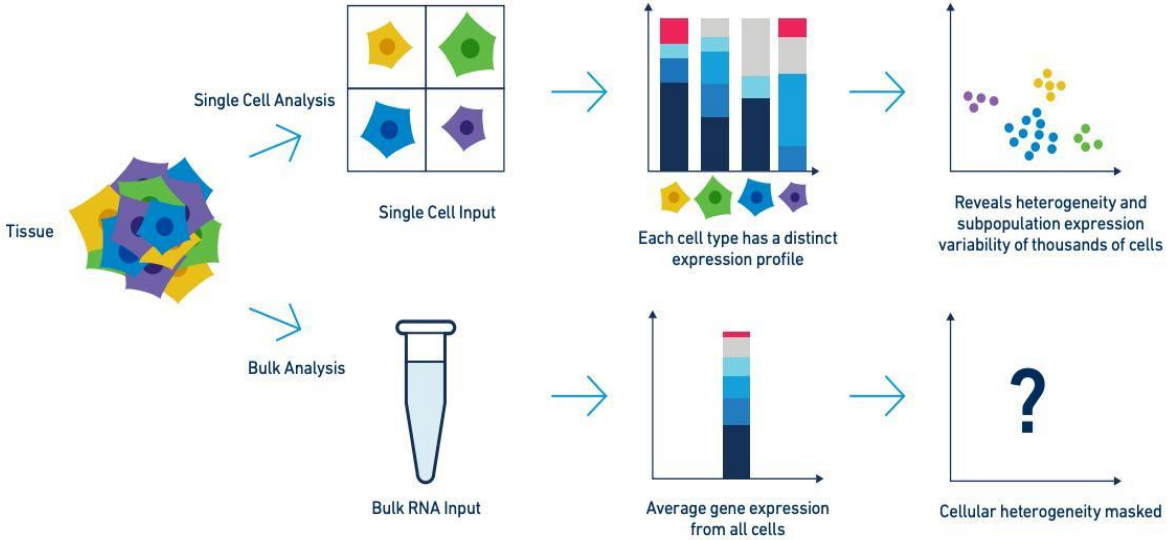

MONASH University
MONASH
BIOINFORMATICS
PLATFORM
Single Cell RNA-seq Workshop
Monash Bioinformatics Platform
GROUP
OF EIGHT
AUSTRALIA

Workshop Summary

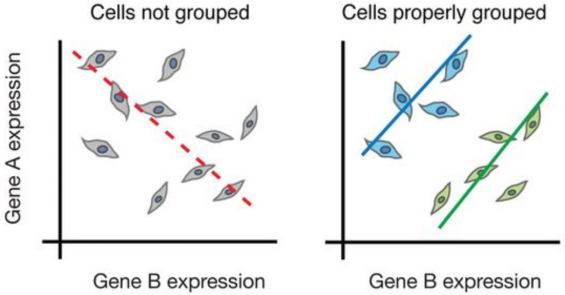
- What are we covering?
  - Basics of using Seurat for single cell analysis
  - QC > Normalisation > Dimension Reduction > Cell Clustering > Differential Expression > Cell type annotation > Dataset Integration
  - What tools to use?

- How exactly?
  - Some theory, then hands-on.
  - Demo, and exercises, helpers floating around the room
  - Based on Seurat tutorial walkthrough
  - Additional content: SingleR and harmony for cell type annotation and dataset integration(time permitting)

