Single cell biology and single cell RNA sequencing

Neo Christopher Chung, Ph.D.

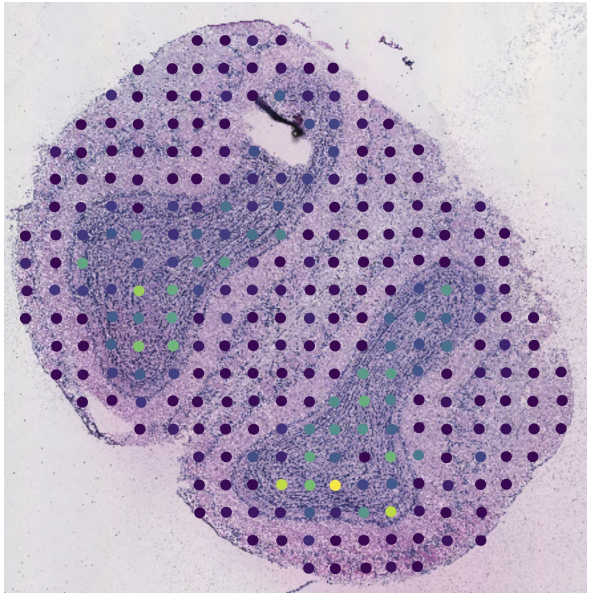
Lecture 7, 1000-719bMSB

“BULK” RNA sequencing

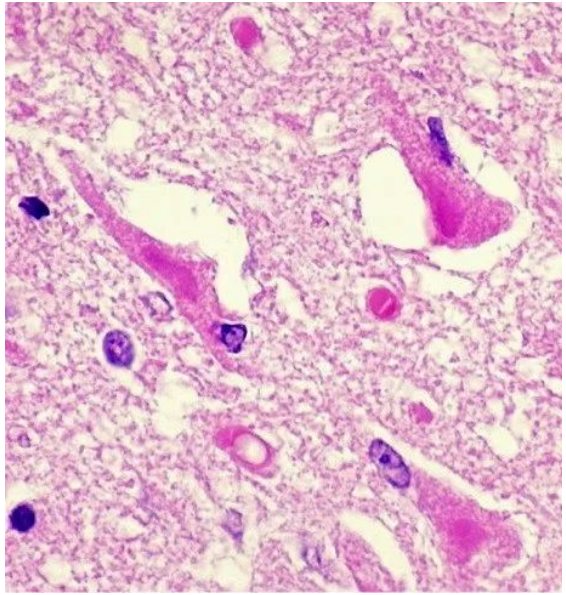
- A major breakthrough (replaced microarrays) in the late 00’s and has been widely used since
- Comparative transcriptomics, e.g. samples of the same tissue from different species
- Quantifying expression signatures from complex diseases
- **Insufficient** for studying heterogeneous systems, e.g. early development studies, complex tissues (brain)
- Does **not** provide insights into the stochastic nature of gene expression

Heterogeneity in Cells

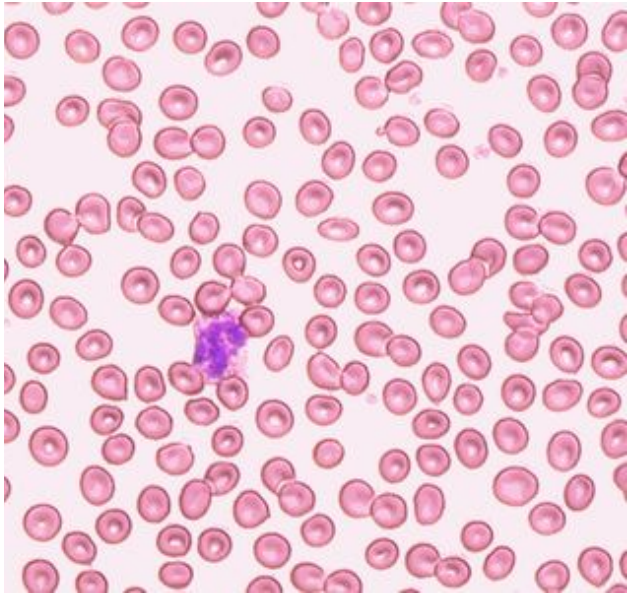
Olfactory bulb



Brain

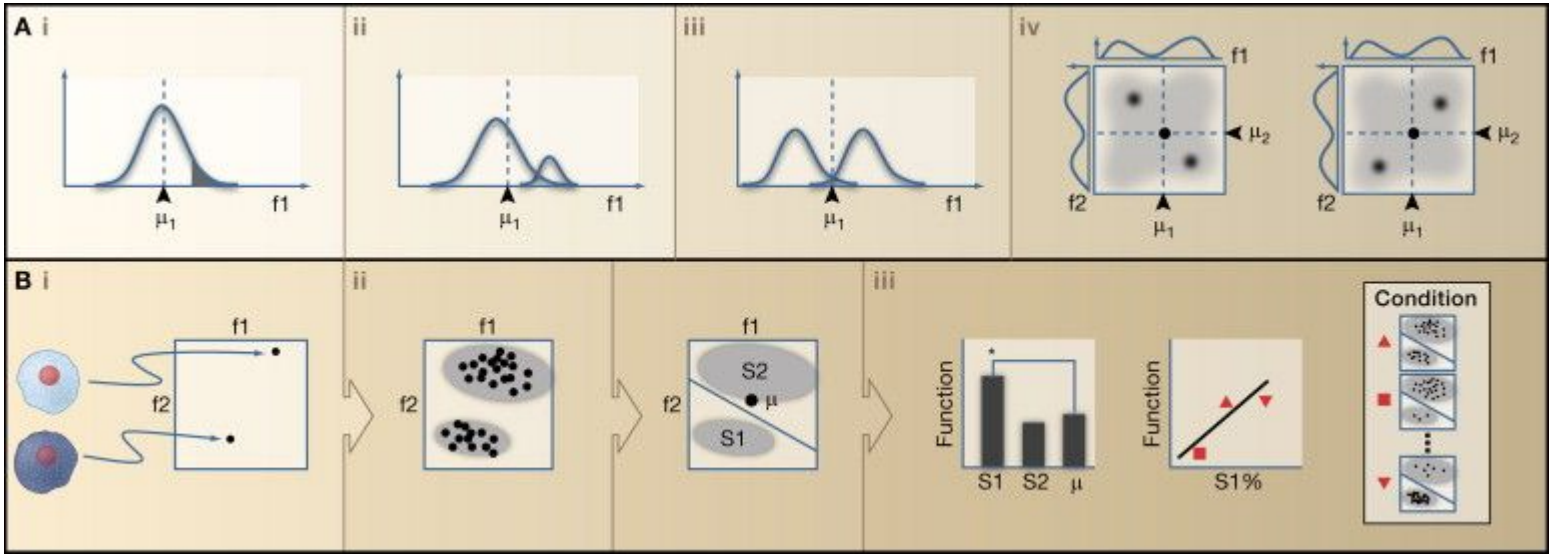


Blood



Ensemble Averages and Heterogeneity

Behaviors of cells in (i) the tail of a distribution (shaded area) or (ii) a small subpopulation (at right) may differ from the remainder of the population or from the “mean” behavior (dashed line at μ1).



