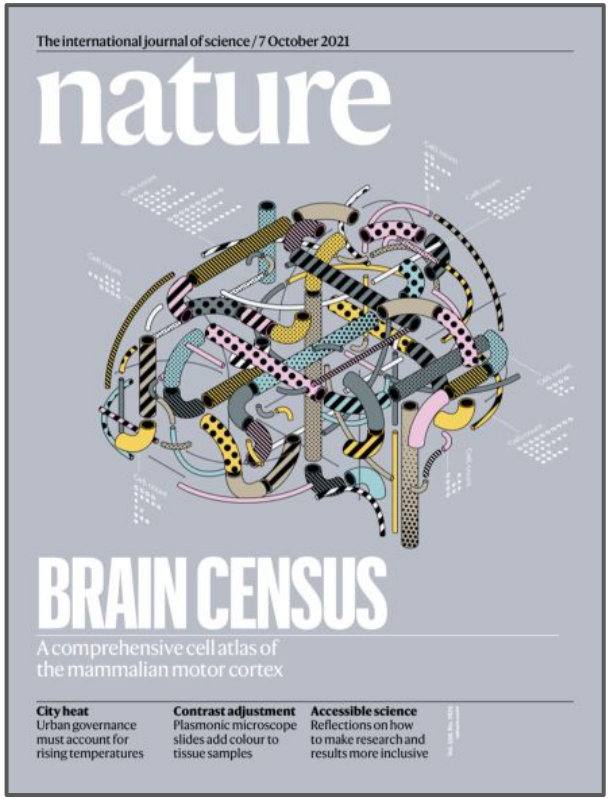
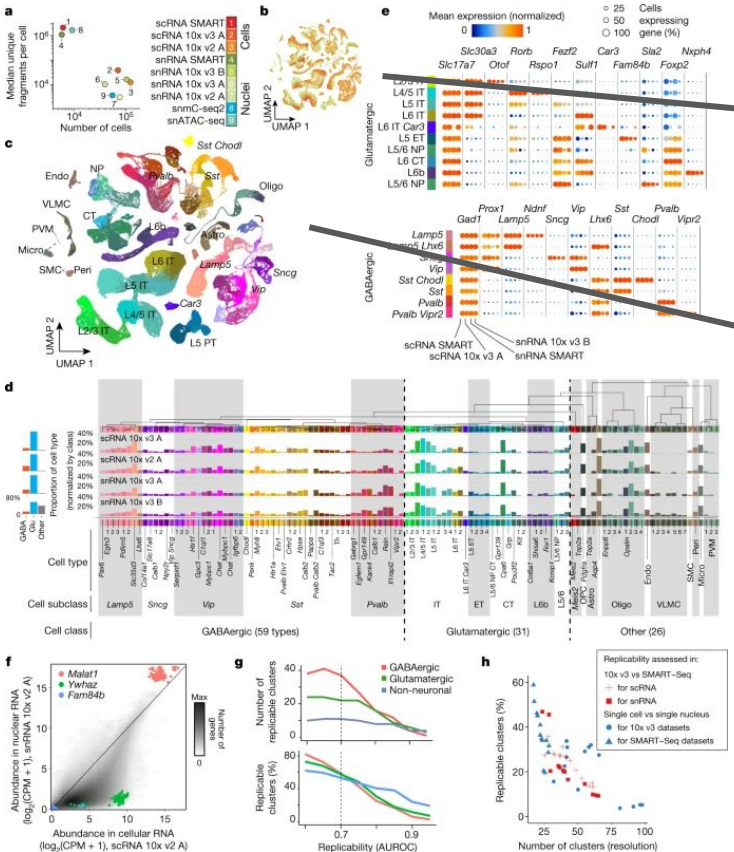
Introduction Lecture

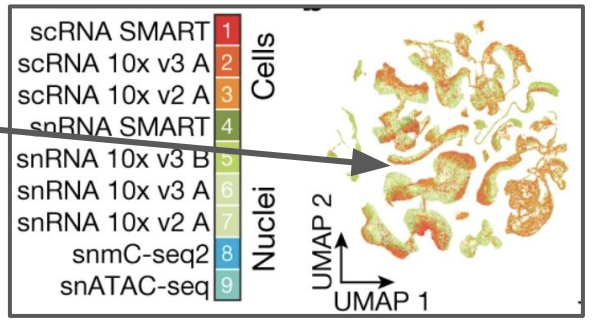
Data Science and AI for Neuroscience Summer School
Tara Chari and Lior Pachter
July 11, 2022

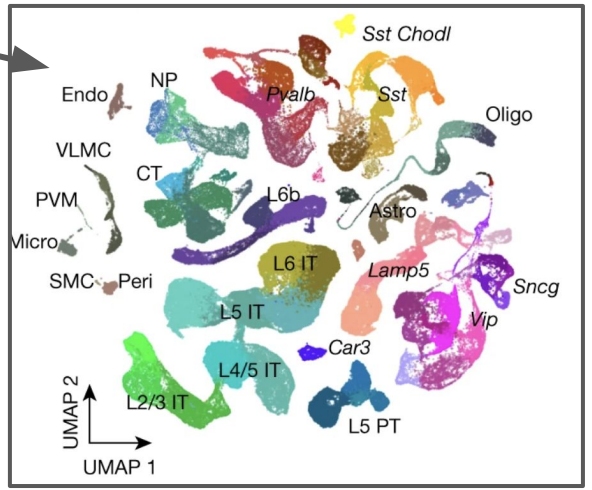
The **B**rain **I**nitiative **C**ell **C**ensus **N**etwork



Multi platform transcriptomic taxonomy of cell types in the MOp

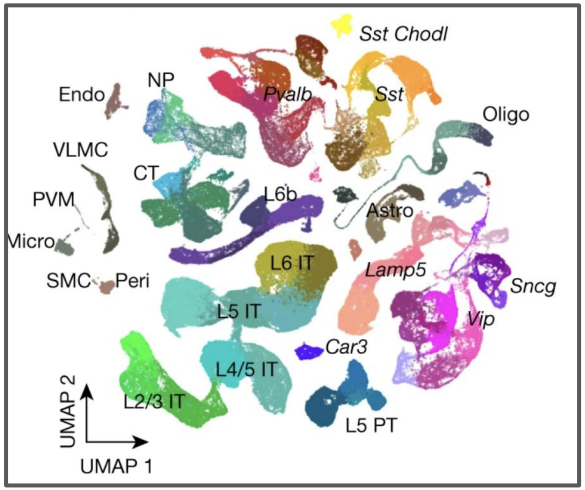






Yao et al., 2021

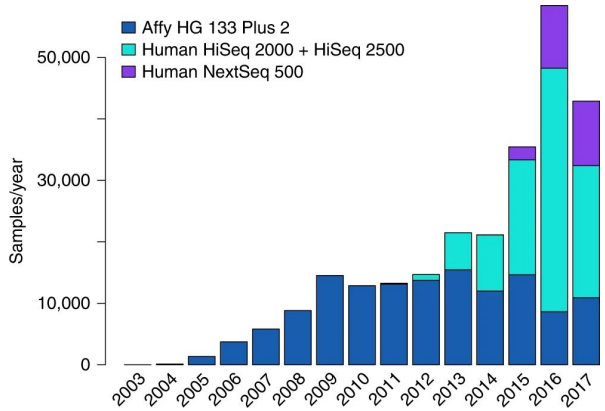
The cell atlas as a “result”



- A single-cell gene RNA-seq experiment produces a gene expression matrix. This is an array of *genes x cells* whose entries record how many times a gene has been observed in a cell.
- Geometrically, the table can be viewed as describing the cells as points in a high-dimensional space.
- A subset of the matrix is used to cluster the cells.
- The matrix is embedded in 2D to produce a visualization of the cells. Popular methods are *non-linear*, and try to optimize preservation of local neighborhoods of cells.
- The cells in the 2D embedding are colored according to their cluster assignments.

Why single-cell RNA-seq?

- A bulk RNA-seq experiment produces a gene expression vector
  - Prior to bulk RNA-seq, bulk measurements of gene expression could be performed using DNA microarrays.
    - (While DNA microarrays are not used much anymore, *methods* for analysis of DNA microarray data are frequently re-discovered (and republished) as single-cell RNA-seq analysis methods).



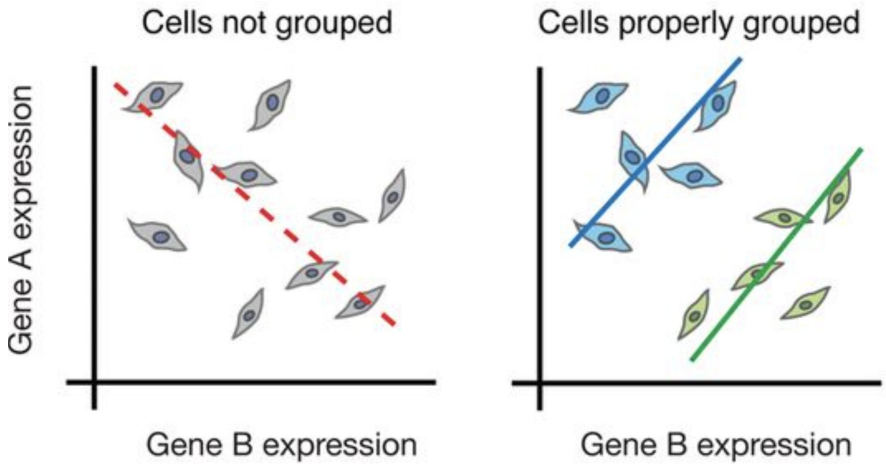
Lachmann et al., 2018

Publicly available RNA-seq samples currently available at GEO/SRA for human and mouse compared to available samples collected with the popular Affymetrix HG U133 Plus 2 platform.

Why *single-cell* RNA-seq?

