Single-cell RNA-sequencing (scRNA-seq): knowing the in and outs of the data generated

Ken Lau, Assistant Professor of Cell and Developmental Biology
ken.s.lau@vanderbilt.edu

CQS Summer Academy (8/13/2018)

http://www.mc.vanderbilt.edu/vumcdept/cellbio/laulab/index.html
Twitter: @KenLauLab

About me and single cells

- Started lab at Vanderbilt in 2013 with focus on single-cell biology of the gut (IBD and colon cancer)

- Multiplex imaging, CyTOF, scRNA-seq

- Training at Toronto/MIT/Harvard on multivariate analysis, mathematical modeling, and tissue systems

- inDrop in lab since August 2016 (first 1cell customer outside of Boston), > 50 samples ran so far > 500 000 cells sequenced; we have two systems

Outline

- Introduction to scRNA-seq techniques
- Discussion on scRNA-seq data issues
- Brief Python introduction

https://github.com/KenLauLab/Discovery_Oriented_Data_Science

scRNA-seq techniques

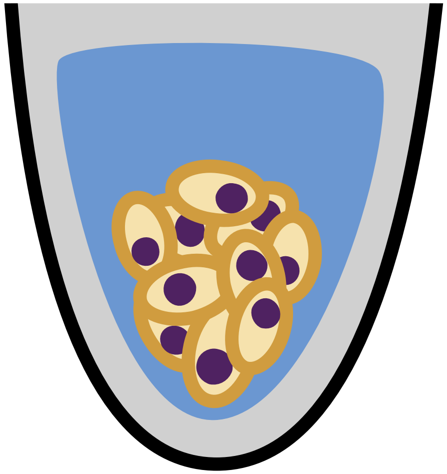
Bulk data versus single-cell data

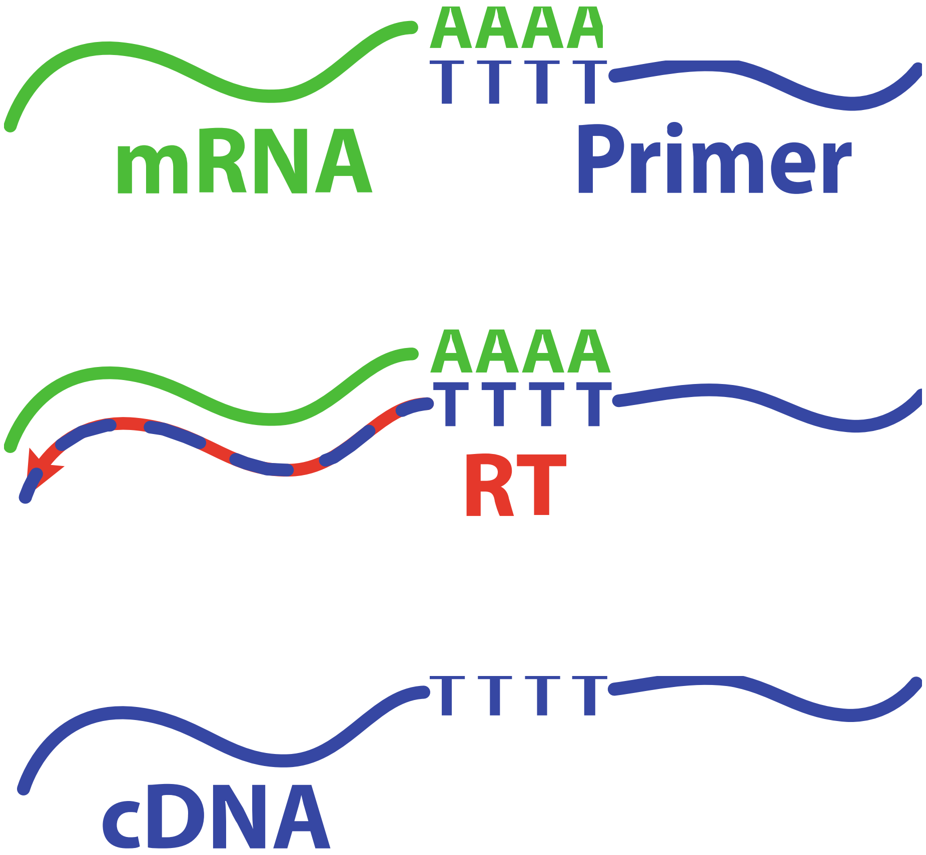
VS.

Bulk

Single-cell

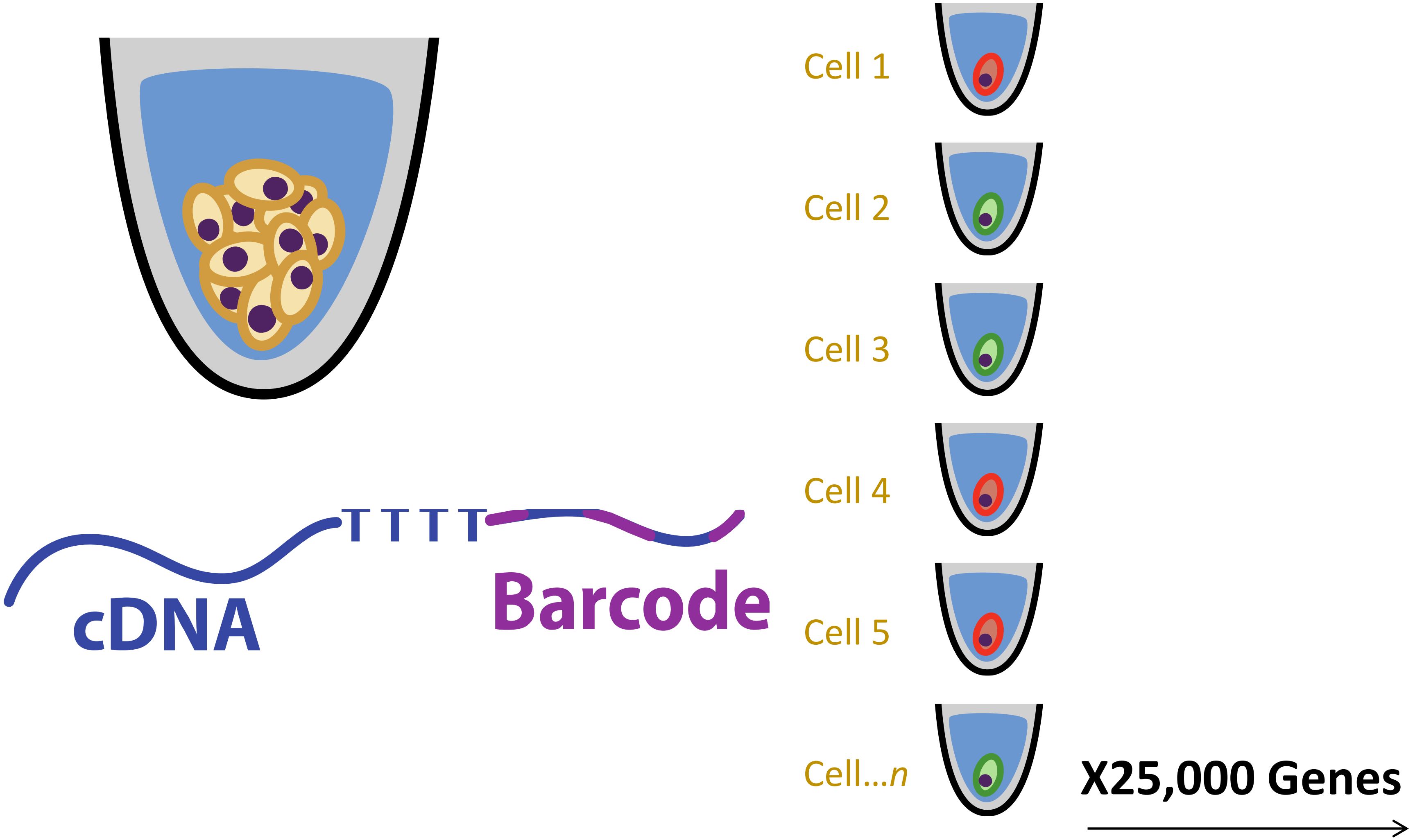
Bulk RNA-seq





1 sample x 25,000 Genes

Single-cell RNA-seq



TTTT

cDNA

Barcode

Cell 1

Cell 2

Cell 3

Cell 4

Cell 5

Cell...*n*

X25,000 Genes

scRNAseq protocols

Cell Encapsulation techniques

- Droplet-Based
- Well-Based
- Microfluidic capture (Fluidigm C1)

Lysis and RT

- Coupled – requires balanced mix
- Uncoupled – enables more aggressive lysis

RNA capture strategies

- Poly dT priming
- Targeting / enrichment

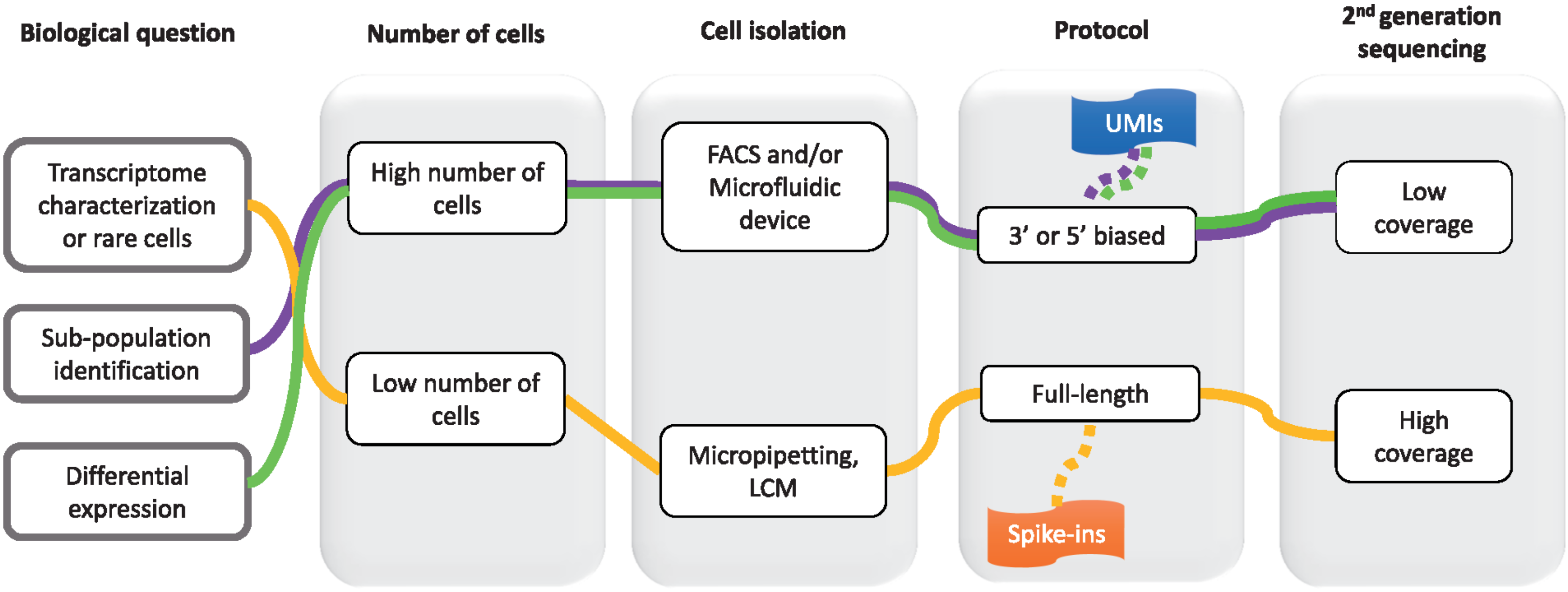
Indexing strategies

- During capture/RT (typically per **cell** indexes - barcodes)
- After RT (typically per **well** indexes)

Amplification strategies

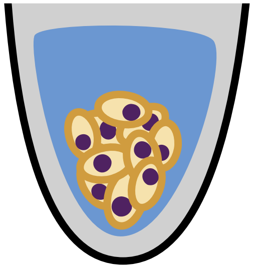
- PCR vs IVT

Typical logic for a scRNA-seq experiment

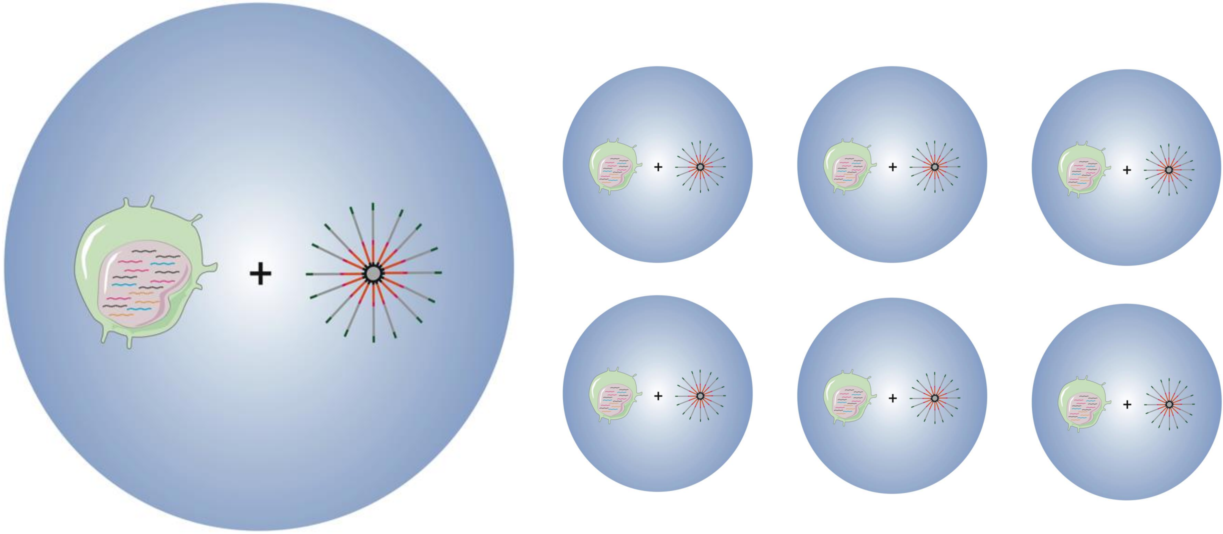


Alessandra Dal Molin, Barbara Di Camillo; How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives, *Briefings in Bioinformatics*, , bby007,

