CellPie: a fast spatial transcriptomics topic discovery method via joint factorization of gene expression and imaging data

Sokratia Georgaka[1,*], William Geraint Morgans[1], Qian Zhao[1], Diego Sanchez Martinez[2], Amin Ali[1,2,3], Mohamed Ghafoor[1], Syed-Murtuza Baker[1], Robert Bristow[1,2], Mudassar Iqbal[1], and Magnus Rattray[1,*]

[1]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester M13 9PL, UK
[2]Cancer Research UK Manchester Institute, Manchester
[3]The Christie NHS Foundation Trust, Manchester

Motivation: Spatially resolved transcriptomics has enabled the study of expression of genes within tissues while retaining their spatial identity. The lack of single-cell resolution for most of the current high-throughput spatial transcriptomics technologies led to the development of *in-silico* methods, to disentangle the spatial profiles of individual cell-types. However, most current approaches ignore useful information from associated imaging data that can help to better resolve cell-types or spatial domains.

Results: We present *CellPie*, a fast, reference-free topic modelling method, based on joint non-negative matrix factorisation between spatial RNA transcripts and histological or molecular imaging features. This synergy of the two modalities can lead to improved single-cell deconvolution and spatial clustering. We assessed *CellPie* in two different tissues and imaging settings, showing an improved accuracy against published deconvolution and clustering methods. In addition, in terms of computational efficiency, *CellPie* outperforms all tested deconvolution methods by at least two orders of magnitude, without the use of GPUs.

Availability:https://github.com/ManchesterBioinference/CellPie

Contact: sokratia.georgaka@manchester.ac.uk

Spatial transcriptomics | Joint non-negative matrix factorisation | data integration | deconvolution

Introduction

In multi-cellular organisms, tissues are complex systems composed of millions of cells, which constitute the building blocks of whole organs. Within tissues, cells vary in type and activity, and their development and function are influenced by interactions with their surroundings. Therefore, dissecting spatial cellular organisation and heterogeneity within the tissue is important for understanding normal tissue function, as well as disease, which often have spatial origins (1).

Cutting-edge technologies, such as single-cell RNA sequencing (scRNA-seq), achieve high-throughput and high-resolution gene expression profiles, which provide us with a powerful insight into the characterisation of the heterogeneity at the transcriptomic level (2, 3). However, due to tissue dissociation, these methods are unable to retain the spatial context of the molecules within the tissue. Array-based spatially resolved transcriptomics methods, such as spatial transcriptomics (ST) (4), which is commercially available as Visium by 10x Genomics, and Slide-seq (5) (Slide-seq V2), allow for molecular profiling while retaining the spatial information of the tissue, at different resolution (6, 7). The barcoded Slide-seq beads ('pucks') provide near single-cell spatial resolution of $10\mu m$ while Visium barcoded spots come with a coarser resolution of $55\mu m$. This means that, depending on the tissue density, each bead/spot is capturing $2-3$ cells for Slide-seq or around 10 cells for Visium.

Disentangling the spatial cell-type composition of each barcoded spot is a key computational challenge in spatial transcriptomics data analysis. Current deconvolution methods are divided between reference-based and reference-free methods. Reference-based deconvolution methods, such as cell2location (8), stereoscope (9), SPOTlight (10) and DestVI (11) require single-cell data, ideally from the same tissue, which is not always available. Alternatively there are reference-free deconvolution methods such as stDeconvolve (12).

Another important limitation of both reference-based and reference-free deconvolution methods is that they solely rely on spot-level gene expression count data, failing to take into account any morphological image features already available through paired histology images. These image-derived features could lead to more accurate deconvolution estimates. A recent Bayesian deconvolution method, GIST (13), attempts to incorporate a cell-type specific informative image-derived prior of specific cell-type abundance estimates. However, GIST does not explicitly use morphological image features extracted from paired histopathology hematoxylin and eosin (H&E) images. Moreover, the computational burden of the aforementioned methods is typically high and in some instances GPUs are essential.

To address these limitations, we present *CellPie*, a flexible and fast reference-free deconvolution tool based on unsupervised multi-modal Non-negative Matrix Factorisation (NMF). NMF-based methods have become popular in the single-cell genomics field because they can be used to discover interpretable sparse features from high-dimensional data. NMF has previously been applied to ST data via a spatially aware model with a Gaussian process prior over the factors in (14, 15).

CellPie jointly models spatial molecular data and paired his-

tological or molecular imaging features, in a simple and computationally efficient way. The output of *CellPie* is a spatial map of spot-wise cell-type abundances (topics) and topic-specific feature (gene) scores which can be used for topic annotation and downstream analysis (see Fig. 1).

We evaluated the ability of *CellPie* to provide a spatial map of cell-types in a coronal section of an adult mouse brain as well as in a human invasive prostate carcinoma sample from Visium platform. Moreover, in the second validation, we demonstrate that by clustering the inferred topic vectors (cell-type proportions) from *CellPie*, we achieve more precise spatial clusters compared to other established clustering methods.

Materials and methods

Joint Non-negative Matrix Factorisation. *CellPie* employs a scalable NMF-based topic modelling approach, which offers an unsupervised joint, low-dimensional representation of multiple modalities. Our approach builds on efficient single matrix factorisation scheme proposed in (16) which has been recently developed for joint factorisation of multi-omics data (intNMF (17)). Here, we adapt the method to jointly factorise spatial gene expression and histology/molecular imaging data. Compared to other decomposition methods, such as Principal Component Analysis (PCA) (18), the element-wise non-negativity constraint of NMF allows for an easier and more intuitive interpretation of the resulting lower rank factors.

The joint factorisation problem for spatial gene expression and image features data is formulated (as in intNMF) as follows:

Given two non-negative input matrices $Y_{\text{rna}} \in {\mathbb{R}_+}^{m \times n}$ and $Y_{\text{img}} \in {\mathbb{R}_+}^{m \times f}$ where m is the number of spatial locations (spots), n is the number of genes and f the number of image features, building on the intNMF formulation, *CellPie* seeks for a common non-negative matrix $W \in {\mathbb{R}_+}^{m \times k}$ and two individual non-negative $H_{\text{rna}} \in {\mathbb{R}_+}^{k \times n}$ and $H_{\text{img}} \in {\mathbb{R}_+}^{k \times f}$ matrices, such that:

$$\begin{aligned} Y_{\text{rna}} &\approx W H_{\text{rna}}, \\ Y_{\text{img}} &\approx W H_{\text{img}}, \end{aligned} \qquad \textbf{(1)}$$

for gene expression counts and image feature data respectively. $k < m, n, f$ is an integer number, which represents the number of topics and must be specified *a priori*.

In *CellPie*, each spot in the gene expression matrix is normalised by the total counts over all genes.

