jazzPanda: A hybrid approach to find spatial marker genes in imaging-based spatial transcriptomics data

Xinyi Jin[1,2], Givanna H. Putri[1,2], Jinming Cheng[3], Marie-Liesse Asselin-Labat[1,2], Gordon Smyth[1,2] and Belinda Phipson[1,2]

[1] The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade VIC 3052 Australia
[2] University of Melbourne, Department of Medical Biology, 1G Royal Parade VIC 3052 Australia
[3] Centre for Biomedical Data Science, Duke-NUS Medical School, Singapore, 169857, Singapore

Abstract

Spatial transcriptomics enables the understanding of the spatial architecture of tissues, providing deeper insight into tissue structure and cellular neighbourhoods. A crucial step in the analysis of spatial data is cell type identification. In single cell RNA-sequencing (scRNA-seq) analysis, cells are clustered according to their transcriptional similarity, and marker genes for each cluster identified. Marker analysis identifies genes highly expressed in each cluster compared to the remaining clusters, and these marker genes are used to annotate clusters with cell types. For spatial data, there are limited software tools for appropriate marker gene detection methods that account for the spatial distribution of gene expression. Tools developed for scRNA-seq ignore spatial information for the cells and genes. We have developed a hybrid approach to prioritize marker genes that uses the spatial coordinates of gene detections and cells making up clusters. We propose a binning approach that effectively "pseudobulks" gene detections and cells within clusters that can then be used as input into linear models for marker analysis. Our approach can account for multiple samples and background noise. We have tested our methods on several public datasets from different platforms including Xenium, CosMx and MERSCOPE. The marker genes detected by our method show strong spatial correlation with the corresponding clusters and have increased specificity compared to other methods. The method is implemented in the jazzPanda R Bioconductor package and is publicly available (`https://bioconductor.org/packages/jazzPanda`).

Background

The next evolution in single cell sequencing is spatially resolved 'omics, which captures information about the spatial architecture of a tissue as well as measuring gene or protein expression at a specified resolution (spot, single cell, sub-cellular level). First introduced by Ståhl et al. [41], the field is developing rapidly towards sub-cellular resolution and enhanced transcript detection efficiency [13, 23]. By studying the transcriptome in the context of the spatial organization of the tissue, greater insights into the structures and their surrounding neighbours can be gained.

Spatially resolved 'omics can help with understanding complex functions and regulatory mechanisms in various tissue microenvironments. For example, a recent study used spatial technology to understand long-term memory formation in axolotl. The authors identified a novel neuron subpopulation and showed its interaction with astrocytes for long-term memory formation [42]. Spatial transcriptomics technology also lends itself to the study of the tumor microenvironment. Arora et al. integrated single cell and spatial transcriptomics data to characterise malignant cells in tumor cores and leading edge regions in oral squamous cell carcinoma. The spatial transcriptional signatures were used to model and predict disease progression and treatment efficacy [3].

Current commercialized spatial transcriptomics technologies can be categorized into two major types, imaging-based and sequencing-based technologies. Commercially available imaging-based technologies such as Xenium [22], CosMx [21], and MERSCOPE [48] take advantage of fluorescence-based microscopy to quantify transcripts and focus on a pre-designed panel of genes. Sequencing-based technologies, such as Slide-seq [39], Visium [22], and Stereo-seq [11], rely on the design of spatial indexing methods to encode the spatial location prior to sequencing. The different technologies represent a trade-off between numbers of genes profiled and cellular resolution. For example, Visium measures spot-level locations at a resolution of 55 μm, which typically captures multiple cells. Slide-seq has barcoded beads at a resolution of 10 μm in diameter. Stereo-seq (also known as STOmics) achieves 500 nm resolution for samples of varying sizes, ranging from 50 mm^2 to 174 cm^2 [11]. Imaging-based technologies can achieve subcellular resolution but are typically limited in the number of genes included in the panel. Sequencing-based technologies generally achieve whole-transcriptome profiling at a lower resolution. However, the field is developing rapidly, with Nanostring recently achieving sub-cellular resolution with a panel of 18,936 genes [23].

Spatial transcriptomics is an emerging field, and various computational tools have been developed to understand and analyse spatial data. There are a number of papers re-

viewing the development of spatial technologies [29, 45] and computational tools [26, 30]. However, gold standard analysis workflows for spatial data are still being developed and it is commonplace to use methods designed for single cell data. A typical analysis workflow for single cell RNA-seq (scRNA-seq) includes quality control, dimension reduction, clustering, cell type annotation, and other downstream analysis such as differential expression analysis between groups of samples. Aspects of this workflow can be applied to analyse spatial transcriptomics data [26], however as the data generation mechanisms are different compared to scRNA-seq data, particularly for imaging based technologies, not all computational tools will be appropriate. For example, library size normalization methods specific to single cell data can remove biological signal in spatial data [7]. Other analysis aspects are unique to spatial data, including cell segmentation and assigning gene detections to cells, identifying spatially variable genes, and local and global neighbourhood analysis.

Similarly to scRNA-seq, a crucial step in the analysis of spatial transcriptomics data is cell type annotation. There are a number of ways this can be done: clustering followed by marker gene analysis to identify genes highly expressed in each cluster compared to the remaining cells [20, 38]; cell type classification based on reference datasets or gene lists [33, 49]; label transfer approaches using well annotated single cell data [2, 40], and machine learning approaches utilizing a smaller subset of well annotated cells [47, 14]. Most, if not all, of these methods ignore the spatial information for the cells and genes, and assume the data is summarised at the single cell or spot level.

Marker genes are useful in that they can be used to annotate clusters with cell types. There are a number of tools developed to find marker genes for scRNA-seq data. Seurat's *FindMarkers* method is one popular approach that finds marker genes by performing Wilcoxon Rank Sum tests on gene expression data from two groups of cells [20]. The limitation of this approach is the inability to adjust for other factors such as batch or sample. Marker gene detection approaches originally developed for scRNA-seq data have been used for the analysis of spatial transcriptomics studies [12], although there has not been thorough benchmarking studies done to show that these methods perform well for spatial data. There is limited literature on appropriate marker gene detection methods that can account for the spatial distribution of gene expression. Qiu et al. developed a method called *saSpatial* that aims to find differentially expressed genes with spatial statistics for spatial transcriptomics data [37]. The proposed method starts with constructing a spatial map integrated with gene expression using the software Arcmap, which provides a compositional representation of five predefined patterns for every spot in the spatial data. For a selected region of interest, *saSpatial* relies on local Moran's I statistics to obtain the pattern composition within the region. They perform a Chi-square test on

the compositional representation of the gene within the region of interest and the other regions [37]. C-SIDE is another method [8] to identify genes with cell type specific differential expression. C-SIDE models gene expression as an additive mixture across cell types of log-linear cell type-specific expression functions. However, this method is tightly integrated with their robust cell type decomposition framework [9], resulting in limited flexibility in input types and restricted adaptability across different analysis pipelines.

In order to address this gap in our own spatial 'omics analysis toolkit, we asked the question: what constitutes a good marker gene? In the single cell field, we have developed statistical methods that identify genes that are more highly expressed in the cluster of interest, compared to the remaining cells. However when we apply these methods directly to sparse and low count spatial omics datasets, where each transcript detection per cell is only one or two for the majority of genes and cells, we find that a large majority of the genes in a gene panel are statistically significant. For our particular application, we defined marker genes as having two desirable properties: 1) the marker gene expression is somewhat unique to the cell type of interest, and 2) there is a strong overlap in the locations of the transcript detections and the cells of the cell type of interest.

In this study, we introduce jazzPanda, a method for identifying marker genes in imaging-based spatial transcriptomics data. Due to the sparsity of imaging-based spatial transcriptomics data, our new method jazzPanda utilises a binning approach whereby transcripts are binned into a hexagonal or rectangular grid space across the tissue. This allows us to transform the x and y spatial coordinates of transcript detections of a gene to a one dimensional gene vector; and similarly transform cells in a cluster to a one dimensional cluster vector. Each element of the gene and cluster vectors corresponds to a specific grid superimposed on the tissue sample, and the binning process effectively spatially "pseudobulks" the detection and cell counts at a specified resolution. We propose two approaches to find spatial marker genes using the transformed coordinates information. The first approach measures the correlation coefficient of the spatial information between every pair of cluster and gene vectors. We then perform permutation to assess the statistical significance of the calculated correlation. However this has the same limitation as a Wilcoxon rank sum test; it is substantially more challenging to account for multiple covariates and samples. The second approach is based on linear modelling of gene and cluster vectors that utilises lasso regularization. The linear modelling approach has the advantage that it can account for multi-sample experimental designs and background noise. We show that our new method controls the Type I error rate through simulation, and we apply our method to six publicly available datasets. Finally we compare our results to the Wilcoxon Rank Sum Test and the moderated t-statistic from the limma package [38]. We also show that jazzPanda is robust to grid size choice and investigate time and memory us-

age. jazzPanda is available as an R Bioconductor package (`https://www.bioconductor.org/packages/release/bioc/html/jazzPanda.html`). All analysis code is available as an analysis website at `https://github.com/phipsonlab/jazzPanda_workflowr` and on GitHub (`https://github.com/phipsonlab/jazzPanda_paper`).

Results

jazzPanda: a new method for spatially aware marker analysis

The foundation of our approach lies on binning transcript detections and cells within clusters to a particular grid resolution across the tissue. Imaging-based spatial transcriptomics technologies have the advantage that they can achieve sub-cellular resolution, as well as providing the spatial coordinates of every transcript that is detected. However, the transcript counts per gene per cell tend to be very low; often zero, one or two counts. Binning allows for greater statistical power as the counts per gene per bin is much larger compared to cell level counts. As mentioned, an ideal marker gene has significantly higher expression in the target cluster compared to the remaining clusters. In the spatial context, this translates to a marker gene showing higher transcript detections within the target cluster region compared to other regions over the physical tissue space. Therefore, we make the assumption that the expression of marker genes will linearly correlate with their corresponding cluster over the two dimensional space. We assume that clustering has already been performed and cluster labels for each cell are available. For optimum performance of jazzPanda, the clusters should be spatially separable to some degree. We propose two approaches to identify the genes that show a strong linear relationship between the cluster and gene vectors.

Figure 1 shows an overview of the jazzPanda method. First, we convert the two-dimensional coordinates to a one-dimensional spatial vector by discretizing the original tissue space with tiles. The tile can be applied in different shapes like rectangles, squares, or hexagons. We define a gene vector (g_i) for every gene, and a cluster vector (x_i) for every cluster, with the length equal to the number of tiles. A gene vector can be obtained by calculating the total number of transcripts in every tile for that gene, regardless of whether the transcript was allocated to a cell. In this manner gene vectors are agnostic to cell segmentation. If the transcript level coordinates are unavailable, the gene vector can be calculated from the gene by cell count matrix and cell level locations. As each cell location can be converted to a one-hot vector, we can multiply the expression value of the gene by the individual cell vector to discretize the gene expression. The gene vector can be constructed by taking the sum of the multiplied values across all cells for each tile. The cluster vectors are calculated by simply counting the number of cells in each tile

for that cluster. We define the marker genes for a cluster as genes whose spatial vectors show a strong linear relationship with the target cluster vector.