(i) Single-cell measurements allow cells (left) to be represented as points in a (high-dimensional) feature space (right). (ii) Cell populations can be partitioned into distinct regions of feature space. This partition may be determined manually or automatically. Illustrated is a decomposition into two subpopulations, S1 and S2

Altschuler and Wu (2010) Cell

Towards the single cell analysis

1. Population-averaged assays are powerful tools in biology, enabling the identification of components and interactions within complex metabolic, signaling, and transcriptional networks.
2. An assumption is that ensemble averages reflect the dominant biological mechanism operating within individual cells in a population.
3. Some models derived from ensemble averages may not represent individual cell function even for a simple bell-shaped distribution of single-cell measurements.
4. Population distributions can also mask the presence of rare or small subpopulations of cells

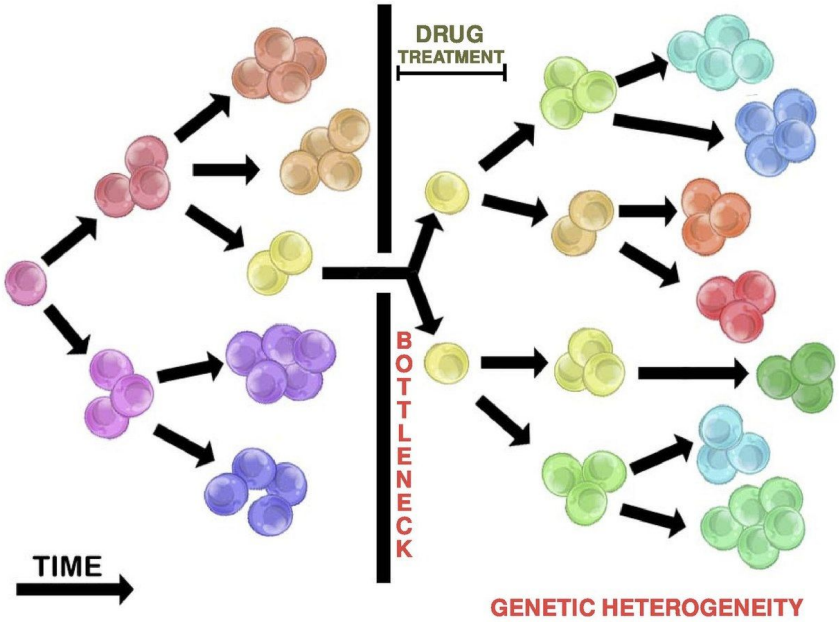
Potentials & Applications

1. Cancer and other complex multifaceted diseases are highly heterogeneous
2. Cellular subpopulations contribute unequally to disease progression or response to therapeutic intervention
3. Heterogeneity poses practical challenges for building accurate clinical models - inference and predictions
4. The dynamics and responses of single cells to drugs can vary widely and could have a large impact

Heterogeneous gene expression

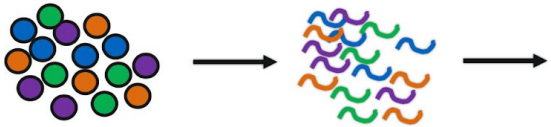
1. Exogenous factors introduce mutations, such as ultraviolet radiation (skin cancers) and tobacco (lung cancer).
2. Genomic instability such as impaired DNA repair mechanisms which can lead to increased replication errors and defects in the mitosis machinery that allow for large-scale gain or loss of entire chromosomes.
3. Mutational tumor heterogeneity refers to variations in mutation frequency in different genes and samples. The etiology of mutational processes can considerably vary between tumor samples from the same or different cancer types
4. Epigenetics impacts how our genome works without altering its sequences, via environmental, developmental, and other external factors
5. Mechanochemical heterogeneity
6. Tumour microenvironments, proliferation, metastatic potential, etc

Treatment resistance



- Drug administration in heterogeneic tumours will seldom kill all tumour cells. Upon a treatment, few drug resistant cells might survive. This allows resistant tumour populations to replicate and grow a new tumour through the branching evolution mechanism
- Due to the genetic differences within and between tumours, biomarkers that may predict treatment response or prognosis may not be widely applicable.
- Current model systems typically lack the heterogeneity seen in human cancers.

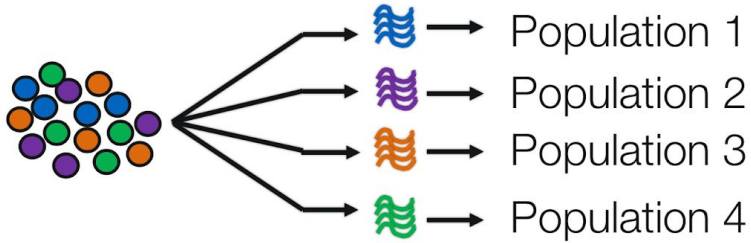
Bulk RNA-seq



average expression level

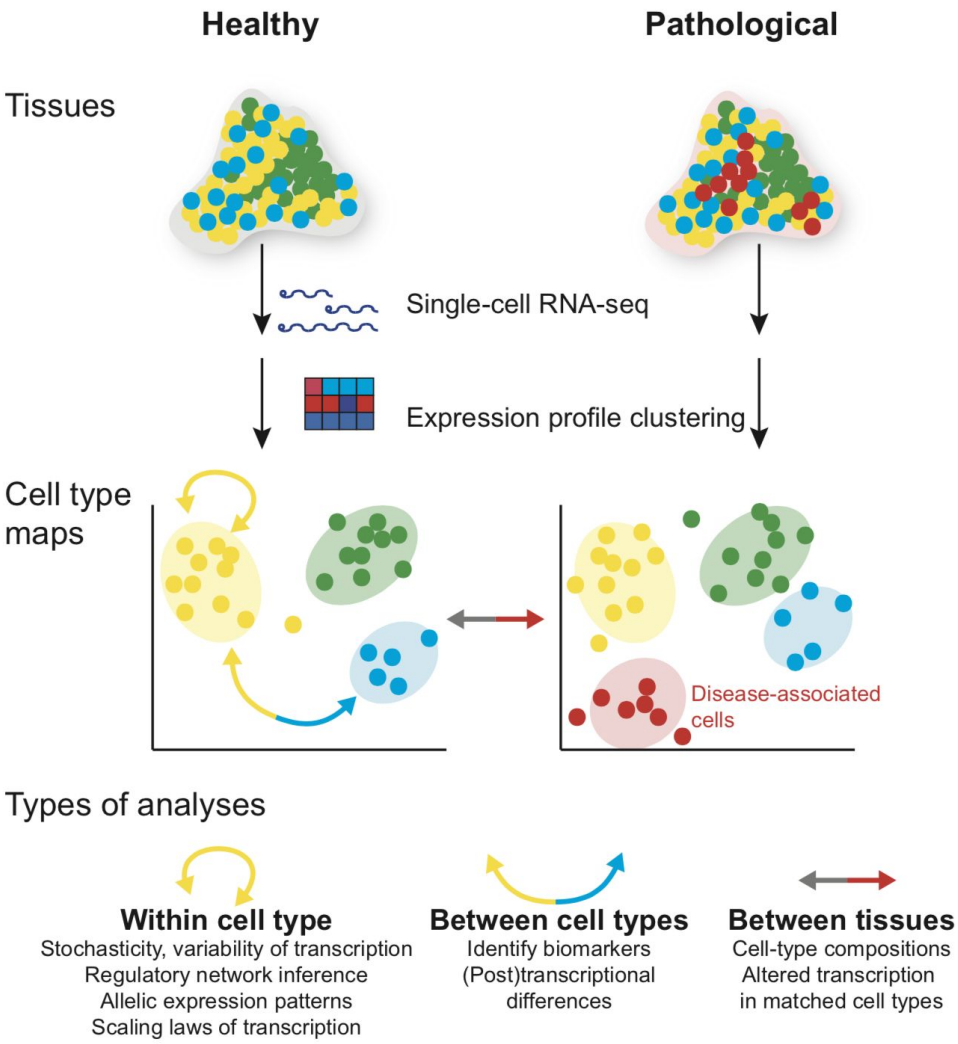
- comparative transcriptomics
- disease biomarker
- homogenous systems

scRNA-seq



- define heterogeneity
- identify rare cell population
- cell population dynamics