The factors W, H_{rna} and H_{img} are calculated by solving the following optimisation problem:

$$\min_{W, H_{\text{rna}}, H_{\text{img}}} \alpha \|Y_{\text{rna}} - W H_{\text{rna}}\|_F^2 + (2-\alpha)\|Y_{\text{img}} - W H_{\text{img}}\|_F^2 ,$$

$$\text{s.t. } W, H_{\text{rna}}, H_{\text{img}} \geq 0 , \qquad \textbf{(2)}$$

where the parameter α assigns a weight to the cost function of each modality (default is set to 1.0 so that each modality is equally weighted) and $\|\cdot\|_F$ is the Frobenius norm. The above non-convex optimisation problem is reduced to an alternating pair of convex optimisations through iterative updates. The accelerated hierarchical least squares (acc-HALS) algorithm (16) is adapted to jointly factorise two matrices, see (17) for details.

The image-based features (Y_{img}) are extracted from H&E or molecule-based (e.g. immunofluorescence) images using *Squidpy* (19). *Squidpy* offers a variety of functions for image feature extraction, including *summary*, *histogram* and *texture*. Specifically, image features are calculated for each spatial location (Visium spot) resulting in a $m \times f$ feature matrix. *CellPie* by default uses *histogram* features for the joint factorisation, however there is an option for a user-defined matrix of image features. Spot-level pixel intensity features (bin-counts) are calculated for each image channel, over the whole image range, reflecting both the tissue architecture as well as cell abundances (see overview of *CellPie* approach in Fig. 1).

Model Selection. To help the user selecting a representative number of topics, k, *CellPie* implements model selection based on the elbow/knee point of the loss as function of topic number (using the "Kneedle" algorithm) (20). However, it is worth pointing out that prior biological knowledge of the tissue is always advantageous and, for optimal results, the rank, k, should be selected in conjunction with this prior knowledge (if available).

Initialisation. Due to the iterative nature of the intNMF algorithm, initialisation of the W,H_{rna} and H_{img} factors is required. Good initial values are essential to guarantee a successful NMF decomposition. Various initialisation methods have been proposed for NMF, including random and non-random strategies. Using random initialisation can lead to convergence at local minima, making it essential to perform multiple restarts to find the best initialisation. An effective, non-random strategy, proposed in (21), is the Non-negative Double Singular Value Decomposition (NNDSVD). This algorithm uses two Singular Value Decomposition (SVD) processes and has been demonstrated to rapidly reduce the approximation error. *CellPie* uses NNDSVD as the default initialisation method. Initialisation methods based on clustering (k-means, Fuzzy C-means or Hierarchical) are also implemented in *CellPie* although, similar to random initialisation methods, clustering methods usually require multiple restarts.

Datasets. Most deconvolution methods use synthetic ST data for validation, as the ground-truth is known for this data. However, it is not obvious how to generate realistic synthetic ST data with paired with synthetic image features, as required in this context. Therefore, to assess the performance of *CellPie*, two benchmarking cases are considered. In the first instance, we use a publicly available 10x Visium adult mouse brain dataset with immunofluorescence (IF) staining. This dataset was chosen because the IF staining of neuronal and glial cells, can serve as ground-truth, facilitating quantitative comparison among various published

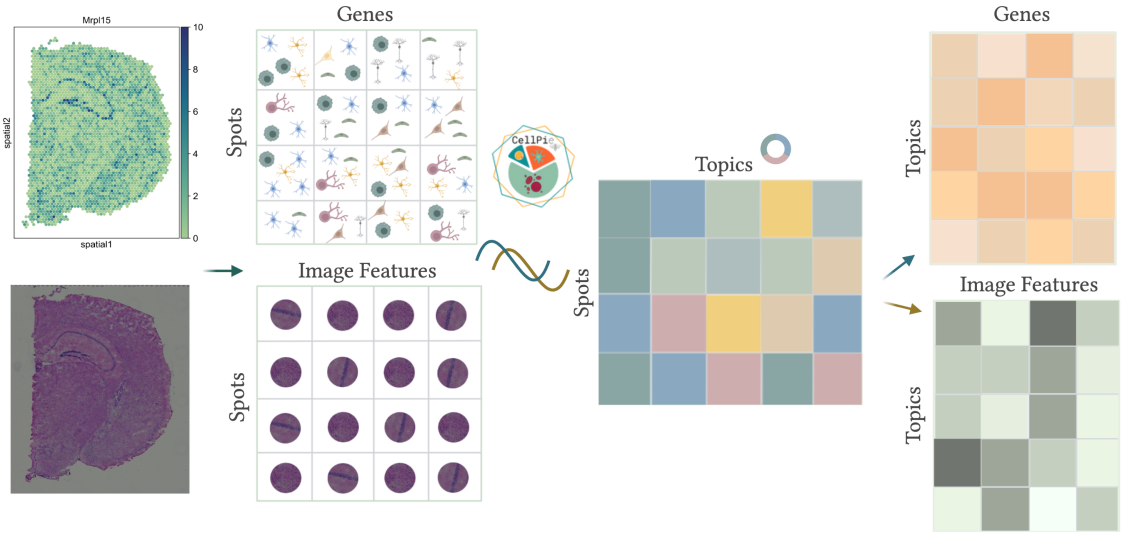

Fig. 1. Graphical overview of *CellPie* method. *CellPie* takes as input spatial gene expression counts (spots by genes) and paired histological or immunofluorescence image features (spots by features) matrices. These two modalities are jointly factorised using joint non-negative matrix factorisation, resulting to three matrices: a shared spots by topics matrix, containing topic (cell-type associated) proportions and two individual matrices, a gene loading and an image loading matrix.

deconvolution methods. The second validation uses another publicly available 10x Genomics Visium dataset of human invasive prostate carcinoma. In this case, the ground-truth is established through annotations by a pathologist. We evaluate the capability of *CellPie* to find topics that resemble the areas annotated by the pathologist and to pinpoint topic specific marker genes.

Parameter settings for competing methods.

stDeconvolve is a reference-free deconvolution method based on latent Dirichlet allocation (LDA). We execute *stDeconvolve* with the default parameters: 'cleancounts' with min.lib.size 100 and min.reads 10, 'restrictCorpus' with removeAbove 1, removeBelow 0.05 and nTopOD 1000. Then we ran the fitLDA and optLDA function with 17 topics. Finally, we filtered out cell-types using the 'getBetaTheta' with perc.filt 0.01 and betaScale 1000. Since *stDeconvolve* is a reference-free deconvolution method, similarly to *CellPie*, we seek the inferred topics that best correlate with the available ground-truth.

Stereoscope is a reference-based deconvolution method which uses probabilistic inference based on negative binomial distribution to infer cell-type proportions. We ran *Stereoscope* on GPUs with the default settings:'max_epochs' 100 for 'RNAStereoscope' and 'max_epochs' 2000 for 'SpatialStereoscope' functions.

Cell2Location is another reference-based spatial method based on a Bayesian model where absolute and relative abundances of cell-types are estimated by linearly decomposing the spatial gene expression matrix into a set of pre-estimated single-cell signatures. We first computed the average gene expression in the single-cell reference dataset (signature) using 'compute_cluster_averages' function. Then we ran Cell2Location using GPUs with the following settings: 'N_cells_per_location' 10, detection_alpha 20, max_epochs 2000, 'train_size' 1. We extracted the 'q05_cell_abundance_w_sf' abundances.

