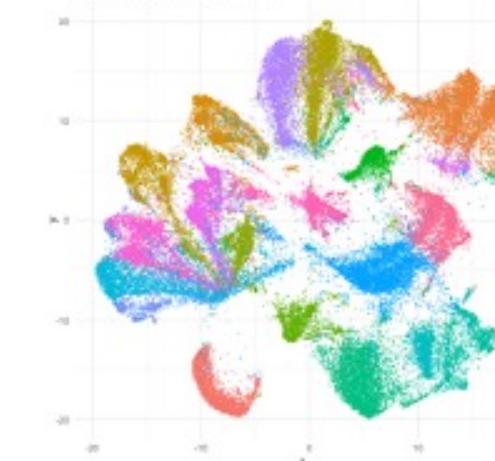


The Single Cell RNA-seq Workflow:

A practical guide to ensure experimental success



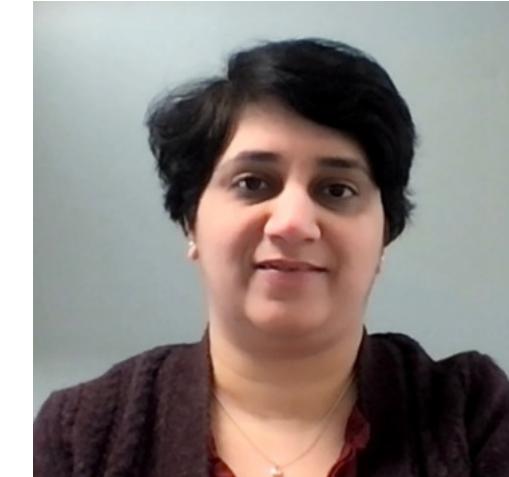
Arpita Kulkarni, Ph.D.
Associate Director, Single Cell Core, Harvard Medical School
arpita_kulkarni@hms.harvard.edu



HMS's Single Cell Core: The Team and Faculty Advisors



Mandovi Chatterjee
Director



Arpita Kulkarni
Assoc. Director



Pratyusha Bala
Assoc. Director
(Spatial Transcriptomics)



Alexa Yeagley
Research Assoc.



Dr. Ollie
scPawlice & Floof



Dr. Allon Klein



Dr. Jeff Moffitt



Dr. Chris Benoist



HMS's Single Cell Core: our mission

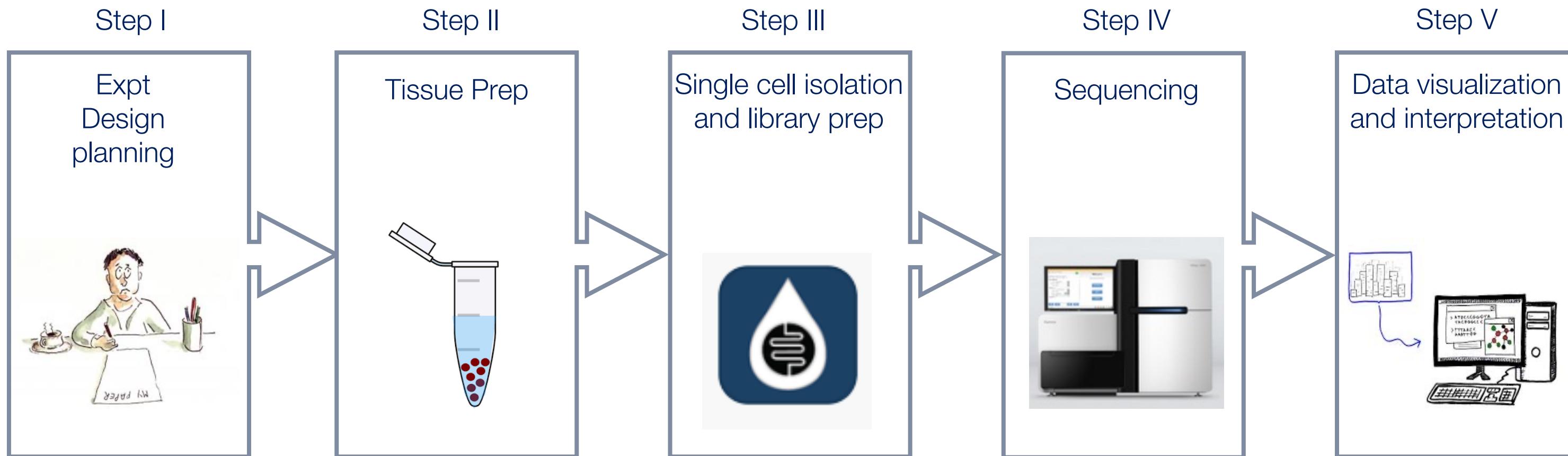
Enable discoveries by guiding in the design, execution & interpretation of single cell and spatial -omics assays

- We are one of the oldest single cell core's on campus
- Open to all, fee-for-service core
- Worked w/ >500 PI's and 50,000+ samples to date



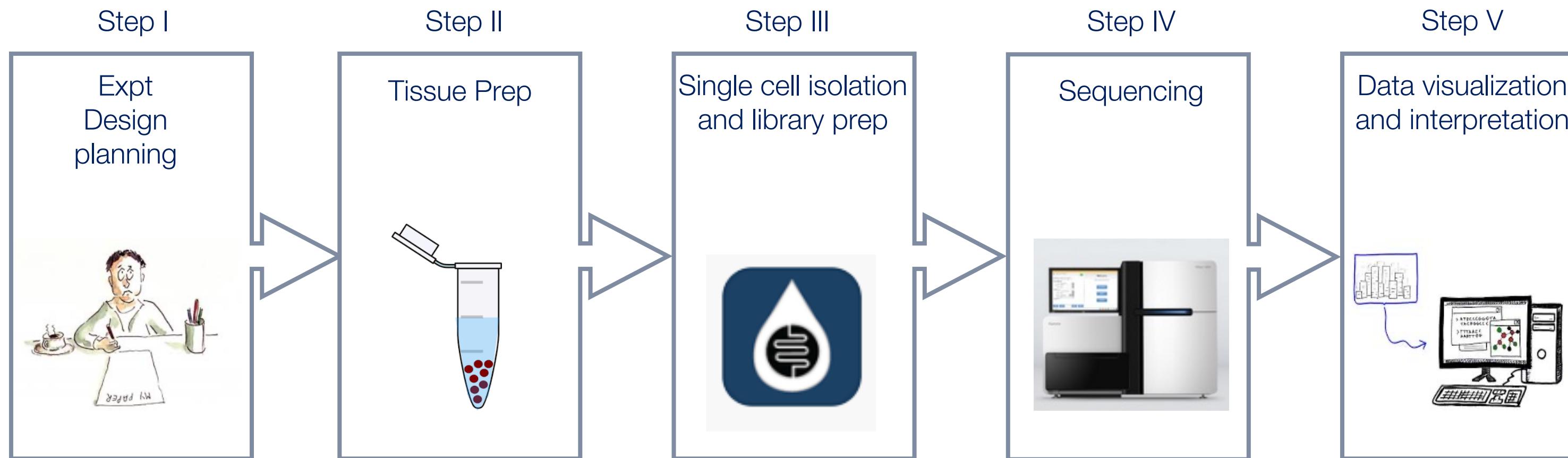
Outline for today's talk (~40mins)

- A background into scRNAseq
- What good scRNAseq data looks like (defining quality metrics for goal setting)
- The scRNAseq workflow

