Downloaded from orbit.dtu.dk on: May 29, 2024

DTU Library

Development and Optimization of tools for High-quality Spatial Single-cell Gene Expression Profiling using High-resolution Spatial Transcriptomics Technology

Li, Mei

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Li, M. (2023). *Development and Optimization of tools for High-quality Spatial Single-cell Gene Expression Profiling using High-resolution Spatial Transcriptomics Technology*. DTU Bioengineering.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Development and Optimization of tools for High-quality Spatial Single-cell Gene Expression Profiling using High-resolution Spatial Transcriptomics Technology

Ph.D. thesis

Mei Li

Department of Biotechnology and Biomedicine (DTU Bioengineering)

Technical University of Denmark

November 2023

SUPERVISORS AND FUNDING

Principal supervisor

Professor Susanne Brix Pedersen, Disease Systems Immunology group, Section for Protein Science and Biotherapeutics, Department of Biotechnology and Biomedicine (DTU Bioengineering), Technical University of Denmark.

Co-supervisor

Professor Yong Hou, European Research Institute, BGI-Research, and the School of Life Sciences, University of Chinese Academy of Sciences.

Funding

The work in this PhD thesis was supported by the BGI-DTU PhD program between Department of Biotechnology and Biomedicine (DTU Bioengineering), Technical University of Denmark, and BGI College, Shenzhen, China.

Projects included in this thesis were supported by grants from: National Key R&D Program of China (2022YFC3400400).

ACKNOWLEDGEMENT

Firstly, I would like to express my heartfelt gratitude to my principal supervisor, Professor Susanne Brix Pedersen, and co-supervisor, Professor Yong Hou, for providing me the opportunity to pursue my PhD at DTU. I am extremely grateful to Professor Susanne Brix for her guidance and support during my doctoral studies. Academically, I have gained a wealth of knowledge in immunology and bioinformatics through her guidance and discussions on various projects. More importantly, I have learned the importance of rigorous academic integrity from them, which will benefit my future career and entire life.

I am also immensely thankful to Rasmus Ibsen Delhi, Carsten Eriksen, Lisbeth Buus, Ellen Magdalena Staudinger, Tommaso Del Buono D'Ondes and other members of the Disease Systems Immunology group. They gave me a lot of help in the PhD study program, courses, and acclimating campus life at the beginning of my doctoral journey. Their unwavering support and camaraderie have made for many cherished memories, often shared during our delightful lunch breaks in the yard of Building 224, which brims with knowledge and innovative ideas.

Further, I would like to thank the BGI-DTU PhD program for providing the PhD opportunity for industrial employees. I would also like to acknowledge my colleagues in BGI-Research, Yong Zhang, Yuxiang Li, and my lovely team members, for their support and helping with the data analysis.

Lastly, I would like to express my deepest appreciation to my family and friends for their unwavering encouragement. I am especially grateful to my husband, Yangbao Liu, for his patience and full support during the writing of my thesis. I felt sorry for the lack of accompany with my daughter during this demanding period. I would also like to give special thanks to my best friend, Qing Zhou, whose rich experience in academic writing has been immensely helpful to me.

TABLE OF CONTENTS

ABSTRACT

Spatially resolved technology is widely recognized as a cutting-edge technology in life sciences. It is increasingly utilized in various areas such as organ development, organism growth, tumor heterogeneity and evolution, as well as clinical translational research. Higher spatial resolution typically implies smaller molecular quantities, whereas observing whole tissues requires a larger field-of-view. The substantial progress in fundamental and translational research come with potential requirements in terms of higher resolution (*i.e.,* at single-cell level) and larger field-of-view, as well as the urgent need for tools to analyze the raw data. To address the data analysis challenges posed by large field-of-view and high resolution spatially resolved technology, this doctoral project is based on Stereo-seq, and aims to provide analysis methods and tools for obtaining high-quality spatial single-cell data, thereby facilitating the application of spatially resolved technology.

In the first study, we developed a framework called StereoCell for high-resolution and large field-of-view spatial transcriptomic data analysis. StereoCell provided a comprehensive and systematic platform for generating high-confidence single-cell spatial data, including image stitching, registration, nuclei segmentation, and molecule labeling. By utilizing better-performing algorithms during image stitching and molecule labeling, StereoCell reduced stitching error and time, and improved the signal-to-noise ratio of single-cell gene expression data compared to existing methods. These improvements were validated in mouse brain tissue and results confirmed that StereoCell produced highly accurate spatial single-cell gene expression profiles, thus facilitating clustering and cellular annotation within the biological tissue.

With recent advancements in Stereo-seq technology, it is now possible to acquire cell boundary information, such as cell membrane/wall staining images. In the second study,

we took advantage of this progress, and updated StereoCell to a new version, STCellbin, which used nuclei staining images as a bridge to procure cell membrane/wall staining images that align with spatial gene expression map. By employing a sophisticated cell segmentation technique, we obtained precise cell boundaries, thereby yielding more reliable profiles of single-cell spatial gene expression. STCellbin was utilized in mouse liver (cell membranes) and *Arabidopsis* seed (cell walls) datasets. This enhanced capability offered valuable insights into the spatial organization of gene expression within cells, contributing to a deeper understanding of tissue biology.

In the third study, we proposed an efficient and adaptive Gaussian smoothing (EAGS) imputation method for high-resolved spatial transcriptomics. The adaptive two-factor smoothing of EAGS created patterns based on the spatial expression information within single cells, as well as adaptive weights for the smoothing of cells in the same pattern, and then utilized the weights to restore the gene expression profiles. The performance efficiency of EAGS were assessed by using simulated and high-resolved spatial transcriptomic datasets of the mouse brain and olfactory bulb. Compared with other competitive methods, EAGS showed higher clustering accuracy, better biological interpretations, and significantly reduced resource consumption from computational processes.

PUBLICATIONS

Manuscripts included in the PhD thesis:

1. **Mei Li**[1,2,†], Huanlin Liu[1,†], Min Li[1,†], Shuangsang Fang[1,3,†], Qiang Kang[1,†], Jiajun Zhang[1,4], Fei Teng[1,4], Dan Wang[5,1], Weixuan Cen[1], Zepeng Li[1], Ning Feng[1], Jing Guo[1], Qiqi He[1], Leying Wang[1], Tiantong Zheng[1], Shengkang Li[1,6], Yinqi Bai[1], Min Xie[4], Yong Bai[1], Sha Liao[1,4], Ao Chen[1,4], Susanne Brix[2,*], Xun Xu[1,7,*], Yong Zhang[1,6,7,*], Yuxiang Li[1,6,7,*]. StereoCell: a highly accurate single-cell gene expression processing software for high-resolution spatial transcriptomics (Under revision in Nature Protocols)
2. Bohan Zhang[1,2,†], **Mei Li**[1,3,†], Qiang Kang[1,†], Zhonghan Deng[1], Hua Qin[2], Kui Su[1], Xiuwen Feng[1], Lichuan Chen[1], Huanlin Liu[1], Shuangsang Fang[2], Yong Zhang[1], Yuxiang Li[1], Susanne Brix[3,*], Xun Xu[1,*]. Generating single-cell gene expression profiles for high-resolution spatial transcriptomics based on cell boundary images (To be submitted)
3. Tongxuan Lv[1,2,†], Ying Zhang[1,†], **Mei Li**[1,4,†], Qiang Kang[1,‡], Shuangsang Fang[1,3], Yong Zhang[1], Susanne Brix[4,*], Xun Xu[1,2,*]. EAGS: efficient and adaptive Gaussian smoothing applied to high-resolved spatial transcriptomics (GigaScience, accepted)

†, equal contributions

*, corresponding authors

Papers that I co-authored during my PhD program, but not included in the thesis:

1. Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, Xin Wang, Huifang Lu, Xi Chen, Xing Liu, Xin Huang, Zhao Li, Yan Hong, Yujia Jiang, Jian Peng, Shuai Liu, Mengzhe Shen, Chuanyu Liu, Quanshui Li, Yue Yuan, Xiaoyu Wei, Huiwen Zheng, Weimin Feng, Zhifeng Wang, Yang Liu, Zhaohui Wang, Yunzhi Yang, Haitao Xiang, Lei Han, Baoming Qin, Pengcheng Guo, Guangyao Lai, Pura Muñoz-Cánoves, Patrick

H. Maxwell, Jean Paul Thiery, Qing-Feng Wu, Fuxiang Zhao, Bichao Chen, **Mei Li,** Xi Dai, Shuai Wang, Haoyan Kuang, Junhou Hui, Liqun Wang, Ji-Feng Fei, Ou Wang, Xiaofeng Wei, Haorong Lu, Bo Wang, Shiping Liu, Ying Gu, Ming Ni, Wenwei Zhang, Feng Mu, Ye Yin, Huanming Yang, Michael Lisby, Richard J. Cornall, Jan Mulder, Mathias Uhle´ n, Miguel A. Esteban, Yuxiang Li, Longqi Liu, Xun Xu, and Jian Wang. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays, *Cell.* 2022 May 12;185(10):1777-1792.e21. https://doi.org/10.1016/j.cell.2022.04.003

2. Mengnan Cheng, Liang Wu, Lei Han, Xin Huang, Yiwei Lai, Jiangshan Xu, Shuai Wang, **Mei Li,** Huiwen Zheng, Weimin Feng, Zirui Huang, Yujia Jiang, Shijie Hao, Zhao Li, Xi Chen, Jian Peng, Pengcheng Guo, Xiao Zhang, Guangyao Lai, Qiuting Deng, Yue Yuan, Fangming Yang, Xiaoyu Wei, Sha Liao, Ao Chen, Giacomo Volpe, Miguel A Esteban, Yong Hou, Chuanyu Liu, Longqi Liu. A Cellular Resolution Spatial Transcriptomic Landscape of the Medial Structures in Postnatal Mouse Brain. *Front Cell Dev Biol*. 2022 May 17. 10:878346. https://doi.org/10.3389/fcell.2022.878346

3. Ao Chen, Yidi Sun, Ying Lei, Chao Li, Sha Liao, Juan Meng, Yiqin Bai, Zhen Liu, Zhifeng Liang, Zhiyong Zhu, Nini Yuan, Hao Yang, Zihan Wu, Feng Lin 1, Kexin Wang, **Mei Li**, Shuzhen Zhang, Meisong Yang, Tianyi Fei, Zhenkun Zhuang, Yiming Huang, Yong Zhang, Yuanfang Xu, Luman Cui, Ruiyi Zhang, Lei Han, Xing Sun, Bichao Chen, Wenjiao Li, Baoqian Huangfu, Kailong Ma, Jianyun Ma, Zhao Li, Yikun Lin, He Wang , Yanqing Zhong, Huifang Zhang, Qian Yu, Yaqian Wang, Xing Liu, Jian Peng 1, Chuanyu Liu, Wei Chen, Wentao Pan, Yingjie An, Shihui Xia, Yanbing Lu, Mingli Wang, Xinxiang Song, Shuai Liu, Zhifeng Wang, Chun Gong, Xin Huang, Yue Yuan, Yun Zhao, Qinwen Chai, Xing Tan, Jianfeng Liu , Mingyuan Zheng, Shengkang Li, Yaling Huang, Yan Hong, Zirui Huang, Min Li, Mengmeng Jin, Yan Li, Hui Zhang, Suhong Sun, Li Gao, Yinqi Bai, Mengnan Cheng, Guohai Hu, Shiping Li, Bo Wang, Bin Xiang, Shuting Li,

Huanhuan Li, Mengni Chen, Shiwen Wang, Minglong Li, Weibin Liu, Xin Liu, Qian Zhao, Michael Lisby, Jing Wang, Jiao Fang, Yun Lin, Qing Xie, Zhen Liu, Jie He, Huatai Xu, Wei Huang, Jan Mulder, Huanming Yang, Yangang Sun, Mathias Uhlen, Muming Poo, Jian Wang, Jianhua Yao, Wu Wei, Yuxiang Li, Zhiming Shen, Longqi Liu, Zhiyong Liu, Xun Xu, Chengyu Li. Single-cell spatial transcriptome reveals cell-type organization in the macaque cortex, *Cell.* 2023 August 17;186: 1-18. https://doi.org/10.1016/j.cell.2023.06.009

4. Xiaojie Qiu, Daniel Y. Zhu, Jiajun Yao, Zehua Jing, Lulu Zuo, Mingyue Wang, Kyung Hoi (Joseph) Min, Hailin Pan, Shuai Wang, Sha Liao, Yiwei Lai, Shijie Hao, Yuancheng Ryan Lu, Matthew Hill, Jorge D. Martin-Rufino, Chen Weng, Anna Maria Riera-Escandell, Mengnan Chen, Liang Wu, Yong Zhang, Xiaoyu Wei, **Mei Li,** Xin Huang, Rong Xiang, Zhuoxuan Yang, Chao Liu, Tianyi Xia, Yingxin Liang, Junqiang Xu, Qinan Hu, Yuhui Hu, Hongmei Zhu, Yuxiang Li, Ao Chen, Miguel A. Esteban, Ying Gu, Douglas A. Lauffenburger, Xun Xu, Longqi Liu, Jonathan S. Weissman, Shiping Liu, Yinqi Bai. Spateo: multidimensional spatiotemporal modeling of single-cell spatial transcriptomics. *bioRxiv*. 2022 December 11. https://doi.org/10.1101/2022.12.07.519417
5. Sha Liao, Yang Heng, Weiqing Liu, Jinqiong Xiang, Yong Ma, Ligen Chen, Xiuwen Feng, Dongmei Jia, Diyan Liang, Caili Huang, Jiajun Zhang, Min Jian, Kui Su, **Mei Li,** Yuin-Han Loh, Ao Chen, Xun Xu. Integrated Spatial Transcriptomic and Proteomic Analysis of Fresh Frozen Tissue Based on Stereo-seq. *bioRxiv*. 2023 April 28. https://doi.org/10.1101/2023.04.28.538364
6. Xing Liu, Chi Qu, Chuandong Liu, Na Zhu, Huaqiang Huang, Fei Teng, Caili Huang, Bingying Luo, Xuanzhu Liu, Yisong Xu, Min Xie, Feng Xi, **Mei Li,** Liang Wu, Yuxiang Li, Ao Chen, Xun Xu, Sha Liao, Jiajun Zhang. StereoSiTE: A framework to spatially and quantitatively profile the cellular neighborhood organized iTME. *bioRxiv.* 2023 May 21. https://doi.org/10.1101/2022.12.31.522366

7. Chun Gong, Shengkang Li, Leying Wang, Fuxiang Zhao, Shuangsang Fang, Dong Yuan, Zijian Zhao, Qiqi He, **Mei Li**, Weiqing Liu, Zhaoxun Li, Hongqing Xie, Sha Liao, Ao Chen, Yong Zhang, Yuxiang Li, Xun Xu. SAW: An efficient and accurate data analysis workflow for Stereo-seq spatial transcriptomics. *bioRxiv*. 2023 August 21. https://doi.org/10.1101/2023.08.20.554064
8. Chao Zhang, Qiang Kang, **Mei Li,** Hongqing Xie, Shuangsang Fang, Xun Xu. BatchEval Pipeline: Batch Effect Evaluation Workflow for Multiple Datasets Joint Analysis. *bioRxiv*. 2023 October 20. https://doi.org/10.1101/2023.10.08.561465 (Gigabyte, accepted)

ABBREVIATIONS

AI	artificial intelligence
cDNA	complementary DNA
cPAL	combinatorial probe anchor ligation
CFW	calcofluor white
CHI	Calinski-Harabasz Index
CPU	Central Processing Unit
CTA	cancer transcriptome atlas
DAPI	4,6-diamidino-2-phenylindole
DBiT-seq	deterministic barcoding in tissue for spatial omics sequencing
DBI	Davies-Bouldin Index
DDT	Distance Distribution Threshold
DNA	deoxyribonucleic acid
DNB	DNA nanoballs
EAGS	efficient and adaptive Gaussian smoothing
EASI FISH	Expansion-Assisted Iterative Fluorescence *In situ* Hybridization
EDA	exploratory data analysis
EEL-FISH	enhanced electronic FISH
ExM	expansion microscopy
FFT	Fast Fourier Transform
FFPE	formalin-fixed paraffin-embedded
FISH	fluorescence *in situ* hybridization
FISSEQ	fluorescence *in situ* sequencing
GB	GigaByte
Geo-seq	geographical position sequencing
GMM	Gaussian Mixture Model
GPR	Gaussian process regression

GPT	Generative Pre-trained Transformer
GPU	graphics processing unit
GUI	Graphical User Interface
HCR	hybridization chain reaction
HDMI	high-definition multimedia interface
HDST	high-definition spatial transcriptomics
HybISS	hybridization-based ISS
H&E	hematoxylin-eosin
ISH	*in situ* hybridization
ISS	*in situ* sequencing
LCM	laser capture microdissection
mIF	multiplex immunofluorescence
MERFISH	Multiplexed Error-Robust FISH
MFWS	multiple Fast Fourier Transform weighted stitching
MIDs	molecular identifiers
MIST	Microscopy Image Stitching Tool
MLCG	molecule labeling using the cell morphology and GMM
NGS	next-generation sequencing
NMF	non-negative matrix factorization
osmFISH	Ouroboros smFISH
PCR	polymerase chain reaction
Pixel-seq	Poly-indexed library-sequencing
PIC	photo-isolation chemistry
QC	Quality Control
RAM	random access memory
RCA	Rolling Circle Amplification
RCTD	Robust Cell Type Decomposition
RF	random forests

RNA-seq	RNA sequencing
ROI	region of interest
RT	reverse transcription
scRNA-seq	Single-cell RNA sequencing
seqFISH	sequential FISH
SEDAL	sequencing with error-reduction by dynamic annealing and ligation
smFISH	single molecule FISH
SOP	standard operating procedures
ssDNA	single strand DNA fluorescence
SRT	Spatially Resolved Transcriptomics
Stereo-seq	spatially enhanced resolution transcriptome sequencing
ST	Spatial transcriptomics
STARmap	spatially-resolved transcript amplicon readout mapping
3D	three dimensions
TIVA	transcriptome *in vivo* analysis
UMAP	uniform manifold approximation and projection.
UMI	unique molecular identifier

1 Introduction

1.1 Single-cell RNA sequencing and spatially resolved transcriptomics

Gene expression is a dynamic process with extensive spatiotemporal heterogeneity within tissues. At present there exists different transcriptomic technologies, including bulk RNA sequencing (RNA-seq)[1], Single-cell RNA sequencing (scRNA-seq)[2], and spatially resolved transcriptomics[3,4] that enable investigations of gene expression in cells and tissues. Traditional bulk RNA-seq captures the transcriptome in a large number of mixed cells, but it can only deliver the average expression level of genes within the sample, and ignores the heterogeneity of gene expression among different cells within a given cell population and between different cell types[1]. scRNA-seq advances our understanding of cellular gene expression to the single-cell level. It provides independent RNA expression profiles for each cell, enabling identification of gene expression differences between cells and the identification of rare cells within heterogeneous cell populations[5]. Although scRNA-seq greatly expands our knowledge of cellular heterogeneity, performing single-cell sequencing requires the dissociation of cells from tissues, leading to the loss of spatial information[3,6]. However, the spatial location of cells implies possible interactions between cells, and such interaction may be directly related to physiological and pathological functions of tissues. Therefore, it is necessary to link gene expression with spatial information to enhance our understanding of tissue function in health and disease[7]. Driven by this demand, Joakim Lundeberg's research group first proposed the concept of spatial transcriptomics in 2016 and published the first spatial transcriptomics technique based on *in situ* capturing of mRNAs[8]. Since then, a series of high-throughput *in situ* RNA detection techniques have been categorized as spatially resolved transcriptomics (SRT). Although these techniques have different principles, they all share a common feature, which is the recording of the spatial location information of detected molecules[9]. SRT can simultaneously obtain spatial information and gene expression data of cells, but

currently, there are only few techniques achieve single-cell resolution. Some techniques achieve high resolution, but the methods for obtaining data are complex, and the data quality is difficult to compare with mature single-cell RNA sequencing, which greatly limits its effectiveness. Just like single-cell sequencing has revolutionized various fields of biology, spatial resolved technology is widely recognized as a new frontier in the life sciences, and with the explosive innovation and growth of the technology, SRT is poised to usher in a new era of biological research and facilitate a more comprehensive understanding of the intricacies of living systems[10].

1.2 Categories of spatially resolved transcriptomics technologies

Investigations of tissues using two-dimensional (2D) or three-dimensional (3D) structural information are referred to as 'spatial biology', which has become the latest frontier of molecular biology[7]. Considering its immense importance and popularity, and potential to provide insights into currently unresolved research questions, spatially resolved transcriptomics (SRT) technology has been recognized as the method of the year in 2020 by Nature Methods[3]. It is worth noting that SRT methods generate data that allow us to identify specific functional regions within the whole genome range by deciphering the diversity of gene expression[11] and address cell type heterogeneity and intercellular communication through the analysis of intrinsic gene expression features and their physical proximity[12,13]. The powerful capabilities of SRT technology in decoding tissue complexity have helped us create maps of key biological processes, such as tissue and even organ development[14,15], as well as pathological mechanisms underlying several human diseases[16,17]. From the method of tissue molecular detection through *in situ* hybridization first reported in 1969[18] to the recent emergence of various SRT technologies based on next-generation sequencing (NGS) [8,19–23], SRT technology has undergone revolutionary developments due to advancements in sequencing technology and single-cell "omics". In terms of the way spatial information is obtained, current SRT technologies can be roughly categorized into four types: region of interest

(ROI) selection[24], *in situ* hybridization (ISH)[18], *in situ* sequencing (ISS)[25], and *in situ* spatial barcoding[8]. The developers of these technologies typically aim to achieve a combination of transcriptome-wide analysis, single-cell resolution, and high gene detection efficiency (**Figure 1**). Despite the growing feasibility of this endeavor, the characteristic of current-era technologies is to balance between these goals.

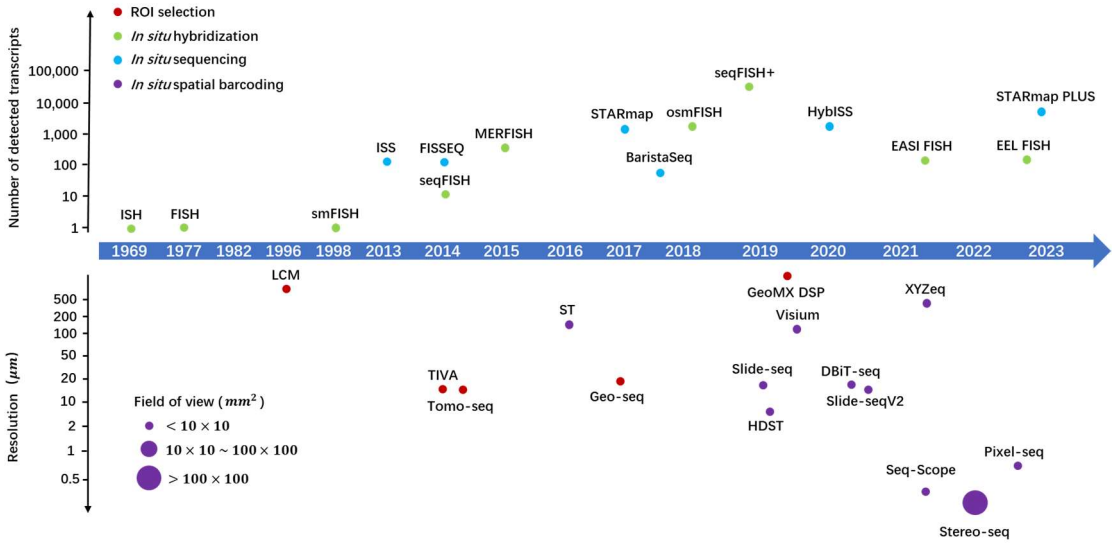

Figure 1: Timeline of the major SRT technologies. The approaches are categorized and distinguished by different colors (red, ROI selection method; green, *in situ* hybridization method; blue, *in situ* sequencing method; purple, *in situ* spatial barcoding method). The x-axes represent the timeline, and the y-axes are divided into two parts: the upper part represents the number of transcripts detected, while the bottom part represents the spatial resolution (in μm) in each of the technologies. The size of purple circles roughly indicates the field-of-view for these methods. Information about the listed methods are provided as follows: ISH[18], FISH[26], smFISH[27], ISS[25], FISSEQ[28], seqFISH[29], MERFISH[30], STARmap[31], BaristaSeq[32], osmFISH[33], seqFISH+[34], HybISS[35], EASI FISH[36], EEL FISH[37], STARmap PLUS[38], LCM[39], TIVA[40], Tomo-seq[41], ST[8], Geo-seq[19], Slide-seq[20], HDST[42], GeoMX DSP[43], Visium[44], DBiT-seq[21], Slide-seqV2[45], XYZeq[46], Seq-Scope[22], Stereo-seq[47], Pixel-seq[23].

1.2.1 ROI selection

Spatial localization could be obtained by selecting and isolating regions of interest (ROI) with known positions and shapes through physical and optical marking of ROIs. The separated ROIs can then be analyzed using complementary DNA (cDNA) microarrays or RNA-seq, or dissociated into single cell suspension for scRNA-seq. Laser capture

microdissection (LCM)[39] enables precise cutting and capturing of ROIs in tissue sections for various omics analyses, including RNA-seq, DNA-seq, epigenetic analysis, of which RNA-seq could obtain gene expression profiles with spatial information, even at the single-cell level. In Tomo-Seq[41], frozen tissues are sequentially sectioned and RNA-seq is performed for each section, resulting in a gene expression atlas coupled with spatial information. Geo-seq[19] involves cutting the tissue into small blocks, followed by RNA-seq of each block, to reconstruct a three-dimensional gene expression atlas using the positional information of the small tissue blocks. STRP-seq is an innovative physical microdissection method that slices adjacent tissue sections into thin strips at different angles and reconstructs the gene expression map in 3D using Tomographer[48]. In contrast, recent studies have utilized light to label ROIs instead of physical dissection. Transcriptome *in vivo* analysis tag (TIVA-tag) [40]and photo-isolation chemistry (PIC)[49] employ photosensitive groups to trigger reverse transcription (RT) *in vivo* or in fixed samples. Nanostring GeoMx[43] uses photocleavable groups similar to TIVA to release fluorescent combinatorial tags bound to detection probes (antibodies or hybridization probes), which are then quantified using Nanostring's nCounter technology. Due to the use of a predefined gene panel instead of poly-A capture, Nanostring provides cancer transcriptome atlas (CTA) gene panels with over 1,800 genes as well as human and mouse whole transcriptome panels with over 18,000 genes. Light can also trigger the attachment of DNA barcodes or barcode combinations, with Light-seq[50] and ZipSeq[51] using light labeling to assign region IDs through crosslinking or hybridization, followed by barcode reassignment of reads to these regions. All ROI selection techniques have their own advantages and limitations: they offer deep profiling approaching bulk sequencing levels, enable whole transcriptome detection, and allow for customization of ROIs from single cells to entire areas. However, their throughput is relatively low (typically less than a few hundred locations) as each selected region must be individually collected and processed.

1.2.2 *In situ* hybridization

The method based on *in situ* hybridization is a descendant of the single molecule FISH (smFISH)[27], which is a technique that decomposes individual mRNA molecules in tissues into sub-diffraction fluorescence spots. smFISH is often regarded as the "gold standard" in RNA quantification methods because it can detect low abundance transcripts with one copy per cell, thus spatial analysis techniques derived from it often have excellent sensitivity[9]. In sequential FISH (seqFISH)[29], fluorescently labeled FISH probes hybridize directly to cellular mRNA and after each cycle, mRNA is removed by deoxyribonucleic acid (DNase) digestion, keeping the mRNA in place. In another typical technique, Multiplexed Error-Robust FISH (MERFISH)[30], the probes include mRNA binding regions and a series of "reporter" regions corresponding to barcode elements, which are detected through secondary hybridization rounds. This effectively makes detection independent of the original RNA molecules, making the scheme more resistant to ribonuclease (RNase) contamination, and due to the removal of only the fluorophores but not the probes, the multiple rounds of hybridization in MERFISH take less time than in seqFISH, making it possible to image hundreds to thousands of targets. Another version of seqFISH[29], seqFISH+[34], follows the same method. Another element introduced by MERFISH and adopted in seqFISH+ is the encoding strategy in information theory to make the combinatorial barcode "error-robust", this mitigates the effects of hybridization failure or non-specific binding, which can be quite significant due to the periodicity of imaging. In recent years, both MERFISH and seqFISH+ have been extended to the whole transcriptome level, suitable for simultaneous detection of mRNA and proteins through oligo-conjugated antibodies, and enhanced by adding signal amplification strategies and tissue embedding and clearing.

Although seqFISH and MERFISH are currently the two main hybrid-based spatial transcriptomics methods available, there are also some other technologies that possess distinctive components. Ouroboros smFISH (osmFISH)[33] is a periodic smFISH method that lacks a barcode scheme and multiplexing capability (only detects a few transcripts

per cycle) to obtain a simpler protocol unaffected by transcript abundance or density[33]. Saber-FISH[52] and Clamp-FISH[53], based on different signal amplification methods, offer significantly improved signal-to-noise ratios compared to the "first-generation" protocols, but have not been shown to detect more than 100 transcripts. SCRINSHOT[54] is a similar method specifically designed to detect approximately 30 transcripts in formalin-fixed paraffin-embedded (FFPE) tissues. A recent approach called enhanced electronic FISH (EEL-FISH)[37] uses electrophoresis to drive cell mRNA onto the surface of a conductive glass slide. The tissue is then removed and mRNA is detected using multiplexed FISH. Tissue removal leads to a significant increase in signal-to-noise ratio and higher speed. Another method, Expansion-Assisted Iterative Fluorescence *In Situ* Hybridization (EASI FISH)[36], aims to address the fact that seqFISH and MERFISH only allow analysis of limited tissue sizes, while being time-consuming and labor-intensive. Lastly, commercial options for some of these techniques are becoming feasible. Many of these solutions offer analysis of thousands of genes and the option to simultaneously analyze dozens of proteins. Most other smFISH-based techniques, such as hybridization-based ISS (HybISS)[35] and split-FISH[55], use barcoding similar to seqFISH or MERFISH. smFISH faces many challenges, which have been addressed through various methods. Rolling Circle Amplification (RCA), branched DNA[56], hybridization chain reaction (HCR)[57], primer exchange reaction, and tissue clearing can improve signal-to-noise ratio. As the gene repertoire increases, it becomes increasingly likely for transcripts to overlap, causing optical crowding. This can be alleviated by expansion microscopy (ExM)[58], imaging only subsets of probes at a time, using computational super-resolution to image highly expressed genes without the use of combinatorial barcoding, and calculating solutions to resolve overlapping points.

