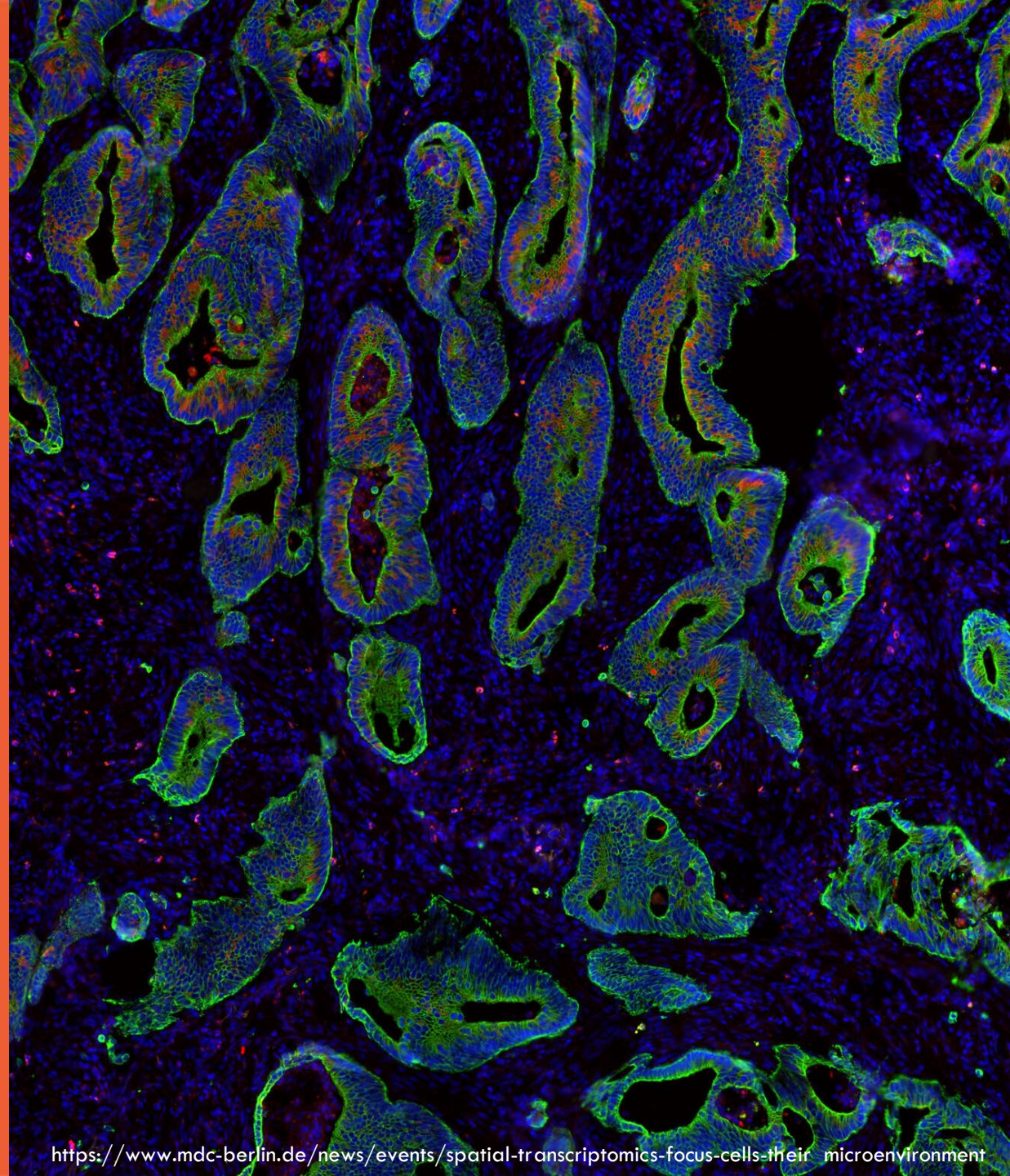
Unlocking single cell spatial omics analyses with SCDNEY

- Sydney Precision Data Science Centre
- Charles Perkins Centre
- School of Mathematics and Statistics
- Faculty of Medicine and Health

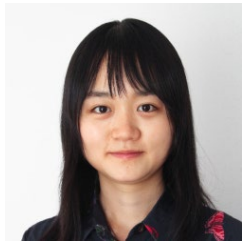
THE UNIVERSITY OF SYDNEY



https://www.mdc-berlin.de/news/events/spatial-transcriptomics-focus-cells-their_microenvironment

Sydney Precision Data Science Centre

Extracting insight from the data deluge



Dr Helen (Xiaohang) Fu
Xiaohang.fu@sydney.edu.au

Daniel Kim
Daniel.kim2@sydney.edu.au

Other Contributors:

- Professor Jean Yang
- Dr Dario Strbenac
- Mr Farhan Ameen
- Mr Alex Qin
- Dr Yue Cao
- Mr Nick Robertson
- Mr Andy Tran

Single-cell methods @ SCDNEY

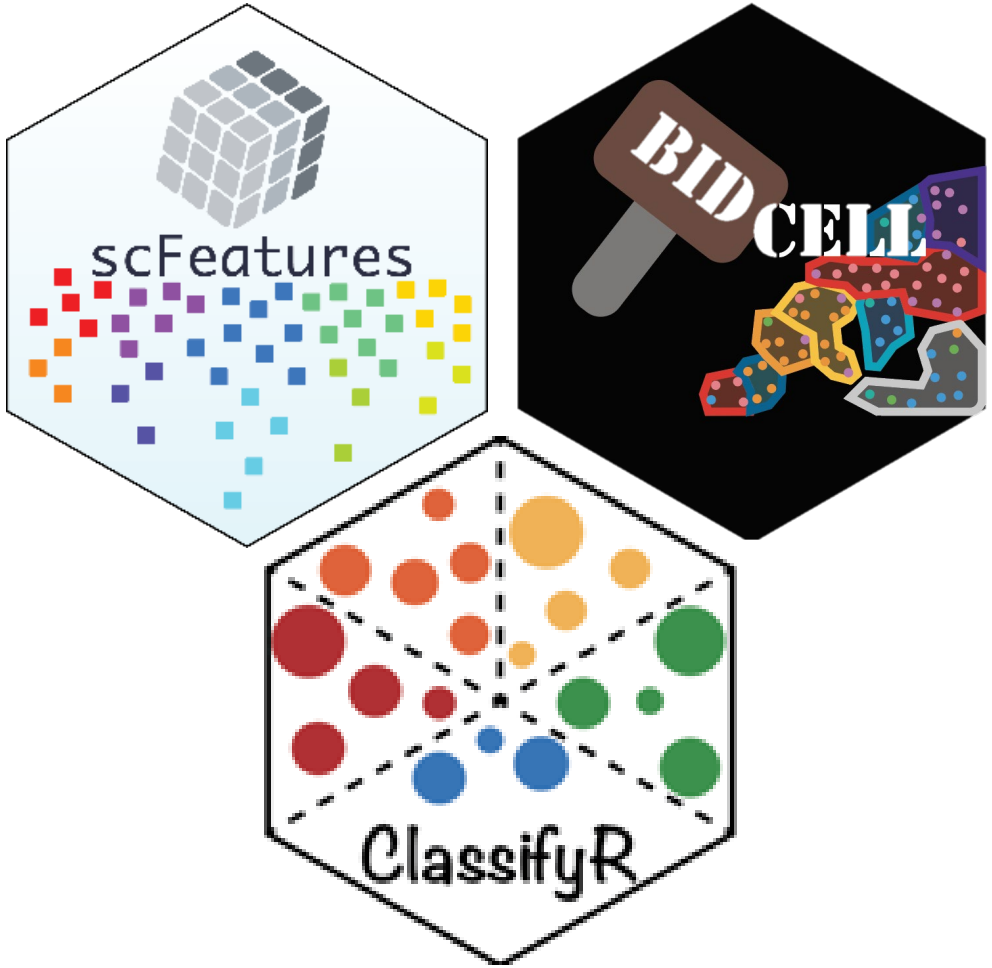
Single Cell Data iNtEgrative analYsis

We have a series of methodologies develop for single cell omics as well as single cell multi-omics data.

https://github.com/SydneyBioX/scdney

Schedule

- **Part I:** Introduction
- **Part II:** Exploring spatial data
- **Part III:** Feature engineering with scFeatures
- **15 min break**
- **Part IV:** Survival analysis with ClassifyR
- **Part V:** Identify cohort heterogeneity
- **Part VI:** Cell segmentation with BIDCell



Configuring Google Cloud

1. Obtain login details here: https://t.ly/eRxii

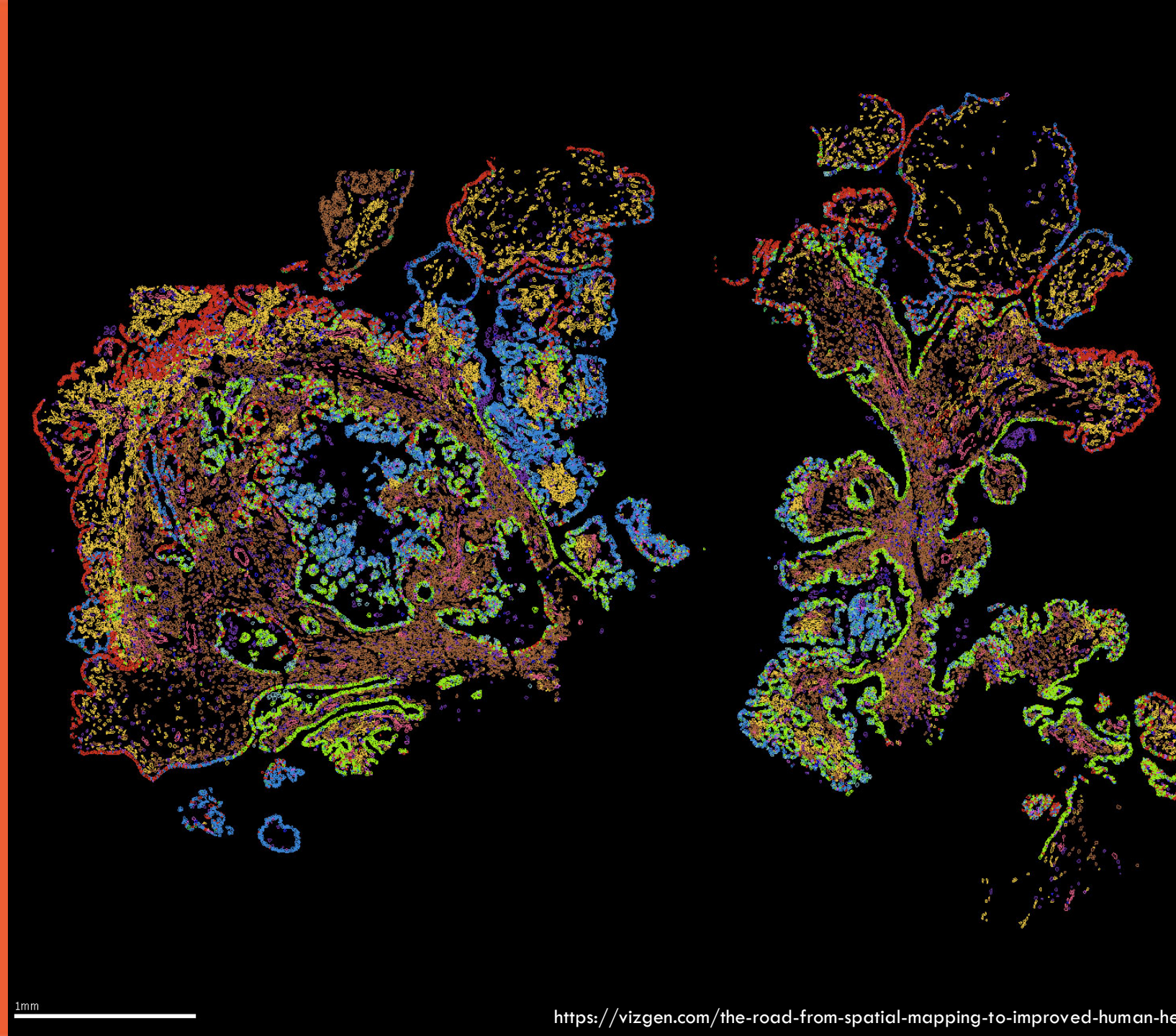
2. Type the machine IP into your browser to get into google cloud
3. Choose a username and password and log in, mark 'x' your selection
4. In RStudio, type the following:

   in console: `system(paste0("cp -r /home/gittmp/* ", getwd()))`

   in terminal: `unzip data.zip -d data`

Materials also on github: https://github.com/SydneyBioX/Cornell_SCDNEY_2024

PART I:
Introduction



https://vizgen.com/the-road-from-spatial-mapping-to-improved-human-he

Spatially resolved technologies

Bulk

Single Cell

Spatial

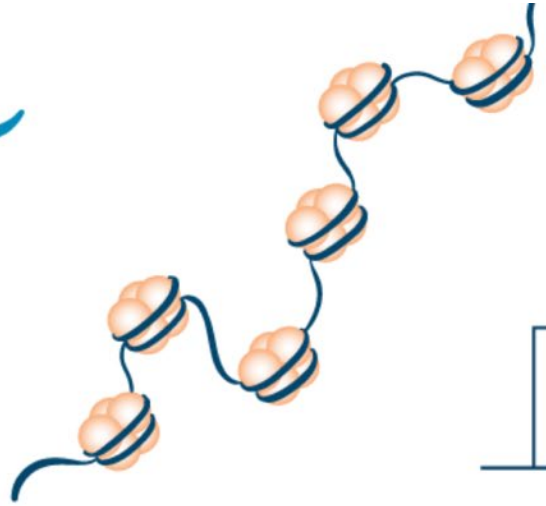
Credit: Bo Xia

Single-cell omics

Genome

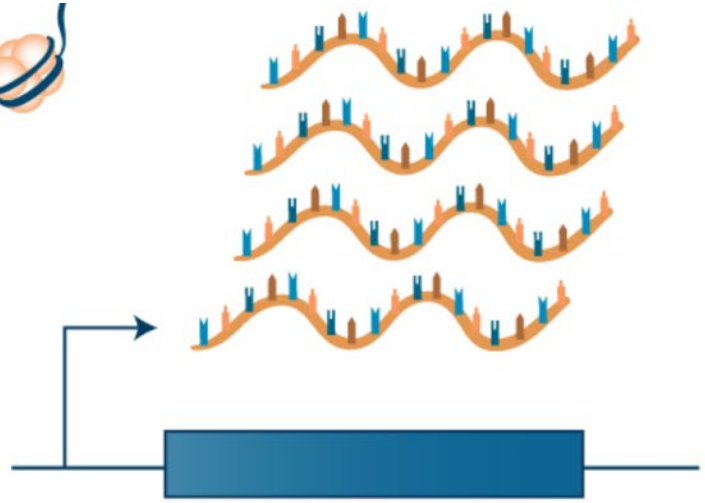
- SNS
- SCI-seq

Chromatin Accessibility



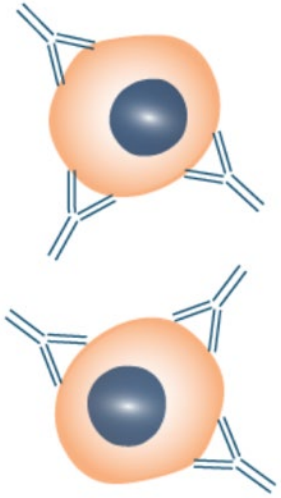
- scATAC-seq
- sciATAC-seq
- scTHS-seq
- 10x Genomics

Gene Expression



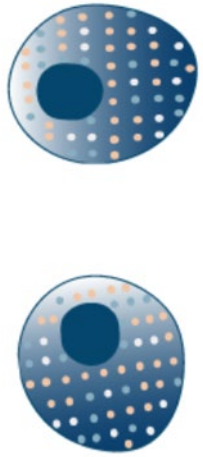
- scRNA-seq
  - Smart-seq2
  - CEL-seq
  - MARS-seq
  - 10x Genomics
  - Drop-seq
  - inDrops

Protein Abundance



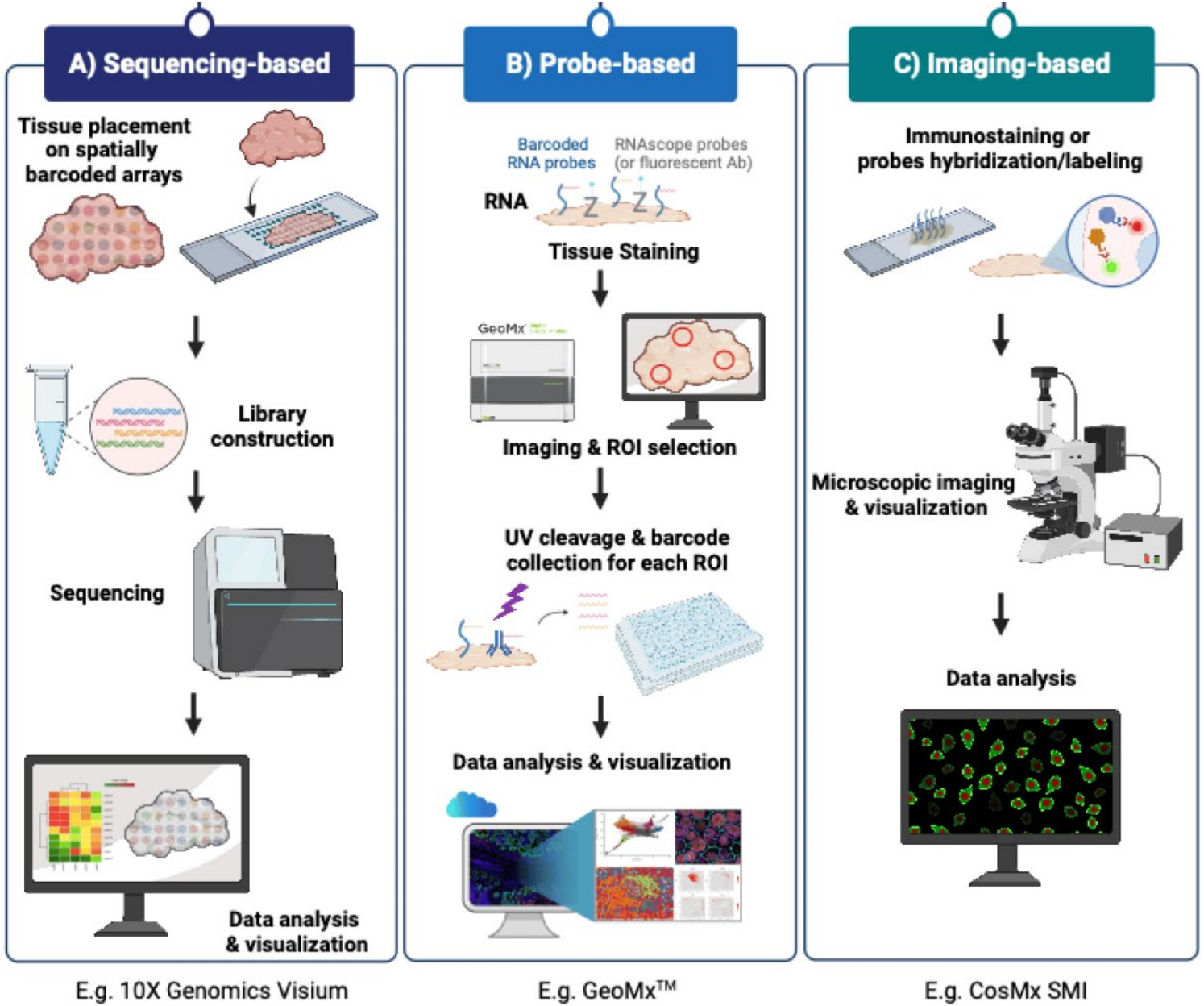
- CITE-seq
- REAP-seq
- Imaging Mass Cytometry (IMC)

Spatial Transcriptomics



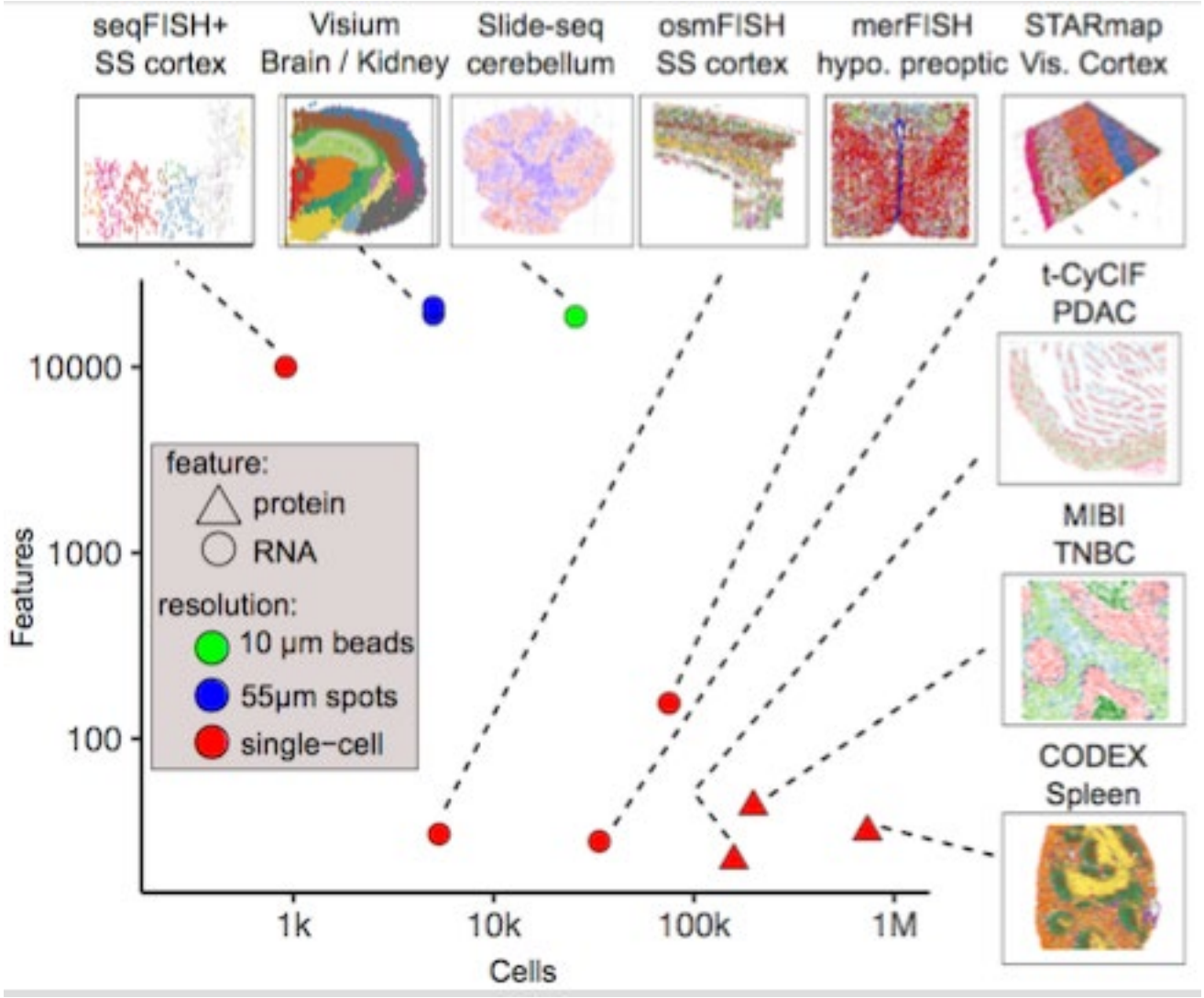
- MERFISH
- smFISH
- STARmap

Efremova, M., Teichmann, S.A. Computational methods for single-cell omics across modalities. *Nat Methods* **17**, 14–17 (2020).



Chen et al. Spatial Transcriptomic Technologies. *Cells* 2023, 12, 2042.

