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.

ABSTRACT

Stereo-seq is a cutting-edge technique for spatially 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 generate 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 images as a bridge to align cell membrane/wall staining images 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.

STATEMENT OF NEED

Spatially 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 [1], tumor heterogeneity [2], cancer evolution [3], 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 [4]. Our previous work offers the one-stop software StereoCell for acquiring high signal-to-noise ratio single-cell spatial gene expression profiles from Stereo-seq data [5]. The image data generated by Stereo-seq used for StereoCell are cell nuclei staining images. 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 [6], CosMx [7], and Xenium [8], several of these techniques still struggle to achieve accurate cell boundary information, as they are based on cell nuclei staining images that can be generated using stains such as 4,6-diamidino-2-phenylindole (DAPI). Hematoxylin-eosin (H&E) and single strand DNA fluorescence (ssDNA) staining images are also commonly used and readily obtainable data. We here implement a procedure based on simultaneous cell membrane/wall and cell nuclei staining using multiplex immunofluorescence (mIF) and

calcofluor white (CFW) staining [9,10], 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 [5], we utilize the cell nuclei staining images as a bridge to align the cell membrane/wall staining images with the spatial gene expression maps, 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 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 data.

IMPLEMENTATION

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 [5]. The stitched cell nuclei and cell membrane/wall staining images are registered using the Fast Fourier Transform (FFT)

algorithm [11]. 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 [12] 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 [13] is set as optional, which can generate a tissue mask to assist in filtering out impurities outside the tissue.

Image stitching

The image stitching steps in STCellbin is consistent with the image stitching steps in StereoCell. The MFWS algorithm [5] is adopted, which calculates the offsets of two adjacent tiles with overlapping areas using FFT [11] 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 [5].

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 [11] is then used to calculate the image offsets (similar to MFWS [5]). To save computing

resources, the two stained images are mean-based subsampled [14] (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 cell nuclei and cell membrane/wall staining images can be registered (Fig. 2D). The second registration is the same as in StereoCell [5], 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 cell 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.

Cell segmentation

The cell segmentation step of STCellbin is performed using Cellpose 2.0 [12] 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.

Molecule labeling

The molecule labeling of STCellbin is the same as the one used in StereoCell in principle. 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 [15]. 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.

RESULTS

Datasets

We adopt two datasets acquired via Stereo-seq technology [4]. 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

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, …, *ns*). Then a rule is set:

$$\begin{cases} \text{s_cell}_i \text{ is segmented correctly} & \text{if } IoU_i > 0.5 \\ \text{s_cell}_i \text{ is segmented incorrectly} & \text{otherwise} \end{cases} \tag{1}$$

where *IoU* is the standard intersection over union metric [16] 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.

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 [17] (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.

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.

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

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.

LIST OF ABBREVIATIONS

DAPI: 4,6-diamidino-2-phenylindole; H&E: hematoxylin-eosin; ssDNA: single strand DNA fluorescence; mIF: multiplex immunofluorescence; CFW: calcofluor white; FFT: Fast Fourier Transform.

DECLARATIONS

Ethics Approval and Consent to Participate

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Key R&D Program of China (2022YFC3400400).

Authors' Contributions

Conceptualization: BZ and ML; Project administration and supervision: SB and XX; Software implementation: ZD, HQ, KS and HL; Data collection and processing: QK, XF and LC; Validation:

QK and ZD. Project coordination: BZ and ML; Manuscript writing and figure generation: BZ, ML and QK; Manuscript review: ML, SF, YZ, YL and SB.

Acknowledgements

We thank China National GeneBank for providing technical support.

REFERENCES

1. Fang S, Chen B, Zhang Y, Sun H, Liu L, Liu S, et al. Computational Approaches and Challenges in Spatial Transcriptomics. *Genom Proteom Bioinf*. 2023;21:24–47.
2. Lu T, Ang CE, Zhuang X. Spatially resolved epigenomic profiling of single cells in complex tissues. *Cell*. 2022;185:4448–4464.
3. Erickson A, He M, Berglund E, Marklund M, Mirzazadeh R, Schultz N, et al. Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature*. 2022;608:360–367.
4. Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*. 2022;185:1777-1792.
5. Li M, Liu H, Li M, Fang S, Kang Q, Zhang J, et al. StereoCell enables highly accurate single-cell segmentation for spatial transcriptomics. *bioRxiv*. 2023.
6. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*. 2015;348:aaa6090.
7. He S, Bhatt R, Brown C, Brown EA, Buhr DL, Chantranuvatana K, et al. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat Biotechnol*. 2022;40:1794-1806.
8. Janesick A, Shelansky R, Gottscho AD, Wagner F, Rouault M, Beliakoff G, 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.