Single Cell Sequencing

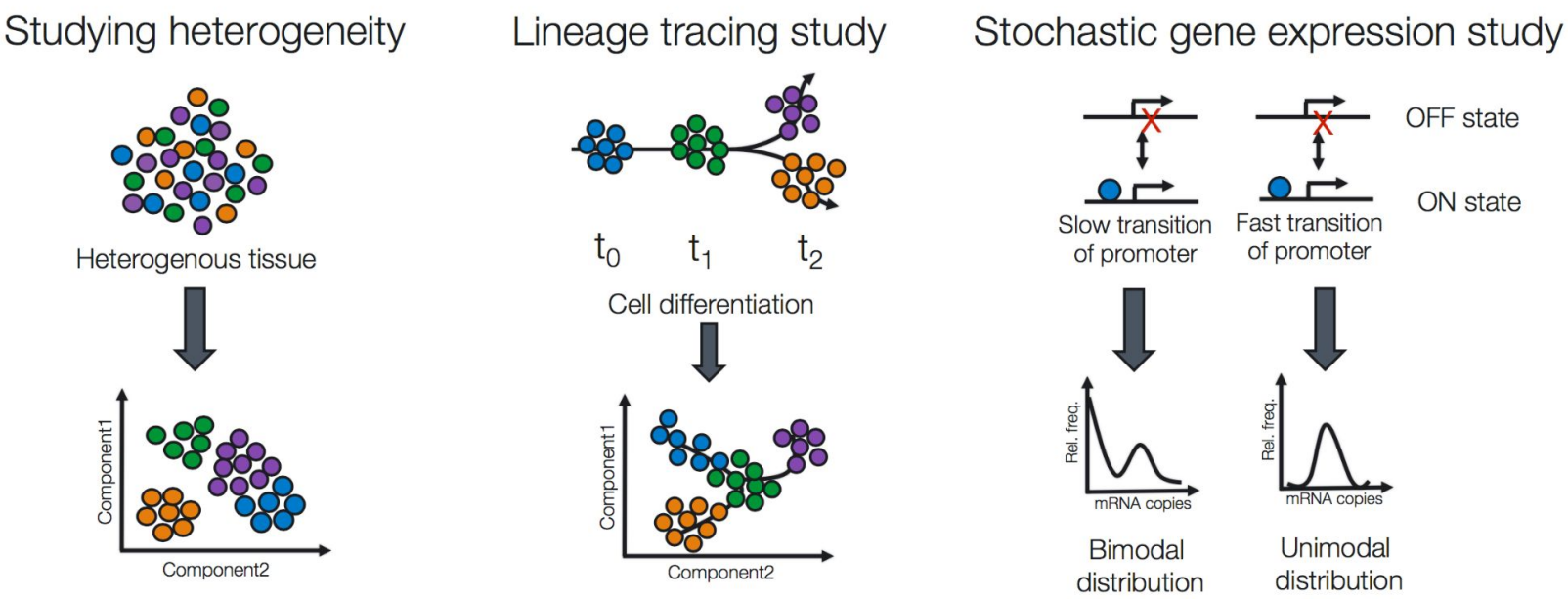


10x Genomics



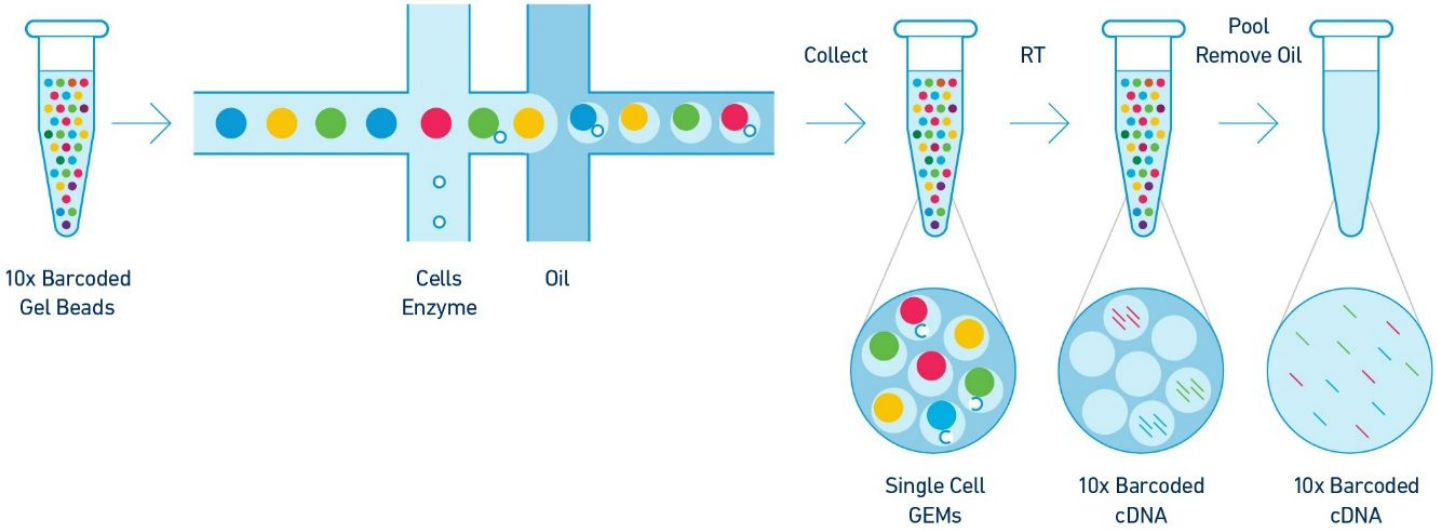
Trapnell, C. Defining cell types and states with single-cell genomics, Genome Research 2015

Single Cell Sequencing



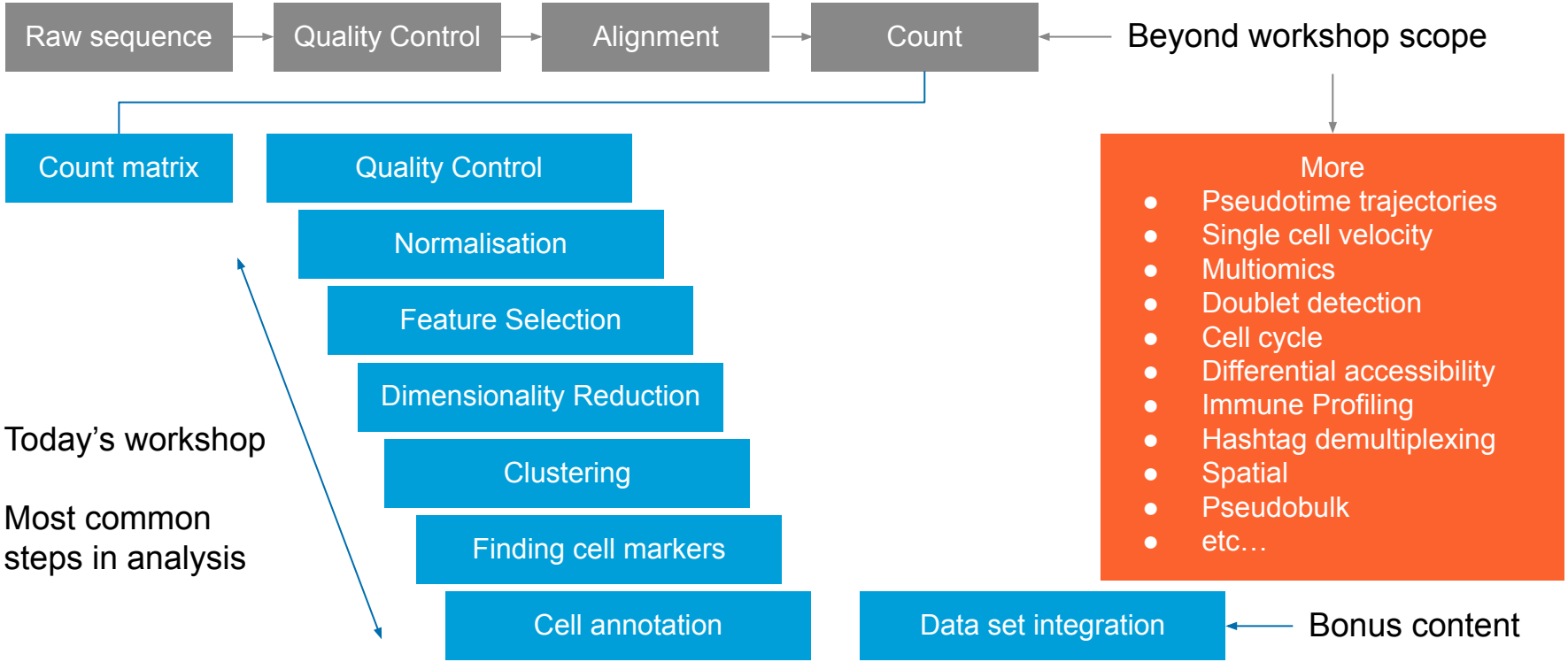
Liu S and Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges, F1000 Research 2016 (doi: 10.12688/f1000research.7223.1)
Junker and van Oudenaarden; Every Cell Is Special: Genome-wide Studies Add a New Dimension to Single-Cell Biology, Cell 2014 (doi: 10.1016/j.cell.2014.02.010)

Single Cell Sequencing

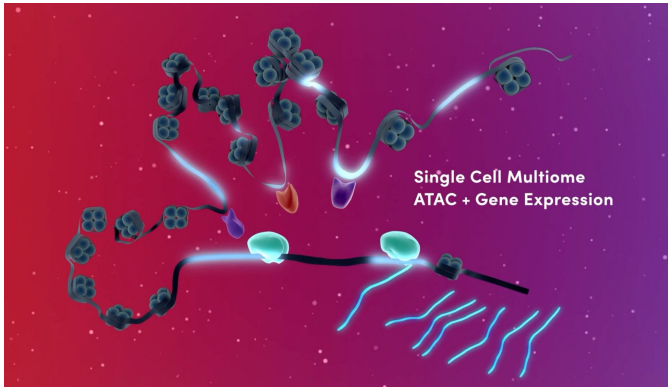


10x Genomics

Single Cell Analysis Workflow

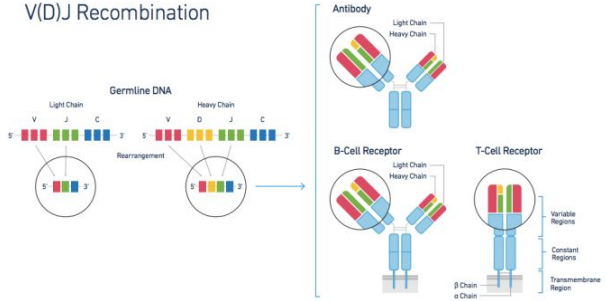


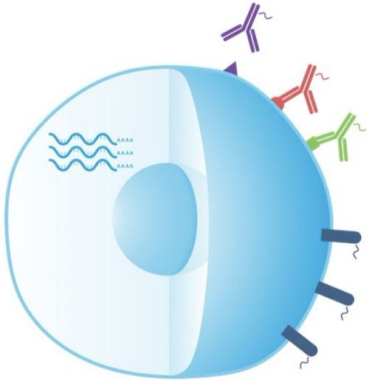
Content We Won't Be Covering

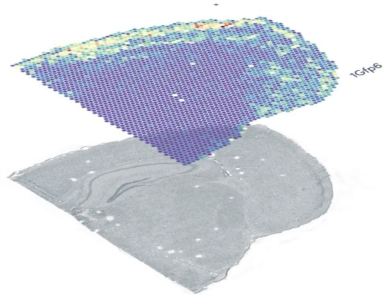


- Different sequencing technologies
- Pipelines for processing raw data
- Single cell immune profiling, atac, cell surface protein expression, spatial, etc
- Further downstream analysis beyond the basic workflow

