Unlocking single cell spatial omics analyses with SCDNEY

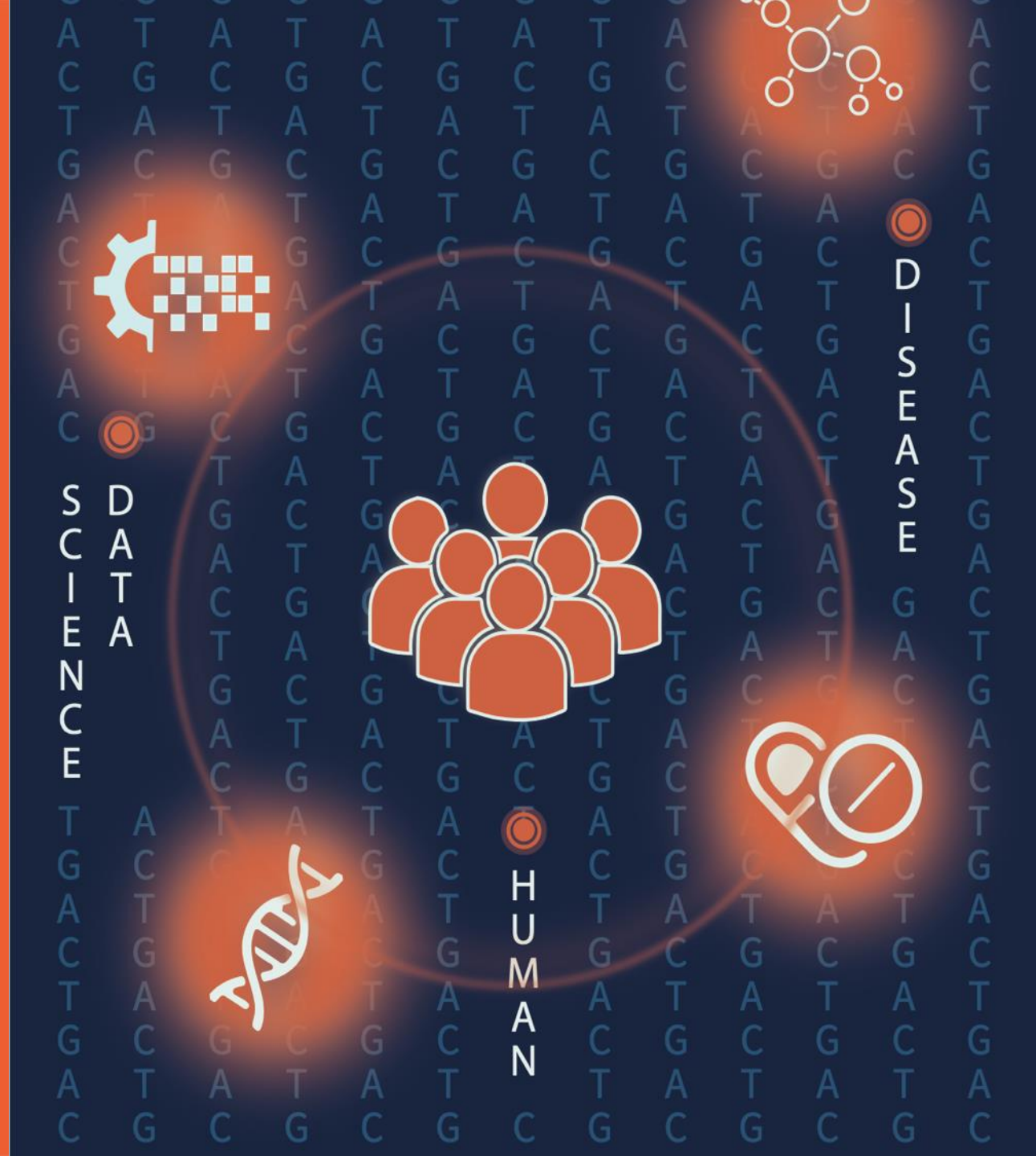
Yue Cao [yue.cao@sydney.edu.au]

Nick Robertson [nicholas.robertson@sydney.edu.au]

Andy Tran [andy.t@sydney.edu.au]

Sydney Precision Data Science Centre
Charles Perkins Centre
School of Mathematics and Statistics

THE UNIVERSITY OF SYDNEY



Workshop presenters

Presenters:

- Dr Yue Cao
- Mr Nick Robertson
- Mr Andy Tran
- Dr Helen Fu
- Dr Ellis Patrick
- Prof Jean Yang

Other Contributors:

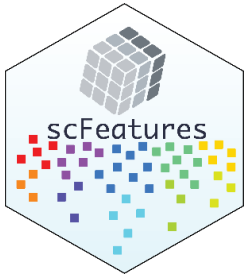
- Dr Dario Strbenac
- Mr Farhan Ameen
- Mr Alex Qin

Sydney Precision Data Science Centre

Extracting insight from the data deluge

Roadmap for the workshop

- Part I: 14:30 – 14:45 Introduction
- Part II: 14:45– 15:20 Exploring spatial data
- Part III: 15:20 – 16:10 Feature engineering with scFeatures
- Break: 16:10 – 16:35
- Part IV: 16:35– 17:10 Survival analysis with ClassifyR
- Part V: 17:10 -17:30 Identify cohort heterogeneity



Configuring Google Cloud

1. Obtain login details by scanning this QR code:

2. Type the machine ID into your browser to get into google cloud

Machine 1: xxxxxxx

Machine 2: xxxxxxxx

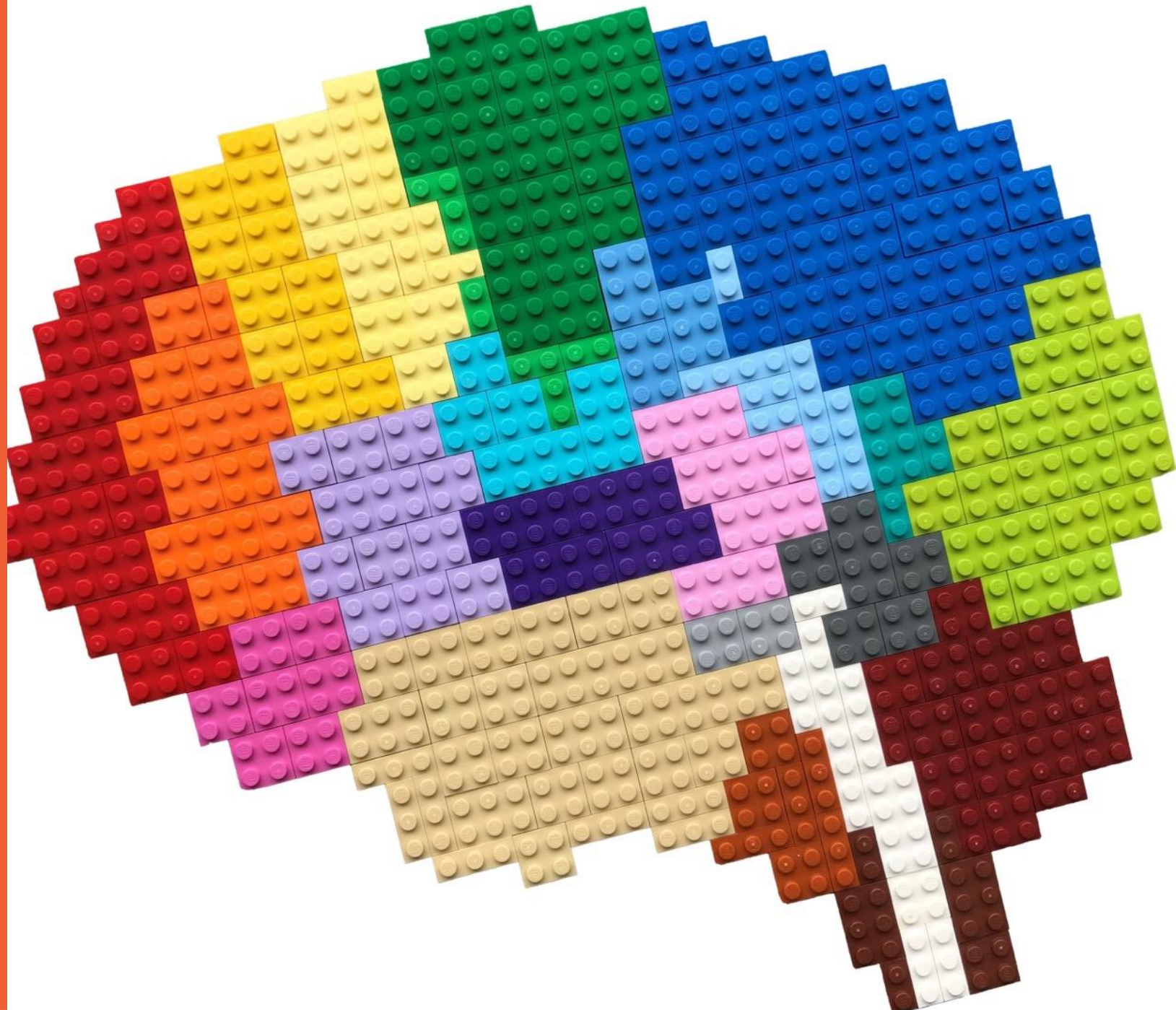
3. Log in with the username and password
4. In RStudio, type the following in the terminal:

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

Materials also on github: https://github.com/SydneyBioX/BiocAsia2023

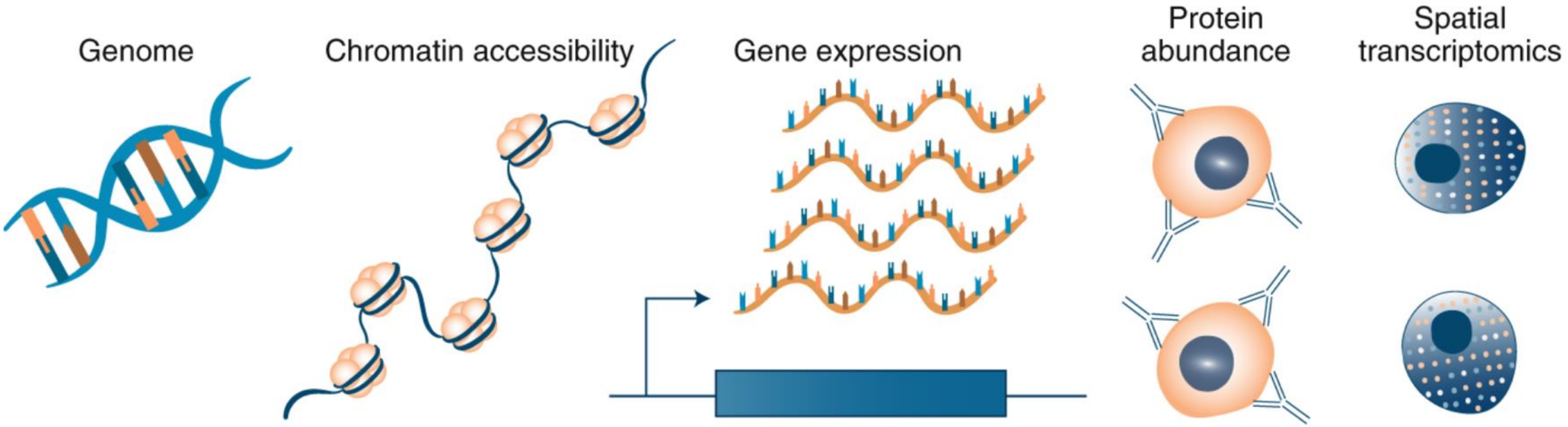
PART I:

Introduction



THE UNIVERSITY OF SYDNEY

Single-cell technologies measure different aspects of cells



- SNS
- SCI-seq

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

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

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

- MERFISH
- smFISH
- STARmap

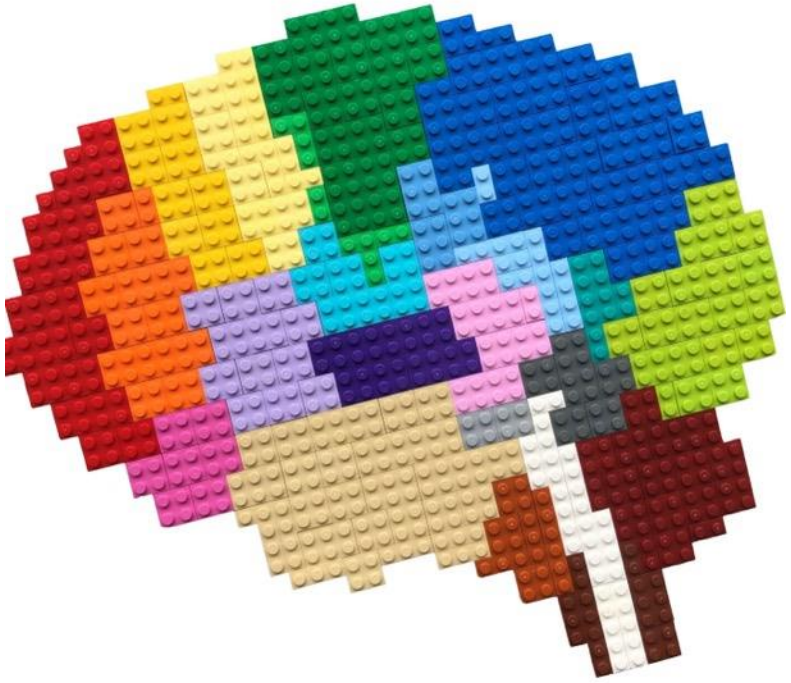
Efremova, M., Teichmann, S.A. Computational methods for single-cell omics across modalities. *Nat Methods* **17**, 14–17 (2020).