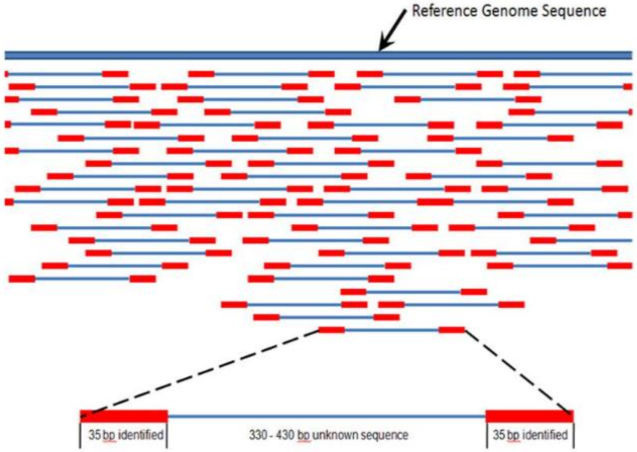
Single cell suspension***



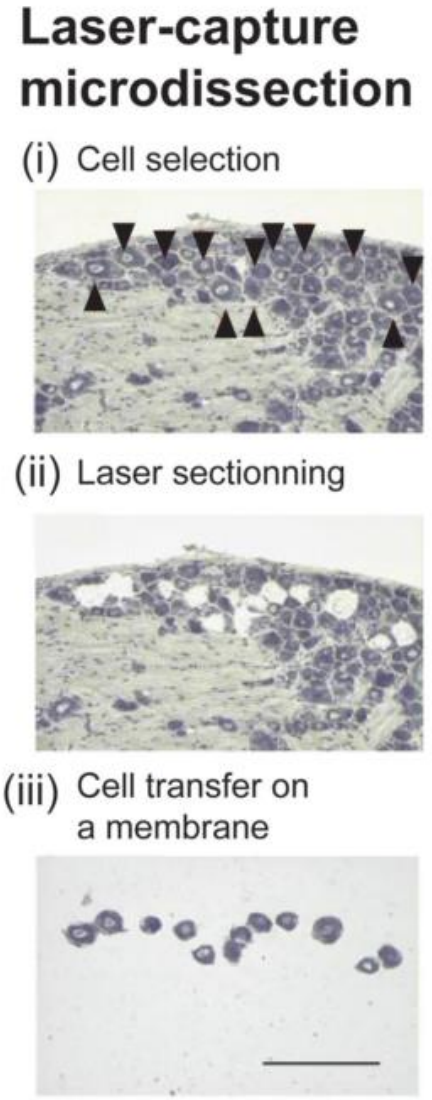
Single-cell encapsulation/ Library preparation



Sequencing and alignment (Bioinformatics I)



Well-known methods to isolate single cells


Laser-capture microdissection
(i) Cell selection
(ii) Laser sectionning
(iii) Cell transfer on a membrane

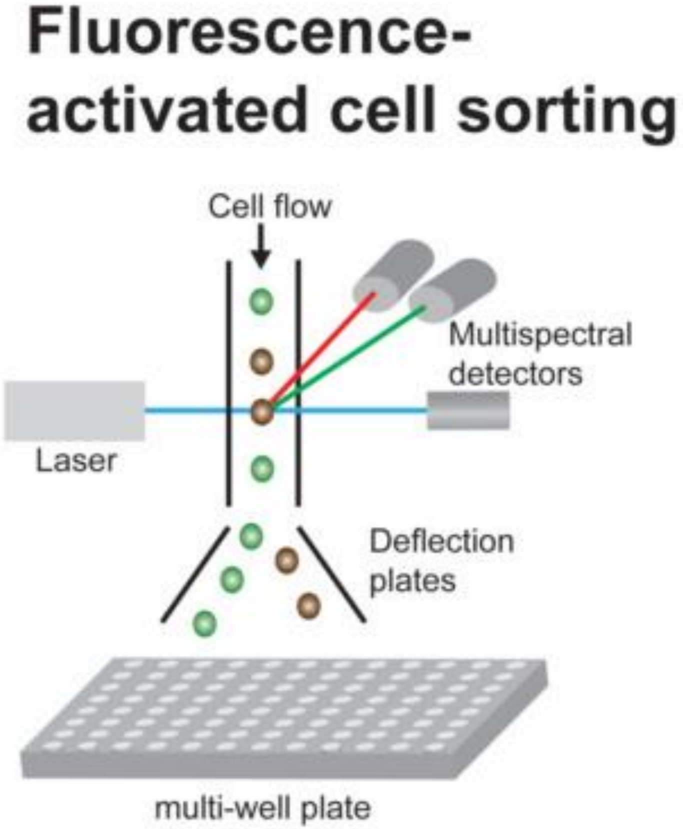
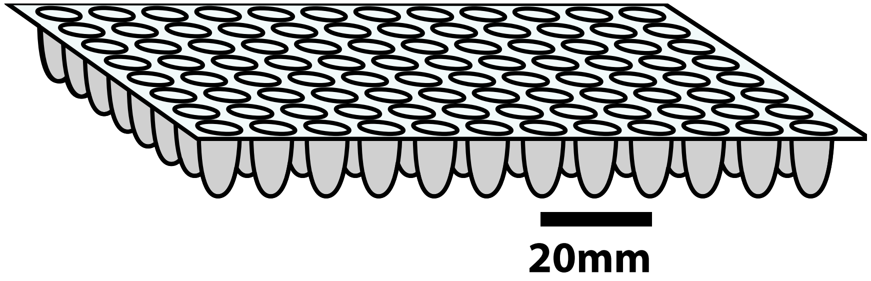
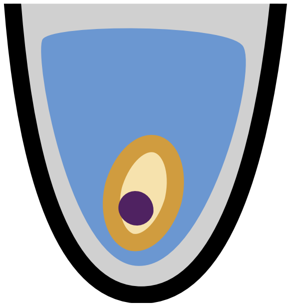


Plate-Based scRNAseq



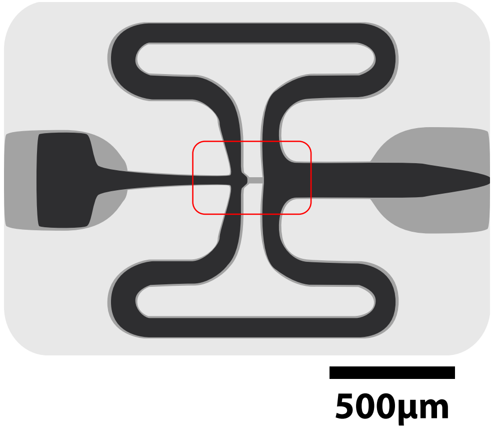


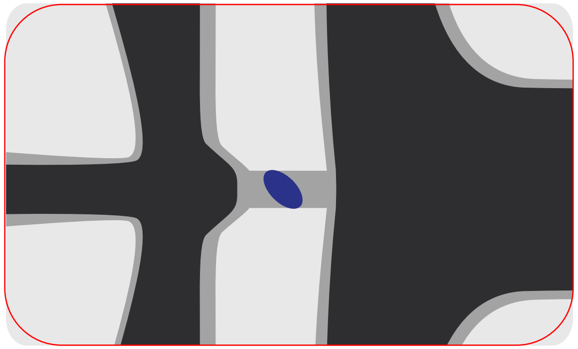
- Isolate RNA, label transcripts using barcoded RT primers (3’ seq) or through template switching library prep (enables full length)

Relative to other platforms:

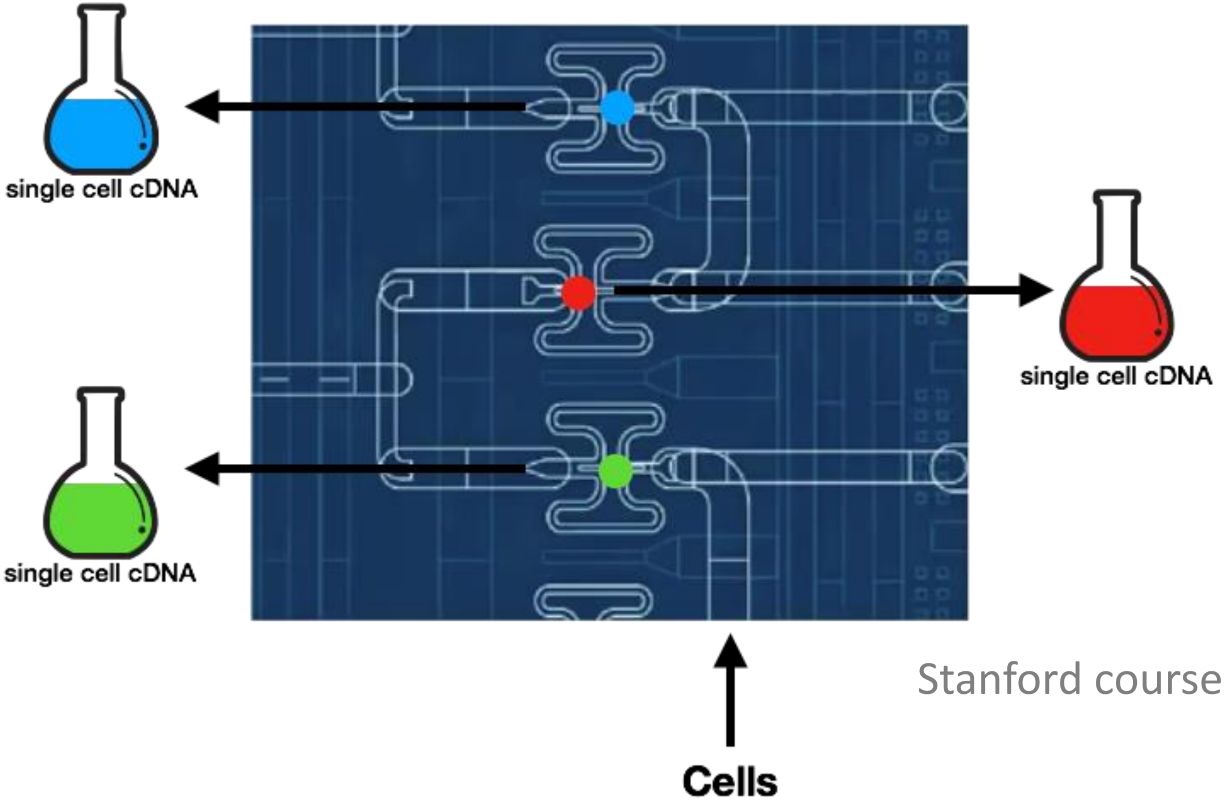
- ~10µl/ cell, <1000 cells (higher volume, lower throughput)
- Deeper sequencing possible - flexibility
- Number of wells become limiting (doublet rate vs. cost)

Microfluidic capture scRNAseq (Fluidigm C1)





Capture chips are built to accommodate specific ranges of cell size



Stanford course

Bead-based capture:
Immobilized sets of indexed primers

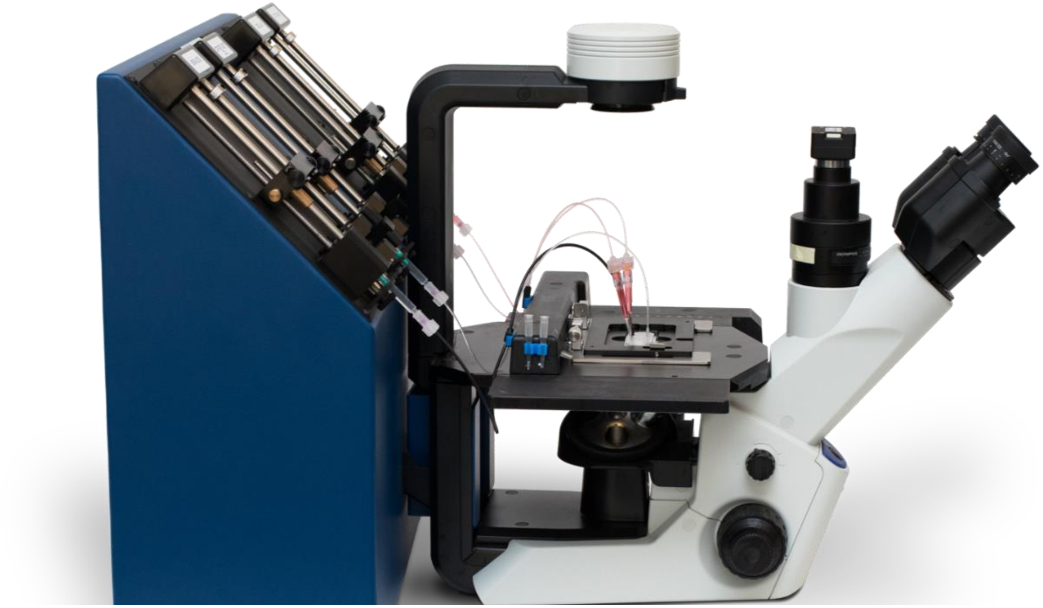
- Each bead is coated with primers containing a barcode unique to that bead – (index for each cell)

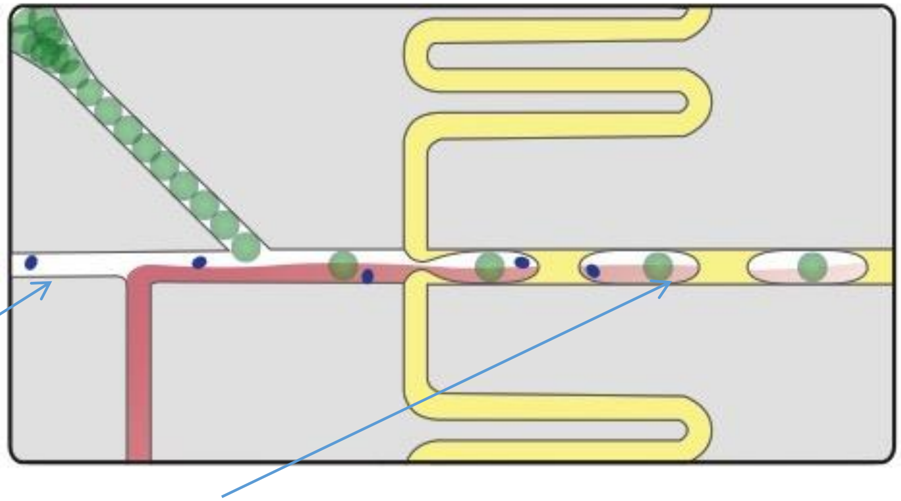
Barcode UMI TTTT

- And a Unique Molecular Identifier (UMI) that uniquely tags each primer – (index for each transcript)