To estimate the Spearman correlation between the ground-truth and the inferred proportions, for both **Stereoscope** and **Cell2Location** we aggregated all the neuronal and glial subtype proportions/abundances into one neuronal and one glial proportion/abundance data-frame.

SpaGCN (22) is a spatial transcriptomics downstream analysis tool which uses a graph convolutional network to integrate gene expression, spatial information and histology. We used **SpaGCN** for clustering in theinvasive prostate carcinoma dataset. We executed **SpaGCN** with the default settings: For the adjacent matrix, $s = 1$ and $b = 49$. Hyperparameter $p = 0.5$, 'search_l' function with start 0.01, end 1000, top 0.01 and max_run 100. For clustering, we used n_clusters 5 based on which the recommended res is 0.3.

stLearn (23) is another tool for spatial transcriptomics downstream analysis which integrates spatial location, image and gene expression information using graph-based clustering. We used **stLearn** for clustering of the invasive prostate carcinoma dataset. We ran **stLearn**'s preprocessing, tilling and clustering analysis with all the arguments following the 'stSME clustering' tutorial provided in stlearn.readthedocs.io.

Results

Coronal section of mouse brain with IF staining. To benchmark the performance of *CellPie* against other published deconvolution methods (reference-free and reference-based), a publicly available 10x Genomics Visium dataset of a coronal section of an adult mouse brain was used (24). Immunofluorescence staining was performed on the tissue section, against anti-NeuN and GFAP antibodies (and DAPI), which represent neuronal and glial specific protein markers, respectively (Fig. 2A). To derive an independent ground-truth, spot-wise average pixel intensity values were extracted for each of the three channels in the immunofluorescence image (Fig. 2B, Fig. 2C), using *squidpy*.

To emphasize the significance of integrating imaging fea-

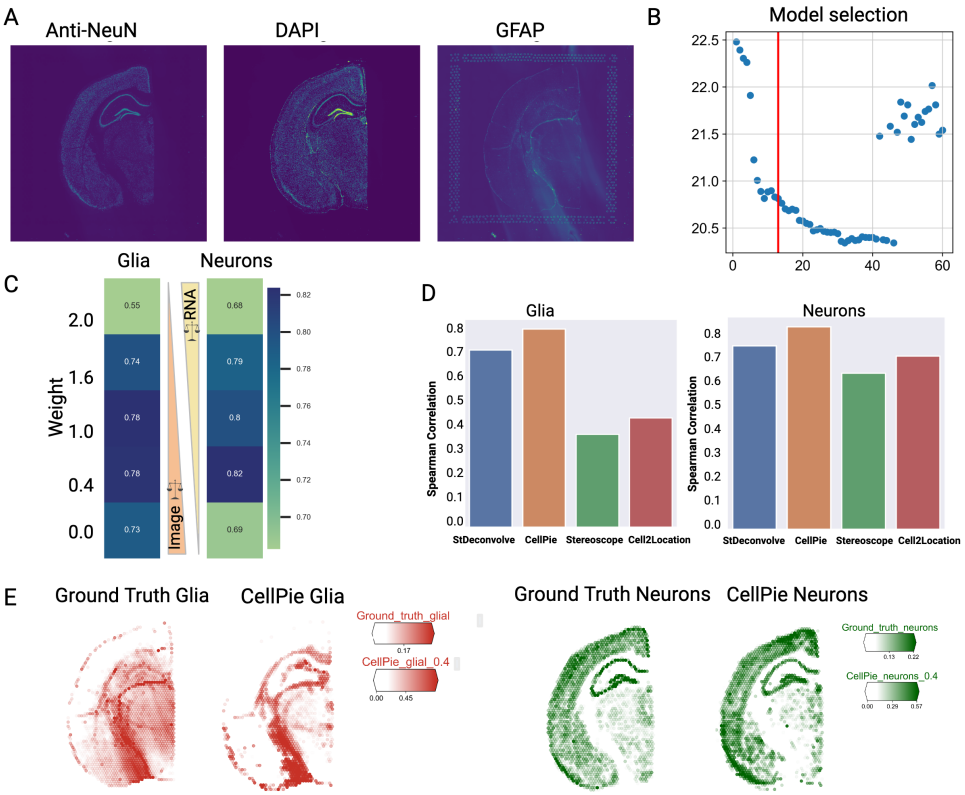

Fig. 2. Validation of *CellPie* using Mouse brain ST data and IF staining, (A) Immunofluorescence image of adult mouse brain coronal section (24), stained with NeuN, GFAP and DAPI. (B) Model selection based on the elbow point of the loss function, where the optimal number of topic found to be $k = 13$. (C) Spearman correlation between *CellPie* and ground-truth for glial and neuronal cell-types, for a range of modality weights ($\alpha = 0.0, 0.4, 1.0, 1.6, 2.0$). (D)Left: Spearman correlation between the ground-truth and the deconvolved proportions of *stDeconvolve*, *Cell2Location*, *CellPie*, and *Stereoscope*, for glia. Right: The same but for neurons. The results reported for *CellPie* correspond to weight $\alpha = 0.4$, based on the correlations values in (C). (E) Left: Intensity values extracted using the GFAP channel of the immunofluorescence image to serve as a ground-truth of glial cell-types and proportions of glial cell-types as estimated using *CellPie*. Right: Similar but for neuronal cell-types, using the NeuN channel as ground-truth.

tures, we executed *CellPie* with $k = 13$ topics over a range of weights ($\alpha = 0, 0.4, 1.0, 1.6, 2.0$), with $\alpha = 0$ corresponding to image only and $\alpha = 2$ RNA only. These topics aim at discerning various glial and neuronal sub-types, such as astrocytes, oligodendocytes, microglial, excitatory and inhibitory neurons and more (supp Fig. 1). For a comprehensive analysis, we aggregated all topics corresponding to glial and neuronal cell-types. This was achieved by combining those topics that demonstrated the highest Spearman correlation in relation to the ground-truth (Fig. 2B,Fig. 2C).

Fig. 2D shows a comparison between *CellPie* and other published deconvolution methods, for a range of *CellPie* image/gene expression weights. When equal modality weight is used ($\alpha = 1$), *CellPie* achieves the best performance amongst the deconvolution methods benchmarked in this study for both neuronal (Spearman correlation 0.82) and glial (Spearman correlation 0.78) cell-types. In the standard single modality NMF limit ($\alpha = 2$) where only spatial transcriptomics data are considered, the Spearman correlation is reduced to 0.68 and 0.55 for neurons and glia, respectively.

