Benchmarking Cell Segmentation Approaches for High-Resolution Spatial Transcriptomics

Master Thesis

Raphael De Gottardi

rdegottardi@ethz.ch

Department of Biosystems Science and Engineering
ETHZ
Statistical Bioinformatics Group
UZH
Genome Informatics Group
FGCZ

Supervisors:
Dr. Hubert Rehrauer, Dr. Paul Gueguen,
Prof. Dr. Mark D. Robinson

August 14, 2025

Abstract

Spatial Transcriptomics technologies have achieved subcellular resolution, with Visium HD providing transcript counts within 2μm$\times 2\mu$m capture squares. By aggregating these squares to model individual cells, this advancement opens the door for single-cell-like analysis with preserved spatial context. However, the standard 8μm$\times 8\mu$m binning approach introduces significant artifacts. This motivated the development of methods that use segmentation on H&E microscopy images to map Visium HD squares to cells. Such approaches are referred to as bin-to-cell methods and their evaluation constitutes the main focus of this work.

Three bin-to-cell methods were evaluated: Bin2Cell and ENACT, both relying on nucleus segmentation and subsequent circular expansion, and a custom experimental method using Xenium segmentation. Xenium segmentation uses specialized staining to segment whole cells in detail. We compared the methods against the 8μm$\times 8\mu$m binning baseline using a lung cancer dataset. Our evaluation framework integrated general QC metrics, segmentation-based evaluation, cell type annotation, and specialized coexpression-based scores.

Our analysis demonstrates that bin-to-cell methods improve results compared to simple binning of squares. Among bin-to-cell methods, Bin2Cell and ENACT showed similar performance. The variation in results depends rather on parameter choice, data characteristics and choice of downstream analysis method than on bin-to-cell method choice. The approach using Xenium segmentation showcases possible improvements when using more accurate segmentation but is not feasible to become a standard tool. Therefore, no universally superior method can be defined. Instead we reveal important insights on prior assumptions on the data, computational efficiency and practical applicability that should be considered when selecting appropriate methods for specific research questions.

This work establishes a robust evaluation framework for bin-to-cell method assessment and provides guidelines for method selection. We integrated the methods into the Functional Genomics Center Zurich ecosystem, facilitating broader access to these tools. The systematic comparison reveals that informed selection of methods based on computational constraints, tissue characteristics, and research requirements enables optimal results. Our findings establish benchmarking standards for the Spatial Transcriptomics community and lay the groundwork for systematic evaluation of future methods.

Contents

Chapter 1

Introduction

1.1 Spatially Resolved Transcriptomics

Recent advances in Spatially Resolved Transcriptomics (SRT), also known as Spatial Transcriptomics, have revolutionized our understanding of complex tissues by enabling high-resolution mapping of gene expression in their native spatial contexts. Named Method of the Year 2020 by Nature Methods [1], this approach enables researchers to tackle tissues with cellular heterogeneity and study tissue architecture. Spatial information in tissue matters when studying complex tissues, for example in cancer research. Cellular arrangement can be crucial in understanding tumor microenvironments and classifying cancer stages as well as tumor types [2] [3]. It has also proven to be important when studying immunological niches [4]. Another field with potential is development, SRT has been used to study mouse development stages [5] as well as organoids [6]. Such examples highlight that in biology, structure informs function and is therefore crucial in understanding underlying concepts. Current SRT methods can be categorized into two types: imaging-based and sequencing-based. A large number of technologies has been developed [7], the methods vary in resolution, transcriptome coverage, sensitivity, and price. In this thesis we focus on two technologies from 10X Genomics (short 10X) which are available at the Functional Genomics Center Zurich (FGCZ): Visium which is a sequencing-based method and Xenium, which is an imaging-based method.

1.1.1 SRT Methods

The current flagship technologies provided by 10X Genomics are Visium HD (released in February 2024 [8]) and Xenium Prime 5K (released in June 2024 [9]). In this thesis we focus particularly on these two technologies. We also show an overview of both SRT methods, summarizing the specifications provided by 10X Genomics [10].

Visium HD

Visium is a widely adopted platform that enables whole transcriptome profiling across diverse tissue types. Visium v1 and v2 have shown promising results in cancer research and immunology [2], [4]. Visium HD promises to usher in a new era of discovery in spatial biology [11]. The reason is the high resolution which is now achievable. Transcripts can be assigned precisely to 2μm by 2μm squares with no gaps in between. It can be used on fresh frozen (FF), formalin fixed paraffin embedded (FFPE) or fixed

frozen tissue. The workflow spans multiple days, it starts with QC on the samples, hematoxylin and eosin (H&E) or immunofluorescence (IF) staining and subsequent high resolution imaging. The probe solution is then added so the probe pairs can hybridize to the RNA and ligate, confirming a positive target hit. The tissue is then placed on a special slide which is coated with a continuous lawn of oligos over a capture area of 6.5mm x 6.5mm. The oligos are barcoded into 2μm x 2μm squares, where each oligo contains its unique UMI, and every oligo in the same square has the same barcode (see fig 1.1). The barcode can later be mapped to its spatial location on the high resolution H&E image. The transfer is facilitated by using a specialized tool (CytAssist) used to minimize processes such as diffusion and fluidic flow. The tissue is permeabilized and the probes can hybridize to the barcoded oligos. At this point the probes and barcodes are connected and can be amplified by PCR, a library is created which can be sequenced using a method of choice. Finally the reads are mapped to probes and barcodes using Space Ranger [12], a software provided by 10X Genomics which returns the barcode-counts-matrix as well as QC metrics of the data. The squares are by default aggregated to different bin sizes (2μm (trivial), 8μm, and 16μm).

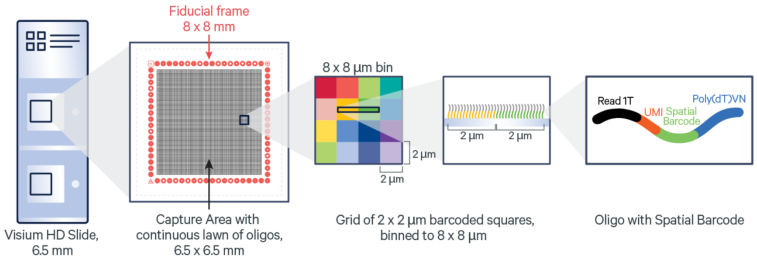

Figure 1.1: Visium HD slide architecture. Adapted from [11].

The most important benefits of Visium HD are that it has a transcriptome wide probe panel, that the probes work with heavily fragmented RNA (allows use of FFPE tissue) and that the squares are smaller than most cells. The technology however also has drawbacks, including the output being limited to counts, lack of SNP, isoform or exogenous gene detection (custom probes can be designed), species restriction to human and mouse and most importantly no straightforward square to cell mapping. The task of aggregating Visium HD squares to cells has been termed bin-to-cell and is the main focus of the thesis.

Xenium Prime 5K

The imaging-based solution provided by 10X is named Xenium and their latest technology (Prime 5K) allows for detection of 5000 genes. Xenium uses probes which will bind to the RNA of interest with both ends. This allows rolling circle amplification of the probe upon ligation. Fluorescent probes are then added, hybridize and are then imaged in cycles. The resulting series of images can then be decoded to obtain the transcript information at nanometer resolution. Additionally the workflow includes cell boundary (ATP1A1, E-Cadherin, CD45) , interior (DAPI, 18S sRNA) and nuclei (DAPI) staining, allowing to obtain cell segmentation at high resolution and confidence [13]. Another powerful benefit is that the workflow is non-destructive, allowing to perform H&E or IF imaging, but also Visium HD or multiplexed protein detection. In

this thesis we used this multimodal segmentation technique to combine the detailed Xenium segmentation with Visium HD data.

Summary Table

Category	Visium HD	Xenium Prime 5K
Plexity	Whole transcriptome (20,000 genes), customizable with spike-ins	Targeted 5000-gene panel, customizable (+100 genes or fully custom 480)
Methodology	Sequencing-based	Imaging-based
Sensitivity	Higher (∼1.5x) total transcript abundance	Higher (∼3.2x) per-gene sensitivity
Capture Area	6.5mm × 6.5mm	22.5mm × 10.5mm
Resolution	Sub-cellular resolution: $2\mu m \times 2\mu m$ capture squares, $8\mu m \times 8\mu m$ bins recommended as starting point for analysis	Nanometer-level resolution
Segmentation	Requires 3rd party tools. (or Space Ranger v4.0+)	Multimodal segmentation: Cell boundary, interior and nucleus staining
Multiomic Capability	Pre-run H&E or IF staining on same section	Non-destructive: Post-run H&E and IF staining, Visium HD possible on same section
Tissue Compatibility	FFPE, fresh frozen, fixed frozen	FFPE, fresh frozen
Data Analysis	Barcode-counts-matrix, preprocessed via Space Ranger, visualized in Loupe Browser	Cell-count-matrix, onboard analysis, visualized in Xenium Explorer
Use Case Recommendation	Broad tissue discovery, atlas construction, gene expression screening	Precision analysis of known genes, zoom-in on regions/cells of interest

Table 1.1: Comparison of Visium HD and Xenium Prime 5K Spatial Transcriptomics Platforms.

1.1.2 Bin-to-Cell Methods

With Visium HD reaching subcellular resolution, the door was opened to use Spatial Transcriptomics data for single cells. However this requires accurate mapping of the 2μm x 2μm squares to cells, which is a challenging task. 10X Genomics recommends to start the analysis with the squares aggregated to 8μm x 8μm bins (8x8) which is close to typical cell sizes, ensuring to contain more transcripts and making it computationally less demanding [11]. However these bins will likely not coincide with the cell shapes, which leads to partial coverage of cells, coverage of specific cell regions (e.g only boundary but not nucleus), coverage of multiple cells and bias towards cell types

matching the bin size. All of those scenarios can impair the power of the analysis and influence the results. Polanski et al. highlight four limitations to the 8x8 bins [14]:

- Large cells will be fragmented by multiple bins while small or non-square cells would have suboptimal overlap with the square 8μm bins.
- Downstream analysis will be heavily biased by the grid properties. The typical distances between cells will be units of 8μm (e.g. 8, 16, 24 etc). This will diminish the ability to infer realistic cell associations, or cell-cell communication
- Boundaries, fine tissue structures (e.g. vessels) or tissue edges will be affected as these will deviate dramatically from a square grid.
- 8μm bins are likely to capture contents of multiple cells, introducing intrinsic doublets drastically hampering the ability to explore the true molecular content of a single cell in an unbiased manner.

Segmentation-Based

To provide a better alternative, bin-to-cell methods have been proposed. The first one was Bin2Cell [15] by Polanski et al. from the Teichmann Lab. They proposed to use nucleus segmentation algorithms on the high resolution microscopy image and then transfer the labels to the underlying 2μm x 2μm squares (see fig. 1.2). The closest neighbors can then be assigned the same label to estimate a circular shape of the cell. More methods and pipelines have emerged, including ENACT [16] and most lately Space Ranger v4.0 [17] by 10X Genomics which includes a similar pipeline as Bin2Cell in its latest release. Such methods enable to study complex tissues which would else be suffering from the artifacts mentioned before, this especially includes heterogeneous tissues, for example tumors. Furthermore, this brings the technology one step closer to having single-cell-like data with additional spatial information.

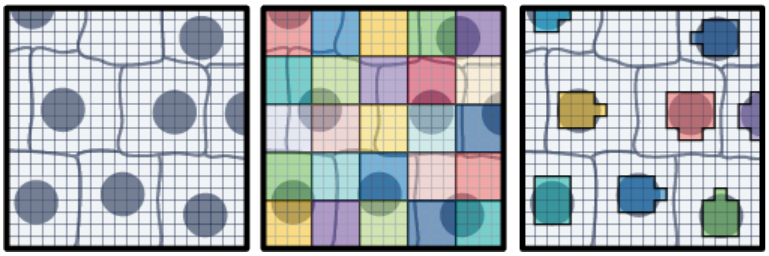

Figure 1.2: Two approaches for binning 2μm x 2μm squares: 8μm x 8μm (middle) or using nucleus segmentation on the H&E image (right). Adapted from [18].

The different methods compared and evaluated in the thesis are described in detail in the methods section. Segmentation-based bin-to-cell methods using the described approach have demonstrated to be powerful in tackling challenging tissues and have come close to the quality of Single Cell Transcriptomics (scRNA-seq) data. However they still suffer from some limitations imposed by the nature of the approach: As the technology requires tissue slices of 3μm to 20μm, part of the cells will be cut open or overlapping each other (or both). For a sliced cell, this means that even in the best case,

part of the transcripts will be missing, in the worst case it will have lost its nucleus and is thus not detected by the segmentation. For overlapping cells, their transcripts will mix and inevitably end up on the same square, this is especially challenging if the cells are of different cell types. The circular expansion of the nuclei imposes assumptions the user needs to be aware of: The cell bodies are assumed to be evenly distributed around the nucleus, and the cells are assumed to be of the same size (one single expansion parameter). This also penalizes cells with irregular or amorphous morphologies e.g. neurons or fibroblasts. These limitations can impair downstream analysis like cell annotation or detection of rare cell types. Using more accurate cell segmentation approaches is therefore expected to be advantageous. However, it remains unclear how critical detailed segmentation is for the downstream results.

Probabilistic Segmentation

An alternative approach to assigning transcripts captured by SRT to cells is probabilistic segmentation. Probabilistic segmentation methods aim to overcome the limitations of traditional image-based segmentation by modeling the spatial distribution of transcripts. These methods are motivated by the need to reduce transcript misassignment, which can confound Spatial Transcriptomics data and lead to false biological interpretations. Baysor [19] exemplifies probabilistic segmentation by optimizing transcript assignment through a mixture model framework, avoiding reliance on image data or supervised training. Proseg, an approach recently proposed by D. Jones et al. [20], uses an unsupervised probabilistic model to infer morphologically plausible cell boundaries and reposition misplaced transcripts.

Deconvolution Methods

At this point it is worth to mention the efforts on deconvolution methods [21], which aim to estimate the cell-type contributions within each Spatial Transcriptomics spot or bin. Although this type of method emerged from the technologies providing resolutions above typical cell sizes, the field is rapidly evolving and has some promising approaches in improving SRT data analysis. Such methods include reference-based approaches which use annotated scRNA-seq datasets to infer known cell-type proportions or reference-free approaches. The latter use matrix factorization or probabilistic models to infer cell types directly from spatial data without prior annotations. Deconvolution would not be needed if true single cell resolutions was available for SRT. For current technologies, deconvolution could help to overcome the limitations of segmentation-based bin-to-cell methods.

1.1.3 Image Segmentation Methods

The above described, current image-based bin-to-cell methods rely on cell segmentation on H&E images. Although it is not clear how important an accurate cell segmentation is, it is safe to assume that a more detailed segmentation is favorable. Cell segmentation is a popular task in biology and has been recently revolutionized by deep learning models. State-of-the-art segmentation tools now employ deep learning architectures, with methods like CellPose [22], StarDist [23], and Mesmer [24] demonstrating superior

performance in identifying cell boundaries from H&E morphology compared to classical watershed-based approaches. However, significant challenges remain in standardizing segmentation performance across different tissue types, staining protocols, and imaging conditions, making robust and generalizable segmentation a possible bottleneck in the analysis pipeline of SRT data. As cell boundaries are hard to resolve with conventional staining methods, many approaches rely on nucleus segmentation and subsequent circular expansion.

StarDist

StarDist is a deep learning-based segmentation method specifically designed for detecting star-convex objects, making it particularly well-suited for cell nucleus segmentation in microscopy images [23]. The method operates on the principle that most cell nuclei can be approximated as star-convex shapes, where any line from the object's center to its boundary remains within the object. StarDist employs a U-Net-like convolutional neural network architecture that simultaneously predicts object probabilities and star-convex polygon distances from potential object centers, enabling direct reconstruction of object boundaries. The models are open and they provide pretrained models for 2D images in H&E and IF. The method is computationally efficient and demonstrates robust generalization across different imaging modalities, tissues and cell types. The restriction to nucleus segmentation is an issue as the objective is most often the whole cell. Also the restriction to star-convex shapes has the drawback that it will perform poorly for highly irregular shapes. However, the star-convex assumption is useful when circularly expanding the segmentation in downstream analysis.

The strong performances across different tissue types and computational efficiency have made StarDist the standard choice for all the bin-to-cell methods published to this point. One could use a custom segmentation, however the only pre-included option in Bin2Cell, ENACT and Space Ranger v4.0 is StarDist. Therefore we did not compare between different nucleus segmentations, assuming that the nucleus segmentation by StarDist is satisfying. We did however compare the nucleus segmentation using StarDist with subsequent circular expansion to the segmentation obtained by the Xenium dataset. As mentioned in the introduction to Xenium, it uses cell boundary (ATP1A1, E-Cadherin, CD45), interior (DAPI + 18S sRNA) and nuclei (DAPI) staining, combined with a high resolution imaging system to provide detailed segmentation results.

1.1.4 Cell Type Annotation Methods

A common task when analyzing scRNA-seq data is cell type assignment. Annotating the cells by type allows to explore single cell-type features as well as the cell-type proportions in a tissue of interest. This becomes even more powerful with Spatial Transcriptomics outputs because they add morphology images and exact locations of the cells. This can be interesting when validating predictions or exploring neighborhood patterns. Traditionally, cell type assignment is done manually. Therefore dimensionality reduction techniques are used (PCA, UMAP) followed by clustering (Leiden, Louvain). Known marker genes are then used to label the clusters manually. This process is time consuming, requires biological expertise and is prone to be subjective. To solve

these limitations, automated cell type assignment tools have emerged. These are often based on modern machine learning methods and trained on large scale reference data sets, two examples are CellTypist [25] and Azimuth [26]. CellTypist employs logistic regression model to predict cell types based on gene expression signatures, while Azimuth uses reference mapping approaches that project query cells onto reference datasets. These tools provide rapid, standardized, and reproducible cell type annotations. They leverage large reference databases (e.g. Human Cell Atlas [27]) and can handle large-scale datasets efficiently. On the other hand, their quality also depends on the reference datasets, which can lead to potential batch effects between query and reference data. Also this risks reduced accuracy when dealing with rare cell types or tissue-specific subtypes not well-represented in training data. For example mapping diseased states using a healthy reference. Spatial context is ignored in current tools, missing opportunities to leverage neighborhood information for improved classification. We explored how the classical but also the modern cell type annotation methods perform on Visium HD data after bin-to-cell mapping and compare them to scRNA-seq datasets.

1.1.5 Benchmarking

The bin-to-cell pipelines are being developed at a very fast rate (similar to the technology itself). The publications include some metrics on the results, however there is still need for more comprehensive and independent evaluation of the performance of such methods. This is highly challenging because every result is tied to the choice of parameters, the tissue quality and type as well as data quality. Also the complexity of the tools grows with the complexity of the technologies. Therefore it is hard for a user to understand how an output is generated from sample to final result. This is also because of the number of assumptions involved and the large amount of parameters that can be set. While these tools are opening the doors to many options for users, it can be hard to find the best fitting tool and choose the parameters appropriately. This makes finding the best technology and analysis method for a given research project increasingly difficult.

1.2 Objective

FGCZ provides access to both Xenium Prime 5K and Visium HD technologies, enabling SRT data acquisition through complementary imaging-based and sequencing-based platforms. The required machines, enough computing power and expertise in analysis of omics data is available. With the fast release of new methods, chemistries and tools, it becomes increasingly hard to maintain the overview of the best practices for each technology. Also in general, there is a lack of (independent) benchmarking tools and reports to assess the best performing pipeline. Therefore a goal of this thesis was to implement bin-to-cell methods within the FGCZ framework, make them available to users and staff and ensure informed, efficient and reproducible pipelines. This includes finding best practices, stating assumptions and finding caveats. The methods were then evaluated using common benchmarking metrics, as well as comparison to an experimental method referred to as 'Xseg', which uses the Xenium segmentation to

perform bin-to-cell assignment. This required manual alignment of the Xenium segmentation to the Visium HD data. Using the Xenium segmentation aimed to evaluate the importance of accurate segmentation when using bin-to-cell methods. The evaluation included common QC metrics, segmentation-based metrics, evaluation of downstream analysis as well as specialized metrics to assess performance of such tools (Spurious relative coexpression analysis and MECR score). The evaluation sheds light on the power of such tools. The results were evaluated and used for downstream analysis to evaluate how close they were to single cell resolution SRT.