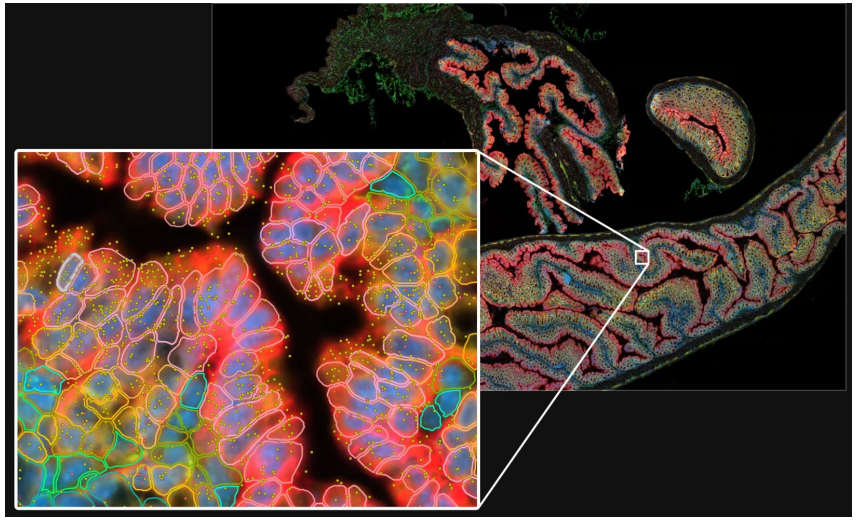
Different types of spatial technology



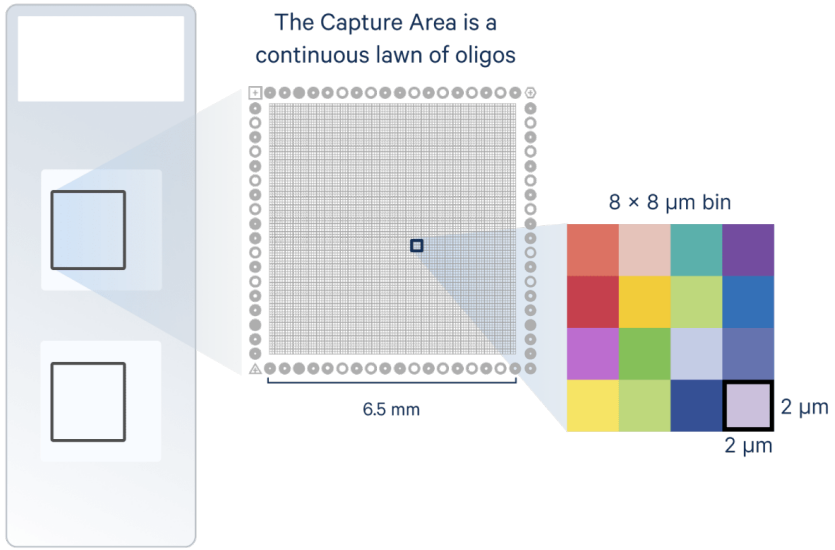
Popular platforms

- 10x Genomics
  - Xenium (< 5,000 genes)
  - Visium HD (whole transcriptome)
- Vizgen
  - MERSCOPE (< 1,000 genes)
- NanoString
  - CosMx (< 6,000 genes and 64 proteins)

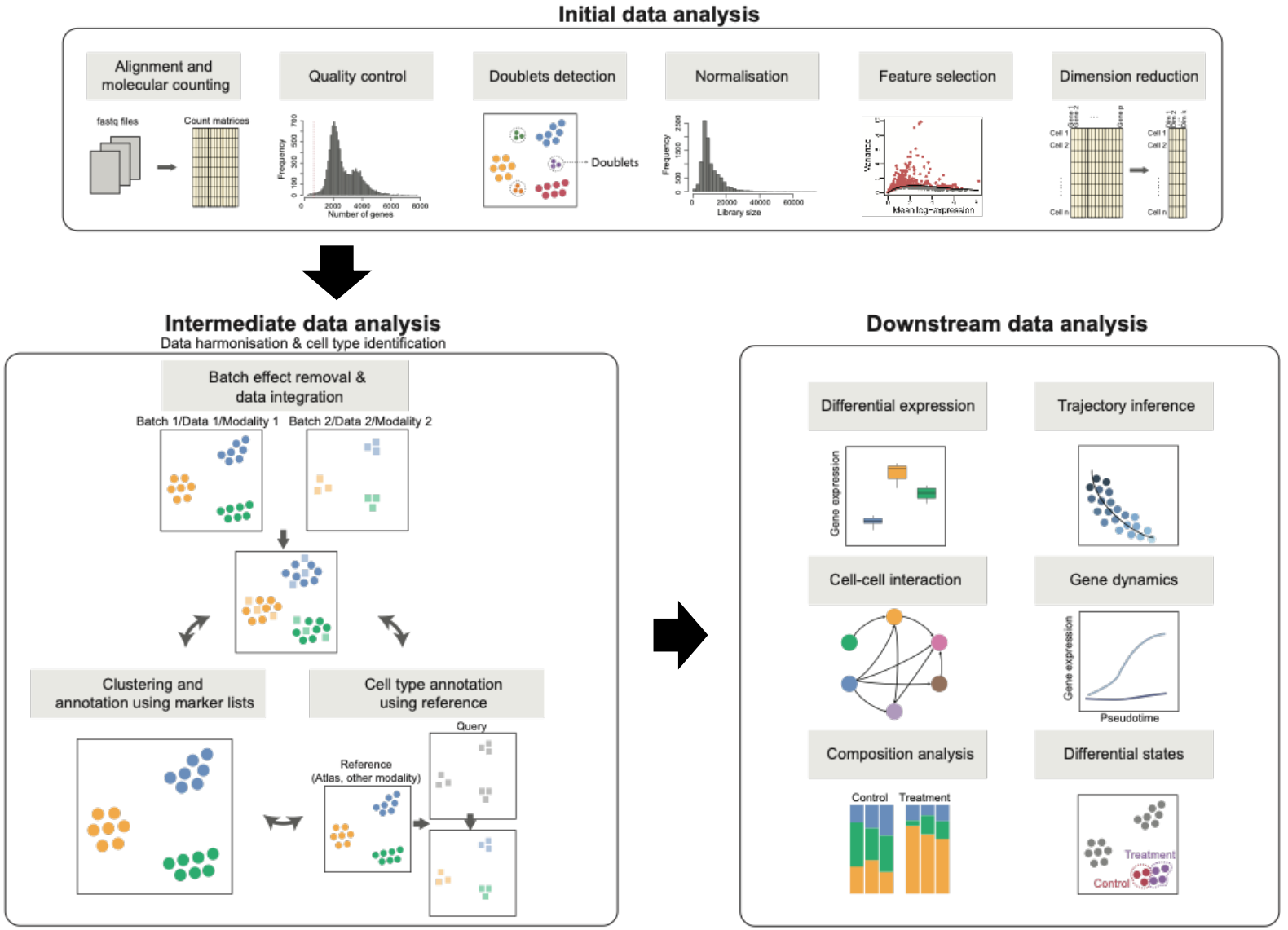
Xenium



Visium HD



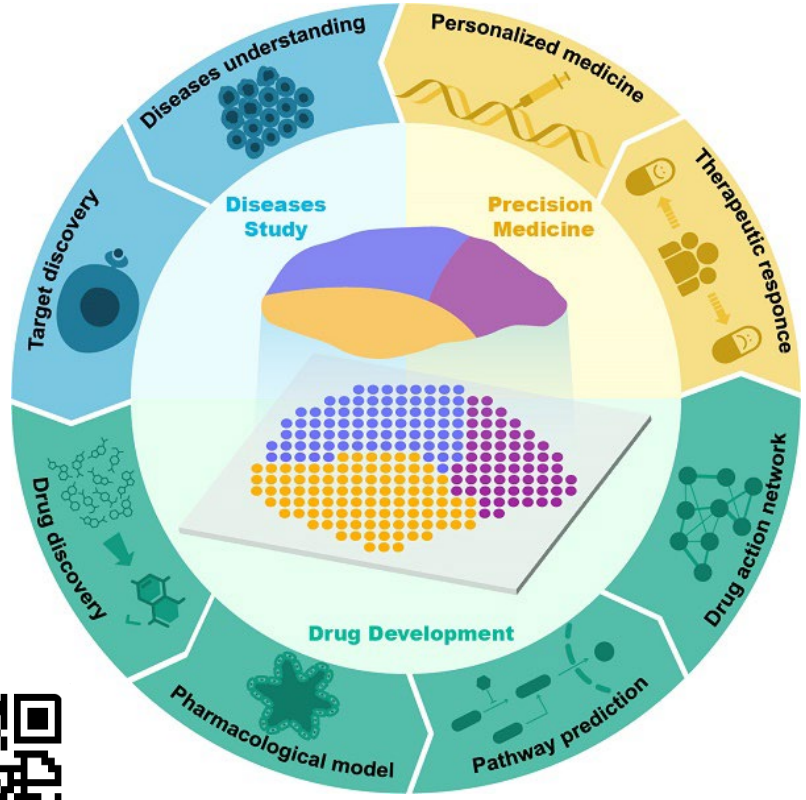
Overview of single-cell data analysis



Thanks to Yingxin Lin

Common questions asked in Spatial omics analysis

- Are there any spatially variable genes in my data?
- Are there any cell-types that co-localise?
- Which spatial features are associated with our outcome of interest or disease subtypes?
- **Can you think of any other questions?**



https://www.thno.org/v14p2946.htm

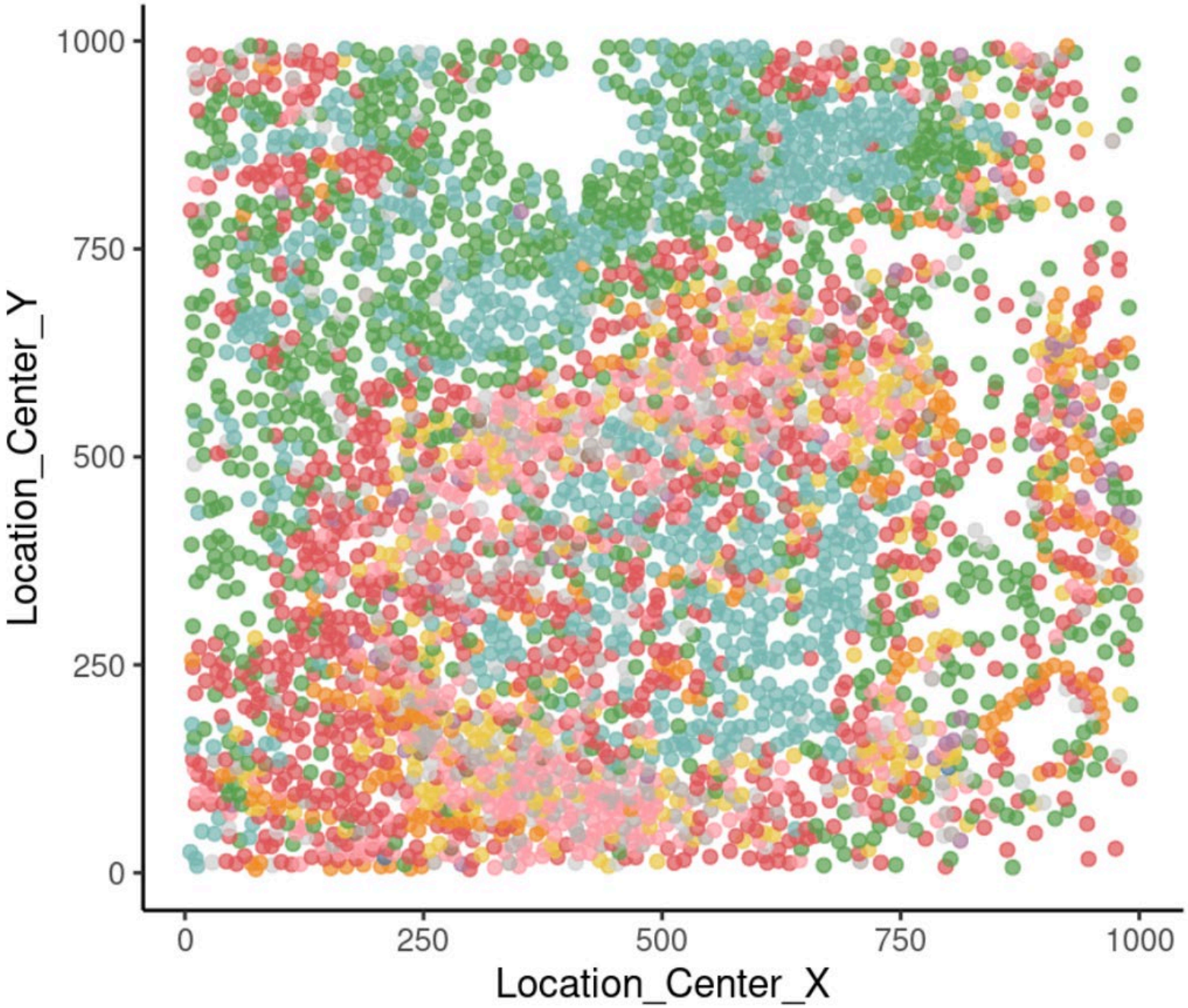
Questions?

THE UNIVERSITY OF SYDNEY

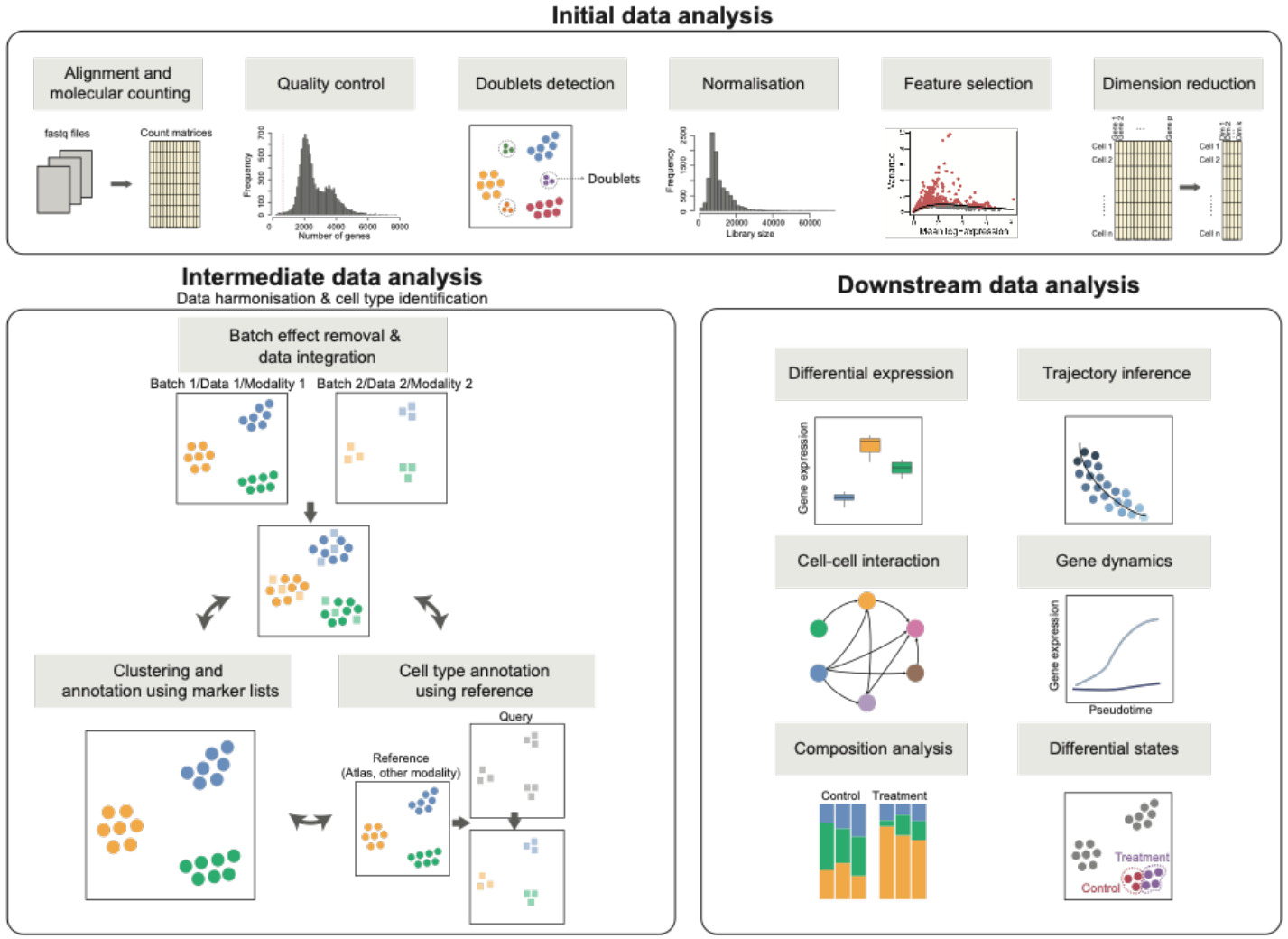
https://vizgen.com/the-road-from-spatial-mapping-to-improved-human-he

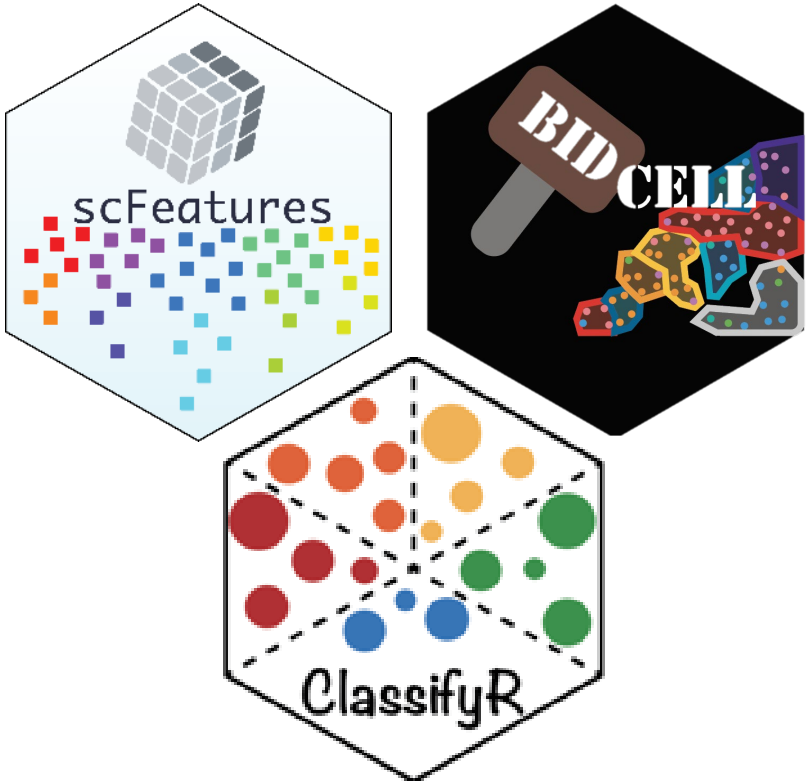
PART II:
Exploring spatial data

THE UNIVERSITY OF SYDNEY



Today...


Initial data analysis
Alignment and molecular counting
fastq files
Count matrices
Quality control
Frequency
Number of genes
Doublets detection
Doublets
Normalisation
Frequency
Library size
Feature selection
Variance
Mean log-expression
Dimension reduction
Cell 1
Cell 2
Cell n
Intermediate data analysis
Data harmonisation & cell type identification
Batch effect removal & data integration
Batch 1/Data 1/Modality 1
Batch 2/Data 2/Modality 2
Clustering and annotation using marker lists
Cell type annotation using reference
Query
Reference (Atlas, other modality)
Downstream data analysis
Differential expression
Gene expression
Trajectory inference
Cell-cell interaction
Gene dynamics
Gene expression
Pseudotime
Composition analysis
Control
Treatment
Differential states
Treatment
Control


scFeatures
BID CELL
ClassifyR

Data – IMC

nature cancer

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature cancer › articles › article

Article | Published: 17 February 2020

Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer

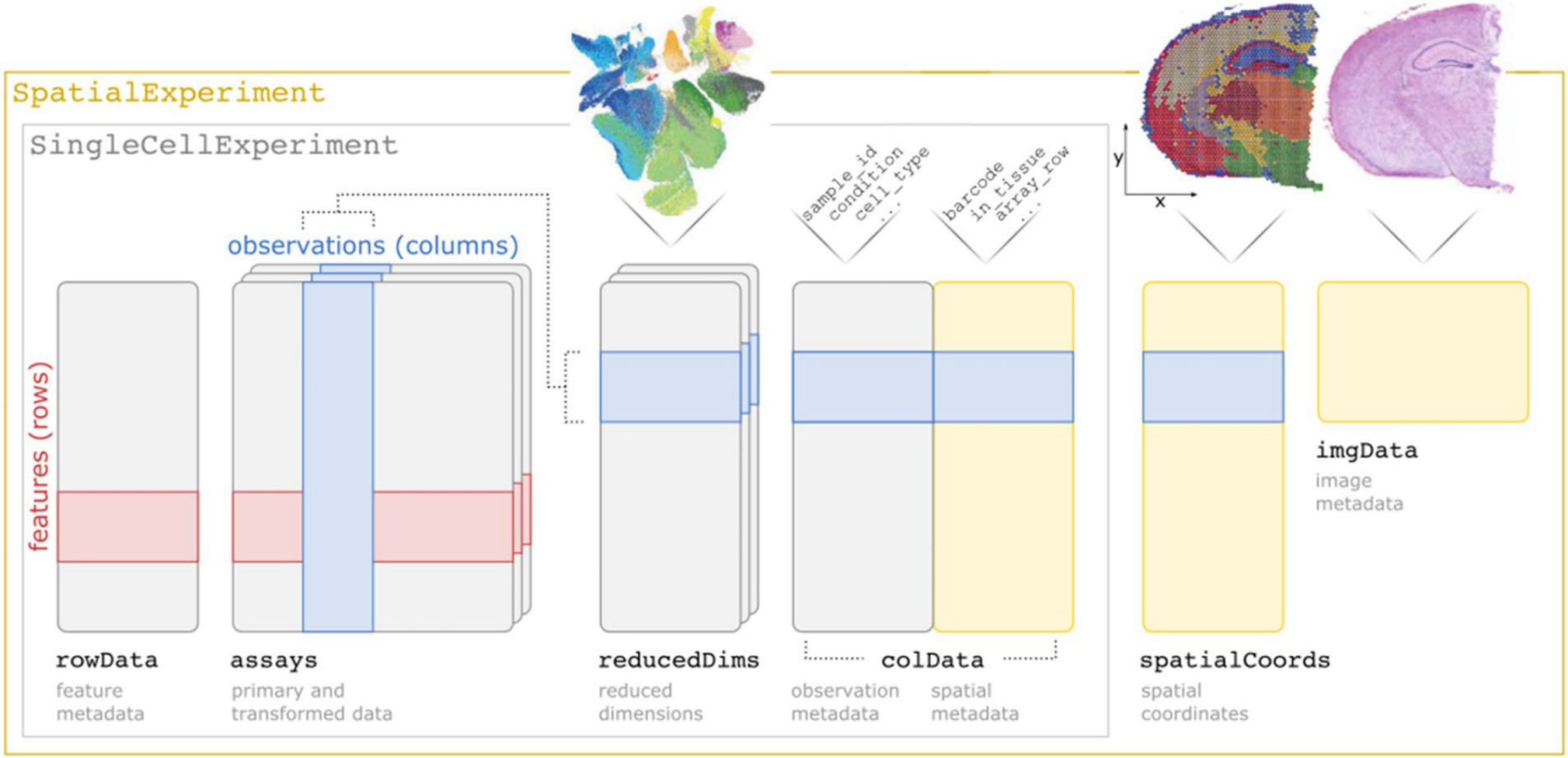
H. Raza Ali, Hartland W. Jackson, Vito R. T. Zanotelli, Esther Danenberg, Jana R. Fischer, Helen Bardwell, Elena Provenzano, CRUK IMAXT Grand Challenge Team, Oscar M. Rueda, Suet-Feung Chin, Samuel Aparicio, Carlos Caldas ✉ & Bernd Bodenmiller ✉

Nature Cancer **1**, 163–175 (2020) | Cite this article

30k Accesses | **132** Citations | **166** Altmetric | Metrics

- Single-cell resolution
- 37 proteins x 483 patients
- In this workshop: 77 patients with no lymph node metastasis
- Outcome of interest: Recurrence

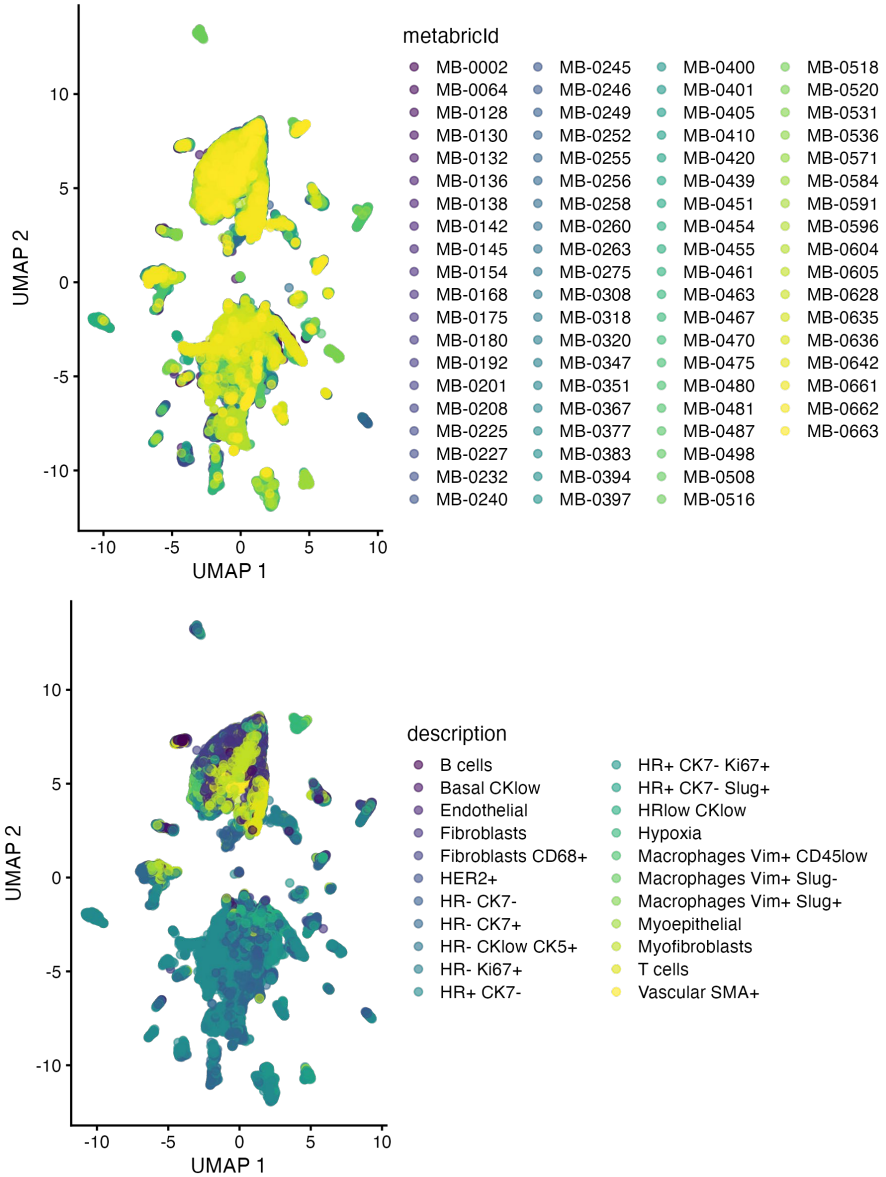
How to represent processed spatial omics data?