Droplet-based encapsulation

- Co-encapsulating cells and beads in thousands of 1-5nL droplets
- Beads carry barcoded poly-T primers to capture RNA
- Encapsulation rate follows Poisson distribution
- Excess of "vessels" to minimize doublets

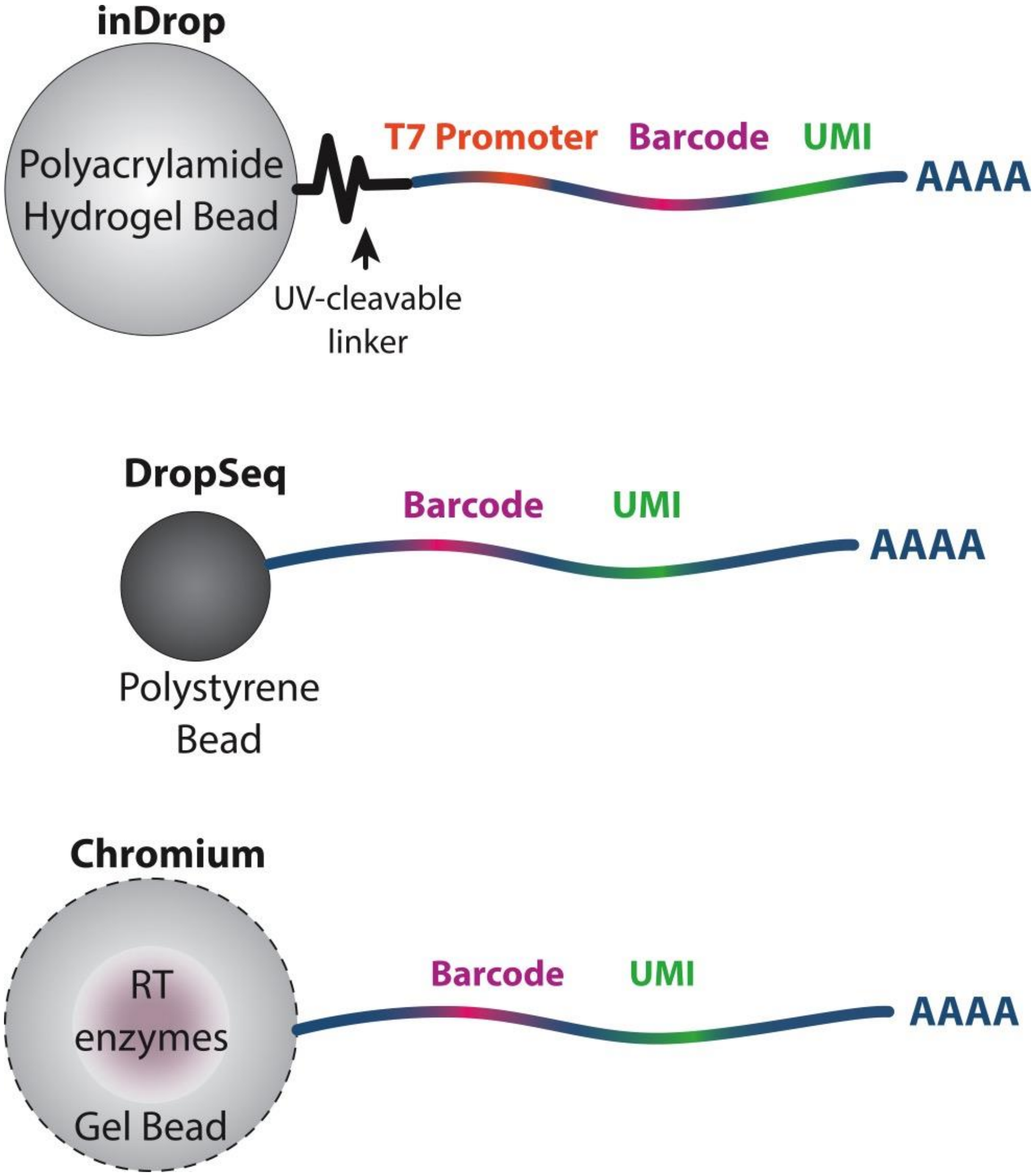




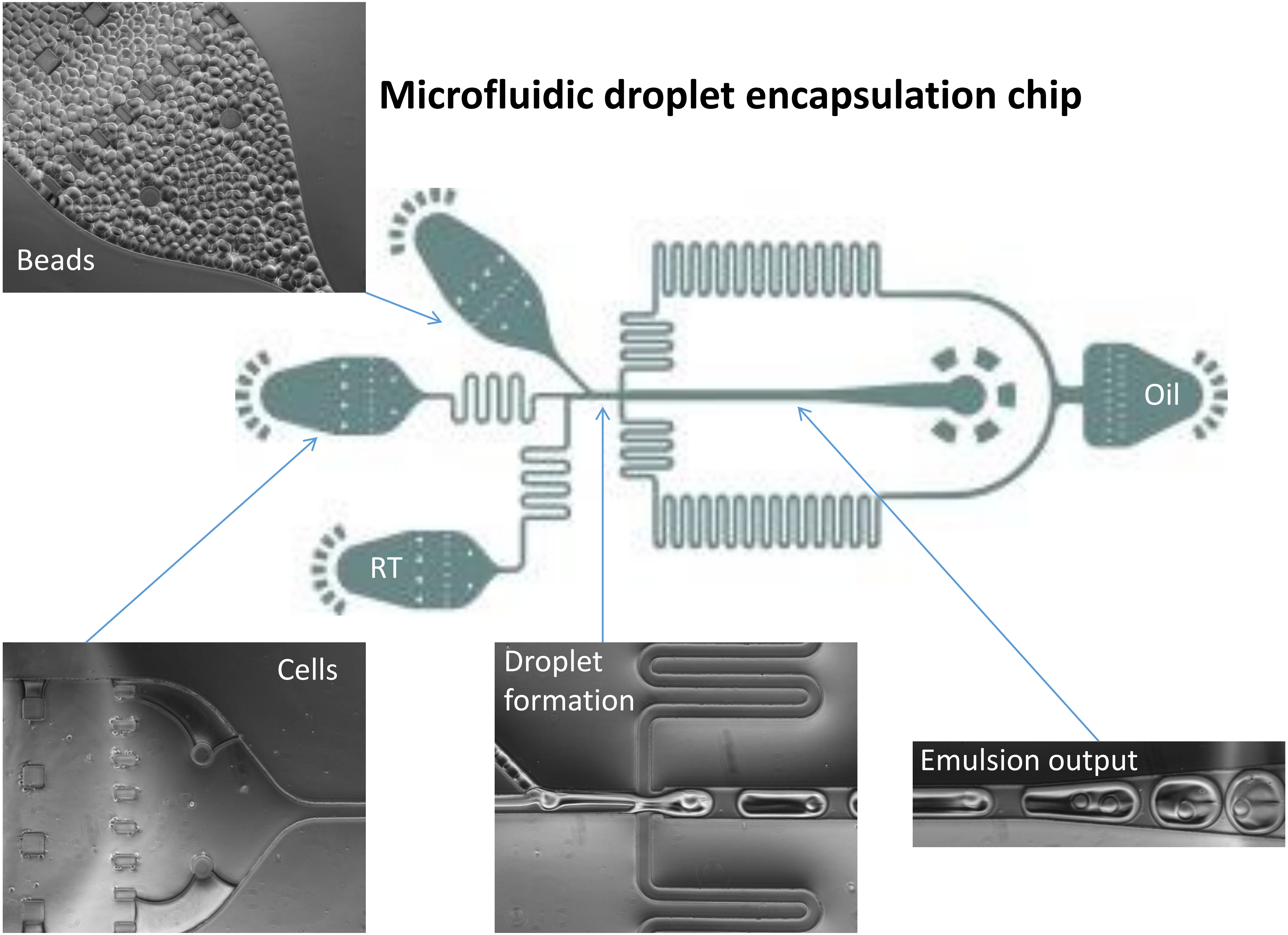
Cells entering/ sec x Beads/droplet = cells captured/ sec

Beads

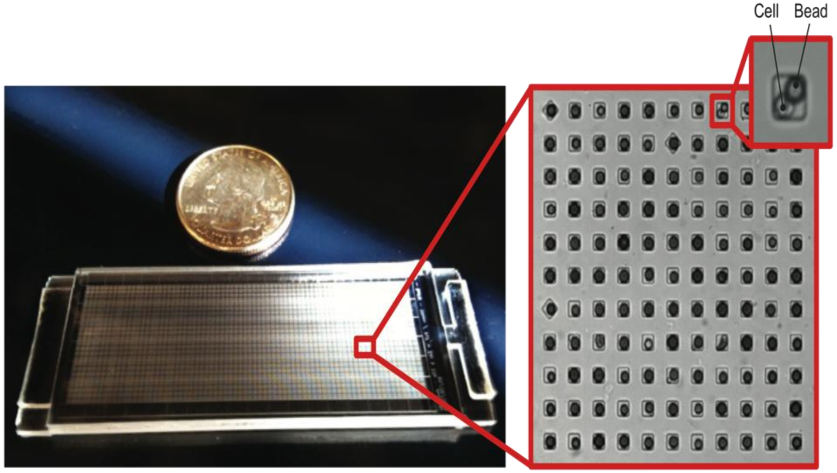
- inDrop (1cellBio)
- Drop-seq (Chemgenes)
- Chromium (10x Genomics)



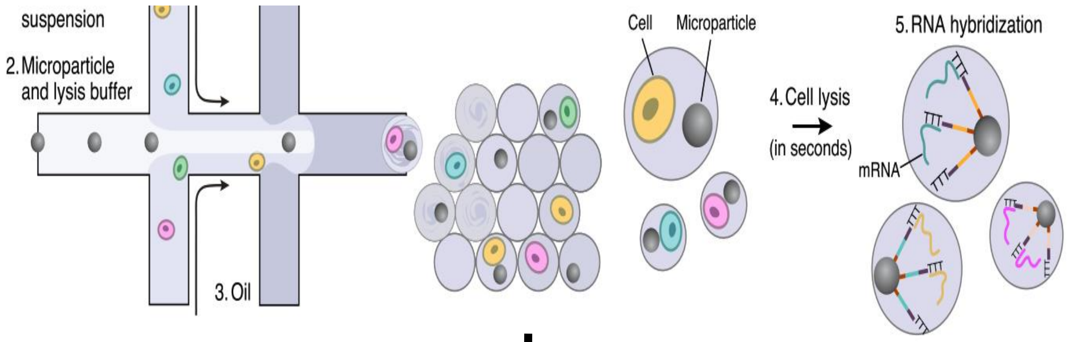
Microfluidic droplet encapsulation chip



Seq-Well - microwell sequencing (Shalek lab)