With the drastic improvement of squares reaching subcellular size, SRT technologies can aim at resolution at the scale of single cells. The capabilities of spatial data have shifted from analyzing tissue regions to cells. ScRNA-seq data has been available for a while and is well established in the field. Analysis tools for scRNA-seq are well documented, accessible and known by experts. Therefore a standing goal for SRT is to come as close to single cell resolution as possible. Single cell resolution SRT would enable the use of scRNA-seq analysis pipelines, having additional imaging and coordinate information. A particular focus is on cell type annotation which we consider the most important downstream task. In this thesis we aimed to evaluate how close current methods are to achieving this goal.

Chapter 2

Methods

2.1 Datasets

10X Genomics has a large number of publicly available datasets using their instruments and technology. The Visium HD datasets include the raw data as well as the outputs after being preprocessed with Space Ranger and are therefore easy to use for downstream pipelines.

2.1.1 Visium HD Post Xenium Application: Lung Cancer Dataset

As part of the ”Post-Xenium In Situ Applications: Immunofluorescence, H&E, Visium v2, and Visium HD” technical note [28], 10X published multiple datasets on a Lung Cancer sample [29]. The released data includes two Xenium datasets (v1 and 5K panel) as well as two corresponding Visium (v2 and HD) datasets using the same tissue slice. Additionally two more adjacent tissue slices were analyzed using Visium HD only as a control. This dataset was published by 10X to demonstrate the capability of post-Xenium applications, but shows some interesting properties that could be used for benchmarking purposes. We especially focused on the Xenium segmentation. For the dataset using the 5K panel, they used all three staining methods available (boundary, internal and nucleus) to obtain a detailed segmentation of the cells. 97.8% of the cells were segmented using boundary or interior staining. It is safe to assume that this segmentation is more accurate than nucleus segmentation and circular expansion, which allows us to use it as a "ground truth". In the following we refer to the datasets captured with Visium HD and Xenium 5K on the same tissue slice as the 'Lung Cancer' dataset. The capture areas only partially overlap (see fig. 2.1) in a region containing roughly 180k cells. The coordinates of the bins, images and segmentation can be transformed to match using affine transformations, partially requiring manual alignment. Janesick et al. performed a similar alignment using previous Xenium and Visium technologies [30]. We combined their helper functions with 10X's Xenium explorer alignment tool to align the datasets.

We additionally explored Visium HD datasets on Mouse Embryo, Mouse Brain, Human Colorectal Cancer and Human Tonsil all by 10X Genomics and publicly available.

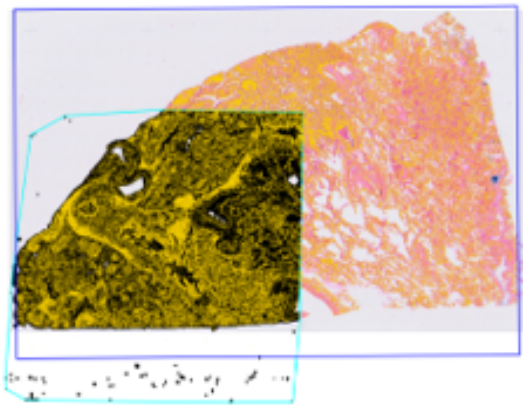

Figure 2.1: Overlay of the Visium HD (light blue frame) post Xenium 5K (dark blue frame) dataset on the Lung Cancer tissue. H&E image in the background, Visium HD bins over tissue in black and Xenium cell segmentation in yellow.

2.1.2 Rhesus Macaque Dataset

Space Ranger v4 was released during the writing of this thesis. We were not able to install and run the pipeline on the Lung Cancer dataset to compare all the methods directly in time. However 10X released a Rhesus Macaque kidney dataset [31] early which was preprocessed using Space Ranger with the segmentation feature. This dataset captures spatial gene expression from fresh frozen kidney tissue of a male Rhesus Macaque using the Visium HD platform. Tissue sections were cryosectioned at 10μm, stained with H&E, and imaged at 20X magnification. We obtained the most important general metrics for this dataset and present a summary of the results as well as some preliminary remarks. Additional metrics using this dataset can be found in the appendix.

2.1.3 ScRNA-seq Reference Dataset

Having Xenium 5K and Visium HD data from the same tissue slice helps comparing results and pipelines. The current state of the art when studying tissues at detailed transcriptomic level is scRNA-seq. A dataset on the Lung Cancer tissue sample is unfortunately not available. Therefore we used a publicly available scRNA-seq dataset from the Curated Cancer Cell Atlas maintained by the Weizmann Institute of Science [32]. The database contains a total of 77 single-cell RNA sequencing datasets from various human cancers including lung. We decided to use a lung adenocarcinoma dataset published by Bischoff et al. [33]. The data covers 120'961 cells across 20 lung adenocarcinoma samples. It was obtained using 10X Chromium single cell and the cells are already labeled by cell type. We used this dataset to compare the results of single cell tools on SRT data and to find common cell-type proportions in similar tissue. In the following, it will be referred to as 'scRNA-seq reference dataset'.

2.2 Considered Pipelines

Bin-to-cell methods were outlined in the introduction. In the following, we summarize the most important pipeline steps for each of the methods considered in the benchmarking. This includes the binning of squares to 8x8 bins as a baseline, both bin-to-cell methods currently available (Bin2Cell and ENACT) as well as an alternative method using the segmentation from the corresponding Xenium assay (named Xseg). We motivate the choice of parameters and document how the methods were implemented to ensure reproducibility and facilitated use. We also motivate the alternative method, using the Xenium segmentation on Visium HD data, and describe how it was realized. A set of visualizations was generated for each pipeline on the same crop region of the Lung Cancer sample.

2.2.1 8x8 Bins

10X Genomics currently recommends to start the analysis with the squares aggregated to 8μm x 8μm bins (8x8) which is close to typical cell sizes. De Oliveira et al. have demonstrated that this method can be used to understand cellular interactions in tumor microenvironments [34]. By default, Space Ranger outputs the barcode-count-matrix binned at 2μm, 8μm, and 16μm bin sizes. The bins are flagged if they overlap tissue regions based on the H&E image. Only flagged bins are used for downstream analysis. Other bin sizes (multiples of 2μm) can also be generated manually with low effort. We consider the 8x8 bins as a baseline method and study the output quality improvements of using more sophisticated bin-to-cell methods. For some metrics we also show the results using 16μm bins for reference as the cell sizes for the Lung Cancer dataset are mostly in between.

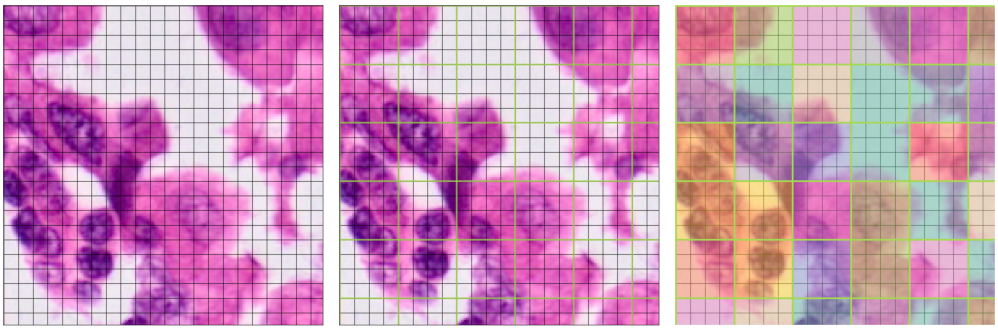

Figure 2.2: Visualization of the $8\mu m \times 8\mu m$ binning process over a small tissue region crop of the Lung Cancer dataset.

2.2.2 Bin2Cell

Bin2Cell [15] (short B2C) is one of the first tools to use segmentation on tissue images to enable smart binning of Visium HD squares to cells. Bin2Cell uses StarDist segmentation on the morphology image as well as on the total gene expression image to find cell boundaries. The labeled squares are then aggregated to cells.

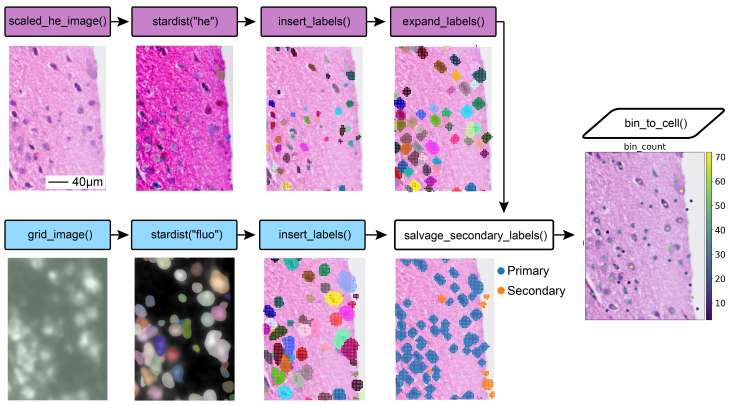

Figure 2.3: Bin2Cell steps, split in segmentation of H&E image (top) and segmentation on GEX (bottom). Adapted from [15].

The input data corresponds to the output from Space Ranger, which is commonly used to preprocess the raw data after sequencing. Bin2Cell requires the binned outputs for the 2μm x 2μm bins, the high resolution morphology H&E image as well as the CytAssist images (lower resolution, used for plotting purpose only). A user can set the common StarDist parameters (e.g.'prob_thresh', 'nms_thresh') as well as some parameters specific to Bin2Cell (e.g.'mpp', 'buffer','max_bin_distance') affecting the segmentation and downstream results. We summarize the most important parameters and the effect of varying them in the appendix.

The pipeline starts with a destriping step, a row and column based normalization of the bin counts to counteract manufacturing artifacts. Normalization will convert the transcript counts to floating point numbers and slightly change the total transcript count. Then, nucleus segmentation using StarDist is performed on the high resolution morphology image. The nucleus segmentation is then projected onto the underlying bins. To obtain cells, the labels are then transferred to unassigned neighboring bins in all directions up to 'max_bin_distance' away from the nucleus. Squares that are equidistant from 2 nuclei are assigned to the closest nucleus square by gene expression proximity in PCA space. Then, a different model of StarDist is used to segment cells from the total gene expression image (GEX). In the total gene expression image, one pixel corresponds to one 2μm x 2μm square and the sum of transcript counts is used as pixel value to obtain a grayscale image. Cells are expected to appear as clouds, similar to the signal obtained in immunofluorescence microscopy images. Therefore cell shapes can be segmented directly without need for circular expansion. Note that the segmentation results can be directly mapped to the bins as they correspond to the pixels. This can enable finding anucleated cells or cells which lost their nucleus in the tissue preparation process. The GEX mask is then overlapped with the H&E mask and each GEX label which does not overlap any H&E label is kept, the rest is ignored. Finally, all bins with the same label are aggregated to one cell and saved to Anndata format, a data structure commonly used for scRNA-seq experiments, including the coordinates of the nucleus center.

For the benchmarking, we decided to keep the 'prob_thresh' and 'nms_thresh' at their Bin2Cell defaults (0.01, 0.05). The resolution of the images, defined by the 'mpp'

parameter was also kept to the default 0.5μm per pixel. We recommend to visually inspect the outputs and make sure they correspond to the expected cell morphologies.

2.2.3 ENACT

ENACT was developed after Bin2Cell by Kamel et al. [16]. The pipeline aims to address shortcomings in Bin2Cell. Bin2Cell always assigns all transcript from one square to the same cell which can be an issue in tissues where cells are tightly packed. Providing four different transcript assignment methods, ENACT offers the opportunity to assign individual transcripts from the same square to different cells. The tool consists of a bin-to-cell pipeline and subsequent cell type assignment of the segmented cells. Like Bin2Cell, ENACT requires the Space Ranger outputs as input. It also starts by first segmenting nuclei on the full-resolution H&E image using StarDist. Instead of mapping the nucleus segmentation to the squares, the segmentation is first expanded up to a parameter ('expand_by_nbins'). As a consequence, many squares will overlap multiple cells. Both squares and predicted cell outlines are then represented as Shapely polygons to perform precise geometric operations for the squares overlapped by multiple cells (red in fig. 2.4).

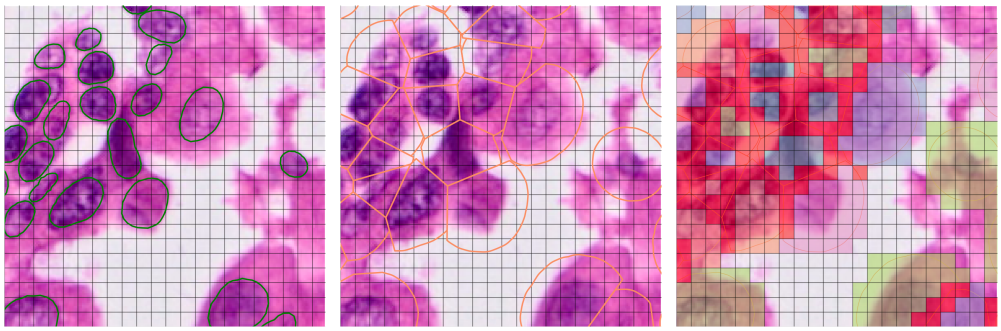

Figure 2.4: Visualization of the ENACT steps. Nucleus segmentation (left), expansion by a fixed distance (middle) and label transfer to squares (right). The red squares overlap multiple cell segmentations and will therefore assign their transcripts based on one of the four assignment methods.

ENACT implements four assignment methods (one naive and three weight-based) to redistribute transcripts from overlapping bins according to their true spatial intersections. We briefly summarize the different methods in table format (see tab. 2.1). A detailed explanation of the methods can be found in the supplementary material of the ENACT paper [16].

We explored all the transcript assignment methods, especially for the specialized scores (see section "Benchmarking Pipeline"). For the main benchmarking we decided to use the Weighted-by-Area results as they are defined as default and have reached the best results in the benchmarking released in the paper. ENACT further integrates support for cell-type annotation tools, including CellAssign, CellTypist and Sargent. These tools and their key properties are summarized in an overview of cell type assignment methods in the appendix (tab. 2.3). CellAssign uses lists of marker genes provided by the user for the expected cell-type populations. This makes it labor intensive and requires biological expertise. Sargent is not integrated into the ENACT pipeline but

could be soon or could be manually integrated. This is why we decided to focus on CellTypist as cell annotation tool. CellTypist is a easy to use and well known automated cell type annotation approach using pre-trained models that deliver accurate classifications without requiring manual curation or extensive parameter tuning. The ENACT outputs contain the segmented polygon coordinates (nuclei and cells), an image of the segmentation boundaries aligned to the H&E image, the cell-count-matrix including cell positions and annotation as well as a TissUUmaps [35] project folder for interactive visualization of the data. The cell-count-matrix can be loaded into Seurat [36] (R) or AnnData [37] (Python) format for downstream spatial analyses with common tools (eg. Scanpy [38], Squidpy [39]).

Method	**Summary**
Naive	Assigns each square only if it overlaps exactly one cell, squares intersecting multiple cells are dropped. Benefits: maximizes precision by avoiding ambiguous assignments, at the cost of discarding squares (sensitivity).
Weighted-by-Area	Allocates transcripts in each square to cells in proportion to the square–cell intersection area (intersection area / square area). This method achieved the best results in the paper benchmarking and is therefore used as default.
Weighted-by-Transcript	Starts by doing the naive assignment to obtain a transcript profile for each cell. Every transcript count in a shared square is then split between the cells based on their transcript profile.
Weighted-by-Cluster	Starts by doing the naive assignment and then grouping cells into clusters via K-means on the naive counts. For each cluster, average gene expression profiles are computed and used to weight transcripts in shared squares: transcripts are apportioned to overlapping cells according to the cluster-level expression of each gene.

Table 2.1: Summary of ENACT's four transcript assignment methods. The selected method used to assign the transcripts in squares overlapped by multiple cells (colored red in fig. 2.4).

2.2.4 Xenium Segmentation

Besides localization of transcripts at nanometer resolution in 3D, Xenium can perform cell segmentation on a highly detailed level. Xenium segmentation uses specialized protein and RNA markers that ensure specificity and sensitivity of the segmentation, as well as robustness to different cell types. Therefore we can safely assume that Xenium segmentation is more precise than nucleus segmentation and subsequent circular expansion. Xenium is nondestructive, which enables capture of Visium HD data after having obtained the detailed segmentation on the same tissue slice. This makes the Lung Cancer dataset very valuable when estimating the importance of accurate segmentation. One can find most of the scale factors between images as well as crops and shifts from the data provided by 10X. However the transfer of Xenium segmentation onto the Visium images is not a standard procedure. Janesick et al. [30] performed a similar alignment when comparing ScFFPE-seq, Visium CytAssist (non HD) and Xenium data on serial FFPE sections. We adapted their protocol including performing manual alignment for the Lung Cancer dataset in order to use the Xenium segmentation

for Visium HD data.

The segmentation outputs are per default saved as a pixel mask. The polygon representation is also generated per default to enable efficient visualization. However it is less precise than the mask and not recommended to be used for segmentation [40]. We decided to nevertheless use the polygon segmentation for downstream analysis. This is because we would need to convert the mask to different coordinates and another resolution which is non-trivial for distinct labels and would in every case also add uncertainty to the cell boundaries. The second reason is that the ENACT pipeline uses polygon representations anyway. The full pipeline for the conversion of the Xenium segmentation to Visium coordinates and data format compatible for ENACT is documented in the GitHub repository.

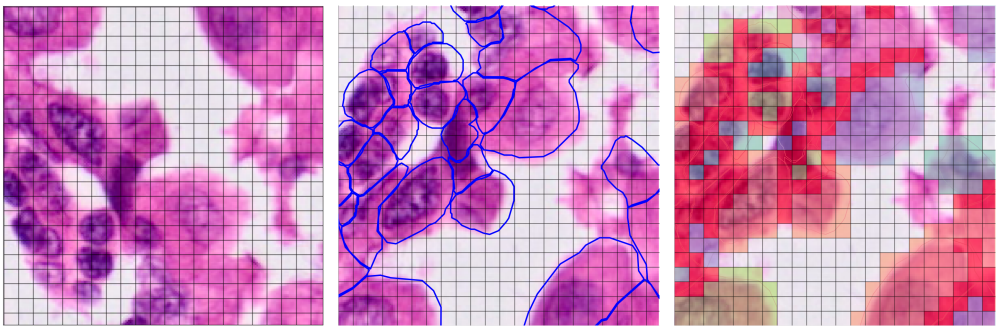

Figure 2.5: Visualization of the Xenium segmentation method steps. The Xenium segmentation of cells is shown in the middle and the label transfer (using ENACT) to squares on the right. The coloring scheme is the same as for ENACT. Note that there is no cell-to-cell mapping between the segmentations of ENACT and Xenium.

Using the polygon representation allowed the embedding of the segmentation into the ENACT pipeline. The nucleus segmentation and expansion steps are skipped and the subsequent bin-to-cell assignment using the Xenium segmentation is done using the ENACT transcript assignment methods (see tab 2.1). The segmentation can still be expanded by setting a positive expansion parameter. This was useful to assess if such an expansion can help to capture diffused transcripts and therefore increase sensitivity.

2.2.5 Summary Table

We summarize the most important information of the methods described in a summary table (2.2).