- While bulk RNA-seq has been popular, there is a problem with analysis of gene expression in bulk: *averaging.*

- Robinson's paradox:



Trapnell et al., 2015

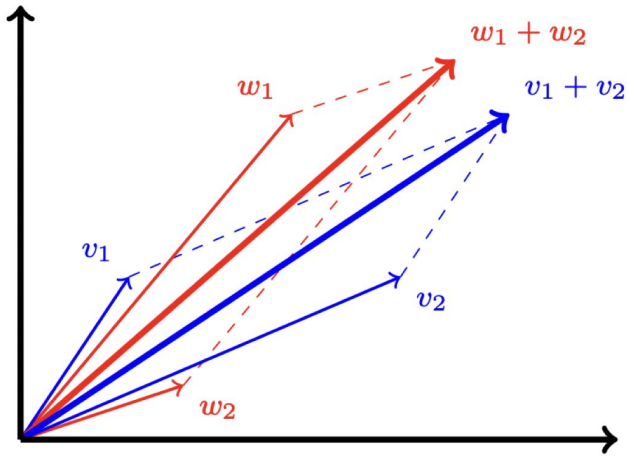
A categorical version: Simpson's reversal

	Democrats		Republicans		
North	**94%**	145/154	85%	138/162	
South	**7%**	7/94	0%	0/10	House
	61%		**80%**		
North	**98%**	45/46	84%	27/32	
South	**5%**	1/21	0%	0/1	Senate
	69%		**82%**		

Votes for and against the Civil Rights Act, 1964

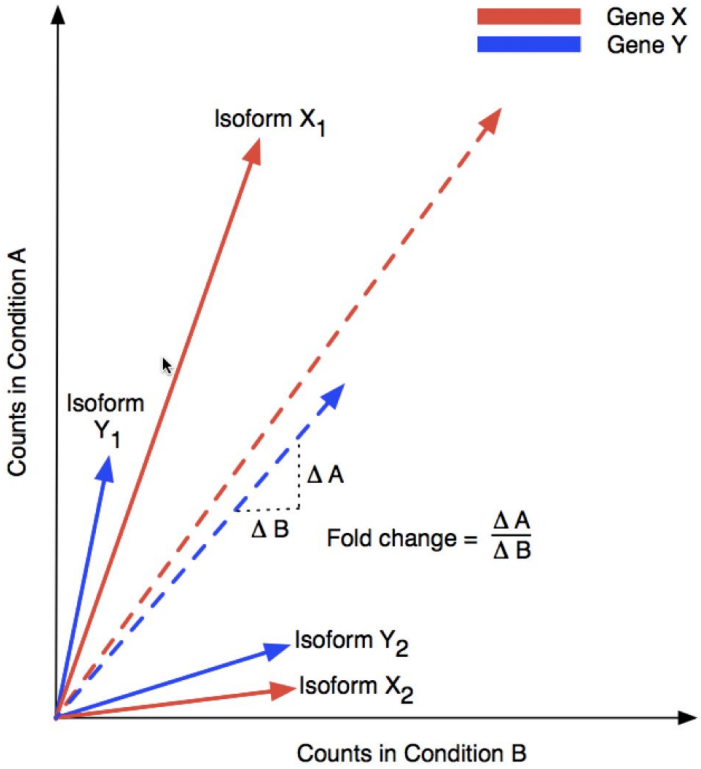
Geometrical interpretation of Simpson's reveral

- The slope (**rate of change**) of v_1 is larger than the slope of w_1; the slope of v_2 is larger than the slope of w_2.
- The slope of v_1+v_2 is less than the slope of w_1+w_2.



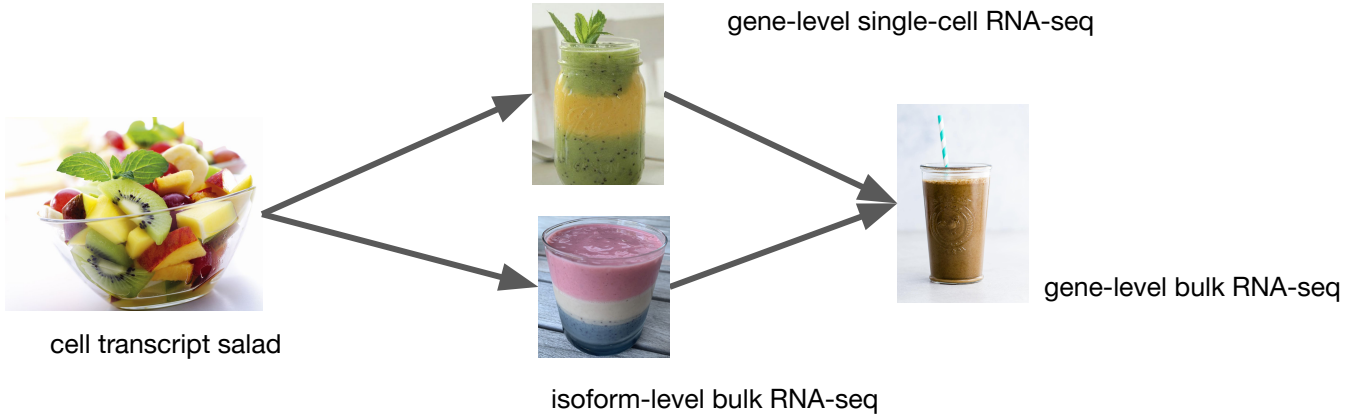
Linusson and Stamps, 2021

The relevance of isoform resolution

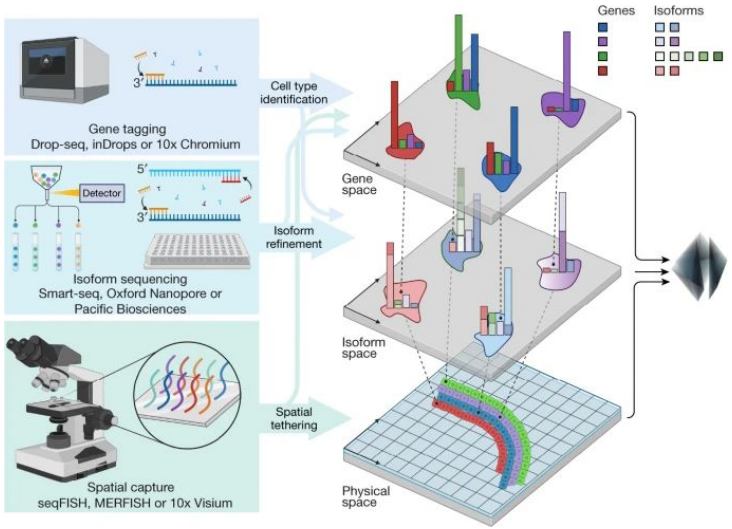


Getting to the bottom is about getting to the relevant story

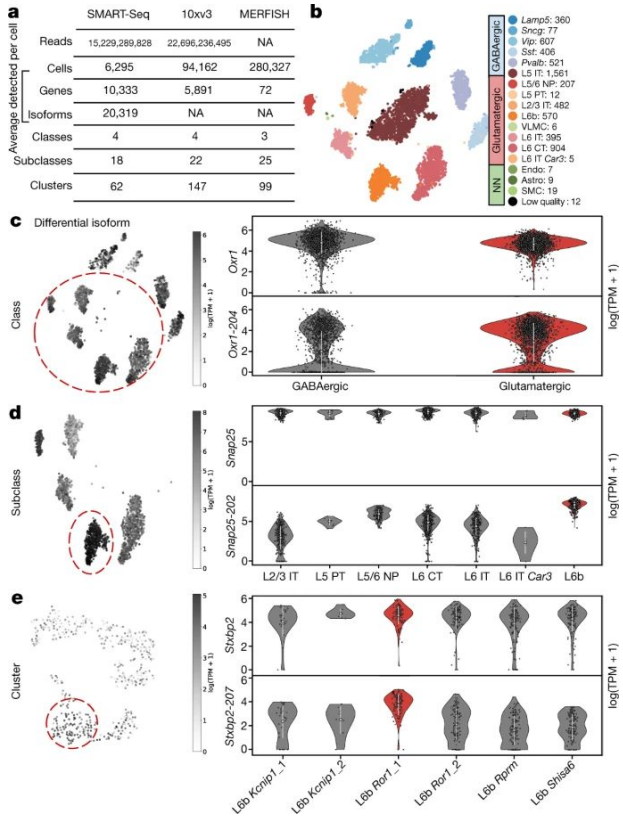
- The civil rights act example shows that lack of resolution may mask the importance and relevance of key confounding variables
  - The question of whether confounding variables are causal has bearing on whether to view the effect as a *paradox* or a *reversal* (Pearl, 2014).
- In terms of gene expression, “getting to the bottom” means
  - Isoform resolution **and** cell-type resolution.



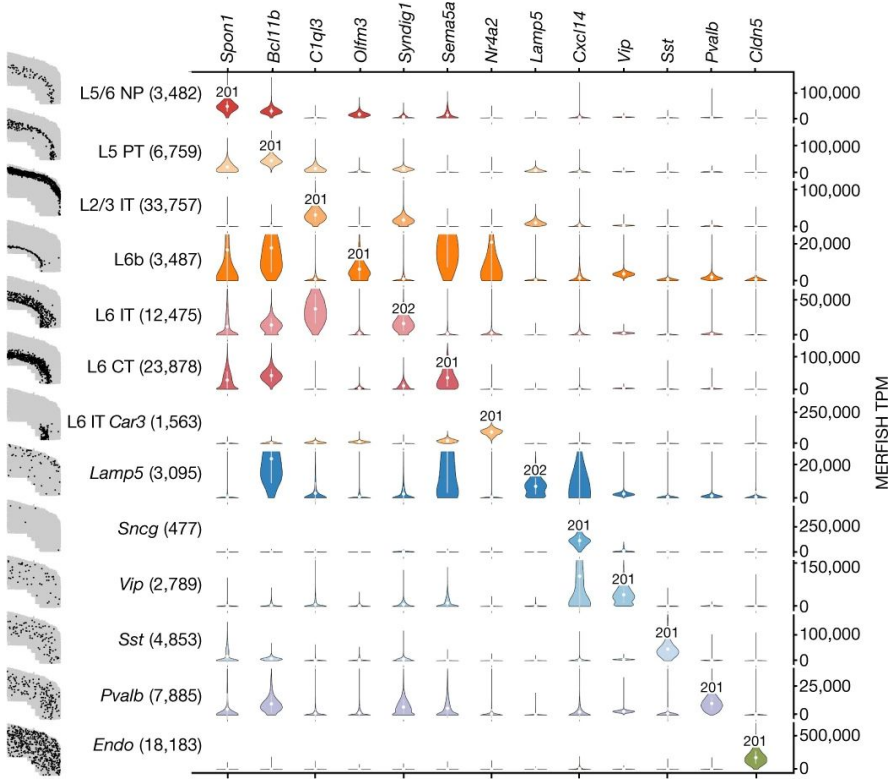
Isoform cell-type specificity in the mouse primary motor cortex



Booeshaghi et al., 2021



An isoform atlas of the MOp



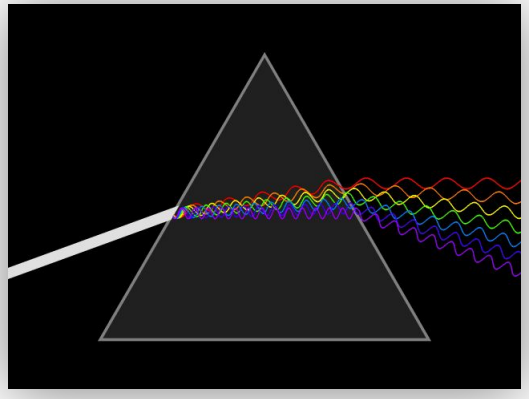
The averaging fallacies and paradoxes

- The dangers of averaging have been noted many times and there are many names that are used to refer to the different types of reversal. Some distinctions and confusions:
  - Continuous vs. categorical explanatory, outcome and confounding variables.
  - Weak (≥) vs. strong (>) reversal.
  - Reversal vs. paradox

- Named versions include Robinson's paradox, Simpson's paradox, suppression, Lord's paradox, amalgamation paradox, Yule-Simpson effect, …
  - The terms "**Simpson's reversal**" or "**Simpson's paradox**" are frequently used to describe many/all of these (even though Simpson's initial paper described a weak reversal for a very specific case).
  - The related term "**ecological fallacy**" refers to the fallacy of making inferences about individuals from group statistics.

