COM SCI C121 Week 7

Aidan Jan

May 16, 2024

Single-cell RNA-seq Technologies and Fundamental Issues

We are given many samples of cells; each sample consists of some number of cells. The data would appear like:

	Sample 1	Sample 2	...	Sample N
Gene 1	20	32		301
Gene 2	100	100		100
...				
Gene p	17	10		43

What we would rather have: for each sample, we extract RNA from *each individual cell* and sequence it. This would get us individual cell results which makes it easier to compare, instead of aggregate results per sample.

Why do we want to keep cell identity?

- Cells are inherently *heterogenous*
- If we assay many cells, we can understand heterogeneity within the heterogenity.
- Given a population, we can understand what "types" of cells exist and what "state" they are in

The ideal (dissociated) single-cell transcriptomics method is:

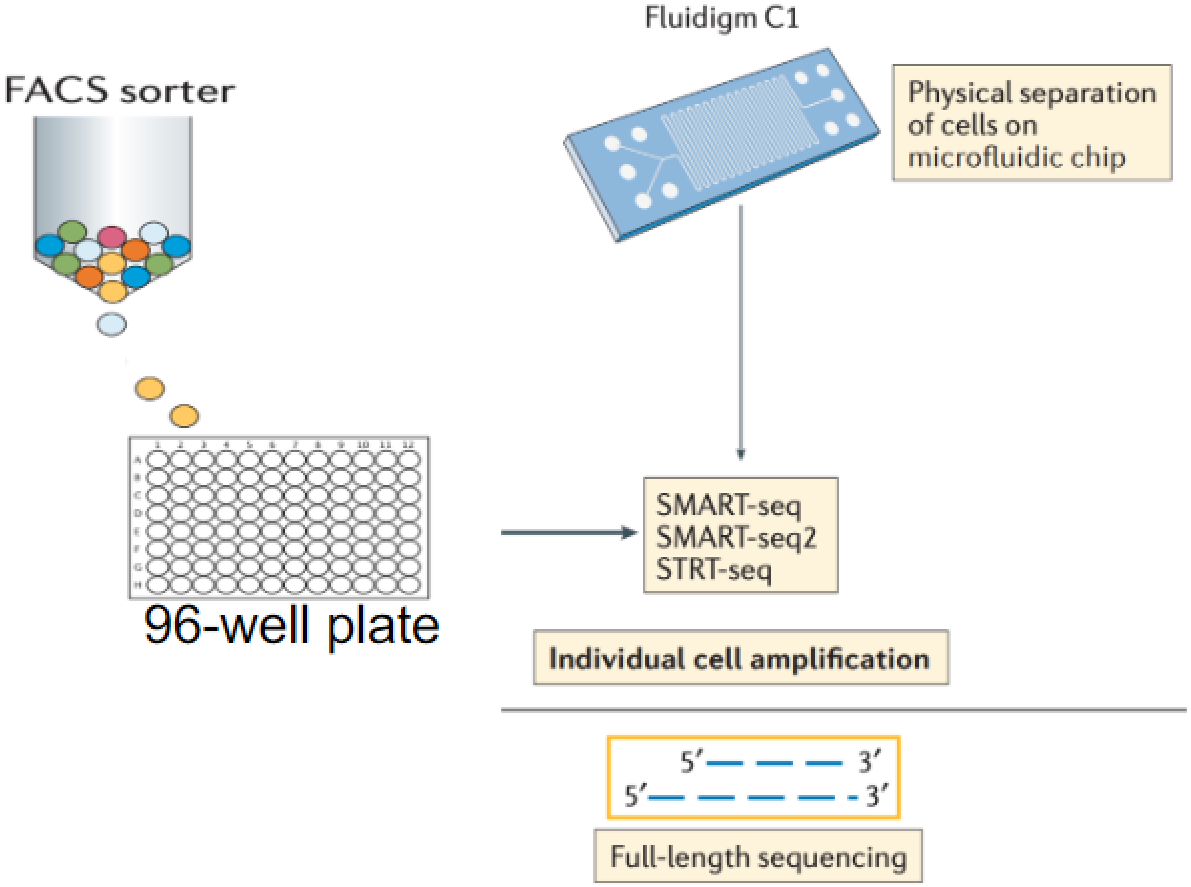
- Universal in terms of cell size, type, and state.
- In situ measurements
- No minimum input of number of cells to be assayed
- Every cell is assayed, i.e., 100% capture rate
- Every transcript in every cell is detected, i.e., 100% sensitivity.
- Every transcript is identified by its full-length sequence
- Transcripts are readily associated to single cells, e.g., no doublets
- Additional measurements of other cell attributes.
- Cost effective per cell
- Easy to use.
- Open source.