Aspect	8×8 bins	Bin2Cell	ENACT	Xenium Segmentation
Purpose	Increase signal per bin	Segmentation and bin-to-cell assignment	Segmentation, bin-to-cell assignment using different methods and cell type assignment	Custom methods to assess importance of accurate segmentation
Input data	Cell-count matrix	H&E image and Cell-count matrix	H&E image and Cell-count matrix	H&E image, Cell-count matrix and coordinate-transformed Xenium segmentation
Image used	None	H&E and GEX image	H&E	Xenium
Segmentation algorithm	None	StarDist (H&E and Fluoro)	StarDist (H&E)	10X (using boundary, interior and nucleus staining)
Bin-to-cell assignment	Naive 16 to 1	Nucleus to square, then square to neighbors	First nucleus expansion then 1 of 4 methods	Used ENACT

Table 2.2: Summary of the bin-to-cell methods considered for the benchmarking.

2.2.6 Space Ranger v4

10X Genomics not only provides the technology for the data generation but also software for basic analysis. We introduced Space Ranger as a preprocessing software. It requires the full-resolution microscopy image, a reference, slide metadata and the FASTQ files from the sequencing. The main output consists of the barcode-count matrix for all squares (2μm) and aggregated bins (8μm and 16μm). Space Ranger v4 was released in July 2025, including their own bin-to-cell method. The implementation relies again on StarDist for nucleus segmentation on the H&E image. 10X used a custom dataset including 170 H&E tissue images covering many preservation methods and tissues to train the custom StarDist model. After segmentation, the barcodes (corresponding to 2μm squares) are assigned to the closest nucleus up until a maximum distance parameter. We did not include Space Ranger in the main benchmarking analysis due to the late release. However, we used the published Rhesus Macaque kidney dataset to explore the results and summarize some preliminary findings. More information, visualizations and the summary table (2.2) including Space Ranger v4 can be found in the appendix and in the GitHub repository.

2.3 Benchmarking Pipeline

To compare the tools available, we ran the bin-to-cell methods described above on the available datasets. All metrics are based on the pipeline outputs, namely the cell-count-matrix and optionally the segmentation outputs. Mainly the Lung Cancer dataset was used because it is the only dataset where we could use the Xenium segmentation and is therefore most valuable when comparing methods. Other datasets were also used for specific comparisons.

2.3.1 Requirements on Computational Resources

While we did not perform a rigorous performance evaluation of the observed models, we summarize some remarks and general insights on the required resources to run the bin-to-cell pipelines. This includes run time and maximal RAM usage as well as remarks on GPU usage. For the evaluation we always ran the pipeline using all genes available (no highly variable gene (hvg) filtering).

2.3.2 General Metrics

We started by checking some of the common QC metrics and how they differ across methods. This includes cell count, total transcript count, unassigned transcripts and number of detected genes as dataset metrics. The total transcript counts are a measure for sensitivity. The fraction of unassigned transcripts has been associated to loss of information, therefore it is preferable to keep this value low [41]. We also looked at per cell distribution metrics which include transcripts per cell (tpc) and number of genes per cell. We also used the coordinates available to show the density and location of the detected cells. Finally we also show the cell representations in UMAP space (first two dimensions) after PCA reduction to the first 50 principle components and colored by Leiden clusters. We run Leiden clustering at resolutions of 0.5, 0.7, 1.0, 1.2 and 1.5 and then choose the one with the number of clusters closest to 12 for best visualization. These general metrics can be computed using an app (in SUSHI [42]) we created which returns the metrics for any cell-count-matrix in an easy to understand report.

2.3.3 Segmentation-Based Metrics

We can compare two types of segmentation: StarDist and Xenium. Bin2Cell and ENACT both use the StarDist public models. So under choice of same parameters, the only difference in terms of segmentation is the additional GEX segmentation in Bin2Cell. Space Ranger v4 will also use StarDist although they trained their own model on a large dataset. Therefore we consider the segmentation outputs from ENACT as sufficient to represent Bin2Cell and Space Ranger v4 segmentations as well. A comparison of the Space Ranger segmentation to ENACT showing this similarity can be found in the appendix. For a detailed comparison of segmentation outputs of single cells, one needs a mapping between segmentations. As not all the cells are captured, some are split in two, some are shifted slightly and they don't share an id if corresponding to each other, it is a hard problem to compare segmentations at a single cell level over a full dataset. Instead we provide general segmentation metrics on the full dataset and

chose to focus on a crop to show differences in segmentation of single cells. For morphological comparison of the segmentation outputs we show an overlay of mass centered polygons from a 1000×1000 pixel crop region of the Lung Cancer dataset sample. This allows to highlight shape characteristics. We also show both segmentations overlaid with the morphology H&E image in a crop region. This allows to assess how well the segmentations match the visible cell shapes. The cell area distribution is also shown. We also compare different expansion parameters, from nuclei only up to expansion by 3 bin sizes.

2.3.4 Annotation-Based Metrics

We presented common cell type annotation methods in the introduction (Manual annotation, Azimuth) and introduced the methods supported by ENACT (CellAssign, CellTypist, Sargent). For the benchmarking, we decided to focus on the comparison of CellTypist to the manual annotation and also considered the Azimuth results to have a third source of annotations. We summarize the key properties of the methods compared in a table (tab 2.3) and provide an overview of all the mentioned methods in the appendix (C.2).

Cell Annotation

Three distinct cell annotation approaches were applied to the bin-to-cell output datasets as well as the scRNA-seq reference dataset to enable comparison of results. The annotation methods included manual annotation following state-of-the-art practices, automated annotation using CellTypist, and reference mapping using Azimuth.

The manual annotation pipeline followed established single-cell RNA-seq analysis practices. Cells were filtered to retain those with at least 10 detected genes, and genes expressed in fewer than 3 cells were removed. The data was normalized to median total counts and log-transformed. The top 4,000 highly variable genes were identified for downstream analysis. Principal component analysis was performed, followed by UMAP embedding. Leiden clustering was applied at multiple resolutions to identify cell clusters. Differentially expressed genes were identified for each cluster using the Wilcoxon rank-sum test. Based on marker gene expression and differentially expressed genes, clusters were manually assigned to cell types and stored in mapping files for reproducibility.

Automated cell type annotation was performed using CellTypist, a logistic regression-based classifier trained on large reference datasets. Two pre-trained models were evaluated: "Human_Lung_Atlas.pkl" and "Cells_Lung_Airway.pkl". The latter was selected for final analyses based on better confidence scores and cell-type proportions matching the reference. Input data was preprocessed according to CellTypist requirements: normalization to 10,000 counts per cell followed by log1p transformation. Confidence scores were extracted to assess prediction quality.

Reference mapping was performed using Azimuth through the panhumanpy interface, which projects query cells onto comprehensive human reference atlases. The model is not tissue specific, and therefore gives more general cell type label annotations. The method computes embeddings in the reference space and uses classification methods

to transfer cell type labels. Final level labels and associated prediction scores were extracted as quality metrics.

Summary Table

	Manual	**CellTypist**	**Azimuth**
Attributes	State of the art	Logistic regression model	Reference mapping
Requires gene markers	Yes	No	No
Model scope	Known markers	Tissue-specific models	Pan-human reference
Pre-trained models	No	Yes (40+ tissues)	Yes (multiple references)
Custom model support	No model	Yes	No
Novel cell type detection	Yes	No	Limited
Scalability	Manual effort	High	High
Included in ENACT	No	Yes	No
Confidence scores	No	Yes	Yes
Platform	Any	Python-based	R and Python supported

Table 2.3: Comparison of cell type assignment methods considered in the benchmarking pipeline.

Selection of Cell Types

To compare annotations, we decided to restrict the choice of possible cell types to a set of 17 cell types commonly found in the lung. Any other predictions would be aggregated to an additional group 'other'. This is non-trivial because one can only know the abundant cell types after having predicted the labels. We therefore stared with the labels from the CellTypist model Cells_Lung_Airway which showed better confidence score distributions than other models. We then ran the predictions on all the outputs to be compared, namely: 8x8 bins, Bin2Cell outputs, ENACT outputs, Xenium segmentation outputs, and the single cell reference dataset. The detailed cell type labels were aggregated into their parent groups (e.g. 'B_memory' and 'B_naive' to 'B') and any label group that did not show more than 1% abundance in any dataset of the above would be considered 'other'. With this mapping, each cell could be mapped to a class across datasets and methods. Each cell type was mapped to a color which is consistent among the results.

Metrics

The confidence scores can be used to compare quality of datasets, of course this will depend on how well the model fits the sample of interest. If the wrong model is

chosen, the confidence scores will be low, regardless of the quality of the dataset. We also compared the cell-type proportions of the different output datasets using the different methods comparing variability between outputs. We also plotted the spatial coordinates of the cells colored by cell types which helps to assess if the cell types cluster in similar tissue regions.

2.3.5 Specialized Scores

Promising approaches to evaluating the quality of segmentation use coexpression of genes within the segmented cells. The correlation of gene expression between related genes is used to assess the specificity of different segmentation methods. If segmentation correctly captures individual cells, genes that are typically coexpressed in a cell type should show strong correlations within those segmented units. Two scores, developed to evaluate coexpression using different approaches were replicated and used to examine the outputs.

Spurious Relative Coexpression Score

Jones et al. developed Proseg [41], a segmentation method that uses probabilistic modeling of transcript distributions. To benchmark segmentation quality, they introduced relative spurious coexpression. This score captures increases in gene coexpression, possibly caused by transcript misassignment when nuclear boundaries are expanded. This metric reflects the intuition that poor segmentation will cause transcripts to be misassigned, especially when neighboring cells are of different cell types. We implemented this metric and used it on the available bin-to-cell outputs to assess segmentation quality. Therefore we followed the algorithm description as closely as possible and adopted all the parameter choices from the original implementation which is available on GitHub (in Julia) [20]. The computation of the scores consists of two steps.

The first step involves taking prior nuclear segmentation, which by assumption has low rates of spurious coexpression, and comparing it to nuclear expansion segmentation, which by assumption has high rates of spurious coexpression. An especially large expansion parameter ($+8\mu$m) is used to match the assumption. Gene pairs that show dramatically higher rates of coexpression in nuclear expansion are considered a spurious pair. As both Bin2Cell as well as ENACT start from nucleus segmentation, we already had the data needed. ENACT additionally labels the expanded cells with the same ids as the nuclei which allows to obtain a direct mapping. Given the similarity between ENACT and Bin2Cell segmentations, we used the ENACT outputs to define the spurious gene pairs. For a cell-nucleus pair to be included, the cell must contain at least 50 transcripts to ensure enough signal. As the coexpression is affected by total transcript counts, the counts of the cells as well as nuclei are resampled to a multinomial distribution with a total count of 50 and probabilities matching the observed proportions for that cell. Let S_i denote the set of cells expressing gene i (count $\geq$1 in the resampled set). The conditional coexpression between genes i and j is defined as

$$C_{ij} = \frac{S_i \cap S_j}{S_j} \tag{2.1}$$

Any pair of genes (i, j) where $C_{ij}/C_{ij}^n \geq 1.5$ is defined as 'spurious', where C_{ij} is the

matrix with cell conditional coexpression and C^n_{ij} the matrix for the nuclei conditional coexpression. We saved the gene pairs for use in the second step. We also save the nucleus conditional coexpression matrix C^n_{ij} which is needed to normalize (therefore 'relative') the coexpression values.

In the second step we compute the relative spurious coexpression distribution for the output datasets. Note that these used different parameters as the data used to define the spurious gene pairs in the first step. The computation of the scores requires the output cell-count-matrix (using the bin-to-cell method of interest) and the information obtained before, namely the spuriously coexpressed gene pairs and the nuclei conditional coexpression matrix C^n_{ij}. We load the data for the methods considered: 16x16 bins, 8x8 bins, Bin2Cell, ENACT, and Xenium segmentation. Here again we make sure that the cells contain at least 50 transcripts and then resample and normalize them. The conditional coexpression matrix for the data C^d_{ij} is then calculated in the same way as above (eq. 2.1) and then normalized with the nuclei matrix. The values of the resulting matrix form the distribution of the relative spurious coexpression D.

$$\mathcal{D} = \{\frac{C^d_{ij}}{C^n_{ij}} \mid i, j = 1, \ldots, g;\ j = 1, \ldots, g\} \tag{2.2}$$

Note that even if we normalized to counteract the effect that larger transcript counts lead to higher coexpression values by resampling, methods with lower transcript counts will still perform better. This happens because cells with few transcripts likely also have less genes detected and therefore are more likely to have collisions when resampling.

MECR Score

In their benchmark study of multiplexed in situ gene expression profiling technologies, Hartman et al. introduce the Mutually Exclusive Co-expression Rate (MECR) [43], a novel specificity metric for benchmarking segmentations of SRT data. Unlike the spurious relative coexpression score, MECR directly quantifies the rate of co-expression between gene pairs known to be mutually exclusive in scRNA-seq data, offering a data-driven measure of molecular misassignment. The MECR can be interpreted as a measure for sensitivity. It is inspired by 'barnyard' analysis, where mixtures of cells from different species are pooled together in the same sample to assess specificity. It is, like the relative conditional coexpression, also based on coexpression scores and also affected by variations in transcript per cell counts. This favors approaches prioritizing specificity over sensitivity like applying minimal expansion or omitting transcripts with low assignment confidence. Therefore the authors suggest to use the total transcripts per cell as a measure of sensitivity and show it as a function of the MECR which is a measure for specificity. Results on the upper left of the plot would be optimal at both and therefore be favored. The MECR score calculation is also a two step process.

The first step consists in finding mutually exclusive gene pairs. These are pairs of genes who are typically not expressed in the same cell type. Common marker genes are great candidates. To find a set of such genes without bias, the mutually exclusive gene pairs are selected using a reference dataset. This dataset should be as close in terms of biology to the target dataset as possible. Optimal would be an adjacent tissue section. Also, a scRNA-seq experiment is favorable over SRT datasets as we can be sure that cells are separated and the sensitivity is higher. We therefore chose to use

the scRNA-seq reference dataset. The dataset is very large and has lots of genes. We therefore decided to focus on differentially expressed genes (DEG), identified using the standard Wilcoxon rank-sum test. Note that the scRNA-seq reference dataset cells are pre-labeled. To be selected as mutually exclusive, a gene must be expressed in at least 25% of its corresponding cell type and in only 1% of all the other cells. We made sure that the selected genes overlap with the ones in our Lung Cancer dataset. The result is a list of gene pairs, any pair corresponding to the same cell type is removed, resulting in a set of mutually exclusive gene pairs. The list of gene pairs can be represented as a matrix G_{ij} with marker genes on the rows and columns where $G_{ij} = 1$ if i and j are defined as mutually exclusive and $G_{ij} = 0$ else. We show an example for a 'barnyard' plot for 3 coexpressed genes and 3 mutually exclusive genes (fig 2.6).

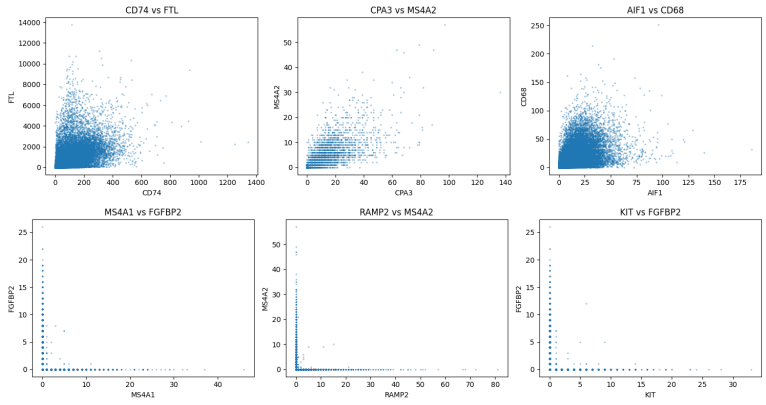

Figure 2.6: Barnyard plots on the scRNA-seq reference dataset for 3 mutually exclusive gene pairs (top) and 3 coexpressed genes (bottom).

The second part is again dataset specific. For each of the output datasets (cell-count-matrix), the coexpression matrix 'Coex' is calculated for the marker genes (eq. 2.3). Only the pairs contained in the mask are kept to form the MECR per gene distribution (eq. 2.4).

$$\text{Coex}_{ij} = \frac{S_i \cap S_j}{S_i \cup S_j} \qquad \text{for } i, j \in \text{DEG} \tag{2.3}$$

$$\text{MECR} = \{Coex_{ij} \mid G_{i,j} = 1\} \tag{2.4}$$

This distribution is a measure of specificity and is shown as box plots. The average of this distribution is then used as an estimation of the MECR for the whole dataset. It can be used to compare datasets on the same tissue, optimally on the same sample as we do. This average MECR for each dataset is plotted versus the transcripts per count value which acts as a proxy for sensitivity.

2.4 Code Availability

The scripts used to run the pipelines as well as the detailed evaluation of methods are publicly available in a GitHub repository. This includes a detailed guide on the transfer and alignment of Xenium segmentation to Visium HD data. All the mentioned input and output datasets are stored on Gstore, FGCZ's storage system, and can be accessed via SUSHI under Project 37785. The code for the SUSHI apps developed is available in the official SUSHI repository. They are based on EzPyz, a package created in the scope of this thesis to handle Python code within SUSHI.

Chapter 3

Results

3.1 Requirements on Computational Resources

All of the methods can be run using a CPU, and none of them specifically require a GPU. However StarDist which is based on TensorFlow can be run on a GPU which can lead to faster completion. Bin2Cell excels in low RAM usage, typically not exceeding 10 GB, even for relatively large dataset. ENACT can also be run on a standard laptop, however only if filtering to 1000 highly variable genes. To run ENACT on the whole set of genes requires around 70GB of RAM, which is easily available at FGCZ but could be limiting for individual users. The runtime is also longer for ENACT. It is important to notice here that ENACT not only includes the bin-to-cell assignment but also the cell type prediction.

Method	Run Time (hh:mm:ss)	Max RAM
Bin2Cell	00:17:20	8.6 GB
ENACT (by Area)	03:03:44	69.3 GB
ENACT (by Transcript)	14:31:37	212.5 GB
ENACT (by Cluster)	15:42:58	222.9 GB
ENACT (Naive)	01:45:45	69.1 GB

Table 3.1: Computational resources required for running the pipelines on the Lung Cancer dataset with default parameters (on all genes, no hvg filtering).

3.2 General Metrics

We show the results obtained when taking general QC metrics of the different bin-to-cell methods on the Lung Cancer dataset. This includes the 8x8 bins the 16x16 bins, the Bin2Cell (b2c) outputs, the ENACT outputs and the custom Xenium segmentation (xseg) pipeline outputs. We start by showing the cell count. Due to normalization of transcript counts (destriping) in ENACT and Bin2Cell, the transcript counts are not necessarily integers and can show effects of rescaling. The plot showcases that the 16x16 bins contain 4 8x8 bins each and that the obtained cell counts after running bin-to-cell methods are between the 8x8 bins and the 16x16 bins. The difference between Bin2Cell and ENACT can be explained by the added GEX segmentation in Bin2Cell. Finally we notice that the Xenium segmentation captured fewer cells than the StarDist-based methods. The 8x8 bin data can be filtered by transcript count, which can be done as

a filtering technique. This is suggested to ignore regions which do not overlap cells, but still contain transcripts at low density. To obtain a number of cells close to the segmentation-based approaches, one needs to remove all the 8x8 bins containing less than 56 (b2c) or 70 (xseg) transcripts per bin.

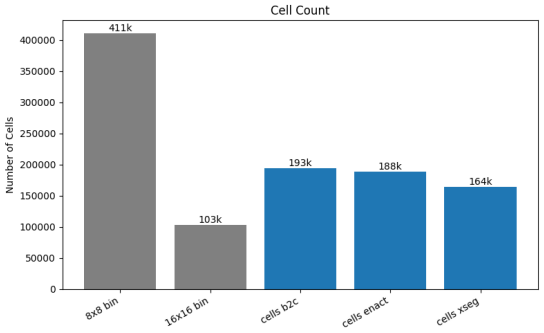

Figure 3.1: Cell counts obtained by method. Aggregated bins are shown in grey (bins are interpreted as cells).