This workshop is based on the assumption that you have 10x gene expression data generated by Cellranger



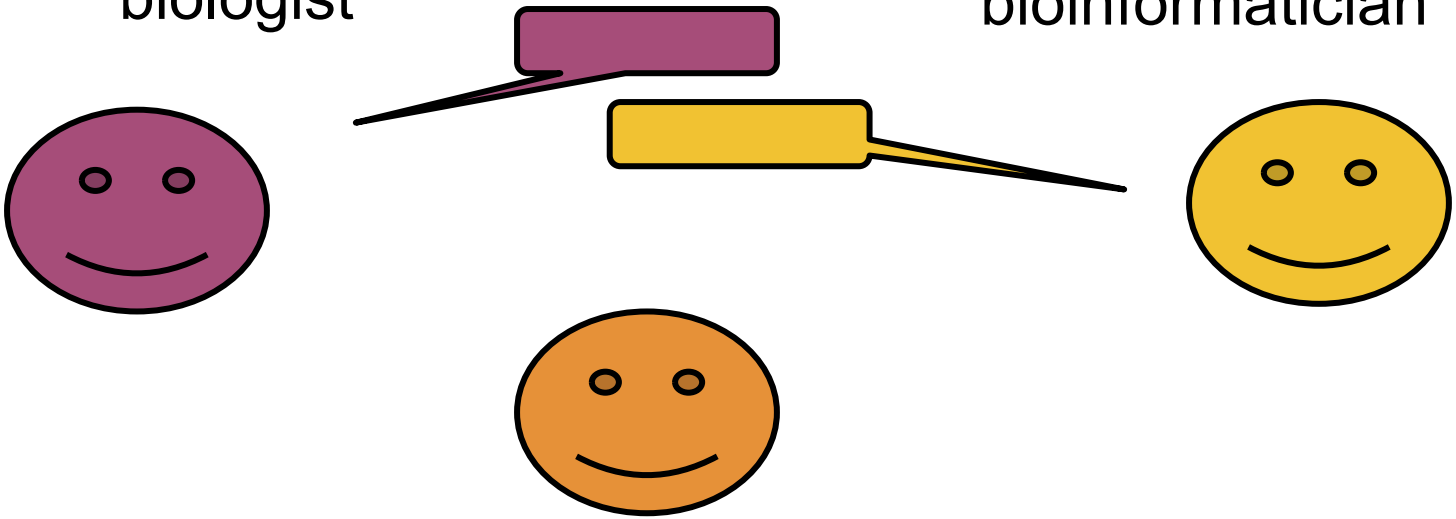




Have a Dialogue Going

Discuss with the biologist

Discuss with the bioinformatician



Or be both

MONASH University

MONASH
BIOINFORMATICS
PLATFORM

The Counts Matrix

GROUP OF EIGHT AUSTRALIA

Raw Data to Counts Matrix

Raw fastq files

Align

Pipeline

Reference Genome

Count + call cells

Pipeline

	AAACATACAACCAC	AAACATTGAGCTAC	AAACATTGATCAGC	AAACCGTGCTTCCG	AAACCGTGTATGCG	AAACGCACTGGTAC	AAACGCTGACCAGT	AAACGCTGGTTCTT	AAACGCTGTAGCCA	AAACGCTGTTTCTG	AAACTTGAAAAACG	AAACTTGATCCAGA	AAAGAGACGAGATA	AAAGAGACGCGAGA	AAAGAGACGGACTT	AAAGAGACGGCATT	AAAGCAGATATCGG	AAAGCCTGTATGCG	AAAGCCTGTCTAG	AAAGTTTGATCACG
AL627309.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AP006222.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-206L10.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-206L10.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LINC00115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOC2L	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
KLHL17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PLEKHN1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-54O7.17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HES4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
RP11-54O7.11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISG15	0	0	1	9	0	1	0	0	0	3	0	0	1	5	0	2	1	0	0	0
AGRN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C1orf159	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TNFRSF18	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TNFRSF4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	12	0	0	0
SDF4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	
B3GALT6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
FAM132A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
UBE2J2	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0

Count Matrix

Analysis

Counts Matrix - From Cellranger

Sparse matrix format only includes non-zero counts

matrix.mtx

```
%%MatrixMarket matrix coordinate real
general
%
32738    2700    2286884
32709    1    4
32707    1    1
32706    1    10
32704    1    1
32703    1    5
32702    1    6
32700    1    10
32699    1    25
32698    1    3
32697    1    8
…
```

features.tsv

```
ENSG00000243485    MIR1302-10
ENSG00000237613    FAM138A
ENSG00000186092    OR4F5
ENSG00000238009    RP11-34P13.7
ENSG00000239945    RP11-34P13.8
ENSG00000237683    AL627309.1
ENSG00000239906    RP11-34P13.14
ENSG00000241599    RP11-34P13.9
ENSG00000228463    AP006222.2
ENSG00000237094    RP4-669L17.10
ENSG00000235249    OR4F29
ENSG00000236601    RP4-669L17.2
ENSG00000236743    RP5-857K21.15
ENSG00000231709    RP5-857K21.1
…
```

barcodes.tsv

```
AAACATACAACCAC-1
AAACATTGAGCTAC-1
AAACATTGATCAGC-1
AAACCGTGCTTCCG-1
AAACCGTGTATGCG-1
AAACGCACTGGTAC-1
AAACGCTGACCAGT-1
AAACGCTGGTTCTT-1
AAACGCTGTAGCCA-1
AAACGCTGTTTCTG-1
AAACTTGAAAAACG-1
AAACTTGATCCAGA-1
AAAGAGACGAGATA-1
AAAGAGACGCGAGA-1
…
```

Counts Matrix

Cell barcodes

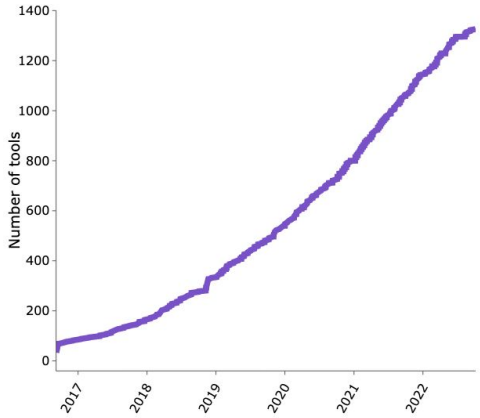
Features

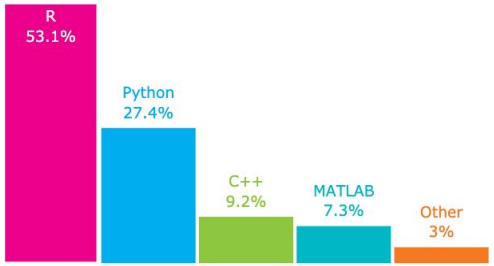
	AAACATACAACCAC	AAACATTGAGCTAC	AAACATTGATCAGC	AAACCGTGCTTCCG	AAACCGTGTATGCG	AAACGCACTGGTAC	AAACGCTGACCAGT	AAACGCTGGTTCTT	AAACGCTGTAGCCA	AAACGCTGTTTCTG	AAACTTGAAAAACG	AAACTTGATCCAGA	AAAGAGACGAGATA	AAAGAGACGCGAGA	AAAGAGACGGACTT	AAAGAGACGGCATT	AAAGCAGATATCGG	AAAGCCTGTATGCG	AAAGGCCTGTCTAG	AAAGTTTGATCACG
AL627309.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AP006222.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-206L10.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-206L10.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LINC00115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOC2L	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
KLHL17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PLEKHN1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-54O7.17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HES4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
RP11-54O7.11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISG15	0	0	1	9	0	1	0	0	0	3	0	0	1	5	0	2	1	0	0	0
AGRN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C1orf159	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TNFRSF18	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TNFRSF4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	12	0	0
SDF4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
B3GALT6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
FAM132A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
UBE2J2	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0