1.2.3 *In situ* sequencing

The method closest to the *in situ* hybridization-based method is the *in situ* sequencing-based method. This series of methods amplifies the target signal *in situ* using padlock

or RCA, and then identifies the base signals using microscopy. The most used sequencing technique in this method is ligation sequencing, rather than the synthetic sequencing that dominates traditional next-generation sequencing. Similarly, these methods achieve subcellular resolution; however, due to the requirement for prior knowledge and the limited cellular space, this method can only capture target transcripts and has limited throughput. The maximum read length is ~30 nucleotides, as longer reads are more difficult to achieve *in situ* than on flow cells due to the influence of many variables. ISS[25] methods obtain spatial transcriptomic information through sequencing, typically using spatial barcoding, gene barcodes (targeted), or *in situ* cDNA short fragments (untargeted). This method relies on ligases to connect two DNA fragments - a primer with a known sequence and a probe - and any mismatched probes will be washed away if they match the template. The probes used are degenerate, except for one or two query bases encoded by color. Subsequently, ISS was commercialized by Cartana in 2013, and barcoded oligonucleotides linked to RNA for multiplexed *in situ* analysis were sequenced with one query base per probe[59], like combinatorial probe anchor ligation (cPAL)[60], for gene barcoding. In cPAL, one base of the gene barcode is queried by each probe. Meanwhile, a technique called fluorescence *in situ* sequencing (FISSEQ)[28] generates RCA amplicons of cDNA molecules through circularization, which are themselves generated by mean of RT of mRNAs by using an oligo(dT) primer. This technique allows for true "untargeted" sequencing in space, although it is only used for very short 3'-end reads and has very low sensitivity (far below 1% of the total cellular transcriptome) due to the low efficiency of *in situ* RT and cDNA circularization. In the spatially-resolved transcript amplicon readout mapping (STARmap)[31], gene barcodes are sequenced by sequencing with error-reduction by dynamic annealing and ligation (SEDAL), which uses two query bases to reject errors, but single-base encoding can also be used. BaristaSeq[32] improves efficiency by optimizing the gap filling padlock probe method and using Illumina synthetic sequencing for detection. Recently, both the lock-based ISS method and the untargeted FISSEQ method have been combined with

expansion microscopy chemistry[61], resulting in substantial improvements in efficiency and signal-to-noise ratio. One of the key innovations of ExSeq[61] is the combination of untargeted, short-cycle ISS with conventional "bulk" sequencing after extracting cDNA present in the tissue. Short sequences serve as different "labels" to match each "bulk" read to specific locations, effectively generating full-length, long reads of ISS that are sufficient for mapping alternative splicing isoforms at subcellular resolution in complex tissues such as the mouse brain.

Both *in situ* hybridization-based methods and *in situ* sequencing-based methods have one commonality: detection is ultimately accomplished through high-resolution microscopic imaging of tissues, and gene expression is measured by effectively counting individual mRNA molecules. While imaging is the most direct method to obtain spatial information, single-molecule microscopy is highly challenging and requires extremely precise instruments and procedures. This technical barrier becomes even more complex due to the need for multiple imaging cycles, as well as extensive sample manipulation through fluid exchange, enzymatic reactions, and other procedures. It is not surprising that image co-registration and feature extraction are among the biggest challenges in analyzing all image-based spatial transcriptomics datasets. Single-molecule microscopy can also be slow as it requires high magnification and resolution, resulting in a very small field of view, necessitating the assembly of numerous small images and image stitching (which itself is not a straightforward process) to obtain any meaningful sample size. Given the diffraction limit, molecular crowding within cells also imposes a limit on the number of molecules that can be resolved. Therefore, the sensitivity of detection is influenced by cell size, with larger cells allowing for deeper analysis.

1.2.4 *In situ* spatial barcoding

Some may argue that if only spatial location can be encoded in sequence space, then the best available tool for generating high-quality transcriptomics data at a very high

speed and throughput is already available in the form of conventional NGS. This is precisely the key insight behind spatial transcriptomics. In ST, ordered oligonucleotide arrays are deposited on a glass/silicon slide using microarray printing technology. Each oligonucleotide includes a handle(adapter) compatible with downstream sequencing reactions, a barcode specific to each spatial location, a random unique molecular identifier (UMI) for correcting polymerase chain reaction (PCR) amplification bias, and a poly(T) sequence for capturing polyadenylated (poly(A+)) mRNAs. Each array spot has a different spatial barcode. Thin tissue sections (~10 μm) are then placed on the array, and cell RNA diffuses onto the barcode oligonucleotides through cellular permeabilization, followed by *in situ* RT to generate spatially indexed cDNA. The latter is then amplified, adapter ligated to generate libraries, and sequenced using standard NGS. Some key advantages include no requirement for prior knowledge, unbiased capture of target molecules, ability to generate long sequencing reads (although still biased towards the 3'-end of transcripts), independence from complex imaging instruments, high speed (processing time effectively independent of sample size), and the potential for parallelization. These features greatly facilitate the commercial applications of this technology[10]. *In situ* spatial barcoding is a recent breakthrough that employs high-density barcode probes fixed on a carrier to capture spatial RNA within tissue sections, offering an unbiased method for capturing the entire transcriptome within tissues.

1.2.4.1 Microarray-based *in situ* spatial barcoding technologies

The breakthrough in SRT technology in the early days was based on the use of glass surfaces or microbeads to connect spatial probes and generate capture carriers. Additionally, by increasing the density of spatial probes on the glass surface or reducing the diameter of the microbeads, higher precision in spatial resolution can be achieved. ST[8] was the first technology based on *in situ* spatial barcoding. It utilizes glass slides with fixed spatial barcodes and poly(T) probes to generate thousands of capture points,

creating capture regions. The spatial probes in ST are consistent within each capture point but variable between two capture points; thus, the minimum distance between capture points determines the spatial resolution. In 2019, 10X Genomics released Visium[44] based on ST, which has a higher spatial resolution of 100 μm and a diameter of 55 μm (center-to-center distance between adjacent capture motions), smaller than ST but still larger than the diameter of most cells. Based on the design, improving spatial resolution further remains a challenge for ST/Visium. Considering the overall strategy for generating capture regions, Slide-seq[20]/Slide-seqV2[45] and high-definition spatial transcriptomics (HDST)[42] are very similar. They both utilize pre-synthesized spatial barcodes and poly(T) probes placed on a slide to form capture regions, instead of directly printing probes onto glass slides in ST. Each bead contains a unique barcode sequence, which serves as the basic unit of spatial resolution (the center-to-center distance between adjacent beads). The spatial resolution is 10 μm in Slide-seq, while it is 2 μm in HDST. If only considering the physical size of cells, such as mouse brain cells of about 10 x 10 μm^2, Slide-seq/Slide-seqV2 and HDST are among the first or few methods that achieve single-cell resolution using barcoding. However, in terms of using single-cell regions as the smallest analysis unit, the capture efficiency of each barcode head is not satisfactory, requiring sufficient capture of RNA information to perform statistically significant calculations. Thus, enhancing RNA capture efficiency will play a crucial role in advancing the development of high spatial resolution SRT techniques.

1.2.4.2 Microfluidics-based *in situ* spatial barcoding technologies

Combining microfluidic technology with SRT technology is a compatible strategy that allows for the alteration of spatial resolution by adjusting channel size, making spatial multi-omics more achievable with the use of spatial barcoding. However, due to limitations in engineering materials, the spatial resolution of this method is difficult to achieve at the single-cell level. Deterministic barcoding in tissue for spatial omics sequencing (DBiT-seq)[62] is the first microfluidic-based technology that measures

mRNA and protein spatial information simultaneously through NGS. Unlike the other techniques, DBiT-seq does not pre-synthesize capture regions with unique spatial barcodes and poly(T) probes. Instead, the barcode oligonucleotides flowing through each channel are linked to target biomolecules. Subsequently, the microfluidic chip is rotated 90° and barcoding is performed a second time, generating a grid where each intersection has a different index. In this process, poly(T) probes bind to mRNAs, meaning that mRNAs in the region channels are linked to the barcodes. The subsequent steps include mRNA-cDNA complex synthesis in RT, followed by tissue digestion, collection of mRNA-cDNA, template amplification, and generation of NGS libraries. The pixel size of the channels determines the spatial resolution, ranging from 10 to 50 μm. Although the spatial resolution of DBiT-seq is already approaching the single-cell level, its capture and spot analysis are still not equivalent to single cells. Additionally, its capture efficiency is relatively high, possibly due to the absence of diffuse RNA from tissue to barcode carriers.

1.2.4.3 Sequencing carrier-based *in situ* spatial barcoding technologies

The combination of NGS "chips" and SRT technology is essential for single-cell resolution spatial analysis, which requires two rounds of sequencing. The carrier can be a sequencing chip or a similar chip that can carry a higher density of spatial probe distribution. In addition, the sequence of spatial probes is not required when attached to the chip surface, but can be sequenced during the attachment process or in the first round of sequencing. Then, tissue sections are placed on the chip surface to capture RNA, and a second round of sequencing is performed. The first round sequencing data is compared to obtain the spatial location of each molecule, thereby achieving single-cell level spatial resolution. Poly-indexed library-sequencing (Pixel-seq)[23] uses a "stamp gel" as a template, which contains clusters of cloned DNA of about 1 μm, including poly(T) probes and spatial barcodes, which can be replicated onto many "replica gels" and capture regions are performed on the slide using poly(T) probes. The

distance between adjacent and specific DNA clusters determines the spatial resolution of Pixel-seq, which is about 1 μm. It achieves spatial single-cell resolution by combining weighted network segmentation and transcript separation into cell masks. Considering that in other *in situ* barcode-based methods, the captured cell layer is the only cell layer close to the replica gel surface, rather than relying on permeable multiple cell layers, the capture efficiency of Pixel-seq is similar to that of permeable SRT methods, for example, an average of about 1000 UMIs in a 10 x 10 μm^2 area of mouse tissue. However, this still needs to be enhanced to capture higher UMIs in a single-cell area to achieve more biological identification. Due to the gel-gel replication strategy, many "replica gels" can be generated from a "stamp gel" template with the same spatial barcode probes, and even "replica gels" can be used as "stamp gels" for subsequent manufacturing rounds, so only one or a few "replica gels" need to be sequenced to obtain spatial barcode information. In contrast, each capture chip of Seq-Scope[22] and Stereo-seq[47] requires separate sequencing, and the Pixel-seq strategy can reduce manufacturing costs and time. Seq-Scope is based on solid-phase amplification and provides a spatial resolution of about 0.5-0.8 μm (average ~0.6 μm), surpassing the previously published *in situ* spatial barcode-based techniques in 2021. The first round of sequencing aims to identify the unique spatial barcodes in the region's physical array, followed by exposing the poly(T) probe domain to generate an HDMI-encoded RNA capture array (capture region) for tissue section sequencing. Similar to ST and HDST, Seq-Scope can perform tissue section sequencing and H&E staining, while in Seq-Scope, H&E staining images can be segmented based on cell boundaries. The images and sequencing data are then merged for further analysis at the level of true single cells. The recently developed Spatially Enhanced Resolution Genomic Sequencing (Stereo-seq)[47] technique is the first method to achieve subcellular resolution and centimeter-scale field of view. It utilizes DNA nanoballs (DNBs) on sequencing chips as spatial barcodes. Stereo-seq is based on a pattern array of DNB with a diameter of 220 nm on a photolithography chip, overcoming the physical distance limitation between spots and

achieving nanoscale spatial resolution or super-resolution. The Stereo-seq chip combines high resolution of DNB and high probe density per unit area, enabling effective capture of spatial transcriptome at the whole-genome level with nanoscale resolution. In addition, by capturing the track line features of the DNB array stained with nucleic acid dyes and taking photographs together with tissue sections before and after fixation and permeabilization reactions, Stereo-seq can accurately obtain tissue cell morphology and sequenced mRNA transcripts, achieving further single-cell segmentation. Overall, with the high-density barcode chips (with large probes and nanoscale resolution), trajectory line features of the DNB chip, high polymer capture efficiency, and appropriate RNA diffusion, Stereo-seq is capable of decoding tissue complexity at the single-cell level with spatial resolution and centimeter-scale field of view.

1.2.5 Trends in the methods of spatially resolved transcriptomics

Ideally, spatial resolved technologies would provide a nearly complete and unbiased profiling of the entire RNA molecule content at subcellular resolution in a short period of time, and be sufficiently reliable and affordable to be applied to many samples. However, currently all techniques require some compromises, so the choice of method depends on the specific needs of each study. Capture efficiency, whole transcriptome, resolution, throughput, large tissues, ease of implementation, robustness, and cost are all key variables in the balancing act of spatial analysis.

1.2.5.1 Capturing efficiency

Despite displaying significantly higher gene detection numbers, for example, in single-cell sized regions, Stereo-seq detected 1450 UMIs (mouse olfactory bulb, 100 μm^2), Seq-Scope detected 848 UMIs (mouse liver, 100 μm^2), and Slide-seqV2 detected 550 UMIs (mouse embryo, ~100 μm^2). Compared to the sensitivity of current single-cell RNA sequencing (scRNA-seq) technologies, such as ~40,000 UMIs (~6,000 genes)

detectable in 10X Chromium, ~8,000 genes detectable in Smartseq2, significant improvements are still required in the SRT technology to enhance cell classification performance. Currently, SRT has proposed a new strategy, which is to utilize poly(A) polymerase to add poly(A) tails *in situ* on RNA, enabling the detection of the entire spectrum of RNA while using barcode-based methods to capture the added poly(A) RNA. Compared to only capturing naturally occurring poly(A) tails, which cannot measure degraded RNA, the latter can improve capture efficiency. RNA degradation poses a significant challenge in some clinical tissues or easily degradable samples.

1.2.5.2 Spatial resolution

From a few mixed cells per capture point in ST/Visium to near single-cell resolution in Slide-seq/Slide-seqV2, HDST, and DBiT-seq, and even to true single-cell resolution based on cell segmentation in Seq-Scope and Stereo-seq; however, further improvements in resolution are still necessary for new insights into cellular spatial gene regulation and biology. Expansion microscopy (ExM) is a potential technology that utilizes expandable gel materials to enhance the spatial resolution of SRT techniques, which has been applied in Ex-Seq and STARmap/STARmap PLUS as *in situ* sequencing methods. For barcode-based methods, expanded spatial transcriptomics combines ExM with Visium, increasing the resolution of Visium arrays from 55 μm to 20 μm and improving the RNA capture efficiency per region. Furthermore, if higher-resolution SRT techniques such as Seq-Scope and Stereo-seq are combined with expansion microscopy, the spatial resolution could be further improved from ~500 nm to ~100 nm or even less than 100 nm, which could improve biological understandings of cellular interplays in tissues.

1.2.5.3 Tissue area

In general, techniques with lower detection efficiency are often more suitable for analyzing larger tissue areas. In current SRTs, tissue sections several millimeters wide,

such as a significant portion of mouse brain coronal sections, can be suitable for ST tissue capture areas, which are considered large, and increasing tissue area and sequencing depth to improve sensitivity will increase sequencing costs[9]. Cartana ISS and HybISS have also been used to depict tissue areas several millimeters wide, but only with about 100 genes. The advantage of HybISS here is strong RCA signals and less optical crowding, which is due to lower detection efficiency and lower magnification, resulting in faster imaging speed. For smFISH techniques, there seems to be a trade-off between the size of tissues and the number of genes. ROI selection techniques are usually used for a small number of ROIs, as selecting a very large number of ROIs and processing them individually without spatial barcoding is labor-intensive.

1.2.5.4 Accessibility of technology

While many new technologies have been developed, most of them have not been disseminated beyond their originating institutions. Among the more widely spread companies, the most popular ones are often commercial platforms such as LCM[39], 10X Visium[44] (formerly ST and Cartana ISS, acquired by 10X), and Nanostring GeoMX[43]. Additionally, many major institutions have core facilities for NGS, reducing the cost of purchasing new equipment and training individual lab personnel if platforms like LCM, Visium, and GeoMX are available. Tomo-seq has also gained popularity, perhaps due to its ease of implementation on standard equipment. In contrast, smFISH-based technologies have not been widely disseminated so far, possibly due to the complexity of homemade fluidic systems, longer imaging times, terabyte-level images, and expensive probes. Some smFISH technologies are being commercialized through automated imaging and fluidic platforms, such as Vizgen's commercialized MERFISH technology and Resolve Biosciences' molecular mapping platform based on smFISH. Moreover, Rebus Esper can be programmed to automate different smFISH technologies and process images online, like Illumina sequencing, and has been used for automated

cycling-osmFISH. With the emergence of new automated commercial platforms, smFISH-based technologies may become more popular, especially if adopted by core facilities. The combination of *in situ* barcoding technology and bulk sequencing technology has been very successful. The availability of commercial choices, the independence from high-precision instruments, and the compatibility with existing workflows have reduced the entry barriers for many groups. Data analysis is also relatively simple because all processing occurs at the level of sequence data, which is a more widely specialized field in biological laboratories compared to image analysis. However, the low resolution and random localization of spatial barcoding relative to the samples means that information can sometimes be averaged across several different types of cells, leading to result confusion. Nevertheless, ST and its derivative technologies have been successfully applied in many studies, particularly in the fields of neuroscience and cancer biology, and improved methods may further drive the adoption of spatial analysis techniques.

1.3 Tools and challenges in spatially resolved transcriptomics data analysis

Compared to single-cell "omics", SRT generates a large amount of data (often many TBs), and the entire data processing is influenced and biased by many factors that without proper analysis tools, an experiment costing thousands of dollars may become completely useless. At the same time, the raw data is sometimes very large, and without specialized hardware and software, they cannot be visualized correctly. Therefore, there are huge challenges in processing of SRT data. The entire process of spatial dataset analysis can be roughly divided into two parts: (1) upstream analysis, which generates a gene expression matrix with spatial location information, which is directly associated with spatial technology itself. And then, selects different resolutions and analyzes different regions based on different applications; (2) downstream analysis, which is usually applicable to gene count matrices and cell or spot locations, and is therefore

largely independent of data collection techniques. These are illustrated in Figure 2, and will be further detailed below.

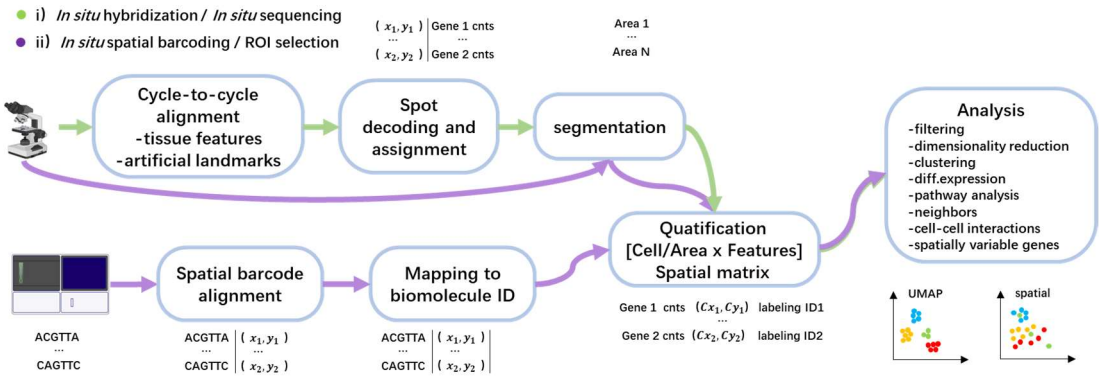

Figure 2: Analysis workflows for spatial datasets. The core acquisition modality can be two main categories: i) imaging-based, including *in situ* hybridization and *in situ* sequencing; ii) next-generation sequencing-based, including *in situ* spatial barcoding and ROI selection. Imaging-based methods involve several preprocessing steps for the field-of-view images captured by the microscope during each cycle, including image stitching, background subtraction, and registration. In single-molecule imaging methods, the spots that represent signals on the images need to be identified and decoded to assign them to corresponding molecular labels. The cell images are then segmented to define areas corresponding to each cell (mask) by either identifying the cell membranes or performing a geometric expansion around the nucleus as a representation of the cytoplasm. Subsequently, the number of spots or signal intensities is integrated within the cell masks to generate an area-by-features matrix. By contrast, NGS-based methods produce datasets in the form of sequencing output, which are preprocessed and parsed to assign each read to a spatial coordinate through the position barcodes and a biomolecule's ID by mapping it to a reference, such as a transcriptome. After obtaining the area-feature expression matrix, subsequent analysis steps typically involve filtering, dimensional reduction, clustering, and identification of differentially expressed markers. Some intermediate results can be visualized directly or based on the dimensionality reduction space of genomic features. Further analyses are performed according to the specific research questions.

1.3.1 Upstream analysis

1.3.1.1 Spatial matrix generation

The so-called "spatial expression matrix" in its wide matrix form is a two-dimensional non-negative matrix, where the row names are genes and the column names are two-dimensional XY coordinate pairs. To reduce storage space, the spatial expression matrix can be processed using a long matrix that only includes the three-dimensional coordinates (X, Y, gene) and the positions of non-zero values. Obtaining the spatial

expression matrix is the first step in upstream analysis of spatial data, with a key step being to decipher the spatial location of each mRNA transcript molecule. For spatial techniques based on ISH and ISS, the raw data consists of fluorescent dot images. The first step is to stitch the images and align cyclical images. After alignment, the images need to be processed for background correction, filtering, and normalization to identify the transcription point signals. The signal points are then connected to match genes, and error correction and detection are applied to obtain the molecular quantity at each location. In this process, image registration is a very important issue that has been extensively discussed elsewhere and becomes more complex in spatial profiling due to the large amount of data. Most methods use easily identifiable "reference landmarks" (fluorescent beads) as features to calculate the registration matrix, which makes registration easier but requires additional experimental processing. Recently, the SpaceTx Consortium has made significant efforts and released a python module called Starfish[63], which has been successfully applied to most techniques such as seqFISH, MERFISH, and ISS. However, it has not been widely adopted yet as it generally does not perform as well as dedicated pipelines for each method.

In spatial barcode-based SRT, there are various methods for obtaining the spatial gene expression matrix. Similar to standard single-cell sequencing techniques, spatial barcoding technique involves two libraries: read 1 containing spatial barcode and UMI sequences, and read 2 containing cDNA sequences. However, the difference lies in the fact that in spatial techniques, barcodes are associated with spatial coordinates rather than cells. It is worth noting that each technique currently has its own method of associating barcodes with spatial coordinates: (a) in DBiT-seq, ST, and other techniques, the spatial position of barcodes is predetermined, and each barcode has a known spatial coordinate[62]; (b) in seq-Scope technique, the synthesis sequencing technology ensures that the corresponding spatial coordinates can be obtained during the synthesis process on the spatial chip HDMI[22]; (c) in Stereo-seq, after loading DNB onto the sequencing chip, a first round is performed to obtain the barcode sequences and corresponding

spatial coordinates of each DNB, followed by a second round of spatial transcriptome experiment and cDNA sequence capture. Once the correspondence between barcodes and coordinates in read 1 is established, read 2 can be aligned to the genome to determine the spatial location of gene expression[47]. Overall, this workflow is consistent with the methods used in single-cell transcriptome analysis.

1.3.1.2 Image processing and cell segmentation

High-resolution SRT datasets typically include immunohistochemistry, hematoxylin and eosin (H&E) staining, and/or nucleic acid staining images to obtain subsequent cell/nucleus segmentation and supplement cell-level biological features. Most techniques based on *in situ* hybridization and *in situ* sequencing have subcellular resolution, but almost all downstream analyses are performed at the single-cell level. Once a molecule is detected in space, it needs to be assigned to a cell, and all measurements within a cell region need to be integrated into a single abundance score for each gene or protein, ultimately creating a cell feature matrix. To achieve this, the original images are segmented to identify regions belonging to individual cells. In the previous section, we mentioned that in the processing of the two aforementioned techniques, a matrix is obtained through image processing, and stitching and registration operations are required. Here, we also need to perform stitching operations on the stained images and align them with the fluorescent images of the expression matrix. For image stitching and registration, ASHLAR[64] provides multiplexing technology with better results, but lacks a good gold standard for evaluation. Image registration tools (such as Spot Detector[65]) provide an automated process for aligning images and spot locations, mainly including alignment, framing, and validation steps. However, with the emergence of high-resolution SRT techniques, a more precise alignment process is needed. For example, in the alignment process of Stereo-seq, the matrix of each expressed DNB point is converted to a grayscale image, and then manually aligned with the nucleic acid staining image using track line features. The

next step is to obtain spatial single-cell/nucleus data based on the segmentation of cell/nucleus positions in the stained images, providing a bridge for achieving spatial single-cell resolution. Image segmentation is a core problem in biological imaging and computer vision, and it has benefited from the rapid development of artificial intelligence (AI) and deep learning in recent years[66–68]. Although cell nucleus segmentation benefits from the regular shape of the nucleus and the availability of universal stains (DNA intercalators), determining the cell membrane boundary is a more complex task. Membrane markers often produce low signals, have poor specificity for the cell membrane, or are not universal for all cell types, so even with the use of AI models, effective segmentation cannot be achieved. While combining multiple labels has shown some promising results, developing a truly universal membrane marker would have a revolutionary impact on the field of spatial transcriptomics. Based on various considerations for achieving cell segmentation, the multi-dimensional mRNA density estimation calculates cell types based on RNA distribution without cell boundary segmentation[65], while Baysor[69] achieves cell boundary recognition and segmentation annotation. Joint segmentation and cell type annotation (JSTA)[70] use cell types defined by scRNA-seq to assist in identifying spatial data.

1.3.1.3 Cell type deconvolution

In order to achieve spatial single-cell analysis in low-resolution SRT technology, it is often necessary to perform deconvolution from scRNA-seq datasets instead of directly performing cell segmentation. One common solution is to integrate SRT with scRNA-seq datasets. It is worth noting that one method currently available to integrate the two datasets are through probabilistic modeling. The expression of individual cells in single-cell data can often be modeled as a negative binomial distribution or a Poisson distribution. Robust Cell Type Decomposition (RCTD)[71], and cell2location[72] use the aforementioned probabilistic models to integrate individual cells and provide

deconvolution of cell types. SPOTlight[73] and SpatialDWLS[74] use topic similarity obtained from non-negative matrix factorization (NMF) dimensionality reduction to determine SRT cell types. CellTrek[75] projects SRT and scRNA-seq data into the same latent space and models them using multivariate random forests (RF)[76]. Tangram[77] is a deep learning method that calculates the correspondence between SRT points or cells after cell segmentation using neural networks. However, the integration strategies of SRT and scRNA-seq datasets still face some challenges, including the possibility of ignoring or losing partial information of rare cell types due to current technological and algorithmic limitations. For example, in low spatial resolution SRT technologies, a capturing point may cover and capture mixed RNA from multiple cells, making it impossible to accurately distinguish the molecules *in situ* of each cell and label them to their corresponding cells, which prevents the subsequent single-cell precision analysis. Moreover, the transcriptomic pattern of rare cell types does not show obvious differences from the major cell types; while, the specific information of rare cell types retained during data integration may be a potential point for further identifying rare cell type-specific markers[4]. For example, RaceID[78] provides a strategy to identify rare cell information from multiple cell populations and has shown good performance[79].

1.3.2 Downstream analysis

1.3.2.1 Combination of spatial and single cell

Given the correlation between scRNA-seq and spatial data, as well as the analysis approach of spatial data in exploratory data analysis (EDA), popular scRNA-seq EDA ecosystems, such as Seurat[80–84], SCANPY[85], and single cell experiment (extended by SpatialExperiment[86]), have all enhanced the functionality of spatial data. For example, they have updated data containers and functionalities to facilitate the visualization of gene expression and cell or point metadata in spatial locations. Dedicated EDA packages specifically for spatial data have also been developed, which feature beautiful graphics and well-documented software, such as Giotto[87] and STUtility[88]. Seurat and

Giotto[87] have also implemented basic methods for identifying spatial variable genes. Additionally, Giotto has implemented methods for identifying cell type enrichment on ST and Visium spots, identifying co-expression of genes, cell type co-localization of gene expression, and identifying spatial regions. Spatial variable genes refer to genes whose expression is correlated with spatial location. These genes are often identified using three methods: Gaussian process regression (GPR)[89] and its extensions for Poisson[90] and negative binomial (NB)[91], Laplacian score[92], and Moran's I. The GPR-based method employes GPR to normalize the rate parameter of gene expression or Poisson/NB gene expression and determine whether the model provides a better fit to the data with or without spatial features. Laplacian score-based methods identify genes that exhibit a stronger association with the spatial neighborhood graph structure in terms of their expression. The positioning of cells can be modeled as a spatial point process, using gene expression as markers, and spatial variable genes can be identified by their correlation with location. The computational cost of fitting GPR models to multiple genes, particularly when using Bayesian methods with Markov chain Monte Carlo, can be high. Similarly, permutation tests used in Laplacian score methods can be time-consuming[9]. In some cases, classical spatial autocorrelation measures such as Moran's I, as implemented in Seurat v3[82] and above, are directly applied to identify spatial variable genes in both GPR-based and Laplacian score-based methods. MERINGUE[93] utilizes a local version of Moran's I, and its significance tests are implemented in well-established geographic spatial software packages, which are simple and quick to execute, but may have lower statistical power compared to model-based methods.

1.3.2.2 Cell-cell communication

The emergence of SRT technology enables the direct measurement of cell communication in specific microenvironments through the analysis of cell-cell interactions, which is more advantageous for analyzing short-distance interactions. Therefore, false-positive results based solely on RNA expression levels are excluded.

Based on the SRT dataset, the expression of ligands and receptors and the determination of spatial distance coexist, allowing the determination of cell-cell interactions. Importantly, the localization of ligands at the corresponding subcellular positions is crucial in the spatial aspect of cell-cell interactions, which cannot be deciphered by standard single-cell RNA sequencing experiments. To address this, Physical Interaction (PIC)-seq methods provide a high-throughput solution to study interactions arising from spatial proximity, while advanced algorithms considering intercellular communication constants can provide additional input. It is worth noting that several computational tools have been developed in this regard, such as CellChat[94] and CellCall[95]. Different tools have applied various methods to identify intercellular communication using the SRT dataset. For example, Gene Graph Convolutional Neural Networks (GCNG)[96] analyze gene interactions between cells, SVCA[97] quantifies the impact of intercellular interactions on gene expression, and DIALOGUE[98] explains multicellular collaborative programs. Additionally, NICHES[99] and HoloNet[100] consider the expression profiles of ligands and receptors to visualize heterogeneous signaling prototypes or develop graph neural network models to explore transcriptional events related to intercellular communication. Finally, the SpaTalk[101] tool models and scores ligand-receptor-target signaling networks based on network biology methods, such as knowledge graph approaches[102], to reveal cell communication in the SRT dataset on a general scale. There are also many other types of downstream analyses that are useful for spatial transcriptomics analysis, including identifying prototype gene patterns, defining spatial regions by transcriptome, inferring gene-gene interactions, subcellular transcript localization, and gene expression estimation from H&E images.

1.3.3 Challenges in spatially resolved transcriptomics data analysis

Although some analysis tools have been developed, with the decreasing cost of SRT technology and its widespread application, more challenges in data analysis will arise. These challenges may include the following: (a) SRT data size is enormous. As high-

density large field-of-view datasets increase, new methods for SRT data storage and data preprocessing speed will be urgently needed in the future. (b) Batch effects often occur between different slices due to different technical batches, operators, or operating methods. Correcting batch differences is a significant challenge that needs to be addressed to compare different slices and obtain analysis conclusions that are more accurate and reliable. It is worth noting that several techniques have been developed to process batch effects in scRNA-seq methods[103–106]. However, considering the complexity of SRT, all these methods require standardization. (c) As the individual slice in SRT cannot represent the whole organ, methods for three-dimensional (3D) reconstruction using SRT data will become increasingly important in the future. Currently, advances in the 3D field are primarily based on the 3D reconstruction of multiple adjacent slices of the same tissue, which relies on the stability of experimental techniques to allow comprehensive analysis of multiple adjacent slices, and also relies on analysis tools capable of 3D reconstruction, such as the probabilistic aligner for spatial transcriptomics (PASTE) in ST experiments[107]. Nevertheless, more efficient analysis methods are still needed in the future to address many challenges faced by current 3D reconstruction, such as cell mapping and reconstruction of links between adjacent slices. In addition, with the rapid development of advanced analysis methods, it is necessary to systematically evaluate these methods to determine their reliability and guide the selection of analysis tools. Many benchmark studies are needed in terms of single-cell resolution and spatial variable gene selection, and new algorithms need to be established for standardized evaluation systems, cell segmentation accuracy, spatial domain recognition, and spatial clustering. So far, existing benchmark studies have shown that some tools perform better in assessing the spatial distribution of RNA transcripts, while others show better performance in cell deconvolution methods or clustering accuracy[108].

