Introduction scRNA-Seq

Thomas D. Otto

Learning aims

- Reflect on different between single cell and bulk RNA-Seq – what are the applications?
- Understand the differences between scRNA-Seq methods and how to apply them
- Explore how to process scRNA-Seq data
- Learn new methods, Seurat, Pseudo time
- Establish your R knowledge
- Critical evaluation of scRNA-Seq

Content

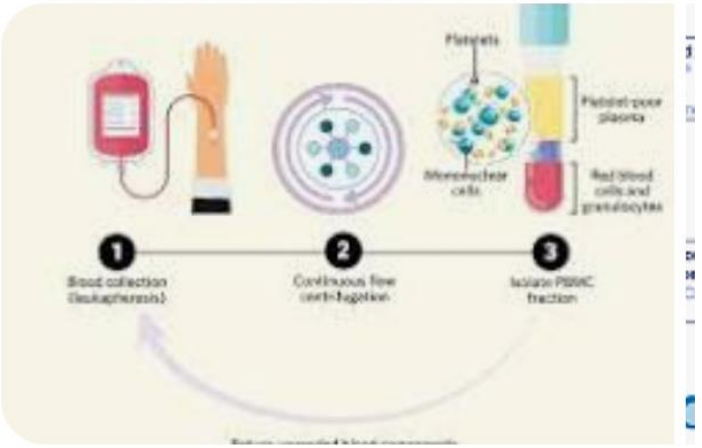
1. Overview: Exercises, Assessment, Lectures, datasets, WHY?
2. Why all of the fuss about single cell? – My approach.
3. Different single-cell technologies
4. Quality control: IGV, Web summary
5. Analysis pipeline

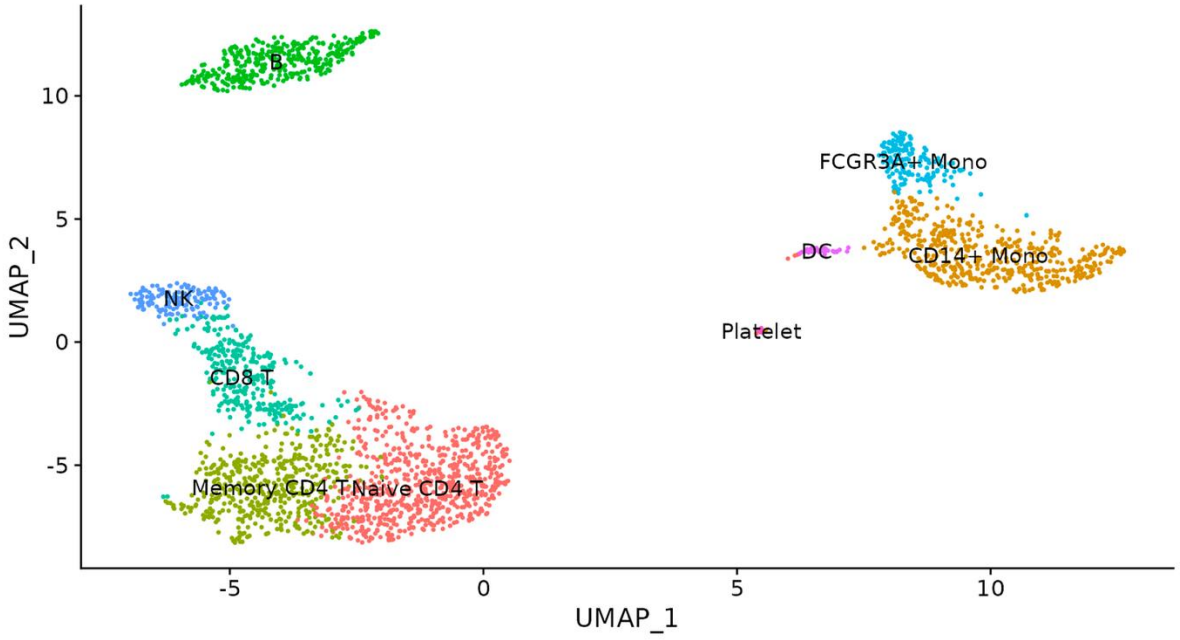
Datasets used: PBMC

Peripheral blood mononuclear cells (PBMC) give selective responses to the immune system and are the major cells in the human body immunity.

National Institutes of Health (NIH) (.gov)
https://www.ncbi.nlm.nih.gov › articles › PMC4673925

Isolated Human Peripheral Blood Mononuclear Cell (PBMC ...





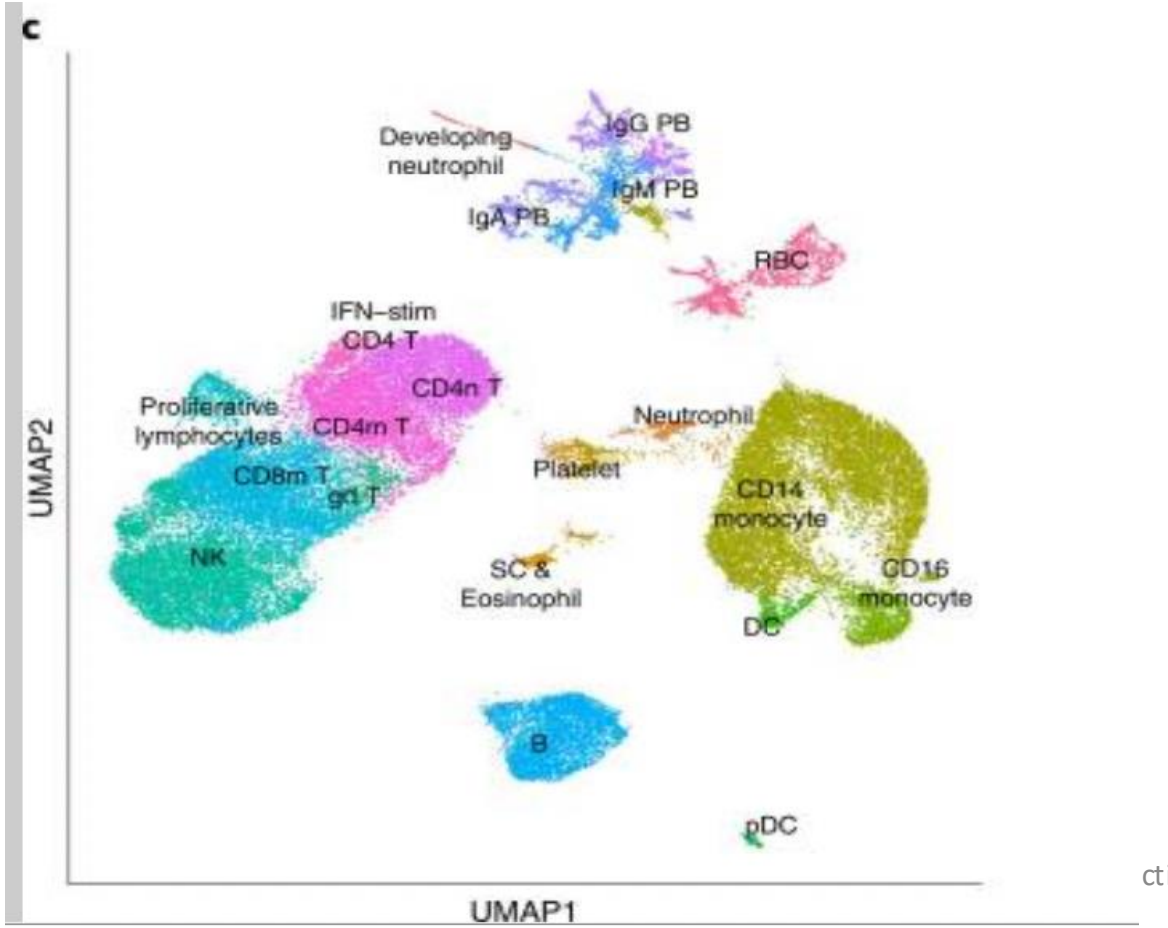
Seurat tutorial

World Health Organization (WHO)
https://www.who.int › Health topics

Coronavirus disease (COVID-19)

Coronavirus disease (COVID-19) is **an infectious disease caused by the SARS-CoV-2 virus**. Most people infected with the virus will experience mild to moderate ...

Coronavirus disease (COVID · 2019 novel coronavirus disease · 22 December 2023



COVID single cell example datase

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7382903/

Discussion, if not clear…

- Why do we do transcriptomics?
- Why should we do scRNA-Seq?
- Is bulk not good enough?

2. My scRNA-Seq thoughts

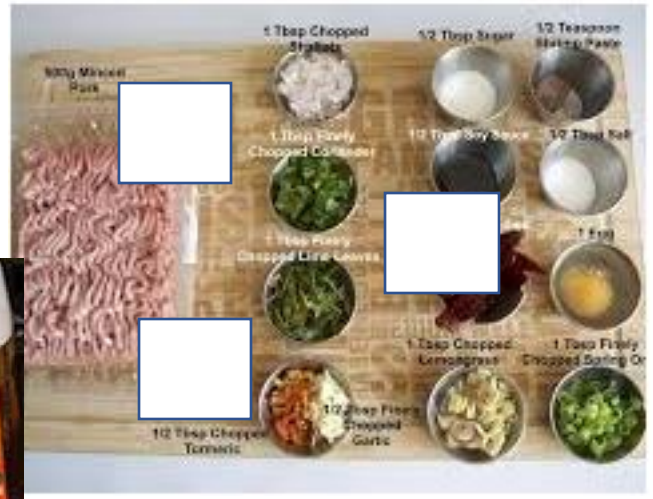
Tastes
different!

How can we
understand
the differences?

My scRNA-Seq thoughts

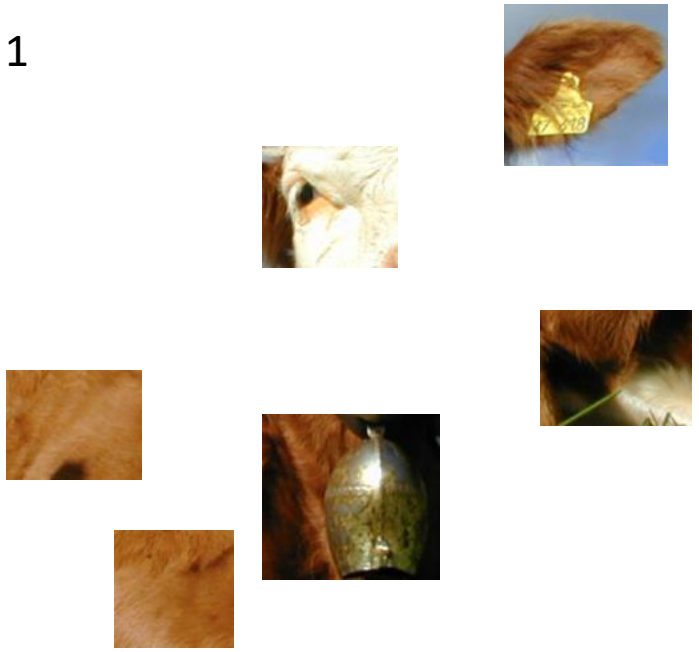
single cell is really noise….

What is similar
What is different

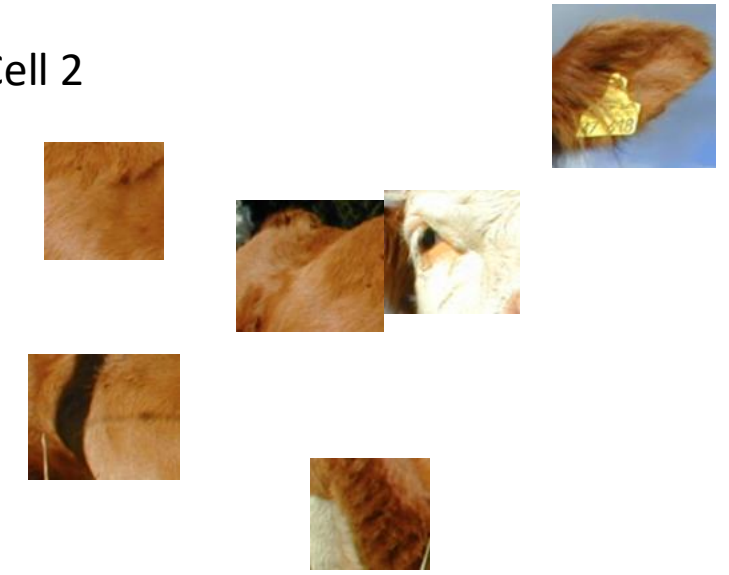


Actually

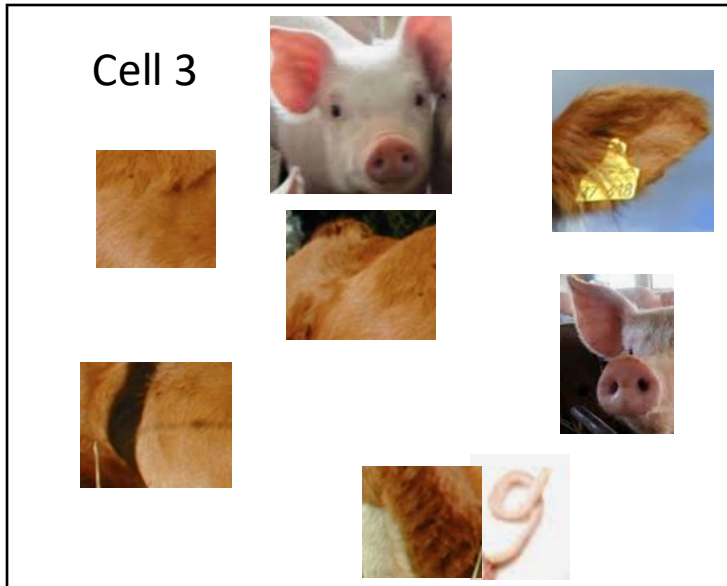
Cell 1



Cell 2



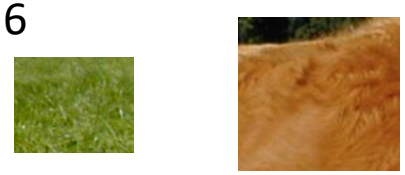
Cell 3



Cell 4

Cell 5

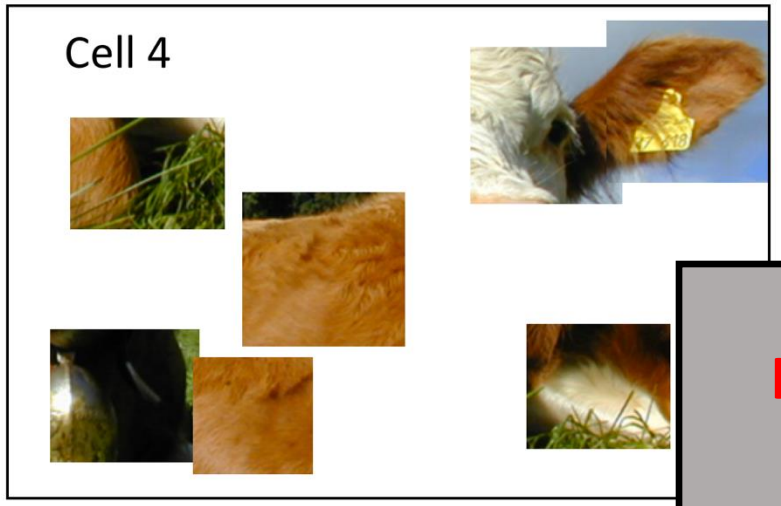
Cell 6



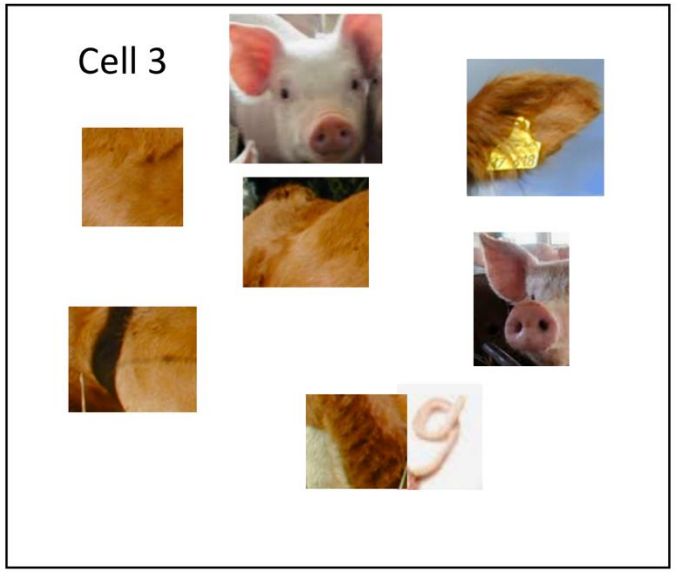
Cell 7

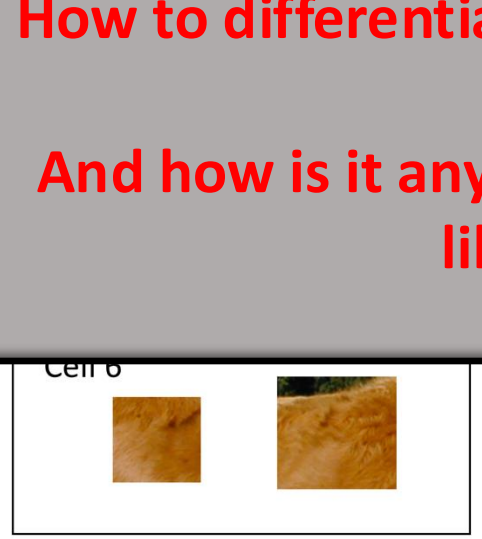
Which gene is highly expressed?

Good cells with 10-40% of the genes "covered" that are expressed

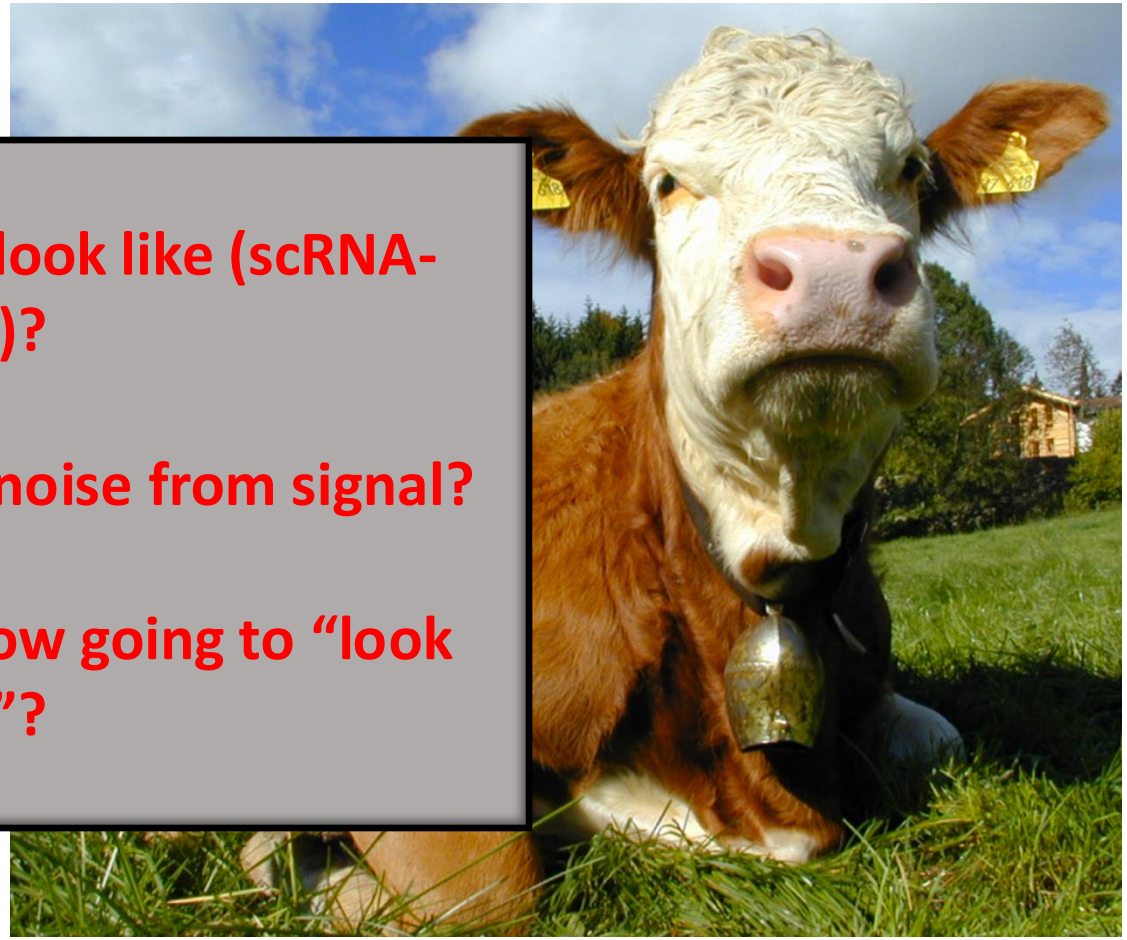


doublets





"bulK"



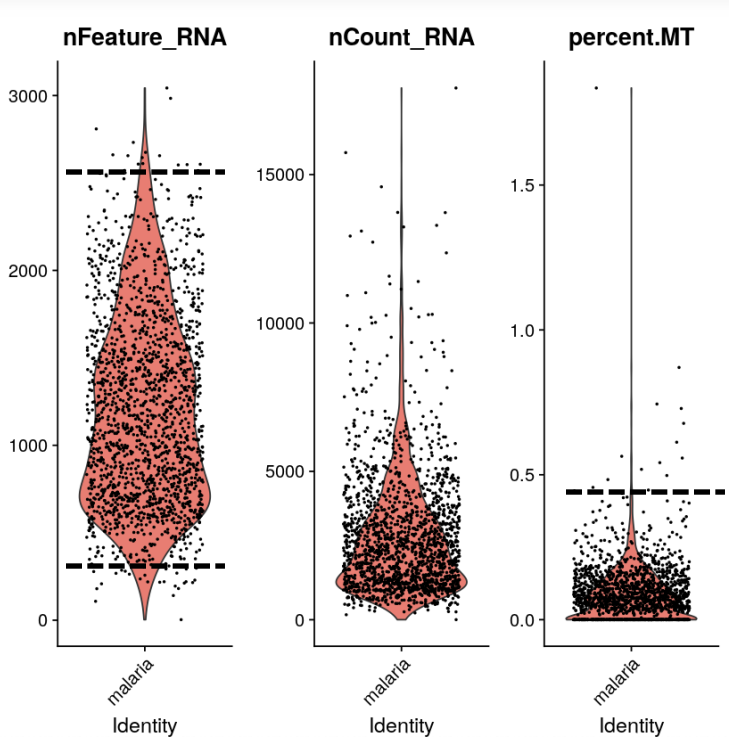
How do dying cell look like (scRNA-Seq)?