https://singlecellcore.hms.harvard.edu/



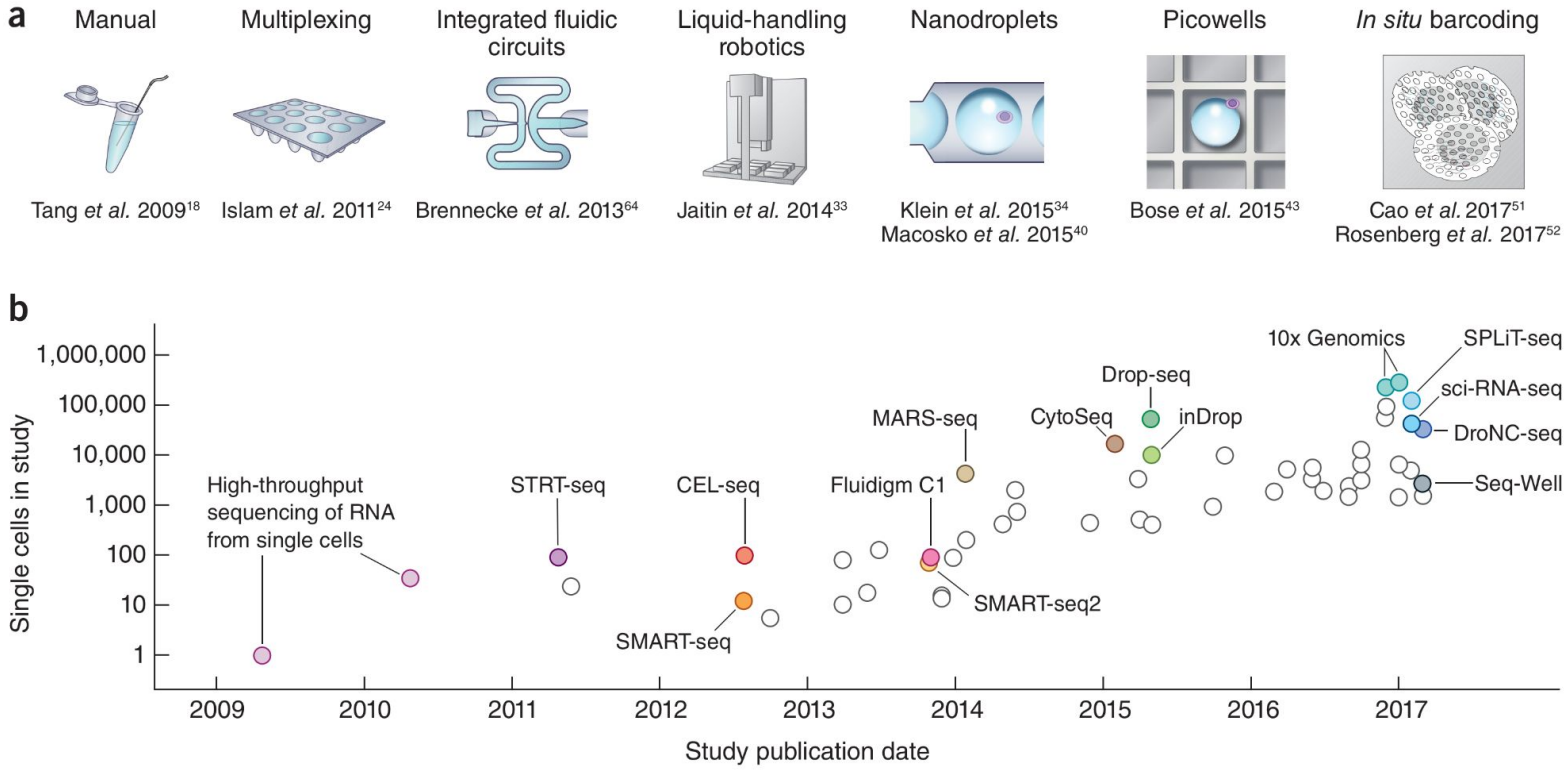
Sandberg, Nature Methods 2014

Single Cell RNA sequencing

- Tang et al. 2009 introduced a **single cell** technology
- Gained widespread popularity ~ 2014 with new protocols and lower sequencing costs
- Measures the **distribution of expression levels** for each gene across a population of cells
- Allows to study new biological questions in which **cell-specific changes in transcriptome are important**, e.g. cell type identification, heterogeneity of cell responses, stochasticity of gene expression, inference of gene regulatory networks across the cells.
- Currently there are several different protocols in use, e.g. SMART-seq2 (Picelli et al. 2013), CELL-seq (Hashimshony et al. 2012) and Drop-seq (Macosko et al. 2015)
- There are also commercial platforms available, including the Fluidigm C1, Wafergen ICELL8 and the 10X Genomics Chromium
- We will focus on drop-let based technology, namely **Drop-seq** and **10X GemCode/Chromium**

Adapted from the course on Analysis of single cell RNA-seq data (Cambridge/Sanger)

ScRNA-seq development

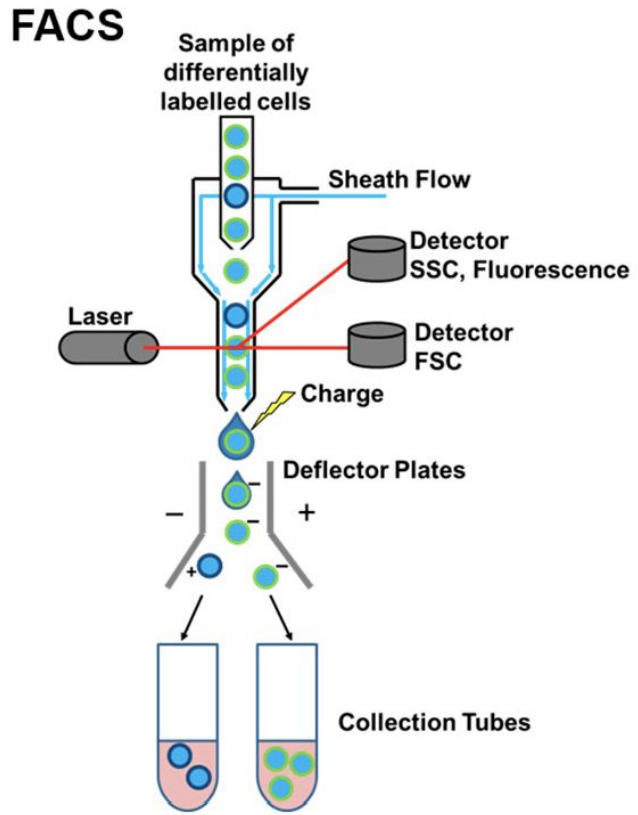


Svensson et al (2018)

SMART-Seq

Most common after a publication by Picelli et al. 2014.

Using flow cytometer (FC)



SMART-Seq

Most common after publication by Picelli et al. 2014.

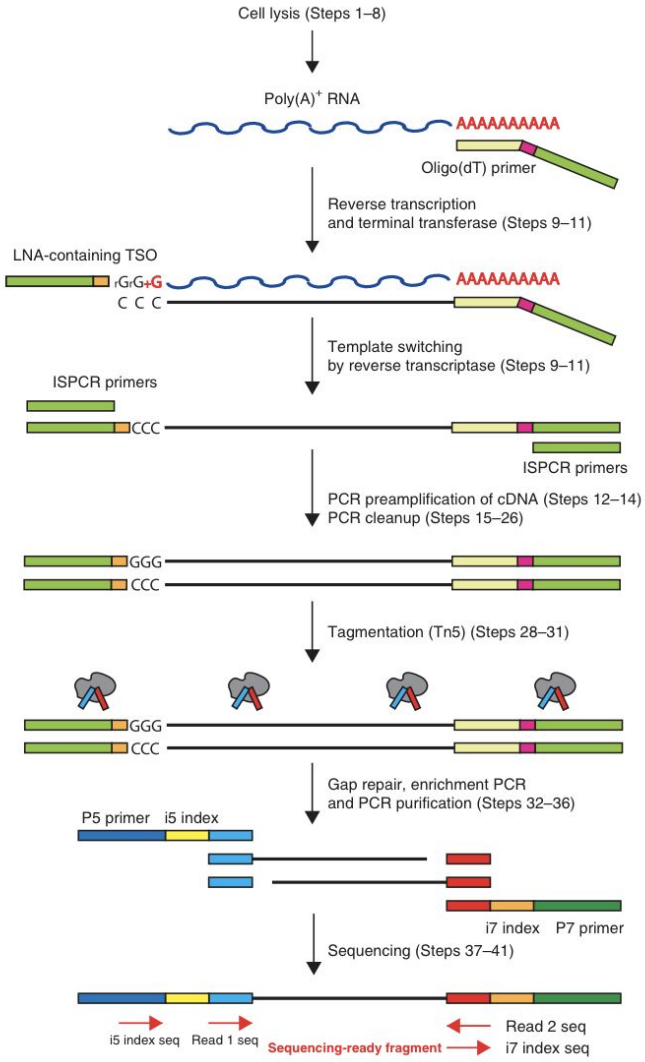
Using flow cytometer (FC)