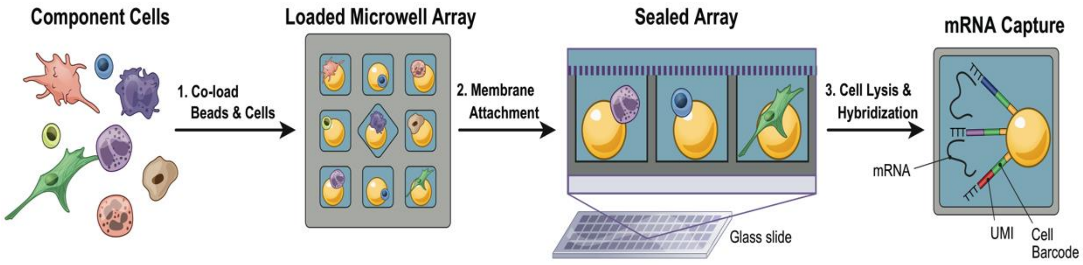


Drop-Seq

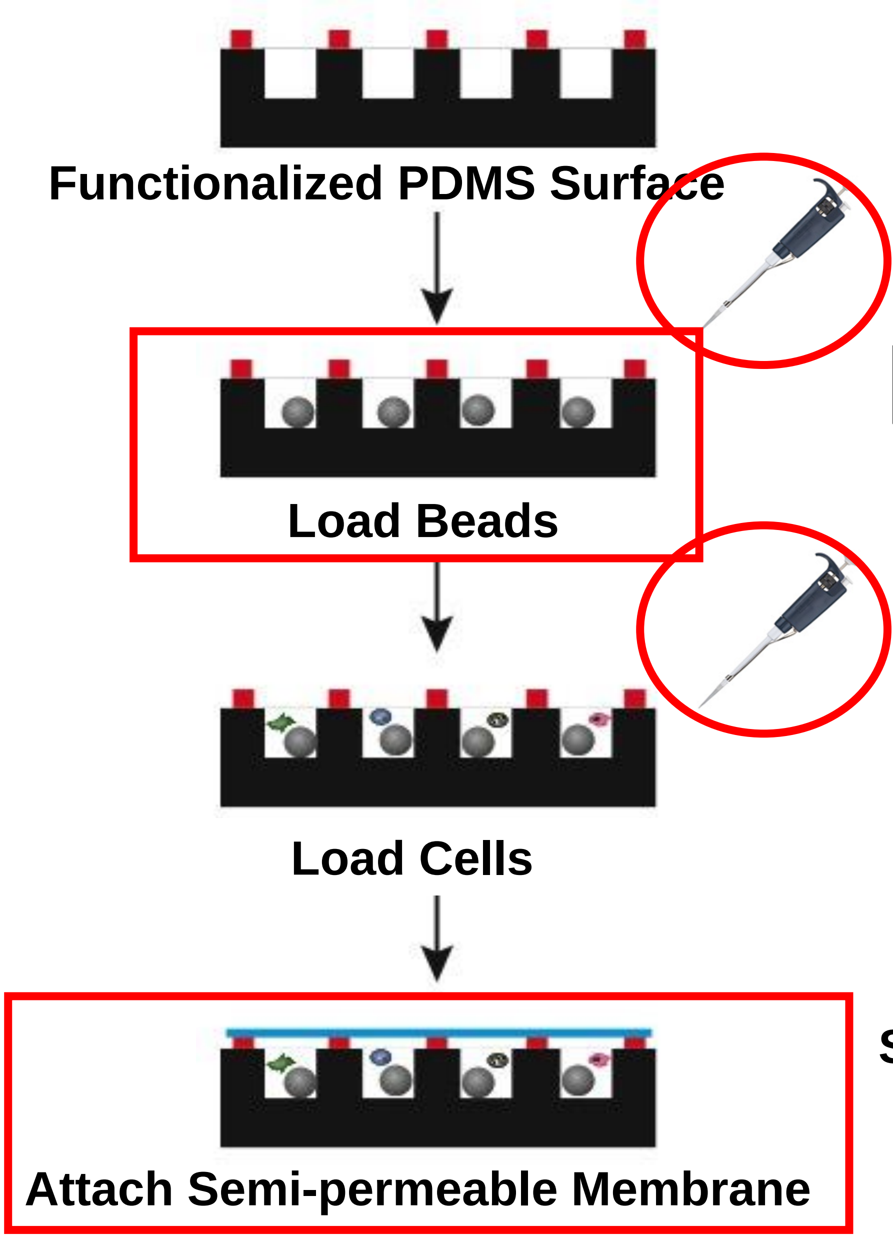


↓

Seq-Well



Seq-Well: Principle



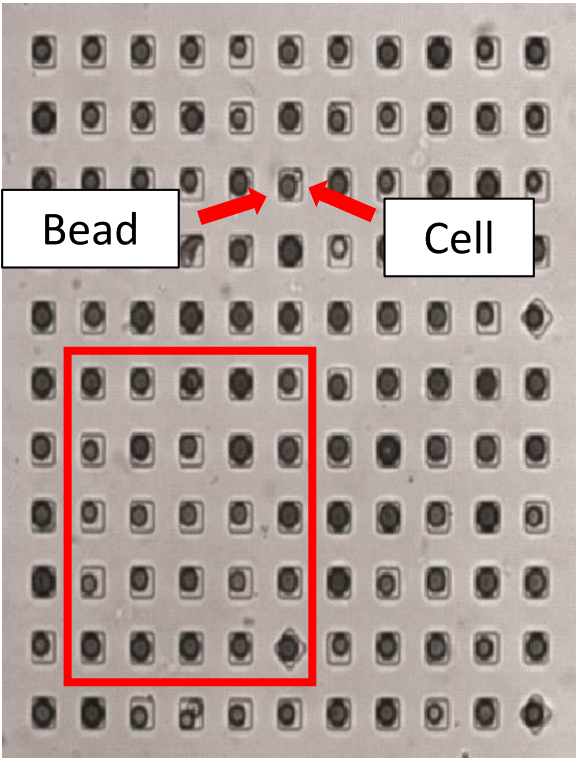
Functionalized PDMS Surface

Load Beads

Load Cells

Attach Semi-permeable Membrane

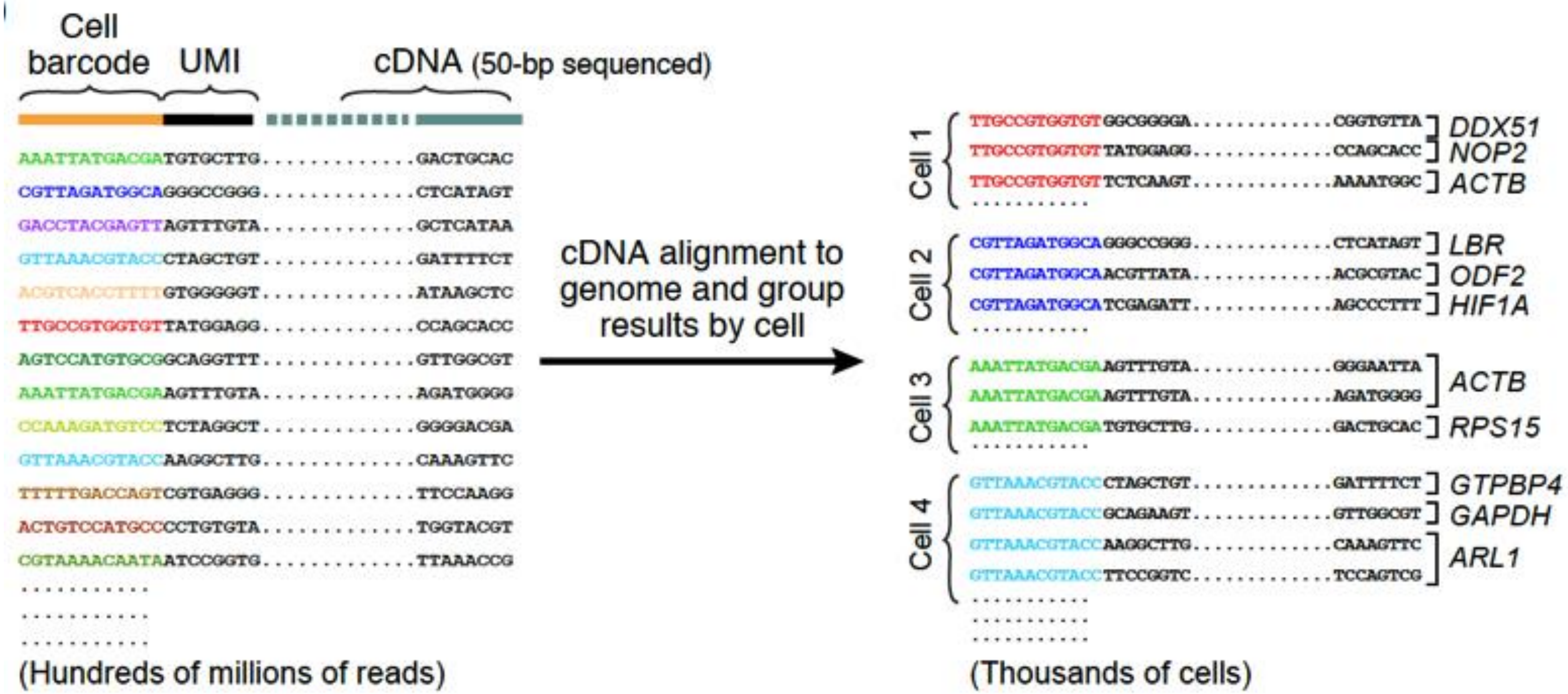
Nanowell Array



Size Exclusion ➔ ≤ 1 bead per well

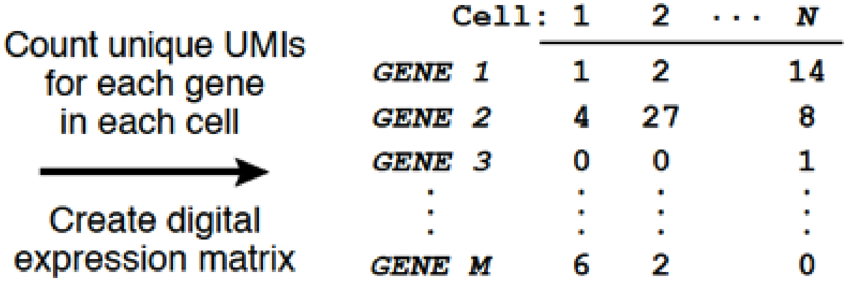
Sealing ➔ ~~Cross Contamination~~

Deconvolving the data


Cell barcode
UMI
cDNA (50-bp sequenced)
AAATTATGACGATGTGCTTG............GACTGCAC
CGTTAGATGGCAGGGCCGGG............CTCATAGT
GACCTACGAGTTAGTTTGTA............GCTCATAA
GTTAAACGTACCCTAGCTGT............GATTTTCT
ACGTCACCTTTTGTGGGGGT............ATAAGCTC
TTGCCGTGGTGTTATGGAGG............CCAGCACC
AGTCCATGTGCGGCAGGTTT............GTTGGCGT
AAATTATGACGAAGTTTGTA............AGATGGGG
CCAAAGATGTCCTCTAGGCT............GGGGACGA
GTTAAACGTACCAAGGCTTG............CAAAGTTC
TTTTTGACCAGTCGTGAGGG............TTCCAAGG
ACTGTCCATGCCCCTGTGTA............TGGTACGT
CGTAAAACAATAATCCGGTG............TTAAACCG
(Hundreds of millions of reads)
cDNA alignment to genome and group results by cell
Cell 1
TTGCCGTGGTGTGGCGGGGA............CGGTGTTA
DDX51
TTGCCGTGGTGTTATGGAGG............CCAGCACC
NOP2
TTGCCGTGGTGTTCTCAAGT............AAAATGGC
ACTB
Cell 2
CGTTAGATGGCAGGGCCGGG............CTCATAGT
LBR
CGTTAGATGGCAACGTTATA............ACGCGTAC
ODF2
CGTTAGATGGCATCGAGATT............AGCCCTTT
HIF1A
Cell 3
AAATTATGACGAAGTTTGTA............GGGAATTA
AAATTATGACGAAGTTTGTA............AGATGGGG
ACTB
AAATTATGACGATGTGCTTG............GACTGCAC
RPS15
Cell 4
GTTAAACGTACCCTAGCTGT............GATTTTCT
GTPBP4
GTTAAACGTACCGCAGAAGT............GTTGGCGT
GAPDH
GTTAAACGTACCAAGGCTTG............CAAAGTTC
GTTAAACGTACCTTCCGGTC............TCCAGTCG
ARL1
(Thousands of cells)

Reads with same barcodes collapse into cells

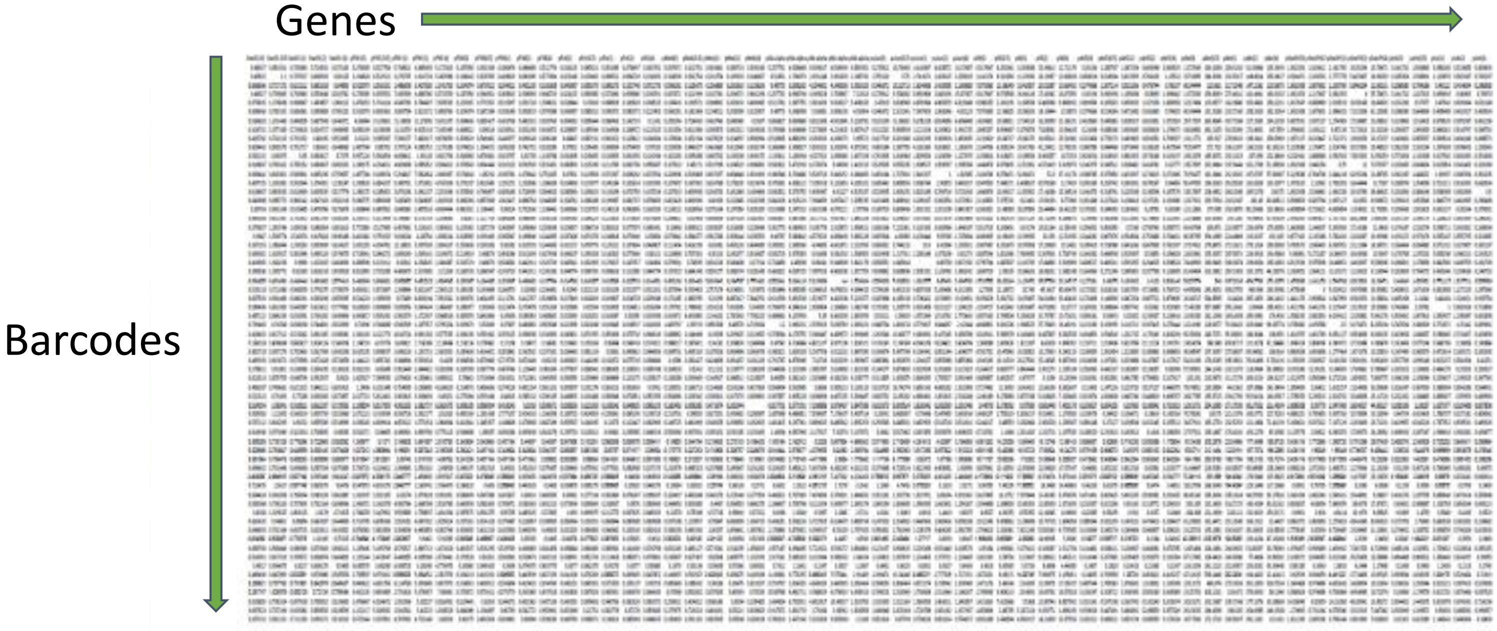
Read with same UMIs collapse into transcript counts


Count unique UMIs for each gene in each cell
Create digital expression matrix

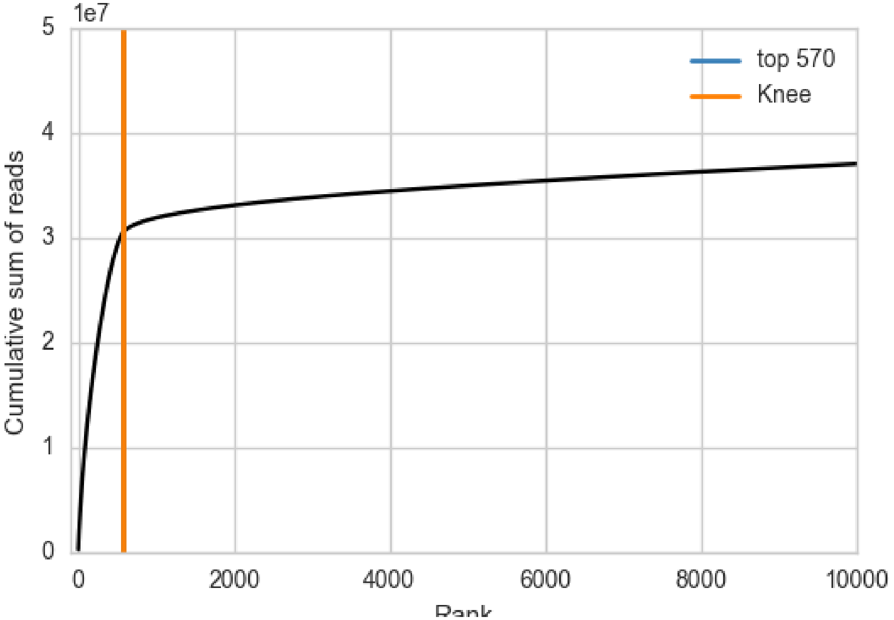
Cell:	1	2	...	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

