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

Introduction

- scRNA-Seq is the 'latest and greatest' transcriptomic technique
- Previously all our analysis involved multiple cells per sample
- All were combined 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
- All were combined 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
- In general, determining heterogeneity of our samples is challenging

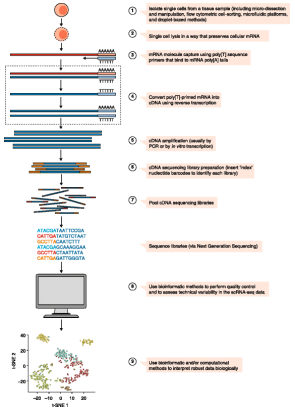
Introduction

- 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?
- 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 assigned to the same cluster/cell-type

Workflow Outline



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

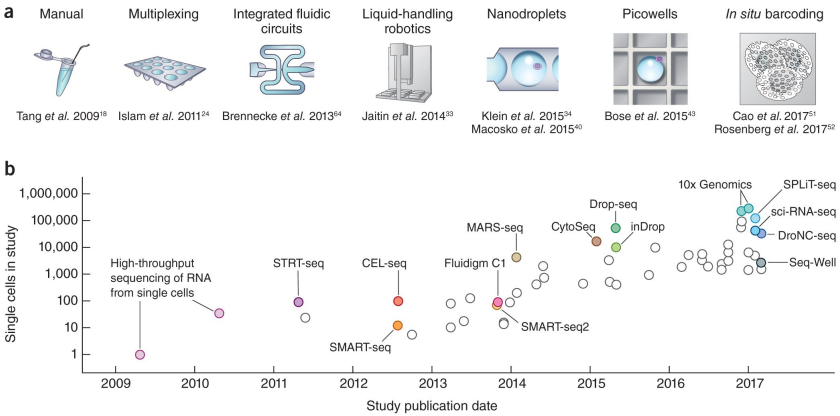
- Bulk RNA-Seq is primarily focussed on differentially expressed (DE) genes
- scRNA-Seq focusses on identifying cell-types within a sample
- How do we discriminate between different cell-types and different cell-states?
- What is the most intelligent approach for identifying DE genes
  - Is it between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

scRNA Protocols

Isolating Individual Cells

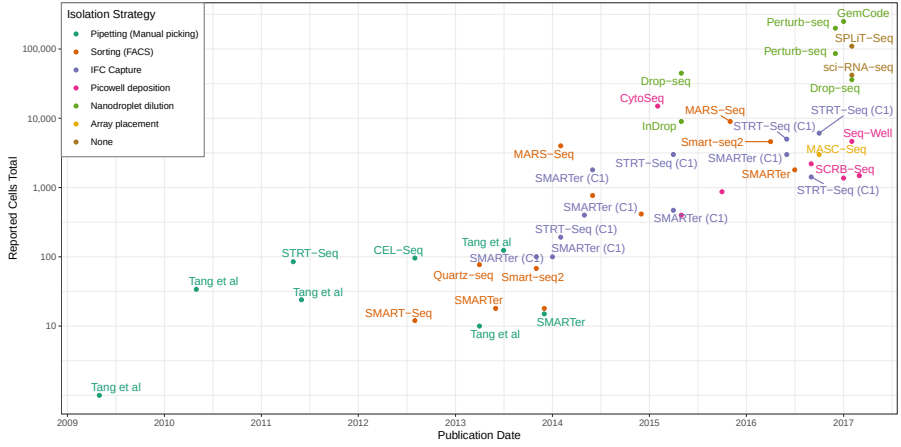
- Early protocols used a dilution series or manual isolation with a microscope (*micromanipulation*)
- Laser Capture Micro-dissection (LCM)
- Fluorescence-Activated Cell Sorting (FACS)
  - Labelled antibodies to specific surface markers
  - MACS is a magnetic-based approach
- Microfluidics/Droplet-based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