Righelli et al, Bioinformatics (2022)

Shila Ghazanfar E: shila.ghazanfar@sydney.edu.au

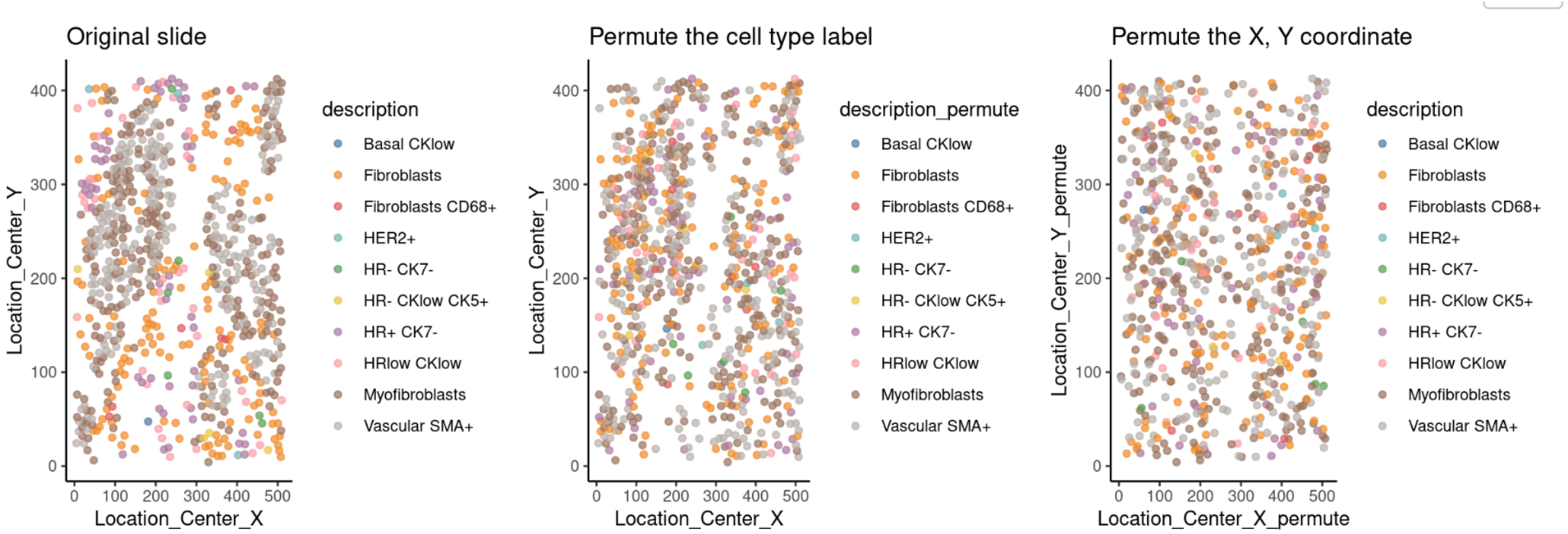
Visualisation



When there is a large number of categories, are dimensionality reduction plots interpretable or misleading due to overplotting?

Visualisation

We can examine whether the slides appear to be **structured** or **randomly distributed** by random permutation

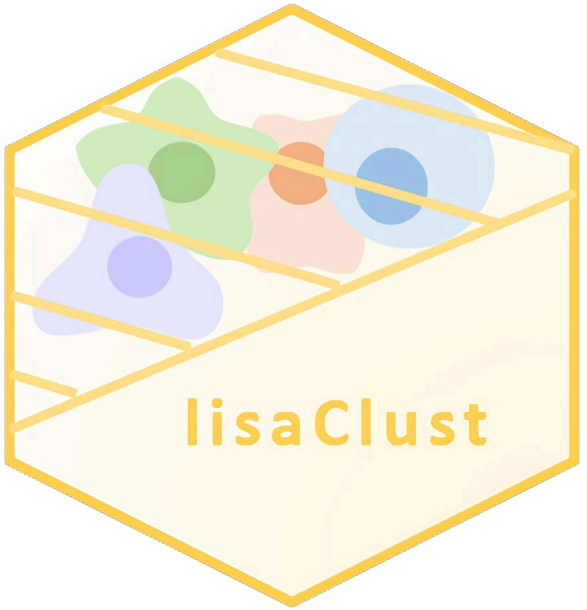


Describe tissue microenvironment and neighbourhood

- We can segment the slides into multiple unique **regions** or **patterns.**

- There are multiple methods to perform this task such as ClusterMap, BASS, lisaClust

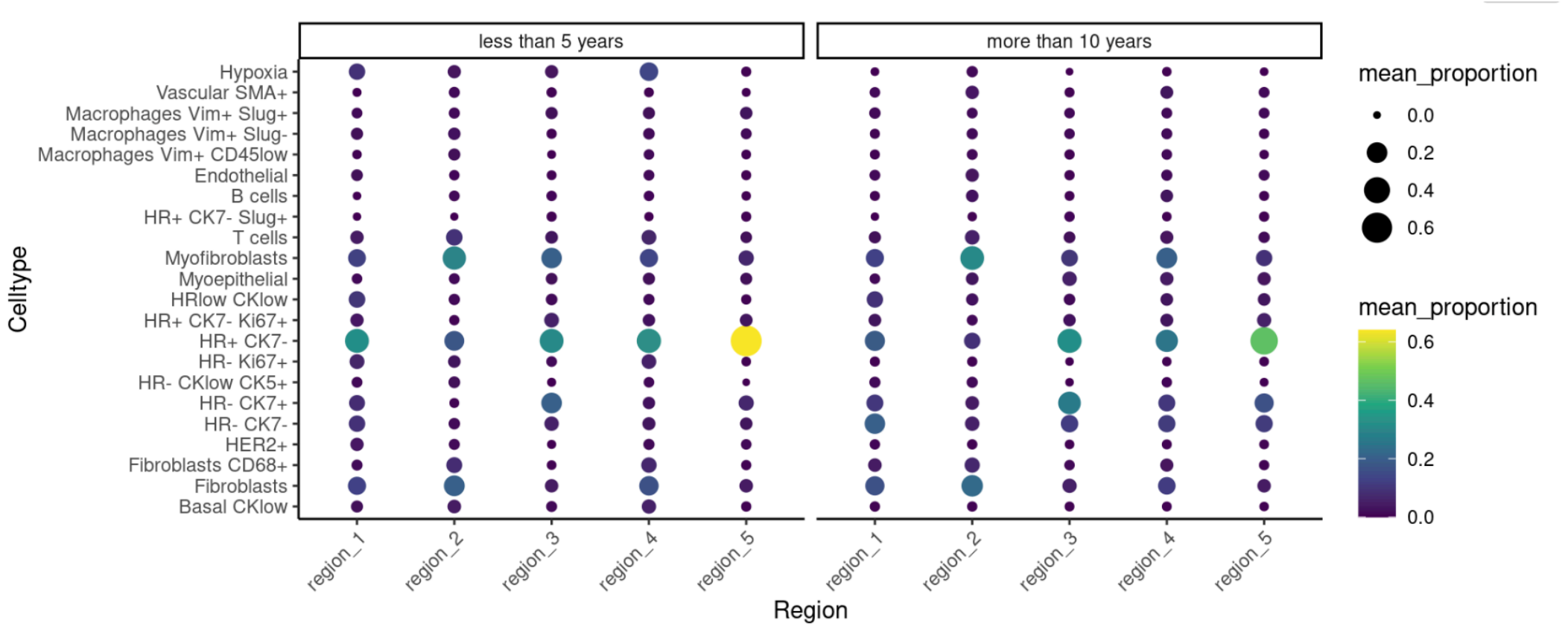
- Today we use **lisaClust** as an example. LisaClust identifies and visualises regions of tissue where **spatial associations** between cell-types is similar.



Describe tissue microenvironment and neighbourhood

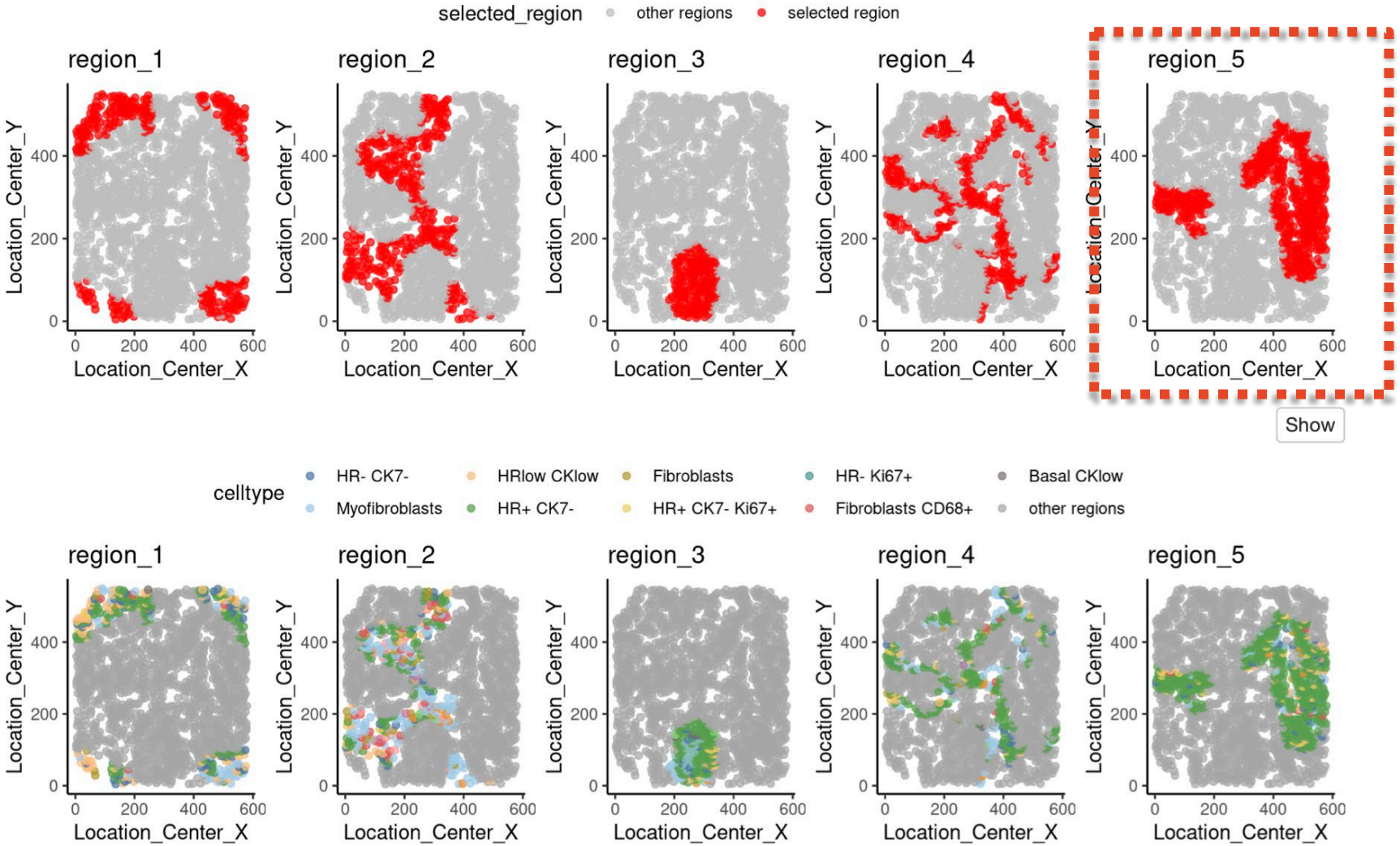
As a case study, we compare individuals with good or poor prognosis.
We define:

- good prognosis as individuals with > 10 years survival and
- poor prognosis as individuals with < 5 years survival.



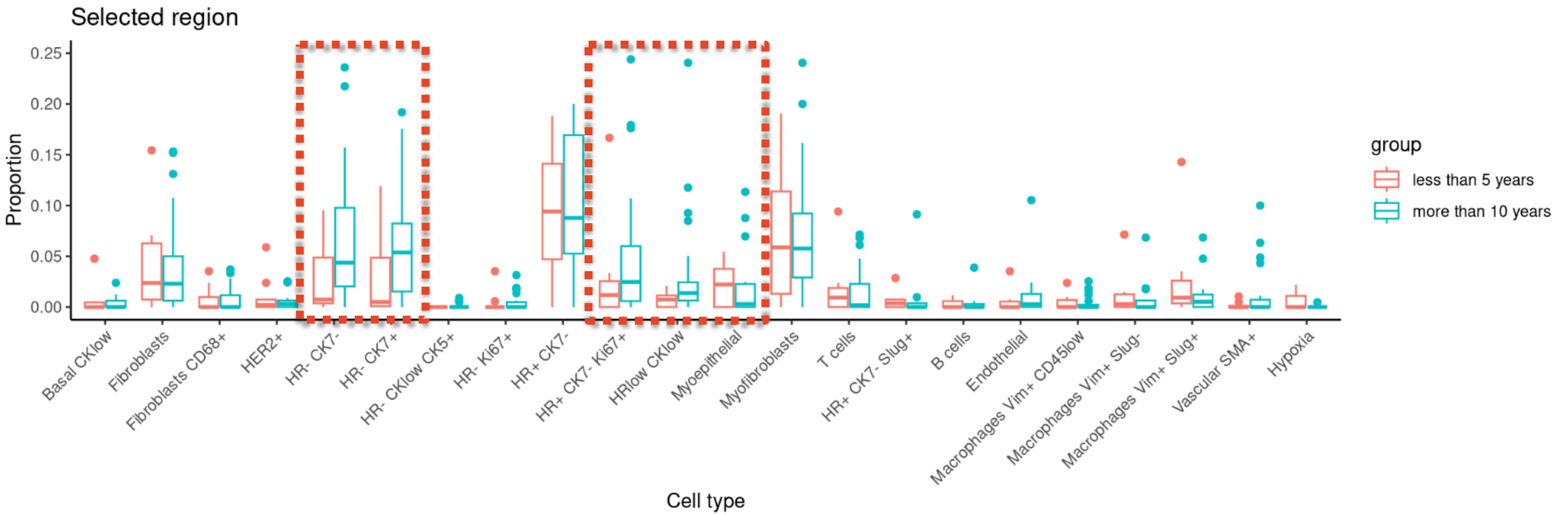
Describe tissue microenvironment and neighbourhood

Visualise the region output by highlighting each **region** and the **cell-types** in each region



Describe tissue microenvironment and neighbourhood

Here we select region 5 and plot boxplot of cell-type proportions across patients, coloured by the condition



Questions?

THE UNIVERSITY OF SYDNEY

https://vizgen.com/the-road-from-spatial-mapping-to-improved-human-he

PART III:
Feature engineering with scFeatures

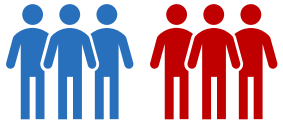
Dr Yue Cao

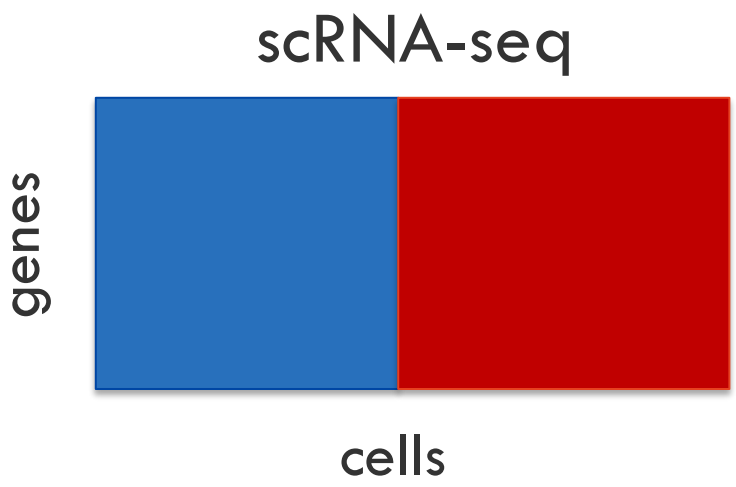
School of Mathematics and Statistics

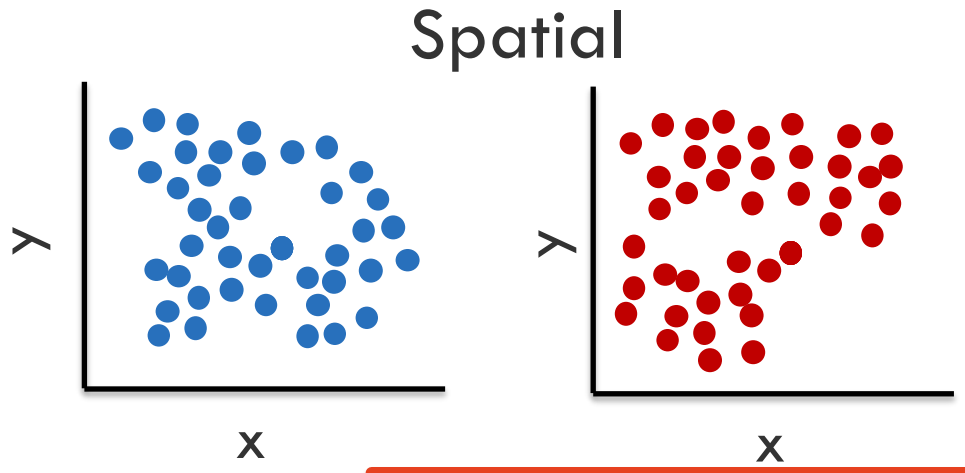
THE UNIVERSITY OF SYDNEY

scFeatures

How do we compare the difference between two conditions?





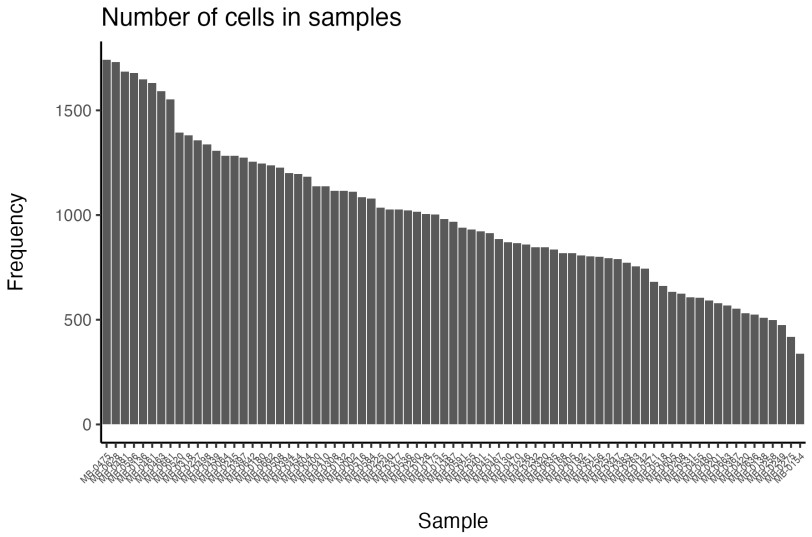


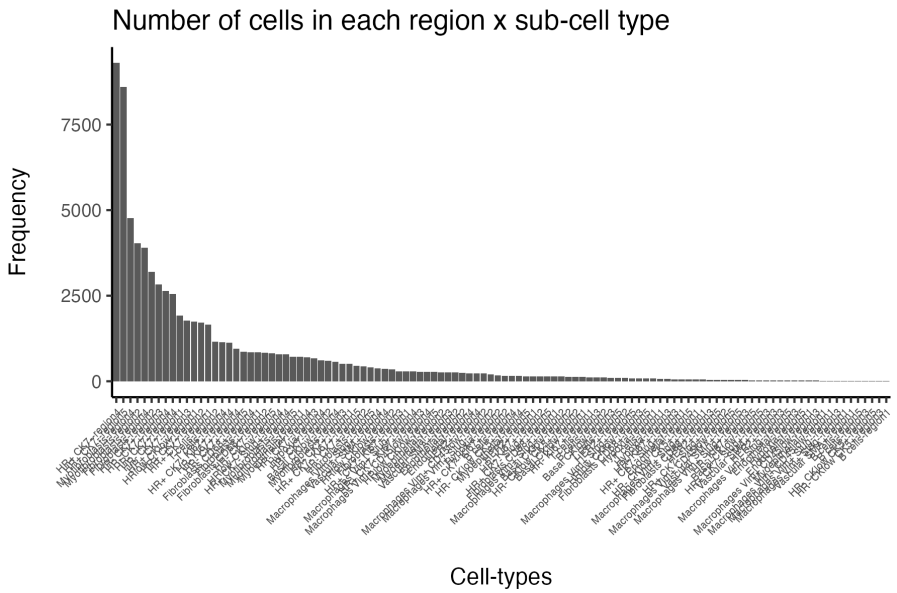
Aggregated gene expression

Cell-type
- Proportions
- Gene expression
- Pathway expression
- Cell-cell communication

Spatial metrics
- L function
- cell type interaction
- Moran's I
- Nearest neighbour correlation

Quality Control



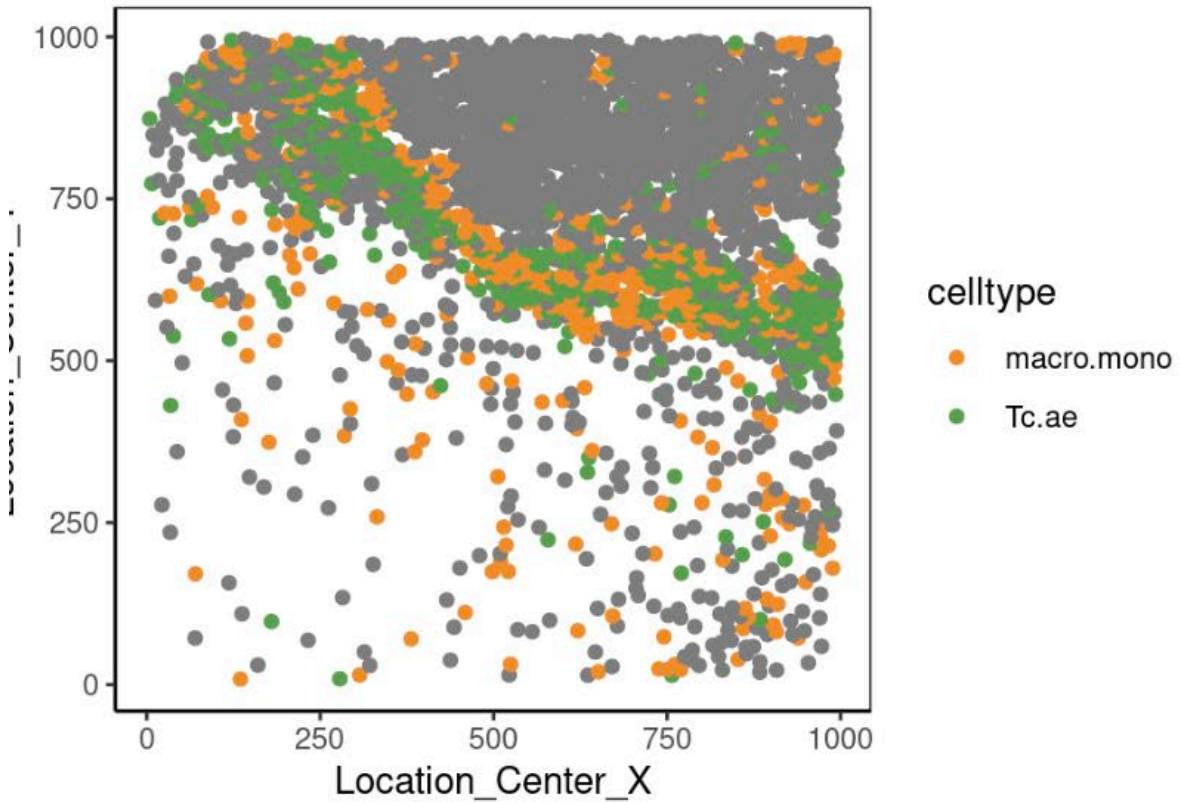


Are there any samples or cell-types you would like to remove from the data? Why or why not?

