOTlight and https://github.com/MarcElosua/SPOTlight_deconvolution_analysis.

The ST and scRNA-seq data has been previously published (11,15,34) and is freely available at the Gene Expression Omnibus (GEO) under GSE133549, and GSE71585 and GSE111672 and n the Genome Sequence Archive under project PRJCA001063. Docker environments are available for R and Rstudio at Docker Hub marcelosua/spotlight_env_r:latest and marcelosua/spotlight_env_rstudio:latest respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors contributions: H.H. and M.E. designed the study. M.E. developed SPOTlight and performed ST and scRNA-seq data analyses. P.N. compiled the PDAC scRNA-seq data and performed clustering and immune cell annotation. E.M. and I.G. supported the data analysis. H.H. and M.E. wrote the manuscript. All authors read and approved the final version.

FUNDING

Ministerio de Ciencia, Innovación y Universidades [SAF2017-89109-P, AEI/FEDER to U.E.]; project (BCLLATLAS) that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme [810287]; Chan Zuckerberg Initiative (in part); Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa and the CERCA Programme/Generalitat de Catalunya; Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III, the Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement; Spanish Ministry of Science and Innovation with funds from the European Regional Development Fund (ERDF) corresponding to the 2014–2020 Smart Growth Operating Program. Funding for open access charge: Ministerio de Ciencia, Innovación y Universidades [SAF2017-89109-P, AEI/FEDER to U.E.].

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024

Conflict of interest statement. None declared.

