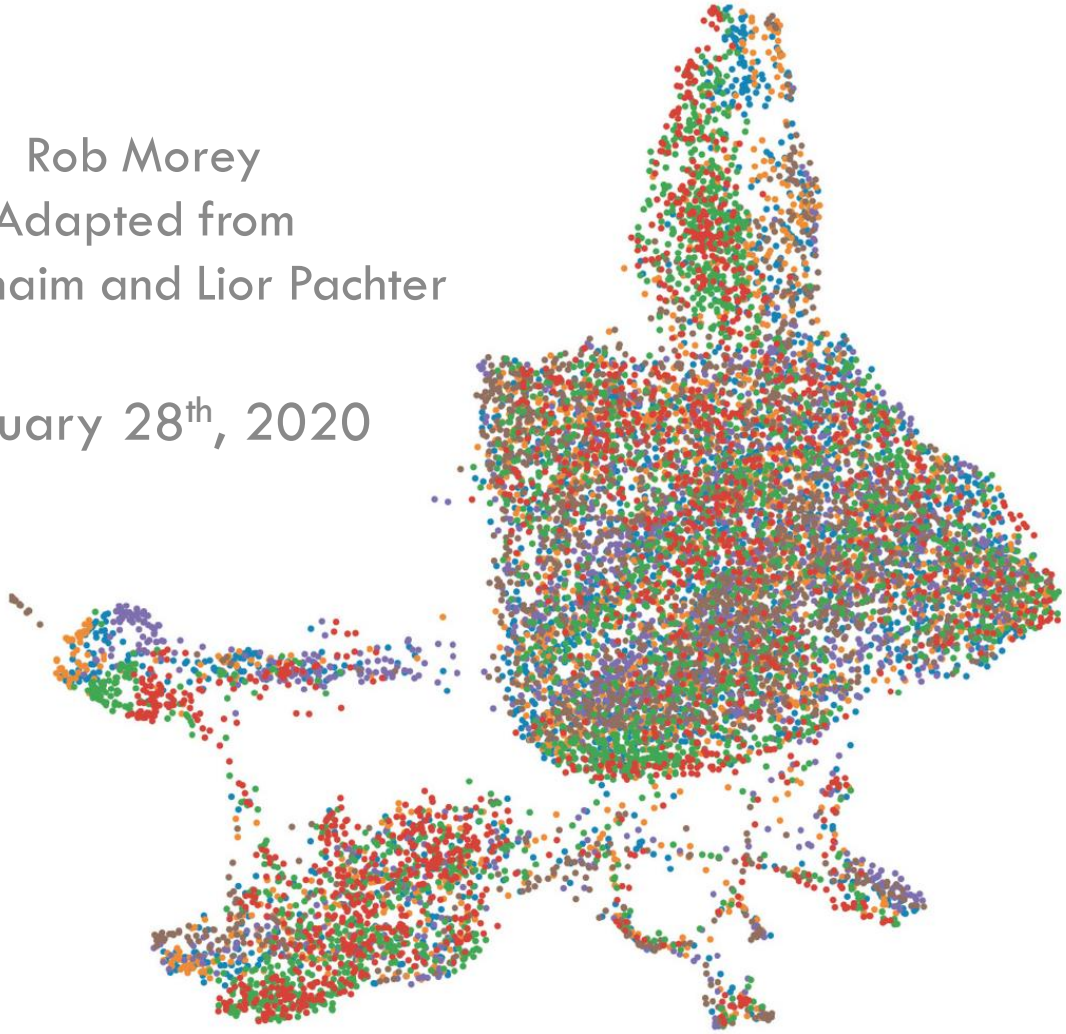
Quantitative Methods in Genetic and Genomics

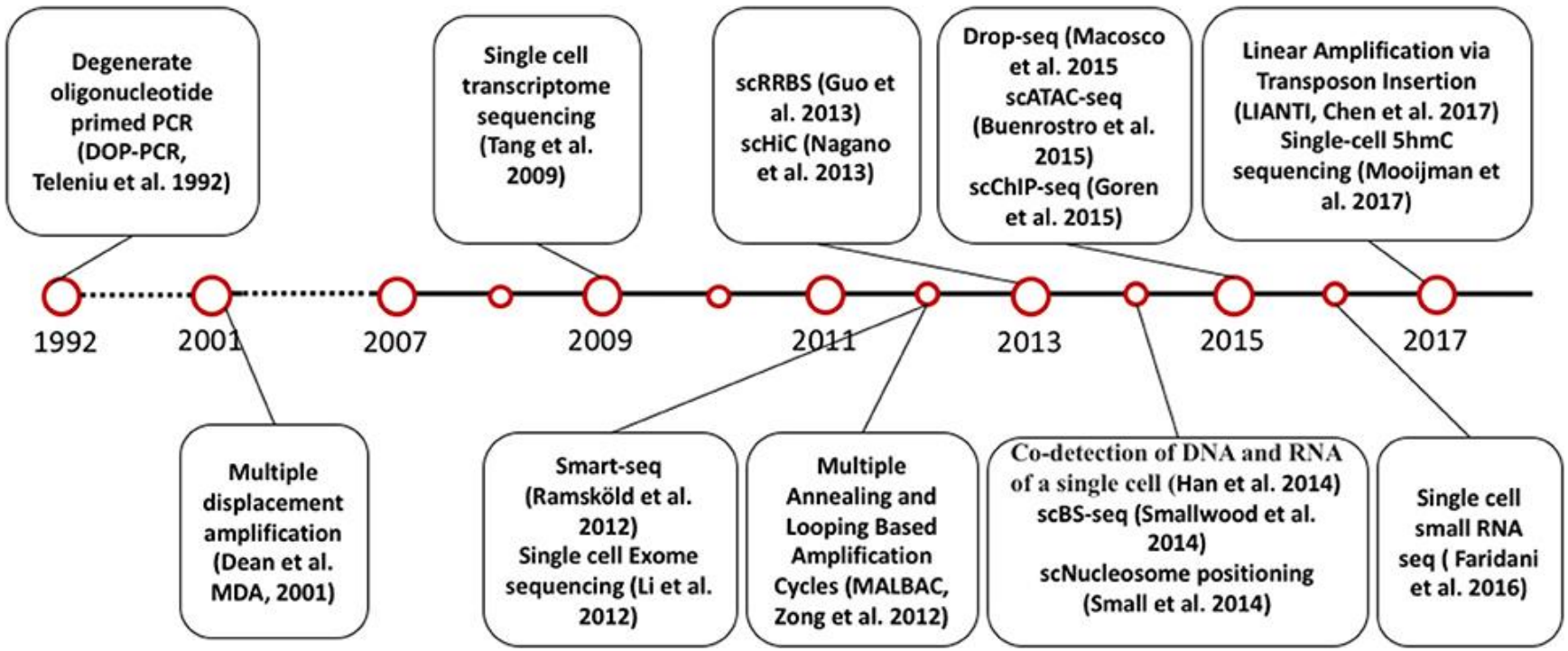
Single Cell Module

Rob Morey
Adapted from
Alex Chaim and Lior Pachter

January 28th, 2020

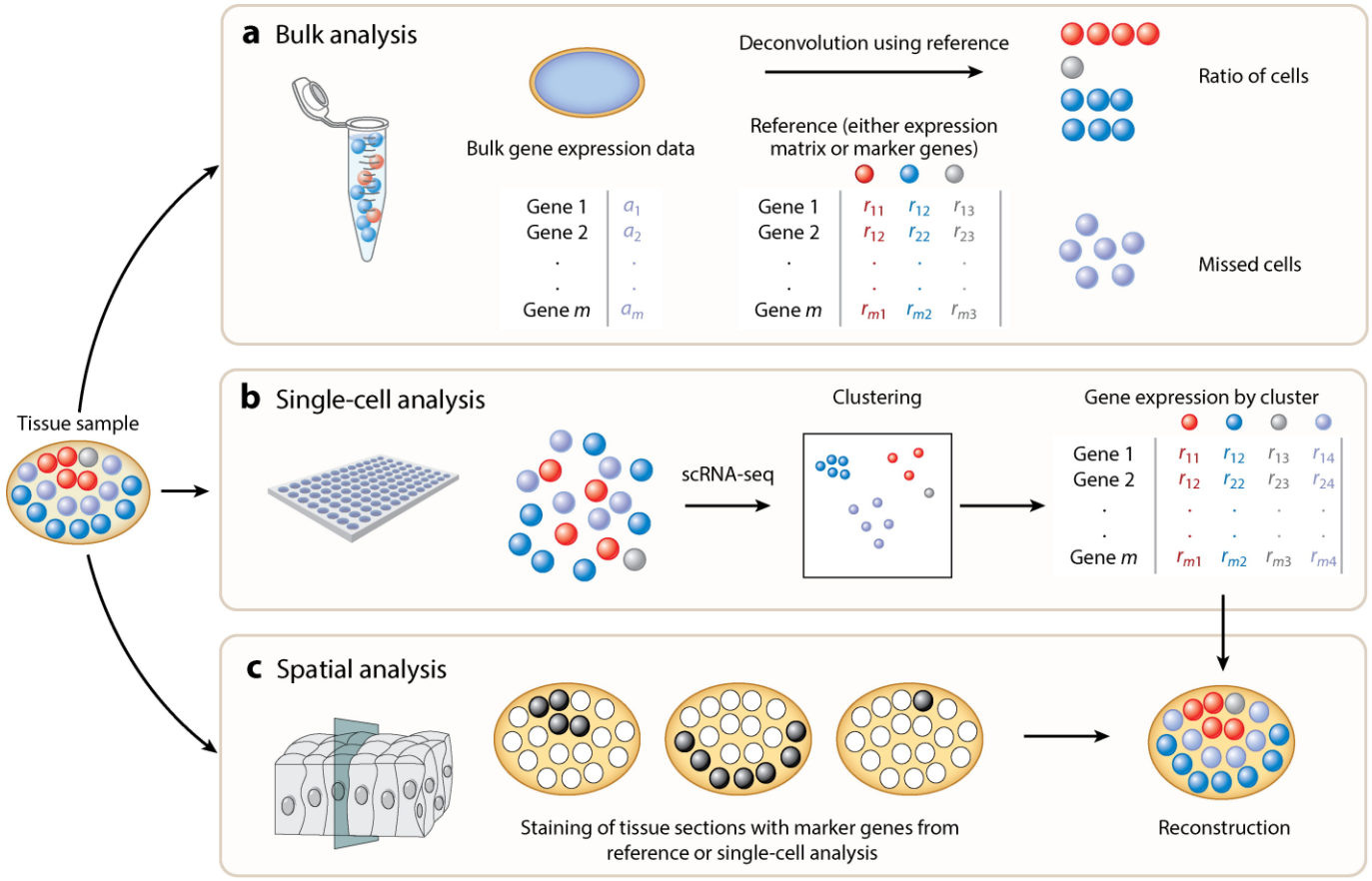


Single Cell Genomics



Youjin et al. 2018. DOI=10.3389/fcell.2018.00028

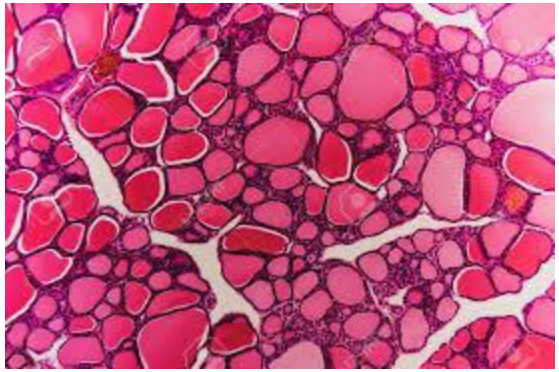
Why single cell transcriptomics?