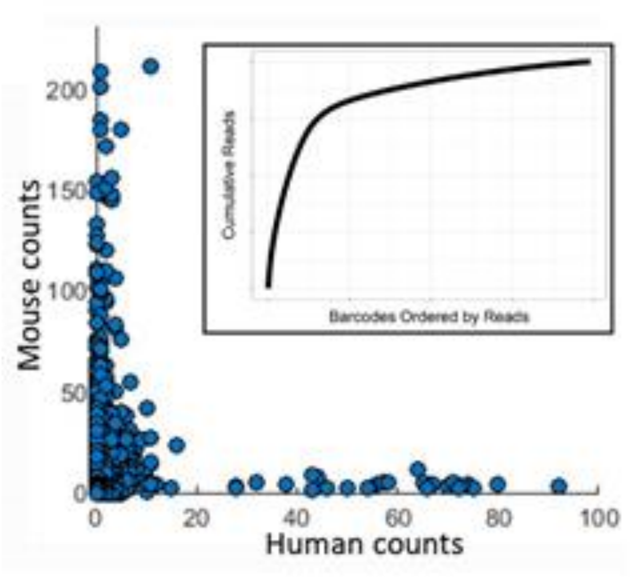
scRNA-seq Data Exploration
(and problems)

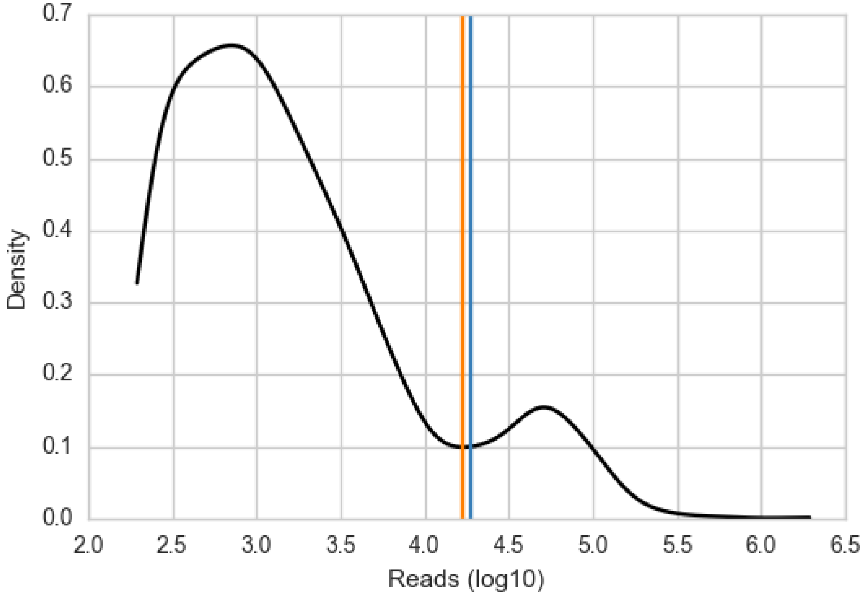
Table of genes and barcodes

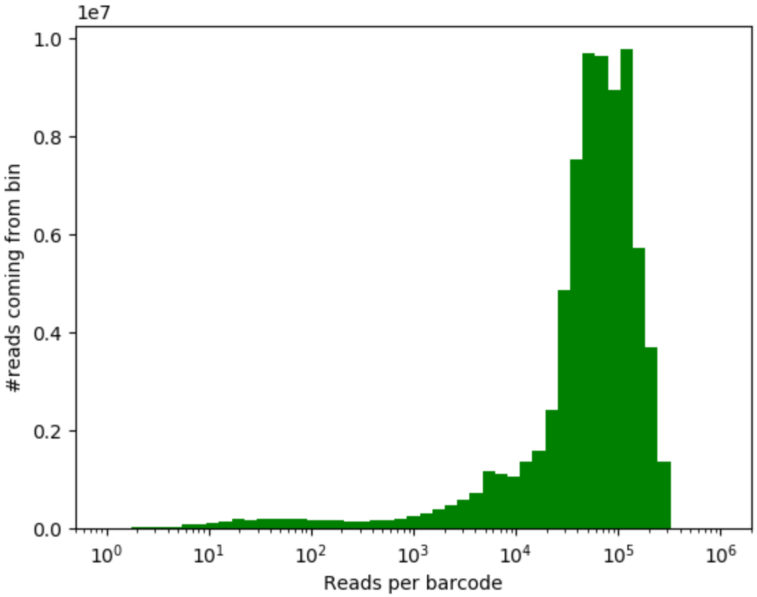


Inflection point method for identifying barcodes with real cells



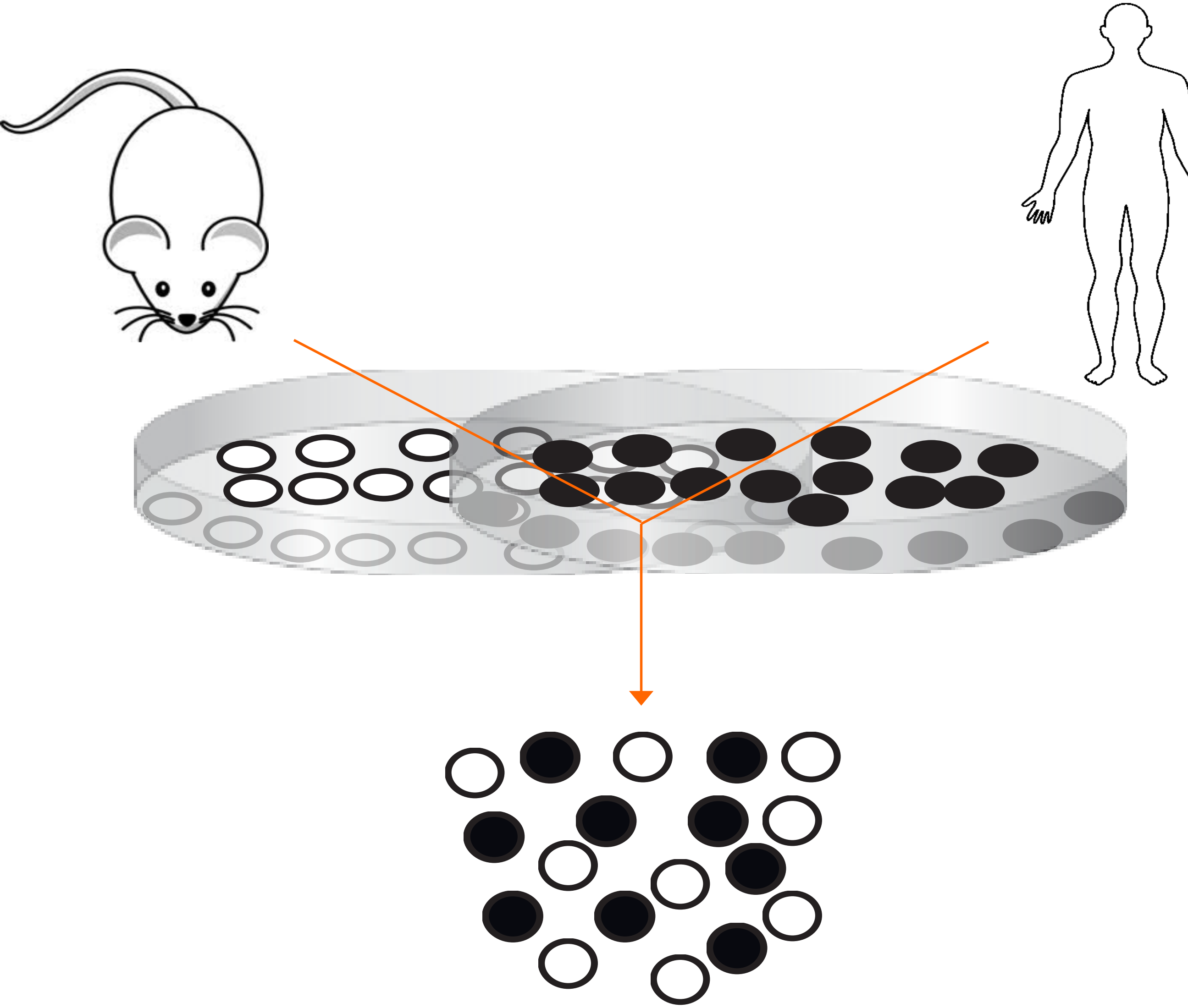


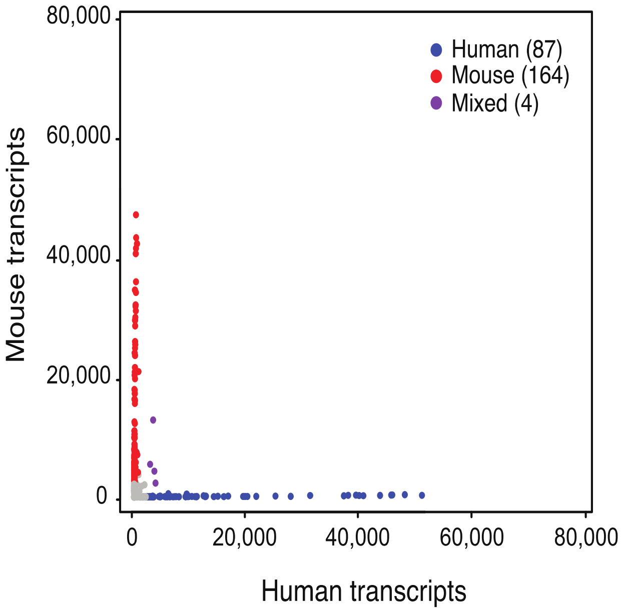




https://cgatoxford.wordpress.com/2017/05/18/estimating-the-number-of-true-cell-barcodes-in-single-cell-rna-seq/

Doublet rate determination





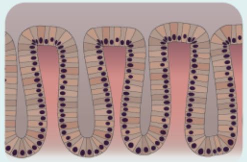
Klein et al, *Cell*, **161**, 2015
Macosko et al, *Cell*, **161**, 2015

Procedure for the isolation of high viability single-cells from tissues

Keep all reagents on ice and perform all procedures at 4˚c, avoid working with overconcentrated tissue/cell solutions

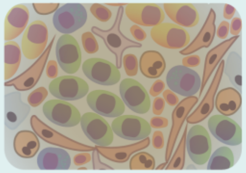
Tissue

**Keep cold and minimize ischemic time*
**Mince if needed*



Wash in PBS Decant small debris

Normal colonic epithelium can be isolated through chelation, while other tissues may require direct mechanical processing to achieve 50-500µm fragments. Care should be taken to remove dead cells during washes to maintain viability

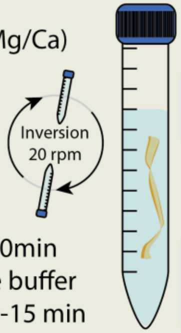


Wash in PBS Decant small debris

The process of isolating tissue fragments should be optimized to the needs of the target cells. This may mean filtering or taking other measures to enrich.

Chelation

10ml DPBS(-Mg/Ca)
3mM EDTA
0.5mM DTT
10mM NAC



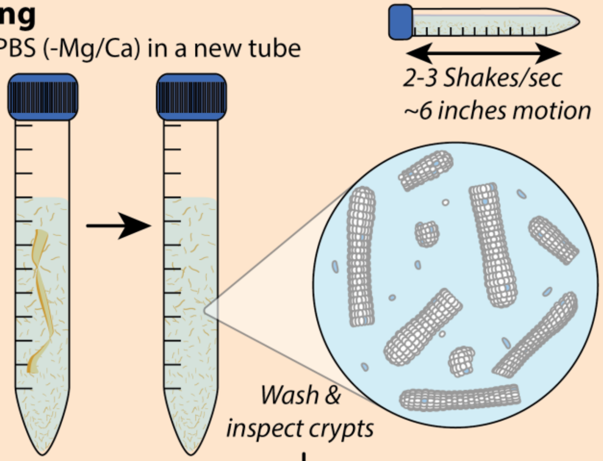
~30-60min
Change buffer every 10-15 min

Use forceps to move tissue

Move back to fresh chelation buffer

Shaking

10ml DPBS (-Mg/Ca) in a new tube

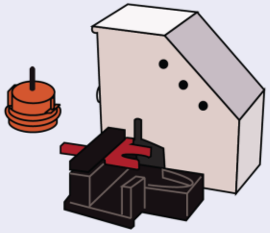


2-3 Shakes/sec
~6 inches motion

Wash & inspect crypts

Medimachine

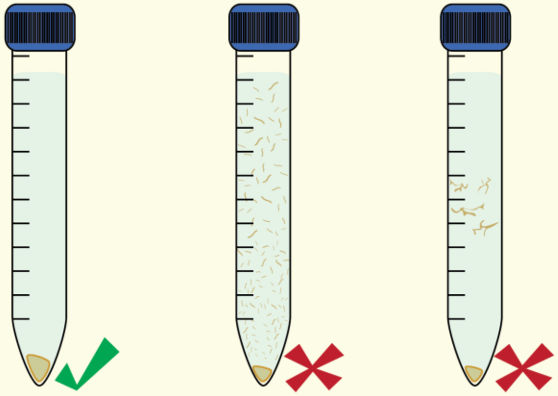
In cold room
Pre-wet 50µm medicon before use
Wash quickly and thoroughly after



**GentleMACS is a potentially less damaging approach we have not yet tested for breaking apart tissue and further dissociating cells.*

Washing

DPBS (-Mg/Ca) spin @~300xg for 2.5min



Good pelleting: (fat or pancreas may still float in first wash)

Sticking to tube; may need more chelation (EDTA)*

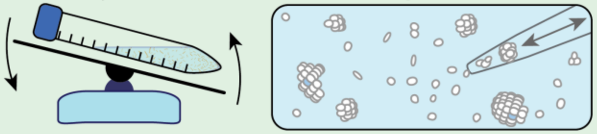
Floating/clumping typically caused by excess mucin or inviable cells**

**Pellet without EDTA prior to re-suspending in protease*
***Optimize spin conditions so that single cells are decanted*
***Dont leave tissue in pellet too long, split if too dense*
***Re-suspend well, but do not over pipette prior to dissociation*