Sort your cells into a plate -- 96 or 384 wells

Lyses individual wells

Starting with mRNA with an oligo DT

Ending with cDNA copies with PCR ends



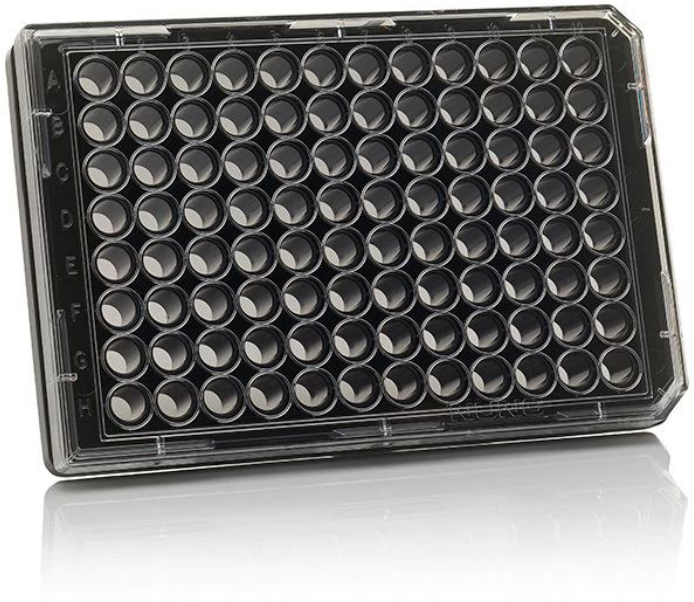
SMART-Seq: Disadvantages

Requires manual pipetting

Few hundreds to thousands of cells

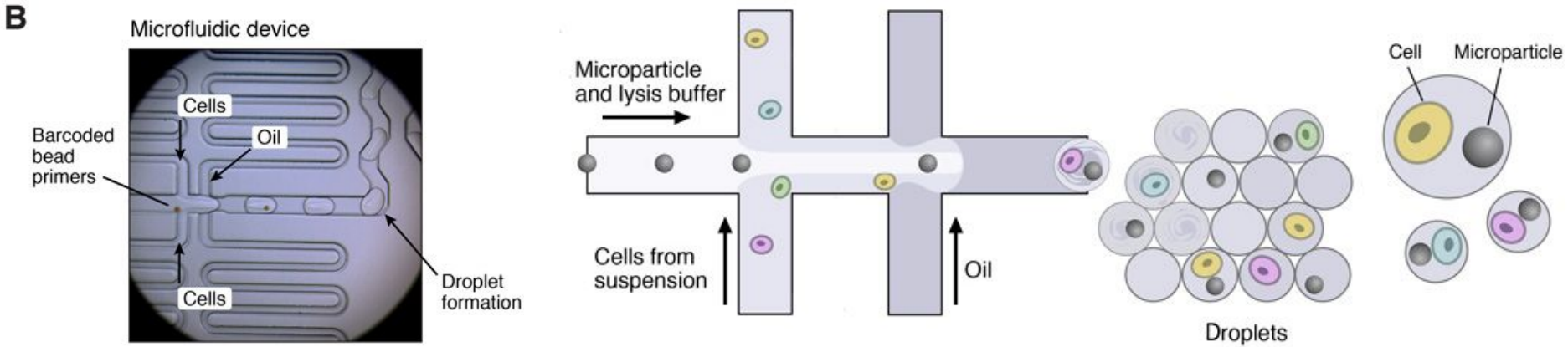
Takes time and money

Very low error rates



DropSeq

(1) prepare a single-cell suspension from a tissue

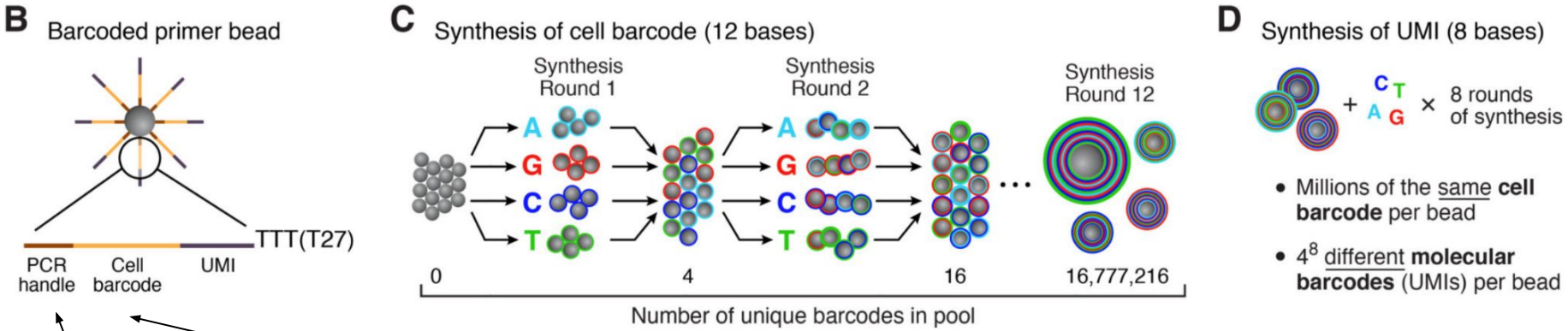


Macosko et al 2015

DropSeq

(1) prepare a single-cell suspension from a tissue

(2) co-encapsulate each cell with a distinctly barcoded microparticle (bead) in a nanoliter-scale droplet



Common sequence

Cell specific

Macosko et al 2015

DropSeq

(1) prepare a single-cell suspension from a tissue

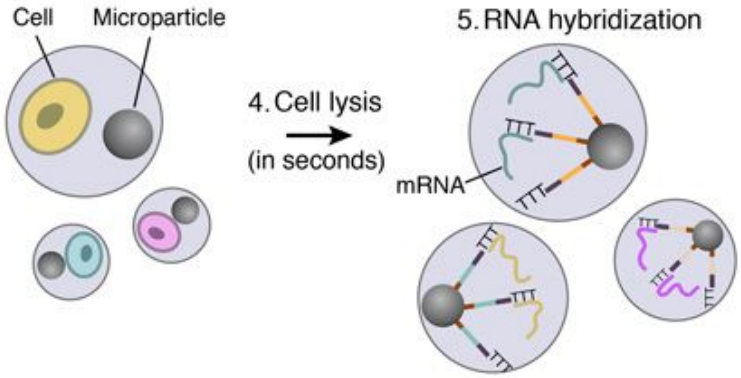
(2) co-encapsulate each cell with a distinctly barcoded (bead) in a nanoliter-scale droplet

(3) lyse cells after they have been isolated in droplets

(4) capture a cell's mRNAs on its companion microparticle, forming STAMPs (Single-cell Transcriptomes Attached to Microparticles)

