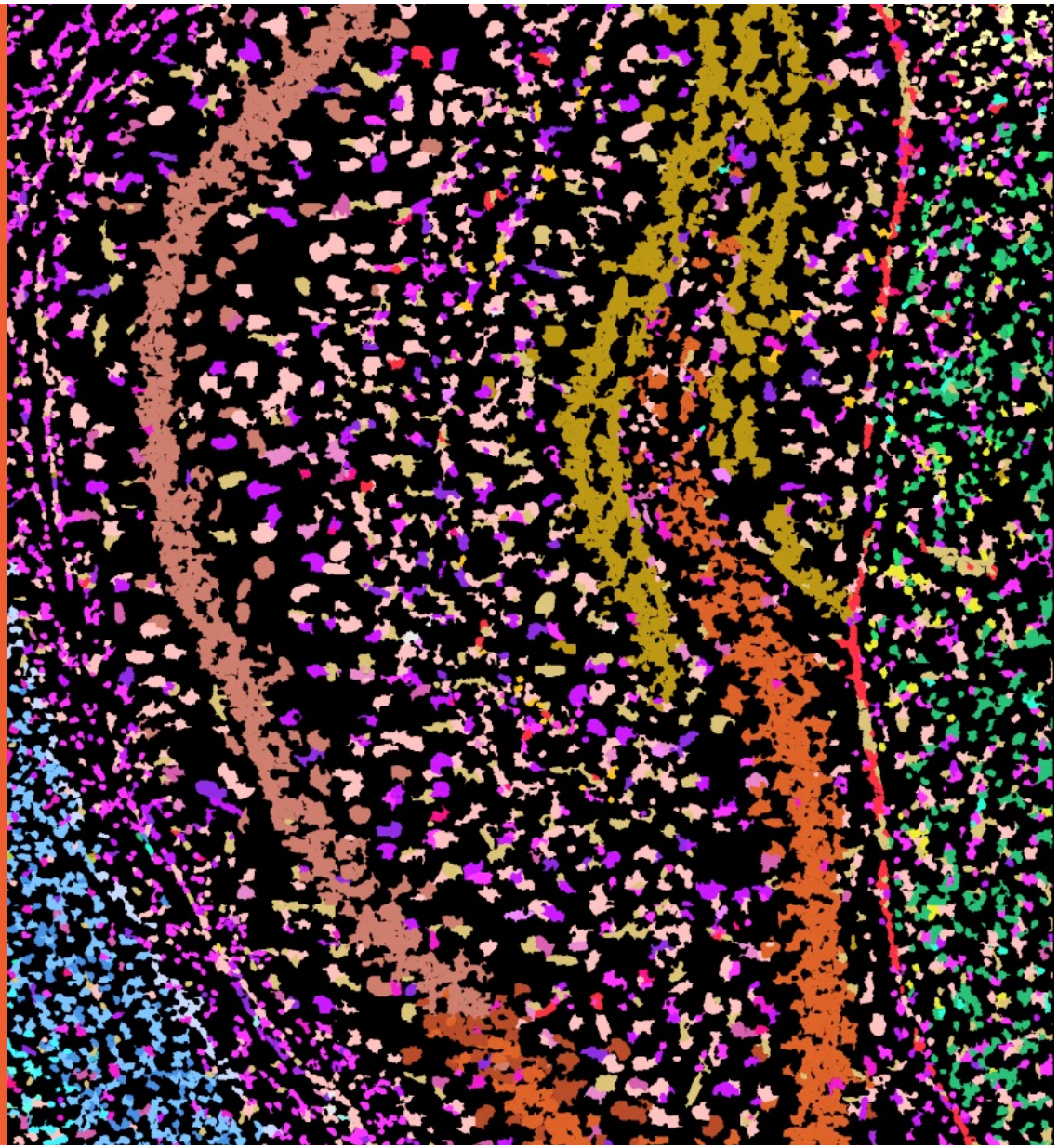
Biologically-informed self-supervised learning for segmentation of subcellular spatial transcriptomics data