Single-cell RNA Technology Development

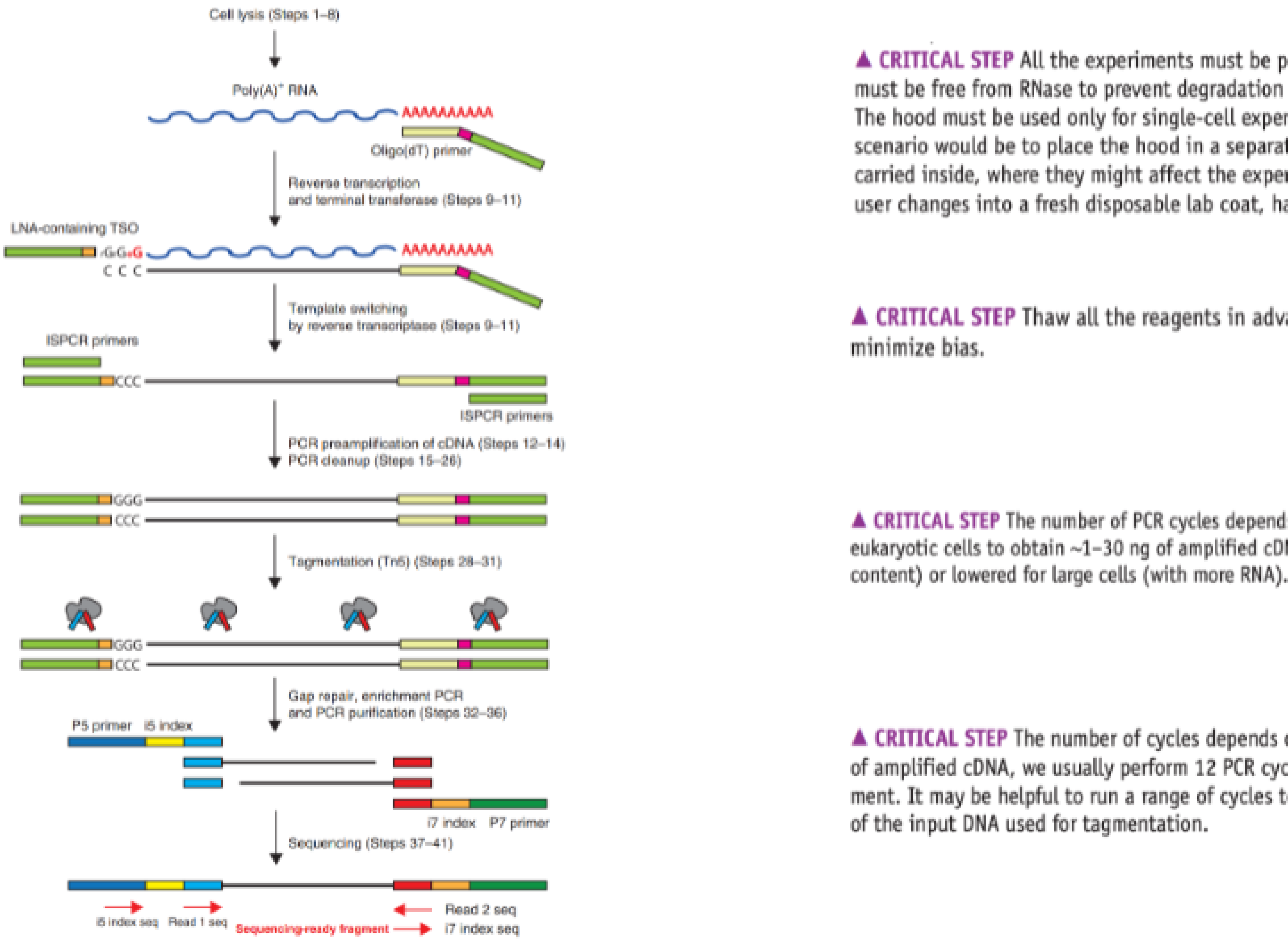
- Studies have been done since around 2013, but since then, especially in years 2018-2020, the technology improved dramatically. The number of studies per year increased from about 2 in 2013 to over 70 in 2020.
- At the same time, many new tools were developed.

