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

scRNA Protocols

Data Analysis

Background

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

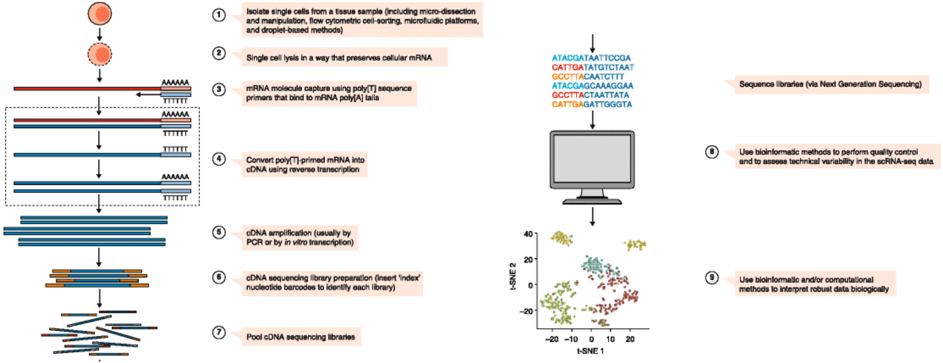
Introduction

- 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



① Isolate single cells from a tissue sample (including micro-dissection and manipulation, flow cytometric cell-sorting, microfluidic platforms, and droplet-based methods)

② Single cell lysis in a way that preserves cellular mRNA

③ mRNA molecule capture using poly[T] sequence primers that bind to mRNA poly[A] tails

④ Convert poly[T]-primed mRNA into cDNA using reverse transcription

⑤ cDNA amplification (usually by PCR or by *in vitro* transcription)

⑥ cDNA sequencing library preparation (insert 'index' nucleotide barcodes to identify each library)

⑦ Pool cDNA sequencing libraries

Sequence libraries (via Next Generation Sequencing)

⑧ Use bioinformatic methods to perform quality control and to assess technical variability in the scRNA-seq data

⑨ Use bioinformatic and/or computational methods to interpret robust data biologically

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

Motivation

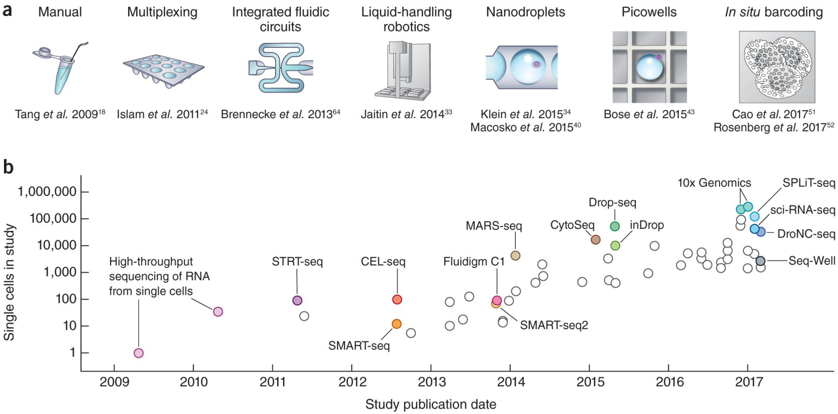
- Bulk RNA-Seq is primarily focussed on differentially expressed (DE) genes
- scRNA-Seq focusses on identifying cell-types within a sample
- How do we discriminate between different cell-types and different cell-states?
- What is the most intelligent approach for identifying DE genes
  - Is it between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

scRNA Protocols

Isolating Individual Cells

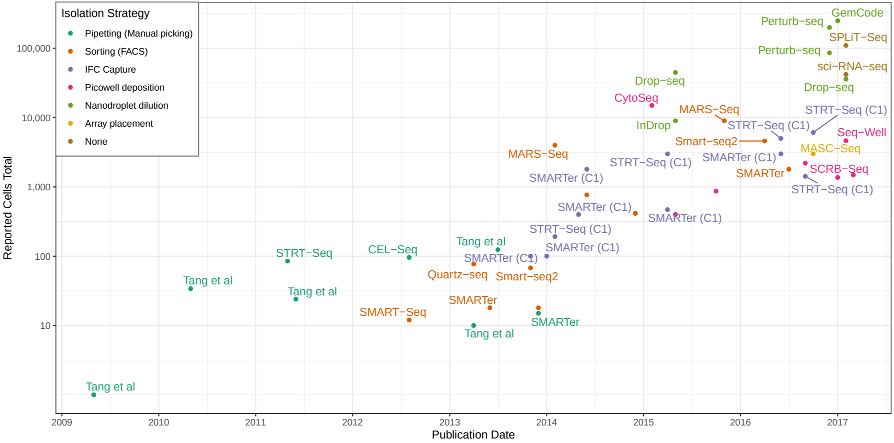
- Early protocols used a dilution series or manual isolation with a microscope (*micromanipulation*)
- Laser Capture Micro-dissection (LCM)
- Fluorescence-Activated Cell Sorting (FACS)
  - Labelled antibodies to specific surface markers
  - MACS is a magnetic-based approach
- Microfluidics/Droplet-based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