The purpose of single-cell RNA-seq

- Decompose tissue or organ expression into its constituent parts
- Identify the different types of cells that comprise tissues and organs
- Examine the molecular biology of cells via their expression signatures
- Determine differentiation trajectories of cells
- Develop expression biomarkers for disease states



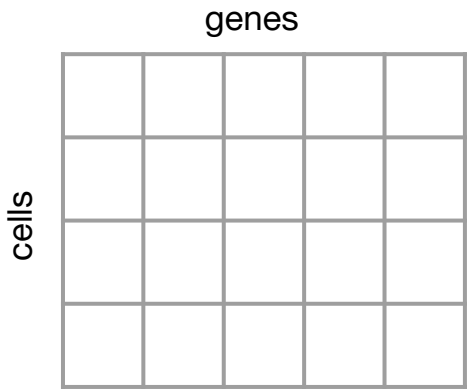
Single-cell RNA-seq as a theme for computational biology

- Data processing requires **efficient data structures** and **scalable algorithms**; many of the tools used are based on ideas and concepts that are applicable to a broad swath of (DNA sequence based) genomics.
- The **statistical challenges** that arise in single-cell RNA-seq data analysis are common throughout the biological sciences.
- **Mathematical** models for the molecular biology of the cell are key to interpretation of single-cell RNA-seq experiments. Elements of the models that are used are present in many other applications of mathematical biology.
- Single-cell RNA-seq **data is plentiful** and for the most part publicly available.
- Single-cell RNA-seq **technologies are becoming omnipresent** in biology labs.

Single-cell RNA-seq

- Single-cell RNA-seq is neither single, cell, nor RNA.
  So what is it?

- Single-cell RNA-seq refers to a group of (constantly improving) technologies and analysis tools that
  - start with an **INPUT** of cells,
  - **OUTPUT** a (proxy for a) gene expression matrix.



What is a gene expression matrix

- Expression: A term referring to the *process* by which *information* in a gene is used to generate protein or non-coding RNA product.

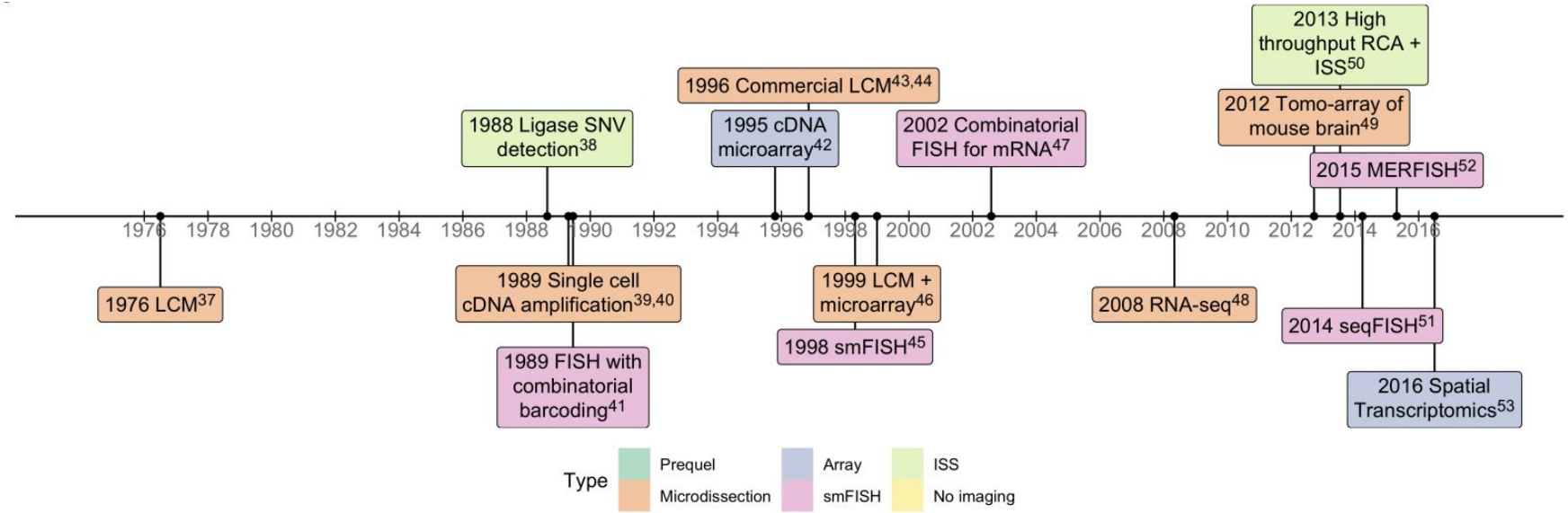
- Gene: A term coined by the Danish botanist Wilhelm Johannsen (1857-1927) in 1909. It has no precise definition, and its meaning has been evolving since it was introduced.

- Matrix: A rectangular array of numbers used to represent a linear map.

The ideal (dissociated) single-cell transcriptomics method

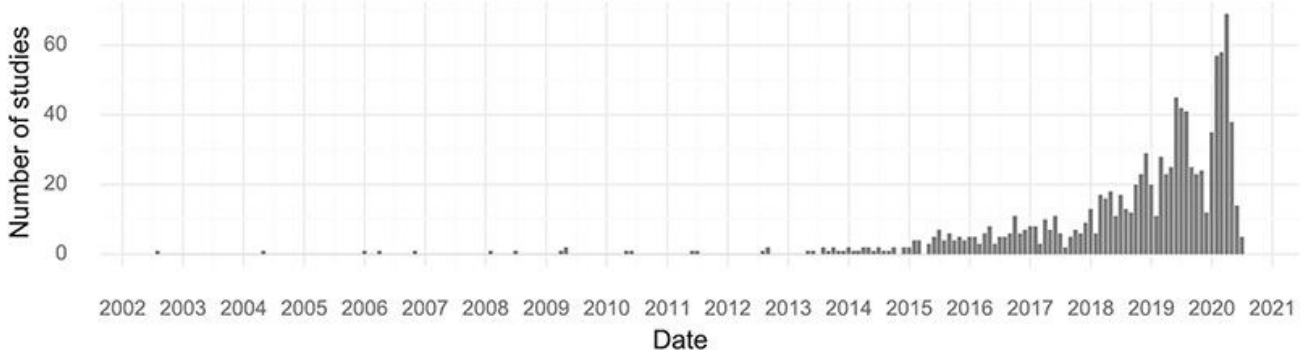
- **Universal** in terms of cell size, type and state.
- ***In situ*** measurements.
- No ***minimum input*** of number of cells to be assayed.
- Every cell is assayed, i.e.100% ***capture rate***.
- Every transcript in every cell is detected, i.e.100% ***sensitivity***.
- Every transcript is identified 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.

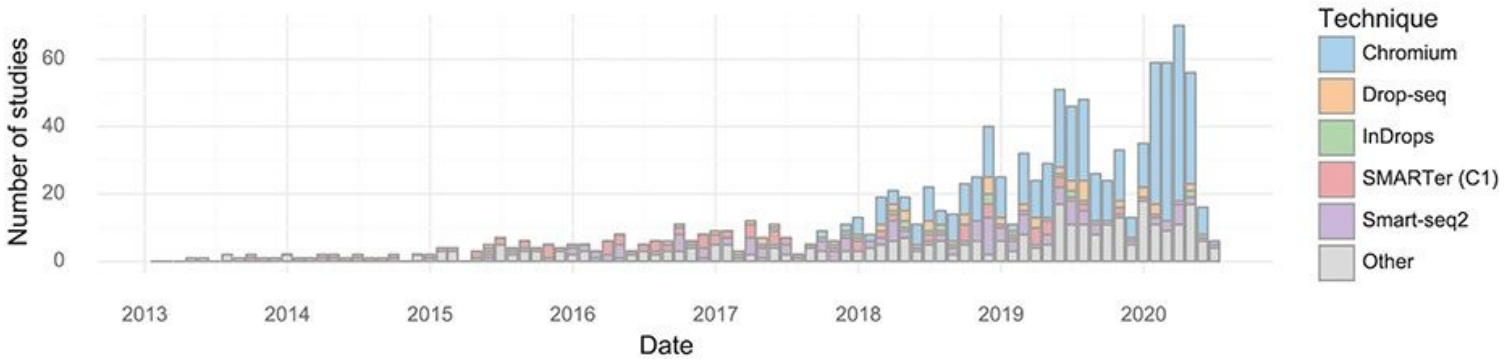
Spatial single-cell RNA-seq



Moses et al. 2021

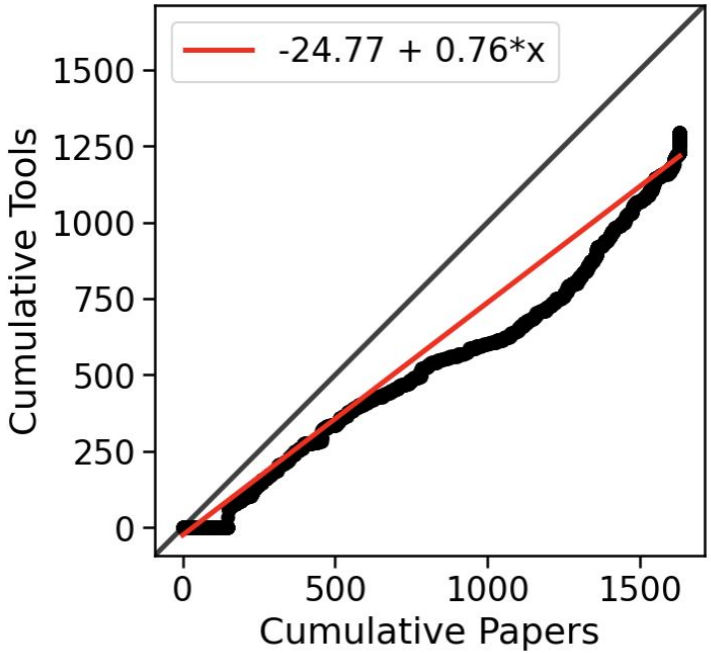
A decade of single-cell RNA-seq technology development





Svensson *et al.* 2020

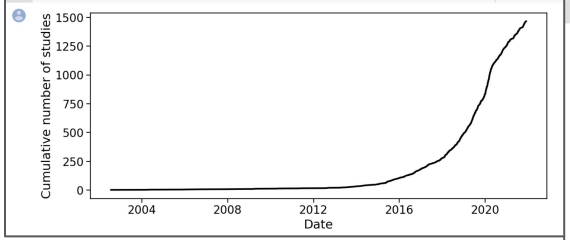
Technology development ⇔ methods development