Prof Jean Yee Hwa Yang

work lead by
Xiaohang (Helen) Fu and Yingxin Lin

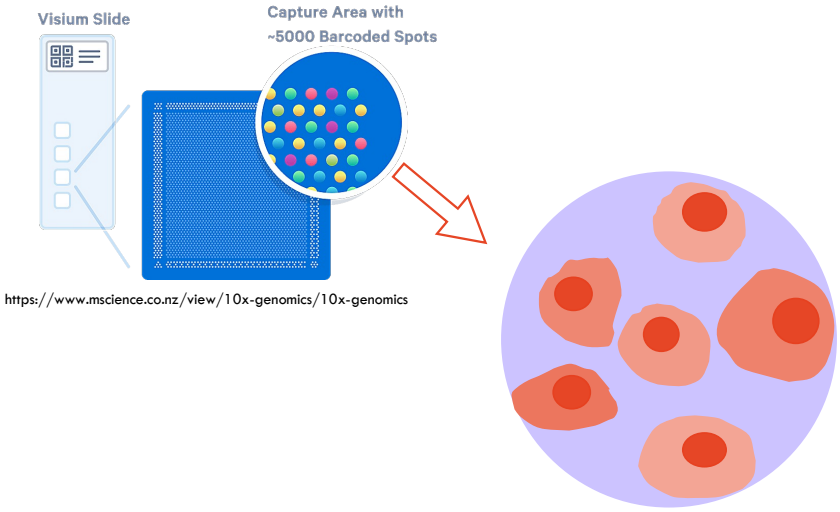
Sydney Precision Data Science Centre

THE UNIVERSITY OF SYDNEY



Spatially resolved transcriptomics

Nature Method of the Year 2020

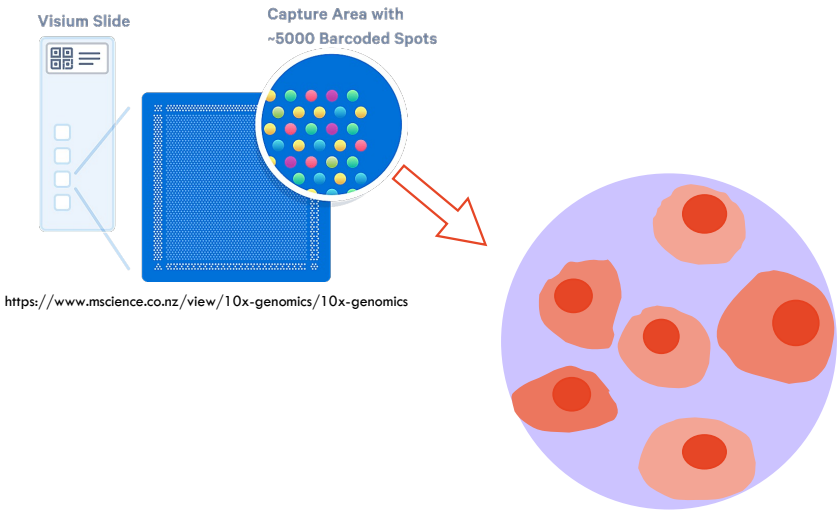


https://www.mscience.co.nz/view/10x-genomics/10x-genomics

1-10 cells in 1 spot

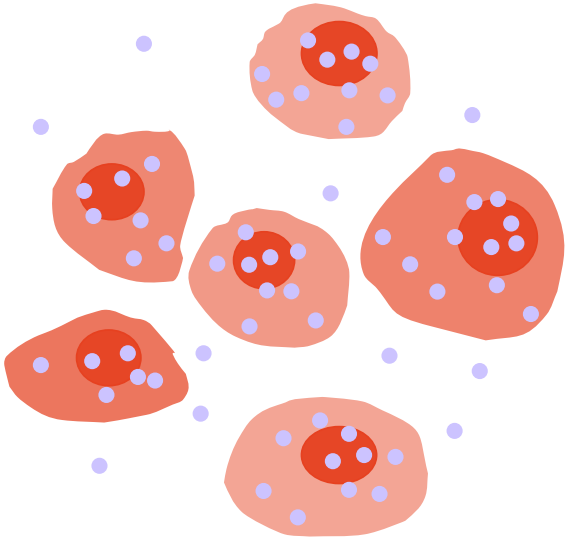
Spatially resolved transcriptomics

Nature Method of the Year 2020



1-10 cells in 1 spot

Subcellular spatially resolved transcriptomics



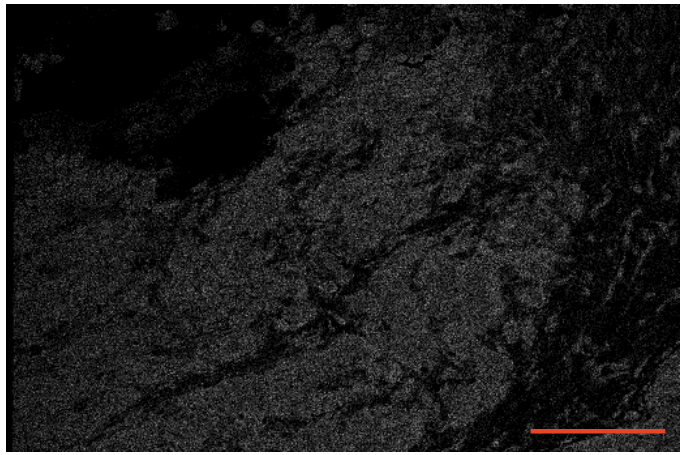
Subcellular detections

10x Genomics Xenium
NanoString CosMx
Vizgen MERSCOPE

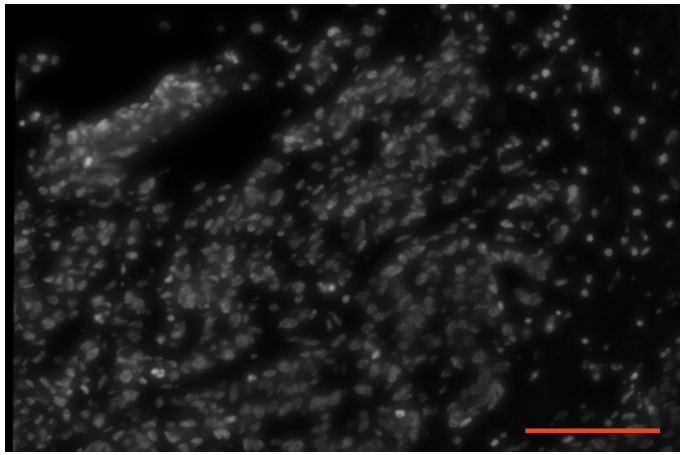
Subcellular spatially resolved transcriptomics (SST)

- Detect expression of hundreds of genes at subcellular resolution *in situ*
- Capture multi-channel **spatial transcriptomic maps** and **DAPI** images

Transcripts (Summed channels)



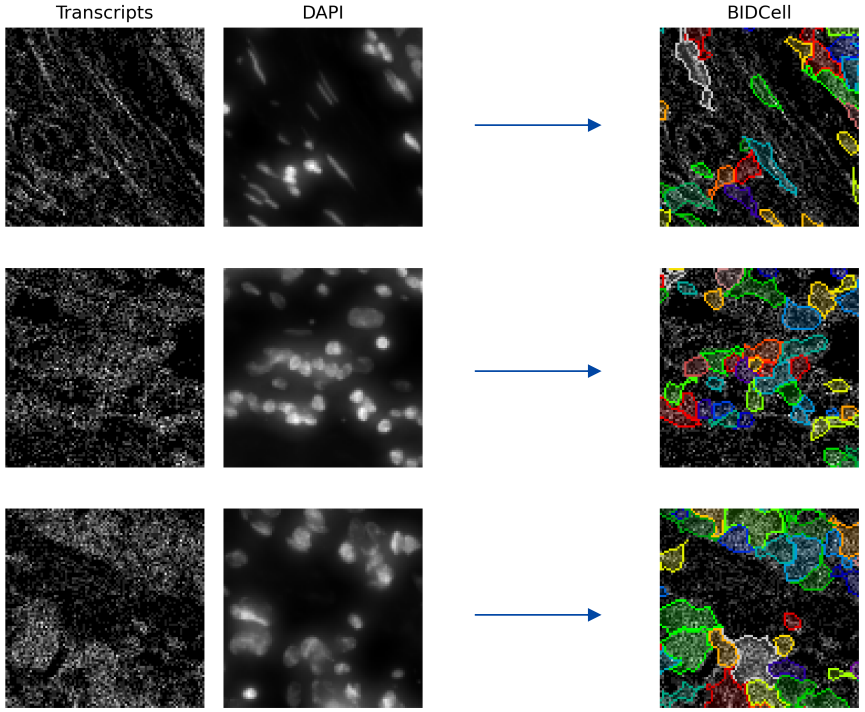
DAPI (Nuclei)



Scale: 100 µm
10x Genomics Xenium (breast cancer, 313 genes)

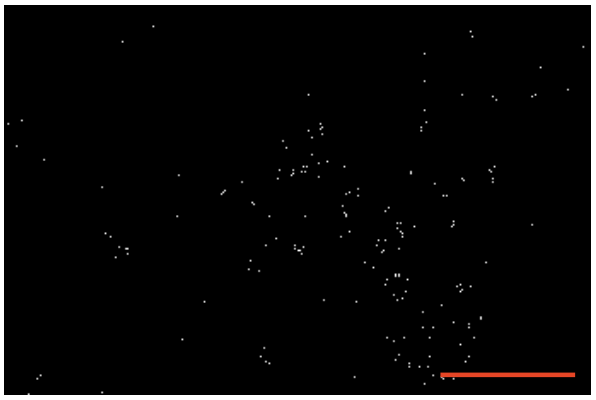
Cell segmentation

- Task: Identify all the pixels belonging to a cell in the form of a mask
- Do this for every unique cell in the image – i.e., instance segmentation
- The segmentation is used to quantify gene expression of each cell, by collecting the detected transcripts that falls inside the pixels belonging to a cell's mask

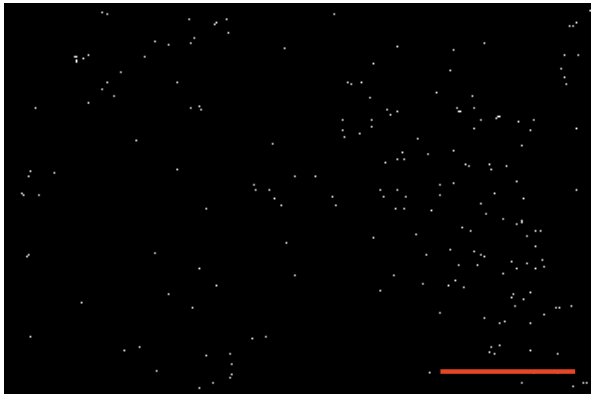


Challenges of SST images

- High dimensionality (hundreds of channels)
- High sparsity within each channel
- Lack of visual boundaries
- Densely-packed together cells
- No ground truth