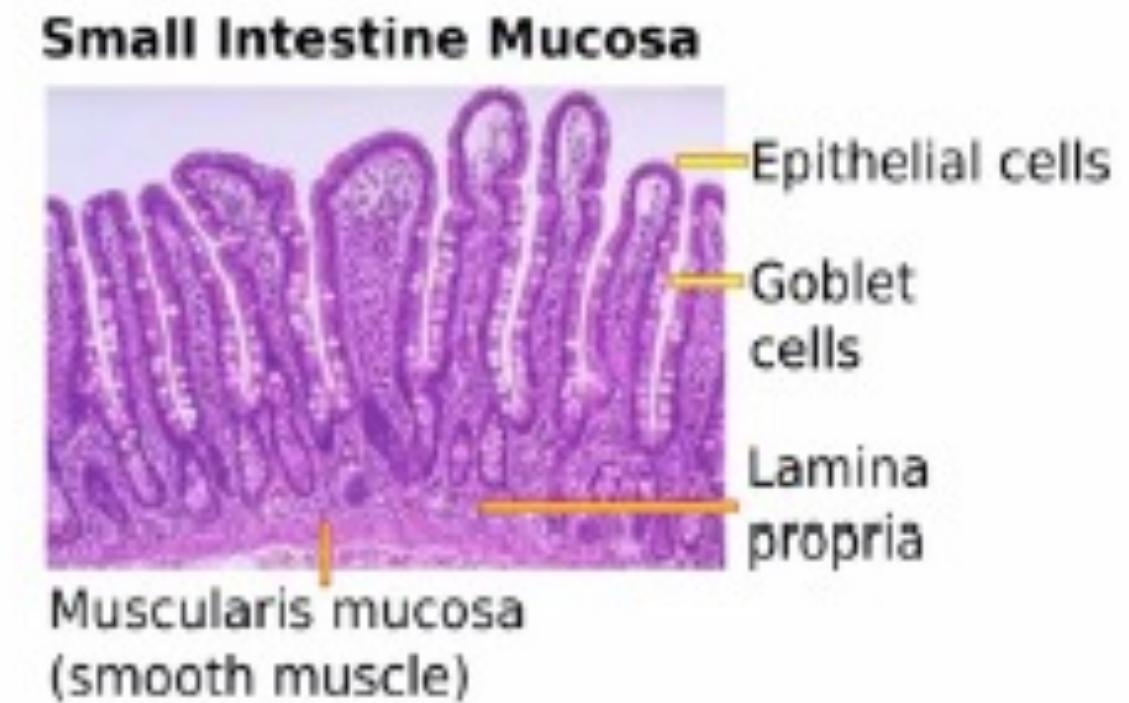
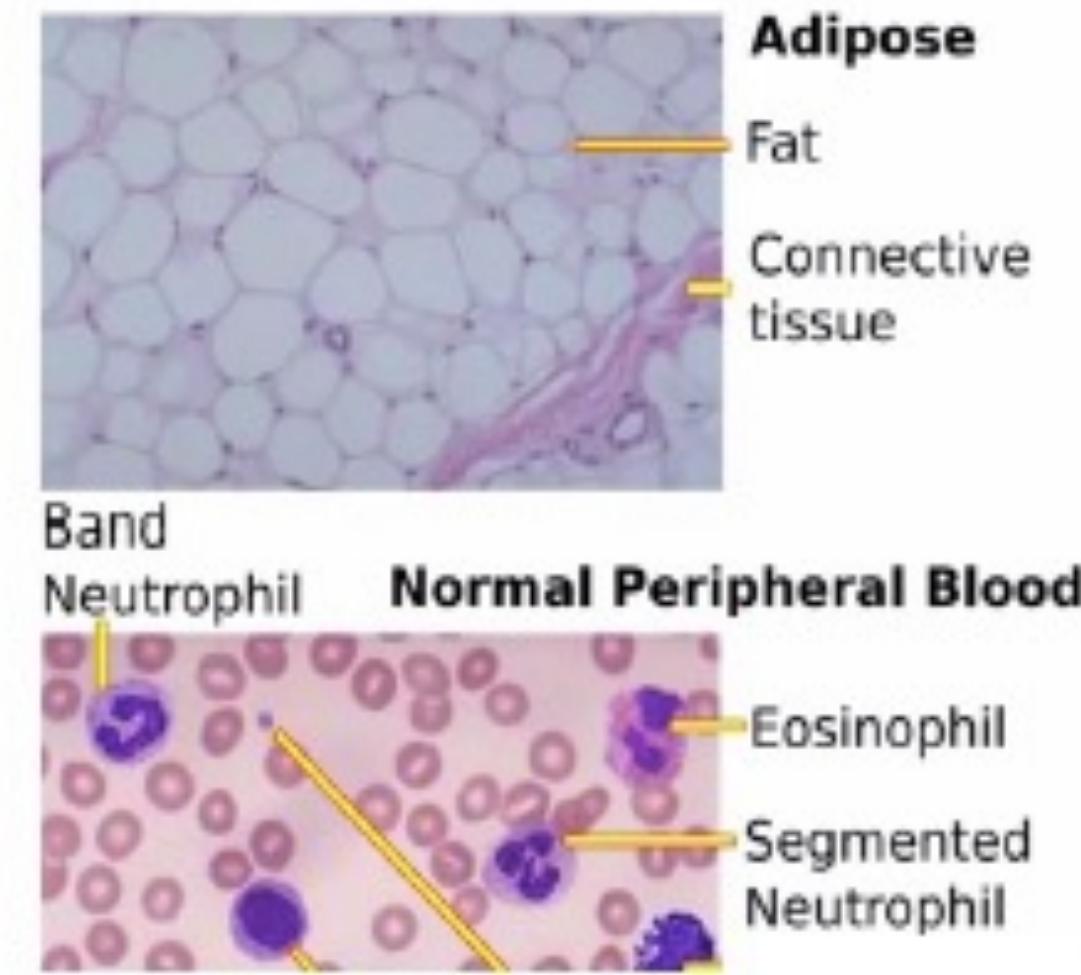
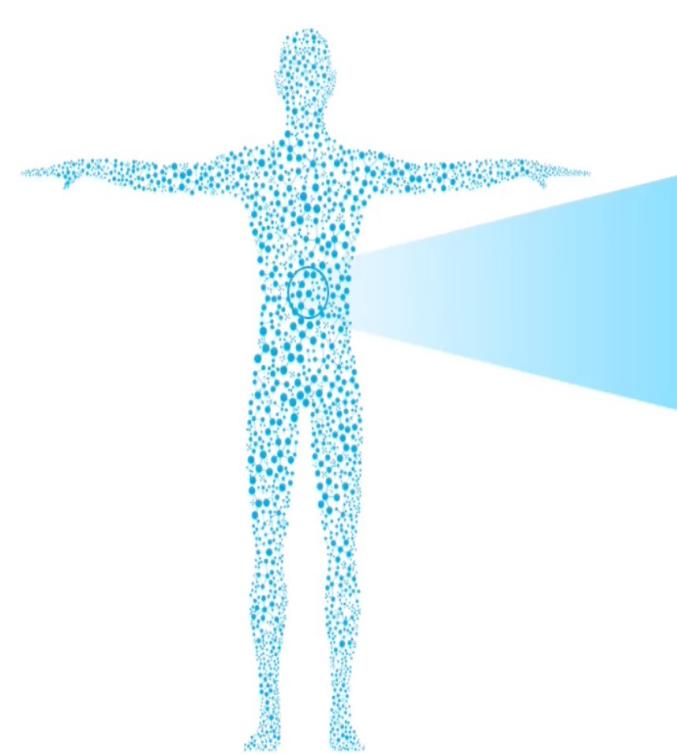


Outline for today's talk (~40mins)

1. This is an “overview” type talk (not comprehensive given the scope and time constraints)
2. Geared towards beginners, to reduce barrier of entry



We know tissues are heterogeneous

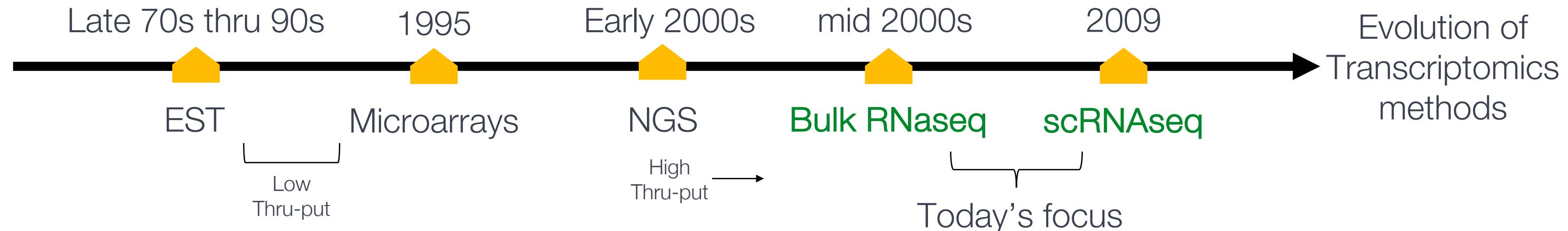


Q. What makes cells (& tissues) different from one another? How do we measure these differences?

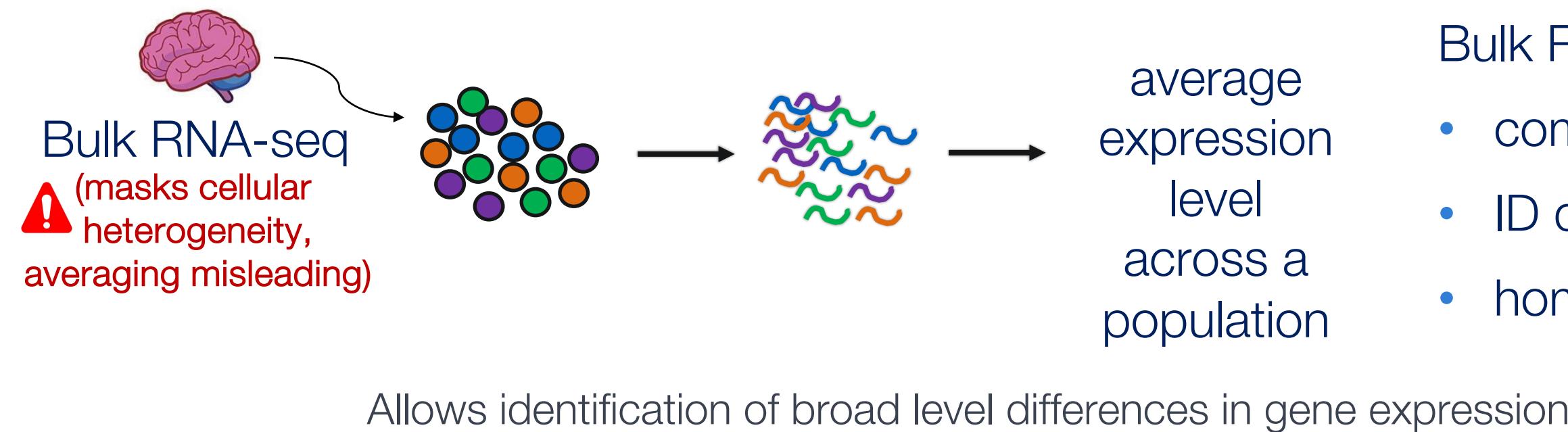
Transcriptomics: Unraveling the nature of cellular identity

DNA → RNA → Protein

Measuring (m)RNA “transcriptomics” is a reasonable & powerful proxy to unravel cellular identity