Protocol Timeline



Taken from V. Svensson, R. Vento-Tormo, and S. A. Teichmann. "Exponential scaling of single-cell RNA-seq in the past decade". In: *Nat Protoc* 13.4 (Apr. 2018), pp. 599–604

Protocol Timeline



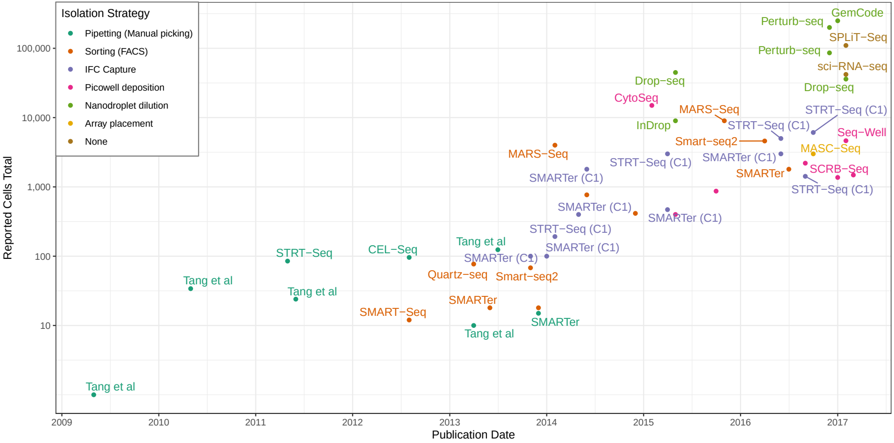
Data taken from Svensson, Vento-Tormo, and Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade"

THE UNIVERSITY *of* ADELAIDE

IFC Capture

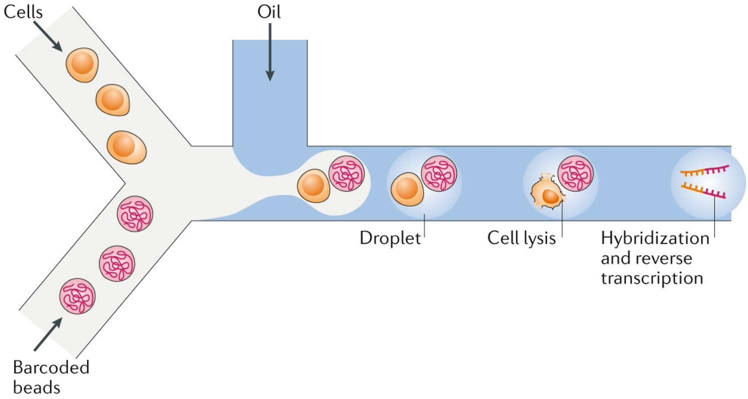
- Integrated Fluidic Circuit (IFC) chips
  - Most common is the Fluidigm C1
- Deliver tiny volumes into ‘reaction chambers’
- Early chips had 96 chambers $\implies$ multiple chips / experiment
- Recent chips handle ∼800 cells

