Analysis of anatomical multi-cellular structures from spatial omics data using *sosta*

Samuel Gunz[1], Helena L. Crowell[2], and Mark D. Robinson[1,*]

[1]Department of Molecular Life Sciences and
SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[2]Centro Nacional de Análisis Genómico, Barcelona, Spain
[*]Correspondence to: mark.robinson@mls.uzh.ch

October 13, 2025

Abstract

Spatial omics technologies enable high-resolution, large-scale quantification of molecular features while preserving the spatial context within tissues. Existing analysis methods largely focus on spatial arrangements of single cells, whereas biological function often emerges from multicellular arrangements. Here, we introduce structure-based analysis of spatial omics data, which focuses on the direct analysis of multicellular, anatomical structures. We illustrate this type of analysis using two publicly available datasets and provide *sosta*, an open-source Bioconductor package for broad community use.

Introduction

Spatial organisation reflects biological function across scales. Cells and tissues tend to self-organize into spatially segregated anatomical structures with distinct biological function, which is often reflected in the morphology and location of such multicellular structures and organs [1] (Figure 1a-b). Recent advances have enabled researchers to quantify molecular features at an unprecedented scale and resolution, while preserving their spatial context [2] (Figure 1c). This motivates a data analysis approach that focuses on the anatomical structure dimension of spatial omics data.

While many computational methods have been proposed to analyse emerging types of spatial omics data [3], they usually focus on the arrangement of individual cells, often within a single tissue or slide; for example, the analysis of cell neighbourhoods, spatial domains, or cell-cell interaction patterns. Moreover, methods to detect spatially variable features or gradients often do not account for heterogeneity at the level of anatomical structures [2, 4]. In addition, few methods exist for the analysis of multi-sample multi-condition spatial datasets [3]. To date, general frameworks and strategies are lacking to perform analysis at the (anatomical) structure level, where multicellular arrangements become the focus of analysis. Some early attempts have been made, such as extracting spatial domains and quantifying geometric features [5, 6].

Here, we introduce a new genre of data analysis, so-called structure-based analysis, whereby (computationally-extracted) anatomical structures represent the basic unit and resolution to analyse the data at (Figure 1c-d). We demonstrate this type of analysis using two datasets: quantifying structural rearrangements during colorectal malignancy transformation; and, recovery of structurally relevant gene expression gradients in human tonsil germinal centres. To build a foundation for such analyses, we also introduce *sosta* (spatial omics structure analysis), an open-source Bioconductor [7] package for reconstruction, characterization and comparison of anatomical structures from spatial omics data across scales.