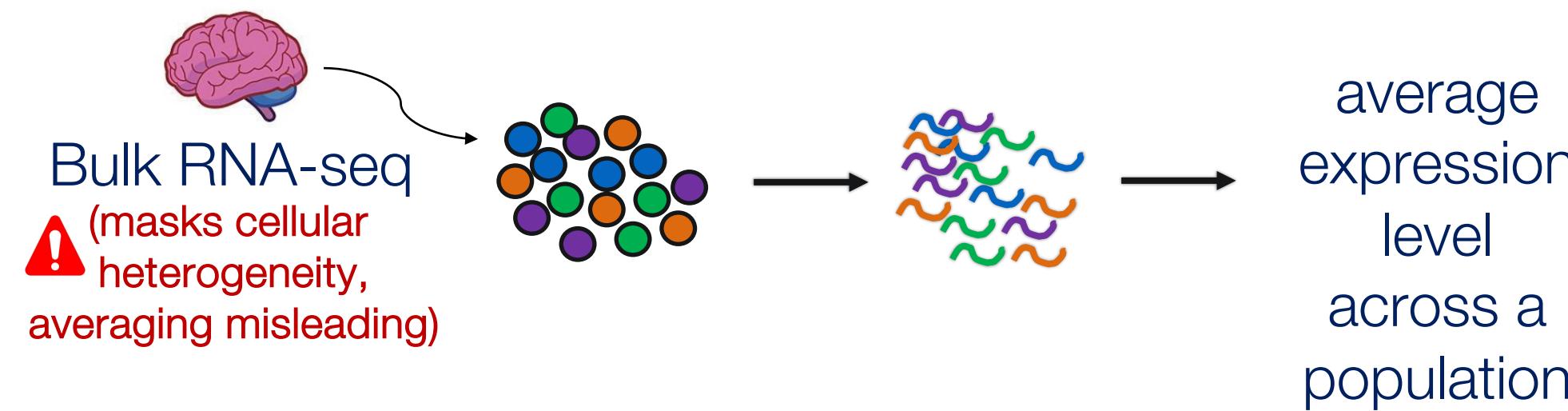
Bulk vs single cell RNA sequencing (scRNA-seq)



Bulk RNA-seq good for -

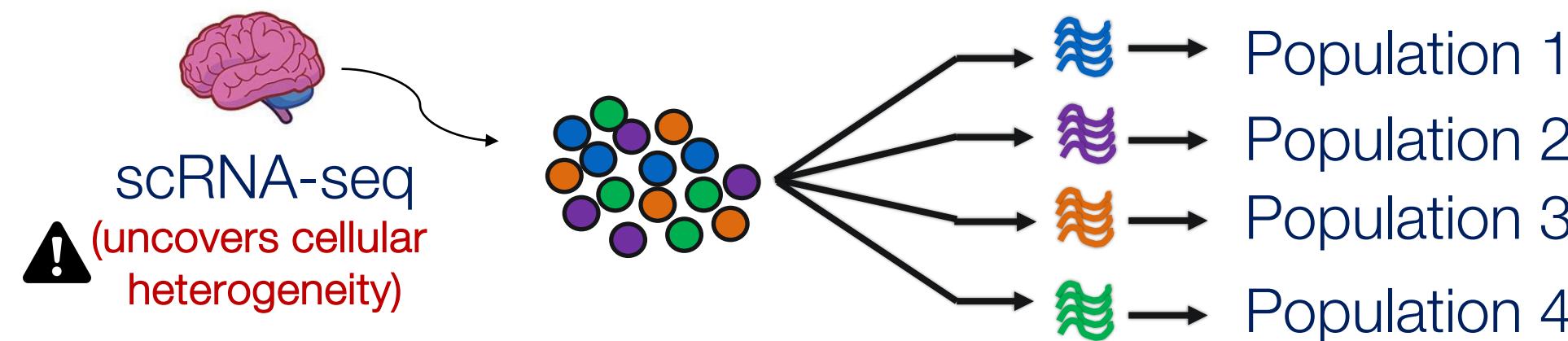
- comparative transcriptomics
- ID disease biomarkers
- homogenous systems

Bulk vs single cell RNA sequencing (scRNA-seq)



Bulk RNA-seq good for -

- comparative transcriptomics
- ID disease biomarkers
- homogenous systems



scRNA-seq good for -

- defining heterogeneity
- identify rare cell population(s)
- cell population dynamics

Allows identification of cell to cell variation in gene expression

Bulk vs scRNA-seq: a difference of resolution



smoothie

Average expression level
- Comparative transcriptomics
- Disease biomarker
- Homogenous systems



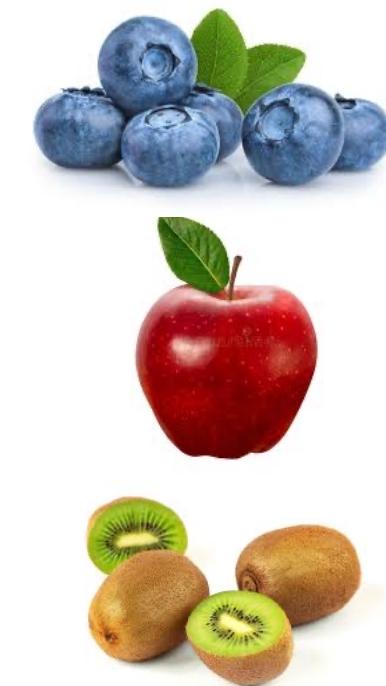
Bulk RNA-seq



Mix-Fruit salad



scRNA-seq



Individual components

Separate populations
- Define heterogeneity
- Identify rare cell populations
- Cell population dynamics

Which technique to use when?

Bulk vs scRNA-seq: not always an either/or situation

Cost effective, good quality data



smoothie

Bulk RNaseq as a –

1. Complementary first step
2. For homogenous systems
3. Broad level differences



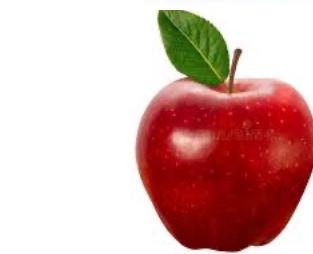
Bulk RNA-seq



Mix-Fruit salad



scRNA-seq



Individual components

- Separate populations
- Define heterogeneity
 - Identify rare cell populations
 - Cell population dynamics

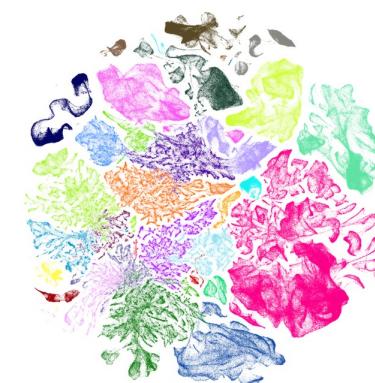
Common applications of scRNA-seq – why use scRNaseq?

a) "cell atlas"-type studies

- Heterogeneous populations

Uncover cellular heterogeneity

e.g. Allen brain atlas,
Tumor environment etc



b) "timeseries"-type studies

- Snapshots in biol. process

Bio. trajectories/cell fate,
Dev timelines,
lineage tracing

e.g. embryogenesis

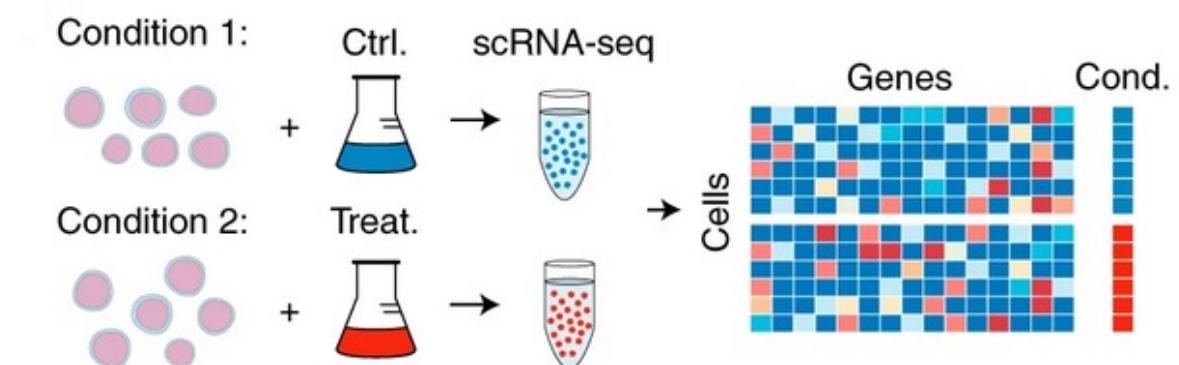


c) "screening"-type studies

- Single cells as individual expt.

