

Background  
oooooo

scRNA Protocols  
oooooooooooooooooooo

Data Analysis  
oooooooooooooooooooo

Trajectory Analysis  
ooooo

# Lecture 10: Single-Cell RNA Sequencing

## BIOINF3005/7160: Transcriptomics Applications

Dr Stephen Pederson

Bioinformatics Hub,  
The University of Adelaide

May 25th, 2020



THE UNIVERSITY  
of ADELAIDE

Background  
oooooo

scRNA Protocols  
oooooooooooooooooooo

Data Analysis  
oooooooooooooooooooo

Trajectory Analysis  
ooooo

## Background

## scRNA Protocols

## Data Analysis

## Trajectory Analysis



THE UNIVERSITY  
*of* ADELAIDE

Background  
●○○○○

scRNA Protocols  
○○○○○○○○○○○○○○○○

Data Analysis  
○○○○○○○○○○○○○○○○

Trajectory Analysis  
○○○○

# Background



THE UNIVERSITY  
of ADELAIDE

## Introduction

- scRNA-Seq is the ‘latest and greatest’ transcriptomic technique
  - Previously all our analysis involved multiple cells per sample
    - Now commonly known as bulk RNA-Seq
  - Large cell numbers during tissue extraction, library preparation etc.
  - Most experiments have **highly** heterogeneous cell populations, e.g.

## Introduction

- scRNA-Seq is the ‘latest and greatest’ transcriptomic technique
  - Previously all our analysis involved multiple cells per sample
    - Now commonly known as bulk RNA-Seq
  - Large cell numbers during tissue extraction, library preparation etc.
  - Most experiments have **highly** heterogeneous cell populations, e.g.
    - Different regions of the brain contain highly specialised cells
    - The immune system is highly complex
    - Cancer samples have both infiltrating and tumour cells

## Introduction

- If a gene is increased 2-fold in expression:
  - Is this 2-fold in 100% of cells?
  - Or is it 4-fold in 50% of cells?
  - Or is it down 2-fold in 25% and up 8-fold in 25% and unchanged in 50%?
- Changes in gene expression can be highly specific to individual cell-types
- Determining heterogeneity of our bulk samples is challenging



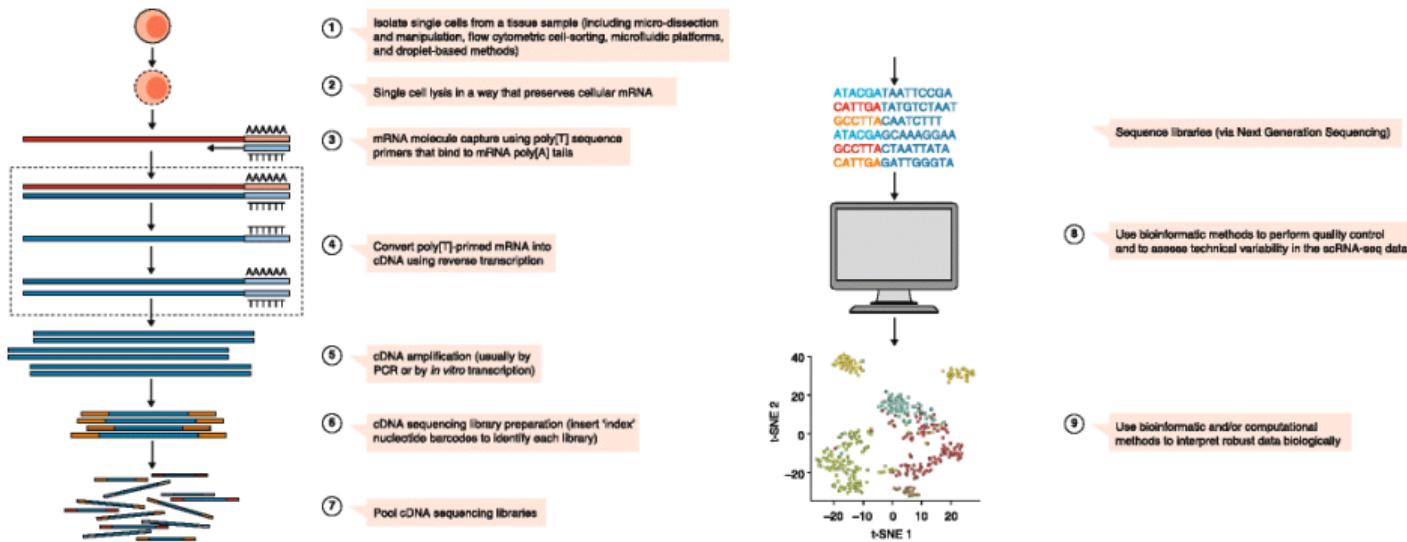
## Introduction

- The most intuitive solution is to obtain RNA from each cell and sequence
  - Reality is much trickier than this

## Introduction

- The most intuitive solution is to obtain RNA from each cell and sequence
  - Reality is much trickier than this
  - How do we characterise which cell is which cell-type?
  - What do we even mean by the term ‘cell-type’?
  - How do we capture as many transcripts from each cell as we can?
    - Missing values are a huge issue in scRNA-seq
  - How do we compare within the same cell-types between experimental groups?
    - e.g. treated and untreated cell types may not be easily assigned to the same cluster/cell-type

## Summarised scRNA Workflow



Taken from A. Haque et al. "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications". In: *Genome Med*. THE UNIVERSITY OF ADELAIDE 9.1 (Aug. 2017), p. 75

## Motivation

- Bulk RNA-Seq is primarily focussed on differentially expressed (DE) genes
- scRNA-Seq focusses on identifying cell-types within a sample
- How do we discriminate between different cell-types and different cell-states?
- What is the most intelligent approach for identifying DE genes
  - Is it between clusters/cell-types  $\implies$  marker genes
  - Is it between the same cell-types under differing treatments/cell-states?