Physical Separation of Cells in Wells

"First generation" single-cell sequencing is low throughput (relatively speaking). It involves manually placing cells into a 96-well plate.



Example: Library preparation for SMART-Seq2



▲ CRITICAL STEP All the experiments must be p
must be free from RNase to prevent degradation
The hood must be used only for single-cell exper
scenario would be to place the hood in a separat
carried inside, where they might affect the exper
user changes into a fresh disposable lab coat, ha

▲ CRITICAL STEP Thaw all the reagents in adv
minimize bias.

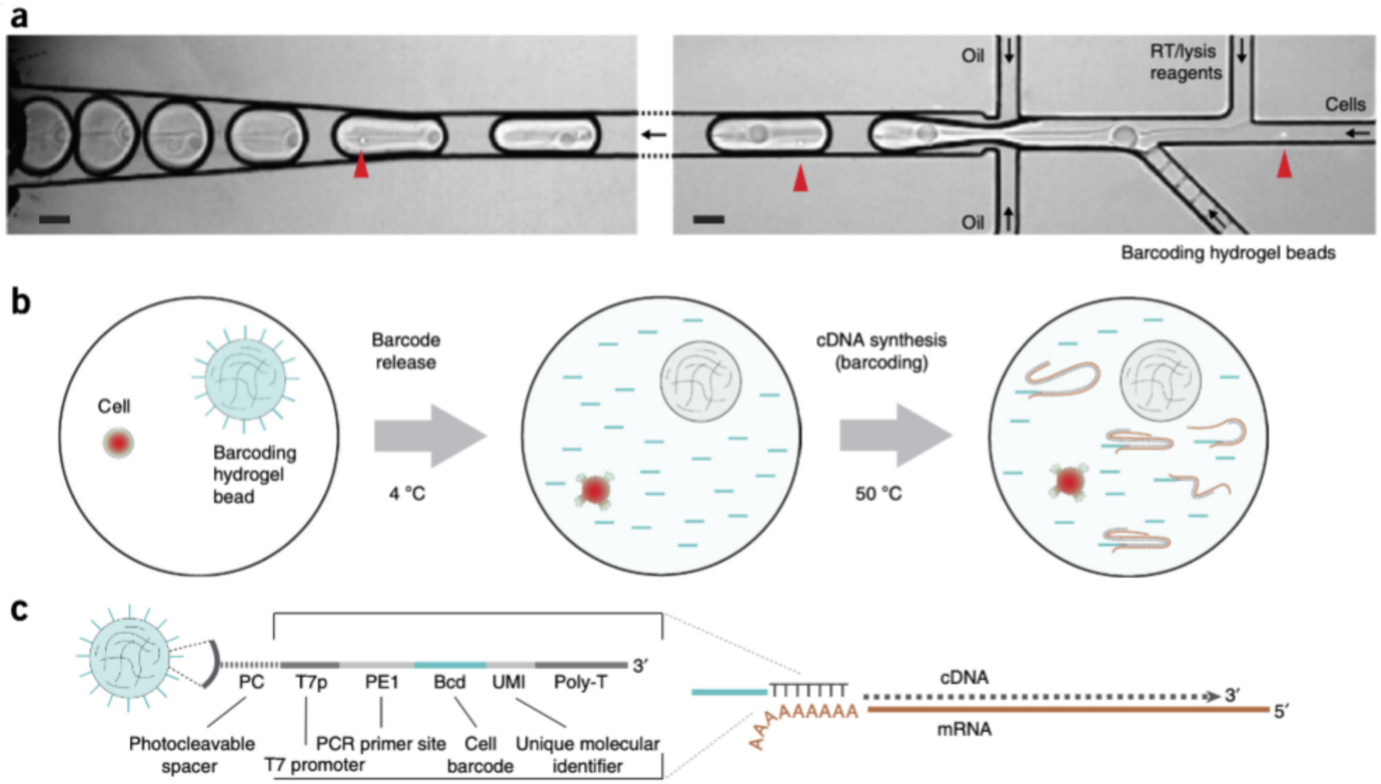
▲ CRITICAL STEP The number of PCR cycles depend
eukaryotic cells to obtain ~1–30 ng of amplified cD
content) or lowered for large cells (with more RNA).