Uncover GEX diff on perturbation

e.g. CRISPR studies,
Therapeutics discovery



Defining metrics for good quality scRNAseq data

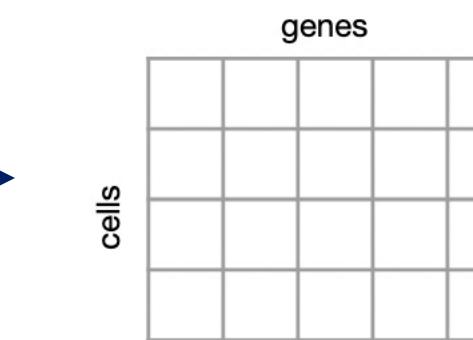
scRNAseq final goals -

- Gain novel biological insights
- publish & advancement of scientific work

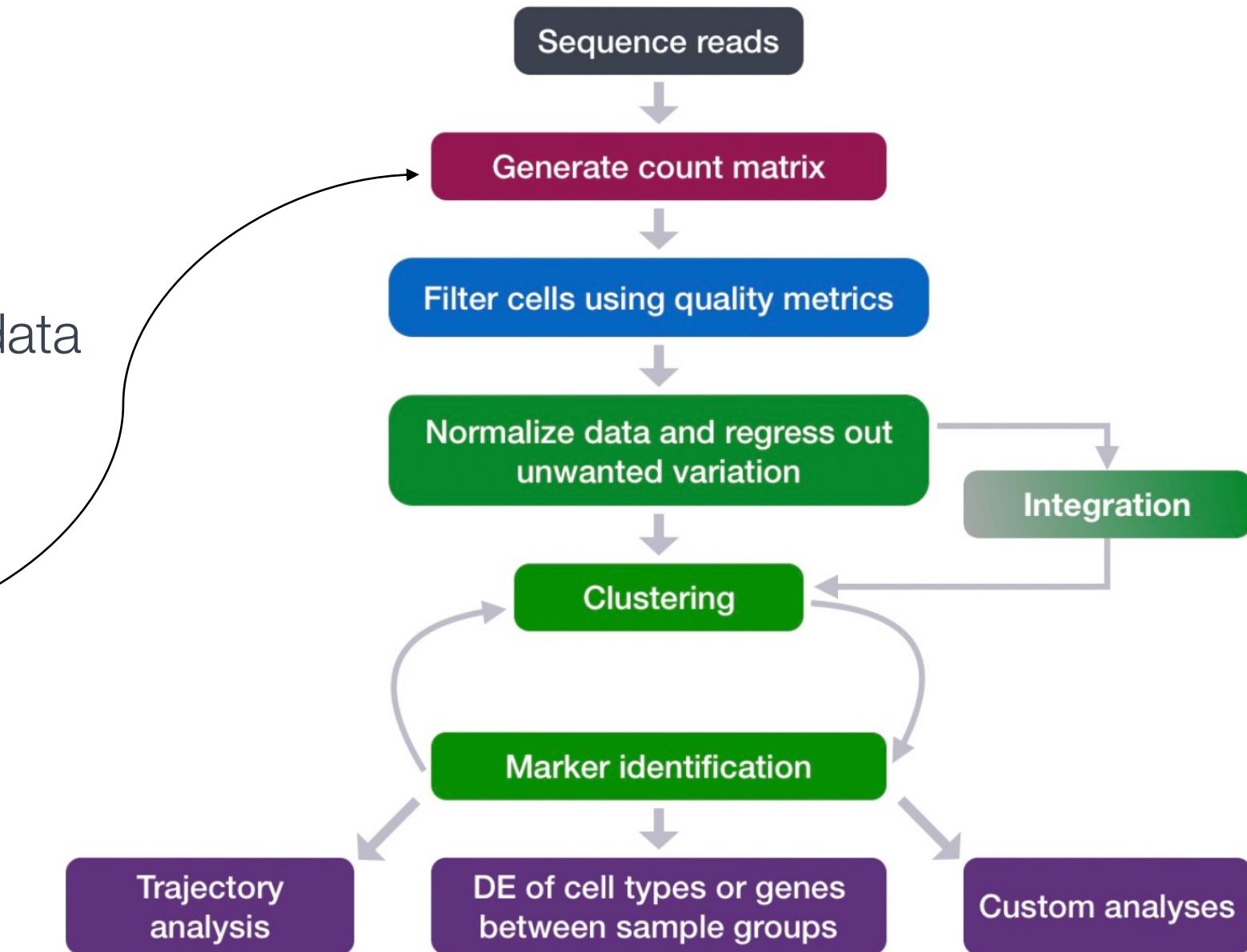
So lets imagine, you've done the expt and have the data



INPUT: cells



OUTPUT: gene expression matrix



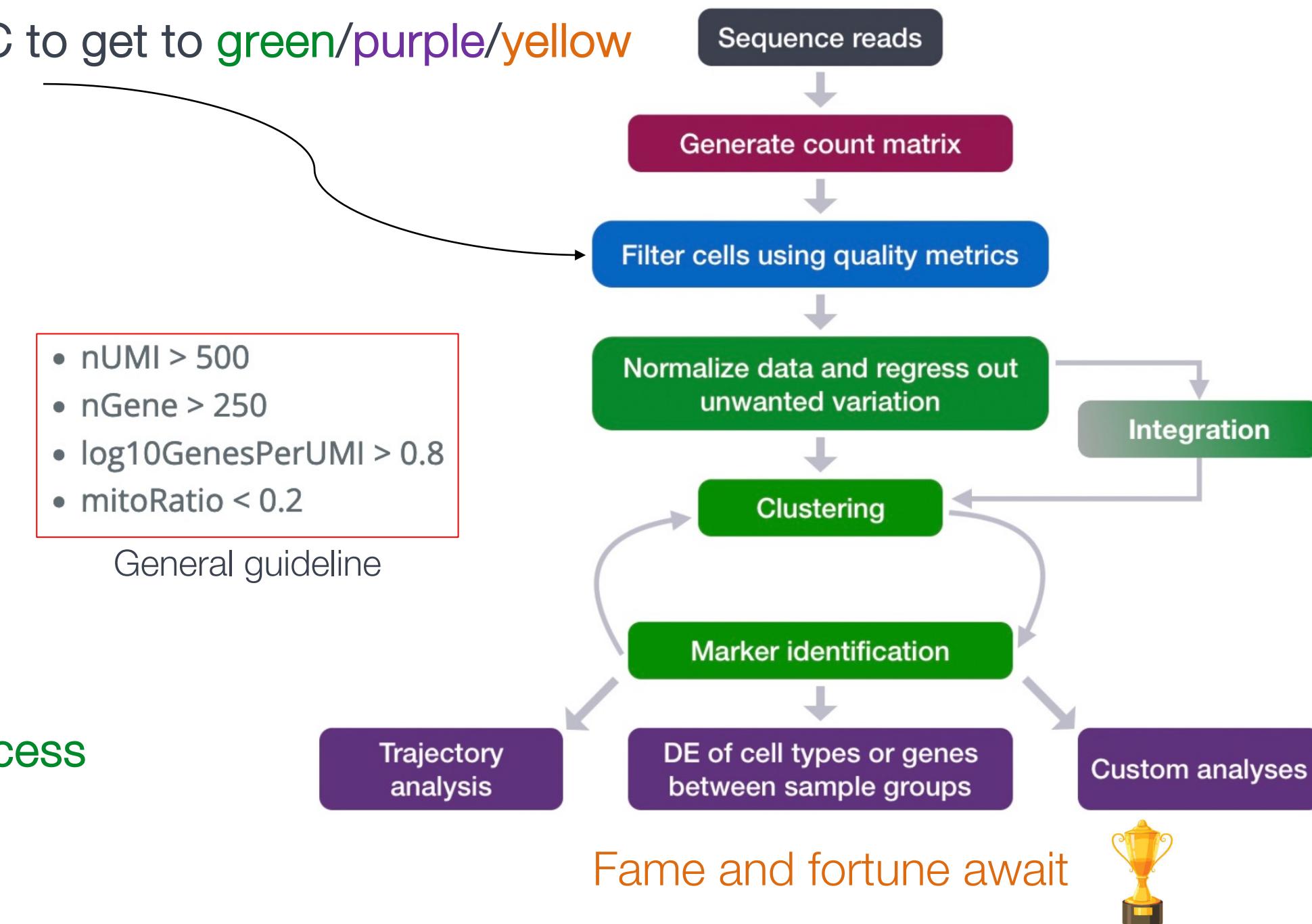
Defining metrics for good quality scRNAseq data

Filtering: Cells have to pass this first QC to get to green/purple/yellow

Quality metrics -

1. Cell counts
2. nUMI counts (transcripts) per cell and nGenes per cell
3. Mitochondrial counts ratio
4. Novelty or complexity score

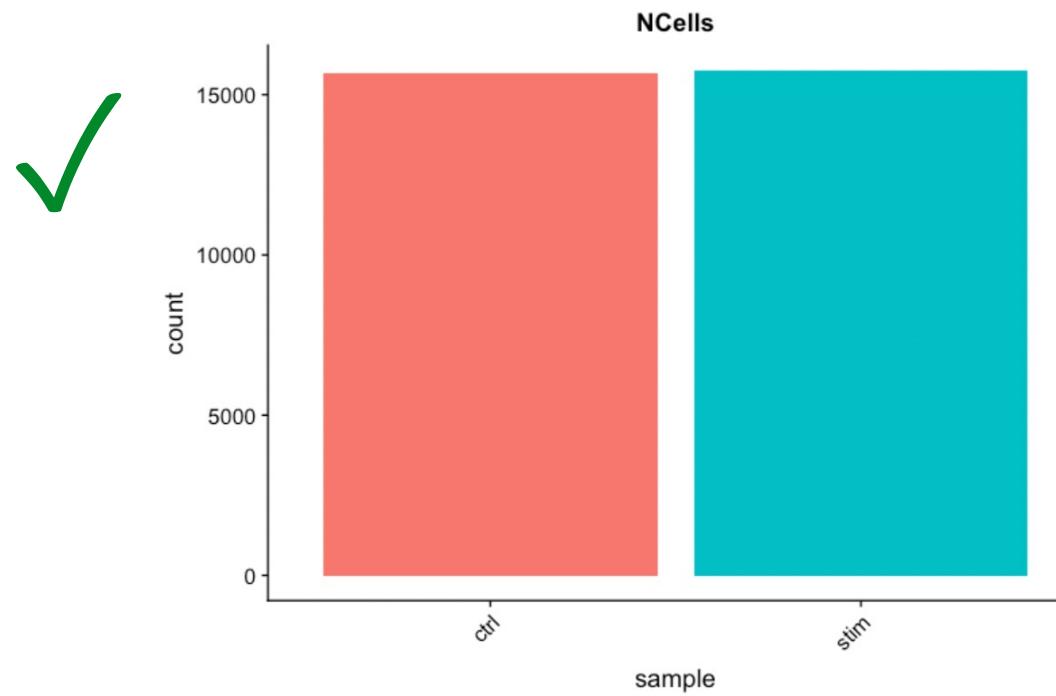
- no metric in isolation
- setting thresholds is an iterative process (starting from lenient to stringent)