Background  
oooooo

scRNA Protocols  
●oooooooooooooooooooo

Data Analysis  
oooooooooooooooooooo

Trajectory Analysis  
ooooo

# scRNA Protocols



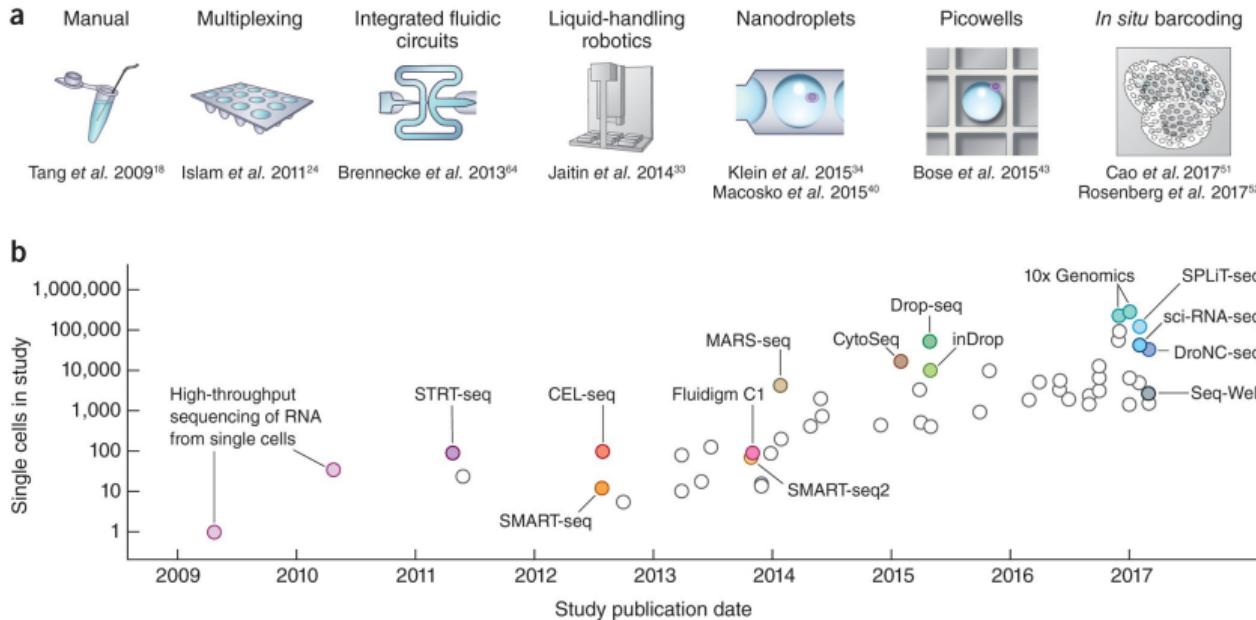
THE UNIVERSITY  
*of* ADELAIDE

## Isolating Individual Cells

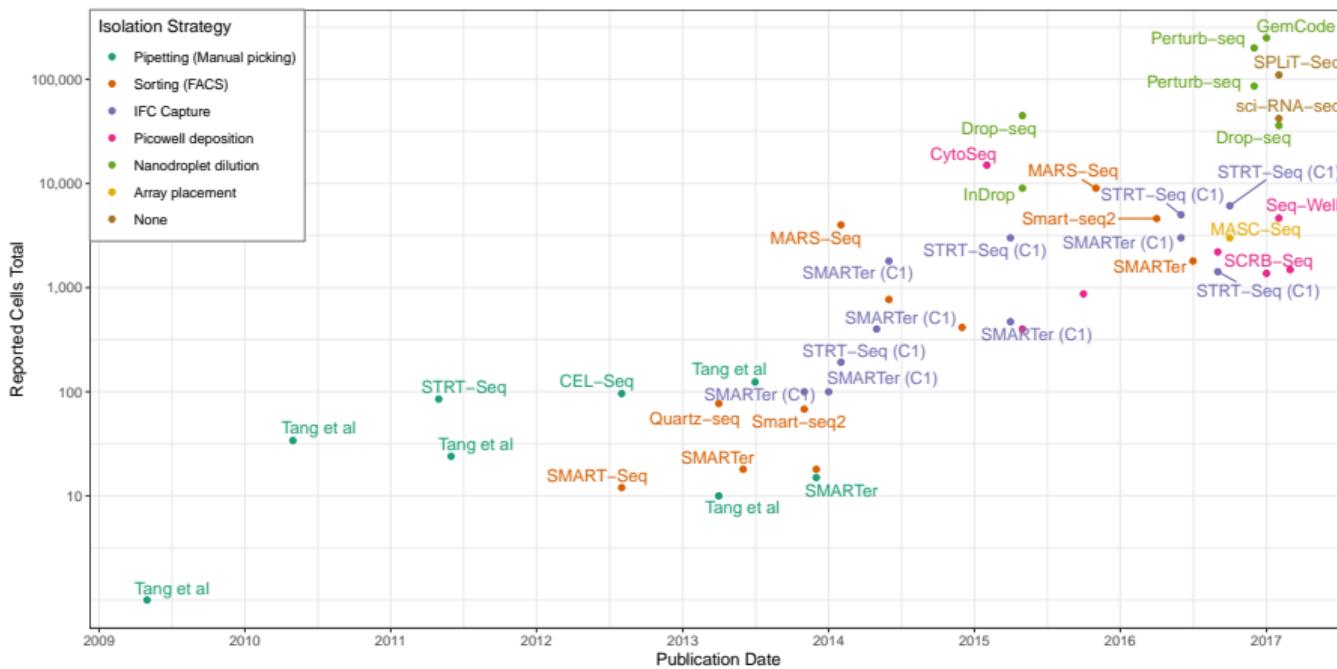
- Early protocols used a dilution series or manual isolation with a microscope (*micromanipulation*)
- Laser Capture Micro-dissection (LCM)
- Fluorescence-Activated Cell Sorting (FACS)
  - Labelled antibodies to specific surface markers
  - MACS is a magnetic-based approach
- Microfluidics/Droplet-based approaches
- Multiple rounds of splitting and pooling



# Protocol Timeline



# Protocol Timeline



Data taken from Svensson, Vento-Tormo, and Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade"

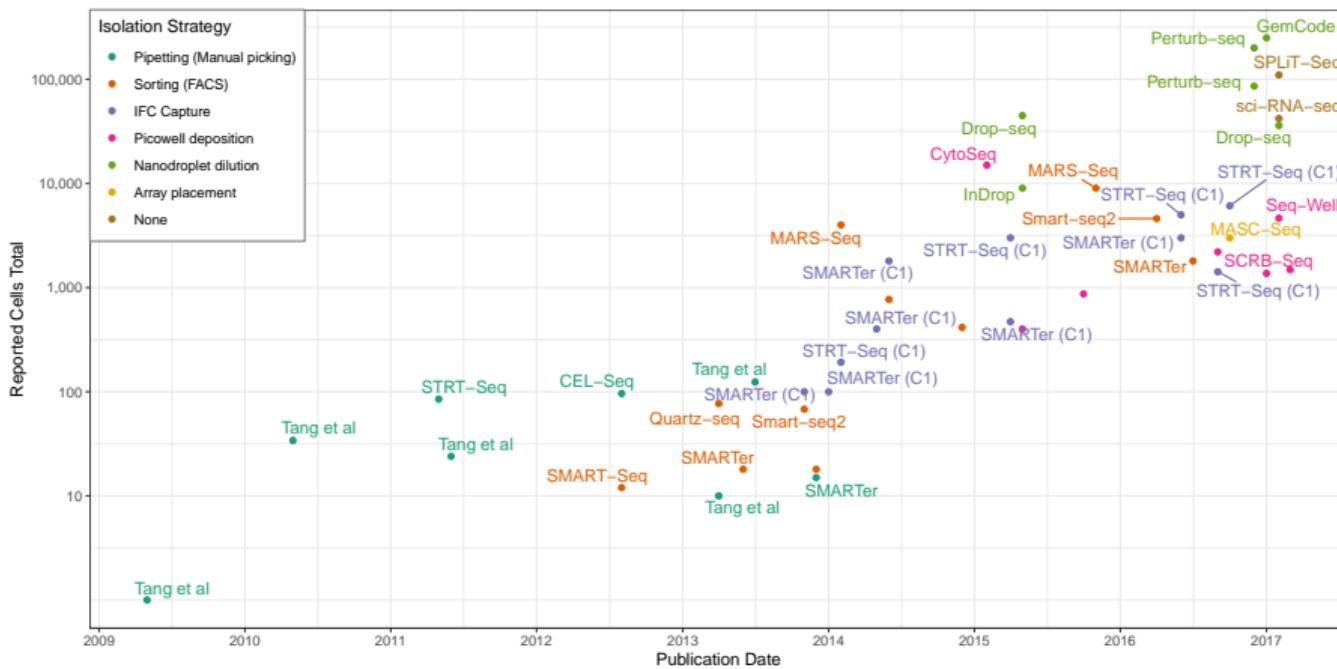


# IFC Capture

- Integrated Fluidic Circuit (IFC) chips
  - Most common is the Fluidigm C1
- Deliver tiny volumes into ‘reaction chambers’
- Early chips had 96 chambers  $\implies$  multiple chips / experiment
- Recent chips handle  $\sim$ 800 cells