Spatially resolved technologies

Bulk

Single Cell

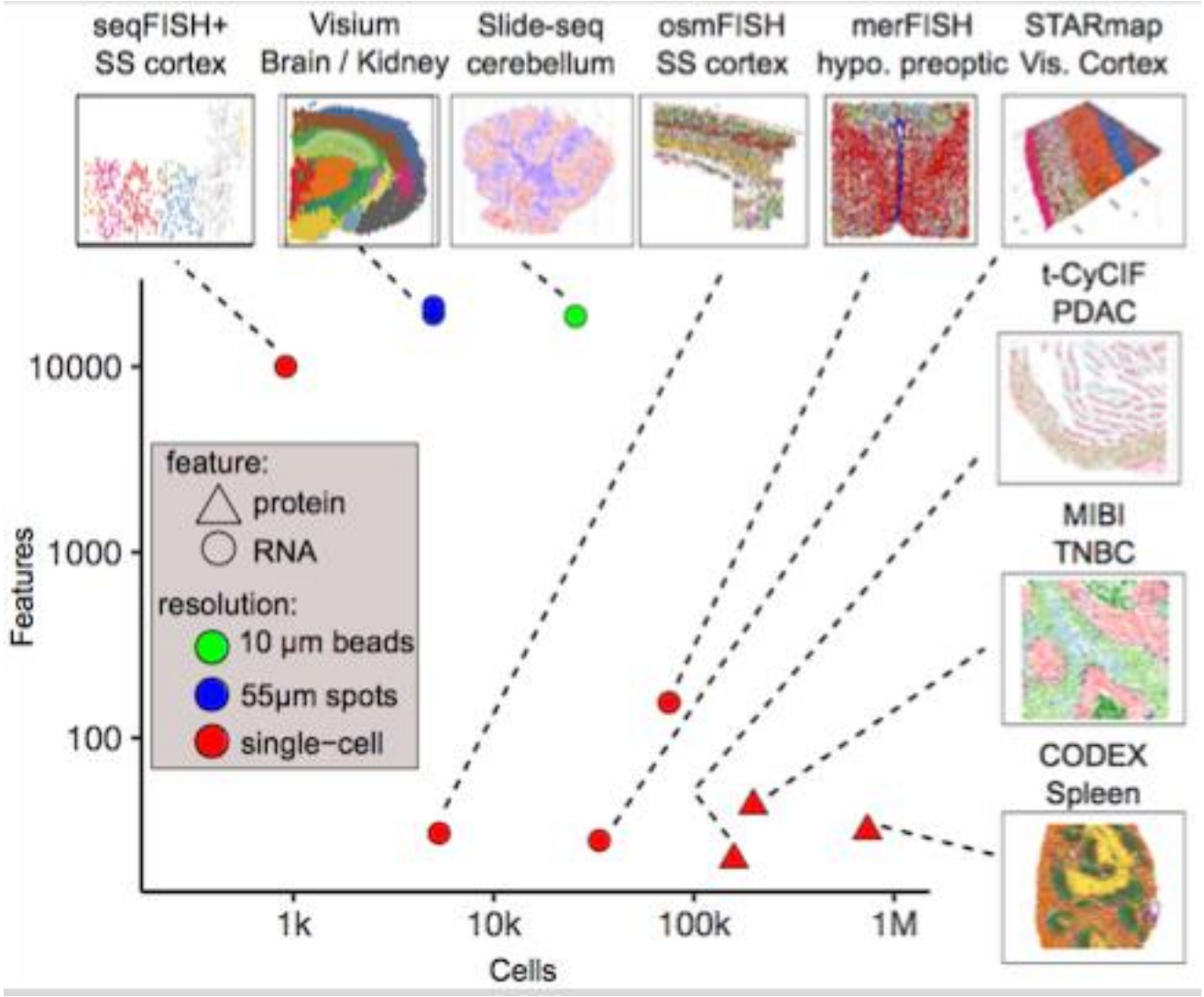
Spatial



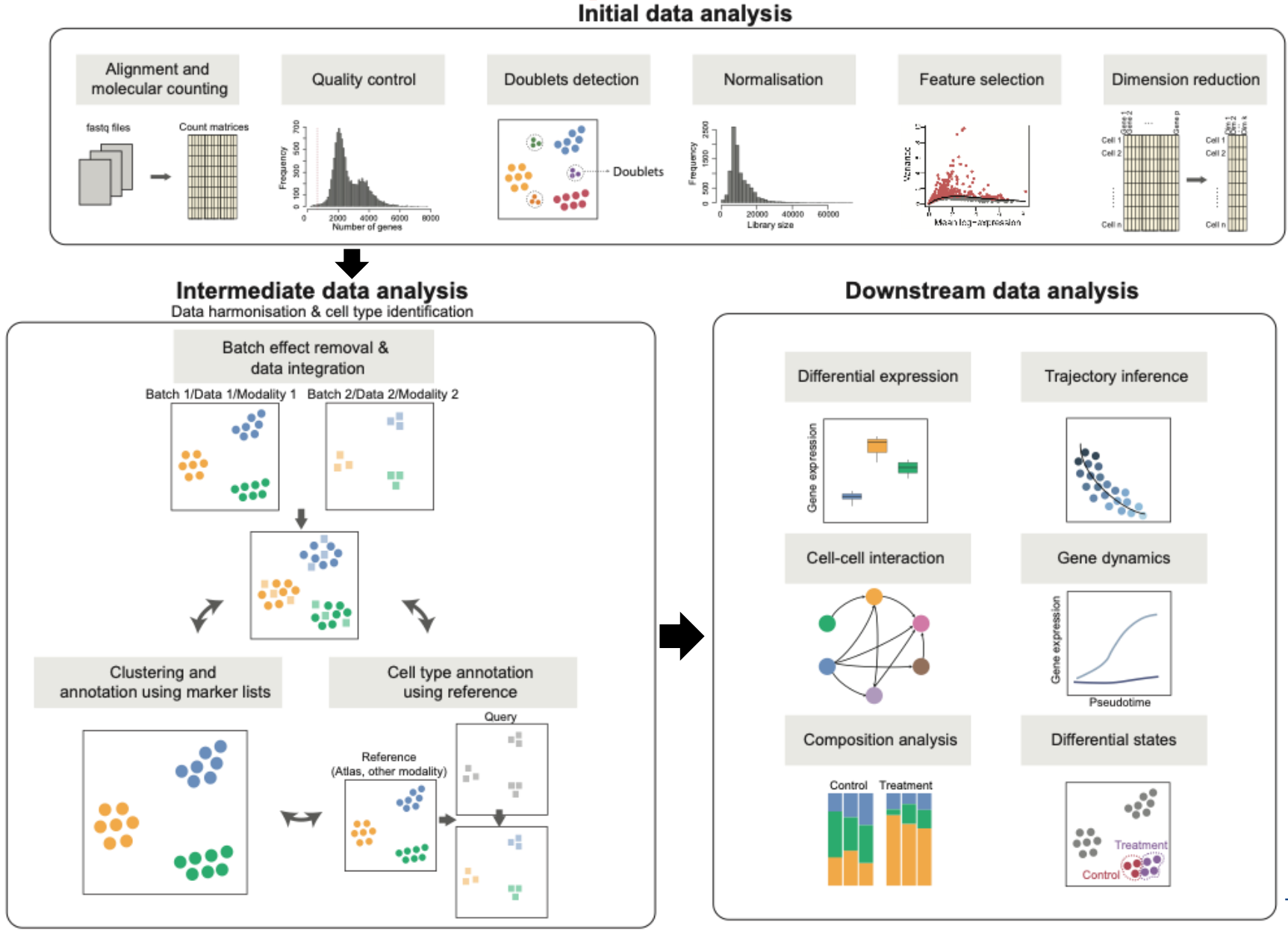
Credit: Bo Xia

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

Different types of spatial technology



Overview of single-cell data analysis



Thanks to Yingxin Lin

Single-cell methods @ Sydney

Single cell data iNtEgrative analYsis

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

scdney 0.2.0 | Packages | Publications

Publications

Methodology and tools

Preprint

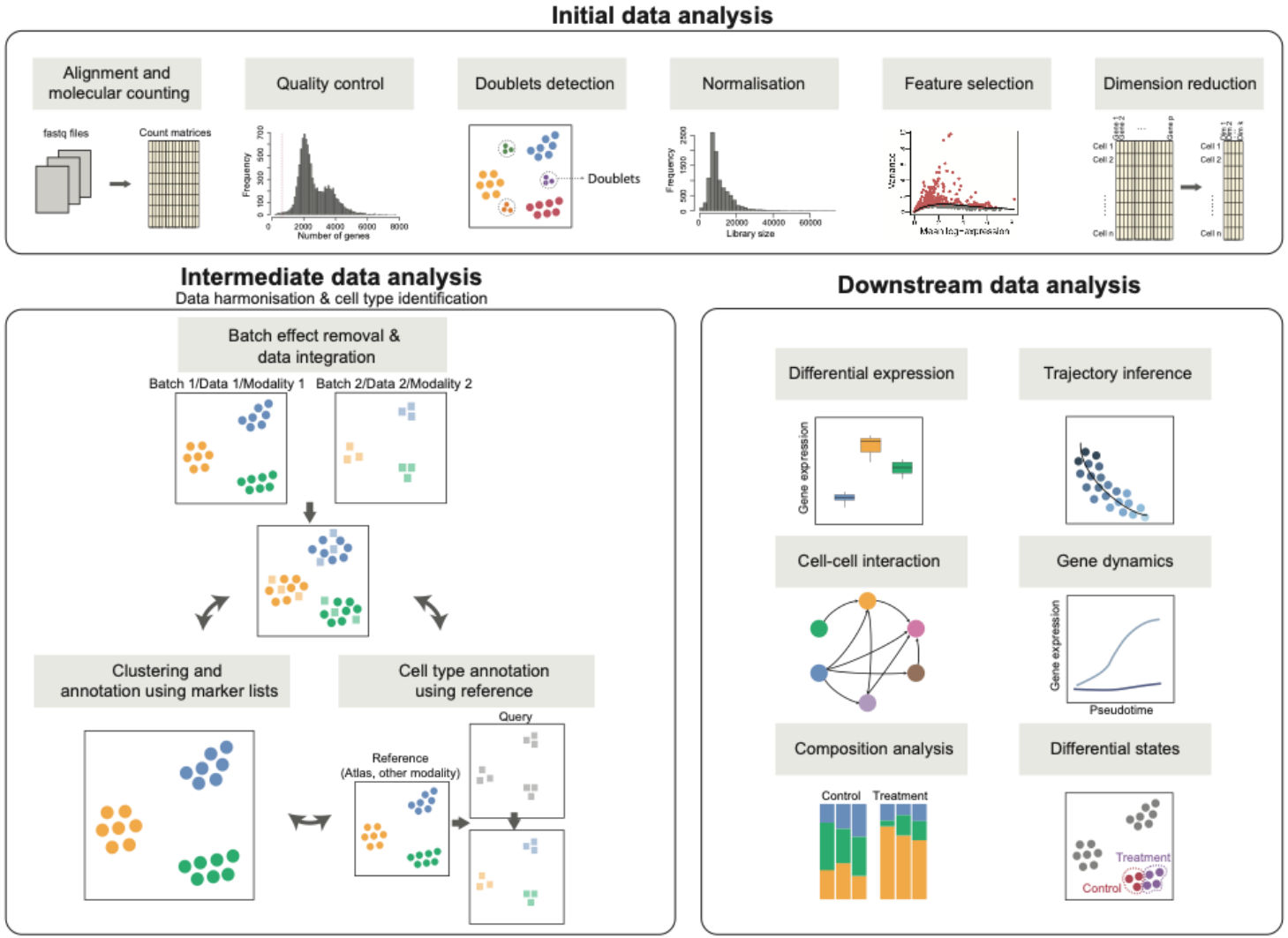
- **Patrick, E., Canete, N. P., Baharlou, H., Iyengar, S. S.,** Harman, A. N., Sutherland, G. T., and **Yang P.** (2021) Spatial analysis for highly multiplexed imaging data to identify tissue microenvironments. *Biorxiv*. [package][shiny][paper]
- **Lin, Y.**, Loo, L., **Tran, A.**, Moreno, C., Hesselson D., Neely, G., and **Yang, J.Y.H.** (2020) Characterization of cell-cell communication in COVID-19 patients. *Biorxiv*. [paper]

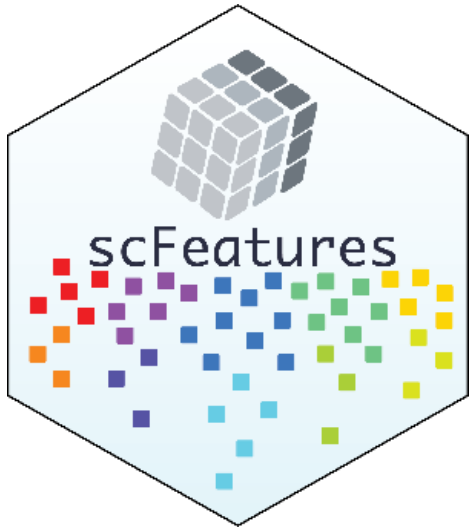
2022

- **Canete, N. P., Iyengar, S. S.**, Wilmott. J. S., **Ormerod, J. T.**, Harman, A. N., and **Patrick, E.**(2021) spicyR: Spatial analysis of in situ cytometry data in R. *Bioinformatics*. [package][shiny][paper]
- **Cao, Y., Lin, Y., Patrick, E., Yang, P., and Yang, J.Y.H.** (2022) scFeatures: Multi-view representations of single-cell and spatial data for disease outcome prediction. *Bioinformatics*. [paper]
- **Lin, Y.**, Wu, T.Y., Wan, S., **Yang, J.Y.H.,** Wong, W.H., and **Wang, Y.X.R.** (2022) scJoint: transfer learning for data integration of single-cell RNA-seq and ATAC-seq. *Nature Biotechnology*. [package][paper]
- **Tran, A., Yang P., Yang, J.Y.H., and Ormerod, J. T.** (2022) scREMOTE: Using multimodal single cell data to predict regulatory gene relationships and to build a computational cell reprogramming model. *NAR Genomics and Bioinformaitics*. [package][paper]

https://sydneybiox.github.io/scdney/

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
Dimension reduction
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


ClassifyR

Interactive Q&A

- In this workshop we will have **interactive question and answer** via Menti

Go to

www.menti.com

Enter the code

8474 3969

Or use QR code

Q0A – experience

How confident are you with **R** and **single cell/spatial analysis?**

Go to

www.menti.com

Enter the code

8474 3969

Or use QR code

Q0B – data

What type of spatial data are you likely to analyse in the future?

Go to

www.menti.com

Enter the code

8474 3969

Or use QR code

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

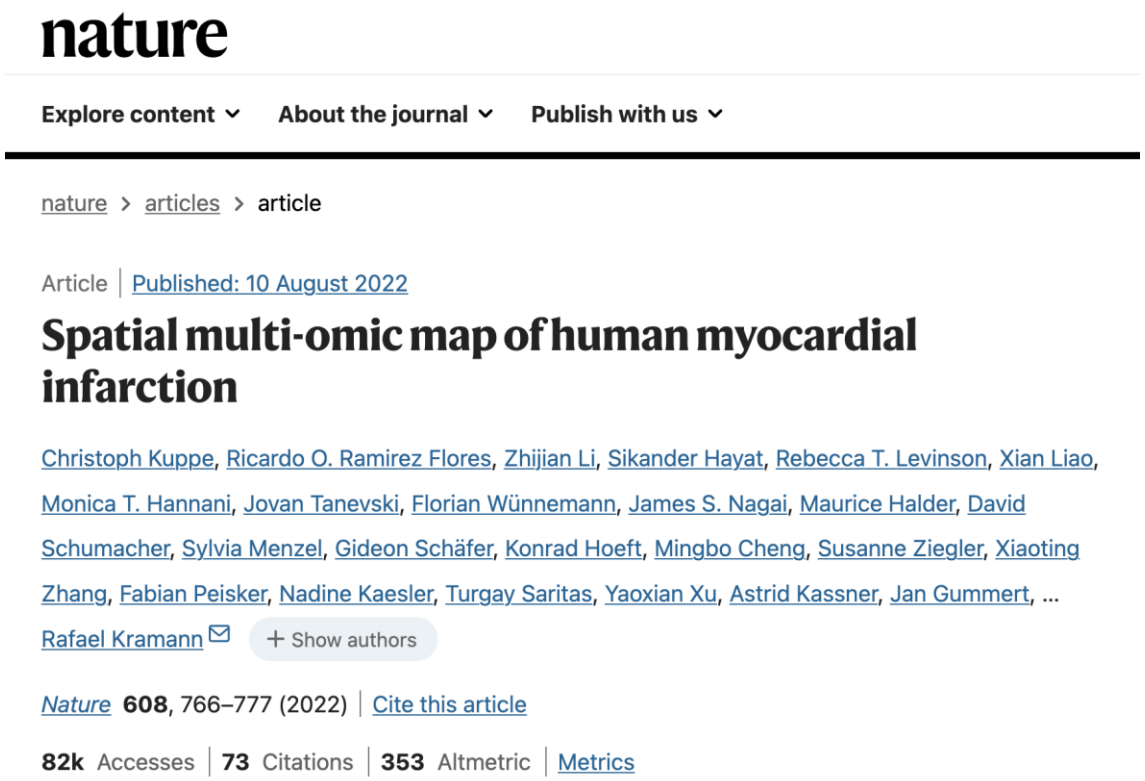
- Single-cell resolution
- Measures 37 proteins

- Full dataset contains 483 patients
- In this workshop we look at a subset 77 patients with no lymph node metastasis.

- Outcome of interest is to estimate risk of recurrence

Data – 10X Visium [optional case study]

- Spot based technology
- > 10,000 genes



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Published: 10 August 2022

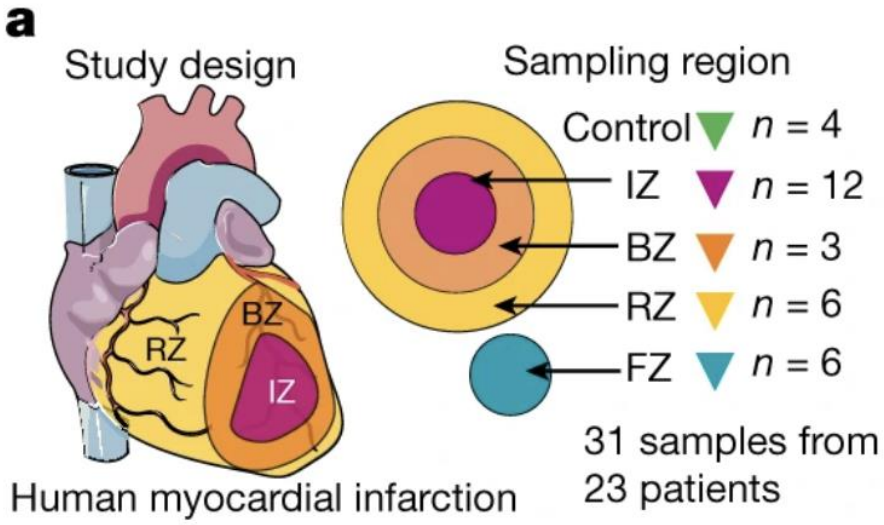
Spatial multi-omic map of human myocardial infarction

Christoph Kuppe, Ricardo O. Ramirez Flores, Zhijian Li, Sikander Hayat, Rebecca T. Levinson, Xian Liao, Monica T. Hannani, Jovan Tanevski, Florian Wünnemann, James S. Nagai, Maurice Halder, David Schumacher, Sylvia Menzel, Gideon Schäfer, Konrad Hoeft, Mingbo Cheng, Susanne Ziegler, Xiaoting Zhang, Fabian Peisker, Nadine Kaesler, Turgay Saritas, Yaoxian Xu, Astrid Kassner, Jan Gummert, ... Rafael Kramann + Show authors

Nature **608**, 766–777 (2022) | Cite this article

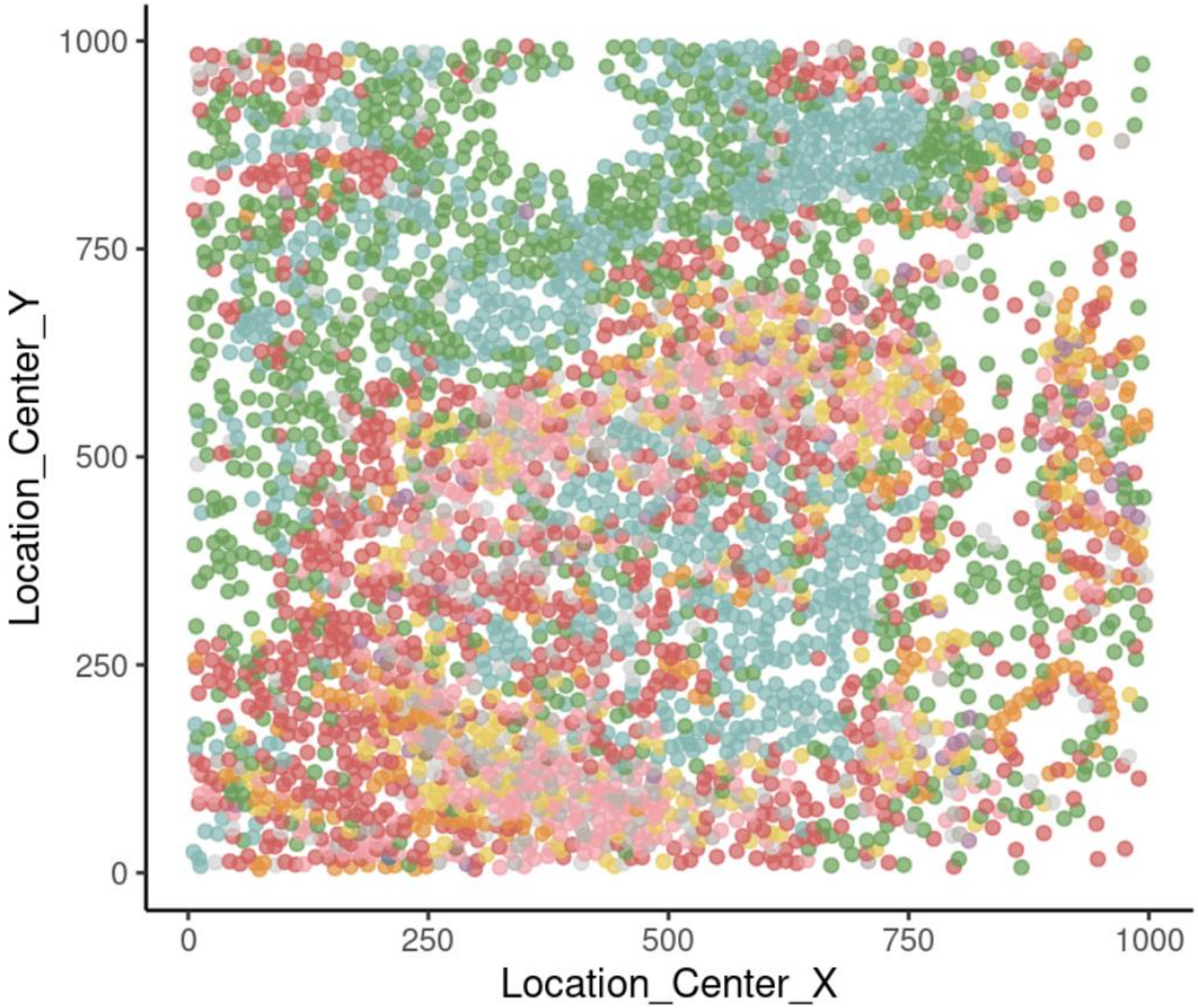
82k Accesses | **73** Citations | **353** Altmetric | Metrics

- Full dataset contains 31 samples
- This workshops compares myogenic (control and broader zone samples) and ischaemic samples (ischaemic zone), a total of 19 samples
- Outcome of interest is to predict myogenic versus iscahemic