Chen X, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:29–51

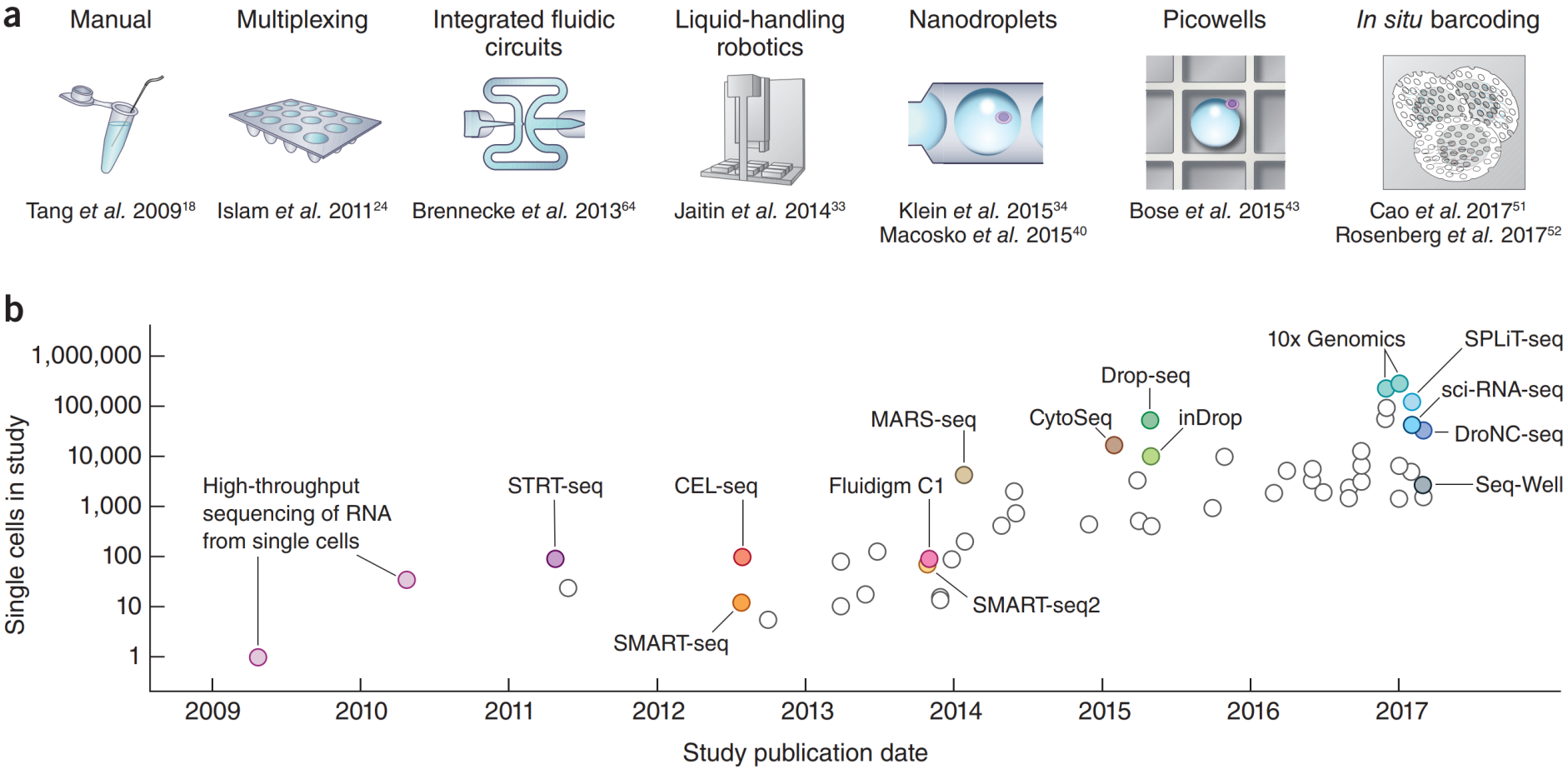
Microarrays or bulk RNA sequencing mask the true distribution of gene expression levels across cells.

How do we go from cell to count matrix?



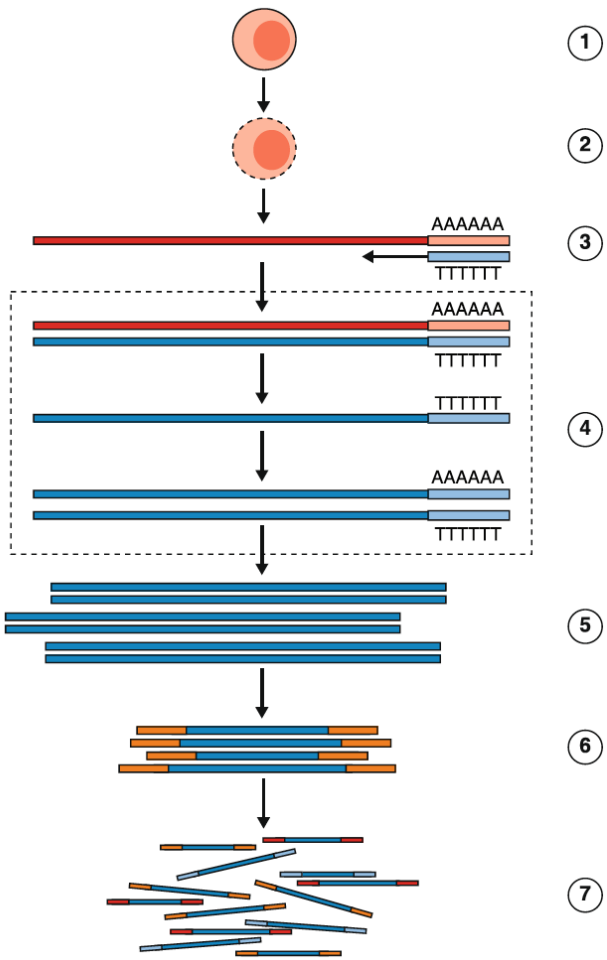
Cell:	1	2	...	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