The detected genes per dataset and distribution of genes per cell are shown. A high number of detected genes is favorable in order to extract the most information out of a sample. Some genes were not captured by any of the bin-to-cell methods, namely 'OR2T27', 'IFNL2', 'VSX1', 'ZP2', 'PRAMEF4', 'PRB4'. 'OR2T27', 'VSX1', 'ZP2' and 'PRB4'. These excluded genes are specific to regions outside the lung and are not expected to be found in this tissue. All the excluded genes have a count of 1 in the dataset.

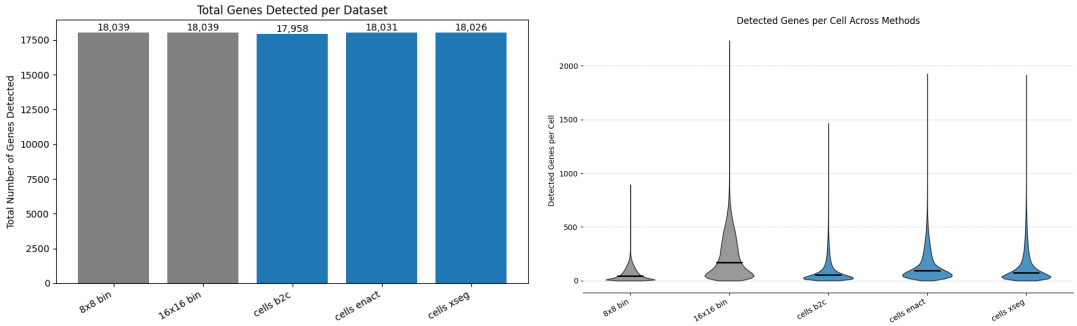

Figure 3.2: Gene detection per dataset (left) and per cell (right). Median shown in black.

Next we show the total transcript counts and the related unassigned transcripts. The 8x8 bins and 16x16 bins cover the whole dataset and thus contain the total counts for the sample. ENACT captured most of the transcripts, leaving only 3.6% unassigned. The fraction of unassigned transcripts can be varied by choosing larger expansion parameters, at the cost of risking to increase noise.

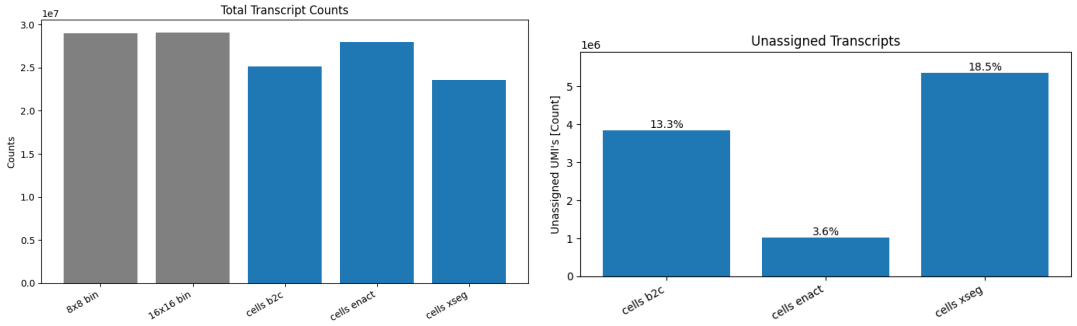

Figure 3.3: Transcripts captured by the methods. The bins (8x8 and 16x16) contain the total amount of detected transcripts.

Dividing the total number of transcripts by the number of cells detected gives the average transcripts per cell. The distribution is also shown. For reference, a scRNA-seq dataset commonly has a mean transcript per cell count in the range of thousands.

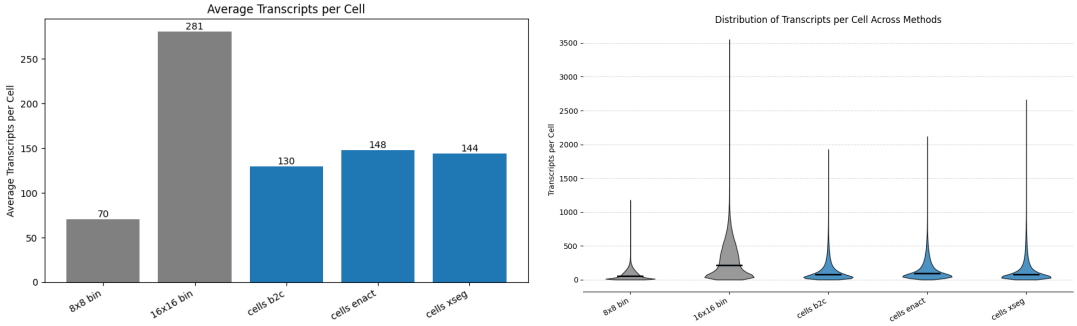

Figure 3.4: Transcript per cell averages and distribution.

Lastly, we use the spatial coordinates to show the spatial distribution of the cells. The hexbins are colored by cell density. The results across methods look similar. For the bins, the density is trivially uniform, thus we show the transcript density instead. Note that Bin2Cell is, due to the GEX segmentation step, the only method to map the debris on the left empty space of the sample to cells.

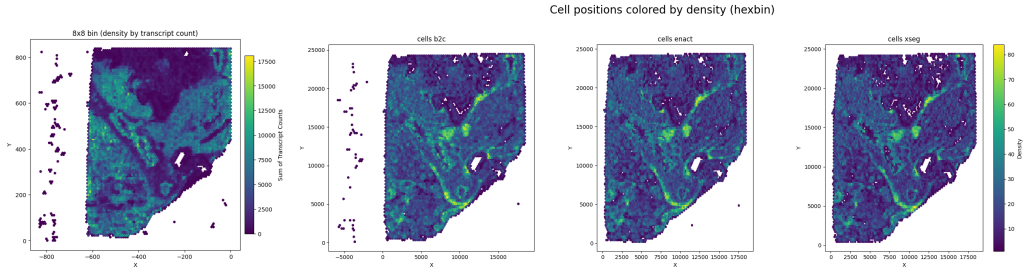

Figure 3.5: Transcript density for the bin dataset and cell density for the bin-to-cell outputs. Methods from left to right: 8x8 Bins, Bin2Cell, ENACT, Xseg.

The UMAP representation is also computed as part of the basic QC evaluation, we cluster the cells using Leiden clustering and show the clusters in the first two UMAP coordinates as well as in spatial coordinates. Notably, the UMAP representation of the bin-to-cell outputs is quite similar across methods and most importantly quite different to the 8x8 bin dataset.

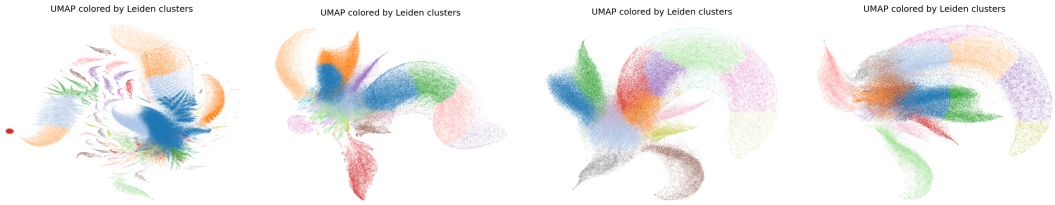

Figure 3.6: UMAP's colored by Leiden clusters. Methods from left to right: 8x8 Bins, Bin2Cell, ENACT, Xseg.

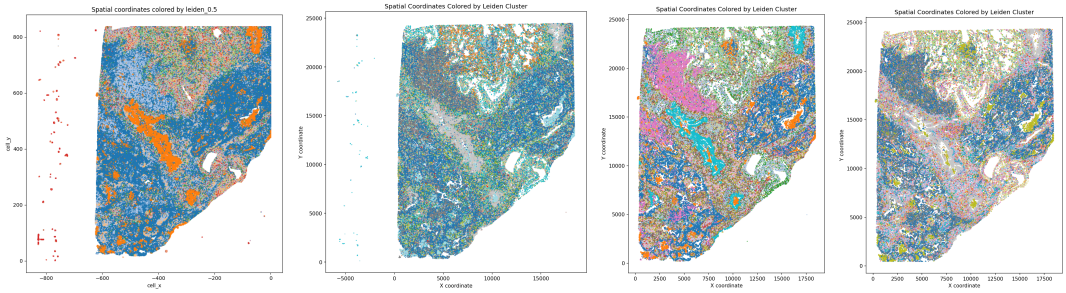

Figure 3.7: Spatial Coordinates colored by Leiden clusters. The colors correspond to the UMAPs above. Methods from left to right: 8x8 Bins, Bin2Cell, ENACT, Xseg.

Space Ranger v4

To compare Bin2Cell and ENACT to the recently released Space Ranger v4, we computed the general metrics for the Rhesus Macaque dataset. We confirm the same differences between Bin2Cell and ENACT. Bin2Cell has a slightly higher cell count due to the GEX segmentation step. ENACT however captures more transcripts, leading to low unassigned transcript rate and most available genes captured. Space Ranger v4 (SR4) detected a higher amount of cells and has similar transcript coverage as ENACT.

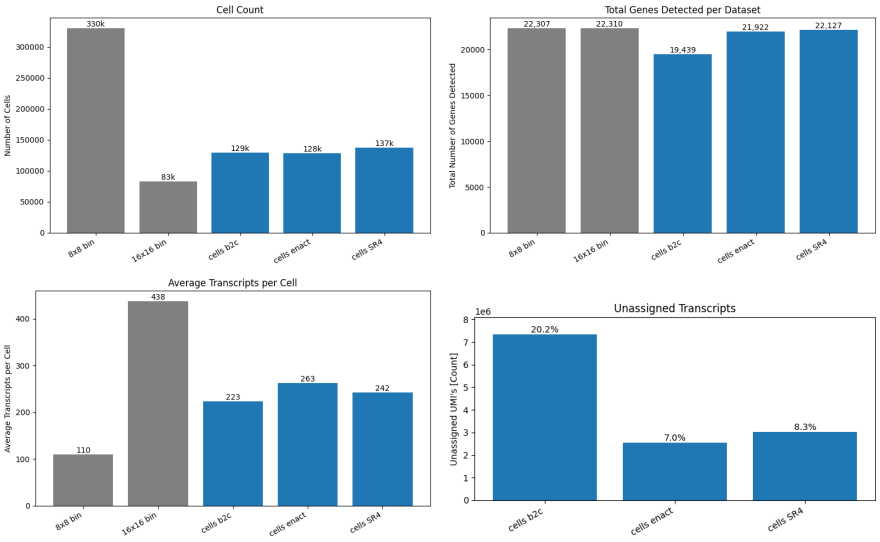

Figure 3.8: General metrics obtained for the Rhesus Macaque dataset. The metrics are used to compare between the bins, Bin2Cell, ENACT and Space Ranger v4. No data is available for Xseg considering this dataset.

A closer look at the segmentation outputs reveals that Space Ranger v4 uses low prob_thresh default parameters. Debris around the sample are apparently segmented as nuclei and in some cells the nucleus is not visible by eye. We show the spatial map of cell positions for comparison of outputs. More analysis results can be found in the appendix.

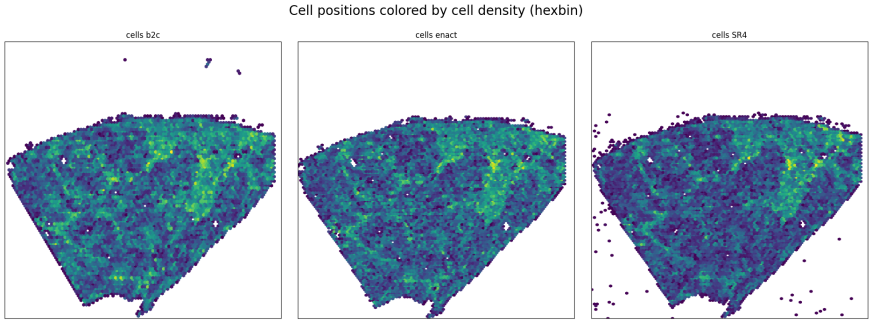

Figure 3.9: Spatial coordinates of the cells in the Rhesus Macaque dataset, visualized as a density map.

3.3 Segmentation-Based Metrics

For the comparison of the expanded StarDist segmentation (ENACT) and the Xenium segmentation we selected a tissue crop. This offers a qualitative overview of the differences between the methods. In this crop, we show the H&E high resolution microscopy image, overlayed with both types of segmentation.

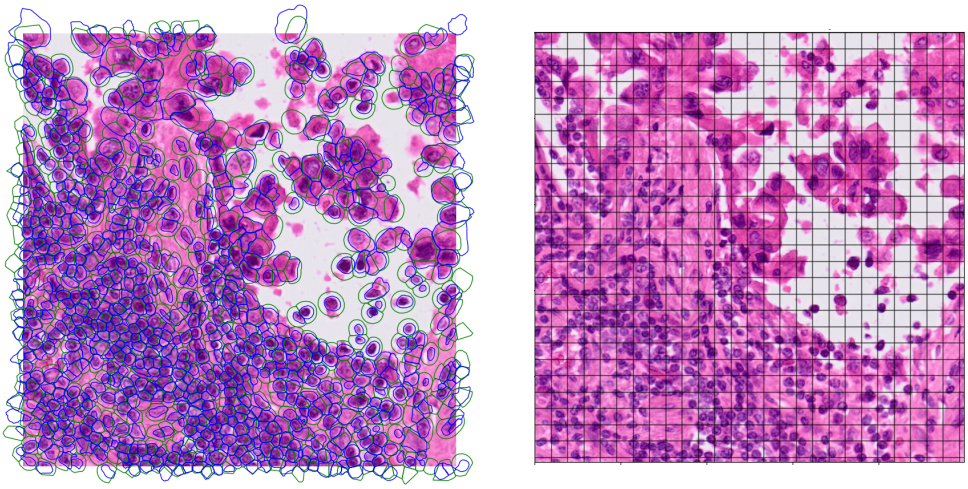

Figure 3.10: Segmentation overlay on a crop of the H&E tissue image (left). Xenium in blue StarDist in green (nuclei expanded by 4μm). A representation (to scale) of the 8μm x 8μm bins is shown for reference (right).

The demeaned polygon shapes for both types of segmentation are stacked and plotted to show segmentation characteristics. The plot only includes cells shown in the crop shown in fig. 3.10. The blue boundaries clearly show the complex shapes segmented by Xenium. The green shapes illustrate the effect of circular expansion: mostly circular boundaries or straight lines and corners where cells touch. Also the size (distribution) of the cells is different.

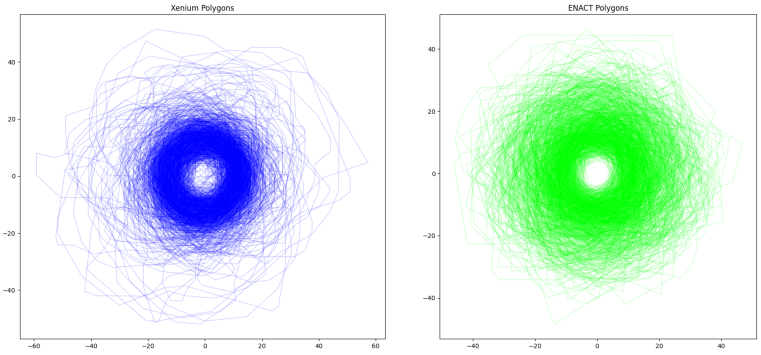

Figure 3.11: Cell segmentation silhouettes for Xenium (blue) and expanded StarDist (green).

The segmentation defines the cell area which can be shown as a distribution over all detected cells. The area of the cells segmented by ENACT does not vary much from the mean. This is expected as the expansion distance is a set parameter. Varying the parameter can be used to shift the distribution towards different cell sizes. Xenium segmentation shows tails on both ends with respect to ENACT, meaning that it detected smaller as well as much larger cells in comparison. The largest segmented cell covers an area of 10739 pixels (486μm^2) in Xenium while only 4975 pixels (225μm^2) in ENACT.

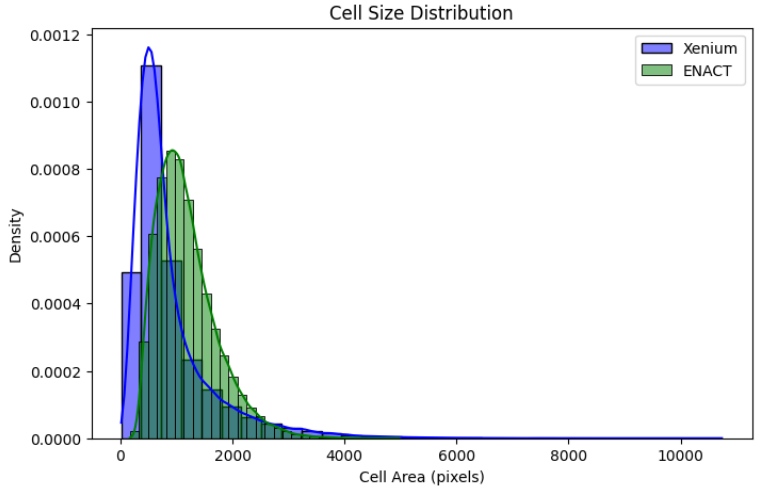

Figure 3.12: Area distribution for both segmentation methods. The distribution includes the whole dataset.

If tissue is detected accurately, the cell area should correlate to the transcripts per cell. We therefore compare the number of underlying squares to the number of transcripts per individual cell. The reason we don't use the segmentation directly is that the mapping from segmentation to output cell is not part of the outputs. The slopes are above the average transcripts per square value of the dataset (4.4), meaning that the segmented cells contain a higher transcript density than random segmentation.

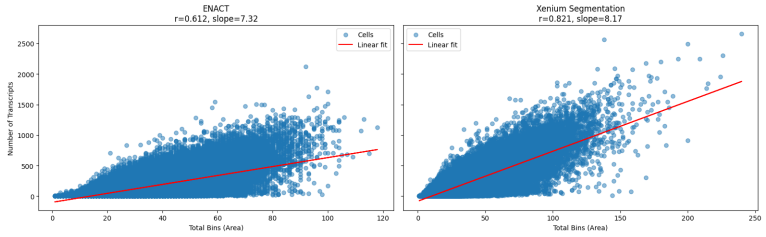

Figure 3.13: Scatter plots showing the relationship between cell area (total bins) and tpc for ENACT (left) and Xenium Segmentation (right).

We summarize some important segmentation-related metrics for both segmentation methods in a table.

Metric	**Xenium**	**ENACT**
Mean cell area	40.65 μm^2	53.36 μm^2
Median cell area	29.59 μm^2	48.89 μm^2
Total area coverage	6.71 mm^2	10.12 mm^2

Additionally we report the intersection over union between both the segmentations. This is computed by considering the whole set of cells as one segmentation as there is no 1-to-1 mapping between single cells. We report the Dice coefficient in the same way.

$$\text{IoU} = 0.467 \tag{3.1}$$

$$\text{Dice} = 0.614 \tag{3.2}$$

3.4 Annotation-Based Metrics

To compare the results of the cell type predictions from SRT data, we decided to analyze a scRNA-seq reference with similar tissue. Cell annotation was performed using three different methods, across all output datasets as well as the scRNA-seq reference. Since the scRNA-seq data does not include the spatial coordinates we can not show them, however we compare the cell type proportions. Also, as the automated annotation tools were designed for scRNA-seq, the reference is used to set a baseline of results to compare the SRT data with. The colors of the simplified cell types are consistent across datasets and annotation methods. The results are presented by annotation method, including manual annotation, CellTypist and Azimuth.

3.4.1 Manual Cell Annotation

The manual annotation was done for three datasets. The scRNA-seq dataset, serving as a reference to see what kinds of cell types can be expected in such tissue and roughly at which proportions. The 8x8 bins are considered a state of the art method for comparison. And the ENACT outputs represent bin-to-cell methods in general. The spatial distribution, colored by cell type shows similarities between the binned and ENACT outputs (fig. 3.14). B-cells, Secretory cells and Macrophages can be found in similar regions in both datasets. For other cells, the cell annotations are not consistent with each other. This is well visible in regions where the 8x8 bins were annotated as fibroblasts (Fibro) while the ENACT cells were annotated as basal cells.