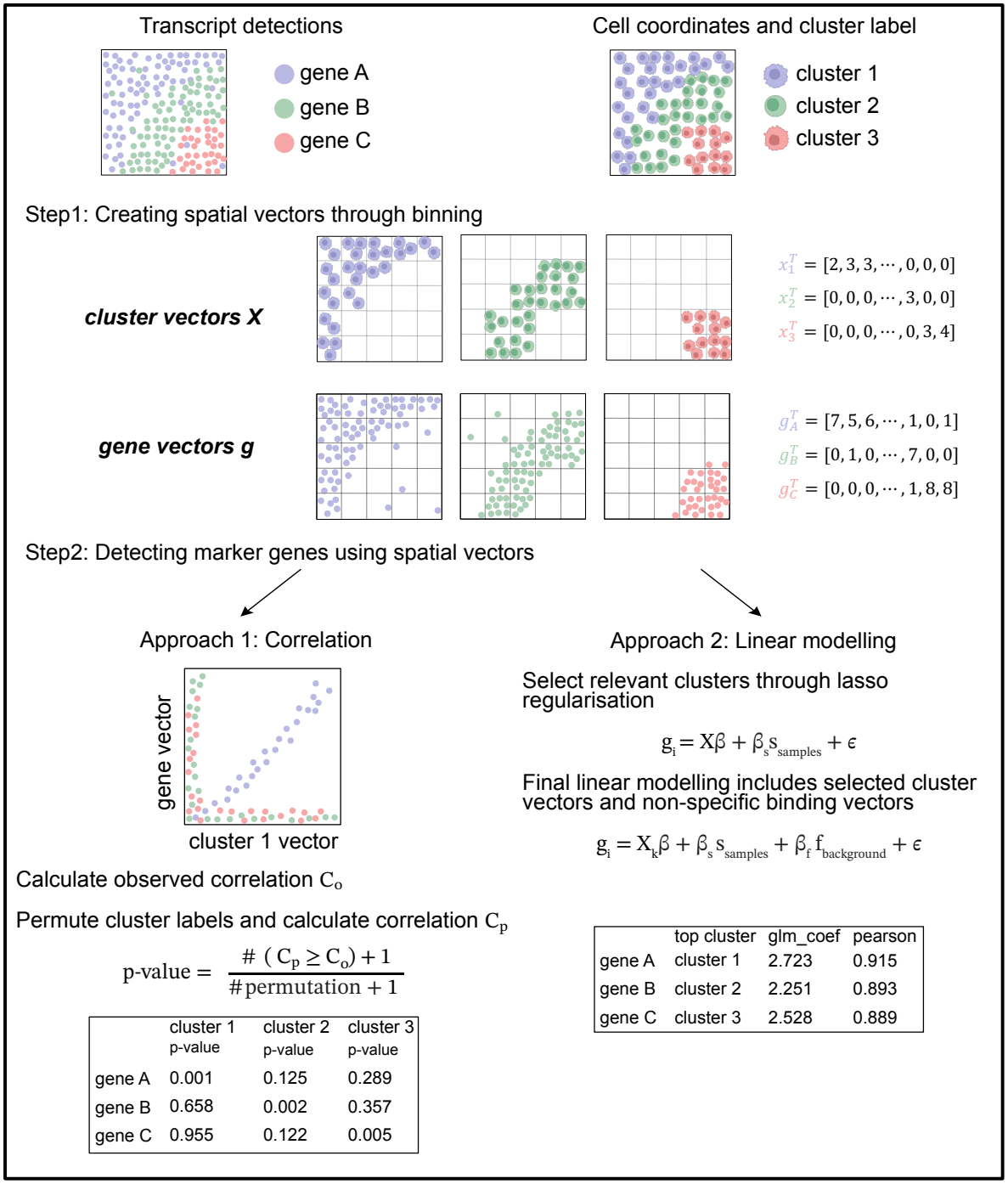

Figure 1: Overview of the jazzPanda approach. In Step 1, spatial vectors at the transcript and cluster level are created using square, rectangular or hexagonal bins. In Step 2, the spatial vectors are used as input for marker analysis using either a permutation framework to obtain statistical significance on the correlation estimates between clusters and gene vectors, or through a linear modelling approach.

We provide two approaches for detecting marker genes based on computed spatial vectors. The first approach is based on correlation. We can compute the correlation for every gene-cluster vector pair, and perform permutation to assign a significance level. The second approach utilises a generalised linear modelling approach. We fit a linear model to explain the gene vector as the linear combination of the cluster vectors. We first apply lasso regularisation to select the most relevant cluster vectors, and the retained cluster vectors form a second linear model for the gene vector. The second linear model contains additional covariates that are important to model, such as background noise. A given gene is defined as a marker gene for its most statistically significant cluster. Details of the statistical model are given in 'Methods' and further information about each approach is detailed below.

Simple correlation approach with permutation

Correlation is one way to measure the strength of the linear relationship between two vectors. If a gene vector has a strong positive correlation with a target cluster vector we can assign that gene as a marker gene for that cluster. We can obtain a significance level for the correlation between the gene and the cluster using permutation in the following manner. For each permutation run, we randomly shuffle the cluster label for the cell and build the cluster vectors with the shuffled cluster labels. We can then recalculate the correlation matrix between every gene vector and the shuffled cluster vectors. This process is repeated a number of times. The p-value for an observed correlation between a cluster and gene vector can be calculated as the proportion of permuted datasets with correlation at least as large as the observed correlation using the simple equation in Phipson & Smyth [32]. The permutation p-values are then adjusted for multiple testing using the Benjamini and Hochberg method [6].

Generalised linear modelling approach

The limitation of our correlation approach is that we are unable to account for additional covariates in a relatively straightforward manner. Assuming our samples are jointly clustered such that the clusters are consistently labelled, an ideal statistical model would be able to account for multiple samples rather than analysing each sample separately. Suppose we have n genes in a dataset, and c spatially separable clusters. Let v denote the length of each constructed spatial vector. Since the original count scale most faithfully preserves linear relationships between the cluster and gene vectors, no further scaling or transformation was performed. Treating the gene vector as the response variable, and the cluster vectors as observed variables, we can build a generalized linear model for each gene.

As every gene vector $\mathbf{g}_i$ has a unique relationship with every cluster vector, we use lasso

regularization to select the k most highly relevant clusters for each gene [16, 43]. Since genes have different expression levels, we estimate the lasso penalty parameter λ_i for each gene independently by cross validation. A small λ_i will lead to a weak penalty and the coefficients will be less shrunk towards zero, and we will retain more candidate clusters (features) for the next step. A larger λ_i will lead to a strong penalty, more shrinkage and fewer candidate clusters. We perform 10 fold cross-validation to automatically select the λ_i value. Given v tiles for a gene vector and all cluster vectors, we apply stratified sampling to group tile values into ten subsets, where each subset contains homogeneous data points that can represent the overall distributions [27]. During the cross validation step, a linear Gaussian model is trained on nine folds of the data and validated on the remaining one fold. This process is repeated ten times and every fold is selected as the validation set. We record the conservative λ_{1se} that is within one standard error of the minimum cross-validation error to avoid over-fitting. Based on the optimal λ_i value, we will obtain multiple clusters with non-zero lasso coefficients. The assumption is that these clusters will capture a more precise feature space and are further used to construct a final generalized linear model. To account for study designs with multiple samples, we introduce a sample level vector (S) in the final model to detect shared marker genes across different samples. This vector captures the sample to sample variability in the feature space and reduces sample-specific effects in the final output.

The output of the final model can be interpreted in two ways. First, we can select the cluster with the minimum p-value and largest (positive) coefficient to define gene $\mathbf{g}_i$ as a unique and strong marker gene for the selected cluster. In this setting, clusters effectively 'compete' for assignment by maximizing how well they explain the gene expression vector. Second, we also output all significant clusters for a given gene based on the final model for investigation of shared marker genes. This approach has a less stringent criteria for defining marker genes for a cluster, with the lasso regularisation enforcing sparsity and selecting the most informative clusters.

Accounting for non-specific binding of RNA molecules

Due to the linear modelling approach we have flexibility to further extend the framework to account for background noise by including additional covariates to represent unwanted patterns. Different spatial platforms have different implementations of negative controls for assessing the quality of the assay and as input into the technical algorithms. There are predominantly two types of negative controls. One type includes barcode sequences that do not match any real probes in the assay, and all the tested platforms - Xenium (negative control codeword), CosMx (falsecode), and MERSCOPE (blank genes) - have such negative controls (Figure 2 a-c, Supplementary Figure S1). The other type of negative

controls are called negative probes, which are generally artificial sequences that are not found in the study organisms to assess non-specific binding in the assay. Both Xenium and CosMx also include negative probes as controls [17, 5].

With our linear modelling approach we are able to account for this source of variability in the data. We use the same binning approach to define background noise vectors ($\mathbf{f}$) for every negative control category. For example, we define two negative control vectors for Xenium. We sum all negative codeword and negative probe detections separately, and include both the negative codeword vector and negative probe vector in the final linear model as a way of capturing the background noise. Including these background vectors in the final linear model as additional covariates can leverage the technical background impact and better control the false discovery rate.

False discovery rate control: null simulations

To better understand the performance of our two approaches, we simulated uniform spatial vectors for genes and negative controls from a real dataset. We randomly selected 100 real genes from the CosMx human liver cancer dataset. For every selected gene, we generated the same number of uniformly distributed points matching the number of transcript detections across the tissue space to simulate a uniform gene. To simulate a uniform background effect, we generated n uniformly distributed points where n equals to the number of detections for each negative control category. We applied jazzPanda to these simulated datasets, using the real cluster vectors from the human liver cancer dataset to test whether the simulated uniform genes will be assigned to any of the clusters as marker genes. This process is repeated 100 times and we applied both the correlation with permutation (no background) and linear modelling approach (with background) to detect marker genes.

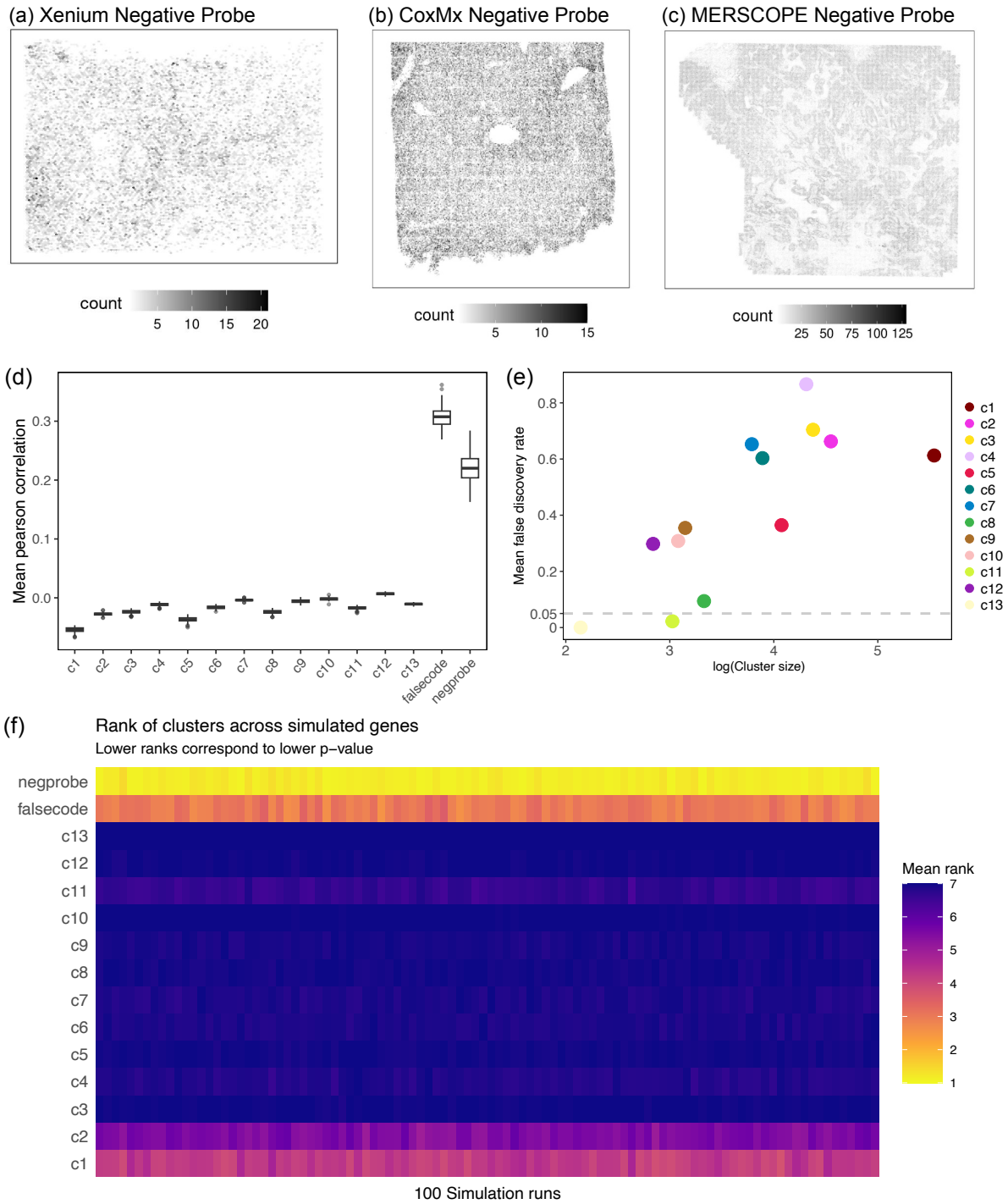

Figure 2: Spatial distribution of negative control probes on imaging based platforms (a)-(c) and evaluating jazzPanda's performance using simulated datasets (d)-(f). (a) Negative control probes for Xenium human breast cancer sample. (b) Negative control probes for CosMx health human liver sample. (c) Negative control probes for MERSCOPE human breast cancer sample. (d) Boxplot of the Pearson correlation between uniformly generated gene vectors, negative control vectors, and cluster vectors. 100 datasets were simulated with 100 uniformly generated genes. (e) Scatter plot showing the relationship between the cluster size and mean false discovery rate (FDR) from p-values calculated using jazzPanda-correlation without a correlation cut-off applied. Mean FDR's calculated across 100 genes and 100 datasets. (f) Mean ranking of cluster and negative control vectors derived from the gene-wise p-values using the linear model approach across each simulation run.

Figure 2d shows the overall Pearson correlation of the simulated gene vectors with the real cluster vectors, as well as with the negative controls. We can see that the uniformly simulated gene vectors show extremely low correlation with the real clusters, as is expected (median correlation is -0.015). The median correlation of the gene vectors with the simulated background falsecode and negprobe vectors is 0.287 and 0.242 respectively (Figure 2d). If a correlation cut-off was simply applied, for example, at 0.2, then none of the genes would be called as marker genes. However, we noted that when we performed permutations to assign p-values to the correlations, an inflated number of significant genes were observed, particularly for larger clusters (Figure 2e). Therefore, in practice it is crucial to examine the correlation together with the permutation p-values. When we apply even a modest cut-off of 0.05 on the Pearson correlation, we can effectively reduce the false positive rate to approximately 5.16% (Supplementary Figure S2).