# Protocol Timeline



Data taken from Svensson, Vento-Tormo, and Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade"



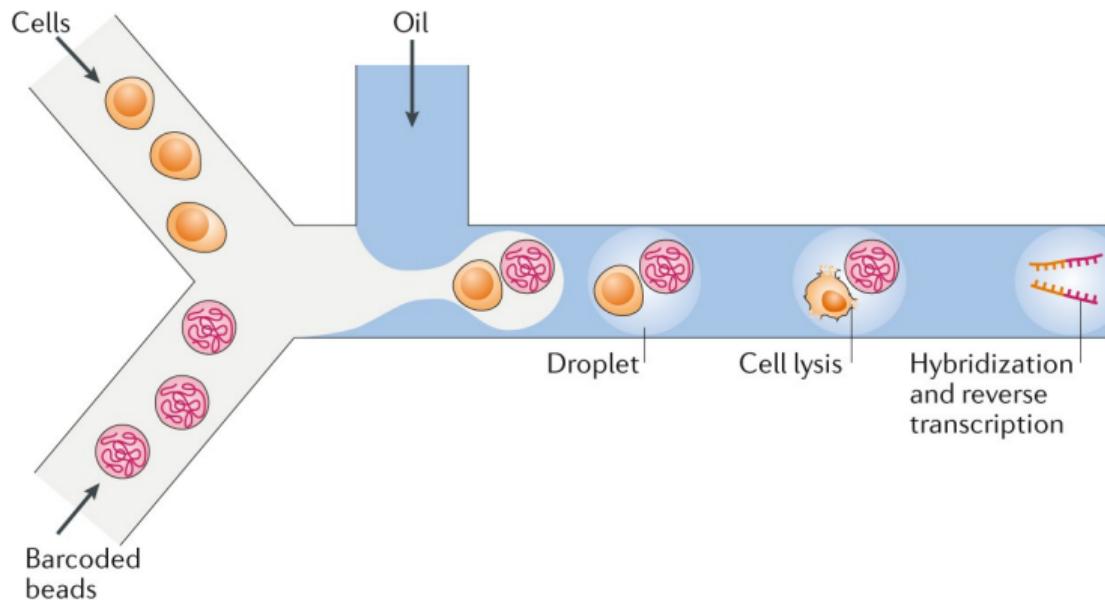
Background  
oooooo

scRNA Protocols  
oooooooo●oooooooooooo

Data Analysis  
oooooooooooooooooooo

Trajectory Analysis  
ooooo

## Droplet-based Approaches

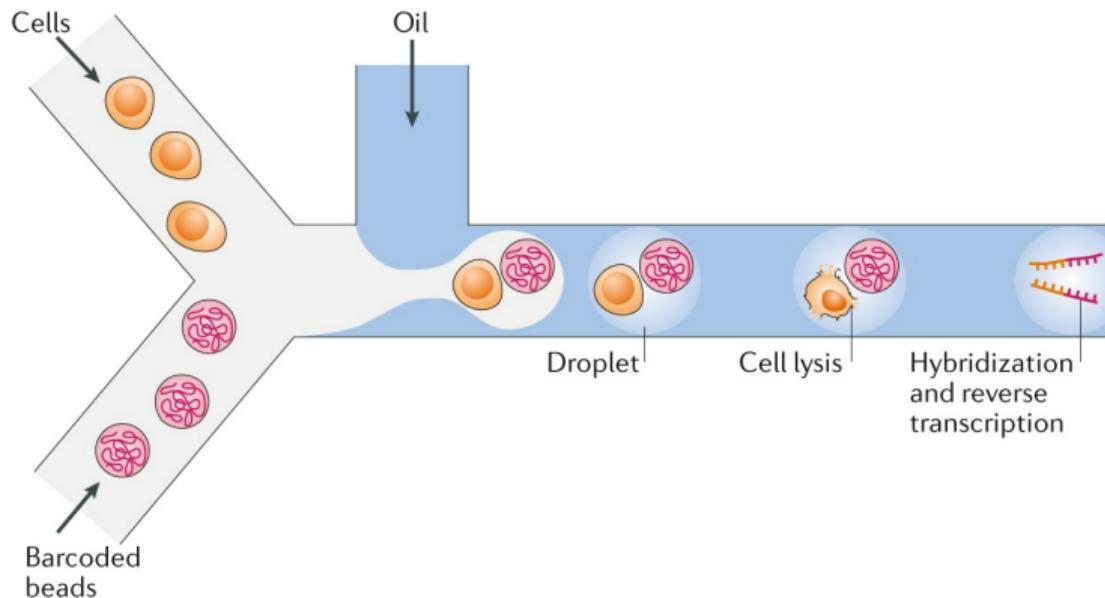


Taken from S. S. Potter. "Single-cell RNA sequencing for the study of development, physiology and disease". In: *Nat Rev Nephrol* 14.8 (Aug 2018), pp. 479–492



THE UNIVERSITY  
of  
ADELAIDE

## Droplet-based Approaches



Flow rate is modelled as a *Poisson* process to minimise doublets

Taken from Potter, "Single-cell RNA sequencing for the study of development, physiology and disease"



## Sequencing Overview

- Individual cells are isolated  $\implies$  how do we sequence?
- Need a method to track which reads come from which cell
- Sequencing is performed on a standard Illumina machine, i.e. multiplexed
- Each cell is essentially an individual library prep
  - Barcodes / UMIs are used for cell / molecule identification
- For bulk RNA-Seq we need  $0.1 - 1\mu\text{g}$  of RNA ( $10^5 - 10^6\text{pg}$ )
  - An individual cell contains 1-50pg RNA



Background  
oooooo

scRNA Protocols  
oooooooo●oooooooooooo

Data Analysis  
oooooooooooooooooooo

Trajectory Analysis  
ooooo

# SMART<sup>1</sup>-Seq (C1)

1. All reagents are in the IFC reaction chambers
2. Cells are lysed
3. polyA RNA reverse transcribed into **full length cDNA**
  - oligo(dT) priming and template switching
4. 12-18 PCR cycles
5. cDNA fragmentation and Adapter ligation