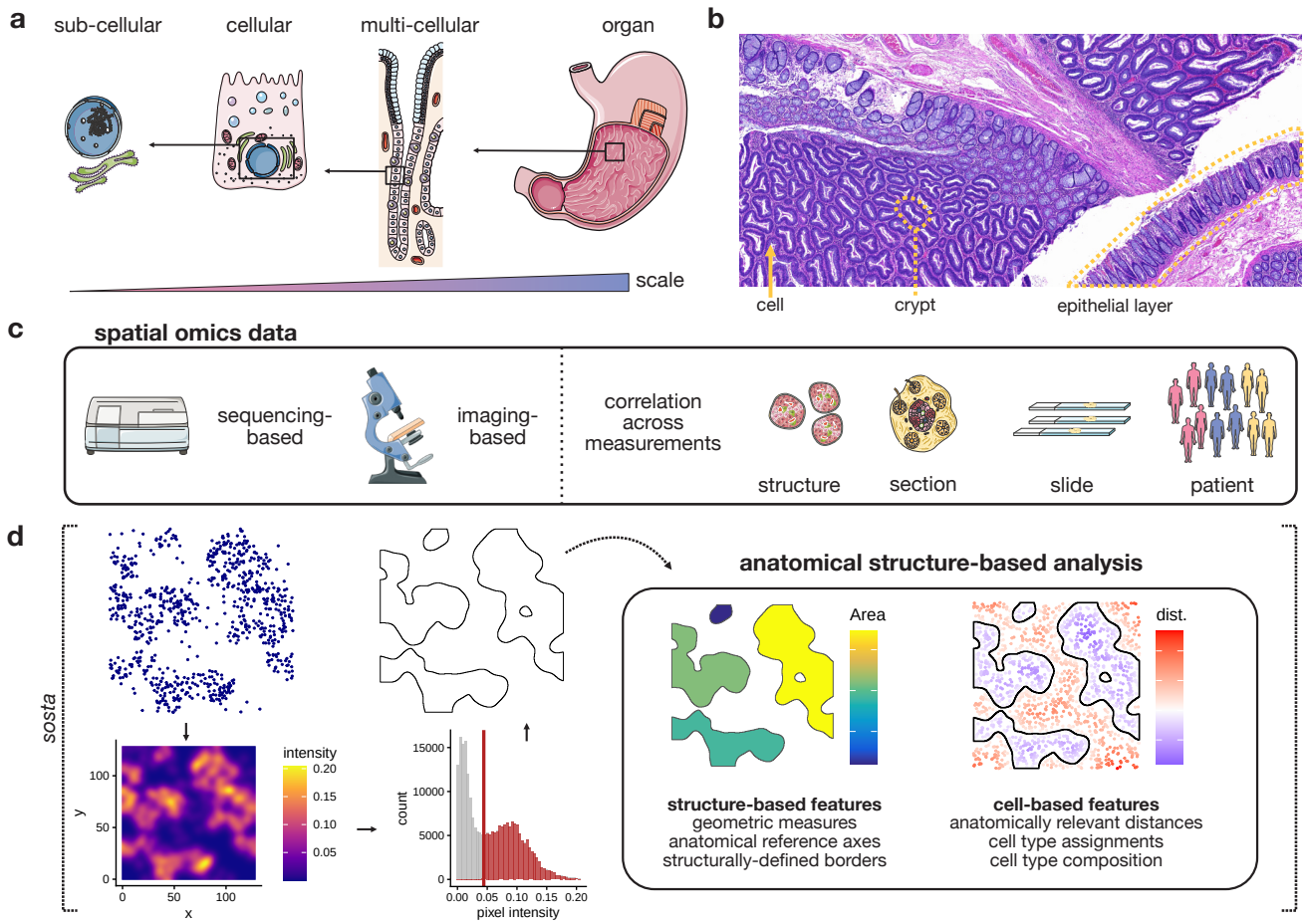


Figure 1: **Structure-based analysis of spatial omics data.** **a**) Spatial arrangements reflect functions across scale, ranging from the sub-cellular to the organ level. **b**) An example histopathology image that illustrates diverse multicellular anatomical structures in the transition from healthy intestinal epithelium to pre-malignant lesions (taken from [8]). **c**) Spatial omics technologies are used to study biological phenomena across different scales, ranging from subcellular to multicellular, with correlation also present at different scales. **d**) Illustrated workflow of density-based reconstruction of anatomical structures using a threshold on the point pattern intensity of cell centroids as implemented in the `sosta` package, followed by the analysis of anatomical structures by quantifying geometric features, or by using anatomical-structure information as a reference scale for downstream analysis. Illustrations in **a** and **c** were taken from Servier Medical Art (`https://smart.servier.com/`), licensed under CC-BY 4.0.

Results

Using our analysis approach, we aim to bridge the gap between classical histopathology and spatial transcriptomics by directly quantifying structural rearrangement and subsequent changes in the transcriptional landscape. To characterize structural differences, we rely on a structure reconstruction method that determines a cut-off on the point pattern density of cells or transcripts implemented in our package *sosta*. This approach allows for flexible and fast "segmentation" of a set of (disconnected) anatomical structures by relying on coordinates of associated features of interest, such as cell types or transcript assignments (Figure 1b, Methods). We then focus on structure-based analysis by quantifying a set of metrics related to the recovered anatomical structures and using this resolution to examine the data (Figure 1d).

To highlight structure-based analysis of spatial transcriptomics data, we first quantified structural changes in the transformation from healthy human colon (HC) tissue to colorectal cancer (CRC) [8] (Figure 2a). This transition from HC crypt epithelial cells to CRC is a process that includes intermediate stages, such as polyps or tubulovillous adenomas (TVAs). These stages are typically identified by histopathology since they are

characterized by discernible cellular morphology and structural rearrangement. The transcriptional differences and trajectory along this transition have been studied in Crowell et al. [8].

To accentuate morphological changes between HC and (pre-)malignant regions, we calculated a range of geometric features using *sosta* and centred on a geometric metric called "fibre width", which represents the width of the outside linear epithelial border (Figure 1c, Suppl. Table S1, Methods). Fibre width increased along the pathological trajectory from HC to TVA to CRC, both at the structure and cell type level (Figure 2d-e). To further study this progression, we compared the association of fibre width with the median pseudo-time per structure as inferred in the original publication for each cell. We observed a strong positive correlation between the median fibre width per epithelial structure and the median pseudo-time of cells within the structure (Spearmans's ρ = 0.46, *p*-value $< 10^{-10}$, Figure 2f). In addition to structural analysis of the epithelial cell layer, we further reconstructed and quantified individual epithelial crypts (Figure 2g). We found that the "shape index", a morphological measure of a shape's regularity, reflects the transition from regular crypts in the normal epithelium to irregular ones in premalignant regions (Spearmans's ρ = 0.22, *p*-value $< 10^{-10}$, Figure 2h). Furthermore, we used the reconstructed crypts to identify local transitions between normal and pre-malignant tissue, i.e., transition crypts. By filtering on the frequency of cell types adjacent to the intestinal crypts, we could recover manually annotated transition crypts and identify *de novo* transition crypts (Methods, Figure 2i, Suppl. Figure S1d-f).

By using structure-based analysis, we gained additional insight about the morphological transition during the progression to CRC. For illustration, this was done on a single slide. However, this type of analysis could also be performed with multiple replicates and across conditions. In this case, it is crucial to account for correlation of the repeated measurements (i.e., structures) on the same experimental unit (i.e., sample) (Figure 1c, see Supplementary Note).

In addition to geometric features, anatomical structures offer a baseline for structure-aware analysis of gene expression. Unlike conventional spatially variable gene analysis, this approach relates spatial variation of gene expression to meaningful anatomical axes. To demonstrate this, we performed structure-based analysis of a spatial transcriptomics dataset of human tonsil tissue [9] (Figure 2e). Relevant anatomical structures in this dataset are germinal centres (GCs), locations where B cells mature along a defined anatomical trajectory: from light to dark zone (L/DZ) [10]. We used our density-based reconstruction approach to define GCs and, within each GC, constructed anatomically relevant axes by: i) identifying anatomical zonations in each GC; and, ii) projecting a matching axis through each GC (Figure 2j-k, Methods). This allowed for quantification of gene expression along a unified anatomically relevant gradient across GCs.