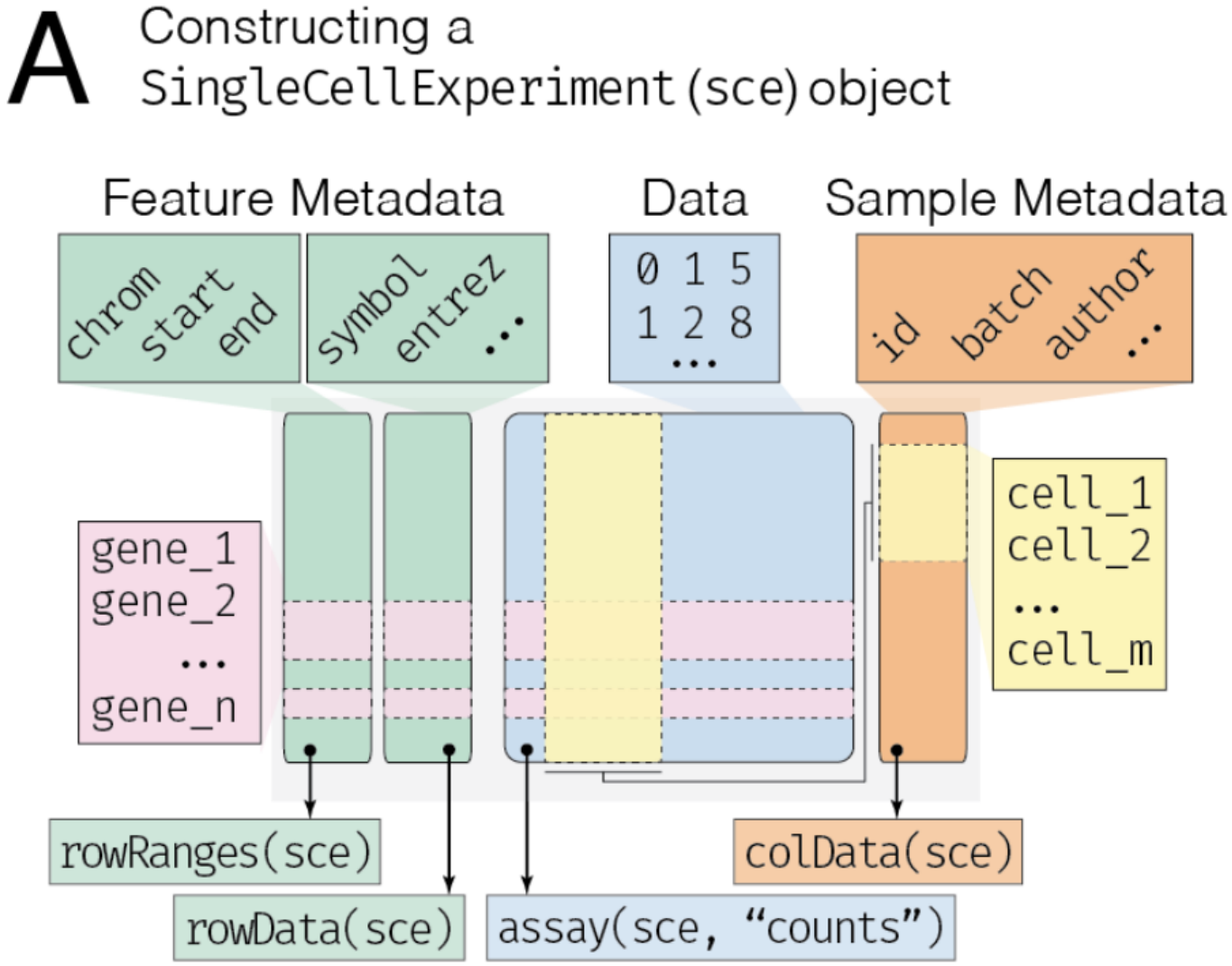
PART I:

Initial data analysis

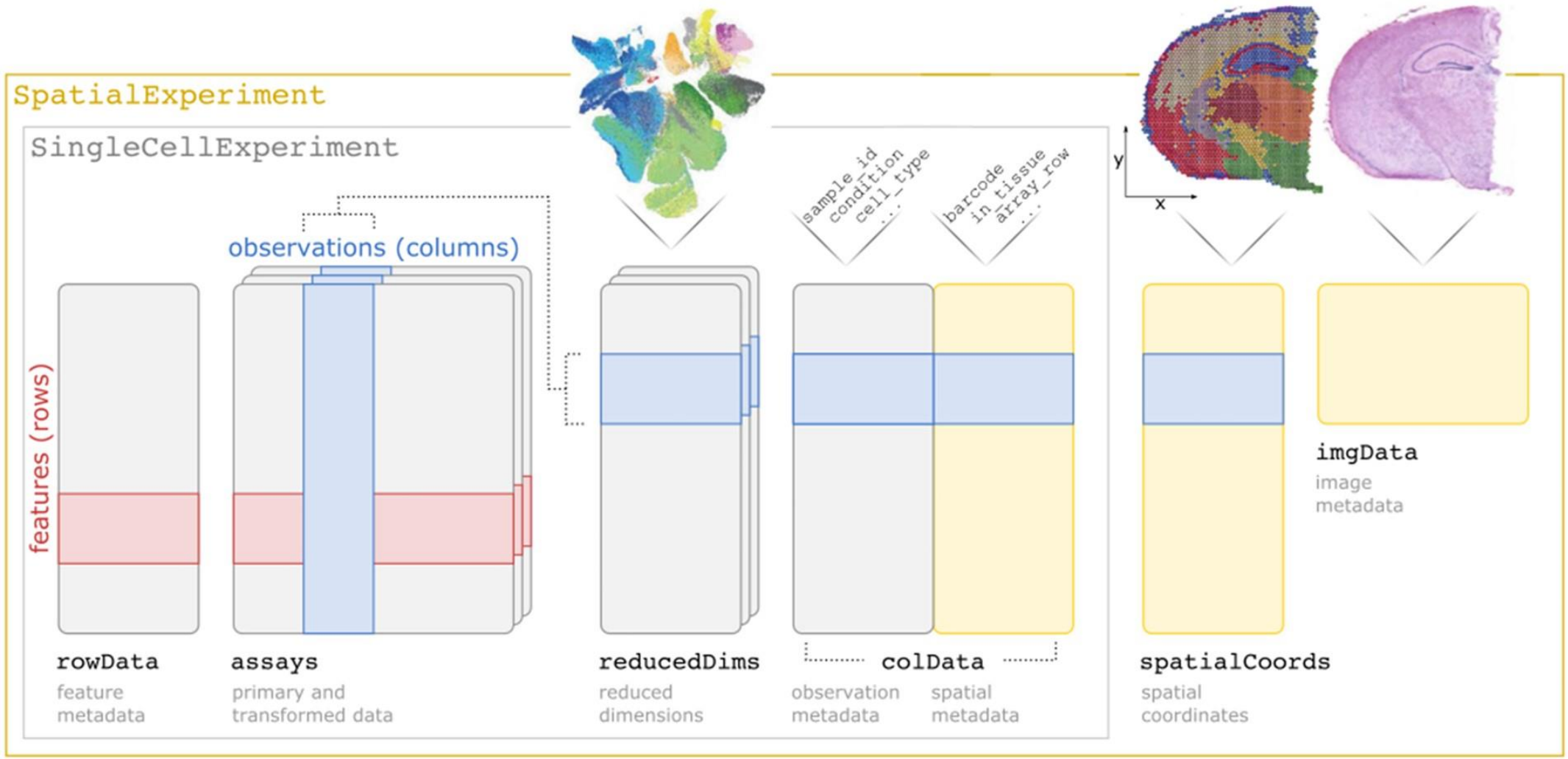


THE UNIVERSITY OF SYDNEY

How to represent single cell data?



How to represent processed spatial omics data?



Righelli et al, Bioinformatics (2022)

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

Data

Data is stored as `SingleCellExperiment` S4 object

```
## Quick look at data
data_sce
```

```
class: SingleCellExperiment
dim: 38 76307
metadata(0):
assays(2): counts logcounts
rownames(38): HH3_total CK19 ... H3K27me3 CK5
rowData names(0):
colnames(76307): MB-0002:345:93 MB-0002:345:107 ... MB-0663:394:577 MB-0663:394:578
colData names(17): file_id metabricId ... x_cord y_cord
reducedDimNames(1): UMAP
mainExpName: NULL
altExpNames(0):
```

Data

The **count matrix** (proteins x cells) can be retrieved using the `logcounts()` function

```{r}
logcounts(data_sce)[1:5, 1:5]
```

	MB-0002:345:93	MB-0002:345:107	MB-0002:345:113	MB-0002:345:114	MB-0002:345:125
HH3_total	2.5410994	2.2699132	1.83124965	2.54153682	1.77142549
CK19	1.1452546	1.2155873	0.18690945	0.41539814	0.09835966
CK8_18	1.9270420	1.7957914	0.68297921	1.77968459	0.27171782
Twist	0.1972571	0.1604647	0.13589493	0.27251430	0.16355599
CD68	0.1092099	0.1396621	0.04580369	0.07365269	0.00000000

Data

- **Metadata** can be retrieved using `colData()`
- The complicated code is just to make the table output look nice and scrollable

```
DT::datatable( data.frame(colData(data_sce))[1:5, ], options = list(scrollX = TRUE))
```

Show 10 entries Search:

	file_id	metabricId	core_id	ImageNumber	ObjectNumber	Fibr
MB-0002:345:93	MB0002_1_345	MB-0002	1	345	93	0
MB-0002:345:107	MB0002_1_345	MB-0002	1	345	107	0.
MB-0002:345:113	MB0002_1_345	MB-0002	1	345	113	0.
MB-0002:345:114	MB0002_1_345	MB-0002	1	345	114	0.
MB-0002:345:125	MB0002_1_345	MB-0002	1	345	125	6

Showing 1 to 5 of 5 entries Previous 1 Next

Data

From the clinical table, we can work out the number of patients in each **stage** and their current **status**

```
print("Stage and death")
table(clinical$Breast.Surgery, clinical$Death, useNA = "ifany")
```

```
[1] "Stage and death"

                     0  1
  BREAST CONSERVING 38 14
  MASTECTOMY        16  6
  <NA>               2  1
```

Data

The number of patients based on **ER status**

```
table(clinical$ER.Status, useNA = "ifany")
```

```
[1] "Number of patients based on ER status"

 neg  pos <NA>
  12   64    1
```

Data

The number of patients based on **grade**

```
table(clinical$Grade , useNA = "ifany")
```

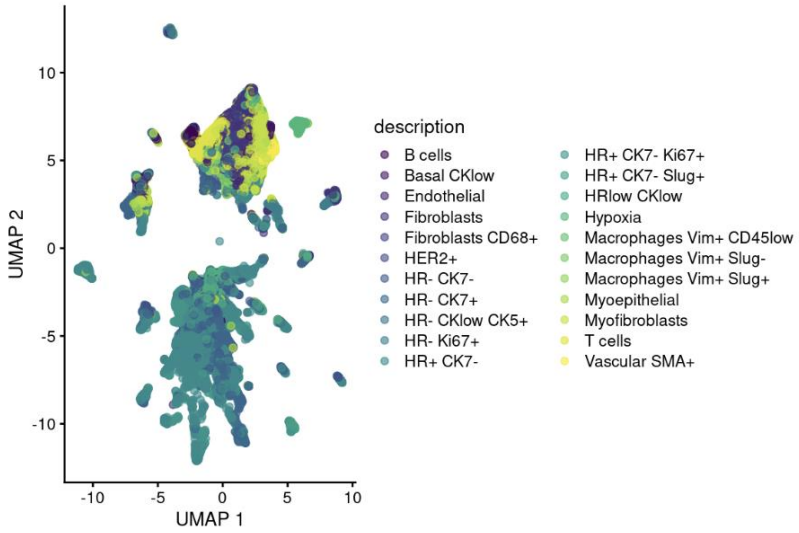
```
[1] "Number of patients based on Grade"

   1    2    3 <NA>
  14   30   28    5
```

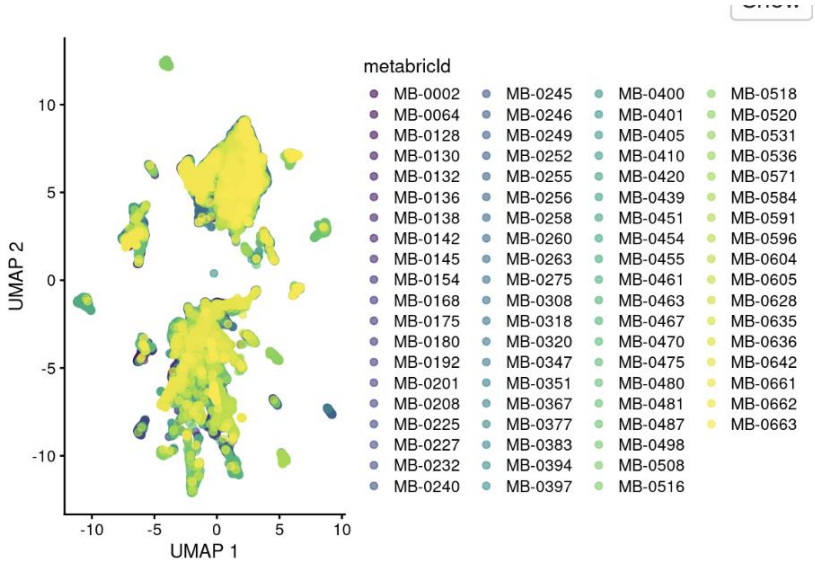
Visualisation

- Typically in single-cell data analysis, we perform **dimension reduction** to project the high dimensional *cell x gene* matrix to **2D space.**
- This allows us to **visualise** various things of interest, such as **distribution of cell types** and **disease outcomes.**

Colour by cell type

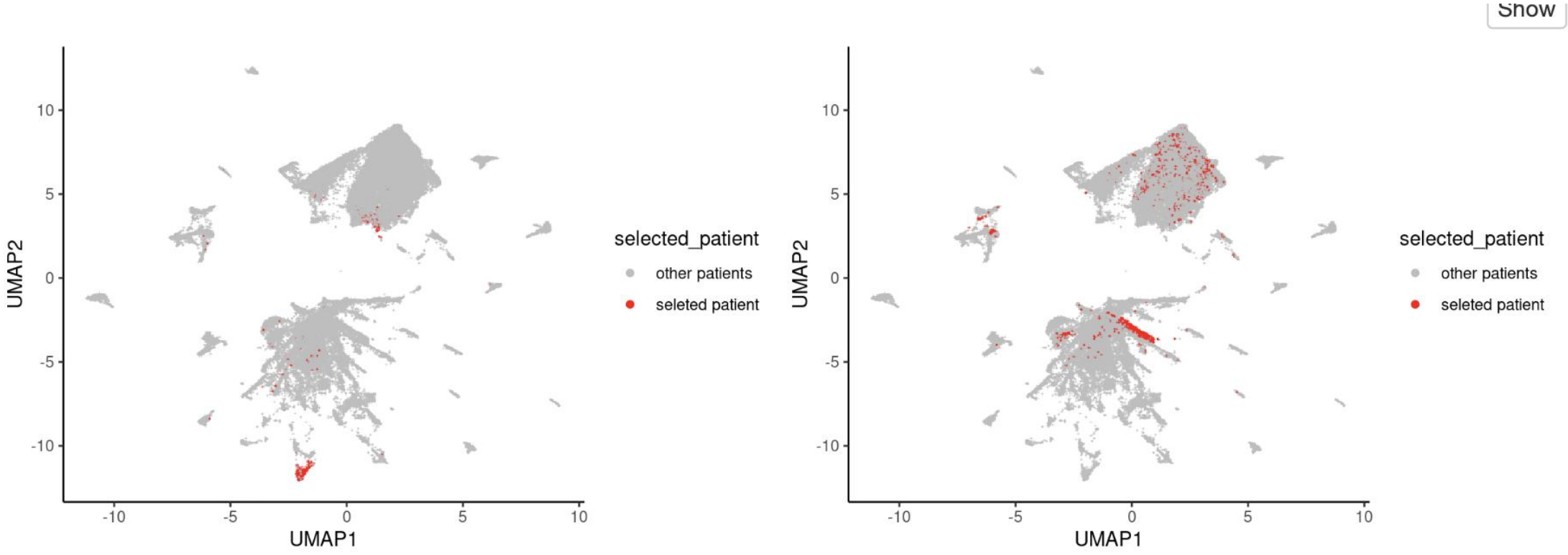


Colour by patient ID



Visualisation

– Another way to plot is highlighting each **region of interest** (eg, the patient ID)



Q1 – visualising cell types

Does each cell type cluster together?

Go to

www.menti.com

Enter the code

8474 3969

Or use QR code

Q2 – visualising patients

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

Go to

www.menti.com

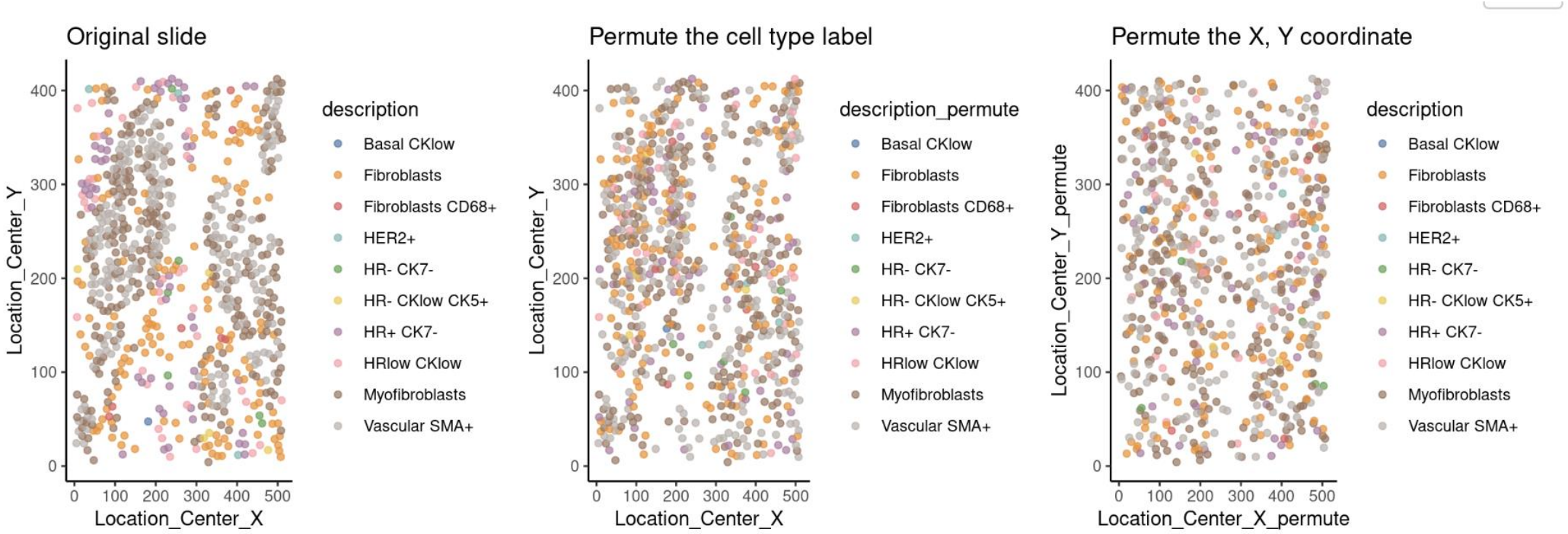
Enter the code

8474 3969

Or use QR code

Visualisation

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



Q3 – visualising cell type

Is there a **structure** in the data or are the cell types **randomly distributed**?