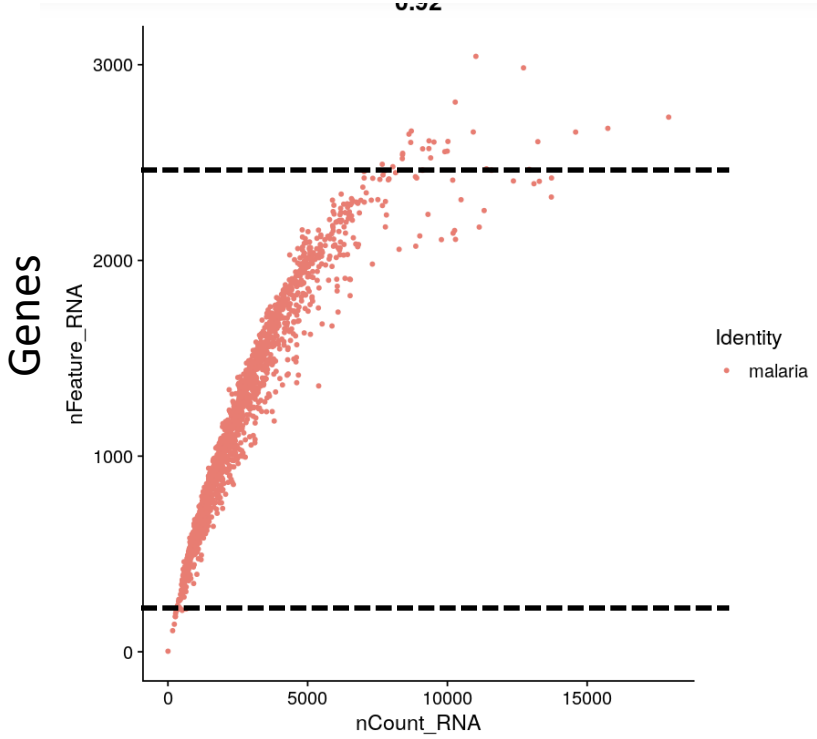
How to differential noise from signal?

And how is it anyhow going to "look like"?

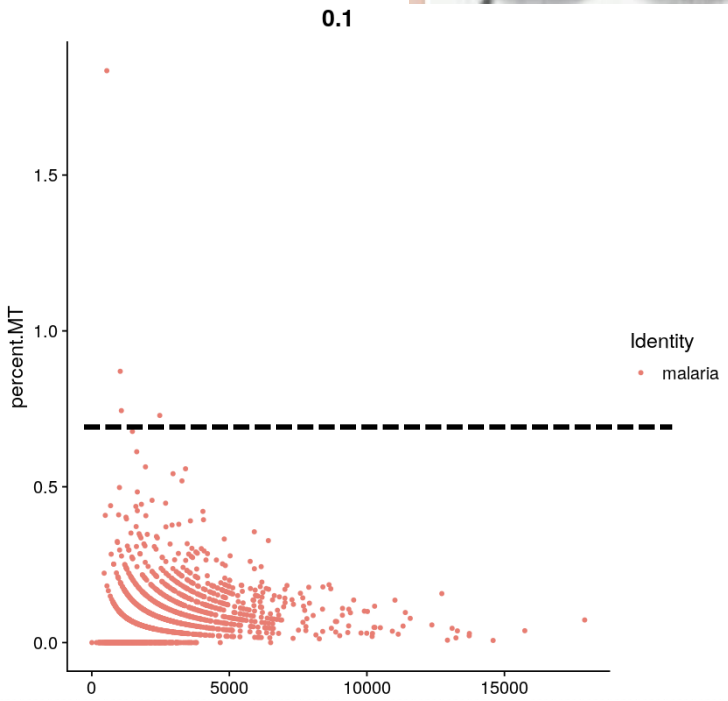
What you need to do!

Genes UMI





UMI



Find Cut-offs

```
BL01 <- subset(BL01, subset = nFeature_RNA > xXx &
            nFeature_RNA < xXx  &
            percent.MT < xXx)
```

https://scampi.mvls.gla.ac.uk

SCAMPI

Single Cell Analysis Methods Presented Interactively

Block Library

- Load Data — Add
- Basic Filtering — Add
- Quality Control Plots — Add
- Quality Controls Filtering — Add
- Identify Highly Variable Genes — Add
- Principal Component Analysis — Add
- Integration — Add
- Run UMAP — Add
- Plot Dimension Reduction — Add

Canvas

Run

Load Data

Dataset: Peripheral Blood Mononuc

Basic Filtering

Minimum Genes Per Cell: 200
Minimum Cells Per Gene: 3

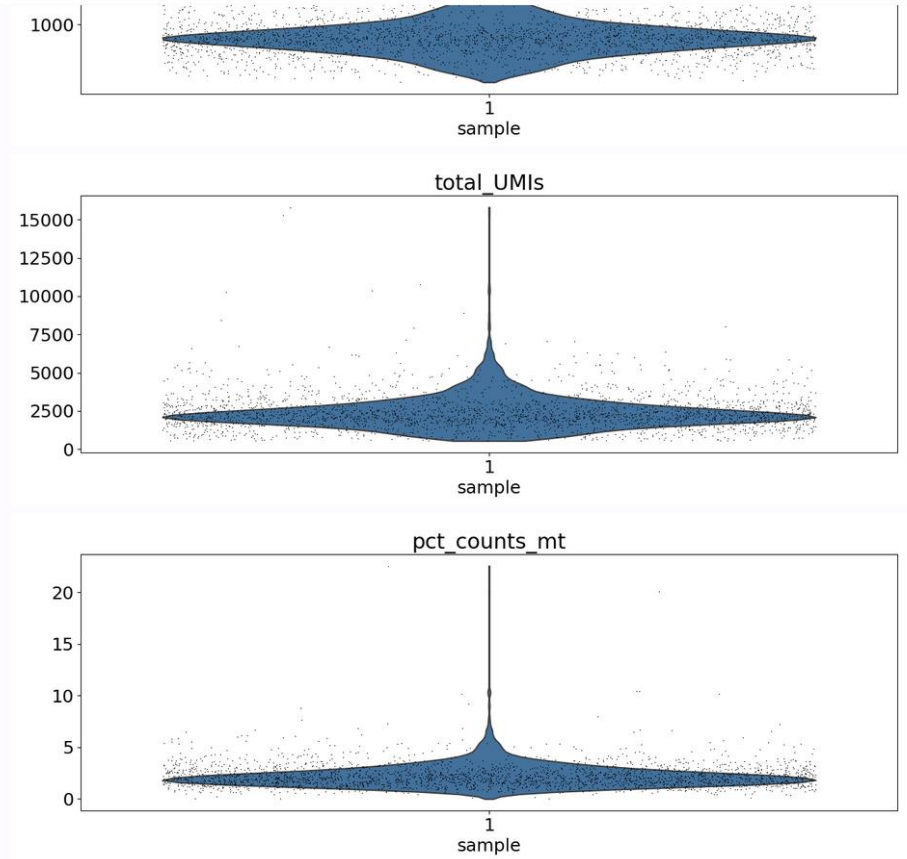
Quality Control Plots

Quality Control Filtering

Sample: 1

Minimum Genes Per Cell: 200
Maximum Genes Per Cell: 2500
Maximum % Mitochondrial Genes: 5

Output

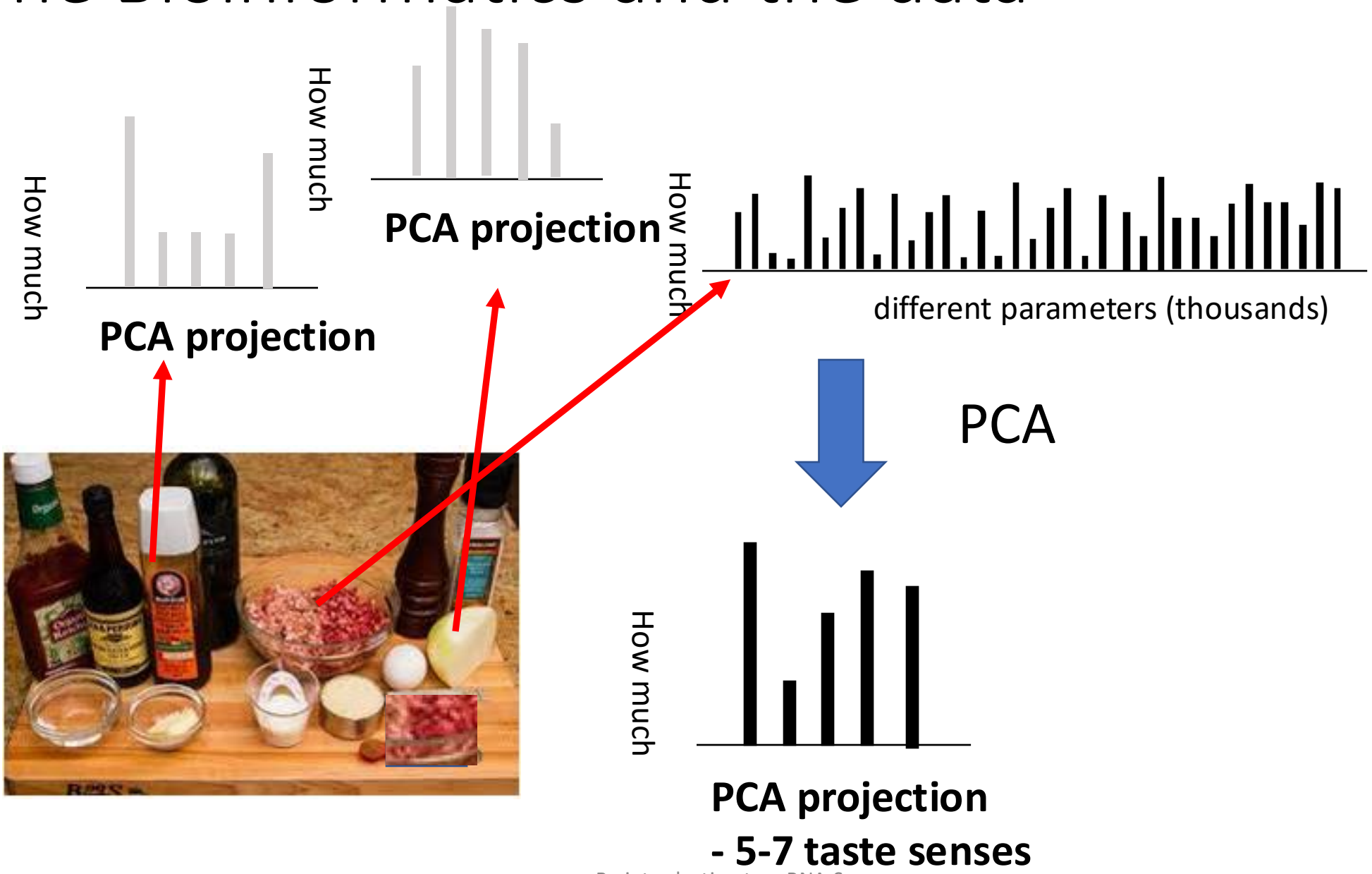


Quality Control Filtering

Object with: 2,638 cells and 13,714 genes

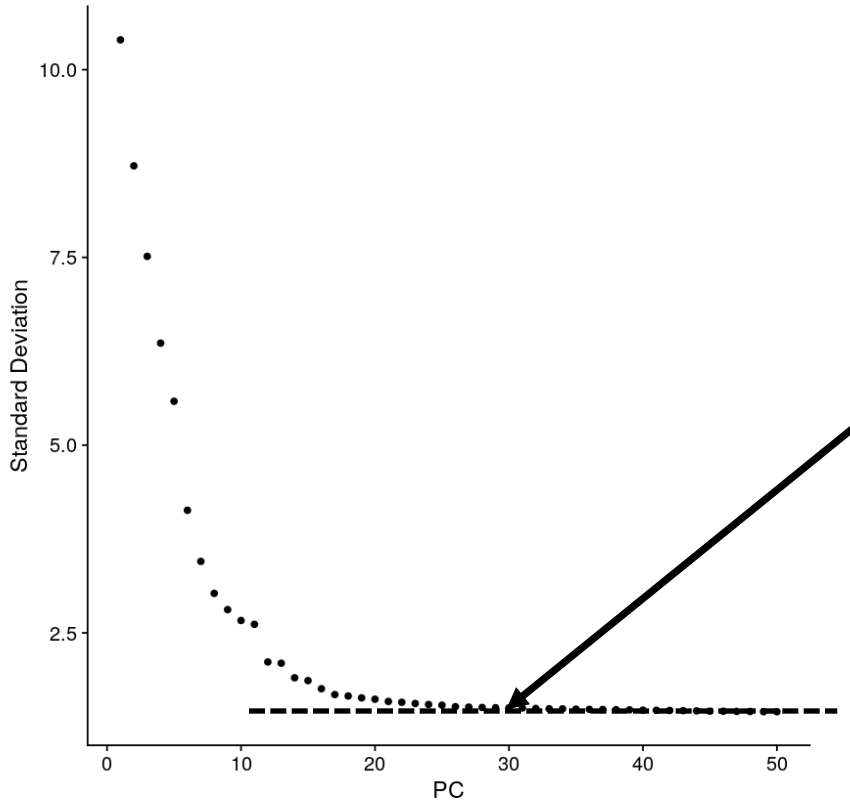
Developed by the Otto Lab

The Bioinformatics and the data



What you need to do - PCA

```
In [9]: ElbowPlot(Combined, ndims = 50)
```




```
In [ ]: # You have to set this variable:
dim=xXx

Combined <- RunUMAP(Combined, reduction = "pca", dims = 1:dim)
Combined <- FindNeighbors(Combined, reduction = "pca", dims = 1:dim)
```

In SCAMPI

Quality Control Filtering

Sample
1

Minimum Genes Per Cell	200
Maximum Genes Per Cell	2500
Maximum % Mitochondrial Genes	5

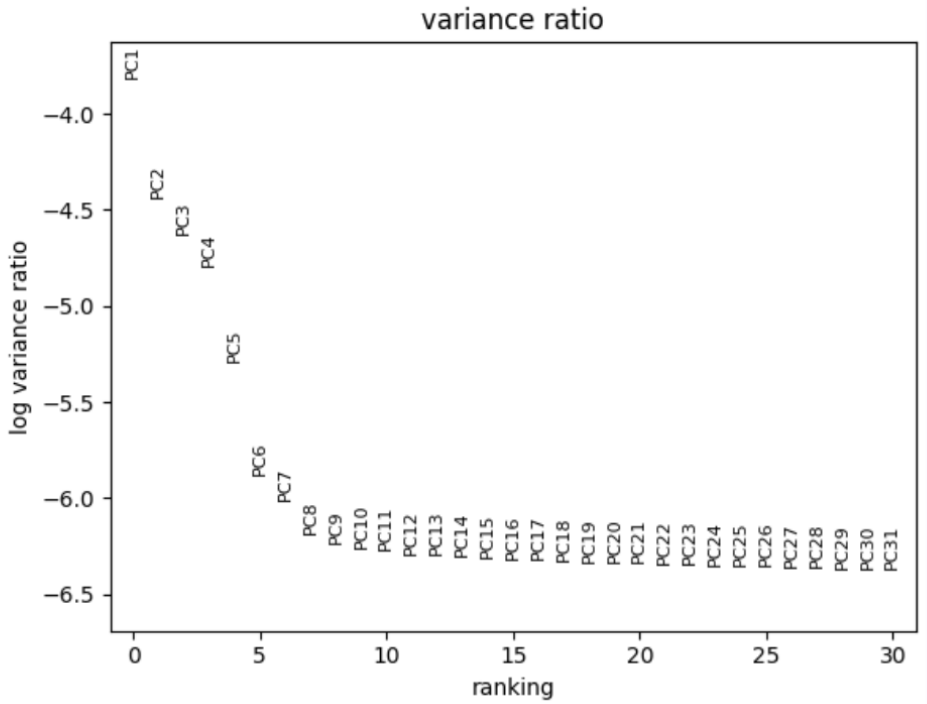
Identify Highly Variable Genes

Minimum Mean	0.0125
Maximum Mean	3
Minimum Dispersion	0.5

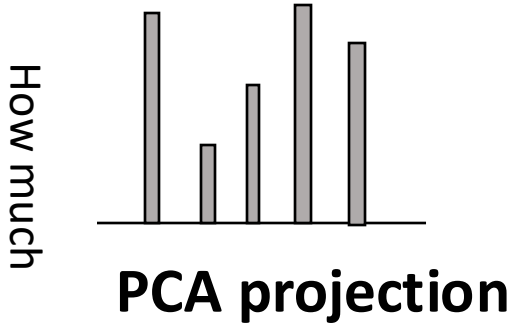
Principal Component Analysis

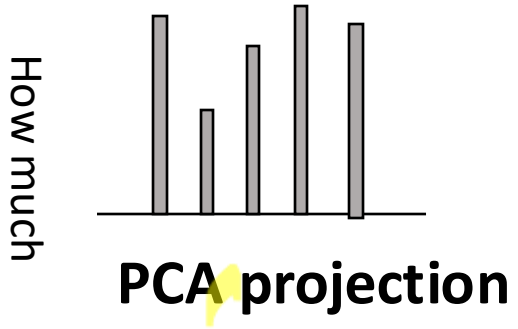
Run UMAP

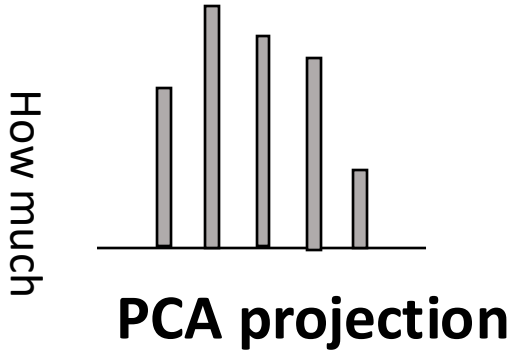
Number of Neighbours	10
Number of Principal Components	40

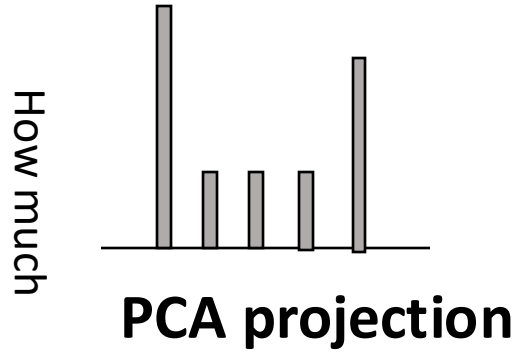


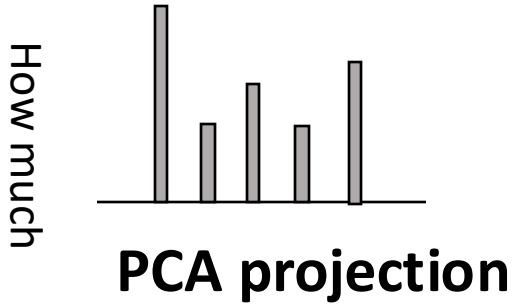
All about visualization (and R)

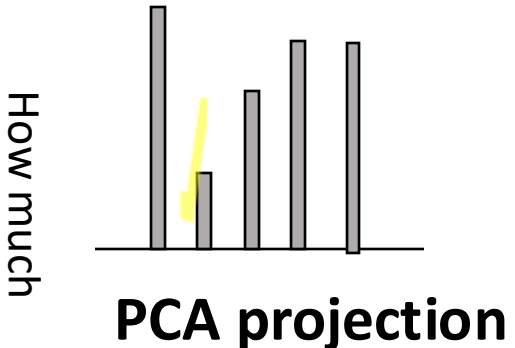






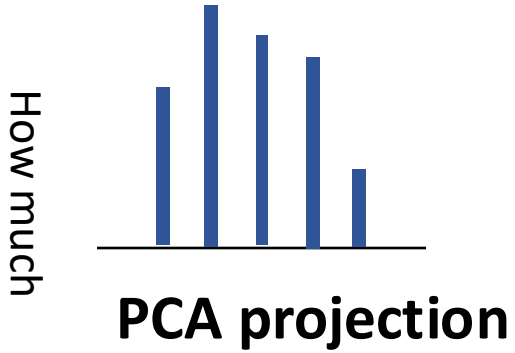






https://en.wikipedia.org/wiki/Taste

All about visualization (and R)