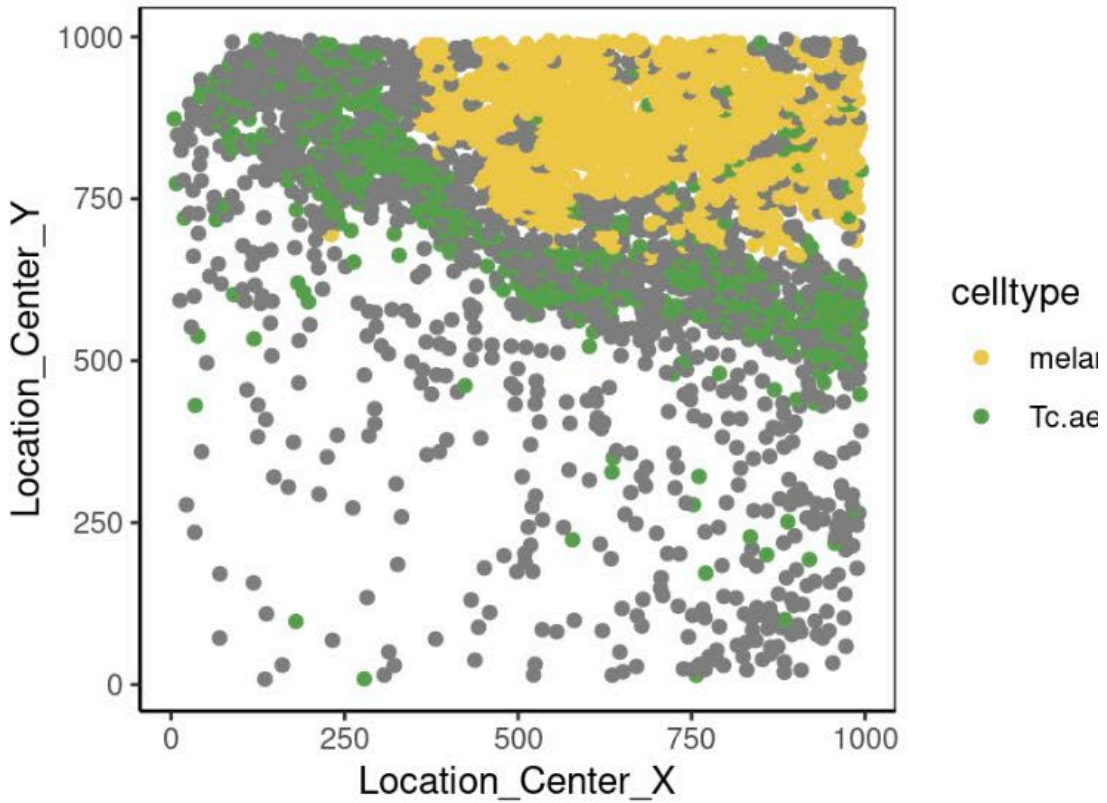
L function

Measure of the clustering of data points

High L value

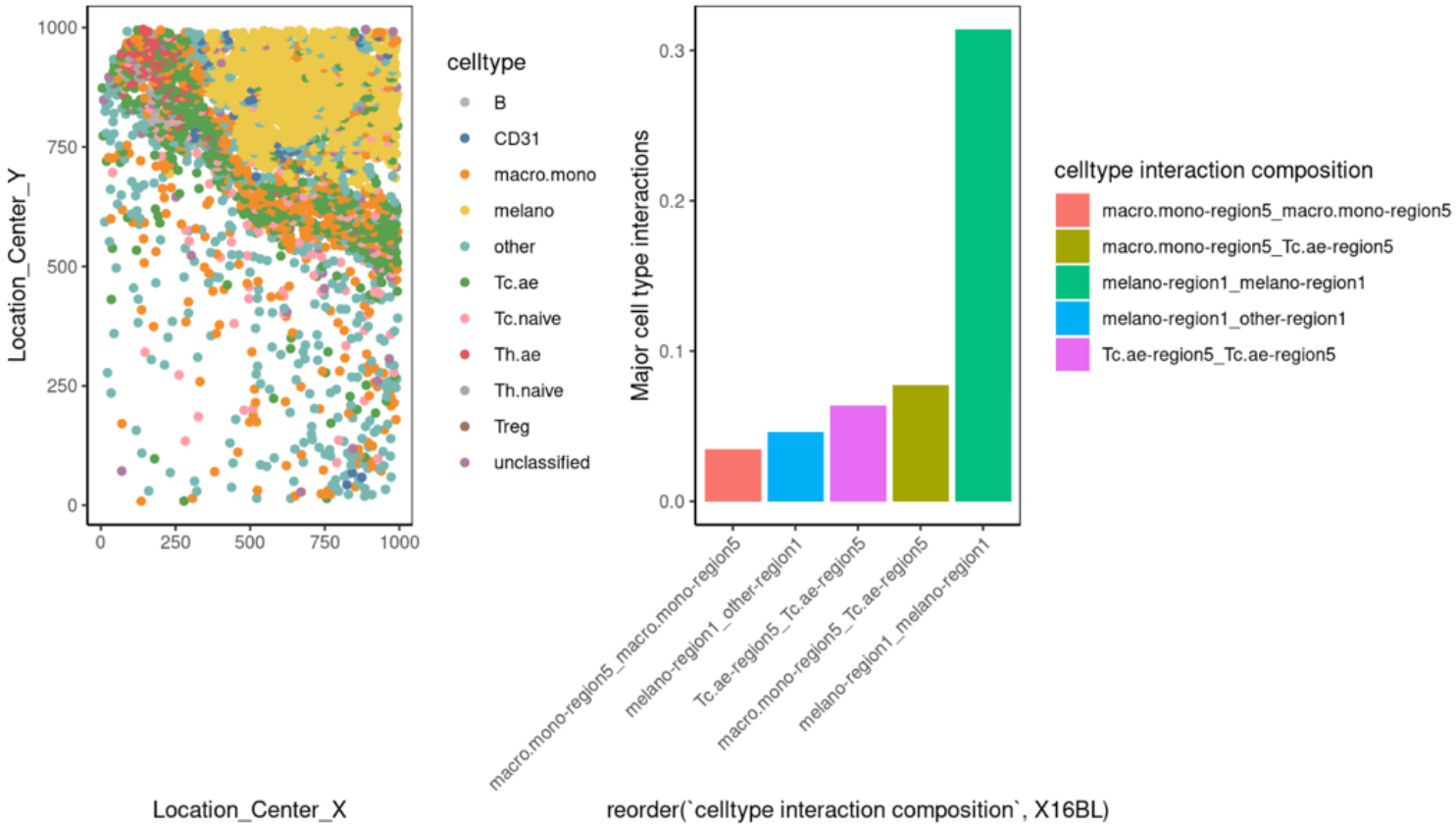


Low L value



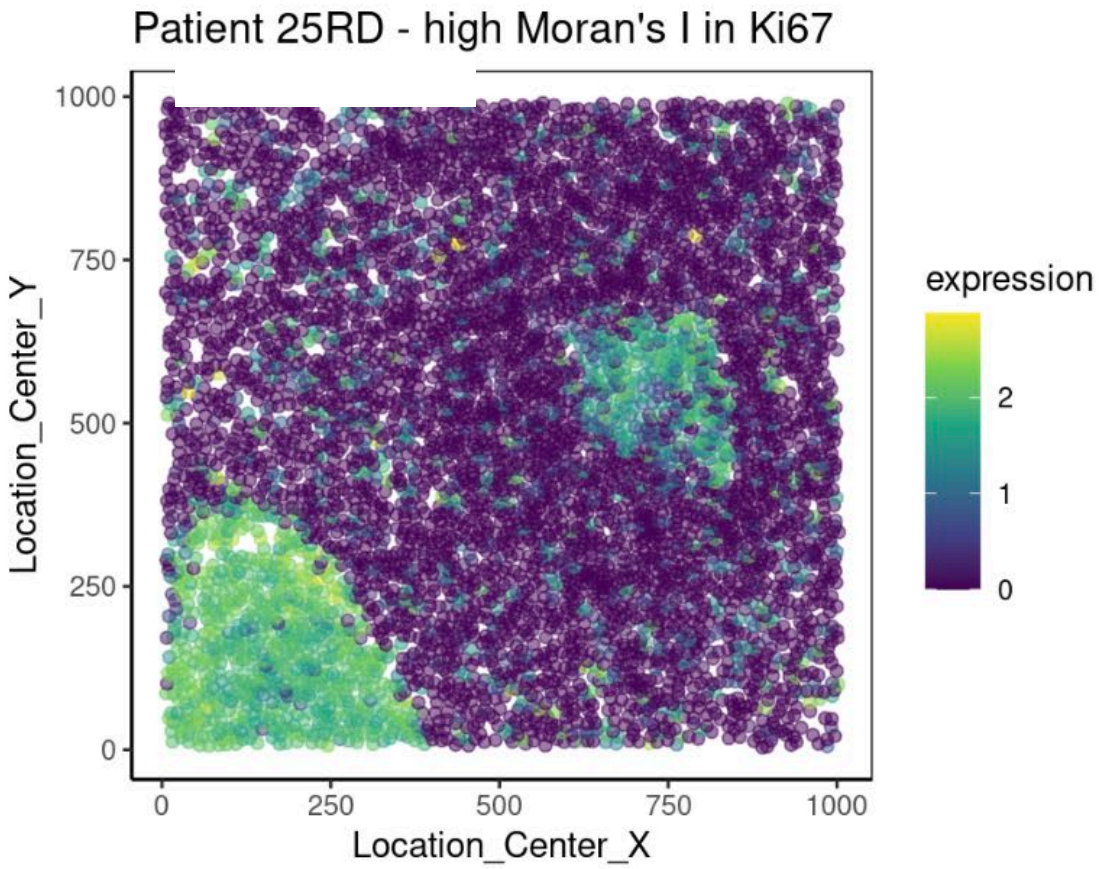
Cell-type interaction composition:

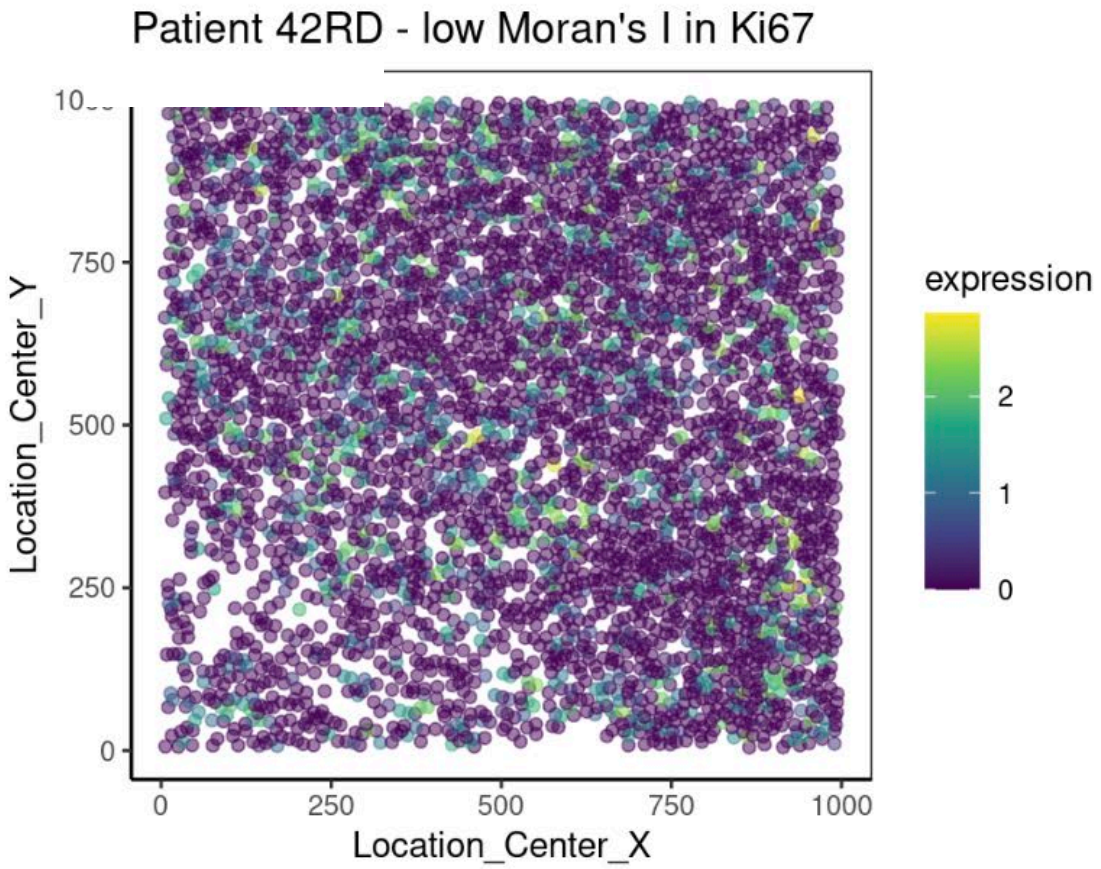
We calculate the **nearest neighbours** of each cell and then calculate the pairs of cell-type based on their nearest neighbours. This allow us to summarise it into a **cell-type interaction composition**.



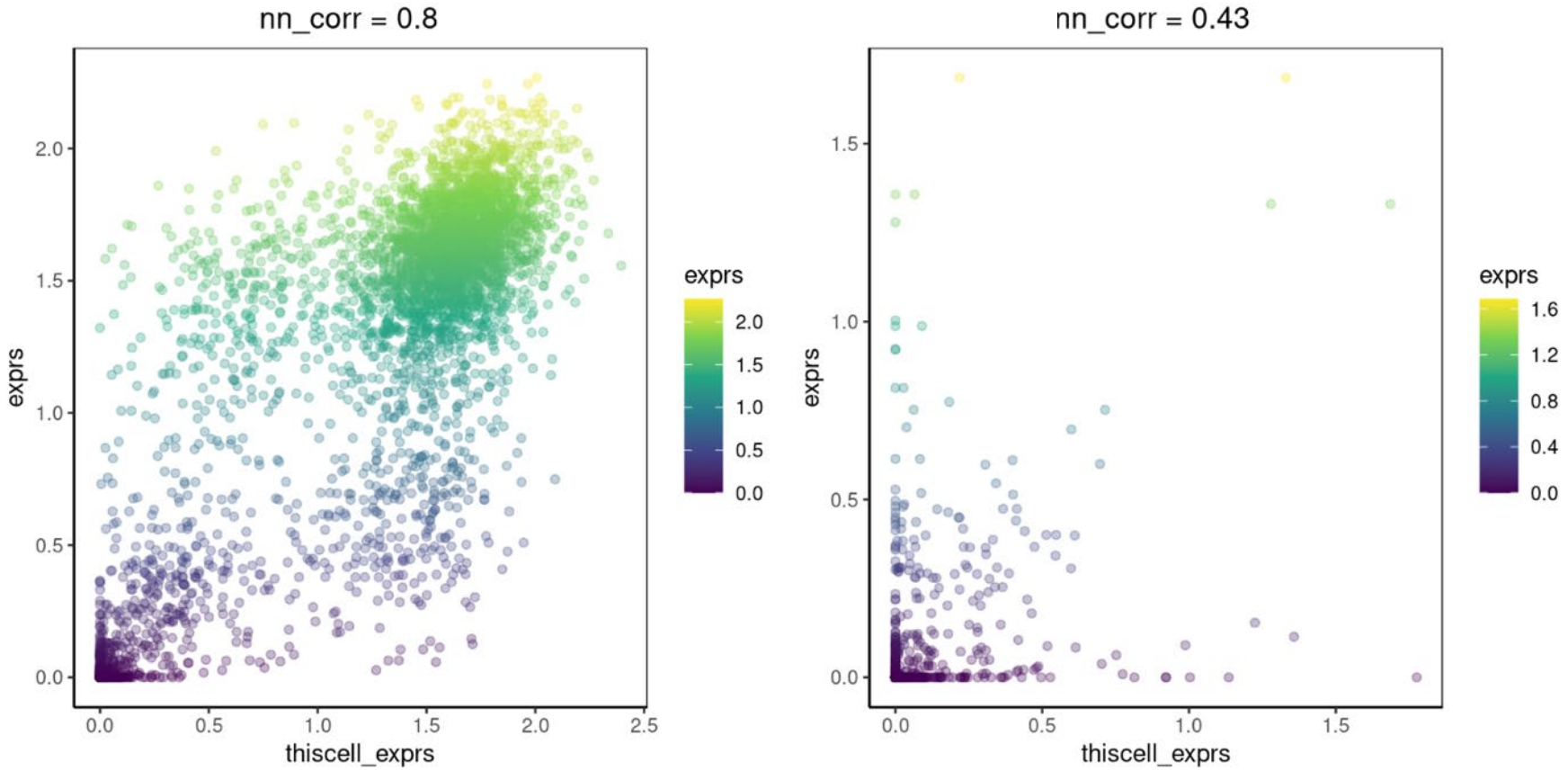
Moran's I

Measure of autocorrelation

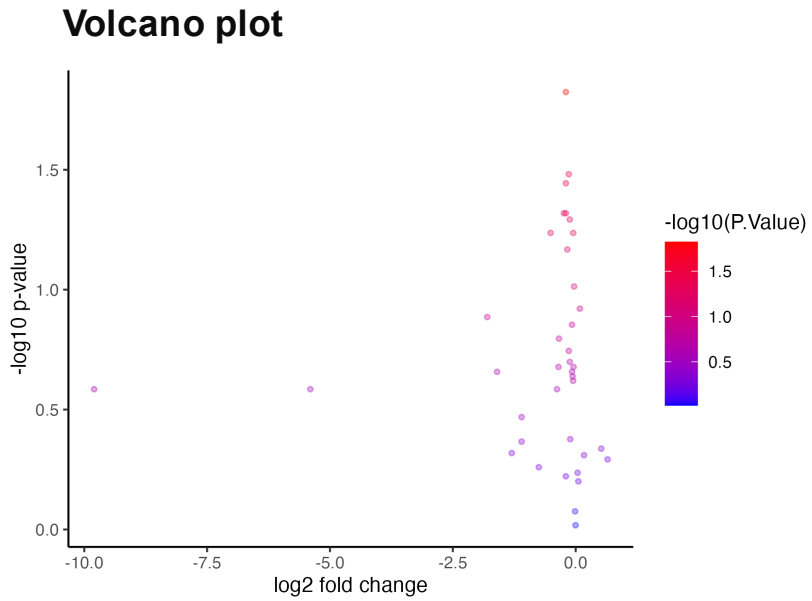




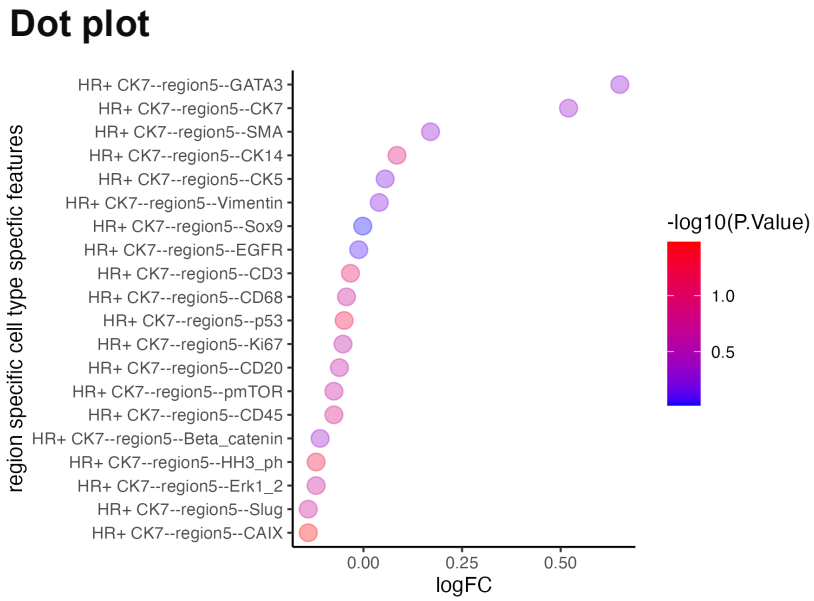
Nearest Neighbor Correlation – illustration with S100 (HER2 gene)



Association Study



VS



Which do you prefer, the volcano plot or dot plot? Why?

What can we do with the features

Association study to identify condition associated **cell types** and **features**

Patient **prediction** with outcome label

Patient **clustering** without outcome label

Break (15 mins)

THE UNIVERSITY OF SYDNEY

PART IV:
Disease classification with ClassifyR

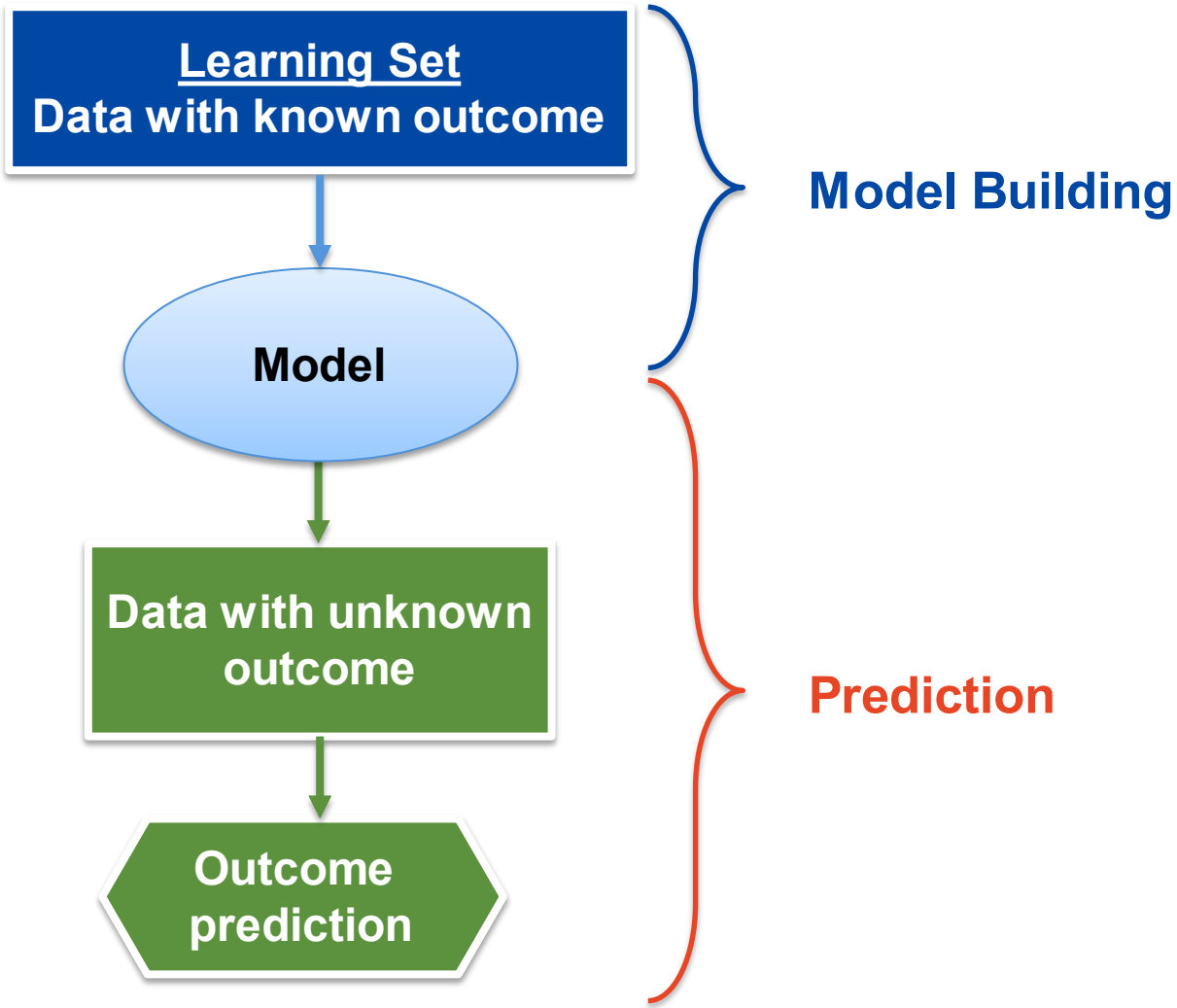
Dr Dario Strbenac

School of Mathematics and Statistics

THE UNIVERSITY OF SYDNEY

ClassifyR

Modelling and Evaluation



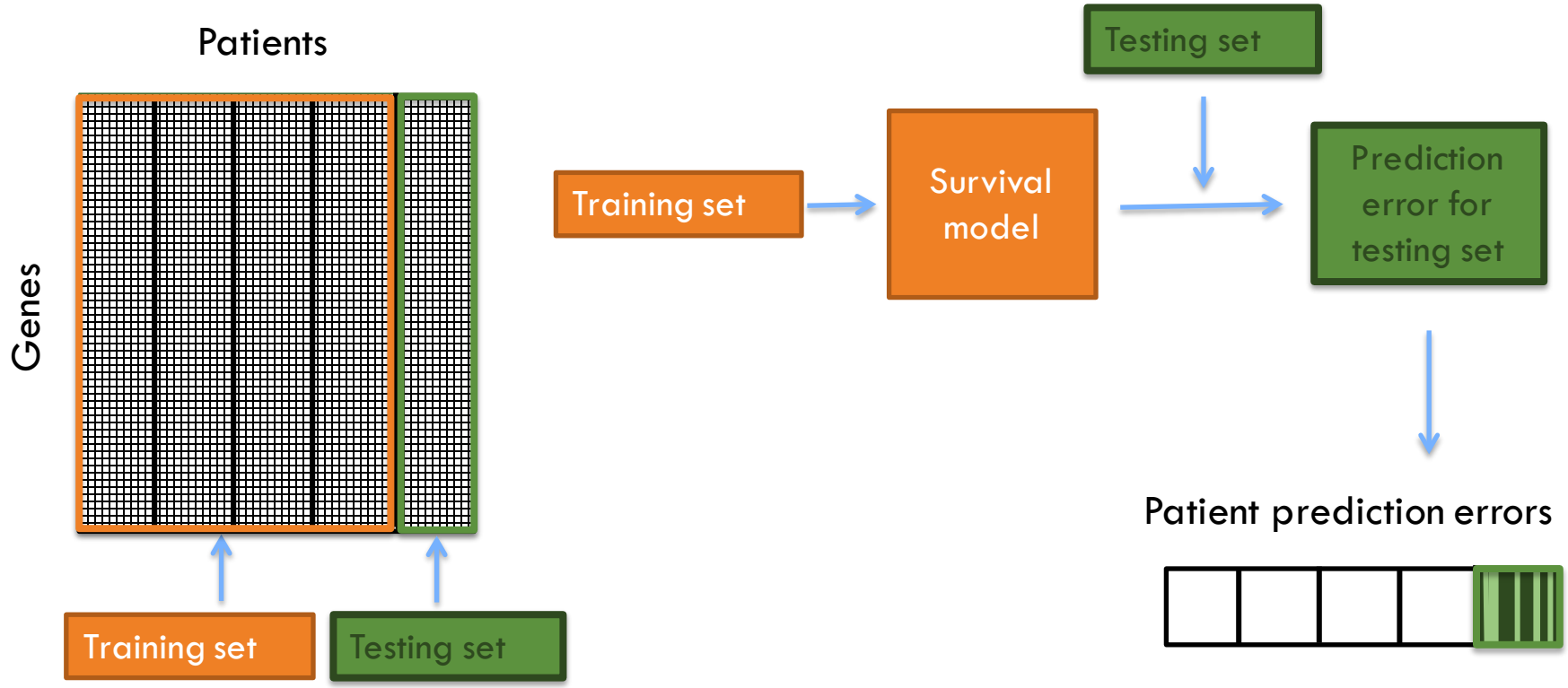
Many choices in model building:

- Algorithm
- Feature selection
- Parameters
- Distance measures
- Aggregation methods
- Many more!