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

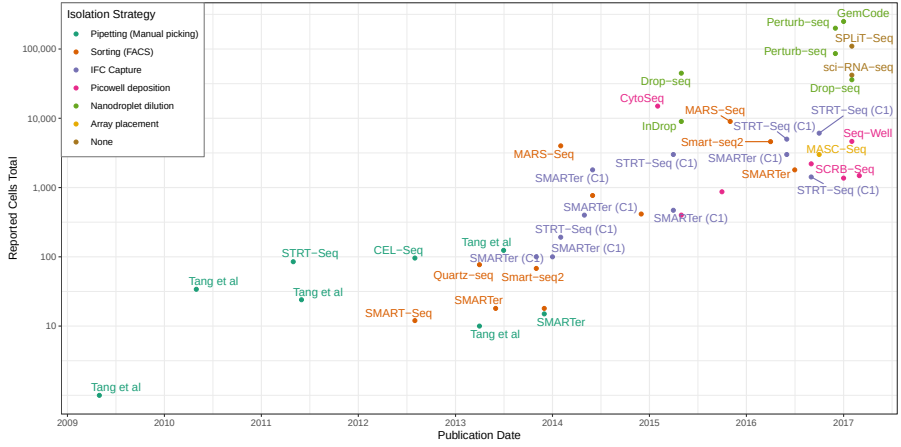


Data taken from Svensson, Vento-Tormo, and Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade"

IFC Capture

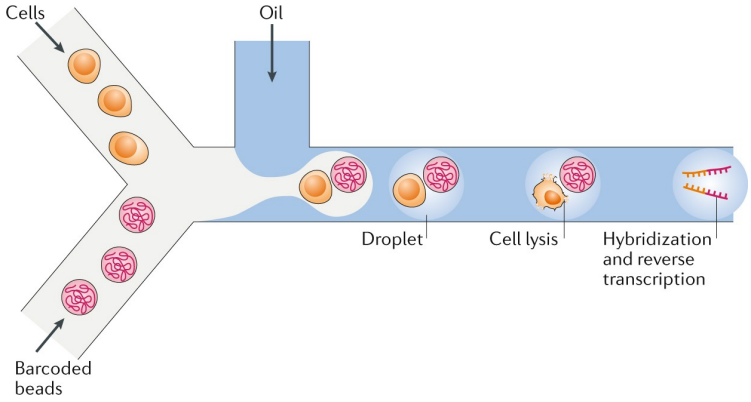
- Integrated Fluidic Circuit (IFC) chips
  - Most common is the Fluidigm C1
- Deliver tiny volumes into 'reaction chambers'
- Early chips had 96 chambers $\implies$ multiple chips / experiment
- Recent chips handle $\sim$800 cells

Protocol Timeline



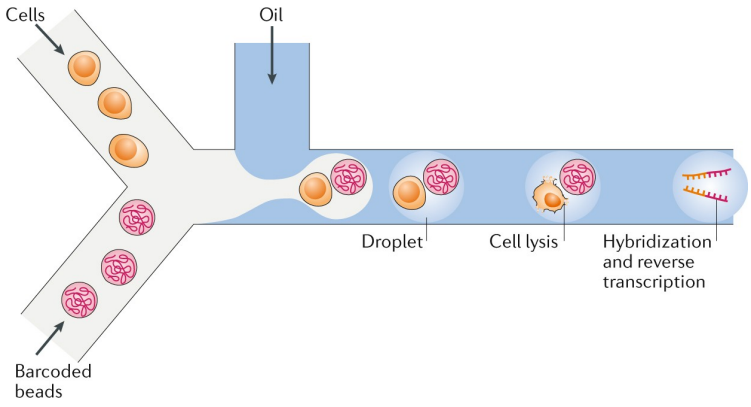
Data taken from Svensson, Vento-Tormo, and Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade"

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

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

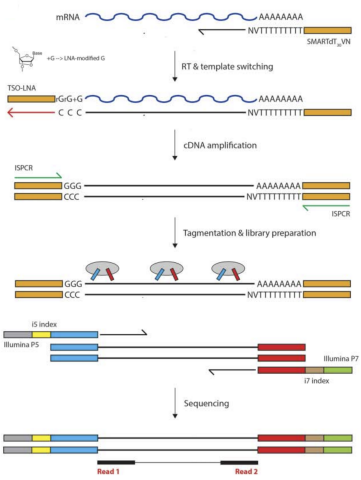


Image from S. Picelli. "Single-cell RNA-sequencing: The future of genome biology is now". In: *RNA Biol* 14.5 (May 2017), pp. 637–650