THE UNIVERSITY  
of ADELAIDE

---

<sup>1</sup>SMART = Switching Mechanism at 5' End of RNA Template

Background  
○○○○○

scRNA Protocols  
○○○○○○○●○○○○○○○○

Data Analysis  
○○○○○○○○○○○○○○○○

Trajectory Analysis  
○○○○

# SMART-Seq (C1)

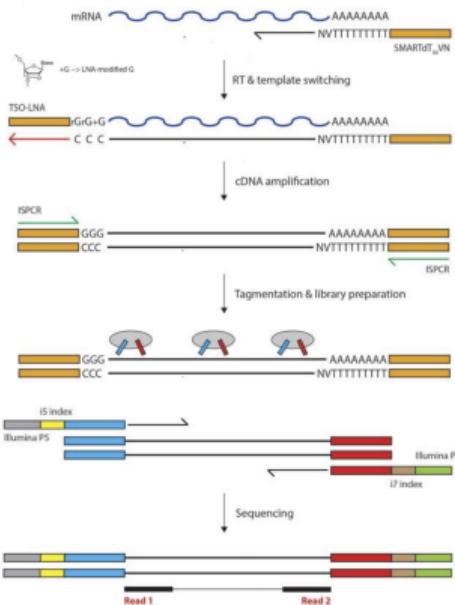


Image from S. Picelli. "Single-cell RNA-sequencing: The future of genome biology is now". In: *RNA Biol* 14.5 (May 2017), pp. 637–650



THE UNIVERSITY  
of ADELAIDE

## Droplet-based Methods

- Popularised by the 10X Genomics Chromium System
- Each gel bead contains the reagents
  - 30nt poly(dT) primer with 16nt 10x Barcode, 12nt UMI<sup>2</sup>
- Illumina primers and restriction enzymes added later

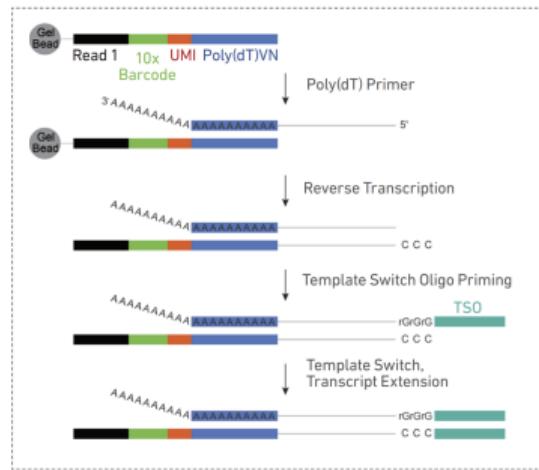
---

<sup>2</sup>Unique Molecular Identifier



# 10X Chromium Protocol

Inside individual GEMs



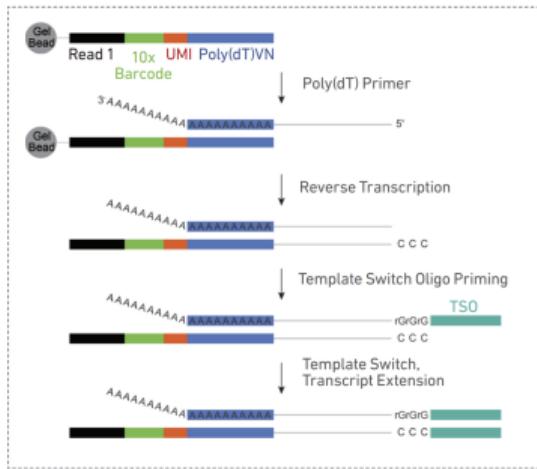
Barcoded, full-length cDNA is pooled then  
PCR amplified



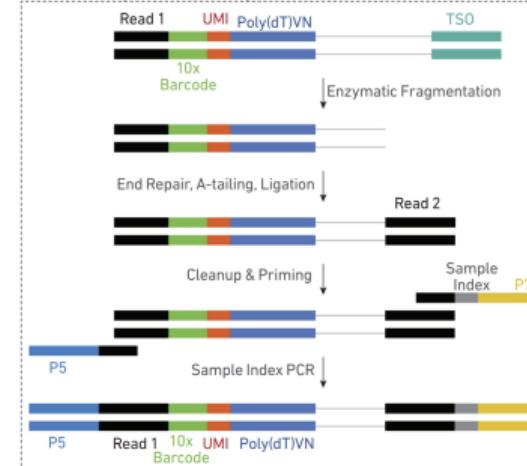
THE UNIVERSITY  
of ADELAIDE

# 10X Chromium Protocol

Inside individual GEMs



Pooled amplified cDNA processed in bulk



Barcoded, full-length cDNA is pooled then PCR amplified



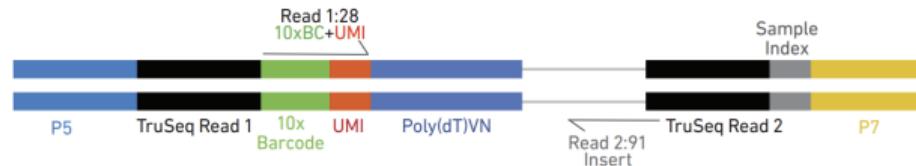
Background  
ooooooo

scRNA Protocols  
oooooooooooo●ooooo

Data Analysis  
oooooooooooooooooooo

Trajectory Analysis  
oooooo

# 10X Chromium Protocol



- Only R2 contains the sequence information
- Only the 3' end is sequenced
- Each template RNA should have one UMI  $\implies$  PCR duplicates can be identified



THE UNIVERSITY  
of ADELAIDE

## Other Variations

### CITE-Seq<sup>3</sup>

- Prior to sorting cells can be 'labelled' with antibody-oligo complexes
- Oligos allow additional recognition of surface proteins
- On cell lysis these oligos are amplified along with RNA



## Other Variations

### SPLIT-Seq<sup>4</sup>

- Cells are split into pools and fixed
- One barcode/pool
- Multiple rounds of pooling and barcoding
- All amplification is *in situ*
- Able to be applied to single nuclei



## Comparison of Methods

Protocol	C1 (SMART-Seq)	SMART-Seq2	10X Chromium	SPLIT-Seq
<i>Platform</i>	Microfluidics	Plate-based	Droplet	Plate-based
<i>Transcript</i>	Full-length	Full-length	3'-end	3'-end
<i>Cells</i>	$10^2 - 10^3$	$10^2 - 10^3$	$10^3 - 10^4$	$10^3 - 10^5$
<i>Reads/Cell</i>	$10^6$	$10^6$	$10^4 - 10^5$	$10^4$