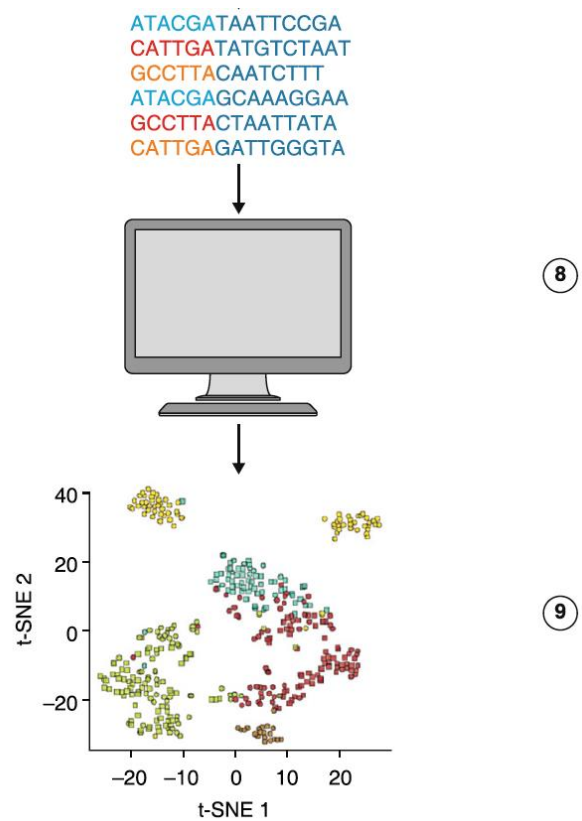
A decade of single-cell RNA-seq



Svensson et al. 2018. DOI:10.1038/nprot.2017.149

General workflow of a single cell (sc) RNA-Seq experiment

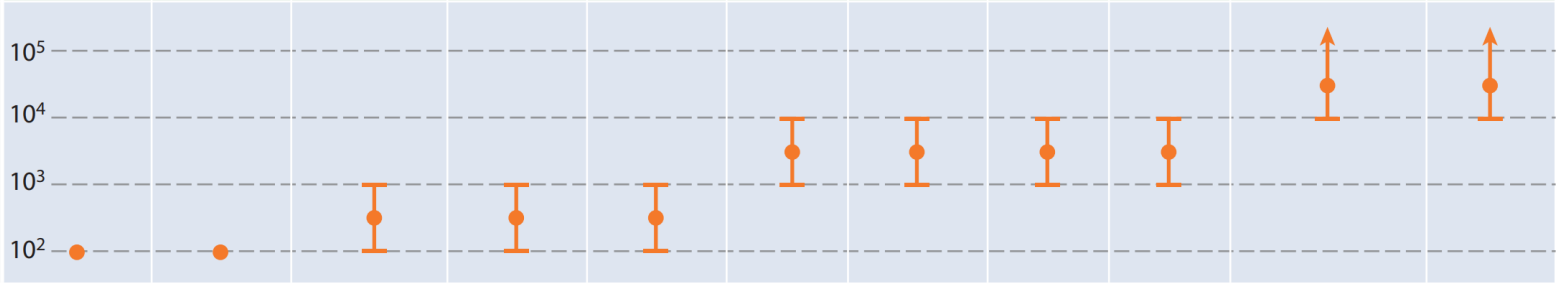




Haque *et al.*, 2017 Genome Medicine

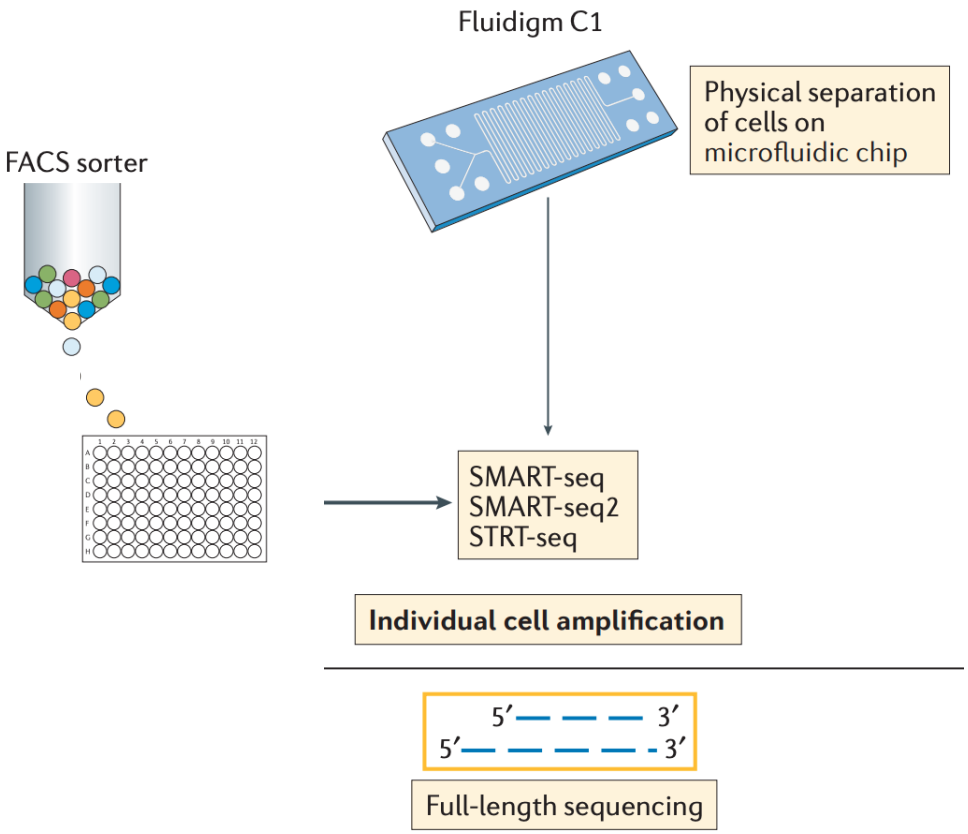
Popular single-cell RNA-seq protocols

	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											



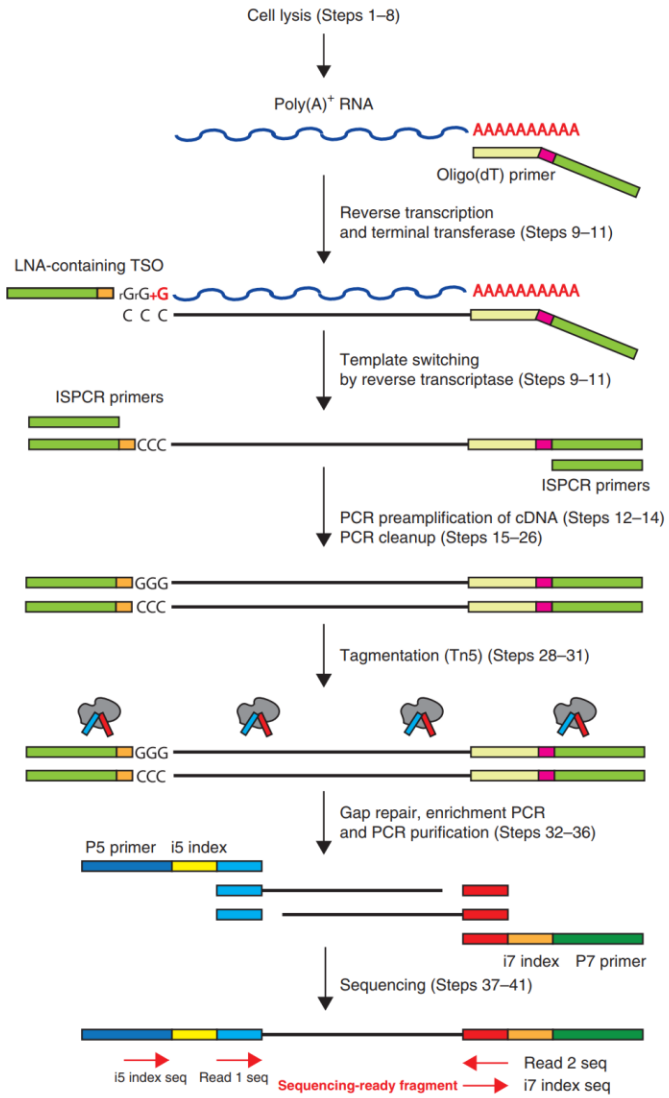
Chen et al. 2018. DOI:10.1146/annurev-biodatasci-080917-013452

Physical separation of cells in wells



Papalexi and Satija, 2017. DOI:10.1038/nri.2017.76

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. DOI:10.1038/nri.2017.76

Single cell technologies (3'-end vs full-length)

Table 1. Cost Efficiency Extrapolation for Single-Cell RNA-Seq Experiments

Method	TPR[a]	FDR[a] (%)	Cell per Group[b]	Library Cost ($)	Minimal Cost[c] ($)
CEL-seq2/C1	0.8	~6.1	86/100/110	~9	~2,420/2,310/2,250
Drop-seq	0.8	~8.4	99/135/254	~0.1	~1,010/700/690
MARS-seq	0.8	~7.3	110/135/160	~1.3	~1,380/1,030/820
SCRB-seq	0.8	~6.1	64/90/166	~2	~900/810/1,080
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

See also Figure 6.

[a]True positive rate and false discovery rate are based on simulations (Figure 6; Figure S9).

[b]Sequencing depth of one, 0.5, and 0.25 million reads.

[c]Assuming $5 per one million reads.

Ziegenhain *et al.*, 2017 Molecular Cell

Table 1. Widely Used Single-Cell Sequencing Methods

Sequencing Method	Starting Cell No.	Cell Separation	Notes	Cell Capture	Transcript Capture	Representative Library Prep Cost per Cell[a]
Fluidigm C1[b]	~1,000 cells	cells capture in size-specific chambers	must know the size of cells of interest; allows for staining and imaging prior to cell rupture	96- or 800-chamber units are available	an average of 6,606 genes/cell (no data on percentage)	$1.70
DropSeq	~150,000 cells/run	droplet-based separation	remains the most cost-effective and most customizable	~5% of cells per run (approximately 7,000 cells)	~10.7% of the cell's transcripts	$0.06
Chromium 10X	~1,700 cells/run	droplet-based separation	the most commercially successful method; almost fully automated	~65% of cells per run (approximately 1,000 cells)	~14% of the cell's transcripts	$0.10
SCI-Seq	~500,000 cells (depends on experimental design)	FACS sorter; cells are never singly isolated	combinatorial indexing of individual methanol-fixed permeable cells	5%–10% of cells	~10%–15% of the cell's transcripts	$0.05–$0.14[c]

All of the methods require the establishment of a cell dissociation technique. The price is highly dependent on the number of cells sequenced, the desired depth of sequencing, and the sequencing platform used. For this table, the prices are at the lower end of the price range for single-cell library prep.

[a]As of July 2018.

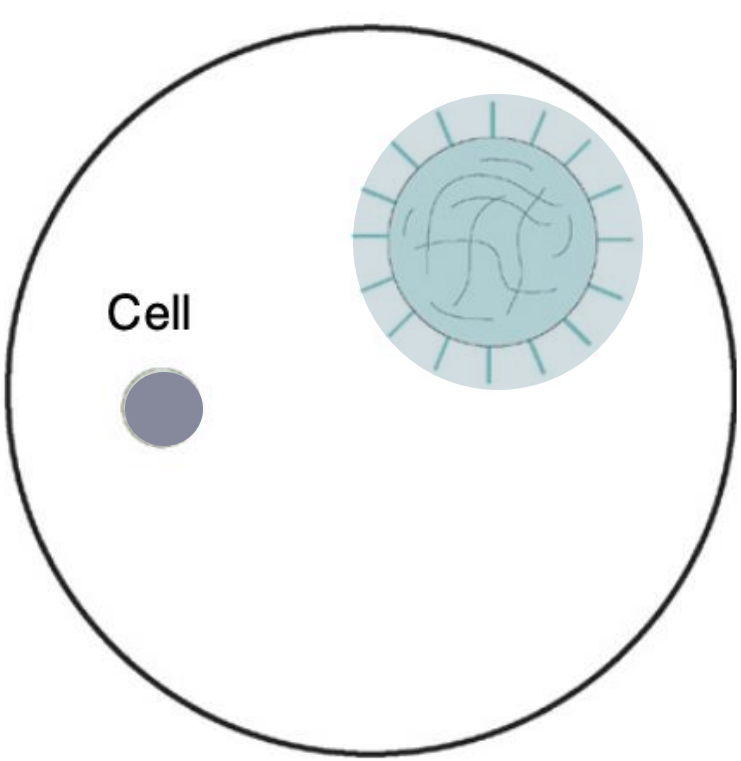
[b]Based on the 800-chamber medium-size isolation unit.

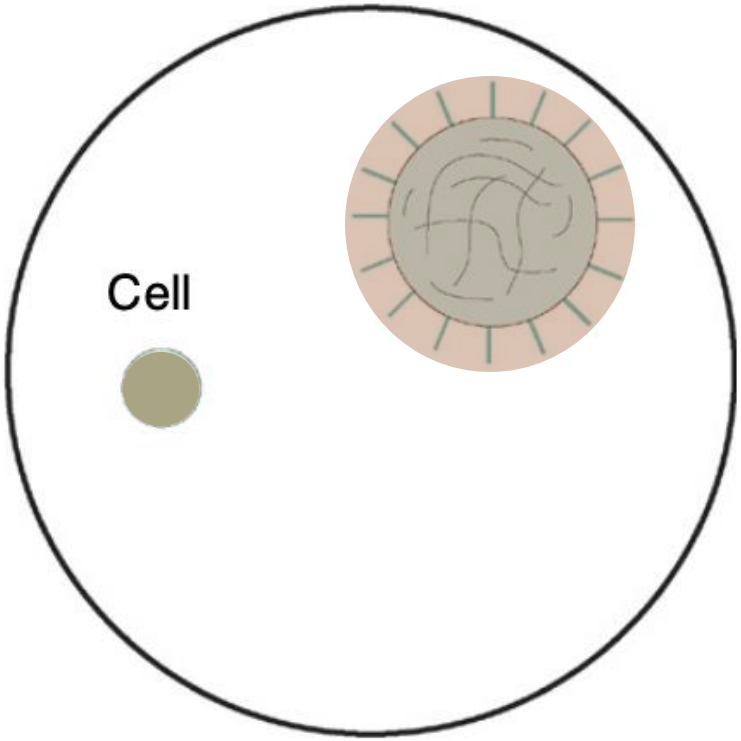
[c]Dependent on how many cells are prepped for sequencing and how many doublets are tolerated.

An Introduction to the Analysis of Single-Cell RNA-Sequencing Data

AlJanahi, Aisha A. et al.
Molecular Therapy - Methods & Clinical Development , Volume 10 , 189 - 196

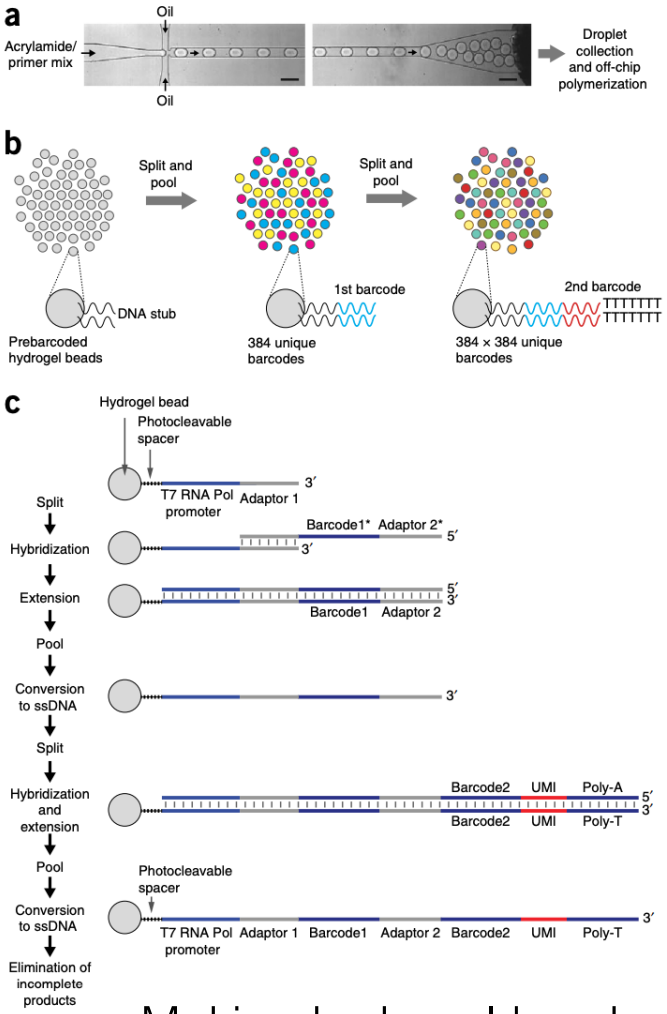
Microfluidic methods: Beads, Cells and Droplets