▲ CRITICAL STEP The number of cycles depends
of amplified cDNA, we usually perform 12 PCR cyc
ment. It may be helpful to run a range of cycles t
of the input DNA used for tagmentation.

SMART-Seq2 meets all the 'requirements' we introduced above. It is universal, in situ measurements, has no minimum input, with 100% capture rate and sensitivity, etc. The only problems is that it is **not** cost effective or easy to use.

The inDrops Approach (Microfluidics)

This is a technique to isolate cells in a faster, more automated process.



This technique places each cell in a droplet, where each one contains a barcoding hydrogel. Ideally, every droplet contains exactly one cell and one barcode. The droplets are separated using oil (nonpolar).

- Ideally, droplets contain exactly one barcode and one cell
- Droplets with a barcode but no cell is okay, it is just ignored since there is no cell.
- Doublets (droplets with more than one cell), and uncaptured cells are bad because they skew the data.
- Splits are when a droplet have one cell but two barcodes, and collisions are when a single barcode is present in two different droplets. Both are bad because they double-count the cells.

Barcode Diversity

- Barcode collisions occur when beads with identical barcode sequences are present in droplets with two different cells.
- The number of available barcode sequences depends on the sequence length L. Sequences of length L can yield up to 4^L barcodes.
- The number of distinct barcodes needed is a function of the number of cells that are to be barcoded.

Assuming that each of the N cells get one barcode at random from a set of M barcodes, the expected number of cells with a unique barcode is given by

$$\mathbb{E}(\text{cells with unique barcode}) = N\left(1 - \frac{1}{M}\right)^{N-1}$$

Proof: If we denote the probability that any specific barcode associates with some cell by p, then $p = 1/M$. The probability that a given barcode is used for some specific set fo k cells is therefore:

$$\mathbb{P}(\#cells = k) = \binom{N}{k} p^k (1-p)^{N-k}$$

where:

- N = number of cells
- M = number of barcodes
- p = probability a specific barcode associated with a cell

Simplifying:

$$\begin{aligned}\mathbb{P}(\#cells = 1) &= \binom{N}{1} p^1 (1-p)^{N-1} \\ &= \binom{N}{1} \frac{1}{M}^1 \left(1 - \frac{1}{M}\right)^{N-1} \\ &= \frac{N}{M} \left(1 - \frac{1}{M}\right)^{N-1}\end{aligned}$$