The pitfall of the permutation approach is that we cannot easily incorporate background signal when calculating p-values. However, with our linear modelling approach we can include the simulated background vectors in the model. When the spatial pattern of a given gene is highly correlated with background patterns, we expect that the linear modelling approach will not identify this gene as a marker gene for any cluster due to the introduction of the background vectors. Figure 2e shows the rank of the clusters based on the final linear model coefficients. Our approach selects the negprobe and falsecode as the most relevant cluster in 89.27% and 10.73% cases respectively. Therefore, the top result with minimum p-values and largest model coefficient is always the negative controls, and the gene will not be assigned as a marker gene to any cluster. This highlights that our linear modelling approach can effectively account for background effect and reasonably control the false discovery rate. However, the Lasso regularisation step may result in other clusters being retained to include in the final model. When we examined these results further, we noticed it was most likely that either one cluster (35.68%) or two clusters (40.99%) were retained. It was very rare to see three (3.95%) or four (1.44%) clusters retained. These selected clusters tend to be large clusters or have more uniformly distributed cells across the tissue space. Although lasso may select additional real clusters, we will always return the background as the most relevant "cluster" for these simulated genes after ordering by minimum p-value and largest model coefficient, which aligns with the genes being more highly correlated with the background vectors (Figure 2d). Therefore we recommend using the linear modelling approach with background included to effectively control the false discovery rate. However, the correlation calculated from the permutation approach may still be useful to examine in conjunction with the output from the linear model.

Performance of jazzPanda on real spatial datasets

Datasets

We tested jazzPanda on six public datasets from three imaged-based spatial technologies (Figures 3-4, Supplementary Figure S3). For Xenium, we included mouse brain (three technical replicates), human breast cancer (two biological replicates) and a human lung cancer dataset. For CosMx we analysed a healthy human liver sample and a human liver cancer sample, and for MERSCOPE we included a human breast cancer sample (Table 1). We examined the correlation-based approach on datasets containing only one sample: Xenium human lung cancer sample, CosMx human healthy liver sample, CosMx human liver cancer sample and the MERSCOPE human breast cancer sample. We performed 5000 permutations and kept genes with significant adjusted p-values and an observed correlation larger than the 75th percentile of the overall gene-cluster correlation for the corresponding cluster. We rank the significant markers by the glm coefficient for the linear modelling approach and rank permutation markers by the observed correlation. Moreover, we applied the linear modelling approach on all the datasets and accounted for multiple samples for multi-sample datasets and background detections. In the following results sections we refer to the permutation approach as jazzPanda-correlation and the linear modelling approach as jazzPanda-glm. The detailed analysis for each dataset is included in Supplementary Material.

For some datasets we have used the provided cell type labels as clusters (Xenium human breast cancer, CosMx human liver cancer, CosMx human healthy liver) as input to jazzPanda. For the remaining three datasets where cell type labels were not available (Xenium mouse brain, Xenium human lung cancer and MERSCOPE human breast cancer), we performed Banksy clustering which incorporates gene expression and spatial information, and manually annotated the clusters using well known marker genes (see Methods for further details).

jazzPanda detects spatially relevant marker genes for single sample datasets

We used the CosMx healthy liver sample as a single sample example to illustrate the spatial and biological relevance of the detected marker genes (Figure 3). This dataset contains 1000 genes and 332,877 cells. We visualised the spatial organisation of this sample in Figure 3a, where each point is one cell and is coloured by the cell type. CosMx provides annotation for 15 major cell types for this sample, and the UMAP indicates transcriptional differences for cell populations (Figure 3b). The cell types differ in abundance with hepatocytes (Hep.4 and Hep.5) accounting for approximately 65% of cells. In contrast, several immune and endothelial cell types typically contain less than 5,000 cells (Figure 3c).

Table 1: Dataset overview

technology	dataset	#samples	#genes	#transcript	#cells	#clusters
Xenium	human breast cancer	2	313	4.51e+07	2.79e+05	9
Xenium	human lung cancer	1	392	6.65e+07	5.31e+05	13
Xenium	mouse brain	3	248	1.41e+08	4.75e+05	9
CosMx	human healthy liver	1	1000	1.85e+08	3.33e+05	15
CosMx	human liver cancer	1	1000	4.78e+08	4.60e+05	13
MERSCOPE	human breast cancer	1	500	4.90e+08	6.49e+05	10

We used 40×40 square bins to define spatial vectors for the cells of every cluster and the transcript detections for every gene. The cluster vectors are computed from the cell coordinates and the gene vectors are computed from the transcript detection coordinates. Taking stellate cells as an example (Figure 3d), using jazzPanda-correlation we detected ten marker genes with adjusted p-value < 0.05 and observed correlation larger than the 75th percentile of the gene-cluster correlations. Our correlation approach detects insulin-like growth factor-binding protein 7 (IGFBP7) as the top gene for stellate cells in the CosMx human healthy liver sample (Figure 3e). The correlation between the gene vector and the cluster vector is 0.904 and the adjusted permutation p-value is 0.0018 (Figure 3f). When we plot the transcript detections of IGFBP7 beside the target cluster, it is clear that the marker gene shows highly concordant spatial patterning with the cluster over the tissue space (Figure 3d-e). We can see a clear linear relationship when we visualise the IGFBP7 gene vector with the stellate cluster vector (Figure 3f). The IGFBP7 gene is widely studied in stellate cells in single cell human liver studies in the literature. It is differentially expressed in hepatic stellate cells [31], and it is used as a specific marker gene for hepatic stellate cells [31]. Supplementary Figure S4 shows the top 3 markers for every cluster identified using jazzPanda-correlation.

Performing the marker analysis using jazzPanda-glm, IGFBP7 is also found as the most significant marker gene for the stellate cells cluster (Supplementary Figure S5). In general, there is some concordance in the top 3 markers between the two approaches, with at least one gene in common for most clusters, however there can be differences in the ranks and the different methods can detect different genes in the top 3 marker list (Supplementary Figures S4, S5). Supplementary Figures S6 and S7 show the linear relationship between the top 3 marker gene vectors and their associated cluster vectors.

In addition to the CosMx healthy liver sample, we performed marker analysis using

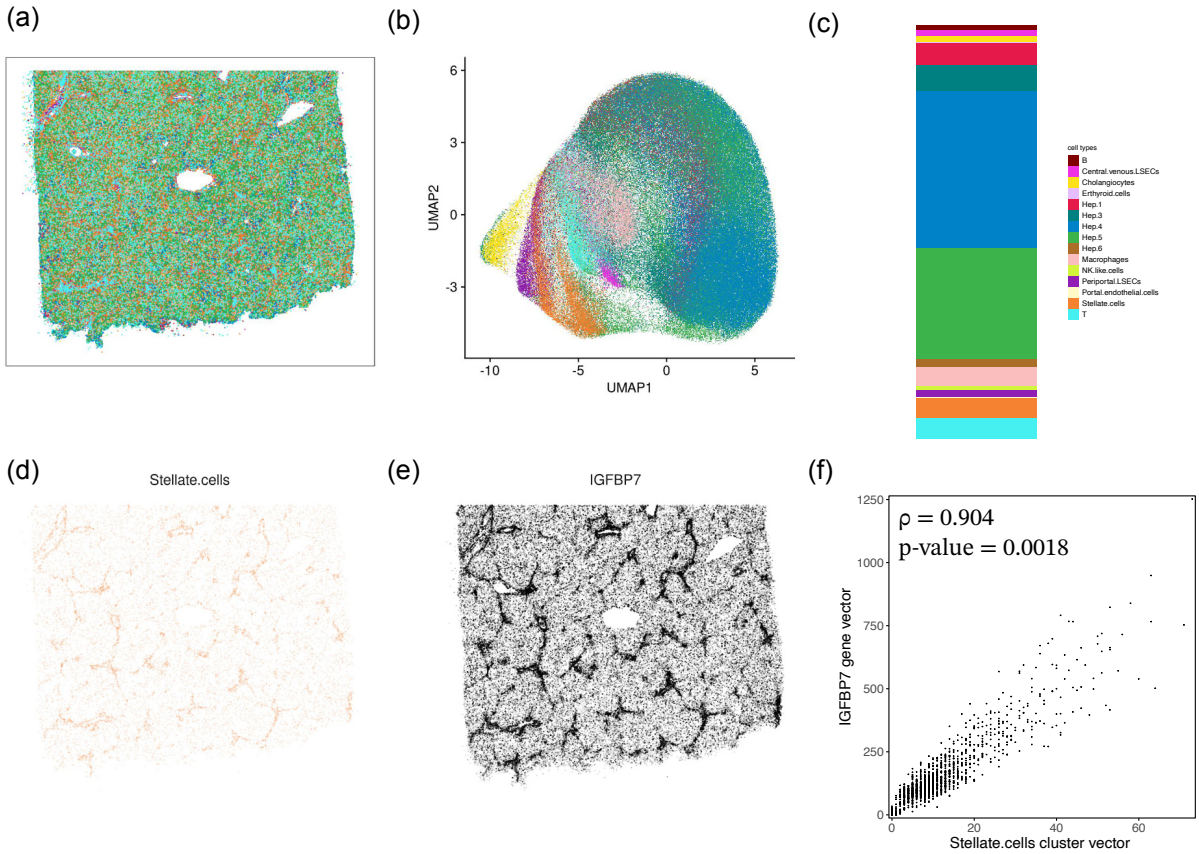

Figure 3: Application of the correlation approach to a CosMx healthy human liver sample. (a) Spatial visualization of cell type coordinates, where each point represents a cell and is coloured by cell type (b) UMAP plot showing cell types (c) Cell type composition of the liver sample (d) Spatial visualization of cell coordinates for Stellate cells (e) Spatial visualisation of transcript coordinates for the top marker gene, IGFBP7, selected based on the smallest adjusted permutation p-value and highest Pearson correlation by the correlation approach (f) The relationship between the Stellate cells cluster vector (x axis) and the marker gene vector (y axis), where each point refers to one grid value. The Pearson correlation between Stellate cells cluster vector and IGFBP7 gene vector is 0.904, and the adjusted permutation p-value is 0.0018.

both jazzPanda-correlation and jazzPanda-glm on the other single sample datasets. Supplementary Figure S8 shows the top 3 marker genes for the Xenium human lung cancer dataset using jazzPanda-correlation and Supplementary Figure S9 shows the top markers using jazzPanda-glm. For strong markers with high correlation and large glm coefficients the two methods usually agree (e.g. MARCO for the alveolar macrophage cluster, LGMN for the inflammatory TAMs). Supplementary Figures S10 and S11 show the linear relationship between the top 3 marker gene vectors and their associated cluster vectors.

Supplementary Figures S12 and S13 show the top 3 markers genes for each cluster for the CosMx Human Liver Cancer dataset for jazzPanda-correlation and jazzPanda-glm respectively. Supplementary Figures S14 and S15 show the linear relationship between

the cluster and gene vectors for the top 3 markers for each cluster. We find that in general one of the three top markers are in common between the two methods. Although the two approaches can rank different genes in the top 3, in general the genes are known markers for the relevant cell type. For example, the top three marker genes for Tumor_2 cells detected by jazzPanda-correlation are NDUFA4L2, IGFBP3 and VEGFA (Figure S12). The top three markers identified by jazzPanda-glm are ENO1, PGK1 and MIF. These top marker genes are highly relevant in hepatocellular carcinoma. NDUFA4L2 is overexpressed in hepatocellular carcinoma, and high expression of NDUFA4L2 is closely associated with tumor metastasis [24, 52]. IGFBP3 plays an important role in cancer cell proliferation, and VEGFA is highly expressed in multiple tumour types in hepatocellular carcinoma [28, 35]. Studies show elevated expression of ENO1, PGK1 and MIF in hepatocellular carcinoma tissues [51, 46, 25]. Thus, while our two approaches can detect different genes as the top 3 markers, the selected markers are all biologically relevant

Supplementary Figures S16 and S17 shows the top 3 marker genes for the MERSCOPE human breast cancer dataset for jazzPanda-correlation and jazzPanda-glm respectively. Again, a similar trend emerges that in general at least 1 of the top 3 marker genes are detected in both methods, and both methods identify marker genes that appear spatially relevant when comparing to the spatial distribution of the cells in the relevant cluster. Taking the T cell cluster as an example, the TRAC and CD3E are ranked as the top marker genes for both approaches, which are known marker genes for annotating T cells in single cell human breast cancer tissues [15]. jazzPanda-correlation identifies CD2 as an additional top marker gene for T cells, which is involved in T-cell immune response [19]. The additional marker gene detected by jazzPanda-glm is CCL5, which is used as a T cell signature gene and known to increase upon T cell activation [4]. Supplementary Figures S18 and S19 show the linear relationship between the top 3 marker gene vectors and their associated cluster vectors.

jazzPanda detects shared marker genes for datasets with multiple samples

Spatial technology is already being used in more complex experiments with multiple samples and groups [10, 18, 36]. Hence it is important that our methods are able to handle multiple biological replicates.