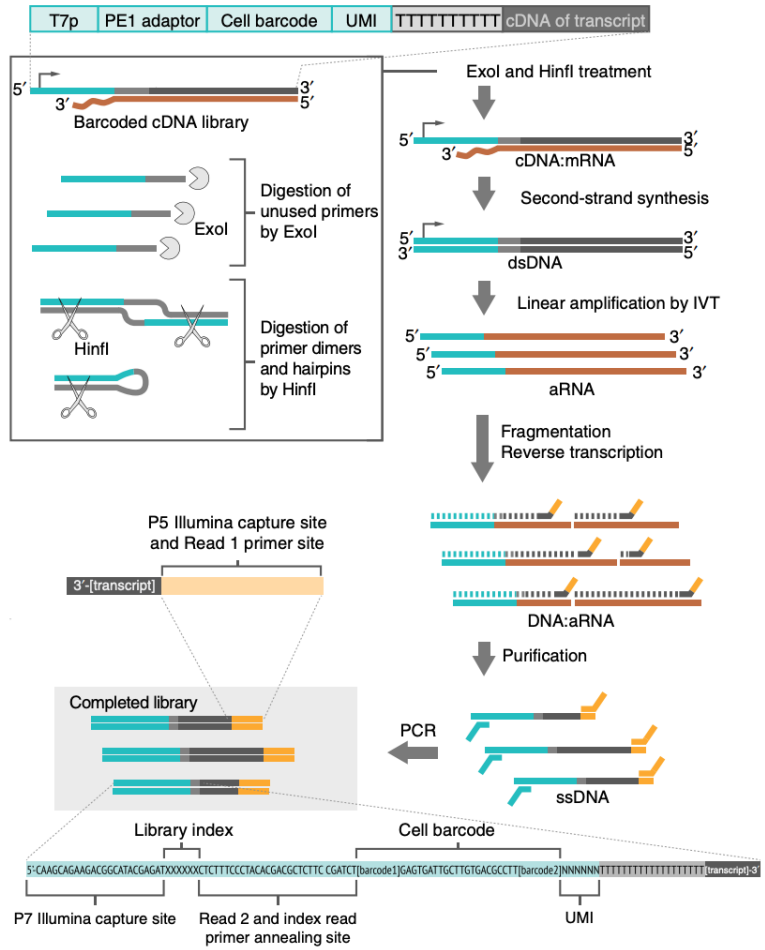


Zilionis et al. 2016. DOI:10.1038/nprot.2016.154

Example: the inDrops protocol



Making hydrogel beads

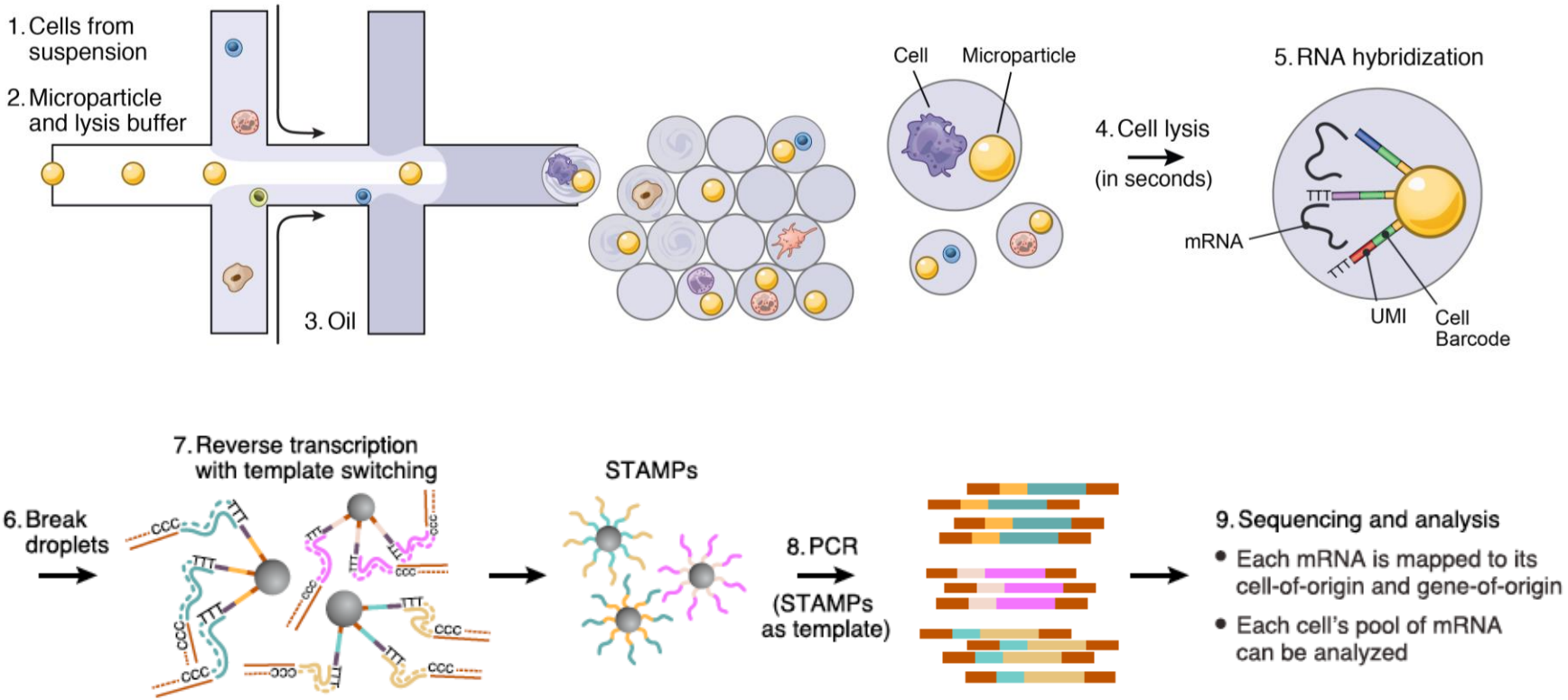


Library preparation

Zilionis et al. 2016, DOI:10.1038/nprot.2016.154

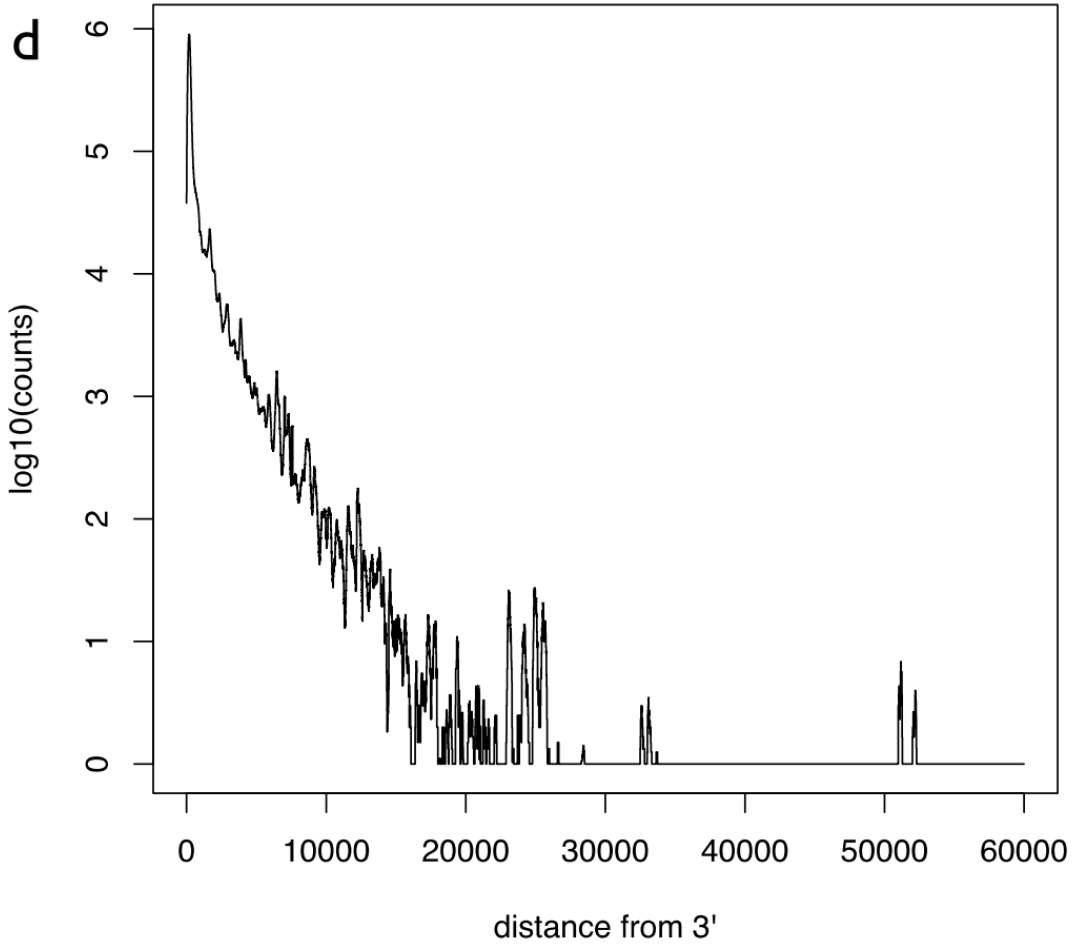
Drop-Seq

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets



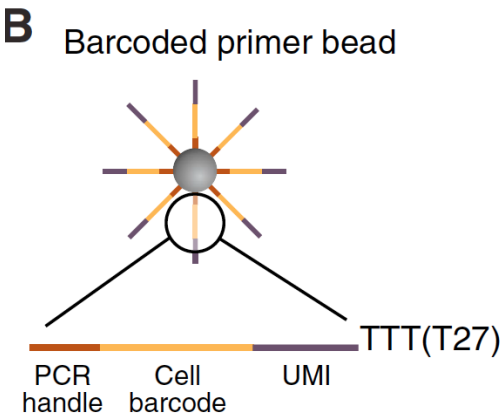
Macosko *et al.*, 2015 Cell

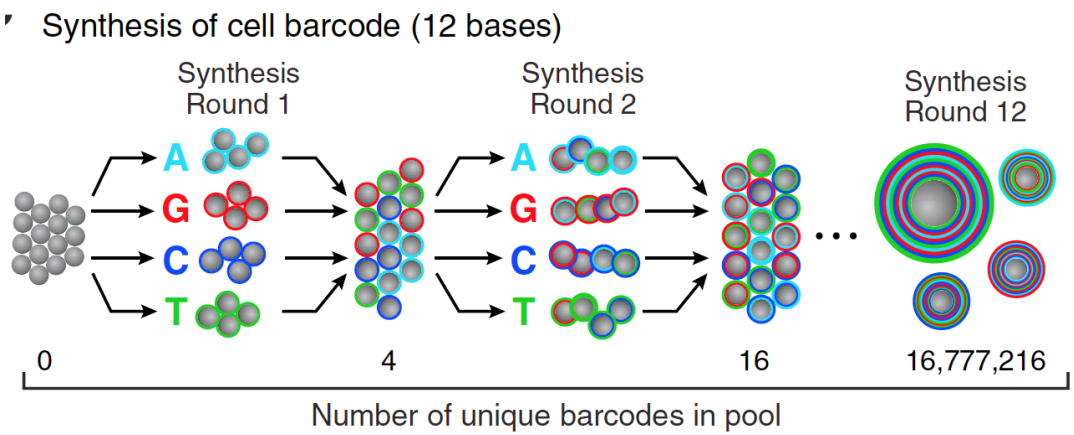
What it means to be “3’ technology”



Ntranos et al. 2019. DOI:10.1038/s41592-018-0303-9

3’ end Capture Beads - UMIs and Bead Barcodes





Synthesis of UMI (8 bases)