Due to how expected values work (e.g., $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$), thus,

$$\mathbb{E}(\text{cells with unique barcode}) = M \cdot \frac{N}{M}\left(1 - \frac{1}{M}\right)^{N-1} = N\left(1 - \frac{1}{M}\right)^{N-1}$$

For N assayed cells and M barcodes, the **barcode collision rate** can be estimated as

$$1 - \frac{\mathbb{E}(\text{cells with a unique barcode})}{\text{number of cells}} = \qquad 1 - \left(1 - \frac{1}{M}\right)^{N-1} \approx 1 - \left(\frac{1}{e}\right)^{\frac{N}{M}}$$

Barcode collisions lead to **synthetic doublets**. Avoiding synthetic doublets requires high **relative barcode diversity**, i.e., a high ratio of M/N.

Droplet Tuning Concepts

- The **capture rate** is 1 - the fraction of cells that are in droplets without any beads.
- The **split rate** is the fraction of droplets with exactly one cell that have more than one bead.
- The **doublet rate** is the fraction of droplets with 1 bead that have more than one cell.

Binomial Distributions for Beads and Cells

- Consider n droplets, each of which have a probability p of containing a single cell. Then the probability that k cells will be captured is

$$\mathbb{P}(\text{number of cells} = \text{k}) = \binom{n}{k} p^k (1-p)^{n-k}$$

- This suggests modeling the number of cells captured with a random variable that follows a Binomial distribution. That is, $X \sim B(n, p)$.
- By the law of rare events, for a large n and small p, $B(n,p)$ is approximated well with the Poisson distribution $Poi(np)$. i.e.

$$\binom{n}{k} p^k (1-p)^{n-k} \approx e^{-\lambda} \frac{\lambda^k}{k!}$$

- This is convenient for many reasons: the expression on the right is easier to evaluate and the parameter λ is readily interpretable as the expected value of a Poisson random variable.
- The expected value of a Poisson random variable is λ.

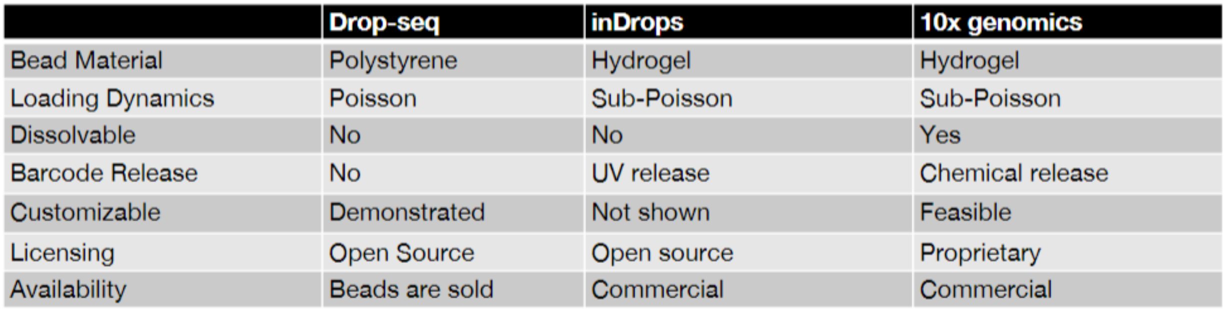
A Poisson Approximation for Beads and Cells

- Load **cells** into droplets at Poisson rate λ.
- Load **beads** into droplets at Poisson rate μ.

$$\mathbb{P}(\text{droplet has } k \text{ cells}) = \frac{e^{-\lambda}\lambda^k}{k!}$$
$$\mathbb{P}(\text{droplet has } j \text{ beads}) = \frac{e^{-\mu}\mu^j}{j!}$$

- The Poisson approximation yields a simple formula for the capture rate: $1 - e^{-\mu}$.
- The split rate estimate is $\frac{(1-e^{-\mu}-\mu e^{-\mu})}{1-e^{-\mu}}$
- This provides a quantitative assessment of the tradeoff between the capture rate and the split rate.