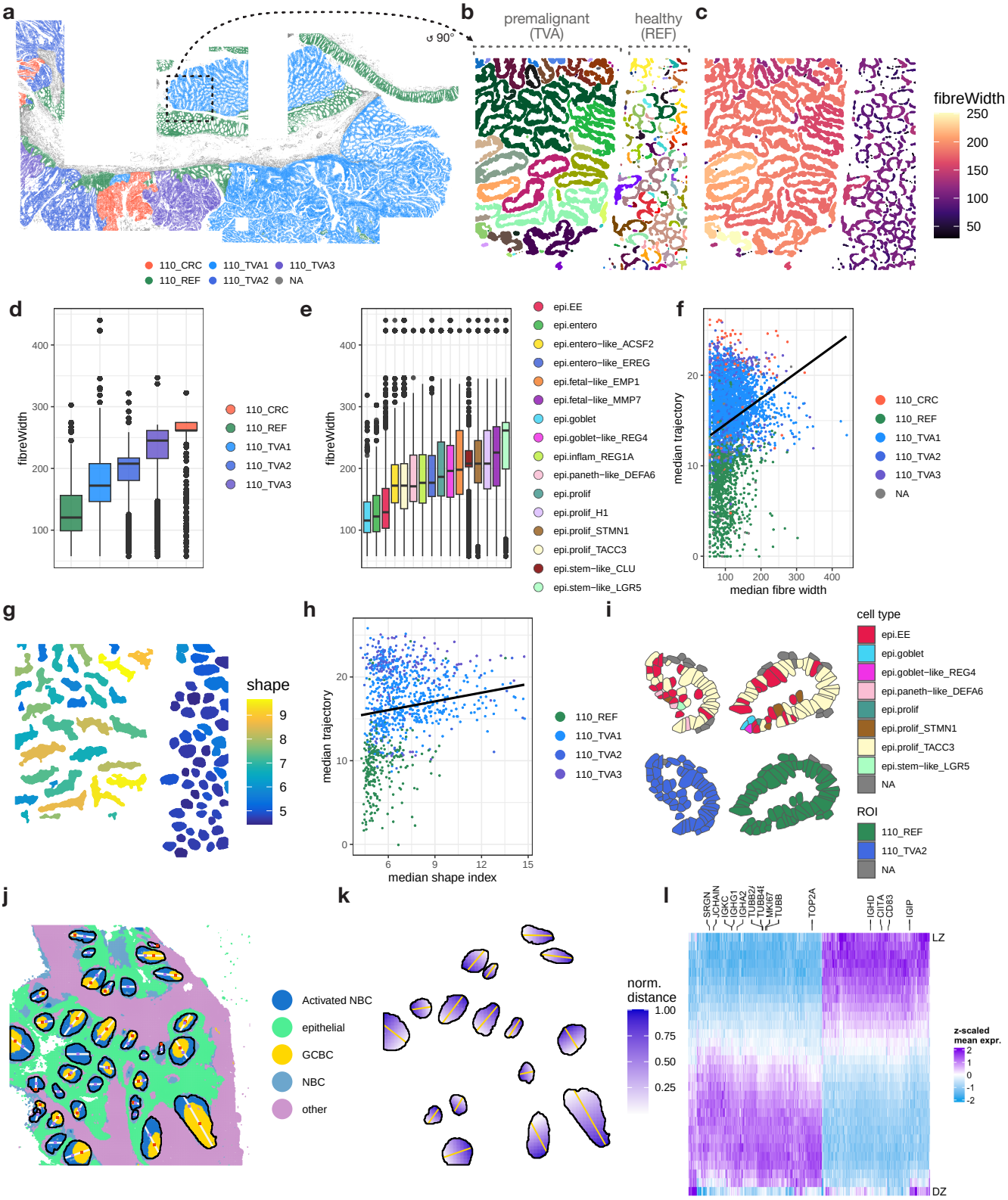
Next, we averaged expression of all GC pixels within binned distances and used the maximal information coefficient (MIC) [11], a metric that can capture linear and non-linear relationships, to identify genes that showed gene expression changes along the LZ-DZ axis. We identified a total of 1569 genes that showed a spatially varying gene expression profile (MIC $>$ 0.999). These included known markers of B cells proliferation, maturation, and antigen presentation (e.g., MKI67, CIITA) [10, 12] (Figure 2l).

In this example, structure-based analysis offered a straightforward approach to recover interpretable and anatomically relevant gene expression gradients without relying on discrete cell types.

Discussion

To date, spatial omics analysis methods have primarily focused on the single-cell level, often failing to fully capture heterogeneity arising from differences in tissue architecture and cellular composition. Previous work has shown that scale is an important factor when analysing spatial omics data [13, 14]. This motivates a structure-based analysis approach that is aware of multicellular anatomical patterns, as has been shown in classic histopathology for decades [15]. Our package *sosta* functions as a modular and extensible tool within the Bioconductor framework to perform structure-based analysis at different scales. We provide comprehensive vignettes to enable a user friendly experience (see `https://bioconductor.org/packages/sosta`).

In addition, our density-based reconstruction approach allows for analysis of anatomical structures ranging from few cells or transcripts such as tertiary lymphoid struc-



(Figure legend on next page.)