Go to

www.menti.com

Enter the code

8474 3969

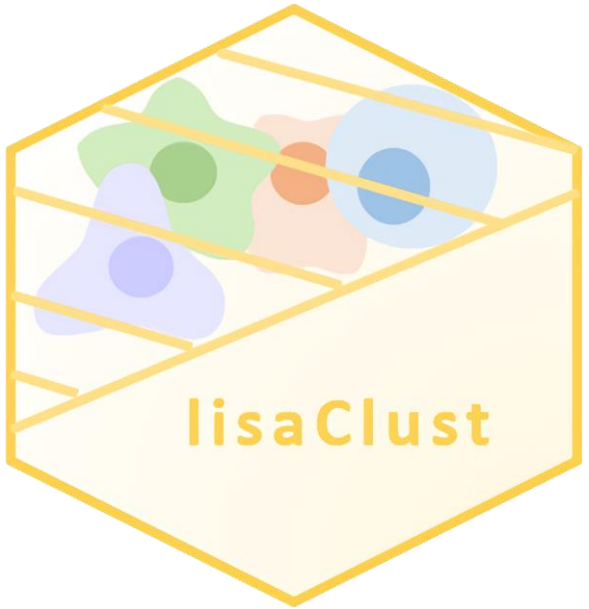
Or use QR code

Describe tissue microenvironment and neighbourhood

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

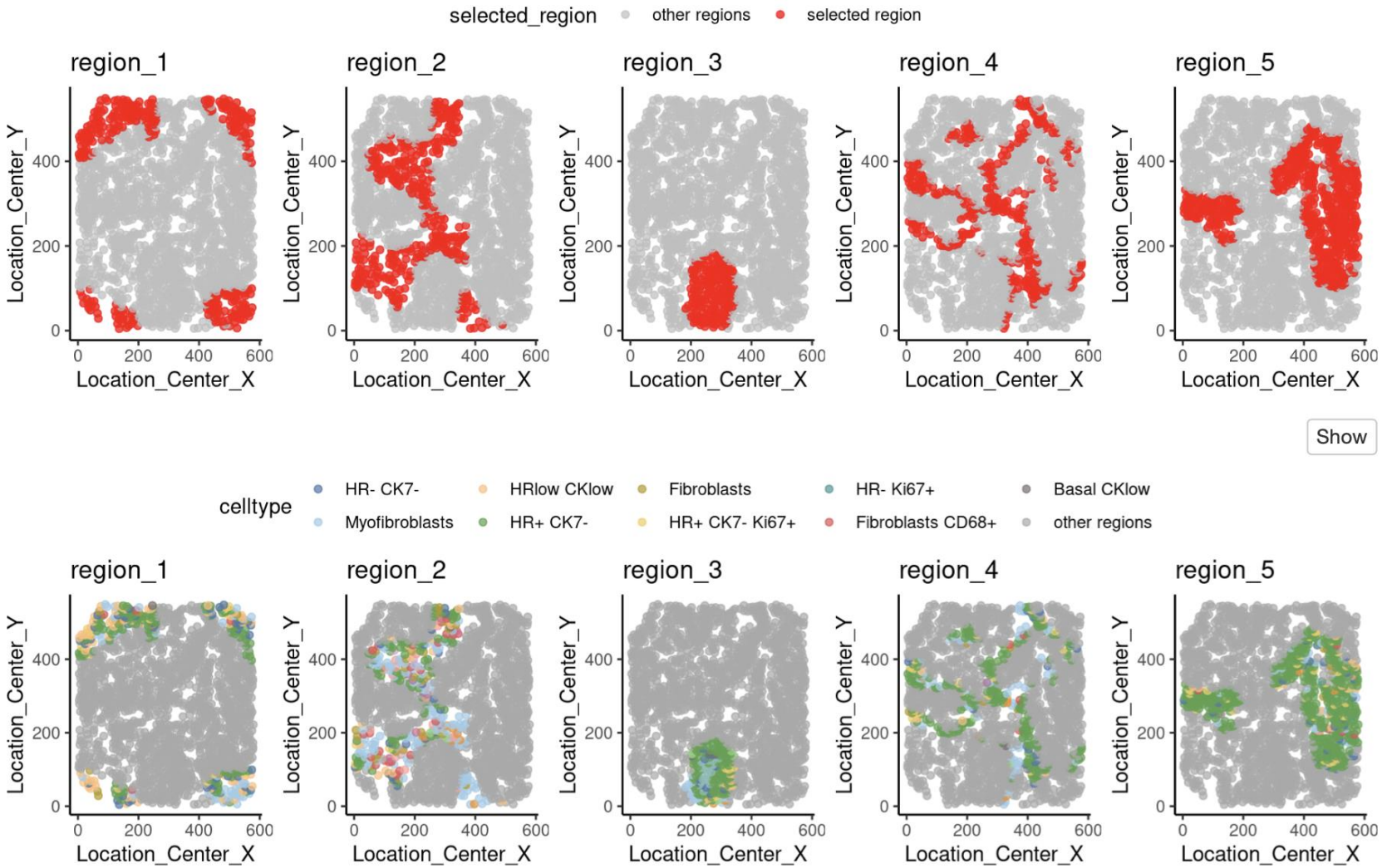
- There are multiple methods to perform this task such as ClusterMap, BASS, lisaClust

- 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

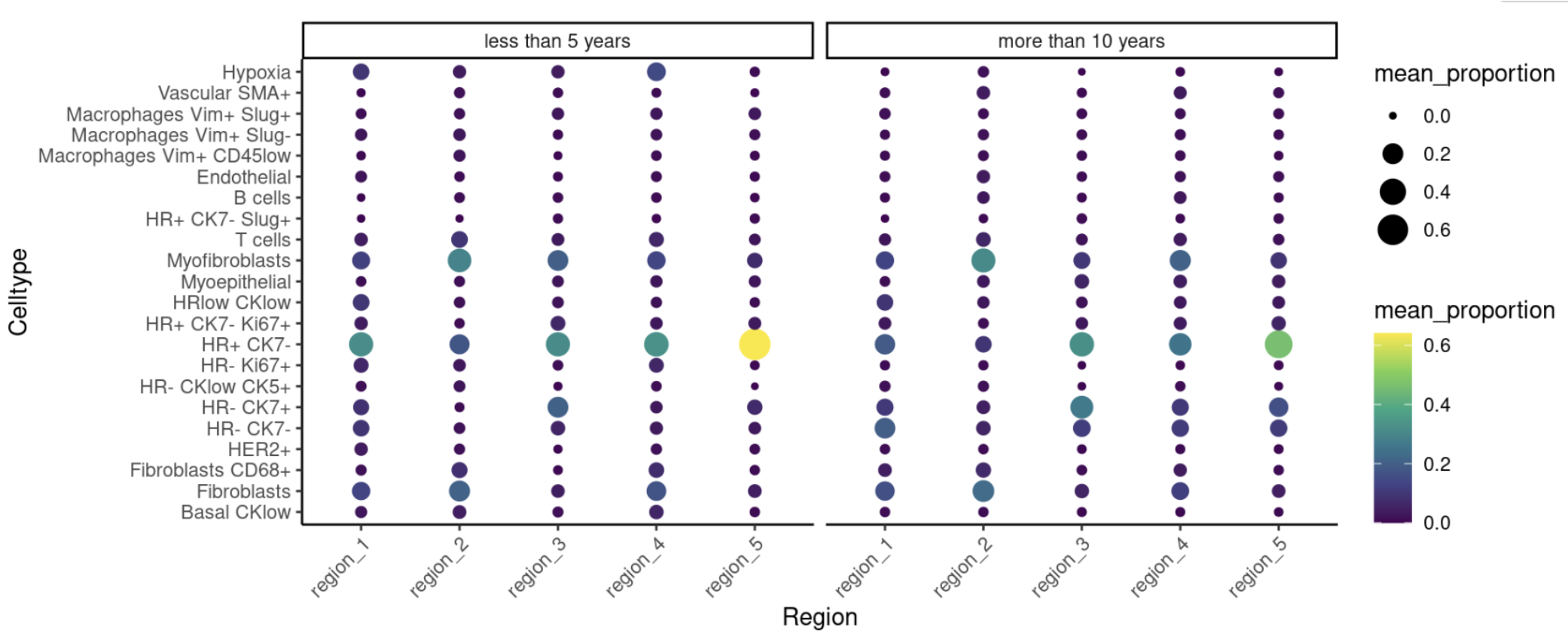
Visualise the region output by highlighting each **region** and the **cell types** in each region



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.



Q4 – visualising response

– Which regions appear to be different between poor prognosis (short–term survival) and good prognosis (long–term survival) ?

Go to

www.menti.com

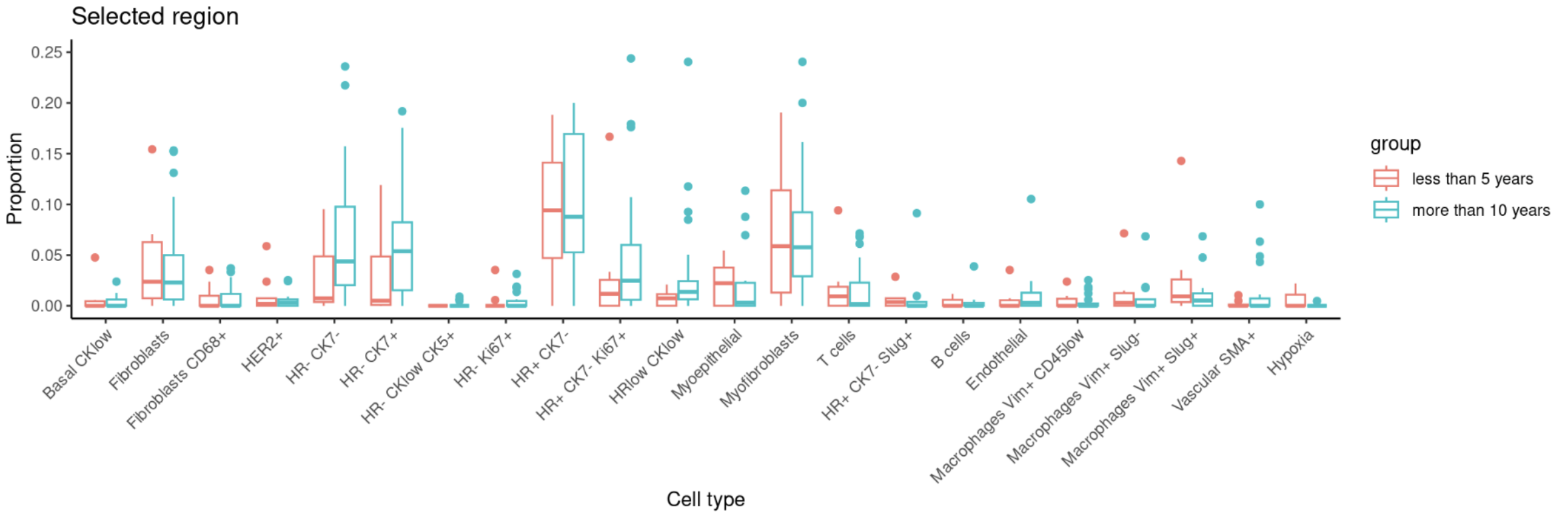
Enter the code

8474 3969

Or use QR code

Describe tissue microenvironment and neighbourhood

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



Q5 – data filtering

Are there any samples or cell types you would like to remove from the data?

Go to

www.menti.com

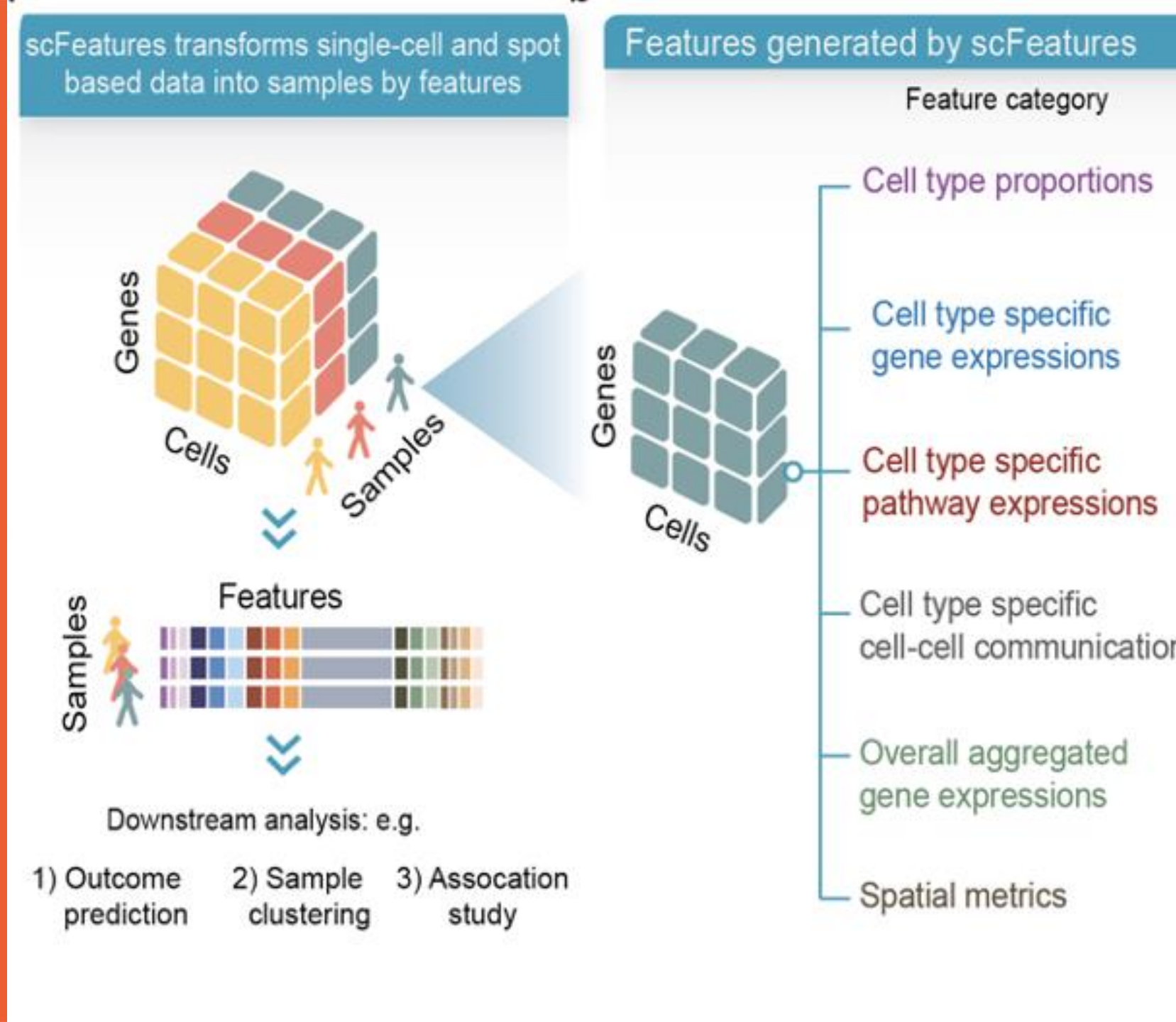
Enter the code

8474 3969

Or use QR code

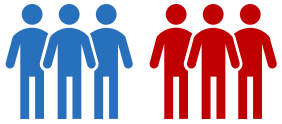
PART II:

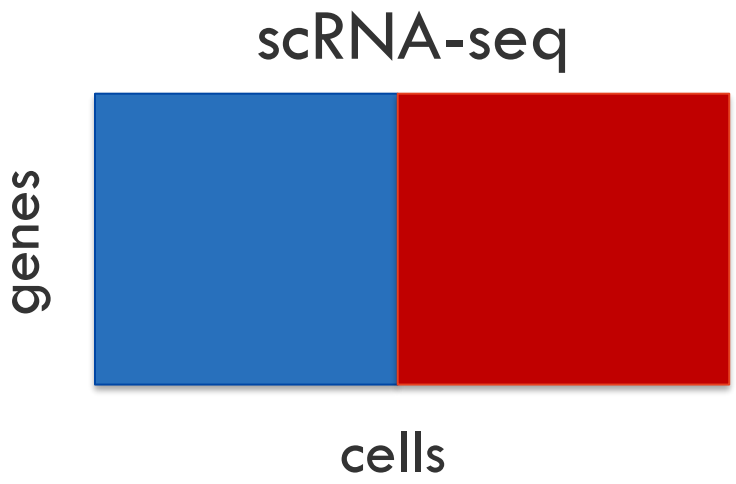
Extracting informative features with scFeatures

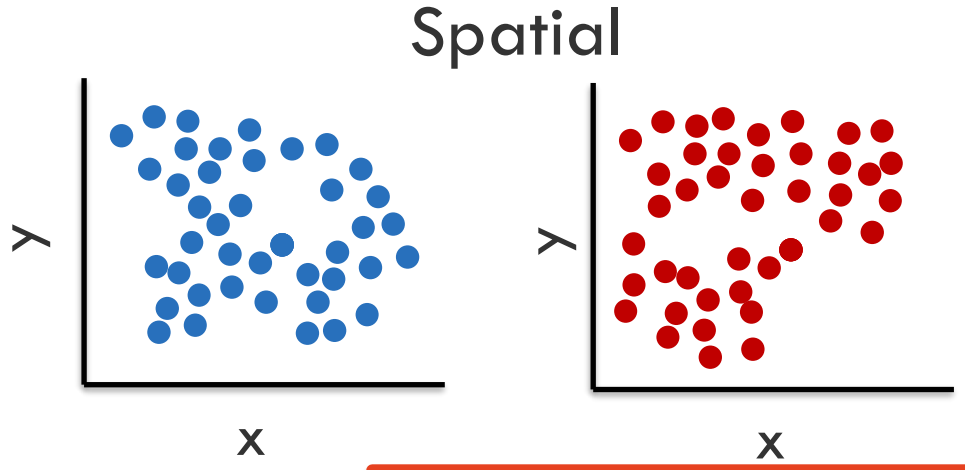


THE UNIVERSITY OF SYDNEY

How do we compare the difference between two conditions?