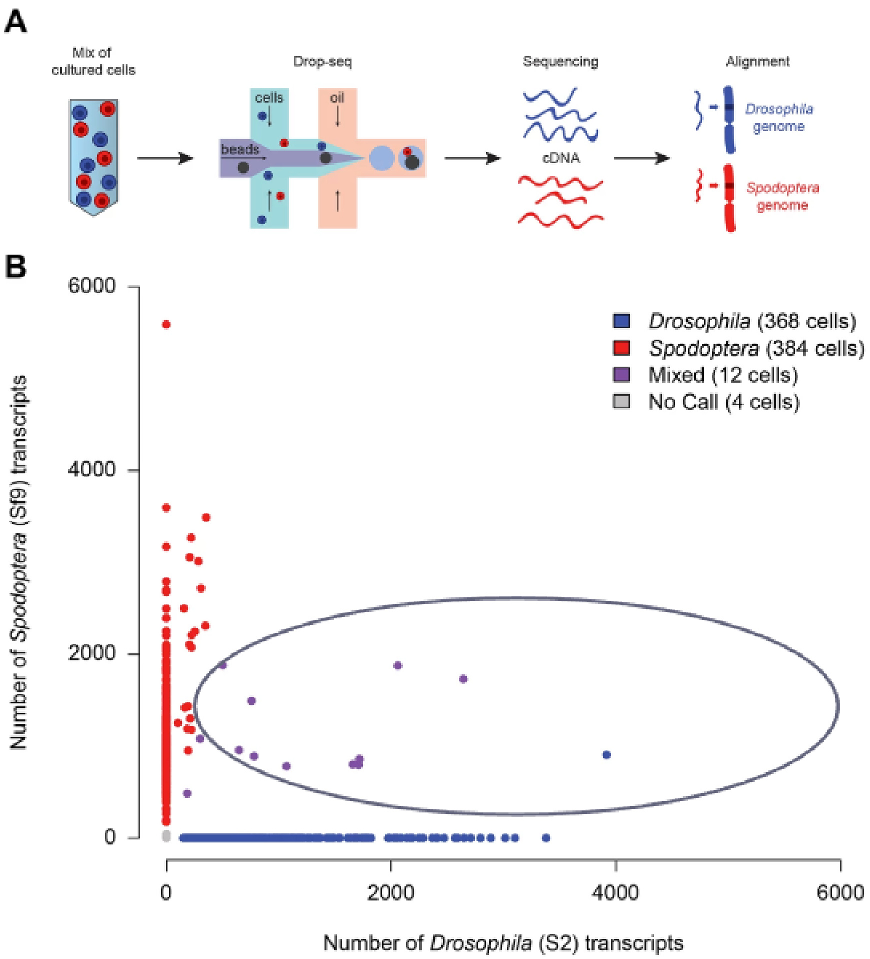
Reducing the Number of Beadless Droplets



	Drop-seq	inDrops	10x genomics
Bead Material	Polystyrene	Hydrogel	Hydrogel
Loading Dynamics	Poisson	Sub-Poisson	Sub-Poisson
Dissolvable	No	No	Yes
Barcode Release	No	UV release	Chemical release
Customizable	Demonstrated	Not shown	Feasible
Licensing	Open Source	Open source	Proprietary
Availability	Beads are sold	Commercial	Commercial

Technical Doublets

- Technical doublets arise when two or more cells are captured in a droplet with a single bead. The technical doublet rate is therefore the probability of capturing two or more cells in a droplet given that at least one cell has been captured in a droplet: $\frac{(1-e^{-\lambda}-\lambda e^{-\lambda})}{1-e^{-\lambda}}$.
- Note that "overloading" a microfliuidics single-cell experiment by loading more cells while keeping flow rates constant will increase the number of technical doublets due to an effective increase in λ.