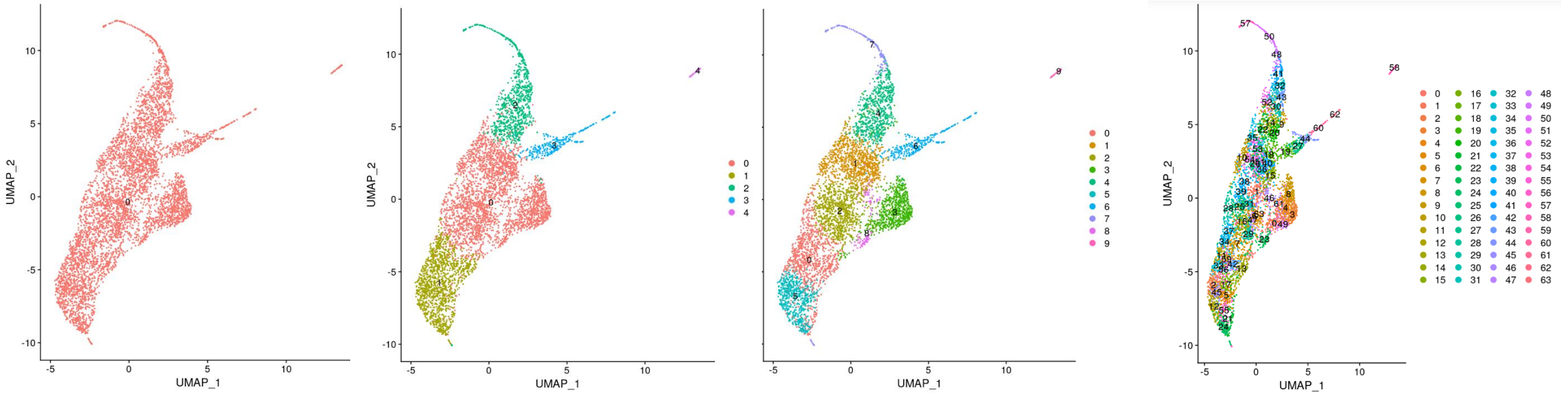


https://en.wikipedia.org/wiki/Taste

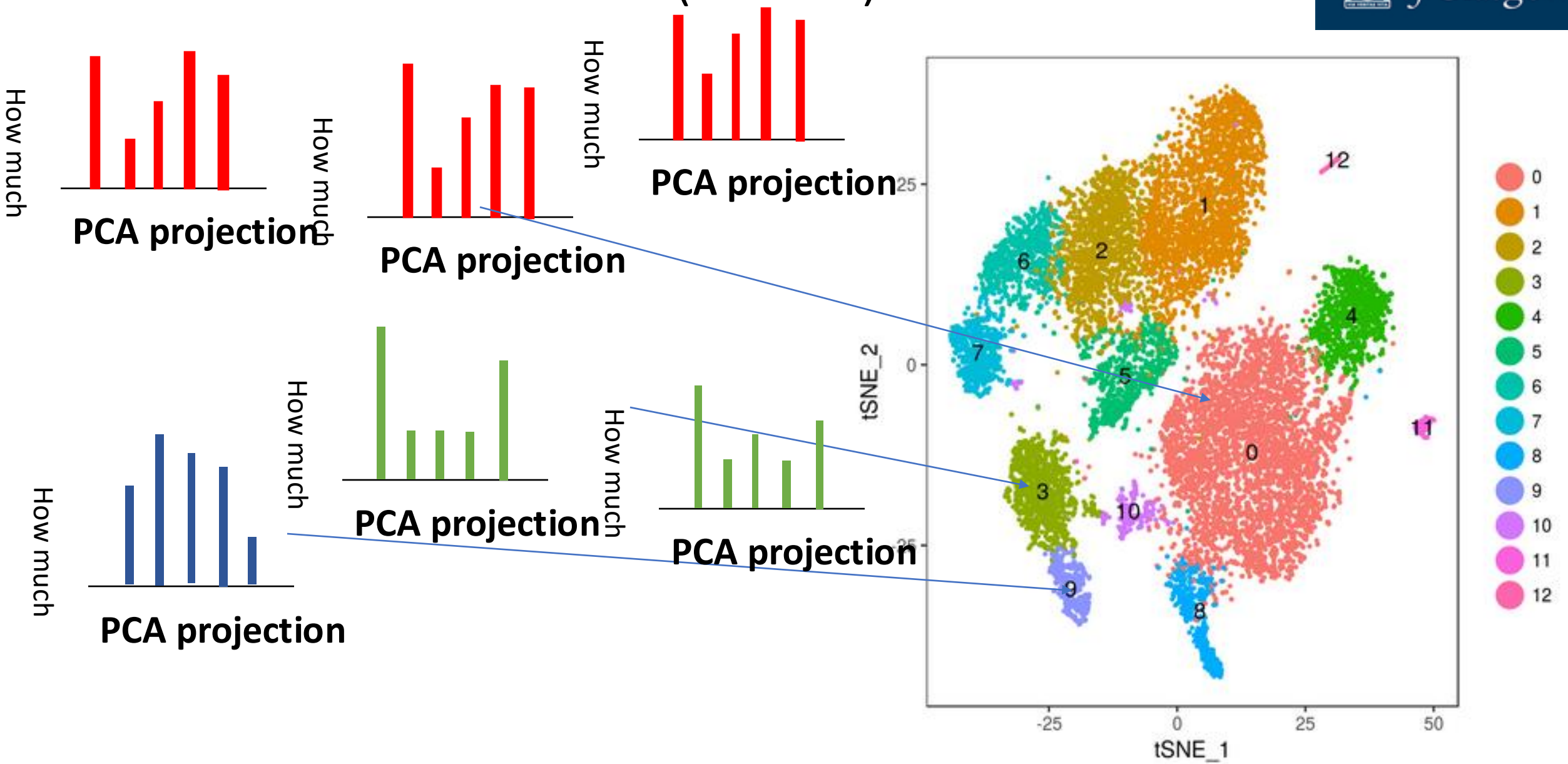
What you need to do - Clustering

```
In [ ]: #let's plot the UMAP.
        DefaultAssay(Combined) <- "integrated"
        # you have to set the resolution!
        Combined <- FindClusters(Combined, resolution = xXx)
```

0 0.1 0.2 10

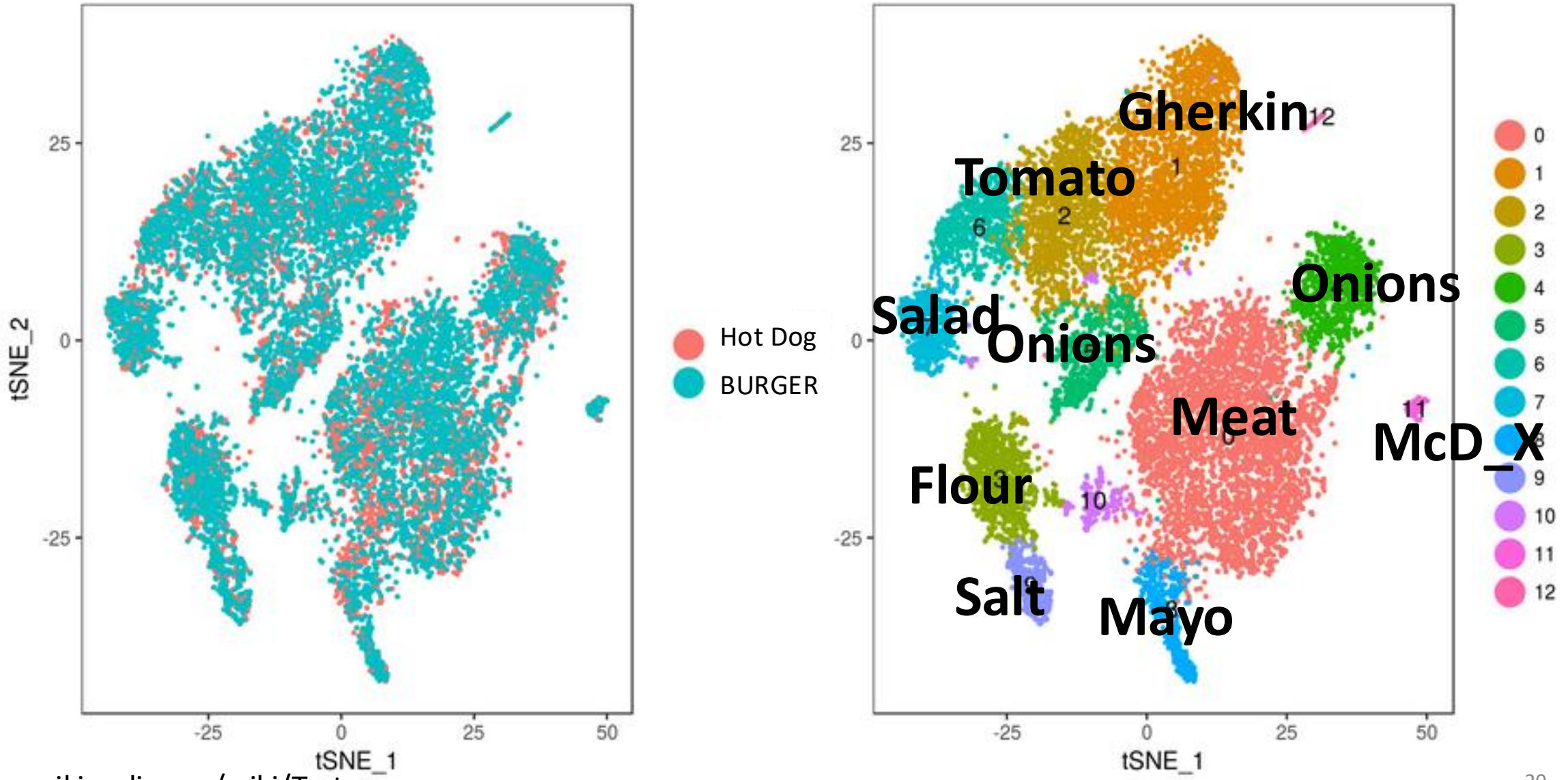


All about visualization (and R)



https://en.wikipedia.org/wiki/Taste

All about visualization (and R)



https://en.wikipedia.org/wiki/Taste

In SCAMPI

Plot Dimension Reduct

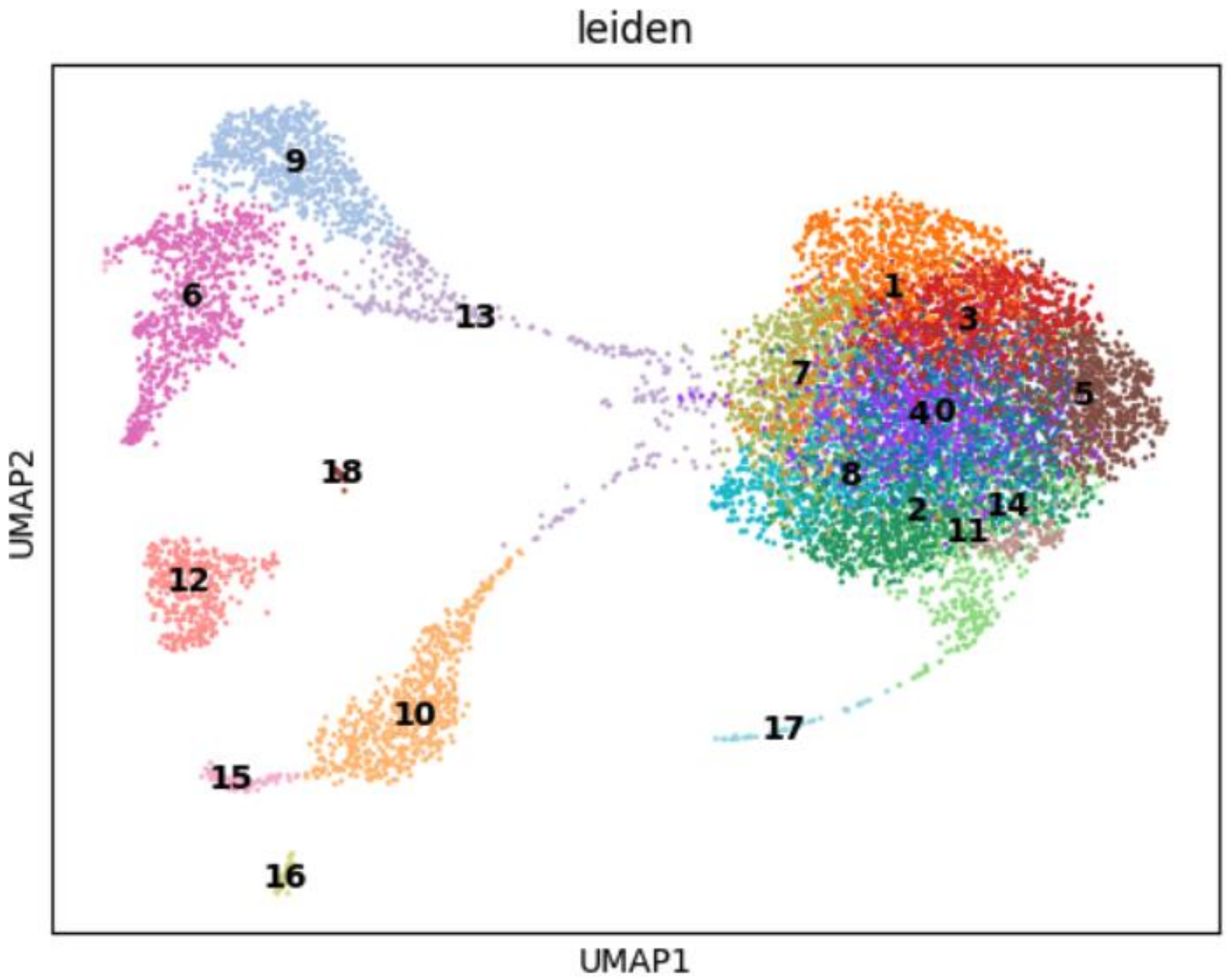
Color By
leiden

Reduction
PCA

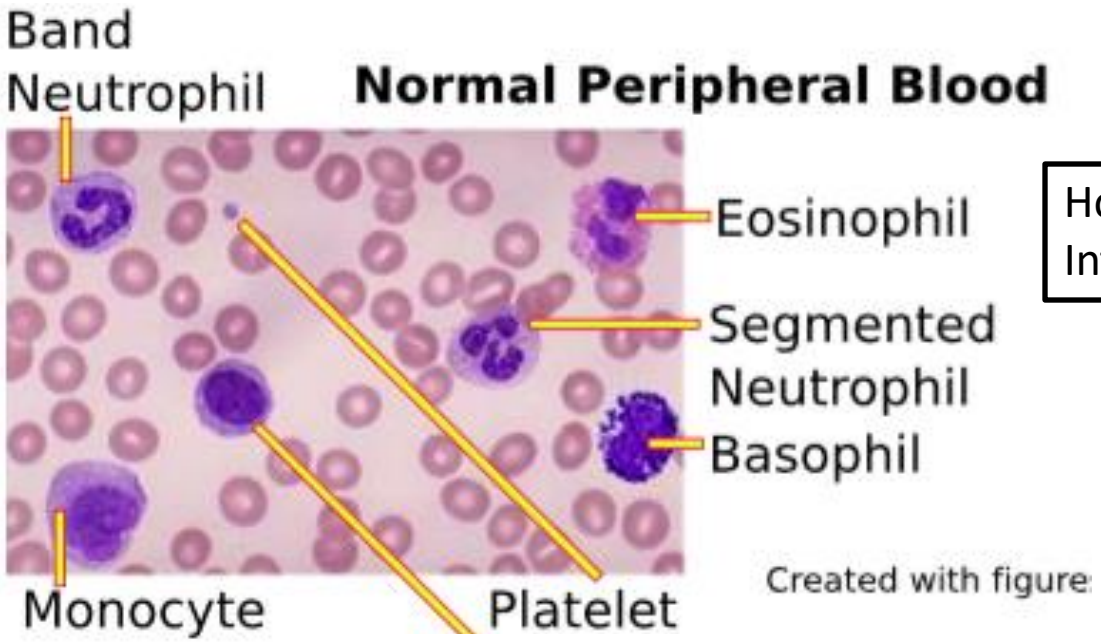
Plot Dimension Reduct

Color By
leiden

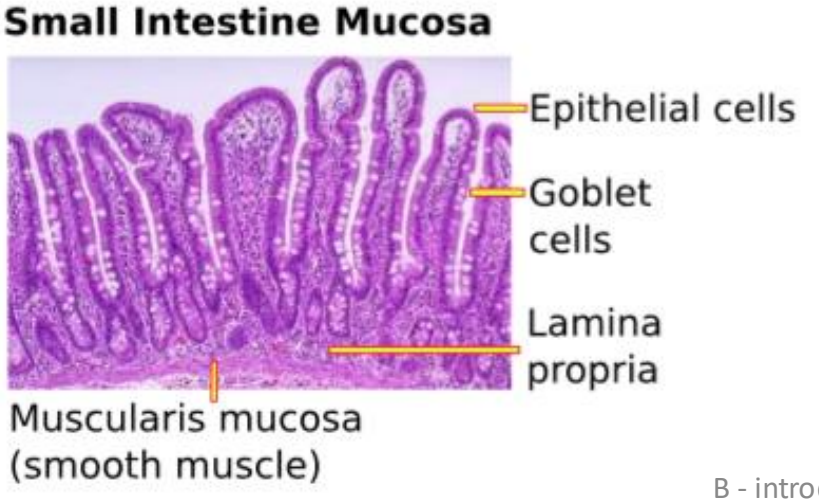
Reduction
TSNE



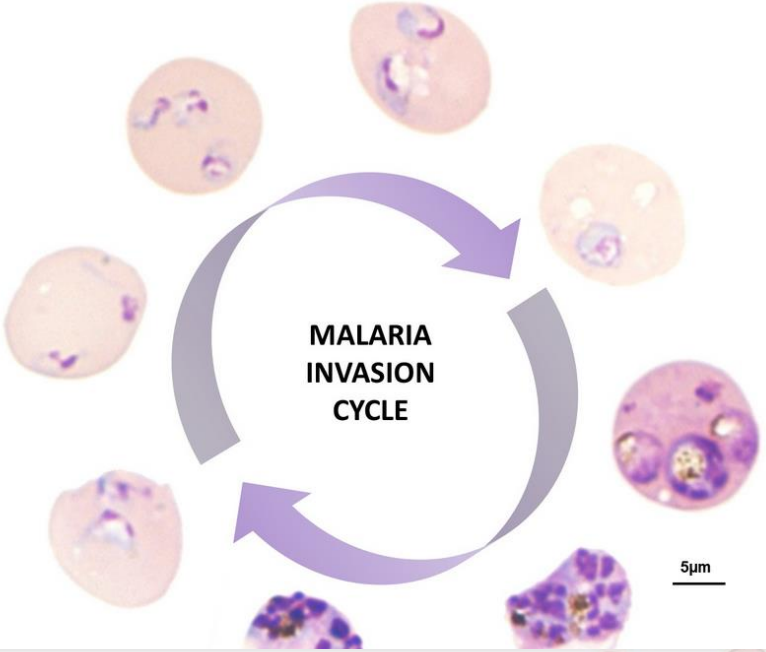
Why so powerful?



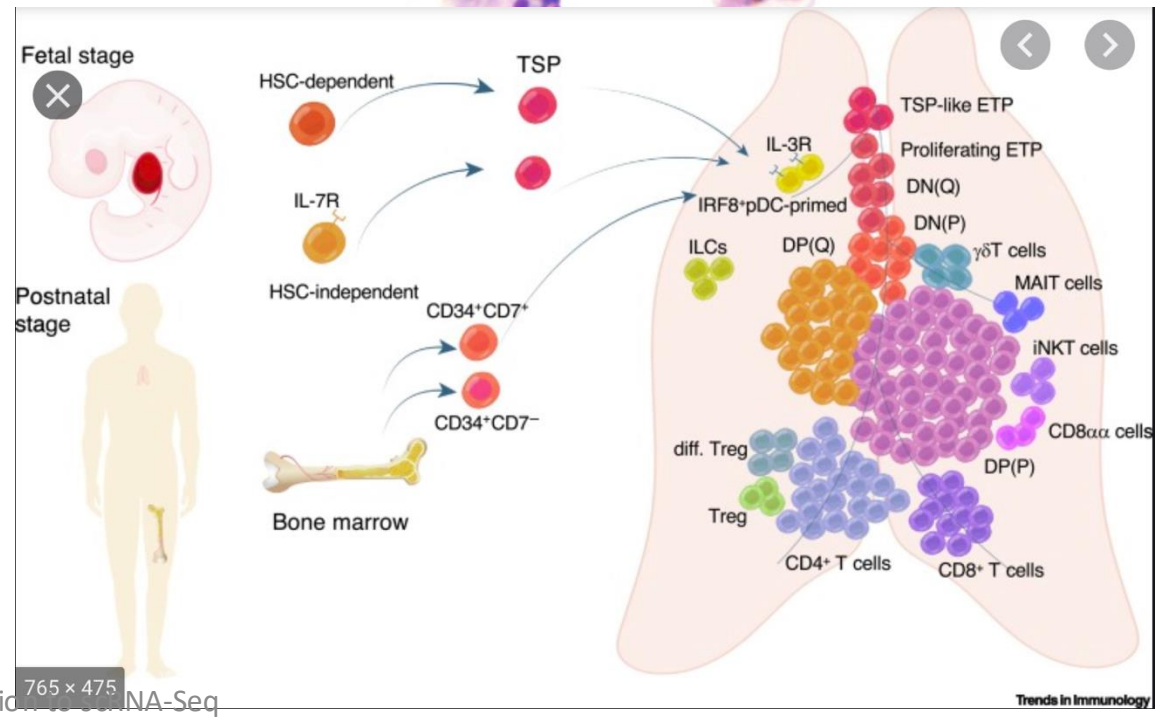
library.med.utah



Host:Pathogen Interaction



Source: nhsbt.nhs.uk

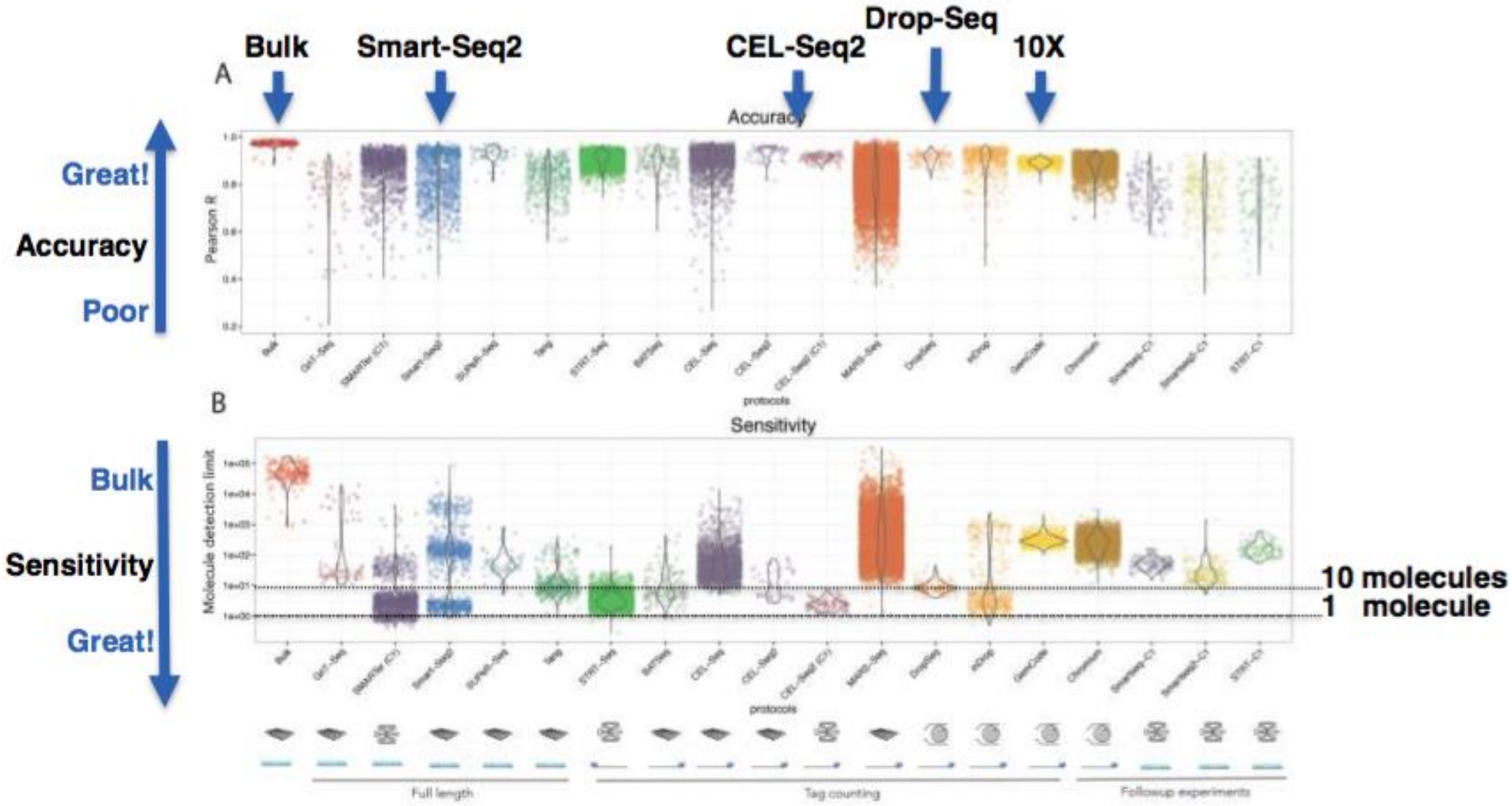


Development of T-cells

3. Different technologies

scRNA-Seq methods

- Show the methods
- a lot, focus o



Again:

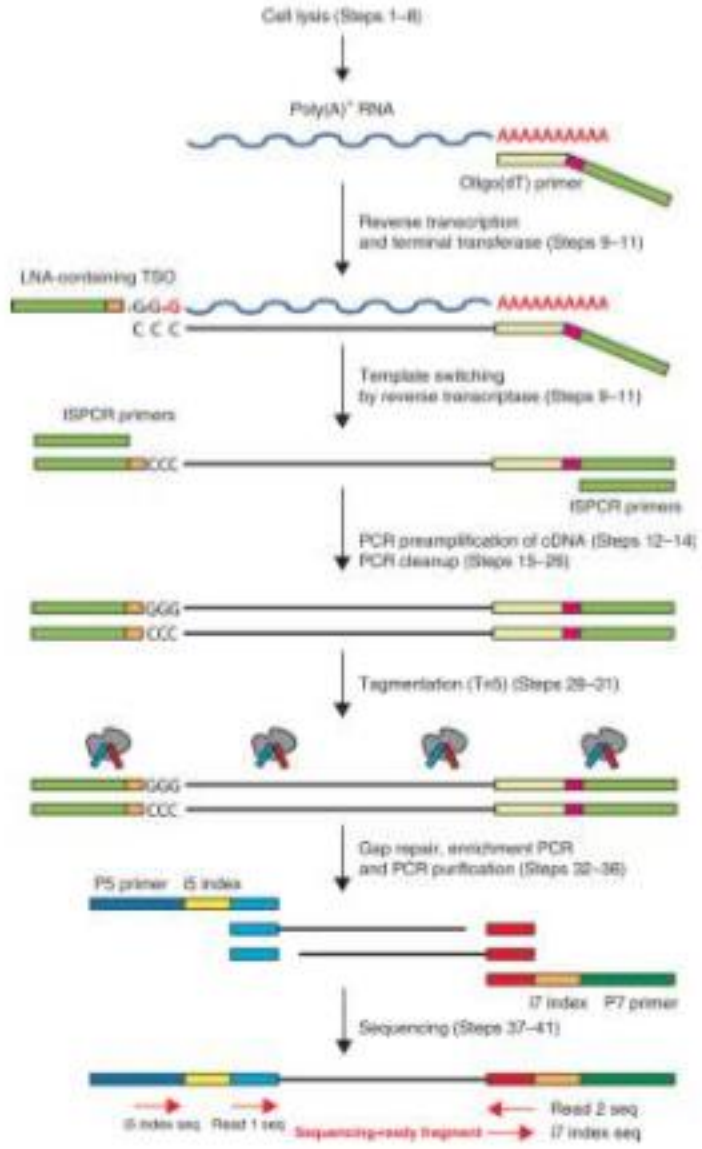
- How can we get to one cell?
- How can we sequence most of the genes of the cell?