We applied jazzPanda-glm to detect shared marker genes for datasets with multiple biological replicates. The Xenium human breast cancer dataset contains two biological replicates, and there are nine shared clusters (Figure 4a-b). Each sample has unique spatial organisation of the different clusters (Figure 4a), and the two samples have distinct cell type composition (Figure 4b). We used 40×40 square bins to define spatial vectors

for this dataset. Xenium has two types of negative controls (probe and codeword targets) to capture technical issues like non-specific binding. We converted the negative control detections into two background vectors with the same binning dimensions. The background vectors are included as covariates in the final linear model, as well as a covariate for sample. Supplementary Figure S1a, d shows the two types of negative control probes for this dataset.

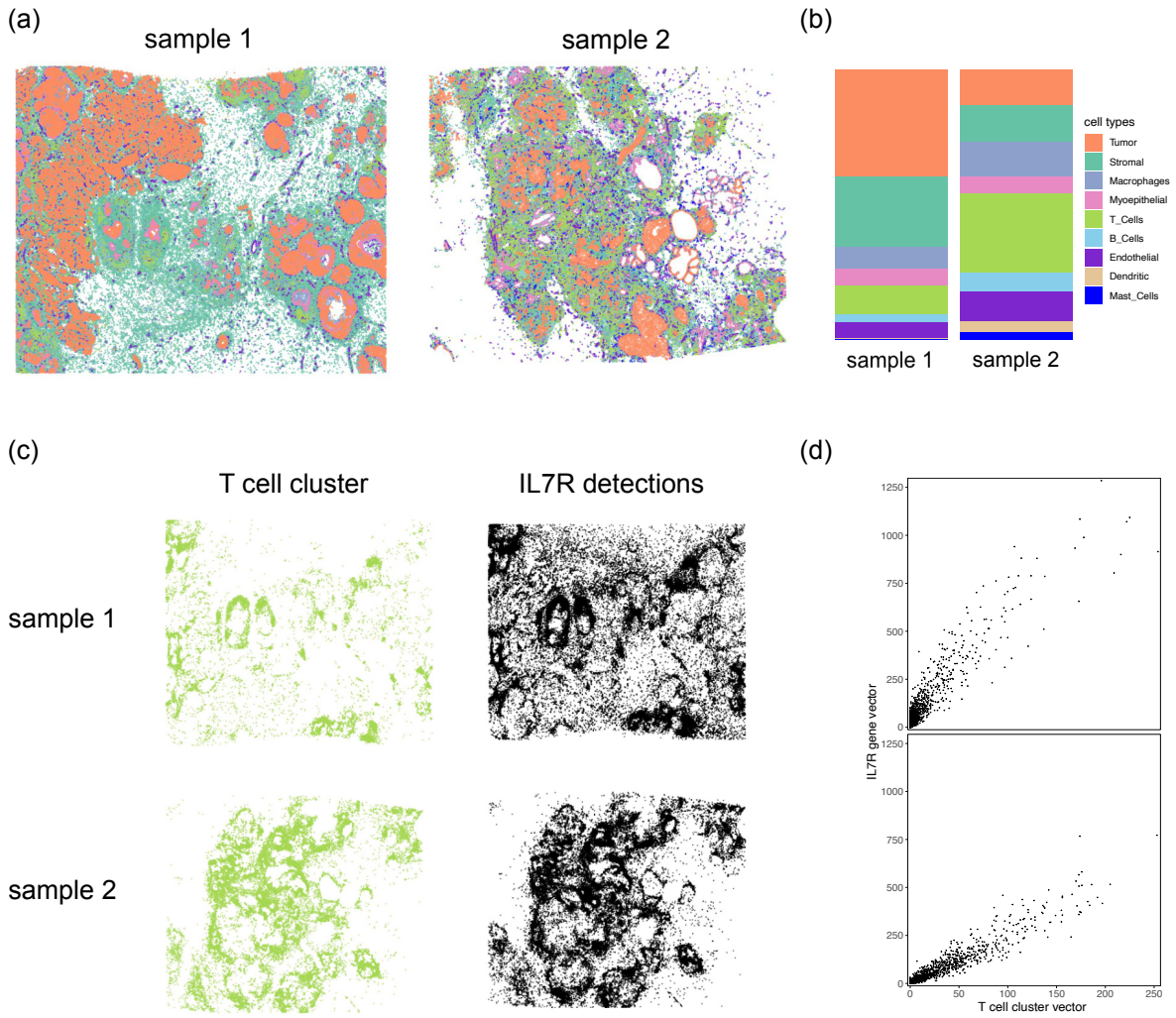

Figure 4: Application of the linear modelling approach to two Xenium HER2+ human breast cancer samples. (a) Spatial visualization of cell type coordinates for each sample, where each point represents a cell and is coloured by cell type (b) Cell type composition annotated per sample (c) Spatial visualization of cell type coordinates for the T cell cluster and the corresponding transcript detections for the top marker gene, IL7R, selected by the linear modelling approach (f) The relationship between the T cell cluster vector (x axis) and the marker gene vector (y axis), each point refers to one grid value.

Figure 4c visualises the top marker gene interleukin 7 receptor (IL7R) detected for T cells for each sample. We can see that the T cell cluster from these two samples have unique spatial patterns. The spatial concordance between the IL7R transcripts and T cell cluster is extremely high within each sample. When we visualise the IL7R gene vector with the

T cell cluster vector, we can see a clear linear relationship for both samples (Figure 4d). IL7R and its ligand IL7 are crucial for T cell development [44]. One study suggests that higher levels of IL7R expression are associated with more aggressive breast tumor [1]. The top 3 markers for the T cell and remaining 8 cell types are shown in Supplementary Figure S20. Supplementary Figure S21 shows the linear relationship between the cluster and gene vectors for the top 3 markers.

The Xenium mouse brain dataset has three replicate samples that are consecutive sections of the same mouse brain, i.e. they are not independent biological replicates. Hence the spatial distribution of the cell types are nearly identical and the cell type composition is very similar between the three replicates (Supplementary Figure S3a). We used jazzPanda-glm to identify marker genes for the 9 provided clusters/cell types. Supplementary Figure S22 shows the top 3 markers genes for each cluster.

Annotating the mouse brain has further complexities as the layers of the brain form domains that may comprise of several cell types, which can prove challenging for annotation of clusters. In general we see that jazzPanda is able to identify genes that overlap spatially with the cluster of interest but there are also clearly shared marker genes identified. For example, Calb1 is identified as a marker for Exc_DG_Granule, however it is also clearly expressed on the periphery of the brain (Supplementary Figure S22). Despite this, there is a strong positive correlation observed for the identified marker genes and their clusters (Supplementary Figure S23).

Comparison of jazzPanda with single cell marker analysis methods

For marker gene analysis in single cell data, a common approach is to use a Wilcoxon Rank Sum Test comparing each cluster to the remaining clusters. One implementation in R is the 'FindAllMarkers' function in the Seurat package. There is also a Python implementation in the Scanpy package. A recent benchmarking study found that simple methods for marker analysis such as the Wilcoxon Rank Sum Test and t-test often outperform other more complex methods [34]. A drawback of the Wilcoxon Rank Sum Test is that it is a non-parametric test and cannot account for additional covariates such as sample or sex. Here we compare the marker genes detected using jazzPanda, Wilcoxon Rank Sum Test and moderated t-tests as implemented in the limma R package. The Wilcoxon Rank Sum Test as implemented in Seurat also incorporates a logFC cut-off of 0.1 as default which we applied to all datasets. For the moderated t-test and jazzPanda-glm we included a sample level covariate in the case of multi-sample datasets.

We compared the marker genes detected using the Wilcoxon Rank Sum Test, moder-

ated t-test and jazzPanda for all six datasets (Figure 5, Supplementary Figures S24-S35). For one-sample datasets, we also ran jazzPanda-correlation. Figure 5 shows the results from selected clusters from a single sample dataset (CosMx healthy liver sample, Figure 5a, c) and a multi-sample dataset (Xenium human breast cancer, Figure 5b, d). We selected large, medium and small clusters in each dataset to examine whether there was any bias with regards to number of cells in the cluster relative to the total number of cells in the dataset. We defined marker genes as follows. For the Wilcoxon Rank Sum Test, marker genes had adjusted p-value < 0.05 with the default logFC cut-off of 0.1; for the moderated t-test, marker genes had adjusted p-value < 0.05; for jazzPanda-correlation, marker genes had observed correlation $> 75^{th}$ percentile of all cluster-gene correlations for that particular cluster and adjusted p-value < 0.05. For jazzPanda-glm, marker genes were selected based on the lasso regression p-value and maximum glm coefficient in the final linear model.

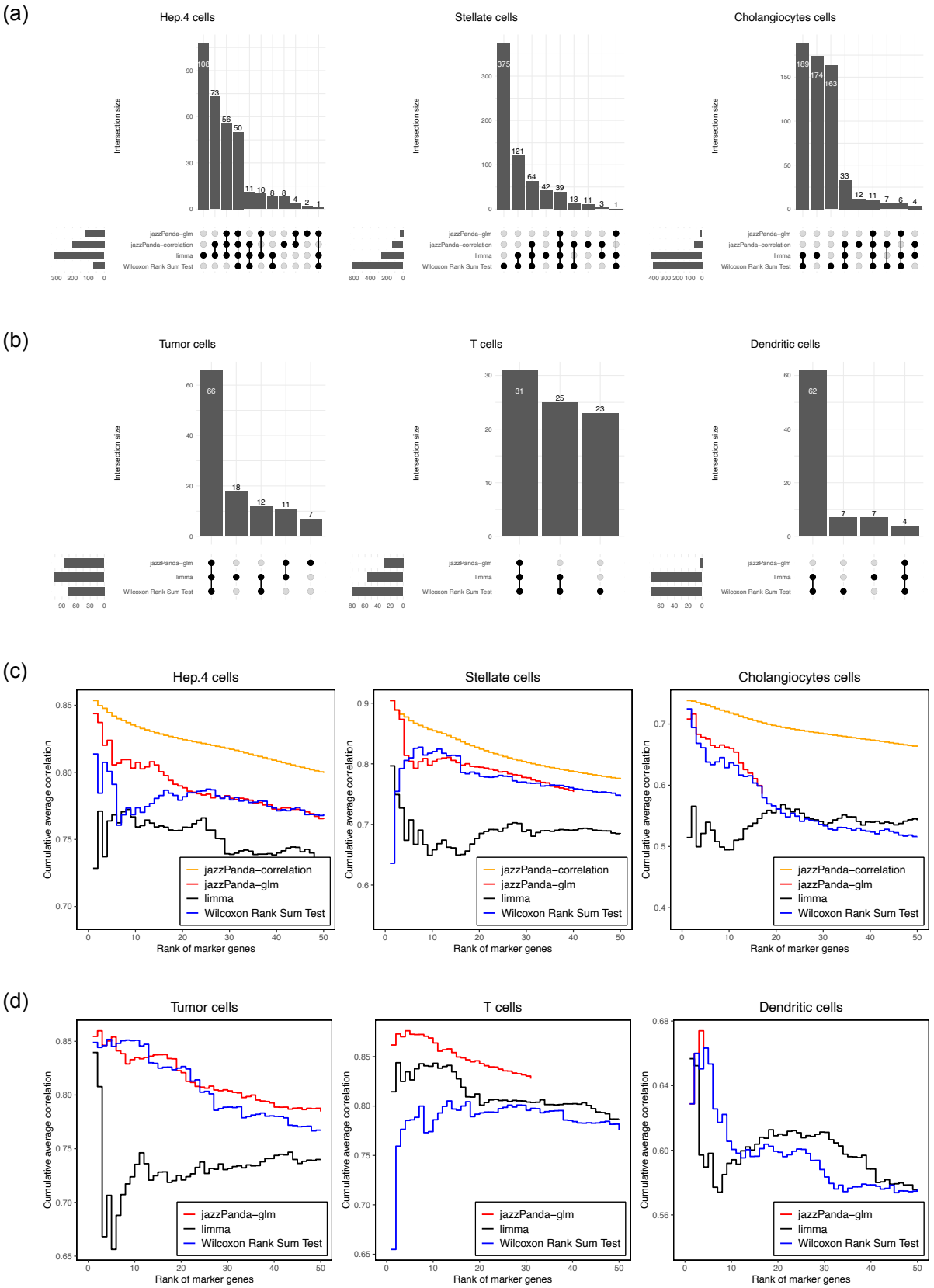

Figure 5: Comparison of marker analysis methods. (a) Upset plots showing overlap of genes detected using different marker analysis methods for a large cluster (Hep.4 cells), a medium cluster (Stellate cells), and a small cluster (Cholangiocyte cells) for the CosMx healthy liver sample. (b) Upset plots showing overlap of genes detected using different marker analysis methods for a large cluster (Tumor cells), a medium cluster (T cells), and a small cluster (Dendritic cells) for the two Xenium breast cancer samples. (c) Cumulative average correlation of top 50 ranked marker genes for each method for the healthy liver sample. (d) Cumulative average correlation of top 50 ranked marker genes for each method for the two breast cancer samples.