9. Liao S, Heng Y, Liu W, Xiang J, Ma Y, Chen L, et al. Integrated Spatial Transcriptomic and Proteomic Analysis of Fresh Frozen Tissue Based on Stereo-seq. *bioRxiv*. 2023.
10. STOmics Decuments. https://en.stomics.tech/resources/documentsdocuments.
11. Duhamel P, Vetterli M. Fast fourier transforms: A tutorial review and a state of the art. *Signal Process*. 1990;19(4):259-299.
12. Pachitariu M, Stringer C. Cellpose 2.0: how to train your own model. *Nat Methods*. 2022;19:1634-1641.
13. Azad R, Asadi-Aghbolaghi M, Fathy M, Escalera S. Bi-Directional ConvLSTM U-Net with densley connected convolutions. *arXiv*. 2019.
14. Levina A, Priesemann V. Subsampling scaling. *Nat Commun*. 2017;8:15140.
15. Reynolds D. Gaussian Mixture Models. *Encycl Biom*. 2009;741:659-663.
16. Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: a generalist algorithm for cellular Segmentation. *Nat Methods*. 2021;18:100-106.
17. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*. 2019;9:5233.

Figure legends

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.

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.

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.

Fig. 1

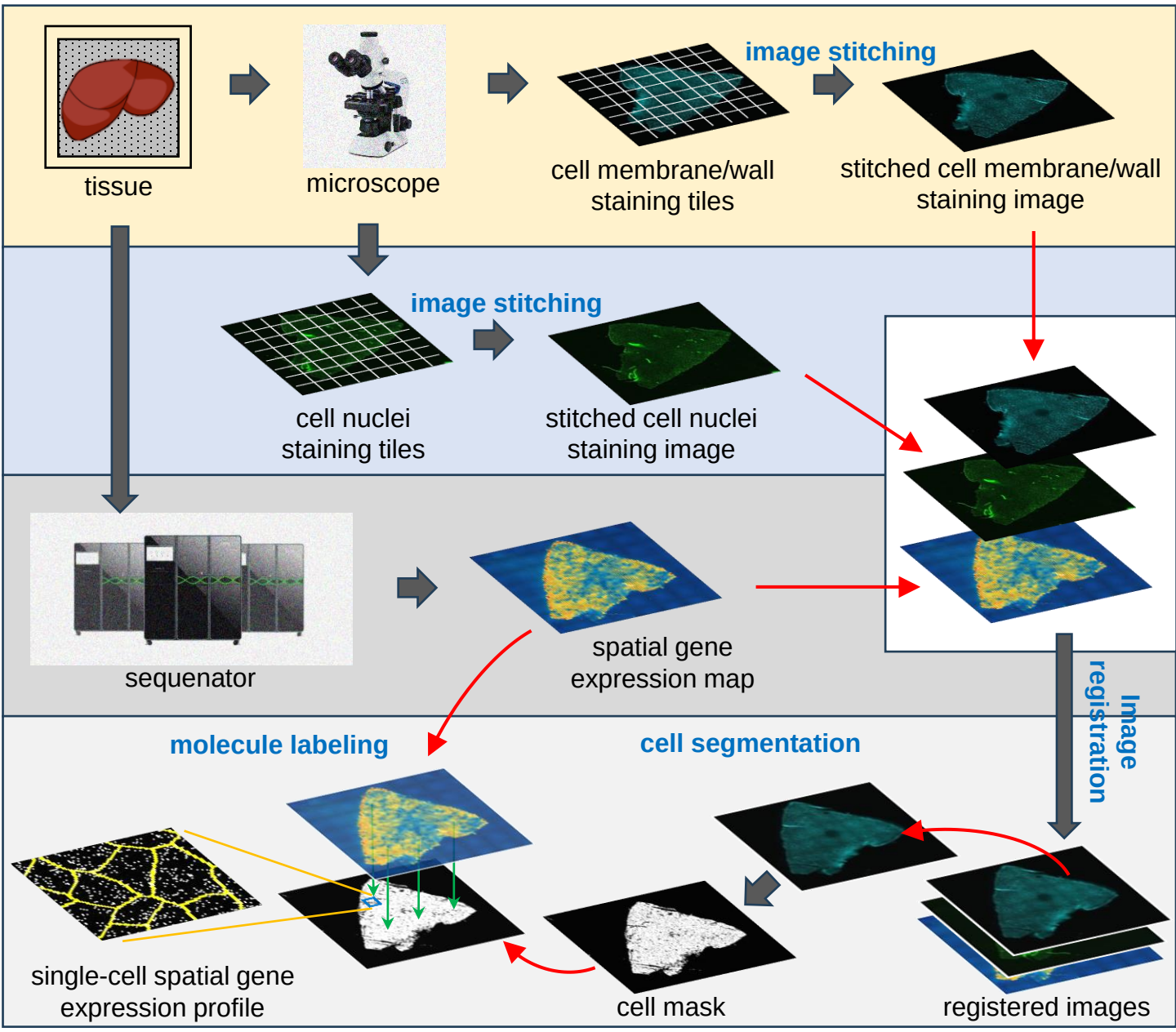

tissue
microscope
image stitching
cell membrane/wall staining tiles
stitched cell membrane/wall staining image
image stitching
cell nuclei staining tiles
stitched cell nuclei staining image
sequenator
spatial gene expression map
molecule labeling
cell segmentation
Image registration
single-cell spatial gene expression profile
cell mask
registered images