Counts

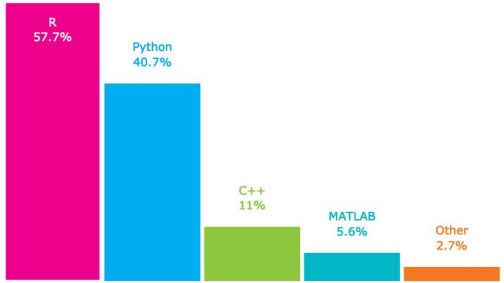

MONASH University
MONASH
BIOINFORMATICS
PLATFORM
Single Cell Analysis Ecosystems
GROUP
OF EIGHT
AUSTRALIA

Huge Number of Single Cell Analysis Tools





Oct 2019



Oct 2022 - source: https://www.scrna-tools.org

With so many tools, how do you then pick which one to use?

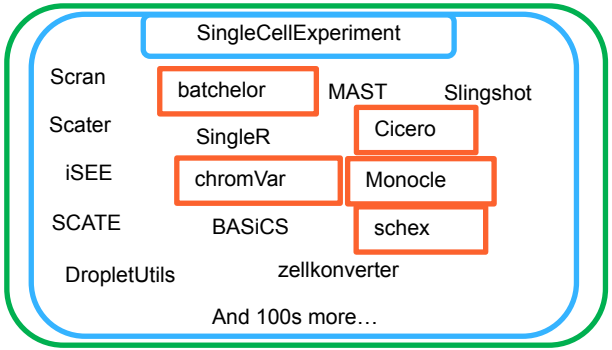
R is very popular, python is catching up

Most tools are capable of performing general analysis tasks - some will have specialised types of analysis only that particular tool can perform

Single Cell Analysis Ecosystems

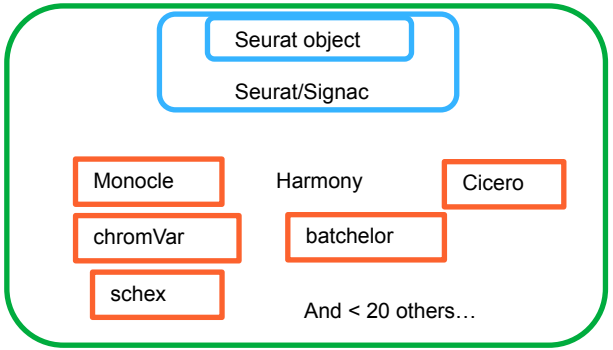
Bioconductor: R

- Repository of many bioinformatics analysis packages
- Single cell packages in Bioconductor make use of the ***singleCellExperiment*** class



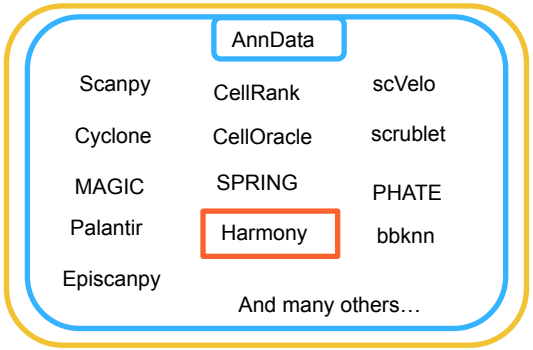
Seurat (Signac): R

- Twin R packages that has decided to make themselves a one-stop shop for most common single cell analysis tasks
- Uses the ***Seurat*** class



Scanpy: Python

- Toolkit for single cell analysis
- Uses the ***anndata*** class
- Large ecosystem of tools that integrate with scanpy

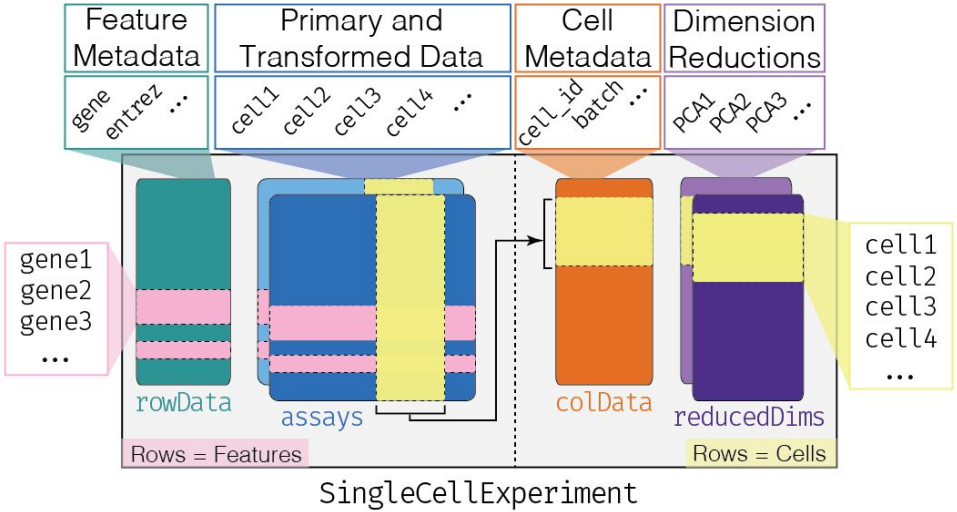


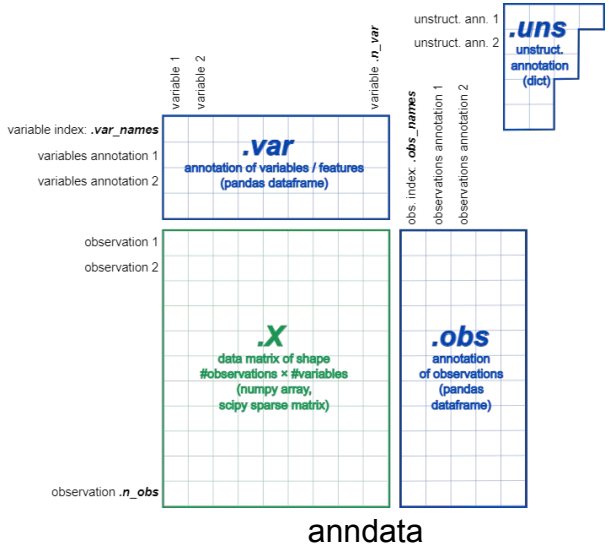
Single Cell Analysis Ecosystems

Each system has a **shared central data structure** for storing and representing single cell data

Conversion between different data structures isn't impossible but it can be difficult

Knowing what analysis you want to do downstream can save you a lot of pain if you use the right data structure upstream.