Figure 2: **Two examples of structure-based analysis of spatial transcriptomics data a**) Overview of section 110 of the Crowell et al. dataset (REF = reference / healthy colon, TVA = TVA regions, CRC = colorectal cancer regions). **b**) Subregion indicated by the dotted box in **a**, rotated counter-clockwise by 90°, that reflects the transition between normal colonic crypts and premalignant TVA regions. Results of the *sosta* density-based reconstruction are randomly coloured to highlight unique structures. **c**) Epithelial structures in b) coloured by their respective fibre width. **d**) Boxplot of the distribution of fibre width per cell (each cell was assigned the fibre width metric of its corresponding epithelial structure) grouped by the assigned region. **e**) Boxplot of per-cell fibre widths grouped by cell types. x-axis ordered by median fibre width. **f**) Association between the median fibre width per epithelial structure and its median (inferred) pseudo-time [8]. **g**) Reconstructed epithelial crypts of b). The colour reflects their corresponding "shape metric". **h**) Association between the median pseudo-time and the median shape index per reconstructed epithelial crypt. **i**) Two examples of *de novo* identified transition crypts using structure-based analysis. **j**) Structure-based analysis of germinal centres (GCs) in the human tonsil [9]; black outlines correspond to reconstructed GCs, colours reflect inferred clusters. Coloured squares correspond to reference points used to reconstruct anatomical reference axes (white lines). **k**) Reference axes through GCs; colours correspond to distances scaled between 0 and 1 within each GC. **i**) Heatmap of 1569 genes with spatially-varying expression within the GCs. Highlighted are known marker genes. LZ = light zone, DZ = dark zone.

tures, up to the analysis of cancer tissue comprising thousands of cells. In these cases, the structure of interest were known in advance, but the approach we use and facilitate is a general one. Moreover, the type of structure-based analyses we have shown are applicable to a diverse set of biological tissues and questions. To date, researchers often use prior biological knowledge to guide selection of metrics that quantify relevant aspects in their data. This offers further opportunities to investigate unsupervised methods to first identify and quantify biologically relevant features at the anatomical structure level [16]. By leveraging existing statistical frameworks for analysis of structure-derived features, we show how structure-based analysis can ease biological discovery. However, researchers must heed the correlation structure of structure-based analysis of spatial omics data, since multiple measurements from the same experimental unit are extracted (see Supplementary Note).

While our framework offers methods to reconstruct, quantify and compare anatomical structures ranging from single to multi-sample multi-condition setups, we acknowledge that there exists a plethora of methods that can be applied in each step of structure-based analysis. For example, a range of methods exists to reconstruct and segment (multi-)cellular structures, the space of possible metrics that quantify anatomical structures is vast and different strategies to detect gradients exist [17]. Nonetheless, our flexible and extensible framework can accommodate future developments in this space.

Methods

Density-based reconstruction of anatomical structures

Reconstruction of anatomical structures was performed by using a cut-off on a point pattern density estimate, where the point pattern corresponds either the centroids of selected cell type(s) or the locations of transcripts of interest (e.g., features marking a cell type, functional pathway, or other biological aspects). As described by Diggle [18], a density image is estimated and given by

$$\hat{\lambda}(u) = e(u)\Sigma_i k(x_i - u)w_i$$

where k is a Gaussian kernel, e(u) is an edge correction factor and w are the weights.

By default, the smoothing kernel is estimated using cross validation on the bandwidth selection as described by Diggle [18]. The density estimation and kernel selection are implemented in the R package `spatstat` [19].

To estimate the cut-off threshold value c for reconstruction, we use $c = \frac{m_1+m_2}{2}$ where m_1 and m_2 are the density values at the first and second mode of the intensity distribution over all pixels in the estimated density image. The cut-off was then manually adjusted. Overall, downstream results were consistent across a broad range of parameter setting (i.e., bandwidth of smoothing kernel and cut-off threshold), indicating robustness of our segmentation approach. However, at certain parameters values we observed substantial deviations, especially in the number of detected structures (Figure S1c).

Functions for density-based reconstruction are implemented in the R/Bioconductor package `sosta`.

Reconstruction and quantification of epithelial crypts in colon tissue

We used annotations of epithelial cells (epi) as determined by the original publication for reconstruction of epithelial structures (Suppl. Figure S1a). Reconstruction was performed using the function `reconstructShapeDensitySPE` (parameters: bndw = 50, thres = $9 \cdot 10^{-5}$, dim = 2500) of `sosta` [20] (Suppl. Figure S1b).

Structure metrics To describe the morphology of the epithelial structures, we calculate the following geometric measurements. The fibre width is given by $w = \frac{P-\sqrt{P^2-16A}}{4}$, where P is the perimeter of structure and A is its area. The shape index is calculated by $s = \frac{P}{\sqrt{A}}$, where P is the perimeter and A the area of the structure. Geometric calculations were conducted using the `sf` package [21]. Association of structure metrics with histological regions, cell types and the pseudo-time was done using the authors' annotations. For association of structure metrics and pseudo-time, we either assigned a structure metric to each cell by finding the intersection of structures and cell centroid coordinates, or calculated the median pseudo-time per structure.

Epithelial crypts To recover crypts, we applied the function `reconstructShapeDensitySPE` of `sosta` to all cells that fell below the density threshold `complement = TRUE` (parameters: bndw = 50, thres = $9 \cdot 10^{-5}$, dim = 2500). To identify transition crypts, we assigned all cells within a certain distance to the crypt boundary to a specific crypt and then calculated the fraction of cell types per crypt. We next considered cell-level annotations from [8] reflecting their association to malignancy (reference, TVA or CRC region), and determined a threshold on composition to identify transition crypts (i.e., crypts containing between 20 and 70 cells, ¿50% epithelial cells, and among those epithelial cells: ¿10% reference and ¿50% TVA or CRC cells, Suppl. Figure S1d-f).