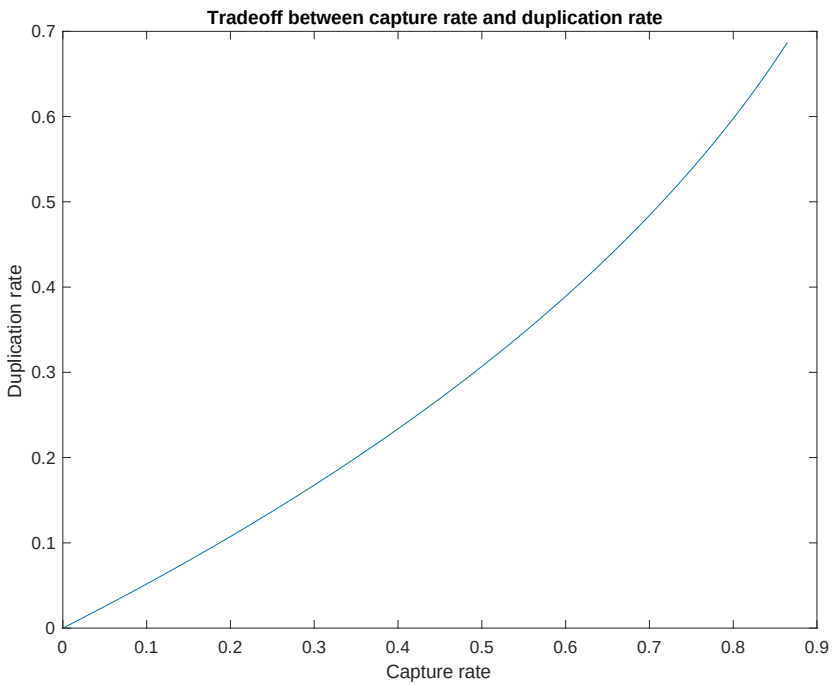
- Millions of the <u>same</u> **cell barcode** per bead
- 4^8 <u>different</u> **molecular barcodes** (UMIs) per bead

- $4^{12} = 16{,}777{,}216$ barcodes
- $4^8 = 65{,}536$ UMIs
- $> 10^6$ oligos/bead

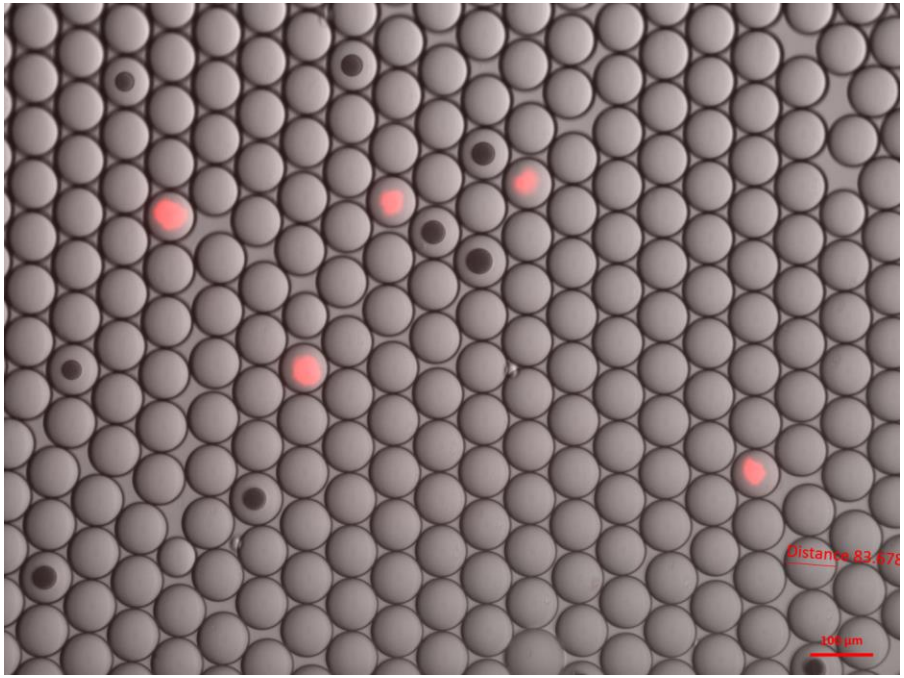
Macosko *et al.*, 2015 Cell

Cell capture and duplication rates

- The cell capture rate is the probability that a droplet has at least one bead = $1 - e^{-\mu}$
- The cell duplication rate is the rate at which captured single cells are associated with two or more different barcodes, which is equal to $\dfrac{(1 - e^{-\mu} - \mu e^{-\mu})}{1 - e^{-\mu}}$.
- This leads to a tradeoff:



Drop-Seq implementation



Capture rate < 10%

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

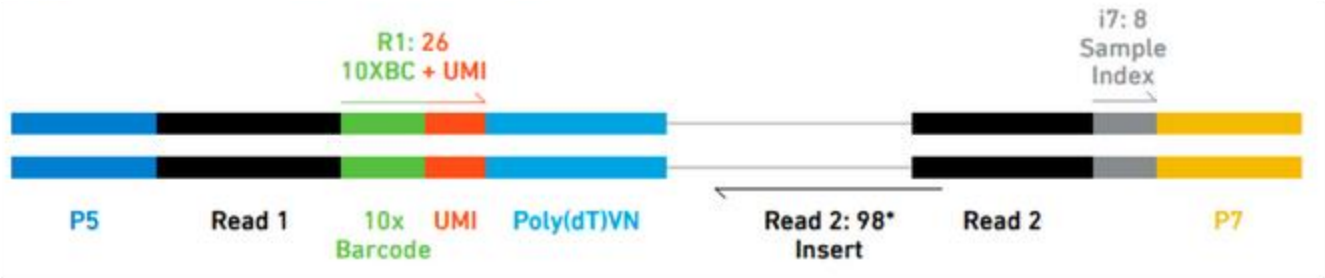
- Load **beads** into droplets at Poisson rate

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

- Load **cells** into droplets at Poisson rate

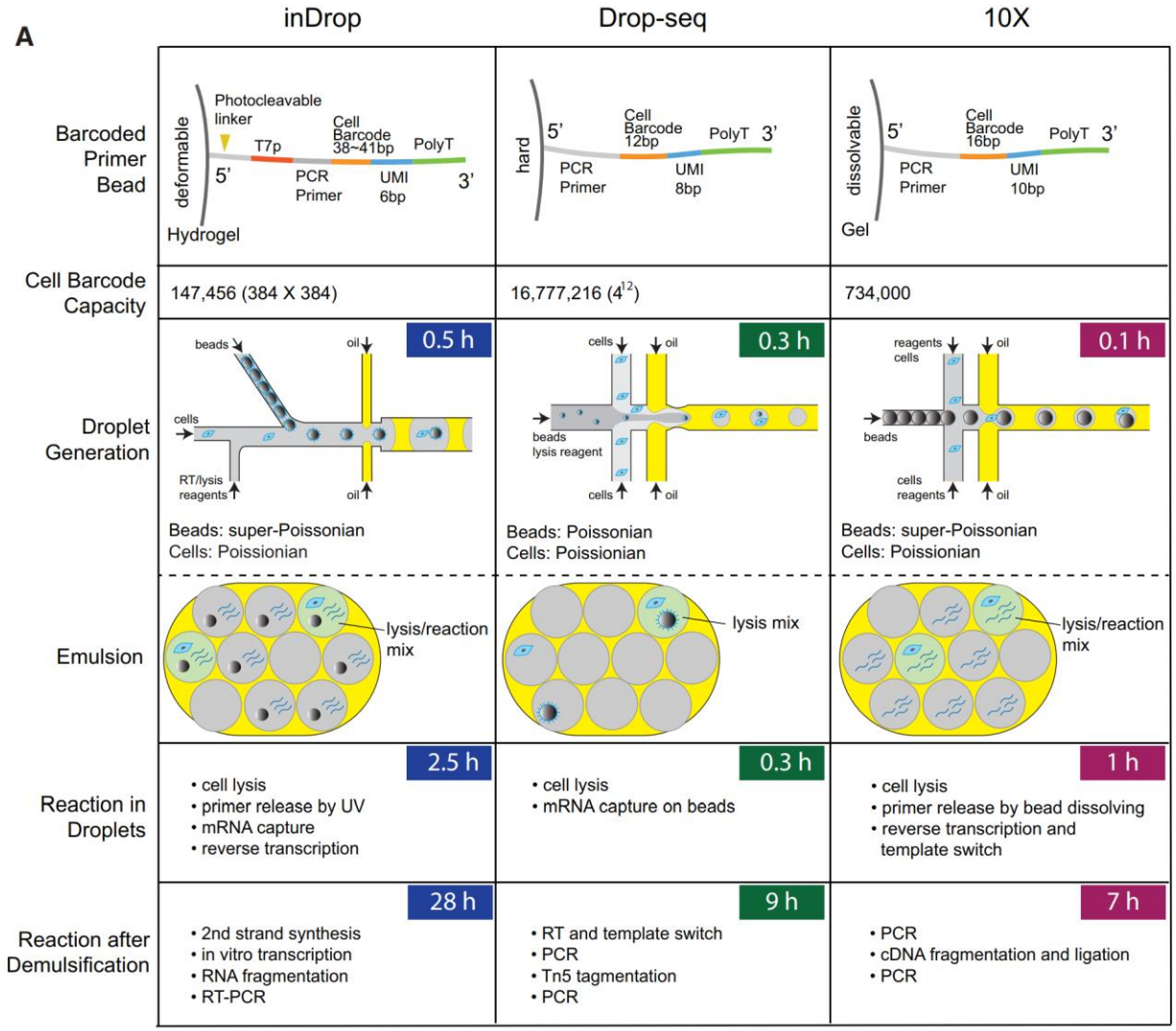
10X now offers 5’-end

Single Cell 3' v2 Gene Expression Library:



Single Cell 5' Gene Expression Library:

Summary of droplet technologies



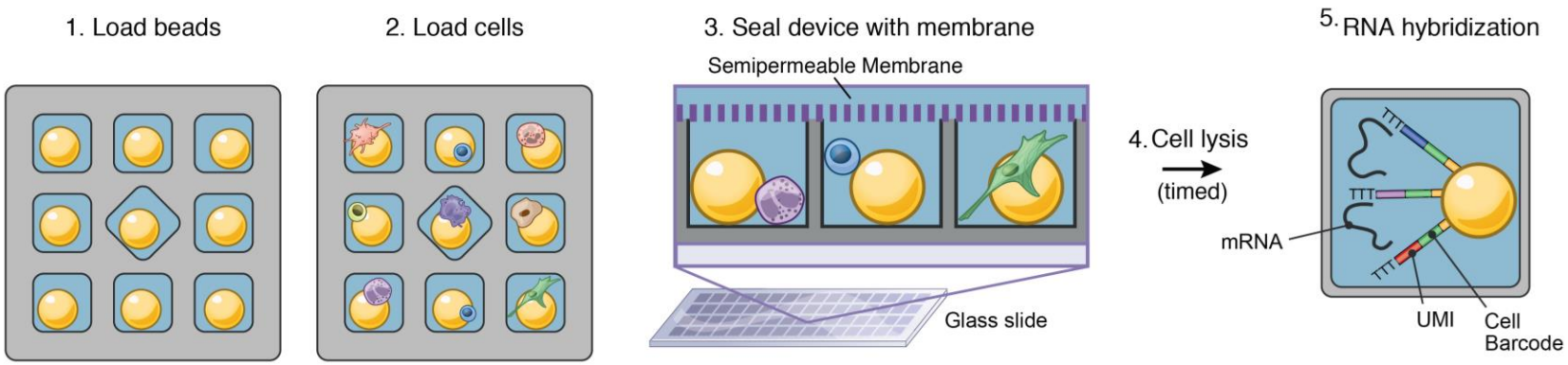
Zhang et al. 2019. DOI:10.1016/j.molce

Seq-Well abrogates the need for droplet-based microfluidic devices

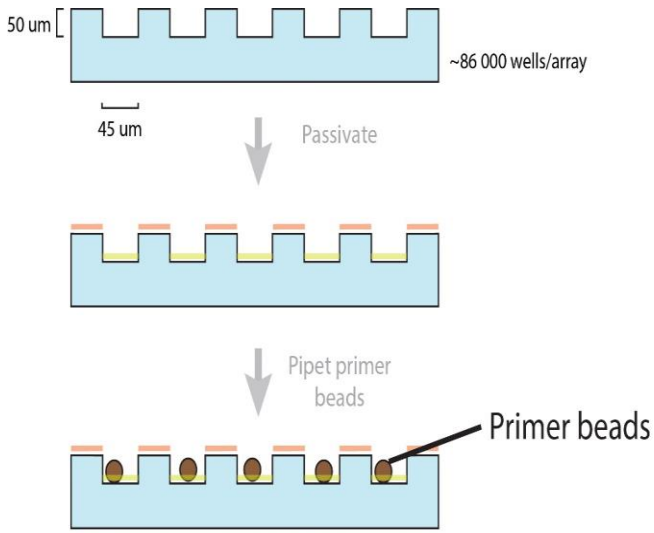
Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput

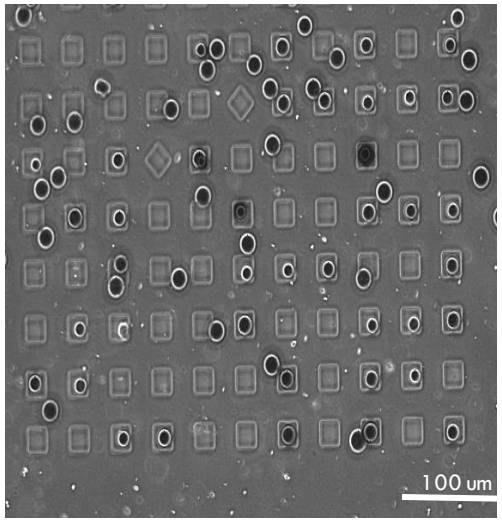
NATURE METHODS | VOL.14 NO.4 | APRIL 2017 |

Todd M Gierahn[1,8], Marc H Wadsworth II[2–4,8], Travis K Hughes[2–4,8], Bryan D Bryson[4,5], Andrew Butler[6,7], Rahul Satija[6,7], Sarah Fortune[4,5], J Christopher Love[1,3,4,9] & Alex K Shalek[2,3,4,9]

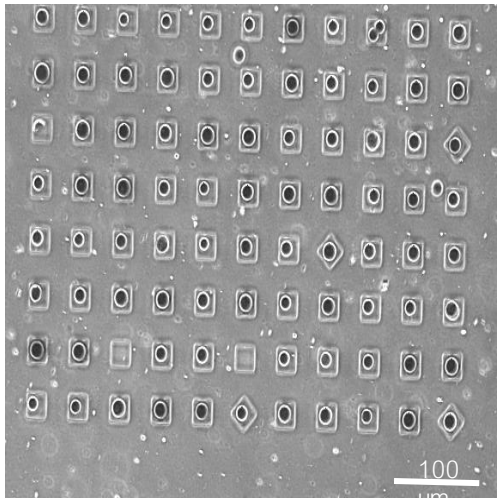


Seq-Well workflow in Images



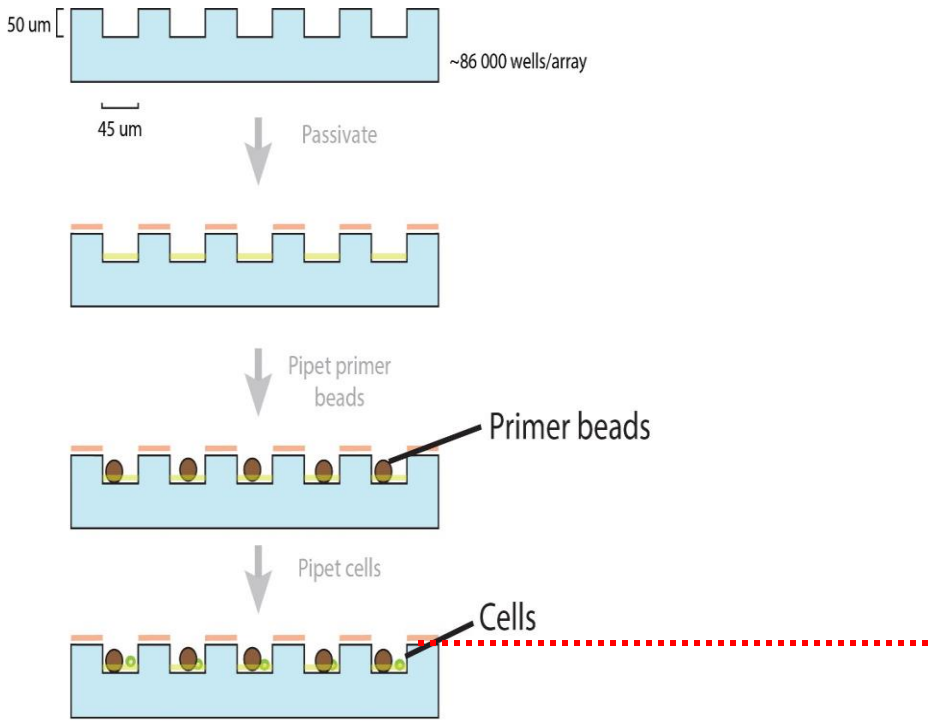


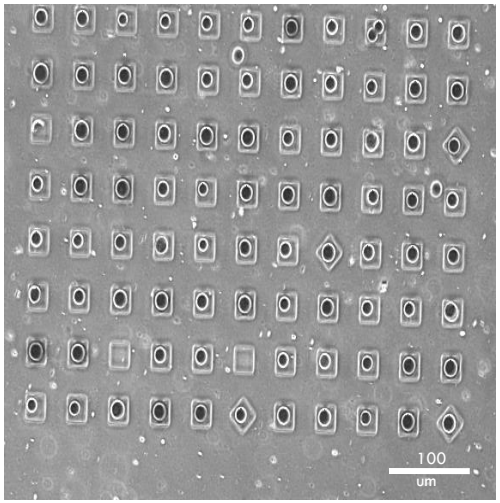
Pre-wash



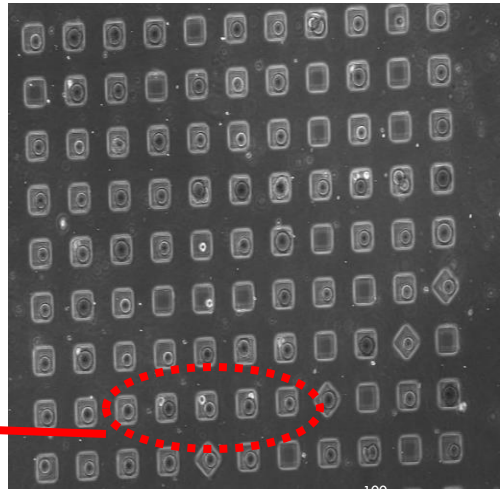
Post-wash + shake

Seq-Well workflow in Images

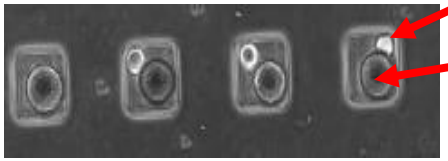




Pre-cell loading



Post-cell loading + shake



Cell

Bead

Capture rate > 50%

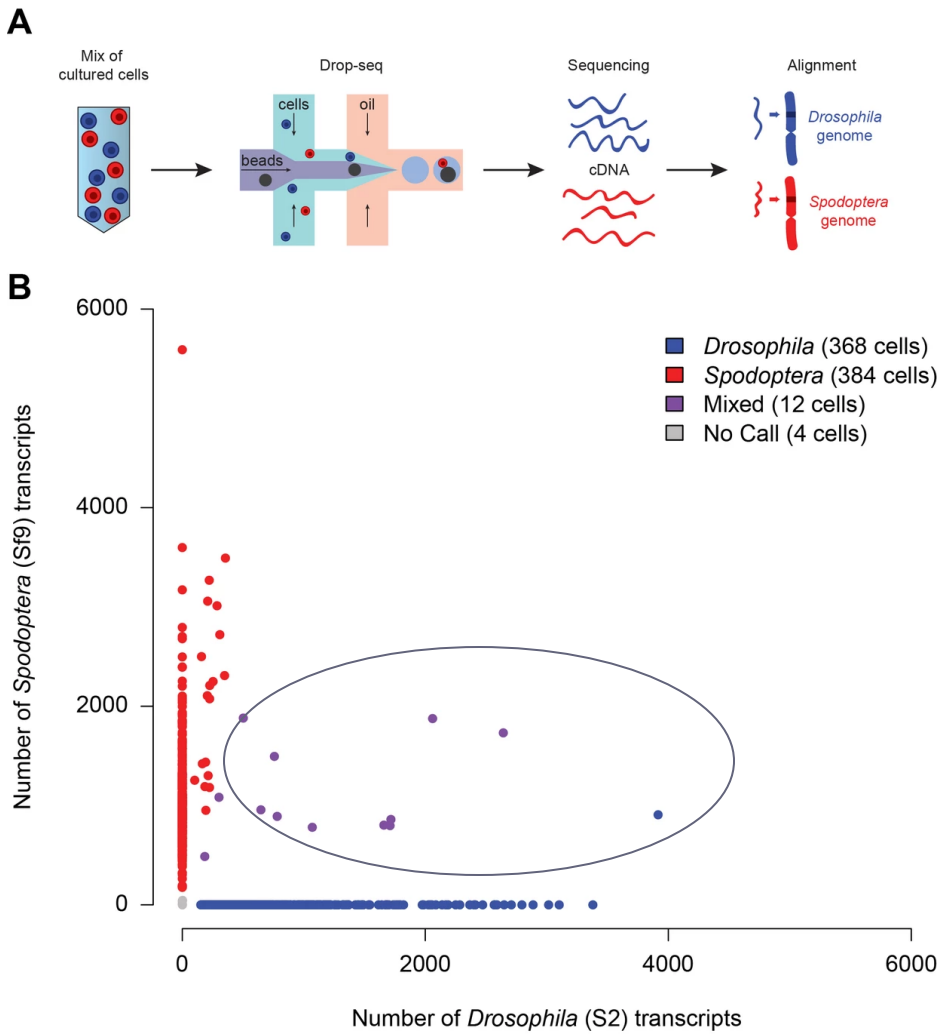
Technical doublets

- **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:

$$\frac{(1 - e^{-\lambda} - \lambda e^{-\lambda})}{1 - e^{-\lambda}}.$$

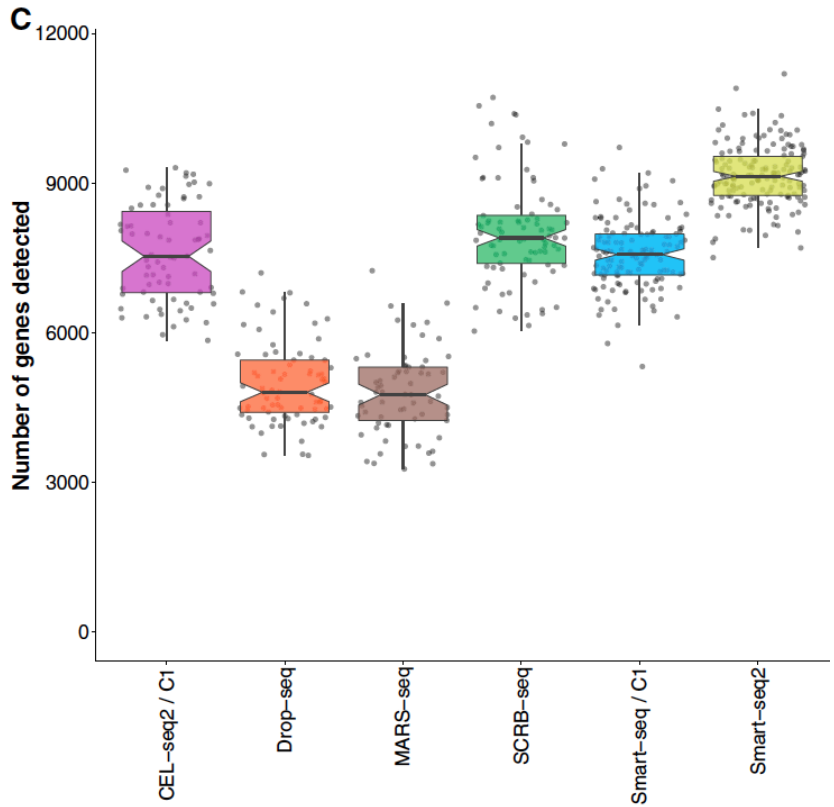
- Note that “overloading” a microfluidics single-cell experiment by loading more cells while keeping flow rates constant will increase the number of technical doublets due to an effective increase in λ and also the number of synthetic doublets due to a decrease in relative barcode diversity.