In summary, the challenges to be addressed include cross-slice data alignment, elimination of batch effects, data normalization, and filling in information gaps between

slices. Other challenges involve (a) integrating ST with other modalities[10]. Given the breakthroughs in spatial multi-omics methods, there is an urgent need to improve an integrated algorithm for understanding spatially resolved data with tissue functionality and how to integrate SRT data with data obtained from other methods such as proteomics and chromatin accessibility[109]. (b) how to identify consistent spatial domains in gene expression patterns and histological features[110]. (c) how to combine SRT data with multiple (rather than single) adjacent tissues that have undergone different slice treatments[107].

1.4 Thesis aim, problems and hypotheses

Based on a thorough exploration of the progress, significance, and challenges of spatial transcriptomics, it was clear that NGS-based barcode spatial technology has developed rapidly during recent years. Due to the widespread use of second-generation sequencing technology, it was decided that there was no need at the time of initiation of the thesis work for additional equipment development, as several available technologies were deemed to be highly commercially viable. The technology used in this thesis work is Stereo-seq technology. Stereo-seq, which is based on mainstream second-generation sequencing technology (DNBSEQ), has the advantage of having a large field of view (centimeter-level) combined with high resolution (nanometer-level). It can also capture the whole-wide transcriptome, and has recently shown to be useful in many research fields. For example, the Stereo-seq technology has been used to analyze the gene spatial expression patterns of organ development in late-stage mouse embryos, draw organ development trajectory maps, and construct spatial-temporal transcriptome maps of mouse organ development[47]. This technology has also been applied to establish high-resolution spatial landscapes of cell types in the regenerating brain of axolotls, revealing new cell types involved in brain regeneration[111]. In addition, Stereo-seq has also created a three-dimensional spatial resolution cell type atlas of the macaque brain, with most sections covering an area of 5 x 3 cm^2, making it possible to study the large-scale

distribution, heterogeneity, and functionality of the whole brain[112]. Moreover, Stereo-seq has played an important role in disease mechanism research. For example, Wu *et al*. used Stereo-seq for the first time to analyze the heterogeneity and microenvironment of intrahepatic cholangiocarcinoma, defining a 500 mm wide region centered on both sides of the tumor boundary and characterizing the cells and transcriptional levels[113]. Zhang *et al*. used Stereo-seq technology to study colorectal adenocarcinoma, describing the characteristics of complex tumor regions and identifying molecular patterns involved in discontinuous inflammatory responses within this region[114]. In addition, Xia *et al*. used Stereo-seq to study the spatial resolution of plant cell landscapes, distinguishing cell subtypes of *Arabidopsis* leaves based on spatial information for the first time, demonstrating the new discovery capabilities of high-resolution Stereo-seq in the field of plant biology[115]. However, compared to other spatial technologies such as *in situ* sequencing, *in situ* hybridization, and microdissection, which have commercialized products such as 10X Xenium[116], Nanostring CosMx[117], and Vizgen MERFISH[30], barcode-based technology has only recently developed its high-resolution, large field-of-view, and urgently needs to accelerate its applications to help researchers speed up their scientific progress. Stereo-seq has been launched and is undergoing rapid iteration and optimization. However, due to its high resolution and large field-of-view, existing data analysis tools cannot effectively solve the problems related to analysis of Stereo-seq-derived.

The overall research aim of the current thesis work was therefore to solve and provide more efficient and accurate analysis tools to support Stereo-seq data analysis, focusing on the addressing the following problems and hypotheses:

1) Problem: The production of a high-quality expression matrix during upstream data processing is too slow and inefficient. Hypothesis: It is possible to implement a complete and fast process for efficient and accurate attainment of single-cell spatial data from large field-of-view, high-resolution NGS barcode-based spatial transcriptomics

technology.

2) Problem: The accuracy of cell segmentation, *i.e.,* obtaining cell boundary information, is non-optimal for obtaining a high-quality gene expression matrix at single-cell resolution from large field-of-view applications. Hypothesis: By implementing other staining and automation procedures, analyses tools for obtaining spatial single-cell transcriptomics data from Stereo-seq can be developed.

3) Problem: For high-resolution spatial resolved technology, due to technical limitations, there is always a low capture rate, resulting in a large amount of empty capture sites and a relatively low overall sum of gene numbers in spatial single-cell transcriptomes, which have a significant impact on downstream analysis. Hypothesis: Improving the gene number level or enhancing the quality of gene expression profiles during data processing will be possible to implement using imputation or filtering methods.

The three hypotheses are addressed individually in the research manuscripts that make up chapter 2-4 that follow below. Each of the chapters are formatted according to journal requirements.

2 StereoCell: a highly accurate single-cell gene expression processing software for high-resolution spatial transcriptomics

Mei Li[1,2,8], Huanlin Liu[1,8], Min Li[1,8], Shuangsang Fang[1,3,8], Qiang Kang[1,8], Jiajun Zhang[1,4], Fei Teng[1,4], Dan Wang[5,1], Weixuan Cen[1], Zepeng Li[1], Ning Feng[1], Jing Guo[1], Qiqi He[1], Leying Wang[1], Tiantong Zheng[1], Shengkang Li[1,6], Yinqi Bai[1], Min Xie[4], Yong Bai[1], Sha Liao[1,4], Ao Chen[1,4], Susanne Brix[2]✉, Xun Xu[1,7]✉, Yong Zhang[1,6,7]✉, Yuxiang Li[1,6,7]✉

[1] BGI Research, Shenzhen 518083, China

[2] Department of Biotechnology and Biomedicine, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

[3] BGI Research, Beijing 102601, China

[4] BGI Research, Chongqing 401329, China

[5] Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, Zhuhai 519087, China

[6] BGI Research, Wuhan 430074, China

[7] Guangdong Bigdata Engineering Technology Research Center for Life Sciences, Shenzhen 518083, China

[8] These authors contributed equally: Mei Li, Huanlin Liu, Min Li, Shuangsang Fang, Qiang Kang

✉e-mail: sbrix@dtu.dk; xuxun@genomics.cn; zhangyong2@genomics.cn; liyuxiang@genomics.cn

Under revision in Nature Protocols.

Abstract

Owing to recent advances in resolution and field-of-view, spatially resolved sequencing, such as Stereo-seq, has emerged as a cutting-edge technology that provides a technical foundation for the interpretation of large tissues at the single-cell level. To generate accurate single-cell spatial gene expression profiles from high-resolution spatial omics data and associated images, a powerful tool is required. Here we present StereoCell, an image-facilitated one-stop software for high-resolution and large field-of-view spatial transcriptomic data of Stereo-seq. StereoCell provides a comprehensive and systematic platform for the generation of high-confidence single-cell spatial gene expression profiles, which includes image stitching, image registration, tissue segmentation, nuclei segmentation and molecule labeling. StereoCell is user-friendly and does not require a specific level of omics and image analysis expertise. StereoCell has contributed to the generation of a mouse organogenesis spatiotemporal transcriptomic atlas and was applied to generate reliable single-cell spatial gene expression profiles from continuous mouse brain slice datasets in previously published works. StereoCell is a fast tool for image and spatial omics data, demonstrated to be capable of handling a mouse brain dataset (131,990,020 molecules and 117 image tiles) in about 80 min on a server with 40-core CPU, 128 GB of RAM and 24 GB of GPU.

2.1 Introduction

Spatially resolved technology generates comprehensive data regarding the distribution of molecules that can be used to identify the location and function of cells within tissues, which helps to broaden our understanding of organ development[106], tumor heterogeneity[118] and cancer evolution[119,120]. This rapidly developing field is focused on obtaining detailed molecular information at single-cell resolution, as well as spatial information per molecule of tissue in a large field-of-view[121]. Single-cell resolution technologies[47,30,20], make it possible to explore spatial omics data at the single-cell, or even subcellular level. When combined with large field-of-view technologies[47], it allows for generation of 3D maps representing biological functions within single cells at the organ level.

2.1.1 Development of the protocol

Our previous work developed the high-resolution and large field-of-view spatially resolved technology Stereo-seq[47], which enabled whole-organ sequencing of most tissues in model organisms (e.g., mouse brain (1cm×1cm) and mouse embryo (1cm×2cm)). The Stereo-seq protocol generates two main outputs: spot-level gene expression data and high-content stained tissue images. The former is based upon detection of unique molecular identifiers captured by each DNA nanoball (DNB) on the sequencing chip while the latter provides detailed image-based (visual) information about the cellular matrix within tissue samples. Solutions that merge the two to obtain accurate spatial gene expression profiles at single-cell level in large field-of view settings will advance spatially resolved technologies and be an important steppingstone for downstream analyses.

Here we present StereoCell, an image-facilitated one-stop software for Stereo-seq transcriptomic data. StereoCell combines large field-of-view tissue images and high-resolution spatial gene expression data to obtain high-confidence single-cell spatial gene expression profiles. StereoCell provides a straightforward and systematic platform, which includes image stitching, image registration, tissue segmentation, nuclei

segmentation and molecule labeling (Fig. 1a). We use the multiple Fast Fourier Transform (FFT)[122] weighted stitching algorithm, named MFWS, for image stitching. MFWS is based on frequency domain information and allows the accurate stitching of high-resolution images in a wide field-of-view, in addition to improvements in cell dislocation errors caused by inaccurate image stitching and low efficiency caused by the huge number of image tiles (Fig. 1b). We transform the spatial gene expression data into a map. The stitched image is registered with the spatial gene expression map based on "track lines" (marker lines designed on Stereo-seq chip), which includes translation, scaling, flipping and rotation (Fig. 1c). We train the deep learning models for tissue segmentation (Fig. 1d) and nuclei segmentation (Fig. 1e) from the registered image. We enable molecule labeling using the cell morphology and Gaussian Mixture Model (GMM)[123] algorithm, named MLCG, which applies GMM to fit the molecules within nuclei mask for accurate assignment of surrounding molecules to the most probable cell, obtaining a gene expression profile at the single-cell level (Fig. 1f).

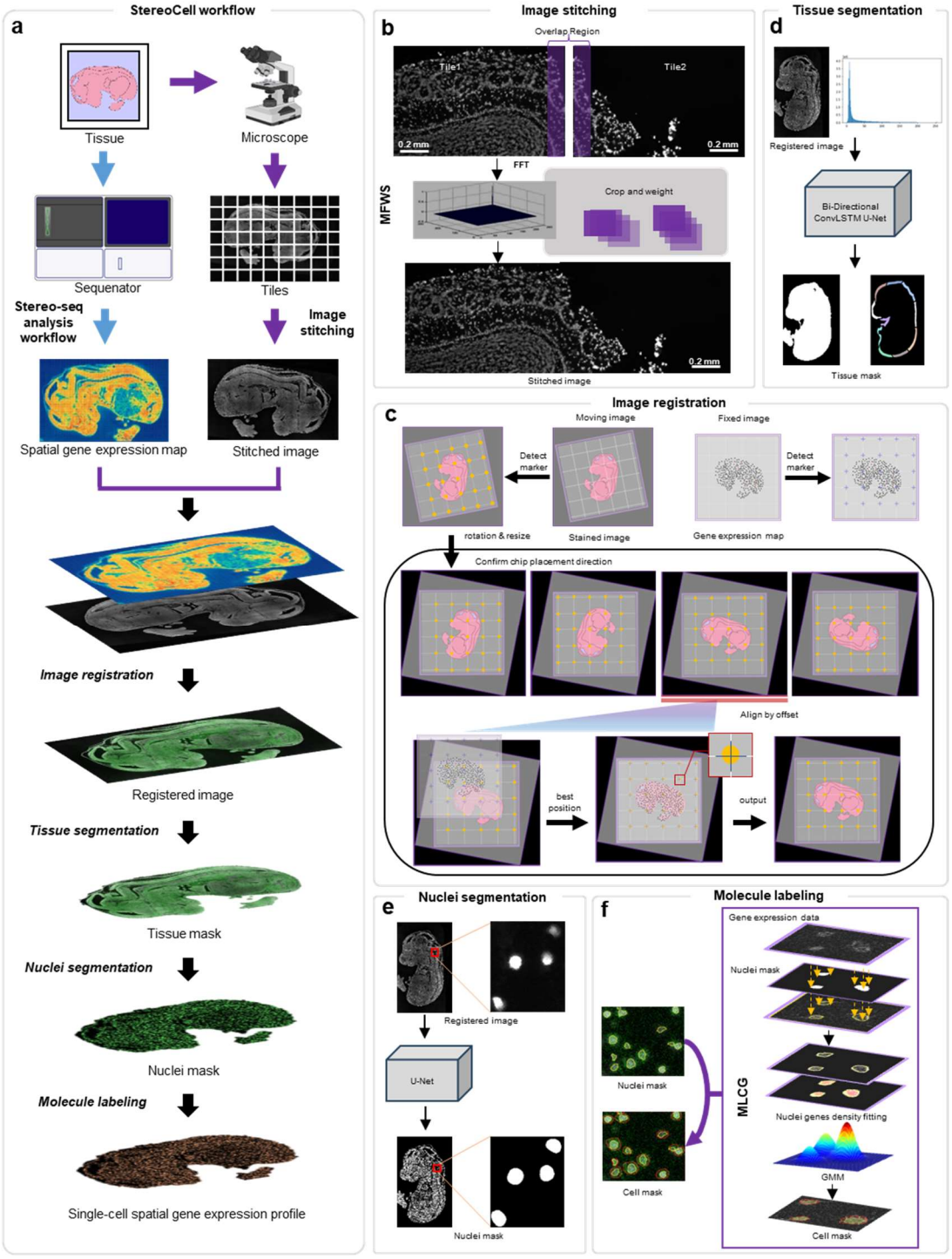

Fig. 1 | The StereoCell workflow and applied algorithms.

a. Schematic overview of the StereoCell software. The spatial gene expression data and the morphological (nuclei-stained) image of tissue are obtained using spatial transcriptomics technology and microscopy, respectively. The image tiles obtained by microscopy are stitched together to generate a large mosaic image of the whole tissue, the spatial gene expression data is transformed to a map, and the stitched image and spatial gene expression map are registered. Tissue and nuclei segmentations are performed on the registered image to obtain the tissue mask and nuclei mask. Molecule labeling is

adopted to obtain single-cell spatial gene expression profile.

b. Cutting the overlap regions of the adjacent image tiles, and calculating the spectral information based on the Fast Fourier Transform (FFT) algorithm. Weighted spectral information of the overlap cuts pairs to take the maximum spectral position as the offset of the adjacent image tiles. The image tiles are stitched using the offset value.

c. Registration with spatial gene expression map and stitched image based on "track lines". The spatial gene expression map is the fixed image, and the stitched image is the moving one. We detect the "track lines" on two images, the "track lines" on the spatial gene expression map represent the template, for the stitched image, we calculate the scale and rotation parameters using the "track lines". We then use the morphological features to get a rough registration of the offset, flip, and 90° rotation. "Track lines" are used to fine-tune registration.

d. Bi-Directional ConvLSTM U-Net is used to obtain the tissue mask on registered image.

e. U-Net is used to obtain the nuclei mask on registered image.

f. Molecules within the nuclear boundary are assigned to a given cell, while molecules outside the nucleus are labeled to the cell that gains the largest probability from Gaussian Mixture Model fitting.

2.1.2 Advantages of the protocol

Although spatially resolved technologies vary in resolution and field-of-view, one common feature is their ability to produce traditional high-content tissue images using dyes such as fluorescence, hematoxylin-eosin (H&E) or 4,6-diamidino-2-phenylindole (DAPI) for cellular or nucleus staining. The generated images provide important information regarding tissue and cell morphology, however, two main difficulties exist: insufficient precision in image mosaics and inaccuracy of molecule labeling.

The first difficulty is limited by the small imaging area, mechanical tolerance, and imperfect calibration of microscopes (e.g., 10×imaging lens, ~1mm×1mm/tile), which makes them sufficient for visual inspection of tissue images, but not accurate enough for quantitative single-cell analytical approaches. For example, a 1cm×1cm object generates approximately 10×10 image tiles, while a 3cm×3cm object generates thousands of image tiles. Existing image stitching methods, such as ASHLAR[64] and MIST[124], can meet most stitching accuracy requirements. However, stitching misalignment is a major issue and it is difficult to collect evaluation datasets to improve the accuracy of image mosaics in a large field-of-view, which is an obstacle in expanding their application.

The second difficulty, which is the key to achieving high-precision single-cell resolution, is classifying the transcripts and other molecules into the owner cell, a procedure termed “molecule labeling”. However, molecule labeling is not an easy task due to the low efficiency of molecule capture and the problem of diffusion of molecules outside cells during wet-lab procedures, which greatly affects the correct annotation of transcripts to the corresponding cell. Recent technology has considered this issue, such as 10X Xenium[116], which includes more transcripts by extending the distance outside the nucleus (15 μm). Since the diameter of immune, stromal, and tumor cells varies, it is however challenging to obtain accurate single-cell data using such a generalized model that introduces varying levels of noise depending on the cell type. Some available methods, including Baysor[69] and Pixel-seq[23], require high-quality data as input, such as the number of captured molecules and their density distribution, which may not be applicable to most sequencing-based spatial omics data. Furthermore, existing spatial data analysis frameworks, such as Seurat[83], RCTD[71], Cell2location[72], Squidpy[125], and Spateo[126], only focus on the downstream analysis tasks or one specific aspect of the spatial data analysis, or consider only cell nuclei segmentation.

StereoCell is the first one-stop software to generate single-cell spatial gene expression profiles for whole transcriptome based on high-resolution and large field-of-view spatially resolved sequencing. For image stitching, StereoCell employs MFWS to reduce stitching errors in large-field-of-view datasets. The morphological tissue image stitching procedure of StereoCell is accurate and reliable for single-cell identification and is flexible and convenient in terms of run time. Its high-precision stitching algorithm is useful for the correction of stitching to almost subcellular precision. For molecule labeling, StereoCell employs MLCG to increase the signal-to-noise ratio of the single-cell spatial gene expression profile, which yields more reliable analysis of cell clustering and annotation. StereoCell has been successfully applied to datasets of various organs (such as brain, heart, embryo, artery, testis, kidney, liver and lymph) from different organisms (such as *Homo sapiens*, *Mus musculus*, *Macaca mulatta* and

Leporidae). The molecule labeling of StereoCell is also applicable to the datasets from multiple platforms other than Stereo-seq when the nuclei mask and spatial gene expression data are given. StereoCell's rich documentation in the form of a functional application programming interface, examples, and tutorial workflows, is easy to navigate and accessible to both experienced developers and beginner analysts[127]. StereoCell has the potential to serve as a bridge between the fields of image analysis and molecular omics, providing a foundation for the development of next generation computational methods for spatially resolved technologies.

2.1.3 Applications of the protocol and comparison with other methods

2.1.3.1 StereoCell processes high-precision morphological tissue images from image tiles

MFWS of StereoCell takes a folder of all image tiles obtained by microscopy and information files as the input, and outputs a stitched mosaic image of whole tissue. The datasets from different field-of-views (4 Stereo-seq mouse brain datasets with the chip sizes of 1cm×1cm, 1cm×2cm, 2cm×2cm and 2cm×3cm respectively and a public dataset[128], and their grid sizes (11,9), (15,21), (25,21), (23,29) and (10,10), respectively) are collated. The image stitching methods, ASHLAR[64] and MIST[124], are used for comparison. For each dataset, the standards are designed (Supplementary materials) to calculate the relative offset error between each two adjacent image tiles and the absolute offset error of the entire stitched image in the stitching results of different methods. The relative offset errors are statistically analyzed by Wilcoxon signed rank test[129] to examine the significant differences. The runtime to obtain the absolute offset error for each method is recorded.

Processing of a morphological image from a tissue slide requires the stitching of an array of multiple image tiles generated by microscopy. Microscopes can automatically capture the image tiles of a tissue one by one and stitch the image tiles together using the built-in stitching method. However, overlapping areas between adjacent image tiles generated during microscope movement may be imprecise due to mechanical tolerance

and imperfect calibration. Such stitching errors are common in mosaic images and need to be removed if the goal is to achieve single-cell resolution in spatial omics applications. As an example, a mouse brain mosaic image on a chip with a size of ~2cm×2cm and a tile size of 2424×2031 pixels are displayed (Fig. 2a), corresponding to a physical size of ~1.2mm×1.0mm. To visualize seams and stitching errors, we select two adjacent image tiles from the image (Fig. 2b). For accurate stitching results, the lower part of tile 1 (dotted area) and the upper part of tile 2 (dotted area) should be accurately overlaid. Shadows that represent inaccurate overlap of cells can be easily seen in the stitching results produced by the microscope (Fig. 2c, left), but MFWS of StereoCell is able to stitch the image tiles accurately, resulting in the absence of shadows (Fig. 2c, right). The stitching results of two image tiles within non-tissue areas (full-colored part) where individual "track lines" can be clearly seen using auto-contrast adjustment as also shown (Fig. 2c). The applied Stereo-seq chip contains periodic "track lines" with a size of 1500 nm (~3 pixels), which is apparent from the image (Fig. 2c, f). We use a collection of data to evaluate the accuracy and efficiency of existing stitching algorithms as compared with MFWS.

We apply MFWS to the public dataset, and the results show that the relative offset errors of MFWS are comparable with those of MIST, both being concentrated within 5 pixels, while ASHLAR has larger offsets (>10-pixels) (Fig. 2d). Moreover, the offset error distribution is much more concentrated by MFWS. Next, we calculate the relative and absolute offset errors on 4 Stereo-seq mouse brain datasets. MFWS is shown to perform significantly better than ASHLAR and MIST with respect to the relative offset errors for all image size combinations (Fig. 2e). A 10-pixel (~5 μm) dislocation roughly corresponds to half a cell (Fig. 2f). In all the datasets, the errors generated by MFWS remain within a maximum of 10-pixels in the tissue area, while the other methods produce a shift greater than 10-pixels, which may misplace the cell (Fig. 2g). As the number of image tiles increases, the run time becomes a significant factor for stitching algorithms. The run time of MFWS is significantly shorter than that of the other

methods, mostly due to the embedded spectral calculation based on FFT (Fig. 2h).

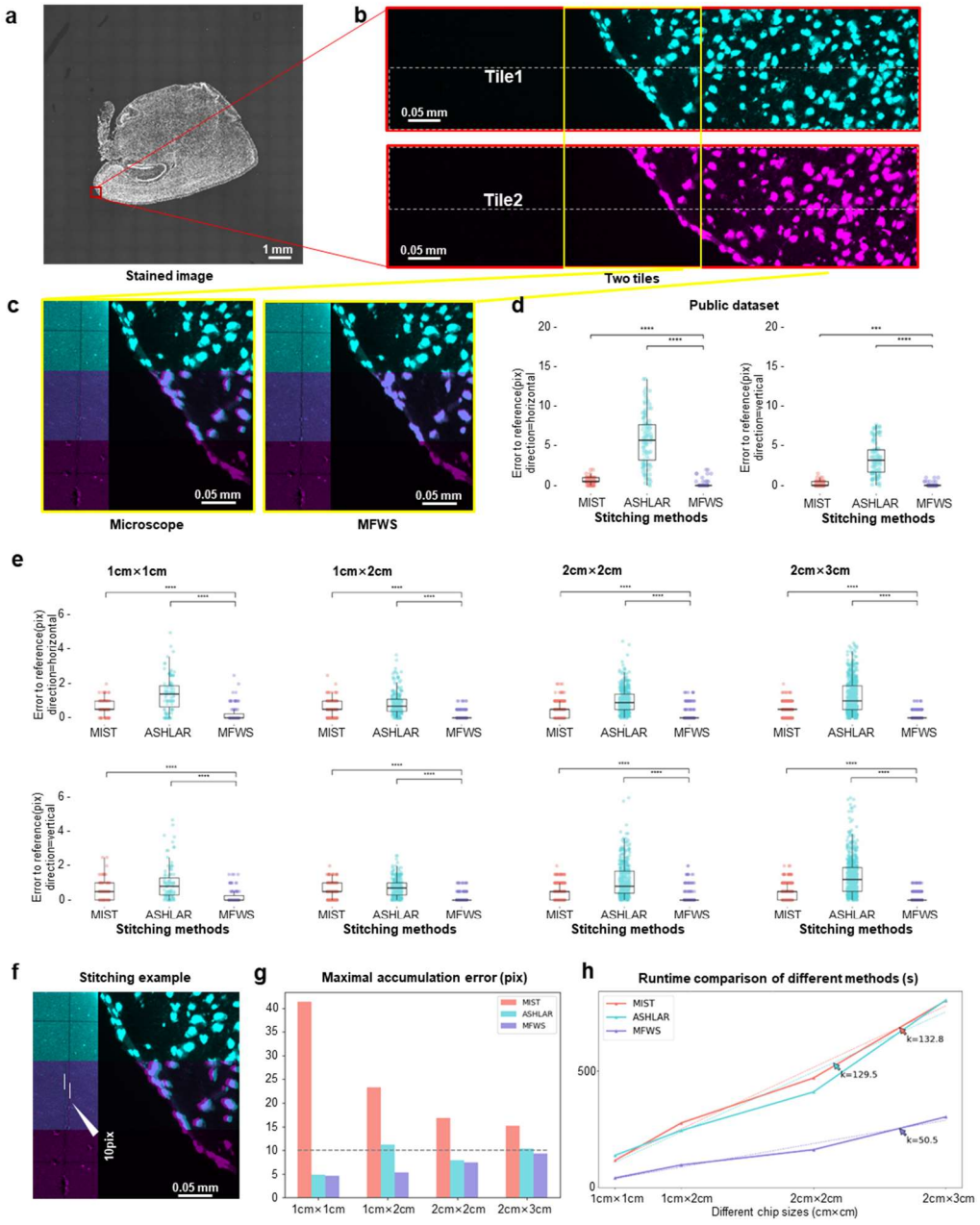

Fig. 2 | StereoCell includes a high-precision mosaic method for image stitching, enabling large field-of-view morphological image analysis at single-cell spatial accuracy.

a. Mosaic image of mouse brain tissue on a 2cm×2cm chip.

b. The red box indicates the edge of the tissue (containing cells and non-tissue), and two neighboring image tiles (in the vertical direction) are shown in different colors. The part that needs to be overlaid by stitching is the lower part of tile 1 (within the dotted area) and the upper part of tile 2 (within the dotted area).

c. The stitching results for the two image tiles (yellow box in **b**) using the stitching method built in to the

microscope (left) and MFWS (right). In each sub-figure, the left part shows the overlap of cells after stitching and the right part shows the "track lines" in the non-tissue area. The contrast of the background was increased to clearly show the "track lines" incorporated into the Stereo-seq chip.

d. Comparison of the relative errors produced by MIST, ASHLAR, and MFWS on a public dataset.

e. Comparison of the relative errors produced by MIST, ASHLAR, and MFWS on Stereo-seq mouse brain datasets, analyzed using different chip sizes from 1cm×1cm to 2cm×3cm, corresponding to number of image tiles from 11×9 to 23×29. The evaluation is based on the ground truth calculated using the "track lines" on the Stereo-seq chip.

f. An example of the stitching error in pixels in the zoomed-in image. A dislocation of 10-pixels roughly corresponds to half a cell.

g. Bar graph illustrating the maximal accumulation of stitching errors produced by MIST, ASHLAR, and MFWS on the datasets from **e**.

h. Line graph displaying the run time of MIST, ASHLAR, and MFWS on Stereo-seq mouse brain datasets.

2.1.3.2 StereoCell generates highly accurate single-cell gene expression profiles of spatial omics datasets

Using StereoCell, we first obtain a nuclei mask and a cell mask, and then output their single-cell spatial gene expression profiles. To demonstrate the improvement in transcript assignment to cells by using the cell mask generated by StereoCell, we here compare the gene expression profile output using each of these masks on Stereo-seq data generated of mouse olfactory bulb, involving analysis of spatial gene expression data containing 37,288,344 molecules and 143 image tiles. The generated profiles are input into Stereopy (v6.0)[130] (a downstream analysis tool) for analysis. The silhouette coefficient[131] is used to evaluate clustering results. The moran's I (calculating by Scanpy[85] package) is used to evaluate spatial correlation. The details of the data analysis process and evaluation metrics are provided in Supplementary materials.

After obtaining the nuclei mask, the molecules located inside the nucleus are assigned to the corresponding cell, and MLCG of StereoCell fits the molecular density distribution to each cell nucleus and re-labels the molecules outside the nucleus to finally generate a cell mask (Fig. 3a). By applying the cell mask, the uniquely expressed genes and total gene counts in a single cell increased by approximately 2.34 and 2.56 times, respectively, as compared to the nuclei mask alone (Fig. 3b). Utilization of the

cell mask provides better overall clustering (Fig. 3c), and its silhouette coefficient is higher than that of the nuclei mask. We manually annotate the cell types by comparing the differentially expressed genes in each cluster with the marker genes of cell types inferred by a reference dataset[132]. Of note, there are fewer scattered points and a more concentrated distribution in the spatial position of the astrocyte layer (Moran's I: 0.40 vs. 0.30) and dopaminergic neuron layer (Moran's I: 0.52 vs. 0.37) after molecule labeling using the cell mask (Fig. 3c). We further explore the marker gene distribution and expression in the annotated cell types, and all the marker genes provided by the reference dataset[132] display higher expression using the cell mask as compared with the nuclei mask (Fig. 3d). We compare subtypes of granule cells discovered using the nuclei mask vs. the cell mask. The nuclei mask is able to identify two subtypes of granule cells: granule cell 0 and granule cell 3, while the cell mask enables identification of three granule cell subtypes: granule cell 0, granule cell 1, and granule cell 3. We explore the expression of the granule cell 1 subtype-marker genes Syt1, Scg2, and Cplx1, among all granule cells. There is no obvious difference in expression when using nuclei mask, while higher expression is found in granule cell 1 as compared with the other two subtypes when using cell mask (Fig. 3e). This supports that the cell mask captures a higher number of transcript signals from rare cells, facilitating better annotation and thereby enhances single-cell spatial resolution.

We estimate and compare the similarity of spatial gene expression between a reference single-cell dataset of the mouse olfactory bulb[26] and our Stereo-seq based mouse olfactory bulb dataset using the StereoCell nuclei mask vs. the cell mask to generate single-cell expression profiles. Tissue cells are auto-annotated by Tangram[77] and correlation analysis is performed between the expression matrixes of single cells using each of the masks and the single-cell sequencing expression profile of the reference[132]. The Spearman correlation coefficient using annotations based on the cell mask is higher (0.4% to 8.9% better) than that based on the nuclei mask. Overall, the single-cell expression matrix generated by the cell mask shows a higher correlation with the single-

cell reference (Fig. 3f), indicating that the cell mask provides a spatial single cell-level gene expression profile closer to that generated by single-cell sequencing.