Cold Protease Dissociation

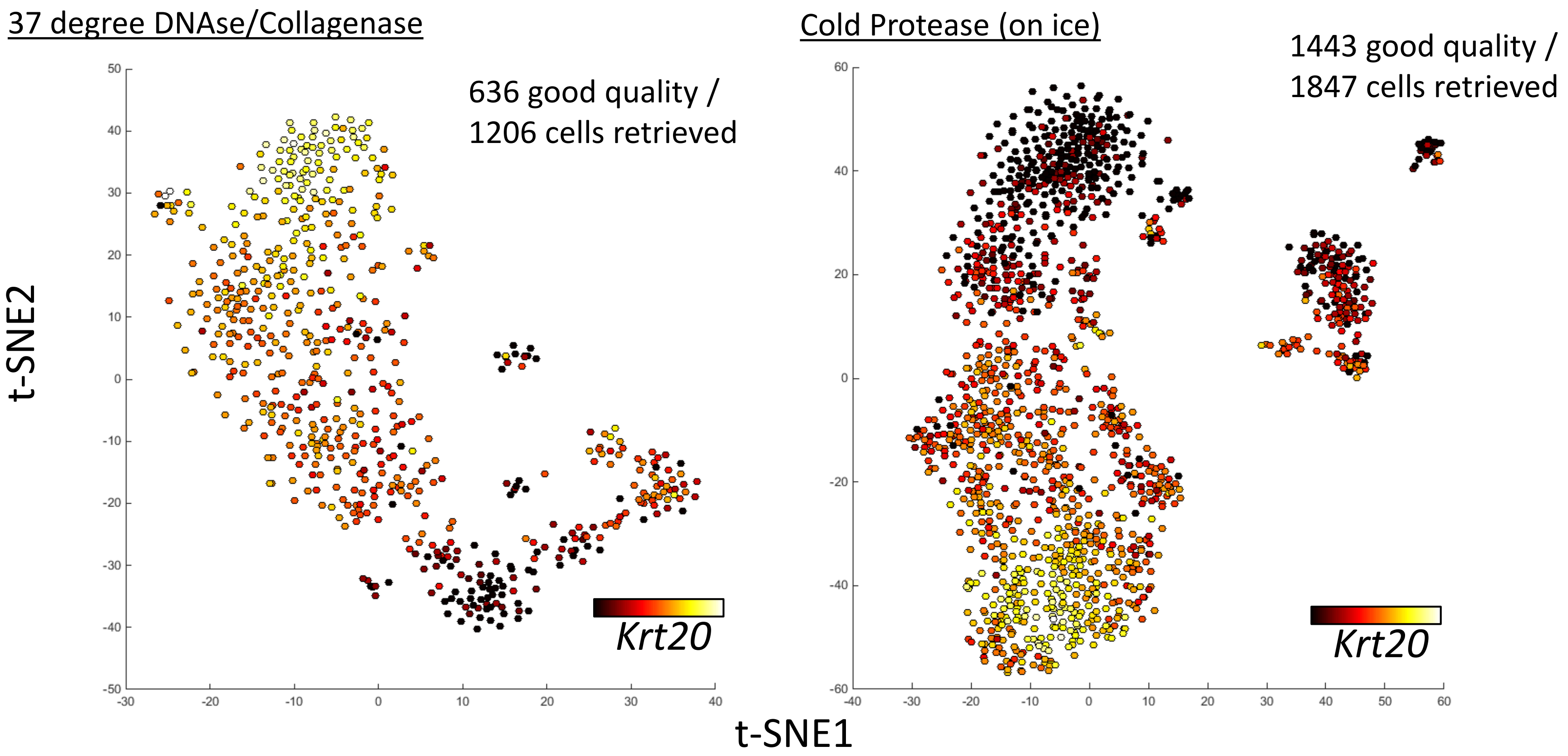
2-5ml DPBS (-Mg/Ca) + 2.5mg/ml DNase + 5mg/ml Subtilisin
**Use 2ml for every 50-100µl of peleted tissue, dont overconcentrate*
**Add buffer to frozen subtilisin aliquot to thaw immediately before use*
**Pipette with a 1000µl tip every 5-10 min and check for singlets*

4-6˚c with gentle motion



Single-cell isolation for scRNA-seq to minimize dissociation artifacts

- cold protease from *Bacillus licheniformis,* soil bacteria from Himalayan glaciers
- enables tissue preparation on ice (at 4 degrees)



37 degree DNAse/Collagenase

636 good quality / 1206 cells retrieved

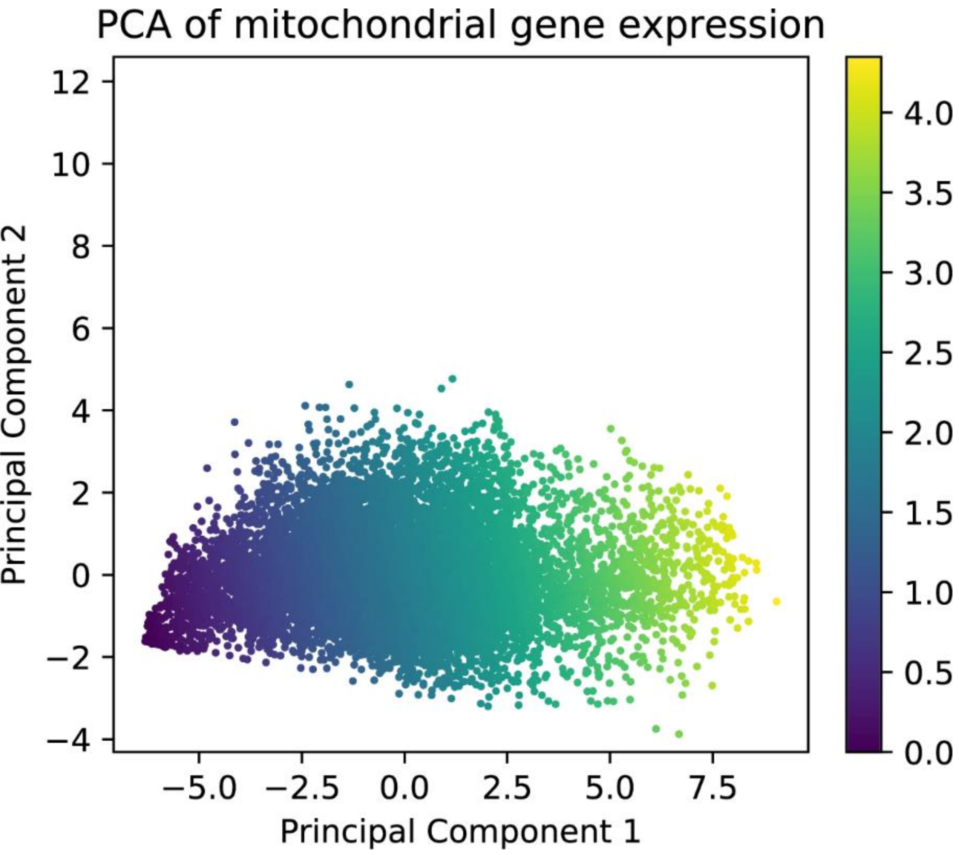
Cold Protease (on ice)

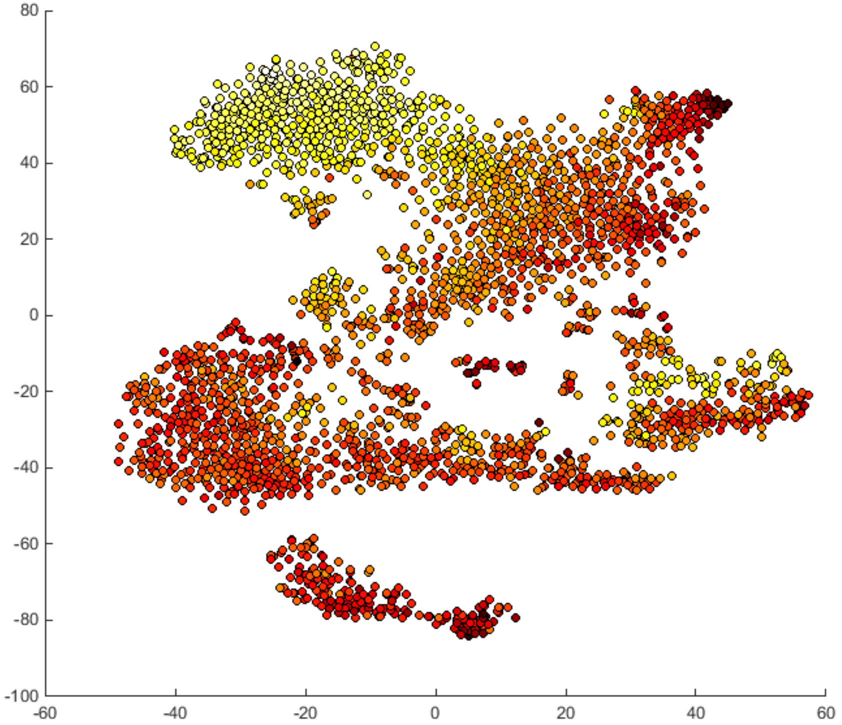
1443 good quality / 1847 cells retrieved

- caveat – the efficacy for retrieving all cell types from all tissues unknown

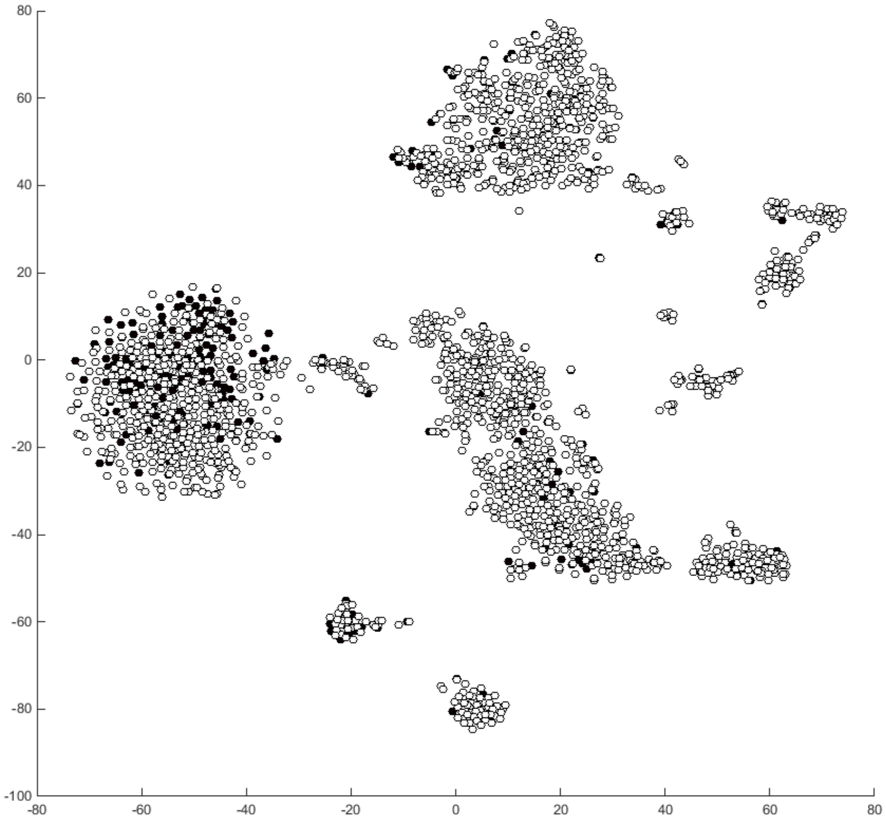
(Adam et al., *Development*, 2018)