(5) reverse-transcribe, amplify, and sequence thousands of STAMPs

(6) use the STAMP barcodes to infer each transcript's cell of origin

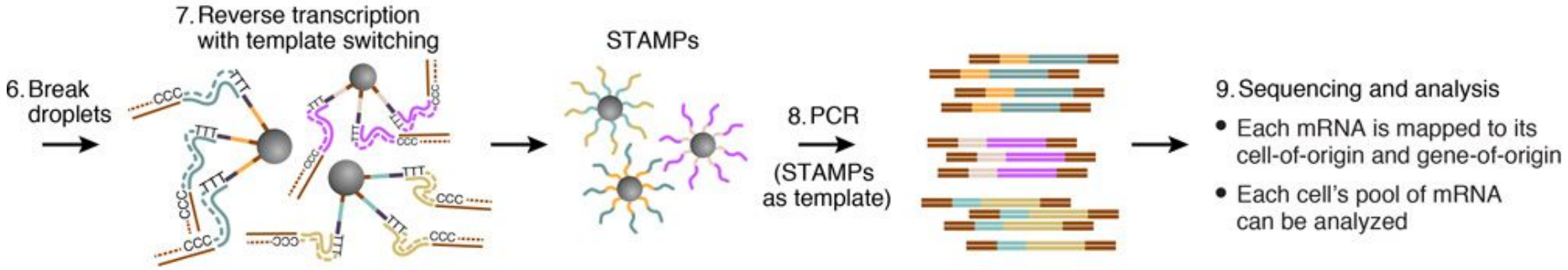


Macosko et al 2015

DropSeq

Each bead has many barcoded Oligo-DT -- unique barcodes can be identified

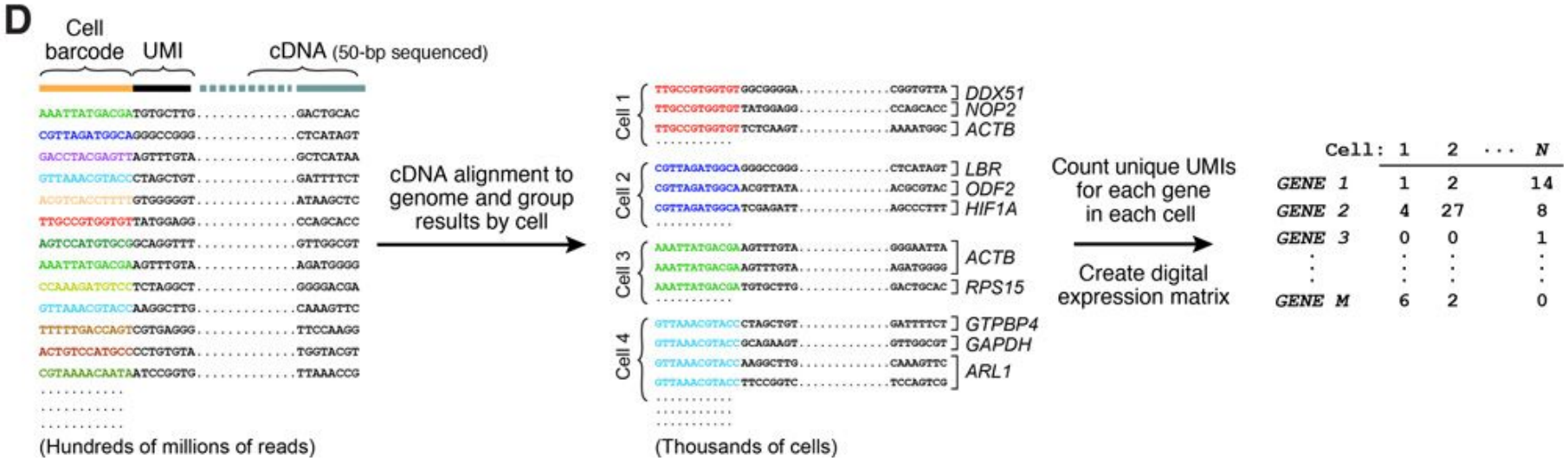
Pulled down those special beads for PCR amplification and sequencing



DropSeq

Each bead has many barcoded Oligo-DT -- unique barcodes can be identified

Pulled down those special beads for PCR amplification and sequencing



Retinal single cell data analysis

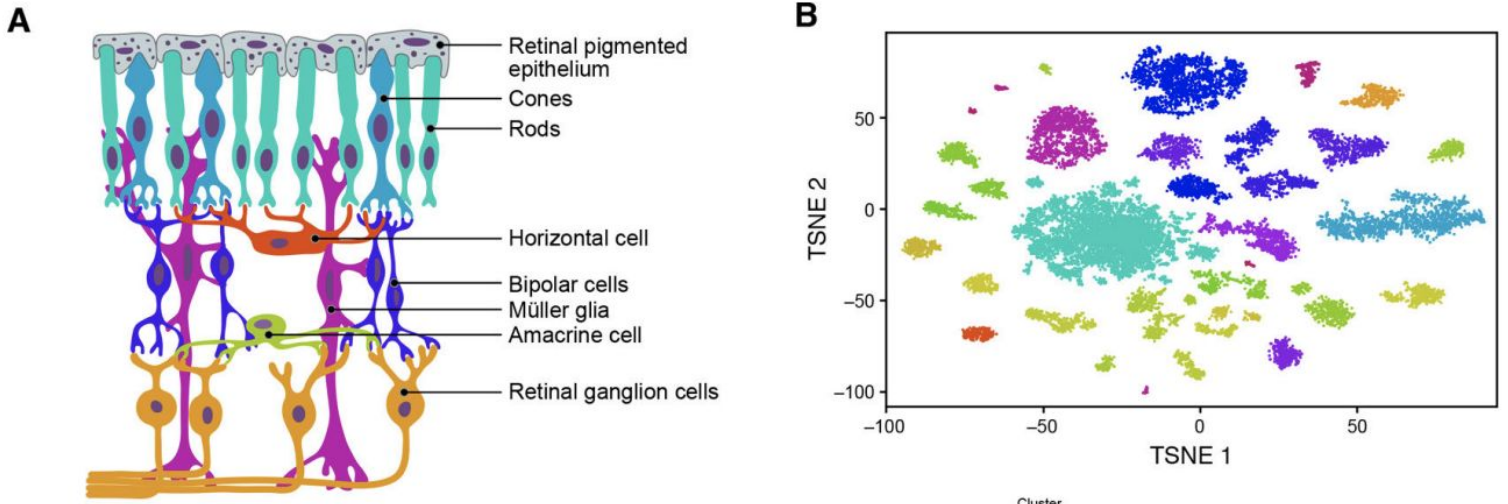
1. PCA on 13,155 cells (training dataset) with >900 genes
2. Find 32 significant PCs through the jackstraw analysis
3. Project each cell onto significant PCs
4. Apply t-Distributed Stochastic Neighbor Embedding (t-SNE)
5. Project remaining 36,145 cells (< 900 genes detected) on 2D t-SNE
6. Density clustering to identified clusters
7. Differential expression testing to confirm that clusters
8. Hierarchical clustering for relating the clusters

Retinal cell landscape

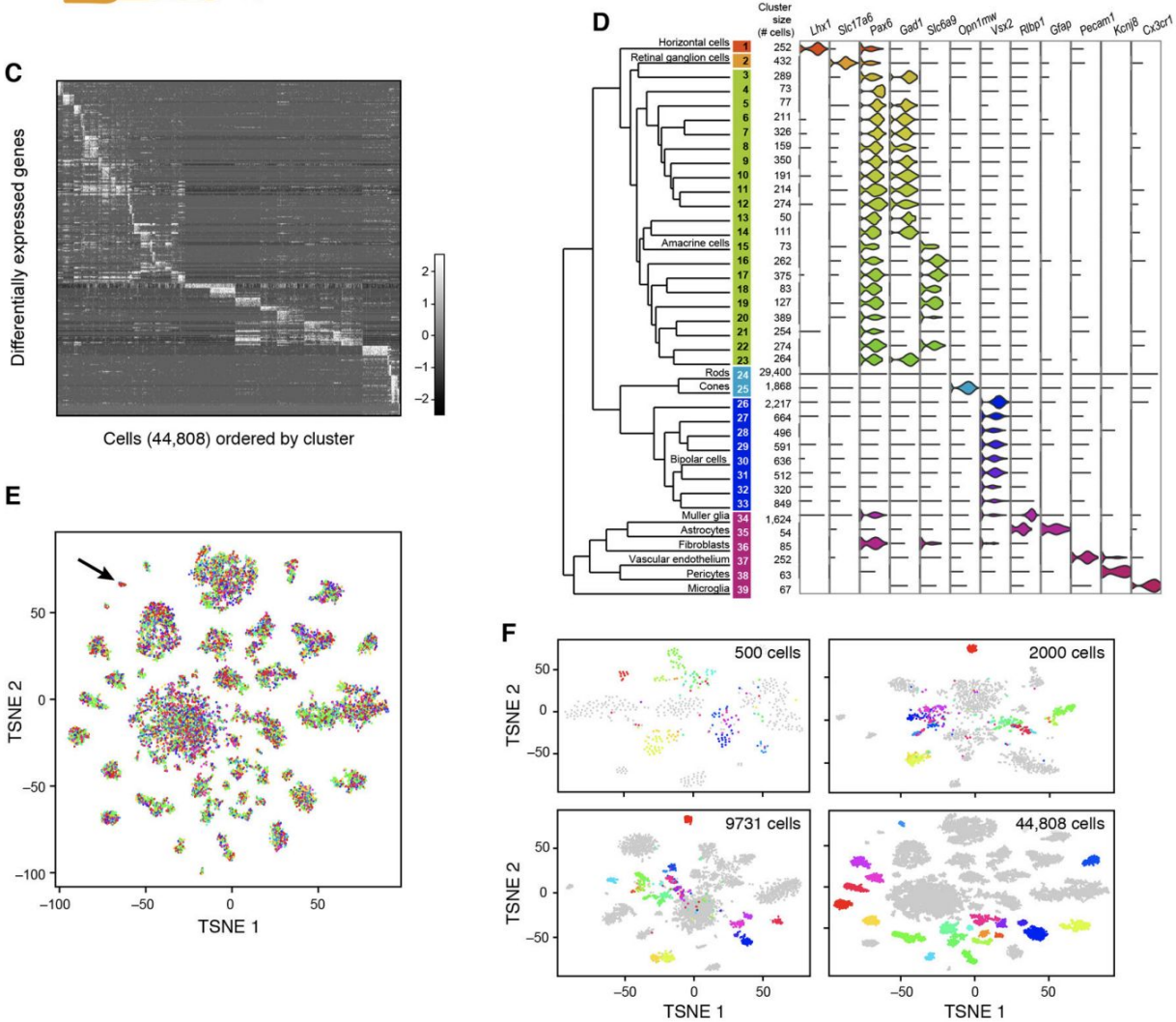
Analyzed transcriptomes from 44,808 mouse retinal cells