Figure 3A illustrates the topics that are most correlated with glial (topics deconv_2 and deconv_4) and neuronal cell-types (topics deconv_3, deconv_7). Among *CellPie*'s outputs is a gene loading matrix, of dimension topics by genes, with a score representing importance of the genes within each topic. This matrix, in conjunction with a single-cell mouse brain cell signature dataset (8) are input into the scanpy.tl.score_genes function from the Scanpy package (25, 26) (with default parameters). This function requires the input of marker genes. To find marker genes for the cell-types present in the mouse brain single-cell dataset, we perform differential gene expression using the 'Wilcoxon' method. The output is a ranking of cell-types with an associated score (the difference of the mean expression of a set of genes and the mean expression of the signature genes) that could be utilised for topic interpretation and downstream analysis. A heatmap of the ranked cell-types for all the topics that are correlated to glia and neurons is shown in Fig. 2B. The expressed cell-types are in keeping with their corresponding topics.

In order to verify whether the topics obtained in the mouse brain have biological significance, we used the top 100 marker genes of each topic (*CellPie*'s gene loading matrix) to perform GO enrichment analysis. The results were consistent with our expectations for cell-type distribution. For example, in Topic 5 and Topic 7, the enriched terms predominantly focus on biological processes related to aerobic respiration and energy metabolism (Fig. 2C). This aligns with the stronger dependency of neurons on aerobic respiration, as previous studies have shown that about eighty per-

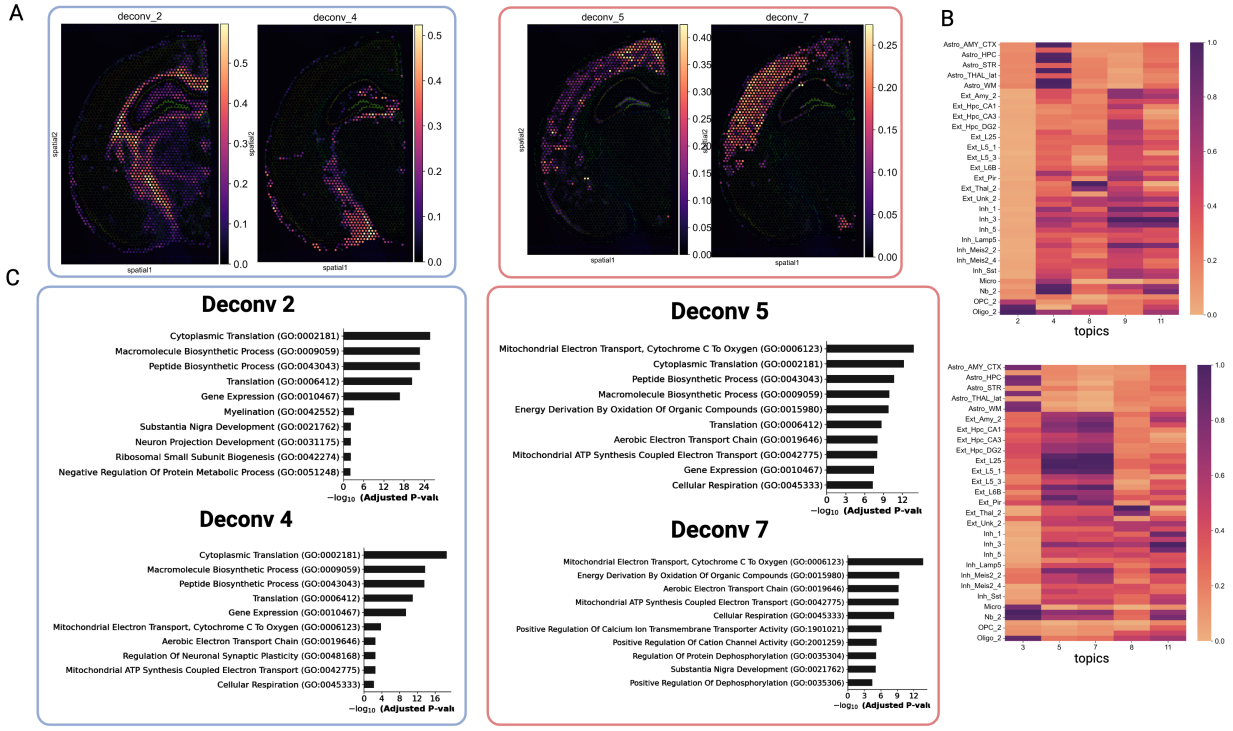

Fig. 3. Analysis of *CellPie*'s gene loading factors. (A) Top scored cell types of the topics used to synthesize the neuronal topic. The scoring has been estimated using a single-cell reference dataset from a coronal mouse brain section in conjunction with the gene list in *CellPie*'s gene loading. (B) Similar as in (A) but for glia topics.

cent of energy is consumed by neurons to maintain action potentials (27). This may suggest that the regions associated with these topics contain many active neurons. The formation and maintenance of synapses depend on interactions between neurons and glial cells (28). Topic 8 includes terms such as "Signal Release From Synapse" and "Neurotransmitter Secretion" implying the presence of a substantial number of synaptic structures in the corresponding locations. Topic 2 is enriched in terms related to "Myelination" and "Neuron Projection Development", Topic 3 is associated with Calcium signaling, and Topic 4 contains terms related to synaptic plasticity, aerobic respiration, and biosynthetic processes, which is consistent with the role of the glia cells in the formation and establishment of neuronal axons (28).

CellPie's computational efficiency is an important feature. On a Macbook Pro with 2.3GHz Quad-Core Intel Core i7 and 16GB memory, the joint NMF executed in only $23s$ for this dataset. This performance facilitates rapid, GPU-free deconvolution, while it provides flexibility for repeated model assessments.

Human prostate cancer data: CellPie topics separate Gleason 3 and Gleason 4 score regions. For our second validation, we use another publicly available 10x Visium data, this time derived from a human prostate sample with adenocarcinoma (29). 10x Genomics supplies a paired H&E image (Fig. 4A) that has been annotated by a pathologist. However, to achieve finer granularity of the invasive carcinoma region, we re-annotated the image (Fig. 4B). Consequently, the entire tumour region was annotated using the Gleason scoring system, labelled as Gleason 3 and Gleason 4. This enables us to evaluate if *CellPie* can identify topics that correspond to these specific Gleason areas.

While image evaluation using IF is relatively straightforward as the extraction of information is done by considering fluorescence intensities, H&E images require the interpretation of colour and tissue morphology. In prostate, two main lineages that are biologically and architecturally different can be seen: the stromal compartment and the epithelial compartment. The stroma is composed of fibroblasts, smooth muscle, nerves, and blood vessels in diverse proportions (30, 31). Prostate carcinoma arises from the glandular epithelial compartment and, unlike other organs, is not graded by individual cells differentiation but mainly by its architectural features using the Gleason grading, from 1 to 5. Due to poor reproducibility and lack of biological support, Gleason 1 and 2 are not reported anymore. Gleason 3 cancers comprise the most differentiated adenocarcinoma, consisting of discrete glandular units with varying sizes and shapes (32). Individual tumor acini have smooth, typically circular edges and intact basement membranes. In contrast, Gleason 4 cancers are composed of poorly formed glandular units with indistinct borders, fused glands, and irregularly infiltrating stroma (32, 33). Intuitively, Gleason 4 tumors exhibit higher degrees of differentiation, increased cancer progression, and therefore stronger prognostic correlations.