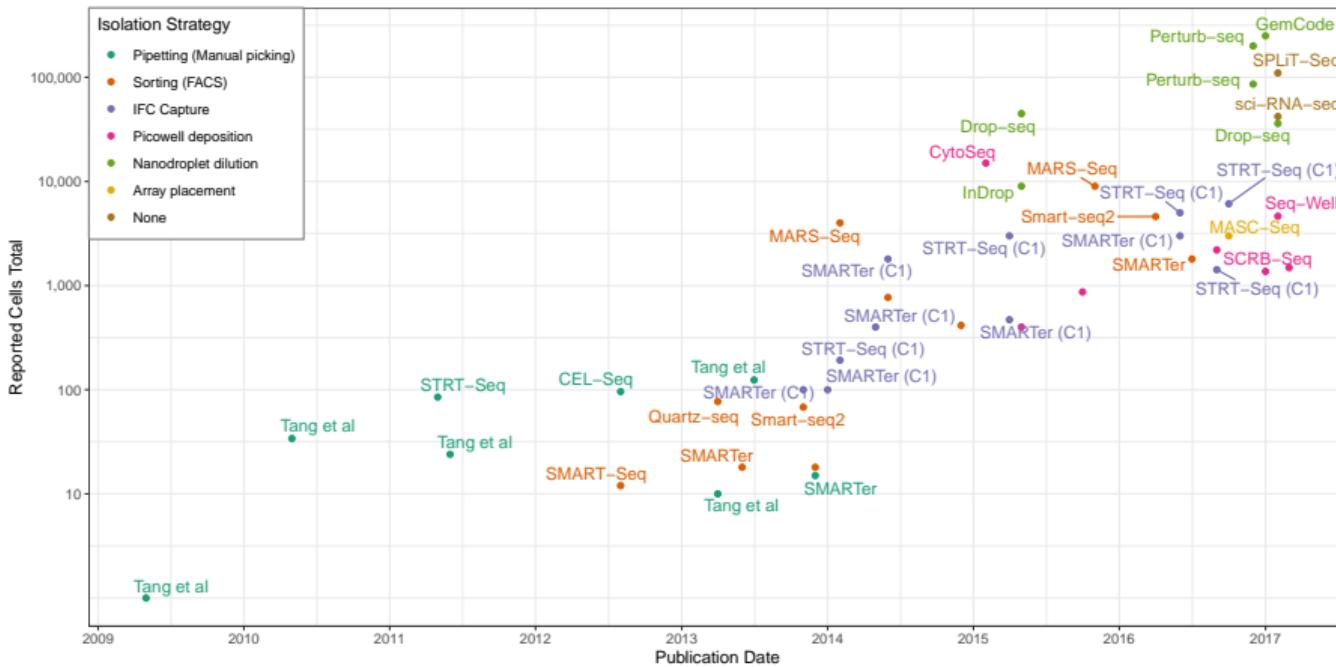
## Comparison of Methods

Protocol	C1 (SMART-Seq)	SMART-Seq2	10X Chromium	SPLIT-Seq
Platform	Microfluidics	Plate-based	Droplet	Plate-based
Transcript	Full-length	Full-length	3'-end	3'-end
Cells	$10^2 - 10^3$	$10^2 - 10^3$	$10^3 - 10^4$	$10^3 - 10^5$
Reads/Cell	$10^6$	$10^6$	$10^4 - 10^5$	$10^4$

Saturation for detection of expressed genes occurs around  $5 \times 10^5$  reads/cell



## Protocol Timeline



Data taken from Svensson, Vento-Tormo, and Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade".

## Technical Challenges

- How to detect intact/viable cells, free RNA etc
- How to ensure only single cells captured, i.e. no doublets
- Unbiased of sampling of RNA molecules (e.g. PCR impacts) and individual cells
  - Large numbers of zero counts for expressed genes
  - Lack of evidence for expression  $\neq$  evidence for lack of expression
- Efficiency of cell capture ( $\sim 50\%$  for 10X)
- How to deal with batch effects
  - Cells from each treatment group are always prepared separately



Background  
oooooo

scRNA Protocols  
oooooooooooooooooooo

Data Analysis  
●oooooooooooooooooooo

Trajectory Analysis  
ooooo

# Data Analysis



THE UNIVERSITY  
*of* ADELAIDE

Background  
oooooo

scRNA Protocols  
oooooooooooooooooooo

Data Analysis  
o●oooooooooooooooo

Trajectory Analysis  
oooo

## Automated Pipelines

- Most pre-processing for 10X data is performed using CellRanger
- Handles demultiplexing, alignment (STAR) and quantification (using UMIs)
  - Full-length transcript methods can utilise kallisto/salmon
- We end up with a **feature-barcode matrix**
  - A **barcode** represents an individual cell (or a set of reactions)
  - A **feature** is commonly thought of as a gene in scRNA-Seq
  - Other single-cell approaches (e.g. scATAC-Seq) are not gene focussed
- Similar to counts from bulk RNA-Seq but with many more columns (cells)



Background  
oooooo

scRNA Protocols  
oooooooooooooooooooo

Data Analysis  
oo●oooooooooooooo

Trajectory Analysis  
oooo

## Filtering

- We need to keep the high quality cells and discard the dubious cells, such as:
  1. Low/High read numbers (library sizes)
  2. Low feature/gene numbers
  3. High proportions of mitochondrial RNA  $\implies$  cells broken prior to lysis



## Filtering

- We need to keep the high quality cells and discard the dubious cells, such as:
  1. Low/High read numbers (library sizes)
  2. Low feature/gene numbers
  3. High proportions of mitochondrial RNA  $\implies$  cells broken prior to lysis
- Also need a method for considering each gene as detectable (Average Counts  $> 1$ )
  - Treatment Groups and Cell-Types are less easily defined *a priori*



## Normalisation

- Cell-specific offsets are once again calculated
  - Each cell is it's own source of variability
- Methods such as TMM are heavily influenced by the large numbers of zero counts
- Pooling and deconvolution:
  1. Perform rudimentary clustering of cells
  2. Normalise across all clusters (TMM assumes most genes are not DE)
  3. Deconvolute cells and normalisation factors
- Calculate log-transformed, normalised expression values (logcounts)



# Clustering

- A key process is grouping similar cells with each other  $\implies$  identifying cell-types
- To speed this up, we often choose the most highly variable genes (HVGs)
- Perform dimensional reduction:
  - PCA is the preferred linear approach, with non-linear approaches being:
  - tSNE (t-Distributed Stochastic Neighbour Embedding)
  - UMAP (Uniform Manifold Approximation and Projection)
- Both tSNE and UMAP are highly sensitive to parameter choice



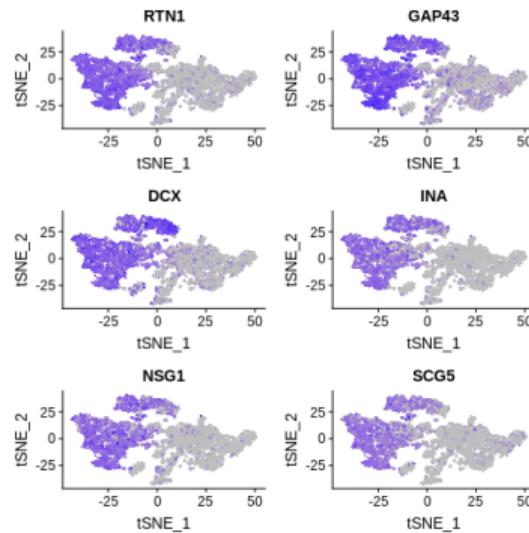
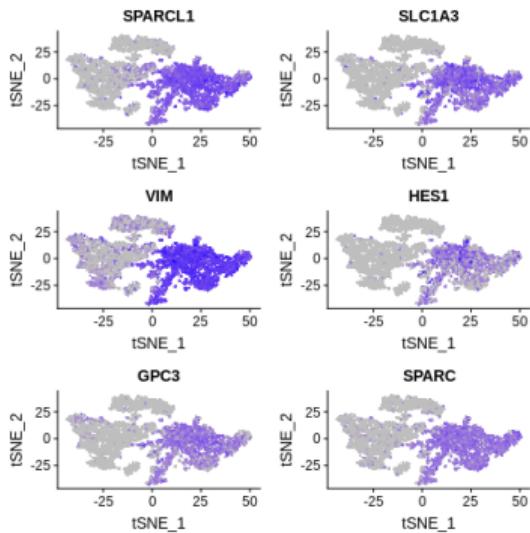
Background  
ooooooo

scRNA Protocols  
oooooooooooooooooooo

Data Analysis  
ooooo●oooooooooooo

Trajectory Analysis  
ooooo

# Clustering



# Clustering

- Formation of clusters allows for *identification of cell-types*
- Is there a “ground truth”?
- Different approaches will provide different results
- Different parameter settings will provide different results
- Each approach could be considered an alternate view-point on the data
  - Some viewpoints reveal particular information
  - Alternate viewpoints reveal different insights
- These are not necessarily contradictory
- Clusters are essentially *artificial constructs* used to represent one or more biological features