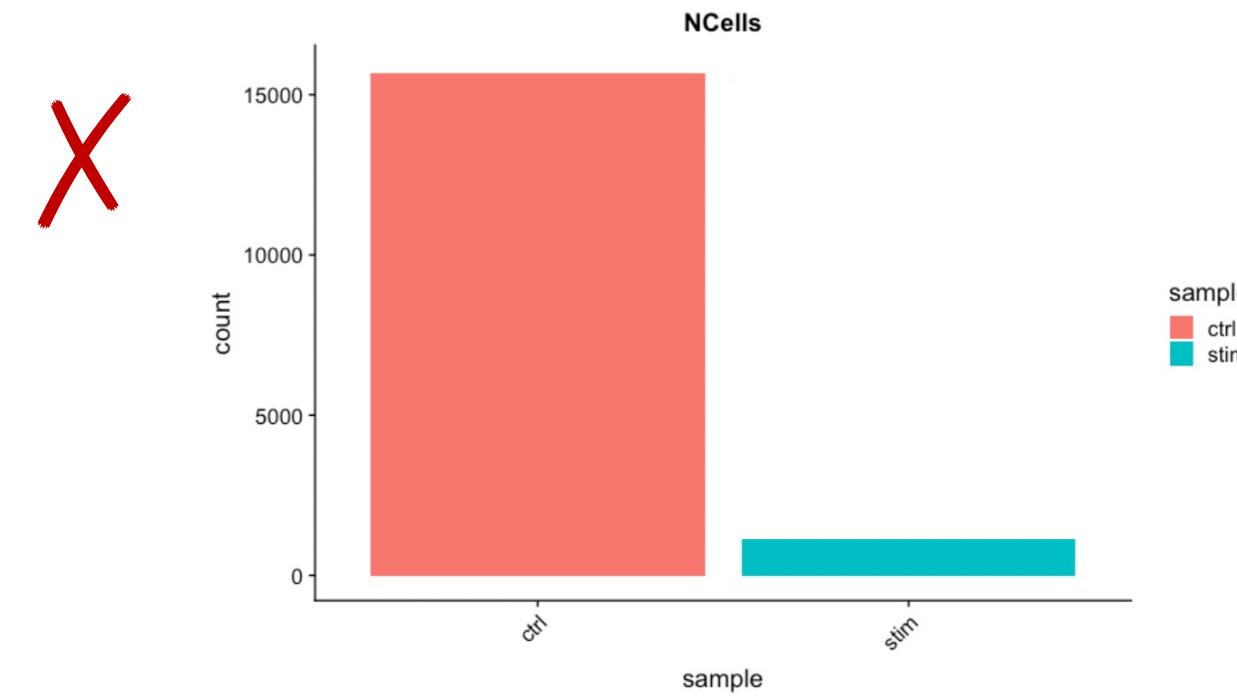
Defining metrics for good quality scRNAseq data

Assessing quality metrics –

1. Cell counts (how many unique cellular barcodes detected, *ideally* same as #cells you started with, but 60-80% cell recovery is good)



E.g. You start w/ 20K cells,
and recover 15K
cell barcodes – **75% recovery!**

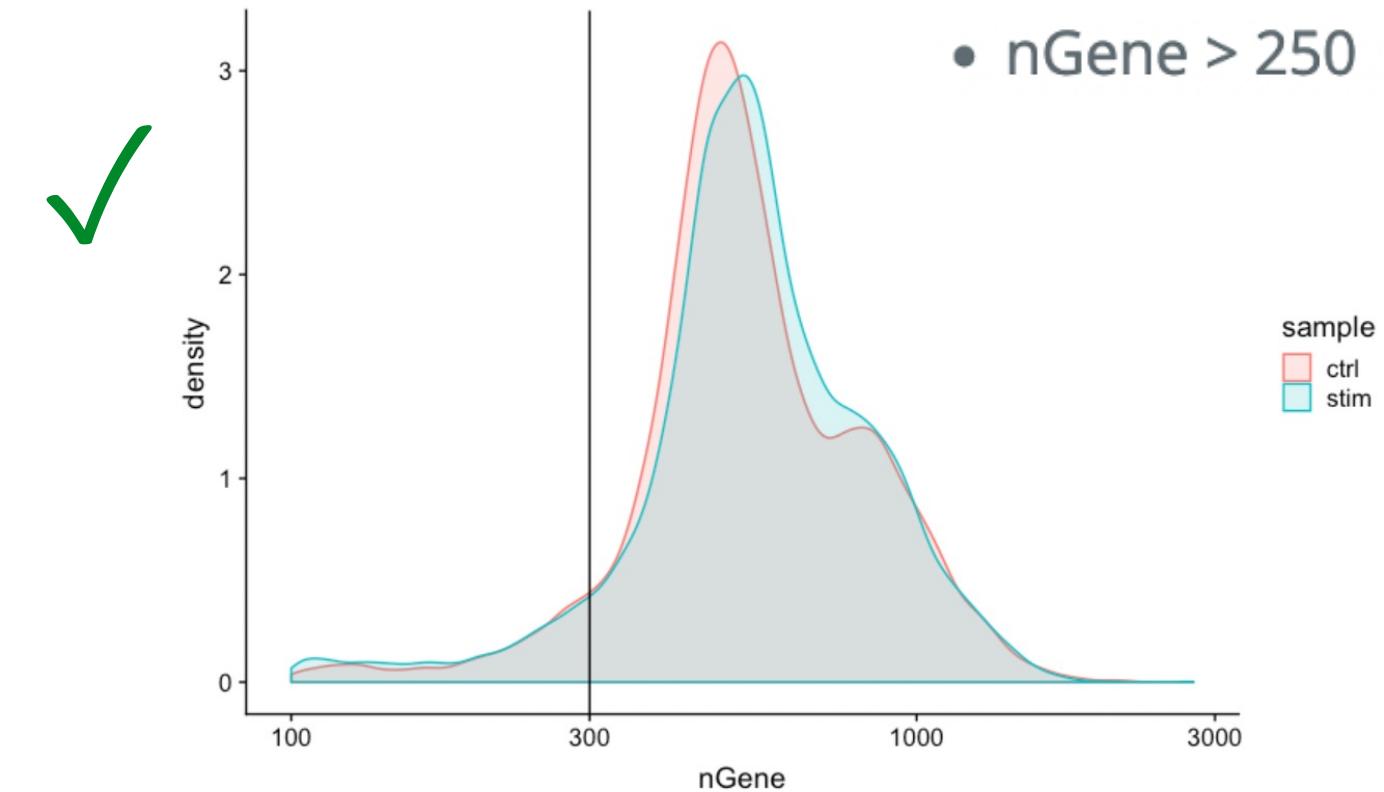
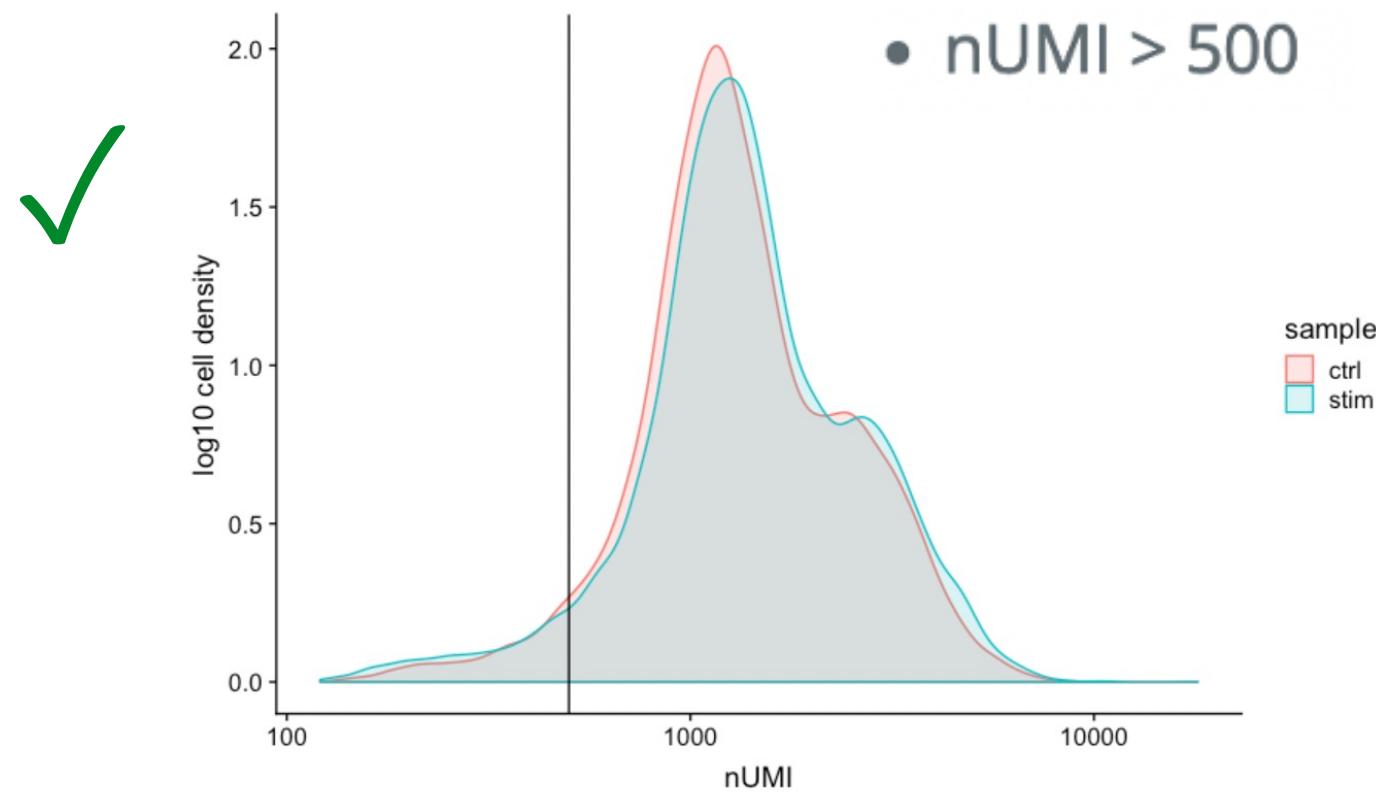