Biological Doublets

- **Biological doublets** arise when two cells form a discrete unit that does not break apart during disruption to form a suspension.
- Biological doublets will not be detected via barnyard plots
- One approach to avoiding biological doublets is to perform nuclear single-cell RNA-seq
- However, biological doublets are not necessarily just a technical nuisance to be avoided. The paper Halpern et al. 2018 utilizes biological doublets of hepatocytes and liver endothelial cells to assign tissue coordinates to liver and endothelial cells via imputation from their hepatocyte partners.

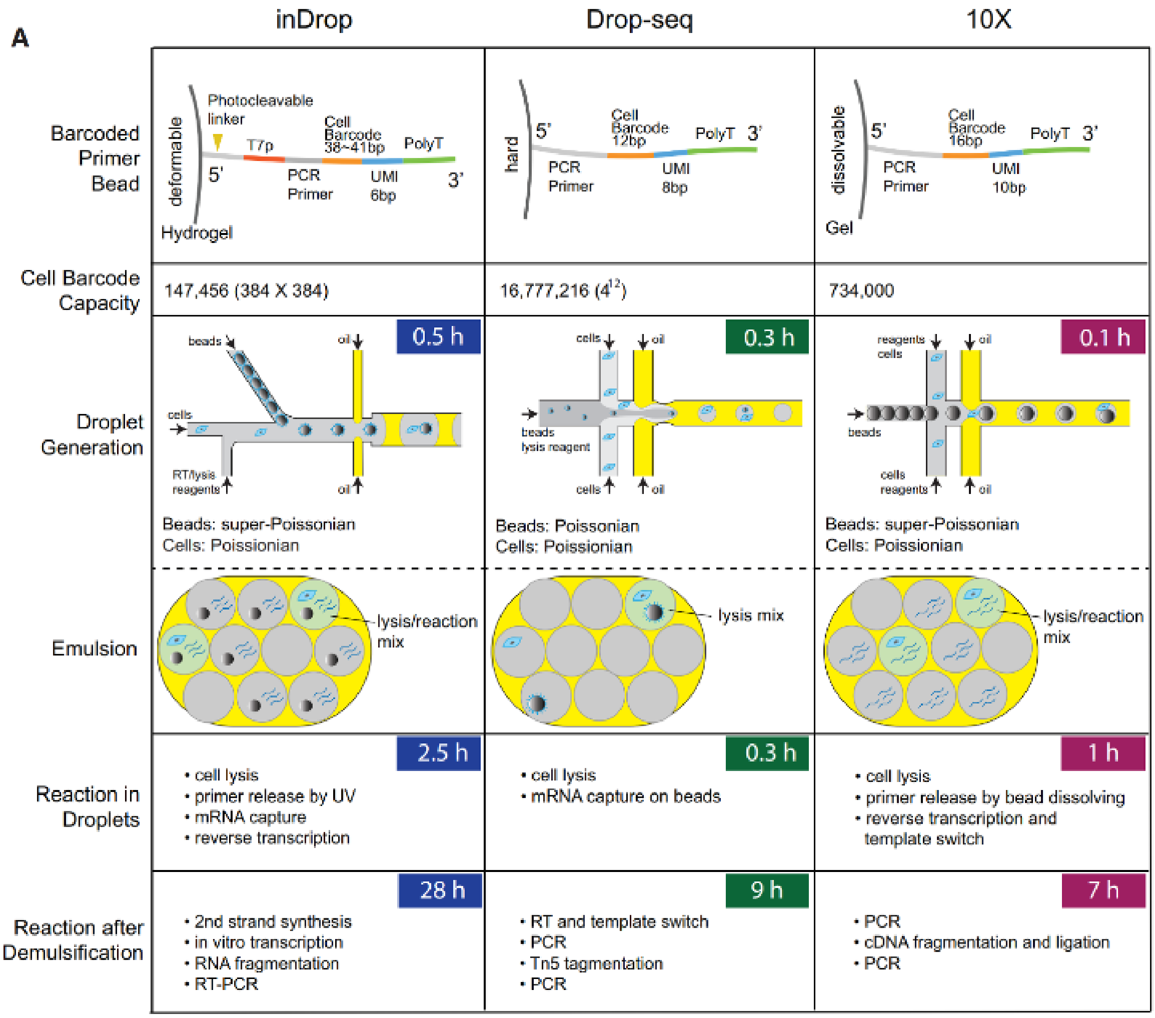
Unique Molecular Identifiers

- Length determined by the diversity calculation and collision rater (same as barcodes).
- *UMI collapsing* refers to the process of using UMIs to avoid double-counting molecules after sequencing.
  - Naive UMI collapsing consists of just counting reads that have the same UMI and cell barcode as a single event
  - UMI collapsing can include collision detection by checking whether reads also originate from the same molecule.

Remarks on Cell Barcodes and UMIs

- Sequencing errors can lead to:
  - incorrectly labeled cells (from cell barcodes).
  - erroneous molecule counts (from UMIs).
- Error correction can be used to address these problems
  - Cell barcode error correction can sometimes be performed using a list of the known cell barcodes in the experiment (technology department).
  - Cell barcode and UMI error correction can be performed by first identifying sequences that likely represent true barcodes (based on frequency).
- Sequencing errors are (sequencing) technology dependent.

Summary of Droplet Single-Cell RNA-seq Methods and Features



	inDrop	Drop-seq	10X
Barcoded Primer Bead			
Cell Barcode Capacity	147,456 (384 X 384)	16,777,216 (4^{12})	734,000
Droplet Generation	0.5 h Beads: super-Poissonian Cells: Poissionian	0.3 h Beads: Poissonian Cells: Poissionian	0.1 h Beads: super-Poissonian Cells: Poissionian
Emulsion			