## Graph-Based Clustering

- Common approaches are *k*-nearest neighbours / shared neighbour weighting
- Relatively efficient computationally
- Uses the reduced dimensional data **not gene expression**
  - Commonly PCA with some optimising for the number of retained PCs
- Represents the similarity between cells as an “edge weight”
- No assumption about ‘shape’ of any clustering
- Clusters are identified using *Community Detection*



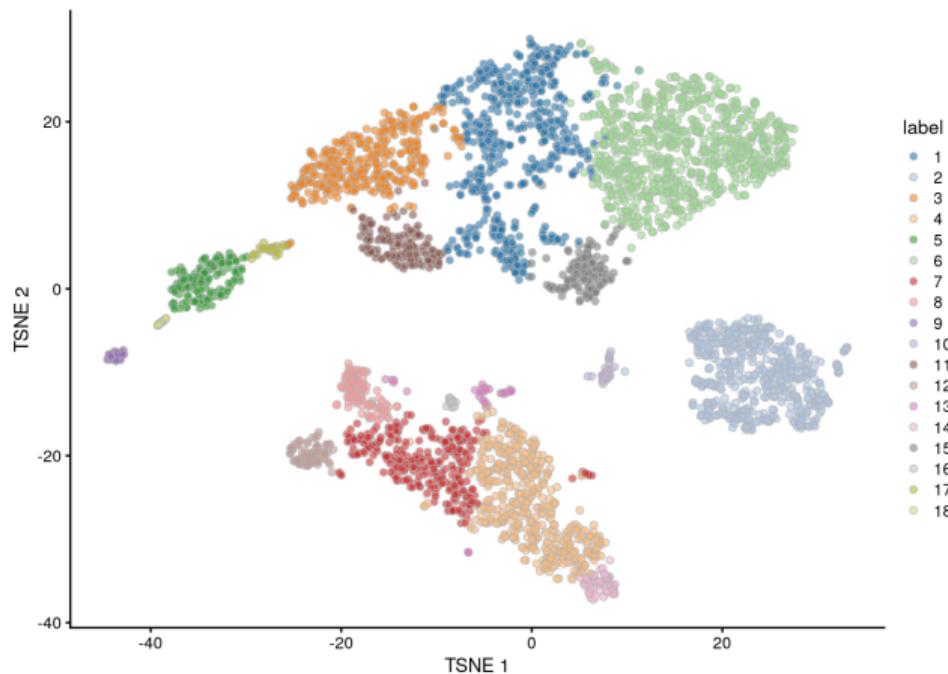
## Background

scRNA Protocols  
oooooooooooooooooooo

## Data Analysis

## Trajectory Analysis

## Visualising Clusters: tSNE



THE UNIVERSITY  
of ADELAIDE

Image taken from Orchestrating Single-Cell Analysis with Bioconductor

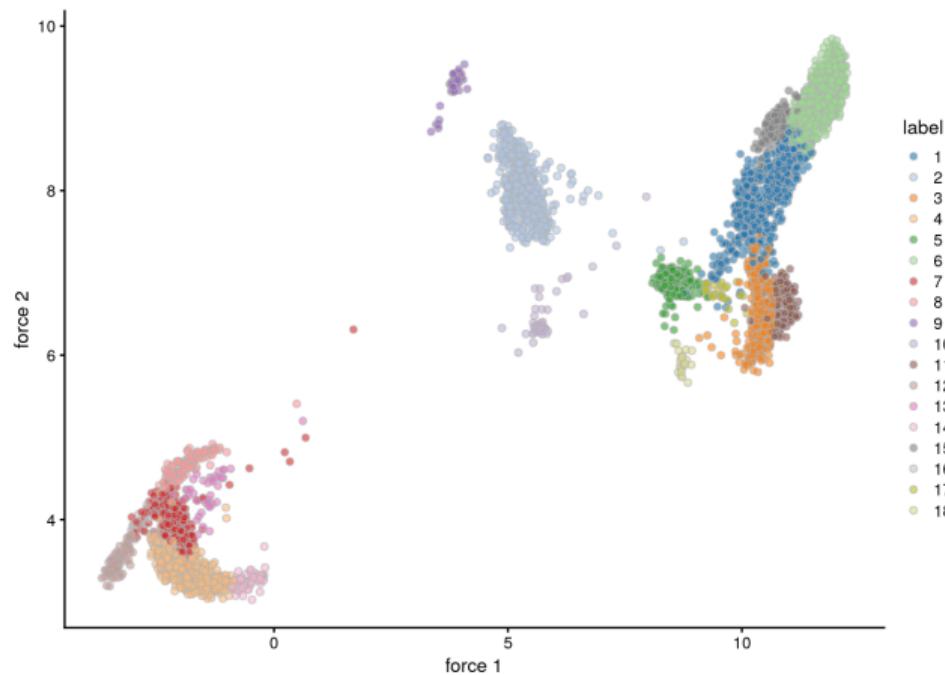
## Background

scRNA Protocols  
oooooooooooooooooooo

## Data Analysis

## Trajectory Analysis

## Visualising Clusters: Force-Directed Layout



THE UNIVERSITY  
of ADELAIDE

Image taken from Orchestrating Single-Cell Analysis with Bioconductor

## Graph-Based Clustering

- Forcing a minimum number of neighbours minimises small clusters
  - Choosing large  $k$  gives fewer larger clusters
- Clustering is performed in high-dimensions (e.g. using 10PCs) but visualised in 2
- Is essentially an exploratory process