Stressed/dying cells

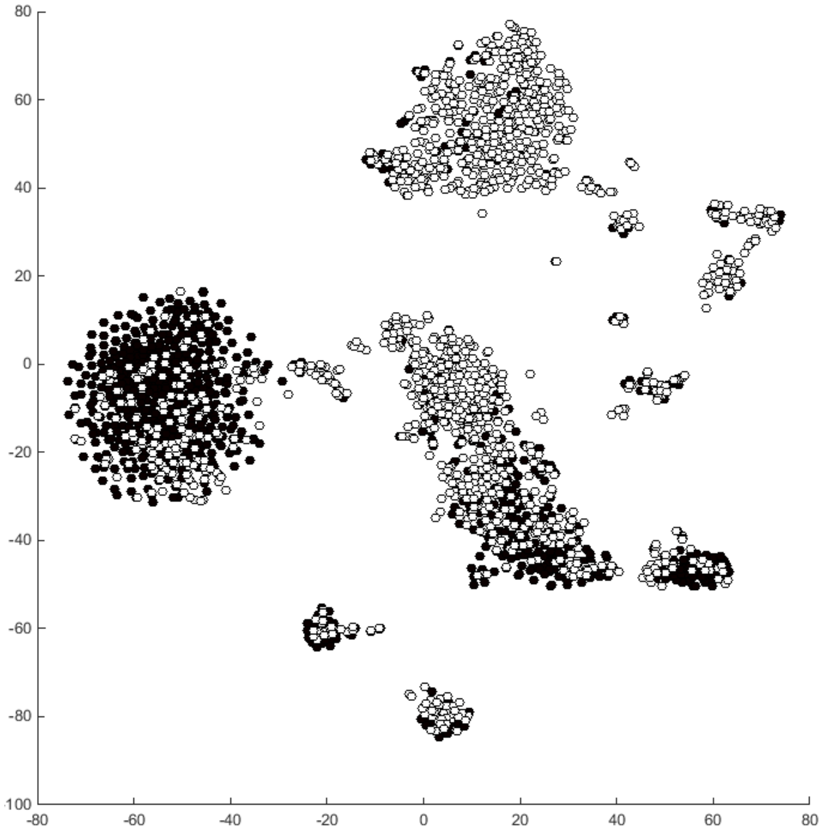




Contaminant from free floating RNA and leaky cell corpses

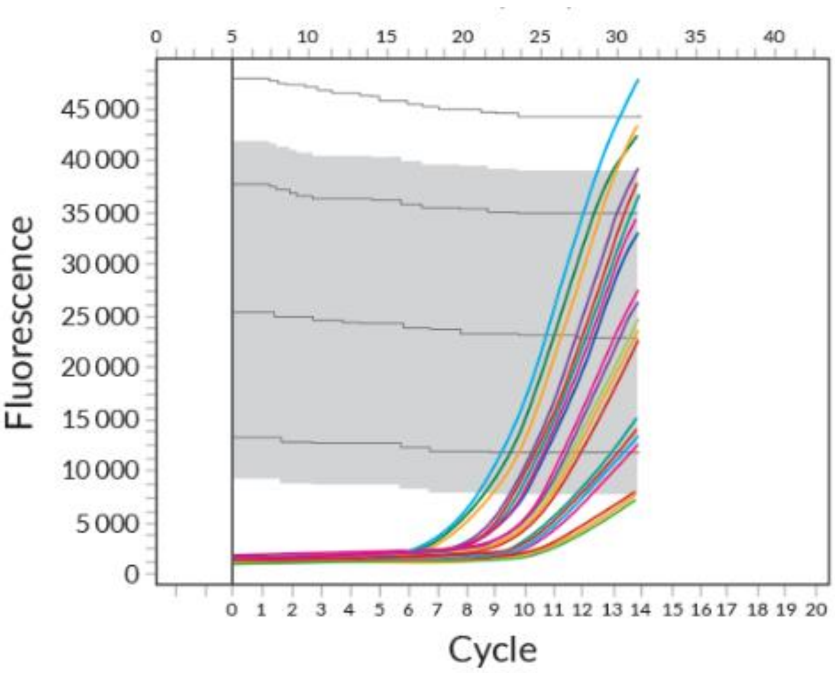


Generous threshold



Strict threshold

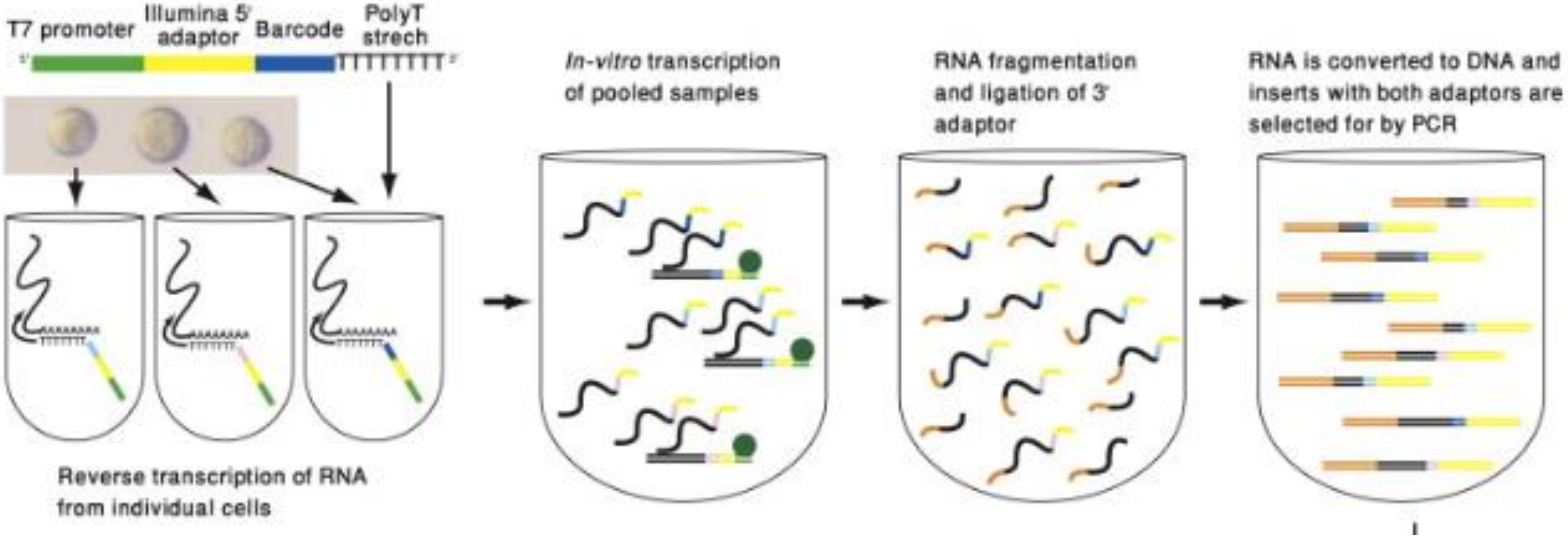
Amplification bias in scRNA-seq



Non-linear amplification

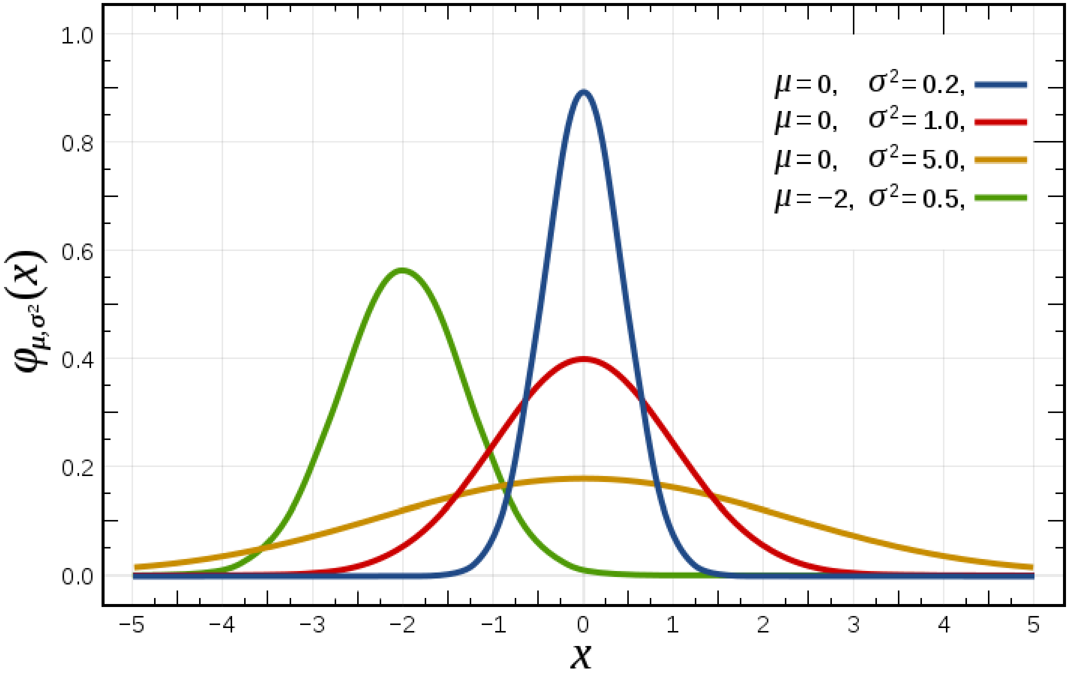
- Highly expressed transcripts are inappropriately represented and replicated
- Sampling of RNA in a cell 1-10%
- Zero inflated data
- Count data – negative binomial distribution

In vitro transcription results in linear amplification of RNA


T7 promoter
Illumina 5' adaptor
Barcode
PolyT strech
5'
TTTTTTTT 3'
Reverse transcription of RNA from individual cells
In-vitro transcription of pooled samples
RNA fragmentation and ligation of 3' adaptor
RNA is converted to DNA and inserts with both adaptors are selected for by PCR

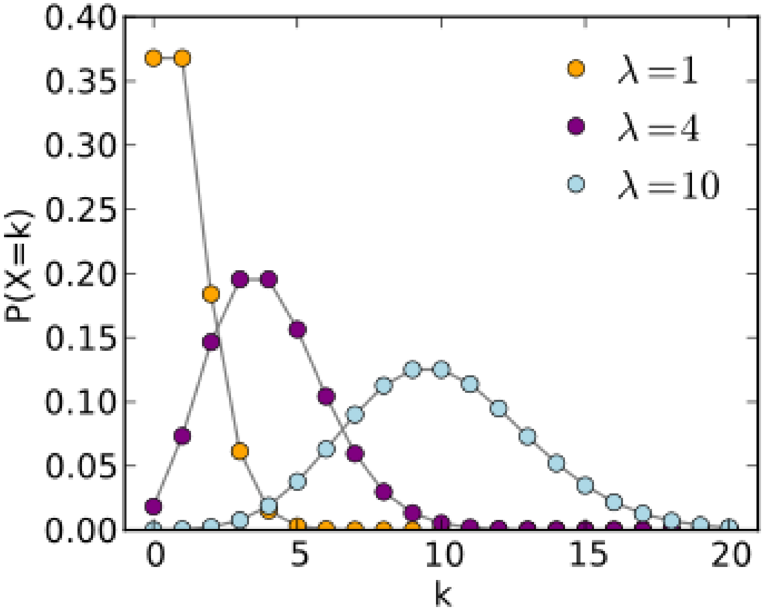
(Itai lab, CEL-seq)

Negative binomial versus Gaussian distributions



Gaussian

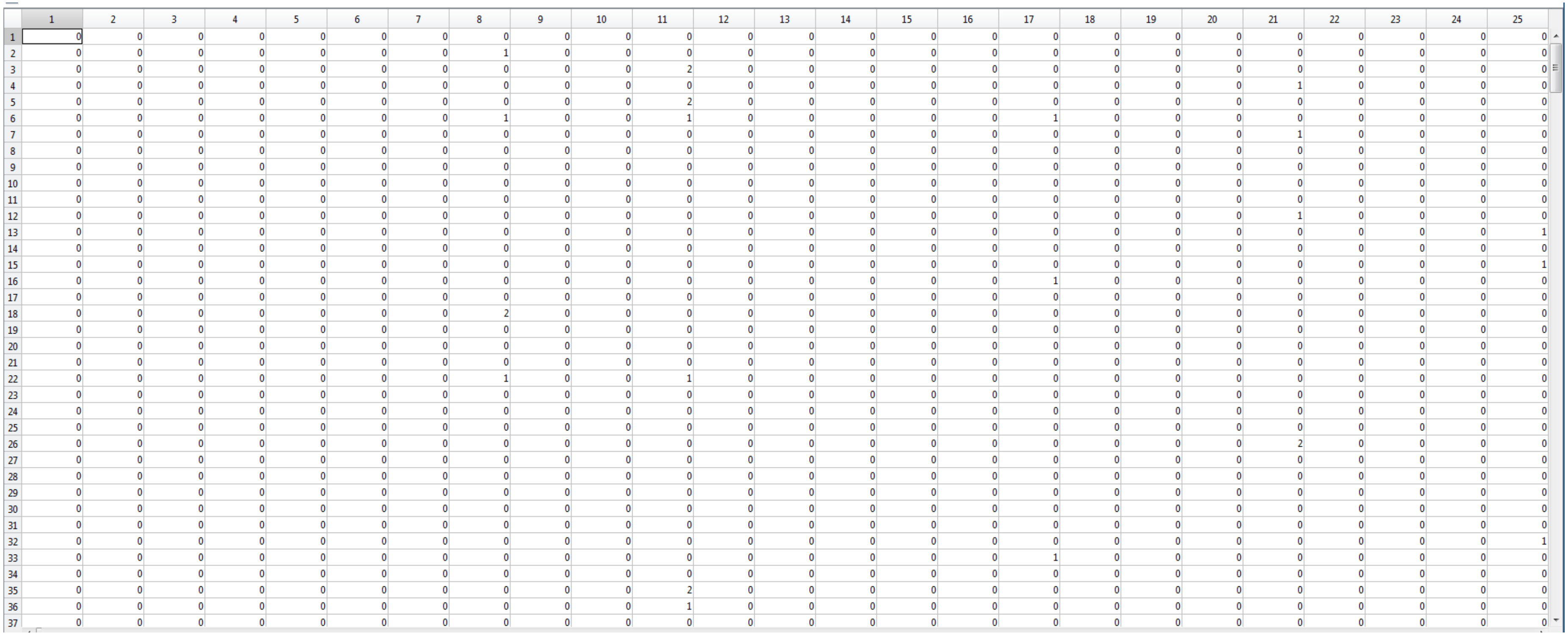
- Continuous variable
- Symmetric



Negative Binomial

- Discrete variable (counts)
- Asymmetric at small means

Zero inflation in scRNA-seq



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- Only 5% of this table is not 0!

What does this mean?

- Statistical tests to determine differentially expressed genes with a zero-inflated negative binomial (ZINB)

- False negatives (genes that are supposed to be expressed by appear as 0)

- Many genes (columns) have low counts due to shallow sampling of transcripts – this means the data are noisy

- Unreliable variables that need to be processed/filtered out prior to downstream analysis

Feature Selection

Variance selection

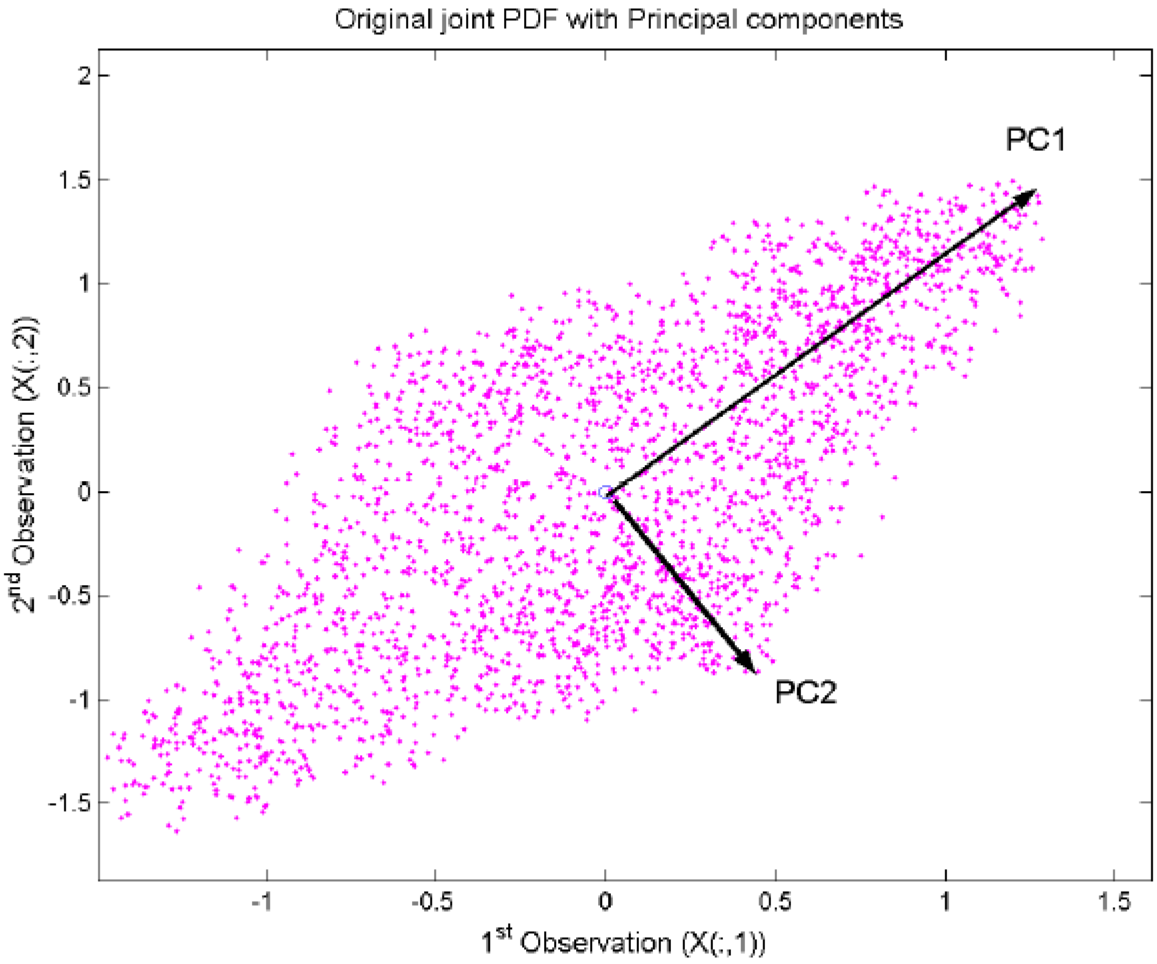
- Easiest – select genes that are the most variable