CD3D



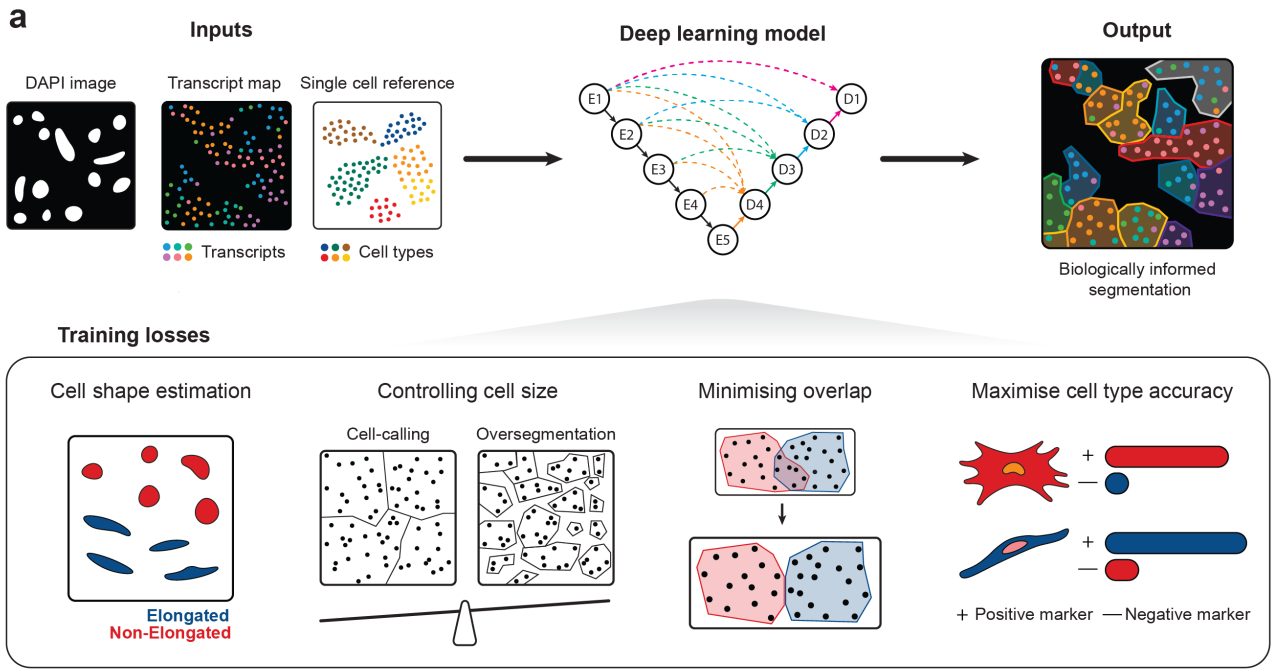
EGFR

Scale: 100 μm
10x Genomics Xenium (breast cancer, 313 genes)

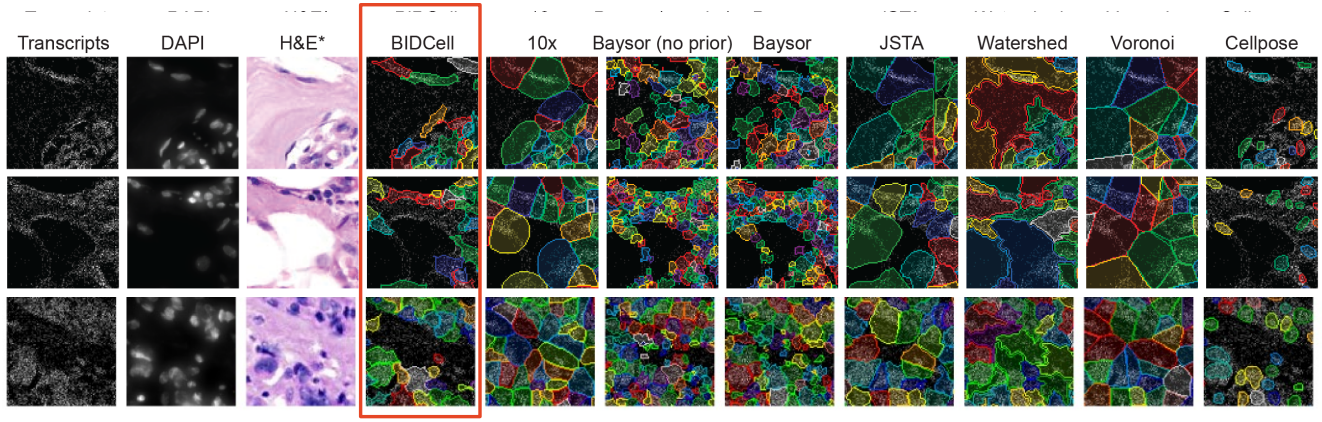
Existing methods

- **Classical**: watershed, Voronoi expansion, dilation from nuclei boundary
  - Spatial gene expression is irrelevant, tend to rely on nuclei/distance-based
- **Clustering/Transcript-based:** Baysor, ClusterMap
  - Disregard cell morphologies, sensitive to hyperparameters, slow, assume cells are homogeneous
- **Deep learning** (e.g., convolutional neural networks (CNNs)): Cellpose
  - SST images are considerably different to other modalities
    - Models pre-trained on other datasets are unsuitable
  - How to learn? There are no cell annotations, and manual annotation is impractical

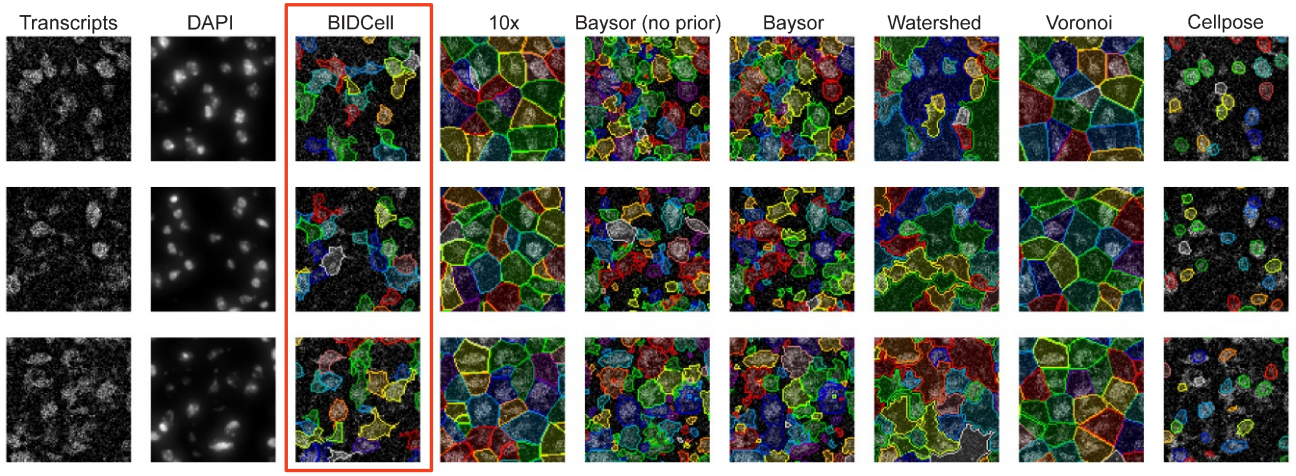
Biologically-informed deep learning-based cell segmentation (BIDCell)



Results



(a) Xenium-BreastCancer1



(b) Xenium-MouseBrain

Cell Segmentation Performance Assessment (CellSPA)

Baseline characteristics

Overall characteristics
cells;
% cells overlapped with nuclei;
% transcript assigned

Cell-level QC metrics
Density; Cell area;
# of genes expressed
# of total transcripts

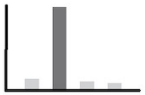
Cell morphology metrics
Elongation; Convexity; ...

Gene-level QC metrics
% of cells expressed

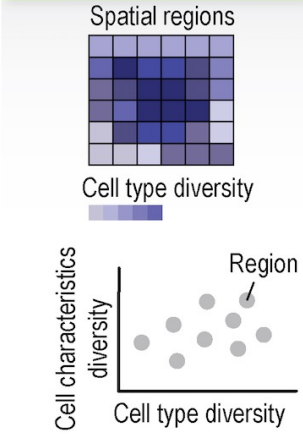
Segmented cell expression purity

Expression concordance with scRNA-seq

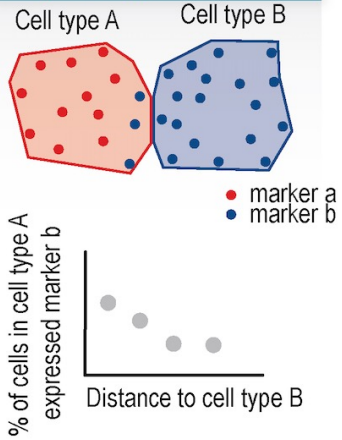
Expression purity in positive/negative markers



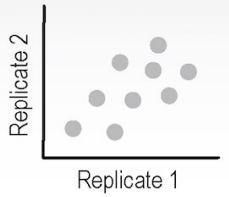
Spatial characteristics diversity



Neighbouring contamination



Replicability



Cell type proportion
Cell-level QC metrics
Gene-level QC metrics
...