  - As many genes as possible!
  - As many cells as possible!

smartSeq2

- Full transcript scRNA-Seq
- Developed for single cell but can performed using total RNA.
- Selects for poly-A tail.
- Full transcript assay.
- – Uses template switching for 5' end capture.
- • Standard illumina sequencing. – Off-the-shelf products.
- Hundreds of samples.
- Often do not see UMI used.

- Poly-A capture with 30nt polyT and 25nt 5' anchor sequence.
- RT adding untemplated C
- Template switching
- Locked Nucleic Acid binds to untemplated C
- RT switches template
- Preamplification / cleanup
- DNA fragmentation and adapter ligation together.
- Gap Repair, enrich, purify.



Equipment

What could be the drawback?

What could be the drawback?

- Few cells
- A bit of work

- On the plus, little dropout
- Coverage of the whole transcript

Drop-seq

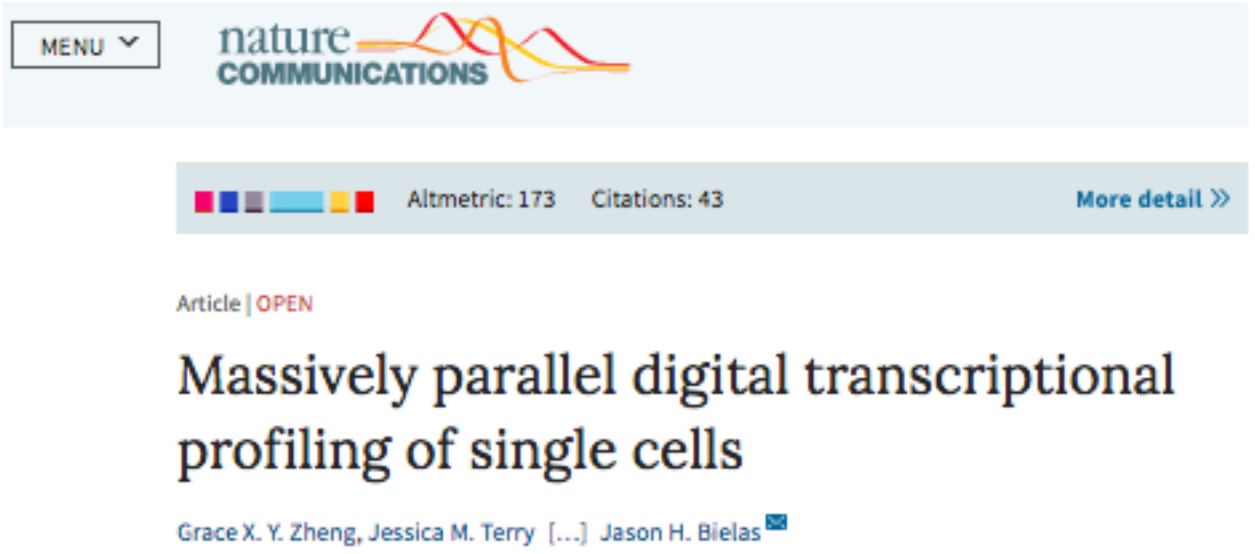
- Moved throughput from hundreds to thousands.
- Droplet-based processing using microfluidics
- Nanoliter scale aqueous drops in oil.
- 3'End
- Bead based (STAMPs).
- Single-cell transcriptomes attached to microparticles.
- Cell barcodes use split-pool synthesis.
- Uses UMI (Unique Molecular Identifier).
- RMT (Random Molecular Tag).
- Degenerate synthesis.

Let’s have a look

- https://ars.els-cdn.com/content/image/1-s2.0-S0092867415005498-mmc8.mp4
- (still working, testing 10/03/2024)

10x massive sequencing

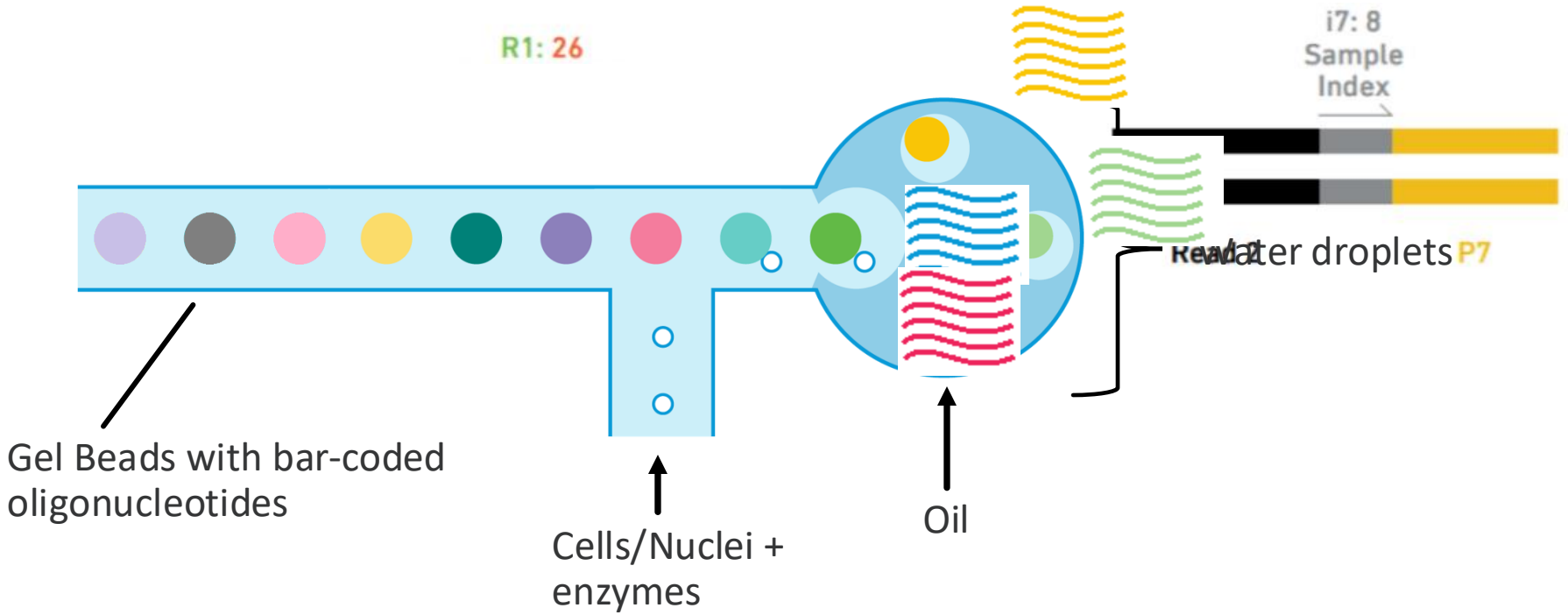
- Droplet-based, 3' mRNA.
  – GEM (Gel Bead in Emulsion)
- Standardized instrumentation and reagents.
- More high-throughput scaling to tens of thousands.
- Less processing time.
- Cell Ranger software is available for install.



Working Principle of **G**elbead **Em**ulsions (**GEM**)

Thanks to Stephen Hague, 10X genomics

Working Principle of **G**elbead **Em**ulsions (**GEM**)



Thanks to Stephen Hague, 10X genomics

10X sequencing (+/-)

- up to 10.000 cells (more possible, but doublets)
- 800-7500 individua
- 500-2500 genes pe expressed in a cells???)
- relative expensive (
- Need to be sequenc 30-50k reads per ce

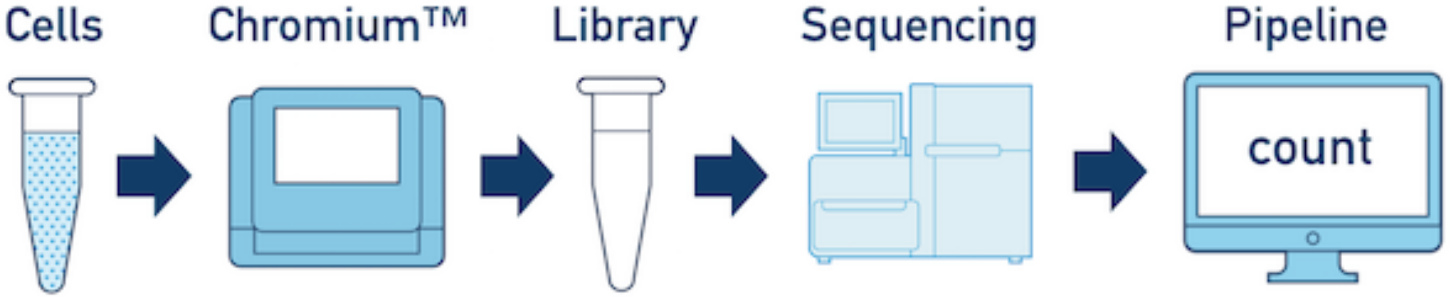
How deep to sequence to get 2000-5000 UMI?

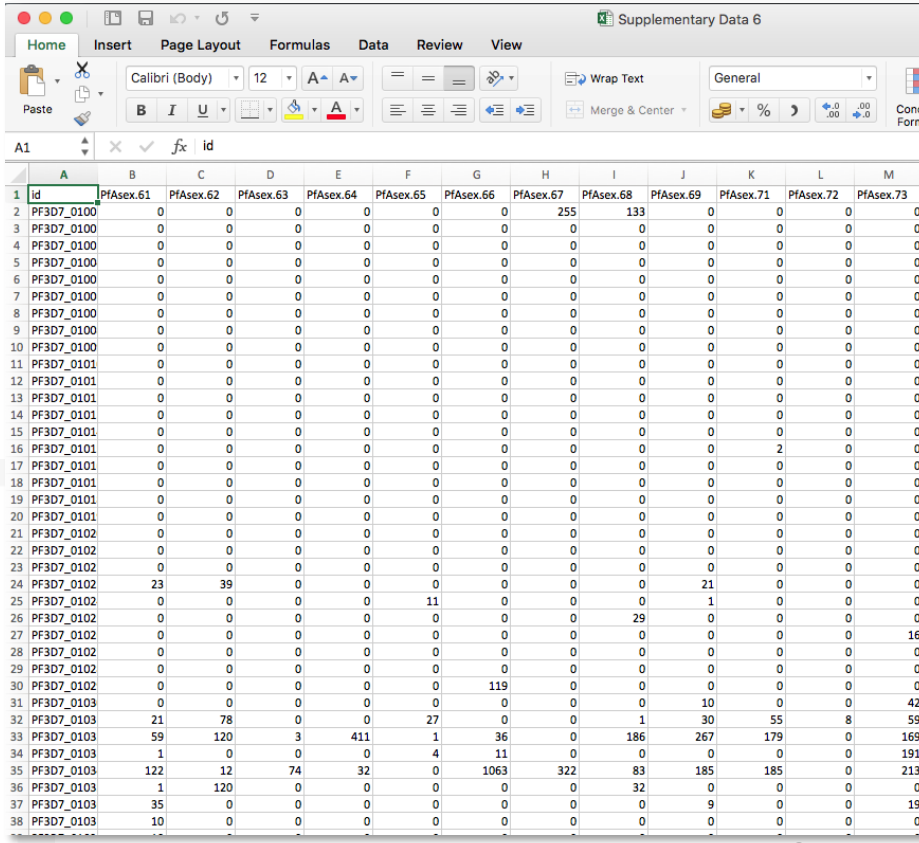
- 3’ ENRICHED
  - Partial coverage
  - No splicing

Output of single cell data is “HUGE”

- A table of ~20 thousands rows and ~10 thousands cells per run
- Each cell (column) is described by ~2,500 genes (row)
- Difficult to visualise this

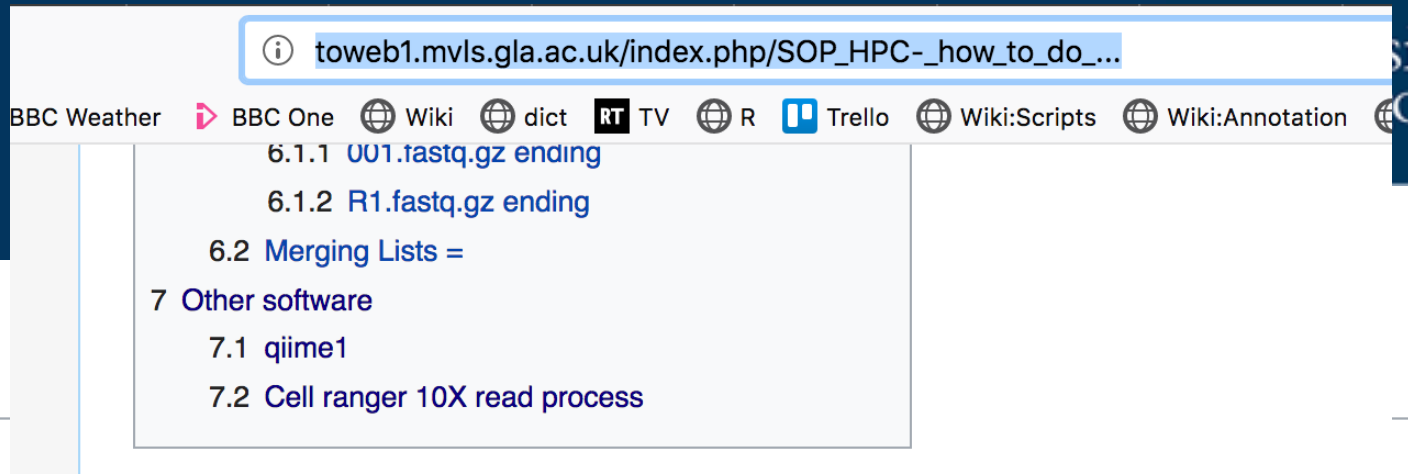
https://www.youtube.com/watch?v=4NAS1qTJmYA





	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	PfAsex.61	PfAsex.62	PfAsex.63	PfAsex.64	PfAsex.65	PfAsex.66	PfAsex.67	PfAsex.68	PfAsex.69	PfAsex.71	PfAsex.72	PfAsex.73
2	PF3D7_0100	0	0	0	0	0	0	255	133	0	0	0	0
3	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0
4	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0
5	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0
6	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0
7	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0
8	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0
9	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0
10	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0
11	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0
12	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0
13	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0
14	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0
15	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0
16	PF3D7_0101	0	0	0	0	0	0	0	0	0	2	0	0
17	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0
18	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0
19	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0
20	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0
21	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	0
22	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	0
23	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	0
24	PF3D7_0102	23	39	0	0	0	0	0	0	21	0	0	0
25	PF3D7_0102	0	0	0	0	11	0	0	0	1	0	0	0
26	PF3D7_0102	0	0	0	0	0	0	0	29	0	0	0	0
27	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	16
28	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	0
29	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	0
30	PF3D7_0102	0	0	0	0	0	119	0	0	0	0	0	0
31	PF3D7_0103	0	0	0	0	0	0	0	0	10	0	0	42
32	PF3D7_0103	21	78	0	0	27	0	0	1	30	55	8	59
33	PF3D7_0103	59	120	3	411	1	36	0	186	267	179	0	169
34	PF3D7_0103	1	0	0	0	4	11	0	0	0	0	0	191
35	PF3D7_0103	122	12	74	32	0	1063	322	83	185	185	0	213
36	PF3D7_0103	1	120	0	0	0	0	0	32	0	0	0	0
37	PF3D7_0103	35	0	0	0	0	0	0	0	9	0	0	19
38	PF3D7_0103	10	0	0	0	0	0	0	0	0	0	0	0

Linux processing



Cell ranger 10X read process

10X bla bla First start an interactive session (start.sh)

Set the path

```
source /export/projects/bioinfo2/to16r/bin/cellranger-2.1.0/sourceme.bash
```

Go to your directory. The first command will get the fastq files from the illumina results. It does the base calling. Normally, polyomics will do that step for you.

```
cellranger mkfastq --id=149 --input-dir=/export/projects/bioinfo2/to16r/Projects/scRNA-Seq_polyomics/DataSet2
/HS319_0085 --samplesheet=HS319_0085_10X.csv
```

```
cellranger count --id=149 --fastq=149/outs/fastq_path/ --transcriptome=/export/projects/bioinfo2/to16r/References/Human
/Cellranger/refdata-cellranger-GRCh38-1.2.0/ --sample=149 &> out.149.txt
```

This will generate the count matrix and also the web output.

Two types of methods

- SmartSeq2:
  - Full length representation
  - Alternative splicing
  - Few cells
  - Good for homogeneous cells - T-cells / parasites
- Drop-Seq / 10x
  - Just counting genes (UMI)
  - A lot of cells – poorer coverage?
  - Good for heterogeneous cells: Blood, joints, brain

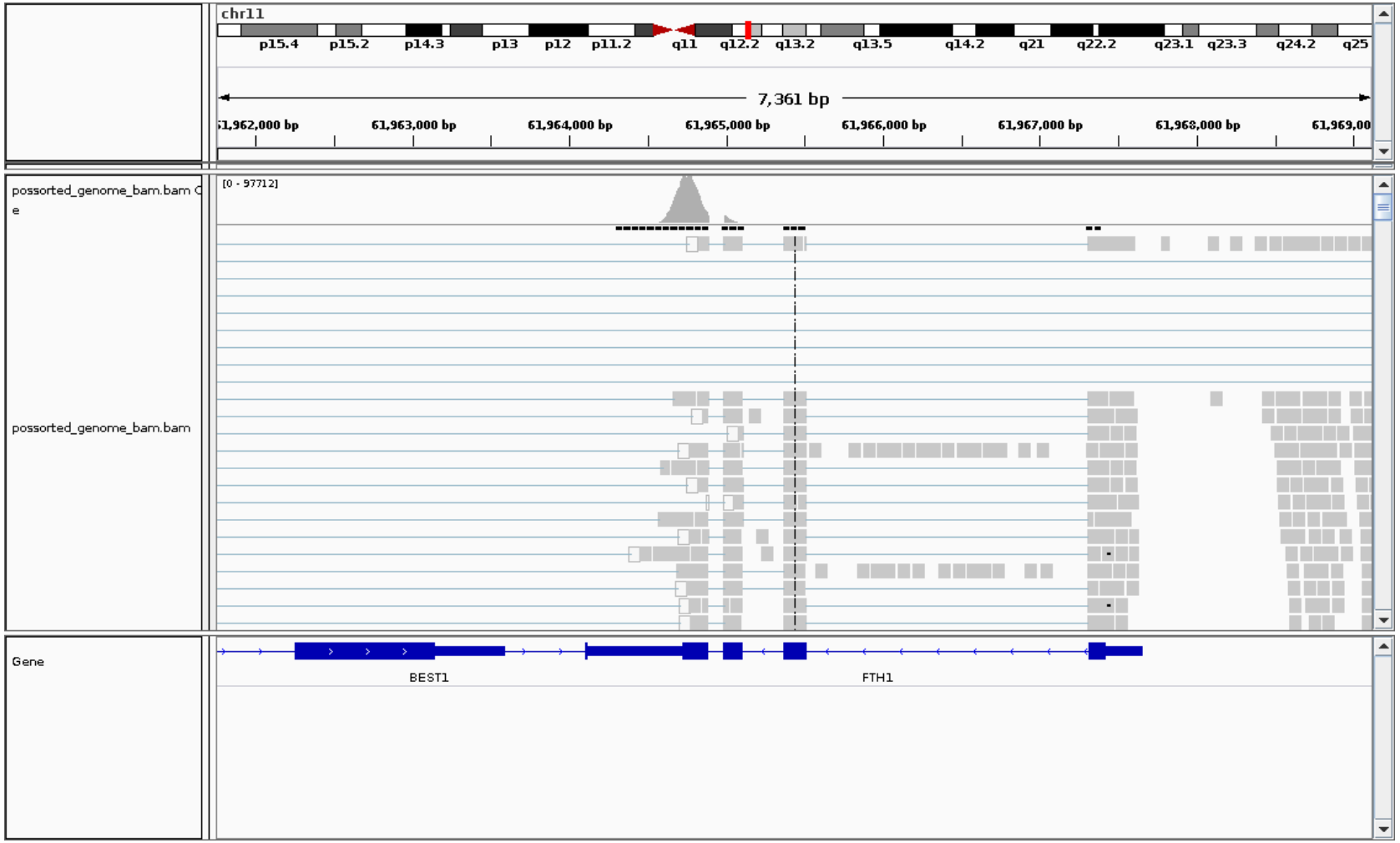
Conclusion 1

- scRNA-Seq is a powerful method to capture the expression profile of individual cell
- Technical challenges, noise, drop out, singletons
- Two main methods for homogeneous versus heterogeneous cell populations
- Expensive, but good?