Found 39 groups

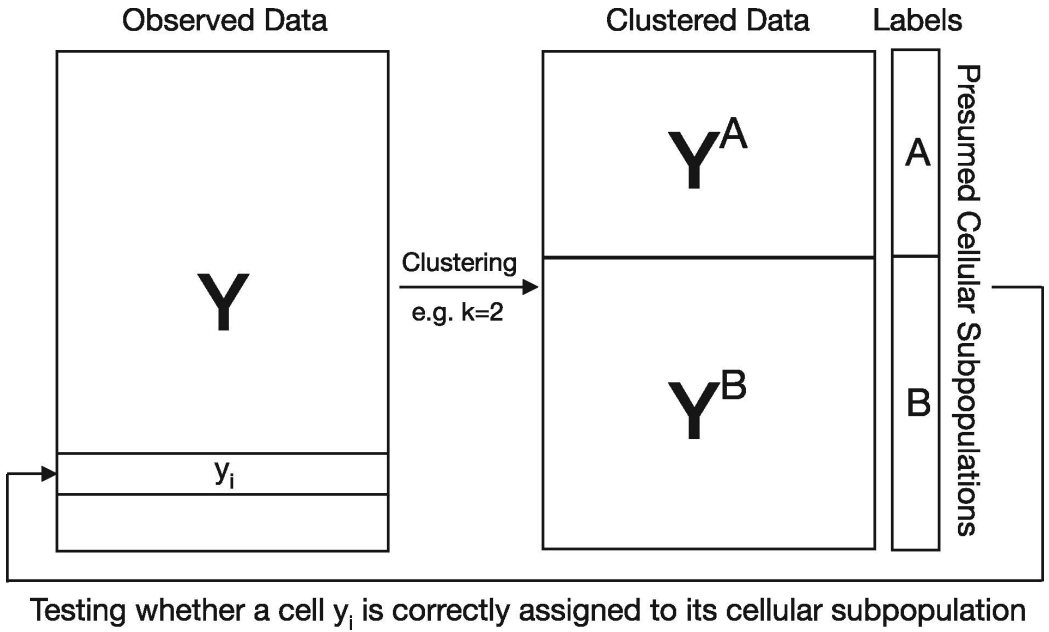
Get known retinal cell classes and novel candidate cell subtypes.



Retinal cell landscape



Circular analysis



In statistics, circular analysis is the selection of the details of a data analysis using the data that is being analysed. It is often referred to as double dipping, as one uses the same data twice. Circular analysis unjustifiably inflates the apparent statistical strength of any results reported and, at the most extreme, can lead to the apparently significant result being found in data that consists only of noise. (Wikipedia

DropSeq: Disadvantages

The biggest issue = Most of beads are empty

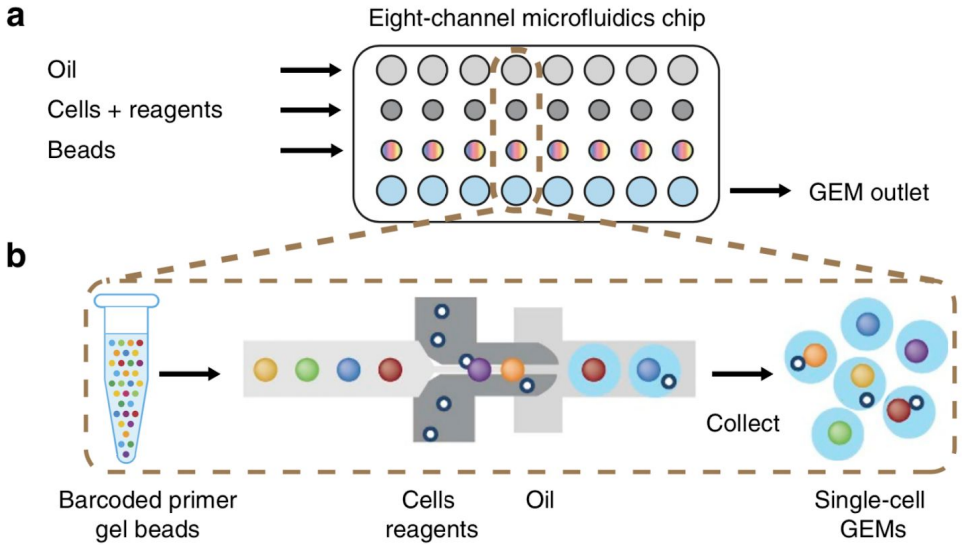
Cells and beads are randomly going through the chip

Some droplets will have more than one cell (doublets or multiplets)

→ thus, an intricate balance between putting a low amount of cells (lowering a chance of multiplets) vs. putting a sufficient amount of cells (getting more beads to have cells)

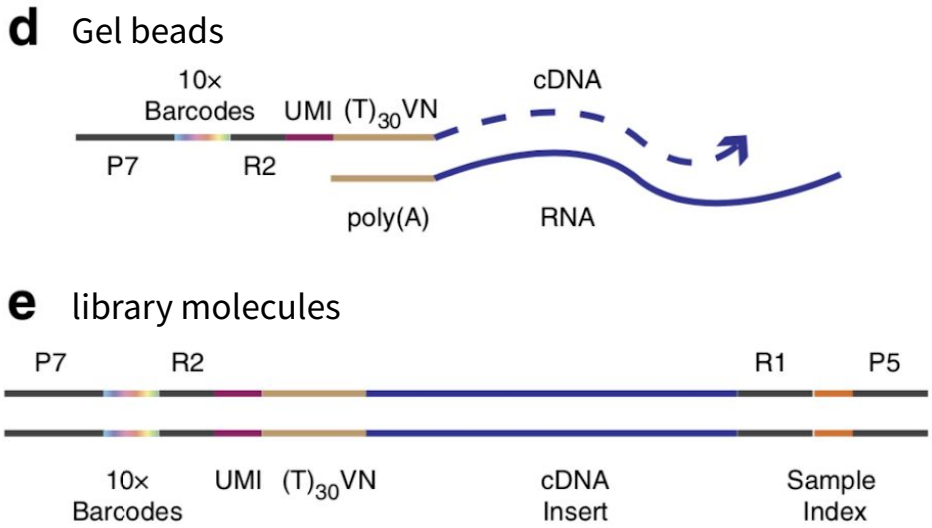
10X Genomics

- Similar to DropSeq
- But microfluidic devices and beads have high fill rate (~80%)
- Each gel bead is functionalized with barcoded oligonucleotides that consists of:
  - (i) sequencing adapters and primers
  - (ii) a 14 bp barcode drawn from ~750,000 designed sequences to index GEMs
  - (iii) a 10 bp randomer to index molecules (unique molecular identifier, UMI) and
  - (iv) an anchored 30 bp oligo-dT to prime polyadenylated RNA transcripts