E.g. 1. You start w/ 10K cells, and recover 15K cell barcodes
2. You start w/ 10K and recover 1K
Was cell count wrong? Is it an issue w/ barcodes?
Likely that junk “cells” present

Defining metrics for good quality scRNAseq data

Assessing quality metrics –

2. UMI (transcripts) per cell and genes per cell



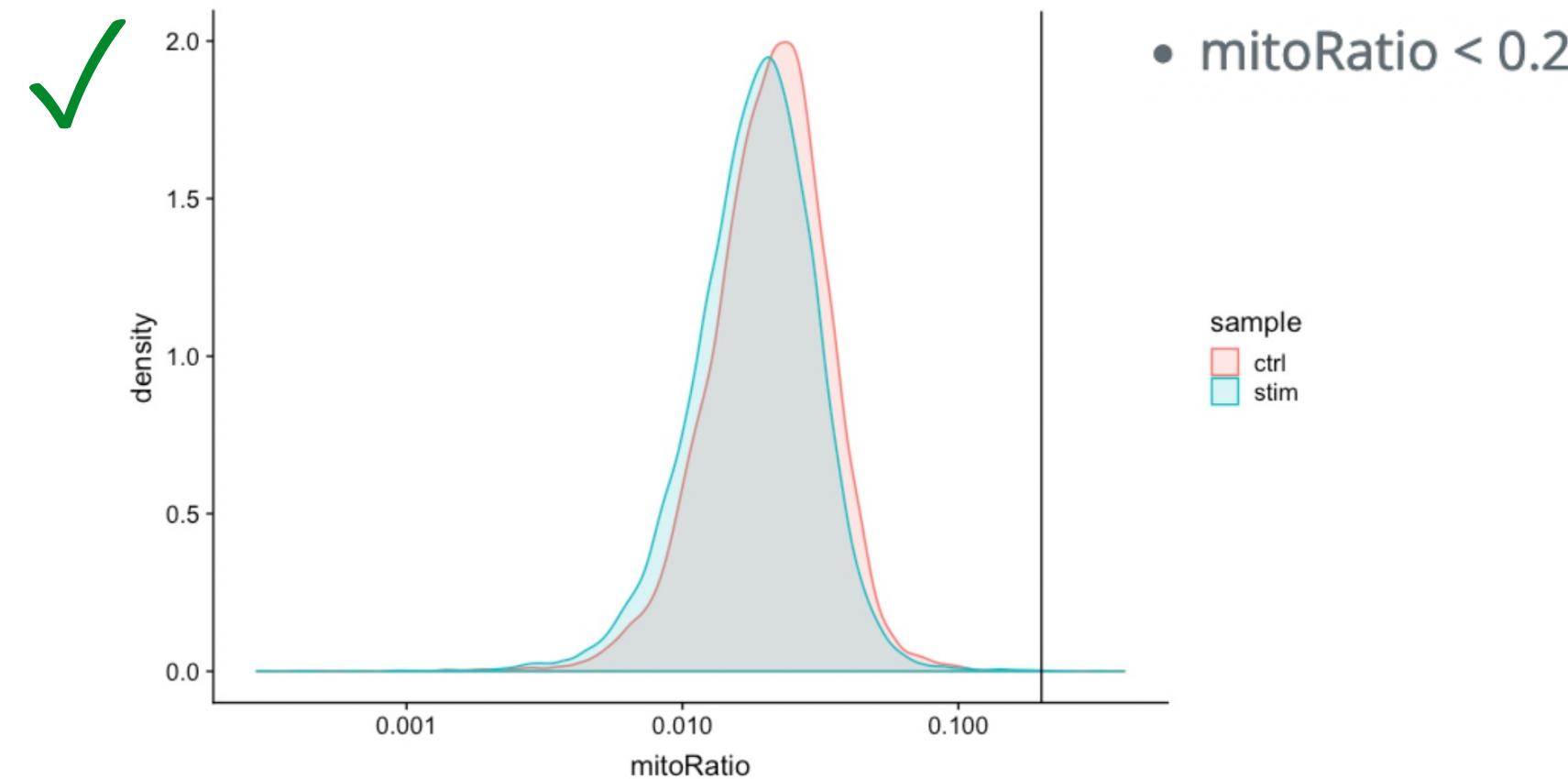
UMIs <500 is bad, between 500-1000/cell is usable, but 1000+ is great!

V low or high UMIs and genes per cell can indicate technical artifacts during prep, or poor quality cells

Defining metrics for good quality scRNAseq data

Assessing quality metrics –

3. Mitochondrial counts ratio (unless expected, high counts indicative of unhealthy, dying or lysed cells)

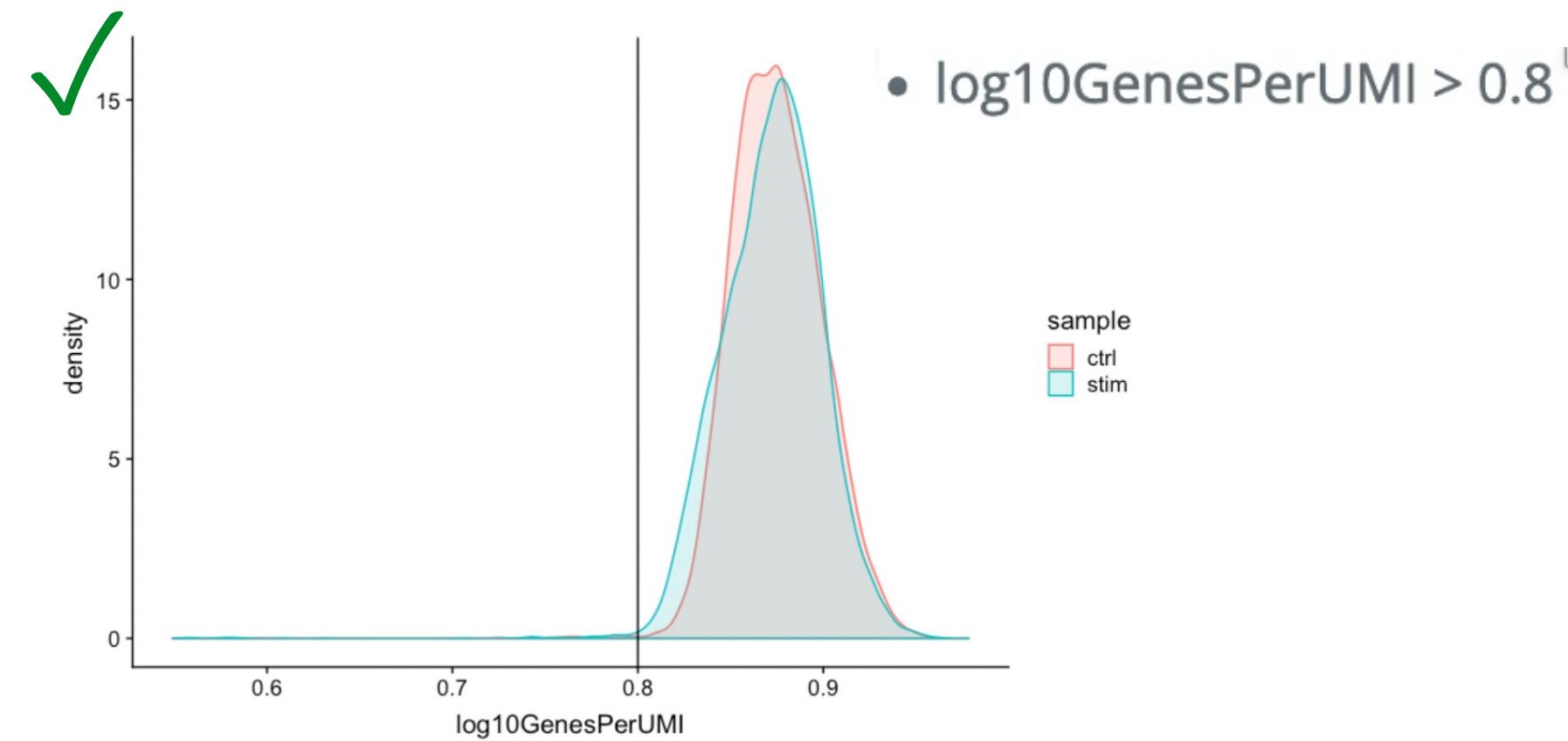


Mitochondrial read counts >0.2 , when sample prep bad

Defining metrics for good quality scRNAseq data

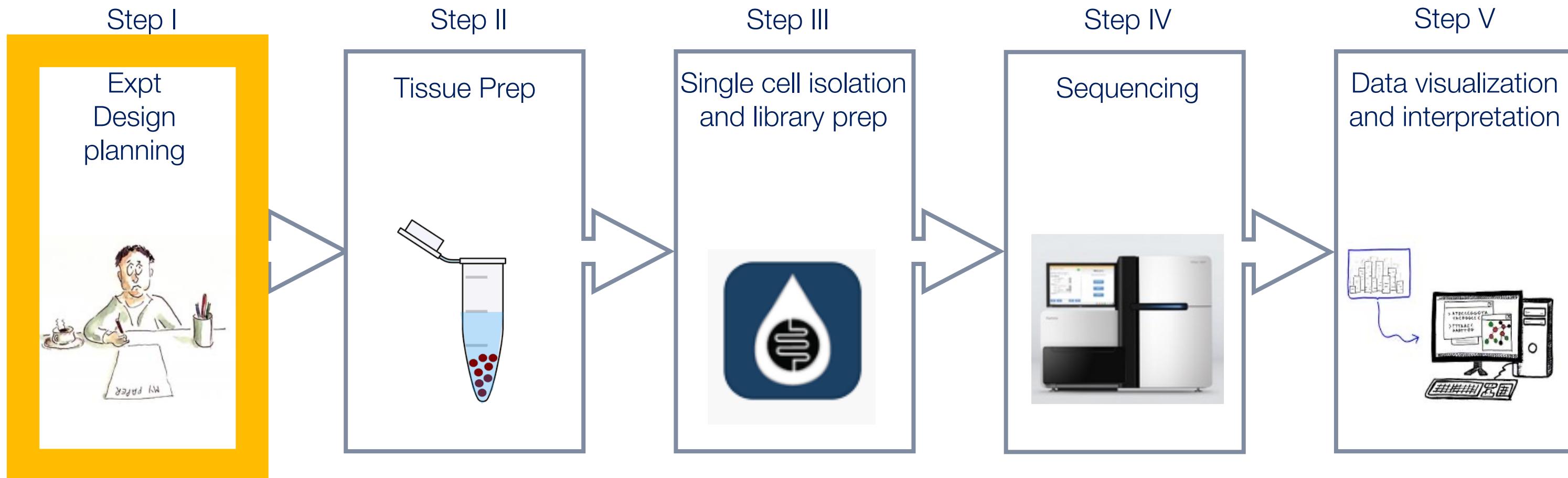
Assessing quality metrics –

4. Novelty or complexity score (more genes detected per UMI, more complex the data, in terms of RNA species)



(low scores indicative of no seq saturation, contamination (if multimodal peaks present), and low complexity cells at sample prep)