Our first observation when comparing the methods was that both jazzPanda approaches generally identified fewer marker genes per cluster compared to Wilcoxon Rank Sum Test and moderated t-test (Figure 5a, b). JazzPanda-glm detected fewer genes than jazzPanda-correlation for the CosMx healthy liver sample (Figure 5a). The Wilcoxon Rank Sum Test and moderated t-test tend to identify a large proportion of the genes in the panel as marker genes for each cluster for both datasets, with the exception of the Hep4.cells in the liver sample, where the Wilcoxon Rank Sum Test has the fewest marker genes detected. This is likely due to the additional logFC threshold that is the default setting for the FindAllMarkers function. A very large marker gene list can make cell type annotation more challenging and may indicate the p-values for Wilcoxon Rank Sum Test and moderated t-tests are too liberal.

Although the Wilcoxon Rank Sum Test and moderated t-tests tend to identify very large sets of genes as markers, we note that both jazzPanda approaches detect marker genes that are a subset of these genes (Figure 5a, b). We also observe that the jazzPanda methods appear to be sensitive to the number of cells in the cluster with more marker genes detected for larger clusters and fewer marker genes detected for smaller clusters, an effect that is more pronounced for jazzPanda-glm. Supplementary Figures S24-S29 show upset plots for all cell types for the six datasets, with similar observations.

While the Wilcoxon Rank Sum Test and moderated t-test tend to identify large sets of marker genes, the top ranked genes are likely to still be biologically relevant. We thus assessed whether the different methods identified genes with high spatial correlation between the marker genes and target cluster. We calculated the spatial correlation of ranked marker genes and their target clusters at the spatial vector level and visualised the cumulative mean correlation of the top 50 markers genes for each method separately (Figure 5c, d). This metric clearly favours jazzPanda-correlation and indeed we see that the top ranked markers based on this approach has the highest cumulative mean correlation (Figure 5c) for a single sample dataset. jazzPanda-glm tends to prioritise genes with high spatial correlation among the top 50 ranked genes compared to moderated t-test and Wilcoxon Rank Sum Test, with the top gene usually exhibiting a high correlation (Figure 5c, d). The performance of the Wilcoxon Rank Sum Test and moderated t-test is more variable for this metric, with the top ranked genes from these methods having high spatial correlation for some clusters and not others. Supplementary Figures S30 - S35 show the cumulative correlation plots for all cell types for the six datasets.

The detected marker genes from our two approaches generally show high consistency. A strong marker gene is expected to have a large correlation with the target cluster with the correlation approach, and a large model coefficient from the linear modelling

approach. Both approaches will detect strong marker genes. However there are some differences between the marker gene results from our two approaches for a one-sample scenario, with jazzPanda-correlation identifying marker genes that are not found using other methods (Figure 5a). This may be due to non-specific binding, which cannot be accounted for using jazzPanda-correlation.

Extension of statistical framework: cluster-cluster and gene-gene correlation

Our approach of binning cells within clusters and transcripts into grids across the two dimensional space forms the basis of a statistical framework that can be used not only for marker gene analysis but for other types of analysis as well. Here we explore how we can use this framework to measure cluster-cluster correlation to get an estimate of global co-location patterns of the cell types within a tissue sample. The same framework can also be applied to estimate gene-gene correlation which will identify genes that are co-expressed in the same location spatially. Negative correlation between two genes implies that the two genes are not co-expressed spatially.

Since the first step in jazzPanda is to calculate spatial vectors for transcripts and cells within clusters, these vectors are readily accessible to feed into further downstream analysis. Here we demonstrate the cluster-cluster and gene-gene correlation analysis on a MERSCOPE human breast cancer sample (Figure 6a). First we calculated the pairwise Pearson correlation for all the clusters (cell types) for the breast cancer sample (Figure 6b). The strongest cluster to cluster correlation is seen within tumour related clusters. The Tumor_luminal_ERBB2 cluster shows correlations of 0.94 and 0.86 with the Tumor_IFNg_APC and Tumor_Cycling_G2M clusters respectively (Figure 6b). The largest negative correlation is observed between the CAF_Mixed and Tumor_IFNg_APC cells (Pearson correlation = -0.31). These relationships are also reflected spatially in the cluster patterns (Figure S16). The positively correlated clusters (Tumor_luminal_ERBB2, Tumor_IFNg_APC and Tumor_Cycling_G2M) show highly overlapping spatial patterns whereas cluster CAF_Mixed occupies different regions compared to the three positively correlated clusters. Another example of co-located cell types is the TAM_C1QC and activated T cells which have a Pearson correlation of 0.64 and appear spatially co-located in Supplementary Figure S16.

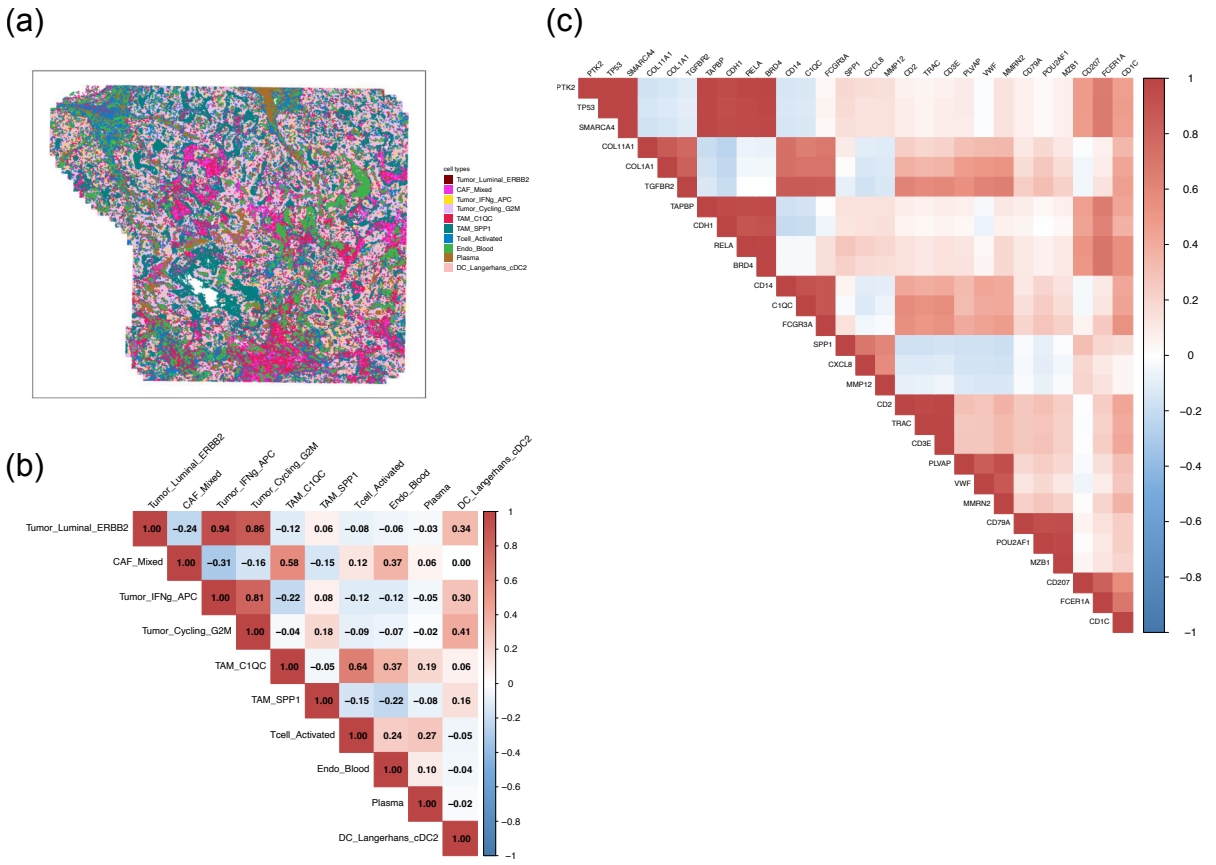

Figure 6: Extension of jazzPanda framework. (a) Spatial visualization of cell type coordinates for a MERSCOPE human breast cancer sample. Every point represents a cell and is coloured by the annotated cell type (b) Pairwise Pearson correlation between cluster spatial vectors (c) Pairwise Pearson correlation between spatial vectors of top 3 marker genes for each cluster.

Similarly we can calculate all pair-wise correlations between the genes in the panel. As a proof of concept we show the Pearson correlation for the top 3 marker genes for the clusters (Figure 6c, see Supplementary Table S1 for Pearson correlation values). As expected, we see strong positive correlation among the top marker genes for the same cluster. For example, the correlation is above 0.98 between PTK2, TP53 and SMARCA4, which are the top markers for the Tumor_luminal_ERBB2 cluster. The correlation patterns observed at the cluster-cluster level in Figure 6b are recapitulated in the gene-gene correlation heatmap of Figure 6c. The marker genes for positively correlated clusters are also correlated (e.g. PTK2 and TAPBP has a correlation of 0.98), and marker genes for negatively correlated clusters show negative correlations (e.g. PTK2 and COL11A1 has a correlation of -0.16). While this is a fairly simple application of gene-gene correlation, this approach could be used to find co-expressed genes more broadly regardless of cell type as well as genes that "repel" each other, i.e. genes that are not expressed at the same spatial coordinates. This statistical framework could potentially be extended to a spatial version of weighted correlation network analysis, a data mining method for

studying biological networks which has been widely used in the study of gene expression [50].

Technical performance

Effect of tile shape and grid length on marker gene analysis

An important parameter that needs to be defined by the user is the grid size to be applied across the tissue space. We wanted to investigate how robust the jazzPanda methods are to choice of grid size as well as the tile shape. We tested grid lengths ranging from 10×10 to 100×100 and compared the identified markers for every cluster for the one-sample CosMx human liver cancer dataset and two-sample Xenium human breast cancer dataset (Figure 7, Supplementary Figures S36-S38).

Figure 7a shows that the detected markers for three clusters of varying sizes from the CosMx human liver cancer dataset are generally consistent under different grid sizes (Figure 7a). The majority of marker genes can be detected as significant for all grid sizes. Supplementary Figures S36, S37, and S38 shows the marker gene consistency under a range of tile lengths for every cluster using the one-sample CosMx human liver cancer data and two-sample Xenium human breast cancer data respectively. It is clear that the detected markers remains consistent and are robust to diverse range of tile lengths when the cluster contains enough cells.

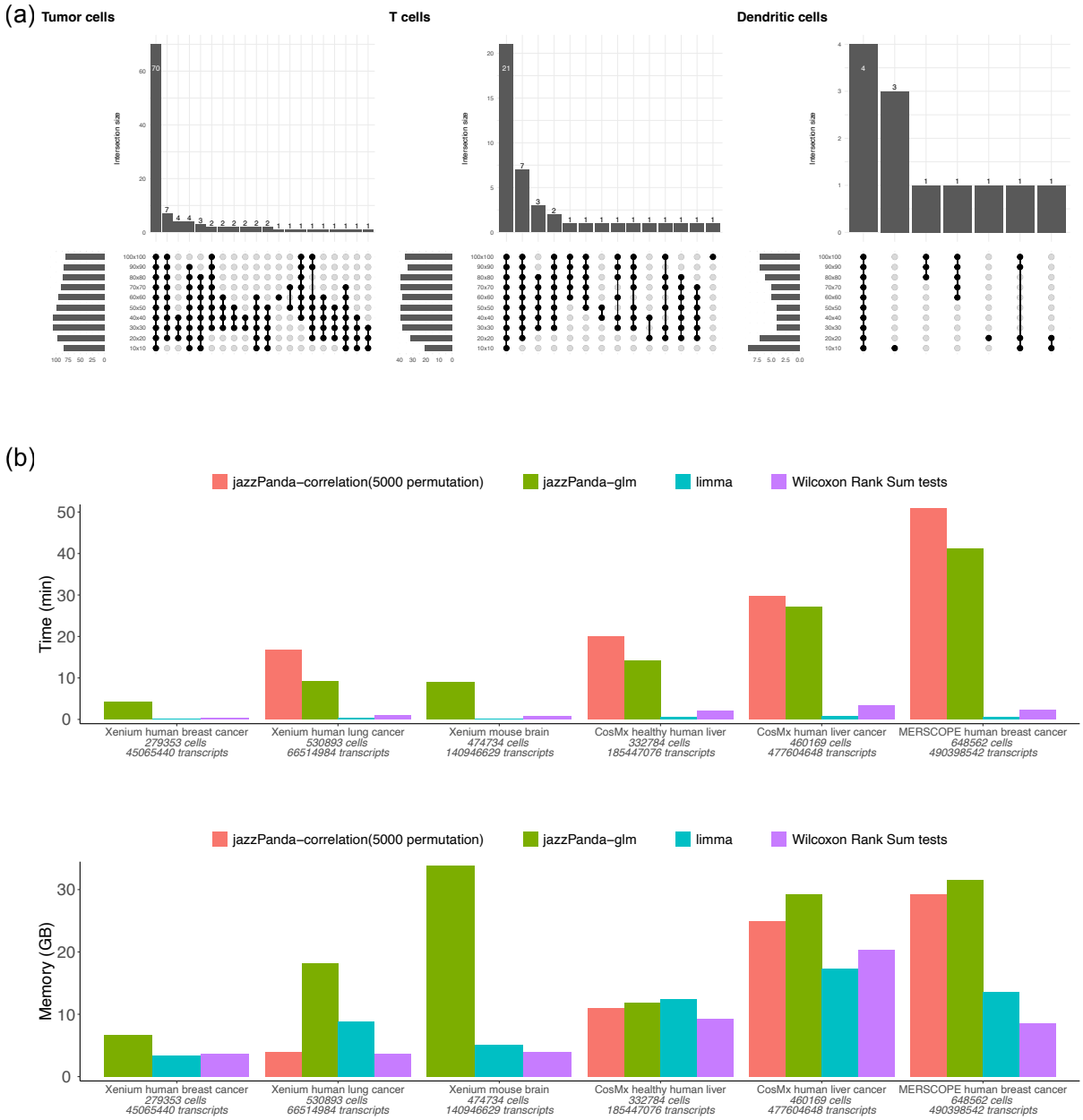

Figure 7: Technical performance of jazzPanda. (a) Upset plot showing consistency of the marker genes identified across a range of tile sizes used to construct spatial vectors from 10^2 (10×10) to 10^4 (100×100) bins. Three cell types of varying sizes were chosen from the CosMx human liver cancer dataset: a large cell type (Tumor cells), a medium cell type (T cells), and a rare cell type (dendritic cells). (b) The time and memory usage for detecting marker genes with different methods across every tested dataset. We evaluated four approaches: jazzPanda-correlation with 5000 permutations, jazzPanda-glm, limma and the Wilcoxon Rank Sum Test.