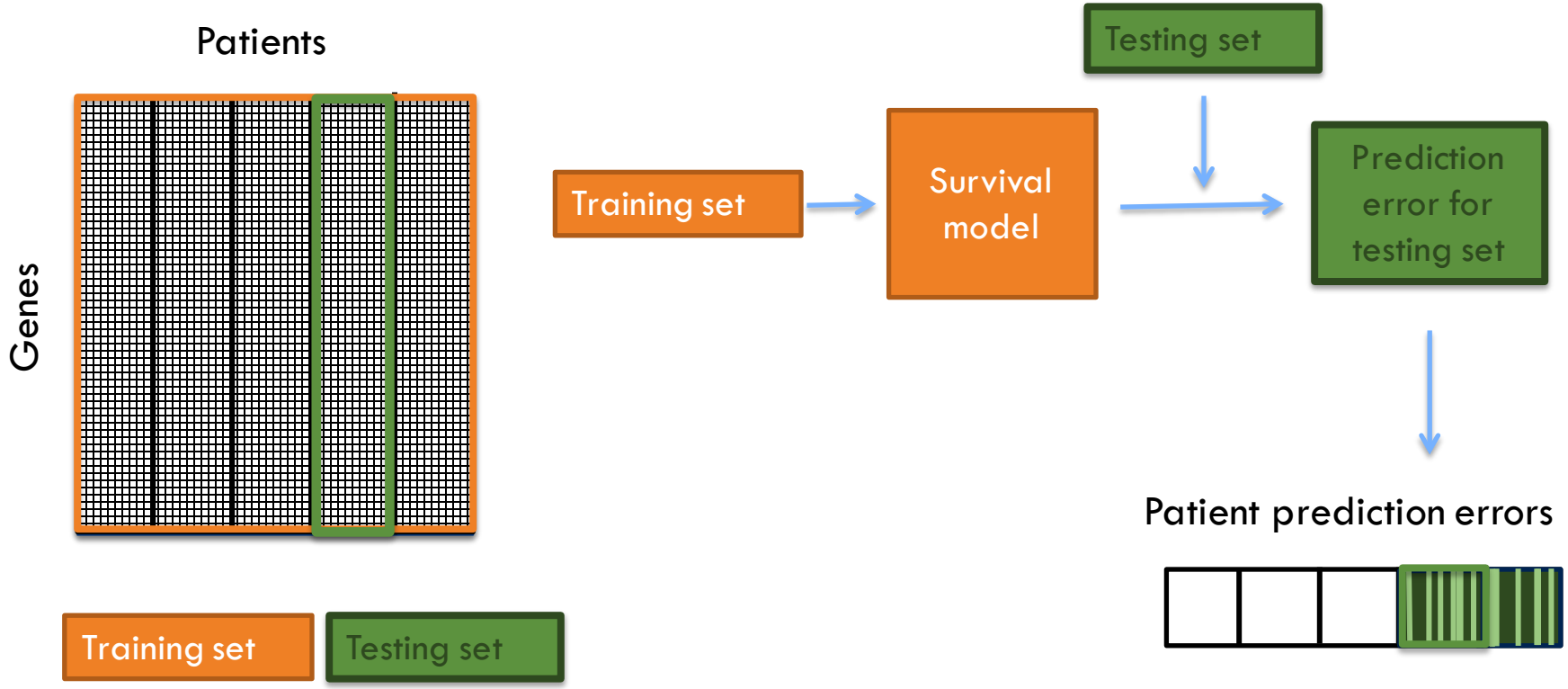
The best choices will vary based on the **data** and **task**.

Any model needs to be **evaluated** for its performance on **future samples**. But we often only have a single data set available.

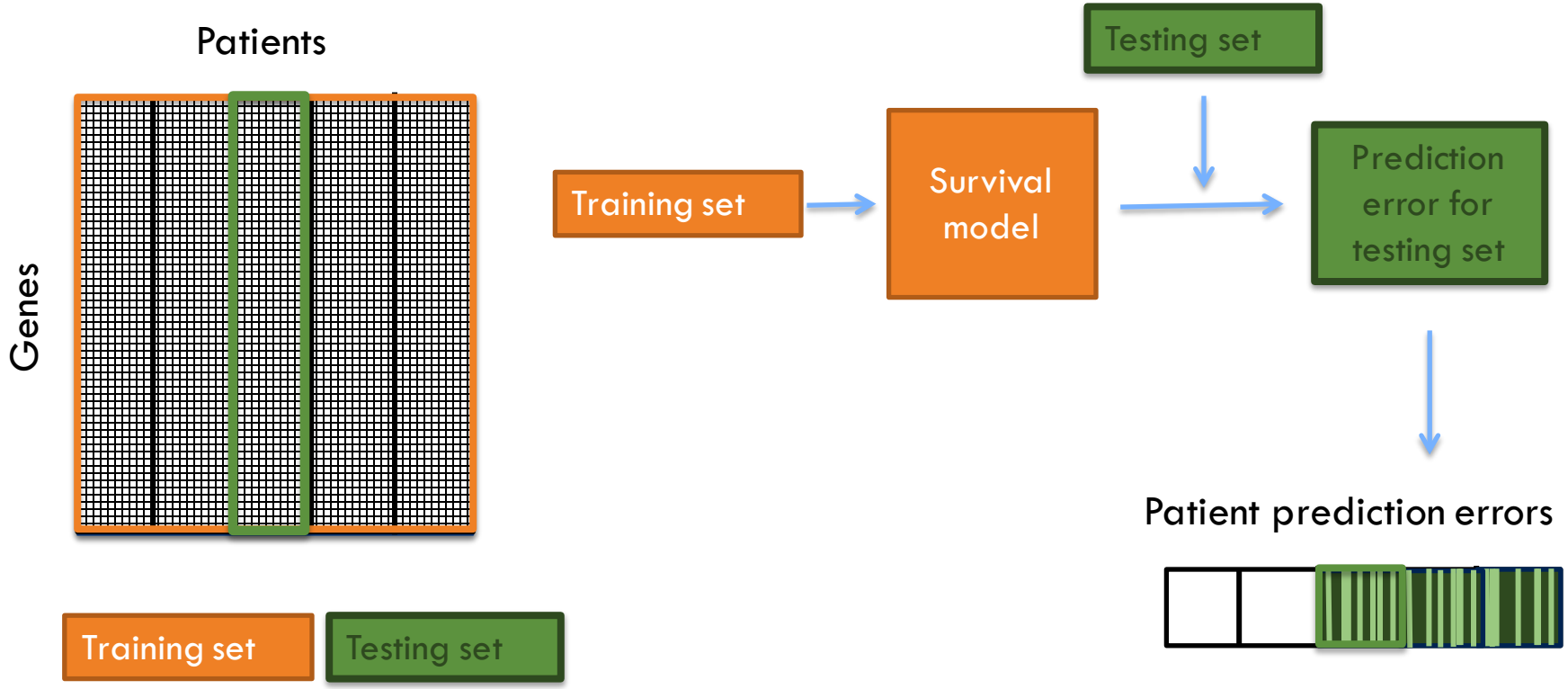
5-fold cross validation



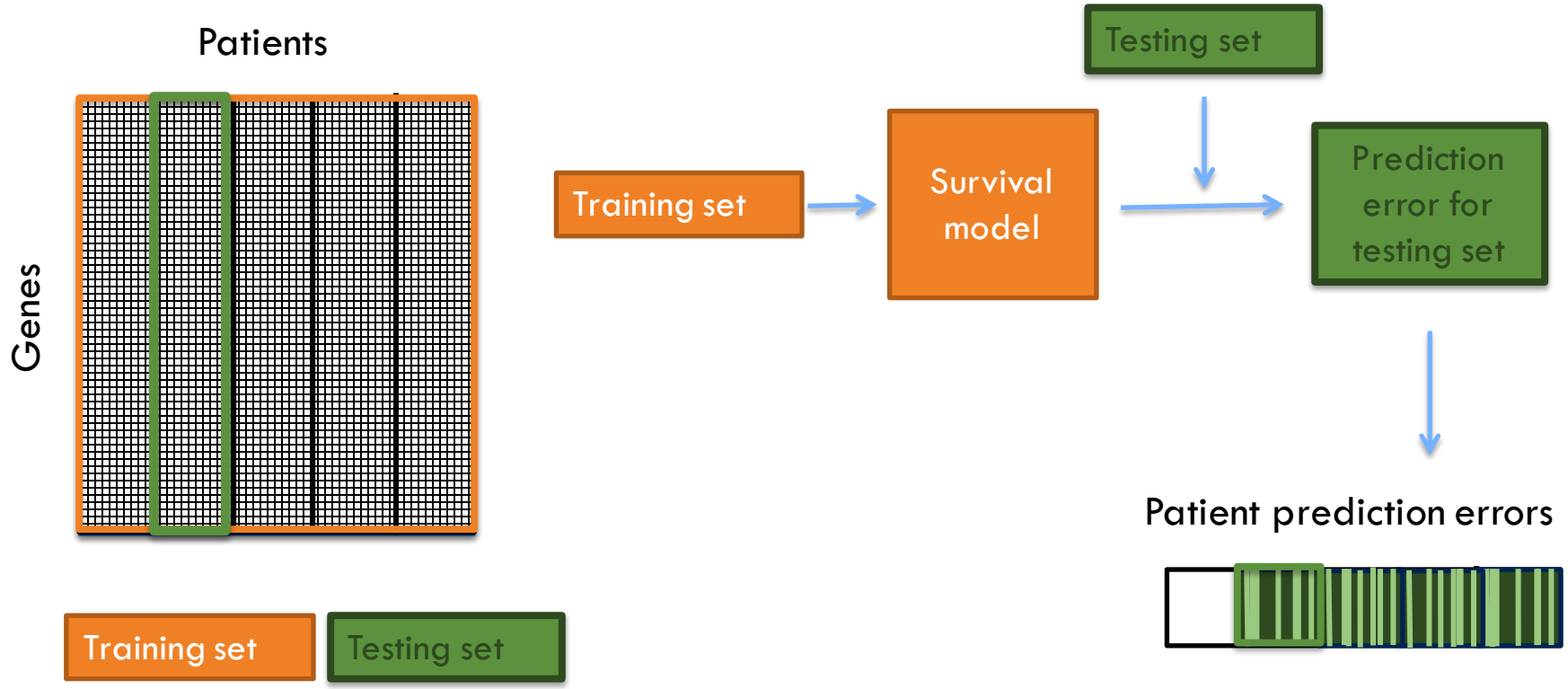
5-fold cross validation



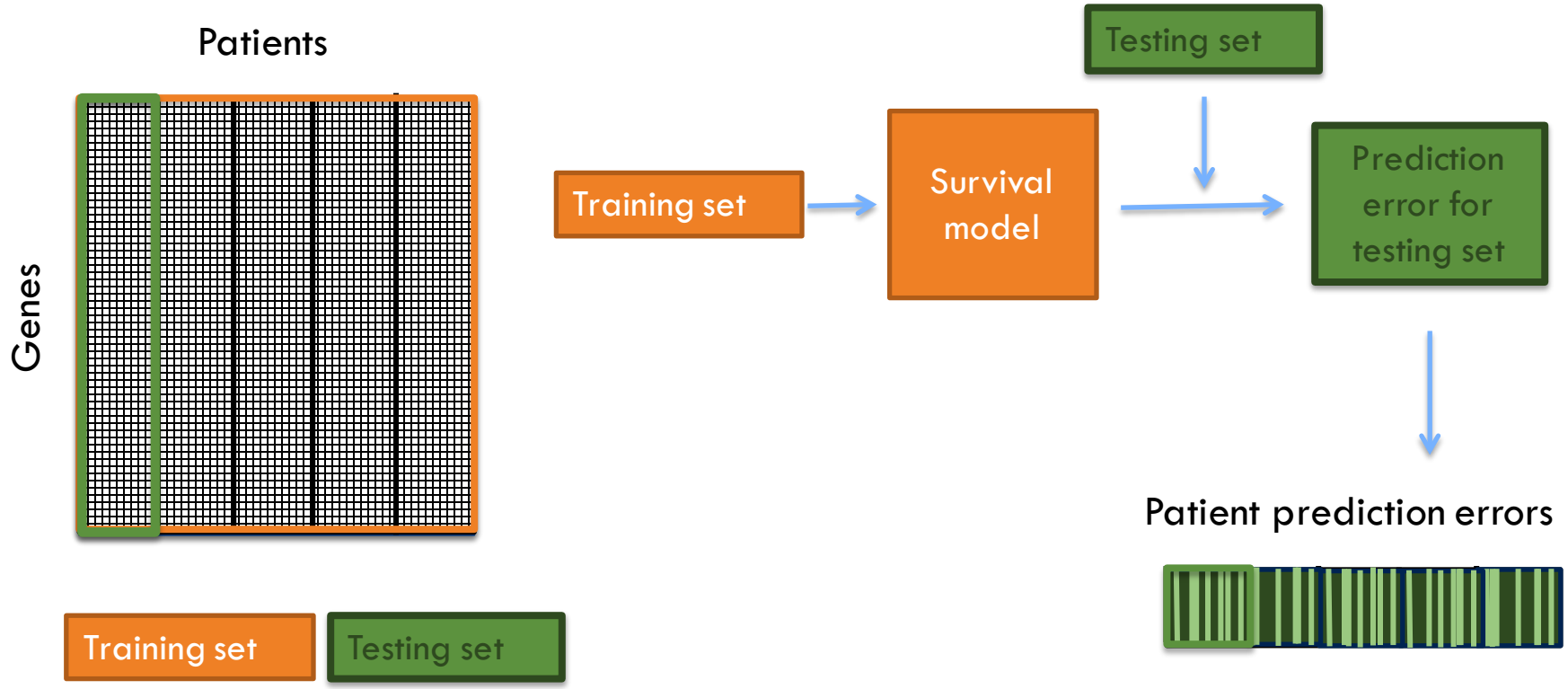
5-fold cross validation



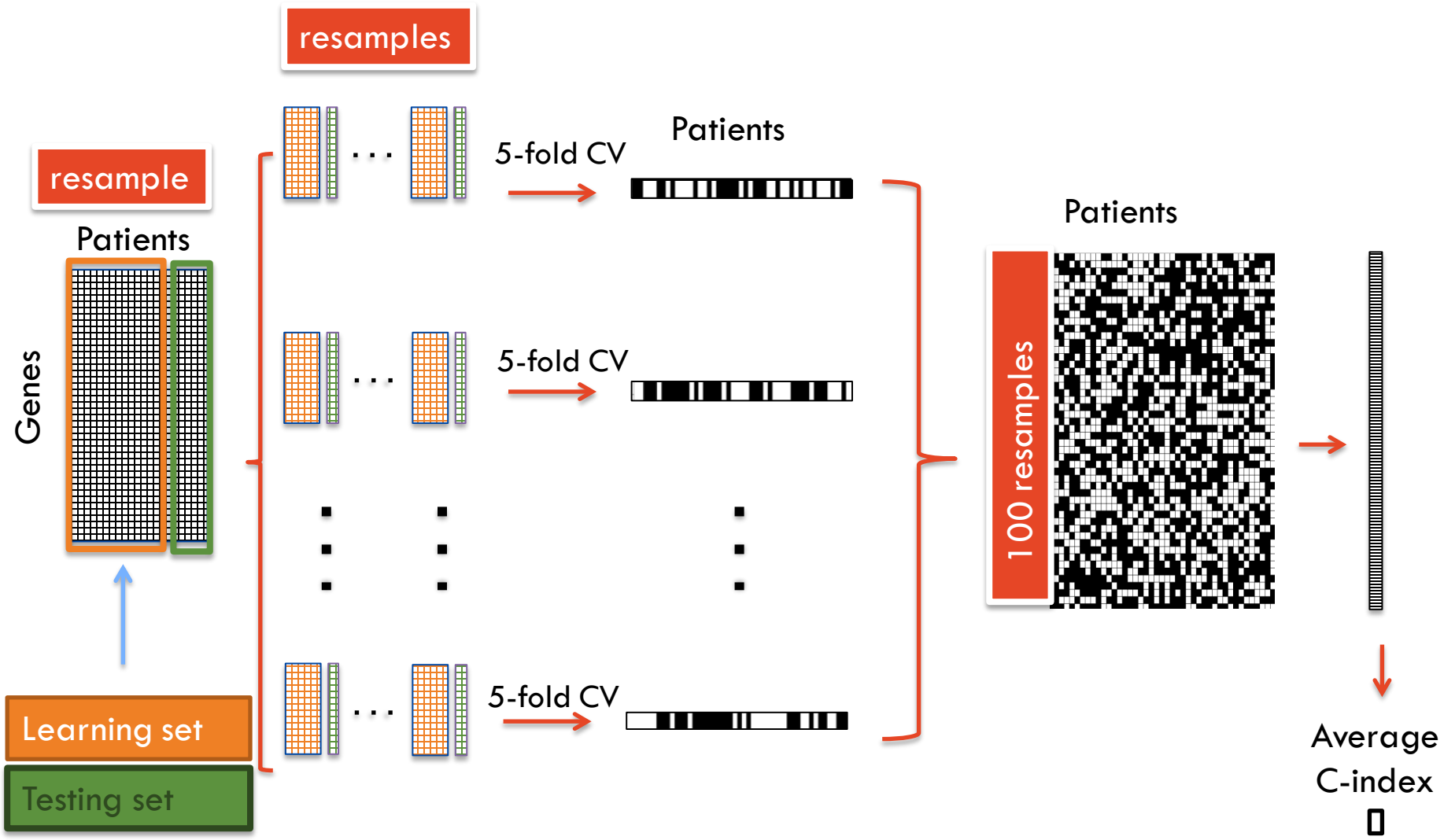
5-fold cross validation



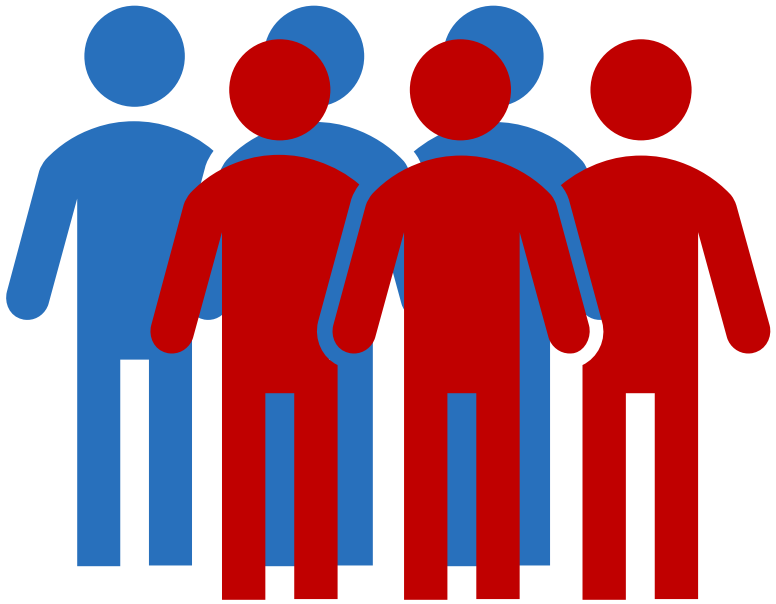
5-fold cross validation



Repeated 5-fold cross validation



Cohort heterogeneity



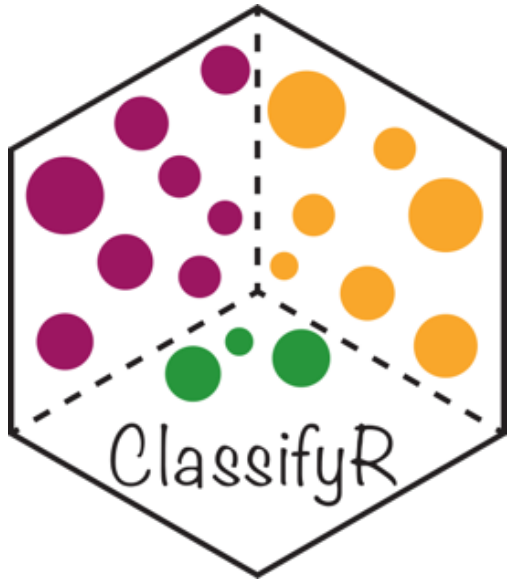
Do you think all samples equally classifiable? Say healthy vs diseased. Why or why not?

ClassifyR: outcome prediction

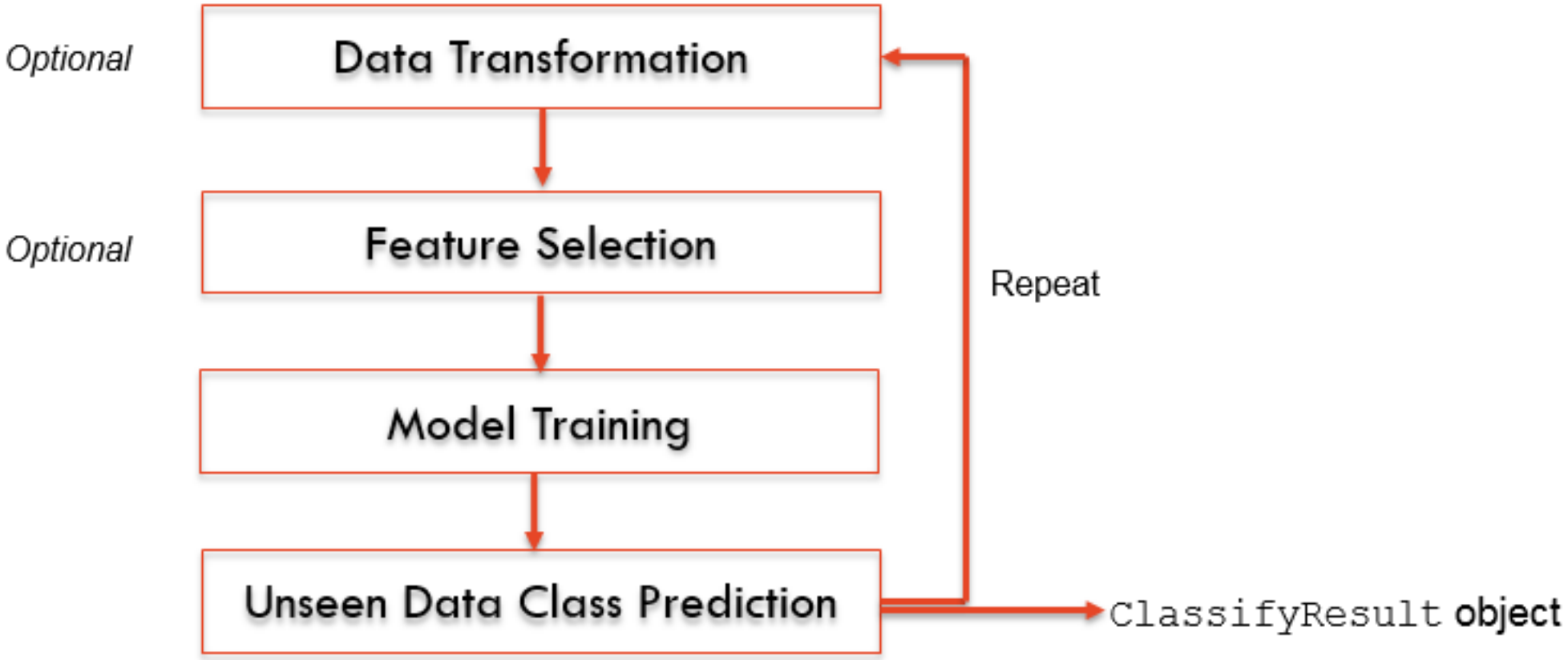
ClassifyR: an R package for performance assessment of classification with applications to transcriptomics FREE

Dario Strbenac ✉, Graham J. Mann, John T. Ormerod, Jean Y.H. Yang Author Notes

Bioinformatics, Volume 31, Issue 11, 1 June 2015, Pages 1851–1853,
https://doi.org/10.1093/bioinformatics/btv066
Published: 01 February 2015 **Article history** ▾



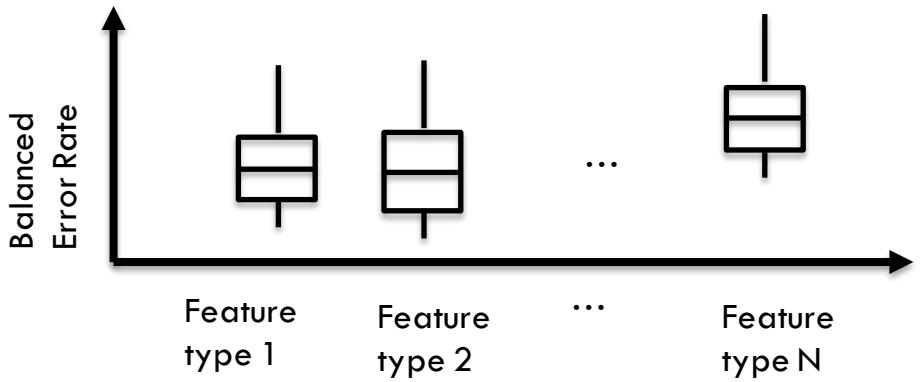
Full cross-validation loops



Multi-view Data Modelling in ClassifyR

Feature type 1 | Feature type 2 | Feature type 3 | ... | Feature type N

Cross-validation



Questions?