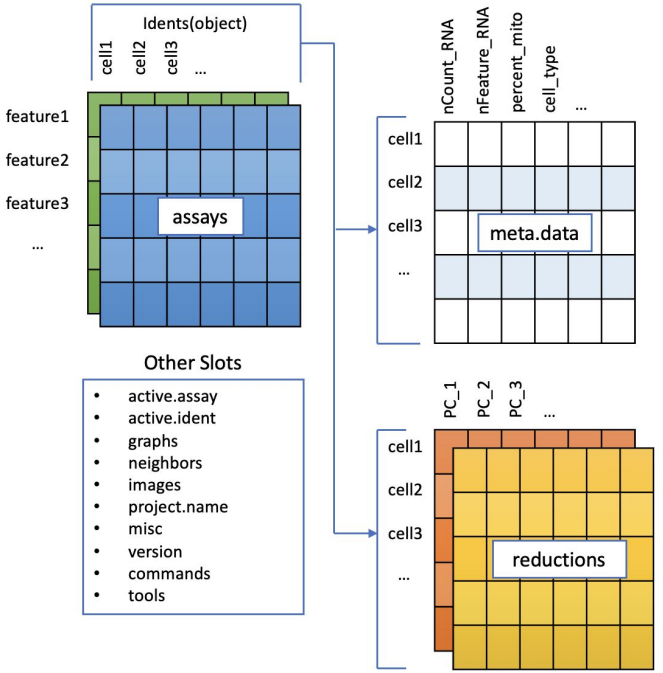


Seurat

- Very popular R package
- Well documented
- Lots of tutorials - we're working through the PBMC tutorial today
- Performs most common single cell analysis tasks

Seurat object - S4 class:

- Container for count matrix, normalised count matrix, dimensionality reduction, cell meta-data, etc
- Can contain more than gene expression data, will also store ATAC peak counts, cell surface protein counts, spatial images, etc



MONASH University

MONASH
BIOINFORMATICS
PLATFORM

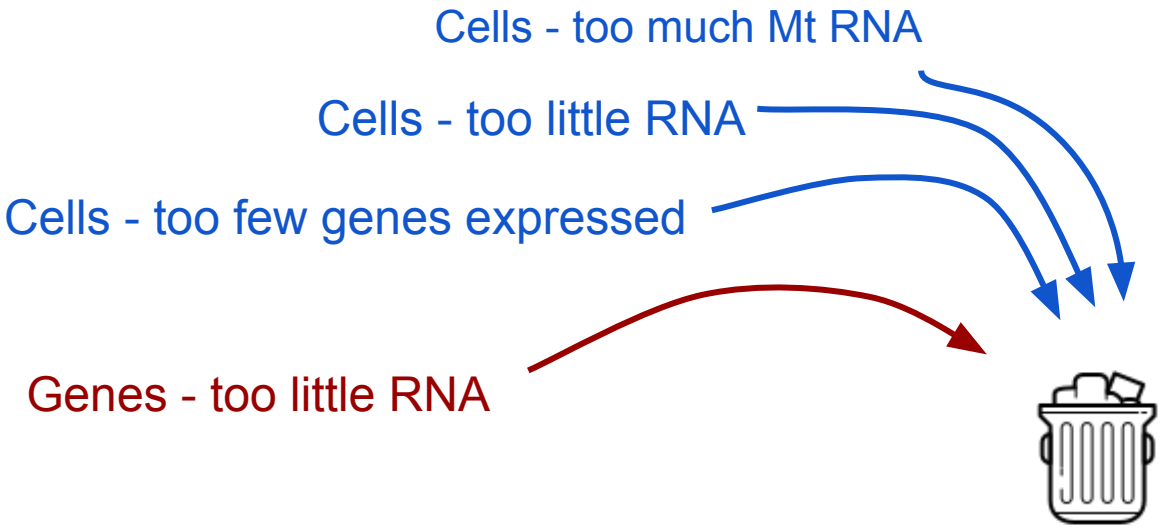
Analysis Workflow

GROUP OF EIGHT AUSTRALIA

Counts Matrix

	AAACATACAACCAC	AAACATTGAGCTAC	AAACATTGATCAGC	AAACCGTGCTTCCG	AAACCGTGTATGCG	AAACGCACTGGTAC	AAACGCTGACCAGT	AAACGCTGGTTCTT	AAACGCTGTAGCCA	AAACGCTGTTTCTG	AAACTTGAAAAACG	AAACTTGATCCAGA	AAAGAGACGAGATA	AAAGAGACGCGAGA	AAAGAGACGGACTT	AAAGAGACGGCATT	AAAGCAGATATCGG	AAAGCCTGTATGCG	AAAGGCCTGTCTAG	AAAGTTTGATCACG
AL627309.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AP006222.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-206L10.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-206L10.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LINC00115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOC2L	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
KLHL17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PLEKHN1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-54O7.17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HES4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
RP11-54O7.11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISG15	0	0	1	9	0	1	0	0	0	3	0	0	1	5	0	2	1	0	0	0
AGRN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C1orf159	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TNFRSF18	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TNFRSF4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	12	0	0
SDF4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
B3GALT6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
FAM132A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
UBE2J2	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0

Quality Control



Filtering the Counts Matrix

	AAACATACAACCAC	AAACATTGAGCTAC	AAACATTGATCAGC	AAACCGTGCTTCCG	AAACCGTGTATGCG	AAACGCACTGGTAC	AAACGCTGACCAGT	AAACGCTGCTTCTT	AAACGCTGTAGCCA	AAACGCTGTTTCTG	AAACTTGAAAAACG	AAACTTGATCCAGA	AAAGAGACGAGATA	AAAGAGACGCGAGA	AAAGAGACGGACTT	AAAGAGACGGCATT	AAAGCAGATATCGG	AAAGCCTGTATGCG	AAAGGCCTGTCTAG	AAAGTTTGATCACG
AL627309.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AP006222.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-206L10.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-206L10.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LINC00115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOC2L	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
KLHL17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PLEKHN1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP11-54O7.17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HES4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
RP11-54O7.11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISG15	0	0	1	9	0	1	0	0	0	3	0	0	1	5	0	2	1	0	0	0
AGRN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C1orf159	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TNFRSF18	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TNFRSF4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	12	0	0
SDF4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
B3GALT6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
FAM132A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
UBE2J2	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0

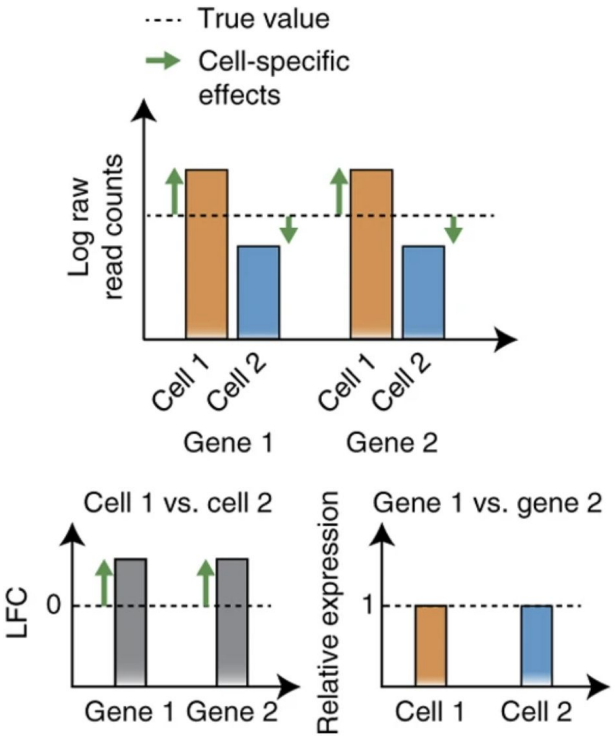
Normalisation

Differences in sequencing coverage between cells due to technical effects (e.g PCR amplification efficiency, amount of mRNA captured, etc)