For generally large clusters like tumor_1 (over than 70%) and tumor_2 (more than 8%), we can see that different tile lengths produce consistent marker analysis results. The same observation can be made for the two-sample Xenium human breast cancer data (Supplementary Figure S36). For small clusters like Central.venous.LSECs cells (less than 0.5%) or Erthyroid.cells (less than 0.1%), fewer marker genes are detected and using different tile lengths may produce different markers. A larger number of tiles can encode richer spatial information, whereas it could reduce the relative significance of the overall pattern as the cell and transcript counts are reduced with increasing tile size. When we increase the number of tiles to the extreme value (e.g., in microns), we will reach close to a binary representation of all the genes and the clusters. It is crucial to avoid such extreme cases as it breaks the global patterns and introduces significant computational load. On the other hand, a smaller number of tiles can better capture the global pattern but can encode more limited spatial detail. A smaller number of tiles can provide increased tolerance of background noise but may sacrifice the loss of spatial information. In practice, we can calculate the average number of cells across tiles for each cluster to guide the tile size selection. We have found that a general rule of thumb is to choose a tile size such that the average number of cells across all tiles is larger than one for all clusters. Moreover, as we use the lasso regularization to select candidate clusters first, the spatial vector cannot be too sparse which ensures a reliable penalty term can be estimated for every gene.

Computational complexity

We examined the time and memory consumption for different methods of identifying marker genes for each dataset, and at different stages in the marker gene analysis (Figure 7b, Supplementary Figures S39-S40). Our tested datasets contain various number of genes and cells. We measured the time and memory consumption of different stages of the jazzPanda framework on an HPC system equipped with 12 CPU cores and up to 300–400 GB of memory.

Figure 7b compares the time taken for jazzPanda-correlation, jazzPanda-glm, moderated t-tests and Wilcoxon Rank Sum Test for the six different datasets. For one sample datasets, jazzPanda-correlation is the slowest (approximately 20 - 50 minutes) as 5000 permutations are performed to obtain statistical significance. jazzPanda-glm can also take roughly 40 minutes for the largest dataset (MERSCOPE human breast cancer). Wilcoxon Rank Sum Test and moderated t-tests are fast in comparison (10 seconds - 4 minutes).

Looking further into why jazzPanda-glm can be slow for large datasets indicated that

building spatial vectors is the most time-consuming step, and the speed mainly depends on the number of transcript detections (Figure S39). As each spatial vector is calculated independently, we have implemented this process in parallel. Users can utilize multiple cores to speed up the overall process when more computational resources are available. The overall time consumption for detecting marker genes with jazzPanda-glm for the smallest dataset takes less than 5 minutes. The CosMx human liver cancer sample contains over 478 million transcripts in total. With parallel execution restricted to 5 CPU cores, jazzPanda-glm and jazzPanda-correlation both completed within 30 minutes.
We also tested tiles in the shape of squares and hexagons to see if there is a time and memory difference. Hexagonal bins are more time consuming to generate compared to square/rectangular bins (Figure S40), however hexagonal bins can more accurately capture spatial information compared to square/rectangular bins for the same total length of tiles. Hence using square versus hexagonal bins represents a trade-off between time and more accurate representation of the spatial information.

In terms of memory consumption, all methods use more memory for larger datasets. Generally jazzPanda-glm requires more memory than other methods ($\sim$ 8-30 GB).

Discussion

Marker analysis is an important step in the analysis workflow of spatial data. We have developed two approaches to find spatially relevant marker genes and have implemented them in the jazzPanda R Bioconductor package. We have tested our approach on multiple publicly available datasets generated on different imaging-based technology: CosMx, Xenium and MERSCOPE. JazzPanda is a hybrid approach that can utilise transcript information directly when creating gene vectors without requiring that transcripts be assigned to cells. This allows more information to be used than solely relying on cell level gene expression measurements. However, if transcript information is not available, the gene vectors can be created from the cell level gene counts matrix.

Through simulation we have shown that it is important to account for background noise using control probes included in the gene panel. By modelling background noise we are able to reduce the number of false positives observed, and from this standpoint the linear modelling approach is preferred to the correlation approach. We have also compared the results from jazzPanda with commonly used methods in single cell analysis, such as Wilcoxon Rank Sum Tests and moderated t-tests. JazzPanda always identifies marker genes that are a smaller subset of the markers identified by single cell approaches, with the marker genes identified showing high spatial relevance with the associated cluster. With a smaller and more unique set of marker genes, the aim is that cell type identifica-

tion is a simpler task.

An important consideration when performing marker analysis with jazzPanda is that it works best when clusters are more spatially separable, particularly when trying to find unique markers as opposed to shared markers. If the data are over-clustered, some clusters may need to be merged to obtain good results. Examining the cluster-cluster correlation can provide some guidance as to whether clusters significantly overlap in their spatial distribution. The user may also examine the top markers for each cluster as opposed to unique markers in the output of jazzPanda.

Through analysis of different datasets with varying cell type composition, we noted that there is an impact of the number of cells in each cluster on the number of significant markers detected for all methods tested. Larger clusters with more cells generally had more power to detect marker genes compared to smaller clusters, and this was consistent across datasets.

Another consideration is the choice of tile shape and size. For tile shape, we noted that hexagon bins can potentially more accurately capture spatial information compared to square or rectangular bins. However this has an increased computational burden and we found our results with square/rectangular bins were good. We have found that a good rule of thumb for choosing the grid size is to choose a grid size such that the average number of cells across all tiles is at least one for all clusters. This also ensures that the estimates of the gene-wise penalty terms are more stable, and that the linearity assumption is satisfied.

Creating gene and cluster spatial vectors is the most time consuming part of the pipeline, however once this has been computed the linear modelling step is fast. Further downstream analysis can be performed on the spatial vectors such as co-localisation of clusters and genes. This framework lends itself to different types of spatial analysis and could potentially be expanded for spatially informed ligand-receptor analysis, and weighted gene co-expression analysis.

The jazzPanda software is flexible as it is designed to be compatible with any clustering method and does not force the user to use any particular analysis workflow. It is designed to be a stand alone analysis compatible with any analysis workflow that includes grouping cells into biologically meaningful entities. This may not necessarily be from the results of clustering algorithms, but could also be applied to niches or any selected region of interest.

Conclusion

jazzPanda is a novel statistical method to perform spatially aware marker analysis in spatial transcriptomics data. Our method produces highly specific marker genes while controlling the false discovery rate. Our methods are available as functions in the jazzPanda Bioconductor R package.

Methods

All analysis code presented in this manuscript can be found at `https://github.com/phipsonlab/jazzPanda_paper`. The analysis website (`https://phipsonlab.io/jazzPanda_workflowr`) was created using the workflowr (1.7.1) R package. The GitHub repository associated with the analysis website is at: `https://github.com/phipsonlab/jazzPanda_workflowr`.

Statistical model framework for jazzPanda

For any particular spatial dataset, we have n genes profiled for each sample s, $s \in \{1, \ldots, S\}$. Each gene i has d_{is} transcript detections profiled across the tissue sample. Each d_{is} detection has an x and y coordinate associated with it. We apply a spatial binning strategy that divides the spatial domain into tiles $V^1, \ldots, V^h, \ldots, V^v$. The tiles can be square, rectangular or hexagonal.

For each gene $i \in \{1, \ldots, n\}$ in sample s, we count the number of detections d_{is} in each tile. We compute a gene vector $\mathbf{g}_{is} \in \mathbb{R}^v$, where each element g_{is}^h of the gene vector represent the sum of the detections of gene i within the boundary of bin V^h in sample s

$$g_{is}^h = \sum_{(x,y) \in V^h} d_{is}$$

For each cluster $j \in \{1, \ldots, c\}$ in sample s, we use the same binning resolution to count the cell numbers in each tile. We compute a cluster vector $\mathbf{x}_{js} \in \mathbb{R}^v$, where each element x_{js}^h of the cluster vector refers to the number of cells in cluster j that fall inside tile V^h in sample s

$$x_{js}^h = \sum_{(x,y) \in V^h} \text{cell}_{js}$$

We can hence define the gene matrix $\mathbf{G}$ and design matrix $\mathbf{X}$ as follows:

$$\mathbf{G} = \begin{bmatrix} \mathbf{g_1} & \cdots & \mathbf{g}_n \end{bmatrix}, \quad \text{where } \mathbf{g_i} = \begin{bmatrix} g_{i1}^1 \\ \vdots \\ g_{iS}^v \end{bmatrix}$$

and

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1} & \cdots & \mathbf{x_c} & \mathbf{s}_1 & \cdots & \mathbf{s}_S \end{bmatrix},$$

$$\text{where } \mathbf{x_j} = \begin{bmatrix} x_{j1}^1 \\ \vdots \\ x_{jS}^v \end{bmatrix}, \quad \mathbf{s}_r = \begin{bmatrix} \delta_{r1} \\ \vdots \\ \delta_{rS} \end{bmatrix}, \quad \delta_{ri} = \begin{cases} 1 & \text{if } i = r, \\ 0 & \text{otherwise.} \end{cases}$$

Here $\mathbf{s}_s$ is an indicator variable for sample.

Each gene vector $\mathbf{g}_i$ for $i \in [1, \ldots, n]$ is modelled as:

$$\mathbf{g}_i = \mathbf{X}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad \text{where } \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{c+s} \end{bmatrix}$$

where $\boldsymbol{\beta}$ are the coefficients estimated via Lasso regression and the λ_i is selected by cross validation

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \|\mathbf{g}_i - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_i \|\boldsymbol{\beta}\|_1 \right\}$$

We keep only those k clusters with non-zero coefficients as X_k:

$$[\mathbf{x_1}, \ldots, \mathbf{x_k}] = \text{cluster columns of } \mathbf{X} \text{ corresponding to } \hat{\boldsymbol{\beta}}_j \neq 0$$

For each gene $\mathbf{g}_i$, we fit a final generalised linear model with the k selected clusters, sample vectors $\mathbf{s}$ and background noise vectors $\mathbf{f}$

$$\mathbf{g}_i = \mathbf{X_k}\boldsymbol{\beta} + \epsilon \quad \text{where} \quad \mathbf{X_k} = \begin{bmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_k & \mathbf{s}_1 & \ldots & \mathbf{s}_s & \mathbf{f}_1 & \ldots & \mathbf{f}_f \end{bmatrix}, \ \epsilon \sim N(0, \sigma^2)$$

For unique marker genes we further assign gene i as a marker gene for the cluster that has the minimum p-value and maximum coefficient.

Datasets

This study tested six publicly available datasets from three platforms. The raw data can be accessed as follows:

- 1) 10x Xenium Mouse brain data: https://www.10xgenomics.com/resources/datasets/fresh-frozen-mouse-brain-replicates-1-standard
- 2) 10x Xenium Human HER2+ breast cancer data: https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast
- 3) 10x Xenium Human lung cancer data: https://www.10xgenomics.com/resources/datasets/xenium-human-lung-preview-data-1-standard
- 4) Nanostring CosMx Human liver healthy and cancer data: https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/human-liver-rna-ffpe-dataset/
- 5) Vizgen MERSCOPE Human breast cancer data:https://info.vizgen.com/ffpe-showcase?submissionGuid=c9a25730-3fe5-444b-bef8-1a74d51ddefb

Analysis

The analysis was performed using R (4.5.1) and the specific package versions can be viewed under the "Session Information" sections of the every script: https://github.com/phipsonlab/jazzPanda_workflowr

Quality control and processing

For three Xenium data, cells with no transcript detections were removed. Transcript showing low quality transcripts such as unassigned transcript detections and those with quality value less than 20 were filtered out. For the two CosMx data, cells labelled as "NotDet" were filtered. We manually converted the transcript detection coordinates in pixels to micrometers by fitting a linear transformation for every FOV. More specifically, we calculated the slope and intercept with two cells in each FOV, and map the pixels to micrometers for every transcript. For MERSCOPE data, cells with less than total 50 detections or more than 2,500 total detections were removed.