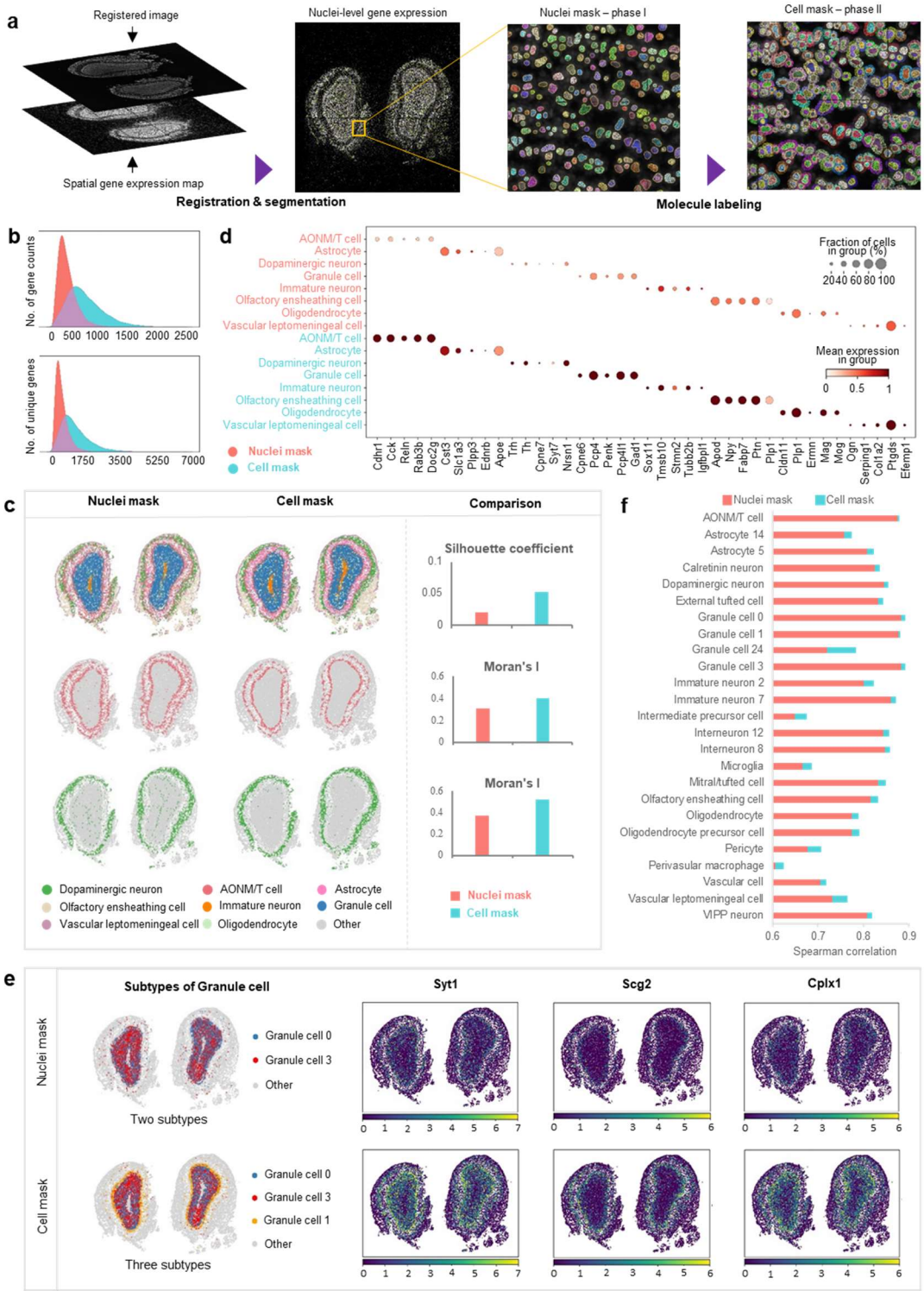

Fig. 3 | StereoCell provides single-cell spatial data with a higher signal-to-noise ratio that facilitates finer cell clustering and annotation of the mouse olfactory bulb.

a. Flowchart of MLCG processing of Stereo-seq mouse olfactory bulb dataset.

b. Comparison of the number of uniquely expressed genes and the total gene count per cell using nuclei mask vs. cell mask on Stereo-seq mouse olfactory bulb dataset. Top: density plot of the total gene counts per cell. Bottom: density plot of the number of uniquely expressed genes per cell.

c. Comparison of the clustering results generated using the "Leiden" algorithm on nuclei mask vs. cell mask. Top row, left: spatial clustering results of mouse olfactory bulb data generated using nuclei mask vs. cell mask; right: comparison of unique cells estimated by Silhouette coefficient. Middle and bottom rows, left: spatial distribution of cells in the inferred clusters of astrocytes and dopaminergic neurons generated using nuclei mask vs. cell mask; right: comparison of the spatial autocorrelation estimated by Moran's I.

d. Comparison of marker gene expression in the main cellular clusters inferred by the "Leiden" algorithm generated using nuclei mask vs. cell mask.

e. Comparison of cellular subtype identification using nuclei mask vs. cell mask. First column: subtypes of granule cells identified using nuclei mask vs. cell mask. Other columns: gene expression heat maps of Syt1, Scg2, and Cplx1 respectively, which have been reported to be marker genes for granule cell 1, using nuclei mask vs. cell mask.

f. Spearman correlation between gene expression in similar cells in tissue slides of the mouse olfactory bulb using nuclei mask vs. cell mask to define the gene expression in single cells and a single-cell reference dataset of the mouse olfactory bulb.

c, **d**, **f**: AONM/T cell: anterior olfactory nucleus mitral/tufted cell, VIPP neuron: vasoactive intestinal peptide positive neuron.

2.1.3.3 StereoCell enables dissection of the structural composition of mouse brain cortex data at single-cell resolution

Non-cell based binning methods where adjacent tissue regions are divided into regions (bins) of specified sizes, are sometimes used in spatial transcriptomics analysis pipelines. To compare the StereoCell cell mask output to the Bin approach, we here apply a Stereo-seq mouse brain dataset containing 131,990,020 molecules and 117 image tiles, where image tiles where split using Bin100, Bin50, Bin20 (Bin$X$ means a bin with $X \times X$ of DNBs[47]). The generated profiles are input into Stereopy (v6.0)[130] for analysis, and more details are provided in Supplementary materials. We export the spatial gene expression map, nuclei-stained image, cell mask and Bin20 outlines to visually demonstrate the segmentation effect of the two methods. We also compare the abilities of StereoCell and the Bin approach to reconstruct known cellular regions in the mouse brain.

The clustering of data generated by StereoCell vs. Bin100, Bin50, and Bin20 is shown (Fig. 4a). It appears that splitting using Bin100 result in difficulties in identifying several important areas in the tissue, such as the brain hub and hippocampus. For Bin50, the tissue cortex and blood cells are poorly identified. Both Bin20 and StereoCell are able to identify important areas in the tissue, but StereoCell obtains the highest silhouette coefficient in evaluating the clustering results using several methods (Fig. 4b). We calculate the uniquely expressed genes and total gene counts of StereoCell vs. the differently sized bins (Fig. 4c). The visualization shows that Bin20 more often divides a single cell into two or more cells, while StereoCell more accurately divides the cell area and is consistent with the actual cell distribution in the tissue (Fig. 4d). Bin20 results in splitting of approximately 90% of the cells, with only ~10% of the nuclei being completely covered in the nuclei-stained image, while only ~2% of the cells are split into two or more cells using StereoCell (Fig. 4e). The resulting single-cell data derived from cell identification using Bin20 and StereoCell were individually annotated using Spatial-ID[133] (Fig. 4f, top) with adolescent mouse brain as a reference[134]. Bin20 annotates 29 different cell types, while StereoCell is able to annotate 37 different cell types. Within the ACTE series, both Bin20 and StereoCell annotate 2 subtypes. In the MEINH series, Bin20 annotates 2 subtypes, while StereoCell annotates 3 subtypes. Within the TEGLU series, Bin20 and StereoCell annotate 10 and 13 subtypes, respectively. In TEINH series, Bin20 annotates 5 subtypes and StereoCell annotates 4 subtypes, which is the only case where fewer subtypes were annotated by StereoCell. StereoCell also annotates 2 subtypes in within the ACNT series and 4 subtypes within the TEINH series. When zooming in on the cortical area (Fig. 4f, bottom), it appears that the staining position of StereoCell and the nucleus are better aligned than seen for Bin20, and the cell state and tissue structure are more in line with the actual brain tissue map. Moreover, in the gear gyrus and cortical regions of the mouse brain, StereoCell performs better than Bin20 in matching the annotation results to the Allen mouse brain dataset (Fig. 4g). Although Bin20 is capable of annotating the

gear gyrus and cortex to align with the location in the Allen mouse brain atlas, the division between layers is not as accurate. Also, expression of marker genes in the gear gyrus (DGGRC2) and cortex (TEGLU3, TEGLU4, TEGLU7, and TEGLU8) is fully captured using the StereoCell algorithm, which is not the case when using Bin20 (Fig. 4h). Overall, StereoCell provides single-cell spatial gene expression profiles with a higher signal-to-noise ratio compared to approaches based on differently sized bins. Moreover, the profile generated using StereoCell align well with cellular annotation within different brain regions and can provide a valuable reference for studies of cellular interaction networks in health and disease.

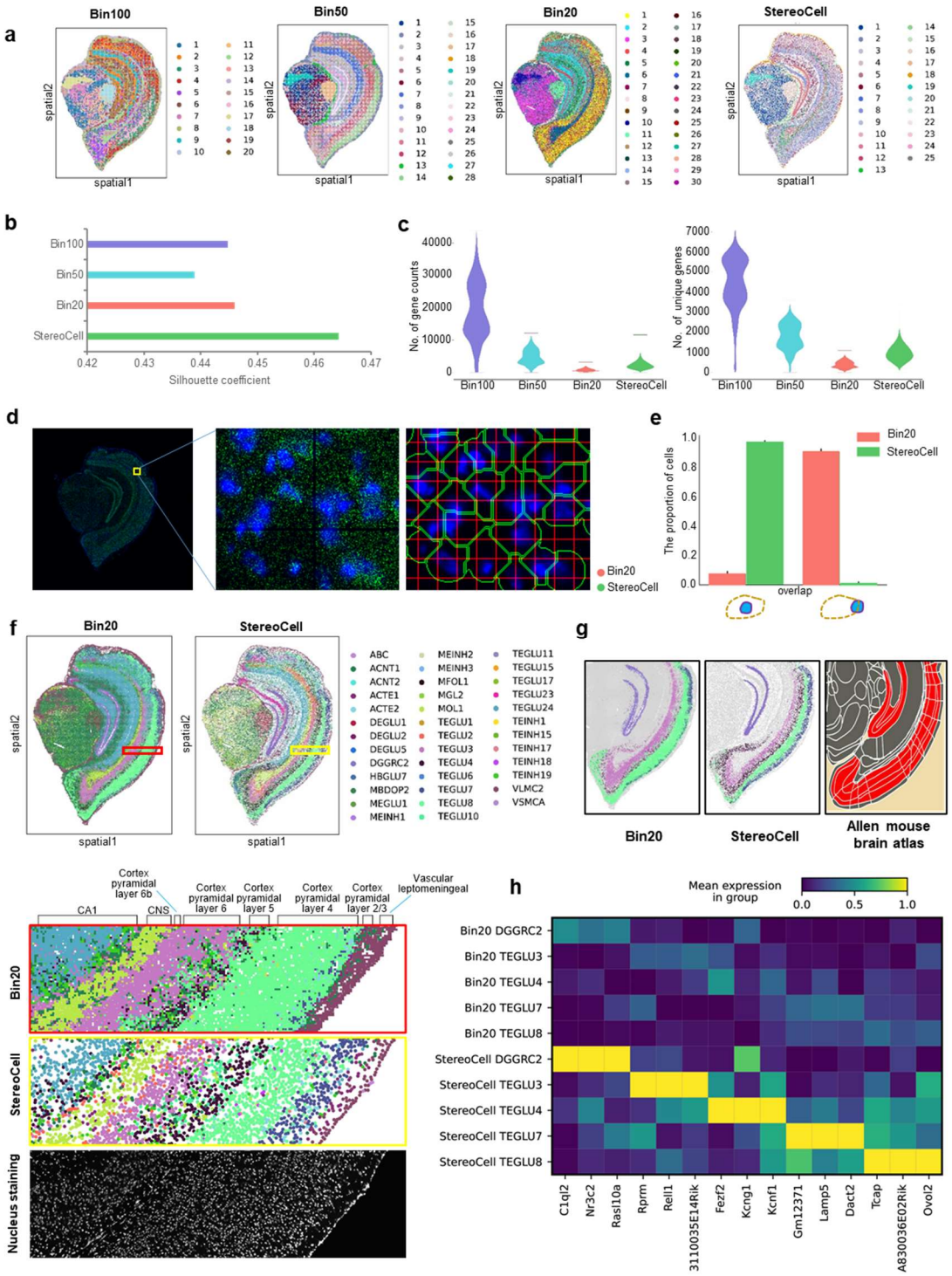

Fig. 4 | Clustering and annotation of mouse brain tissue by StereoCell vs. differently sized bins.

a. Identification of gene expression clusters using differently sized bins vs. StereoCell on Stereo-seq mouse brain dataset.

b. Silhouette coefficient evaluation of clustering performance.

c. Violin plots displaying the total gene counts and number of uniquely expressed genes per cell captured using differently sized bins vs. StereoCell.

d. Comparison of the cell boundary generated using Bin20 and StereoCell. The first image shows the

spatial gene expression map and nuclei-stained image, which are merged in green and blue, respectively. The second image displays an enlarged map of the local area from the first image. The third image is a comparison of the cell boundaries obtained using Bin20 (red) vs. StereoCell (green).

e. Comparison of the proportion of cell nuclei covered or intersected using Bin20 vs. StereoCell.

f. Comparison of end-to-end annotation results of Bin20 vs. StereoCell. Colors represent the same cell type in the Bin20 and StereoCell annotation charts. The enlarged local comparison map of the cortex shows the spatial distribution and boundary information of different cortical cells (bottom sub-figure).

g. Comparison of cell annotation based on Bin20 vs. StereoCell using the Allen mouse brain atlas as a reference. The spatial distribution of the gear gyrus (DGGRC2, black) and the cortex (TEGLU3, TEGLU4, TEGLU7, and TEGLU8) is shown.

h. Heat map displaying the consistency of marker gene expression within the different substructures in **g** (gear gyrus (DGGRC2) and cortex (TEGLU3, TEGLU4, TEGLU7, and TEGLU8) layers)). The upper part: results of Bin20, lower part: results of StereoCell.

2.1.4 Limitations

StereoCell has quality requirements for the input data. For Stereo-seq transcriptomic data, the gene expression counts in each Bin200 should not be less than 5000. For the image tiles taken by the microscope, the "track lines" clarity is required to pass our image quality control.

2.2 Materials

2.2.1 Datasets

We provide a demo dataset for testing StereoCell, which has been used for the experiment of dissecting the structural composition of mouse brain cortex data at single-cell resolution and includes spatial gene expression data with 131,990,020 molecules and 117 image tiles. The demo dataset can be downloaded on our Github repository (given in the Code availability section).

2.2.2 Methods

2.2.2.1 Image stitching

The image tiles are stitched into a large mosaic image of the whole tissue by MFWS (Fig. 1b). The file name of each image tile needs to reflect the row and column, such as "0000_0001.tif" that reflects a tile at row 0 and column 1. Firstly, we calculate the actual overlap value with the neighbors in the horizontal and vertical directions for each image tile. A pair of image tiles in the vertical direction is taken as example. (i) The

overlapping regions of adjacent image tiles are marked as A_f and A_m respectively, and they are divided into N sub-region images as marked $\{f_i\}$ and $\{m_i\}$ respectively. (ii) The FFT[122] algorithm is applied to obtain $2\times N$ frequency spectra based on $\{f_i\}$ and $\{m_i\}$. (iii) The sub-region images in A_f and A_m are paired, such as f_i and m_i, and N cross-power spectra are obtained using the formula for image cross-power calculation. Theoretically, any one of these spectra can reflect the true offset information. (iv) Due to the low signal-to-noise ratio of some sub-regions in the previous step, the calculated value differs greatly from the actual value. The overall accuracy of the algorithm can be significantly improved by weighted enhancement of partial blocks and reduction of residual sub-blocks. Our weighting coefficient is based on the peak value of the mutual power spectrum because experiments show that the greater the peak value, the higher the precision and reliability of the overlap value. A unique cross-power spectrum is obtained after the weighted superposition of N spectra. Then, the corresponding spatial domain correlation graph can be obtained through Inverse FFT. The coordinate of the peak value represents the required offset value. (v) Carrying out the above steps for each pair of adjacent image tiles to generate the offset matrices O_H and O_V. Secondly, transformation of local coordinates to stitching coordinates is required, since image stitching demands the unique coordinates of each tile in the reference system, *i.e.,* the global coordinate matrix L. In the O_H and O_V matrixes, offsets corresponding to low-confidence values are not credible that need to be eliminated. However, it leads to the existence of single or multiple connected domains. Therefore, when obtaining the coordinates, we first find the center of each connected domain, complete the splicing in the connected domain according to the relative relationship among the neighbors, and then use the experience value to fill multiple connected domains to complete the splicing of the entire image. Thirdly, according to the image tiles and the coordinate matrix L, the size of the mosaic image is obtained, the value is traversed, and the seam fusion is completed synchronously during the stitching process.

2.2.2.2 Image registration

The spatial gene expression data is read and transformed into a map, in which the intensity value of each pixel is proportional to the total count of molecules of all genes expressed at position DNB. The stitched image is registered to the transformed map (Fig. 1c). Sequencing chips with sample tissue attached, imaging, and sequencing are designed with periodic "track lines" (horizontal and vertical) to assist base calling and image registration. The "track lines" are displayed as dark straight lines in both the stitched image and transformed map. Image registration is achieved using the transformed map as a template and performing an affine transformation on the stitched image. The "track lines" are detected in the transformed map and stitched image by line searching algorithms, and the intersections of the horizontal and vertical "track lines" are located. The scaling parameter is calculated by comparing the length of the line segments between intersections within the transformed map and stitched image. The rotation angle is defined as the difference between the horizontal "track lines" and the horizontal direction of the image coordinate system. For the offset between them, since the "track line" patterns are periodic, we first calculate the center of gravity of the tissue regions within the transformed map and stitched image, roughly match them together, and then match the "track lines" intersections to fine-tune the image registration.

2.2.2.3 Tissue segmentation

The tissue can be segmented from the registered image. The tissue segmentation process consists of two steps: preprocessing and model inference (Fig. 1d). In addition, a self-scoring mechanism is added to optimize the processing time and eliminate the effort of manually checking the results. The histogram enhancement method is used to improve the separation of tissue and background during the preprocessing, and then, Bi-Directional ConvLSTM U-Net[135] is used to predict tissue masks during the model inference. Compared with the original U-Net[136], the following improvements are incorporated. (i) The convolution sequence is replaced with a densely connected convolution in the final step of the encoding path to encourage feature reuse. (ii) The

feature maps extracted from the corresponding encoding path are combined with the previous decoding up-convolutional layer in a non-linear manner. (iii) The batch normalization is added to accelerate the convergence of the network. When dealing with a large amount of unstable data, the model cannot achieve ideal results with very complex or poor-quality tissue images, and the self-evaluation mechanism help to filter these potentially unsatisfactory results. When the registered image possesses a high gray value, we use the threshold value obtained in the histogram enhancement step as a reference value and assume that the pixel points higher than this value are valid pixels. The effective pixel densities of the tissue inside the mask and in the area surrounding the mask are calculated in sections, and the ratio of the two values is used as the basis for evaluation.

2.2.2.4 Nuclei segmentation

The nuclei can be determined from the registered image due to nuclei staining. Nuclei segmentation consists of three steps (Fig. 1e). The first step is image preprocessing. The median filtering is employed to smooth the noise that may be present in the input image. To both enhance the cell and homogenize the image background, the output of the median filtering operation is input into the pixel-wise subtraction process. This step facilitates the relative conspicuity of nuclei against the background. The second step is segmentation model inference. We use U-Net[136] for segmenting the cells. Some optimizations are made to U-Net that, Residual U-Net[137] is used as a feature extractor and the Pyramid Squeeze Attention[138] module is used instead of the 3×3 convolution in the bottleneck as the basic feature-extracting unit of Residual U-Net, which ensures the model to pay more attention to the highlighted feature representation. The third step is mask post-processing. To obtain a more reliable segmentation, the mask generated by the segmentation model is input to the area filtering operation. Further, the opening operation (erode and dilatate) is applied to revise the cells shape and boundaries.

2.2.2.5 Molecule labeling

The single-cell spatial gene expression profile can be obtained by combining the nuclei

mask and spatial gene expression data (Fig. 1f). Firstly, the morphological tissue image is used to identify nuclear boundaries as described in nuclei segmentation. The specific transcripts that are contained within each nucleus are then assigned. A GMM[123] algorithm is used to estimate the probability of each transcript belonging to a given cell based on the nuclei mask. It is performed by modeling each cell as a GMM distribution, combining its spatial position, transcript count, and density. Simultaneously, we also estimate the probability scores of transcripts in the acellular region. In our model, the probability of assignment of a transcript to a given cell declines progressively with distance from the cell center, but increases with transcript count and density. A mixture model is a probabilistic model that can be used to represent a population distribution with K sub-distributions. GMM can be regarded as a model composed of K single Gaussian models. Taking a single cell as an example, based on the result of nuclei segmentation, we can locate the spatial position of the nucleus. Firstly, GMM is used to fit the molecular distribution of the current cells. We expand the nuclei mask boundary and captured as much as possible of the true distribution of data in cells rather than nuclei to develop the cell mask. According to the empirical value of the cell area, we finally determine the fitting range of 100pixels×100pixels. Subsequently, GMM is used to calculate the probability score of extracellular molecules within the fitting range. Finally, according to the maximum and minimum probability scores within the fitting range, the quartile is used as the adaptive threshold of the current cell. When the probability score is greater than the threshold, the molecule is re-divided into cells. In this way, molecules are assigned to their corresponding cell with high confidence.

2.2.3 Software

Operating system: Windows or Linux

Source package and environment management system: Conda

Script writing: Python v3.8

2.2.4 Hardware

StereoCell can be used on desktops, laptops, workstations, computer clusters or cloud

computing platforms. **CRITICAL:** To ensure the normal operation of StereoCell, we recommend a hardware configuration of at least 8-core CPU and 16 GB of RAM. The use of GPU can greatly accelerate the tissue and nuclei segmentation steps. The amount of disk storage required depends on the input dataset. Our experiment on mouse brain dataset is performed on a server with 40-core CPU, 128 GB of RAM and 24 GB of GPU, and we reserve at least 6 GB of disk storage for all results.

2.3 Procedure

CRITICAL: Before using StereoCell, make sure that "Conda" has been installed, and the "Python" environment has been configured and activated (detailed steps are given in the GitHub repository). StereoCell can be performed in one-stop, and any one step of StereoCell can also be performed individually according to users' requirements. Taking the demo dataset (Stereo-seq mouse brain dataset) as an example, running StereoCell one-stop process by command:

```
python stereocell.py
--tiles_path /data/SS200000135TL_D1
--gene_exp_data /data/SS200000135TL_D1.gem.gz
--output_path /result
--chip_no SS200000135TL_D1
```

where `--tiles_path` is the path of all image tiles, `--gene_exp_data` is the compressed file of spatial gene expression data, `--output_path` is the output path, and `--chip_no` is the chip number of the Stereo-seq transcriptomic data.

2.3.1 Image quality control ● Timing ~1.5 min

1. **CRITICAL:** The image tiles are read and the clarity of their "track lines" are detected by image quality control. Only the detection has passed, the program proceeds to the next step. The detection result prints: "Image QC: PASS" if passed, and "Image QC: FAIL" if failed. This step can also be performed individually by command:

```
python qc.py
--tiles_path /data/SS200000135TL_D1
```

--chip_no SS200000135TL_D1

where `--tiles_path` is the path of all image tiles, and `--chip_no` is the chip number of the Stereo-seq transcriptomic data.

2.3.2 Image stitching ● Timing ~2 min

2. **CRITICAL:** The stitched image is output as an intermediate result in StereoCell one-stop process. This step can also be performed individually by command:

   python stitching.py
   --tiles_path /data/SS200000135TL_D1
   --output_file /result/stitched_image.tif

   where `--tiles_path` is the path of all image tiles, and `–output_file` is the stitched image.

2.3.3 Image registration ● Timing ~3.5 min

3. **CRITICAL:** The registered image is output as an intermediate result in StereoCell one-stop process. This step can also be performed individually by command (since the registration requires "track lines", an additional image quality control is performed when performing this step individually):

   python registration.py
   --image_file /result/stitched_image.tif
   --output_file /result/registered_image.tif
   --gene_exp_data /data/SS200000135TL_D1.gem.gz
   --chip_no SS200000135TL_D1

   where `--image_file` is the stitched image, `--output_file` is the registered image, `--gene_exp_data` is the compressed file of spatial gene expression data, and `--chip_no` is the chip number of the Stereo-seq transcriptomic data.

2.3.4 Tissue segmentation ● Timing ~10 s

4. **CRITICAL:** The tissue segmentation (tissue mask) image is output as an intermediate result in StereoCell one-stop process. This step can also be performed individually by command:

   python segmentation.py
   --type tissue
   --image_file /result/registered_image.tif
   --output_file /result/tissue_mask.tif

   where `--type` can be "tissue" or "nuclei", `--image_file` is the registered image, and `--output_file` is the tissue mask image.

2.3.5 Nuclei segmentation ● Timing ~9 min

5. **CRITICAL:** The nuclei segmentation (nuclei mask) image is output as an intermediate result in StereoCell one-stop process. This step can also be performed individually by command:

   python segmentation.py
   --type nuclei
   --image_file /result/registered_image.tif
   --output_file /result/nuclei_mask.tif

   where `--type` can be "tissue" or "nuclei", `--image_file` is the registered image, and `--output_file` is the nuclei mask image.

2.3.6 Nuclei mask filtering ● Timing ~10 s

6. **CRITICAL:** Nuclei mask filtering is an optional step. Due to poor image quality or overflow of some molecules during gene capturing, there may be some impurities outside the tissue, which may be misclassified within the nuclei mask. This step uses the tissue mask image to filter out the impurities outside the tissue in the nuclei mask image, which can improve the quality of nuclei masks. The filtered nuclei mask image is output to replace the original nuclei mask image (changing the name of the output file can prevent the filtered image from replacing the original image). This step is not included in the StereoCell one-stop process and it can be performed individually by the following command if required:

   python filtering.py
   --nuclei_mask /result/nuclei_mask.tif
   --tissue_mask /result/tissue_maks.tif
   --output_file /result/nuclei_mask.tif

   where `--nuclei_mask` is the nuclei mask image, `--tissue_mask` is the tissue mask image, and `--output_file` is the filtered nuclei mask image.

2.3.7 Molecule labeling ● Timing ~63 min

7. **CRITICAL:** The single-cell spatial gene expression profiles based on nuclei mask and cell mask, respectively, are output. This step can also be performed individually by command:

   python labeling.py
   --image_file /result/nuclei_mask.tif
   --gene_exp_data /data/SS200000135TL_D1.gem.gz
   --output_path /result

   where `--image_file` is the nuclei mask image, `--gene_exp_data` is compressed file of spatial gene expression data, and `--out_path` is the output path.

2.4 Troubleshooting

Troubleshooting advices are found in Table 1.

Table 1 | Troubleshooting table

Step	Problem	Possible reason	Solution
1	IndexError: list index out of range	The information files or weight files are missing.	Make sure the input path contains a folder of all image tiles and two information files (1.ini and info.ini). Or make sure the weight files have been moved into the specified folder according to our GitHub repository instructions.
3	TypeError: only size-1 arrays can be converted to Python scalars	The inappropriate version of the "pyvips" package is installed.	Uninstall the "pyvips" package and reinstall it with version of 2.2.1, use "PyPI" for Windows system, and use "Conda" for Linux system.
2-6	IndexError: list index out of range	The expanded name of the input or output file is missing	Give the expanded name of the input or output file (such as ".tif") when executing the command.
	IsADirectoryError: [Errno 21] Is a directory: '…'	The file name of the input or output is missing.	Give the file name of the input or output instead of just a path when executing the command.

2.5 Timing

Step 1, image quality control: ~1.5 min

Step 2, image stitching: ~2 min

Step 3, image registration: ~3.5 min

Step 4, tissue segmentation: ~10 s

Step 5, nuclei segmentation: ~9 min

Step 6, Nuclei mask filtering: ~10 s

Step 7, Molecule labeling: ~63 min

2.6 Anticipated results

A total of six files are output from StereoCell (see details in Table 2). Due to the time spent in generating the cell mask image, it is not included as output files from the StereoCell one-stop process. When any step of StereoCell is performed individually, the corresponding file is generated as an output. When an error occurs in a certain step during program execution, it will not affect the intermediate results output by the previous steps. That is, after the error is resolved, the intermediate results can be used to continue performing subsequent steps to obtain the final result.

For the generated single-cell spatial gene expression profiles, both "nuclei_mask_profile.txt" and "cell_mask_profile.txt" files contain 5 columns, namely "geneID", "x", "y", "MIDCount" and "CellID", which represent the ID of gene, coordinate x, coordinate y, gene expression count and the ID of assigned cell respectively. In "nuclei_mask_profile.txt", "CellID" of "0" means that the molecule is outside the nuclei, *i.e.,* this molecule is not assigned to any cell. In "cell_mask_profile.txt", a large number of molecules outside the nuclei are assigned to cells through MLCG, thus "CellID" of these molecules is no longer "0", and the remaining molecules with "CellID" of "0" are removed to facilitate downstream analysis.

Table 2 | Output details

Step	Output file	Description
1 Image quality control	NA	No file output and just print "Image QC: PASS" or "Image QC: FAIL"
2 Image stitching	stitched_image.tif	The stitched image

3 Image registration	registered_image.tif	The registered image
4 Tissue segmentation	tissue_mask.tif	The tissue mask image
5 Nuclei segmentation 6 Nuclei mask filtering	nuclei_mask.tif	The nuclei mask image
7 Molecule labeling	nuclei_mask_profile.txt and cell_mask_profile.txt	The single-cell spatial gene expression profiles based on nuclei mask and cell mask, respectively

2.6.1 Data availability

The data that support the findings of this study have been deposited into Spatial Transcript Omics DataBase (STOmics DB) of China National GeneBank DataBase (CNGBdb) with accession number STT0000027: https://db.cngb.org/stomics/project/STT0000027.

2.6.2 Code availability

StereoCell is used to perform the workflow analysis in this paper. Code, demo dataset and graphical user interface software are available at our GitHub repository with the detailed documentation: https://github.com/STOmics/StereoCell/tree/dev.

2.7 Supplementary Materials

The standard design in comparison experiment of image stitching methods.

For each dataset, a standard is designed according to the size of each image tile and the translation parameter set during the microscope shooting, and the standard is fine-tuned manually to correct mechanical errors according to the overlap area between each two adjacent image tiles. These standards are used to calculate the relative offset error between adjacent image tiles in the stitching results of different methods as follows:

$$re_{i,j} = \sqrt{\left(\left|x_i - x_j\right| - x_s\right)^2 + \left(\left|y_i - y_j\right| - y_s\right)^2}$$

where $re_{i,j}$ is the relative offset error between the i-th and j-th image tiles (adjacent

image tiles), (x_i, y_i) and (x_j, y_j) are the coordinates of the image tiles in the stitching result, and (x_s, y_s) is the coordinate of standard.

According to the "track lines" designed on the Stereo-seq chip, a template of stitched image can be obtained for each Stereo-seq dataset (the public dataset has no "track lines" that its template is not obtained). These templates are used to calculate the absolute offset error in the results of different methods as follows:

$$ae = \sum_{k=1}^{n} \sqrt{(xr_k - xt_k) + (yr_k - yt_k)}$$

where *ae* is the absolute offset error, (*xr*, *yr*) and (*xt*, *yt*) are the coordinates of a marker point in the stitching result and template respectively, *k* means the *k*-th marker point (*k* = 1, 2, …, *n*), and there are total *n* marker points obtained according to the "track lines".