Reaction in Droplets	2.5 h • cell lysis • primer release by UV • mRNA capture • reverse transcription	0.3 h • cell lysis • mRNA capture on beads	1 h • cell lysis • primer release by bead dissolving • reverse transcription and template switch
Reaction after Demulsification	28 h • 2nd strand synthesis • in vitro transcription • RNA fragmentation • RT-PCR	9 h • RT and template switch • PCR • Tn5 tagmentation • PCR	7 h • PCR • cDNA fragmentation and ligation • PCR

What Single-cell RNA-seq is not

- It is not **cell**: While measurements are made from RNA molecules in cells, not *all* RNA molecules are captured. In fact, very few RNA molecules are captured. Therefore, what is measured does not provide a complete picture of the RNA inside cells.

- It is not **single**: Since measurements from cells are incomplete, claims are for the most part restricted to groups of cells, rather than individual cells.
- It is not **RNA**: Single-cell RNA-seq largely consists of sampling cDNA molecules from a cDNA library, which serves as a proxy for the (captured) RNA content in cells.

So what are we missing?

We got:

- Universal in terms of cell size, type, and state.
- No minimum input of number of cells to be assayed
- Easy to use.
- Open source.

and kind of got

- Transcripts are readily associated to single cells, e.g., no doublets
- Additional measurements of other cell attributes.

We don't got:

- In situ measurements
- Every cell is assayed, i.e., 100% capture rate
- Every transcript in every cell is detected, i.e., 100% sensitivity.
- Every transcript is identified by its full-length sequence
- Cost effective per cell

So the Data is Complicated. What can we do?

- Have a mix of cells and identify "cell types"
  - Often done using clustering + low dimension projection/embedding
  - We will talk about UMAP, t-SNE, etc.
- Identify genes that are associated with differentiation, aka "pseudotime"
- Differential expression of the "difference" between cell-types... sort of $\ddot\smile$
- Highly-parallel perturbation experiments with CRISPR
- We are starting to see multimodal technologies (e.g., RNA + chromatin). Can we see dynamics?
- Imputation of unobserved genes
- Alternative splicing? Technology will get there...