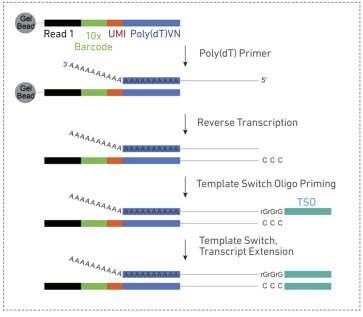
Droplet-based Methods

- 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

Inside individual GEMs

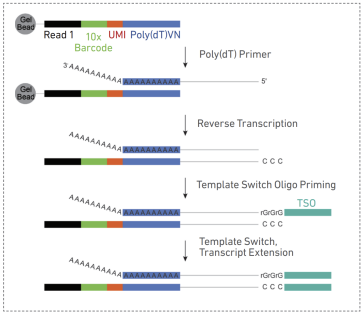


Barcoded, full-length cDNA is pooled then PCR amplified

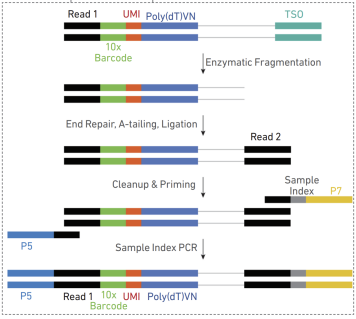
Images from 10X Genomics CG000204_ChromiumNextGEMSingleCell3_v3.1_Rev_D.pdf

10X Chromium Protocol

Inside individual GEMs



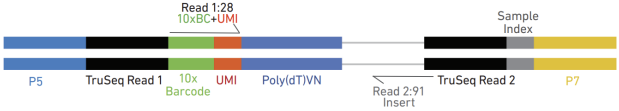
Pooled amplified cDNA processed in bulk



Barcoded, full-length cDNA is pooled then PCR amplified

Images from 10X Genomics CG000204_ChromiumNextGEMSingleCell3_v3.1_Rev_D.pdf

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

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

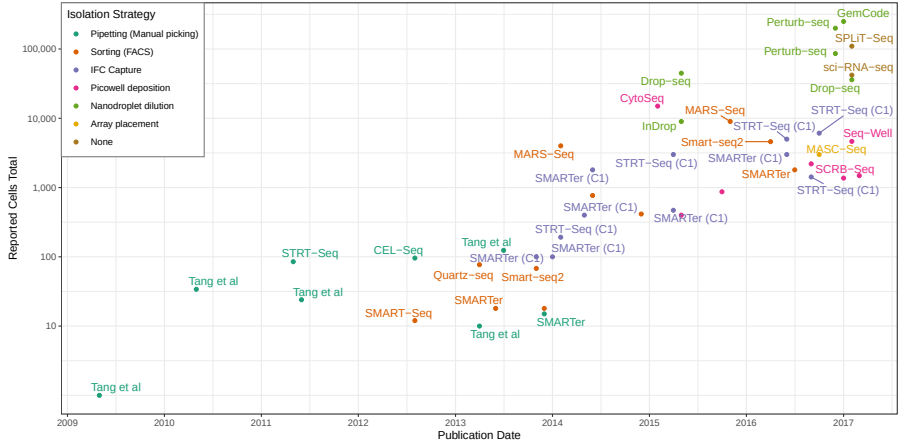
[4] Split-Pool Ligation-based Transcriptome Sequencing

Comparison of Methods

Protocol	**C1 (SMART-Seq)**	**SMART-Seq2**	**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"

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 processed 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

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)

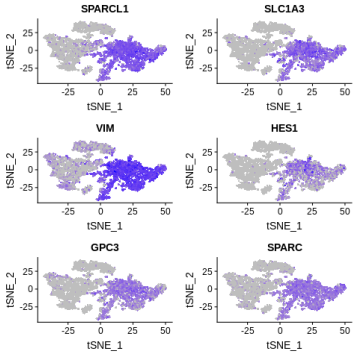
Normalisation

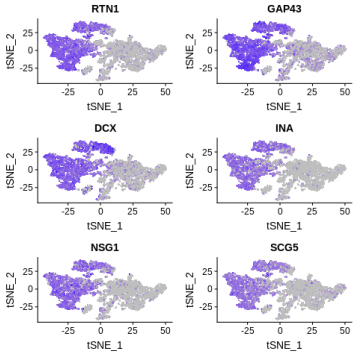
- Cell-specific offsets are once again calculated
  - Each cell is it's own source of variability
- Methods such as TMM are heavily influenced by the large numbers of zero counts
- Pooling and deconvolution:
  1. Perform rudimentary clustering of cells
  2. Normalise across all clusters (TMM assumes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

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
  - tSNE (t-Distributed Stochastic Neighbour Embedding)
  - UMAP (Uniform Manifold Approximation and Projection)

Clustering





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

Spatial Transcriptomics