Protocol Timeline



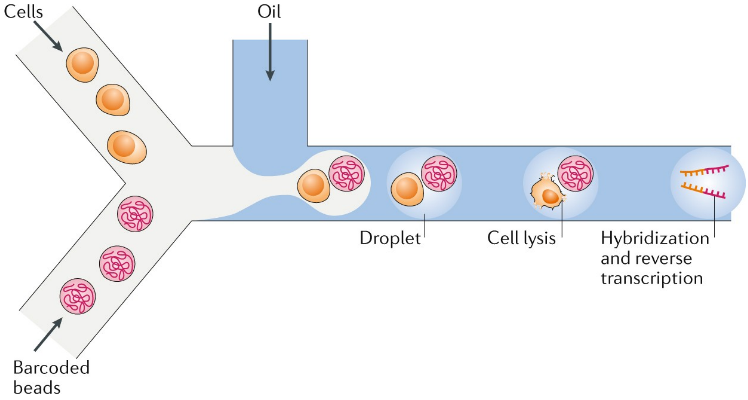
Data taken from Svensson, Vento-Tormo, and Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade"

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

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 $0.1 - 1\mu$g of RNA ($10^5 - 10^6$pg)
    - An individual cell contains 1-50pg RNA

SMART[1]-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

[1] SMART = Switching Mechanism at 5' End of RNA Template

SMART-Seq (C1)

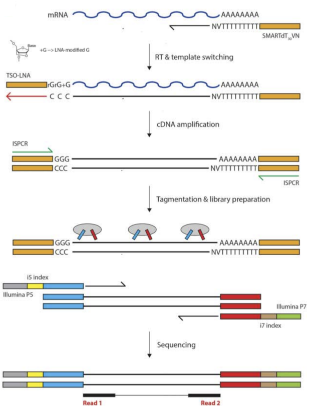


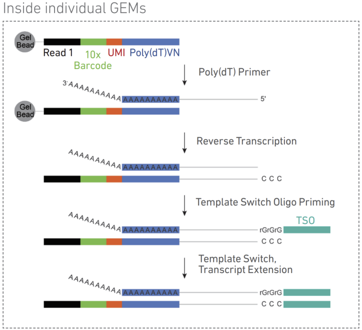
Image from S. Picelli. “Single-cell RNA-sequencing: The future of genome biology is now”. In: *RNA Biol* 14.5 (May 2017), pp. 637–650

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[2]
- Illumina primers and restriction enzymes added later

[2] Unique Molecular Identifier

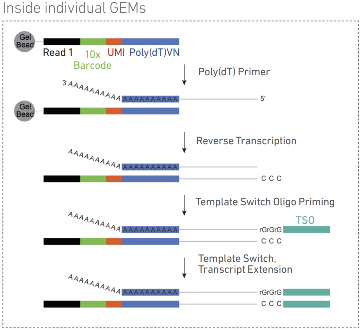
10X Chromium Protocol

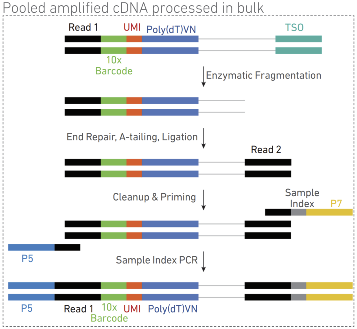


Barcoded, full-length cDNA is pooled then PCR amplified

Images from 10X Genomics CG000204_ChromiumNextGEMSingleCell3_v3.1_Rev_D.pdf

10X Chromium Protocol



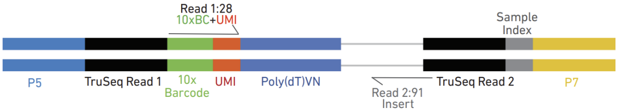


Barcoded, full-length cDNA is pooled then PCR amplified

Images from 10X Genomics CG000204_ChromiumNextGEMSingleCell3_v3.1_Rev_D.pdf

THE UNIVERSITY of ADELAIDE

10X Chromium Protocol



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

THE UNIVERSITY
of ADELAIDE

Other Variations

CITE-Seq[3]

- 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

[3] Cellular Indexing of Transcriptomes and Epitopes by sequencing

Other Variations

SPLIT-Seq[4]

- 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

[4] Split-Pool Ligation-based Transcriptome Sequencing

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

Data sourced from Haque et al., “A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications”