Cell Segmentation Performance Assessment (CellSPA)

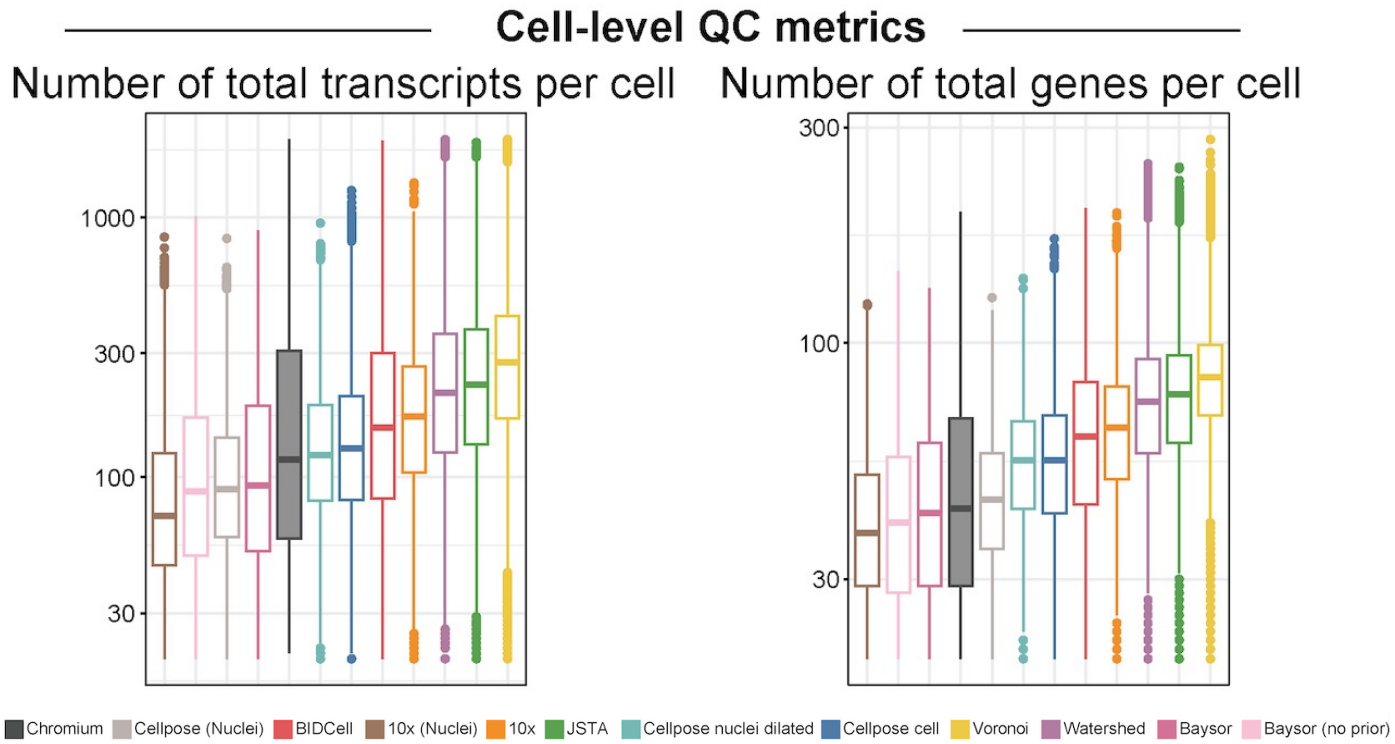
Baseline characteristics

Overall characteristics
cells;
% cells overlapped with nuclei;
% transcript assigned

Cell-level QC metrics
Density; Cell area;
# of genes expressed
# of total transcripts

Cell morphology metrics
Elongation; Convexity; ...

Gene-level QC metrics
% of cells expressed



Cell Segmentation Performance Assessment (CellSPA)

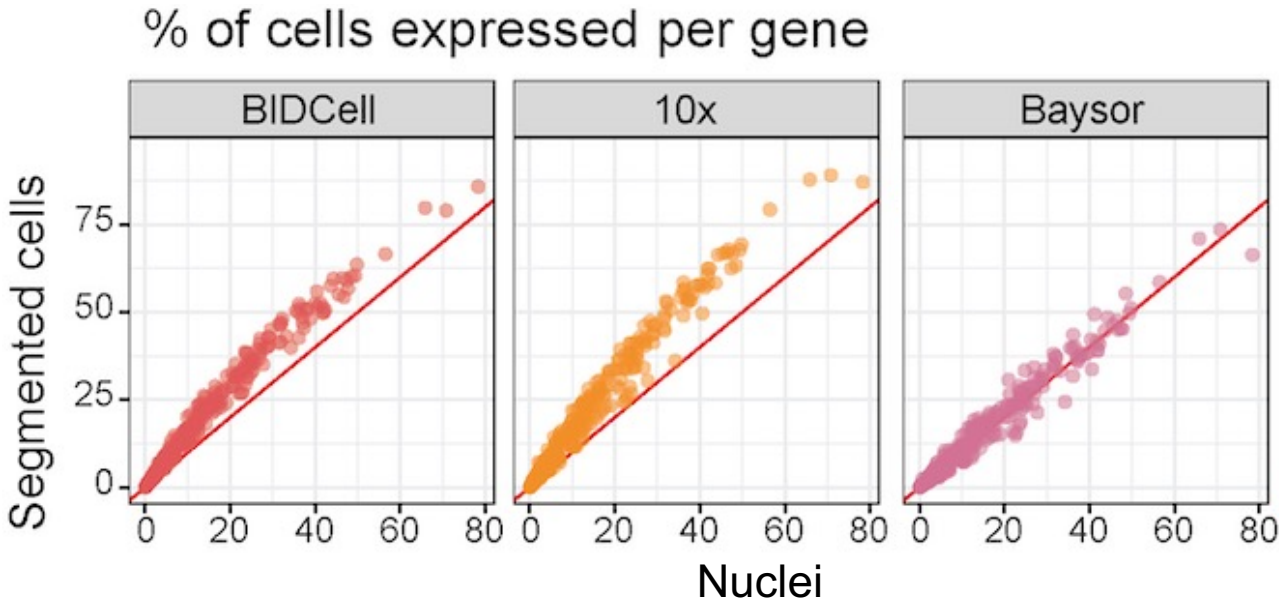
Baseline characteristics

Overall characteristics
cells;
% cells overlapped with nuclei;
% transcript assigned

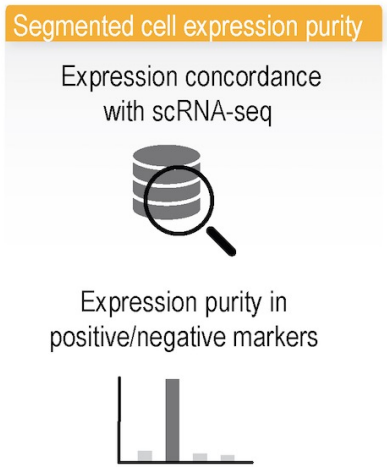
Cell-level QC metrics
Density; Cell area;
# of genes expressed
# of total transcripts

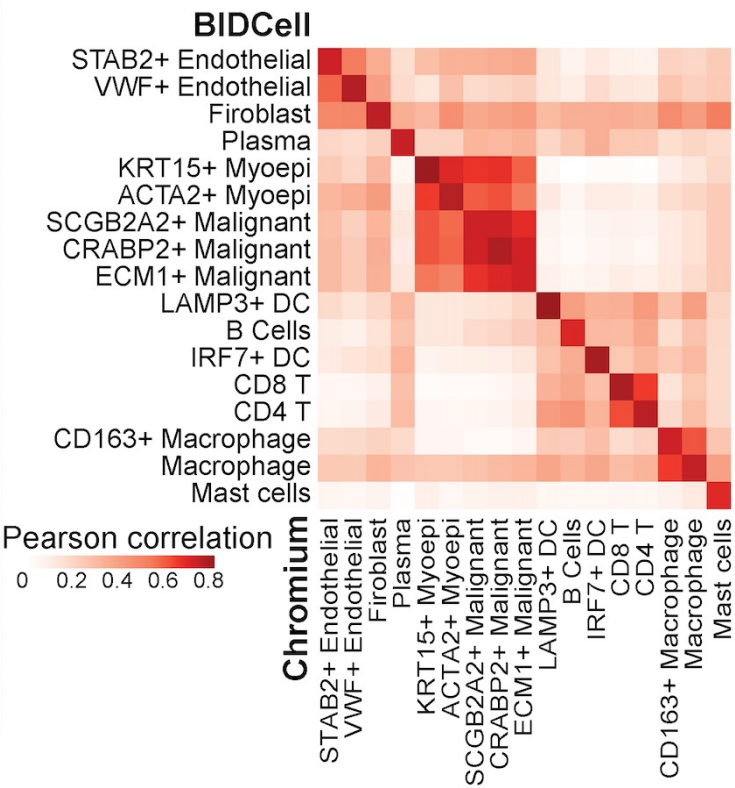
Cell morphology metrics
Elongation; Convexity; ...