Clustering

Xenium mouse brain data: We applied Banksy (1.4.0) pipeline on the log normalised data (*NormalizeData* function from *Seurat* package). For each sample, adjacency graphs were computed with k_geom = 15 and then joined. The Banksy PCA and UMAP were computed on the joined data with lambda 0.2. Clustering was performed using *clusterBanksy* with resolutions 0.1 and 0.5, followed by *connectClusters* to refine cluster relationships.

Resolution 0.1 was selected for cell type annotation.

Xenium human breast cancer data: We process the two samples separately, and align them based on the shared cell types. For sample 1, we used the provided cell type labels and refined them by collapsing subtypes (e.g., CD4/CD8 T cells and T cell–tumor into T_Cells, macrophage subclusters into Macrophages, myoepithelial subclusters into Myoepithelial cells, dendritic subclusters into Dendritic cells, and invasive/DCIS tumor subtypes into Tumor). For sample 2, we perform Louvain clustering with resolution 0.5, and annotate the resulting clusters with known marker genes. This produced 14 cell types, which were then grouped into broader categories for alignment with sample 1 (e.g., B and Plasma into B_Cells, and T/NK clusters into T_Cells). Finally, cells from nine shared cell types in each sample were joined for testing jazzPanda.

Xenium human lung cancer data: The provided graph-based clustering were used and clusters showing similar spatial patterns were grouped. For example, clusters 7, 10, 11, and 17 were grouped into cluster 6, cluster 14 was reassigned to 4, cluster 21 to 18, clusters 20 and 22 to 15, and clusters 19 and 23 to 5.

MERSCOPE human breast cancer data: We used a similar Banksy pipeline as for the Xenium mouse brain data. Clustering was tested at resolutions 0.5 and 0.8 (*clusterBanksy*), and refined using *connectClusters*. We selected resolution 0.5 for cell type annotation, resulting in 10 clusters.

CosMx human healthy liver and the CosMx liver cancer: The provided cell types were used as cluster labels for these two datasets. Several cell types were grouped as one cluster. For example, Antibody.secreting.B.cells and Mature.B.cells were grouped as B cells, CD3+.alpha.beta.T.cells and gamma.delta.T.cells.1 as T cells, and Non.inflammatory. macrophages and Inflammatory.macrophages were joined as Macrophages cells.

Simulation

We performed the simulation study with the CosMx human liver cancer data to evaluate the false discovery rate. We randomly selected 100 real genes from the panel, and simulate the random pattern for every gene using a near-uniform spatial design rather than fully random placement. We constructed a $\sqrt{n} \times \sqrt{n}$ grid spanning the spatial window of the tissue, where n is the number of total transcript detection for a selected gene. Independent uniform jitter was then added to each coordinate to introduce randomness while reducing the local clustering typical of purely random placement. This same process was applied for generating background control coordinates, ensuring consistent spatial structure across

simulations. After generating the transcripts and background detections, we applied jazzPanda-correlation and jazzPanda-glm with 40×40. This process is repeated 100 times to obtain reliable evaluations of the methods.

jazzPanda

We used different length of square bins to create spatial vectors (*get_vectors*) for Xenium human breast cancer (40×40), Xenium human lung cancer (70×70), Xenium mouse brain (50×50), CosMx human healthy liver (40×40), CosMx human liver cancer (40×40) and MERSCOPE human breast cancer data (50×50). We tested the linear modelling approach (*lasso_markers*) on every dataset, and the correlation (*compute_permp*) approach on dataset with one sample only. To specify the background vectors in the linear modelling approach, the negative control detections were converted to spatial vectors (*create_genesets*). These background vectors were binned in the same strategy as for clusters and genes, ensuring consistency in the spatial representation. For Xenium data, we defined negative control probe and codeword vectors. Similarly for CosMx data, two background vectors were calculated, namely negative probes and falsecode. One background vector was defined from blank genes for the MERSCOPE data.

Wilcoxon Rank Sum Test test

The Wilcoxon Rank Sum Test test was applied through the *FindMarkers* function from *Seurat* R package (5.3.0). Markers genes are selected with significant adjusted p-value and logFC threshold of 0.1.

limma

To apply limma for finding marker genes, the count data was first normalised with the function *normCounts* from *speckle* (1.8.0) R package. The design matrix was specified with one cluster as one group and the contrast matrix was built to compare each cluster with the average of all others. The *limma* (3.64.1) pipeline were applied (*lmFit*, *contrasts.fit* and *eBayes*) and the result was summarised with *decideTests*.

Marker gene interpretation

ChatGPT5 was used to assist marker gene interpretation for Xenium human lung cancer (`https://chatgpt.com/share/68d0b7d5-0840-800f-8c2f-7d1574459ede`), Xenium mouse brain (`https://chatgpt.com/share/68d0b7ef-087c-800f-935d-8d89ed410cac`) and MERSCOPE human breast cancer (`https://chatgpt.com/share/68d3c262-a098-800f-90ca-5f39022d8131`) data.

Effect of the tile length on marker gene selection

We compared the marker gene consistency with different tile lengths for the one-sample CosMx human liver cancer and the two-sample Xenium human breast cancer data. We tested a range of tile lengths, from 10×10 to 100×100 square bins.

Correlation-based approach: We used the *compute_permp* function with different tile lengths to evaluate the marker gene consistency for the correlation based approach. A gene is believed as a marker gene with a significant adjusted p-value and a observed correlation greater than 0.05. Marker genes were compared across a range of tile length for every cluster.

Linear modelling based approach: For a given tile length, the spatial vectors were created for every cluster and every gene (*get_vectors*), as well as for the negative controls (*create_genesets*). The top result is used where one gene is regarded as a marker gene for its most relevant cluster, which were then summarised for visualisation.

Time and memory usage

All analysis was executed on the WEHI HPC cluster (SLURM scheduler). Every job is allocated with 12 cores (Intel(R) Xeon(R) CPU), 200-400 GB memory. SLURM job arrays were used for the simulation and computational complexity examination. We used the function *peakRAM* (1.0.2) in R to measure the time and peak memory usage.

Computational complexity relative to the number of cells and transcripts

To evaluate the time complexity on calculating the spatial vectors relative to the number of cells and transcripts, we incrementally generated random points between 10^6 and 10^9 (Figure S39). Every number of transcript detections were tested repeated for five times. We measured the time and memory usage for building spatial vectors with package *peakRAM* (1.0.2).

Computational complexity relative to the tile shape and length

We tested the time and memory usage of calculating 1000 genes vectors with a diverse range of bin lengths. The function *get_vectors* was examined with square bins between 10^2 (10×10) and 10^4 (100×100) and hex bins with side length 620, 310, 207, 155, 124, 103, 89, and 78. The time and peak memory usage on creating spatial vectors and building linear models was measured with package *peakRAM* (1.0.2). Every tile length was repeated measured five times for a more reliable evaluation.

Availability of data and materials

The method is implemented in the open-source Bioconductor R package **jazzPanda** `https://bioconductor.org/packages/jazzPanda`. All raw and processed datasets used in this study are publicly available via Zenodo `https://zenodo.org/records/18149456`

Acknowledgements

We thank the WEHI Research Computing Platform for providing support and access to their High-Performance Computing facility.

Funding

BP is supported by National Health and Medical Research Council Investigator grant GNT2025641. XJ is supported by a CSL Translational Data Science Scholarship, University of Melbourne Research Scholarship, and WEHI Scientific Excellence Scholarship. This work was made possible through Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS. GS is supported by National Health and Medical Research Council Investigator grant GNT2025645.

References

[1] M. A. A. Al-Rawi, K. Rmali, G. Watkins, R. E. Mansel, and W. G. Jiang. Aberrant expression of interleukin-7 (IL-7) and its signalling complex in human breast cancer. *European Journal of Cancer*, 40(4):494–502, March 2004.

[2] Dvir Aran, Agnieszka P. Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P. Naikawadi, Paul J. Wolters, Adam R. Abate, Atul J. Butte, and Mallar Bhattacharya. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163–172, February 2019. Publisher: Nature Publishing Group.

[3] Rohit Arora, Christian Cao, Mehul Kumar, Sarthak Sinha, Ayan Chanda, Reid McNeil, Divya Samuel, Rahul K. Arora, T. Wayne Matthews, Shamir Chandarana, Robert Hart, Joseph C. Dort, Jeff Biernaskie, Paola Neri, Martin D. Hyrcza, and Pinaki Bose. Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nature Communications*, 14(1):5029, August 2023. Publisher: Nature Publishing Group.

[4] Elham Azizi, Ambrose J. Carr, George Plitas, Andrew E. Cornish, Catherine Konopacki, Sandhya Prabhakaran, Juozas Nainys, Kenmin Wu, Vaidotas Kiseliovas, Manu Setty, Kristy Choi, Rachel M. Fromme, Phuong Dao, Peter T. McKenney, Ruby C. Wasti, Krishna Kadaveru, Linas Mazutis, Alexander Y. Rudensky, and Dana Pe'er. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*, 174(5):1293–1308.e36, August 2018.

[5] Joseph M Beechem, Patrick Danaher, Hiromi Sato, Claire Williams, and Michael Rhodes. Evaluating the Technical Performance of Single-Cell Spatial Molecular Imaging Technologies.

[6] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. Publisher: [Royal Statistical Society, Wiley].

[7] Dharmesh D. Bhuva, Chin Wee Tan, Agus Salim, Claire Marceaux, Marie A. Pickering, Jinjin Chen, Malvika Kharbanda, Xinyi Jin, Ning Liu, Kristen Feher, Givanna Putri, Wayne D. Tilley, Theresa E. Hickey, Marie-Liesse Asselin-Labat, Belinda Phipson, and Melissa J. Davis. Library size confounds biology in spatial transcriptomics data. *Genome Biology*, 25(1):99, April 2024.

[8] Dylan M. Cable, Evan Murray, Vignesh Shanmugam, Simon Zhang, Luli S. Zou, Michael Diao, Haiqi Chen, Evan Z. Macosko, Rafael A. Irizarry, and Fei Chen. Cell

type-specific inference of differential expression in spatial transcriptomics. *Nature methods*, 19(9):1076–1087, September 2022.

[9] Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4):517–526, April 2022. Publisher: Nature Publishing Group.

[10] Jiang Chang, Junting Lu, Qingyi Liu, Tao Xiang, Shaosen Zhang, Yonglin Yi, Dongxu Li, Tianyuan Liu, Zeyuan Liu, Xinjie Chen, Zhenghao Dong, Cainan Li, HanZhang Yi, Siqi Yu, Luwei Huang, Fangfei Qu, Mengdi Wang, Dehe Wang, Hao Dong, Guoyu Cheng, Liang Zhu, Jiachen Li, Chenying Li, Pujie Wu, Xiaoting Xie, Andrew E. Teschendorff, Dongxin Lin, Xiaoqun Wang, and Chen Wu. Single-cell multi-stage spatial evolutional map of esophageal carcinogenesis. *Cancer Cell*, 43(3):380–397.e7, March 2025. Publisher: Elsevier.

[11] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, Xin Wang, Huifang Lu, Xi Chen, Xing Liu, Xin Huang, Zhao Li, Yan Hong, Yujia Jiang, Jian Peng, Shuai Liu, Mengzhe Shen, Chuanyu Liu, Quanshui Li, Yue Yuan, Xiaoyu Wei, Huiwen Zheng, Weimin Feng, Zhifeng Wang, Yang Liu, Zhaohui Wang, Yunzhi Yang, Haitao Xiang, Lei Han, Baoming Qin, Pengcheng Guo, Guangyao Lai, Pura Muñoz-Cánoves, Patrick H. Maxwell, Jean Paul Thiery, Qing-Feng Wu, Fuxiang Zhao, Bichao Chen, Mei Li, Xi Dai, Shuai Wang, Haoyan Kuang, Junhou Hui, Liqun Wang, Ji-Feng Fei, Ou Wang, Xiaofeng Wei, Haorong Lu, Bo Wang, Shiping Liu, Ying Gu, Ming Ni, Wenwei Zhang, Feng Mu, Ye Yin, Huanming Yang, Michael Lisby, Richard J. Cornall, Jan Mulder, Mathias Uhlén, Miguel A. Esteban, Yuxiang Li, Longqi Liu, Xun Xu, and Jian Wang. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*, 185(10):1777–1792.e21, May 2022.

[12] Jiaji George Chen, Joselyn Cristina Chávez-Fuentes, Matthew O'Brien, Junxiang Xu, Edward Ruiz, Wen Wang, Iqra Amin, Irzam Sarfraz, Pratishtha Guckhool, Adriana Sistig, Guo-Cheng Yuan, and Ruben Dries. Giotto Suite: a multi-scale and technology-agnostic spatial multi-omics analysis ecosystem. *bioRxiv: The Preprint Server for Biology*, page 2023.11.26.568752, November 2023.

[13] Mengnan Cheng, Yujia Jiang, Jiangshan Xu, Alexios-Fotios A. Mentis, Shuai Wang, Huiwen Zheng, Sunil Kumar Sahu, Longqi Liu, and Xun Xu. Spatially resolved transcriptomics: A comprehensive review of their technological advances, applications, and challenges. *Journal of Genetics and Genomics*, page S1673852723000759, March 2023.

[14] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, August 2024. Publisher: Nature Publishing Group.