Doublet detection: the barnyard plot

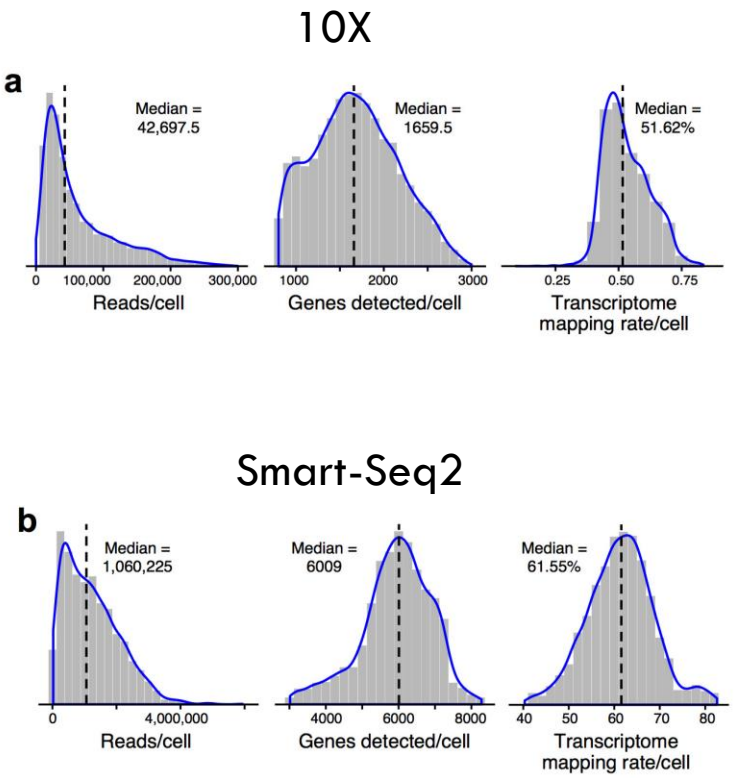


Croset et al. 2018. DOI:10.7554/eLife.34550

3’-end vs full-length Inter-platform comparison

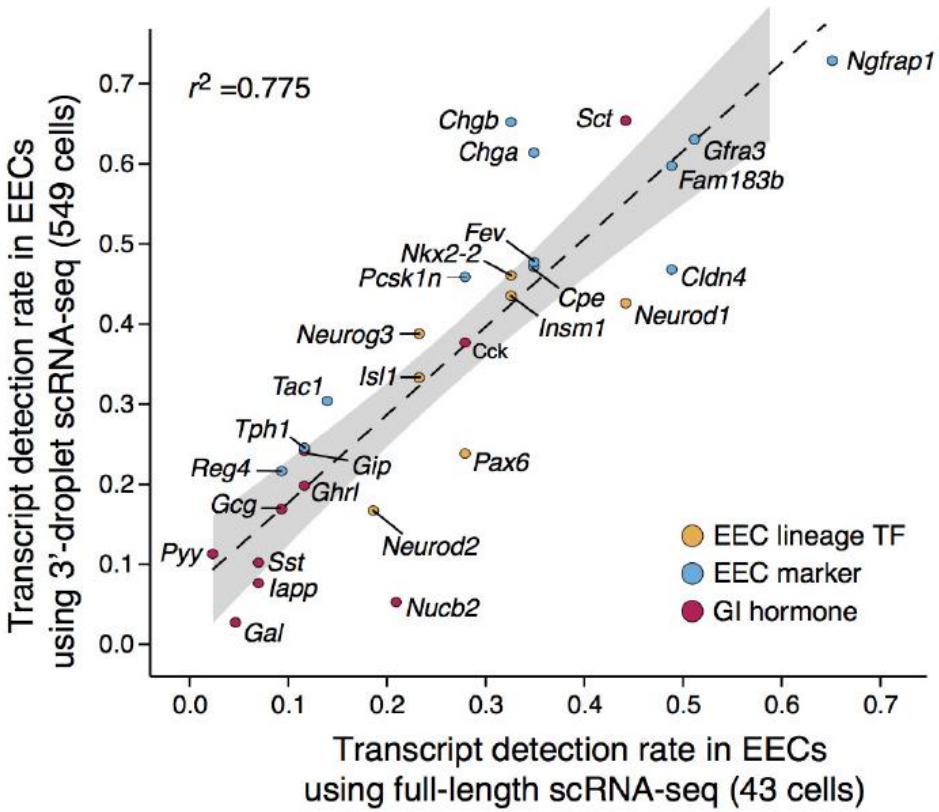


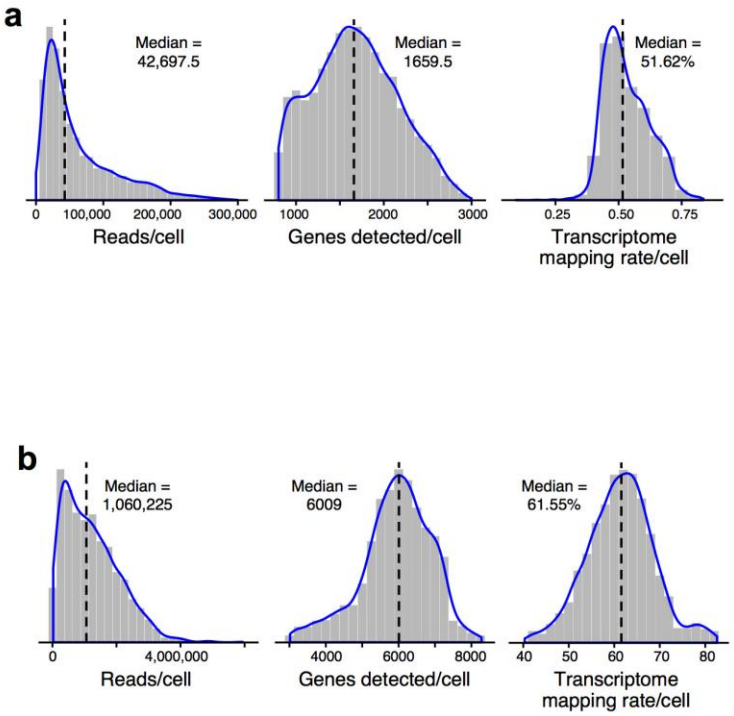
Ziegenhain *et al.*, 2017 Molecular Cell