Steps in a scRNAseq work flow

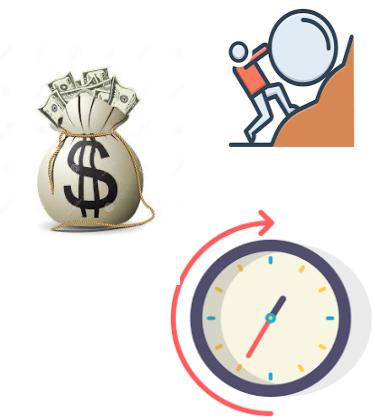


Goal: Put a well-thought-out, holistic plan in place, because single cell expts require clarity from the get-go

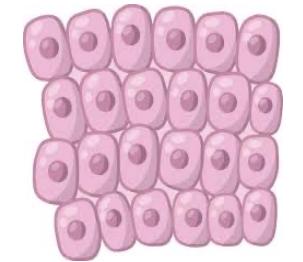
Experimental design considerations: taking stock



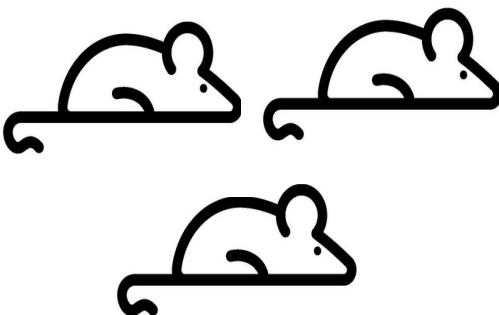
End goal:
Hypothesis testing
Publication
Grant-writing
New study or conti



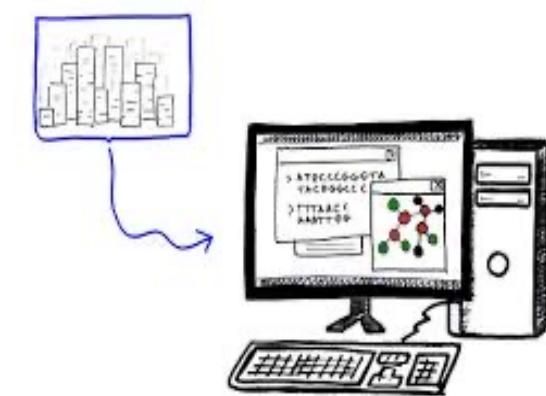
Resources:
Funding/budget
Time
Effort/Team



Sample type:
Cells vs nuclei
Fresh/frozen/fixed
Abundance
Access

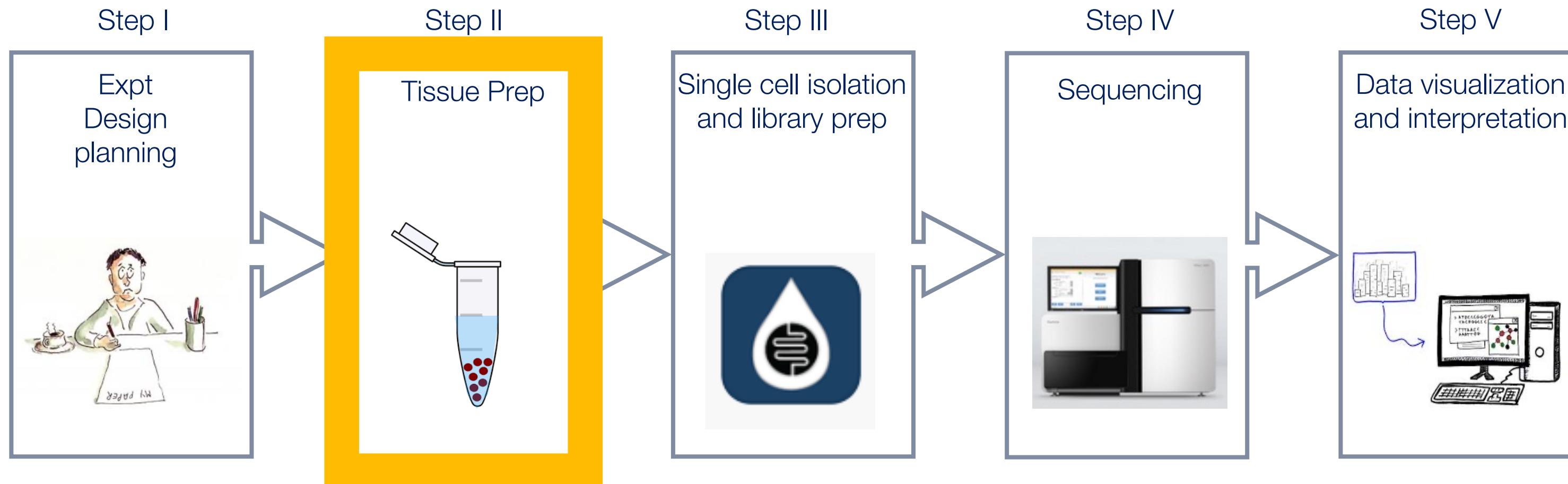


Scale:
Technical &
biological replicates
Comparative groups
Platform choice
#cells barcoded



**Bioinformatics &
analyses capabilities:**
Cloud storage
Computing power
Bioinformaticians
Analysis pipelines
Sequencing costs

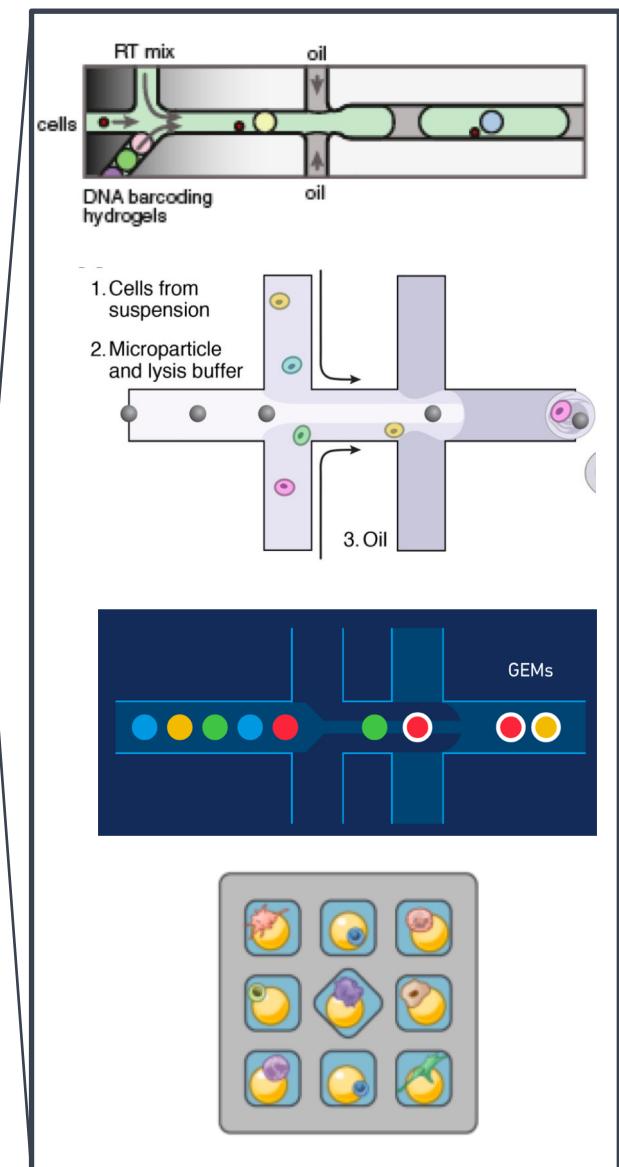
Steps in a scRNAseq work flow



Goal: Get high quality, viable, single cell suspension from sample, assess prep and perform QC