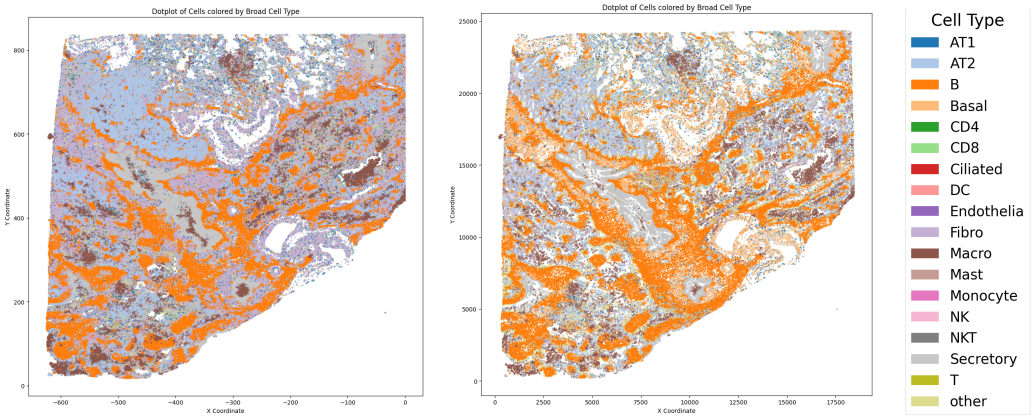

Figure 3.14: Spatial distribution of the annotated cells after manual annotation. The left image shows the annotation for the 8x8 bins and the right image for the ENACT outputs.

The cell type proportions highlight further differences between the predictions. It is important to keep in mind that the reference dataset was obtained from a different tissue sample. The largest contrast is visible in the amount of T cells. Also multiple cell types were only found in the scRNA-seq dataset, including ciliated cells, mast cells, monocytes, natural killer cells (NK) and dendritic cells (DC). The scRNA dataset confirms that a large proportion of immune cells can be expected for such tissue.

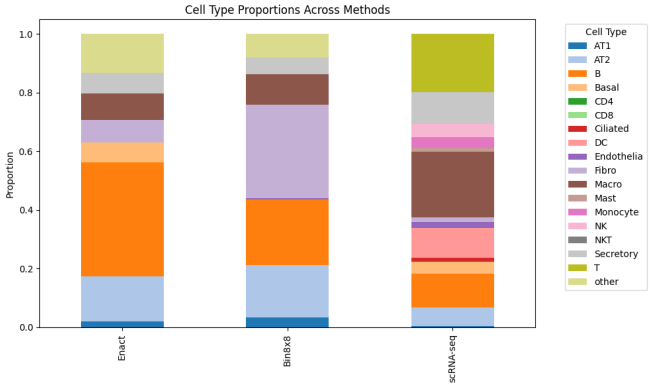

Figure 3.15: Predicted cell-type proportions for the manually annotated datasets.

3.4.2 CellTypist

CellTypist is the annotation method we examined most carefully. We considered different CellTypist models, datasets from different tissues, different modalities and of course different bin-to-cell methods. CellTypist provides two models which can be used to annotate lung tissue: Human_Lung_Atlas and Cells_Lung_Airway. Both the models were used on the bin-to-cell outputs but returned different cell-type proportions. The Human_Lung_Atlas model annotated the majority of cells as Basal resting cells while Cells_Lung_Airway preferred to annotate the cells as macrophages. We show the proportions of annotated cell types for the ENACT outputs as example.

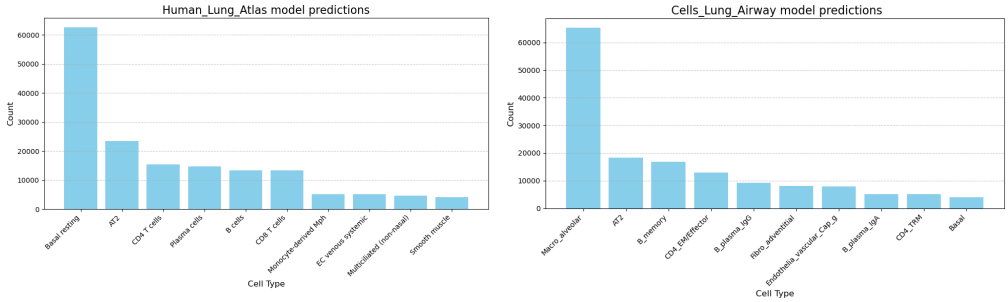

Figure 3.16: Human_Lung_Atlas model left, Cells_Lung_Airway model right. The predictions for the ENACT outputs are shown, the pattern was similar across bin-to-cell output datasets. Only the top 10 cell types are shown.

To define which model made the more plausible predictions, we compared the cell-type proportions to the scRNA-seq data and also checked the confidence scores on the predictions. Both models have high confidence scores on the scRNA-seq reference and both models predict a high proportion of (alveolar) macrophages. Also the confidence scores of the Cells_Lung_Airway model on the bin-to-cell outputs was consistently higher across methods. Therefore further results are shown with the Cells_Lung_Airway model predictions only.

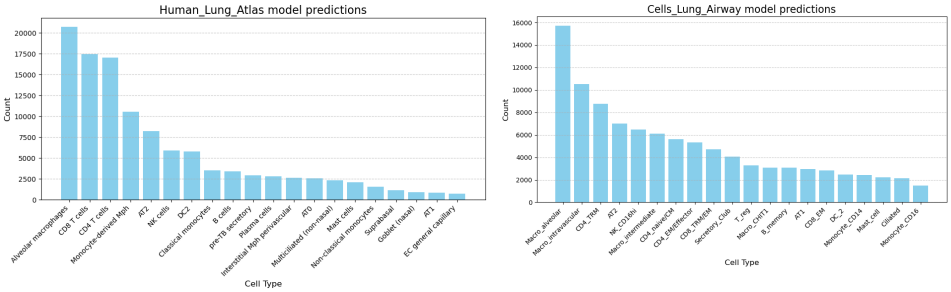

Figure 3.17: Cell-type proportions after running CellTypist on the scRNA-seq reference. Human_Lung_Atlas model left and Cells_Lung_Airway right. Only the top 20 cell types are shown.

The confidence scores for the cells each form a distribution which we show as a ridge plot. The four bin-to-cell methods are shown as well as a Visium HD dataset ENACT on tonsil tissue, segmented by ENACT and the scRNA-seq reference dataset. The confidence scores are rather low for the bin-to-cell methods. The high confidence scores show that neither Visium HD as technology nor ENACT as a bin-to-cell method are solely responsible for the low confidence scores. The scRNA-seq results also show that tumor tissue does not necessarily lead to low confidence scores. A spatial map of low confidence scores as well as additional metrics is reported in the appendix. The main difference between the datasets with low and high confidence scores is the transcript count per cell. The bin-to-cell methods on the Lung Cancer dataset have average transcripts per cell between 130 and 148 and the 8x8 bins 70 tpc. The Tonsil cells segmented using ENACT average at 5250 tpc and the scRNA reference at 7997 tpc.

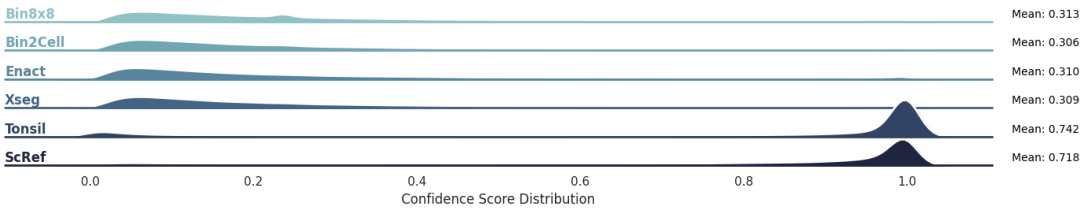

Figure 3.18: Confidence score distribution on the CellTypist predictions. Model used: Cells_Lung_Airway.

The results for cell type annotation using CellTypist are shown as spatial distributions. We also show the corresponding cell type proportions. The results are consistent across methods. Even for the 8x8 bins, the spatial locations of cell types mostly match the results from bin-to-cell methods. The large amount of 8x8 bins compared to cells is visible as higher density. In regions where less cells were segmented, CellTypist predicts macrophages, this is also reflected in the large proportion of macrophages for the 8x8 bins method. The experimental Xseg method leads to proportions very similar to the 8x8 bins. The cell type proportions are overall similar across the bin-to-cell methods. The main differences can be seen in number of macrophages and B-cells.

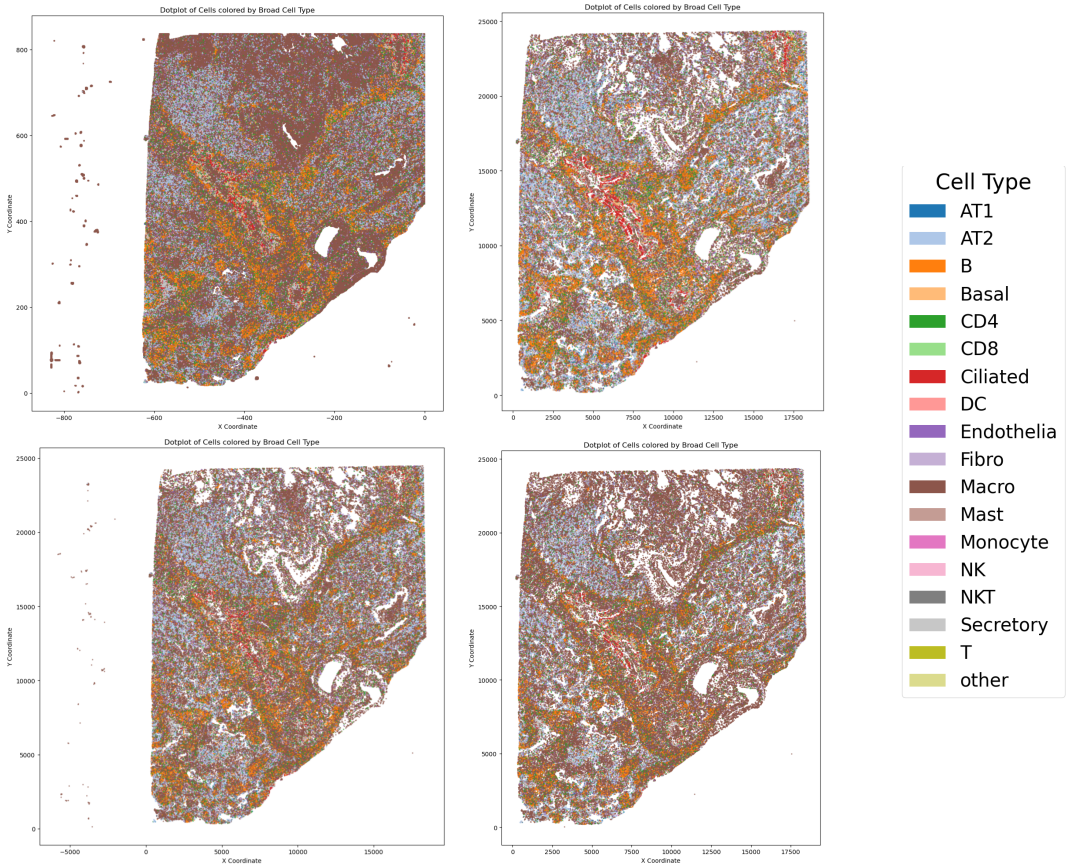

Figure 3.19: Spatial locations of the cells colored by predicted cell type. 8x8 Bins (tl), Bin2Cell (bl), ENACT (tr), Xenium segmentation (br).

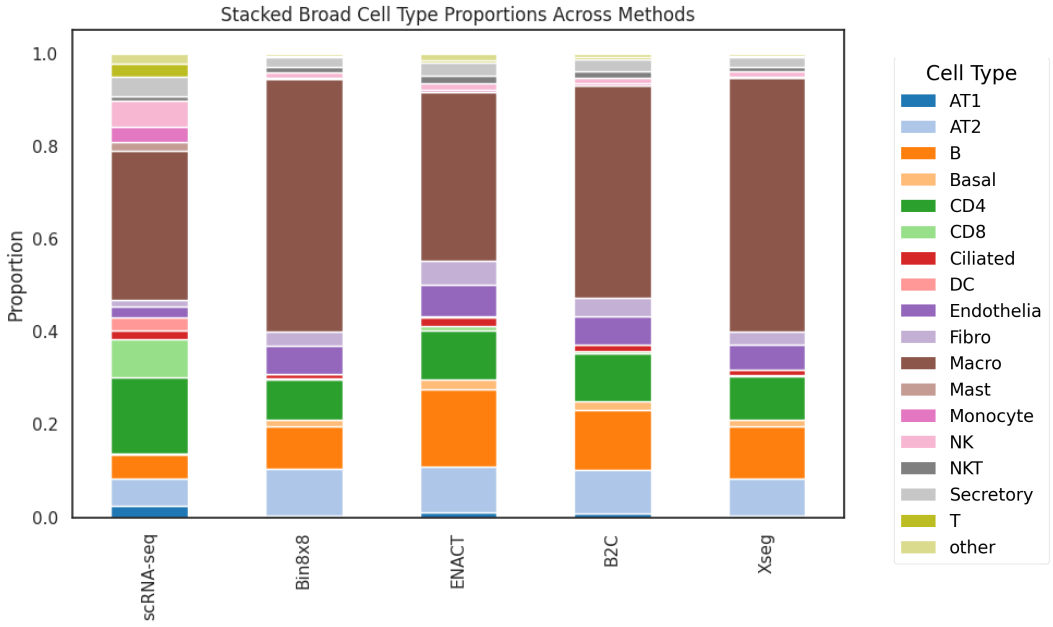

Figure 3.20: Cell-type proportions for the cell type predictions by CellTypist.

3.4.3 Azimuth

Azimuth was used to have a second method performing the cell type annotation to compare the variability within annotation methods. Azimuths model is panhuman. As a consequence, less predicted cell types mapped to the ones selected for comparison, leading to many categorized as 'other'. Also some cell types were not predicted at all in the bin-to-cell datasets, namely basal, ciliated, NK and NKT cells. Again the spatial distribution of cell types predicted is consistent across methods. For the 8x8 bins, regions where the bin-to-cell methods did not segment any cells (white) are filled with cells predicted as 'other'. The proportions are again similar, however the diversity of cell types is clearly low for all the SRT datasets, compared to the scRNA-seq reference. The predictions include mostly B-cells, T-Cells and secretory cells.

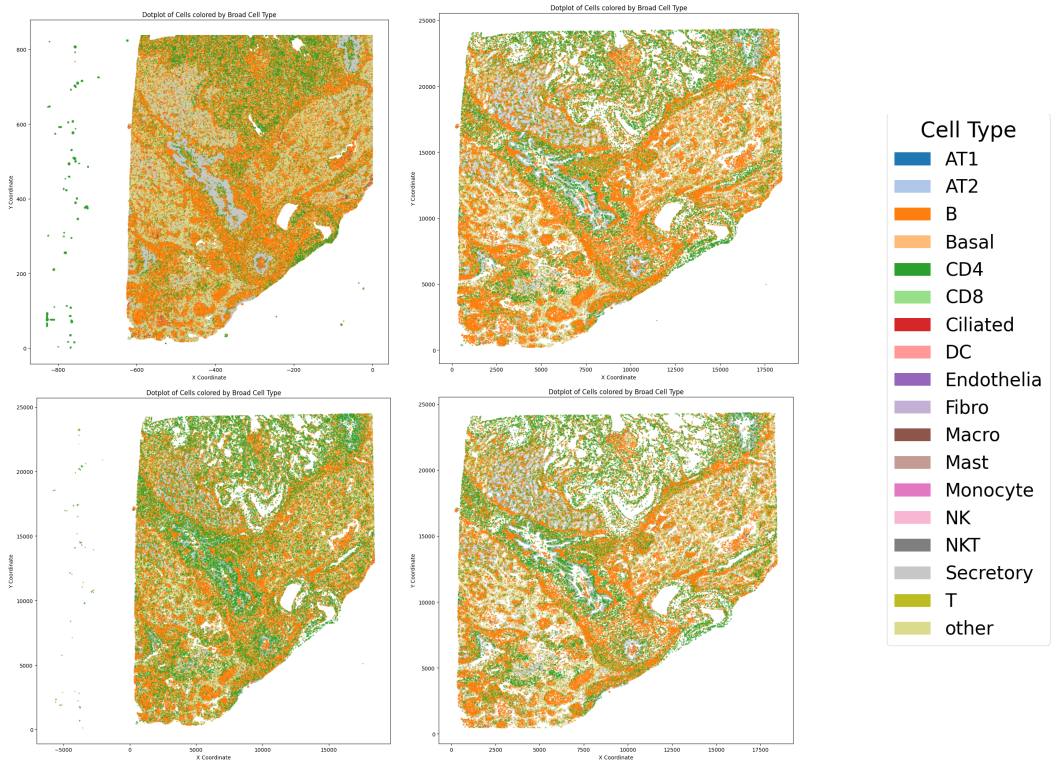

Figure 3.21: Spatial locations of the cells annotated by the Azimuth model. 8x8 Bins (tl), Bin2Cell(bl), ENACT (tr), Xenium segmentation (br).

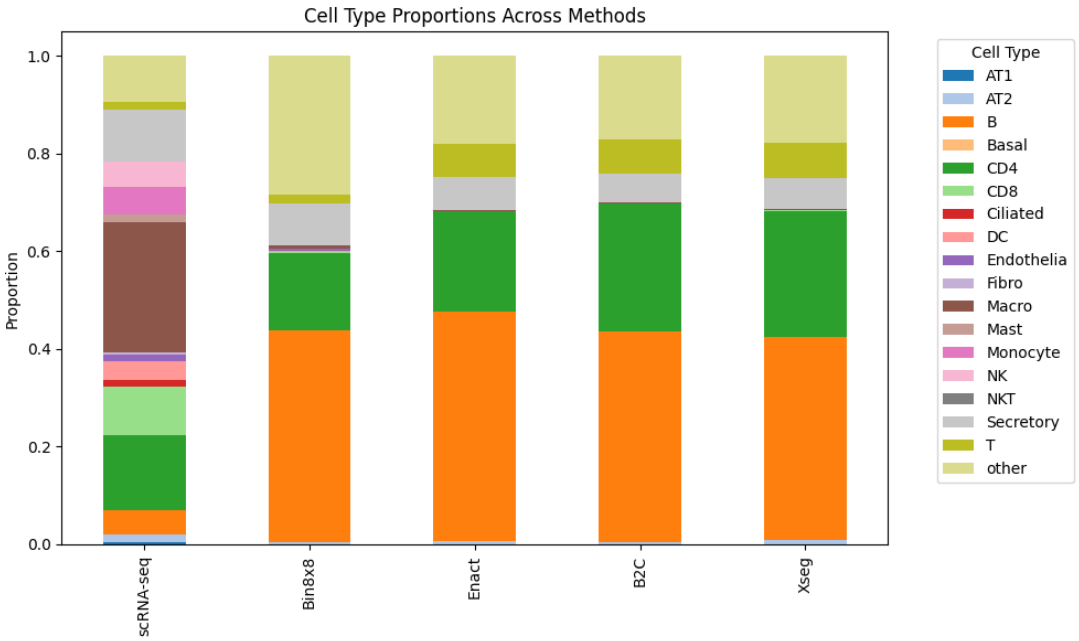

Figure 3.22: Cell-type proportions for the predictions by Azimuth.

3.5 Coexpression-Based Scores

Both specialized scores presented in the methods section were evaluated for at least the four bin-to-cell methods. They are both based on coexpression of genes so the results are similar. Especially two assumptions are shared and need to be kept in mind when interpreting the results. The first is that genes are only perceived as expressed or not expressed without considering the magnitude of expression. Second, ENACTs 'Weighted-by-Area' method, which is the default, will split the transcript counts of a bin overlapped by multiple cells to all of the cells. This increases the signal of genes expressed in boundary regions, effectively leading to a 'blurring' effect. When considering coexpression this inevitably leads to higher scores.