Reconstruction and analysis of anatomically relevant gradients in germinal centres of the human tonsil

Preprocessing of the 10x Visium HD human tonsil dataset followed recommendations of the online book OSTA (`https://lmweber.org/OSTA/`). We used the package `SpotSweeper` [22] for initial quality control (Suppl. Figure S2a). Library-size normalization, highly-variable gene selection, and dimensionality reduction was performed using packages `scater` [23] and `scrapper` [24]. We applied `Banksy` to perform spatial clustering of the data [25] (parameters: lambda = 0.5, k = 50, resolution = 0.8, Suppl. Figure S2b). We used `scrapper` and `scater` for aggregation of genes in clusters. Annotation of anatomical regions was done using the function `findMarkers` in `scran` [26], followed by the function `SingleR` [27] to match marker genes to the reference dataset Massoni-Badosa et al. [12] (Suppl. Figure S2c).

Identification of germinal centres Germinal centres (GCs) were identified by combining the clusters "germinal centre B cells" (GCBCs) and "activated naïve B cells" (NBCs). Reconstructions were performed using `reconstructShapeDensitySPE` of `sosta` (parameters: thres = $3 \cdot 10^{-4}$) (Suppl. Figure S2d). To define anatomically relevant reference axes, we computed centroids of the GCBC cluster (Figure 2j, red squares) and NBC clusters (Figure 2j, violet squares) and connected them with a linear segment (Figure 2j, white line). Reference anchor points were determined as the side where the linear axis intersected the GC outline on the NBC side. A perpendicular line through each anchor was then calculated to establish a normalized radial distance measure within each GC.

Identification of gene expression gradients For subsequent analysis, we selected 16 germinal centres with both GCBC and NBC zones (Figure 2k). The normalized distances within each GC were binned into 30 bins and then aggregated into pseudo-bulk profiles using `scrapper`. Mutual information coefficients (MIC) were calculated using `minerva` [28] to determine genes with a (non)-linear association with the spatial position along the GC axes. Genes with a MIC > 0.999 were selected for visualisation in Figure 2l.

Heatmaps of distance-binned and pseudo-bulked expression profiles were generated using `pheatmap` [29] and `ComplexHeatmap` [30]. Exemplary light zone (LZ) and dark zone (DZ) markers in Figure 2l were selected based on published datasets [10, 12].

Code and data availability

Analyses were conducted using R version 4.5.0 and Bioconductor version 3.21 [7]. Results were visualised using `ggplot2` [31]. `sosta` is freely available from Bioconductor (v 3.21 or higher) `https://bioconductor.org/packages/sosta`.
Code to reproduce the analyses here is available at `https://github.com/sgunz/sosta-manuscript`. All datasets used are publicly available. The CosMx TVA dataset is available at `https://zenodo.org/records/15574384` [32]. The IMC pancreatic islets dataset is accessible through Bioconductor's `ExperimentHub` via the R package `imcdatasets` [33]. The Visium HD human tonsil data is available from the 10x Genomics Datasets resource [9].

Acknowledgments

We thank all members of the Robinson lab for helpful comments and feedback. We want to thank Martin Emons for continuos feedback and support, Izaskun Mallona and Giulia Moro for helpful insights into anatomy and biology, and Bianca Dumitrascu for helpful comments on the analysis.

Medical illustrations in Figure 1 were adapted from Servier Medical Art `https://smart.servier.com/`, licensed under CC-BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`).

Author contributions

SG: Conceptualisation, Methodology, Analysis, Visualisation, Software, Writing - Original Draft; **HLC:** Conceptualisation, Methodology, Writing - Original Draft, Supervision; **MDR:** Conceptualisation, Methodology, Writing - Original Draft, Supervision, Funding acquisition.

Funding

This project was supported by the Swiss National Science Foundation (SNSF) project grant 310030_204869 and the University Research Priority Program Evolution in Action at the University of Zurich to MDR. HLC recognizes support of SNSF project grant 222136.

Conflict of interest

We declare no conflict of interest.

References

[1] Katy Börner et al. "Anatomical structures, cell types and biomarkers of the Human Reference Atlas". en. In: *Nature Cell Biology* 23.11 (Nov. 2021), pp. 1117–1128. DOI: 10.1038/s41556-021-00788-6.

[2] Dario Bressan, Giorgia Battistoni, and Gregory J. Hannon. "The dawn of spatial omics". In: *Science* 381.6657 (Aug. 2023), eabq4964. DOI: 10.1126/science.abq4964.

[3] Lambda Moses and Lior Pachter. "Museum of spatial transcriptomics". en. In: *Nature Methods* 19.5 (May 2022), pp. 534–546. DOI: 10.1038/s41592-022-01409-2.

[4] Jian Hu et al. "Statistical and machine learning methods for spatially resolved transcriptomics with histology". In: *Computational and Structural Biotechnology Journal* 19 (July 2021), pp. 3829–3841. DOI: 10.1016/j.csbj.2021.06.052.

[5] Marco Varrone et al. "CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity". en. In: *Nature Genetics* (Dec. 2023), pp. 1–11. DOI: 10.1038/s41588-023-01588-4.

[6] A. M. Sequeira et al. *GRIDGENE: Guided Region Identification based on Density of GENEs – a transcript density-based approach to characterize tissues by spatial transcriptomics*. en. Aug. 2025. DOI: 10.1101/2025.08.14.670318.