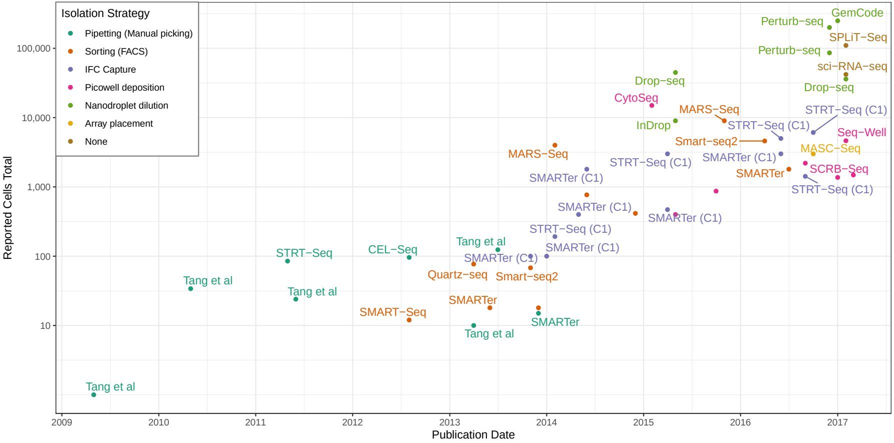
Comparison of Methods

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 5×10^5 reads/cell

Data sourced from Haque et al., "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications"

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

Data Analysis

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)

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 $\Longrightarrow$ 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*

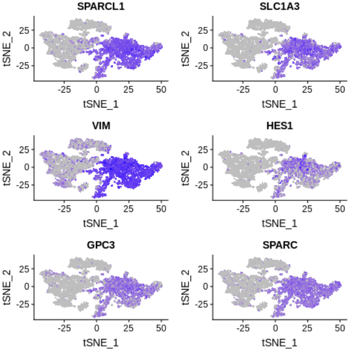
Normalisation

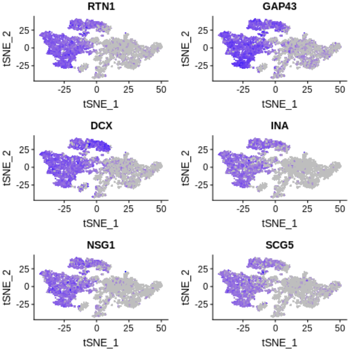
- Cell-specific offsets are once again calculated
    - Each cell is it's own source of variability
- Methods such as TMM are heavily influenced by the large numbers of zero counts
- Pooling and deconvolution:
    1. Perform rudimentary clustering of cells
    2. Normalise across all clusters (TMM assumes 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 ⟹ 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

Clustering





Images from `https://scrnaseq.readthedocs.io/en/latest/ecellranger.html`

Clustering

- Formation of clusters allows for *identification of cell-types*
- Is there a "ground truth"?
- Different approaches will provide different results
- Different parameter settings with 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*