3.5.1 Spurious Relative Coexpression Score

We present the spurious relative coexpression scores for two bin sizes (8x8 and 16x16) as well as the three bin-to-cell outputs (B2C, ENACT and Xseg). The distribution of relative spurious coexpression scores obtained for each gene pair is shown in a violin plot. Lower relative spurious coexpression rates suggest higher-quality segmentation.

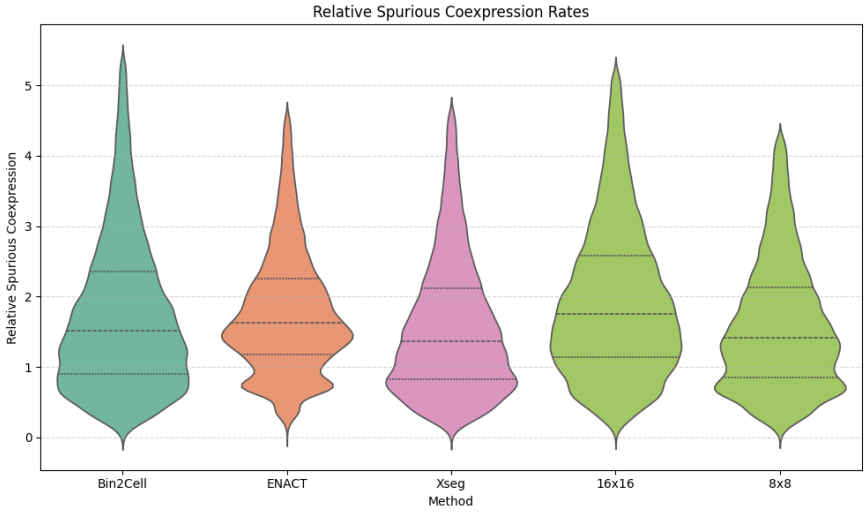

Figure 3.23: Relative spurious coexpression distributions.

The scores suggest that using the Xenium segmentation seems to be the best option among the bin-to-cell methods. However it is also important to note that the 8x8 bins as a crude method also achieve best scores. This can be explained by the sensitivity-specificity trade-off. The score is a measure for specificity so lowering sensitivity can improve performance. This is well visible in the comparison of 8x8 bins and 16x16 bins, they use the same 'algorithm' but vary in sensitivity and specificity. We therefore additionally show a plot of the mean relative spurious coexpression versus the mean transcript count per cell. High quality data can therefore be found in the top left of the plot and the Xenium segmentation shows clear benefits compared to the 8x8 bins.

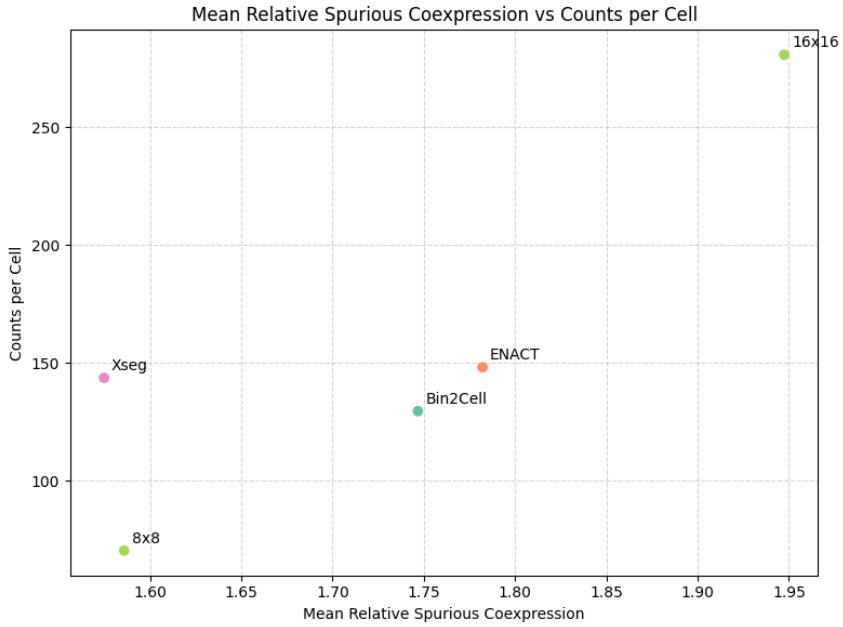

Figure 3.24: Mean relative spurious coexpression distributions (specificity) by mean transcript counts per cell (sensitivity).

3.5.2 MECR Score

The set of mutually exclusive genes was found using the scRNA-seq reference dataset. Across methods, the barnyard plots show very low coexpression of MS4A2 and RAMP2, a gene pair defined as mutually exclusive. On the other hand FTL and CD74, a pair not defined as mutually exclusive is highly coexpressed. Such a plot can be generated for any pair of genes. In case of very important genes for downstream analysis (e.g. marker genes) this can help determine data quality. The overall differences across datasets are quantified by the MECR Scores.

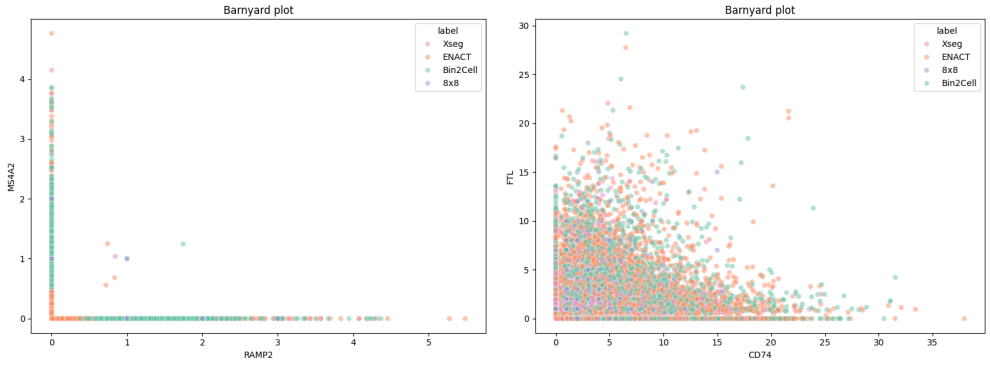

Figure 3.25: Barnyard plots for a gene pair defined as mutually exclusive (left) and a coexpressed gene pair for reference (right).

The MECR scores are shown as a barplot for each method using default parameters as well as all the bin sizes. Again, it is important to note that cells with lower transcript counts will have a lower MECR score. This is shown by the difference between 16x16, 8x8 and 2x2 bins. The MECR scores can not be compared across different samples.

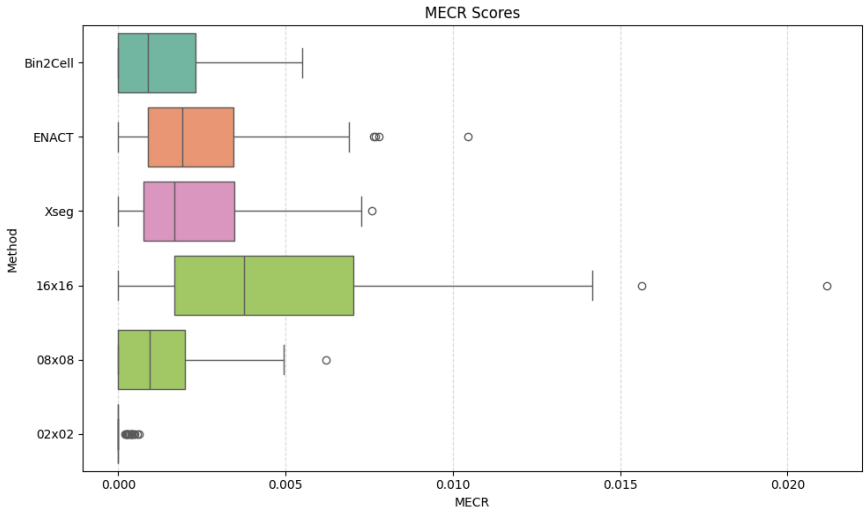

Figure 3.26: MECR for the bin-to-cell outputs and for the binning methods for reference.

Hartman et al. identified that the average MECR can be used as a measure of specificity. It is therefore important to put the results in relation to sensitivity. The mean

counts per cell can be used as a sensitivity proxy. When possible, we ran pipelines with different parameters to explore various specificity-sensitivity trade-offs. The optimum would be to have high sensitivity and high specificity. In this case a high transcript per cell count and a low average MECR score is desired. For this evaluation, not only the default methods at different expansion parameters were used but also the different transcript assignment methods available for ENACT (see table 2.1). The lowest expansion parameter always corresponds to the lowest point of the series and for most methods represents the nucleus segmentation without expansion. For the Xseg method, the segmentation starts at whole cell size so only further expansion could be considered.

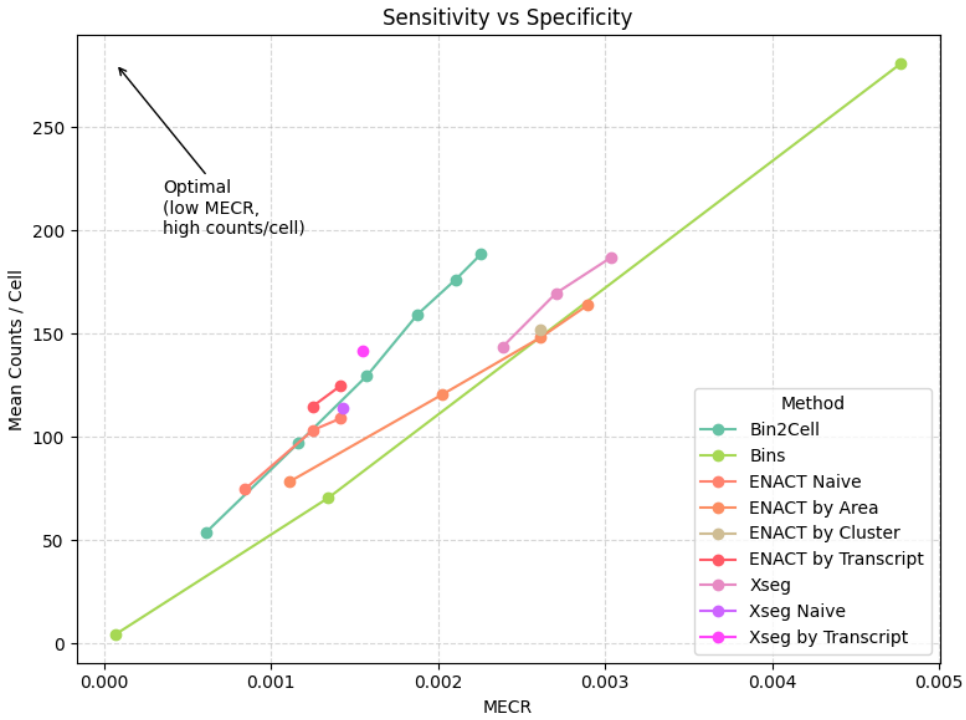

Figure 3.27: Mean counts per cell act as a proxy for sensitivity and the average MECR acts as a proxy for specificity. Different expansion parameters allow to express sensitivity as a function of specificity and find an optimal method.

The binned outputs act as a baseline and all explored methods show an improvement relative to this baseline. Only ENACT using 'Weighted-by-Area' and on the largest expansion falls below. The best performing methods are Xseg using the 'Weighted-by-Transcript' transcript assignment and ENACT using the 'Weighted-by-Transcript' transcript assignment. Bin2Cell also shows remarkably good scores across expansion sizes.

Chapter 4

Discussion

This thesis provides a comprehensive comparison of bin-to-cell methods for Visium HD data. We evaluated three distinct approaches: Bin2Cell, ENACT, and custom Xenium segmentation and compared them to the current state of the art, namely binning squares to $8\mu m \times 8\mu m$ bins. We also used a scRNA-seq dataset to highlight key differences. Our analysis across multiple evaluation metrics reveals important insights in computational efficiency, practical applicability and data requirements. This supports an informed method selection for different research scenarios.

4.1 Interpretation of Results

4.1.1 Performance Comparison of Bin-to-Cell Methods

Each bin-to-cell method offers distinct advantages depending on the specific requirements and constraints of the analysis. The evaluation using general QC metrics, segmentation-based metrics, annotation-based metrics, and coexpression-based scores gives an overview of what a user can expect from such methods, and what not.

Bin2Cell demonstrated remarkable computational efficiency while maintaining satisfactory performance across metrics. As the only method utilizing GEX representation in addition to nuclear segmentation, Bin2Cell offers unique advantages when nuclei are lost or absent from the data. However we noticed that this feature can also classify debris as cells. The requirement that a segmented cell is only kept if it does not overlap with the H&E segmentation at all is also too restrictive to detect cells, especially in tightly packed tissue regions. The method is easy to understand and provides results within short run time including extensive intermediate visualizations. Our results show that Bin2Cell achieves results similar to other methods while using the simplest algorithm, making it particularly suitable for resource-constrained environments or fast data exploration.

ENACT offers the most complete solution, its monolithic framework supports reproducible and well defined workflows. While being computationally more demanding than Bin2Cell, it offers an all-in-one solution to obtain annotated cells from Visium HD squares. The method captured most transcripts while maintaining reasonable cell counts. Using a large expansion parameter can help recover diffused transcripts. The spatial distribution analysis showed that ENACT can produce biologically meaningful cell arrangements. However, our coexpression analysis revealed that ENACT's default 'Weighted-by-Area' method introduces a 'blurring' effect at cell boundaries,

where transcript counts are split between neighboring cells, leading to artificially elevated coexpression scores. The 'Weighted-by-Cluster' method suffers from the same issue, while using the 'Naive' or the 'Weighted-by-Transcript' methods showed the best results. The motivation based on the problem that transcripts from a single bin may belong to different cells is important to be tackled. However the methods proposed do not seem to lead to significant improvement considering downstream results like cell annotation. The 'blurring' by splitting transcript count between cells is especially important to consider when studying genes at an expressed/not expressed level as for example for marker genes.

Xseg, an experimental method using Xenium segmentation offered many insights to assess the effect of accurate segmentation. This is particularly interesting because it uses the same algorithm as ENACT only with different segmentation. This means that every result obtained can be used to study the effect of detailed segmentation. This custom approach outperformed the other methods in the coexpression-based metrics and resulted in biologically meaningful results in downstream analysis. The method excels when cells cannot be accurately approximated as circular shapes, overcoming the limitation of current bin-to-cell methods. However, the analysis requires expensive Xenium data acquisition or similar sophisticated segmentation methods, limiting its practical applicability.

4.1.2 Segmentation Accuracy

The segmentation comparison between expanded Stardist (ENACT) and Xenium segmentation revealed substantial differences in cell shape representation and size distribution, reflecting genuine biological differences in segmentation approaches. Xenium's detection of both smaller and larger cells compared to ENACT's more uniform size distribution confirms that circular expansion may not adequately capture the full spectrum of cell morphologies present in complex tissues. However the use of specialized staining to obtain accurate segmentation can be expensive. Furthermore Bin2Cell as well as ENACT cells resulted in similar metrics and annotations. The variability between results was higher between samples and method parameters rather than between the methods themselves. We advise users to visually confirm the accuracy of segmentation results and exploring various parameters to find the best fitting results. We also highlight the importance of obtaining good quality data, which is the only way to obtain satisfying results, regardless of the quality of the segmentation. Also an estimation of expected transcripts per count, considering the quality needed to tackle a research question should be done before any measurement.

4.1.3 Cell Type Annotation

Our annotation experiments using CellTypist and Azimuth revealed significant challenges when applying scRNA-seq-trained models to Spatial Transcriptomics data. The low confidence scores using both automated cell annotation tools are not trivial to explain. The Visium HD Tonsil dataset processed by ENACT and the scRNA-seq reference both obtained high confidence score annotations with the corresponding models. This excludes Visium HD as a source of data, ENACT as a bin-to-cell method and diseased tissue as unique causes for the low scores. We assume that the reason for the low

confidence scores is either a combination of the above or probably rather the low transcript density of the Lung Cancer dataset. While the best bin-to-cell method reached 148 tpc, most scRNA and the most sensitive Visium HD assays (after bin-to-cell assignment) reach tpc values in the thousands. The counts are normalized before the models are used to annotate, however the normalization can not account for the complex transcriptomic profile expressed by the cells. The transcript density and presence of marker genes not only depends on the quality of the data but also of the selected tissue type and region.

4.1.4 Sensitivity-Specificity Trade-Offs

A critical insight from our analysis is the inherent sensitivity-specificity trade-off present in bin-to-cell methods. The coexpression-based metrics clearly illustrate this: while 8x8 bins achieved excellent specificity scores, they suffer from low sensitivity. Conversely, methods using larger expansion parameters tend to show increased spurious coexpression (reduced specificity). The MECR curves demonstrate that optimal performance lies in balancing these competing factors, with Xseg achieving the best overall results where transcript count per cell is considered alongside specificity metrics. The optimal balance also depends on the application and research question as well as on the quality of data available. We therefore recommend to run the pipelines with different expansion parameters and visualize the results using the presented metrics.

4.2 Limitations

The thesis faced various limitations in time, data availability, software compatibility, biological expertise and in the rapid evolution of technologies and methods. We summarize and justify important constraints, and suggest approaches for further work.

4.2.1 Dataset Constraints

This thesis' primary limitation is the reliance on the Lung Cancer dataset for the main analysis. The choice was necessitated by data quality requirements, as it is currently the only publicly available Visium post Xenium dataset. Efforts to include a Visium post Xenium skin dataset from FGCZ failed due to insufficient transcript counts. The transcript density of the Lung Cancer dataset is also rather low even compared to other Visium HD datasets. The exclusive use of diseased tissue introduces a potential confounding factor. Cancer tissues may exhibit cells with altered gene expression patterns, and infiltration by immune cells is common. The large proportion of cells labeled as immune cells by the annotation methods was at first surprising. It was however confirmed by consistently high proportions across annotation methods as well as similar proportions in the reference dataset.

4.2.2 Methodological Limitations

The absence of a ground truth is a common challenge in benchmarking. Although we employed multiple complementary metrics and compared results across methods,

establishing a definitive ranking would be overly ambitious. This limitation is particularly relevant when interpreting the superiority of one method over another, as different metrics may capture different aspects of transcript assignment. Furthermore the rapid development of Spatial Transcriptomics method leads to changes in the method landscape, which could not be accounted for. SpaceRanger v4 was released during the work on the thesis and not implemented in time, limiting our comparative analysis to the Rhesus Macaque dataset only. The development of new methods and technologies presents an ongoing challenge for method comparisons, as new tools emerge with potentially improved capabilities. The lack of direct comparison between the custom Xenium segmentation pipeline and SpaceRanger v4 represents a missed opportunity to evaluate the most current method comprehensively.

4.2.3 Possible Improvements and Outlook

As only StarDist was included in current bin-to-cell pipelines, we did not compare it to other segmentation methods. The exploration of alternative segmentation approaches, such as comparing StarDist to Cellpose for nucleus or even whole-cell segmentation, could provide insights into segmentation optimization. Segmentation models can be expected to further improve in future. To obtain a more comprehensive evaluation of the methods, the most important aspect would be to consider more datasets on different tissue types. However the limitation to the considered datasets will always remain as samples can vary a lot between tissue types and in sample quality. Also data acquisition is expensive and labor intensive. The development of synthetic datasets with known ground truth segmentation could provide valuable benchmarking resources for method comparison. Such datasets would enable systematic evaluation of method performance across varying cell shapes and densities, tissue architectures, and transcript abundance levels. Looking one step further, integration of multiple SRT data modalities or including protein expression data with Spatial Transcriptomics could be interesting for tissue characterization. Such modalities could not only improve the insights obtained, but also help to validate the results.