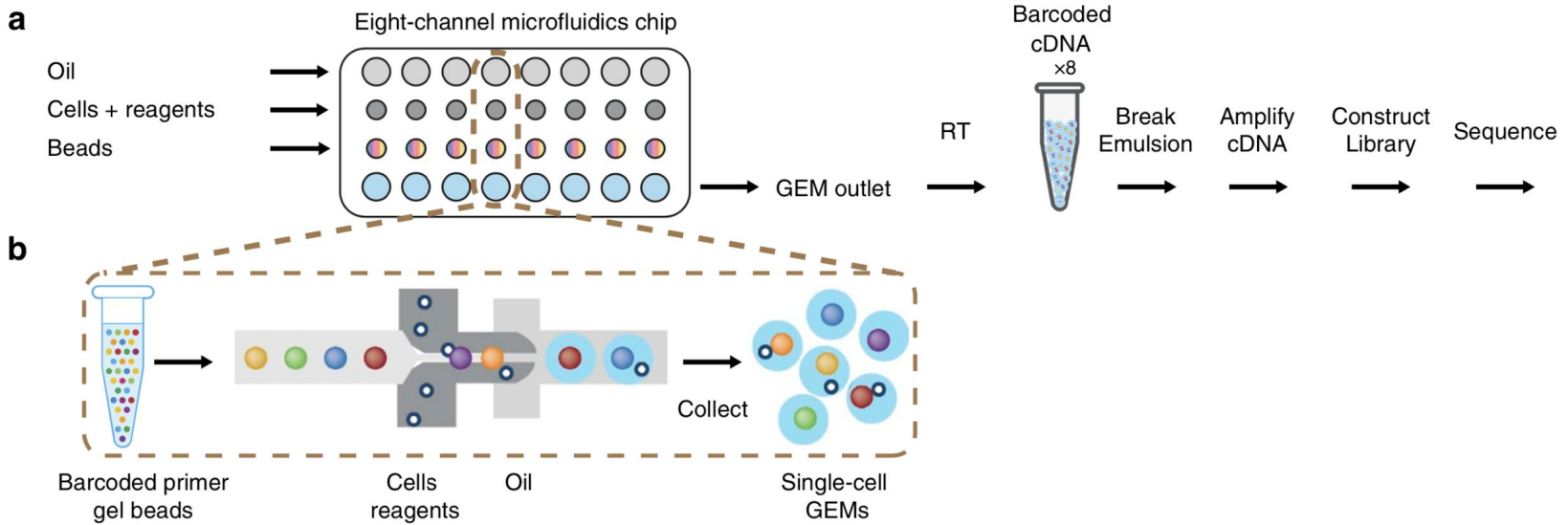
Zheng et al (2017) Nature Comm

cDNA library construction



- Each **resulting cDNA molecule** contains a UMI and shared barcode per GEM, and ends with a template switching oligo at the 3′ end.
- Barcodes and sample indices allow pooling and sequencing of multiple libraries on a next-generation short read sequencer.
- Parallel capture of thousands of cells in each of the 8 channels for scRNA-seq analysis.

10X Genomics

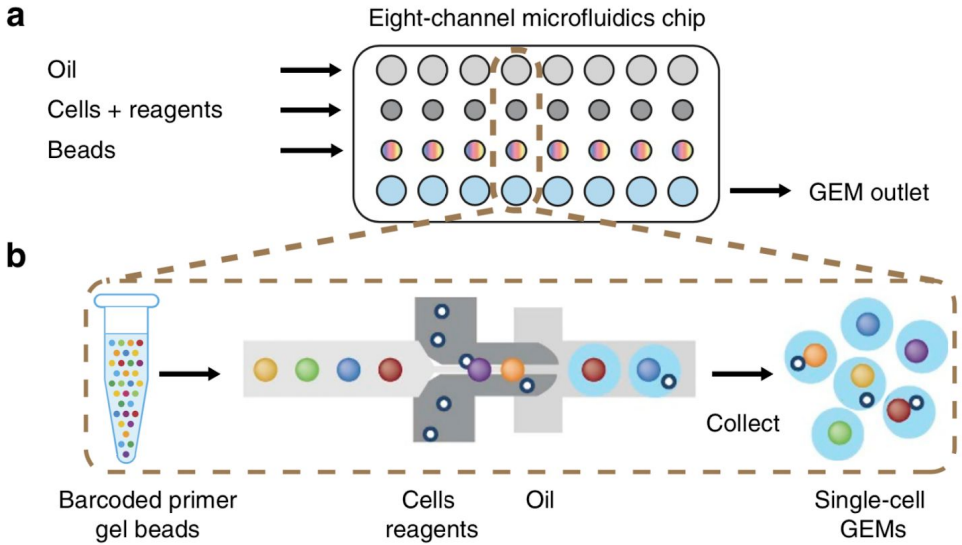


- Cell lysis begins immediately after encapsulation. Gel beads dissolve and release their oligonucleotides for reverse transcription of polyadenylated RNAs.
- The emulsion is broken and barcoded cDNA is pooled for PCR amplification, using primers complementary to the switch oligos and sequencing adapters.
- Amplified cDNAs are sheared, and adapter and sample indices are incorporated into finished libraries, which are compatible with next-generation short-read sequencing.

Zheng et al (2017) Nature Comm

>200,000 single cells in one day

- Within each microfluidic channel, ~100,000 GEMs (but more commonly <50,000) are formed per ~6-min run, encapsulating thousands of cells in GEMs.
- Amplification of cDNA and library construction takes a half day



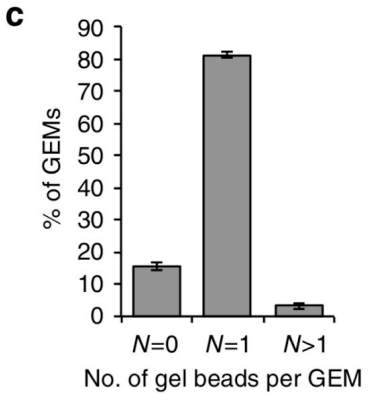
Zheng et al (2017) Nature Comm

Overall experimental steps in scRNA-seq

1. **Single-cell suspension** is generated in a process called **single-cell dissociation** in which the tissue is digested
2. **Single-cell isolation** is performed differently depending on the experimental protocol.
    a. plate-based techniques isolate cells into wells on a plate
    b. droplet-based methods rely on capturing each cell in its own microfluidic droplet
3. Each well or droplet contains the necessary chemicals to break down the cell membranes and perform **library construction**. Library construction is the process in which the intracellular mRNA is captured, reverse-transcribed to cDNA molecules and amplified.
4. mRNAs from each cell can be labelled with **cellular barcodes**. Furthermore, many experimental protocols also label captured molecules with a **unique molecular identifier** *(UMI)*. Cellular cDNA is amplified before sequencing to increase its probability of being measured. UMIs allow us to distinguish between amplified copies of the same mRNA molecule and reads from separate mRNA molecules transcribed from the same gene.
5. These libraries are pooled together (***multiplexed***) for *sequencing*. Sequencing produces read data, which undergo quality control, grouping based on their assigned barcodes (***demultiplexing***) and alignment in read processing pipelines. For UMI-based protocols, read data can be further demultiplexed to produce counts of captured mRNA molecules (*count data*).