[7] Wolfgang Huber et al. "Orchestrating high-throughput genomic analysis with Bioconductor". en. In: *Nature Methods* 12.2 (Feb. 2015), pp. 115–121. DOI: 10.1038/nmeth.3252.

[8] Helena L. Crowell et al. *Tracing colorectal malignancy transformation from cell to tissue scale*. en. June 2025. DOI: 10.1101/2025.06.23.660674.

[9] *Visium HD Spatial Gene Expression Library, Human Tonsil (Fresh Frozen), Ultima Sequencing, HD Spatial Gene Expression dataset analyzed using Space Ranger 3.1.2*. en. Feb. 2025.

[10] Gabriel D. Victora et al. "Identification of human germinal center light and dark zone cells and their relationship to human B-cell lymphomas". en. In: *Blood* 120.11 (Sept. 2012), pp. 2240–2248. DOI: 10.1182/blood-2012-03-415380.

[11] David N. Reshef et al. "Detecting Novel Associations in Large Data Sets". In: *Science* 334.6062 (Dec. 2011), pp. 1518–1524. DOI: 10.1126/science.1205438.

[12] Ramon Massoni-Badosa et al. "An atlas of cells in the human tonsil". English. In: *Immunity* 57.2 (Feb. 2024), 379–399.e18. DOI: 10.1016/j.immuni.2024.01.006.

[13] Eleftherios Zormpas et al. "Mapping the transcriptome: Realizing the full potential of spatial data analysis". en. In: *Cell* (Dec. 2023), S0092867423012199. DOI: 10.1016/j.cell.2023.11.003.

[14] Martin Emons et al. "Harnessing the potential of spatial statistics for spatial omics data with pasta". In: *Nucleic Acids Research* 53.17 (Sept. 2025), gkaf870. DOI: 10.1093/nar/gkaf870.

[15] Raphael Rubin et al. *Rubin's Pathology: Clinicopathologic Foundations of Medicine*. en. Lippincott Williams & Wilkins, 2008.

[16] Valentyna Zinchenko et al. "MorphoFeatures for unsupervised exploration of cell types, tissues, and organs in volume electron microscopy". In: *eLife* 12 (Feb. 2023). Ed. by Louis K Scheffer et al., e80918. DOI: 10.7554/eLife.80918.

[17] Phillip B. Nicol et al. *Identifying spatially variable genes by projecting to morphologically relevant curves*. en. Nov. 2024. DOI: 10.1101/2024.11.21.624653.

[18] Peter Diggle. "A Kernel Method for Smoothing Point Process Data". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 34.2 (1985), pp. 138–147. DOI: 10.2307/2347366.

[19] Adrian Baddeley and Rolf Turner. "spatstat: An R Package for Analyzing Spatial Point Patterns". en. In: *Journal of Statistical Software* 12 (Jan. 2005), pp. 1–42. DOI: 10.18637/jss.v012.i06.

[20] Samuel Gunz and Mark D. Robinson. *sosta: A package for the analysis of anatomical tissue structures in spatial omics data*. manual. 2025. DOI: 10.18129/B9.bioc.sosta.

[21] Edzer Pebesma. "Simple Features for R: Standardized Support for Spatial Vector Data". en. In: *The R Journal* 10.1 (2018), p. 439. DOI: 10.32614/RJ-2018-009.

[22] Michael Totty, Stephanie C. Hicks, and Boyi Guo. "SpotSweeper: spatially aware quality control for spatial transcriptomics". en. In: *Nature Methods* 22.7 (July 2025), pp. 1520–1530. DOI: 10.1038/s41592-025-02713-3.

[23] Davis J. McCarthy et al. "Scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R". In: *Bioinformatics (Oxford, England)* 33.8 (2017), pp. 1179–1186. DOI: 10.1093/bioinformatics/btw777.

[24] Aaron Tin Long Lun and Jayaram Kancherla. "Powering single-cell analyses in the browser with WebAssembly". en. In: *Journal of Open Source Software* 8.89 (Sept. 2023), p. 5603. DOI: 10.21105/joss.05603.

[25] Vipul Singhal et al. "BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis". en. In: *Nature Genetics* 56.3 (Mar. 2024), pp. 431–441. DOI: 10.1038/s41588-024-01664-3.

[26] Aaron T.L. Lun, Davis J. McCarthy, and John C. Marioni. "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor". en. In: *F1000Research* 5 (Oct. 2016), p. 2122. DOI: 10.12688/f1000research.9501.2.

[27] Dvir Aran et al. "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage". en. In: *Nature Immunology* 20.2 (Feb. 2019), pp. 163–172. DOI: 10.1038/s41590-018-0276-y.

[28] Davide Albanese et al. "A practical tool for maximal information coefficient analysis". In: *GigaScience* 7.4 (Apr. 2018), giy032. DOI: 10.1093/gigascience/giy032.

[29] Raivo Kolde. *pheatmap: Pretty Heatmaps*. June 2025.

[30] Zuguang Gu. "Complex heatmap visualization". en. In: *iMeta* 1.3 (2022), e43. DOI: 10.1002/imt2.43.

[31] Hadley Wickham. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York, 2016.

[32] Helena Lucia Crowell. *Supplementary Materiel to: Tracing colorectal malignancy transformation from cell to tissue scale*. June 2025. DOI: 10.5281/zenodo.15574384.