- Variance = genes that are most different across all cells

$$\sigma^2 = \sum \frac{(X - \mu)^2}{N}$$

- Rank genes by top 500 most variable, for example, discard the rest

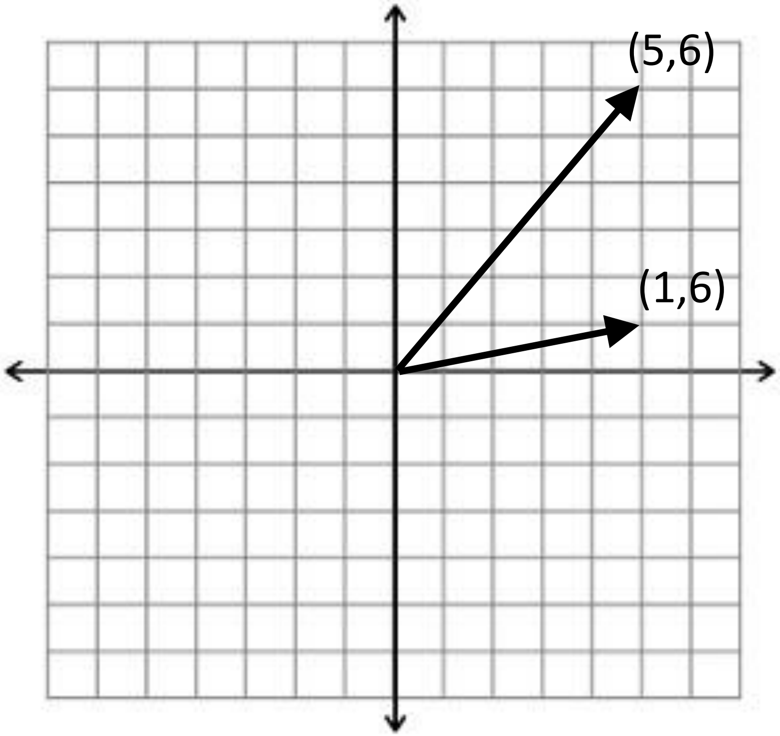
Highly variable axes = Principal Components



PCA (Principal Component Analysis)

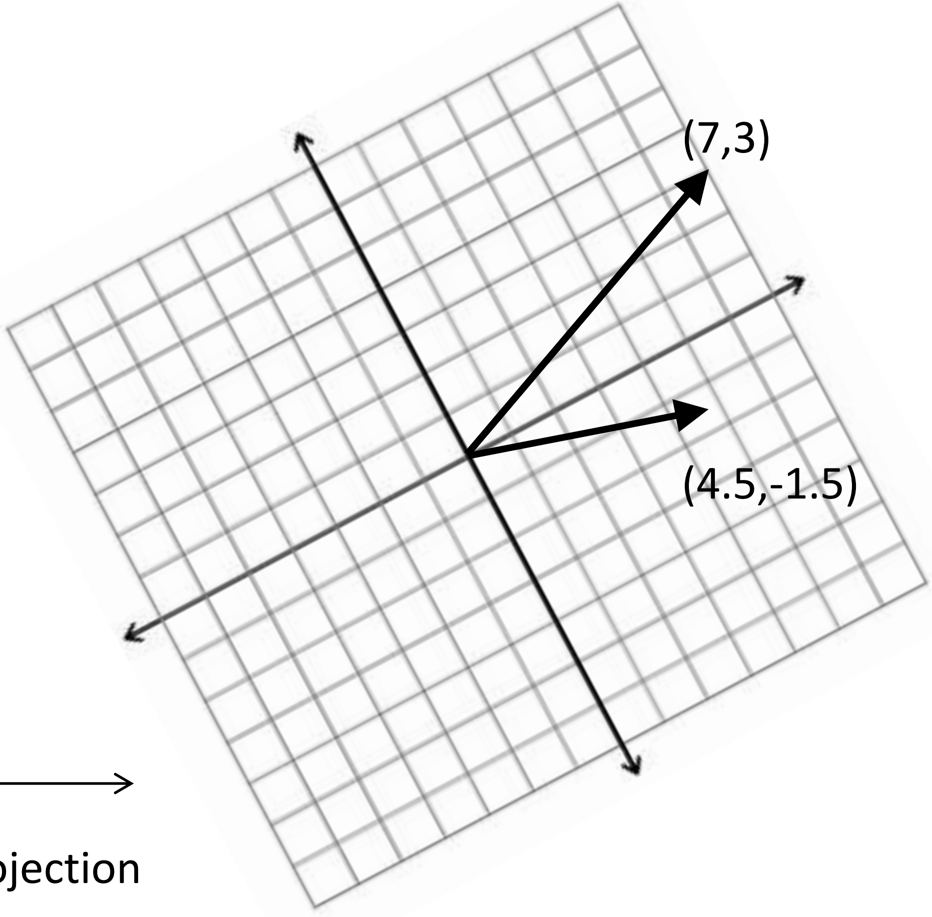
- Principle of PCA is to maximize the Variance of X with the least amount of principal components (latent variables)
- What is variance? Spread of the data, information content, change etc.
- Variance is the covariance of a dataset with itself, i.e. Var(X) = Cov(X,X) → Maximize
- What are principal components? Linear combinations of original variables – linear transformation

Vectors and projections



→

Projection



<u>Basis Set</u>

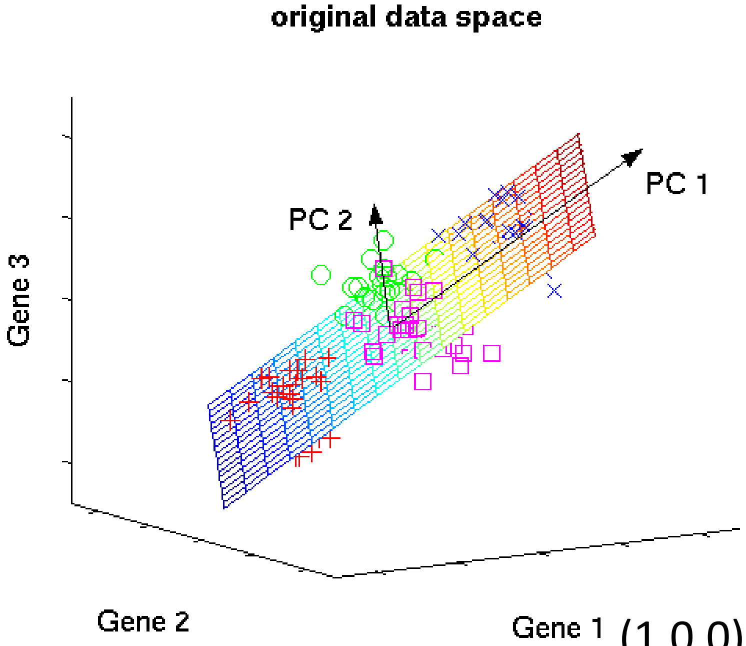
(1,0) →
(0,1) ↑

<u>Basis Set</u>

(7,4)
(-3.5,7)

PCA as a dimension reduction tool

(0,0,1)



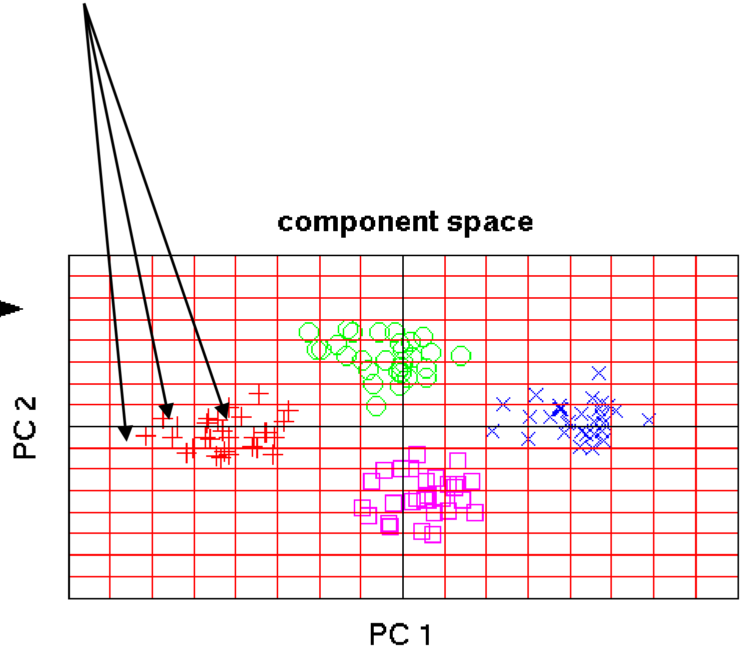
(0,1,0)

(1,0,0)

PCA

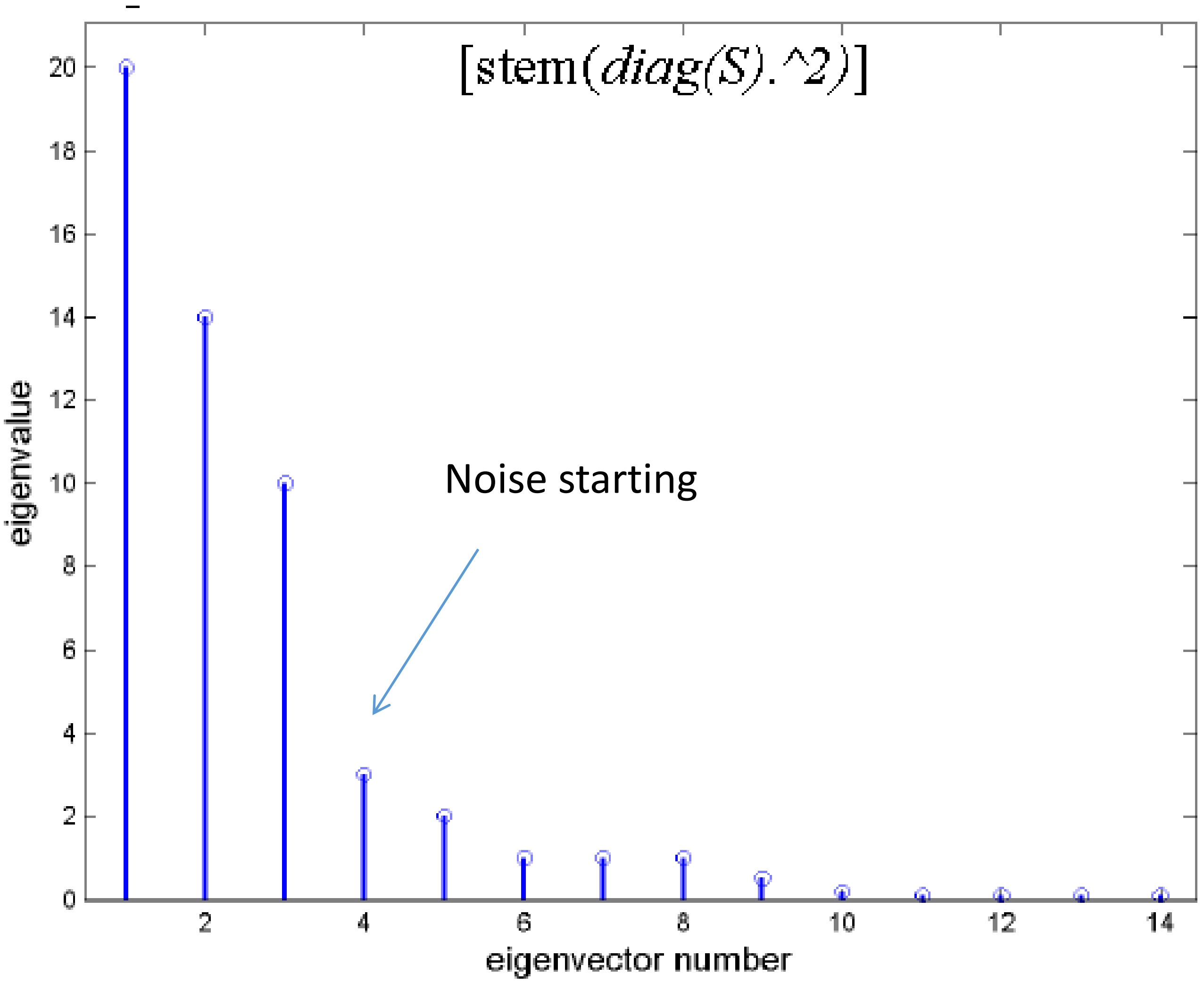
(0.3 Gene1, 0.2 Gene2, 0.5 Gene 3)

Score



(0.1 Gene1, 0.8 Gene2, 0.2 Gene 3)

Selection of principal components



<u>PRINCIPAL COMPONENTS CAN BE ORDERED BY EIGENVALUES (VARIANCE CAPTURED)</u>

Simple summary of this simple feature selection procedure

- Keep most variable genes (over all cells) for downstream analysis, discard rest as noise

- PCA identifies super axes (Principal Components) that are combinations of the original variables that captures the most variance in the data (by eigenvalues)

- Orthogonal – no duplicate or redundant axes – so will only have a few of them