Produced on July 11, 2022 with a Google Colab notebook

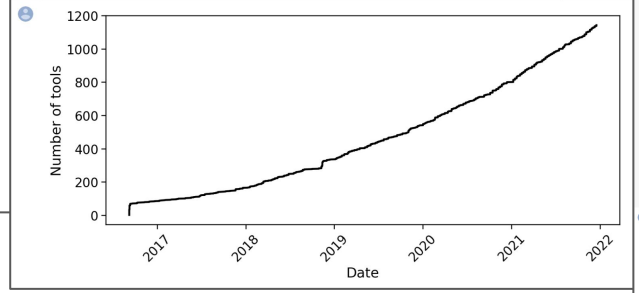
Google Colab notebook for this lecture

```
#@title Growth of single-cell RNA-seq
df = pd.read_csv('http://nxn.se/single-cell-studies/data.tsv', sep='\t')
# converts string to date format, can only be run once!
df['Date'] = pd.to_datetime(df['Date'], format='%Y%m%d')

# converts string of reported cells total to float, can only be run once!
df['Reported cells total'] = df['Reported cells total'].str.replace(',', '').map(float)

# plot number of studies over time
fig, ax = plt.subplots(figsize=(12, 5))

papers = pd.read_csv('http://nxn.se/single-cell-studies/data.tsv', sep='\t')
papers['Datetime'] = pd.to_datetime(papers['Date'], format='%Y%m%d')
papers = papers.sort_values("Date")
papers["count"] = 1

x = papers.Datetime
y = papers["count"].groupby(papers.Datetime.dt.time).cumsum()

ax.plot(x, y, color="k")
ax.set_xlabel("Date")
ax.set_ylabel("Cumulative number of studies")

plt.show()
```



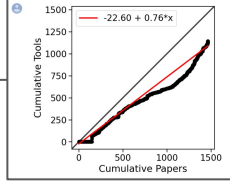
Svensson et al., 2020

```
#@title Growth of single-cell tools { run: "auto" }
tools = pd.read_csv('https://raw.githubusercontent.com/Oshlack/scRNA-tools/master/database/tools.tsv', sep='\t')
tools["Datetime"] = pd.to_datetime(tools["Added"])
tools = tools.sort_values("Added")
tools["count"] = 1

fig, ax = plt.subplots(figsize=(12, 5))

x = tools.Datetime
y = tools["count"].groupby(tools.Datetime.dt.time).cumsum()

ax.plot(x, y, color="k")
ax.set_xlabel("Date")
ax.set_ylabel("Number of tools")
ax.tick_params(axis='x', rotation=45)


plt.show()
```

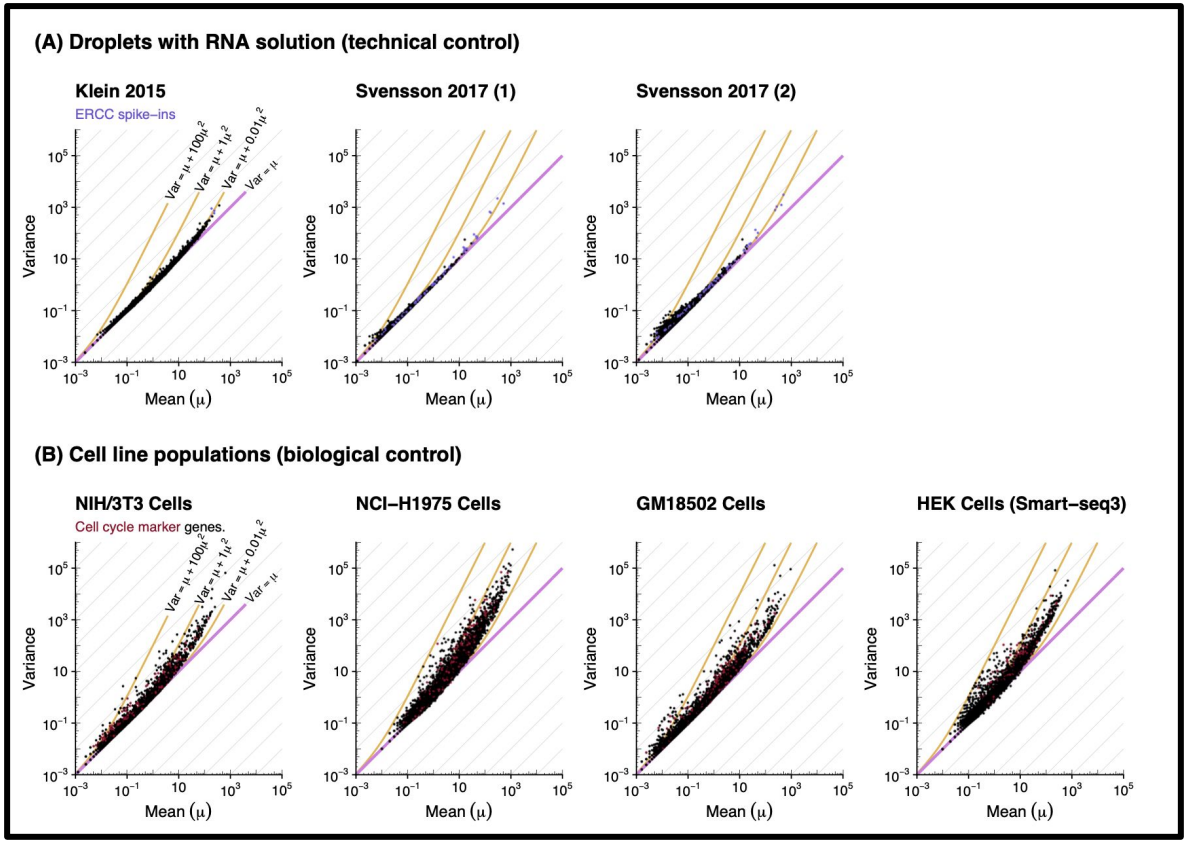


Zappia et al., 2018

```
#@title scRNA-seq tools vs. studies linear regression
date_papers = papers.groupby("Datetime")["count"].sum()
date_tools = tools.groupby("Datetime")["count"].sum()
dates = pd.date_range(start='7/26/2002', end='01/01/2025')
combined = pd.DataFrame(index=dates)
combined["tool_counts"] = combined.index.map(date_tools)
combined["paper_counts"] = combined.index.map(date_papers)
combined = combined.fillna(0)
combined["Datetime"] = combined.index.values

fig, ax = plt.subplots(figsize=(5,5))

x = combined["paper_counts"].groupby(combined.Datetime.dt.time).cumsum()
y = combined["tool_counts"].groupby(combined.Datetime.dt.time).cumsum()

ax.scatter(x, y, color="k")
regr = linear_model.LinearRegression()
x = x.values[:, np.newaxis]
regr.fit(x, y.values)
xx = np.linspace(0, max(x), 200)
yy = regr.intercept_ + regr.coef_*xx
ax.plot(xx, yy, color="r", label=f"{regr.intercept_:.2f} + {regr.coef_[0]:.2f}*x")

lims = [np.min([ax.get_xlim(), ax.get_ylim()]),  # min of both axes
        np.max([ax.get_xlim(), ax.get_ylim()])]  # max of both axes
ax.plot(lims, lims, 'k-', alpha=0.75, zorder=0)
ax.set_aspect('equal')
ax.set_xlim(lims)
ax.set_ylim(lims)

ax.set_xlabel("Cumulative Papers")
ax.set_ylabel("Cumulative Tools")
ax.legend()

plt.show()
```



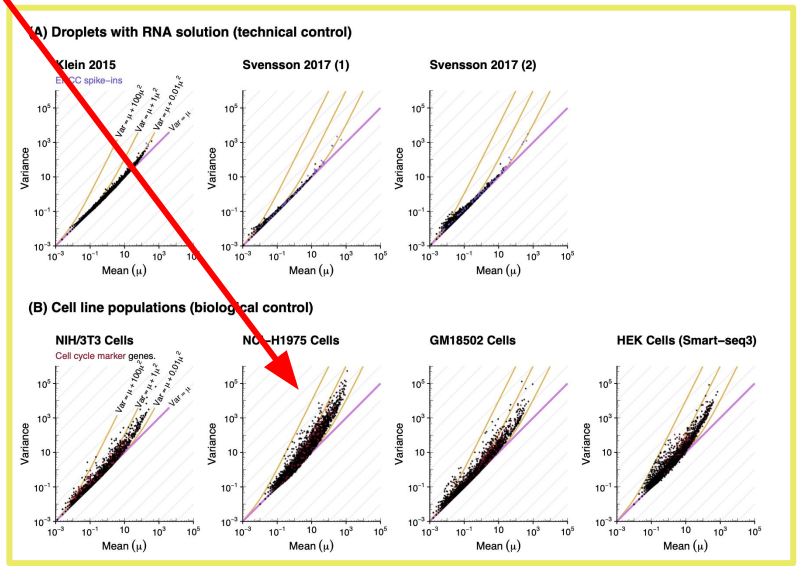
Lots of tools do not translate to lots of understanding



Ahlmann-Eltze and Huber, 2021

The variance is quadratic in the mean

- The negative binomial distribution yields a variance that is quadratic in the mean, namely $V = \mu + \phi \cdot \mu^2$, where μ is the mean and Φ is a parameter called the *dispersion* or *shape* parameter.



A question

- Is the observed variability due to:
    - **technical noise?**
    - **biological stochasticity?**

What do you believe?

Quantitative Biology

A **stochastic model for** bursty **transcription** coupling the telegraph model to a naive RNA transcription model yields a steady state **negative binomial distribution** for molecule counts.

Computational Biology

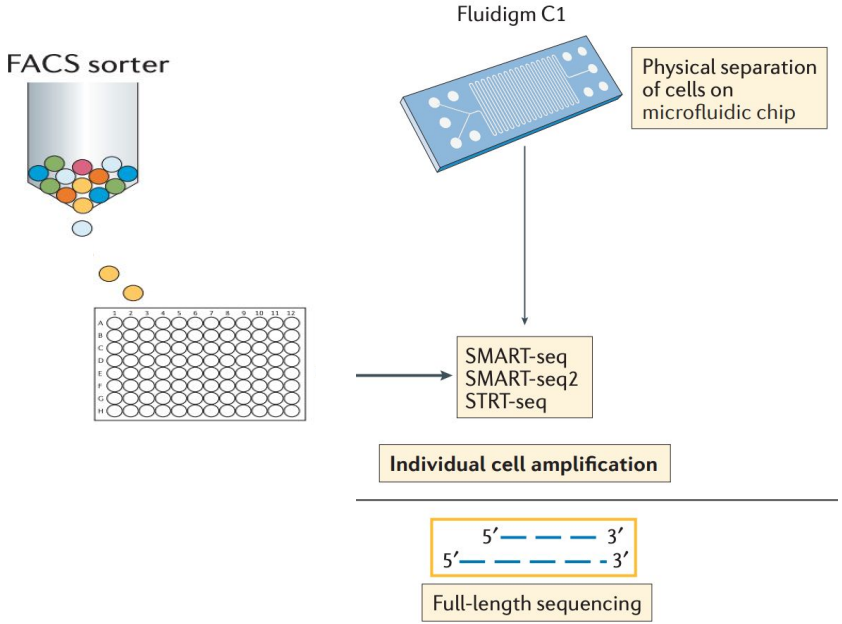
An overdispersed sampling model for RNA molecule counts from single-cell RNA-seq due to **technical variation** yields a **negative binomial distribution** for molecule counts.