- Need to analyse the data

4. Quality control (QC)

View of mapped file in IGV (FTH1)



Position in genome

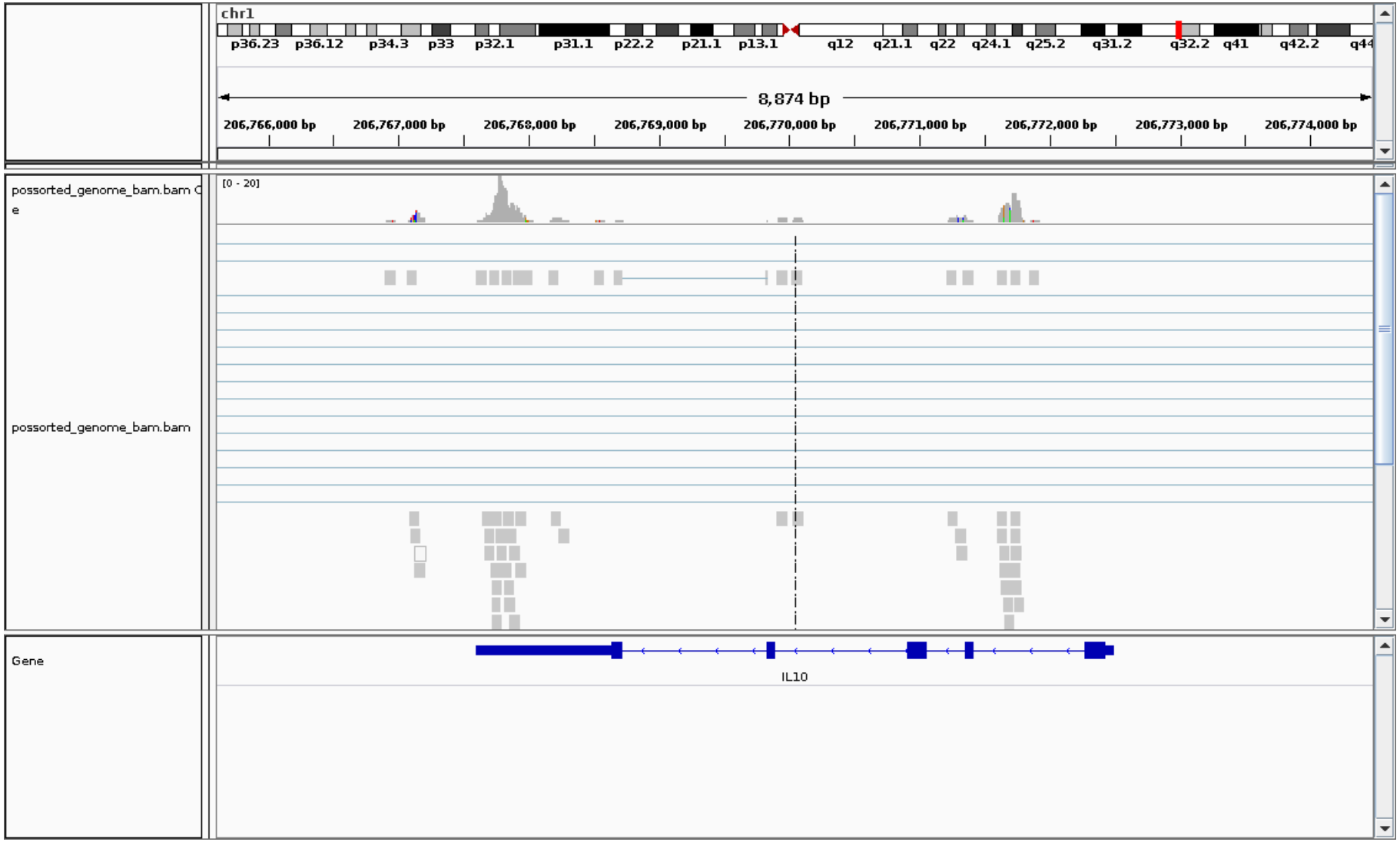
Coverage plot

Mapped reads

2 genes
UTR, Introns, direction

FTH1 is well expressed. Clear peak at 3' end

IGV – IL10



Low expressed
Noise in intron

QC: Web summary

- https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k

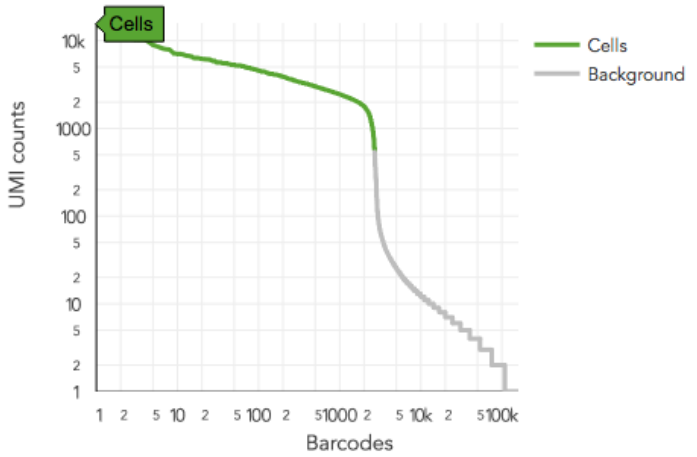
Number of Cells
2,700

Mean Reads per Cell	Median Genes per Cell
68,881	817

Sequencing

Number of Reads	185,980,783
Valid Barcodes	96.3%
Reads Mapped Confidently to Transcriptome	66.0%
Reads Mapped Confidently to Exonic Regions	69.4%
Reads Mapped Confidently to Intronic Regions	15.9%
Reads Mapped Confidently to Intergenic Regions	3.5%
cDNA PCR Duplication	94.0%
Q30 Bases in Barcode	70.9%
Q30 Bases in Read 1	93.0%
Q30 Bases in Sample Index	97.2%
Q30 Bases in UMI	96.0%

Cells



Number of Cells	2,700
Fraction Reads in Cells	96.8%
Mean Reads per Cell	68,881
Median Genes per Cell	817
Median UMI Counts per Cell	2,197

Sample

Name	pbmc3k
Description	Peripheral blood mononuclear cells (PBMCs) from a healthy donor
Transcriptome	hg19

Better run (more recent)

10,194
Estimated Number of Cells

69,859
Mean Reads per Cell

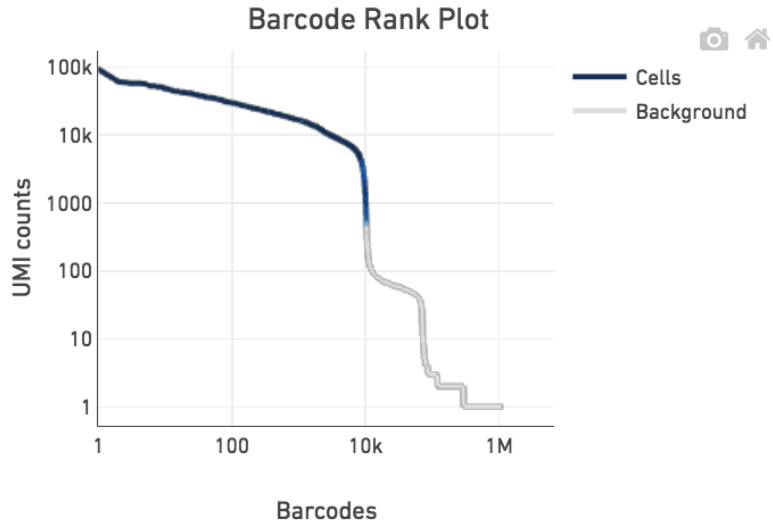
2,152
Median Genes per Cell

Sequencing

Number of Reads	712,144,074
Valid Barcodes	98.3%
Valid UMIs	99.9%
Sequencing Saturation	74.1%
Q30 Bases in Barcode	95.6%
Q30 Bases in RNA Read	92.5%
Q30 Bases in UMI	95.3%

Mapping

Cells



Estimated Number of Cells	10,194
Fraction Reads in Cells	95.6%
Mean Reads per Cell	69,859
Median Genes per Cell	2,152
Total Genes Detected	25,199
Median UMI Counts per Cell	7,697

Buzz group task

- To understand better the output of CellRanger we will form groups of 3-4 students
- You have ~1 minute to discuss, if
  - How good is the run?
  - Is the number of cells/genes good?
  - Is the number of reads high enough?
  - What could be improved?
  - Would it need a rerun?

Discuss 1

- How good is the run?
- Is the number of cells/ genes good?
- Is the number of reads high enough?
- What could be improved?
- Would it need a rerun?

The analysis detected some issues. Details »

Alert	Value	Detail
⚠ **Low Fraction Reads in Cells**	59.1%	Ideal > 70%. Application performance may be affected. Many of the reads were not assigned to cell-associated barcodes. This could be caused by high levels of ambient RNA or by a significant population of cells with a low RNA content, which the algorithm did not call as cells. The latter case can be addressed by inspecting the data to determine the appropriate cell count and using --force-cells.

Estimated Number of Cells
694

Mean Reads per Cell	Median Genes per Cell
159,922	2,276

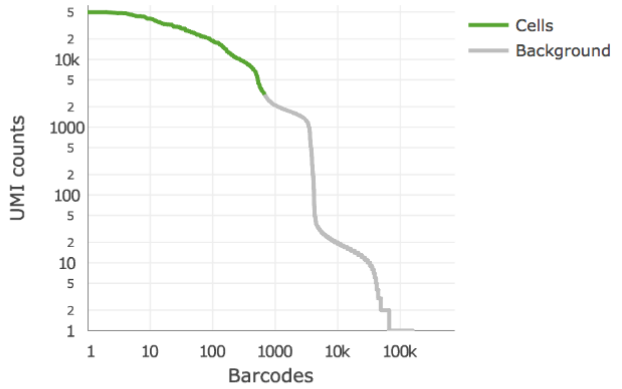
Sequencing

Number of Reads	110,986,438
Valid Barcodes	97.3%
Sequencing Saturation	74.1%
Q30 Bases in Barcode	94.6%
Q30 Bases in RNA Read	79.7%
Q30 Bases in Sample Index	92.0%
Q30 Bases in UMI	92.7%

Mapping

Reads Mapped to Genome	83.2%
Reads Mapped Confidently to Genome	79.7%
Reads Mapped Confidently to Intergenic Regions	3.7%
Reads Mapped Confidently to Intronic Regions	26.2%
Reads Mapped Confidently to Exonic Regions	49.8%
Reads Mapped Confidently to Transcriptome	46.8%
Reads Mapped Antisense to Gene	1.3%

Cells



Estimated Number of Cells	694
Fraction Reads in Cells	59.1%
Mean Reads per Cell	159,922
Median Genes per Cell	2,276
Total Genes Detected	18,110
Median UMI Counts per Cell	8,789

Sample

Name	JB-plus
Description	
Transcriptome	GRCh38
Chemistry	Single Cell 3' v2
Cell Ranger Version	2.1.0

Discuss 2

- How good is the run?
- Is the number of cells/ genes good?
- Is the number of reads high enough?
- What could be improved?
- Would it need a rerun?

5,875
Estimated Number of Cells

17,312
Mean Reads per Cell

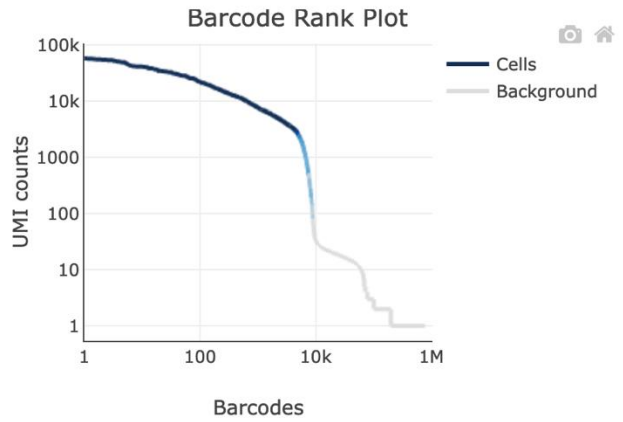
Sequencing

Number of Reads	101,710,132
Number of Short Reads Skipped	0
Valid Barcodes	98.2%
Valid UMIs	100.0%
Sequencing Saturation	43.3%
Q30 Bases in Barcode	96.1%
Q30 Bases in RNA Read	95.9%
Q30 Bases in UMI	96.1%

Mapping

Reads Mapped to Genome	97.1%
Reads Mapped to Genome (Pchab)	0.2%
Reads Mapped to Genome (Mmus)	97.0%
Reads Mapped Confidently to Genome	72.4%
Reads Mapped Confidently to Genome (Pchab)	0.2%
Reads Mapped Confidently to Genome (Mmus)	72.2%
Reads Mapped Confidently to Intergenic Regions	2.1%

Cells



Estimated Number of Cells	5,875
Estimated Number of Cells (Pchab)	221
Estimated Number of Cells (Mmus)	5,669
Fraction Reads in Cells	88.0%
Fraction Reads in Cells (Pchab)	94.3%
Fraction Reads in Cells (Mmus)	87.9%
Mean Reads per Cell	17,312
Median UMI Counts per Cell (Pchab)	203
Median UMI Counts per Cell (Mmus)	4,292
Median Genes per Cell (Pchab)	140
Median Genes per Cell (Mmus)	1,004
Total Genes Detected (Pchab)	3,117
Total Genes Detected (Mmus)	18,771

Sample

Discuss 3

- How good is the run?
- Is the number of cells/ genes good?
- Is the number of reads high enough?
- What could be improved?
- Would it need a rerun?

Summary | Gene Expression

21,566
Estimated Number of Cells

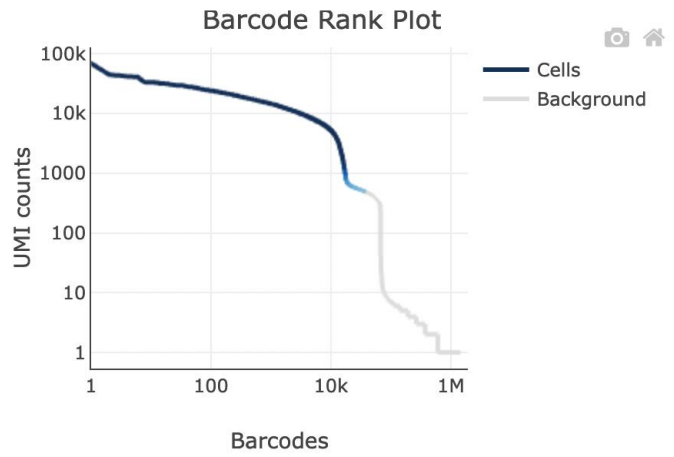
37,218
Mean Reads per Cell

1,499
Median Genes per Cell

Sequencing

Number of Reads	802,638,492
Number of Short Reads Skipped	0
Valid Barcodes	97.6%
Valid UMIs	99.8%
Sequencing Saturation	75.2%
Q30 Bases in Barcode	96.2%
Q30 Bases in RNA Read	92.8%
Q30 Bases in UMI	95.6%

Cells



Estimated Number of Cells	21,566
Fraction Reads in Cells	84.1%
Mean Reads per Cell	37,218
Median UMI Counts per Cell	4,778
Median Genes per Cell	1,499
Total Genes Detected	5,633

Cell ranger Web summary - summary

- Think about what is the best number of cells to aim for!
- How many reads do map the reference?
- Where enough reads sequenced?
- How many genes are in each cells expressed – and UMI?

5. Analysis pipeline

- Many tools to process scRNA-Seq, including
  - QC
  - Normalization
  - Detecting confounder
  - Clustering
  - Visualization
  - Identifying marker genes
  - Differentially expressed genes
  - Pseudo time

- … over the next week

https://satijalab.org/seurat/articles/get_started.html

scRNA-Seq analysis pipelines

One sample:

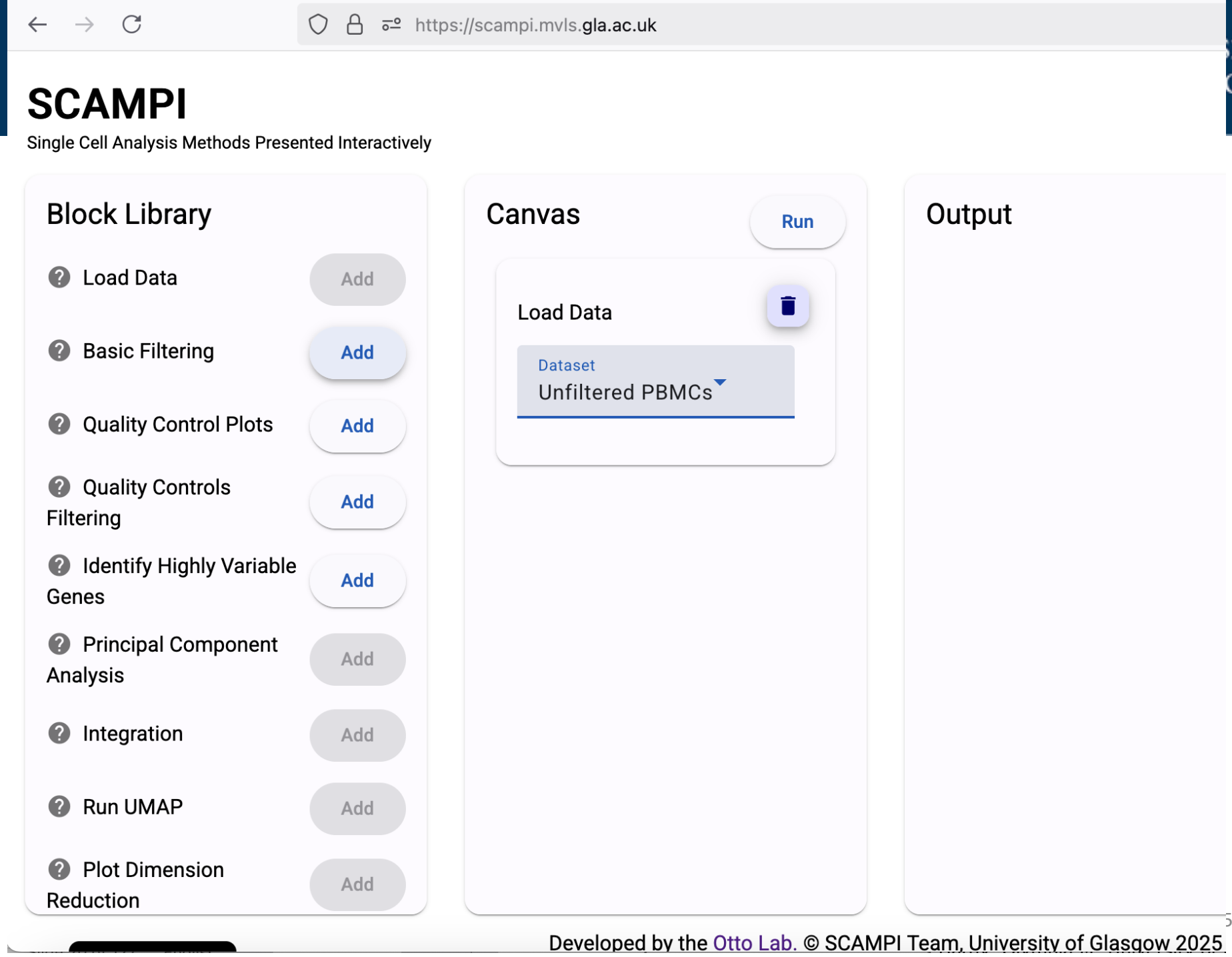
- QC
- Normalisation
- Dimension reduction
- Clustering
- Visualisation
- Identifying marker genes
- Pseudo time

Two samples

- As before, plus
- Integration
- Differential expression analysis

SCAMPI

- Homework…



Analysis

- Seurat in R
- SCANPY, in pyton

- We will do in the exercise the Seurat one

Seurat - Load data (PBMC datasets - immune cells in blood)

```
library(dplyr)
library(Seurat)
library(patchwork)

# Load the PBMC dataset
pbmc.data <- Read10X(data.dir = "../data/pbmc3k/filtered_gene_bc_matrices/hg19/")
# Initialize the Seurat object with the raw (non-normalized data).
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells = 3, min.features = 20
pbmc
```

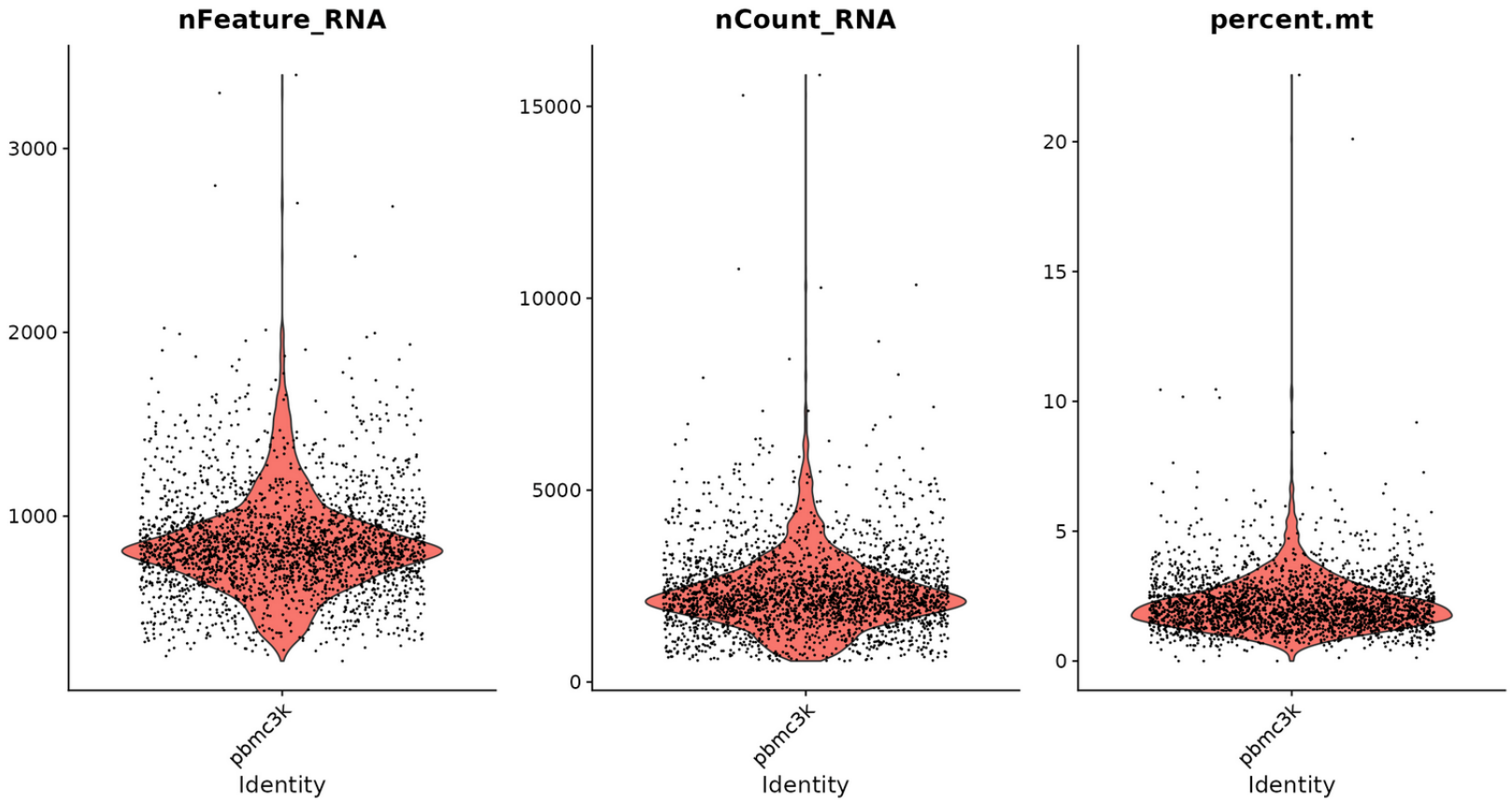
```
## An object of class Seurat
## 13714 features across 2700 samples within 1 assay
## Active assay: RNA (13714 features, 0 variable features)
```