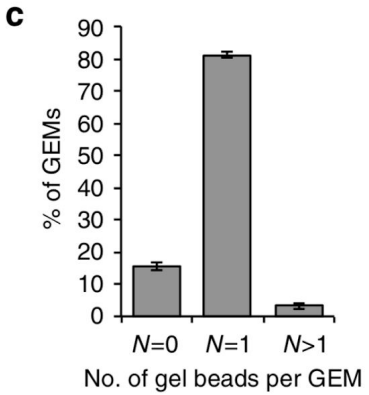
Efficiency vs. error

- The input amount (*a number of cells) into the scRNA-seq platforms dictates the efficiency and the error rate
- More cells you put, less “empty” gel beads
- More cells you put, more gel beads with more than 1 cell



Doublets and Multiplets

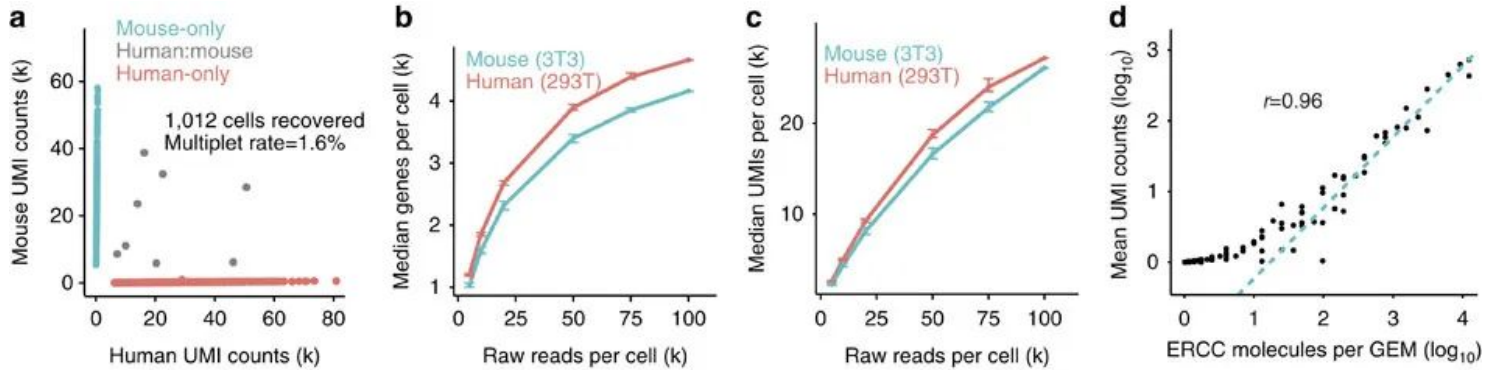
- Given that a large number of single cells are automatically captured, a single droplet may contain two or more single cells.
- Known as doublets or multiplets, they may induce biologically irrelevant gene expression profiles in scRNA-seq studies.
- Zheng et al. (2017) inferred a 3.1% multiplet rate for this mixture experiment.
- Furthermore, for ~ 10000 single cells, Zheng et al. (2017) reported > 8% multiplet rates that approximately linearly increase with the recovered cell number.
- Such contaminations by multiplets are ubiquitous in high-throughput scRNA-seq platforms (Andrews and Hemberg, 2018).



A mixture of human and mouse cells

- Loaded a mixture of ~1,200 human (293T) and ~1,200 mouse (3T3) cells
- Sequenced the library on the Illumina NextSeq 500 to yield ~100k reads per cell.
- 1,012 GEMs contained cells, of which 482 and 538 contained reads that mapped to the human and mouse transcriptome
- > 80% of UMI counts were associated with cell barcodes

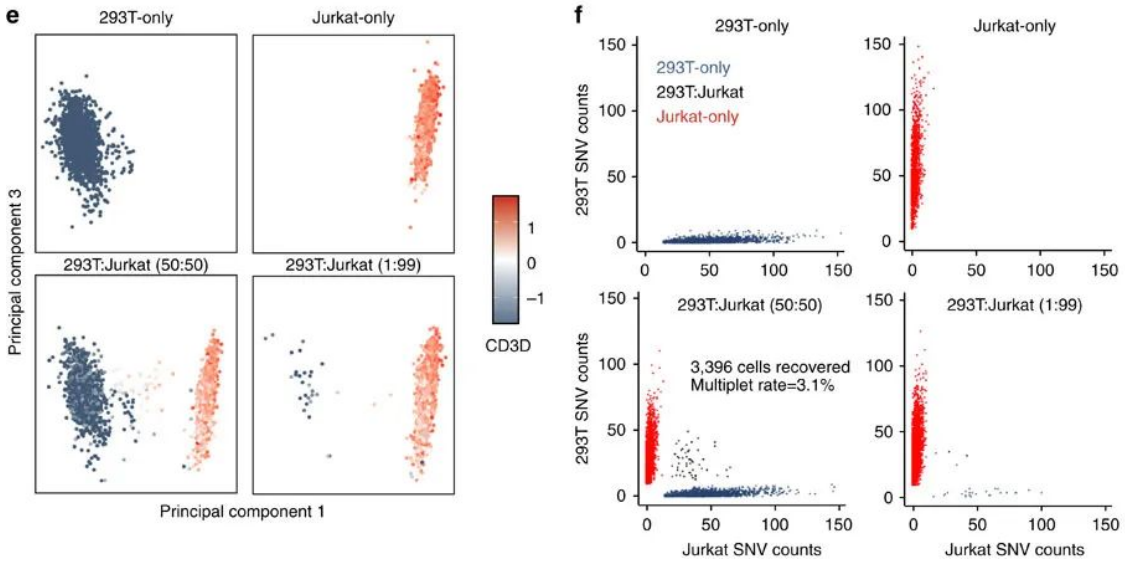
- Multiplet: 8 cell-containing GEMs had a substantial fraction of human and mouse UMI counts (see **fig. a** below)
- Cell capture rate: ~50%, the ratio of the number of cells detected by sequencing and the number of cells loaded
- Cross-talk: 0.9%, a mean fraction of UMI counts from the other species was 0.9% in both human and mouse GEMs



Zheng et al (2017) Nature Comm

A mixture of 293T and Jurkat cell lines

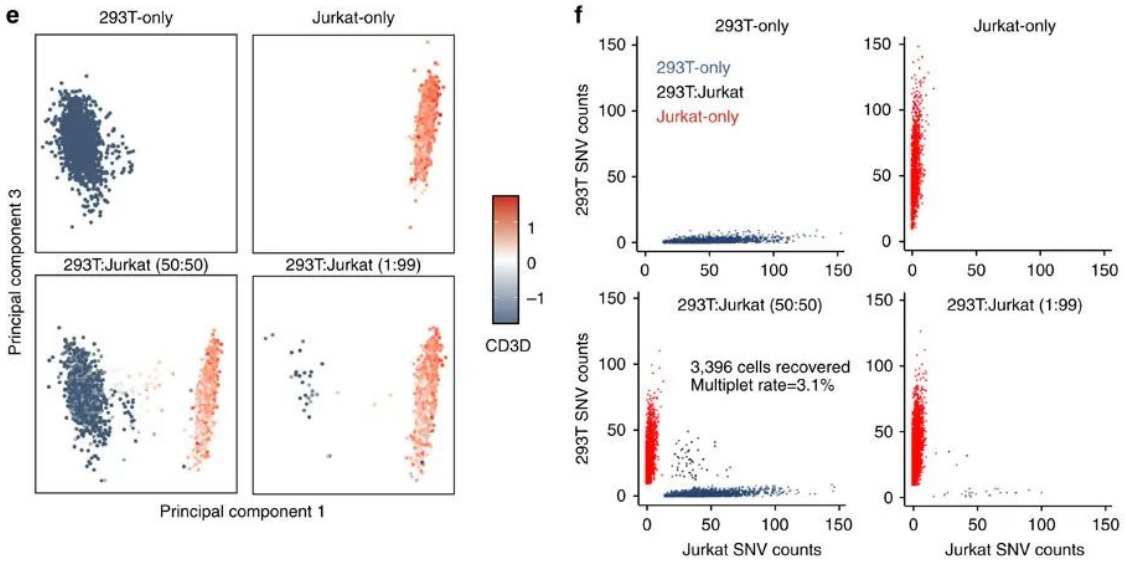
- Jurkat cells: preferentially expressing *CD3D*, a male cell line
- 293T cells: preferentially expressing *XIST*, a female cell line
- Mixed at different ratios
- Multiplet: ~3%
- On average, there are ~350 SNVs detected in each 293T or Jurkat cell
- 45% 293T cells primarily (96%) harbored 293T-enriched SNVs
- 50% Jurkat cells primarily (94%) harbored Jurkat-enriched SNVs



Zheng et al (2017) Nature Comm

A mixture of 293T and Jurkat cell lines

- Jurkat cells: preferentially expressing *CD3D*, a male cell line
- 293T cells: preferentially expressing *XIST*, a female cell line
- Mixed at different ratios
- Multiplet: ~3%
- On average, there are ~350 SNVs detected in each 293T or Jurkat cell
- 45% 293T cells primarily (96%) harbored 293T-enriched SNVs
- 50% Jurkat cells primarily (94%) harbored Jurkat-enriched SNVs



Zheng et al (2017) Nature Comm