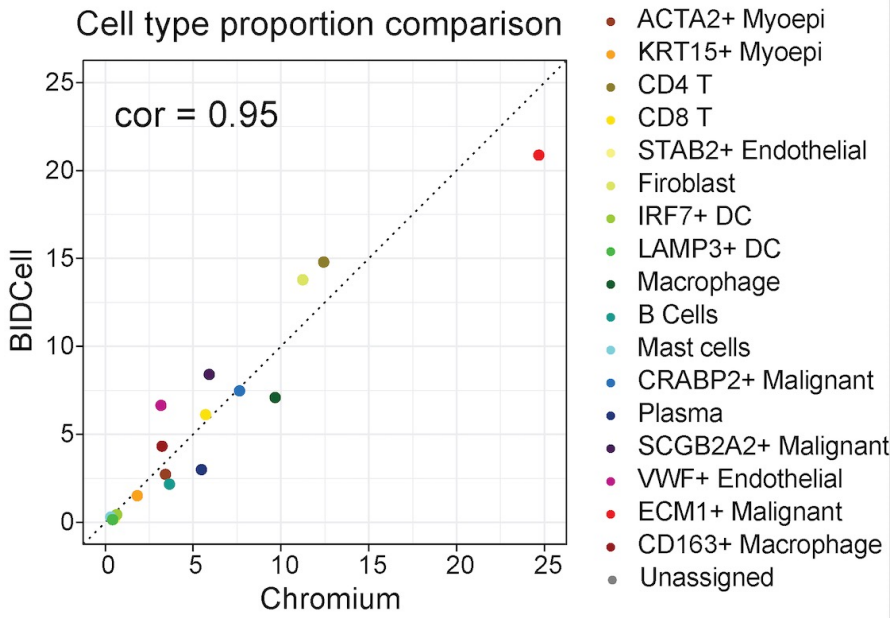
Gene-level QC metrics
% of cells expressed



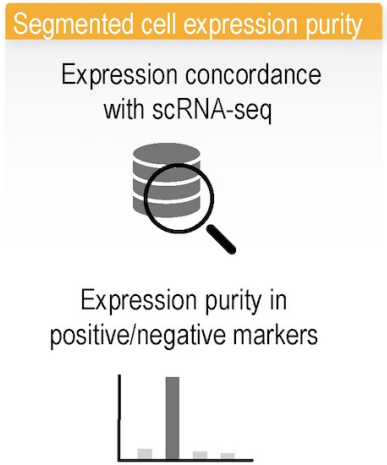
Cell Segmentation Performance Assessment (CellSPA)