Data analysis process and evaluation metrics in experiment of generating single-cell spatial gene expression profile on Stereo-seq mouse olfactory bulb dataset.

The profiles are input into Stereopy (v6.0) to calculate the total gene counts and number of uniquely expressed genes through the quality control function. During the filtering process, the cells with fewer than 150 expressed genes and more than 5% mitochondrial genes are removed, and genes present in less than 3 cells are also removed. The profiles are then normalized using the "SCTransform" function. The differentially expressed genes are summarized by Principal Component Analysis (PCA) to reduce the data dimensionality. With these settings, we run the uniform manifold approximation and projection (UMAP) algorithm to obtain 2D data projections, followed by the "Leiden" clustering to identify all clusters within the dataset. The silhouette coefficient for evaluating clustering results is calculated as follows:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where, *S*(*i*) is the silhouette coefficient, *a*(*i*) indicates the average distance between the *i*-th sample and other samples in its cluster, and *b*(*i*) is the average distance between the *i*-th sample and the samples in other clusters. Silhouette coefficient provides

information on how similar a given cell is to other similar cells/bins (cohesion) in comparison with non-similar cells (separation). The moran's I for evaluating spatial correlation as follows:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{i,j} z_i z_j)}{\sum_{i=1}^{n} z_i^2}$$

where, I is the moran's I, z_i is the deviation of the attribute of factor i from its mean value, $w_{i,j}$ is the spatial weight between factors i and j, n is equal to the factor integration, and S_0 is the aggregation of all spatial weights.

Data analysis process and evaluation metrics in experiment of dissecting the structural composition on Stereo-seq mouse brain dataset.

The generated profiles are input into Stereopy (v6.0), the total gene counts and number of uniquely expressed genes are calculated, the cells with fewer than 200 expressed genes and more than 5% mitochondrial genes are removed, and the genes present in less than 3 cells are removed. The profiles are normalized using the "SCTransform" function. The differentially expressed genes are summarized using PCA and the 2D data projections are obtained by the UMAP algorithm, followed all clusters within the dataset are identified by the "Leiden" clustering.

3 Generating single-cell gene expression profiles for high-resolution spatial transcriptomics based on cell boundary images

Bohan Zhang[1,2,†], Mei Li[1,3,†], Qiang Kang[1,†], Zhonghan Deng[1], Hua Qin[2], Kui Su[1], Xiuwen Feng[1], Lichuan Chen[1], Huanlin Liu[1], Shuangsang Fang[2], Yong Zhang[1], Yuxiang Li[1], Susanne Brix[3,*], Xun Xu[1,*]

1 BGI Research, Shenzhen 518083, China

2 BGI Research, Beijing 102601, China

3 Department of Biotechnology and Biomedicine, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

† Contributed equally.

* Corresponding author. E-mail: sbrix@dtu.dk, xuxun@genomics.cn.

To be submitted.

Abstract

Stereo-seq is a cutting-edge technique for spatial resolved transcriptomics that combines subcellular resolution with centimeter-level field-of-view, serving as a technical foundation for analyzing large tissues at the single-cell level. Our previous work presents the first one-stop software that utilizes cell nuclei staining images and statistical methods to obtain high-confidence single-cell spatial gene expression profiles for Stereo-seq data. With recent advancements in Stereo-seq technology, it is possible to acquire cell boundary information, such as cell membrane/wall staining images. To take advantage of this progress, we updated our software to a new version, named STCellbin, which utilizes the cell nuclei staining image as a bridge to acquire cell membrane/wall staining images that align with spatial gene expression maps. By employing an advanced cell segmentation technique, accurate cell boundaries can be obtained, leading to more reliable single-cell spatial gene expression profiles. Experimental results verify the application of STCellbin on mouse liver (cell membranes) and *Arabidopsis* seed (cell walls) datasets. The improved capability of capturing single cell gene expression profiles by this update results in a deeper understanding of the contribution of single cell phenotypes to tissue biology.

Availability & Implementation: The source code of STCellbin is available at https://github.com/STOmics/STCellbin.

3.1 STATEMENT OF NEED

Spatial resolved single cell transcriptomics enables the generation of comprehensive molecular maps that provide insights into the spatial distribution of molecules within the single cells that make up tissues. This groundbreaking technology offers insights into the location and function of cells in various tissues, enhancing our knowledge of organ development[106], tumor heterogeneity[118], cancer evolution[119], and other biological mechanisms. Resolution and field-of-view are two critical parameters in spatial transcriptomics. High resolution enables detailed molecular information at the single-cell level, and large field-of-view facilitates the creation of complete 3D maps that represent biological functions at the organ level. Stereo-seq simultaneously achieves subcellular resolution and a centimeter-level field-of-view, providing a technical foundation for obtaining comprehensive spatial gene expression profiles of whole tissues at single-cell level[47]. Our previous work offers the software StereoCell for acquiring high signal-to-noise ratio single-cell spatial gene expression profiles from Stereo-seq data[139]. The image data generated by Stereo-seq for StereoCell consists of a nucleus staining image. However, there is a big difference between cell nuclei and cell boundary staining images, based on cell membrane/wall staining, in terms of the ability to capture robust and precise cell specific gene expression profiles. Despite the widespread use of spatial techniques, such as MERFISH[30], CosMx[117], and Xenium[116], several of these techniques still struggle to achieve accurate cell boundary information, as they are based on nuclei staining images that can be generated using stains such as 4,6-diamidino-2-phenylindole (DAPI). With STCellbin, we here implement a procedure based on simultaneous cell membrane/wall and cell nuclei staining using multiplex immunofluorescence (mIF) and calcofluor white (CFW) staining[140,141], to automatically acquire more accurate cell boundary information and thereby obtain more reliable single-cell spatial gene expression profiles.

In STCellbin, we have retained the image stitching, tissue segmentation and molecule labeling steps from StereoCell and improved the image registration and cell

segmentation steps. As the cell membrane/wall staining images miss the “track line” information, which is the key in the image registration step [139], we utilize the cell nuclei staining image as a bridge to align the cell membrane/wall staining image with the spatial gene expression map, upon which we obtain the registered cell boundary information in the cell segmentation step. Based on the cell boundaries information, we directly assign the molecules to their corresponding cells, obtaining single-cell spatial gene expression profiles. We here applied STCellbin on mouse liver (cell membrane) and *Arabidopsis* seed (cell wall) datasets, and confirm the accuracy of cell segmentation. This update offers a comprehensive workflow to obtain reliable single-cell spatial gene expression profiles based on cell membrane/wall information, providing support and guidance for related scientific investigations, particularly those based on Stereo-seq.

3.2 IMPLEMENTATION

3.2.1 Overview of STCellbin

The process of STCellbin includes image stitching, image registration, cell segmentation and molecule labeling (Fig. 1). The Stereo-seq spatial gene expression data, cell nuclei and cell membrane/wall staining image tiles are input into STCellbin. The stitched cell nuclei and cell membrane/wall staining images are obtained through the MFWS algorithm[139]. The stitched cell nuclei and cell membrane/wall staining images are registered using the Fast Fourier Transform (FFT) algorithm[122]. The spatial gene expression data is transformed into a map, this map and a stitched cell nuclei staining image are registered based on “track lines”. Thus, the registration of the gene expression map and cell membrane/wall staining image is implemented. Cell segmentation is performed on the registered cell membrane/wall staining image by Cellpose 2.0[68] to obtain the cell mask. The molecules are assigned to their corresponding cells according to the cell mask to obtain the single-cell spatial gene expression profile. The tissue segmentation step based on Bi-Directional ConvLSTM U-Net[135] is set as optional, which can generate a tissue mask to assist in filtering out

impurities outside the tissue.

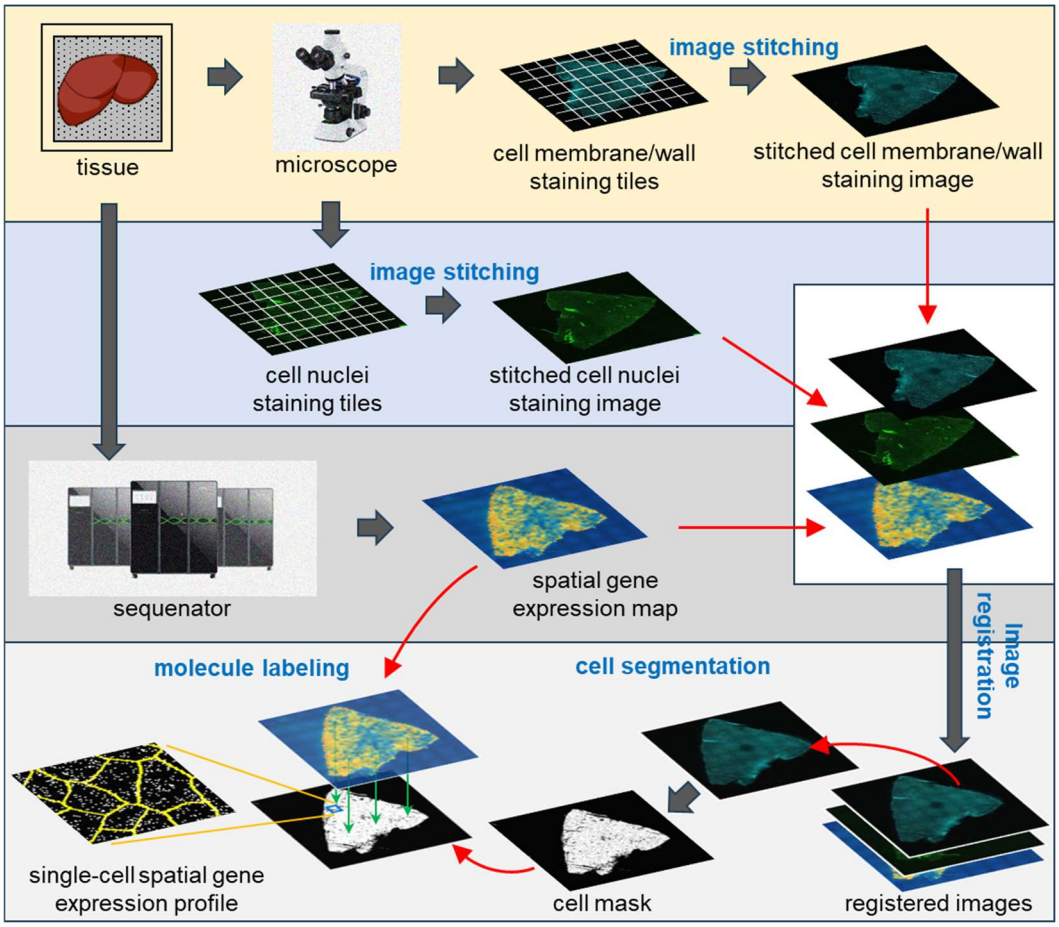

Figure 1. **Overview of STCellbin.** The cell nuclei and cell membrane/wall staining image tiles are stitched into individual large images respectively. The spatial gene expression map and stitched cell membrane/wall staining image are registered with the stitched cell nuclei staining image as a bridge. The cell mask is directly obtained from the registered cell membrane/wall staining image by cell segmentation. The single-cell spatial gene expression profile is obtained by overlaying the generated cell mask and the gene expression map.

3.2.2 Image stitching

The image stitching steps in STCellbin is consistent with the image stitching steps in StereoCell. The MFWS algorithm[139] is adopted, which calculates the offsets of two adjacent tiles with overlapping areas using FFT[122] to stitch these two tiles, and extends this process to all tiles. The relative error, absolute error and running time of MFWS have been verified in our previous work[139].

3.2.3 Image registration

The image registration of STCellbin includes two steps. The first is the registration of the stitched cell nuclei and stitched cell membrane/wall staining images. The two stained images are taken by the same microscope at the same magnification, which ensures that they have similar sizes and no large difference in rotation. Therefore, the key of the registration is to calculate the image offsets. The cell nuclei staining image is fixed, and the size of the cell membrane/wall staining image is adjusted to be consistent with the cell nuclei staining image by cutting and zero-padding (Fig. 2A). FFT[113] is then used to calculate the image offsets (similar to MFWS[139]). To save computing resources, the two stained images are mean-based subsampled[142] (Fig. 2B), the offsets of the subsampled images are calculated (Fig. 2C), and these offsets are restored to the scale of the original images so that the nuclei and cell membrane/wall staining images can be registered (Fig. 2D). The second registration is the same as in StereoCell[139], that is, the spatial gene expression data is transformed into a map, and then this map is registered with the stitched cell nuclei staining image based on “track lines”. This registration fixes the spatial gene expression map and performs scaling, rotating, flipping and translating on the stitched cell nuclei staining image. Since the cell nuclei and cell membrane/wall staining images have been registered, the same operations (scaling, rotating, flipping and translating) are repeated on the cell membrane/wall staining image (Fig. 2E), that is, the cell membrane/wall staining image and spatial gene expression map can be registered using the nuclei staining image as a bridge. STCellbin also has compatibility with registration requirements of specific images. When utilizing staining images produced with a multi-channel microscope, it is possible to omit the registration between these images, and the image stitching parameters can be the same for all channel images. Moreover, the registration can handle the case of multiple mIF staining images taken from identical tissues using the same microscope when there is only a difference in offsets among these images.

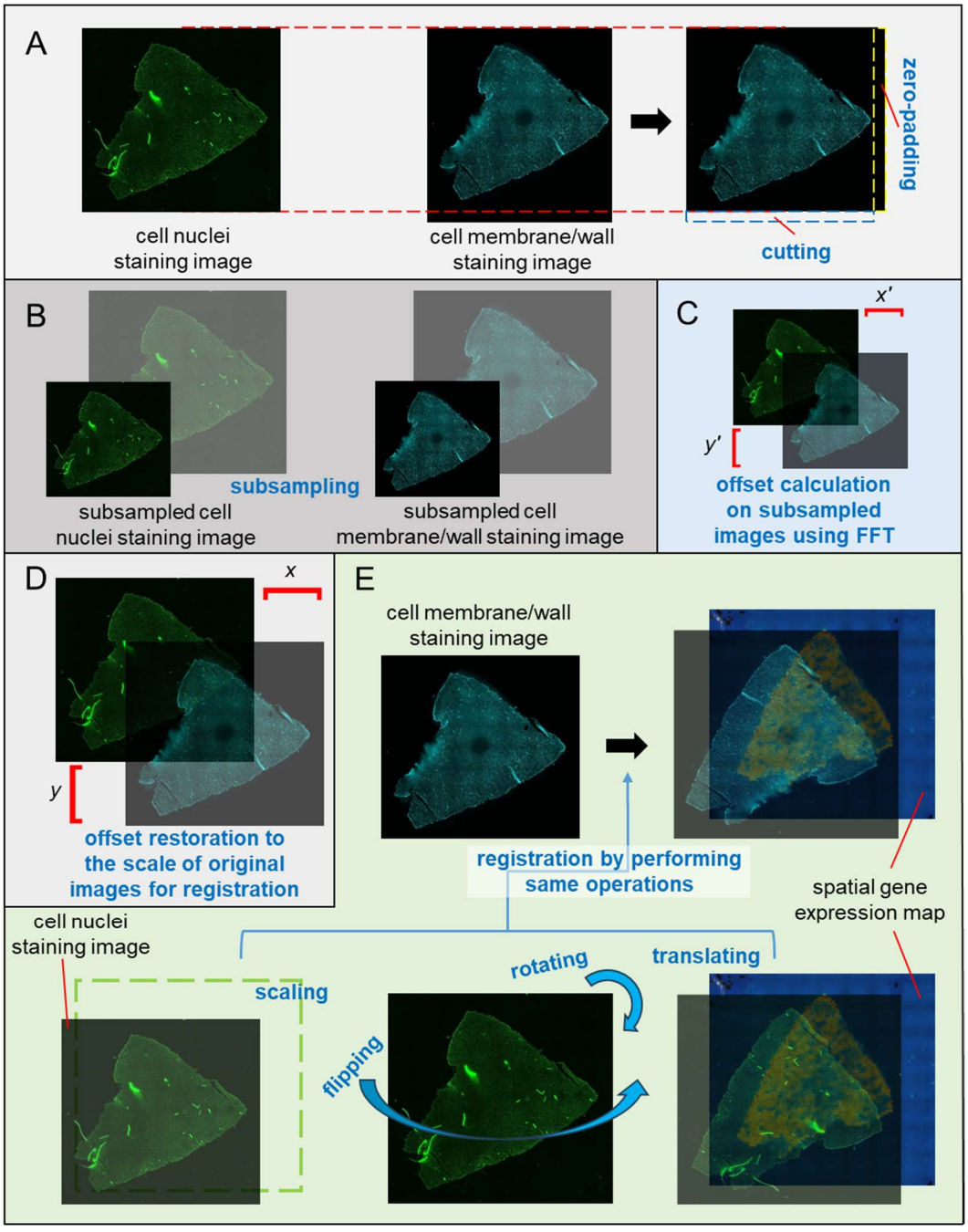

Figure 2. Registration of the cell membrane/wall staining image and spatial gene expression map using the cell nuclei staining image as a bridge. A. Size of the cell membrane/wall staining image is adjusted to be consistent with the cell nuclei staining image. B. Cell nuclei and cell membrane/wall staining images are subsampled. C. Calculating the offsets of the subsampled images. D. Restoring the offsets to the scale of original images for registration. E. Registering the spatial gene expression map and cell nuclei staining image by performing scaling, rotating, flipping and translating, and registering the spatial gene expression map and cell membrane/wall staining image by performing the same operations.

3.2.4 Cell segmentation

The cell segmentation step of STCellbin is performed using Cellpose 2.0[68] with some adjustments. The model architecture of Cellpose 2.0 and its weight files “cyto2” are downloaded. Due to the large size of staining images derived from Stereo-seq data, Cellpose 2.0 cannot be executed smoothly using normal hardware configurations. To circumvent this issue, the staining images are therefore cropped into multiple tiles with overlapping areas to perform cell segmentation and record the coordinates of tiles. The overlapping areas rescue cells at the border of the tiles from being cropped. To obtain the best results, segmentations with different values of the cell diameter parameter are performed independently, and the result with the largest sum of cell areas is retained. All the segmented tiles are assembled into the final segmented result according to the recorded coordinates. Moreover, when selecting the tissue segmentation option, an additional step is executed to apply a filter on the cell mask using the tissue mask, resulting in a filtered segmented outcome.

3.2.5 Molecule labeling

The molecule labeling of STCellbin is in principle the same as the one used in StereoCell. StereoCell assigns molecules in the cell nuclei to the cell by using the cell nuclei mask, and then assigns molecules outside the cell nuclei to the cells with the highest probability density using Gaussian Mixture Model[123]. STCellbin assigns molecules to the cells directly based on the cell mask, while the process of assigning molecules outside the cell is included as an option. The latter decision was made as the cell membranes/walls are usually tightly packed, with only a few molecules outside the cells, and the assignment of these molecules takes a lot of time. Thus, we generally do not recommend this option, and the users can use it according to actual requirements.

3.3 RESULTS

3.3.1 Datasets

We adopt two datasets acquired via Stereo-seq technology[47] One is a mouse liver dataset, a tissue that offers cell boundary information via cell membranes, as in all mammalian tissues. The other dataset is derived from seeds of the plant *Arabidopsis*, a tissue that provides cell boundary information based on rigid cell walls. More details of the two datasets are shown in Table 1.

Table 1. **Details of two datasets used for evaluation of cell boundary information**

Detail	Mouse liver dataset	*Arabidopsis* seed dataset
Data source	A slice of liver	Slices of multiple seeds
Cell nuclei dye	DAPI	ssDNA
Cell membrane/wall dye	mIF	CFW
Number of molecules	16,177,288	62,884,637
Number of cell nuclei staining image tiles	99	126
Number of cell membrane/wall staining image tiles	99	117

Abbreviations: CFW: calcofluor white; DAPI: 4,6-diamidino-2-phenylindole; mIF: multiplex immunofluorescence; ssDNA: single strand DNA.

3.3.2 Evaluation of cell segmentation performance

To evaluate the cell segmentation performance of STCellbin, we designed a ground truth based on a manual markup of the cells according to their cell membranes/walls based on the staining images. The number of cells from ground truth is named *ng*. The number of cells segmented by STCellbin is named *ns*. For each STCellbin segmented cell (s_cell_i), there must be a corresponding cell from ground truth (m_cell_i), where i is the index of the cell ($i = 1, 2, \ldots, ns$). Then a rule is set:

$$\begin{cases} s\_cell_i \text{ is segmented correctly} & \text{if } IoU_i > 0.5 \\ s\_cell_i \text{ is segmented incorrectly} & \text{otherwise} \end{cases} \quad (1)$$

where IoU is the standard intersection over union metric[66] set as:

$$IoU_i = ao_i / au_i \tag{2}$$

where ao_i is the area of overlap between s_cell$_i$ and m_cell$_i$, and au_i is the area of union of these two cells. Then the precision (*Pre*) and recall (*Rec*) are adopted:

$$Pre = nc / ns \tag{3}$$

$$Rec = nc / ng \tag{4}$$

where *nc* is the number of cells correctly segmented by STCellbin.

3.3.3 Generation of single-cell spatial gene expression profiles utilizing cell membrane/wall staining images

STCellbin was next applied to the mouse liver and *Arabidopsis* seed datasets. For each dataset, the input includes a file of spatial gene expression data, a folder of cell nuclei staining image tiles, and a folder of cell membrane/wall staining image tiles. Through the steps of image stitching, image registration, cell segmentation (the option of tissue segmentation is selected), and molecule labeling, the single-cell spatial gene expression profiles are generated as the output.

Given the substantial amount of work required for manual cell marking and limited clarity in certain regions of the staining images, we selected the areas with the best image data from the two datasets for presentation of the segmentation results. When using staining images with different dyes, STCellbin effectively identifies cell membranes/walls for cell segmentation, yielding cell masks that exhibit acceptable agreement with the manually marked results (Fig. 3A). This capability offers significant time and cost savings in practical applications. STCellbin demonstrates reliable identification of cells in both mammalian and plant tissues with a detection rate (*ns*/*ng*) of over 93.6%, and correctly segments most of them (Fig. 3B, left). Using the *Arabidopsis* seed dataset, STCellbin achieves a precision of 60.5% and a recall of 56.7%, while in the mouse liver dataset, it achieves a precision of 74.1% and a recall of 70.5% (Fig. 3B, right).

By employing STCellbin, the Stereo-seq spatial gene expression data includes an attribute of “CellID”, that is, the molecules are assigned to their originating cell to obtain single-cell gene expression profiles with spatial information (Fig. 3C, left). Cell area, number of unique genes per cell and number of gene counts per cell are statistically analyzed based on the data generated from mouse liver and the two *Arabidopsis* seeds with the most accurate segmentation profiles (Fig. 3C, right). By utilizing the obtained single-cell spatial gene expression profiles, clustering analysis was performed using the Leiden algorithm[143] (Fig. 3D). The resulting clusters of cells are spatially mapped within the tissue (Fig. 3D, left hand side for each tissue), allowing for the observation of their specific positions. From the Umaps, it is apparent that the different cell types are effectively distinguished (Fig. 3D, right hand side for each tissue). The spatial location of the different cell types will positively influence a series of downstream analyzes such as cellular annotation in less well-studied tissues.

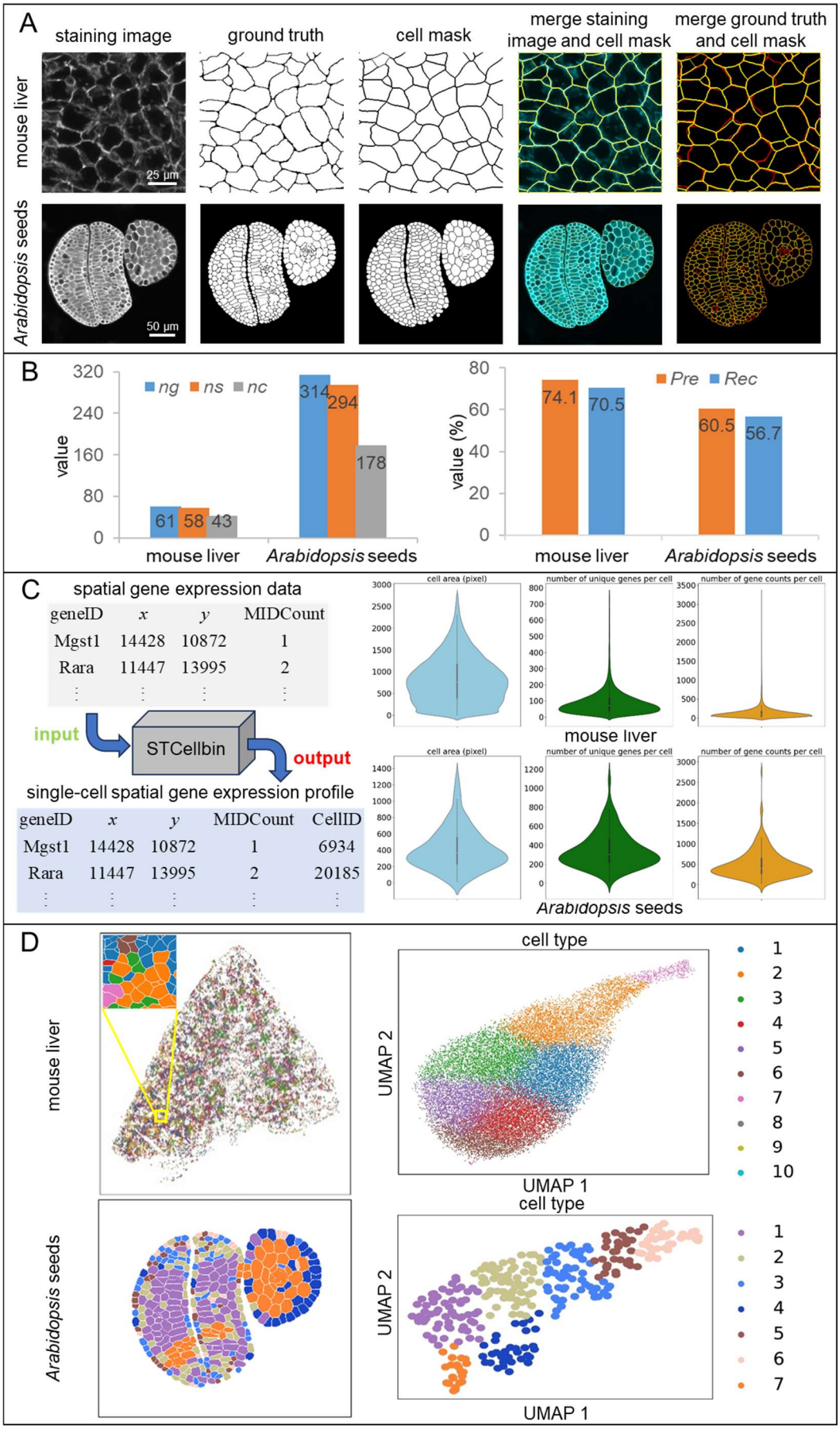
A
staining image
ground truth
cell mask
merge staining image and cell mask
merge ground truth and cell mask
mouse liver
Arabidopsis seeds
25 μm
50 μm
B
ng
ns
nc
value
320
240
160
80
0
314
294
178
61
58
43
mouse liver
Arabidopsis seeds
Pre
Rec
value (%)
80
60
40
20
0
74.1
70.5
60.5
56.7
mouse liver
Arabidopsis seeds
C
spatial gene expression data
geneID x y MIDCount
Mgst1 14428 10872 1
Rara 11447 13995 2
input
STCellbin
output
single-cell spatial gene expression profile
geneID x y MIDCount CellID
Mgst1 14428 10872 1 6934
Rara 11447 13995 2 20185
cell area (pixel)
number of unique genes per cell
number of gene counts per cell
mouse liver
Arabidopsis seeds
D
mouse liver
Arabidopsis seeds
cell type
UMAP 2
UMAP 1
1
2
3
4
5
6
7
8
9
10

Figure 3. Results of STCellbin on mouse liver and *Arabidopsis* seed datasets. **A**. Results of cell segmentation, where in the merged images, cell masks are set in yellow, staining images are set in cyan, and ground truths are set in red. **B**. Evaluation of segmentation performance. **C**. Generation of single-cell spatial gene expression profile, and statistics of cell areas, gene number per cell and gene expression per cell. **D**. Clustering results (left) and Umaps (right) from generated single-cell spatial gene expression profiles of a slice of mouse liver and two *Arabidopsis* seeds.

3.3.4 Discussion

Accurate identification of cell boundaries plays a crucial role in generating single-cell resolution in spatial omics applications. Based on previous work in StereoCell using cell nuclei staining images to generate single-cell spatial gene expression profiles, this STCellbin update extends the capability to automatically process Stereo-seq cell membrane/wall staining images for identification of cell boundaries that facilitates downstream analyses. We also showcase a few examples of the performance of cell membrane/wall segmentation in STCellbin. Currently, the tools for cell nuclei and cell membrane/wall segmentation can be independently executed, allowing users to choose the more suitable solution for their specific applications. In future work, these two techniques can be combined by training a deep learning model that is compatible with any staining image type, thereby achieving more accurate results.

3.4 AVAILABILITY OF SOURCE CODE AND REQUIREMENTS

- Project name: STCellbin
- Project home page: https://github.com/STOmics/STCellbin
- Operating system(s): Platform independent
- Programming language: Python
- Other requirements: Python 3.8
- License: MIT License
- RRID: SCR_024438

3.5 DATA AVAILABILITY

The data that support the findings of this study have been deposited into Spatial

Transcript Omics DataBase (STOmics DB) of China National GeneBank DataBase (CNGBdb) with accession number STT0000048: https://db.cngb.org/stomics/project/STT0000048.

4 EAGS: efficient and adaptive Gaussian smoothing applied to high-resolved spatial transcriptomics

Tongxuan Lv[1,2,†], Ying Zhang[1,†], Mei Li[1,4,†], Qiang Kang[1,‡], Shuangsang Fang[1,3], Yong Zhang[1], Susanne Brix[4,*], Xun Xu[1,2,*]

[1] BGI-Shenzhen, Shenzhen 518103, China

[2] College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[3] BGI-Beijing, Beijing 100101, China

[4] Department of Biotechnology and Biomedicine, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

[†] These authors contributed equally as the first authors.

[‡] Senior author.

[*] Corresponding: sbrix@dtu.dk, xuxun@genomics.cn

Accepted in GigaScience[144].

Abstract

Background: The emergence of high-resolved spatial transcriptomics (ST) has facilitated the research of novel methods to investigate biological development, organism growth, and other complex biological processes. However, high-resolved and whole transcriptomics ST datasets require customized imputation methods to improve the signal-to-noise ratio and the data quality.

Findings: We propose an efficient and adaptive Gaussian smoothing (EAGS) imputation method for high-resolved ST. The adaptive two-factor smoothing of EAGS creates patterns based on the spatial location and expression information of the cells, generates adaptive weights for the smoothing of cells in the same pattern, and then utilizes the weights to restore the gene expression profiles. We assessed the performance and efficiency of EAGS using simulated and high-resolved ST datasets of mouse brain and olfactory bulb.

Conclusions: Compared with other competitive methods, EAGS shows higher clustering accuracy, better biological interpretations, and significantly reduced computational consumption.