Popular single-cell RNA-seq protocols

	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3′	5′	3′	3′	3′	3′	3′	3′	3′	3′
Number of cells per assay											

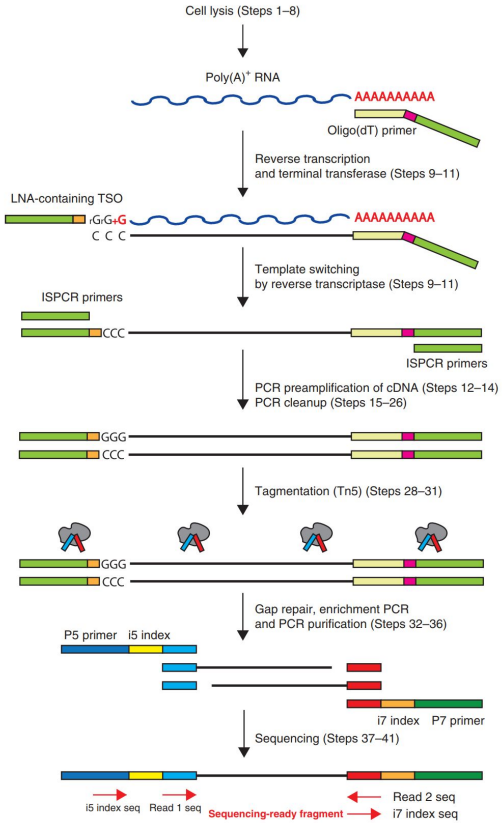
Chen et al. 2018

Physical separation of cells in wells



Papalexi and Satija, 2017

Example: library preparation for SMART-Seq2



▲ **CRITICAL STEP** All the experiments must be performed under a UV-sterilized hood with laminar flow, and all the surfaces must be free from RNase to prevent degradation of RNA and from DNA to prevent cross-contamination from previous samples. The hood must be used only for single-cell experiments up to (but excluding) the cDNA amplification step (Step 12). An ideal scenario would be to place the hood in a separate room with a positive air pressure to prevent any contaminants from being carried inside, where they might affect the experiments. The room should be equipped with a garmenting area in which the user changes into a fresh disposable lab coat, hair net, dust mask, shoe covers and vinyl gloves (powder-free).

▲ **CRITICAL STEP** Thaw all the reagents in advance and assemble the RT mix while performing denaturation (Step 7) to minimize bias.

▲ **CRITICAL STEP** The number of PCR cycles depends on the input amount of RNA. We typically use 18 cycles for single eukaryotic cells to obtain ~1–30 ng of amplified cDNA. The number of cycles can be increased for smaller cells (with less RNA content) or lowered for large cells (with more RNA).

▲ **CRITICAL STEP** The number of cycles depends on the amount of DNA used for tagmentation. If we are starting from 100 pg of amplified cDNA, we usually perform 12 PCR cycles. The optimal number of cycles depends on the sample and the experiment. It may be helpful to run a range of cycles to determine the best conditions. Above are cycling guidelines on the basis of the input DNA used for tagmentation.

Picelli et al., 2014

Example: SMART-Seq2 performance

- 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 assigned correctly to cells, e.g. no* ***doublets****.*
- *Additional multimodal measurements.*
- **Cost** effective per cell.
- Easy to use.
- Open source.

Cost is complicated to measure

Method	TPR[a]	FDR[a] (%)	Cell per Group[b]	Library Cost ($)	Minimal Cost[c] ($)
Smart-seq/C1	0.8	~4.9	150/172/215	~25	~9,010/9,440/11,290
Smart-seq2 (commercial)	0.8	~5.2	95/105/128	~30	~10,470/11,040/13,160
Smart-seq2 (in-house Tn5)	0.8	~5.2	95/105/128	~3	~1,520/1,160/1,090

Given the number of cells needed to reach 80% power as simulated above for three sequencing depths (Figure 6C), we calculated the minimal costs to generate and sequence these libraries. For example, at a sequencing depth of one million reads, SCRB-seq requires 64 cells per group to reach 80% power. Generating 128 SCRB-seq libraries costs ~260$ and generating 128 million reads costs ~640$. Note that the necessary paired-end reads for CEL-seq2/C1, SCRB-seq, MARS-seq, and Drop-seq can be generated using a 50-cycle sequencing kit, and, hence, we assume that sequencing costs are the same for all methods.

Ziegenhain et al. 2017

Sequencing costs are also complicated

Sequencer	read type	cost per lane UC rates	average reads (millions)	bases	Gb	cost /Gb	cost/million fragments
HiSeq3000/4000	SR50	$1,127	340	50	17	$66.29	$3.31
HiSeq3000/4000	SR100	$1,400	340	100	34	$41.18	$4.12
HiSeq3000/4000	PE100	$2,346	340	200	68	$34.50	$6.90
HiSeq3000/4000	PE150	$2,500	340	300	102	$24.51	$7.35
HiSeq2500 Rapid Mode	SR50	$1,127	140	50	7	$161.00	$8.05
HiSeq2500 Rapid Mode	SR100	$1,450	140	100	14	$103.57	$10.36
HiSeq2500 Rapid Mode	SR150	$1,759	140	150	21	$83.76	$12.56
HiSeq2500 Rapid Mode	PE100	$2,346	140	200	28	$83.79	$16.76
HiSeq2500 Rapid Mode	PE150	$2,720	140	300	42	$64.76	$19.43
HiSeq2500 Rapid Mode	PE250	$3,460	140	500	70	$49.43	$24.71

UC Davis Genome Center facility, January 2019. https://genomecenter.ucdavis.edu/

Microfluidic methods

- Macosko et al., **Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets**, 2015. DOI:10.1016/j.cell.2015.05.002

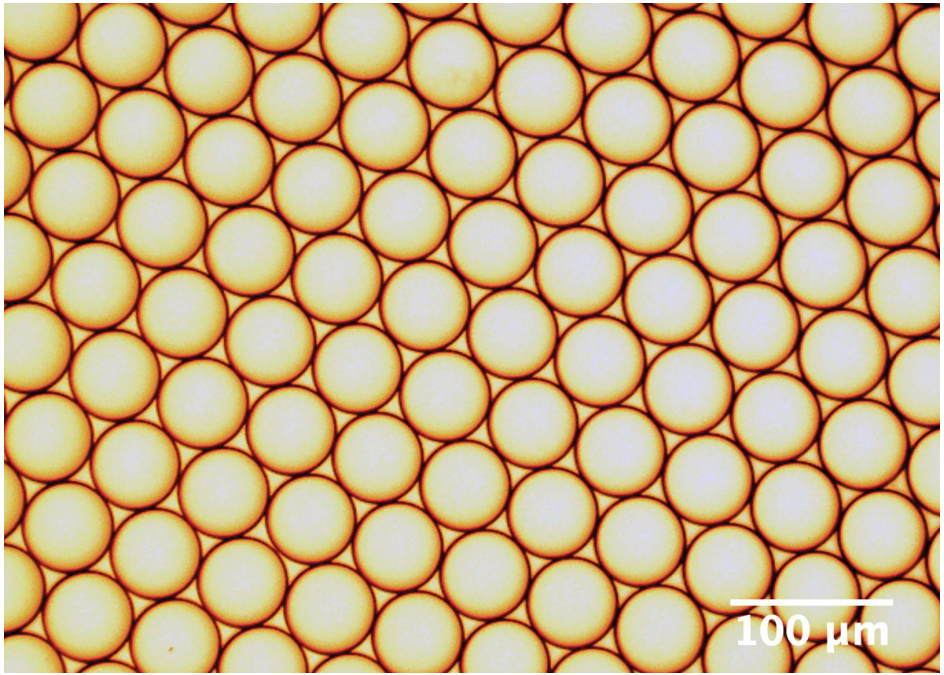
Drop-seq

- Klein et al., **Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells**, 2015. DOI:10.1016/j.cell.2015.04.044

inDrops

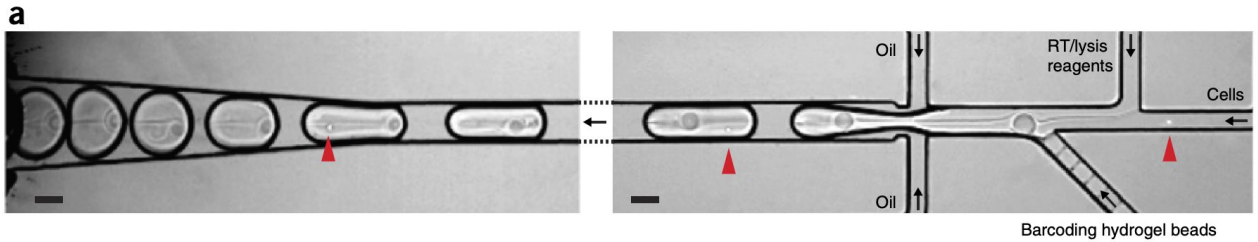
- Song, Chen, Ismagilov, Reactions in droplets in microfluidic channels, 2006. DOI:10.1002/anie.200601554
- Guo, Rotem, Heyman and Weitz, Droplet microfluidics for high-throughput biological assays, 2012. DOI:10.1039/C2LC21147E