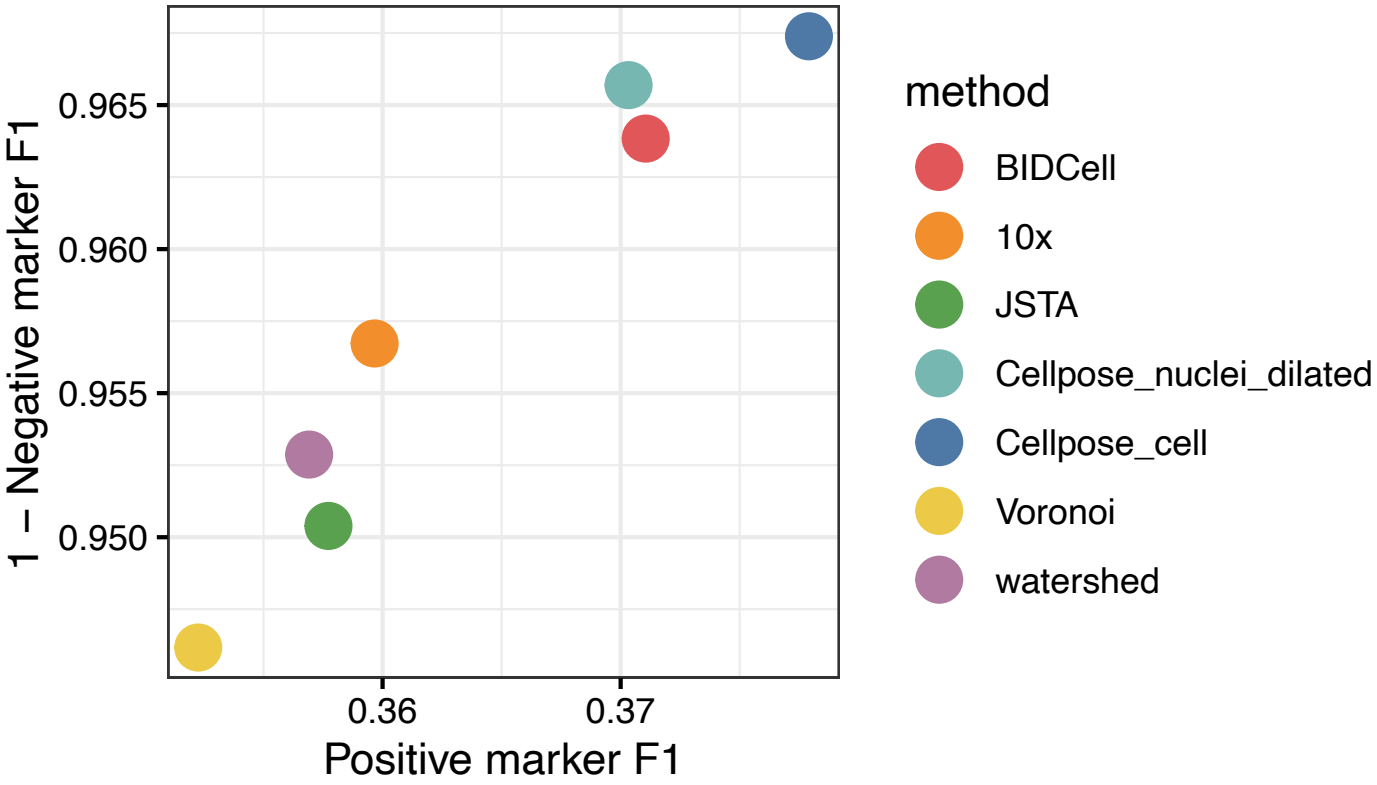




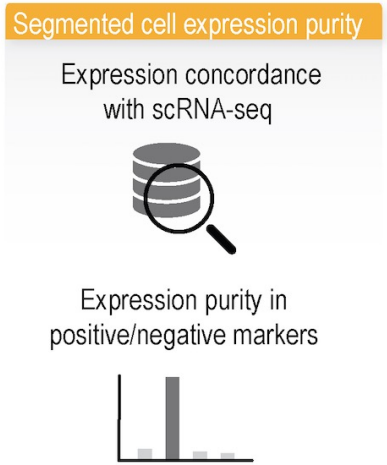


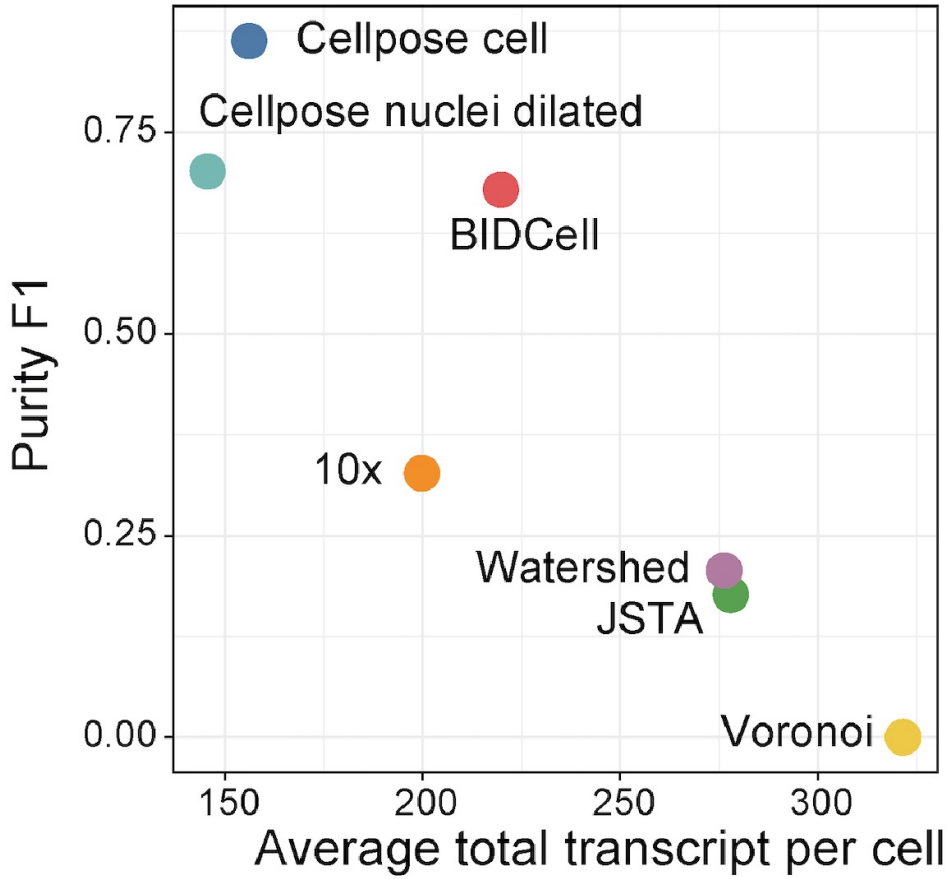
Evaluation – Trade-off between expression purity and size of cell



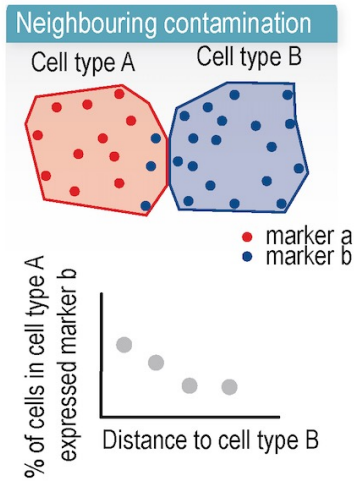


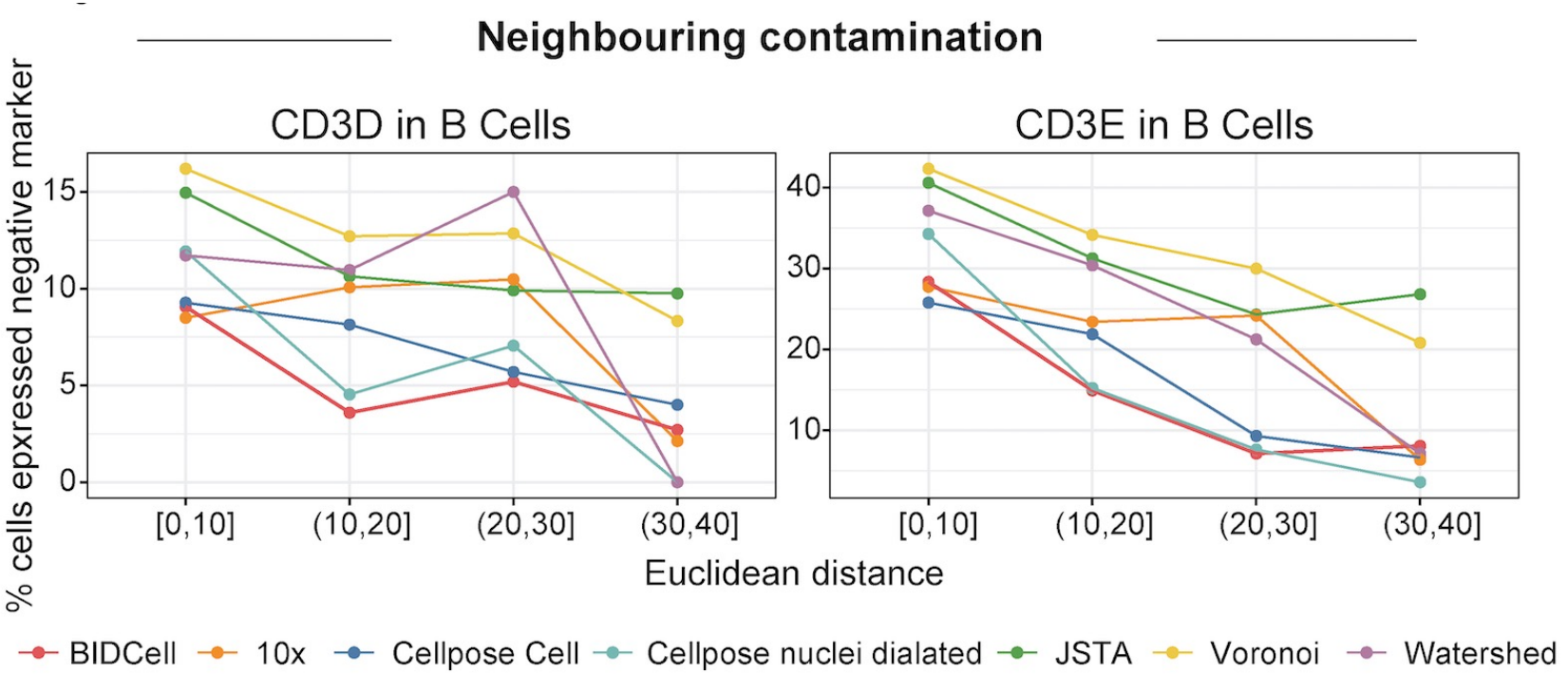
Cell Segmentation Performance Assessment (CellSPA)



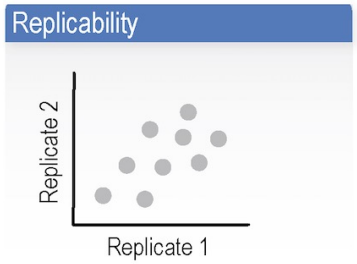


Cell Segmentation Performance Assessment (CellSPA)