[15] Shuning Ding, Niu Qiao, Qingchen Zhu, Yiwei Tong, Shengyue Wang, Xiaosong Chen, Qiang Tian, Yichuan Xiao, and Kunwei Shen. Single-cell atlas reveals a distinct immune profile fostered by T cell-B cell crosstalk in triple negative breast cancer. *Cancer Communications*, 43(6):661–684, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cac2.12429.

[16] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33:1–22, February 2010.

[17] 10x Genomics. Overview of Xenium Algorithms - Official 10x Genomics Support.

[18] Gege Gui, Molly A. Bingham, Julius R. Herzog, Abigail Wong-Rolle, Laura W. Dillon, Meghali Goswami, Eddie Martin, Jason Reeves, Sean Kim, Arya Bahrami, Hermann F. Degenhardt, George Zaki, Prajan Divakar, Edward C. Schrom, Katherine R. Calvo, Christopher S. Hourigan, Kasper D. Hansen, and Chen Zhao. Single-cell spatial transcriptomics reveals immunotherapy-driven bone marrow niche remodeling in AML. *Science Advances*, 11(28):eadw4871, July 2025. Publisher: American Association for the Advancement of Science.

[19] Jinil Han, Yoon-La Choi, Haein Kim, Jun Young Choi, Se Kyung Lee, Jeong Eon Lee, Joon-Seok Choi, Sarah Park, Jong-Sun Choi, Young Deug Kim, Seok Jin Nam, Byung-Ho Nam, Mi Jeong Kwon, and Young Kee Shin. MMP11 and CD2 as novel prognostic factors in hormone receptor-negative, HER2-positive breast cancer. *Breast Cancer Research and Treatment*, 164(1):41–56, July 2017.

[20] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021. Publisher: Elsevier.

[21] Shanshan He, Ruchir Bhatt, Carl Brown, Emily A. Brown, Derek L. Buhr, Kan Chantranuvatana, Patrick Danaher, Dwayne Dunaway, Ryan G. Garrison, Gary

Geiss, Mark T. Gregory, Margaret L. Hoang, Rustem Khafizov, Emily E. Killingbeck, Dae Kim, Tae Kyung Kim, Youngmi Kim, Andrew Klock, Mithra Korukonda, Alecksandr Kutchma, Zachary R. Lewis, Yan Liang, Jeffrey S. Nelson, Giang T. Ong, Evan P. Perillo, Joseph C. Phan, Tien Phan-Everson, Erin Piazza, Tushar Rane, Zachary Reitz, Michael Rhodes, Alyssa Rosenbloom, David Ross, Hiromi Sato, Aster W. Wardhani, Corey A. Williams-Wietzikoski, Lidan Wu, and Joseph M. Beechem. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature Biotechnology*, 40(12):1794–1806, December 2022. Number: 12 Publisher: Nature Publishing Group.

[22] Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A. Morrison, Michelli F. Oliveira, Jordan T. Sicherman, Andrew Kohlway, Jawad Abousoud, Tingsheng Yu Drennon, Seayar H. Mohabbat, and Sarah E. B. Taylor. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353, December 2023. Publisher: Nature Publishing Group.

[23] Rustem Khafizov, Erin Piazza, Yi Cui, Michael Patrick, Evelyn Metzger, Daniel McGuire, Dwayne Dunaway, Patrick Danaher, Margaret L. Hoang, Andrew Grootsky, Megan Vandenberg, Shanshan He, Rachel Liu, Michael McKean, Michael Rhodes, and Joseph M. Beechem. Sub-cellular Imaging of the Entire Protein-Coding Human Transcriptome (18933-plex) on FFPE Tissue Using Spatial Molecular Imaging, December 2024. Pages: 2024.11.27.625536 Section: New Results.

[24] Robin Kit-Ho Lai, Iris Ming-Jing Xu, David Kung-Chun Chiu, Aki Pui-Wah Tse, Larry Lai Wei, Cheuk-Ting Law, Derek Lee, Chun-Ming Wong, Maria Pik Wong, Irene Oi-Lin Ng, and Carmen Chak-Lui Wong. NDUFA4L2 Fine-tunes Oxidative Stress in Hepatocellular Carcinoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 22(12):3105–3117, June 2016.

[25] Yunxi Liao, Chenyang Wu, Yang Li, Jinhua Wen, and Dongyu Zhao. MIF is a critical regulator of mononuclear phagocytic infiltration in hepatocellular carcinoma. *iScience*, 26(8):107273, July 2023.

[26] Boxiang Liu, Yanjun Li, and Liang Zhang. Analysis and Visualization of Spatial Transcriptomic Data. *Frontiers in Genetics*, 12, January 2022. Publisher: Frontiers.

[27] Sharon L. Lohr. *Sampling: Design and Analysis*. Chapman and Hall/CRC, New York, 3 edition, November 2021.

[28] Yang Ma, Chen-chen Han, Yifan Li, Yang Wang, and Wei Wei. Insulin-like growth factor-binding protein-3 inhibits IGF-1-induced proliferation of human hepatocellular carcinoma cells by controlling bFGF and PDGF autocrine/paracrine loops. *Biochemical and Biophysical Research Communications*, 478(2):964–969, September 2016.

[29] Jeffrey R. Moffitt, Emma Lundberg, and Holger Heyn. The emerging landscape of spatial profiling technologies. *Nature Reviews Genetics*, 23(12):741–759, December 2022. Number: 12 Publisher: Nature Publishing Group.

[30] Han-Eol Park, Song Hyun Jo, Rosalind H. Lee, Christian P. Macks, Taeyun Ku, Jihwan Park, Chung Whan Lee, Junho K. Hur, and Chang Ho Sohn. Spatial Transcriptomics: Technical Aspects of Recent Developments and Their Applications in Neuroscience and Cancer Research. *Advanced Science*, 10(16):2206939, June 2023.

[31] Valéry L. Payen, Arnaud Lavergne, Niki Alevra Sarika, Megan Colonval, Latifa Karim, Manon Deckers, Mustapha Najimi, Wouter Coppieters, Benoît Charloteaux, Etienne M. Sokal, and Adil El Taghdouini. Single-cell RNA sequencing of human liver reveals hepatic stellate cell heterogeneity. *JHEP reports: innovation in hepatology*, 3(3):100278, June 2021.

[32] Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9:Article39, 2010.

[33] Hannah A. Pliner, Jay Shendure, and Cole Trapnell. Supervised classification enables rapid annotation of cell atlases. *Nature Methods*, 16(10):983–986, October 2019. Publisher: Nature Publishing Group.

[34] Jeffrey M. Pullin and Davis J. McCarthy. A comparison of marker gene selection methods for single-cell RNA sequencing data. *Genome Biology*, 25:56, February 2024.

[35] Yuchen Qi, Yinghui Song, Mengting Cai, Jianwen Li, Zhangtao Yu, Yuhang Li, Junkai Huang, Yu Jiang, Chuang Peng, Bo Jiang, and Sulai Liu. Vascular endothelial growth factor A is a potential prognostic biomarker and correlates with immune cell infiltration in hepatocellular carcinoma. *Journal of Cellular and Molecular Medicine*, 27(4):538–552, February 2023.

[36] Xuyu Qian, Kyle Coleman, Shunzhou Jiang, Andrea J. Kriz, Jack H. Marciano, Chunyu Luo, Chunhui Cai, Monica Devi Manam, Emre Caglayan, Abbe Lai, David Exposito-Alonso, Aoi Otani, Urmi Ghosh, Diane D. Shao, Rebecca E. Andersen, Jennifer E. Neil, Robert Johnson, Alexandra LeFevre, Jonathan L. Hecht, Nicola

Micali, Nenad Sestan, Pasko Rakic, Michael B. Miller, Liang Sun, Carsen Stringer, Mingyao Li, and Christopher A. Walsh. Spatial transcriptomics reveals human cortical layer and area specification. *Nature*, 644(8075):153–163, August 2025. Publisher: Nature Publishing Group.

[37] Zhihua Qiu, Shaojun Li, Ming Luo, Shuanggen Zhu, Zhijian Wang, and Yongjun Jiang. Detection of differentially expressed genes in spatial transcriptomics data by spatial analysis of spatial transcriptomics: A novel method based on spatial statistics. November 2022. Publisher: Frontiers.

[38] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, April 2015.

[39] 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 (New York, N.Y.)*, 363(6434):1463–1467, March 2019.

[40] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, June 2019. Publisher: Elsevier.

[41] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (New York, N.Y.)*, 353(6294):78–82, July 2016.

[42] Wenfei Sun, Zhihui Liu, Xian Jiang, Michelle B. Chen, Hua Dong, Jonathan Liu, Thomas C. Südhof, and Stephen R. Quake. Spatial transcriptomics reveal neuron–astrocyte synergy in long-term memory. *Nature*, 627(8003):374–381, March 2024. Publisher: Nature Publishing Group.

[43] J. Kenneth Tay, Balasubramanian Narasimhan, and Trevor Hastie. Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software*, 106:1–31, March 2023.

[44] Chunli Wang, Lingzu Kong, Seokmin Kim, Sunyoung Lee, Sechan Oh, Seona Jo, Inhwan Jang, and Tae-Don Kim. The Role of IL-7 and IL-7R in Cancer Pathophysiology and Immunotherapy. *International Journal of Molecular Sciences*, 23(18):10412, September 2022.

[45] Cameron G. Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14(1):68, June 2022.

[46] Huijun Xie, Guihui Tong, Yupei Zhang, Shu Liang, Kairui Tang, and Qinhe Yang. PGK1 Drives Hepatocellular Carcinoma Metastasis by Enhancing Metabolic Process. *International Journal of Molecular Sciences*, 18(8):1630, July 2017.

[47] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10):852–866, October 2022. Publisher: Nature Publishing Group.

[48] Zizhen Yao, Cindy T. J. van Velthoven, Michael Kunst, Meng Zhang, Delissa McMillen, Changkyu Lee, Won Jung, Jeff Goldy, Aliya Abdelhak, Matthew Aitken, Katherine Baker, Pamela Baker, Eliza Barkan, Darren Bertagnolli, Ashwin Bhandiwad, Cameron Bielstein, Prajal Bishwakarma, Jazmin Campos, Daniel Carey, Tamara Casper, Anish Bhaswanth Chakka, Rushil Chakrabarty, Sakshi Chavan, Min Chen, Michael Clark, Jennie Close, Kirsten Crichton, Scott Daniel, Peter DiValentin, Tim Dolbeare, Lauren Ellingwood, Elysha Fiabane, Timothy Fliss, James Gee, James Gerstenberger, Alexandra Glandon, Jessica Gloe, Joshua Gould, James Gray, Nathan Guilford, Junitta Guzman, Daniel Hirschstein, Windy Ho, Marcus Hooper, Mike Huang, Madie Hupp, Kelly Jin, Matthew Kroll, Kanan Lathia, Arielle Leon, Su Li, Brian Long, Zach Madigan, Jessica Malloy, Jocelin Malone, Zoe Maltzer, Naomi Martin, Rachel McCue, Ryan McGinty, Nicholas Mei, Jose Melchor, Emma Meyerdierks, Tyler Mollenkopf, Skyler Moonsman, Thuc Nghi Nguyen, Sven Otto, Trangthanh Pham, Christine Rimorin, Augustin Ruiz, Raymond Sanchez, Lane Sawyer, Nadiya Shapovalova, Noah Shepard, Cliff Slaughterbeck, Josef Sulc, Michael Tieu, Amy Torkelson, Herman Tung, Nasmil Valera Cuevas, Shane Vance, Katherine Wadhwani, Katelyn Ward, Boaz Levi, Colin Farrell, Rob Young, Brian Staats, Ming-Qiang Michael Wang, Carol L. Thompson, Shoaib Mufti, Chelsea M. Pagan, Lauren Kruse, Nick Dee, Susan M. Sunkin, Luke Esposito, Michael J. Hawrylycz, Jack Waters, Lydia Ng, Kimberly Smith, Bosiljka Tasic, Xiaowei Zhuang, and Hongkui Zeng. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature*, 624(7991):317–332, December 2023. Publisher: Nature Publishing Group.

[49] Allen W. Zhang, Ciara O'Flanagan, Elizabeth A. Chavez, Jamie L. P. Lim, Nicholas Ceglia, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany Hewitson, Daniel Lai, Anja Mottok, Clementine Sarkozy, Lauren Chong, Tomohiro Aoki, Xuehai Wang, Andrew P. Weng, Jessica N. McAlpine, Samuel Aparicio, Christian Steidl, Kieran R. Campbell, and Sohrab P. Shah. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nature Methods*, 16(10):1007–1015, October 2019. Publisher: Nature Publishing Group.

[50] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4:Article17, 2005.

[51] Lihong Zhang, Tao Lu, Ye Yang, and Liangfeng Hu. -enolase is highly expressed in liver cancer and promotes cancer cell invasion and metastasis. *Oncology Letters*, 20(5):152, November 2020.

[52] Qiuxiang Zheng, Cuiping Lu, Lian Yu, Ying Zhan, and Zhiyong Chen. Exploring the metastasis-related biomarker and carcinogenic mechanism in liver cancer based on single cell technology. *Heliyon*, 10(6):e27473, March 2024.