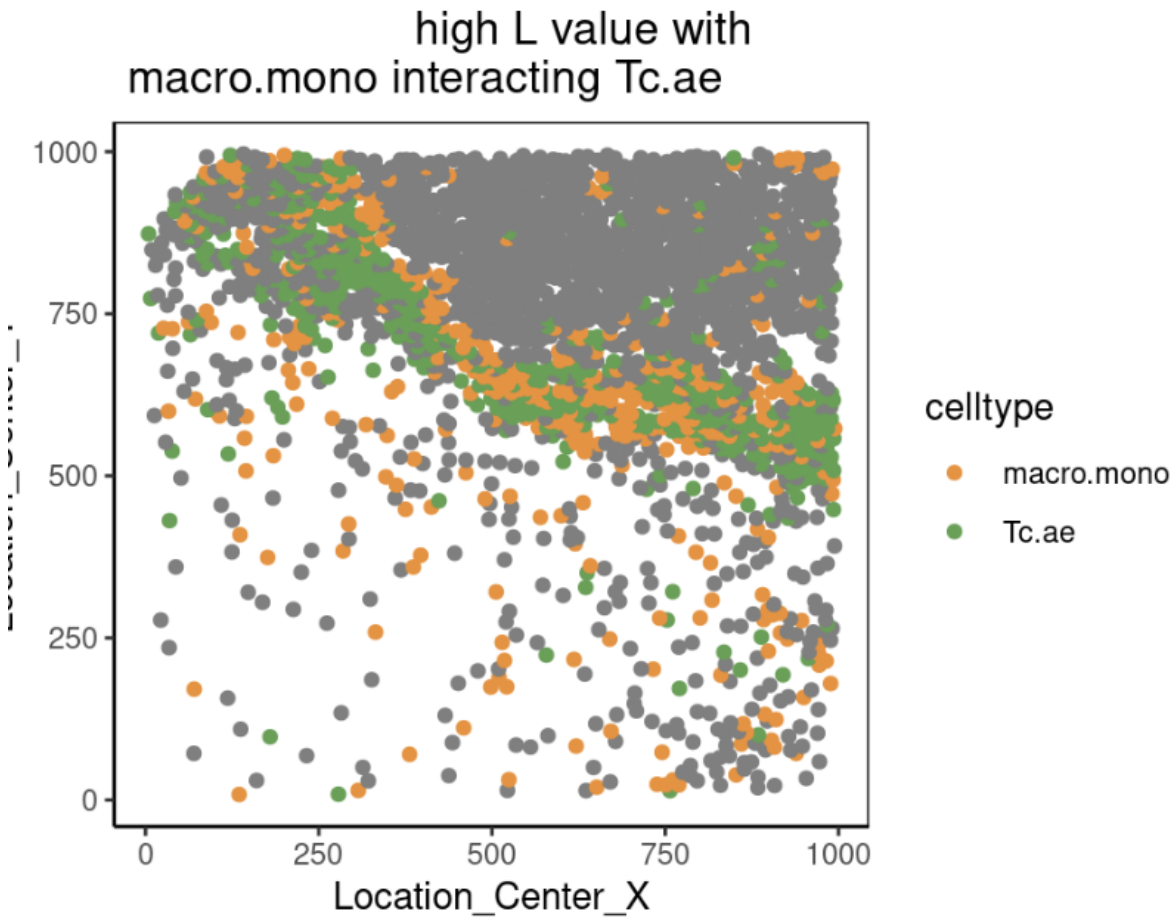


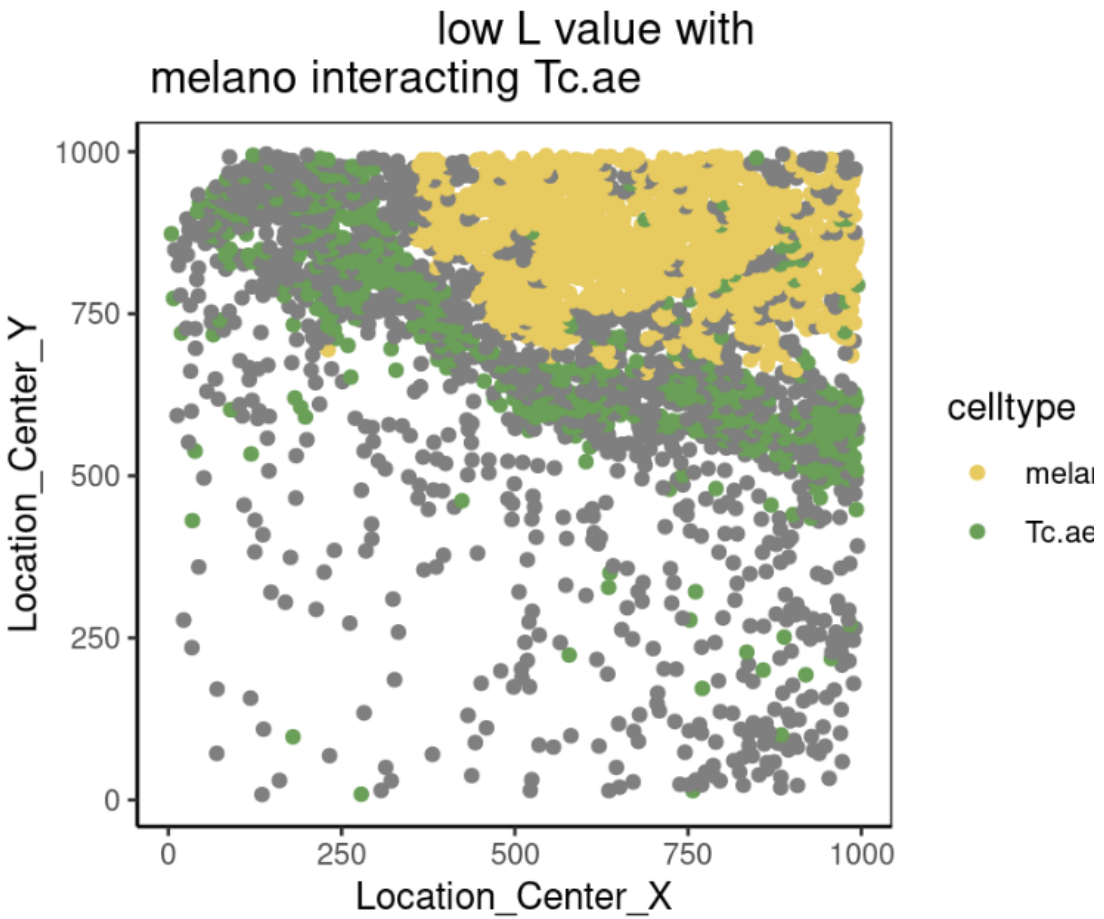
Aggregated gene expression

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

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

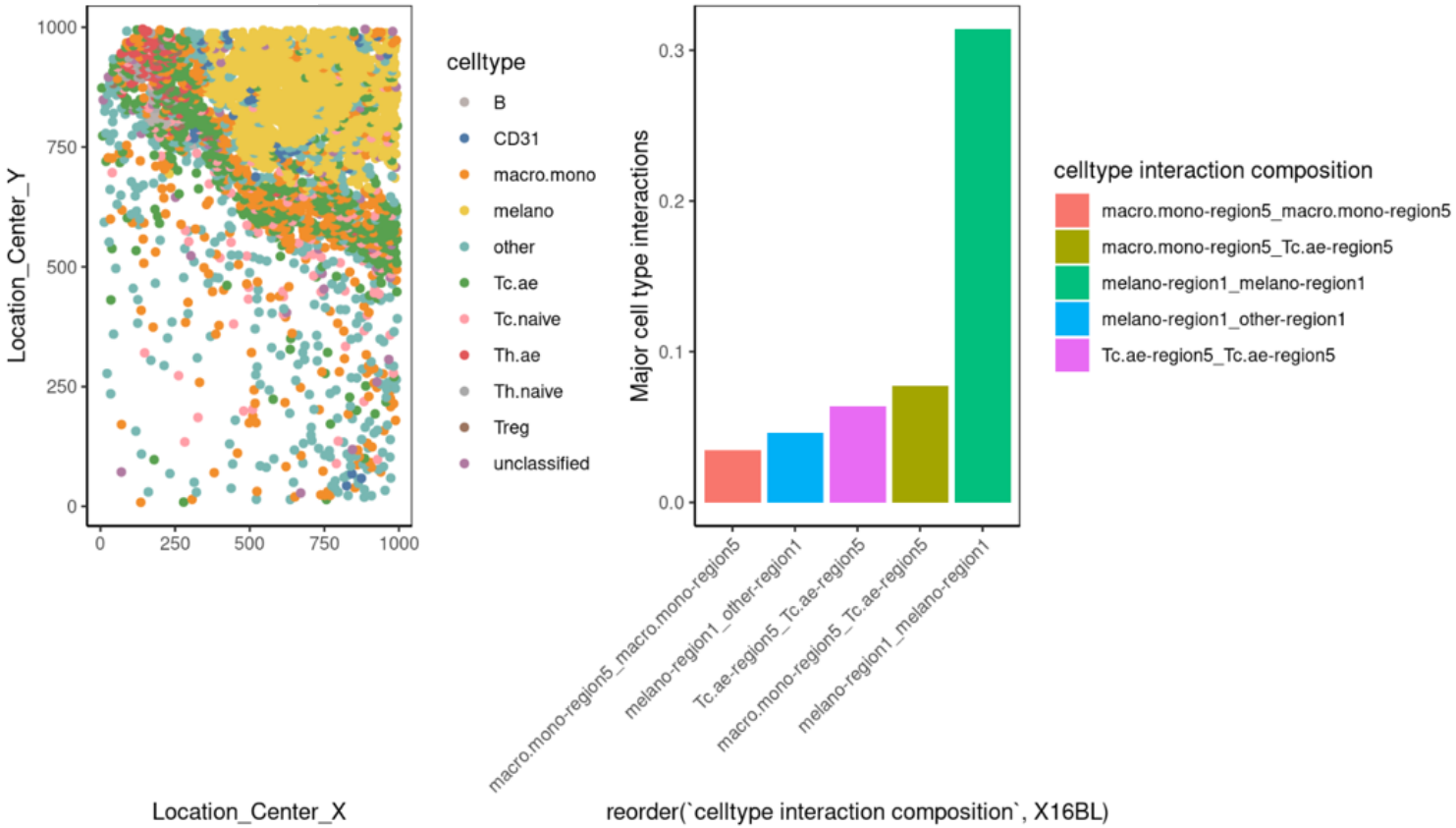
L function


high L value with
macro.mono interacting Tc.ae
1000
750
500
250
0
0
250
500
750
1000
Location_Center_X
celltype
macro.mono
Tc.ae

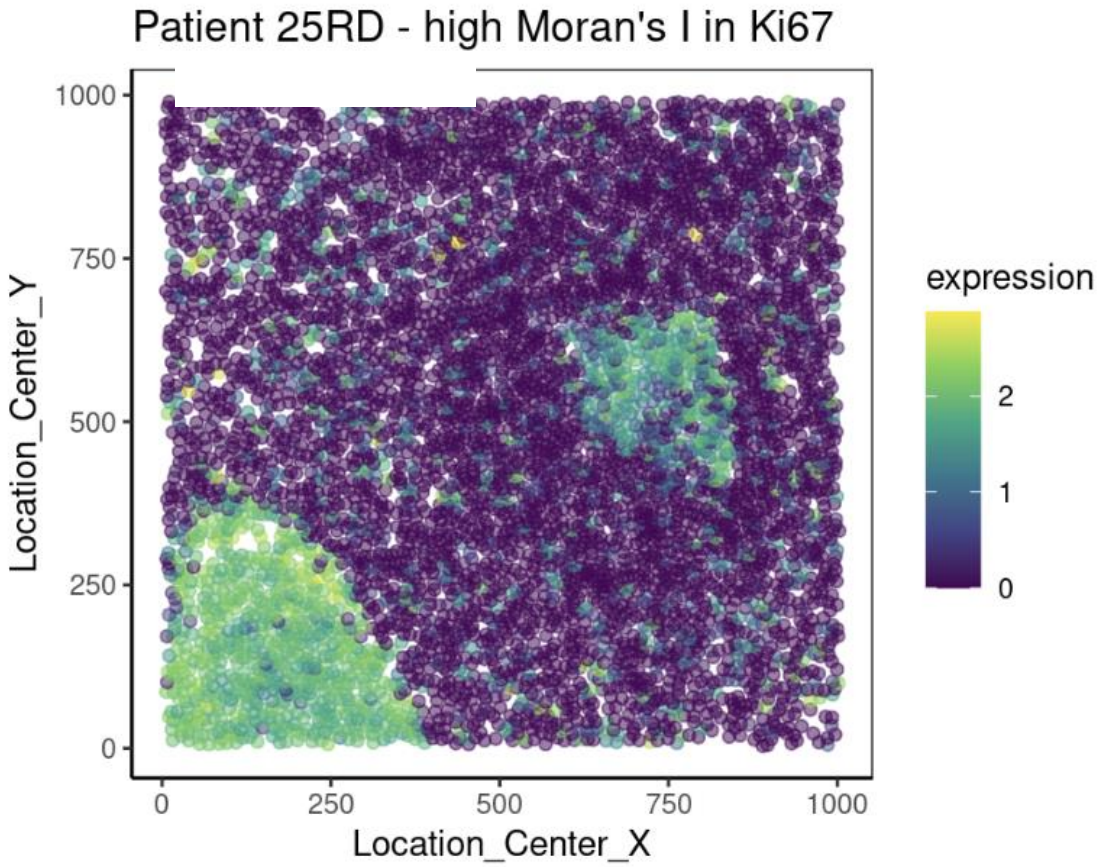

low L value with
melano interacting Tc.ae
1000
750
500
250
0
Location_Center_Y
0
250
500
750
1000
Location_Center_X
celltype
Tc.ae

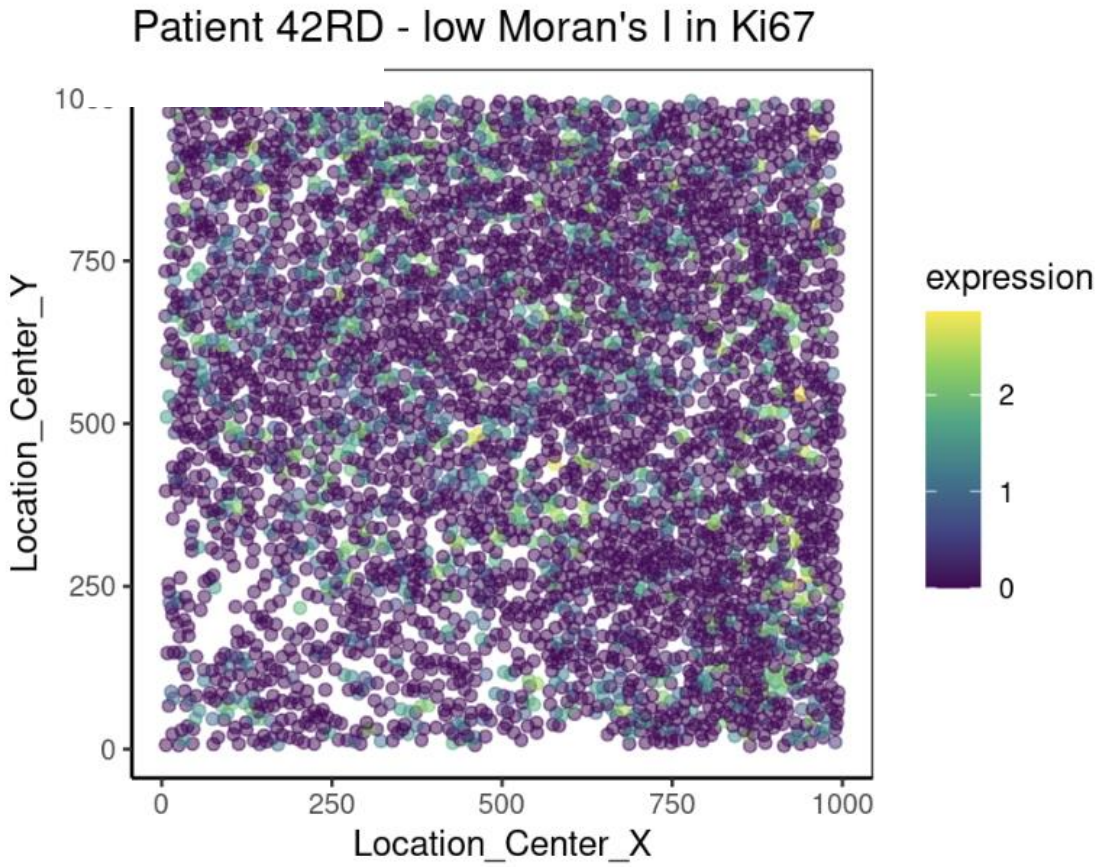
Cell type interaction composition:

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**.