Fig. 2

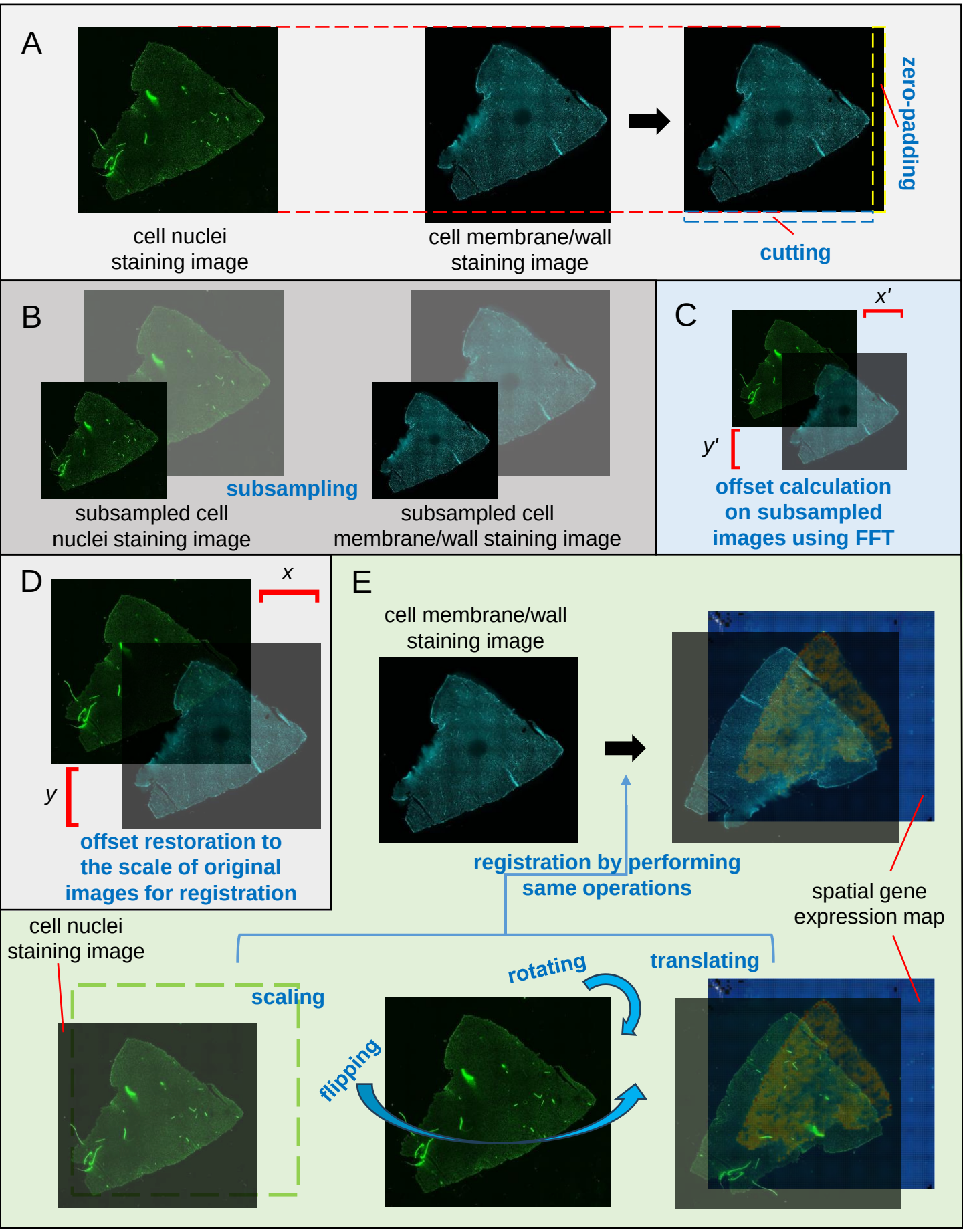
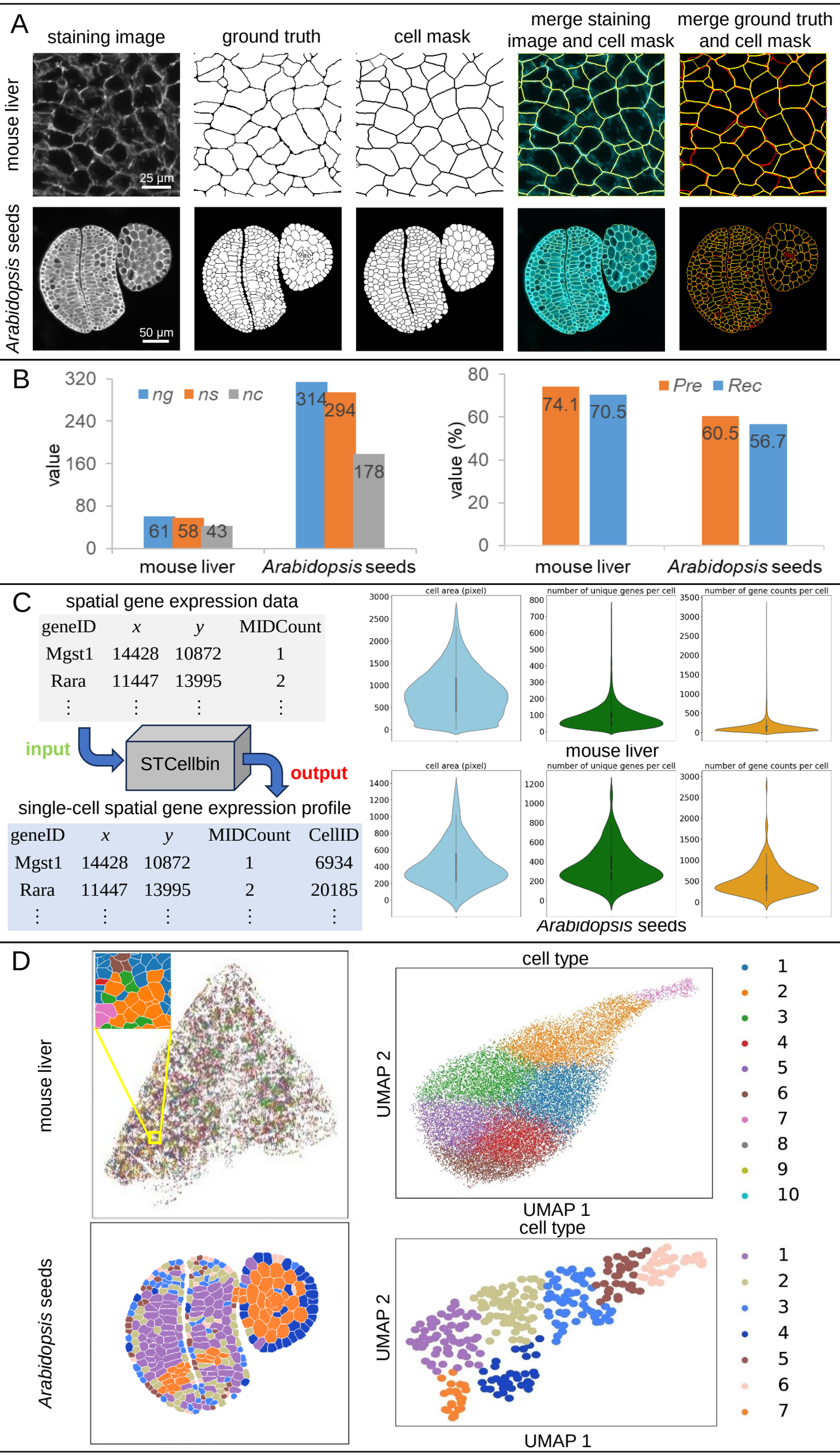

A
cell nuclei staining image
cell membrane/wall staining image
zero-padding
cutting
B
subsampling
subsampled cell nuclei staining image
subsampled cell membrane/wall staining image
C
x'
y'
offset calculation on subsampled images using FFT
D
x
y
offset restoration to the scale of original images for registration
E
cell membrane/wall staining image
registration by performing same operations
spatial gene expression map
cell nuclei staining image
scaling
rotating
translating
flipping

Fig. 3


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
61
58
43
314
294
178
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