Haber *et al.*, 2017 Nature

3’-end vs full-length Inter-platform comparison

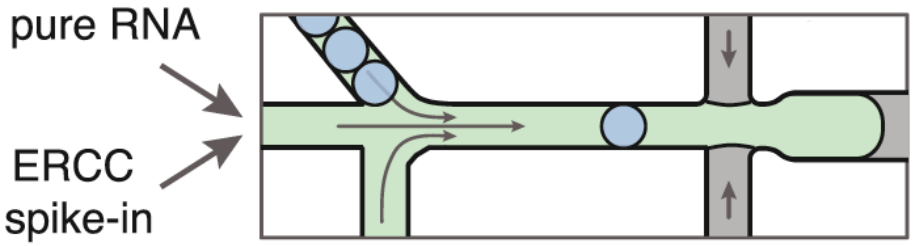




Haber *et al*., 2017 Nature

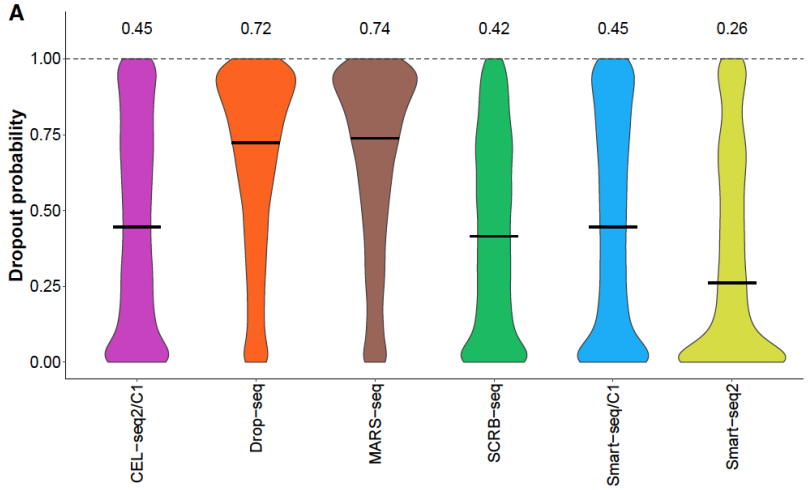
Dropout rates
ERCC spike-ins

- ERCC
  - External RNA Controls Consort

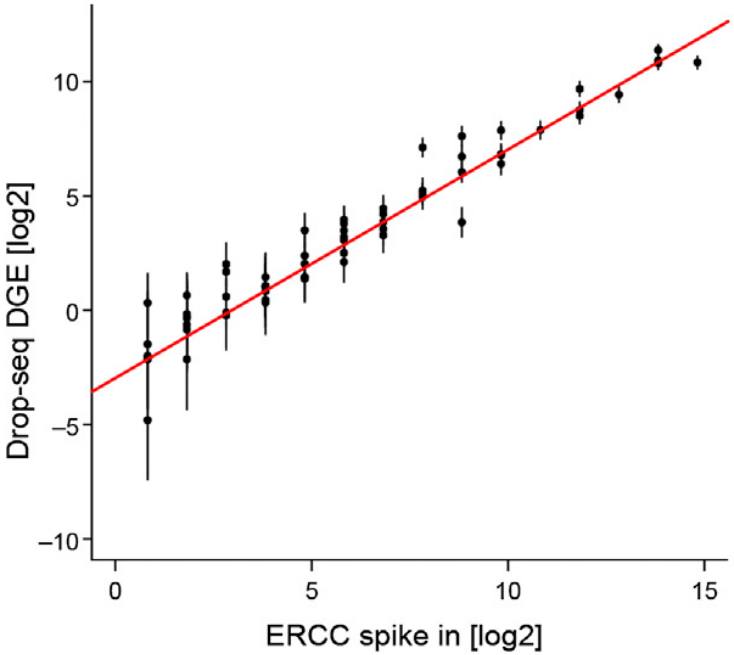


Klein *et al.*, 2015 Cell

- Inter-platform comparison
  - Dropout rate

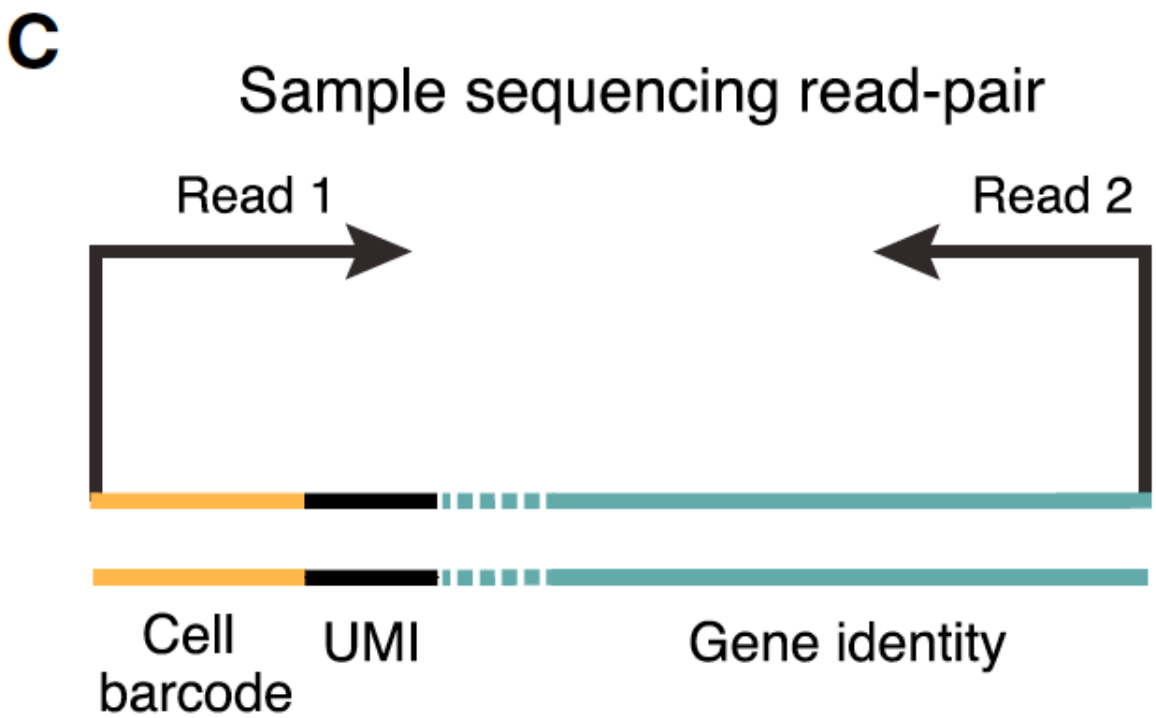


Ziegenhain *et al.*, 2017 Molecular Cell



Macosko *et al.*, 2015 Cell

3’-end based technologies Post-sequencing



SAM/BAM File

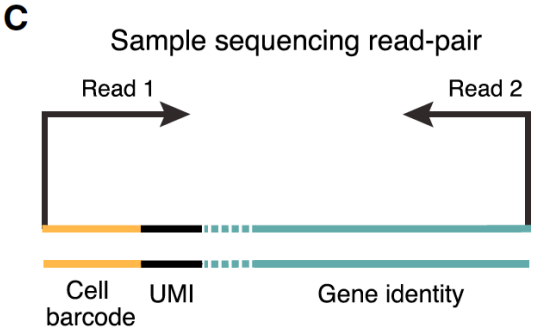
No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

- BAM files are binary, you must use samtools to view them

Bioinformatics. 2009 Aug 15;25(16):2078-9. Epub 2009 Jun 8. The Sequence Alignment/Map format and SAMtools.

Analysis pipeline - (dropseqtools)

- Goal = read1+read2 -> Counts Matrix
- Fastqs
- Unmapped queryname BAM
- Tag cell barcodes
- Tag molecular barcodes
- Trim 5’ primer sequence
- Trim 3’ polyA sequence
- STAR alignment
- Sort STAR alignment in queryname order
- Merge STAR alignment tagged SAM to recover cell/molecular barcodes
- Add gene/exon and other annotation tags



From beads … through mRNA capture … sequencing … to counts matrix

- Split and Pool x 12 = Bead Barcode

  …PCRHANDLEJJJJJJJJJJJJJJJJ

- Equimolar Mix of four bases x 8 = UMI

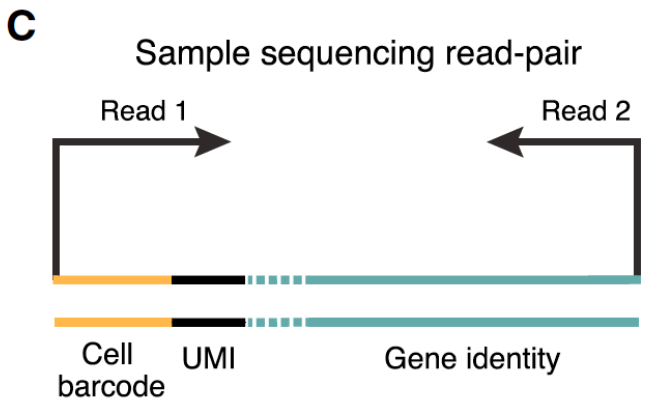
  …PCRHANDLEJJJJJJJJJJJJJJJJNNNNNNNN

- T x 30 = polyT

  …PCRHANDLEJJJJJJJJJJJJJJJJNNNNNNNNTTTT...TTTTT

- mRNA hybridization + RT + …

  …PCRHANDLEJJJJJJJJJJJJJJJJNNNNNNNNTTTT...TTTTT...mRNAmRNAmRNAmRNAmRNA



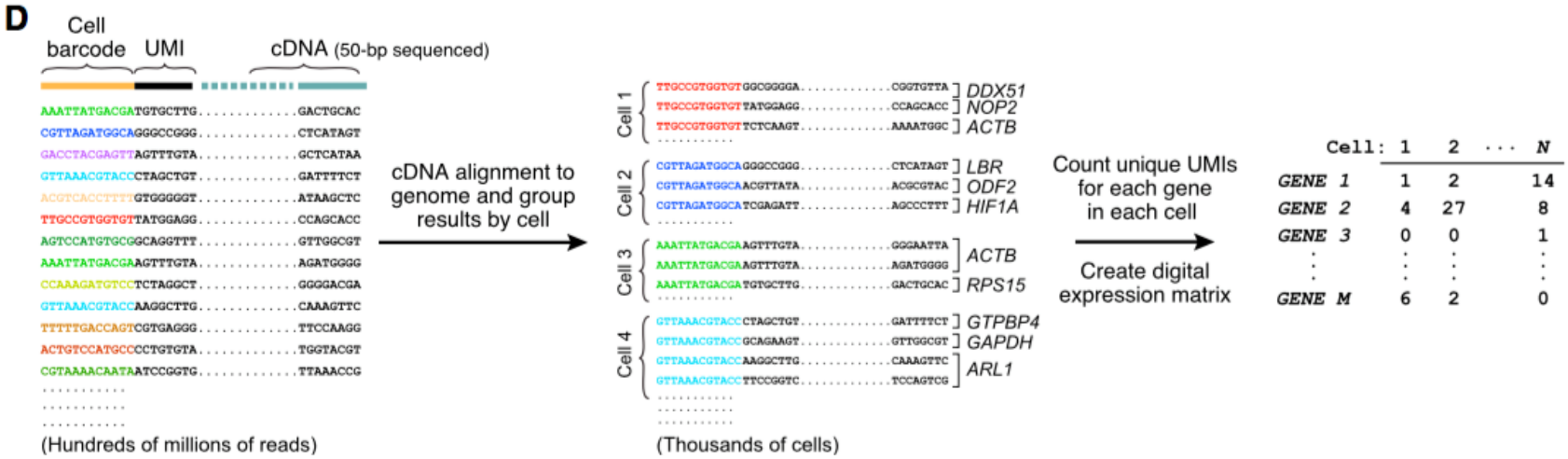
Read1

...JJJJJJJJJJJJJJJJNNNNNNNNTTTTTTT

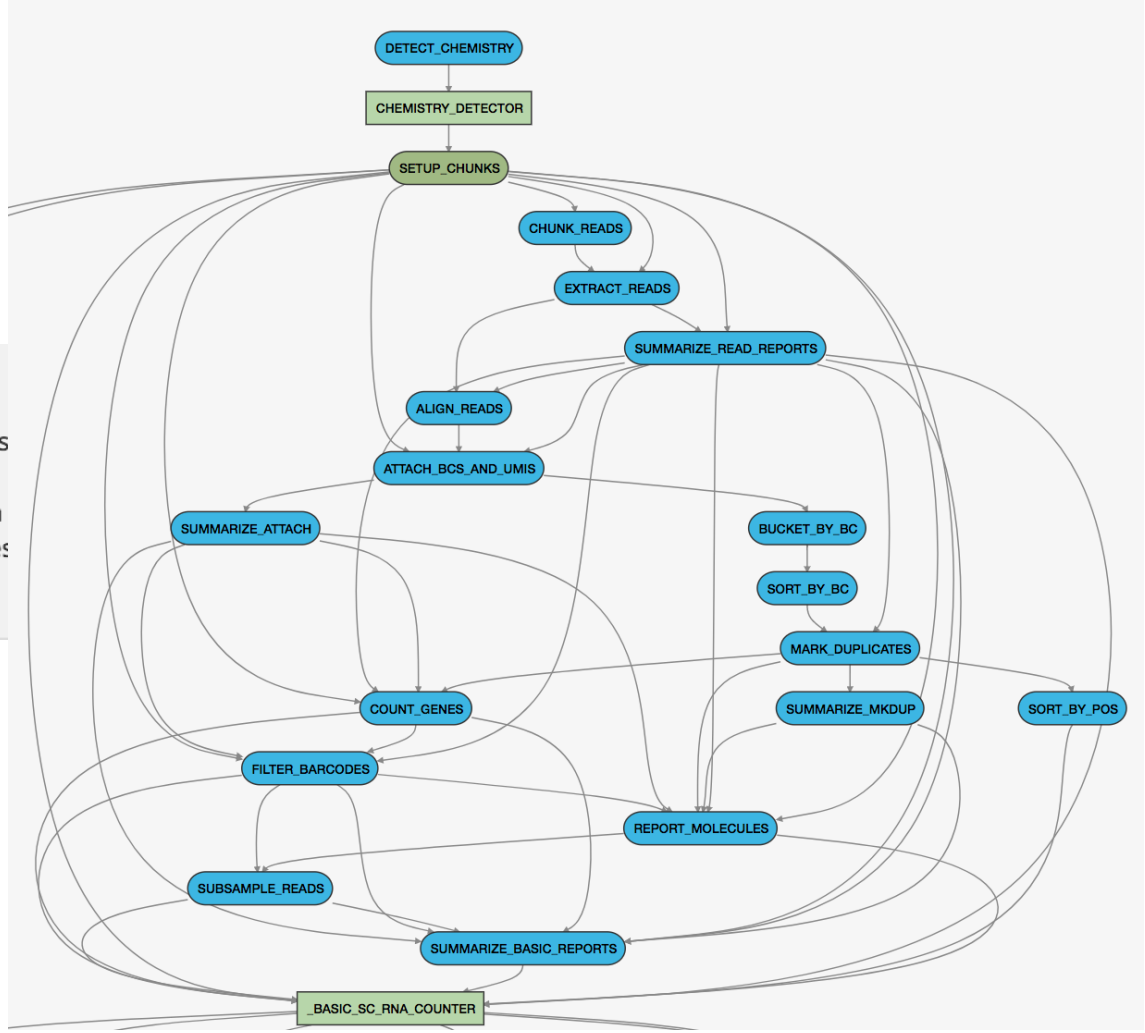
Read 2

...mRNAmRNAmRNAmRNAmRNA

3’-end based technologies - Post-sequencing



10X - cell ranger pipeline

SC_RNA_COUNTER_CS

This is the main Cell Ranger pipeline that performs alignment, de-duplication, and filtering, and generates gene-cell matrices and gene expression analysis. This outer wrapper pipeline prepares files for final output.