The optimal number of topics, based on model selection, was found to be $k = 11$ (Fig. 4C, supp. Fig. 3). In addition, *CellPie* was executed across a range of modality weights α spanning from 0 (image only) to 2 (RNA only). For each specific modality weight, we compared the resulting topics against the pathologist's annotated regions (ground-truth).

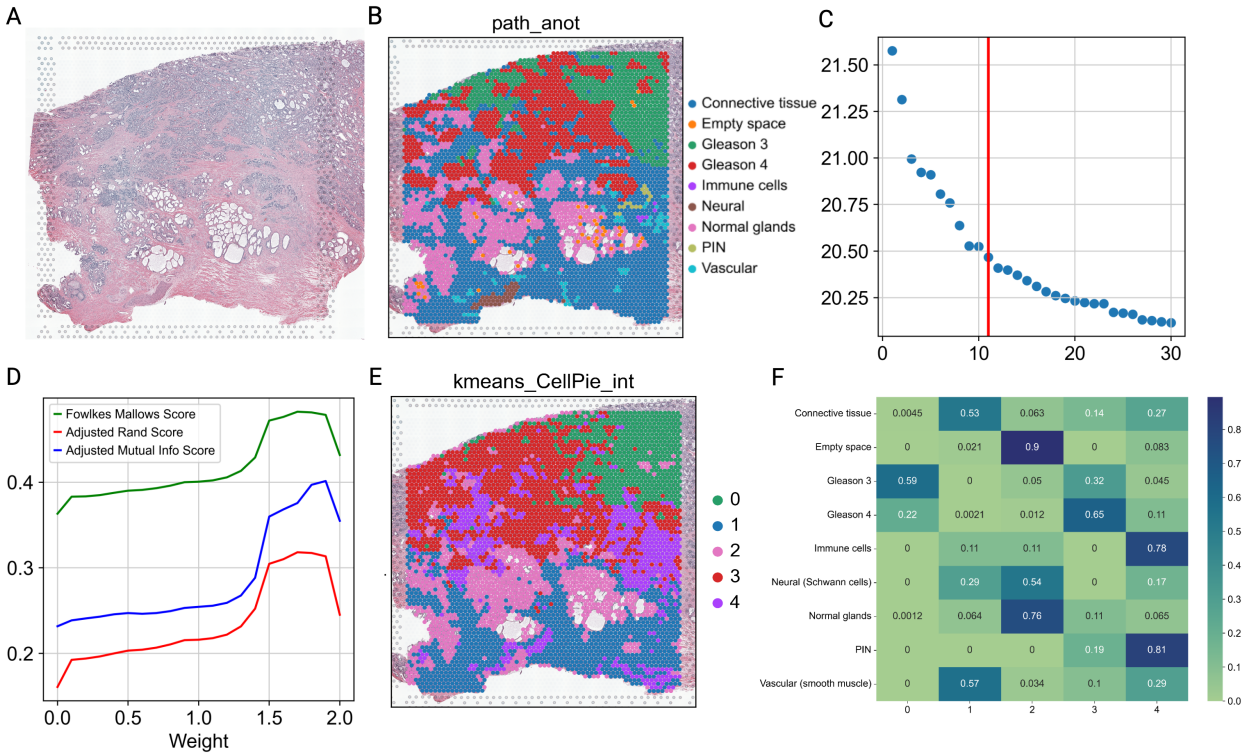

Fig. 4. Validation of *CellPie* on human prostate cancer data, (A) H&E image of human prostate tissue with adenocarcinoma (FFPE) obtained from 10x Genomics (29). (B) H&E image overlaid with pathologist annotations. (C) Model selection, where the optimal number of topics found to be $k = 11$. (D) Evaluation of *CellPie*'s topic discovery accuracy against pathologist's ground-truth across a range of different modality weights, using Fowlkes Mallows, Adjusted Rand Index and Adjusted Mutual Info scores. (D) *CellPie*'s topics, clustered using k-means algorithm with $n = 5$ clusters. (F) Crosstab heatmap showing similarity between pathologist's annotated regions and the clusters in (E).

This comparison was achieved by performing k-means clustering on the topic proportions, setting the number of clusters to 5. To assess the performance, we employed three performance metrics: Fowlkes-Mallows, Adjusted Rand Index and Mutual Information scores, as illustrated in Fig. 4D. According to this figure, peak performance is observed at a modality weight of $\alpha = 1.7$. However, a decline in accuracy is seen at $\alpha = 2.0$. This latter weight is equivalent to the standard single NMF considering only the spatial tramscriptomics data. This highlights the significance of the joint modelling of both image features and transcriptomics data.

Fig. 4C shows the relationship between the ground-truth and the *k-means* clustering of *CellPie*'s topics (Fig. 4B), using a normalised crosstab heatmap. According to the heatmap, cluster 0 is mostly associated with the Gleason 3 region, while the Gleason 4 region is linked to cluster 3. Cluster 1 is a combination of the Connective tissue, Vascular and Neural regions. Cluster 2 is mainly connected to the Normal glands and Neural areas, while cluster 4 depicts Immune cells, PIN and vascular regions.

We compared the clustering results of *CellPie* using both the optimal weight ($\alpha = 1.7$ - kmeans_CellPie_int) and the equivalent to single NMF weight ($\alpha = 2.0$ - kmeans_CellPie_0). There results were then benchmarked against the *k-means* clustering on gene expression data (within the PCA space, with 50 components) and against two published spatial transcriptomics-specific clustering methods, namely, *SpaGCN* and *stLearn*. A qualitative comparison is shown in Fig. 4D and Fig. 5A, where *CellPie*, *stlearn* and *k-means* capture both the connective tissue and normal glands areas, while in *SpaGCN*, these areas are shown as a mixture. *stlearn* is the only one which captures the nerve area as an individual cluster. However, when it comes to distinguishing between the Gleason 3 and Gleason 4 areas, none of the established methods are effective. In constrast, *CellPie* depicts these tumour regions as two separate, distinct clusters.

Figure 5B presents the Adjusted Rand Index (ARI) for the methods previously mentioned, offering a comparison of their overall clustering accuracy. *stlearn* exhibits better performance compared to *CellPie* in terms of ARI. However, when narrowing the focus to just the Gleason 3 and Gleason 4 areas, *CellPie* outperforms all the other methods (Fig. 5C), based on the precision metric. This is defined as the number of spots correctly predicted as Gleason 3 or Gleason 4 by the method over the total number of spots predicted as being in the union of Gleason 3 and Gleason 4.