[33] Nicolas Damond et al. "A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry". en. In: *Cell Metabolism* 29.3 (Mar. 2019), 755–768.e5. DOI: 10.1016/j.cmet.2018.11.014.

[34] Helena L. Crowell et al. "The shaky foundations of simulating single-cell RNA sequencing data". In: *Genome Biology* 24.1 (Mar. 2023), p. 62. DOI: 10.1186/s13059-023-02904-1.

Metric	**Formula**	**Function / Meaning**
Area (A)	A	
Perimeter (P)	P	Length of the structure boundary.
Compactness (C)	$C = \frac{4\pi A}{P^2}$	Normalized measure of compactness (1 = perfect circle).
Major Axis (a)	a	Length of the long axis of the minimum bounding rectangle.
Minor Axis (b)	b	Length of the short axis of the minimum bounding rectangle.
Eccentricity (E)	$E = \frac{b}{a}$	Ratio of minor to major axis. 1 = circular, 0 = elongated.
Circularity / Roundness (R)	$R = \frac{4\pi A}{P_c^2}$	Measures roundness using the convex hull perimeter P_c. Higher = more circular.
Solidity (S)	$S = \frac{A}{A_c}$	Ratio of polygon area to convex hull area A_c. Measures convexity/density (1 = convex).
Curl	$\text{Curl} = 1 - \frac{a}{L_f}$,	Shape bending/curvature measure. L_f is the effective fibre length along the main axis.
Fibre Length (L_f)	$L_f = \frac{4A}{P - \sqrt{P^2 - 16A}}$	Characteristic length of the shape along the main axis.
Fibre Width (W_f)	$W_f = \frac{A}{L_f}$	Characteristic width of the shape perpendicular to the main axis.

Table S1: List of geometric features calculated by *sosta*.