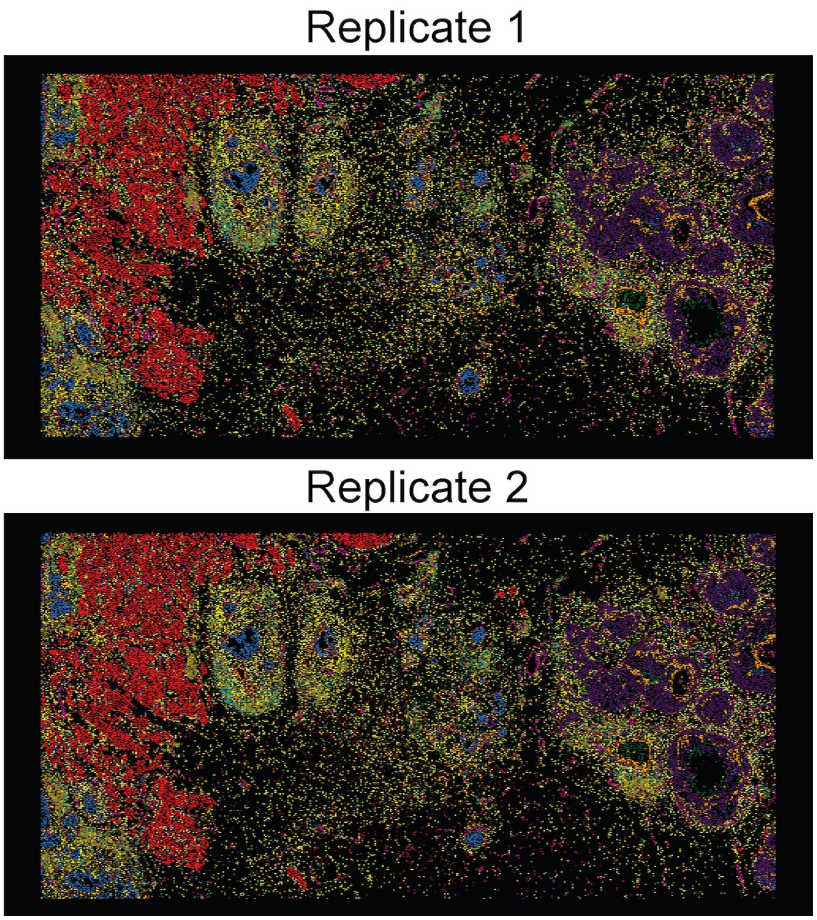


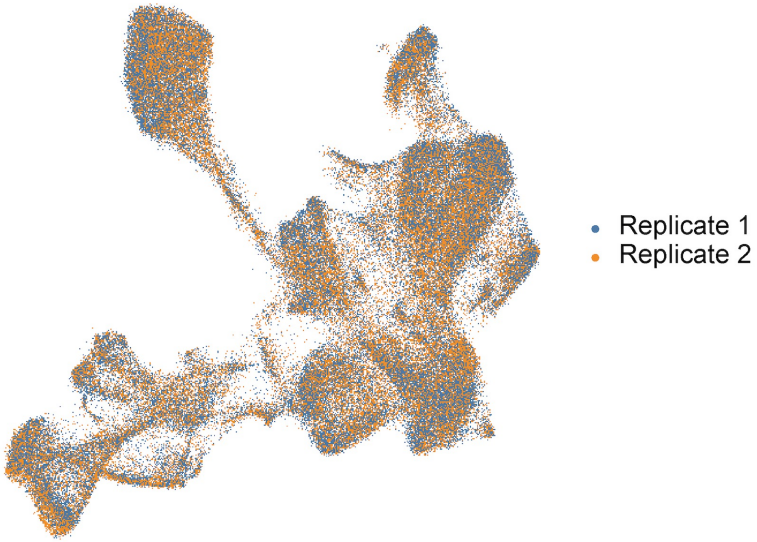


Cell Segmentation Performance Assessment (CellSPA)

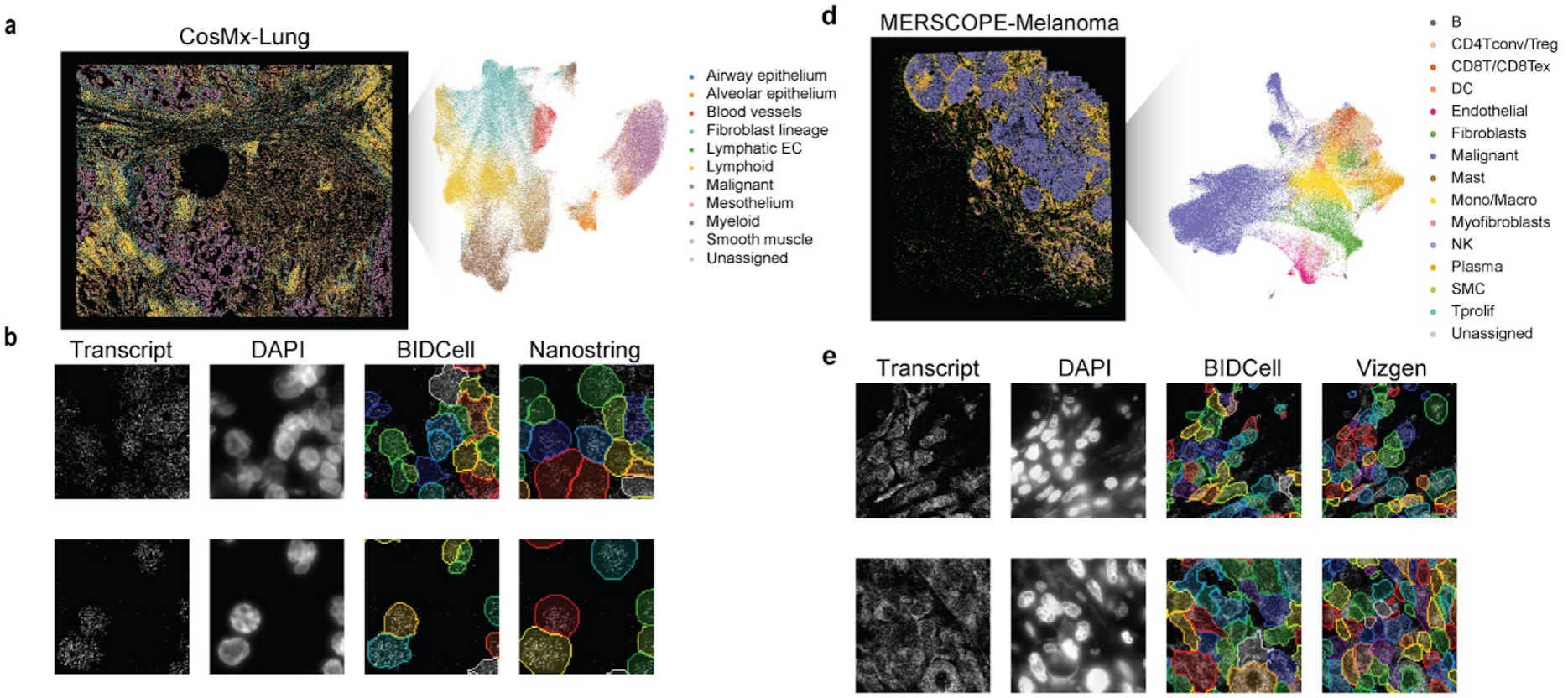


Cell type proportion
Cell-level QC metrics
Gene-level QC metrics
...