To further annotate the topics with cell-types, we interrogate the *CellPie*'s gene loading matrix in conjunction with a modified cell signature obtained from (34), with the addition of markers for B-cells (35) and neurons (using scanpy.tl.score_genes function). In Fig. 6B and Fig. 6C, the topic interpretation of the 10x visium invasive prostate cancer is shown. For example, topics deconv_0, deconv_2 and deconv_6 were in keeping with the epithelia compartment based on their score. Deconv_0 represents the glandular area of the prostate covering both benign (luminal epithelial - LE)

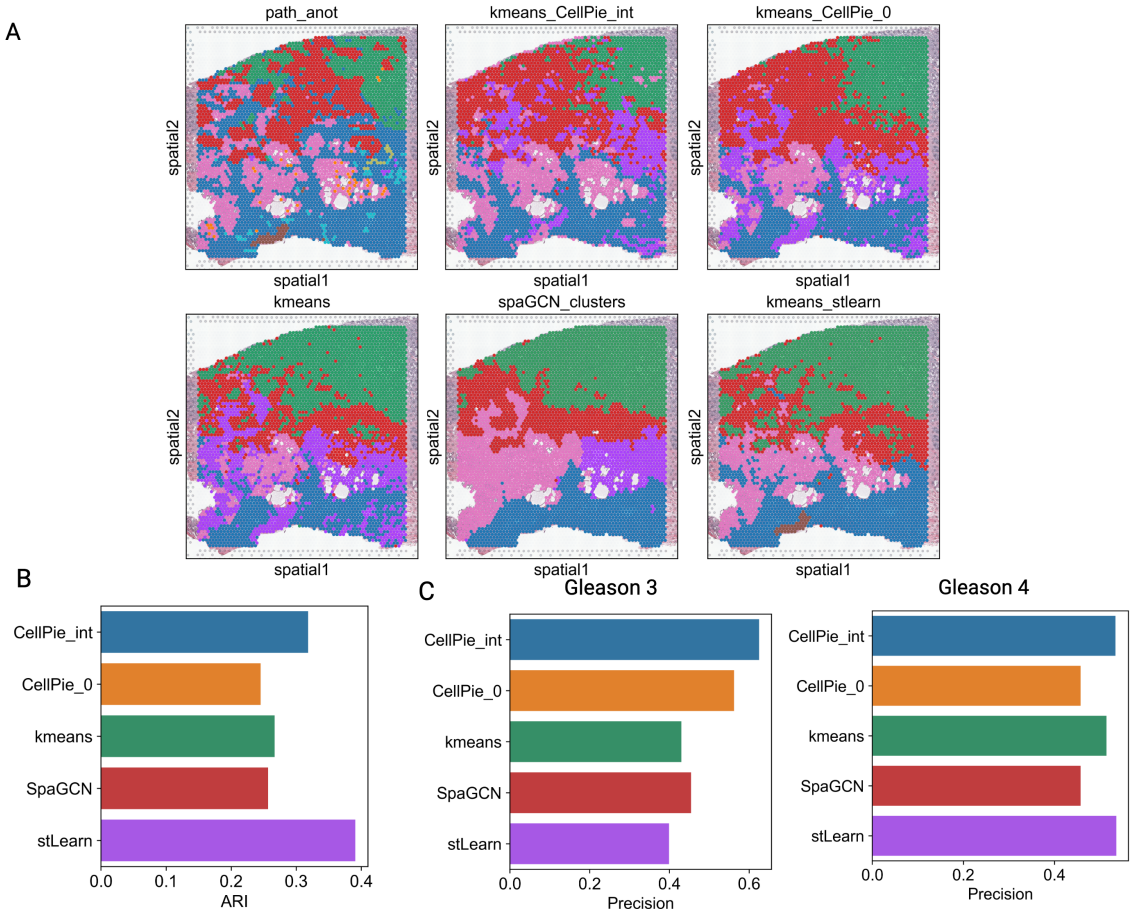

Fig. 5. Validation of *CellPie* on human prostate cancer data, (A) Comparison of *CellPie*'s topics, clustered using k-means algorithm, against other published clustering methods. The images correspond to: first row: pathologist's ground truth, *CellPie* with $\alpha = 1.7$, *CellPie* with only gene expression data ($\alpha = 2$), second row: kmeans clustering of row spatial transcriptomics counts data, *SpaGCN* clusters and *stLearn* clusters. (B) Adjusted Rand Index (ARI) for the methods in (A). (C) Precision metric for the Gleason 3 and Gleason 4 regions, defined as the number of spots correctly identified as Gleason 3 or Gleason 4 by *CellPie* and other methods, over the total number of spots identified as Gleason 3/Gleason 4 by the method.

and malignant (tumour) area, with high expression of *KLK2*, *KLK3* and *ACPP*, genes associated with prostate glandular epithelium marker. The top 3 genes in deconv_2 (*MSMB*, *ACPP* and *AZGP1* usually expressed in benign glandular tissue) and deconv_6 (*PLA2G2A*, *SPON2*, *KLK2* and *KLK3* usually expressed in prostate cancer) showed that the separation between benign and malignant tissue respectively were indeed in keeping with the pathologist's annotation. Within the stromal compartment, we showed that smooth muscle cells corresponded with deconv_1. This is in keeping with the high gene expression of *MYL9*, *ACTG2*, *ACTA2* and *TAGLN*, all involved in muscle contraction.

To further validate the biological significance of the regions identified by our method at the molecular level, we utilized Gene Ontology (GO) analysis for gene identification within the 11 topics (Fig. 6D).

Topic 6 (deconv_6) specifically exhibits high expression in Gleason 3 rather than Gleason 4. Apart from genes related to the androgen receptor (AR) pathway which is enriched in the prostatic epithelial cells, it encompasses various biological processes associated with differentiation features, such as protein translation as well as lipidic metabolism signal to prostate epithelium. Interestingly, the term "Low-Density Lipoprotein Particle Remodeling" is consistent with the recent study which found the interaction of PRSS2 with a receptor (low-density lipoprotein-related receptor protein 1) stimulated prostate tumour growth and progression (36). Conversely, Topic 0 (deconv_0), while widely distributed in cancerous tissues, shows higher expression in Gleason 4. The enriched genes in this topic, aside from the AR pathway, are predominantly involved in RNA and protein processing, including alternative splicing and protein folding. These two regulatory processes are widely considered to play crucial roles in the development of prostate cancer, such as alternative splicing of AR gene (37) and protein folding facilitated by HSPs (heatshock proteins) (38), while lacking differentiation characteristics.