Foundation: monodispersed emulsions



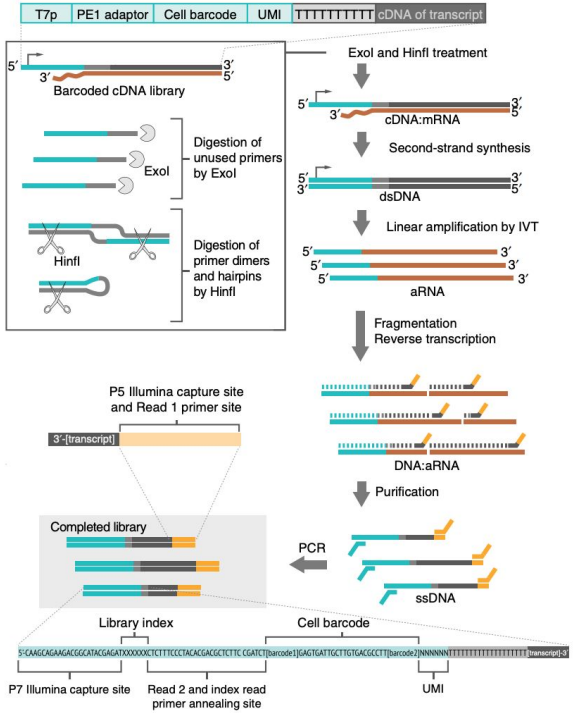
Booeshaghi et al. 2019

Example: the inDrops approach



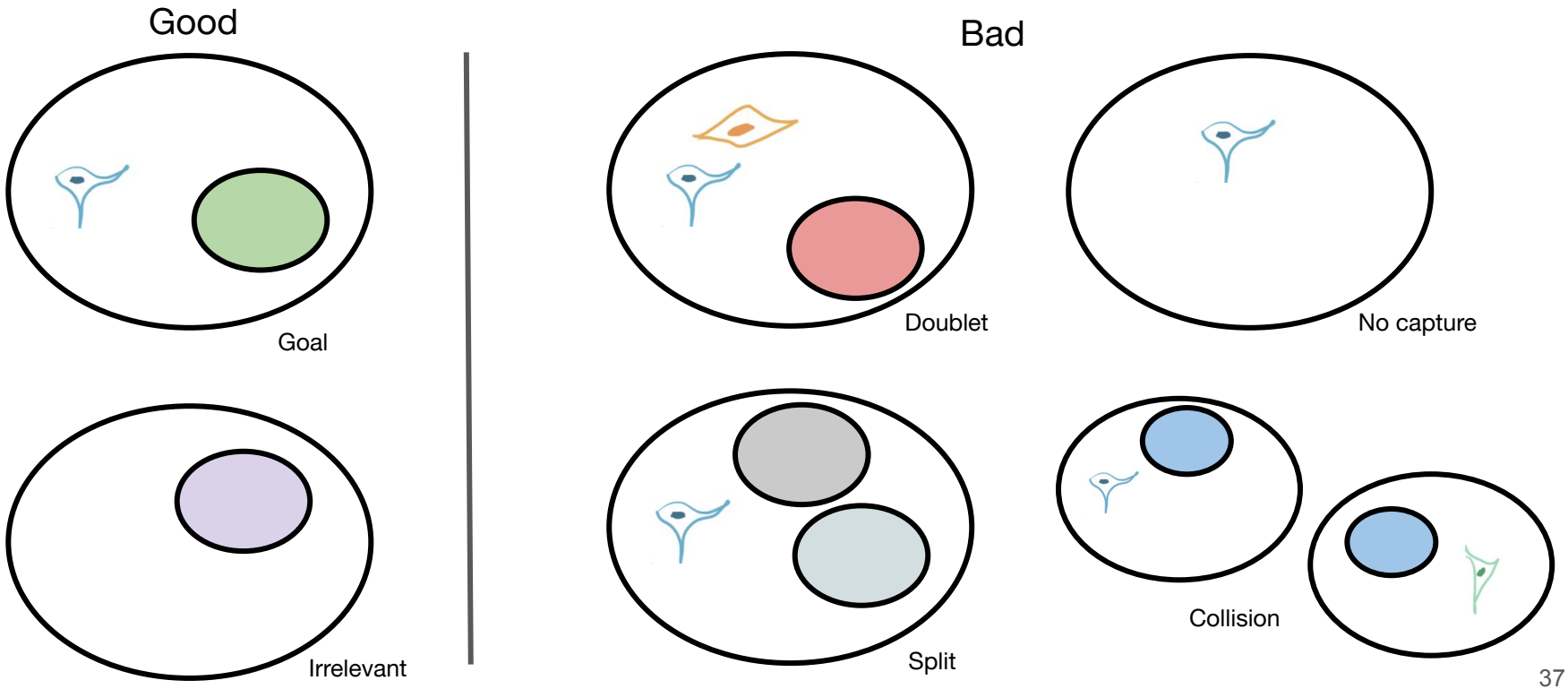
Zilionis et al. 2016

Example: the inDrops protocol



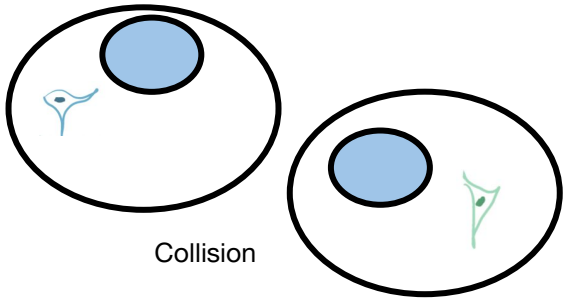
Library preparation [Zilionis et al. 2016]

Beads, Cells and Droplets



Barcode diversity

- ***Barcode collisions*** occur when beads with identical barcode sequences are present in droplets with two different cells.



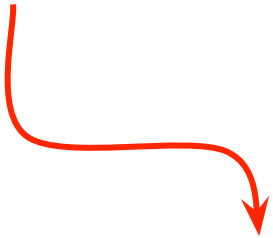
- The number of available barcode sequences depends on the sequence length *L*. Sequences of length *L* can yield up to 4^L barcodes.

- The number of distinct barcodes needed is a function of the number of cells that are to be barcoded.

Estimating the number of cells that will be uniquely barcoded

- Assuming that each of *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

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



Expected value is a generalization of weighted average, and is a number associated to a *random variable*.

Random variables

- A random variable is neither random nor a variable…
  - A random variable is a *function* from a sample space of a probability space, to the real numbers (or more generally a measurable space).

- Example: the sample space is the set { present, absent } (for instance the presence or absence of a specific barcode in a droplet), and the random variable assigns the real number 1 to “present” and the real number 0 to “absent”. This is an example of an *indicator* random variable.

Estimating the number of cells that will be uniquely barcoded

- Assuming that each of *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 of *k* cells is therefore

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

Estimating the number of cells that will be uniquely barcoded

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

Expected value of a random variable

- The expectation of a discrete random variable with outcomes x_1, x_2, x_3, … x_n each occuring with probability p_1, p_2, p_3, … p_n respectively is

$$\mathbb{E}[X] = \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k.$$

- If X_j is an indicator random variable corresponding to barcode *j,* i.e. X_j is 1 if barcode *j* is used for only one cell, and 0 otherwise, then

$$\mathbb{E}[X_j] = \frac{N}{M}\left(1 - \frac{1}{M}\right)^{N-1}.$$

The seemingly magic linearity of expectation

- If *X* and *Y* are two random variables, then

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

This is true ***regardless of whether X and Y are independent***.

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

Barcode collisions

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

$$1 - \frac{\mathbb{E}(\text{cells with a unique barcode})}{\text{number of cells}}$$

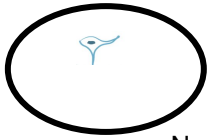
$$= 1 - \left(1 - \frac{1}{M}\right)^{N-1} \approx \quad 1 - \left(\frac{1}{e}\right)^{\frac{N}{M}}.$$

.

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

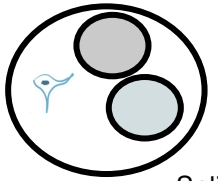
Droplet tuning concepts

- The ***capture rate*** is 1 - the fraction of cells that that are in droplets without any beads.



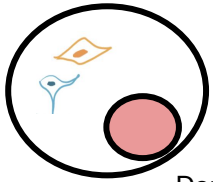
No capture

- The ***split rate*** is the fraction of droplets with exactly one cell that have more than one bead.



Split

- The ***doublet rate*** is the fraction of droplets with 1 bead that have more than one cell.



Doublet

Binomial distributions for beads and cells

- Consider *n* droplets, each of which has 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, $X \sim B(n,p)$.

The law of rare events

- For large *n* and small *p*, the Binomial distribution *B(n,p)* is approximated well with the Poisson distribution *Pois(λ = 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.

Expected value of a Poisson random variable

If *X* is a Poisson random variable, i.e. X ~ Pois(λ), then the expected value of *X* is given by

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!}$$

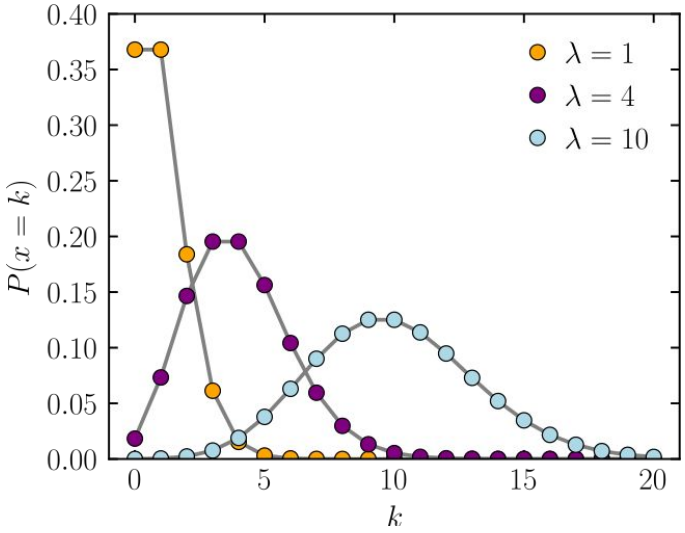
which is equal to λ.

A Poisson approximation for beads and cells

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

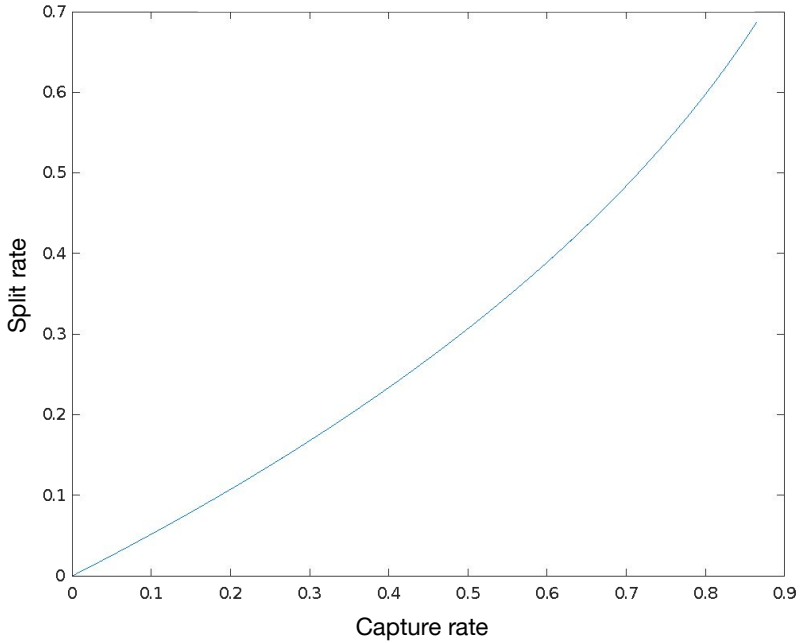
$$\mathbb{P}(\text{droplet has } k \text{ cells}) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