4.3 Value for Users and Staff at FGCZ

With the creation of two apps, Bin2Cell and ENACT were integrated into FGCZ's data analysis platform SUSHI. We added to the standard outputs some figures showing the interim and final results on a crop region which can be selected by the user and which is expanded automatically if not overlapping tissue. Also, we provide a standard analysis jupyter notebook to analyze and explore the results for both methods. An additional app, AnndataReportApp, can be run on any cell-count matrix and summarize simple QC metrics, visualizations and preliminary DEG analysis. These apps can be used on any Space Ranger output by the click of a button. At the same time, they are still very close to the pipelines presented in the corresponding papers. This allows any user who knows the pipeline to use it without requiring knowledge about this work (although recommended). This acts as a contribution to reliable, reproducible and persistent data analysis in research.

4.4 General Guidelines for Usage of Bin-to-Cell Methods

Regardless of the chosen method, we emphasize the critical importance of running multiple pipelines using different parameters and comparing results. The approach by Bin2Cell and ENACT (and also Space Ranger v4) of using nucleus segmentation and subsequent circular expansion implies important assumptions on the data. The substantial differences observed between results across our evaluation metrics demonstrate that parameter choice and disregard of assumptions can significantly impact conclusions. We summarize the most important keypoints:

- Ensure tissue characteristics align with method assumptions. The evaluated bin-to-cell methods are unsuitable for multinucleated cells (e.g., hepatocytes), anucleated cells (e.g., red blood cells), cells with non-circular morphologies (e.g., neurons), or tissues with extreme size heterogeneity (e.g., lymphocytes and megakaryocytes in bones). Additionally, these methods assume centrally located nuclei, which may be violated in tissues such as skeletal muscle.
- Explore different parameter combinations, check the segmentation outputs and adjust the parameters to obtain best matching segmentation results (detailed parameter guidance provided in Appendix A).
- For resource-limited environments, Bin2Cell provides the most efficient solution while maintaining satisfactory performance across evaluation metrics.
- Accurate segmentation is especially important in tissues with heterogeneous cell type distribution e.g. immune cells in tumors.
- Select annotation methods capable of detecting target cell populations, incorporating appropriate single-cell reference datasets when necessary. When available, confidence scores should be maximized through method and parameter optimization.
- Check the transcripts per cell distribution, typically the higher the better. Also make sure the marker genes corresponding to cell types of interest are found in the data and plot their spatial locations to ensure they cluster in the expected regions.
- Most analysis tools were developed for single-cell RNA sequencing data. Machine learning-based approaches may show reduced performance when applied to spatial transcriptomics data, necessitating cautious interpretation and preference for models trained on SRT data if available.

4.5 Conclusion

This comprehensive evaluation of bin-to-cell methods for Visium HD data provides crucial insights for the Spatially Resolved Transcriptomics community. Our analysis demonstrates that bin-to-cell methods can improve results compared to simple binning of squares. Within bin-to-cell methods, we saw similar results for Bin2Cell and ENACT. The variation in results depends rather on parameter choice, data characteristics and choice of downstream analysis method than on bin-to-cell method choice. Xseg overcomes limitations in segmentation quality but is not feasible to become a standard tool. Therefore, no universally superior method can be defined. Instead we reveal important insights on prior assumptions on the data, computational efficiency and practical applicability that should be considered when selecting appropriate methods for specific research questions.

Bin2Cell emerges as the best choice for resource-constrained environments and is easy to interpret because of its minimal complexity. The results are consistent with the results from the other methods and provide satisfactory performance across most evaluation metrics. The method's unique ability to utilize GEX representation can provide advantages when nuclear staining quality is suboptimal or when nuclei are absent from certain tissue regions, however it can also add noise to the results. This again highlights the importance of visual inspection and validation of results.

ENACT demonstrates sophisticated functionality with extensive parameter control, making it well-suited for detailed exploratory analyses where computational resources are available and methodological flexibility is prioritized. However, users must understand of the method's assumptions and parameter impacts. Particularly the 'blurring' effect emerging from splitting transcript counts across cells can compromise the results of downstream analyses.

Xenium segmentation achieves superior specificity when available, offering the most accurate cell boundary detection particularly beneficial for tissues with complex morphologies that cannot be adequately approximated by circular expansion approaches. However the performance improvements do not outweigh the increase in cost related to obtain such a segmentation.

We have seen that the nucleus segmentation with circular expansion can achieve satisfactory results and therefore offers a convenient and efficient solution. We also summarized cases in which such methods fail to capture the complexity of the problem. Cell segmentation methods on H&E images are still expected to improve which could add great value to bin-to-cell methods.

An important insight concerns the inherent sensitivity-specificity trade-offs present in all bin-to-cell methods. The coexpression-based metrics demonstrate that, for bin-to-cell methods, optimization for one aspect comes at the expense of the other. Balancing these factors according to the specific research objective is therefore required. The expansion parameter for both ENACT and Bin2Cell offers an option to explore this trade-off and find the optimal results. The MECR curves illustrate that optimal performance requires careful consideration of both transcript capture rates and segmentation accuracy, with Xenium segmentation (using 'Weighted-by-Transcript') achieving the best overall compromise when both metrics are considered simultaneously.

Our cell type annotation revealed significant challenges when applying scRNA-seq-

trained models to Spatial Transcriptomics data. This highlights fundamental differences between these data modalities, especially in sensitivity. The low confidence scores obtained for the Lung Cancer outputs underscore the importance of tissue-model matching and data quality considerations. The largest differences were observed between annotation methods, mostly in cell type proportion but also in spatial organization. While the variability across bin-to-cell methods, even for the binned outputs, was low. The results were therefore highly dependent of the annotation method chosen, rather than the bin-to-cell method. These findings suggest that automated cell type annotation on Spatially Resolved Transcriptomics data requires more sophisticated approaches.

The systematic evaluation framework developed in this thesis establishes important quality control standards for bin-to-cell method assessment. The integration of general QC metrics, segmentation-based metrics, annotation-based metrics, and coexpression-based scores provides an approach to method validation and comparison. It also summarizes metrics to assess data quality and confirm compliance with assumptions, crucial for downstream analysis. This work establishes a robust framework for informed method selection based on project requirements while also providing quality control metrics for result assessment.

The rapid evolution of Spatial Transcriptomics technologies and analysis methods requires continuous evaluation and adaptation of bin-to-cell approaches as new tools and methods emerge. The evaluation metrics used in this thesis provide a solid foundation for systematic assessment of emerging methods and technologies.

Ultimately, this work contributes to Spatially Resolved Transcriptomics as a field by providing benchmarking for method selection, establishing quality control standards, and facilitating broader access to analysis tools by practical implementation. The insights gained from this evaluation inform the practical application of Spatial Transcriptomics approaches at FGCZ.

Bibliography

[1] "Method of the Year 2020: spatially resolved transcriptomics," Jan. 2021. [Online]. Available: https://www.nature.com/collections/dfibfggefc

[2] G. Efthymiou, A. Saint, M. Ruff, Z. Rekad, D. Ciais, and E. Van Obberghen-Schilling, "Shaping Up the Tumor Microenvironment With Cellular Fibronectin," *Frontiers in Oncology*, vol. 10, p. 641, 2020.

[3] L. L. van der Woude, M. A. J. Gorris, A. Halilovic, C. G. Figdor, and I. J. M. de Vries, "Migrating into the Tumor: a Roadmap for T Cells," *Trends in Cancer*, vol. 3, no. 11, Nov. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405803317301929

[4] e. a. Madissoon E, "A spatially resolved atlas of the human lung characterizes a gland-associated immune niche," *Nature Genetics*, vol. 55, no. 1, pp. 66–77, Jan. 2023, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41588-022-01243-4

[5] T. Lohoff, S. Ghazanfar, A. Missarova, N. Koulena, N. Pierson, J. A. Griffiths, E. S. Bardot, C.-H. L. Eng, R. C. V. Tyser, R. Argelaguet, C. Guibentif, S. Srinivas, J. Briscoe, B. D. Simons, A.-K. Hadjantonakis, B. Göttgens, W. Reik, J. Nichols, L. Cai, and J. C. Marioni, "Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis," *Nature Biotechnology*, vol. 40, no. 1, pp. 74–85, Jan. 2022, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41587-021-01006-2

[6] Z. He, A. Maynard, A. Jain, T. Gerber, R. Petri, H.-C. Lin, M. Santel, K. Ly, J.-S. Dupré, L. Sidow, F. Sanchis Calleja, S. M. J. Jansen, S. Riesenberg, J. G. Camp, and B. Treutlein, "Lineage recording in human cerebral organoids," *Nature Methods*, vol. 19, no. 1, pp. 90–99, Jan. 2022, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41592-021-01344-8

[7] M. Cheng, Y. Jiang, J. Xu, A.-F. A. Mentis, S. Wang, H. Zheng, S. K. Sahu, L. Liu, and X. Xu, "Spatially resolved transcriptomics: a comprehensive review of their technological advances, applications, and challenges," *Journal of Genetics and Genomics*, vol. 50, no. 9, pp. 625–640, Sep. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1673852723000759

[8] M. N. et al., "Visium HD enables whole transcriptome spatial profiling at single cell scale resolution in FFPE tissues." [Online]. Available: https://www.10xgenomics.com/library/7db6a2

[9] 10x Genomics, "10x Genomics Launches 5,000-Plex Gene Panel for Xenium." [Online]. Available: https://www.prnewswire.com/news-releases/10x-genomics-launches-5-000-plex-gene-panel-for-xenium-302158495.html

[10] J. Azevedo, "Optimizing your spatial transcriptomics research with Visium HD and Xenium Prime 5K." [Online]. Available: https://www.10xgenomics.com/blog/optimizing-your-spatial-transcriptomics-research-with-visium-hd-and-xenium-in-situ

[11] O. Habern, "Your introduction to Visium HD: Spatial biology in high definition." [Online]. Available: https://www.10xgenomics.com/blog/your-introduction-to-visium-hd-spatial-biology-in-high-definition

[12] 10x Genomics, "10x genomics space ranger v3.1.1," https://github.com/10XGenomics/spaceranger, 2025, version 3.1.1. A set of analysis pipelines for processing Visium spatial transcriptomics data.

[13] 10x Genomics, "Xenium in situ multimodal cell segmentation: Workflow and data highlights," https://cdn.10xgenomics.com/image/upload/v1710785020/CG000750_XeniumInSitu_CellSegmentation_TechNote_RevA.pdf, 2024, document Number: CG000750.

[14] K. Polański, R. Bartolomé-Casado, I. Sarropoulos, C. Xu, N. England, F. L. Jahnsen, S. A. Teichmann, and N. Yayon, "Supplementary material: Bin2cell reconstructs cells from high resolution visium hd data," *Bioinformatics*, vol. 40, no. 9, p. 4, 2024.

[15] K. Polański, R. Bartolomé-Casado, I. Sarropoulos, C. Xu, N. England, F. L. Jahnsen, S. A. Teichmann, and N. Yayon, "Bin2cell reconstructs cells from high resolution visium hd data," *Bioinformatics*, vol. 40, no. 9, p. btae546, 2024.

[16] M. e. a. Kamel, "ENACT: End-to-End Analysis of Visium High Definition (HD) Data," *Bioinformatics*, vol. 41, no. 3, p. btaf094, Mar. 2025. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaf094

[17] X. Genomics, "Space Ranger." [Online]. Available: https://www.10xgenomics.com/support/software/space-ranger/latest

[18] "Nuclei Segmentation and Custom Binning of Visium HD Gene Expression Data." [Online]. Available: https://www.10xgenomics.com/analysis-guides/segmentation-visium-hd

[19] V. Petukhov, R. J. Xu, R. A. Soldatov, P. Cadinu, K. Khodosevich, J. R. Moffitt, and P. V. Kharchenko, "Cell segmentation in imaging-based spatial transcriptomics," *Nature Biotechnology*, vol. 40, no. 3, pp. 345–354, Mar. 2022, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41587-021-01044-w

[20] D. C. Jones, "dcjones/SpuriousCoexpression.jl," Jun. 2025, original-date: 2024-10-08T00:17:31Z. [Online]. Available: https://github.com/dcjones/SpuriousCoexpression.jl

[21] L. C. Gaspard-Boulinc, L. Gortana, T. Walter, E. Barillot, and F. M. G. Cavalli, "Cell-type deconvolution methods for spatial transcriptomics," *Nature Reviews Genetics*, pp. 1–19, May 2025, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41576-025-00845-y

[22] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, no. 1, pp. 100–106, Jan. 2021, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41592-020-01018-x

[23] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, "Cell detection with star-convex polygons," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, 2018, pp. 265–273.

[24] N. F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, T. Dougherty, C. C. Fullaway, B. J. McIntosh, K. X. Leow, M. S. Schwartz, C. Pavelchek, S. Cui, I. Camplisson, O. Bar-Tal, J. Singh, M. Fong, G. Chaudhry, Z. Abraham, J. Moseley, S. Warshawsky, E. Soon, S. Greenbaum, T. Risom, T. Hollmann, S. C. Bendall, L. Keren, W. Graf, M. Angelo, and D. Van Valen, "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning," *Nature Biotechnology*, vol. 40, no. 4, pp. 555–565, Apr. 2022, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41587-021-01094-0

[25] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, "Cell detection with star-convex polygons," in *Cell Detection with Star-convex Polygons*, 2018, pp. 265–273, iSSN: 0302-9743, 1611-3349 arXiv:1806.03535 [cs]. [Online]. Available: http://arxiv.org/abs/1806.03535

[26] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, and R. Satija, "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, no. 13, pp. 3573–3587.e29, Jun. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867421005833

[27] R. et al., "The Human Cell Atlas White Paper," Oct. 2018, arXiv:1810.05192 [q-bio]. [Online]. Available: http://arxiv.org/abs/1810.05192

[28] 10x Genomics, "Post-xenium in situ applications: Immunofluorescence, h&e, visium v2, and visium hd technical note," https://cdn.10xgenomics.com/image/upload/v1726606356/support-documents/CG000709_Post-Xenium_TechnicalNote_RevC.pdf, 2024, document Number CG000709, Rev C.

[29] 10x Genomics, "Visium hd cytassist gene expression human lung cancer (post xenium experiment)," https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-human-lung-cancer-post-xenium-expt, 2024, spatial transcriptomics dataset generated using Visium HD CytAssist on human lung cancer tissue post Xenium analysis.

[30] A. Janesick, R. Shelansky, A. D. Gottscho, F. Wagner, S. R. Williams, M. Rouault, G. Beliakoff, C. A. Morrison, M. F. Oliveira, J. T. Sicherman, A. Kohlway, J. Abousoud, T. Y. Drennon, S. H. Mohabbat, S. E. B. Taylor, and 10x Development Teams, "High resolution mapping of the tumor

microenvironment using integrated single-cell, spatial and in situ analysis," *Nature Communications*, vol. 14, no. 1, p. 8353, Dec. 2023. [Online]. Available: https://doi.org/10.1038/s41467-023-43458-x

[31] X. Genomics, "Visium HD 3' Gene Expression Library, Rhesus Macaque Kidney (Fresh Frozen)." [Online]. Available: https://www.10xgenomics.com/datasets/visium-hd-three-prime-rhesus-kidney-fresh-frozen

[32] M. Tyler, A. Gavish, C. Barbolin, R. Tschernichovsky, R. Hoefflin, M. Mints, S. V. Puram, and I. Tirosh, "The Curated Cancer Cell Atlas provides a comprehensive characterization of tumors at single-cell resolution," *Nature Cancer*, vol. 6, no. 6, pp. 1088–1101, Jun. 2025, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s43018-025-00957-8

[33] P. Bischoff, A. Trinks, B. Obermayer, J. P. Pett, J. Wiederspahn, F. Uhlitz, X. Liang, A. Lehmann, P. Jurmeister, A. Elsner, T. Dziodzio, J.-C. Rückert, J. Neudecker, C. Falk, D. Beule, C. Sers, M. Morkel, D. Horst, N. Blüthgen, and F. Klauschen, "Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma," *Oncogene*, vol. 40, no. 50, pp. 6748–6758, Dec. 2021, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41388-021-02054-3

[34] M. F. d. Oliveira, J. P. Romero, M. Chung, S. R. Williams, A. D. Gottscho, A. Gupta, S. E. Pilipauskas, S. Mohabbat, N. Raman, D. J. Sukovich, D. M. Patterson, and S. E. B. Taylor, "High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer," *Nature Genetics*, vol. 57, no. 6, pp. 1512–1523, Jun. 2025, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41588-025-02193-3

[35] N. Pielawski, A. Andersson, C. Avenel, A. Behanova, E. Chelebian, A. Klemm, F. Nysjö, L. Solorzano, and C. Wählby, "TissUUmaps 3: Improvements in interactive visualization, exploration, and quality assessment of large-scale spatial omics data," *Heliyon*, vol. 9, no. 5, p. e15306, May 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844023025136

[36] A. Gribov, M. Sill, S. Lück, F. Rücker, K. Döhner, L. Bullinger, A. Benner, and A. Unwin, "SEURAT: visual analytics for the integrated analysis of microarray data," *BMC medical genomics*, vol. 3, p. 21, Jun. 2010.

[37] I. Virshup, S. Rybakov, F. J. Theis, P. Angerer, and F. A. Wolf, "anndata: Annotated data," Dec. 2021. [Online]. Available: https://www.biorxiv.org/content/10.1101/2021.12.16.473007v1

[38] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, no. 1, p. 15, Feb. 2018. [Online]. Available: https://doi.org/10.1186/s13059-017-1382-0

[39] G. Palla, H. Spitzer, M. Klein, D. Fischer, A. C. Schaar, L. B. Kuemmerle, S. Rybakov, I. L. Ibarra, O. Holmberg, I. Virshup, M. Lotfollahi, S. Richter, and F. J. Theis, "Squidpy: a scalable framework for spatial omics analysis," *Nature Methods*, vol. 19, no. 2, pp. 171–178, Feb. 2022, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41592-021-01358-2

[40] X. Genomics, "Understanding Xenium Outputs." [Online]. Available: https://www.10xgenomics.com/support/software/xenium-onboard-analysis/latest/analysis/xoa-output-understanding-outputs

[41] D. C. Jones, A. E. Elz, A. Hadadianpour, H. Ryu, D. R. Glass, and E. W. Newell, "Cell simulation as cell segmentation," *Nature Methods*, vol. 22, no. 6, pp. 1331–1342, Jun. 2025, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41592-025-02697-0

[42] M. Hatakeyama, L. Opitz, G. Russo, W. Qi, R. Schlapbach, and H. Rehrauer, "SUSHI: an exquisite recipe for fully documented, reproducible and reusable NGS data analysis," *BMC Bioinformatics*, vol. 17, no. 1, p. 228, Jun. 2016. [Online]. Available: https://doi.org/10.1186/s12859-016-1104-8

[43] A. Hartman and R. Satija, "Comparative analysis of multiplexed in situ gene expression profiling technologies," *bioRxiv: The Preprint Server for Biology*, Jan. 2024.

[44] Teichlab, "Bin2cell notebook demo," https://nbviewer.org/github/Teichlab/bin2cell/blob/main/notebooks/demo.ipynb, 2024, jupyter notebook demonstrating segmentation-bin2cell for VisiumHD data.

[45] Teichlab, "Bin2cell documentation," https://bin2cell.readthedocs.io/en/latest/index.html, 2024, readthedocs documentation of the bin2cell package.