Background  
ooooooo

scRNA Protocols  
oooooooooooooooooooo

Data Analysis  
oooooooooooo●oooo

Trajectory Analysis  
ooooo

# Graph-Based Clustering

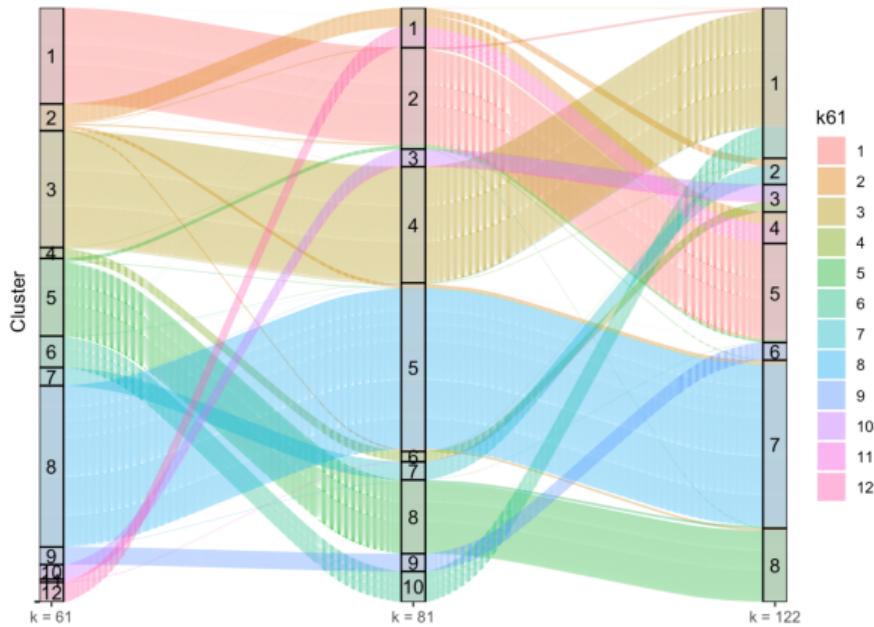


Image from Junwei Wang, Masters Thesis, 2019



THE UNIVERSITY  
of ADELAIDE

## Alternative Clustering Methods

- We can use  $k$ -Means  $\implies$  assumes  $k$  multi-dimensional spheres
- $k$  explicitly sets the number of clusters



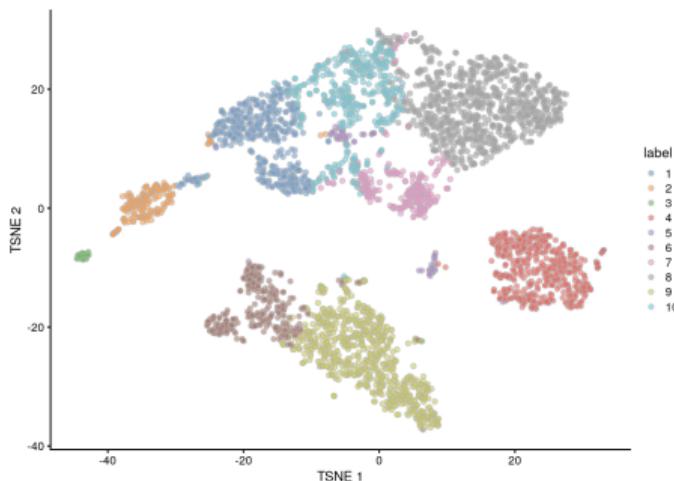
Background  
○○○○○

scRNA Protocols  
○○○○○○○○○○○○○○○○

Data Analysis  
○○○○○○○○○●○○○

Trajectory Analysis  
○○○○

## Alternative Clustering Methods



Setting  $k = 10$



## Background

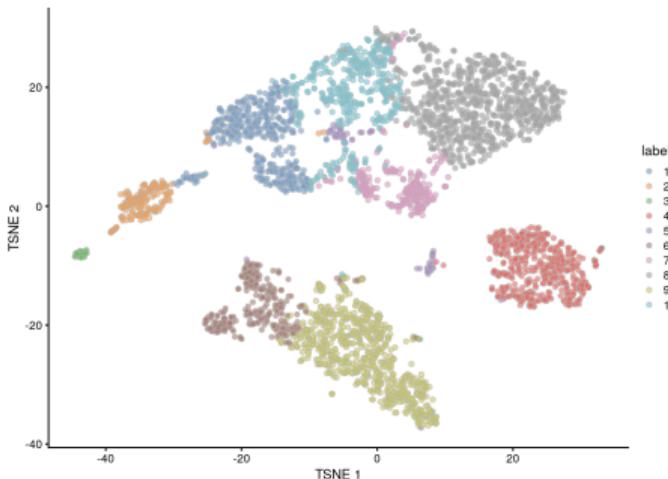
## scRNA Protocols



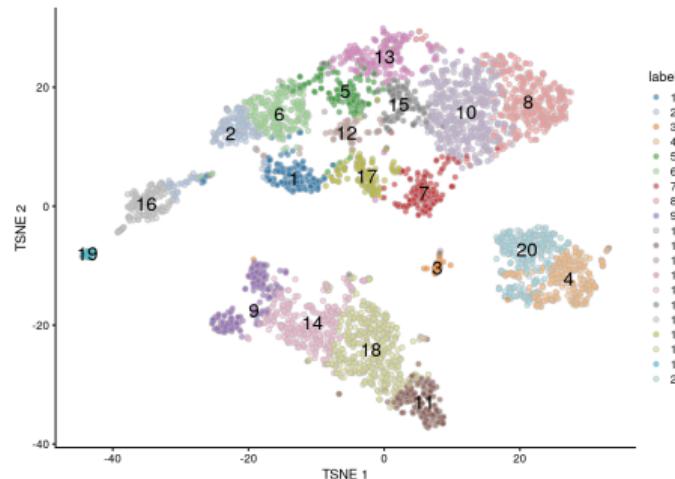
## Data Analysis

## Trajectory Analysis

## Alternative Clustering Methods



Setting  $k = 10$



Setting  $k = 20$

## Marker Selection

- An alternative perspective to differential expression  $\implies$  marker gene selection
- We find which genes *define* one or more clusters  $\implies$  identify known/unknown cell types
- Can also use known markers from CITE-Seq to identify cell-types
- Each cluster needs to be compared to all other clusters
  - Can use *t*-tests, limma/voom, edgeR
  - For unique markers, choose the maximal p-value across all comparisons