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



Cell capture and duplication rates

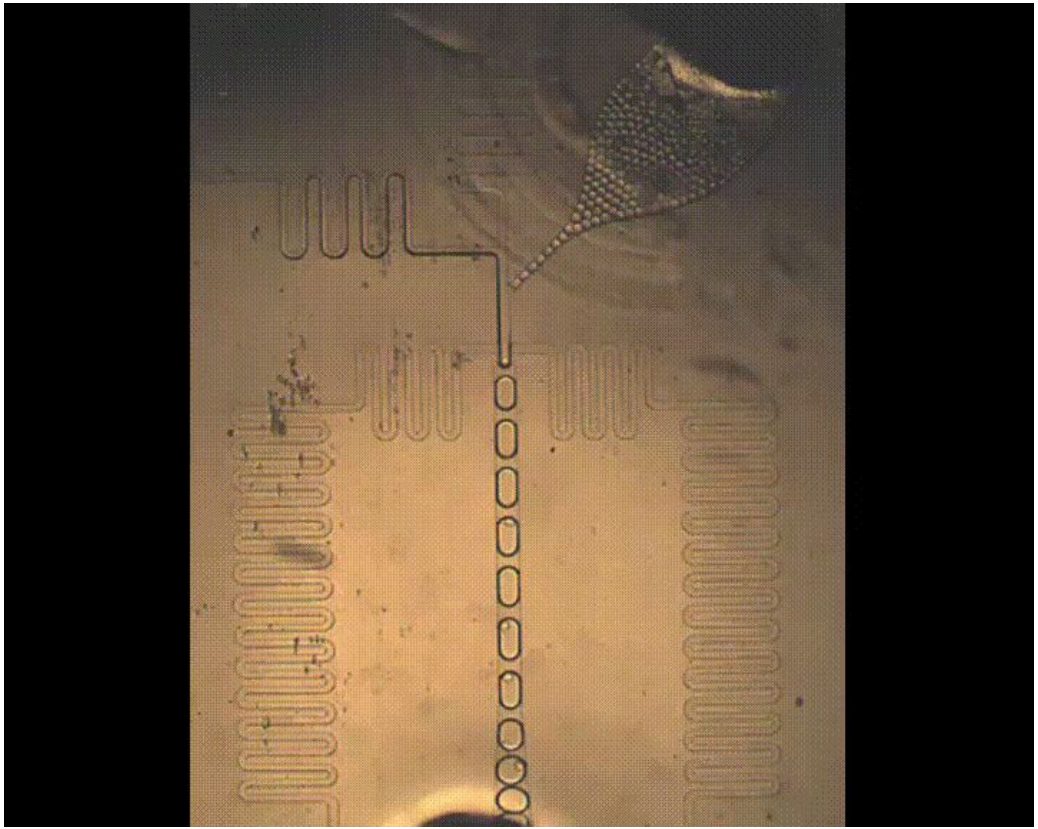
- The Poisson approximation yields a simple formula for the capture rate: $1 - e^{-\mu}$.
- The split rate estimate is $\frac{(1 - e^{-\mu} - \mu e^{-\mu})}{1 - e^{-\mu}}$
- This provides a quantitative assessment of the tradeoff between the capture rate and the split rate.



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

Sub-Poisson (sometimes called super-Poisson) loading



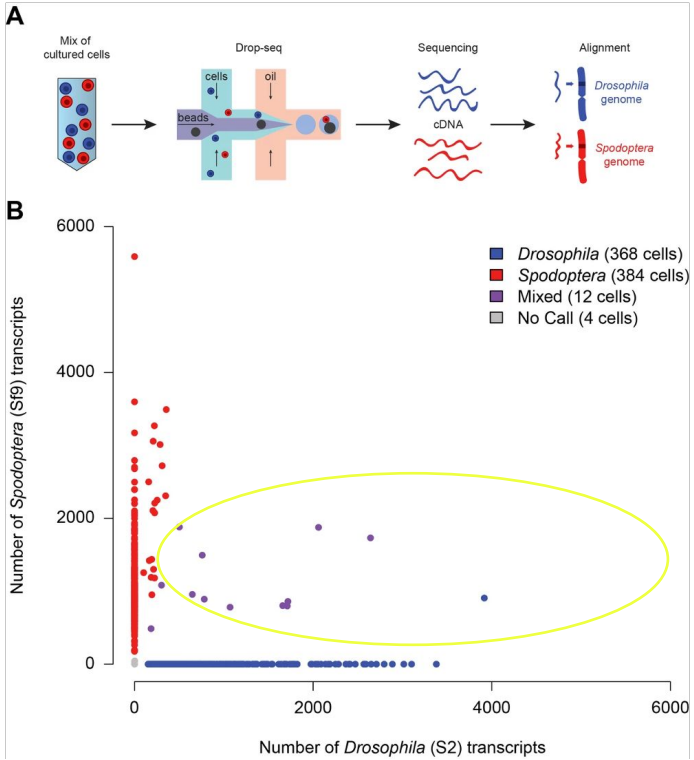
Technical doublets

- **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 microfluidics 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.

Doublet detection: the barnyard plot



Croset et al. 2018.

Bloom's correction

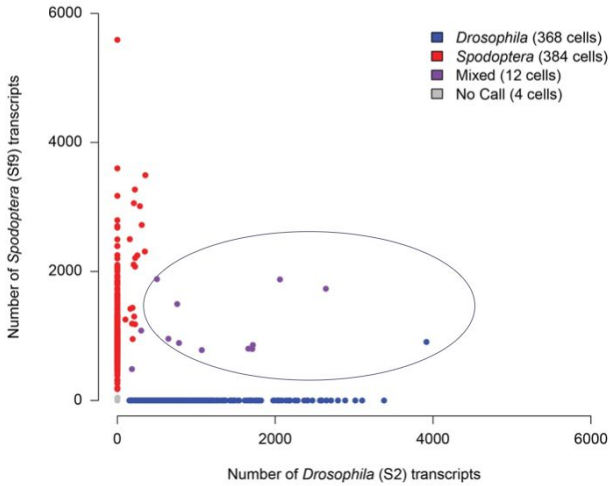
- Total number of droplets N, barnyard axes measured at N_1 and N_2, and observed doublets $N_{1,2}$.

$$N = \frac{N_1 N_2}{N_{1,2}},$$

$$\mu_1 = -\ln\left(\frac{N - N_1}{N}\right),$$

$$\mu_2 = -\ln\left(\frac{N - N_2}{N}\right).$$

$$\begin{aligned} M &= \frac{\Pr(c \geq 2)}{\Pr(c \geq 1)} \\ &= 1 - \frac{(\mu_1 + \mu_2)\, e^{-\mu_1 + \mu_2}}{1 - e^{-\mu_1 - \mu_2}}. \end{aligned}$$

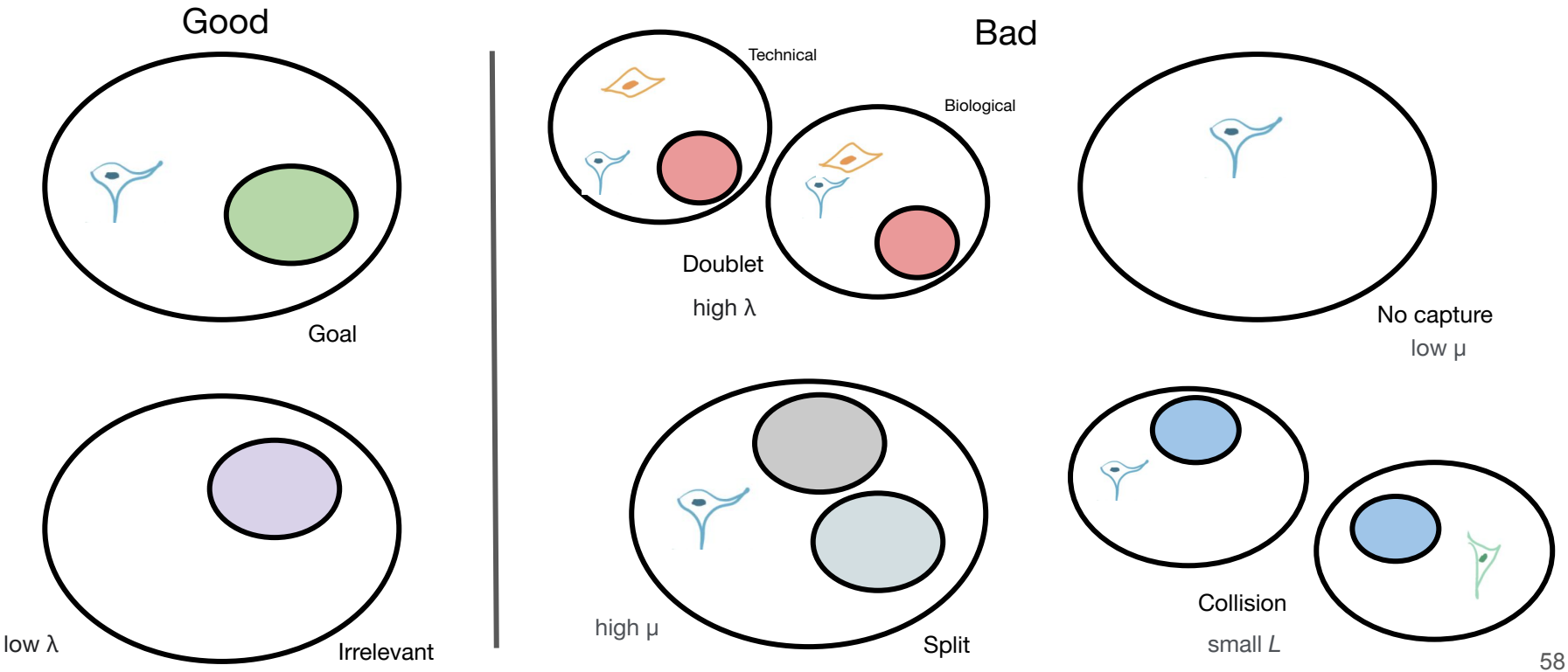


Bloom, 2018.

Biological doublets

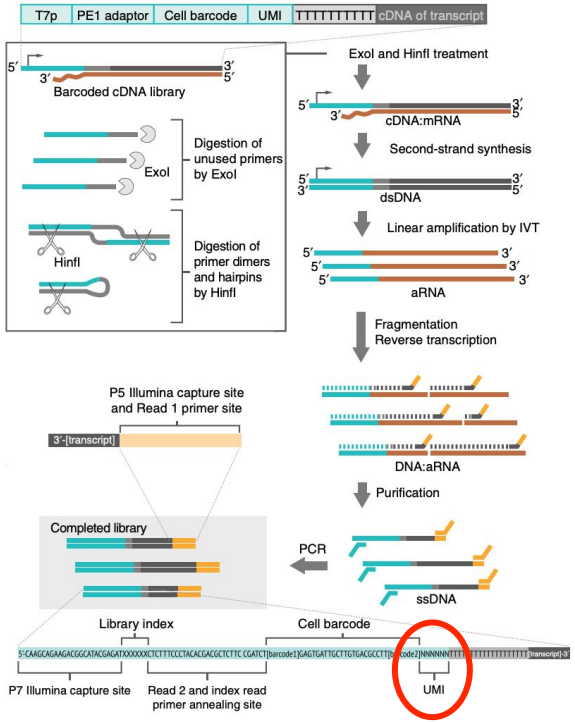
- **Biological doublets** arise when two cells form a discrete unit that does not break apart during disruption to form a suspension.
- Biological doublets will not be detected via barnyard plots.
- One approach to avoiding biological doublets is to perform nuclear single-cell RNA-seq: Habib et al. 2017.
- 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 endothelial cells via imputation from their hepatocyte partners.