THE UNIVERSITY OF SYDNEY

https://vizgen.com/the-road-from-spatial-mapping-to-improved-human-he

PART V:
Identifying cohort heterogeneity with ClassifyR

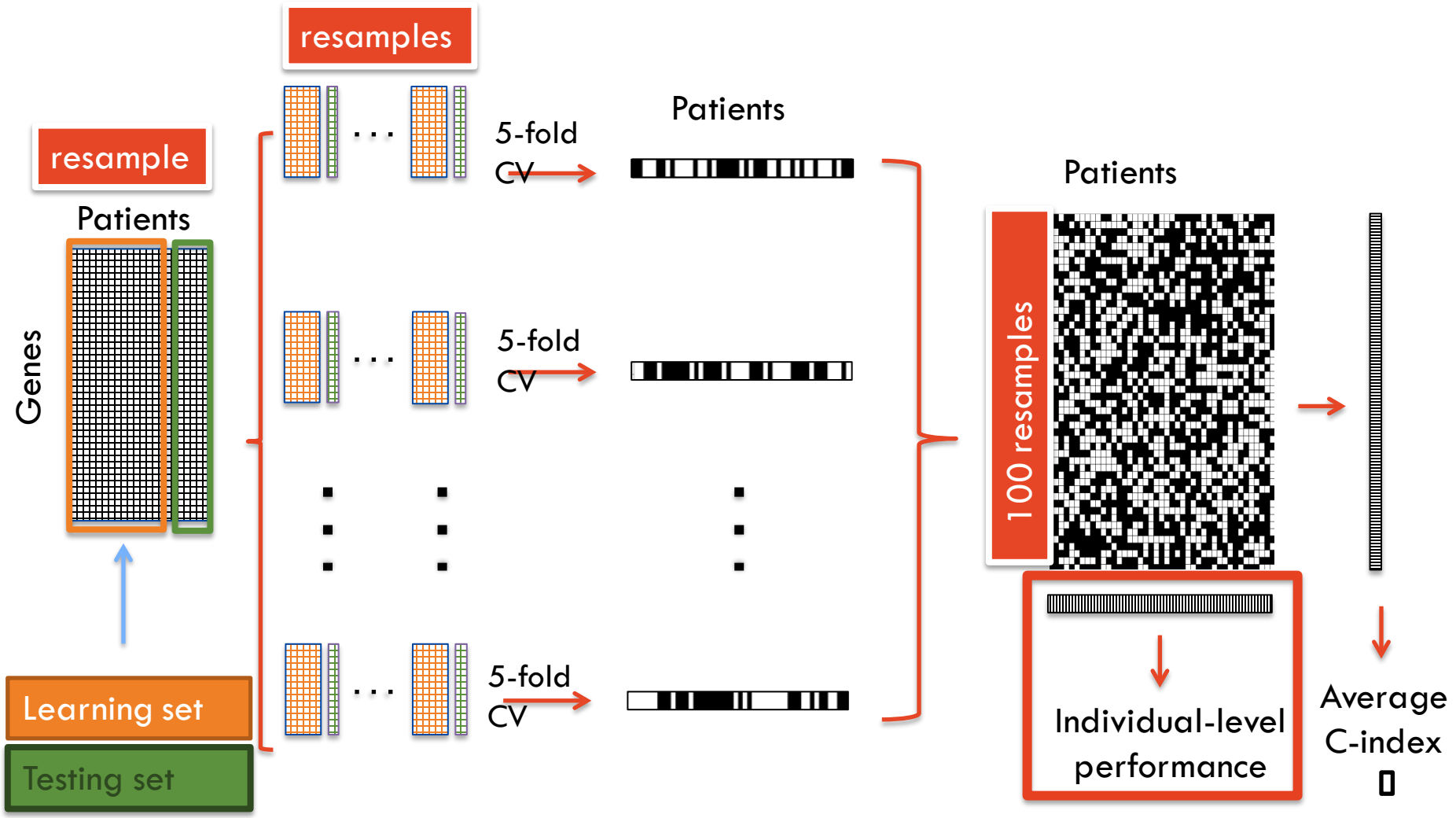
THE UNIVERSITY OF SYDNEY

Identifying cohort heterogeneity

- Important for precision medicine
- Personally tailored treatments
- Most modelling methods are assessed by overall accuracy: 80% of samples correctly classified
- What about the classification performance of individuals?

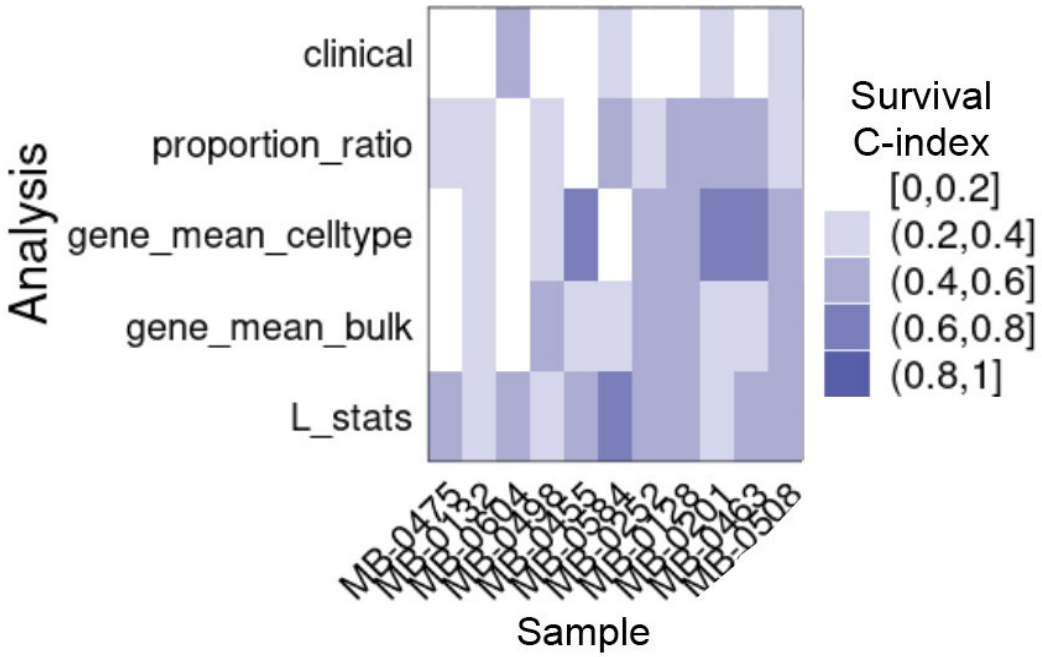
Identifying cohort heterogeneity

Human cohorts are often very **heterogenous**. It can be useful to explore cohort heterogeneity by looking at **individual-level performance.**



Hard vs easy cases (Cohort heterogeneity)

- The **individual-level performance** can be visualised in one line of code using the `samplesMetricMap` **function**



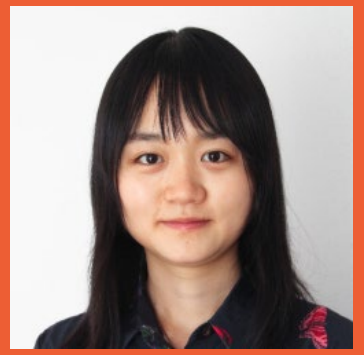
Are spatial features important for predicting recurrence? Does it hold for all individuals?

Questions?

THE UNIVERSITY OF SYDNEY

https://vizgen.com/the-road-from-spatial-mapping-to-improved-human-he

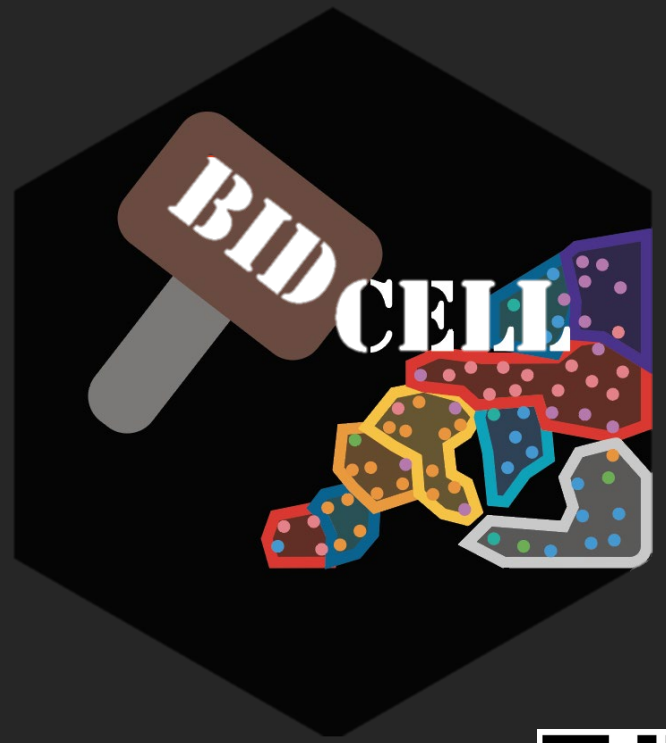
PART VI:
Cell segmentation with BIDCell



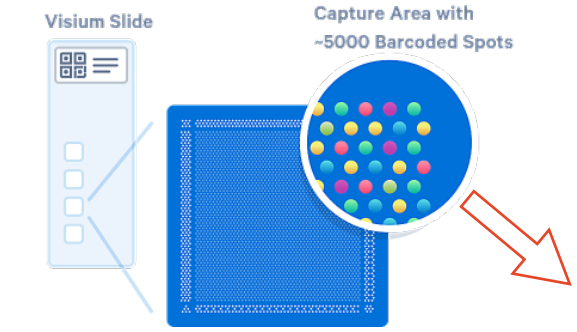
Dr Helen (Xiaohang) Fu

School of Mathematics and Statistics

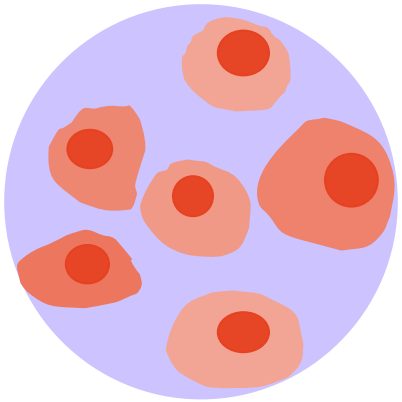
THE UNIVERSITY OF SYDNEY



Spatially resolved transcriptomics

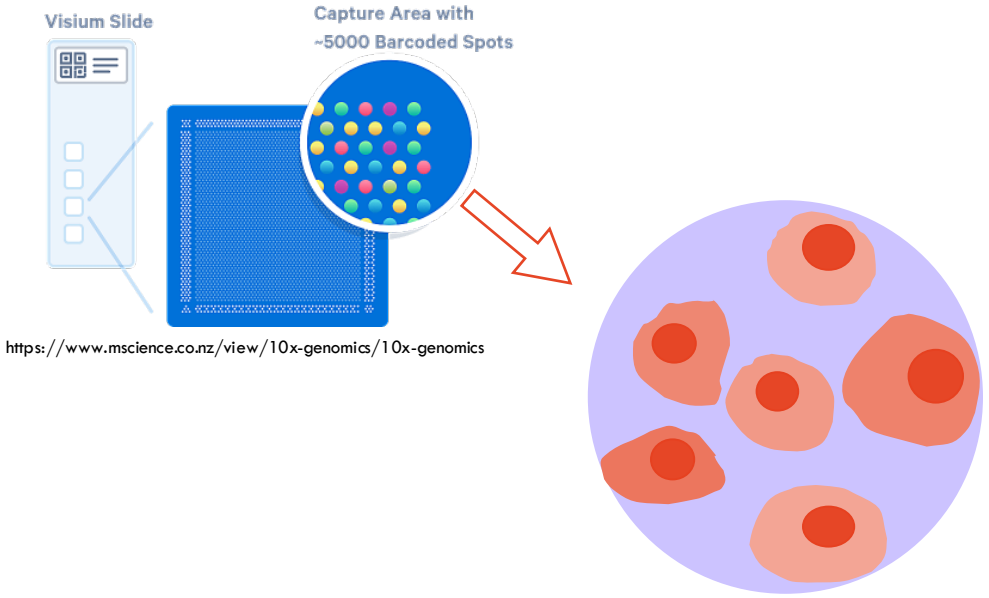


https://www.mscience.co.nz/view/10x-genomics/10x-genomics



1-10 cells in 1 spot

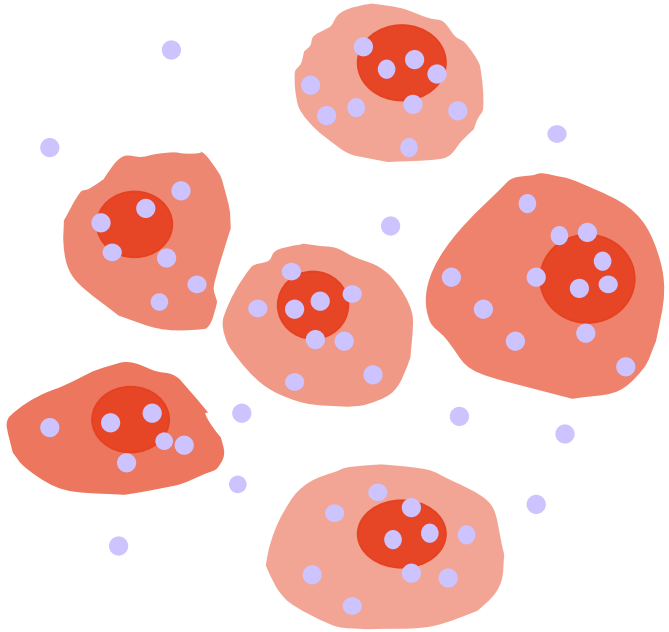
Spatially resolved transcriptomics



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

1-10 cells in 1 spot

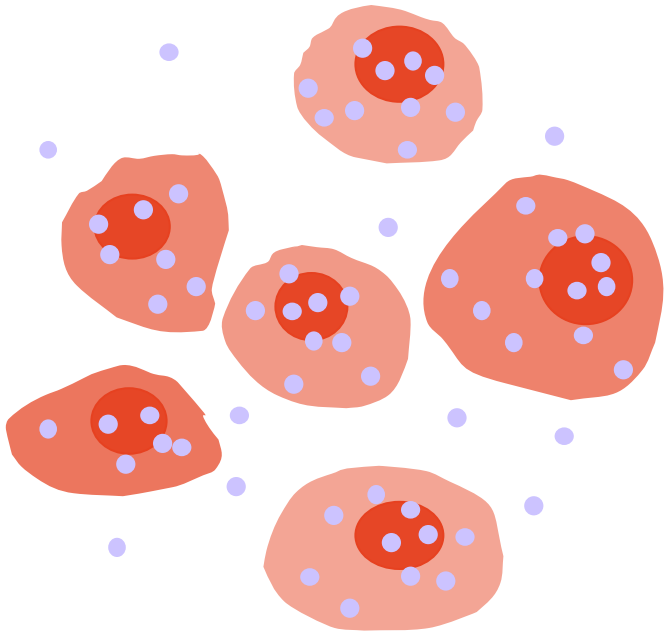
Subcellular spatially resolved transcriptomics (SST)



Subcellular detections

10x Genomics Xenium
NanoString CosMx
Vizgen MERSCOPE

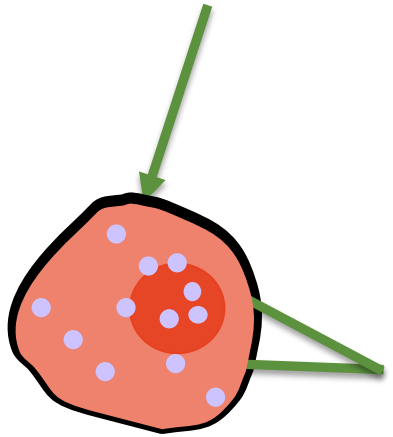
Subcellular spatially resolved transcriptomics (SST) → Cell segmentation



Subcellular detections

10x Genomics Xenium
NanoString CosMx
Vizgen MERSCOPE

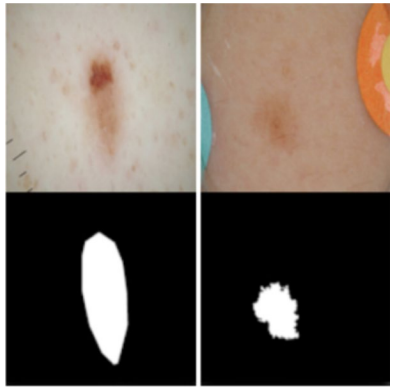
1. Generate a mask that predicts the cell body
2. Identify which transcripts are inside the cell mask
3. Quantify the cells gene expression with the identified transcripts



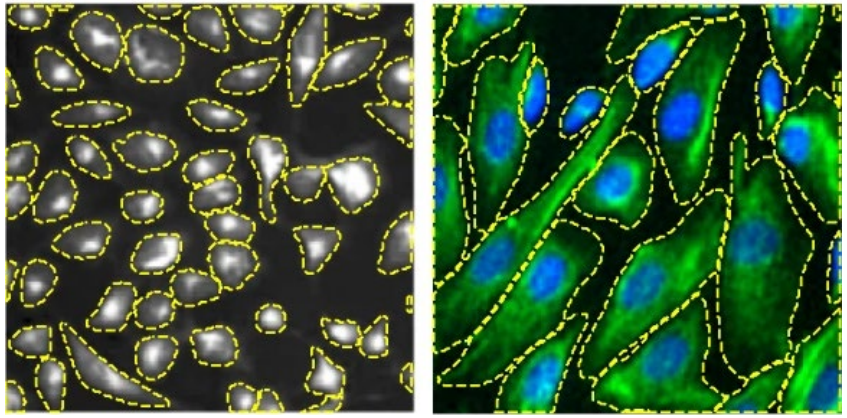
Object segmentation

- Segmentation is a computer vision problem
- Visual features can help us determine the boundaries of objects, and detect/learn what features usually occur together
  - e.g., **colour, edges, contrast**

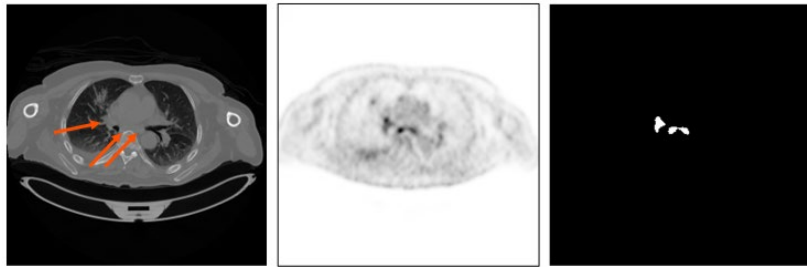
Chen et al. (2019)



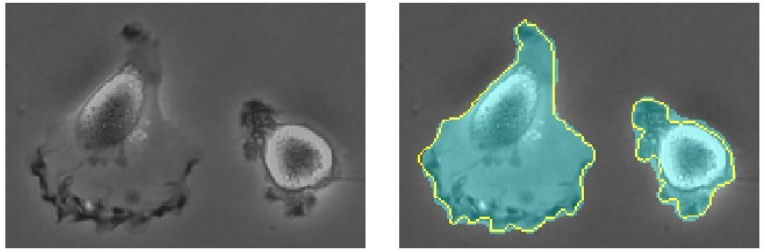
Codella et al. (2018)



Stringer et al. (2021)



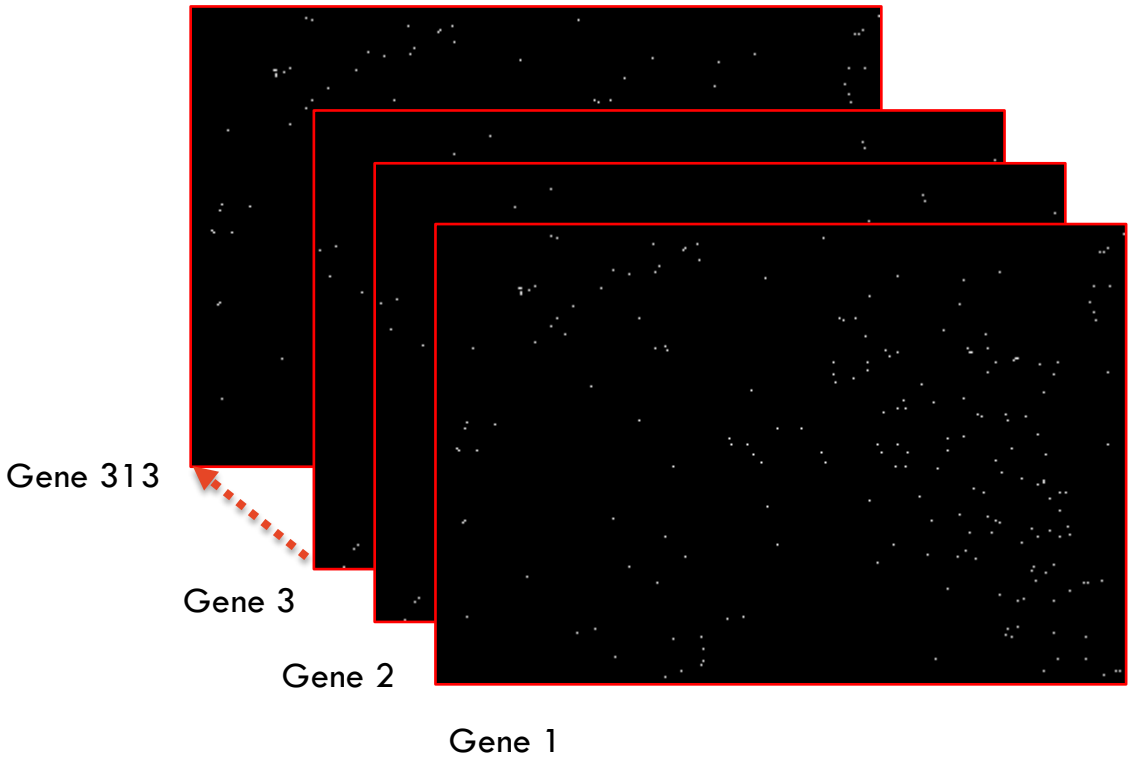
Fu et al. (2021)



Ronneberger et al. (2015)

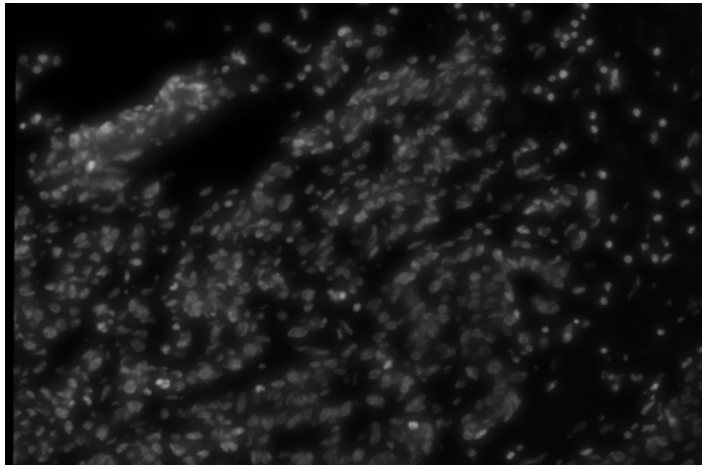
Challenges of SST images

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

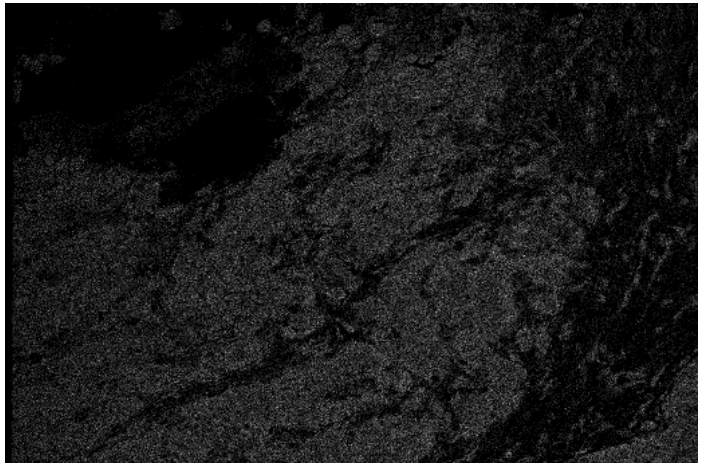


Summed channels

DAPI (Nuclei)



Transcripts (Summed channels)



10x Genomics Xenium (breast cancer, 313 genes)

Existing methods

Classical:

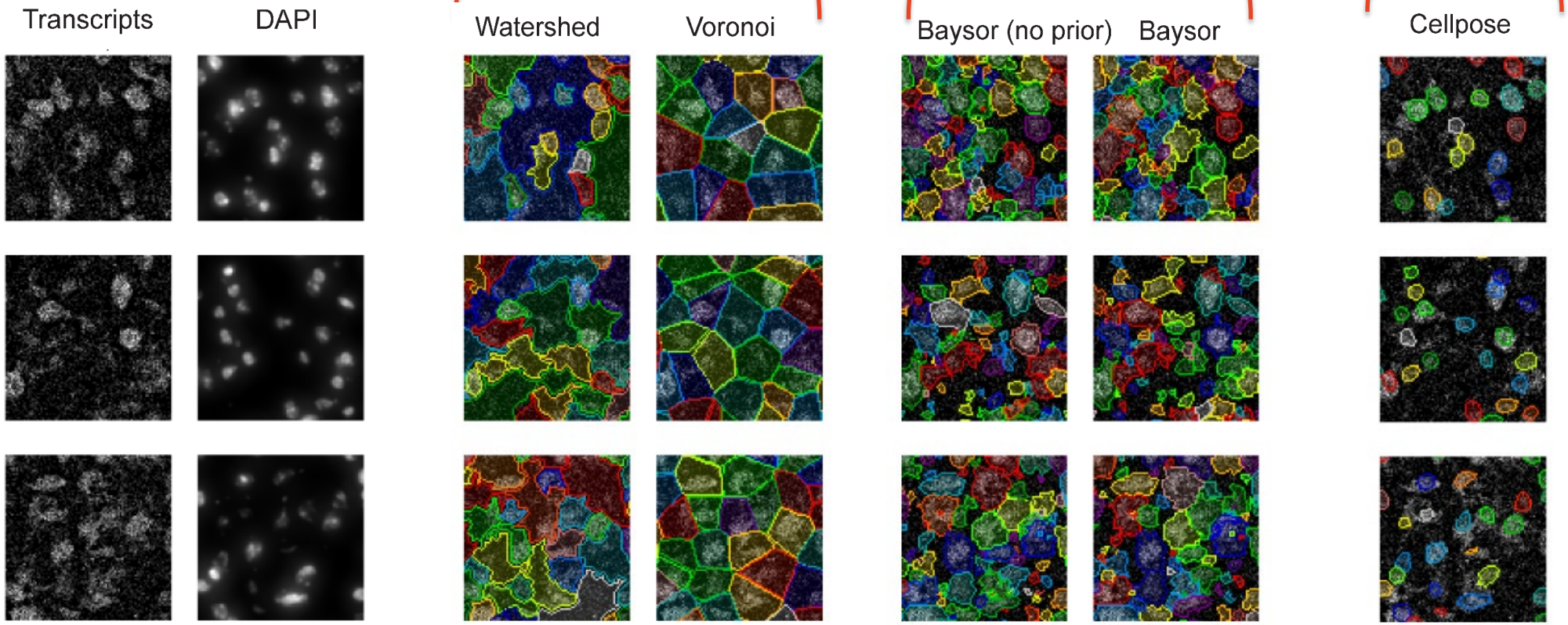
- Dilation from nuclei boundary
- Spatial gene expressions not fully leveraged