Furthermore, in Topic 1 (deconv_1), which exhibits high expression in connective tissue, the enriched GO terms include "Homotypic cell-cell adhesion", "Supramolecular Fiber Organization", and "Actin Filament Organization". These terms highlight smooth muscle and collagen fibre intra and extracellular organization, which are key components of the connective tissue found in the prostate. Topic 2 (deconv_2), which shows high expression in adjacent glandular tissue, lacks enrichment in the AR signaling pathway. Instead, it

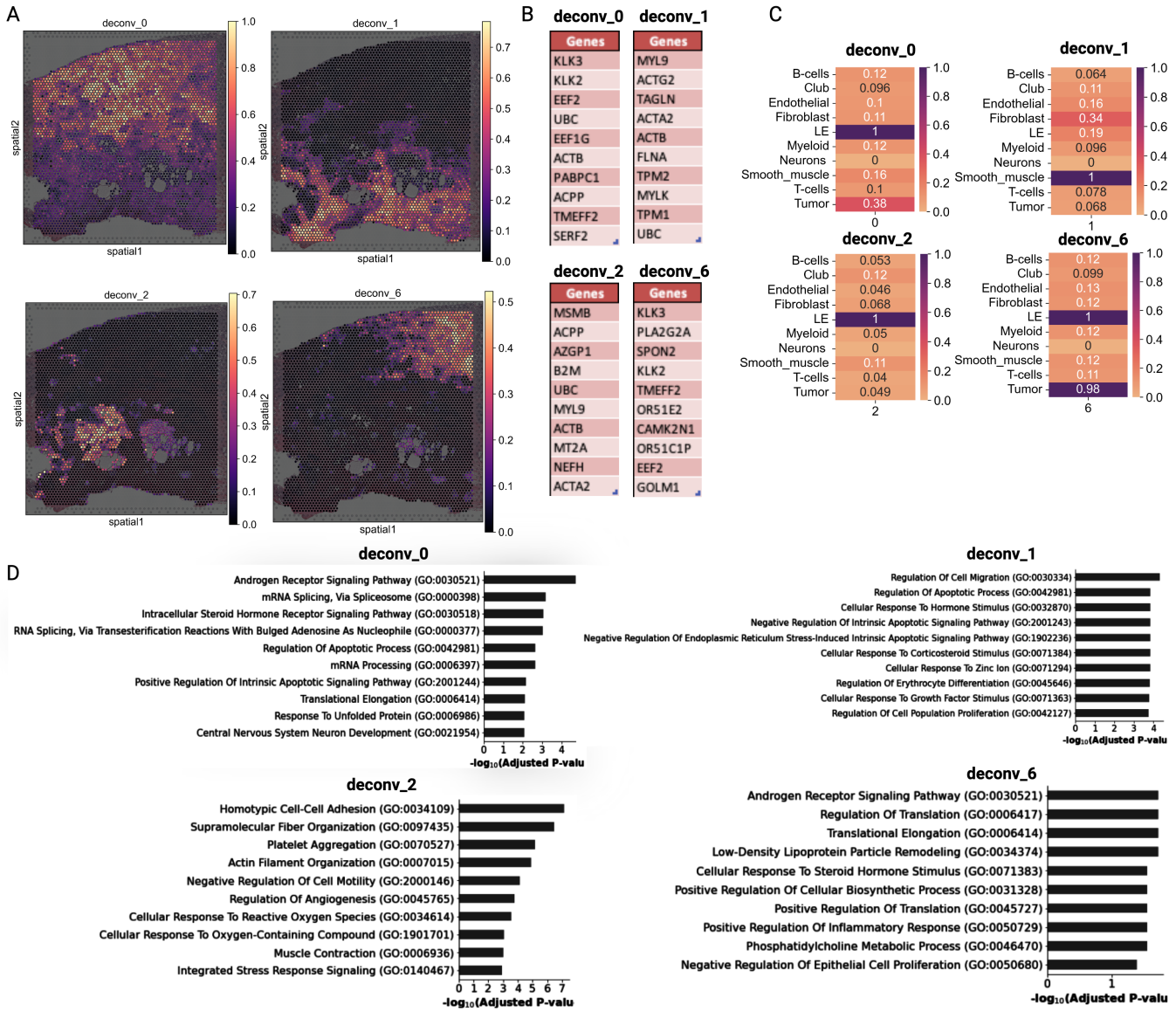

Fig. 6. Biological interpretation of the prostate cancer results: (A) Four of the eleven *CellPie* topics associated with tumour region (deconv_0, deconv_6), connective tissue (deconv_1) and normal glands (deconv_2). (B) Top ten marker genes for each of the topics present in (A). (C) Annotation of cell-types present in topics (A) using a single-cell reference dataset. (D) Gene ontology performed using the top 100 topic related markers for the topics shown in (A).

is enriched in genes related to steroid hormone response and stress signals regulating cell apoptosis, implying a potential balance between differentiation and microenvironmental changes induced by cell-cell communication.

Conclusion

In this paper, we presented *CellPie*, a method based on a rapid joint NMF framework. *CellPie* integrates spatial gene expression counts data and molecular or histological imaging features. This integration helps disentangling multi-cellular spatial transcriptomics data and aids in the identification of distinct spatial compartments. We evaluated *CellPie* using two distinct datasets derived from different species (mouse and human). These datasets are paired with different imaging settings (IF and H& E), and encompass both healthy and diseased tissues. We demonstrated that, when benchmarked for deconvolution, *CellPie*'s accuracy surpassed that of other published spatial deconvolution methods. In the application to human prostate cancer data, *CellPie* emerged as the sole method capable of distinguishing between Gleason 3 and Gleason 4 tumour regions. Moreover, analysis of gene loading matrices for both our benchmark examples were in good agreement with the available literature. By adjusting the relative weight of transcriptomics and image data, we showed how jointly analysing both modalities provides enhanced performance over considering each one in isolation.

CellPie includes image features derived from the high resolution ('hires') Visium images, which have a resolution approximately at 2000×2000 pixels. There are also images available with even higher resolutions, e.g. 27000×25000 pixels. Leveraging these ultra-high resolution images might potentially lead to a better performance. However, feature extraction from such images using *Squidpy* proves to be com-

putationally intense, requiring many hours and memory resources, thus limits *CellPie*'s performance.

One limitation to note is that *CellPie* does not explicitly model the spatial nature of the data, e.g. the neighbourhood relationships between spots. Hence, a spatially aware version of the joint NMF algorithm would be an interesting future direction. Furthermore, *CellPie* amd the intNMF algorithm (17), which *CellPie* is based on, could be extended to jointly factorise more than two matrices simultaneously. This enhancement would allow for integration of several omics datasets possesing a shared dimension (e.g. spatial proteomics).

Acknowledgments

We thank Alkmini Damkou, a PhD student in Simons lab (TUM-NCB) for our fruitful discussions on Mouse Brain results.

Funding

MR and SG are supported by a Wellcome Trust award (204832/B/16/Z). SG and MI acknowledges funding from the University of Manchester's Wellcome Institutional Strategic Support Fund (Wellcome ISSF) grant (204796/Z/16/Z).

Data availability

The *CellPie* code and the notebooks to reproduce the analysis are available as an open-source python package in https://github.com/ManchesterBioinference/CellPie. The prostate invasive carcinoma pathologist's annotations are also available on that repo. All the data used in this paper is publicly available to download from 10x Genomics website (24, 29).