PBMC is the R-object that holds all the data

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Quality control

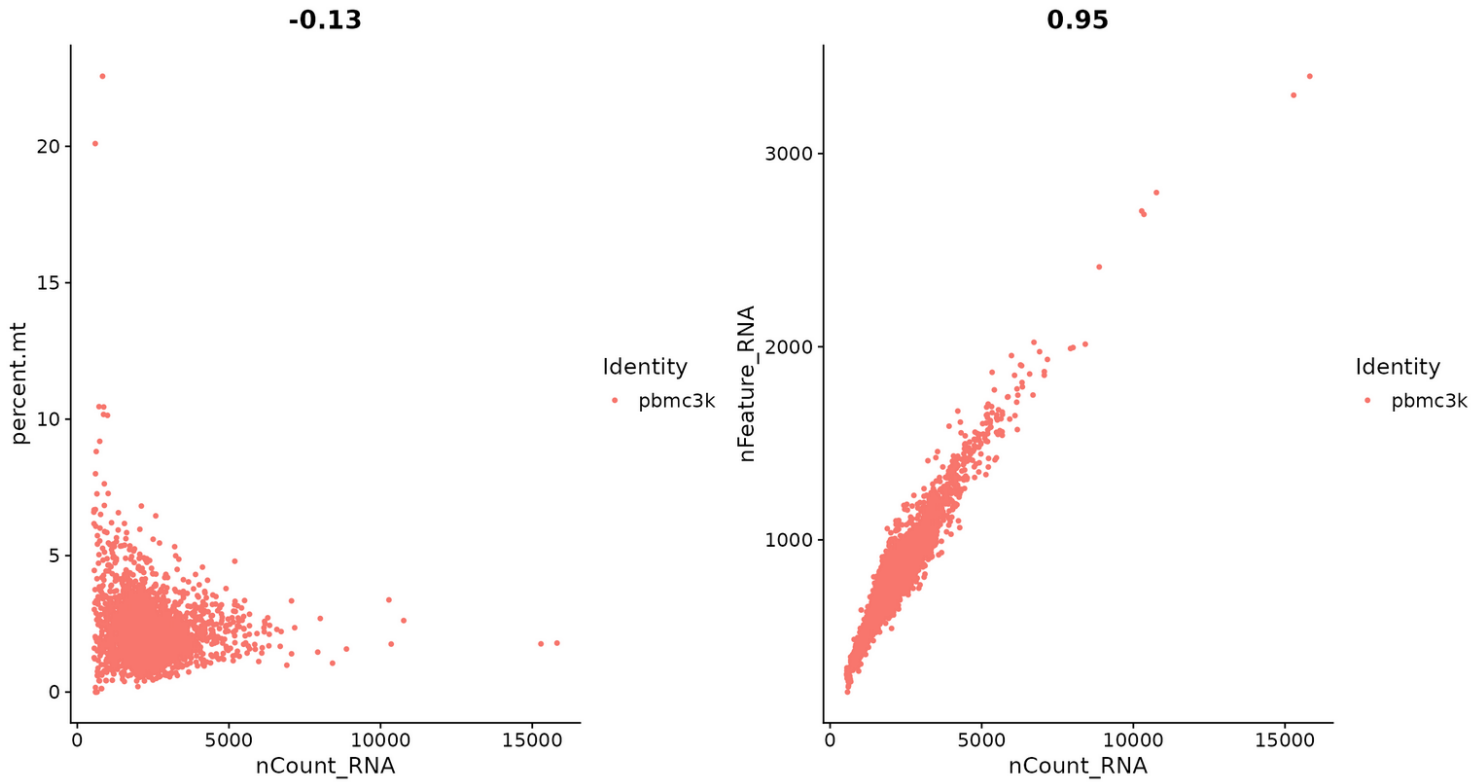
```
# Visualize QC metrics as a violin plot
VlnPlot(pbmc, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
```



Why is the expression of mitochondria interesting?

nFeature_RNA – are the genes per cell; nCount_RNA are the UMI **NOT the reads!!**
Percent.mt is the percentage of expression explained by the mitochondria expressing genes (dead cells?)

Quality control II




```
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5)
```

Get rid of doublets and potential dead cells

Normalisation

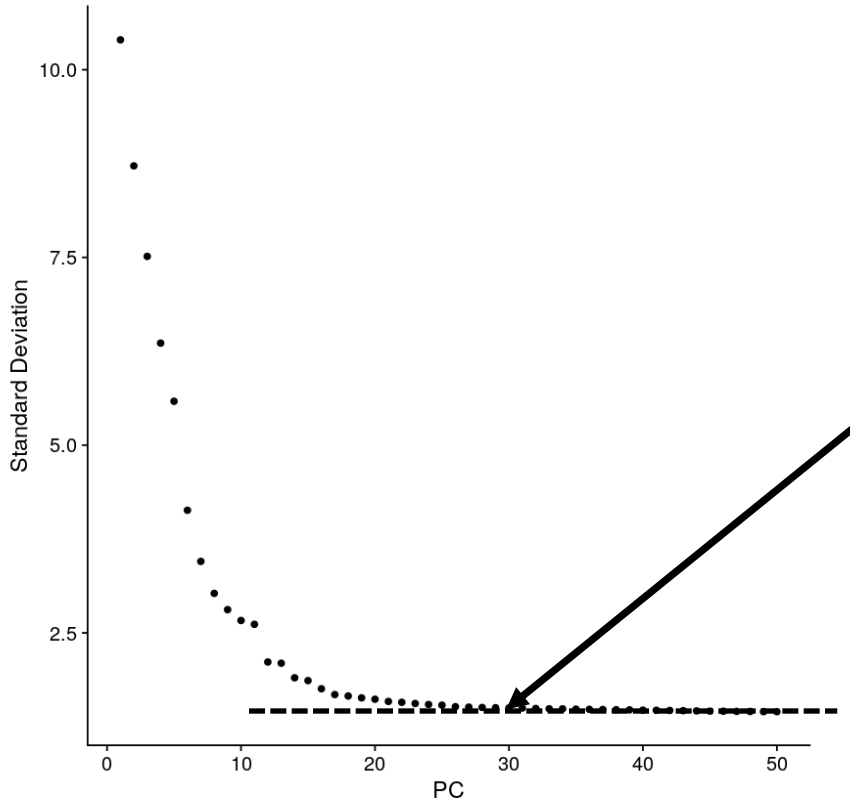
After removing unwanted cells from the dataset, the next step is to normalize the data. By default, we employ a global-scaling normalization method “LogNormalize” that normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default), and log-transforms the result. In Seurat v5, Normalized values are stored in `pbmc[["RNA"]]$data`.

```
pbmc <- NormalizeData(pbmc, normalization.method = "LogNormalize", scale.factor = 10000)
```

There are different methods to normalise, like sc_transform and scran, but we won’t do this here.

Dimension reduction

```
In [9]: ElbowPlot(Combined, ndims = 50)
```




```
In [ ]: # You have to set this variable:
dim=xXx

Combined <- RunUMAP(Combined, reduction = "pca", dims = 1:dim)
Combined <- FindNeighbors(Combined, reduction = "pca", dims = 1:dim)
```

graph-based clustering approach

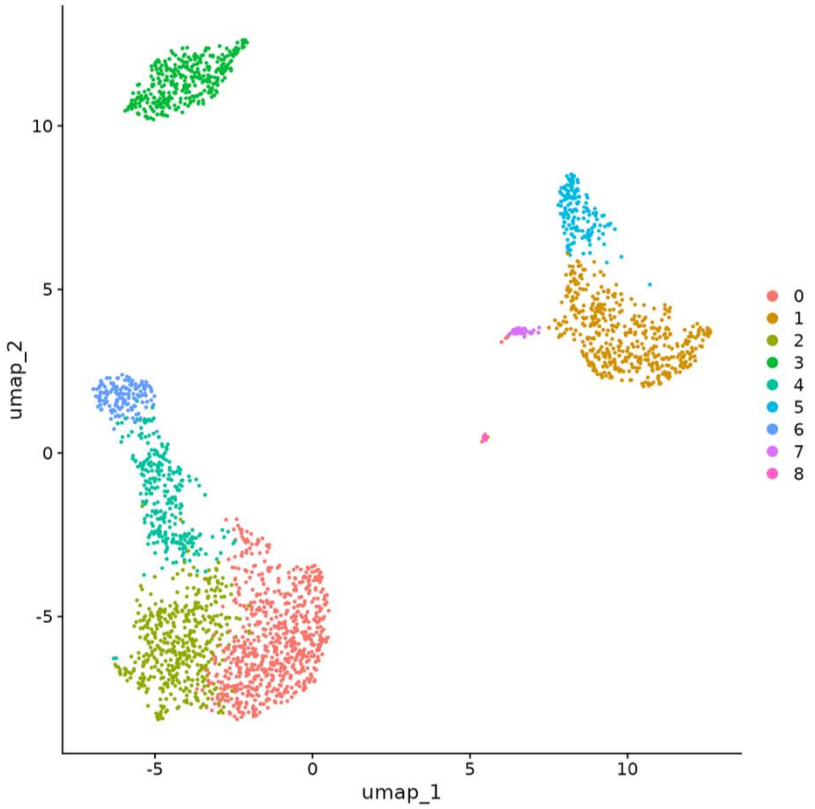
```
pbmc <- FindNeighbors(pbmc, dims = 1:10)
pbmc <- FindClusters(pbmc, resolution = 0.5)
```

So are we don't see anything!

Run non-linear dimensional reduction (UMAP/tSNE)

```
pbmc <- RunUMAP(pbmc, dims = 1:10)
```

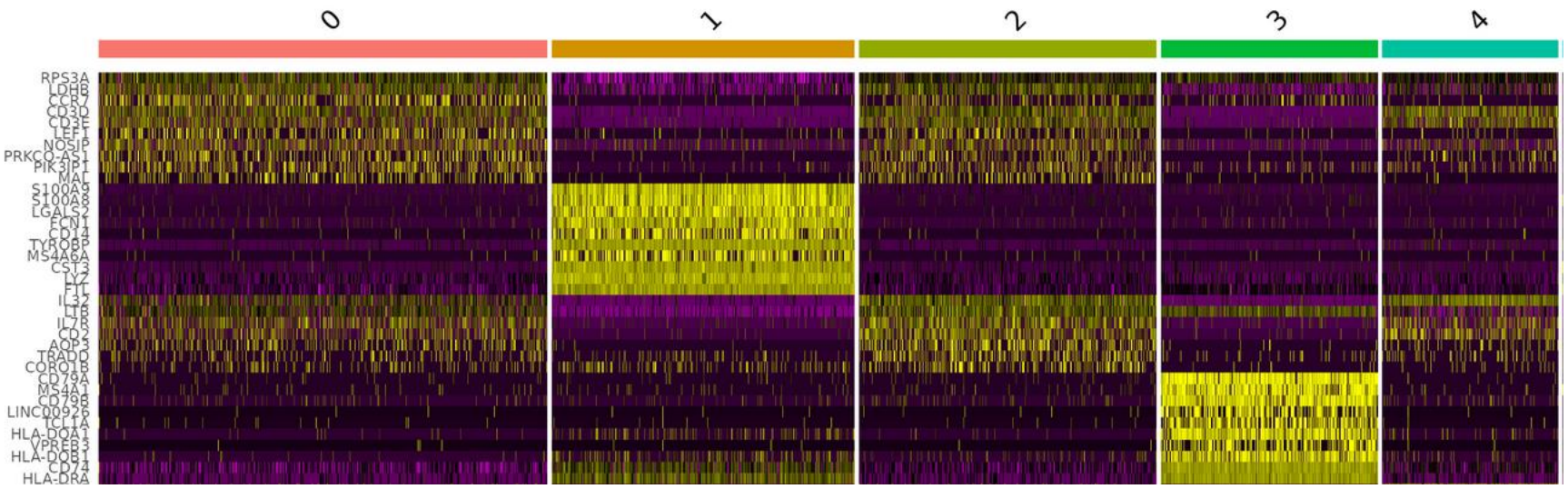
```
# note that you can set `label = TRUE` or use the LabelClusters function to help label
# individual clusters
DimPlot(pbmc, reduction = "umap")
```



Finding differentially expressed features (cluster biomarkers)

```
# find markers for every cluster compared to all remaining cells, report only the positive ones
pbmc.markers <- FindAllMarkers(pbmc, only.pos = TRUE, min.pct = 0.25, logfc.threshold = 0.25)
pbmc.markers %>% group_by(cluster) %>% top_n(n = 2, wt = avg_log2FC)
```

```
top10 <- pbmc.markers %>% group_by(cluster) %>% top_n(n = 10, wt = avg_log2FC)
DoHeatmap(pbmc, features = top10$gene) + NoLegend()
```

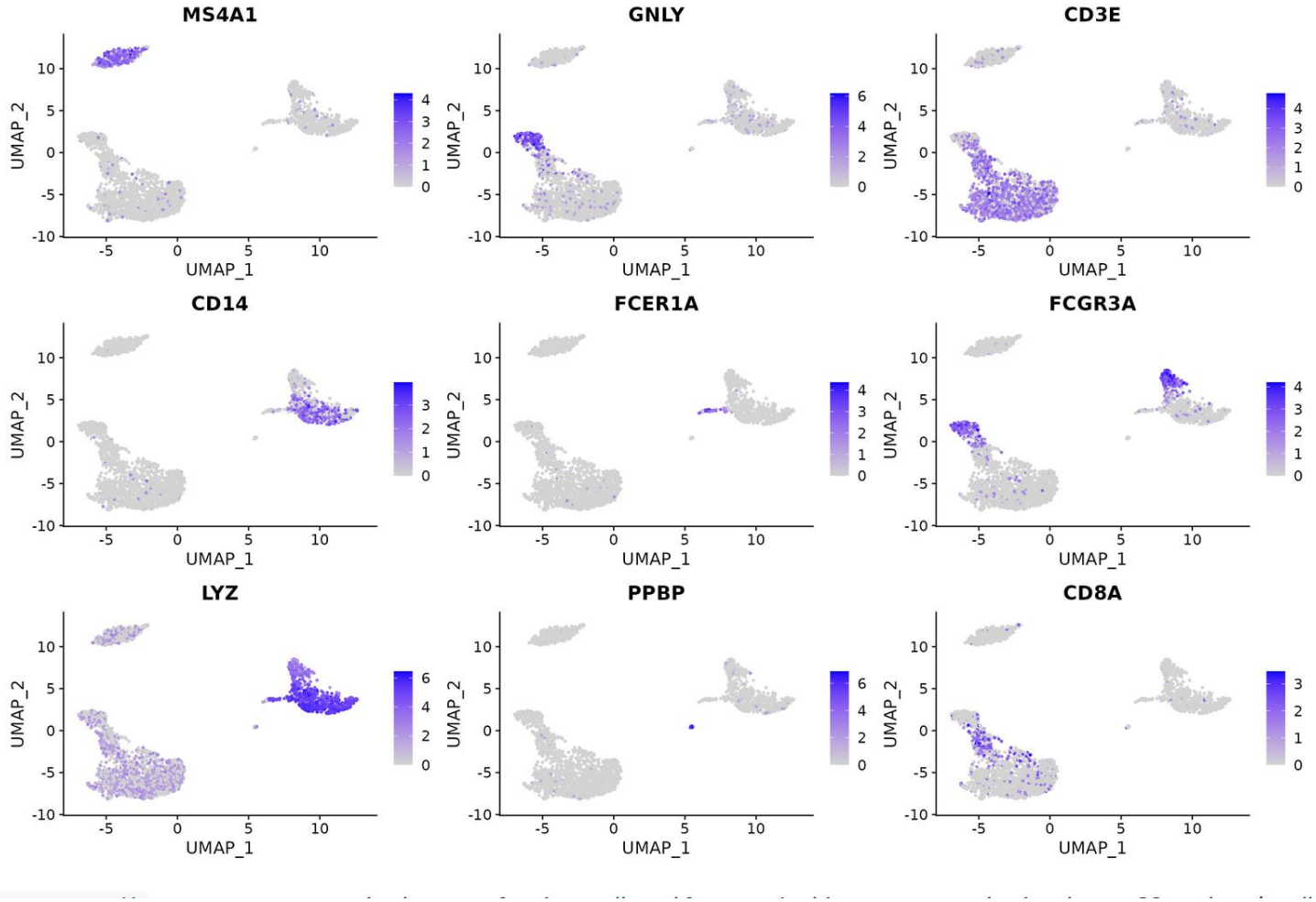


Markers	Cell Type
IL7R, CCR7	Naive CD4+ T
CD14, LYZ	CD14+ Mono
IL7R, S100A4	Memory CD4+
MS4A1	B
CD8A	CD8+ T
FCGR3A, MS4A7	FCGR3A+ Mono
GNLY, NKG7	NK
FCER1A, CST3	DC
PPBP	Platelet

The task here will be to read/investigate (or know), which marker genes are specific for what tissues.

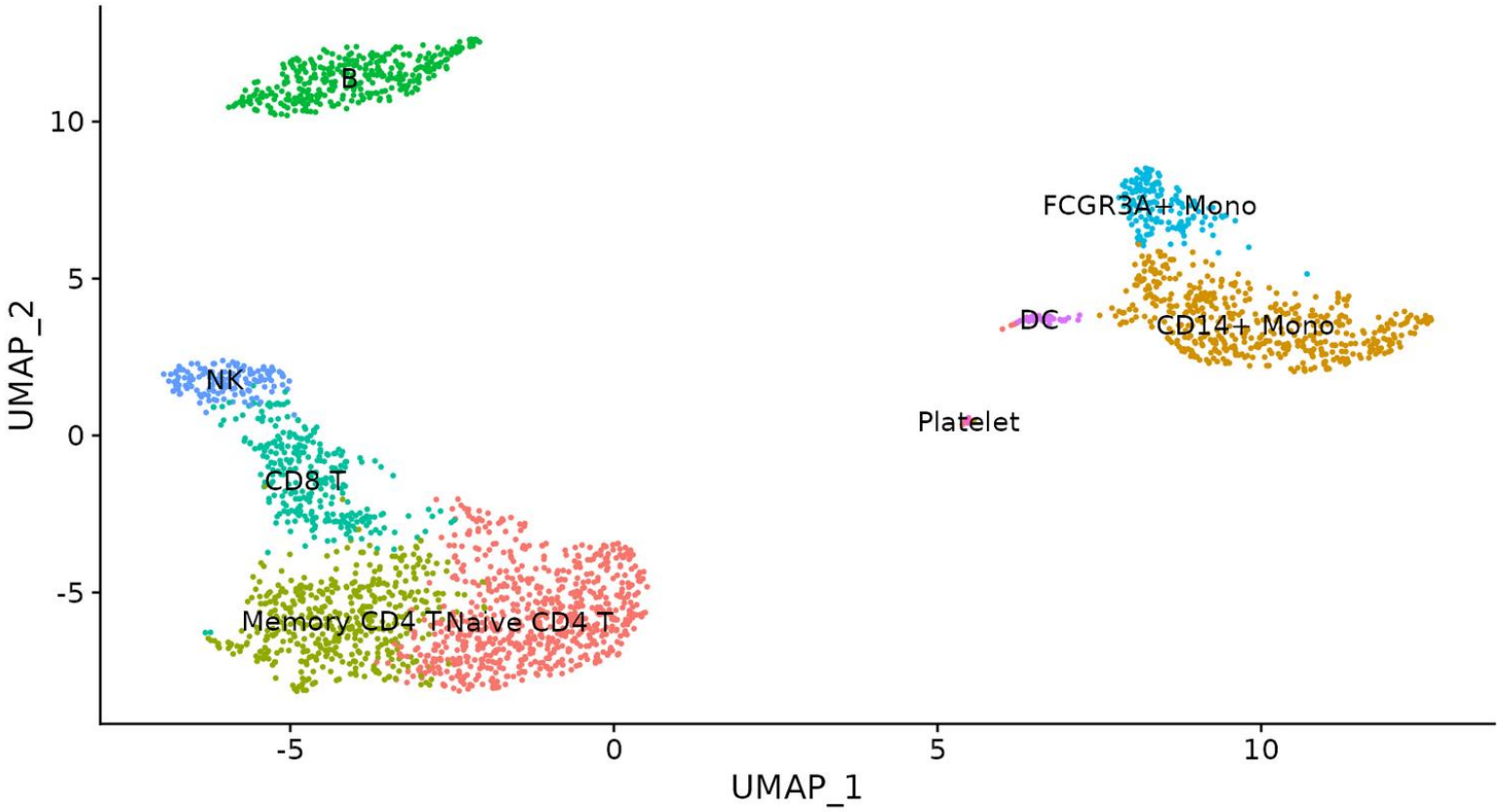
Known markers

```
FeaturePlot(pbmc, features = c("MS4A1", "GNLY", "CD3E", "CD14", "FCER1A", "FCGR3A", "LYZ", "PPBP",
    "CD8A"))
```



Annotation of clusters

```
new.cluster.ids <- c("Naive CD4 T", "CD14+ Mono", "Memory CD4 T", "B", "CD8 T", "FCGR3A+ Mono",
    "NK", "DC", "Platelet")
names(new.cluster.ids) <- levels(pbmc)
pbmc <- RenameIdents(pbmc, new.cluster.ids)
DimPlot(pbmc, reduction = "umap", label = TRUE, pt.size = 0.5) + NoLegend()
```



Be careful; we will use different codes.