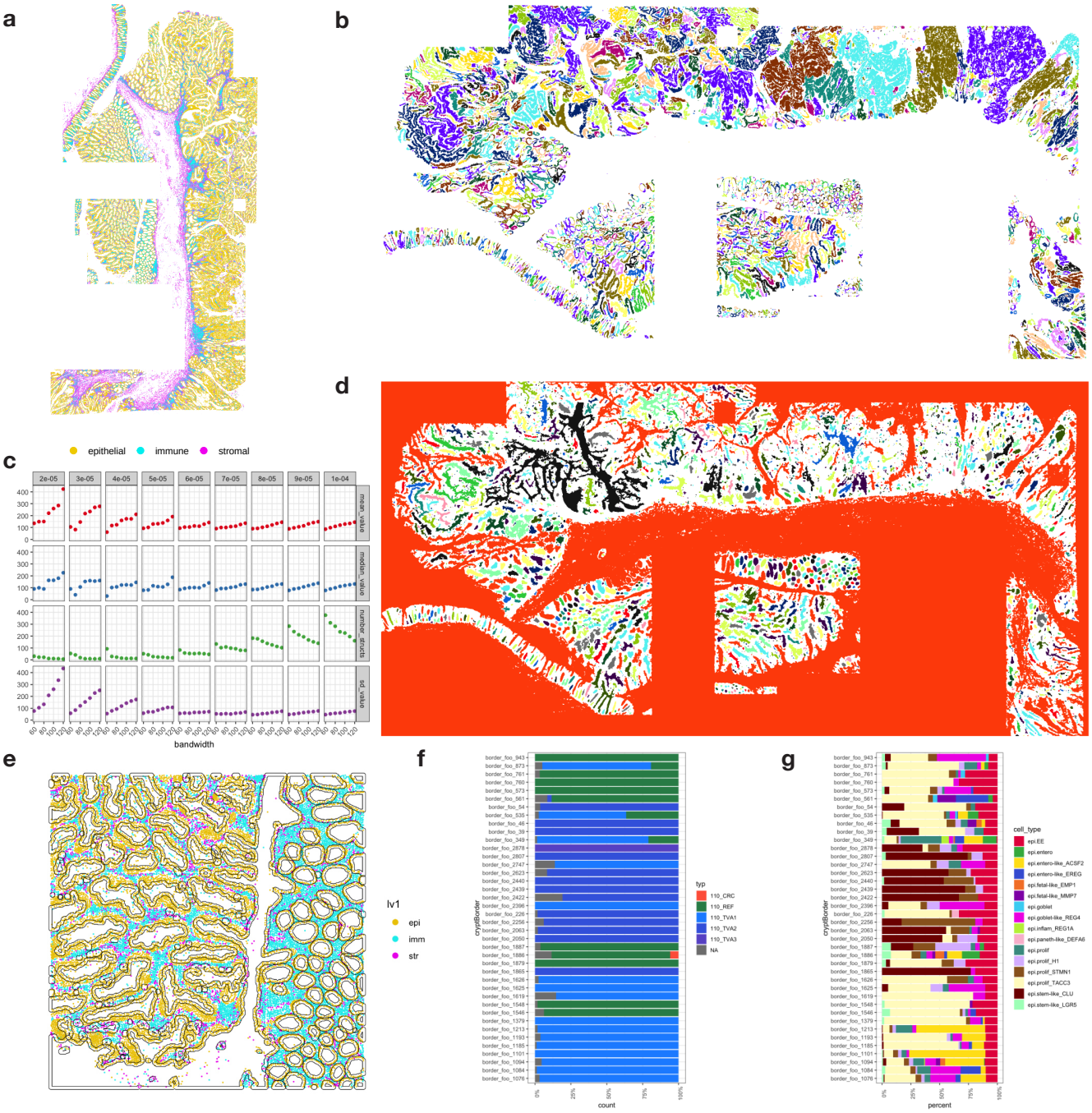


Figure S1: **Structure-based analysis of epithelial structures in the Crowell et al. dataset.** **a**) Slide 110 of the Crowell et al. dataset. Colours correspond to cell type annotations of the original publication. **b**) All reconstructed epithelial structures using the density-based reconstruction method of `sosta`. **c**) All reconstructed inter-epithelial (crypt) regions before filtering. **c**) Impact of parameters bandwidth (x) and threshold (facets) in reconstruction using `reconstructShapeDensitySPE` on the the mean, median and standard deviation of fibre width, and the number of recovered structures in a representative region (c.f. Fig. 2b). **e**) Expansion of crypts to assign cells to crypts before selection of transition crypts. A representative region was chosen for illustration. **f**) Composition of manually annotated regions in identified transition crypts and **g**) cell type composition of identified transition crypts. CRC = colorectal cancer, REF = reference (healthy tissue), TVA = tubulovillous adenoma.

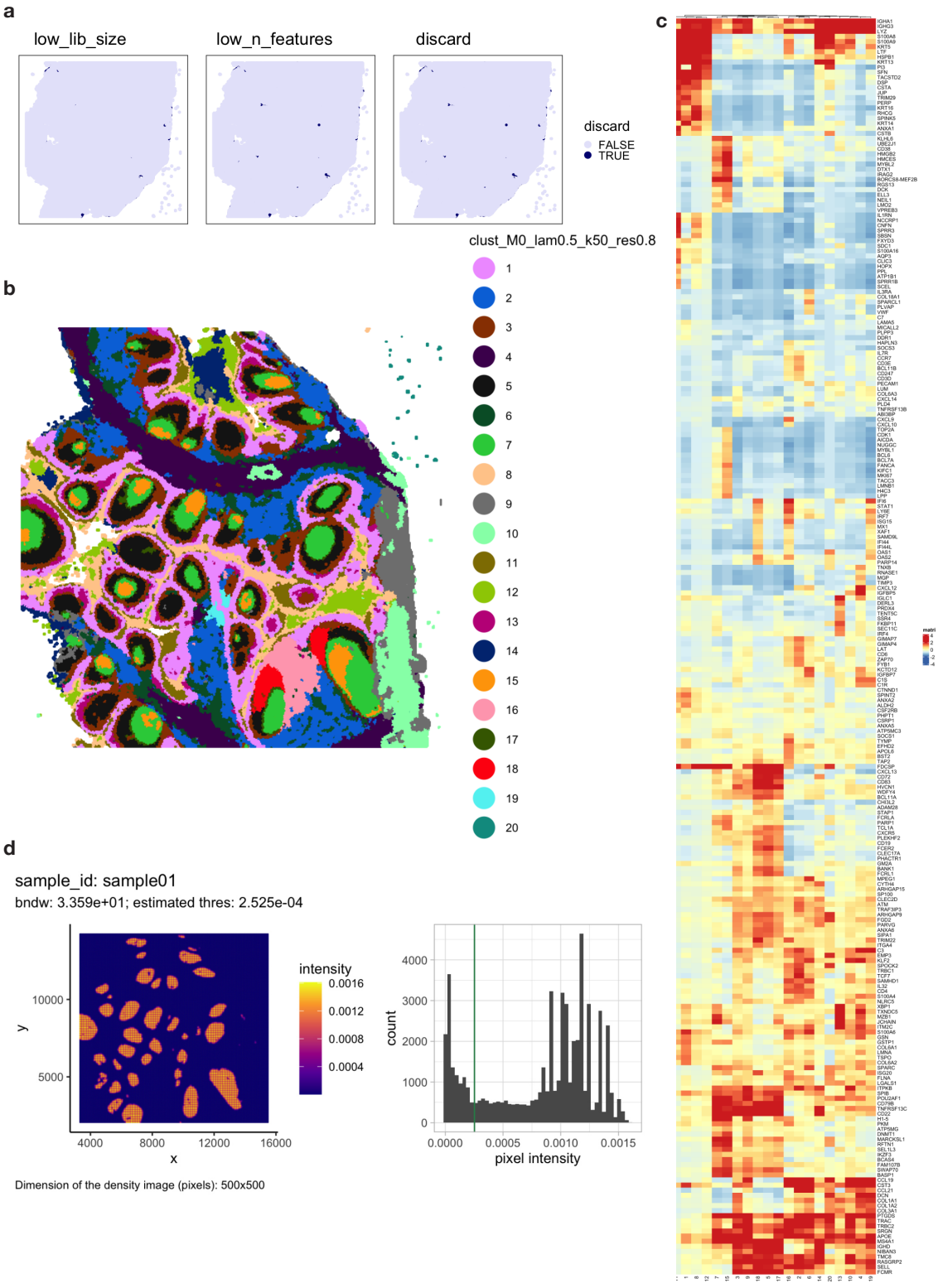


Figure S2: **Structure-based analysis of germinal centres in the human tonsil a**) Results of quality control using `SpotSweeper` [22] and different quality control (left to right: library size below threshold, number of features below threshold, combined flagged spots to discard). **b**) Results of clustering using `Banksy` [25]. **c**) Identified marker genes for the *Banksy* clusters using `findMarkers` of `scater` [23]. **d**) Density image showing density of GC cells with reconstruction parameters estimated by `sosta` [20].