Visualising Clusters: tSNE

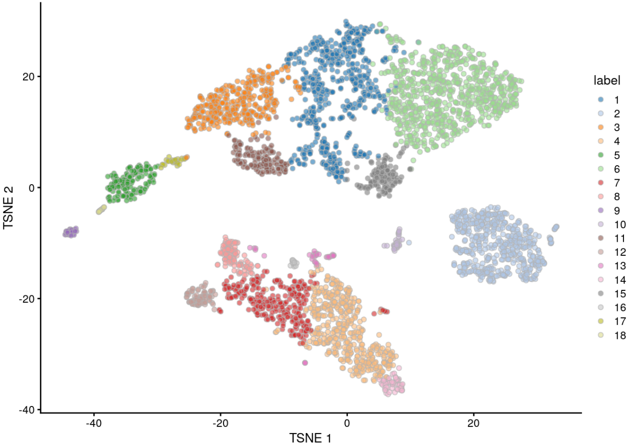


Image taken from Orchestrating Single-Cell Analysis with Bioconductor

Visualising Clusters: Force-Directed Layout

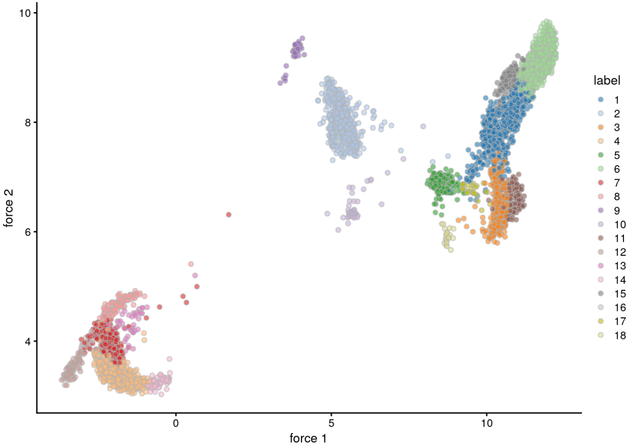


Image taken from Orchestrating Single-Cell Analysis with Bioconductor

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

Graph-Based Clustering

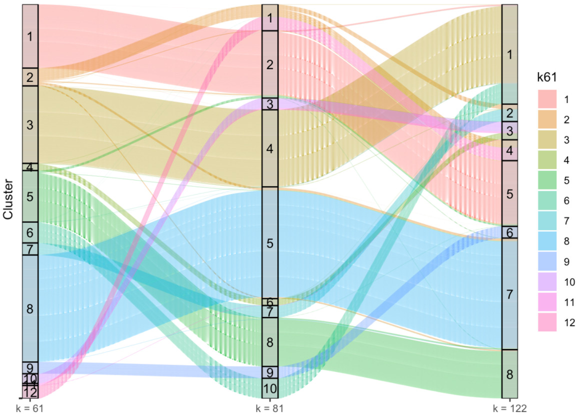
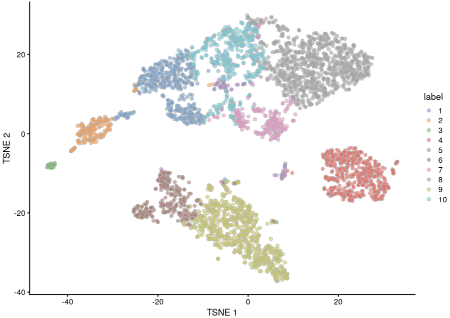


Image from Junwei Wang, Masters Thesis, 2019

Alternative Clustering Methods

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

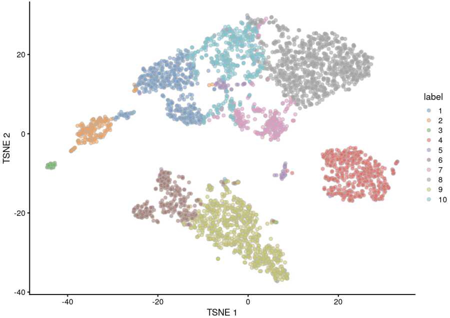
Alternative Clustering Methods



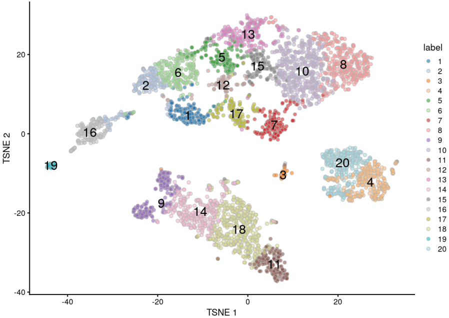
Setting $k = 10$

Images taken from Orchestrating Single-Cell Analysis with Bioconductor

Alternative Clustering Methods



Setting $k = 10$



Setting $k = 20$

Images taken from Orchestrating Single-Cell Analysis with Bioconductor

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

Marker Selection

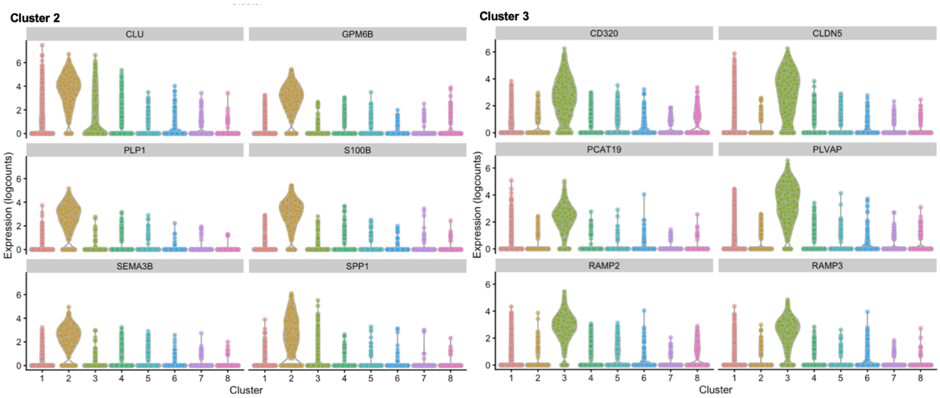


Image modified from Junwei Wang, Masters Thesis, 2019

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