Obviously, you need to adapt those for each dataset.

That’s it

- We quality controlled, clustered and annotated a PBMC dataset.

- But that is bowring, how can we compare two different groups?

Differential expression between two conditions

- Comparing stimulated PBMC or Covid patients with healthy
- For each cluster, we want to know the statistically differentially expressed genes between the two conditions – think about “multiple test” corrections
- Findallmarkers function – set the test; MAST is a good one, and talk about more on Monday
- People argue that pseudo bulking each sample for each clusters (if sufficient replicates) allows better results

Integration:

```
ifnb <- IntegrateLayers(object = ifnb, method = CCAIntegration, orig.reduction = "pca", new.reduc
tion = "integrated.cca",
    verbose = FALSE)
```

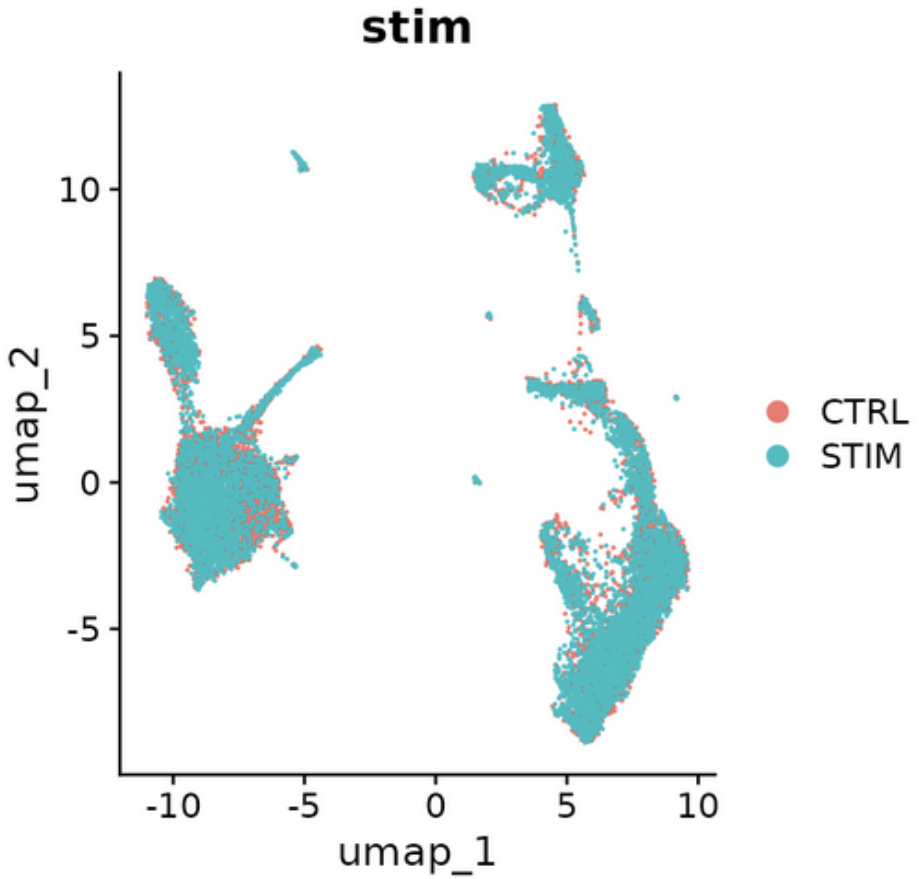
This part is tricky and we will look at other methods.

https://satijalab.org/seurat/articles/integration_introduction

Differential expression

```
ifnb$celltype.stim <- paste(ifnb$seurat_annotations, ifnb$stim, sep = "_")
Idents(ifnb) <- "celltype.stim"
b.interferon.response <- FindMarkers(ifnb, ident.1 = "B_STIM", ident.2 = "B_CTRL", verbose = FALS
E)
head(b.interferon.response, n = 15)
```

```
##                 p_val avg_log2FC pct.1 pct.2     p_val_adj
## ISG15   5.387767e-159  5.0588481 0.998 0.233 7.571429e-155
## IFIT3   1.945114e-154  6.1124940 0.965 0.052 2.733468e-150
## IFI6    2.503565e-152  5.4933132 0.965 0.076 3.518260e-148
## ISG20   6.492570e-150  3.0549593 1.000 0.668 9.124009e-146
## IFIT1   1.951022e-139  6.2320388 0.907 0.029 2.741772e-135
## MX1     6.897626e-123  3.9798482 0.905 0.115 9.693234e-119
## LY6E    2.825649e-120  3.7907800 0.898 0.150 3.970885e-116
## TNFSF10 4.007285e-112  6.5802175 0.786 0.020 5.631437e-108
## IFIT2   2.672552e-108  5.5525558 0.786 0.037 3.755738e-104
```



How is that different to the previous DESeq2 work?

And then – understand difference!

- Have a good look at the list of DE genes
- Do enrichment analysis

If you like Python – SCAMPI was done with SCANP[Y|I]

University of Glasgow

```python
# Import necessary libraries
import scanpy as sc
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Set up Scanpy settings
sc.settings.verbosity = 3  # Verbosity level (0: errors, 1: warnings, 2: info, 3: hints)
sc.settings.set_figure_params(dpi=80, facecolor='white')  # Set figure parameters

# Load single-cell RNA-Seq data (e.g., from a 10x Genomics dataset)
adata = sc.read_10x_mtx(
    'path/to/filtered_feature_bc_matrix',  # Path to the folder containing matrix.mtx, genes.tsv, and barcodes.tsv
    var_names='gene_symbols',              # Use gene symbols as variable names
    cache=True                             # Cache the data for faster loading
)
```

University of Glasgow

```python
# Preprocessing
# 1. Filter cells and genes
sc.pp.filter_cells(adata, min_genes=200)  # Filter cells with fewer than 200 genes
sc.pp.filter_genes(adata, min_cells=3)   # Filter genes detected in fewer than 3 cells

# 2. Normalize data
sc.pp.normalize_total(adata, target_sum=1e4)  # Normalize total counts per cell to 10,000
sc.pp.log1p(adata)  # Log-transform the data

# 3. Identify highly variable genes
sc.pp.highly_variable_genes(adata, min_mean=0.0125, max_mean=3, min_disp=0.5)
adata = adata[:, adata.var.highly_variable]  # Keep only highly variable genes

# 4. Scale the data
sc.pp.scale(adata, max_value=10)

# Dimensionality reduction (PCA)
sc.tl.pca(adata, svd_solver='arpack')

# Compute neighborhood graph
sc.pp.neighbors(adata, n_neighbors=10)

# Clustering (Leiden algorithm)
sc.tl.leiden(adata)  # Cluster cells

# UMAP visualization
sc.tl.umap(adata)  # Compute UMAP
sc.pl.umap(adata, color=['leiden'], legend_loc='on data')  # Plot UMAP with clusters

# Marker gene identification
sc.tl.rank_genes_groups(adata, 'leiden', method='t-test')  # Find marker genes for each cluster
sc.pl.rank_genes_groups(adata, n_genes=25, sharey=False)  # Visualize top marker genes

# Save the results
adata.write('path/to/output.h5ad')  # Save the AnnData object to a file

# Optional: Export results to CSV
pd.DataFrame(adata.obs).to_csv('path/to/clusters.csv')  # Export cluster assignments
pd.DataFrame(adata.var).to_csv('path/to/genes.csv')    # Export gene information
```

Where to tweak the parameters?

- Select cells, exclude doublets & dead cell (more tomorrow)
- Need to reduce dimensionality – PCA - but how much?
- Need to cluster data – solved graph-based clustering, but which resolution?
- How to annotate clusters?

- We are going to cover these further, but here, grasp a general idea how they work
- A lot of literature for further reading available

Take home message here:

- scRNA-Seq is a moving field in terms of methods
- Need to reduce dimensionality – PCA - but how much?
- Need to cluster data – solved graph-based clustering, but which resolution?
- How to visualise data: UMAP – but it is just a visualisation methods

- We are going to cover these further, but here, grasp a general idea how they work
- A lot of literature for further reading available

- In our practical we are using R, however Python with ScanPy, very powerfull as well, and several member of my team prefer it.

Show case SCAMPI

- Home work – try a scampi workflow
- https://scampi.mvls.gla.ac.uk/

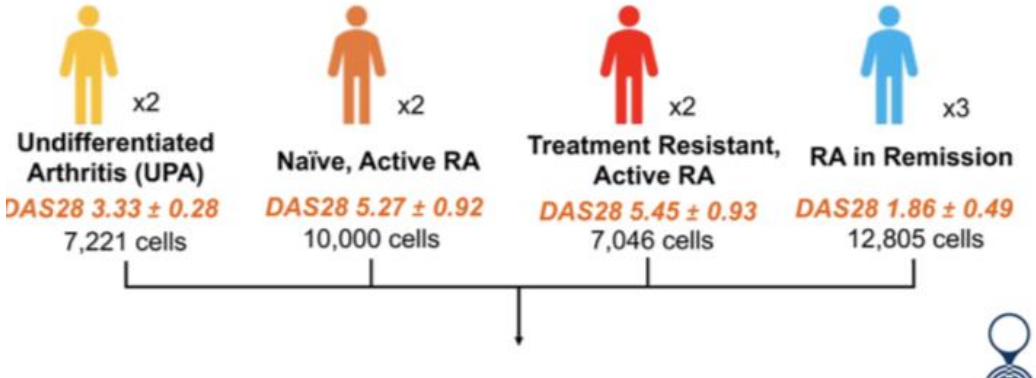
6. Examples

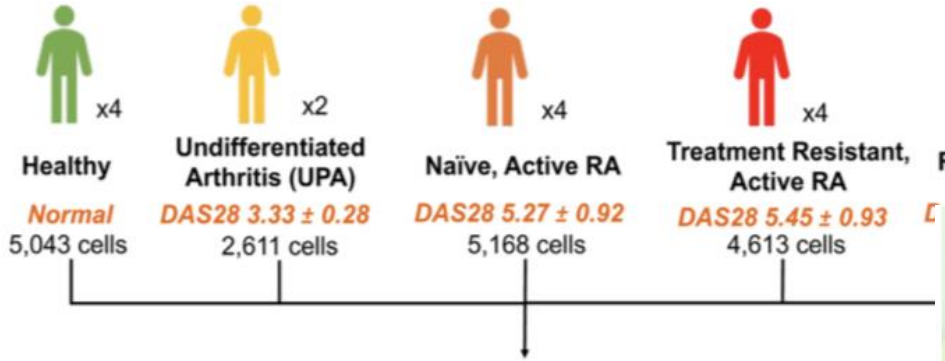
- Role of macrophages in Rheumatoid Arthritis
- Cell atlas ParaCell

(both projects lead by previous bioinformatics students)

Remission in rheumatoid arthritis

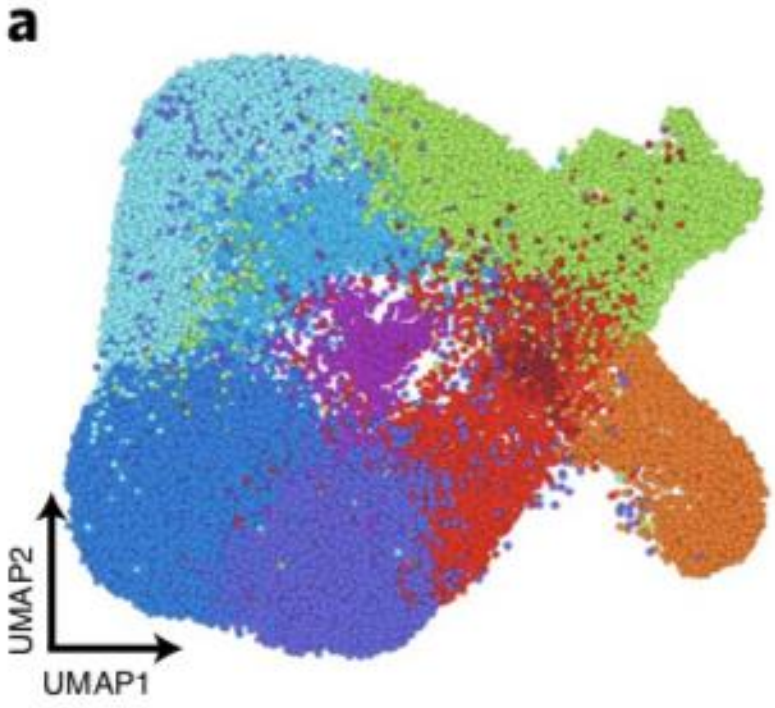
- Mariola Kurowska-Stolarska and Stefano Alivernini: Some people stay in remission, other flare – what is the role of macrophages?
- Did scRNA-seq with 10X chromium



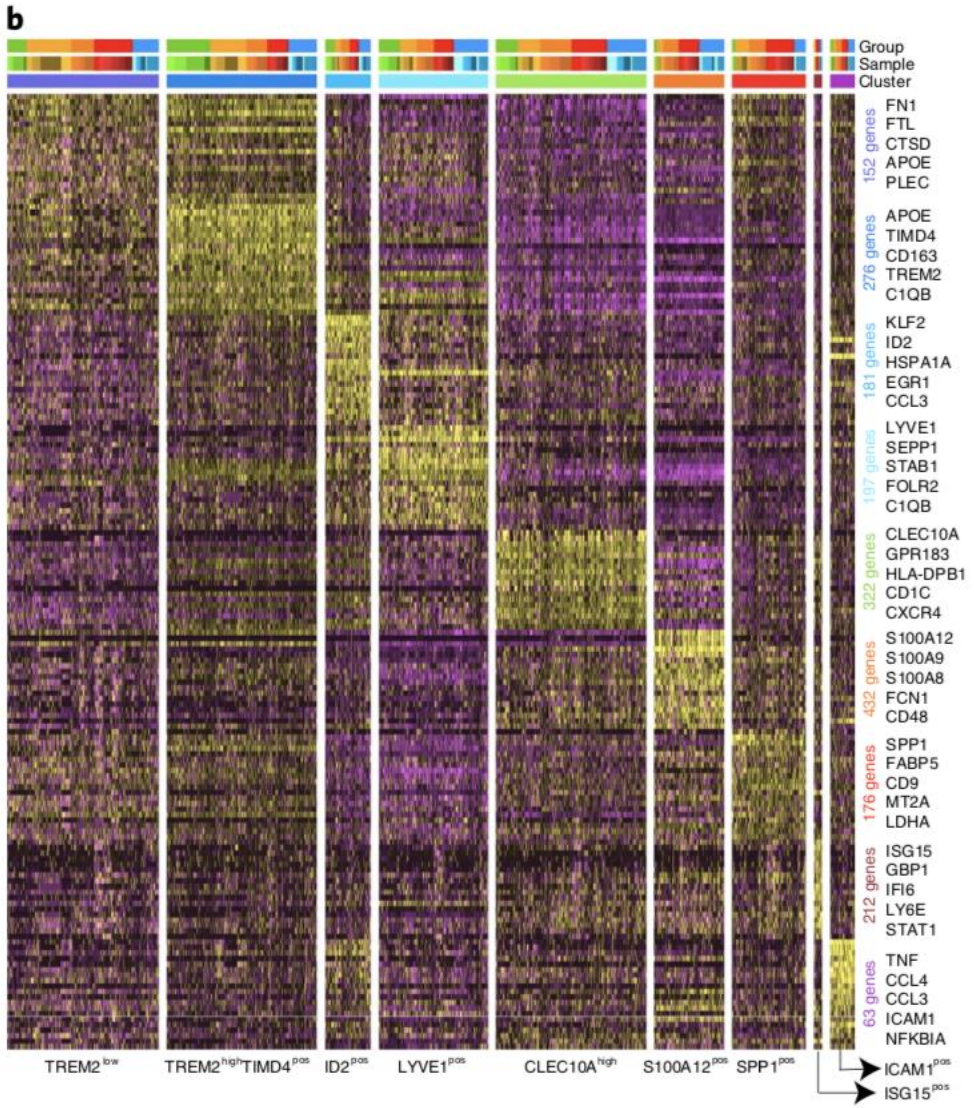


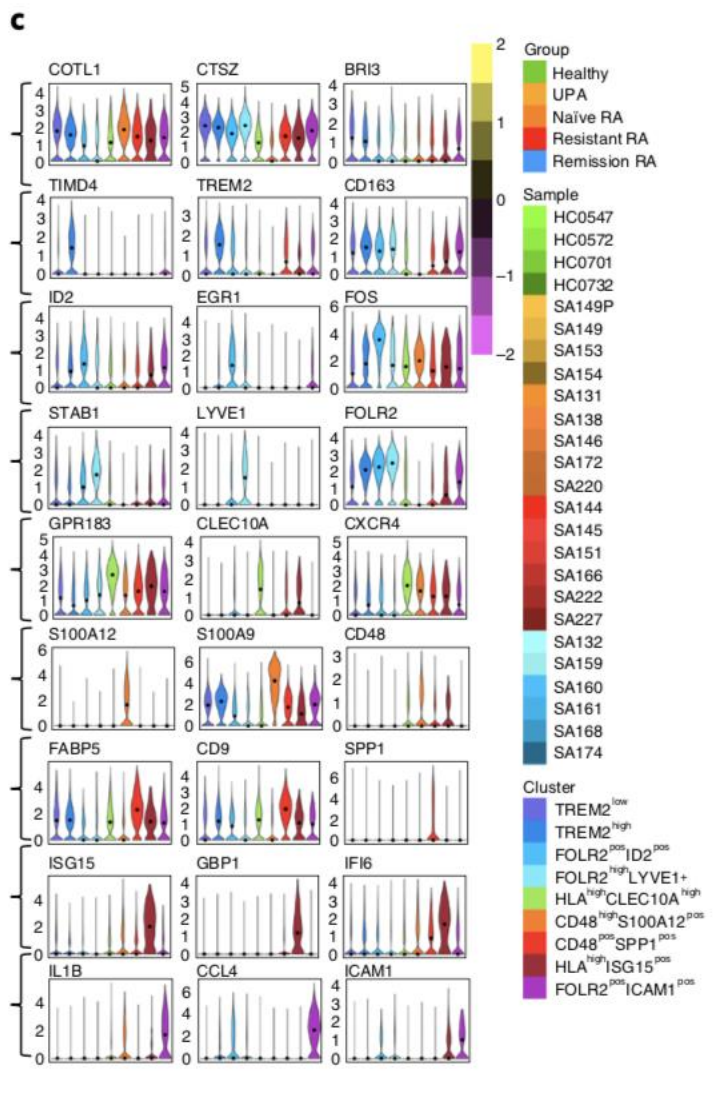
PhD Student Lucy MacDonald

Macrophages subtypes are associated to disease state

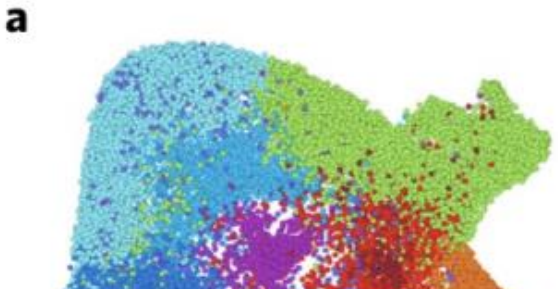


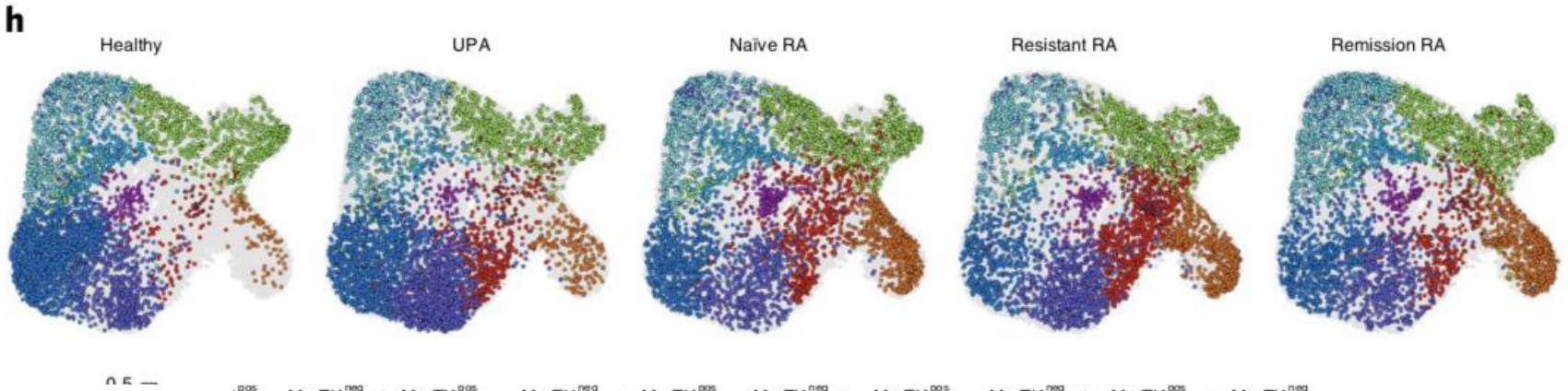
Population	Subpopulation	Cluster
MerTK^{pos}	TREM2^{pos}FOLR2^{pos}	TREM2^{pos}TIMD4^{pos}CD163^{high}
		TREM^{low}
	FOLR2^{high}TREM2^{neg}	ID2^{pos}
		LYVE1^{pos}
		ICAM1^{pos}
MerTK^{neg}	HLA^{high}CD48^{pos}	ISG15^{pos}
		CLEC10a^{high}
	CD48^{pos}	S100A12^{pos}
		SPP1^{pos}