1. Xiaowei Zhuang. Spatially resolved single-cell genomics and transcriptomics by imaging. *Nature methods*, 18(1):18–22, 2021.
2. Xinmin Li and Cun-Yu Wang. From bulk, single-cell to spatial rna sequencing. *International Journal of Oral Science*, 13(1):1–6, 2021.
3. Alexander F Schier. Single-cell biology: beyond the sum of its parts. *Nature Methods*, 17(1):17–20, 2020.
4. Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.
5. Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.
6. Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546, 2022.
7. Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.
8. Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology*, 40(5):661–671, 2022.
9. Alma Andersson, Joseph Bergenstråhle, Michaela Asp, Ludvig Bergenstråhle, Aleksandra Jurek, José Fernández Navarro, and Joakim Lundeberg. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications biology*, 3(1):1–8, 2020.
10. Marc Elosua-Bayes, Paula Nieto, Elisabetta Mereu, Ivo Gut, and Holger Heyn. Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic acids research*, 49(9):e50–e50, 2021.
11. Romain Lopez, Baoguo Li, Hadas Keren-Shaul, Pierre Boyeau, Merav Kedmi, David Pilzer, Adam Jelinski, Ido Yofe, Eyal David, Allon Wagner, et al. Destvi identifies continuums of cell types in spatial transcriptomics data. *Nature biotechnology*, pages 1–10, 2022.
12. Brendan F Miller, Feiyang Huang, Lyla Atta, Arpan Sahoo, and Jean Fan. Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nature communications*, 13(1):1–13, 2022.
13. Asif Zubair, Richard H Chapple, Sivaraman Natarajan, William C Wright, Min Pan, Hyeong-Min Lee, Heather Tillman, John Easton, and Paul Geeleher. Cell type identification in spatial transcriptomics data can be improved by leveraging cell-type-informative paired tissue images using a bayesian probabilistic model. *Nucleic Acids Research*, 2022.
14. F William Townes and Barbara E Engelhardt. Nonnegative spatial factorization. *arXiv preprint arXiv:2110.06122*, 2021.
15. Joseph Bergenstråhle, Ludvig Larsson, and Joakim Lundeberg. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC genomics*, 21(1):1–7, 2020.
16. Nicolas Gillis and François Glineur. Accelerated multiplicative updates and hierarchical als algorithms for nonnegative matrix factorization. *Neural computation*, 24(4):1085–1105, 2012.
17. William Geraint Morgans, Andrew Sharrocks, and Mudassar Iqbal. Scalable joint non-negative matrix factorisation for paired single cell gene expression and chromatin accessibility data. *bioRxiv*, 2023. doi: 10.1101/2023.09.25.559293.
18. Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374 (2065):20150202, 2016.
19. Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L Ibarra, Olle Holmberg, Isaac Virshup, et al. Squidpy: a scalable framework for spatial omics analysis. *Nature methods*, 19(2):171–178, 2022.
20. Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011.
21. Christos Boutsidis and Efstratios Gallopoulos. Svd based initialization: A head start for nonnegative ma-

trix factorization. *Pattern recognition*, 41(4):1350–1362, 2008.

22. Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11):1342–1351, 2021.
23. Duy Pham, Xiao Tan, Jun Xu, Laura F Grice, Pui Yeng Lam, Arti Raghubar, Jana Vukovic, Marc J Ruitenberg, and Quan Nguyen. stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *BioRxiv*, 2020.
24. 10X Genomics. https://www.10xgenomics.com/resources/datasets/adult-mouse-brain-section-2-coronal-stains-dapi-anti-gfap-anti-neu-n-1-standard-1-1-0, .
25. F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.
26. Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.
27. Renaud Jolivet, Pierre J. Magistretti, and Bruno Weber. Deciphering neuron-glia compartmentalization in cortical energy metabolism. 1:4. ISSN 1662-6427. doi: 10.3389/neuro.14.004.2009.
28. Nicola J. Allen and David A. Lyons. Glia as architects of central nervous system formation and function. 362(6411):181–185. ISSN 0036-8075. doi: 10.1126/science.aat0473.
29. 10X Genomics. https://www.10xgenomics.com/resources/datasets/human-prostate-cancer-adenocarcinoma-with-invasive-carcinoma-ffpe-1-standard-1-3-0, .
30. John E McNeal. The zonal anatomy of the prostate. *The prostate*, 2(1):35–49, 1981.
31. Gervaise H Henry, Alicia Malewska, Diya B Joseph, Venkat S Malladi, Jeon Lee, Jose Torrealba, Ryan J Mauck, Jeffrey C Gahan, Ganesh V Raj, Claus G Roehrborn, et al. A cellular anatomy of the normal adult human prostate and prostatic urethra. *Cell reports*, 25(12):3530–3542, 2018.
32. Hugh J. Lavery and Michael J. Droller. Do gleason patterns 3 and 4 prostate cancer represent separate disease states? 188(5):1667–1675. ISSN 1527-3792. doi: 10.1016/j.juro.2012.07.055.
33. Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, 40(2):244–252, 2016.
34. Hanbing Song, Hannah NW Weinstein, Paul Allegakoen, Marc H Wadsworth, Jamie Xie, Heiko Yang, Ethan A Castro, Kevin L Lu, Bradley A Stohr, Felix Y Feng, et al. Single-cell analysis of human primary prostate cancer reveals the heterogeneity of tumor-associated epithelial cell states. *Nature communications*, 13(1):1–20, 2022.
35. Isabel Heidegger, Georgios Fotakis, Anne Offermann, Jermaine Goveia, Sophia Daum, Stefan Salcher, Asma Noureen, Hetty Timmer-Bosscha, Georg Schäfer, Annemiek Walenkamp, et al. Comprehensive characterization of the prostate tumor microenvironment identifies cxcr4/cxcl12 crosstalk as a novel antiangiogenic therapeutic target in prostate cancer. *Molecular Cancer*, 21(1):1–20, 2022.
36. Lufei Sui, Suming Wang, Debolina Ganguly, Tyler P. El Rayes, Cecilie Askeland, Astrid Børretzen, Danielle Sim, Ole Johan Halvorsen, Gøril Knutsvik, Jarle Arnes, Sura Aziz, Svein Haukaas, William D. Foulkes, Diane R. Bielenberg, Arturas Ziemys, Vivek Mittal, Rolf A. Brekken, Lars A. Akslen, and Randolph S. Watnick. PRSS2 remodels the tumor microenvironment via repression of tsp1 to stimulate tumor growth and progression. 13(1):7959. ISSN 2041-1723. doi: 10.1038/s41467-022-35649-9. Number: 1 Publisher: Nature Publishing Group.
37. Ken-Ichi Takayama, Takashi Suzuki, Tetsuya Fujimura, Yuta Yamada, Satoru Takahashi, Yukio Homma, Yutaka Suzuki, and Satoshi Inoue. Dysregulation of spliceosome gene expression in advanced prostate cancer by RNA-binding protein PSF. 114 (39):10461–10466. ISSN 1091-6490. doi: 10.1073/pnas.1706076114.
38. Versha Dahiya and Gargi Bagchi. Non-canonical androgen signaling pathways and implications in prostate cancer. 1869(12):119357. ISSN 1879-2596. doi: 10.1016/j.bbamcr.2022.119357.