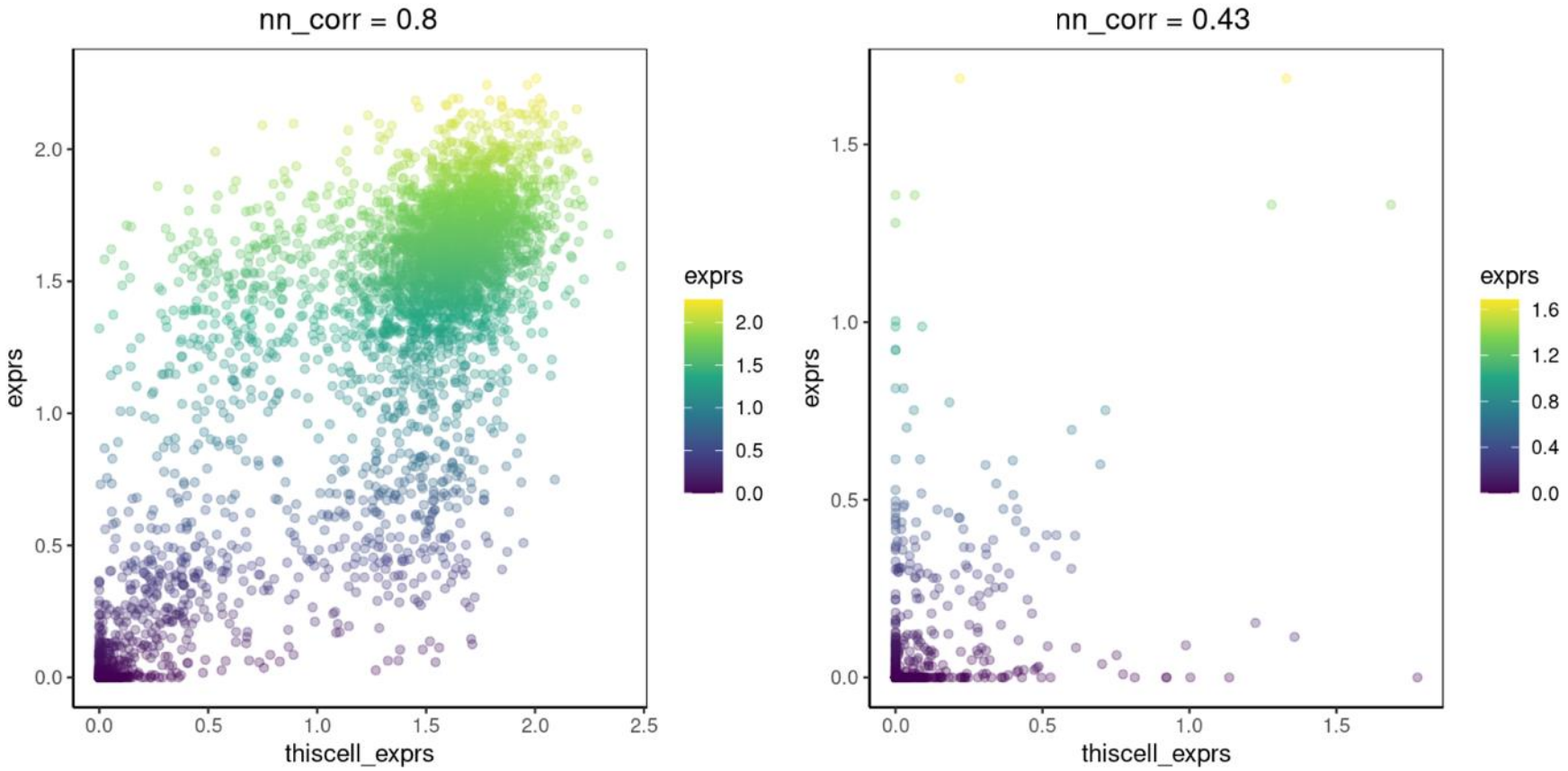


Moran's I

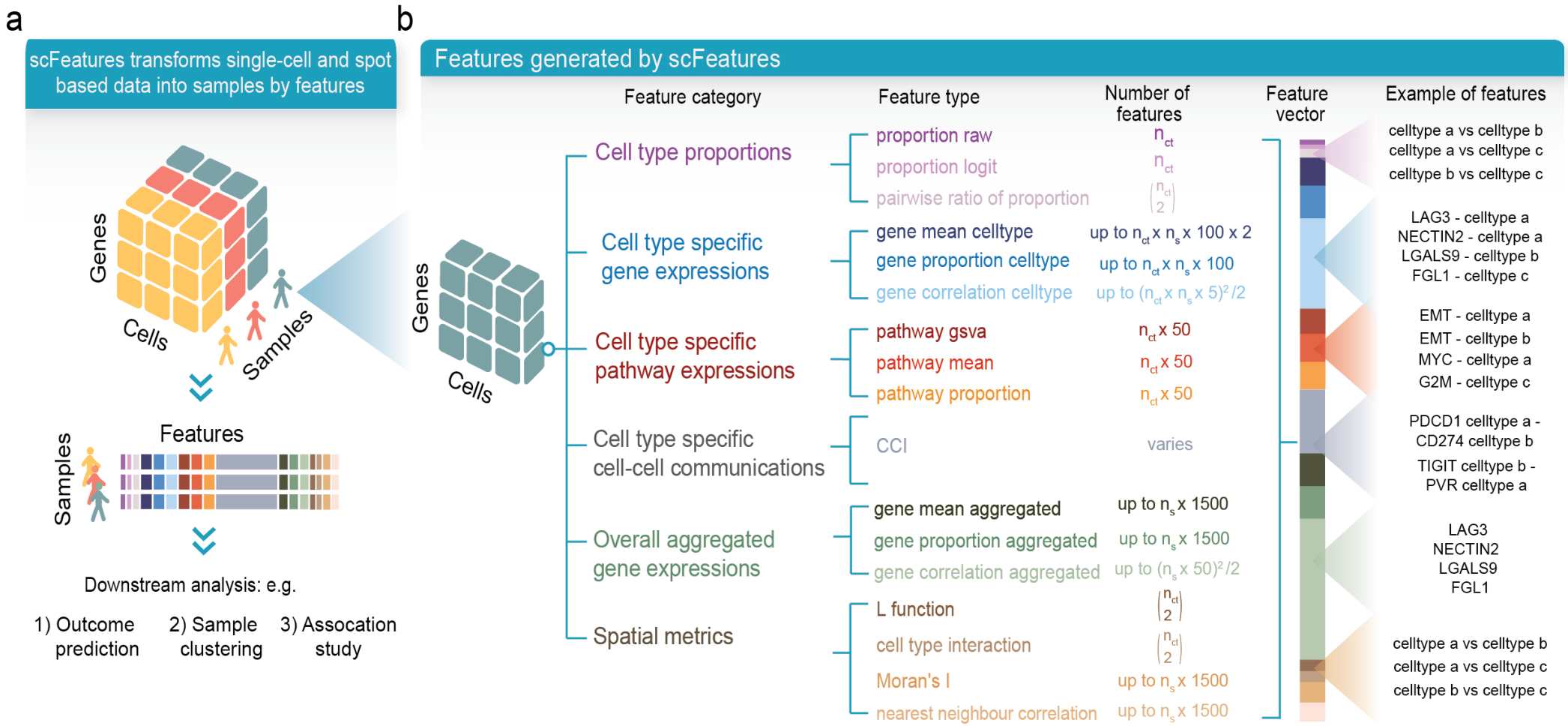

Patient 25RD - high Moran's I in Ki67
Location_Center_Y
Location_Center_X
0
250
500
750
1000
expression
2
1
0


Patient 42RD - low Moran's I in Ki67
Location_Center_Y
Location_Center_X
0
250
500
750
1000
expression
2
1
0

Nearest Neighbor Correlation – illustration with S100



scFeatures – tool for creating interpretable molecular representation of individuals for single-cell and spatial omics data



Creating molecular representations of patients

Given **a single-cell data** (gene x cell matrix), **cell type label, sample label** and **X, Y spatial coordinates,** all feature types can be generated in **one line of code.**

```
# here, we specify that this is a spatial proteomics data
# scFeatures support parallel computation to speed up the process
scfeatures_result <- scFeatures(IMCmatrix, type = "spatial_p",
                                sample = sample, celltype = celltype, spatialCoords = spatialCoords,
                                ncores = 32)
```

Creating molecular representations of patients

Allows user **customisation**, such as data type, species, genes/genesets of interest

Usage

```
scFeatures(
  data = NULL,
  sample = NULL,
  celltype = NULL,
  spatialCoords = NULL,
  spotProbability = NULL,
  feature_types = NULL,
  type = "scrna",
  ncores = 1,
  species = "Homo sapiens",
  celltype_genes = NULL,
  aggregated_genes = NULL,
  geneset = NULL
)
```

Inspecting generated features

All generated feature types are stored in a **matrix** of *samples x features*.

```
type(scfeatures_result)
```

```
[1] "list"
```

```
```{r}
# we have generated a total of 13 feature types
names(scfeatures_result)
```
```

```
 [1] "proportion_raw"       "proportion_logit"     "proportion_ratio"     "gene_mean_
 [5] "gene_prop_celltype"   "gene_cor_celltype"    "gene_mean_bulk"       "gene_prop_
 [9] "gene_cor_bulk"        "L_stats"              "celltype_interaction" "morans_I"
[13] "nn_correlation"
```

```
```{r}
lapply(scfeatures_result, dim)
# each row is a sample, each column is a feature
```
```

```
$proportion_raw
[1]  77 110

$proportion_logit
[1]  77 110

$proportion_ratio
[1]   77 5995

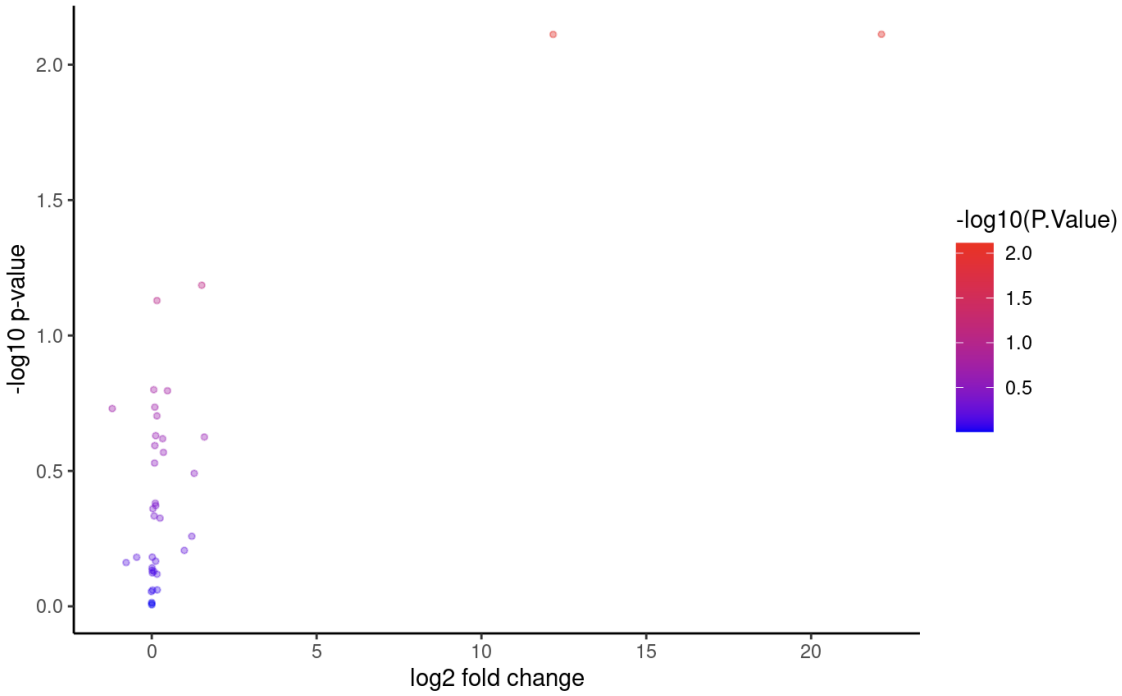
$gene_mean_celltype
[1]   77 4180

$gene_prop_celltype
[1]   77 4180

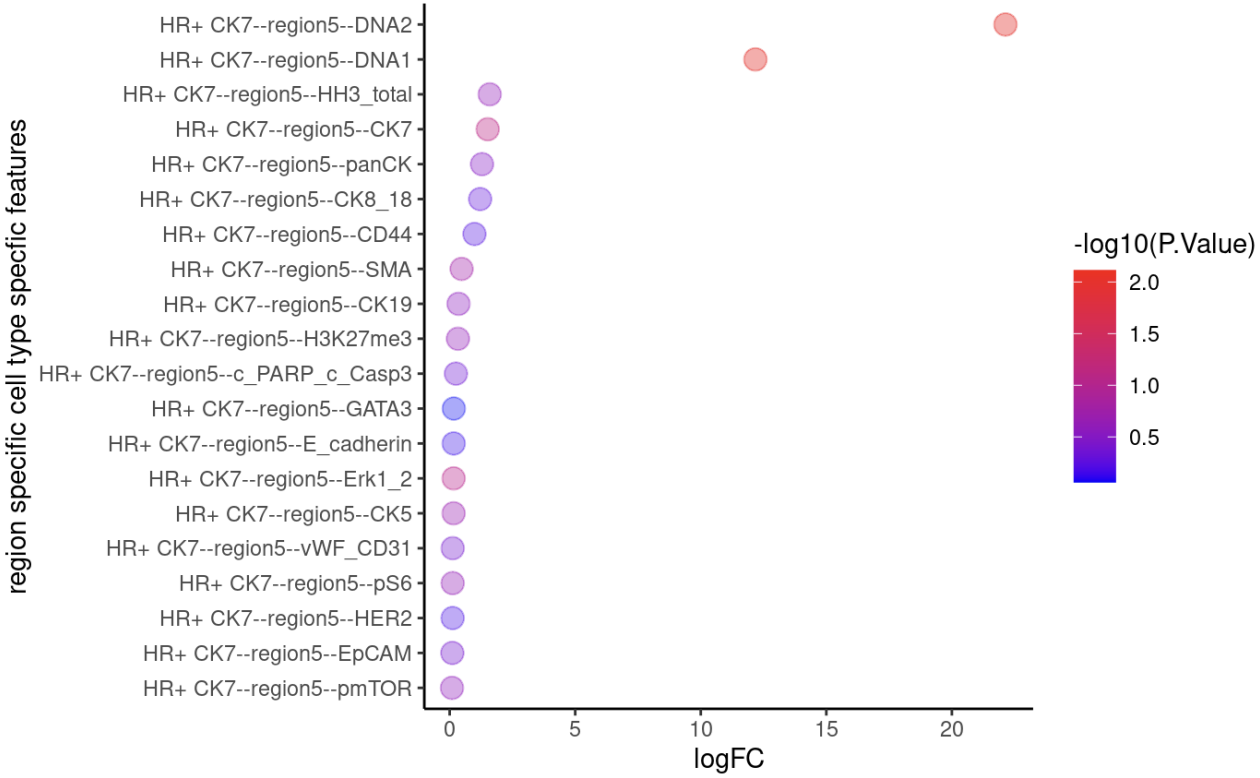
$gene_cor_celltype
[1]    77 27958
```

Visualise scFeatures generated features

Volcano plot



Dot plot



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

We will resume at 16:35

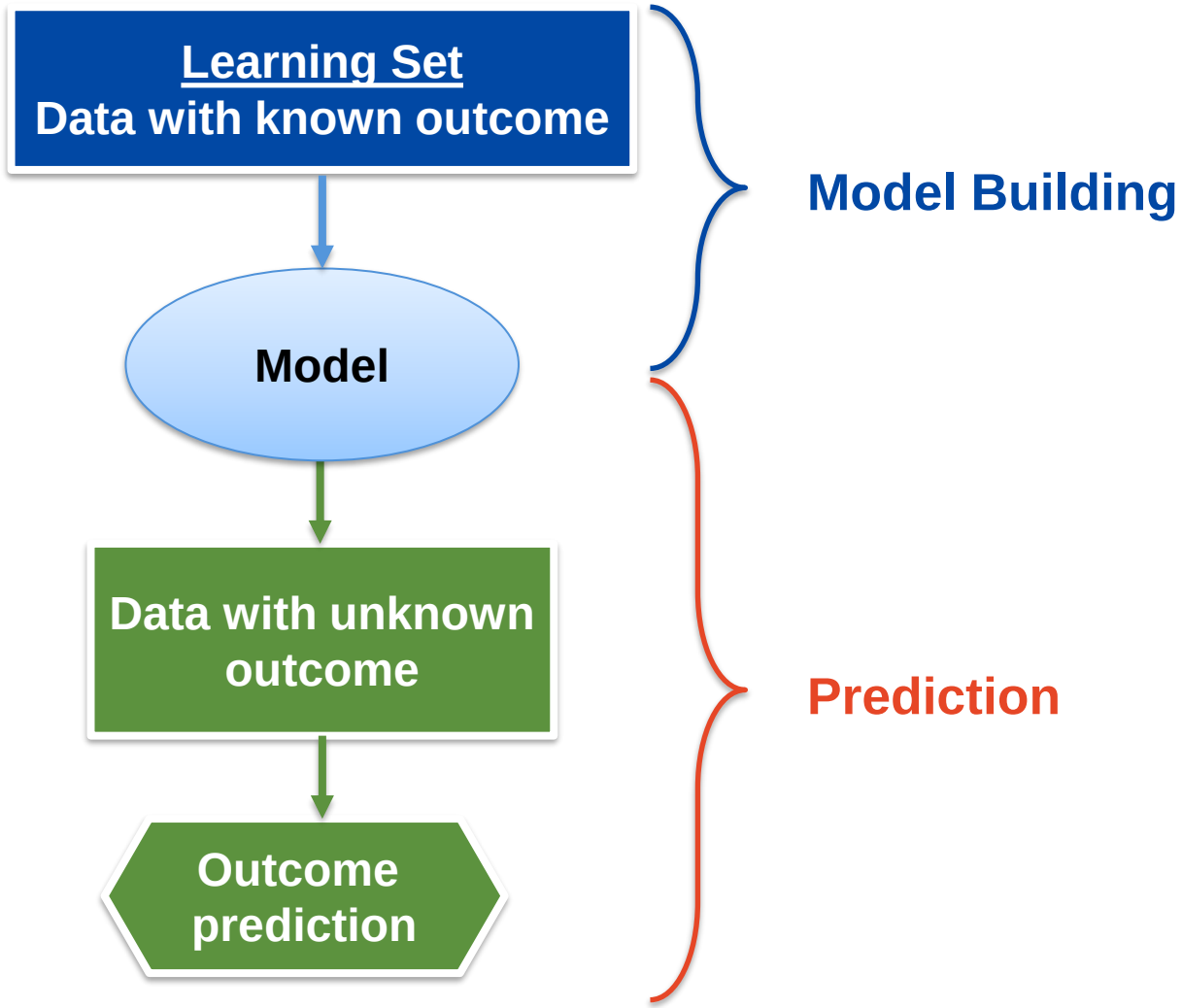
THE UNIVERSITY OF SYDNEY

PART III:

Disease classification with ClassifyR

THE UNIVERSITY OF SYDNEY

Modelling and Evaluation



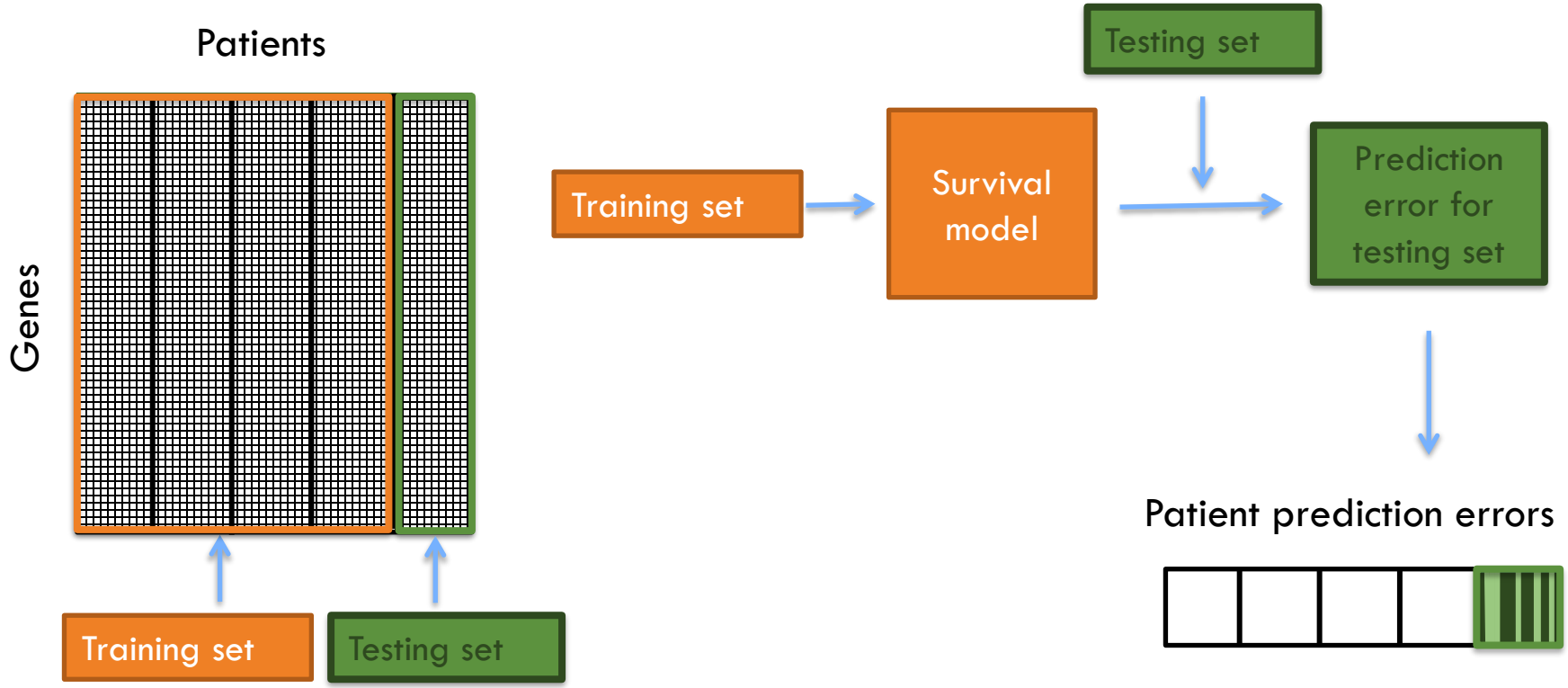
Many choices in model building:

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

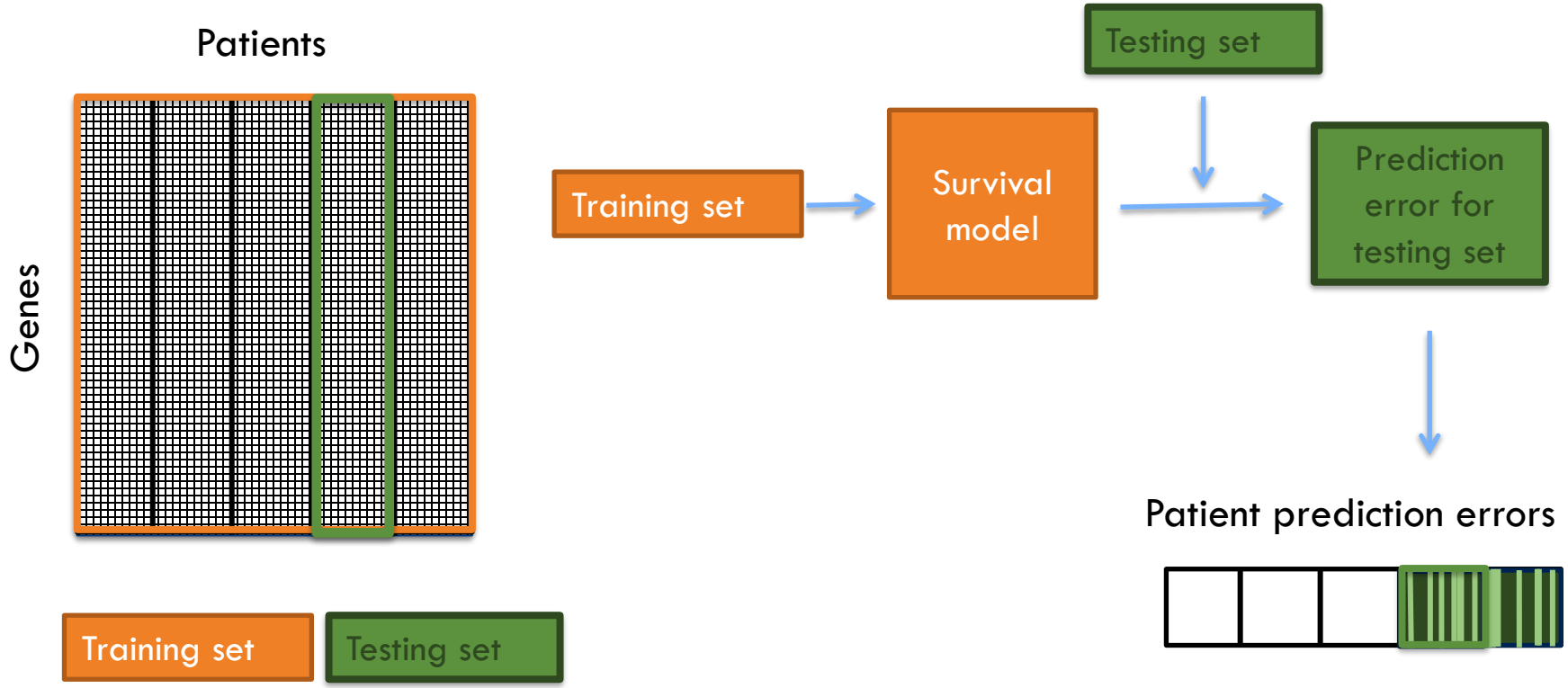
The best choices will vary based on the **data** and **task**.