Clustering/Transcript:

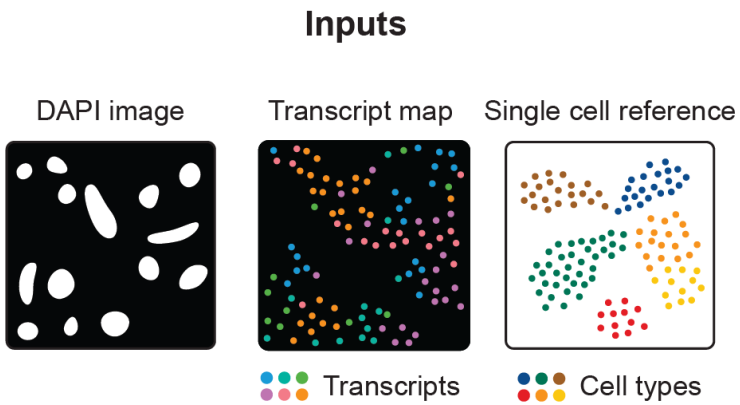
- Assumptions, e.g., homogeneity
- Cell morphologies not considered

Deep Learning:

- Models pre-trained on other datasets are unsuitable
- Requires ground truth annotations



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

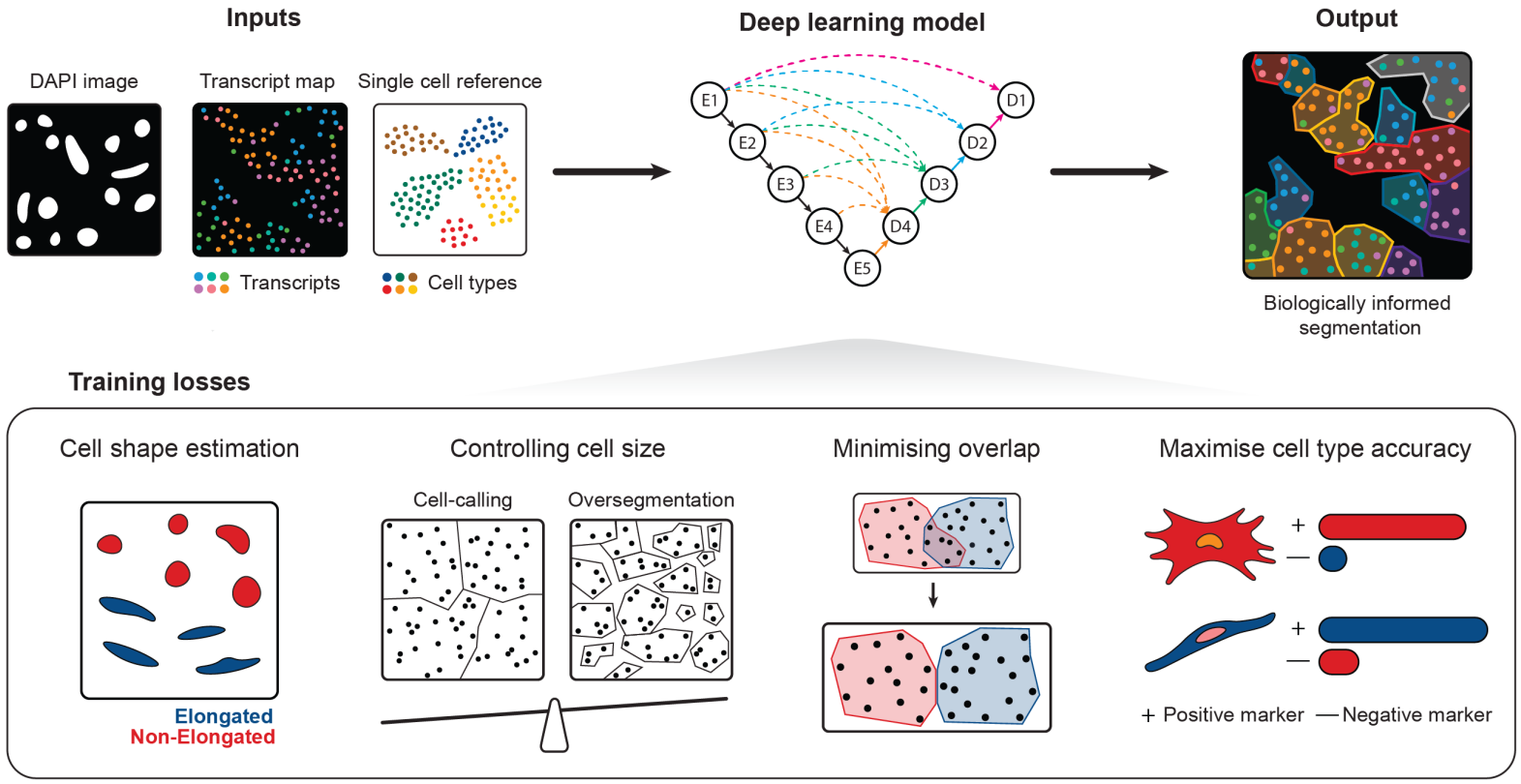


Biologically-informed deep learning-based cell segmentation (BIDCell)

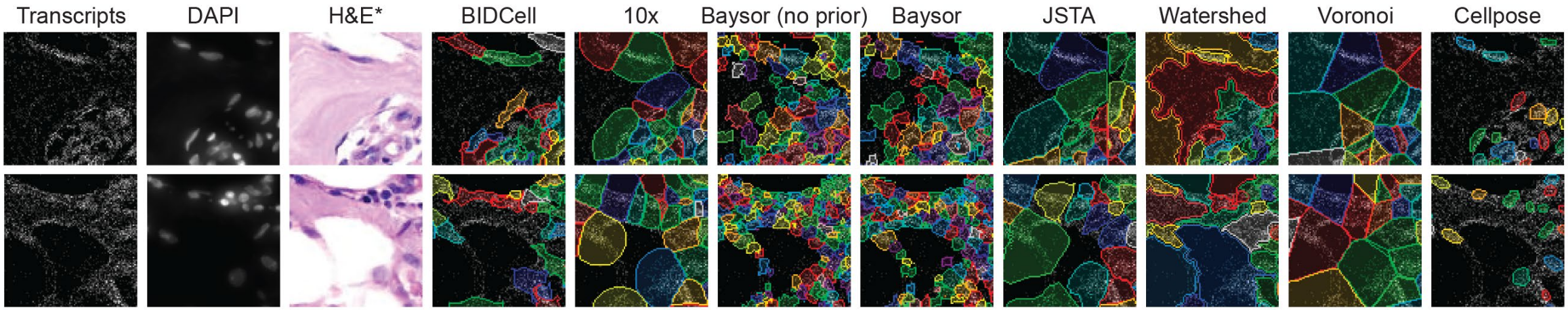
Self-supervised – No ground truth required

Unique loss functions:

1. Allow for varied cell shapes
2. Control over (small) and under (big) segmentation
3. Informed by marker genes from public databases



Example segmentations



Practical considerations

- Linux system with a **12GB** NVIDIA GTX Titan V GPU, Intel(R) Core(TM) i9-9900K CPU @ 3.60GHz with 16 threads, and 64GB RAM

- Xenium-BreastCancer1, with 34.4 million transcripts and 313 genes:

  - Training: 10 minutes
  - Inference: 50 minutes
  - Postprocessing: 30 minutes
  - **Total: 1.5 hours**

- **No tuning of loss function weights:** self-learning, adaptive

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

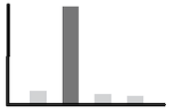
Cell morphology metrics
Elongation; Convexity; ...

Gene-level QC metrics
% of cells expressed

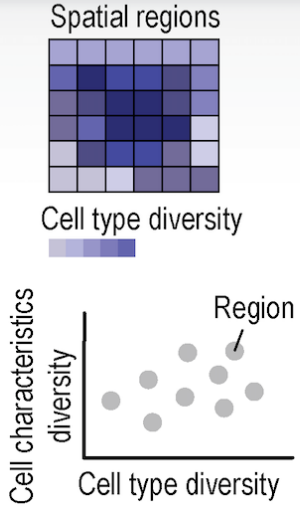
Segmented cell expression purity

Expression concordance with scRNA-seq

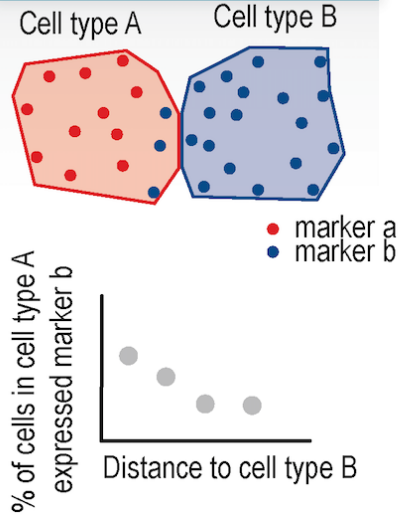
Expression purity in positive/negative markers



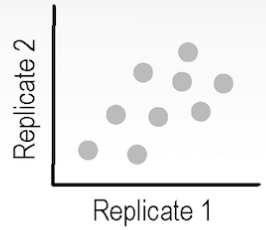
Spatial characteristics diversity



Neighbouring contamination



Replicability



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

QC metrics

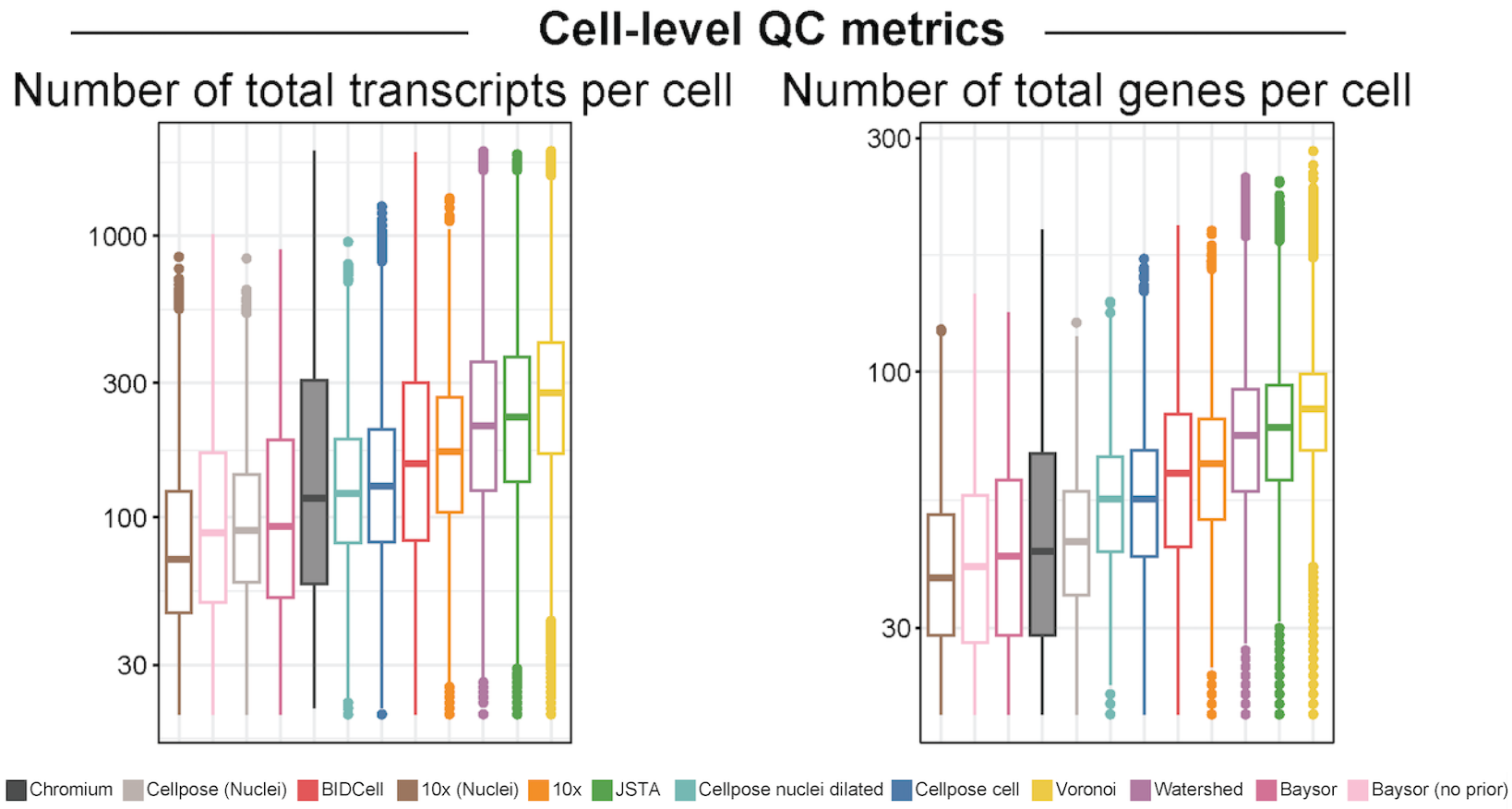
Baseline characteristics

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

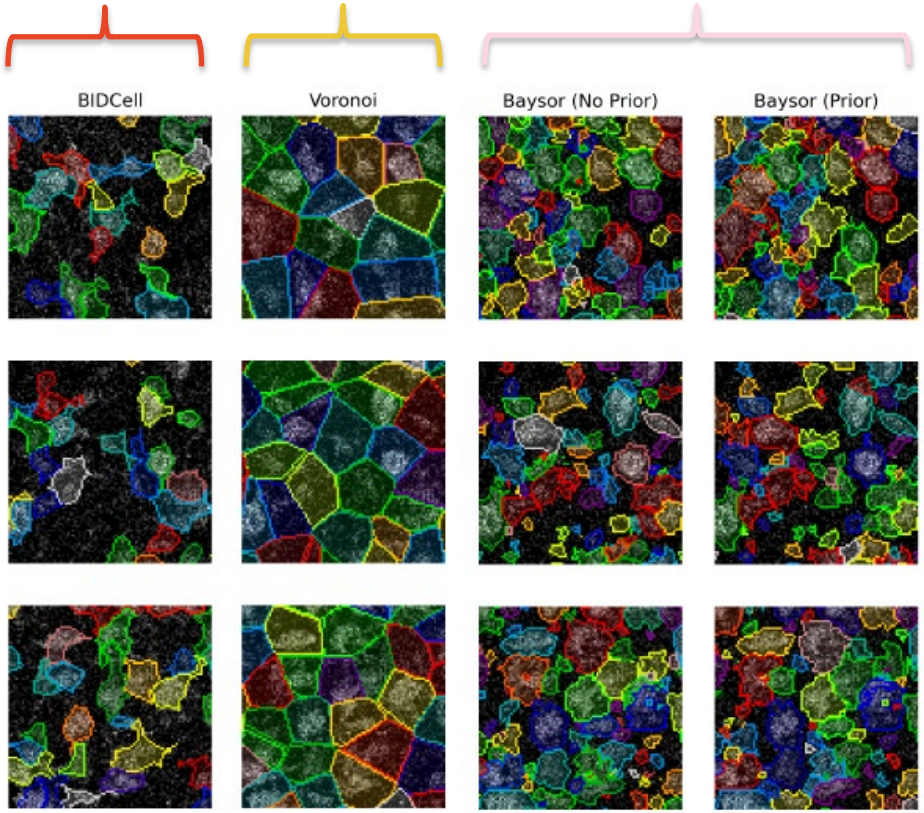
Cell-level QC metrics
Density; Cell area;
# of genes expressed
# of total transcripts

Cell morphology metrics
Elongation; Convexity; ...

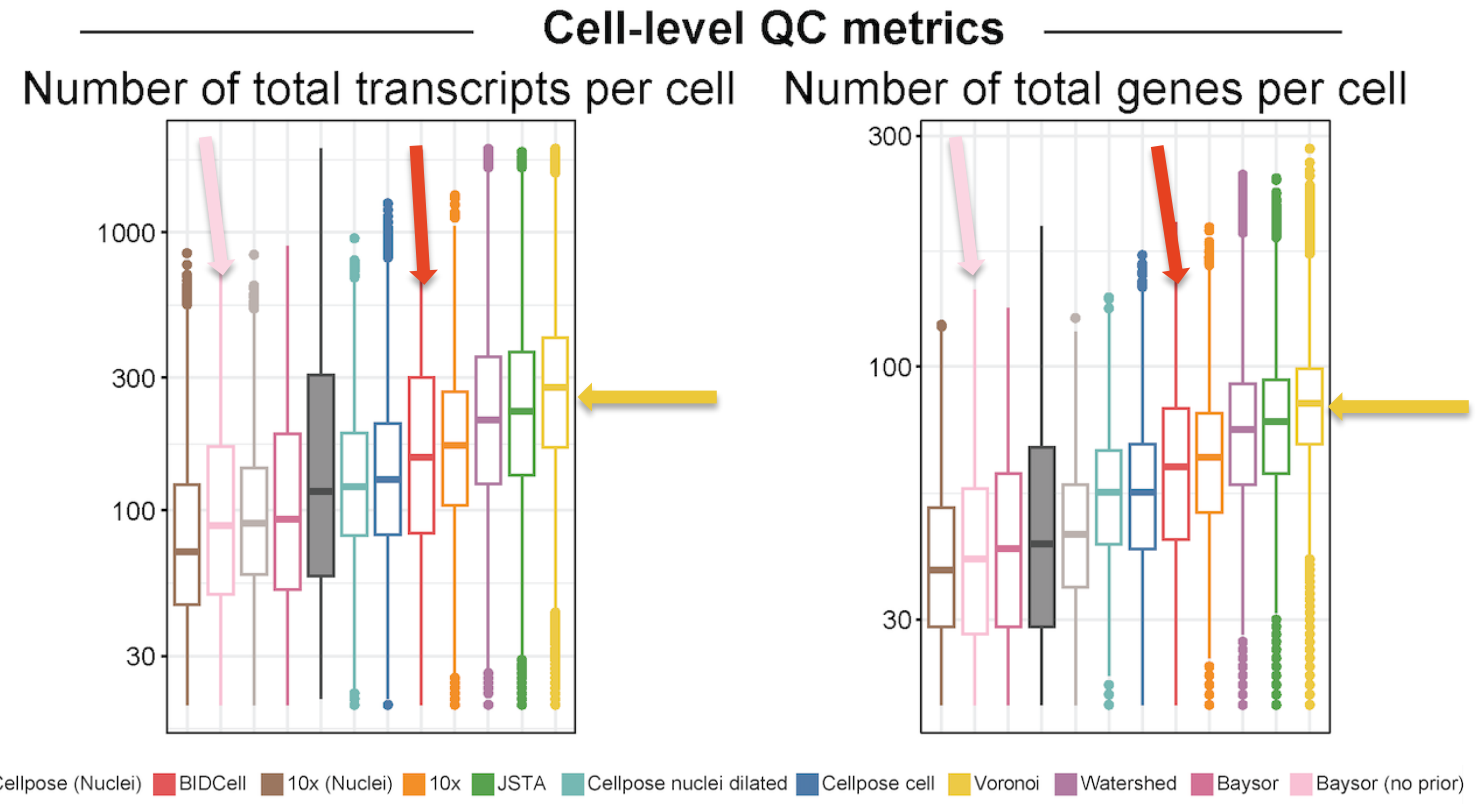
Gene-level QC metrics
% of cells expressed