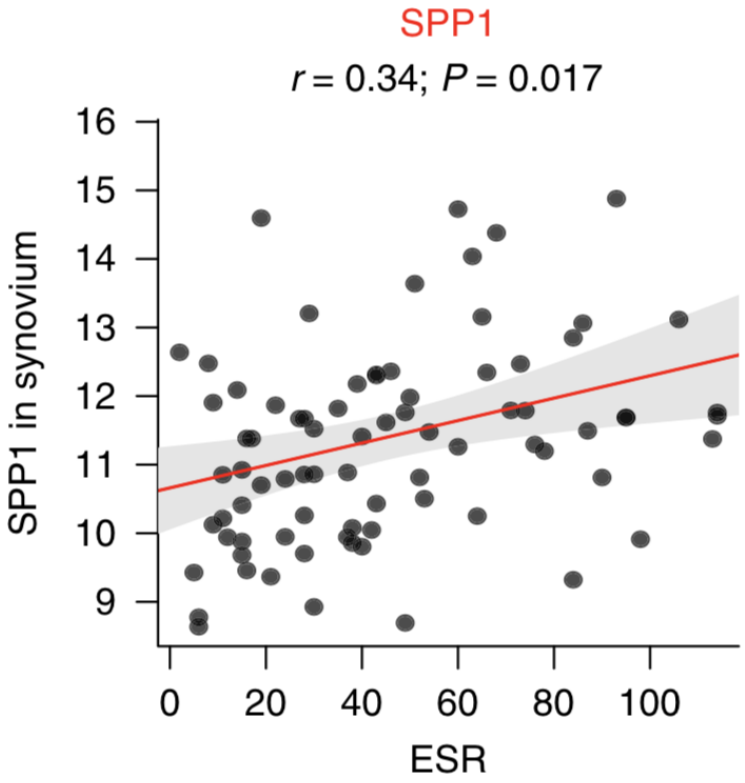


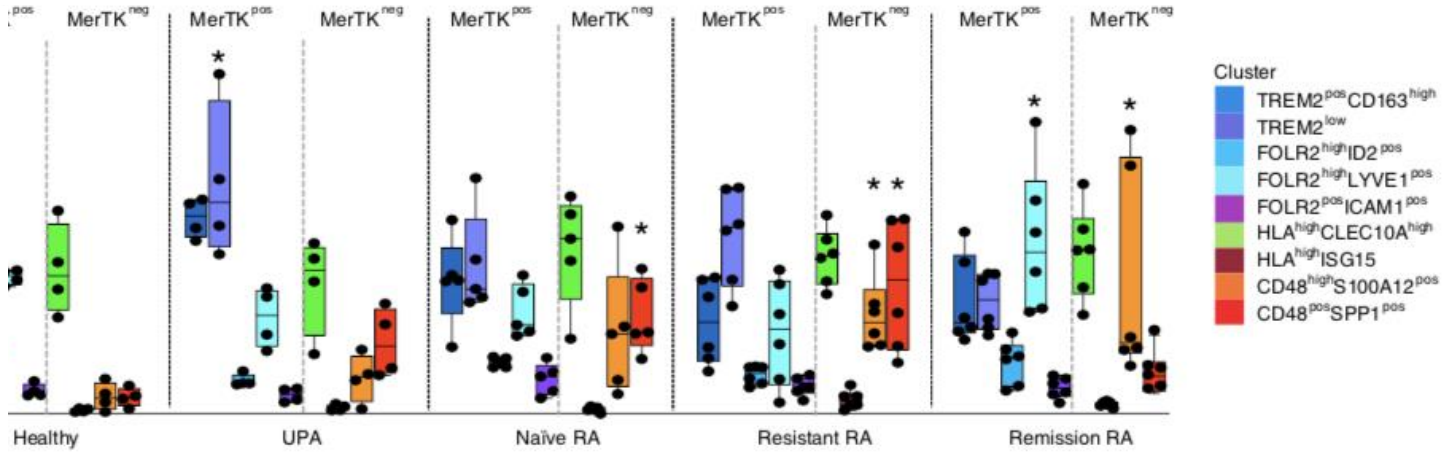


Macrophages subtypes are associated to disease state









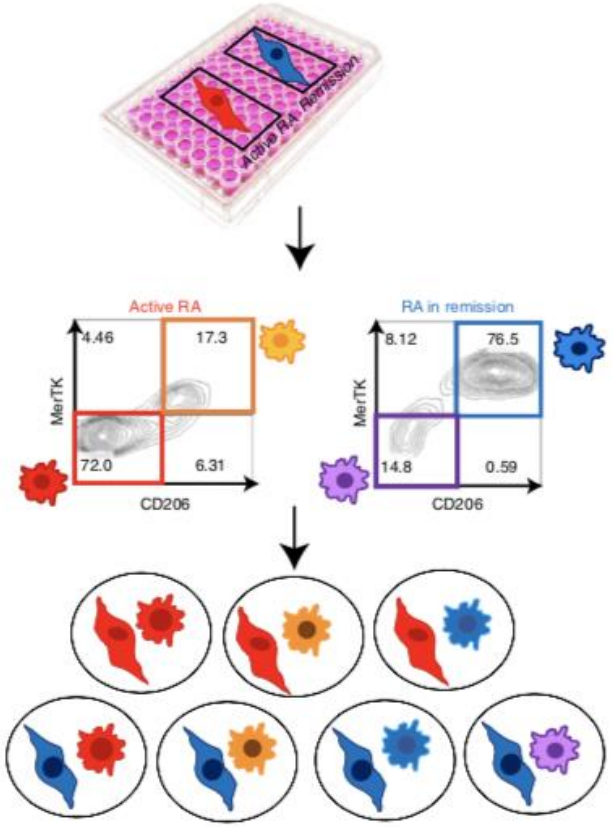
Two-tailed Spearman's correlation between synovial expression of SPP1 with disease activity in the PEAC cohort (IMID-Bio-UK - *n* = 90)

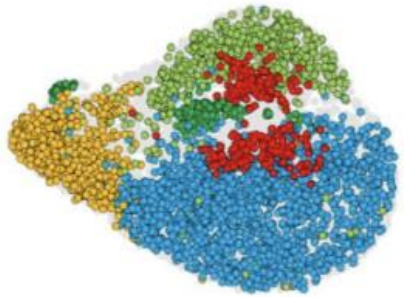
Stimulation of primary fibroblast-like synoviocytes (FLS) with different phenotype macrophages

Step 1. Seeding active RA and remission FLS (3,000 cells per well)

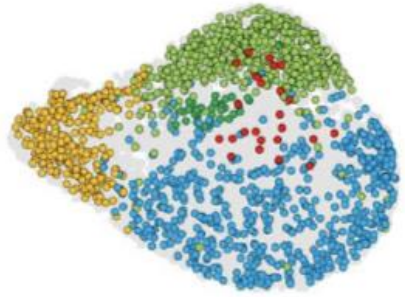
Step 2. Adding MerTK^{neg} or MerTK^{pos} STMs from active or remission RA (3,000 cells) to FLS wells

Step 3. Targeted scRNA-seq after 48-h STM–FLS coculture (BD rhapsody system)

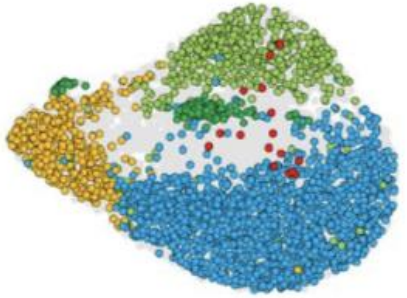




(Condition 2) FLS + MerTK^{neg} STM from RA active



(Condition 3) FLS + MerTK^{pos} STM from RA active



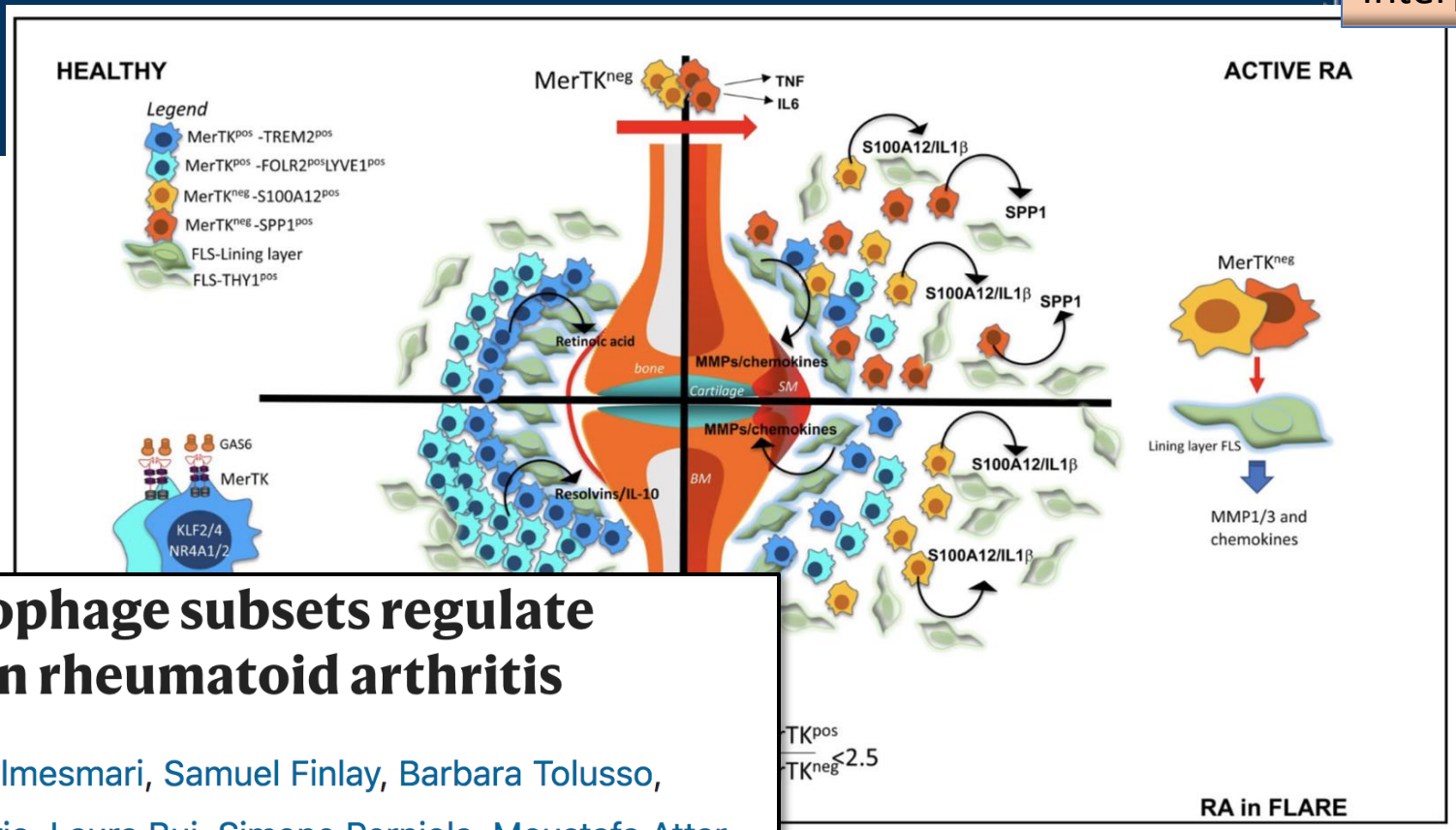
(Condition 4) FLS + MerTK^{pos} STM from RA remission

MerTKnegCD206neg and MerTKposCD206pos STMs induce inflammatory and repair responses, respectively in fibroblast

Summary

Interpretation

- Amazing!



Distinct synovial tissue macrophage subsets regulate inflammation and remission in rheumatoid arthritis

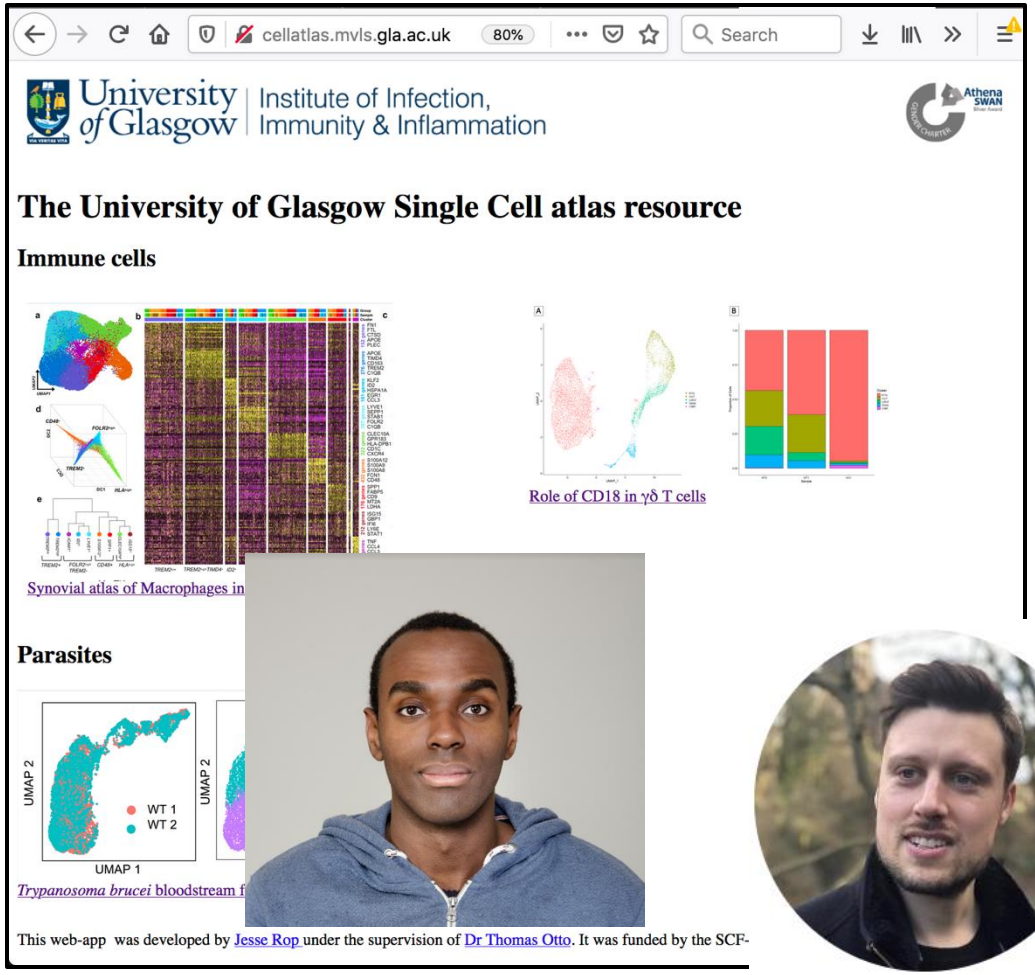
Stefano Alivernini ✉, Lucy MacDonald, Aziza Elmesmari, Samuel Finlay, Barbara Tolusso, Maria Rita Gigante, Luca Petricca, Clara Di Mario, Laura Bui, Simone Perniola, Moustafa Attar, Marco Gessi, Anna Laura Fedele, Sabarinadh Chilaka, Domenico Somma, Stephen N. Sansom, Andrew Filer, Charles McSharry, Neal L. Millar, Kristina Kirschner, Alessandra Nerviani, Myles J. Lewis, Costantino Pitzalis, Andrew R. Clark, Gianfranco Ferraccioli, Irina Udalova, Christopher D. Buckley, Elisa Gremese, Iain B. McInnes, Thomas D. Otto ✉ & Mariola Kurowska-Stolarska ✉ -Show fewer authors

Nature Medicine **26**, 1295–1306(2020) | Cite this article

flammation and remission in rheumatoid arthritis. The **HEALTHY**

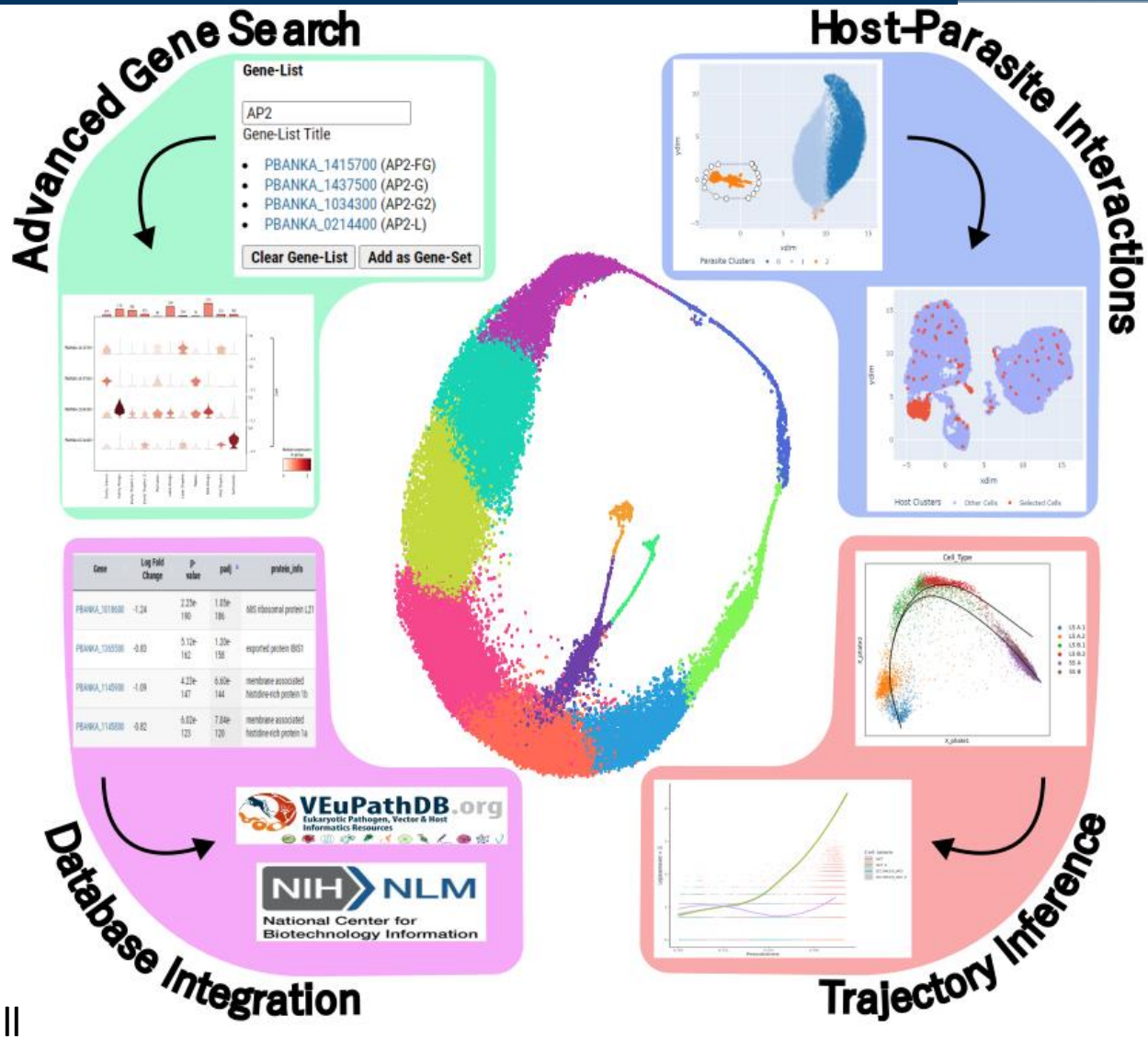
Answer/Solution

Data accessibility - ParaCell



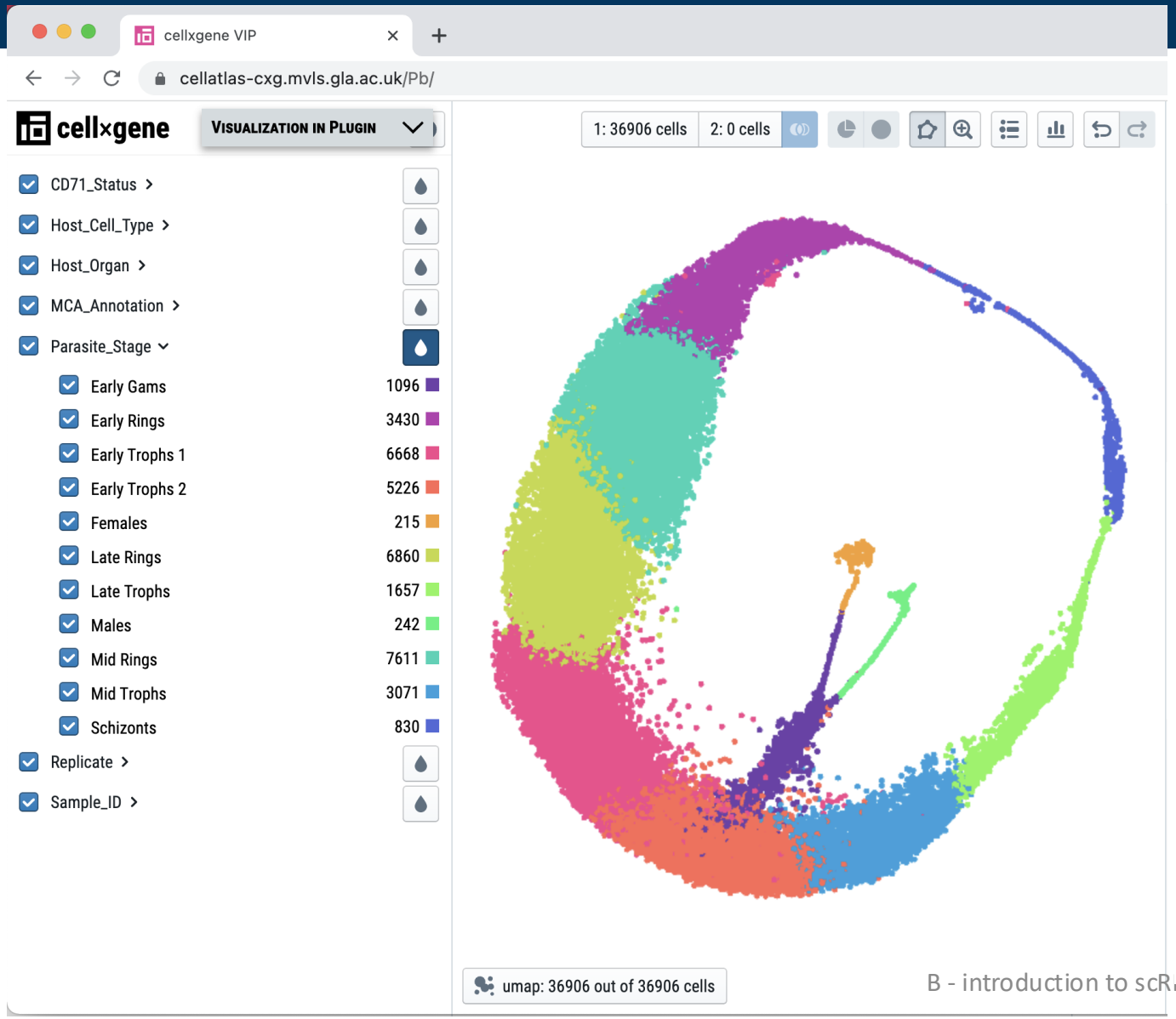
Edward Agboraw

William Haese-Hill



http://cellatlas.mvls.gla.ac.uk/

Cellxgene - paraCell



JOURNAL ARTICLE

paraCell: a novel software tool for the interactive analysis and visualization of standard and dual host–parasite single-cell RNA-seq data

Edward Agboraw, William Haese-Hill, Franziska Hentzschel, Emma Briggs, Dana Aghabi, Anna Heawood, Clare R Harding, Brian Shiels, Kathryn Crouch, Domenico Somma ...
Show more

Nucleic Acids Research, Volume 53, Issue 4, 28 February 2025, gkaf091, https://doi.org/10.1093/nar/gkaf091
Published: 20 February 2025 **Article history** ▾

Show example

- http://cellatlas.mvls.gla.ac.uk/

Conclusion

- Presentation of the pipeline from QC, normalisation, clustering, visualization and annotation.
- scRNA-Seq is noise, and many statistical & computational methods are used
- The pipelines are quite well defined, however, some important parameters need to be set.

Exercise

- SCAMPI – getting the workflow