Library size normalisation: divide counts in a cell by the total counts for that cell

Typically scaled by multiplying by 10000 and log transformed in Seurat

Seurat also has SCTransform - regularised negative binomial model



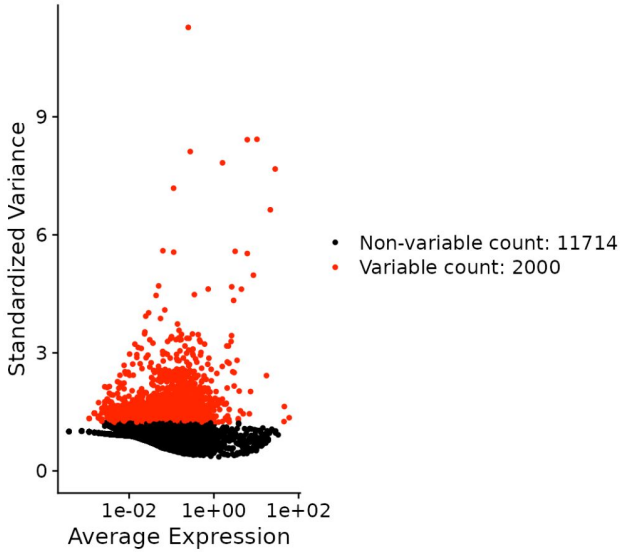
Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* **14**, 565–571 (2017)

Feature Selection

Select genes that have high variability to focus on interesting biological signal for downstream steps that aggregate or cluster cells based on similarity

Not all genes necessarily contain useful information, some contain random noise

Seurat's strategy is to pick the topmost variable 2000 genes

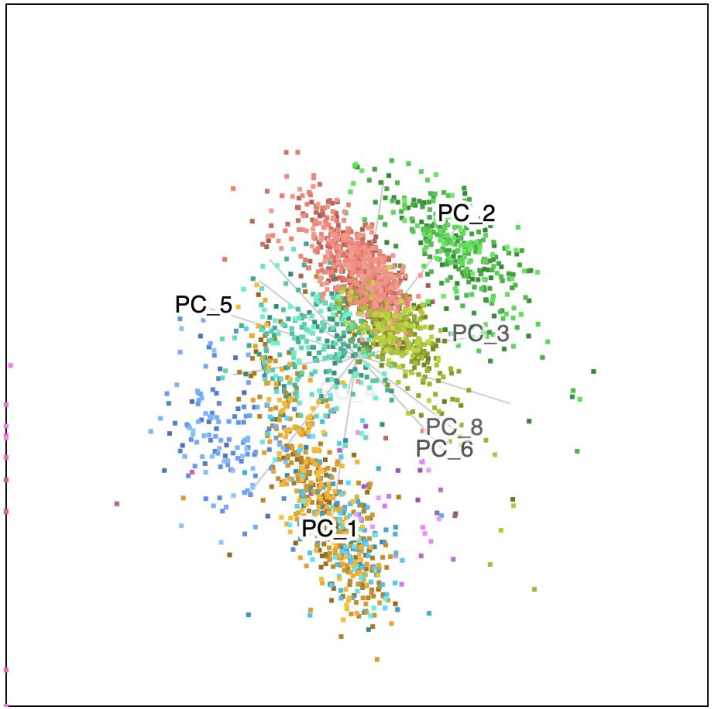


Dimensionality Reduction - PCA

If we view each gene as a dimension, cells inhabit a gene-space with **1000s of dimensions**.

Using Principal Components Analysis, most of the variation in a dataset can usually be summarized into **10s of components**.

This is convenient for many algorithms, but still difficult to visualize...



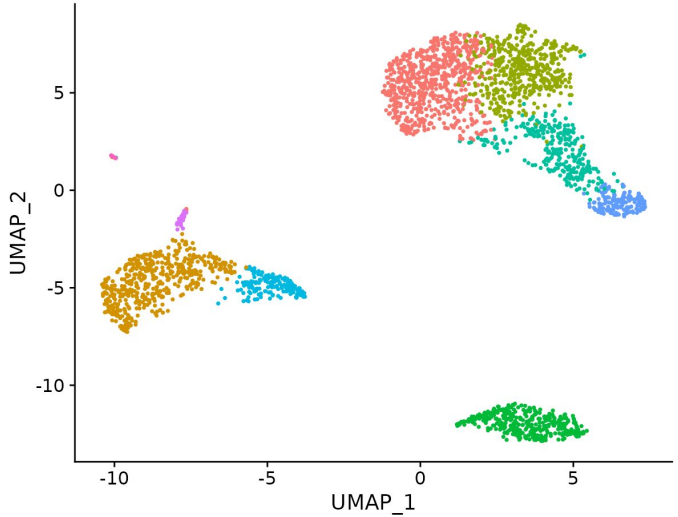
Dimensionality Reduction - UMAP

UMAP provides a further dimensionality reduction step from 10s of PCs to **2 dimensions** that can be easily visualized.

UMAP is a non-linear dimensionality reduction method.

- Very good at showing the structure of the data.
- May arbitrarily warp and tear the data to present it in 2D.

The UMAP layout is a useful map on which other data can be shown, such as clusters or the expression of particular genes.



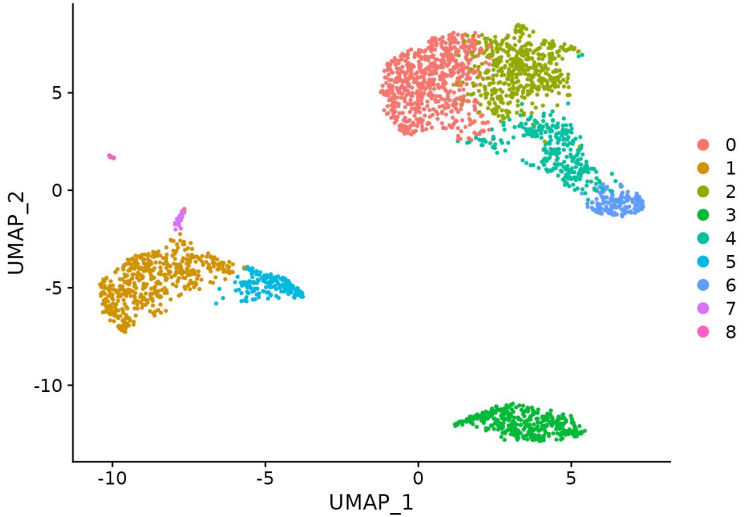
Clustering

Unsupervised learning technique to define groups of cells with similar expression profiles

Highly dependent on parameters chosen

Methods:

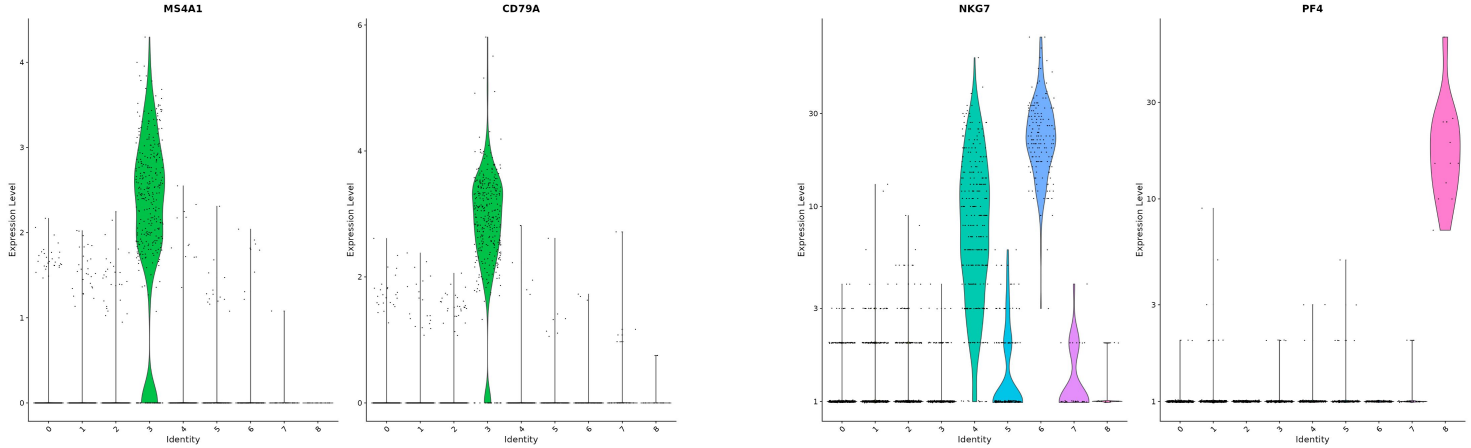
- Graph-based - Seurat - louvain/leiden algorithm
- K-means - loupe browser



Cluster Marker Identification

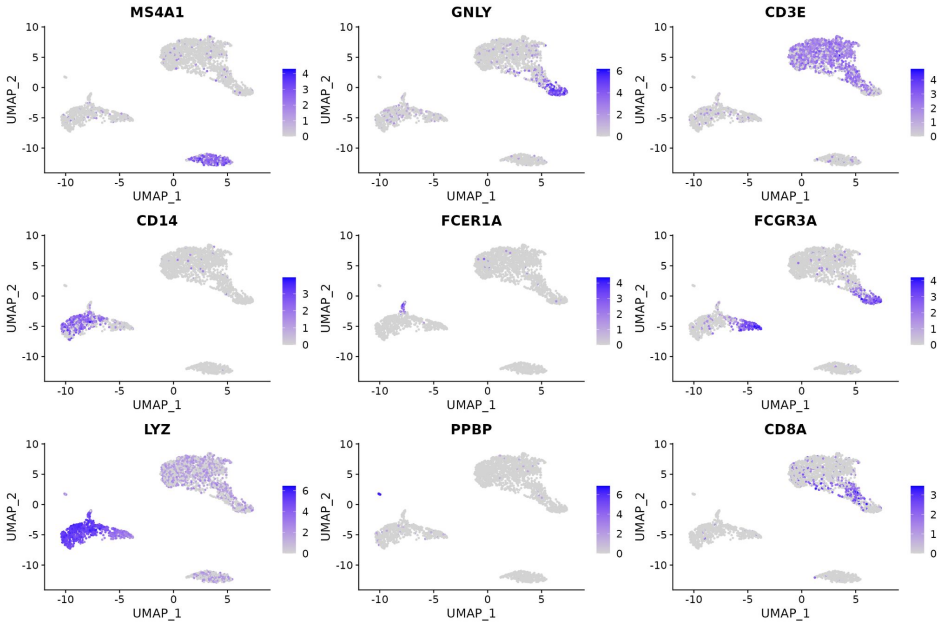
Identify genes that drive the differences between clusters

Use differential tests to get potential marker gene lists - Seurat defaults to the Wilcoxon test but implements several others e.g bimod, roc, t-test, DESeq2, poisson, etc



Manual Cell Annotation

- Use known cluster markers expression to determine cell type
- Requires domain specific knowledge
- Identifying cell types is probably the most time consuming step if you are working with uncharacterised cells
- Capturing cell surface protein expression can help



Automated Cell Annotation

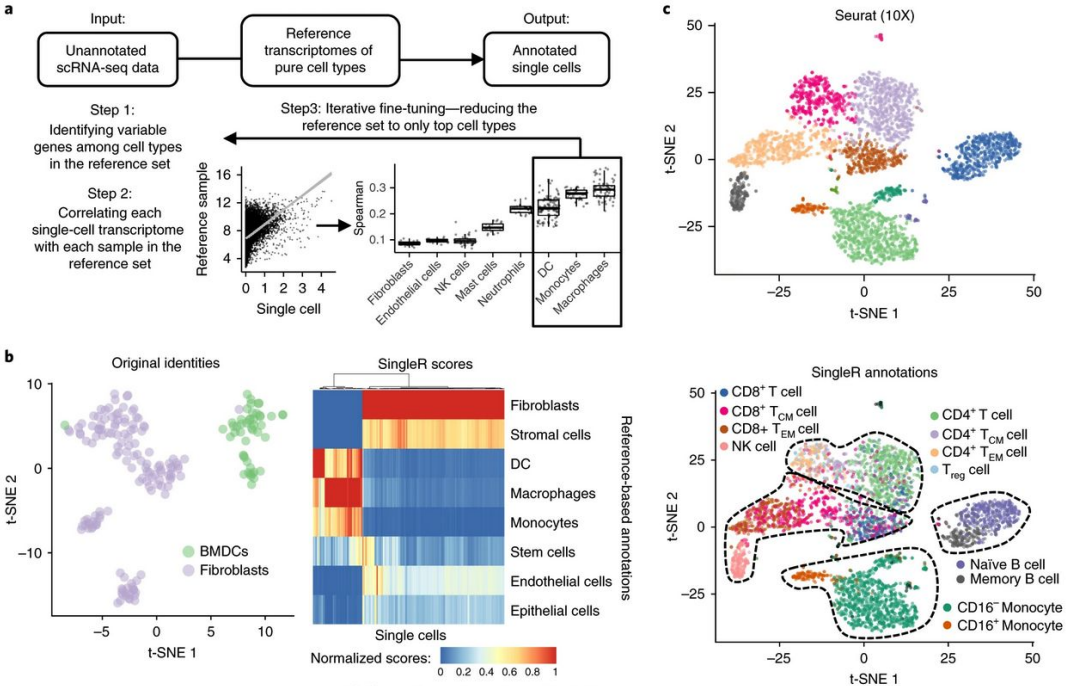
Use a reference dataset to annotate new data

Compare the expression profile of the new dataset against the reference and classify cells in the new dataset

Tools:

- singleR
- Azimuth (Seurat)
- scMatch
- scPred
- Garnett
- etc

Works well when your dataset is represented in the reference



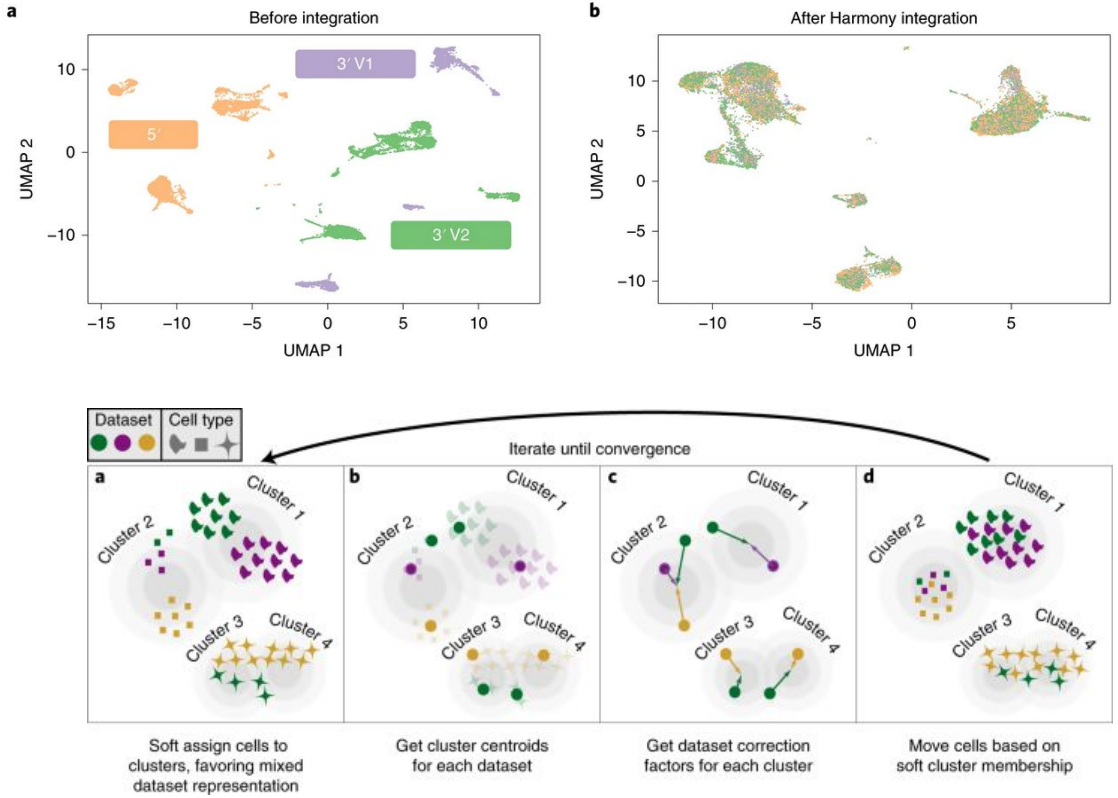
Aran, D., Looney, A.P., Liu, L. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163–172 (2019).

Dataset Integration/Batch Correction

Samples from different experiments can have substantial batch effects that need to be corrected before the data can be jointly analysed

Tools:

- Seurat
- Harmony
- batchelor
- bbknn
- etc



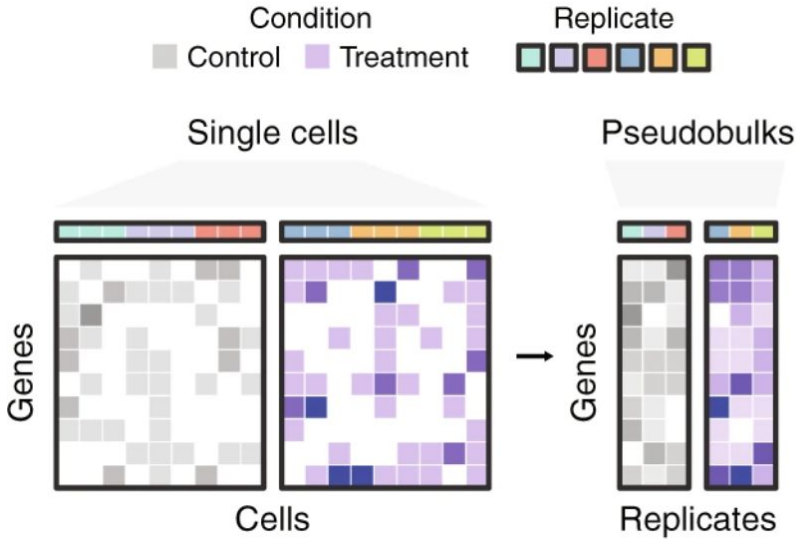
Replication

Original thought was that multiple cells provided replication

However, this is still an n = 1 and biological replicates are required to get valid p.values

Pseudobulking within a sample is reported to outperform single cell DE methods

Single cell DE methods are reported to have a bias towards highly expressed genes even when their expression remains unchanged



Squair, J.W., Gautier, M., Kathe, C. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* 12, 5692 (2021)


MONASH University
MONASH
BIOINFORMATICS
PLATFORM
Let's Get Started
GROUP
OF EIGHT
AUSTRALIA

Workshop Reminders

Ask questions - either in person or on Slack!

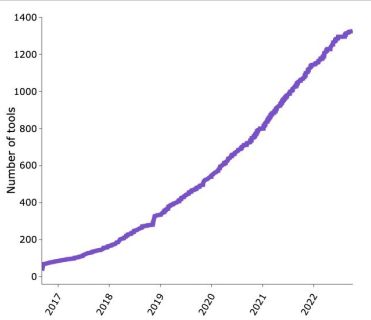
Ask for help if you need it - helpers are in the room to provide help!

Sticky notes - teal - don't need help, pink - need help

The learning material is freely available online - come back to it anytime

We run a help session every Friday @ 3pm - drop by if you have more questions

Single cell analysis is an ever growing and evolving field - this workshop covering how to use Seurat is just the tip of the iceberg



More

- Pseudotime trajectories
- Single cell velocity
- Multiomics
- Doublet detection
- Cell cycle
- Differential accessibility
- Immune Profiling
- Hashtag demultiplexing
- Spatial
- Pseudobulk
- etc…

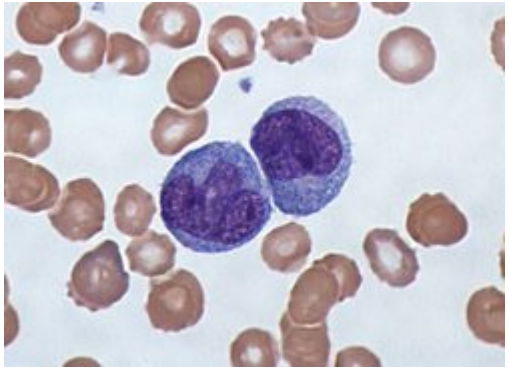
Tutorial

Human Peripheral Blood Mononuclear Cells (PBMC) freely available from 10X Genomics

There are 2,700 single cells that were sequenced on the Illumina NextSeq 500

PBMC - typically a mixture of lymphocytes and monocytes, commonly used in immunology research

Tutorial goal: identify the cell types in this PBMC sample



Further Resources

- Orchestrating Single Cell Analysis: this is one of the most comprehensive resource on learning about single cell analysis. It utilises the Bioconductor ecosystem but is well worth reading even if you stick with Seurat or use scanpy
- Seurat & Signac websites: lots of documentation on how to use these packages
- Ming Tang's scRNA analysis notes: huge list of single cell tools, tutorials, papers, etc organised by topic
- Awesome single cell: another extensive list of single cell tools, tutorials, papers, etc organised by topic
- scRNA-tools: a database that catalogues tools for analysing single-cell data