QC metrics



(b) Xenium-MouseBrain



Whole cell vs. nuclei expression

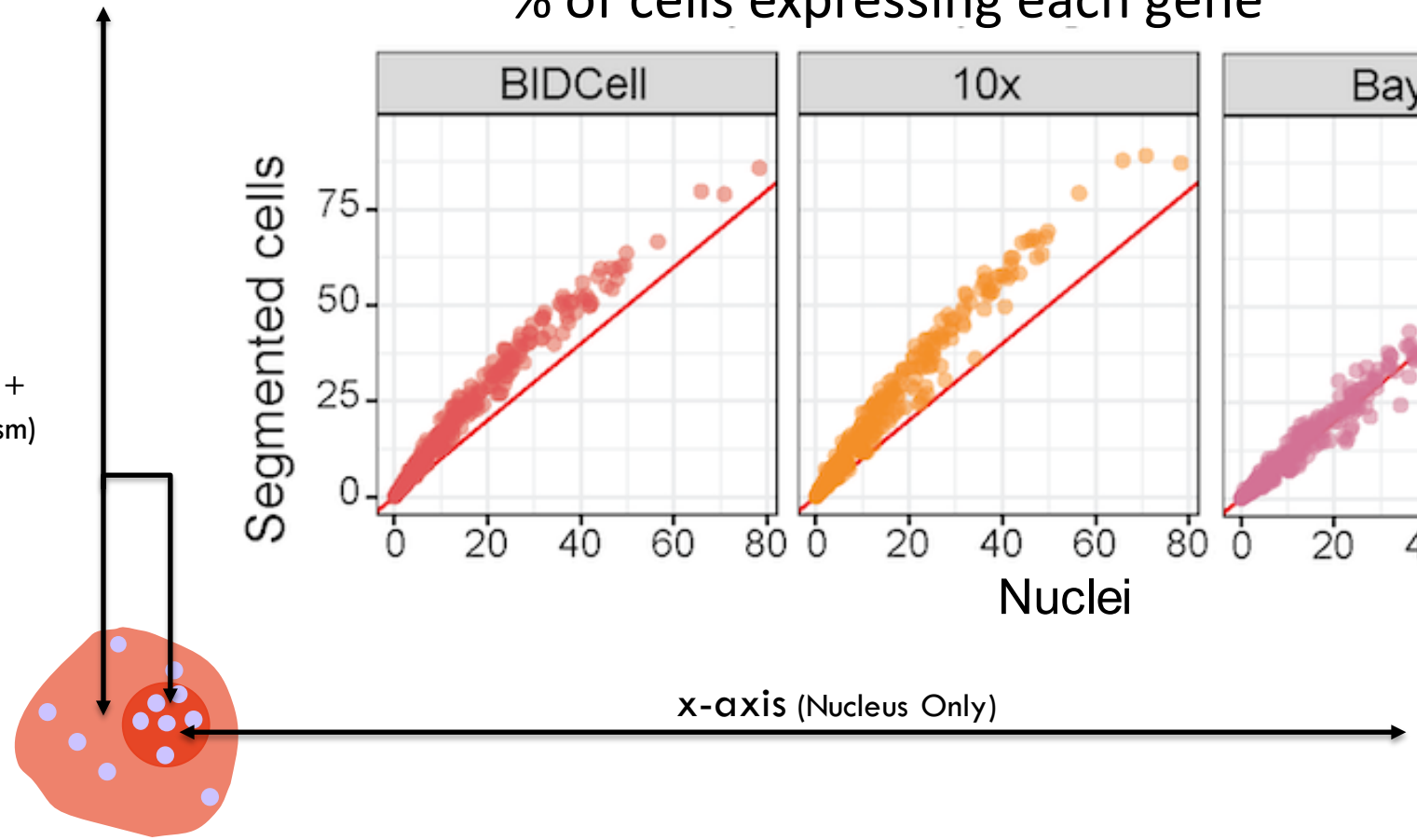
Baseline characteristics

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

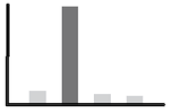


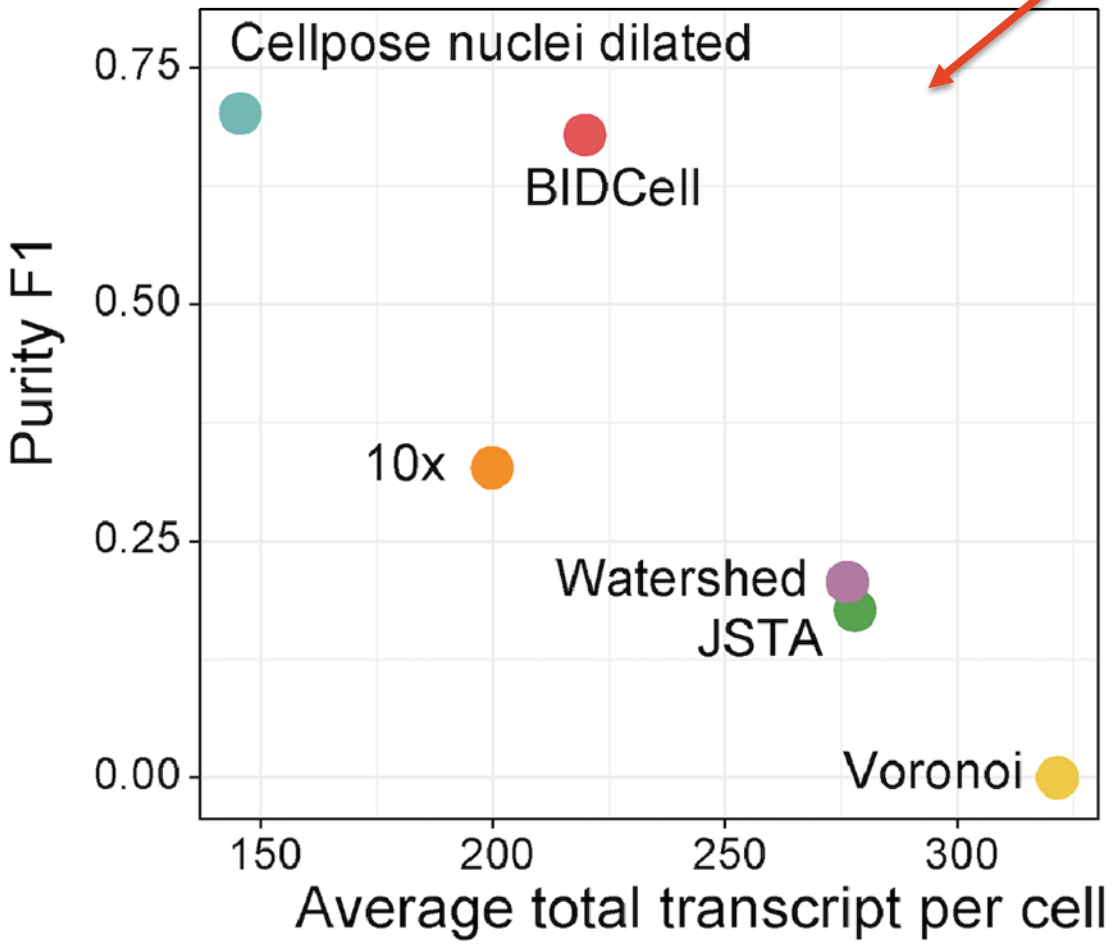
Size vs. purity trade-off

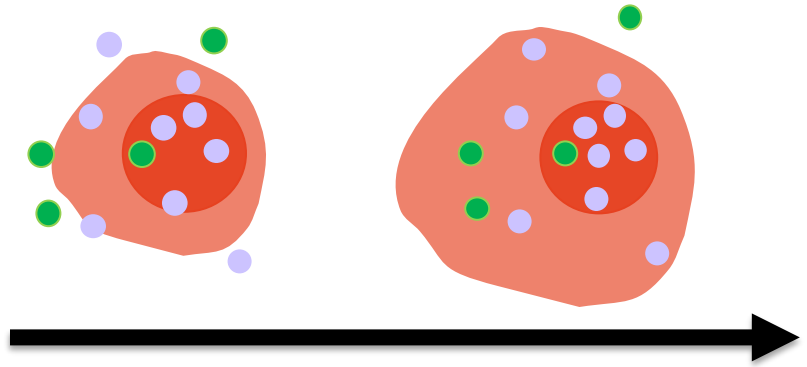
Segmented cell expression purity

Expression concordance with scRNA-seq

Expression purity in positive/negative markers







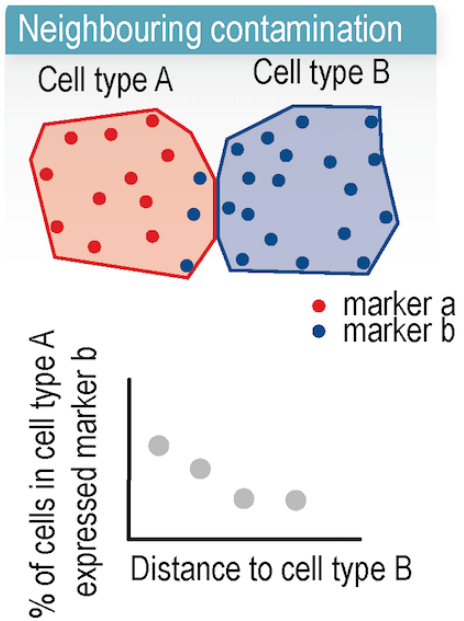
Increase cell size

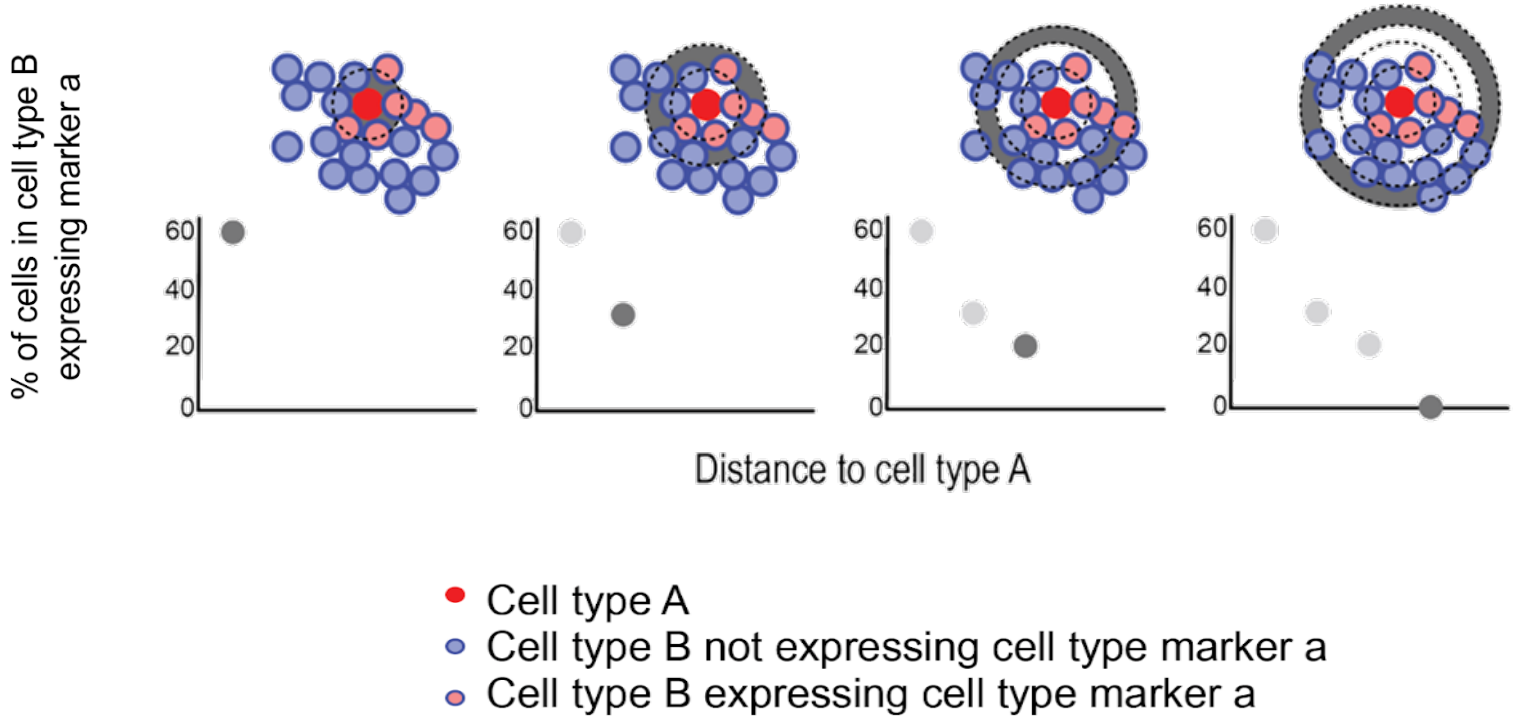
Decrease in purity

Negative Marker

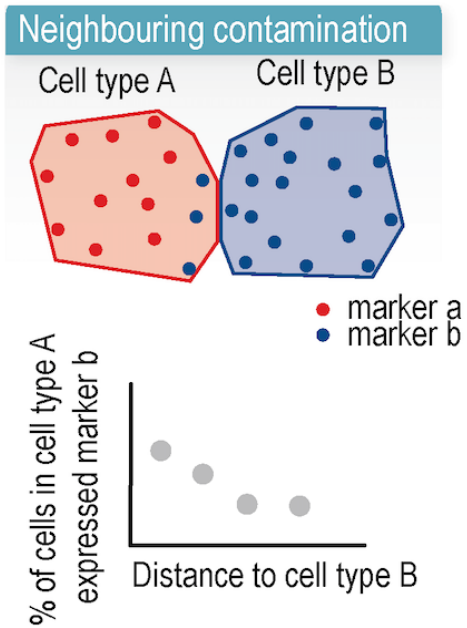
Positive Marker

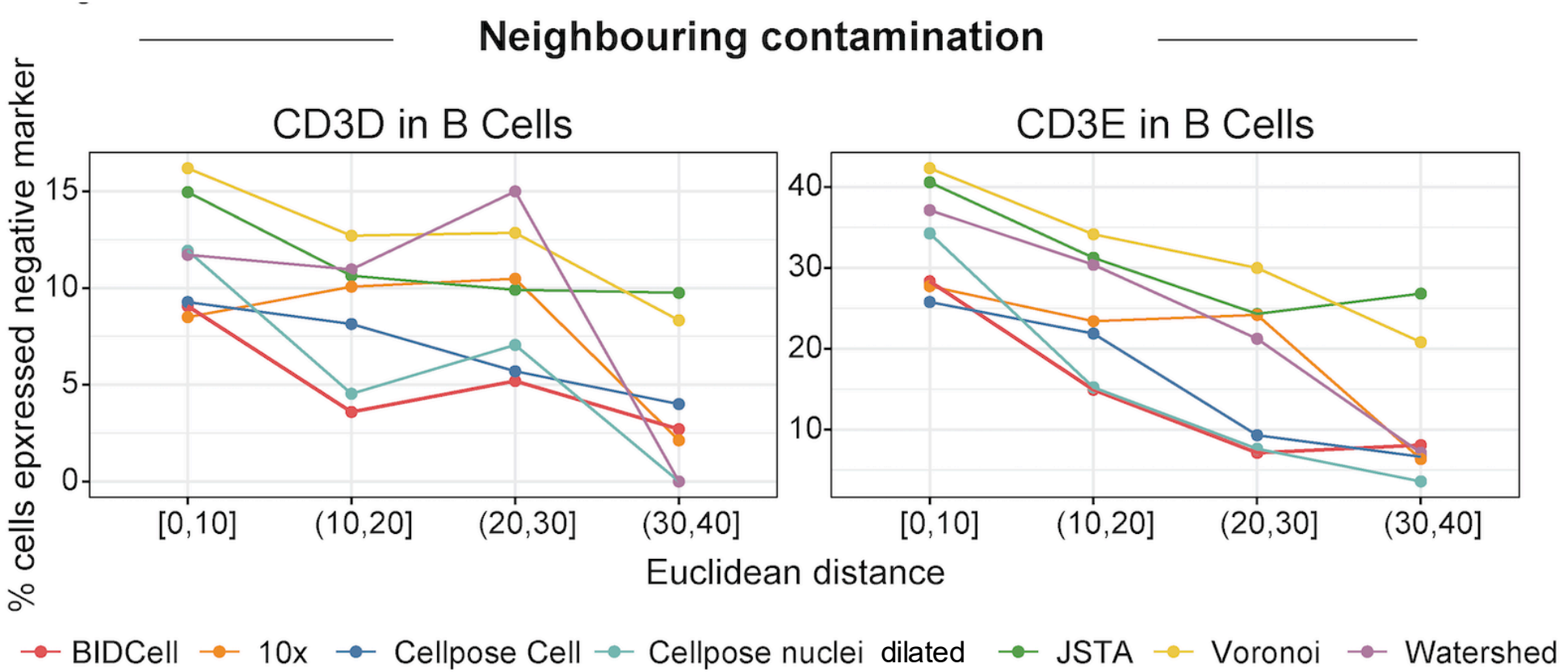
Neighbouring contamination



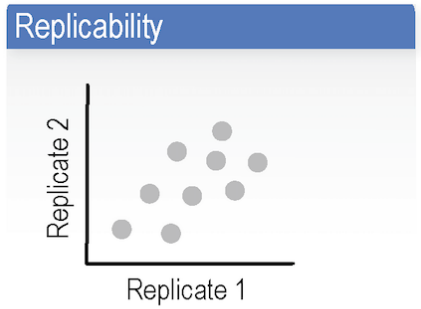


Neighbouring contamination

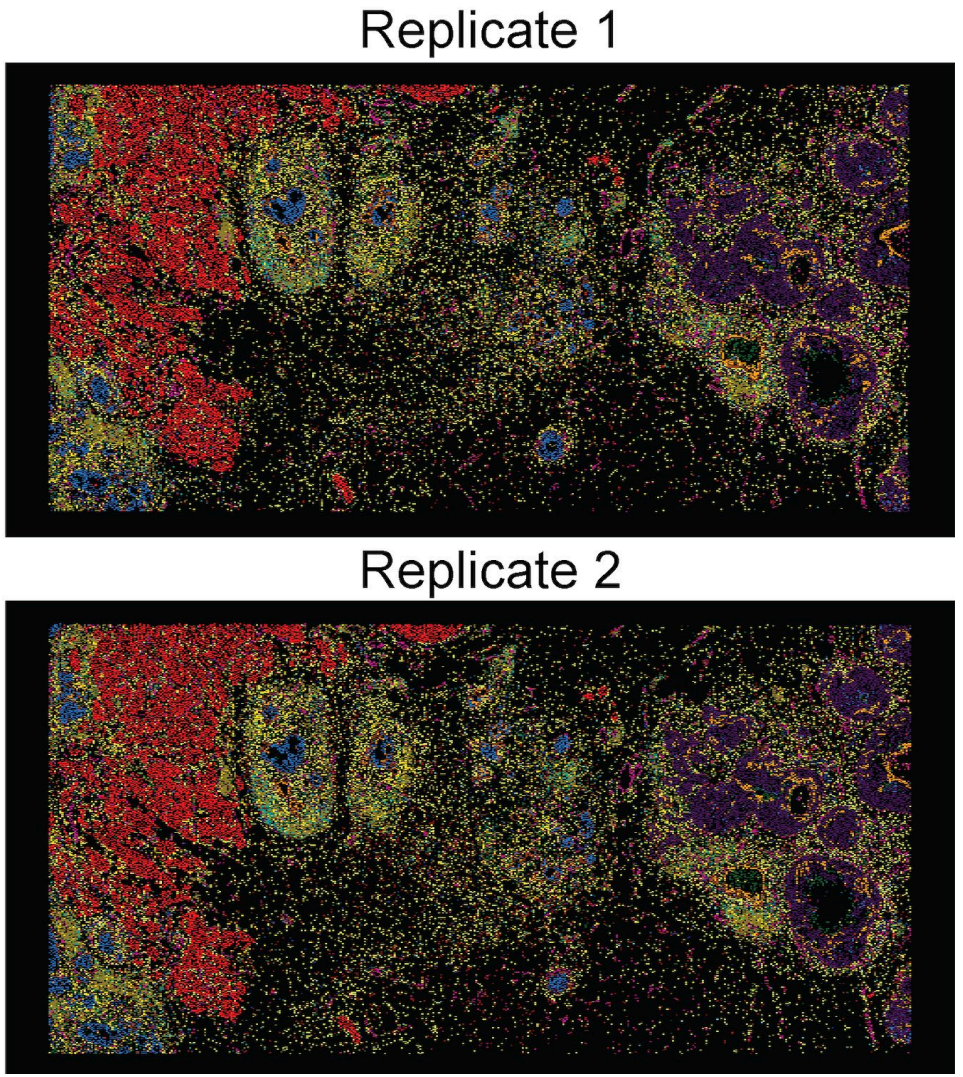


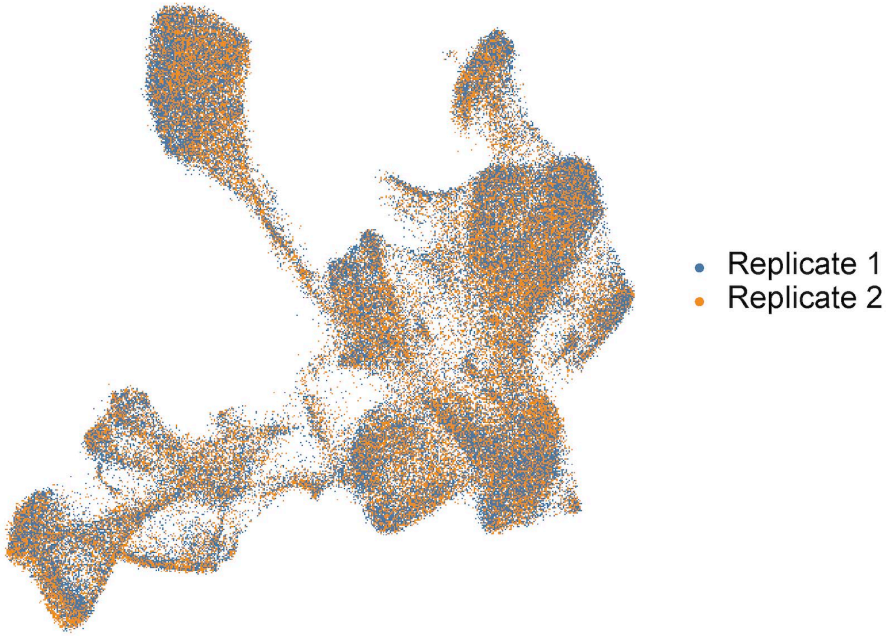


Replicability

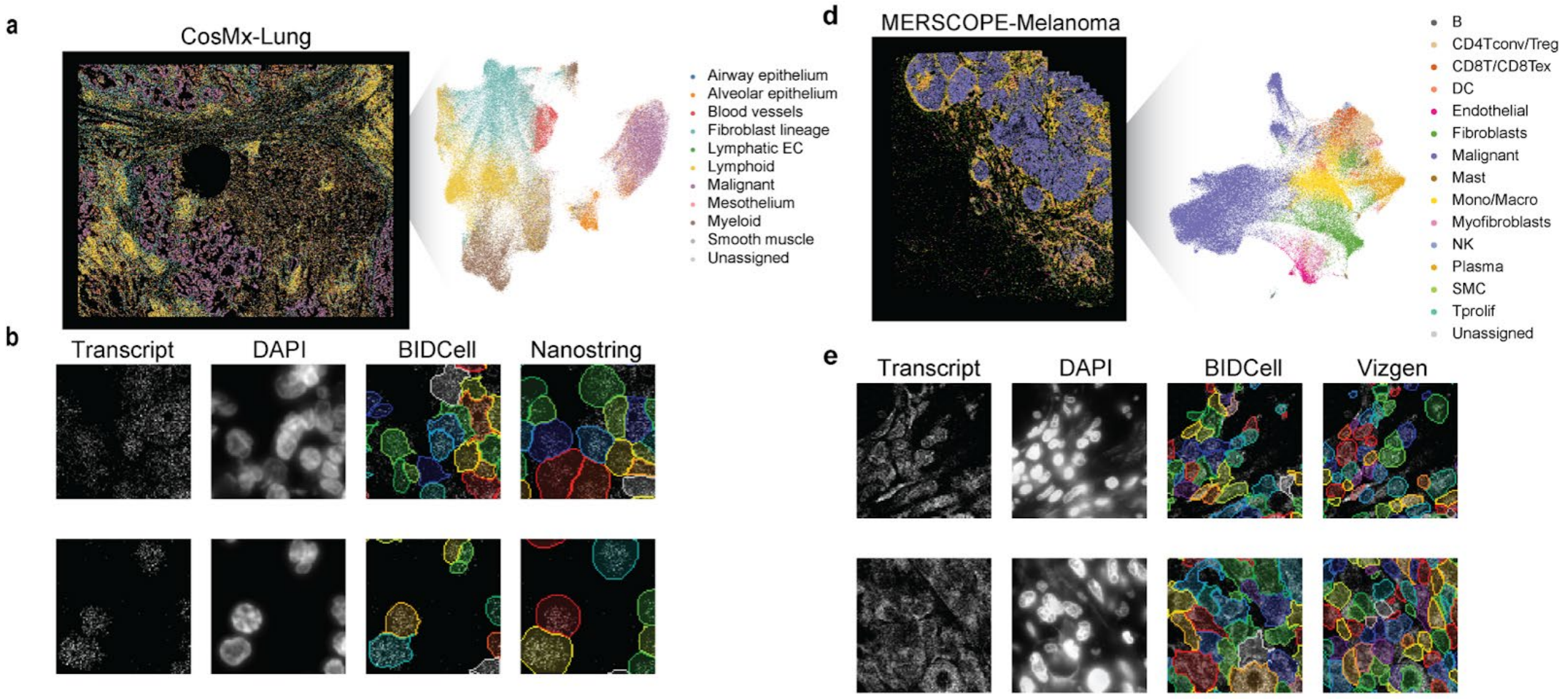


Cell type proportion
Cell-level QC metrics
Gene-level QC metrics
...



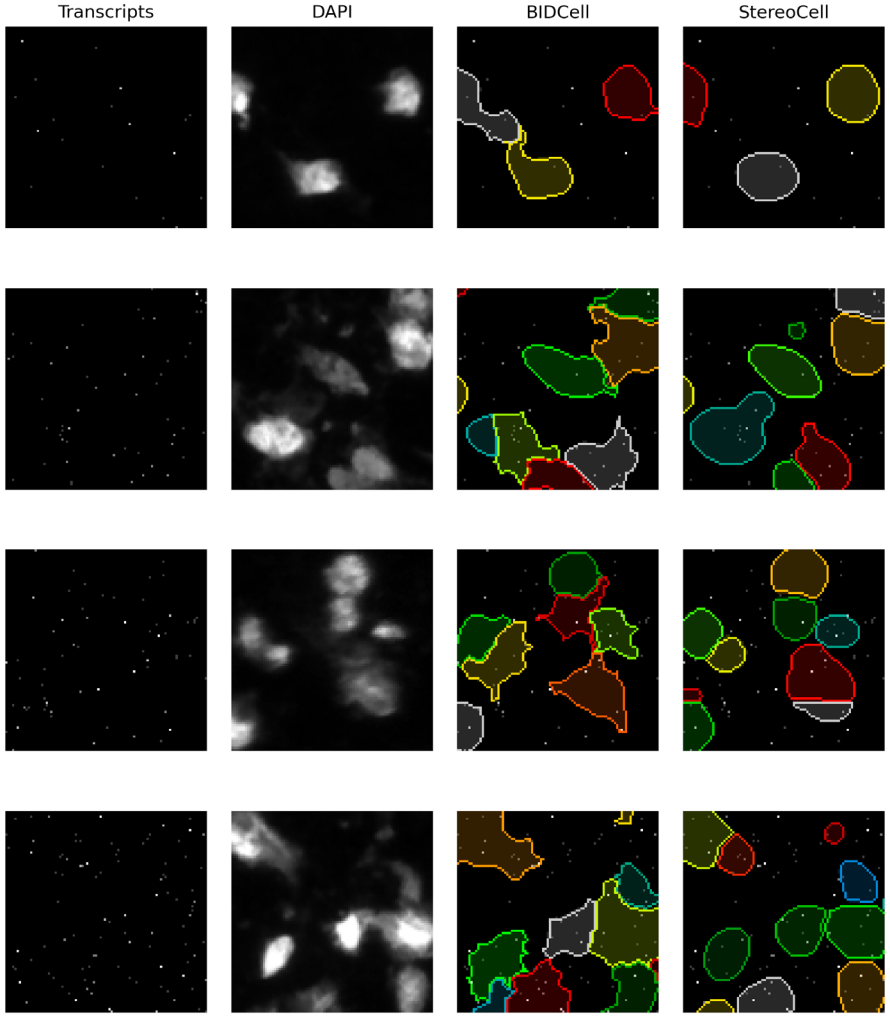


Flexibility



Flexibility

Stereo-seq: Adult mouse brain dataset



Chen, A. et al. Cell, 2022.
Li, M. et al., bioRxiv, 2023.

More details from…

nature communications

Article

https://doi.org/10.1038/s41467-023-44560-w

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

Received: 13 June 2023

Accepted: 13 December 2023

Published online: 13 January 2024

Check for updates

Xiaohang Fu[1,2,3,4,5,8], Yingxin Lin[1,3,4,5,8], David M. Lin[6], Daniel Mechtersheimer[1,3,4], Chuhan Wang[2,3,5], Farhan Ameen[1,3,4], Shila Ghazanfar[1,3,4], Ellis Patrick[1,3,4,5,7], Jinman Kim[2,3,5] & Jean Y. H. Yang[1,3,4,5]

Acknowledgements

Yingxin Lin
Yue Cao
Beilei Bian
Chuhan Wang
Farhan Ameen
Daniel Mechtersheimer
Nick Robertson
Matthew Shu
Hao Wang
Andy Tran
Lijia Yu
Sydney Precision Data Science Centre

Jean Yang
Ellis Patrick
Shila Ghazanfar
Jinman Kim (Computer Science)
Dave Lin (Cornell)
Dinny Graham (Centre for Cancer Research)
Nirmala Pathmanathan (Westmead Breast Cancer Institute)
Agus Salim (University of Melbourne)

THE UNIVERSITY OF SYDNEY

USyd-Cornell Partnership Collaboration Awards

Australian Government
National Health and Medical Research Council
NHMRC

Australian Government
Australian Research Council

D²4H
Laboratory of Data Discovery for Health
醫衞大數據深析實驗室

Contact

- Dr Helen (Xiaohang) Fu [xiaohang.fu@sydney.edu.au]
- Daniel Kim [Daniel.kim2@sydney.edu.au]

Thanks to *USyd-Cornell Partnership Collaboration Awards*

We would love to hear your feedback!

QR code for workshop feedback:

Find out more about scdney:
https://sydneybioxgithub.io/scdney/

THE UNIVERSITY OF SYDNEY