Appendix A

Parameter Impact and Exploration

A.1 Bin2Cell

We summarize the most important parameters and the effect of varying them as described by the StarDist authors, in the Bin2Cell demo notebook [44] and documentation [45]

- 'nms_thresh': Non-maximum suppression threshold. Increasing 'nms_thresh' requires more of the putative objects to be overlapping for them to be merged into a single label, which may help a bit in denser regions.
- 'prob_thresh': For each proposed cell, StarDist will use a sigmoid activation to simulate an object probability output. The threshold will decide which cells are kept and which segmentations are rejected [23]. Lowering 'prob_thresh' to make the calls less stringent is recommended by the Bin2Cell authors (see example in fig. A.1). There is a prob thresh for the StarDist model used for the H&E image as well as one for the GEX image.
- 'mpp': (by default 0.5) This stands for microns per pixel and translates to how many micrometers are captured in each pixel of the input. The StarDist segmentation outputs will vary with this parameter. The model might find small structures as nucleoli. In a similar sense, if the number of pixels per nuclei is too low ("low resolution image") the model might find aggregates of cells or other structures as nuclei. StarDist was trained on images close to 0.3 um/px.
- 'buffer': (by default 150) The image will be cropped to the area overlapping the bins passed as input, plus a buffer of 'buffer' pixels on each side.
- 'max_bin_distance': (by default 2) Defines how far away from the nucleus a bin can be to be assigned to it.

For example, cell types that natively express low levels of RNA but are sparse would gain from more expansion. Similarly, cells that express high levels of RNA and are in a dense environment would gain from less/no expansion.(B2C supplementary mat) - maybe add to guidelines.

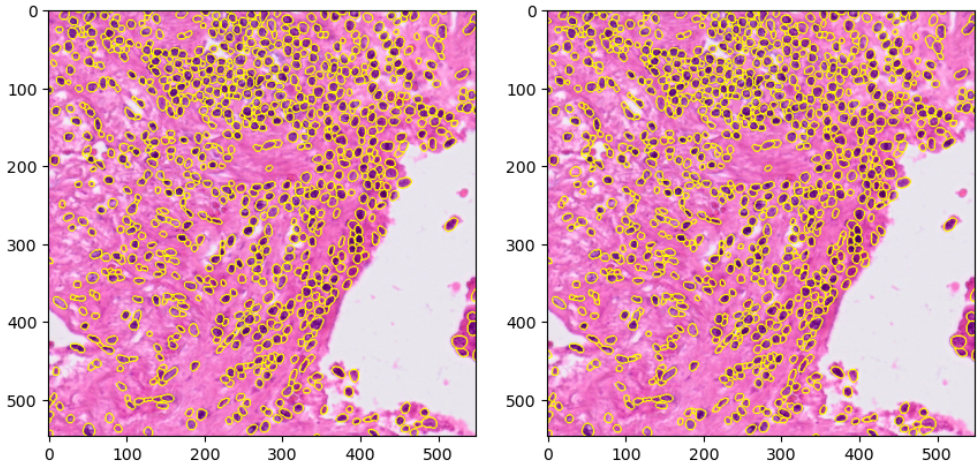

Figure A.1: Nucleus segmentation output after running StarDist with default parameters (left) and with reduced prob_thresh (right). Reducing the prob_thresh leads to more structures being detected as nuclei. Reducing it too much will lead to wrong calls by the model.

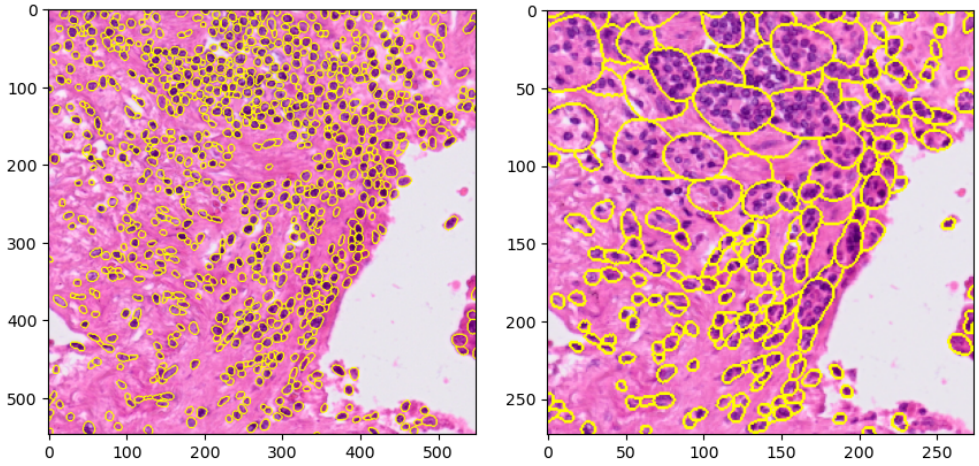

Figure A.2: Nucleus segmentation output after running StarDist with default parameters (left) and with increased mpp (right). Increasing the mpp changes the scale assumed by the model and can caus the model to cluster groups of nuclei which are tightly packed to one single nucleus.

These examples show how important an appropriate choice of segmentation parameters is. The segmentation results should always be verified to make sure the results are satisfying.

A.2 ENACT

Choosing the right expansion parameter is crucial when using bin-to-cell tools. We explored different expansion parameters for the ENACT pipeline and visualize the

effect on the results. In dark green, the nucleus segmentation is shown, the expanded segmentations corresponding to 1,2 and 3 bin sizes are also shown for comparison.

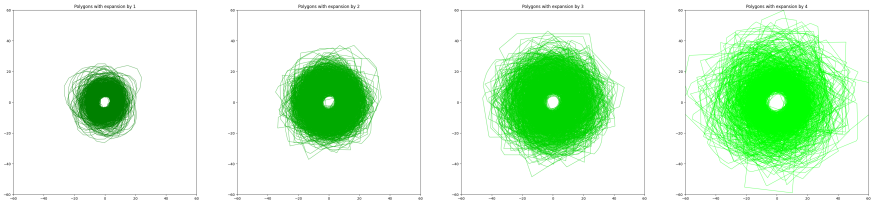

Figure A.3: ENACT segmentation shapes for different expansion parameters.

Showing the various segmentations over the corresponding tissue highlights the effects of circular expansion. Small cells with no neighboring nuclei will expand into empty space. On one hand, including noise from diffused transcripts, on the other corrupting cell size and transcript density metrics. Tightly packed cells do not expand much and expand into space-filling shapes with straight line borders and corners. Both circular and polygonal contours do not reflect typical cell shapes.

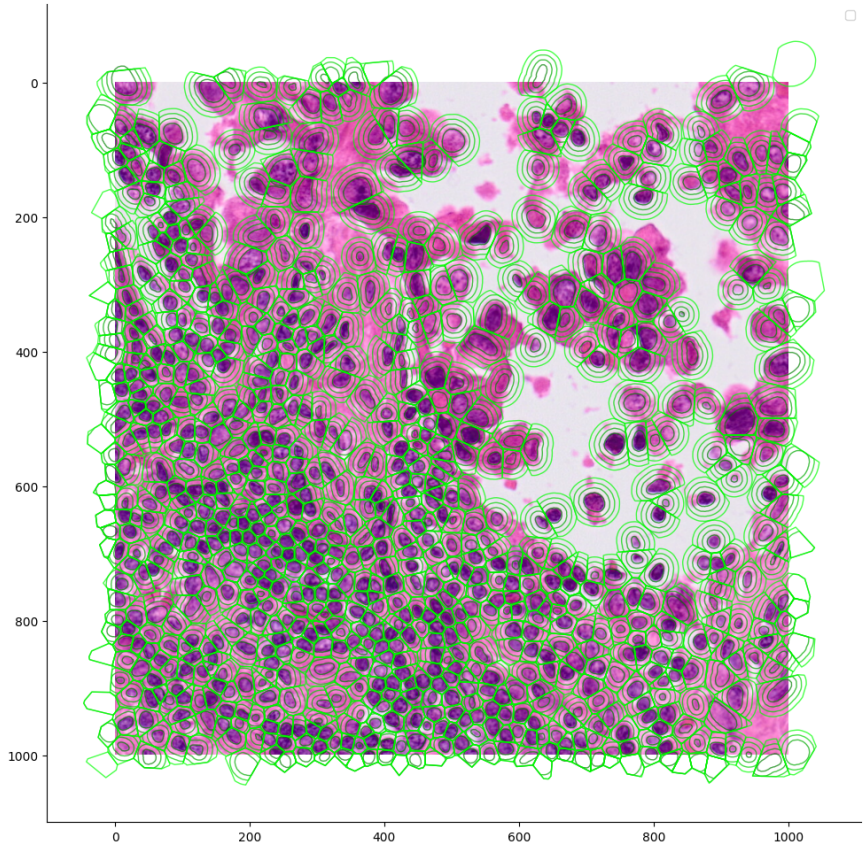

Figure A.4: Crop of tissue region of the Lung Cancer sample with ENACT segmentations corresponding to different expansion parameters. Nuclei in dark green and expansion by equivalent of 1,2 or 3 bins. Axes are ticked with relative pixel coordinates (pixel size: 0.212μm).

Appendix B

Space Ranger v4 Outputs on the Rhesus Macaque Dataset.

Some results on general metrics were included into the main thesis results. We provide some further familiar plots obtained using the Space Ranger v4 segmentation outputs. SpaceRanger v4 assigns the squares to cells directly from the nucleus segmentation, similar to Bin2Cell. Additionally they generate a polygon representation of the aggregated squares which represents the cell segmentation and allows visualization. We ran the segmentation metrics pipeline analogous to the protocol for the Lung Cancer tissue.

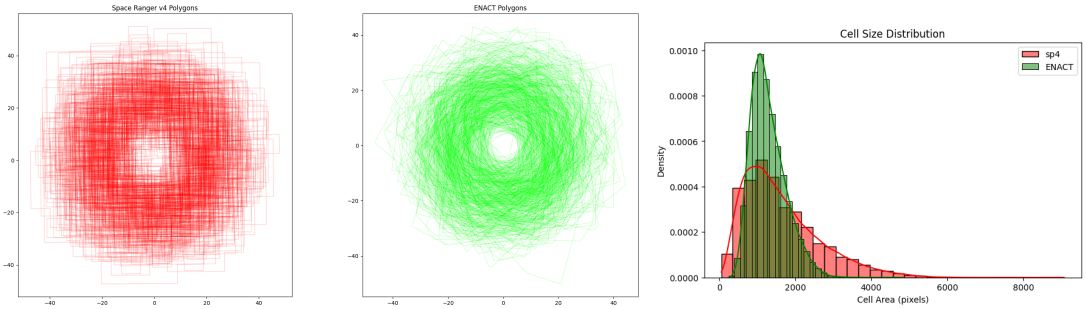

Figure B.1: Comparison metrics for the segmentations by Space Ranger v4 (red) and ENACT on a crop of the Rhesus Macaque kidney dataset. Shapes are compared on the left and area distribution on the right.

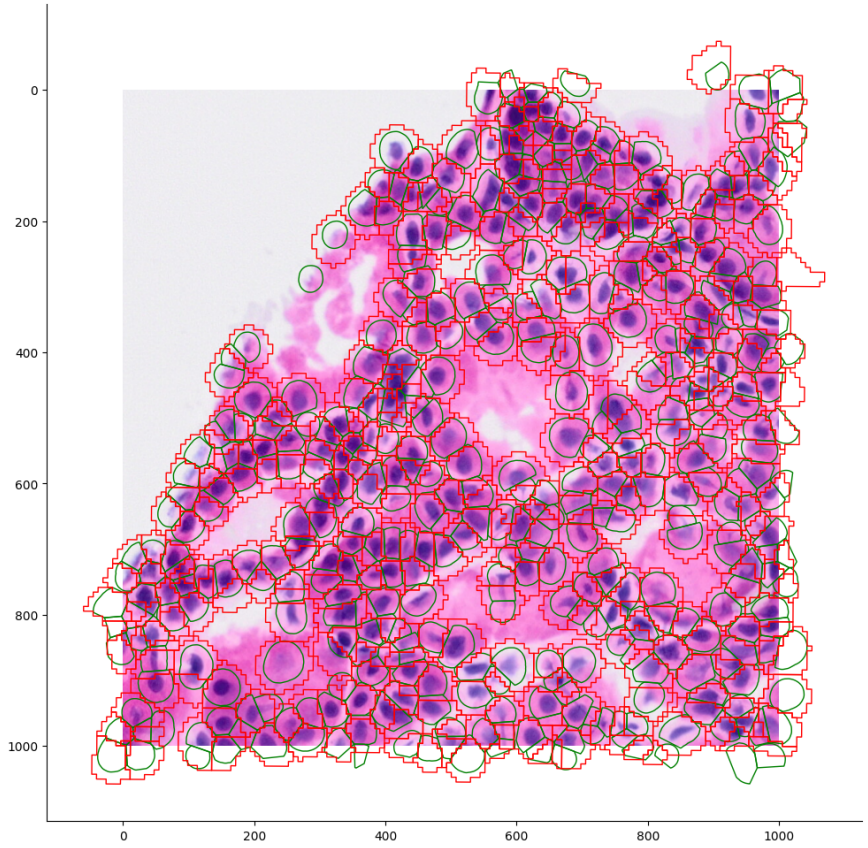

Figure B.2: Segmentation outputs on a crop of the Rhesus Macaque kidney dataset. Space Ranger v4 segmentation is shown in red and ENACT in green. Axes are ticked with relative pixel coordinates (pixel size: 0.212μm).

Appendix C

Summary Tables

Aspect	8×8 bins	Bin2Cell	ENACT	Xenium Segmentation (Xseg)	Space Ranger v4
Purpose	Increase signal per bin	Segmentation and bin-to-cell assignment	Segmentation, bin-to-cell assignment using different methods and cell type assignment	Custom methods to assess importance of accurate segmentation	Cell segmentation and direct transcript assignment from reads
Input data	Cell-Count matrix	H&E image and Cell-Count matrix	H&E image and Cell-Count matrix	H&E image, Cell-Count matrix and coordinate-transformed Xenium segmentation	H&E image, FASTQ files, and slide metadata
Image used	None	H&E and GEX image	H&E	Xenium	H&E
Segmentation algorithm	None	StarDist (H&E and Fluoro)	StarDist (H&E)	Xenium: using boundary, interior and nucleus staining	Custom StarDist (trained on proprietary data)
Bin-to-cell assignment	Naive 16 to 1	Nucleus to bin, then bin to neighbors	First nucleus expansion then 1 of 4 methods	Used ENACT	Based on proximity to the nearest nucleus

Table C.1: Table of Bin-to-Cell Methods Including Space Ranger v4.

Tool	Attributes	Requires Gene Markers	Advantage	Disadvantage	Notes
CellAssign	Probabilistic model	Yes	Custom "model" design via marker genes Probabilistic confidence scores	Requires well-defined marker genes (biological knowledge)	Good for supervised annotation when marker genes are available. Written in R.
CellTypist	Logistic regression model Optional majority voting	No	Pretrained models available Fast inference Can train custom models Confidence scores available	Confidence Scores not present in ENACT output.	Input: normalized and log-transformed expression matrix.
Sargent	Scoring-based method using cell-type-specific gene markers	Yes	Tailored for spatial data like Visium HD Handles low transcript counts well Supports both positive and negative markers	Integration to ENACT not available yet	Single-cell-based algorithm (cells do not compete) no transformation or clustering required
Azimuth Annotation	Integrated pipeline: normalization, visualization, annotation, DEG analysis	No	Uses reference atlases robust across datasets includes QC and visualization steps	May not generalize well to cancer datasets (excluded from training)	Panhuman model
Manual: UMAP + Leiden + DGE	State of the art Cell type assignment. e.g. used in Proseg paper[41]	No	High interpretability Often used as gold standard for validation	Labor-intensive subjective less scalable	Good when gene panels are small or well known.

Table C.2: Table of cell type assignment methods.

Appendix D

CellTypist Confidence Scores

We show the spatial distribution of confidence scores for different bin-to-cell methods.

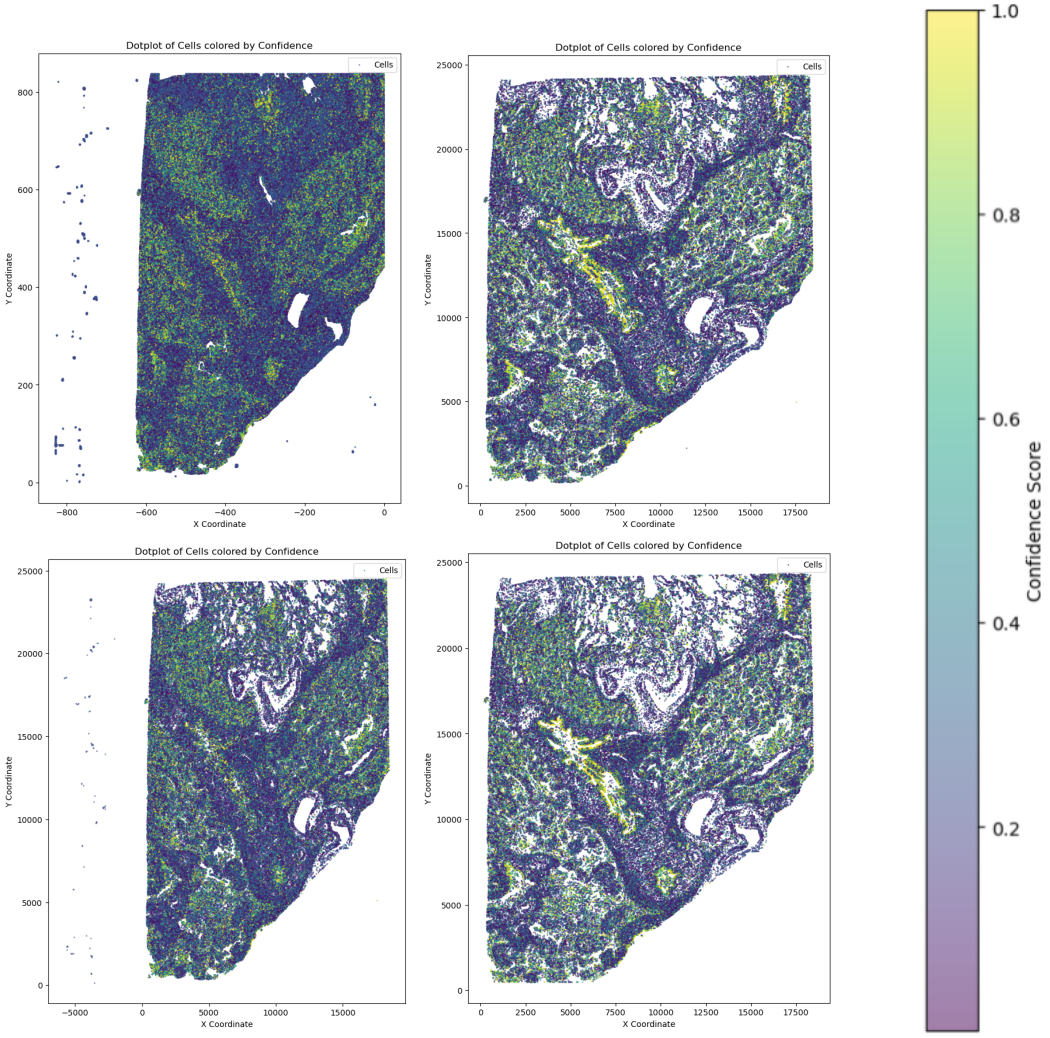

Figure D.1: Spatial locations of the cells colored by confidence score of CellTypist annotation. 8x8 Bins (tl), Bin2Cell (bl), ENACT (tr), Xenium segmentation (br).

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. **In consultation with the supervisor**, one of the following two options must be selected:

☐ I hereby declare that I authored the work in question independently, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

☒ I hereby declare that I authored the work in question independently. In doing so I only used the authorised aids, which included suggestions from the supervisor regarding language and content and generative artificial intelligence technologies. The use of the latter and the respective source declarations proceeded in consultation with the supervisor.

Title of paper or thesis:

Benchmarking Cell Segmentation Approaches for High-Resolution Spatial Transcriptomics

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
De Gottardi	Raphael

With my signature I confirm the following:
- I have adhered to the rules set out in the Citation Guidelines.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Zurich, 14. August 2025	R. De Gottardi

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] For further information please consult the ETH Zurich websites, e.g. https://ethz.ch/en/the-eth-zurich/education/ai-in-education.html and https://library.ethz.ch/en/researching-and-publishing/scientific-writing-at-eth-zurich.html (subject to change).