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

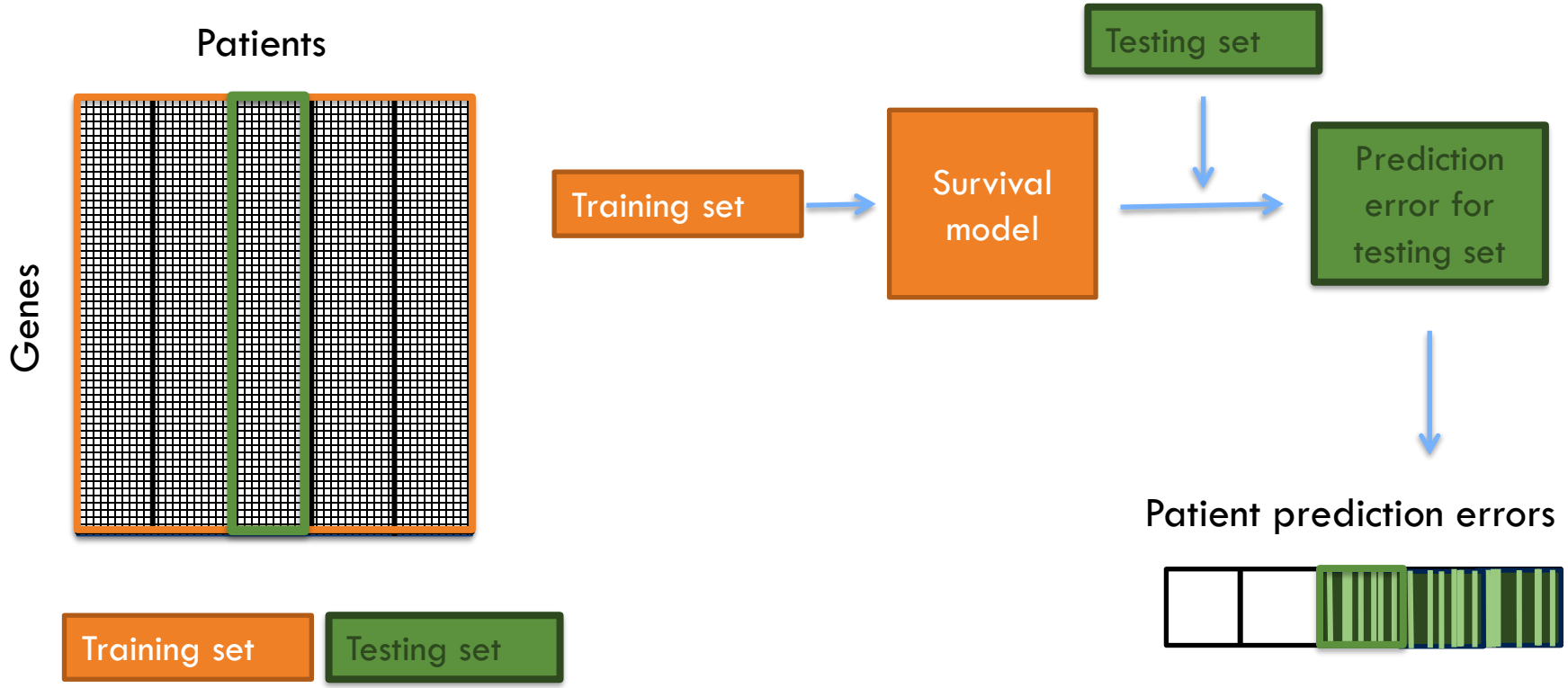
5-fold cross validation



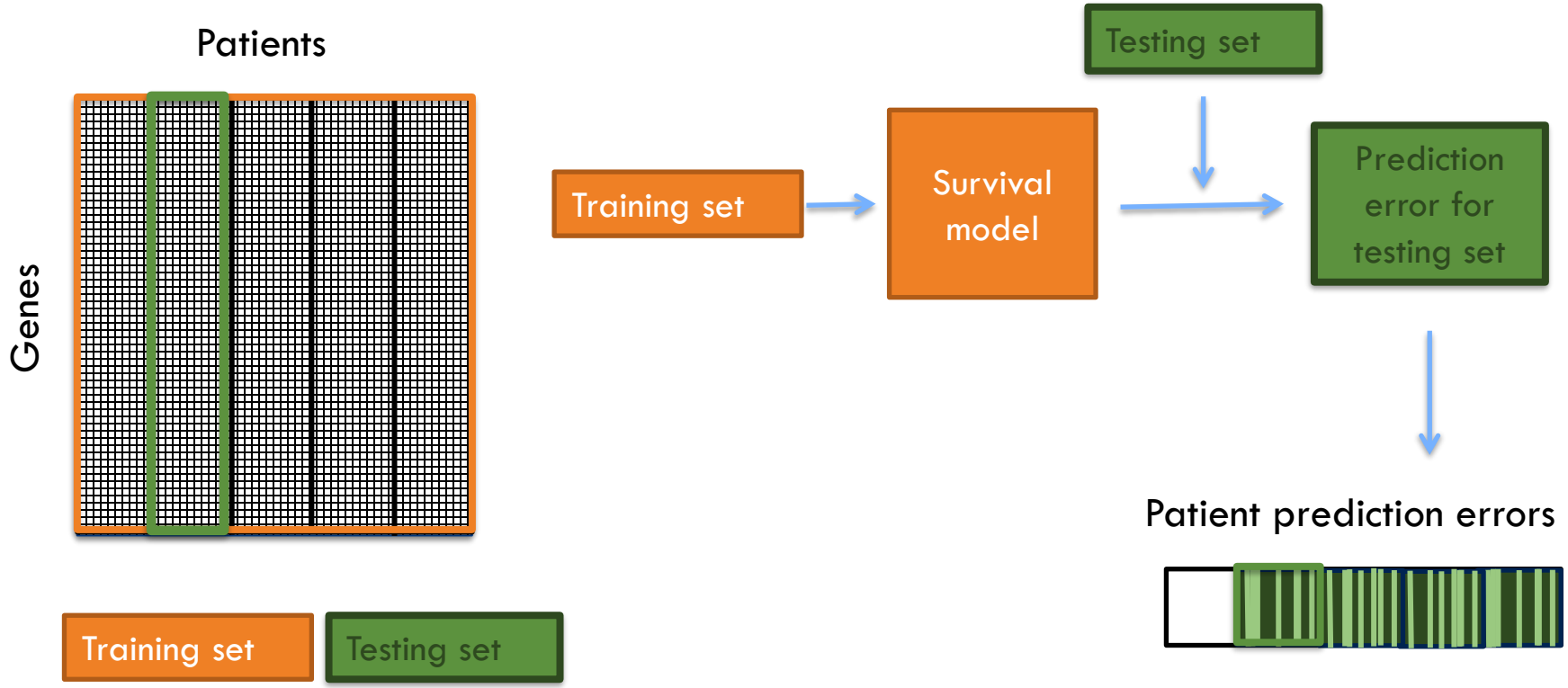
5-fold cross validation



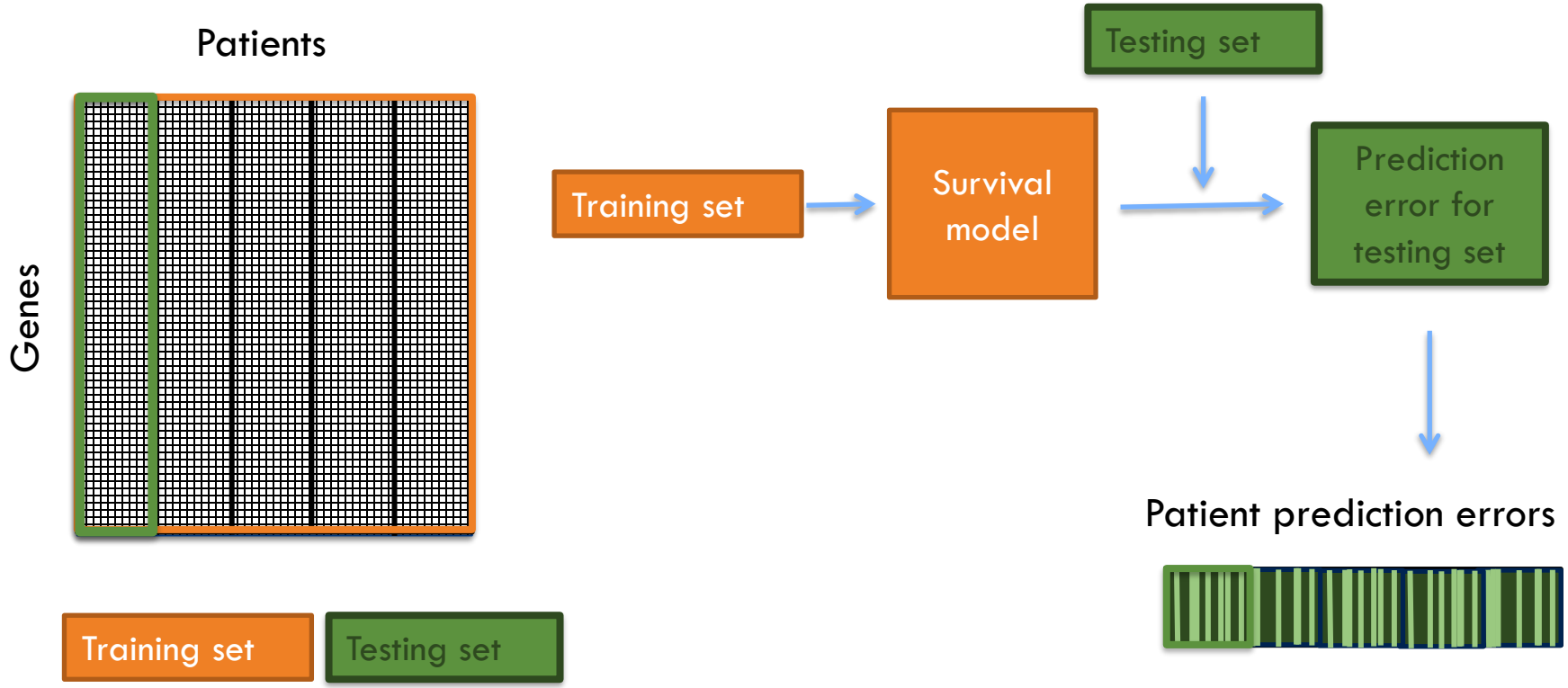
5-fold cross validation



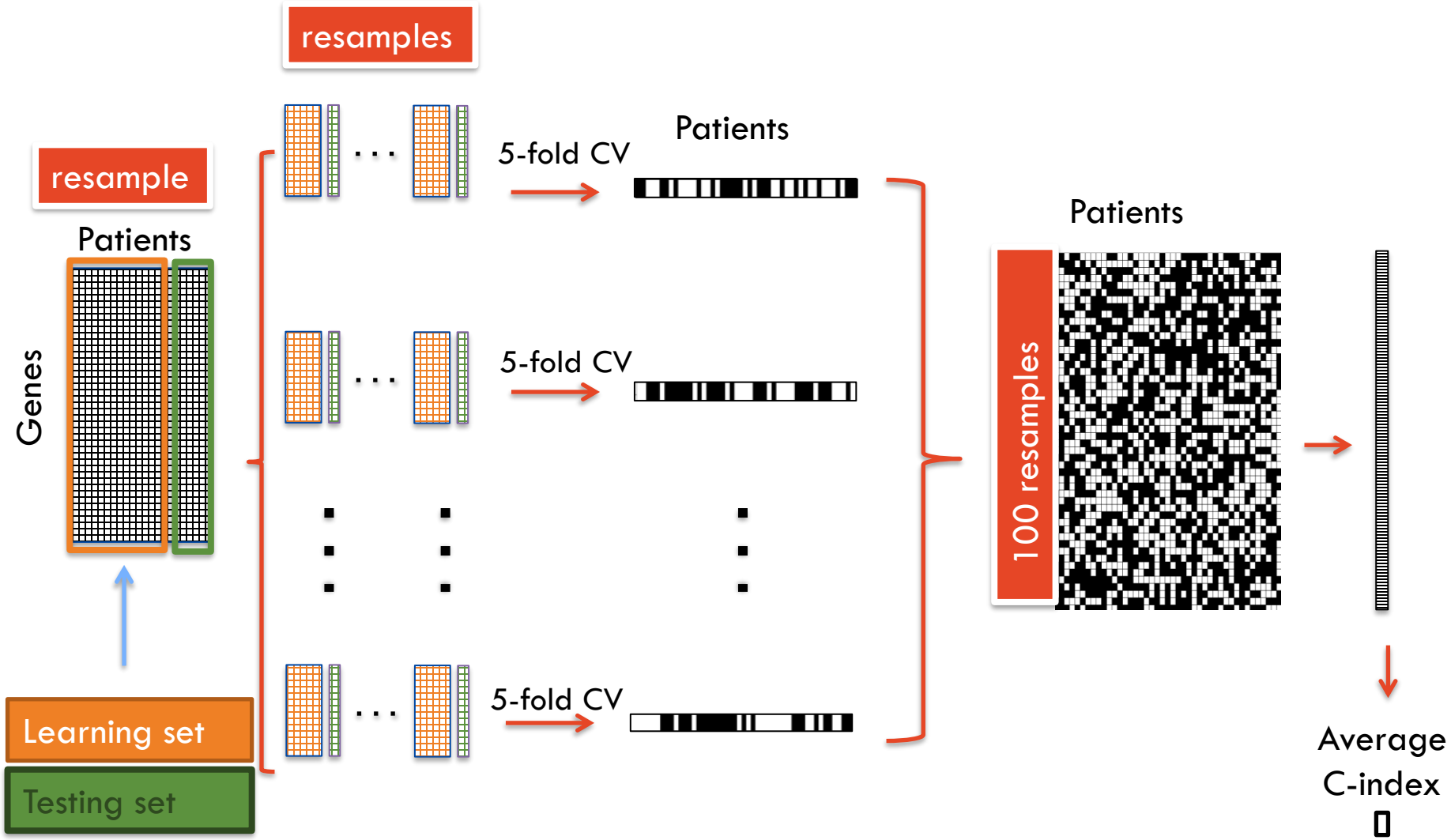
5-fold cross validation



5-fold cross validation



Repeated 5-fold cross validation

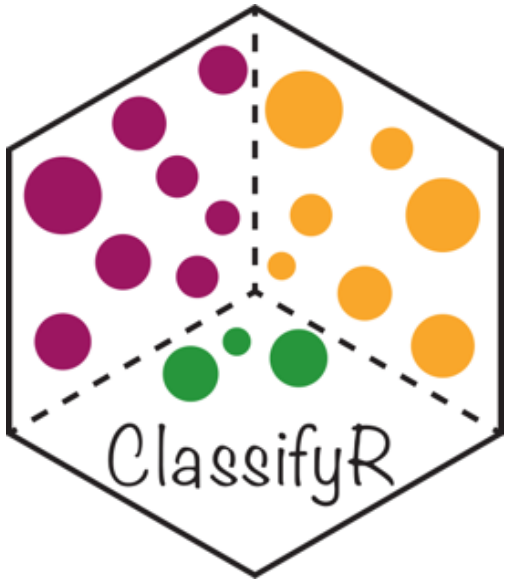


ClassifyR: outcome prediction

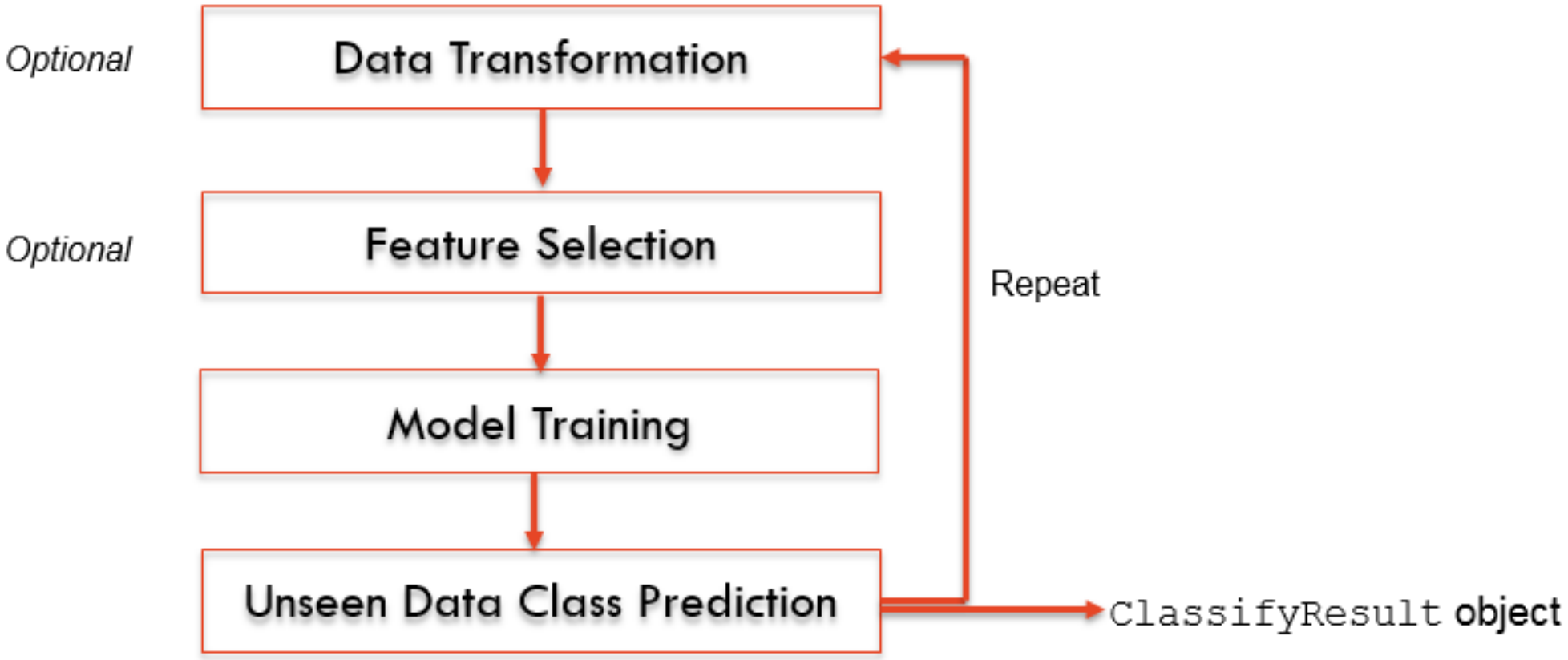
ClassifyR: an R package for performance assessment of classification with applications to transcriptomics FREE

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** ▾



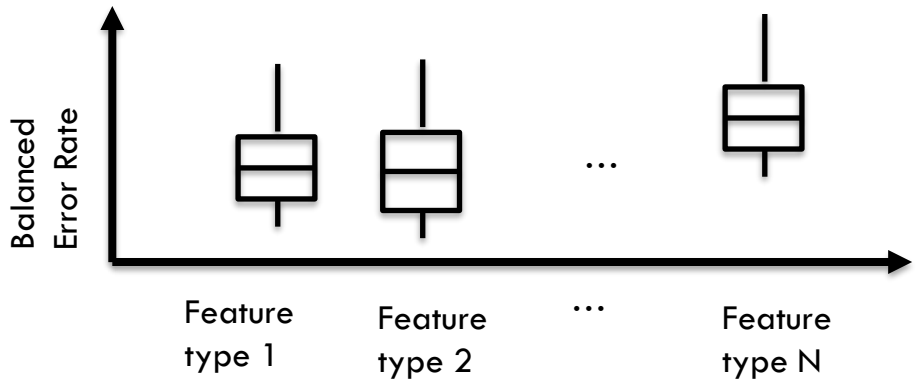
Full cross-validation loops



Multi-view Data Modelling in ClassifyR

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

Cross-validation



`crossValidate`: Cross-validation of multi-view modelling

`crossValidate` allows **direct input** of popular data containers:

- MultiAssayExperiment
- dataframe
- matrix
- a list of matrix and dataframe

Outcome is flexible so that different data sets can be handled with minimal data wrangling. Can be:

- Factor (plain tabular data only).
- Name or index of column containing a factor.
- Object of type `Surv` (plain tabular data only).
- A pair of column names containing time and status information.

Survival analysis

```
usefulFeatures <- c("Breast.Tumour.Laterality", "ER.Status", "Inferred.Menopausal.State", "Grade", "Siz
e")
nFeatures <- append(list(clinical = 1:3), lapply(scfeatures_result, function(metaFeature) 1:5))
clinicalAndOmics <- append(list(clinical = clinical), scfeatures_result)

### generate classfyr result

classifyr_result <- crossValidate(clinicalAndOmics, c("timeRFS", "eventRFS"),
                    extraParams = list(prepare = list(useFeatures = list(clinical = usefulFeatures))),
                    nFeatures = nFeatures, nFolds = 5, nRepeats = 5, nCores = 5)
```

Survival analysis

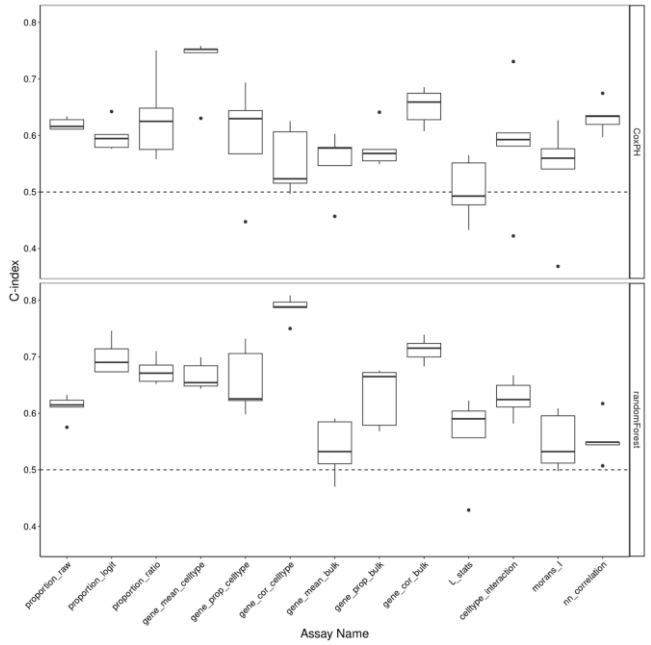
- By default, `classifyR` uses **cox model** for survival analysis.
- We can also specify other models.

```
survForestCV <- crossValidate(clinicalAndOmics, outcome, nFeatures = nFeatures,
                              classifier = "randomForest",
                              nFolds = 5, nRepeats = 5, nCores = 5)
```

Visualising the model performance

- The **model performance** can be visualised in one line of code using the `performancePlot` **function**

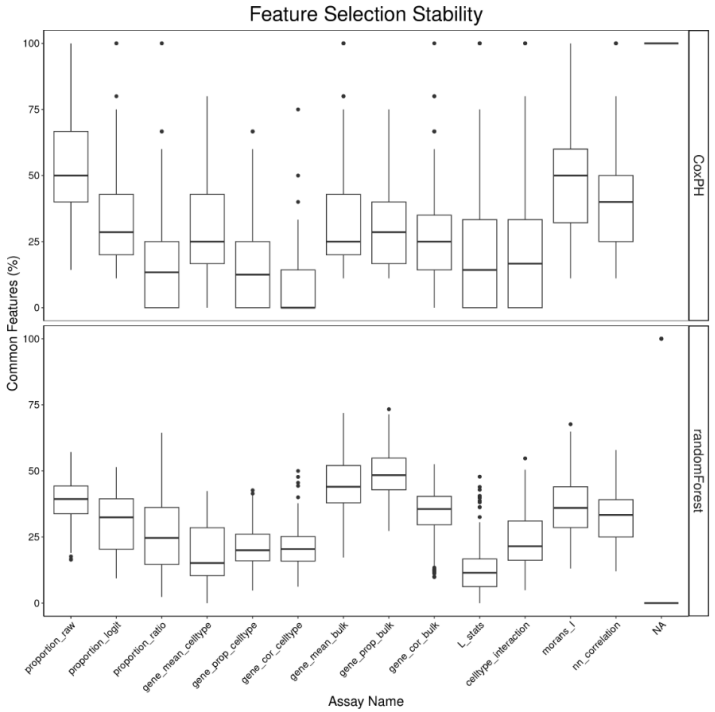
```
performancePlot(append(classifyr_result, survForestCV),
                characteristicsList = list(x = "Assay Name", row = "Classifier Name"),
                orderingList = list("Assay Name" = names(scfeatures_result))) + tilt
```



Visualising the feature stability

- The **feature stability** can be visualised in one line of code using the `selectionPlot` **function**

```
selectionPlot(append(classifyr_result, survForestCV),
              characteristicsList = list(x = "Assay Name", row = "Classifier Name"),
              orderingList = list("Assay Name" = names(scfeatures_result))) + tilt
```



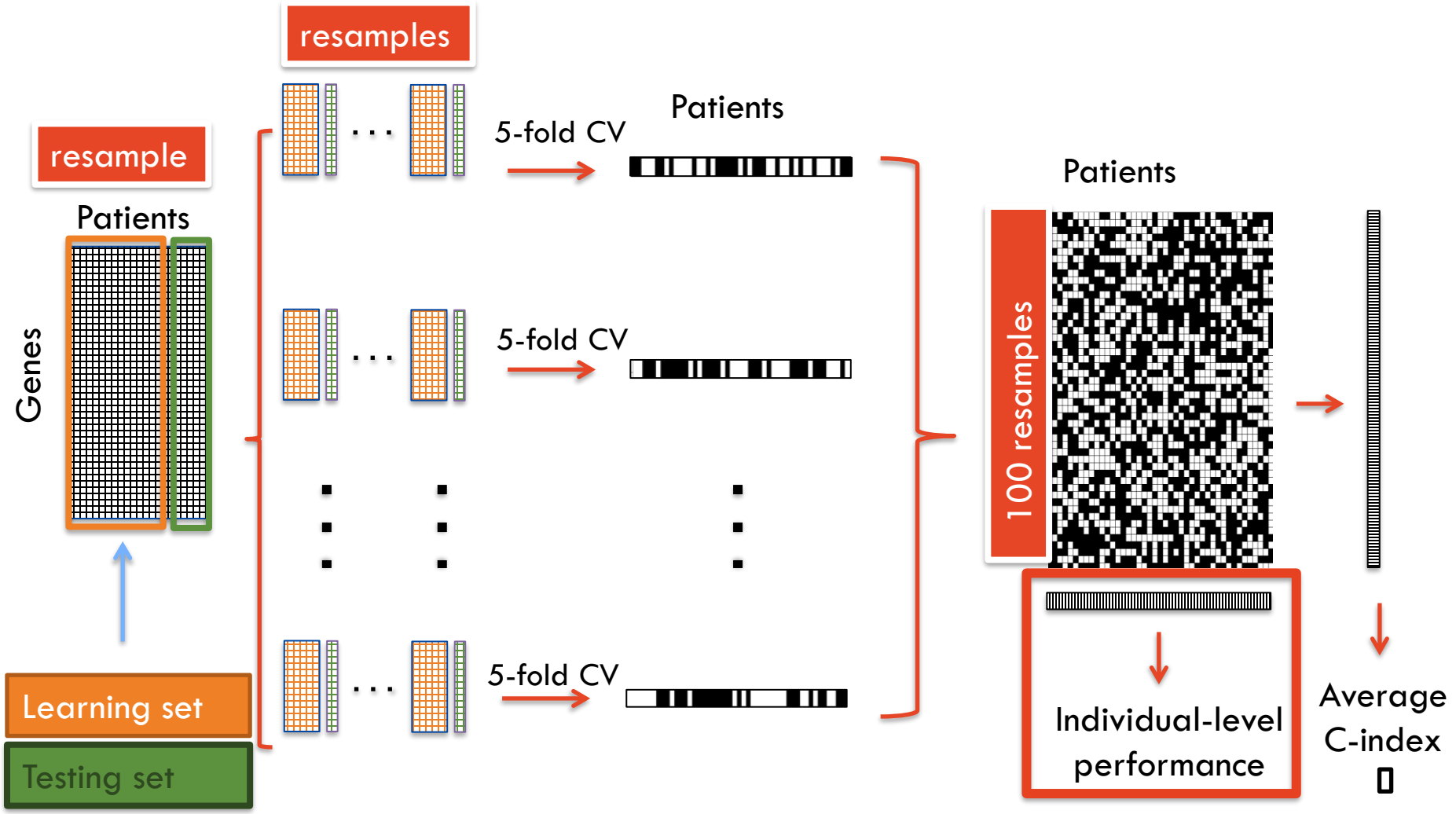
PART IV:

Identifying cohort heterogeneity with ClassifyR

THE UNIVERSITY OF SYDNEY

Identifying cohort heterogeneity

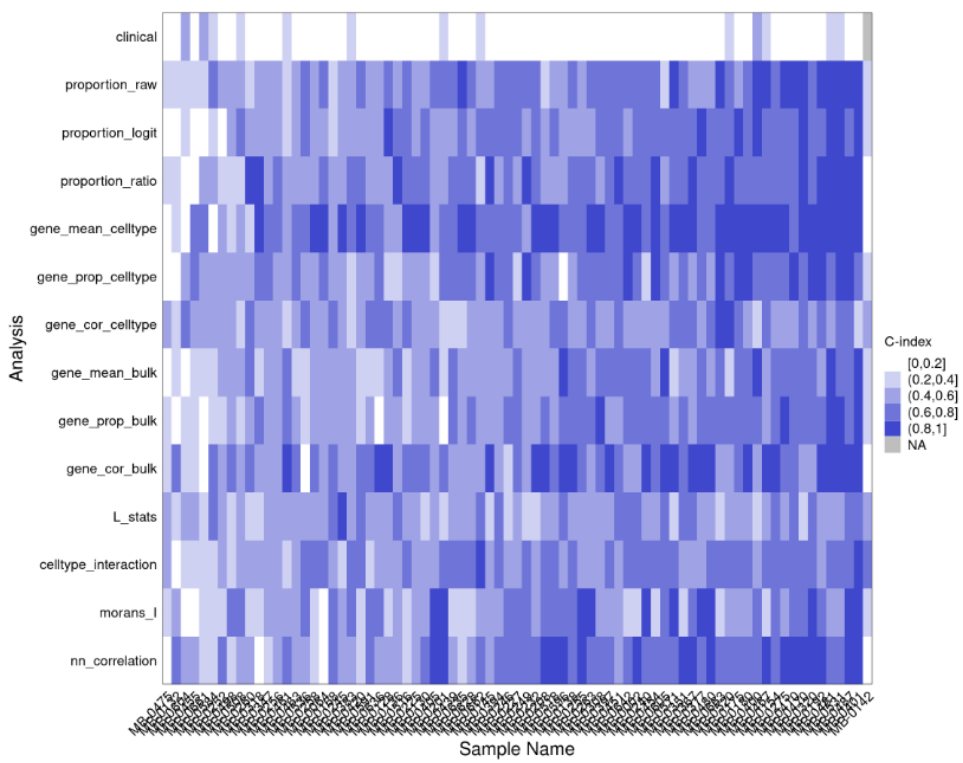
Human cohorts are often very **heterogenous**. It can be useful to explore cohort heterogeneity by looking at **individual-level performance.**



Visualising the individual-level performance

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

```
library(grid)
samplesMetricMap(classifyr_result)
```



Q6 – model selection

Is the highest predictive performance the only way to choose the best model or can other models be better for other reasons?

Go to

www.menti.com

Enter the code

8474 3969

Or use QR code

Q7 – visualising features

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

Go to

www.menti.com

Enter the code

8474 3969

Or use QR code

Contact

- Dr Yue Cao [yue.cao@sydney.edu.au]
- Mr Nick Robertson [nicholas.robertson@sydney.edu.au]
- Mr Andy Tran [andy.t@sydney.edu.au]

We would love to hear your feedback!

https://forms.office.com/r/XbbcDM7b2b

THE UNIVERSITY OF SYDNEY

Find out more about scdney:
https://sydneybiox.github.io/scdney/