Summary


Good
Goal
low λ
Irrelevant
Bad
Technical
Biological
Doublet
high λ
No capture
low μ
high μ
Split
Collision
small L

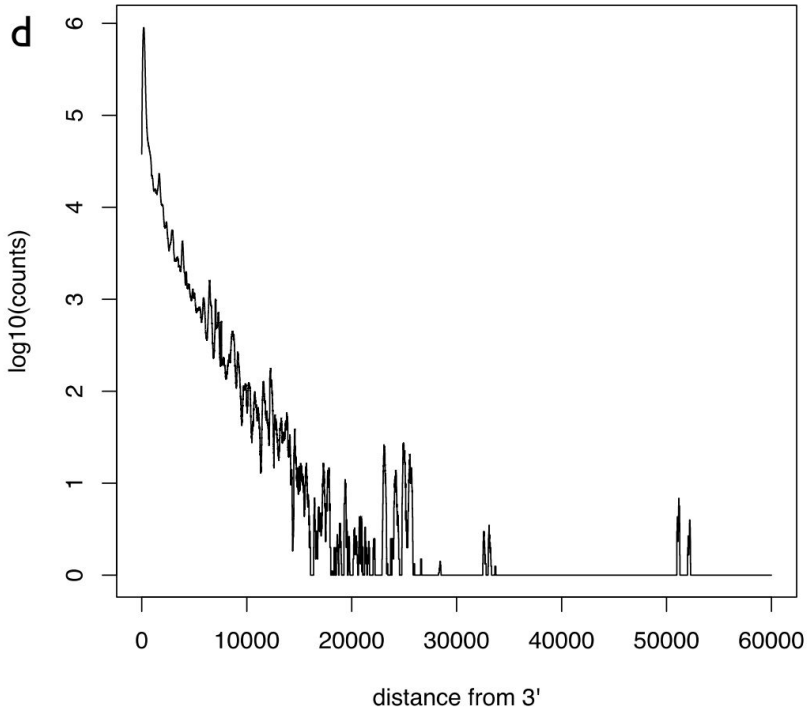
Unique **M**olecular **I**dentifiers

- Length determined by the diversity calculation and collision rate (same as barcodes).
- *UMI collapsing* refers to the process of using UMIs to avoid double-counting molecules after sequencing.
  - Naïve 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.



Zilionis et al. 2016

What it means to be “3’ technology”

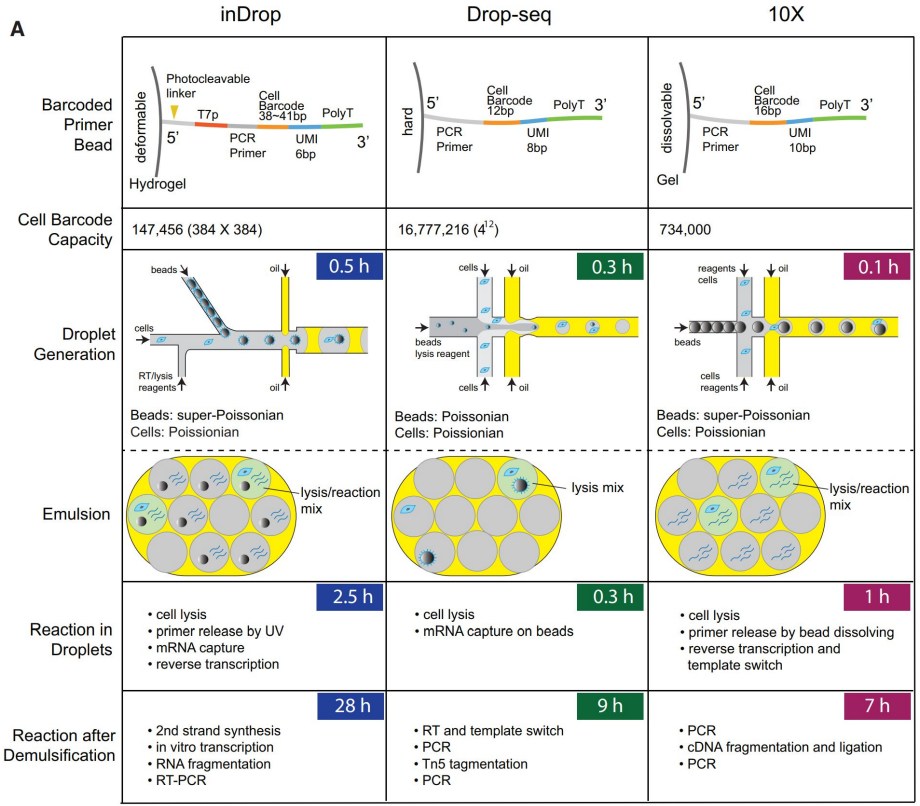


Ntranos *et al.* 2019.

Remarks on cell barcodes and UMIs

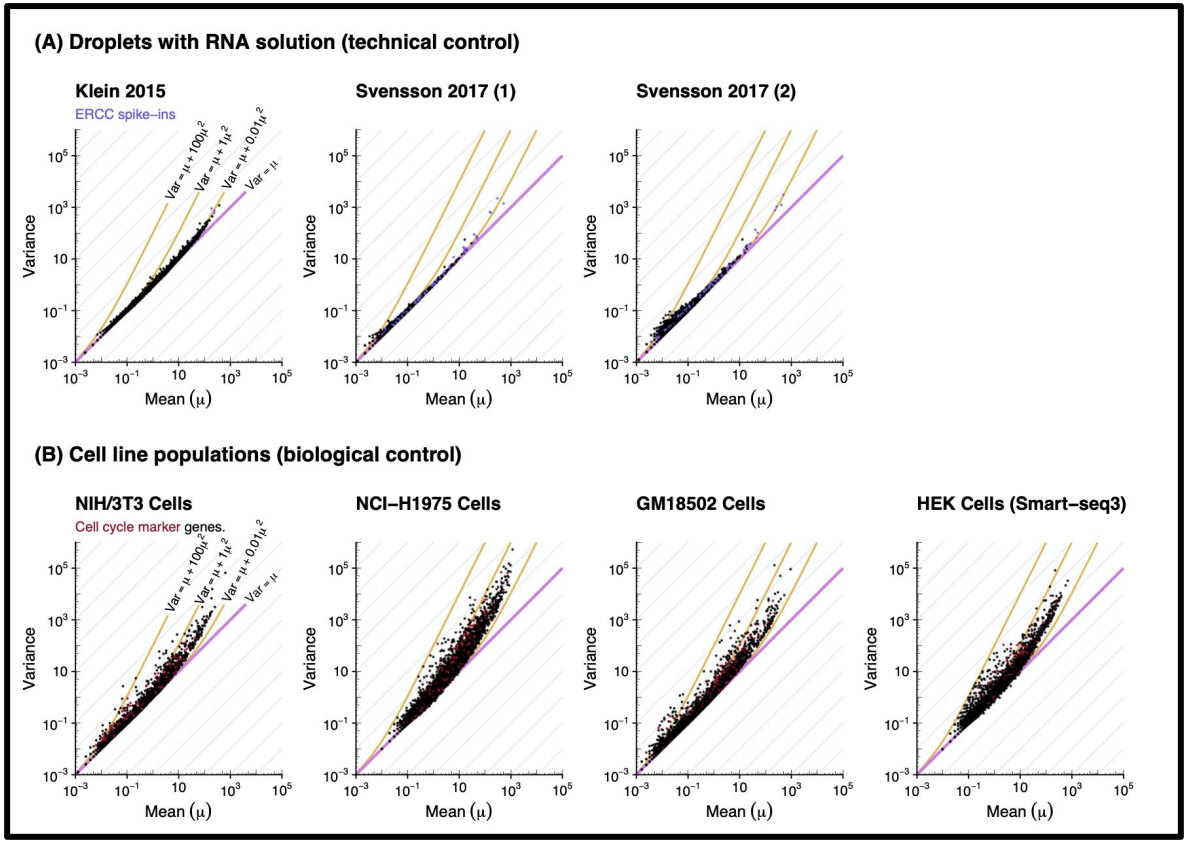
- Sequencing errors can lead to:
  - incorrectly labeled cells (from cell barcodes).
  - erroneous molecule counts (from UMIs).
- Error correction can be used to address these problems
  - Cell barcode error correction can sometimes be performed using a list of the known cell barcodes in the experiment (technology dependent).
  - 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



Zhang et al. 2019.

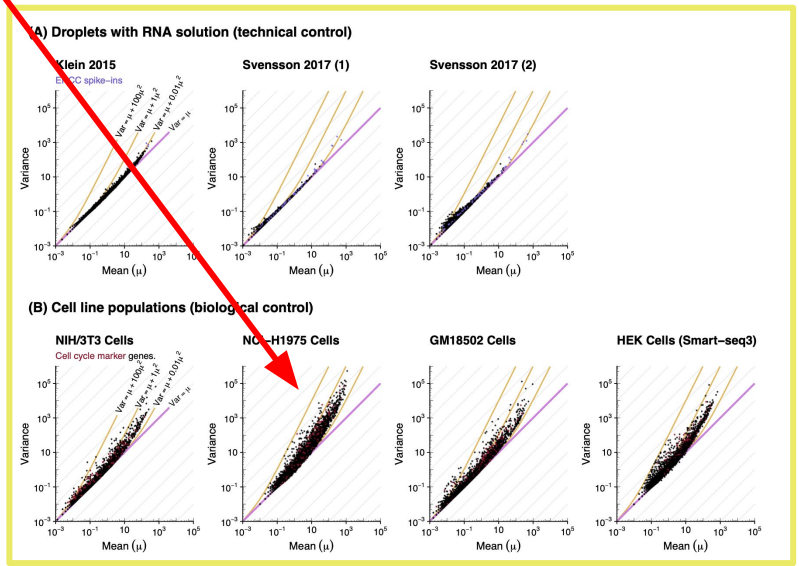
Variation in gene expression measurements



Ahlmann-Eltze and Huber, 2021

The variance is quadratic in the mean

- The negative binomial distribution yields a variance that is quadratic in the mean, namely $V = \mu + \phi \cdot \mu^2$, where μ is the mean and Φ is a parameter called the *dispersion* or *shape* parameter.



A question

- Is the observed variability due to:
    - **technical noise?**
    - **biological stochasticity?**

What do you believe?

Quantitative Biology

A **stochastic model for** bursty **transcription** coupling the telegraph model to a naive RNA transcription model yields a steady state **negative binomial distribution** for molecule counts.

Computational Biology

An overdispersed sampling model for RNA molecule counts from single-cell RNA-seq due to **technical variation** yields a **negative binomial distribution** for molecule counts.

About Google Colab

- Launch a free (Google account required) computer in the cloud and use it to run Jupyter or R notebooks:
- Technical Specifications
- Welcome to Colaboratory
- Overview of Colaboratory Features
- Markdown Guide