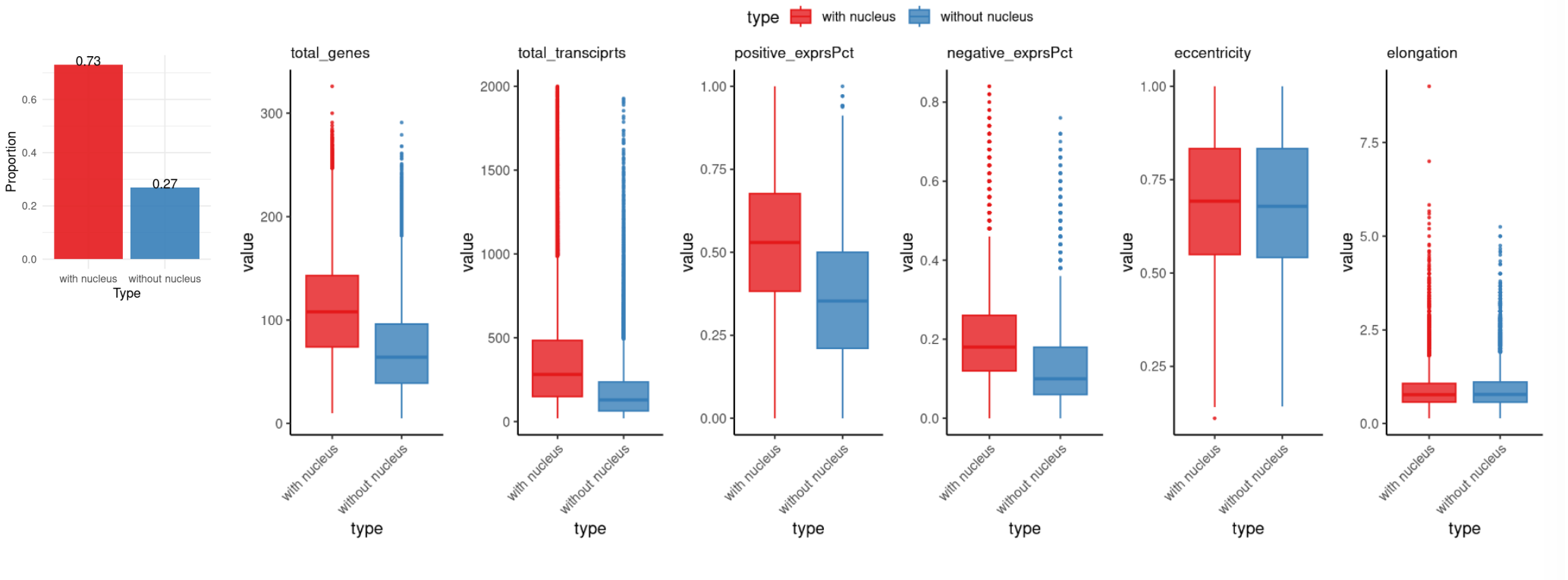


Flexibility

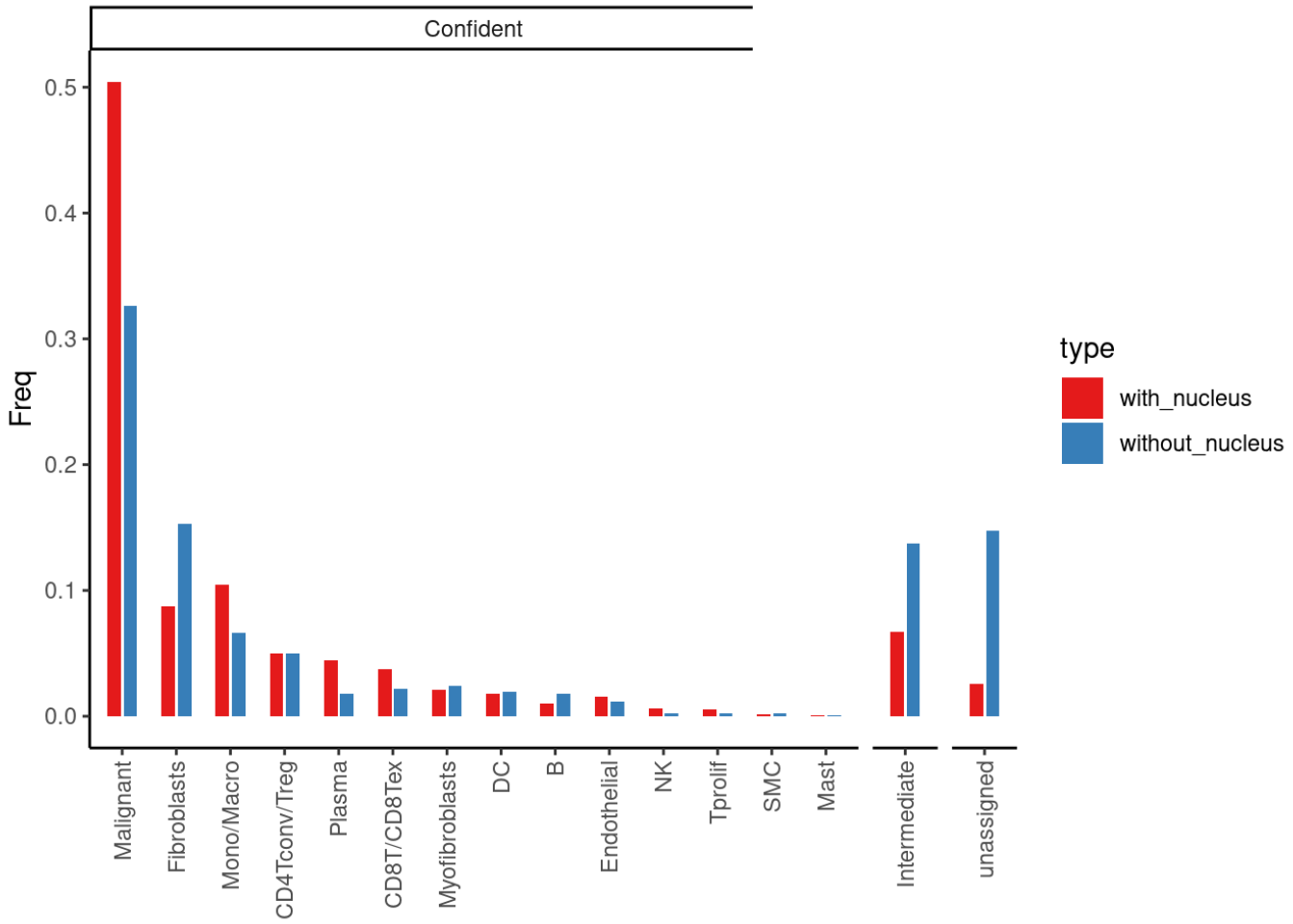


Worked in Stero-seq from BGI

BIDCell2 - What is the difference ?



BIDCell2 - What is the difference ?



Lin etal Molecular System Biology 2020

Data and Code

Manuscript:
https://www.biorxiv.org/content/10.1101/2023.06.13.544733v1

Code:
- BIDCell training and inference in https://github.com/SydneyBioX/BIDCell

- We provide our CellSPA framework in https://github.com/SydneyBioX/CellSPA.

Acknowledgements

University of Sydney

- **Yingxin Lin**
- Yue Cao
- **Helen Fu**
- Dario Strbenac
- Andy Tran
- Xiangnan Xu
- Yunwei Zhang
- Tian Lan

Ellis Patrick

- Elijah Willie
- Adam Chan

Rachel Wang

John Ormerod

- **Daniel Mechtersheimer**
- **Chuhan Wang**
- **Farhan Ameen**

All members of
Sydney Precision Data
Science Centre

Jean Yang

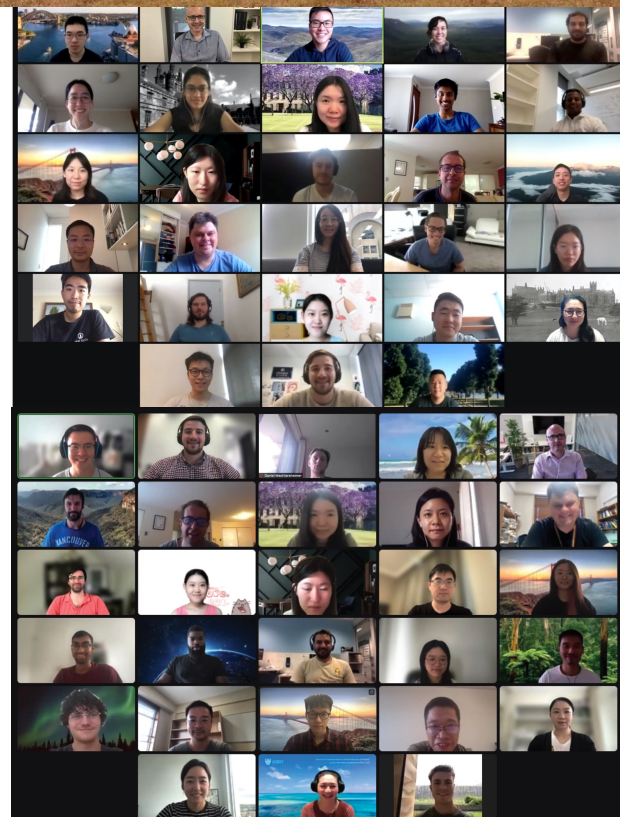
Jinman Kim (Computer Science)

Shila Ghazanfar

David Lin (Cornell)

Pengyi Yang

- Hani Kim
- Taiyun Kim
- Lijia Yu
- Chunlei Liu
- Daniel Kim
- Carissa Chen



THE UNIVERSITY OF SYDNEY

Australian Government
National Health and Medical Research Council
NHMRC

Australian Government
Australian Research Council

D24H
Laboratory of Data Discovery for Health
醫衛大數據深析實驗室 ®