REFERENCES

1. Berglund,E., Maaskola,J., Schultz,N., Friedrich,S., Marklund,M., Bergenstråhle,J., Tarish,F., Tanoglidi,A., Vickovic,S., Larsson,L. *et al.* (2018) Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.*, **9**, 2419.
2. Moncada,R., Barkley,D., Wagner,F., Chiodin,M., Devlin,J.C., Baron,M., Hajdu,C.H., Simeone,D.M. and Yanai,I. (2020) Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.*, **38**, 333–342.
3. Thrane,K., Eriksson,H., Maaskola,J., Hansson,J. and Lundeberg,J. (2018) Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res.*, **78**, 5970–5979.
4. Maniatis,S., Äijö,T., Vickovic,S., Braine,C., Kang,K., Mollbrink,A., Fagegaltier,D., Andrusivová,Ž., Saarenpää,S., Saiz-Castro,G. *et al.* (2019) Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, **364**, 89–93.
5. Asp,M., Giacomello,S., Larsson,L., Wu,C., Fürth,D., Qian,X., Wärdell,E., Custodio,J., Reimegård,J., Salmén,F. *et al.* (2019) A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, **179**, 1647–1660.
6. Maynard,K.R., Collado-Torres,L., Weber,L.M., Uytingco,C., Barry,B.K., Williams,S.R., Catallini,J.L., Tran,M.N., Besich,Z., Tippani,M. *et al.* (2020) Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. bioRxiv doi: https://doi.org/10.1101/2020.02.28.969931, 28 February 2020, preprint: not peer reviewed.
7. Ståhl,P.L., Salmén,F., Vickovic,S., Lundmark,A., Navarro,J.F., Magnusson,J., Giacomello,S., Asp,M., Westholm,J.O., Huss,M. *et al.* (2016) In: *Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics*. American Association for the Advancement of Science.
8. 10x Genomics (2019) Visium Spatial Transcriptomics.
9. Vickovic,S., Eraslan,G., Salmén,F., Klughammer,J., Stenbeck,L., Schapiro,D., Äijö,T., Bonneau,R., Bergenstråhle,L., Navarro,J.F. *et al.* (2019) High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods*, **16**, 987–990.
10. Rodriques,S.G., Stickels,R.R., Goeva,A., Martin,C.A., Murray,E., Vanderburg,C.R., Welch,J., Chen,L.M., Chen,F. and Macosko,E.Z. (2019) Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, **363**, 1463–1467.
11. Tasic,B., Menon,V., Nguyen,T.N., Kim,T.K., Jarsky,T., Yao,Z., Levi,B., Gray,L.T., Sorensen,S.A., Dolbeare,T. *et al.* (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.*, **19**, 335–346.
12. Kotliar,D., Veres,A., Nagy,M.A., Tabrizi,S., Hodis,E., Melton,D.A. and Sabeti,P.C. (2019) Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife*, **8**, e43803.
13. Pascual-Montano,A., Carazo,J.M., Kochi,K., Lehmann,D. and Pascual-Marqui,R.D. (2006) Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**, 403–415.
14. Gaujoux,R. and Seoighe,C. (2010) A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, **11**, 367.
15. Mereu,E., Lafzi,A., Moutinho,C., Ziegenhain,C., McCarthy,D.J., Álvarez-Varela,A., Batlle,E., Sagar, Grün,D., Lau,J.K. *et al.* (2020) Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.*, **38**, 747–755.
16. Parekh,S., Ziegenhain,C., Vieth,B., Enard,W. and Hellmann,I. (2018) zUMIs - a fast and flexible pipeline to process RNA sequencing data with UMIs. *GigaScience*, **7**, giy059.
17. Wang,X., Park,J., Susztak,K., Zhang,N.R. and Li,M. (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.*, **10**, 380.
18. Newman,A.M., Steen,C.B., Liu,C.L., Gentles,A.J., Chaudhuri,A.A., Scherer,F., Khodadoust,M.S., Esfahani,M.S., Luca,B.A., Steiner,D. *et al.* (2019) Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.*, **37**, 773–782.
19. Gong,T. and Szustakowski,J.D. (2013) DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics*, **29**, 1083–1085.
20. M,D., A,T., E,U., Y,L., Cm,P., F,Z. and J,Y. (2021) SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.*, **22**, 416–427.
21. Cable,D.M., Murray,E., Zou,L.S., Goeva,A., Macosko,E.Z., Chen,F. and Irizarry,R.A. (2020) Robust decomposition of cell type mixtures in spatial transcriptomics. bioRxiv doi: https://doi.org/10.1101/2020.05.07.082750, 08 May 2020, preprint: not peer reviewed.
22. Sherman,T.D., Gao,T. and Fertig,E.J. (2020) CoGAPS 3: Bayesian non-negative matrix factorization for single-cell analysis with asynchronous updates and sparse data structures. *BMC Bioinformatics*, **21**, 453.
23. Allen Institute for Brain Science. Cell Types Database: RNA-Seq Data.
24. 10x Genomics (2019) Public Datasets: Spatial Gene Expression.
25. Lein,E.S., Hawrylycz,M.J., Ao,N., Ayres,M., Bensinger,A., Bernard,A., Boe,A.F., Boguski,M.S., Brockway,K.S., Byrnes,E.J. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.
26. Nieto,P., Elosua-Bayes,M., Trincado,J.L., Marchese,D., Massoni-Badosa,R., Salvany,M., Henriques,A., Mereu,E., Moutinho,C., Ruiz,S. *et al.* (2020) A single-cell tumor immune atlas for precision oncology. bioRxiv doi: https://doi.org/10.1101/2020.10.26.354829, 26 October 2020, preprint: not peer reviewed.
27. Azizi,E., Carr,A.J., Plitas,G., Cornish,A.E., Konopacki,C., Prabhakaran,S., Nainys,J., Wu,K., Kiseliovas,V., Setty,M. *et al.* (2018) Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell*, **174**, 1293–1308.
28. Yost,K.E., Satpathy,A.T., Wells,D.K., Qi,Y., Wang,C., Kageyama,R., McNamara,K.L., Granja,J.M., Sarin,K.Y., Brown,R.A. *et al.* (2019) Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat. Med.*, **25**, 1251–1259.
29. Li,H., van der Leun,A.M., Yofe,I., Lubling,Y., Gelbard-Solodkin,D., van Akkooi,A.C.J., van den Braber,M., Rozeman,E.A., Haanen,J.B.A.G., Blank,C.U. *et al.* (2019) Dysfunctional CD8 T cells form a proliferative, dynamically regulated compartment within human melanoma. *Cell*, **176**, 775–789.
30. Cao,J., Spielmann,M., Qiu,X., Huang,X., Ibrahim,D.M., Hill,A.J., Zhang,F., Mundlos,S., Christiansen,L., Steemers,F.J. *et al.* (2019) The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, **566**, 496–502.
31. Han,X., Zhou,Z., Fei,L., Sun,H., Wang,R., Chen,Y., Chen,H., Wang,J., Tang,H., Ge,W. *et al.* (2020) Construction of a human cell landscape at single-cell level. *Nature*, **581**, 303–309.
32. Svensson,V., Beltrame,E., da,V. and Pachter,L. (2019) Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. bioRxiv doi: https://doi.org/10.1101/762773, 09 September 2019, preprint: not peer reviewed.
33. Cembrowski,M.S., Wang,L., Sugino,K., Shields,B.C. and Spruston,N. (2016) Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons. *eLife*, **5**, e14997.
34. Peng,J., Sun,B.F., Chen,C.Y., Zhou,J.Y., Chen,Y.S., Chen,H., Liu,L., Huang,D., Jiang,J., Cui,G.S. *et al.* (2019) Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.*, **29**, 725–738.
35. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M. *et al.* (2017) The human cell atlas. *eLife*, **6**, e27041.

Downloaded from https://academic.oup.com/nar/article/49/9/e50/6129341 by guest on 23 April 2024