Keywords: spatial transcriptomics; imputation; gaussian smoothing; adaptive weight

4.1 Introduction

Recent advances in barcode-based spatial transcriptomics (ST) technology include 10X Visium[44], Slide-Seq[20,45] and high-definition spatial transcriptomics[42]. These advances made it feasible to provide expression profile information of entire genes, which is extremely important for comprehending biological functions and interaction networks[106,145]. High-resolved ST is an essential technical support for analyzing complex biological problems, as the function of complex biological tissues is closely related to the location of the transcriptional expression events within the tissue. However, cell localization and identification are limited by technical factors, such as the chip capture area, the sequencing depth, and the resolution. Spatially enhanced resolution transcriptome sequencing (Stereo-seq)[47] is a new ST technology based on DNA nanoballs. Stereo-seq provides the highest resolution (500 nm) among all currently available ST technologies. Such breakthrough in resolution allows researchers to perform genome-wide analyses of gene expression at the capture site (spot) with a single-cell or even sub-cellular resolution. Wang *et al*.[146] applied Stereo-seq to the 3D reconstruction of the ST of *Drosophila* embryos and larvae, providing a spatial- and temporal-resolved transcriptomic map of the whole organism across the developmental stages for *Drosophila* research. Liu *et al*.[147] reconstructed the developmental trajectory of zebrafish embryos during their development by analyzing Stereo-seq and scRNA-seq datasets from different time points.

Barcode-based high-resolved ST technology captures fewer genes at a single sequencing site (spot) than low-resolution ST technologies, such as 10X Visium[44] leading to high sparsity of the complete gene expression profile. In certain cell cycle phases, some cells do not express a set of genes whose expression thus appears to be null. In addition, amplification bias, cell cycle, library creation, and poor RNA capture rates cause some genes to be expressed but not captured by DNA nanoballs; such genes are called “dropout”[148]. Such biases adversely affect downstream analyses, such as clustering, cellular interaction analyses, and pseudo-temporal reconstructions[149,150],

when the raw data is directly processed.

Various imputation methods have been proposed to solve the “dropout” in gene expression for scRNA-seq datasets[151]. These imputation methods can be broadly classified into 3 categories according to their principles. The first category smooths or diffuses the levels of gene expression in cells with comparable expression patterns to correct (typically) all values (zero and non-zero). MAGIC imputes the missing data on scRNA-seq datasets based on the Markov chains of adjacent domains and recovers gene expression of the characterized cells by data diffusion[152]; DrImpute finds similar cells by consensus clustering and pools their gene expression values to estimate the loss[153]. The second category models the gene expression profile with an existing probabilistic statistical model to simulate the distribution of genes. SAVER assumes that each gene in each cell follows a Poisson-Gamma distribution (a negative binomial distribution) and estimates prior parameters to recover the expression of the missing genes using Poisson LASSO regression methods[154]. Scimpute constructs a mixed Gamma-Normal distribution based on the gene expression profile and uses a non-negative least squares regression model, sc-transform (R package), to perform the imputation[155]. The third category uses deep learning methods to capture the potential spatial representation of cells and reconstruct the expression matrix. DCA is an auto-encoder that predicts the parameters of the selected distribution to generate estimates[156]. These methods offer practical recommendations for single-cell imputation; however, these methods do not account for spatial information in ST datasets, and the methods based on specialized statistical models cannot be applied to the high sparsity of high-resolved ST datasets.

In recent years, ST-based imputation methods have been presented. The method Sprod first projects gene expression onto a potential space, connects nearest neighbor cells to construct patterns, and then learns the denoising matrix using a shared minimization of the graph’s Laplacian smoothing term and reconstruction errors[157]. For ST data without pathology images, Sprod provides cluster-based pseudo-images, but it does not accurately reflect the actual cell clustering situation. STAGATE introduces a

graph attention auto-encoder to construct a spatial neighbor network based on sequencing spots, and then, it introduces a distribution of the spatial neighbor network in the middle layer of the self-encoder to learn the correlation of neighboring sequencing spots and subsequently obtains the recovered gene expression profile by decoder[105]. However, the labels processed based on a specific clustering method are not completely consistent with the reality of the biological organization. It has been noted that the self-attention layer of the network does not consider the interaction between spot pairs and the information about the graphical structure of the spots[158].

To address these problems, we propose an efficient and adaptive Gaussian smoothing (EAGS) method, which is applied to high-resolved ST data. EAGS be derived from the fact that the spatial location of cells in biological tissues has a close relationship with their microenvironment, and the gene expression levels of cells within the same microenvironment are similar[155,159]. EAGS constructs different patterns based on cell expression profiles and cell location information to generate a similarity matrix. The similarity matrix then assesses cellular similarity within expression profiles to recover true biosignatures. By refining the information from proximal cells using adaptive smoothing weights and generating new gene expression profiles, the "dropout" is reduced. The resulting dataset provides RNA abundances more accurately than the original gene expression profile and preserves more of the true biological signal. EAGS enables the usage of high-sparsity ST datasets since it is independent of prior statistical models of the expression preconditioning the gene expression profiles. More crucially, EAGS could be used for large-scale ST datasets without requiring a lot of operating memory since it does not call for the computation of parameters for a pre-defined model, skipping most of the iterative process. We here applied EAGS to the simulated and high-resolved ST datasets of mouse brain and olfactory bulb, and compared it with widely used imputation methods to evaluate its efficacy in terms of fewer "zeros" in the gene expression profiles, improved cell annotation, and spatial organization replication.

4.2 Methods

4.2.1 The workflow of EAGS

In EAGS, the original expression matrix with the single-cell resolution was first used to generate patterns based on expression and spatial information. Then, the tight relationship between cells was established usispang two distinct patterns. Finally, the smoothing weights calculated from the patterns were used to define the level of smoothing for each cell, and were applied to recalculate the gene expression.

4.2.2 Datasets

There were two methods to generate gene expression profiles from Stereo-seq *in situ* captured data. One was to acquire the spatial location information of various cells by conducting cell identification and segmentation on the optical stained image, and then match the cell in the image to the sequencing spots with spatial coordinates[47,139]. The other one was to take consecutive $X \times X$ bins as units (considered as cells), where each bin (bin$X$) contains the total gene expression of $X \times X$ spots[47]. We used the mouse brain[47] and olfactory bulb datasets at single-cell resolution[139], which were generated by the first method and included 61,857 and 33,272 cells respectively. The *in situ* hybridization (ISH) images of the signature genes from the mouse brain were obtained to help compare the impacts of smoothing[134,160]. We also used another mouse olfactory bulb dataset generated by the second method, which contained 812 units of Bin140[161].

The above two categories of gene expression profile with spatial information were pre-processed with the Scanpy toolbox (V1.9.1; RRID:SCR_018139) to remove low-quality signals that might be blended into the gene expression data[85,162]. For the first category, firstly, we filtered genes based on expression in at least 10 cells: those genes were kept. Next, cell outliers were filtered using gene expression: cells expressing at least 300 MID counts were kept. The 2% highest MID counts in all cells were subtracted from the overall number of MIDs across all cells in the gene expression profile. Finally, the coordinates of the spatial position information of the cells and the log-transformed and normalized gene expression profiles were employed as input to EAGS. For the

second category, we filtered genes based on expression in at least 10 cells: those genes were kept. Next, cell outliers were filtered using gene expression: cells expressing at least 300 MID counts were kept.

4.2.3 Pattern construction

Since "similar cells" in organisms with comparable molecular microenvironments express their genes similarly, the regions with identical expression patterns may originate from the same cell type or from the same biological tissue location[155,159]. Using "similar cells" to supplement the information of a particular spot is feasible. Based on spatial location data and gene expression profiles, we constructed two patterns to divide the cells on an ST slice's gene expression profile into several clusters. A comprehensive description of these two pattern styles is given below:

Definition 1 (Gene Expression Pattern): If $\boldsymbol{P}_e(i)$ is the gene expression domain of $\boldsymbol{Cell}_i$ for ST data, then:

$$\forall \boldsymbol{Cell}_j \in \boldsymbol{P}_e(i),\ \forall \boldsymbol{Cell}_k \in \boldsymbol{P}_g - \left(\boldsymbol{P}_e(i) \cup \{\boldsymbol{Cell}_i\}\right),\ s.t. {d_{ij}}^e < {d_{ik}}^e \tag{1}$$

where $\boldsymbol{Cell}_i$, $\boldsymbol{Cell}_j$ and $\boldsymbol{Cell}_k$ are different cells, $\boldsymbol{P}_g$ is the global pattern of gene expression, ${d_{ij}}^e$ and ${d_{ik}}^e$ are the distance between $\boldsymbol{Cell}_i$ and $\boldsymbol{Cell}_j$, and $\boldsymbol{Cell}_i$ and $\boldsymbol{Cell}_k$, respectively.

Balltree is a binary tree data structure that performs well on high-dimensional datasets, especially for Fast Nearest-neighbor Search on high-dimensional datasets[163,164]. The complete gene expression profile is separated into many different subspaces by Balltree. Then, the Euclidean distances between cells are calculated separately. Assuming the pre-normalized gene expression profile still contains m cells, the unsupervised nearest neighbor network toolkit (scikit-learn) is used to extract the n-dimensional principal component data and creates the low-dimensional information

matrix ($\boldsymbol{LDIM}_{(m,n)}$) for the gene expression profile, as shown in Algorithm 1[165]. Then, the neighboring cell matrix is constructed based on the K-Nearest Neighbors network as in Algorithm 2, forming the Expression Neighbor Matrix ($\boldsymbol{ENM}_{(m,m)}$). Different definitions are given depending on whether $\boldsymbol{Cell}_j$ can be attributed to the gene expression pattern of $\boldsymbol{Cell}_i$:

$$ENM_{(i,j)}=\begin{cases}1,\ j=i\\ 1,\boldsymbol{Cell}_j\in\boldsymbol{P}_e(i)\\ 0,\boldsymbol{Cell}_j\notin\boldsymbol{P}_e(i)\end{cases} \tag{2}$$

where $ENM_{(i,j)}$ defines whether $\boldsymbol{Cell}_j$ is within the gene expression pattern $\boldsymbol{P}_e(i)$ of $\boldsymbol{Cell}_i$, if $ENM_{(i,j)}$ = 1, $\boldsymbol{Cell}_j$ belongs to the expression pattern of $\boldsymbol{Cell}_i$; if $ENM_{(i,j)}$ = 0, it does not.

Algorithm 1. Builds the tree structure of Balltree

Balltree is built using a divide-and-conquer method. Initially, Balltree has only one (root) node and all data points are assigned to it. At each step, the partition corresponding to each node is split into two sub-partitions. For a partition p_i, the splitting procedure is as follows:

Step 1: Find the centroid of the node points in $\boldsymbol{LDIM}_{(m,n)}$. Reducing an n-dimensional matrix to a two-dimensional plane, the centroid of the node is centroid 1.

Step 2: Select the farthest point from centroid 1 in p_i as the first (left) child pivot p_i^L.

Step 3: Select the farthest point from p_i^L as the second (right) child pivot p_i^R.

Step 4: Assign each data point p_i to the partition whose pivot is closer.

Step 5: Assign the new sub-partitions as children of v_i in Balltree, *i.e.,* v_i^R and v_i^L.

Algorithm 2. Using Balltree to find the nearest Neighbor of each cell

Input: Balltree structure $nbrs$, nearest neighbor num k , test point t, Current node n

Output: Expression Neighbor Matrix (***ENM***)

Algorithm: $ball-tree-research(nbrs,k,t,n)$

if $distance(t,node.pivot)$ - $node.radius \geq max(q)$:

    return;

if node in leaf-node set:

    Add $node.pivot$ to Q refresh q

    If $length(Q) > k$:

        Remove the point furthest from the test point

        Refresh q

else:

    $ball-tree-research(nbrs,k,t,node.son1)$

    $ball-tree-research(nbrs,k,t,node.son2)$

end if

return ***ENM***

The difference between ST and scRNA-seq datasets is that ST dataset provides the spatial coordinate position of each sequencing site (spot). After StereoCell processing, ST data are spots with a single-cell resolution where every spot corresponds to a single physical cell with spatial coordinates[139]. Cells in adjacent regions of histological sections are more likely to come from the identical microenvironment and belong to similar or identical cell types than cells from other areas. Therefore, we offer the spatial neighborhood pattern as a reference and classify the cluster of cells that are physically adjacent to a specific cell as its “spatial neighborhoods”:

Definition 2 (Spatial Neighbor Pattern): If $\boldsymbol{P}_s(i)$ is the spatial neighbor pattern of $\boldsymbol{Cell}_i$ for ST data, then:

$$\forall \boldsymbol{Cell}_j \in \boldsymbol{P}_s(i),\ s.t. d_{ij}^{\ s} \le \tau_s \tag{3}$$

where $d_{ij}^{\ s}$ is the spatial distance between $\boldsymbol{Cell}_i$ and $\boldsymbol{Cell}_j$, and τ_s represents the maximum spatial distance of $\boldsymbol{P}_s(i)$ of $\boldsymbol{Cell}_i$.

Since the spatial distribution of ST dataset is a two-dimensional plane space, the Euclidean distance can serve as a useful measure of spatial location between cells in a low-dimensional environment. Therefore, the spatial distance Matrix ($\boldsymbol{SDM}_{(m,m)}$) is constructed by computing the Euclidean distance. Furthermore, since ST chips of the Stereo-seq platform vary in size, EAGS fine-tunes the weight value for different chip sizes while calculating Euclidean distances.

4.2.4 Adaptive weight calculation

Cells can be used as smoothing factors for $\boldsymbol{Cell}_i$, and must satisfy both the gene expression pattern and the spatial neighbor pattern belonging to $\boldsymbol{Cell}_i$. A cell acting as the smoothing factor is more similar in gene expression to the smoothed cell than to other cells in the overall expression profile. EAGS defines the nearest neighbor contribution matrix ($\boldsymbol{NCM}_{(m,m)}$) for a ST dataset containing m cells as follows:

$$\boldsymbol{NCM}_{(m,m)} = \boldsymbol{SDM}_{(m,m)} \square \boldsymbol{ENM}_{(m,m)} \tag{4}$$

where $\boldsymbol{NCM}_{(m,m)}$ is the dot product obtained by multiplying the corresponding elements of the $\boldsymbol{SDM}_{(m,m)}$ and $\boldsymbol{ENM}_{(m,m)}$ matrices. The non-zero value $\boldsymbol{NCM}_{nonzero}$ part of the $\boldsymbol{NCM}_{(m,m)}$ is selected as the parameter for smoothing weights, and $\boldsymbol{NCM}_{nonzero}$ is a G-dimensional row vector, where the condition $G \le M \times M$ is satisfied. The p^{th}

percentile of the $\boldsymbol{NCM}_{nonzero}$ along the specified axis is calculated by the following method:

$$(G-1)\times p^{th} = c + t \tag{5}$$

where G represents the number of vectors of $\boldsymbol{NCM}_{nonzero}$. c and t represent the integer and fractional parts of the calculation result, respectively. The Distance Distribution Threshold (DDT) is defined as follows:

$$DDT = (1-t)\times \boldsymbol{NCM}_{nonzero}[c] + t \times \boldsymbol{NCM}_{nonzero}[c+1] \tag{6}$$

where the calculated c and t obtain the p^{th} percentile DDT along the specified axis of the $\boldsymbol{NCM}_{nonzero}$. The calculation of the adaptive weights is based on the $\boldsymbol{NCM}_{nonzero}$:

$$\boldsymbol{GS}_{new} = \mathrm{GS}(\boldsymbol{NCM}) = a \times e^{-\frac{(\boldsymbol{NCM}-b)^2}{2\times\mu^2}} \tag{7}$$

where $\boldsymbol{GS}_{new}$ is the degree of smoothing information and is an adaptive weight determined by the degree of similarity between the cells in the pattern's framework, and $\mathrm{GS}()$ is used to calculate the adaptive weights. The precise smoothing weight contribution between cells is calculated as follows:

$$\mu = \sqrt{-\frac{(DDT-b)^2}{2\times\ln\left(\frac{gs}{a}\right)}} \tag{8}$$

where gs is a hyperparameter that characterizes the overall smoothness of the reference gene expression profile, which represents the overall smoothness of the entire chip. For a 1×1 cm ST chip of the Stereo-seq platform, gs is set to 0.95. μ is the smooth weight that varies around the gs, and characterizes the overall contribution level of cells in both the $\boldsymbol{P}_e(i)$ and $\boldsymbol{P}_s(i)$ to $\boldsymbol{Cell}_i$.

DDT in Eq. (6) refers to the similarity distance between cells generated based on the Gene Expression Pattern and the Spatial Neighbor Pattern in the entire gene expression distribution matrix, that as a global benchmark reference for information

distribution and can be characterized as the distribution of the gene expression matrix from the overall level. μ calculated in Eq. (8) refers to the standardized parameters of the Gaussian model. The value of μ calculated by DDT can make the smoothed gene expression matrix more consistent with the preset distribution, such $\boldsymbol{GS}_{new}$ can be measured with the help of some existing expression quantities, genes are complemented without changing the overall expression profile.

4.2.5 Smooth

The raw gene expression profile can be processed after $\boldsymbol{GS}_{new}$ and raw expression $\boldsymbol{E}_{origin}$ have been obtained:

$$\boldsymbol{E}_{GS}\left(x\right)=\frac{\sum_{i\in P_A(i)}\boldsymbol{GS}_{new}\left(R\left(i,x\right)\right)\times\boldsymbol{E}_{origin}\left(i\right)+\boldsymbol{E}_x}{\sum_{i\in P_A(i)}\boldsymbol{GS}_{new}\left(R\left(i,x\right)\right)+1} \tag{9}$$

where $\boldsymbol{E}_{GS}$ represents the level of gene expression after adaptive weight smoothing, $P_A(i)$ represents all cells in the region where cell x is smoothed, $\boldsymbol{E}_x$ represents the original gene expression of the smoothed cell. The whole process can be represented by Algorithm 3.

Algorithm 3. Calculate weights and perform smoothing

Input: Expression Neighbor Matrix $\boldsymbol{ENM}$, Spatial Distance Matrix $\boldsymbol{SDM}_{(m,m)}$, Origin expression matrix $\boldsymbol{E}_{origin}$, Hyperparameter gs

Output: Smooth expression Matrix $\boldsymbol{E}_{(GS)}$

Step 1: Calculating the K-nearest-neighbor cell Euclidean distance distribution.

Step 2: Smooth threshold takes the percentile value x of the distance distribution and requires a value from 0.2 to 1.

Step 3: Using Eq. (6) to back-calculate the magnitude of μ at this time; preset gs = 0.95.

Step 4: The Gaussian weights at other distances are calculated by substituting μ values into Eq. (7).

Step 5: Re-weighted summation based on the newly calculated Gaussian weights and the original expressions.

If relying entirely on the cells in the $P_A(i)$ as smoothing factors without using the origin gene expression of the smoothed $\boldsymbol{Cell}_i$, Eq. (10) can be further streamlined as:

$$\boldsymbol{E}_{GS}(x) = \frac{\sum_{i \in P_A(i)} \boldsymbol{GS}_{new}\left(R(i,x)\right) \times \boldsymbol{E}_{origin}(i)}{\sum_{i \in P_A(i)} \boldsymbol{GS}_{new}\left(R(i,x)\right)} \tag{10}$$

where $\boldsymbol{E}_{GS}$ is completely calculated from the expression level of cells in $P_A(i)$, regardless of the gene expression of $\boldsymbol{Cell}_i$.

4.2.6 Evaluation method

We used the imputation error by calculating the L2 norm of the difference between the smoothed matrix and ground truth (L2-error)[166]. We used Calinski-Harabasz Index (CHI) and the Davies-Bouldin Index (DBI) to evaluate the significance of the differences in intra-class and extra-class similarity of the clustering results. We used Moran's I and Geary's C to calculate the correlation of cellular marker genes in the gene expression space of the data before and after smoothing[167].

Imputation error by calculating the L2 norm

L2-error is used to compute the difference between two matrix vectors by calculating the Euclidean distance between each corresponding element of the two matrices separately. A lower L2-error represents a higher degree of similarity between the two matrices, indicating that the method performs better. It is defined as follows:

$$\text{L2-error} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(Y_{i,j}\right)^2} - \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(X_{i,j}\right)^2} \tag{11}$$

where $Y_{i,j}$ represents the reference gene expression matrix, and $X_{i,j}$ represents the smoothed gene expression matrix. L2-error is mainly used to compare the difference between the reference expression matrix with "ground truth counts" and the smoothed expression matrix.

Calinski-Harabasz Index

The CHI computes the sum of squares of the distances between points in the class and the class center to determine how closely a class is related[168]. The higher the CHI, the higher the similarity between cells of the same type in the cell population, indicating that this method performs better. It is defined as:

$$\mathrm{CHI}(k) = \frac{\mathrm{tr}\left(\boldsymbol{B}_q\right)}{\mathrm{tr}\left(\boldsymbol{W}_q\right)} \times \left(\frac{h-q}{q-1}\right) \tag{12}$$

where h is the number of training samples, q is the number of categories, $\boldsymbol{B}_q$ is the between-category covariance matrix, $\boldsymbol{W}_q$ is the within-category data covariance matrix, and $\mathrm{tr}()$ is the trace calculation function.

Davies-Bouldin Index

The DBI finds the maximum by calculating the quotient of the sum of the average intra-class distances of any two classes within the sample set and the distance between the centers of the two clusters[169]. The lower the DBI, the higher the similarity between cells of the same type in the cell population, indicating that this method performs better. It is defined as:

$$\mathrm{DBI} = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j}\left(\frac{\sigma_i + \sigma_j}{d\left(\boldsymbol{c}_i, \boldsymbol{c}_j\right)}\right) \tag{13}$$

where n is the number of categories, $\boldsymbol{c}_i$ is the center of the ith category, σ_i is the average distance from all points of the ith category to the center, $d\left(\boldsymbol{c}_i, \boldsymbol{c}_j\right)$ is the

distance between the center points $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$, and $\max()$ is the maximum function.

Moran's I

Moran's I is a global autocorrelation statistic for certain metrics on a graph. It is commonly used in spatial data analysis to evaluate autocorrelation on two-dimensional grids[170]. The higher the Moran's I, the stronger the spatial autocorrelation of the cell population, indicating that the method performs better. It is defined as:

$$\text{Moran's I} = \left(\frac{N}{W}\right) \times \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(w_{ij} \times \left(x_i - \bar{x}\right) \times \left(x_j - \bar{x}\right)\right)}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \tag{14}$$

where N is the number of spatial units indexed by i and j, x is the variable of interest, $\bar{x}$ is the mean of x, w_{ij} are the elements of a matrix of spatial weights with zeros on the diagonal, and W is the sum of all w_{ij}.

Geary's C

Geary's C is a measure of spatial autocorrelation that attempts to determine if observations of the same variable are spatially autocorrelated globally (rather than at the neighborhood level)[171]. The lower the Geary's C, the stronger the spatial autocorrelation of the cell population, indicating that the method performs better. It is defined as:

$$\text{Geary's C} = \frac{(N-1) \times \sum_i \sum_j \left(\boldsymbol{w}_{ij} \times \left(x_i - x_j\right)\right)}{2 \times S_0 \times \sum_i \left(x_i - \bar{x}\right)^2} \tag{15}$$

where $\boldsymbol{w}_{ij}$ is the $i\text{-}th$ row of the spatial weight matrix with zeros on the diagonal, and S_0 is the sum of all the weights.

4.3 Results

4.3.1 Overview of EAGS

We use previously generated the datasets of mouse brain and olfactory bulb as inputs to EAGS[133,139]. The acquisition process of these data is: stereo-seq[47] is used to capture the ST data of the mouse brain and mouse olfactory bulb *in situ* and record the position information of the sequencing spot, just like the data generation process in "Datasets" subsection, and then StereoCell[139] is used to generate ST data at single-cell resolution with spatial information. After obtaining the ST dataset at single-cell resolution, the entire gene expression profile is normalized and smoothed[172], as shown in Fig. 1A.

EAGS constructs two styles of patterns based on the input gene expression information and spatial information, respectively. These two patterns are used to identify similar cells within the pattern, as shown in Fig. 1B. Next, EAGS adaptively generates smoothing weights based on the difference between similar cells and their genes' expression, then utilizes these weights as a reference to complement the expression of similar cells.

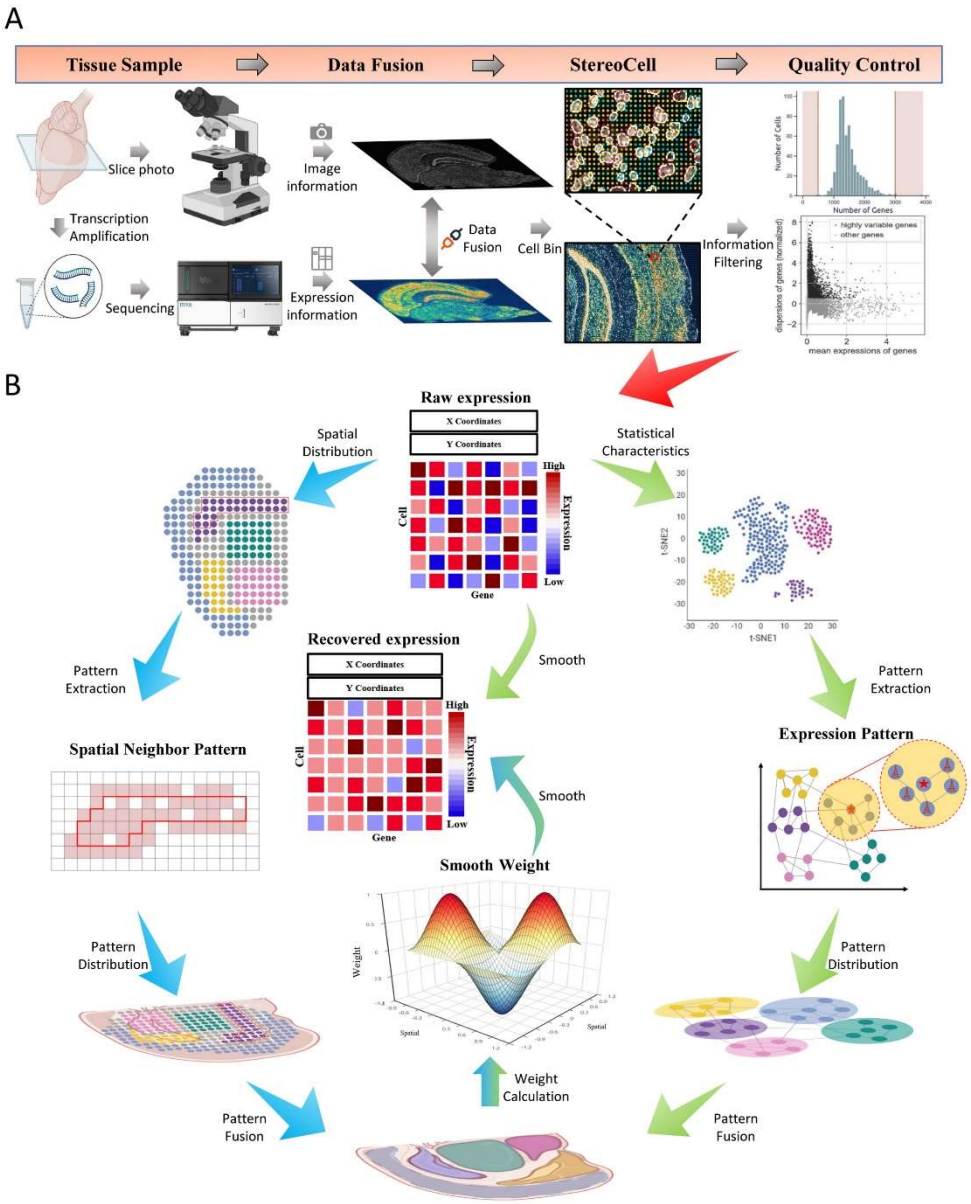

Figure 1: Workflow of EAGS. (A) Data generation process for the input of EAGS. (B) The EAGS method calculates the nearest neighbor information based on the gene expression pattern and spatial information. Then, EAGS adaptively generates smoothing weights and outputs the smoothed results.

4.3.2 EAGS performs better smoothing by adaptive weighting

We use the mouse brain dataset to evaluate EAGS with adaptive weight. The results are compared to the outputs of EAGS with fixed weights. As the mouse brain dataset's adaptive weight value is 19,001, the fixed value weights are set to 25,000 and 15,000. We use Spatial-ID to annotate cell types in order to assess the potential of EAGS to improve the cell annotation power and restore the true levels of gene expression[133]. Fig.

2 shows all the results of the subsequent analysis with the adaptive and the fixed weights. The cell annotation results of EAGS using an adaptive weight compared to a fixed weight generated a cell-type spatial map with clearer tissue outlines and more annotated cell subtypes (Fig. 2A).

Based on our cell annotation results, the CHI of the EAGS smoothing results with adaptive and fixed weights are calculated. Next, Geary's C and Moran's I of the common cell types in the annotation results are calculated (Figs. 2B and C). The results based on the adaptive weight cell annotation show a significant improvement in spatial autocorrelation compared to the others. Also, within the same type of cell annotation, the level of intra-class autocorrelation is higher.

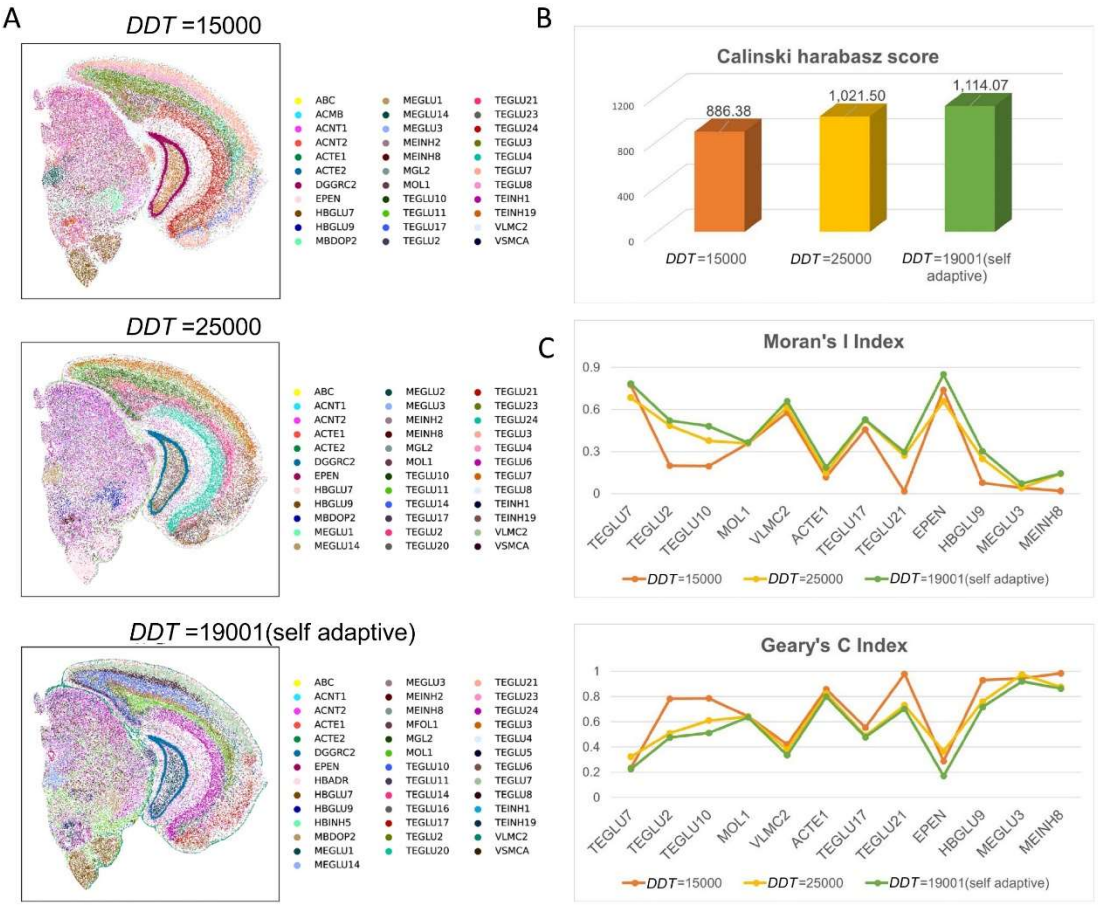

Figure 2: Results of EAGS with adaptive and fixed weights. (A) Spatial cell type map for cell annotation with Spatial-ID using different weights for smoothing results. (B) The smoothing results with different weights are annotated with Spatial-ID cells. The Calinski-Harabasz Index is calculated using cell labels. (C) After the cell annotation using Spatial-ID with different weights, Geary's C and Moran's

I are calculated from annotation results.