Background  
○○○○○

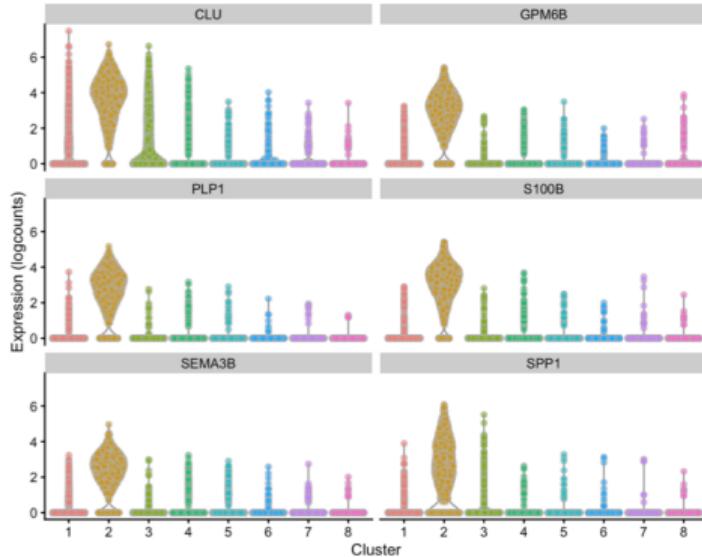
scRNA Protocols  
○○○○○○○○○○○○○○○○

Data Analysis  
○○○○○○○○○○○●○

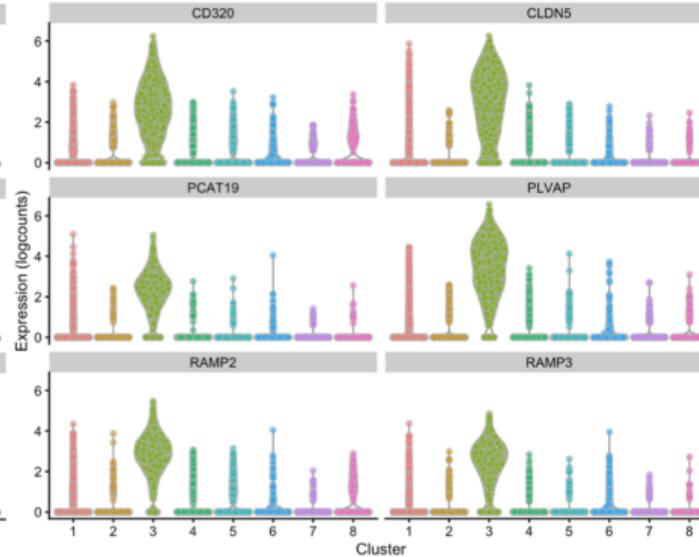
Trajectory Analysis  
○○○○○

# Marker Selection

Cluster 2



Cluster 3



Background  
oooooo

scRNA Protocols  
oooooooooooooooooooo

Data Analysis  
oooooooooooooooooooo●

Trajectory Analysis  
ooooo

## Marker Selection

- Often needs close discussion with biologist
- Relies on their expertise and knowledge of existing markers
- Still much scope for identifying new marker genes and cell-types



THE UNIVERSITY  
of ADELAIDE

Background  
oooooo

scRNA Protocols  
oooooooooooooooooooo

Data Analysis  
oooooooooooooooooooo

Trajectory Analysis  
●oooo

# Trajectory Analysis



THE UNIVERSITY  
*of* ADELAIDE

# RNA Dynamics

- With scRNA, we have dynamic information about each specific cell
- We can see which transcripts are being actively up/down-regulated
- Many introns also have poly(A) sites which are unintended targets for amplification
- If an RNA is being *actively* transcribed we may see introns
- If an RNA *has been processed* and exported we only see exons



# RNA Dynamics

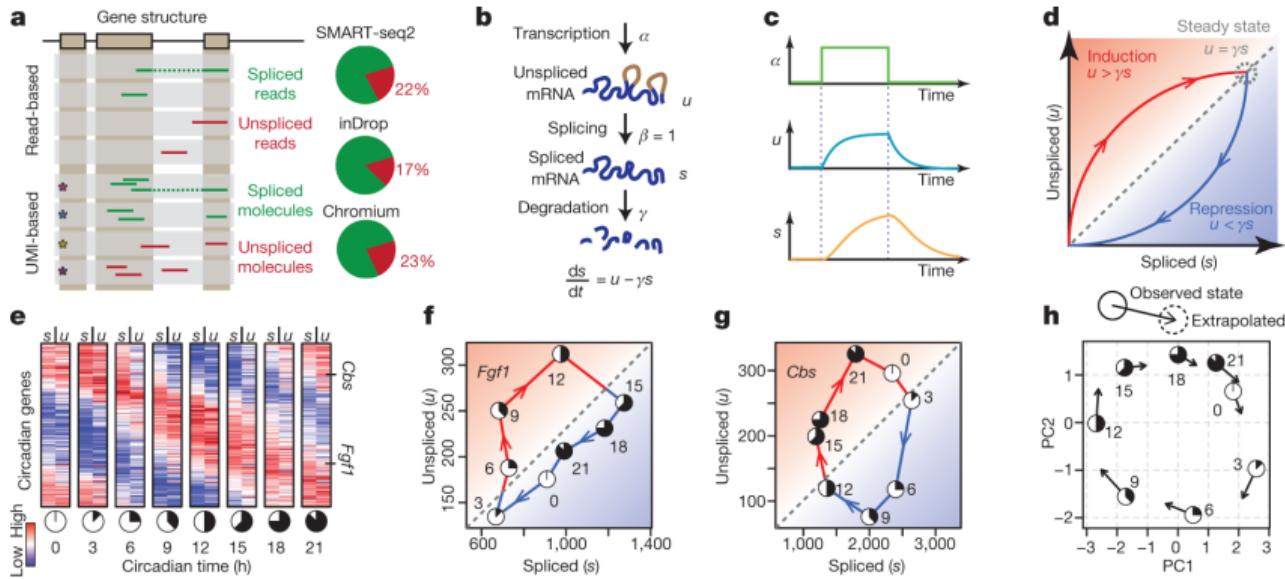


Image taken from G. La Manno et al. "RNA velocity of single cells". In: *Nature* 560.7719 (Aug. 2018), pp. 494–498

# RNA Velocity

- RNA Velocity is defined as the rate of change in spliced RNA with time
- We obtain a value for each gene in each sample
- Use dimensional reduction techniques  $\implies$  project a velocity field onto clusters



# RNA Velocity

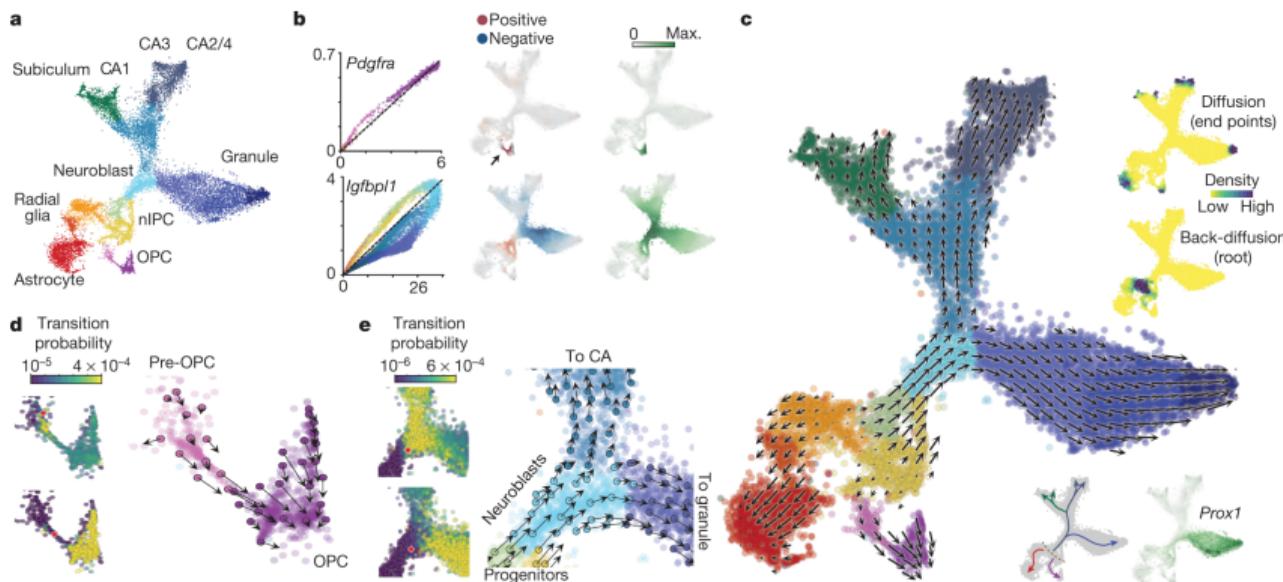


Image taken from La Manno et al., "RNA velocity of single cells"



THE UNIVERSITY  
of ADELAIDE