4.3.3 EAGS smooths gene expression with better performance on simulated ST dataset

We collected Bin140 specification for the mouse olfactory bulb ST dataset[161], and selected the top 2000 highly variable genes as the reference input of ScDesign3 to construct a simulation space group with "ground truth counts"[173]. To simulate the "dropout" phenomenon during the sequencing process, we randomly drop the simulated ST dataset expression to varying degrees and add different proportions of noise. EAGS, MAGIC[152], kNN-smoothing[77], SPCS[159] and STAGATE[105] are used to impute the processed ST dataset, then L2-error with the "ground truth counts" matrix and DBI are calculated respectively. The results are shown in Table 1.

Table 1. Results on simulated ST dataset with different proportions of dropout and noise.

Dataset	Method	10% noise		20% noise		30% noise	
		L2-error	DBI	L2-error	DBI	L2-error	DBI
30% dropout	MAGIC	473.0679	4.7818	524.2593	4.4600	562.8816	4.2533
	kNN-smoothing	381.4241	4.4481	448.7622	4.2175	504.8468	3.9928
	SPCS	322.4659	4.6961	390.2348	4.3379	460.2699	4.2569
	STAGATE	757.3749	11.3146	790.5977	10.4536	791.5948	4.7064
	EAGS	**313.4211**	**4.1926**	**379.4374**	**3.9912**	**449.6919**	**3.9589**
50% dropout	MAGIC	496.2557	4.9018	557.0938	4.4717	590.9306	4.2881
	kNN-smoothing	389.2189	5.2899	458.8456	4.4637	517.1581	4.7092
	SPCS	319.4318	4.8357	398.8399	4.3223	475.0918	4.2884
	STAGATE	705.9759	8.7298	747.5357	76.9083	872.2147	9.3826
	EAGS	**310.2873**	**4.6808**	**386.7170**	**4.3201**	**459.1241**	**4.2065**
70% dropout	MAGIC	506.0679	6.3969	571.8875	5.9789	600.8655	5.7213
	kNN-smoothing	390.3494	7.7208	477.8630	6.6270	532.8365	6.1928

SPCS	321.1818	**5.9028**	413.9192	5.7685	488.5445	5.6784
STAGATE	850.6077	13.1108	814.2477	15.8843	693.4574	7.2281
EAGS	**312.1247**	6.0768	**395.1362**	**5.6497**	**467.0692**	**5.6410**

From Table 1, in the simulated datasets with 30%- and 50%-dropout, L2-error and DBI values obtained by EAGS are always the lowest, regardless of the proportion of noise. When the proportion of dropout is 70%, DBI obtained by EAGS is suboptimal with 10%-noise (only higher than that of SPCS), and the results obtained by EAGS are the best on the other cases. In general, EAGS performs better on different simulated ST datasets and shows obvious advantages in improving intra-cell similarity and consistency with the "ground truth counts" compared with other methods.

4.3.4 EAGS smooths gene expressions for better characterizing the spatial expression patterns of mouse brain

We perform cell annotation on mouse brain data before and after EAGS smoothing using Spatial-ID[133]. The annotation results are shown in Fig. 3A. The mouse brain cell annotation based on data smoothed by EAGS return a clearer tissue structure, and more cell types can be annotated. To further assess the improvement provided by EAGS in cell annotation, we also perform cell annotation with Tangram[77], a technique for merging spatial data types with single cell/single nucleus RNA sequencing data and for cell type annotation. As shown in Fig. 3B, the CHI and DBI are calculated for the spatial autocorrelation of cell types with the gene expression profiles after Tangram and Spatial-ID cell annotation. These results show that EAGS smoothing provides significantly better results in cell-type annotations.

Fig. 3C shows the results of cell annotation using Spatial-ID and the spatial map of Allen Mouse Brain Atlas of corresponding cell types[134,160]. TEGLU24, TEGLU7 and MEINH2 are important cell types in the Hippocampus, Cortex and Midbrain dorsal respectively, and DGGRC2 is the important cell type in the Midbrain ventral and Dentate gyrus. These cell types are more consistent with the Allen's spatial expression

map of cell types after EAGS smoothing. To verify the smoothing effect, Moran's I and Geary's C are calculated for cells with different cell number ratios using the raw or the EAGS smoothed dataset (Fig. 3D). To determine whether the correlation between the above cell types and their marker genes improved after smoothing, the ratios of the number of annotated cell types to their corresponding non-zero marker gene expressions are computed. The ability of EAGS to restore true biological signals is shown in Fig. 3E. Our results show that EAGS contributes to enhancing the cellular features of the mouse brain as well as the spatial autocorrelation and intraclass similarity of the gene expression patterns.

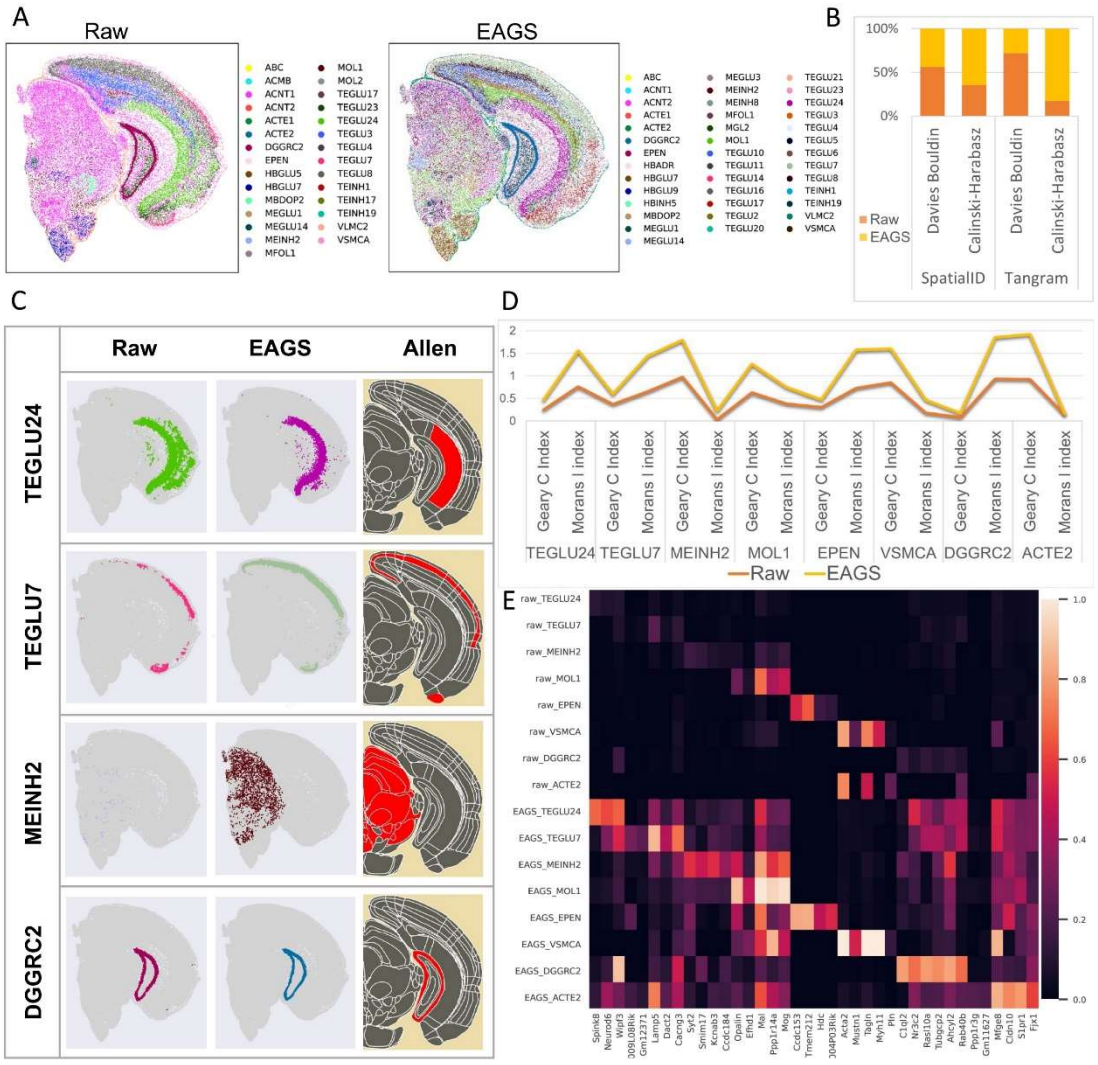

Figure 3: Comparisons between the analysis results obtained from data before and after EAGS

smoothing. (A) Spatial cell type maps of the mouse brain using Spatial-ID cell annotation of raw and EAGS smoothed data. (B) Davies-Bouldin and Calinski-Harabasz Indexes calculated using Spatial-ID and Tangram annotation results obtained from raw and EAGS smoothed data. (C) Comparison of the spatial map and Allen Mouse Brain Atlas obtained from raw and EAGS smoothed dataset. (D) Comparison of Moran's I and Geary's C cell annotation types obtained from raw and EAGS smoothed dataset. (E) Heatmap of non-zero ratio between the number of cell types and their marker genes obtained from raw and EAGS smoothed dataset.

4.3.5 EAGS improves spatial patterns and downstream analyses of gene expression data

EAGS is compared with the imputation methods, MAGIC[152], STAGATE[105] and kNN-smoothing[174], on ST mouse brain dataset (SPCS cannot be executed successfully because the sparsity of this dataset is high). The cell type spatial maps of different imputation methods using Spatial-ID as reference are shown in Fig. 4A (left). EAGS return more cell types and more prominent outlines than other methods. The results of MAGIC are very unbalanced in terms of the number of cell types, with a large number of cell annotations that did not match the true values[134,160]. The annotations of the Midbrain dorsal, the Midbrain ventral, and the Dentate gyrus are mixed using MAGIC. The results of STAGATE show fewer cell types. Also, STAGATE do not result in well-organized cell type distributions in the Hippocampus and Cortex. The cell type boundaries of cell annotation after kNN-smoothing processing are blurred, and different types of cells are mixed. In order to avoid the impact of data sparsity on the interpretability of the results, the input data of the cell annotation is the 50th-dimensional principal component of different imputation results; the Uniform Manifold Approximation and Projection (UMAP) of the annotated results is shown in Fig. 4A (middle). The cell type spatial maps, consisting of cell types that are highly represented and annotated by the three methods, are shown in Fig. 4A (right). Fig. 4B shows the CHI derived from data processed using one of the four methods. After cell annotation, CHI[168] calculated by the cell annotation label using EAGS shows higher spatial autocorrelation than other three methods. EAGS obtain higher Moran's I and Geary's C than other methods (Fig. 4C). Additionally, the spatial maps of a few marker genes

based on their expression are generated (Fig. 4D). The gene expression profiles smoothed by EAGS agree with the Allen's ISH image better than the other methods.

To evaluate the efficiency of high-resolved ST data, we run EAGS, MAGIC, STAGATE, kNN-smoothing, Scimpute and Drimpute three times on ST mouse brain dataset and monitor the average run time. STAGATE is run using a GPU. For the sake of fairness, in this running time comparison, all methods used the CPU uniformly. EAGS require the shortest run time, taking 3,484 seconds, while MAGIC takes 4,109 seconds, kNN-smoothing takes 4,739 seconds, and the other methods need a large memory consumption and cannot reach their final output in an acceptable time.

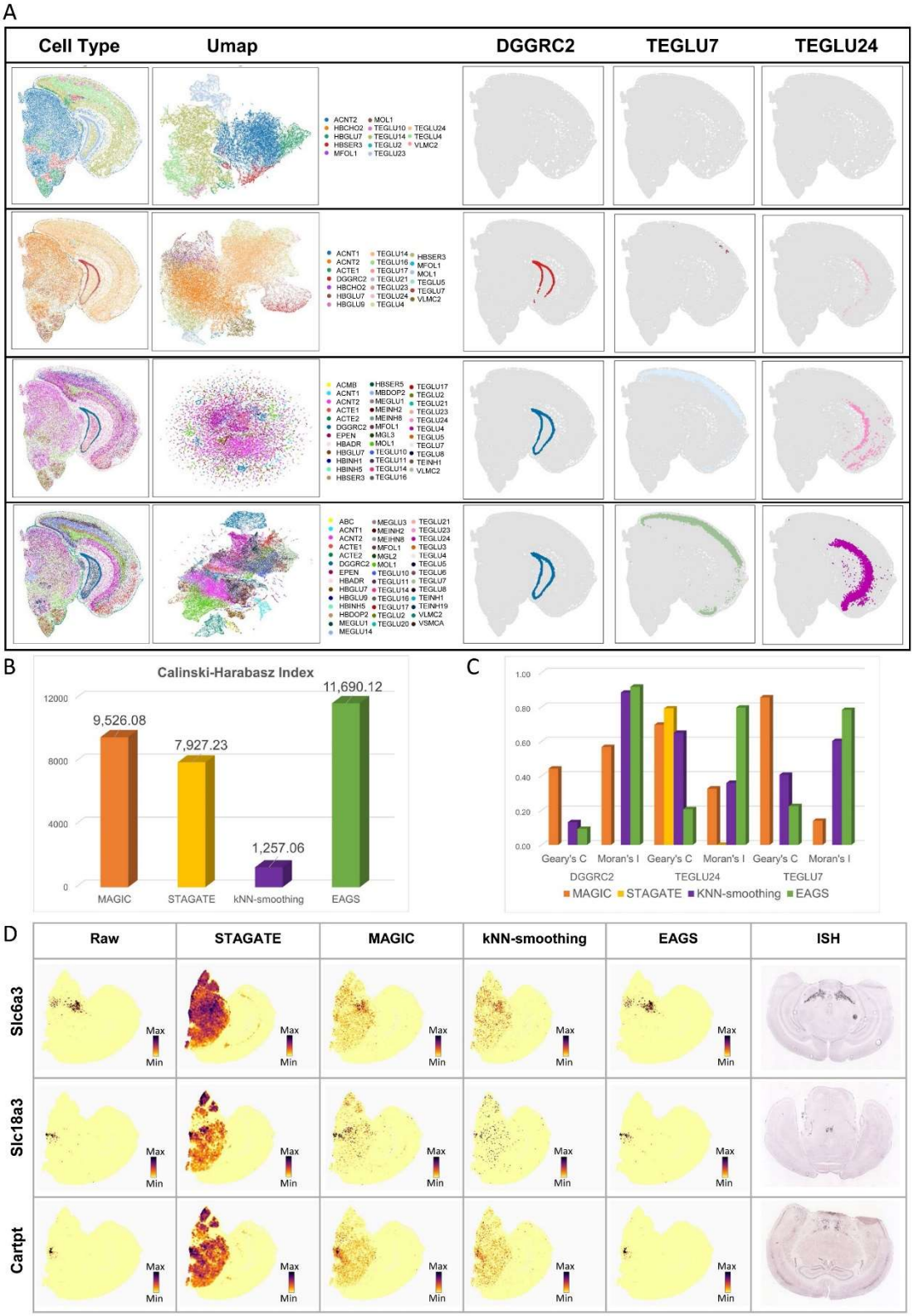

Figure 4: Comparison of different imputation methods. (A) left: Spatial maps of cell types using Spatial-ID cell annotations and four different imputation methods. middle: UMAP dimensionality reduction using Spatial-ID cell annotation and different imputation methods. right: Individual cell type spatial maps after cell annotation and different imputation methods. (B) Calinski-Harabasz Index

calculated using cell labels after Spatial-ID cell annotations and different imputation methods. (C) Moran's I and Geary's C for the DGGRC2, TEGLU7, TEGLU24 cell types. (D) Marker gene heatmaps and Mouse Brain Atlas obtained using different imputation methods.

4.3.6 EAGS application to high-resolved ST dataset of other biological tissues

To verify EAGS's adaptability to high-resolved ST data, we next apply EAGS to mouse olfactory bulb dataset. We generate the mouse olfactory bulb spatial cell map with cell type annotations (Fig. 5A) and the UMAP with cell annotation labels (Fig. 5B). The cell-annotated spatial map of the EAGS results show a clearer outline of the cells in the mouse olfactory bulb (Fig. 5A). The results of EAGS in UMAP form easily distinguishable clusters in the transcriptome space, and the clusters of different cell types have a low degree of overlap (Fig. 5B). We then calculate the CHI and DBI of the results generated without and with EAGS. EAGS can generate the results with higher intraclass similarity. Also, cells belonging to the same annotation type are closer to each other when the data has been smoothed by EAGS (Fig. 5C). Next, we count the cell types with a high proportion of Tangram cell labels to generate a spatial cell map and make a heatmap of the expression of the corresponding marker genes for different types of cells (Fig. 5D). Then we classify the sources of different cell labeling results and calculate Geary's C and Moran's I. The cell type annotation profile generated through dataset smoothed by EAGS is clearer. Also, the corresponding marker gene expression is more concentrated, and the cell types have higher Geary's C and Moran's I if the data has been processed using EAGS. These results indicate a stronger spatial autocorrelation in the transcriptome space.

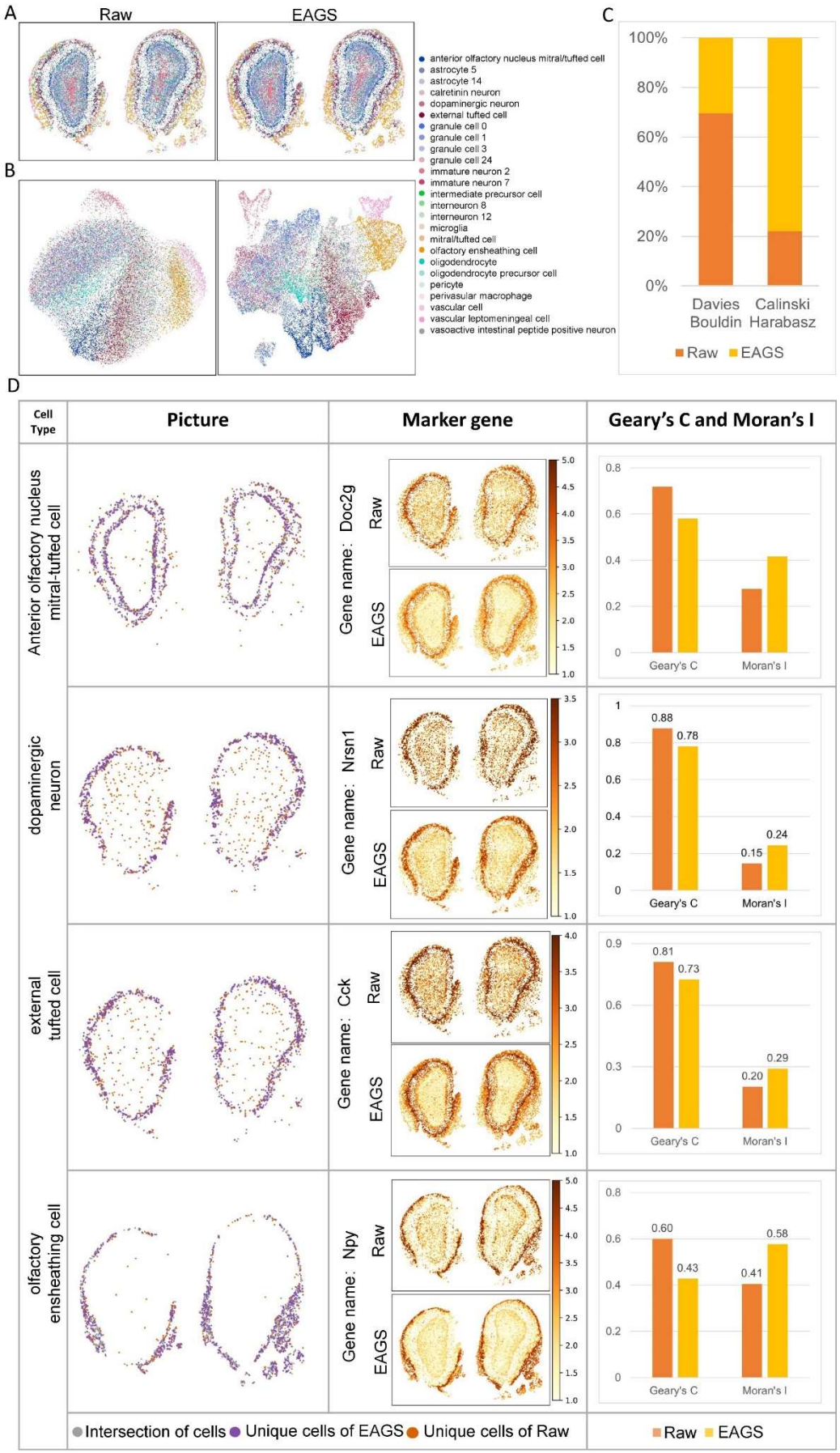
A
Raw
EAGS
B
anterior olfactory nucleus mitral/tufted cell
astrocyte 5
astrocyte 14
calretinin neuron
dopaminergic neuron
external tufted cell
granule cell 0
granule cell 1
granule cell 3
granule cell 24
immature neuron 2
immature neuron 7
intermediate precursor cell
interneuron 8
interneuron 12
microglia
mitral/tufted cell
olfactory ensheathing cell
oligodendrocyte
oligodendrocyte precursor cell
pericyte
perivasular macrophage
vascular cell
vascular leptomeningeal cell
vasoactive intestinal peptide positive neuron
C
100%
80%
60%
40%
20%
0%
Davies Bouldin
Calinski Harabasz
Raw
EAGS
D
Cell Type
Picture
Marker gene
Geary's C and Moran's I
Anterior olfactory nucleus mitral-tufted cell
Gene name: Doc2g
Raw
EAGS
Geary's C
Moran's I
dopaminergic neuron
Gene name: Nrsn1
0.88
0.78
0.15
0.24
external tufted cell
Gene name: Cck
0.81
0.73
0.20
0.29
olfactory ensheathing cell
Gene name: Npy
0.60
0.43
0.41
0.58
Intersection of cells
Unique cells of EAGS
Unique cells of Raw
Raw
EAGS

Figure 5: EAGS application to mouse olfactory bulb data. (A) Cell-annotated spatial map of data before and after EAGS smoothing. (B) Cell-annotated Umap of dataset before and after EAGS smoothing (C) Davies-Bouldin and Calinski-Harabasz Indexes of mouse olfactory bulb data. (D) How the annotation results of the main cell types of the mouse olfactory bulb differ between data without and with EAGS smoothing. We also show the heatmap of the marker genes of different cell types, and Moran's I and Geary's C indexes of the corresponding types. Cells annotated before and after smoothing (grey), cells annotated by EAGS alone (purple), and cells annotated by pre-treatment data alone (orange) are displayed on the left side; the expression heatmap of marker genes corresponding to different cell types are shown on the middle; the Moran's I and Geary's C indices are shown on the right side.

4.4 Discussion

EAGS defines patterns based on expression and spatial information. Specifically, it selects similar elements from the intersection between cells of two patterns, ensuring a reliable source of information is borrowed between smoothed cells and similar cells. The main source of smoothing information for EAGS is the smoothing weights adaptively generated based on gene expression profiles. EAGS considers the overall expression level to generate weights, avoids the appearance of a single edge value, and effectively ensures the reliability of information borrowed between cells. This allows to recover authentic cellular signals with improved intracellular similarity and spatial autocorrelation. For example, the expression of the Cartpt gene in Fig. 4D is scattered in the original data heatmap, with more noise appearing, and the matching degree with Allen's ISH image is low. EAGS smoothing consider the reliability of adjacent information. After EAGS smoothing, a lot of noise is eliminated, more Cartpt genes are expressed in the correct cells, which is more in line with Allen's ISH image, and the aggregation of Cartpt expression is significantly improved. Furthermore, EAGS improves the quality of raw data as it recovers the original biological signals by smoothing cell expression information. The dimensional space is adjusted to ensure the hidden correlation between cells. As it does not depend on a specific statistical model, EAGS does not adjust from the low-dimensional space of the expression profile, thus ensuring the hidden correlation between cells. More importantly, EAGS does not require pre-defined expression models, numerous iterations to obtain the model

parameters, or multiple training sessions on the deep learning model framework of the GPU platform. Consequently, EAGS significantly reduces computational costs and offers a significant execution advantage over other methods. Finally, because of the general applicability of smoothing, EAGS is suitable for different ST data.

It should be noted that the EAGS model is based on the premise that "neighboring" cells in the spatial microenvironment of biological tissues are more similar, which is applicable to most developmental tissue systems. However, for complex microenvironments with high biological heterogeneity (such as tumor microenvironment), this assumption will be challenged. EAGS may result in many false positive signals. When it is necessary to perform EAGS on complex tumor microenvironment samples, when calculating the adaptive Gaussian smoothing weight, the sample may need to be partitioned according to different situations, and gaussian weight will need to be calculated for different areas.

4.5 Conclusion

We propose EAGS, a method for smoothing high-resolved ST datasets that performs two-factor smoothing and adaptive weighting on raw gene expression profiles. EAGS significantly improves computing efficiency, reduces "dropout" in ST data, recovers the expression of true biological signals, and restores the spatial patterns of tissues. In the future, we will explore the false positive signals produced by EAGS imputation strategies, as well as downstream analyses of datasets after imputation.

4.6 Availability of Source Code and Requirements

Project name: EAGS: efficient and adaptive Gaussian smoothing

Project home page: https://github.com/STOmics/EAGS

Operating system(s): Platform independent

Programming language: Python

Other requirements: Python 3.8 or higher

License: MIT License

RRID: SCR_024399

BiotoolsID: EAGS

4.7 Data availability

The mouse brain dataset at single-cell resolution is available in STOmics DB of China National Gene Bank (CNGB) (accession code: "STT0000022")[175]. The mouse olfactory bulb at single-cell resolution is available in STOMICS DataBase (accession code: "STT0000027")[175]. The mouse olfactory bulb data for the Bin140 specification is available in China National Gene Bank (CNGB) (accession code: "CNP0001543") [47]. The ST data at single-cell resolution with spatial information is available in Zenodo [176]. An archival copy of the code and supporting data is available via the GigaScience repository, GigaDB[177].

5 Overall discussion

In the past five years, spatial transcriptomics has been widely regarded as a new frontier in life sciences, opening a new chapter for biological research and enabling deeper understanding of the complexity of living systems[10]. However, despite the stimulating innovation and gradual maturity of spatially resolved technology, researchers still need to invest a considerable amount of effort in seeking solutions to empower and optimize the usability of the techniques. Single-cell sequencing technology has already profoundly changed many fields of biology[2], and spatial resolved technology provides molecular *in situ* information within cells, which will drive the next generation of scientific discoveries[3]. In this regard, obtaining high-quality single-cell level spatial data in spatial transcriptomics will be the foundation for all of this. Therefore, this thesis mainly centered around optimization of tools on how to obtain high-quality single-cell spatial data.

In Chapter 2, we presented the development of StereoCell to enable data processing for production of a high-quality expression matrix. StereoCell was developed as an image-assisted cell segmentation framework for high-resolution and large-field-of-view spatial transcriptomics, providing a complete and systematic solution for obtaining high-confidence spatial single-cell level data. We implemented a tissue morphology image stitching method to ensure precise and reliable single-cell level accuracy, which can be used flexibly and conveniently. By combining FFT-based high-precision stitching methods with benchmark datasets based on chip track lines, we can improve the stitching accuracy to the subcellular level. Additionally, we have implemented a novel molecular labeling method based on GMM and cell nucleus segmentation, which can generate single-cell spatial gene expression data with higher signal-to-noise ratio, thereby obtaining more reliable cell clustering analysis and end-to-end annotation results. Furthermore, we have optimized the image processing module, including chip track line detection, cell nucleus segmentation, and tissue

segmentation. StereoCell also comes with an auxiliary software called CellbinStudio[127], which includes a graphical user interface for manual image stitching, registration, and segmentation. StereoCell provides rich documentation, including functional application programming interfaces, examples, and tutorial workflows, making it easy to use and access without specific levels of omics and image analysis expertise, whether for experienced developers or beginners. The automated processing of StereoCell is specifically designed for Stereo-seq data, as it relies on the production of images, such as ssDNA staining, DAPI staining, or H&E staining, following the Stereo-seq standard operating procedures (SOP). These images are captured to highlight the track line features on the chip and achieve precise registration. To obtain high-precision registration at the cellular level, close collaboration is required in various processes, including imaging equipment, staining SOPs, and chip/slide design, for each SRT platform such as Xenium[116] and CosMx[117]. However, the specific methods employed by commercial products are not disclosed. To ensure accurate registration results, the use of cell-level markers is essential. Therefore, the track line-assisted registration solution of Stereo-seq presents a new model inspired by sequencing chips. However, the fully automated StereoCell is only applicable to Stereo-seq technology or sequencing chips with markers similar to Stereo-seq technology (e.g., Illumina[178] sequencing chips). New marker detection methods need to be developed to replace the track lines in StereoCell. Other methods proposed by StereoCell, such as image stitching, cell segmentation, tissue segmentation, and molecular labeling, can be used independently. These methods rely on common image analysis steps and are currently suitable for processing ssDNA and DAPI images. Efforts are being made to adapt them for H&E images, providing a unified solution for nuclear staining image processing. ASHLAR[64] is the current representative solution for image-based SRTs in terms of image stitching and registration. However, practical considerations, such as running large field-of-view images efficiently, solving cumulative errors, and evaluating datasets, will need to be considered. ASHLAR is primarily applicable to image-based

SRTs and may not meet the requirements for technology based on *in situ* spatial barcoding-based SRTs. There is a significant modal difference between expression maps and staining images, and tissue heterogeneity is high. Even manual registration sometimes falls short, necessitating reliance on fixed marker features to obtain reliable results. StereoCell utilizes advanced cell/tissue segmentation frameworks like Cellpose[68] and DeepCell[67], which are trained and optimized for Stereo-seq data. These frameworks are the best choice for Stereo-seq data or specific requirements (e.g., high cell mask coverage) but may not be as suitable as existing methods (e.g., Cellpose) for general versatility or other types of research (e.g., bacteria). Molecular cell labeling is an essential step in SRT data analysis. The simplest method currently employed involves geometric binning techniques, such as using squares or circles based on a tissue-based cell size, or employing deconvolution statistics to determine cell components (as shown in Chapter 1.3.1.3). StereoCell introduces a more accurate "cellbin" concept based on the position of cell nuclei. While the position of the cell nucleus can be determined, its exact boundary cannot. To infer the cell labeling of molecules, StereoCell combines the distribution of molecules with the positions of cell nuclei, increasing the number of molecules in cells. However, this method is prone to introducing noise and has low computational efficiency. For large field-of-view data, it may even cause program crashes, similar to other methods like Baysor[69] and SCS[179]. To address these issues, an accelerated version has been developed that directly extends the cell nucleus mask. This approach does not consider the distribution characteristics of molecules but improves efficiency by more than 10 times, serving as an initial method for analysis. During the development and optimization of StereoCell, the complexity of the actual situation was found to be greater than anticipated. Factors such as tissue diversity, experimental operation instability, and the various types of images produced by microscopes directly impact data quality and difficulties with data processing. To address this, StereoCell has been modularized and provides Graphical User Interface (GUI) manual assistance software[127]. This software allows for manual

adjustments before integration, and StereoCell supports module upgrades and results replacements, which greatly enhances the scalability and compatibility of the entire process. However, due to its compatibility, more effort is required to ensure accuracy. In the future, there are plans to integrate StereoCell into a platform similar to the foundation model in Generative Pre-trained Transformer (GPT) fields, where researchers can fine-tune the tissue segmentation, cell segmentation, and molecular labeling models according to their specific data. This will allow for achieving the highest accuracy and simultaneous application in various types of research.

With the advancement of spatial technology to encompass single cells, a method for capturing of true cell boundaries images was required. In Chapter 3, we upgraded StereoCell to STCellbin to develop a more accurate cell binning solution that can generate single-cell spatial gene expression profiles using Stereo-seq staining images of cell membranes/cell walls. We retained the steps of image stitching, tissue segmentation, and improved the steps of image registration and cell segmentation. Since the staining images of cell membranes/cell walls lack "track lines", in the image registration step, we used the staining images of cell nuclei (DAPI) as a bridge to align the images of cell membranes or walls with the spatial gene expression map, thereby obtaining registered cell boundary information. In the cell segmentation step, we used advanced technique to segment the staining images of cell membranes/cell walls, thus obtaining cell boundaries. Based on the cell boundary information, we directly assigned molecules to the corresponding cells, obtaining single-cell spatial gene expression profiles. We applied STCellbin to mouse liver (cell membrane) and *Arabidopsis* seed (cell wall) datasets and confirmed the accuracy of cell segmentation. This update provides a comprehensive and universal processing workflow, which can be used to process any imaging information of cell membranes/cell walls produced by spatial technology, to obtain reliable single-cell spatial gene expression profiles based on cell membrane/cell wall information. STCellbin is a result derived from the inheritance of the advantages of StereoCell. Its prominent feature lies in its ability to provide genuine

cellular boundary information, which serves as the gold standard for cellbin[9]. It offers a strategy for acquiring a single-cell spatial matrix using mIF images. However, the automation aspect is limited to Stereo-seq data, and the concepts of registration and segmentation can be referred to separately. From an alternative perspective, when simultaneously obtaining the results of both StereoCell and STCellbin, the preferred choice or the means to achieve the best outcome is to adopt the results of STCellbin. In the scheme of cell segmentation, such as DeepCell[67], when images of both cell nuclei and cell membranes are inputted simultaneously, the results of cell membranes and cell nuclei are directly combined to produce a single final mask (it was discovered during testing that the results based on cell membranes are the most accurate). In the liver data utilized in our study, the results derived from cell membranes are relatively satisfactory, and most cells have been successfully obtained (cells are densely arranged), making it acceptable to not consider the results of cell nuclei. However, in real-life scenarios, mIF can only stain specific cells, and some cells may remain unstained. In general, combining cell nuclei and cell membranes to identify cell positions and boundaries would offer greater reliability, and this represents a potential future direction.

After obtaining a spatial expression map, the quality of these data directly affects downstream analysis. In Chapter 4, we implemented efficient and adaptive Gaussian smoothing (EAGS) to enhance the ability to capture signals from transcripts in Stereo-seq data. EAGS defines a processing mode based on both gene expression and spatial information. Specifically, it selects similar elements from the intersection of the two modes' units, ensuring reliable borrowing of information between smooth units and similar units. The main source of EAGS smoothing information is the adaptive generation of smoothing weights based on gene expression profiles, considering the overall expression level to generate weights, avoiding the occurrence of single-edge values, and effectively ensuring the reliability of intercellular information borrowing. This allows for the recovery of true cell signals and improved intracellular similarity and spatial autocorrelation. After EAGS smoothing, a large amount of noise is

eliminated, and more specific genes are left in the correct cells. In addition, EAGS improves the quality of the original data by smoothing the cell expression information and adjusts the dimensional space to ensure the hidden correlations between cells. Since it does not rely on specific statistical models, EAGS does not adjust from the low-dimensional space of the expression profile, ensuring the hidden correlations between cells. Importantly, EAGS does not require pre-defined expression models, multiple iterations to obtain model parameters, or multiple training on a GPU platform's deep learning model framework. Therefore, compared to other methods, EAGS significantly reduces computational costs and provides significant operational advantages. Finally, due to its general applicability, EAGS is suitable for different ST data, here, mainly the mouse brain and olfactory bulb data of Stereo-seq were tested. EAGS is a statistical model-based method, it should be noted that the model is based on the premise that "adjacent" cells in the biological tissue spatial microenvironment are more similar than “distant” cells, which is applicable to most developmental tissue systems. However, this assumption will be challenged in complex microenvironments with high biological heterogeneity, such as tumor microenvironments. In such cases, EAGS may result in many false positive signals. When applying EAGS to complex tumor microenvironment tissue, it may be necessary to segment the samples and calculate Gaussian weights for different regions based on different situations when computing adaptive Gaussian smoothing weights. The complexity and diversity of spatial resolved technology also results in various specific characteristics of the acquired signals. Currently, due to the bottleneck of capture efficiency, signal loss exists in most cases, so imputation is an important but controversial preprocessing method[4]. In the future, we will explore the false positive signals generated by the EAGS interpolation strategy and the downstream analysis of the interpolated dataset.

In general, the developed image processing methods and spatial data processing frameworks are compatible with other spatial single-cell sequencing technologies. In addition, the joint analysis of multimodal and multi-omics data, such as obtaining the

boundary of the cell nucleus from the cell nucleus image, obtaining the true cell boundary information from the cell membrane/wall image, and even the clustering of gene expression within single cells across a tissue (molecular aggregation), can assist in obtaining cell boundary information. The combined use of the above information will play a positive role in the accuracy of cell positioning and cell area delineation. Moreover, in real-world scenarios, apart from the cell nuclei, cell walls, and cell membranes depicted in StereoCell and STCellbin, numerous microscopy images of diverse types offer researchers abundant insights into tissue biology[9,10]. Presently, the processing techniques solely utilize the cell positions and cell boundary information derived from aligned images, lacking in-depth integration and analysis of the comprehensive microscopy image data. This aspect merits further exploration in future research activities. With the development of spatial technology, we not only focus on the transcriptome, but also on spatial multi-omics, which is an important direction of development[10]. The focus in this thesis was on the generation of tools for capturing and processing of high-quality spatial single-cell data, but in the future the proteome, epigenome, and metabolome, will add on additional layers of information that will need to be integrated to obtain more complete maps of single cells within tissues, and how they interact during different developmental stages and conditions.

6 Overall conclusion

Integrating the advantages of spatial resolved technology and single-cell sequencing to obtain accurate and reliable spatial single-cell atlases has been a challenging and crucial task. As part of this thesis work, three analytical tools were developed with the overall aim of providing more efficient and accurate tools to support analysis of Stereo-seq generated data. The three tools, StereoCell, STCellbin and EAGS, were each implemented to provide different features:

To obtain high-confidence spatial single-cell level transcriptomic data maps from large field-of-view tissue images and high-resolution single-cell gene expression data, we first developed the processing pipeline StereoCell. The pipeline provides a systematic platform for accurate single-cell spatial data acquisition, including image stitching, registration, cell nucleus segmentation, and molecular labeling. Compared to existing methods, StereoCell offers improved algorithms for reducing stitching errors and time during the image stitching and molecular labeling processes, and also enhances the signal-to-noise ratio of single-cell gene expression data. StereoCell is designed to be user-friendly and does not require a specific level of expertise in omics and image analysis. It has showed instrumental in creating a comprehensive spatiotemporal transcriptomic atlas of mouse organogenesis and has been successfully used to generate reliable single-cell spatial gene expression profiles from continuous datasets of mouse brain slices in previous studies[47]. StereoCell is a high-speed tool for analyzing image and spatial omics data. It has demonstrated its ability to handle a large mouse brain dataset (consisting of 131,990,020 molecules and 117 image tiles) in approximately 80 minutes on a server equipped with a 40-core CPU, 128 GB of RAM, and 24 GB of GPU.

To obtain accurate information about cell boundaries for generating more reliable single-cell spatial gene expression profiles, we developed the tool STCellbin. It is an update of StereoCell with implementation of the cell membrane/cell wall staining

images. As part of the image registration step, we utilize cell nucleus staining images as a bridge to align the cell membrane/cell wall staining images with the spatial gene expression profiles, thereby obtaining registered cell boundary information. In the cell segmentation step, we employ an advanced technique to segment the cell membrane/cell wall staining images and obtain cell boundaries. Based on the cell boundary information, we directly assign molecules to their corresponding cells to obtain single-cell spatial gene expression profiles. When applied to mouse liver (cell membrane) and *Arabidopsis* seed (cell wall) datasets, STCellbin demonstrates improved accuracy of cell segmentation, providing valuable insights into spatial organization with this enhanced functionality.

For smoothing high-resolution ST data, and apply dual-factor smoothing and adaptive weighting to the original gene expression profiles, the method EAGS was developed. EAGS was found to significantly improve computational efficiency while reducing "dropout" in ST data, restoring the expression of true biological signals, and restoring the spatial patterns of the tissue.

Altogether, this thesis work has generated knowledge ending up in development of three analytical tools that are incorporating in-depth research from how to obtain single-cell spatial data to how to improve the quality of matrix data, involving image processing techniques, spatial transcriptomics feature mining, and statistics. These solutions have the potential to become a bridge between image analysis and molecular omics fields, providing a foundation for the development of computational methods for next-generation spatially resolved technologies.

7 References

1. Levsky, J. Gene expression and the myth of the average cell. *Trends Cell Biol.* **13**, 4–6 (2003).
2. Method of the Year 2013. *Nat. Methods* **11**, 1–1 (2014).
3. Method of the Year 2020: spatially resolved transcriptomics. *Nat. Methods* **18**, 1–1 (2021).
4. Cheng, M. *et al.* Spatially resolved transcriptomics: a comprehensive review of their technological advances, applications, and challenges. *J. Genet. Genomics* **50**, 625–640 (2023).
5. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
6. Strell, C. *et al.* Placing RNA in context and space – methods for spatially resolved transcriptomics. *FEBS J.* **286**, 1468–1481 (2019).
7. Crosetto, N., Bienko, M. & Van Oudenaarden, A. Spatially resolved transcriptomics and beyond. *Nat. Rev. Genet.* **16**, 57–66 (2015).
8. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
9. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546 (2022).
10. Bressan, D., Battistoni, G. & Hannon, G. J. The dawn of spatial omics. *Science* **381**, eabq4964 (2023).
11. Tian, L., Chen, F. & Macosko, E. Z. The expanding vistas of spatial transcriptomics. *Nat. Biotechnol.* **41**, 773–782 (2023).
12. Grisanti Canozo, F. J., Zuo, Z., Martin, J. F. & Samee, Md. A. H. Cell-type modeling in spatial transcriptomics data elucidates spatially variable colocalization and communication between cell-types in mouse brain. *Cell Syst.*

13, 58-70.e5 (2022).

13. Moffitt, J. R., Lundberg, E. & Heyn, H. The emerging landscape of spatial profiling technologies. *Nat. Rev. Genet.* (2022) doi:10.1038/s41576-022-00515-3.
14. He, Z. *et al.* Lineage recording in human cerebral organoids. *Nat. Methods* **19**, 90–99 (2022).
15. Lohoff, T. *et al.* Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat. Biotechnol.* **40**, 74–85 (2022).
16. Parigi, S. M. *et al.* The spatial transcriptomic landscape of the healing mouse intestine following damage. *Nat. Commun.* **13**, 828 (2022).
17. Cross, A. R. *et al.* Spatial transcriptomic characterization of COVID-19 pneumonitis identifies immune circuits related to tissue injury. *JCI Insight* **8**, e157837 (2023).
18. Gall, J. G. & Pardue, M. L. FORMATION AND DETECTION OF RNA-DNA HYBRID MOLECULES IN CYTOLOGICAL PREPARATIONS. *Proc. Natl. Acad. Sci.* **63**, 378–383 (1969).
19. Chen, J. *et al.* Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat. Protoc.* **12**, 566–580 (2017).
20. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
21. Liu, Y. *et al.* High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell* **183**, 1665-1681.e18 (2020).
22. Cho, C.-S. *et al.* Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* **184**, 3559-3572.e22 (2021).
23. Fu, X. *et al.* Polony gels enable amplifiable DNA stamping and spatial transcriptomics of chronic pain. *Cell* **185**, 4621-4633.e17 (2022).
24. Meier-Ruge, W. *et al.* The laser in the Lowry technique for microdissection of freeze-dried tissue slices. *Histochem. J.* **8**, 387–401 (1976).
25. Ke, R. *et al. In situ* sequencing for RNA analysis in preserved tissue and cells. *Nat.*

Methods **10**, 857–860 (2013).

26. GEORGE T. RUDKIN & B. D. STOLLAR. High resolution detection of DNA–RNA hybrids *in situ* by indirect immunofluorescence. *Nature* **265**, 472–473 (1977).
27. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of Single RNA Transcripts *in Situ*. *Sci. New Ser.* **280**, 585–590 (1998).
28. Lee, J. H. *et al.* Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10**, (2015).
29. Shah, S. *et al.* Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. *Cell* **174**, 363-376.e16 (2018).
30. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
31. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
32. Chen, X., Sun, Y.-C., Church, G. M., Lee, J. H. & Zador, A. M. Efficient *in situ* barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Res.* (2017) doi:10.1093/nar/gkx1206.
33. Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).
34. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).
35. Gyllborg, D. *et al.* Hybridization-based *In Situ* Sequencing (HybISS): spatial transcriptomic detection in human and mouse brain tissue. *biorxiv* (2020) doi:10.1101/2020.02.03.931618.
36. Wang, Y. *et al.* EASI-FISH for thick tissue defines lateral hypothalamus spatio-molecular organization. *Cell* **184**, 6361-6377.e24 (2021).
37. Borm, L. E. *et al.* Scalable *in situ* single-cell profiling by electrophoretic capture of mRNA using EEL FISH. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-

01455-3.

38. Zeng, H. *et al.* Integrative *in situ* mapping of single-cell transcriptional states and tissue histopathology in a mouse model of Alzheimer's disease. *Nat. Neurosci.* (2023) doi:10.1038/s41593-022-01251-x.
39. Emmert-Buck, M. R. *et al.* Laser Capture Microdissection. *Science* **274**, 998–1001 (1996).
40. Lovatt, D. *et al.* Transcriptome *in vivo* analysis (TIVA) of spatially defined single cells in live tissue. *Nat. Methods* **11**, 190–196 (2014).
41. Junker, J. P. *et al.* Genome-wide RNA Tomography in the Zebrafish Embryo. *Cell* **159**, 662–675 (2014).
42. Vickovic, S. *et al.* High-definition spatial transcriptomics for *in situ* tissue profiling. *Nat. Methods* **16**, 987–990 (2019).
43. Merritt, C. R. *et al.* Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat. Biotechnol.* **38**, 586–599 (2020).
44. Ji, A. L. *et al.* Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell* **182**, 497-514.e22 (2020).
45. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
46. Lee, Y. *et al.* XYZeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. *Sci. Adv.* **7**, eabg4755 (2021).
47. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777-1792.e21 (2022).
48. Schede, H. H. *et al.* Spatial tissue profiling by imaging-free molecular tomography. *Nat. Biotechnol.* **39**, 968–977 (2021).
49. Honda, M. *et al.* High-depth spatial transcriptome analysis by photo-isolation chemistry. *Nat. Commun.* **12**, 4416 (2021).
50. Kishi, J. Y. *et al.* Light-Seq: light-directed *in situ* barcoding of biomolecules in

fixed cells and tissues for spatially indexed sequencing. *Nat. Methods* **19**, 1393–1402 (2022).

51. Hu, K. H. *et al.* ZipSeq: barcoding for real-time mapping of single cell transcriptomes. *Nat. Methods* **17**, 833–843 (2020).
52. Kishi, J. Y. *et al.* SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. *Nat. Methods* **16**, 533–544 (2019).
53. Rouhanifard, S. H. *et al.* ClampFISH detects individual nucleic acid molecules using click chemistry–based amplification. *Nat. Biotechnol.* **37**, 84–89 (2019).
54. Sountoulidis, A. *et al.* SCRINSHOT enables spatial mapping of cell states in tissue sections with single-cell resolution. *PLOS Biol.* **18**, e3000675 (2020).
55. Goh, J. J. L. *et al.* Highly specific multiplexed RNA imaging in tissues with split-FISH. *Nat. Methods* **17**, 689–693 (2020).
56. Battich, N., Stoeger, T. & Pelkmans, L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat. Methods* **10**, 1127–1133 (2013).
57. Shah, S., Lubeck, E., Zhou, W. & Cai, L. *In Situ* Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92**, 342–357 (2016).
58. Chen, F., Tillberg, P. W. & Boyden, E. S. Expansion microscopy. *Science* **347**, 543–548 (2015).
59. Liu, S. *et al.* Barcoded oligonucleotides ligated on RNA amplified for multiplexed and parallel *in situ* analyses. *Nucleic Acids Res.* **49**, e58–e58 (2021).
60. Shendure, J. *et al.* Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Sci. New Ser.* **309**, 1728–1732 (2005).
61. Alon, S. *et al.* Expansion sequencing: Spatially precise *in situ* transcriptomics in intact biological systems. *Science* **371**, eaax2656 (2021).
62. Liu, Y. *et al.* High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell* **183**, 1665-1681.e18 (2020).

63. Weinstein, J. STARFISH ENTERPRISE: RNA GOES SPATIAL. *NATURE* **572**, 549–551 (2019).

64. Muhlich, J. L. *et al.* Stitching and registering highly multiplexed whole-slide images of tissues and tumors using ASHLAR. *Bioinformatics* **38**, 4613–4621 (2022).

65. Wong, K., Navarro, J. F., Bergenstråhle, L., Ståhl, P. L. & Lundeberg, J. ST Spot Detector: a web-based application for automatic spot and tissue detection for spatial Transcriptomics image datasets. *Bioinformatics* **34**, 1966–1968 (2018).

66. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).

67. Greenwald, N. F. *et al.* Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2022).

68. Pachitariu, M. & Stringer, C. Cellpose 2.0: how to train your own model. *Nat. Methods* **19**, 1634–1641 (2022).

69. Petukhov, V. *et al.* Cell segmentation in imaging-based spatial transcriptomics. *Nat. Biotechnol.* **40**, 345–354 (2022).

70. Littman, R. *et al.* Joint cell segmentation and cell type annotation for spatial transcriptomics. *Mol. Syst. Biol.* **17**, e10108 (2021).

71. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **40**, 517–526 (2022).

72. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).

73. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* **49**, e50–e50 (2021).

74. Dong, R. & Yuan, G.-C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol.* **22**, 145 (2021).

75. Wei, R. *et al.* Spatial charting of single-cell transcriptomes in tissues. *Nat. Biotechnol.* **40**, 1190–1199 (2022).

76. Robert E. Schapire. *Random Forests*. vols 5–32 (Kluwer Academic Publishers, 2001).

77. Biancalani, T. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).

78. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).

79. Herman, J. S., Sagar & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379–386 (2018).

80. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

81. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

82. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

83. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

84. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01767-y.

85. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

86. Righelli, D. *et al.* SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. *Bioinformatics* **38**, 3128–3131 (2022).

87. Dries, R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial

expression data. *Genome Biol.* **22**, 78 (2021).

88. Bergenstråhle, J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* **21**, (2020).
89. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
90. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).
91. BinTayyash, N. *et al.* Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments. *Bioinformatics* **37**, 3788–3795 (2021).
92. Govek, K. W., Yamajala, V. S. & Camara, P. G. Clustering-independent analysis of genomic data using spectral simplicial theory. *PLOS Comput. Biol.* **15**, e1007509 (2019).
93. Miller, B. F., Bambah-Mukku, D., Dulac, C., Zhuang, X. & Fan, J. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. *Genome Res.* **31**, 1843–1855 (2021).
94. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).
95. Zhang, Y. *et al.* CellCall: integrating paired ligand–receptor and transcription factor activities for cell–cell communication. *Nucleic Acids Res.* **49**, 8520–8534 (2021).
96. Yuan, Y. & Bar-Joseph, Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biol.* **21**, 300 (2020).
97. Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J. & Stegle, O. Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis. *Cell Rep.* **29**, 202-211.e6 (2019).
98. Jerby-Arnon, L. & Regev, A. DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data. *Nat. Biotechnol.* **40**, 1467–1477

(2022).

99. Medaglia, C. *et al.* Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. *Science* **358**, 1622–1626 (2017).
100. Li, H., Ma, T., Hao, M., Wei, L. & Zhang, X. Decoding functional cell–cell communication events by multi-view graph learning on spatial transcriptomics. *bioRxiv* (2022) doi:10.1101/2022.06.22.496105.
101. Shao, X. *et al.* Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with SpaTalk. *Nat. Commun.* **13**, 4429 (2022).
102. Mohamed, S. K., Nounu, A. & Nováček, V. Biological applications of knowledge graph embedding models. *Brief. Bioinform.* **22**, 1679–1693 (2021).
103. Berglund, E. *et al.* Automation of Spatial Transcriptomics library preparation to enable rapid and robust insights into spatial organization of tissues. *BMC Genomics* **21**, 298 (2020).
104. Cheng, J., Liao, J., Shao, X., Lu, X. & Fan, X. Multiplexing Methods for Simultaneous Large-Scale Transcriptomic Profiling of Samples at Single-Cell Resolution. *Adv. Sci.* **8**, 2101229 (2021).
105. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **13**, 1739 (2022).
106. Fang, S. *et al.* Computational Approaches and Challenges in Spatial Transcriptomics. *Genomics Proteomics Bioinformatics* **21**, 24–47 (2023).
107. Zeira, R., Land, M., Strzalkowski, A. & Raphael, B. J. Alignment and integration of spatial transcriptomics data. *Nat. Methods* **19**, 567–575 (2022).
108. Li, B. *et al.* Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat. Methods* **19**, 662–670 (2022).
109. Kriebel, A. R. & Welch, J. D. UINMF performs mosaic integration of single-cell

multi-omic datasets using nonnegative matrix factorization. *Nat. Commun.* **13**, 780 (2022).

110. Hu, J. *et al.* SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
111. Wei, X. *et al.* Single-cell Stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science* **377**, eabp9444 (2022).
112. Chen, A. *et al.* Single-cell spatial transcriptome reveals cell-type organization in the macaque cortex. *Cell* **186**, 3726-3743.e24 (2023).
113. Wu, L. *et al.* An invasive zone in human liver cancer identified by Stereo-seq promotes hepatocyte–tumor cell crosstalk, local immunosuppression and tumor progression. *Cell Res.* **33**, 585–603 (2023).
114. Zhang, R. *et al.* Spatial transcriptome unveils a discontinuous inflammatory pattern in proficient mismatch repair colorectal adenocarcinoma. *Fundam. Res.* **3**, 640–646 (2023).
115. Xia, K. *et al.* The single-cell stereo-seq reveals region-specific cell subtypes and transcriptome profiling in Arabidopsis leaves. *Dev. Cell* **57**, 1299-1310.e4 (2022).
116. Janesick, A. *et al.* High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and *in situ* analysis of FFPE tissue. *biorxiv* (2022) doi:10.1101/2022.10.06.510405.
117. He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat. Biotechnol.* **40**, 1794–1806 (2022).
118. Lu, T., Ang, C. E. & Zhuang, X. Spatially resolved epigenomic profiling of single cells in complex tissues. *Cell* **185**, 4448-4464.e17 (2022).
119. Erickson, A. *et al.* Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature* **608**, 360–367 (2022).
120. Grünwald, B. T. *et al.* Spatially confined sub-tumor microenvironments in pancreatic cancer. *Cell* **184**, 5577-5592.e18 (2021).

121. Zhou, D. C. Spatially restricted drivers and transitional cell populations cooperate with the microenvironment in untreated and chemo-resistant pancreatic cancer. *Nat. Genet.* **54**, 1390–1405 (2022).
122. Duhamel, P. & Vetterli, M. Fast fourier transforms: A tutorial review and a state of the art. *Signal Process.* **19**, 259–299 (1990).
123. Reynolds, D. Gaussian Mixture Models. *Encycl Biom* **741**, 659–663 (2009).
124. Chalfoun, J. *et al.* MIST: Accurate and Scalable Microscopy Image Stitching Tool with Stage Modeling and Error Minimization. *Sci. Rep.* **7**, 4988 (2017).
125. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
126. Qiu, X. *et al.* Spateo: multidimensional spatiotemporal modeling of single-cell spatial transcriptomics. *biorxiv* (2022) doi:10.1101/2022.12.07.519417.
127. CellbinStudio. https://github.com/STOmics/StereoCell/tree/dev.
128. Phase_Image_Tiles. https://isg.nist.gov/BII_2015/Stitching/Phase_Image_Tiles.zip.
129. Taheri, S. M. & Hesamian, G. A generalization of the Wilcoxon signed-rank test and its applications. *Stat. Pap.* **54**, 457–470 (2013).
130. Stereopy. https://stereopy.readthedocs.io/en/latest/index.html.
131. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
132. Brann, D. H. *et al.* Non-neuronal expression of SARS-CoV-2 entry genes in the olfactory system suggests mechanisms underlying COVID-19-associated anosmia. *Sci. Adv.* **6**, eabc5801 (2020).
133. Shen, R. *et al.* Spatial-ID: a cell typing method for spatially resolved transcriptomics via transfer learning and spatial embedding. *Nat. Commun.* **13**, 7640 (2022).
134. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014.e22 (2018).

135. Azad, R., Asadi-Aghbolaghi, M., Fathy, M. & Escalera, S. Bi-Directional ConvLSTM U-Net with Densley Connected Convolutions. in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* 406–415 (IEEE, 2019). doi:10.1109/ICCVW.2019.00052.
136. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) vol. 9351 234–241 (Springer International Publishing, 2015).
137. Zhang, Z., Liu, Q. & Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **15**, 749–753 (2018).
138. Zhang, H., Zu, K., Lu, J., Zou, Y. & Meng, D. EPSANet: An Efficient Pyramid Squeeze Attention Block on Convolutional Neural Network. Preprint at http://arxiv.org/abs/2105.14447 (2021).
139. Li, M. *et al.* StereoCell enables highly accurate single-cell segmentation for spatial transcriptomics. *biorxiv* (2023) doi:10.1101/2023.02.28.530414.
140. STOmics Stereo-seq Transcriptomics Set for mIF User Manual_A20210427. https://cdn-newfile.stomics.tech/static-hash/file/STOmics%20Stereo-seq%20Transcriptomics%20Set%20for%20mIF%20User%20Manual_A20210427.pdf.
141. Liao, S. *et al.* Integrated Spatial Transcriptomic and Proteomic Analysis of Fresh Frozen Tissue Based on Stereo-seq. *biorxiv* (2023) doi:10.1101/2023.04.28.538364.
142. Levina, A. & Priesemann, V. Subsampling scaling. *Nat. Commun.* **8**, 15140 (2017).
143. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
144. Tongxuan Lv *et al.* EAGS: efficient and adaptive Gaussian smoothing applied to high-resolved spatial transcriptomics. *GigaScience* (2023) doi:https://doi.org/10.1093/gigascience/giad097.

145. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* **22**, 627–644 (2021).

146. Wang, M. *et al.* High-resolution 3D spatiotemporal transcriptomic maps of developing Drosophila embryos and larvae. *Dev. Cell* **57**, 1271-1283.e4 (2022).

147. Liu, C. *et al.* Spatiotemporal mapping of gene expression landscapes and developmental trajectories during zebrafish embryogenesis. *Dev. Cell* **57**, 1284-1298.e5 (2022).

148. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).

149. Ly, L.-H. & Vingron, M. Effect of imputation on gene network reconstruction from single-cell RNA-seq data. *Patterns* **3**, 100414 (2022).

150. Xu, J. *et al.* Evaluating the performance of dropout imputation and clustering methods for single-cell RNA sequencing data. *Comput. Biol. Med.* **146**, 105697 (2022).

151. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21**, 218 (2020).

152. Van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729.e27 (2018).

153. Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. & Garry, D. J. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* **19**, 220 (2018).

154. Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**, 539–542 (2018).

155. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).

156. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).

157. Wang, Y. *et al.* Sprod for de-noising spatially resolved transcriptomics data based on position and image information. *Nat. Methods* **19**, 950–958 (2022).

158. Park, W., Chang, W., Lee, D., Kim, J. & Hwang, S. GRPE: Relative Positional Encoding for Graph Transformer. Preprint at http://arxiv.org/abs/2201.12787 (2022).

159. Liu, Y. *et al.* SPCS: a spatial and pattern combined smoothing method for spatial transcriptomic expression. *Brief. Bioinform.* **23**, bbac116 (2022).

160. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).

161. Chao Zhang *et al.* spatiAlign: An Unsupervised Contrastive Learning Model for Data Integration of Spatially Resolved Transcriptomics. *biorxiv* (2023) doi:10.1101/2023.08.08.552402v2.

162. Virshup, I. *et al.* The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat. Biotechnol.* **41**, 604–606 (2023).

163. Omohundro, S. M. *Five Balltree Construction Algorithms*. (1989).

164. Kumar, N., Zhang, L. & Nayar, S. What Is a Good Nearest Neighbors Algorithm for Finding Similar Patches in Images? in *Computer Vision – ECCV 2008* (eds. Forsyth, D., Torr, P. & Zisserman, A.) vol. 5303 364–378 (Springer Berlin Heidelberg, 2008).

165. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

166. Chen, S., Yan, X., Zheng, R. & Li, M. Bubble: a fast single-cell RNA-seq imputation using an autoencoder constrained by bulk RNA-seq data. *Brief. Bioinform.* **24**, bbac580 (2023).

167. Desgraupes, B. Clustering Indices. *Univ. Paris Ouest Lab Modal'X* **1**, 34 (2013).

168. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat. - Theory Methods* **3**, 1–27 (1974).

169. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).

170. Moran, P. A. P. Notes on Continuous Stochastic Phenomena. *Biometrika* **37**, 17 (1950).

171. Geary, R. C. The Contiguity Ratio and Statistical Mapping. *Inc. Stat.* **5**, 115 (1954).

172. Chen, G., Ning, B. & Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.* **10**, 317 (2019).

173. Song, D. *et al.* scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01772-1.

174. Wagner, F., Yan, Y. & Yanai, I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *biorxiv* (2017) doi:10.1101/217737.

175. Xu, Z. *et al.* STOmicsDB: a database of Spatial Transcriptomic data. *biorxiv* (2022) doi:10.1101/2022.03.11.481421.

176. Lv1, T., Zhang, Y., Li, M., & Qiang Kang. EAGS: efficient and adaptive gaussian smoothing applied to high-resolved spatial transcriptomics (Version 1) [Data set]. *Zenodo 2023* doi:https://doi.org/10.5281/zenodo.7906815.

177. Xu X, Lv T, Zhang Y, *et al*. Supporting data for "EAGS: efficient and adaptive Gaussian smoothing applied to high-resolved spatial transcriptomics." *GigaScience Database 2023*. http://dx.doi.org/10.5524/102457

178. Illumina. https://www.illumina.com/.

179. Chen, H., Li, D. & Bar-Joseph, Z. SCS: cell segmentation for high-resolution spatial transcriptomics. *Nat. Methods* **20**, 1237–1243 (2023).