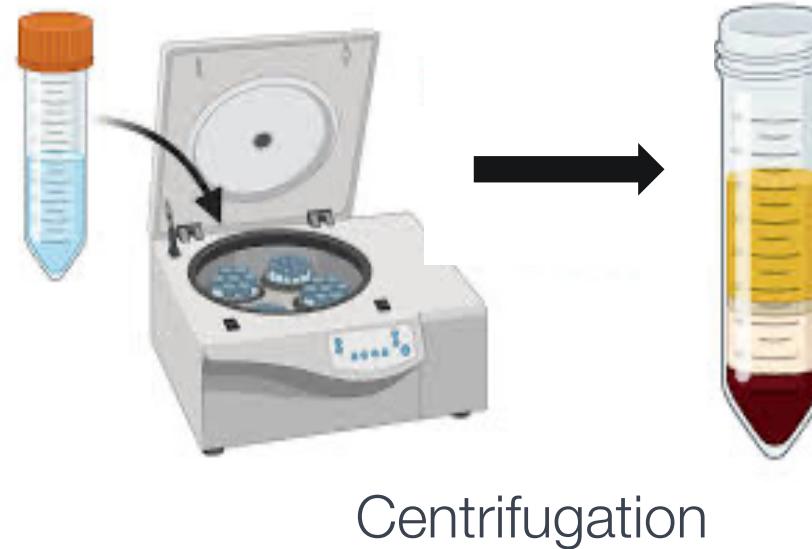
Sample prep protocol varies by cell type and logistics

- Enzyme-based dissociation- trypsin, collagenase, liberase, accutase.
 - Gentle washes.
 - Dead cell removal kit, filtering out the debris.
 - Density gradient (Ficoll, Optiprep)
- Solid tissue**
- Adherent cell culture**
- Suspended cell culture**
- Liquid tissue**
- Dissociation**
- Enrichment**
- Quality check**
- Note: the final sample buffer should be devoid of heparin, Ca^{2+} , Mg^{2+} and EDTA
- Magnetic bead selection (MACS)
 - FAC sorting
- Cell concentration
 - Cell Viability (Trypan Blue)



Dissociation: making a single cell suspension

1. Mechanical methods – cutting, shearing, laser dissections, FACS



Centrifugation



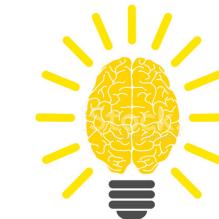
Vortexing



Pipette mixing

No universal protocol

Requires optimization of protocol for every tissue/cell type



2. Enzymatic dissociation



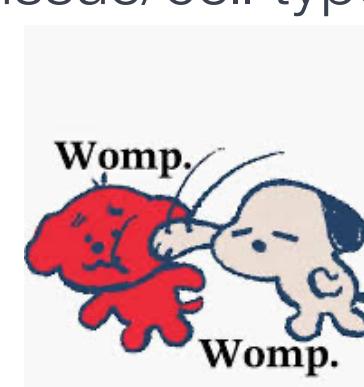
Enzymes (trypsin, collagenase, liberase, accutase)



Dissociator

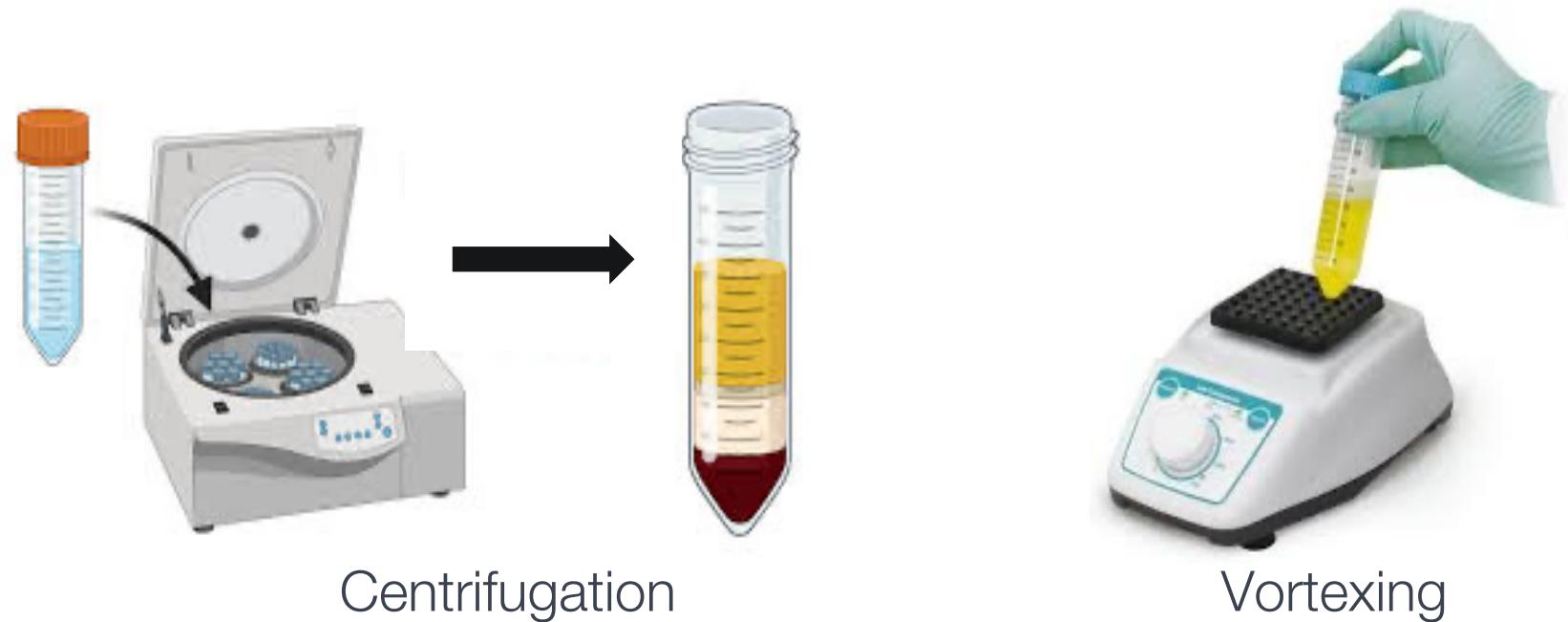
3. Combination

4. Automated



Dissociation: making a single cell suspension

1. Mechanical methods – cutting, shearing, laser dissections, FACS



2. Enzymatic dissociation

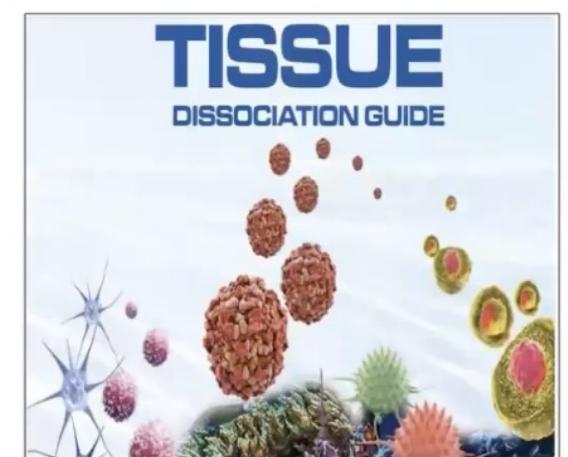
3. Combination

4. Automated



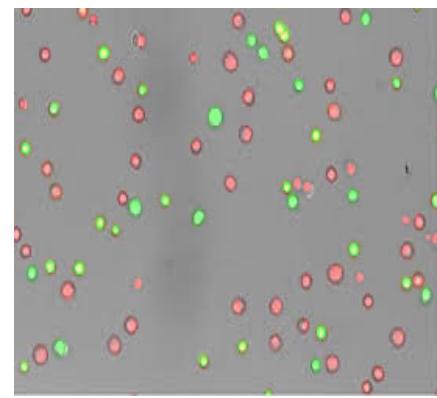
How to find a protocol-

- Publications/Literature
- Technology websites
- Customer support
- Online resources
- Talk to experts
- Use ready-to-use dissociators
- Trial-n-error



<https://www.technologynetworks.com/cell-science/how-to-guides/tissue-dissociation-guide-334918>

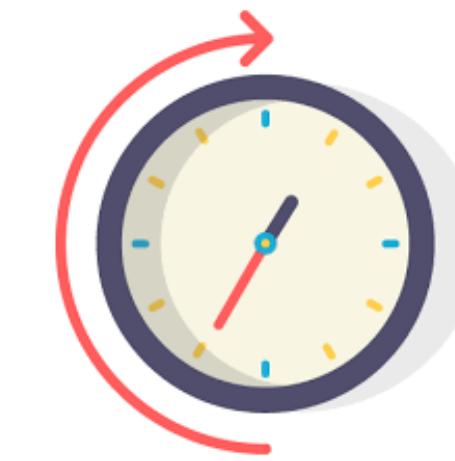
Factors affecting sample preparation (& data quality): top 5



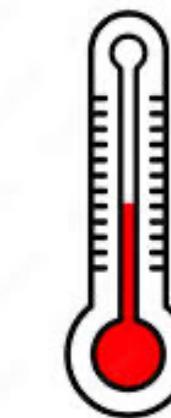
Cell Viability



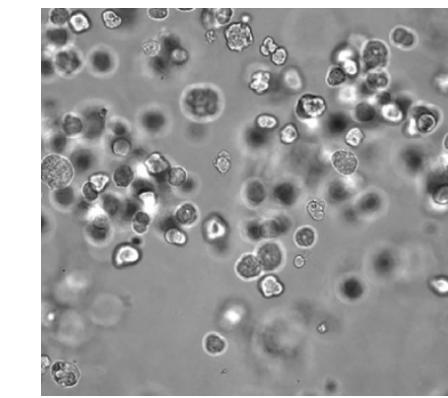
Cell Count



Time



Temp

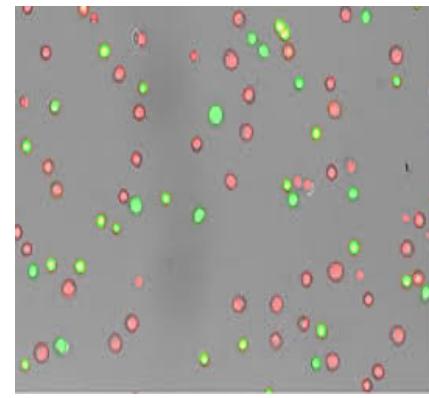


Quality

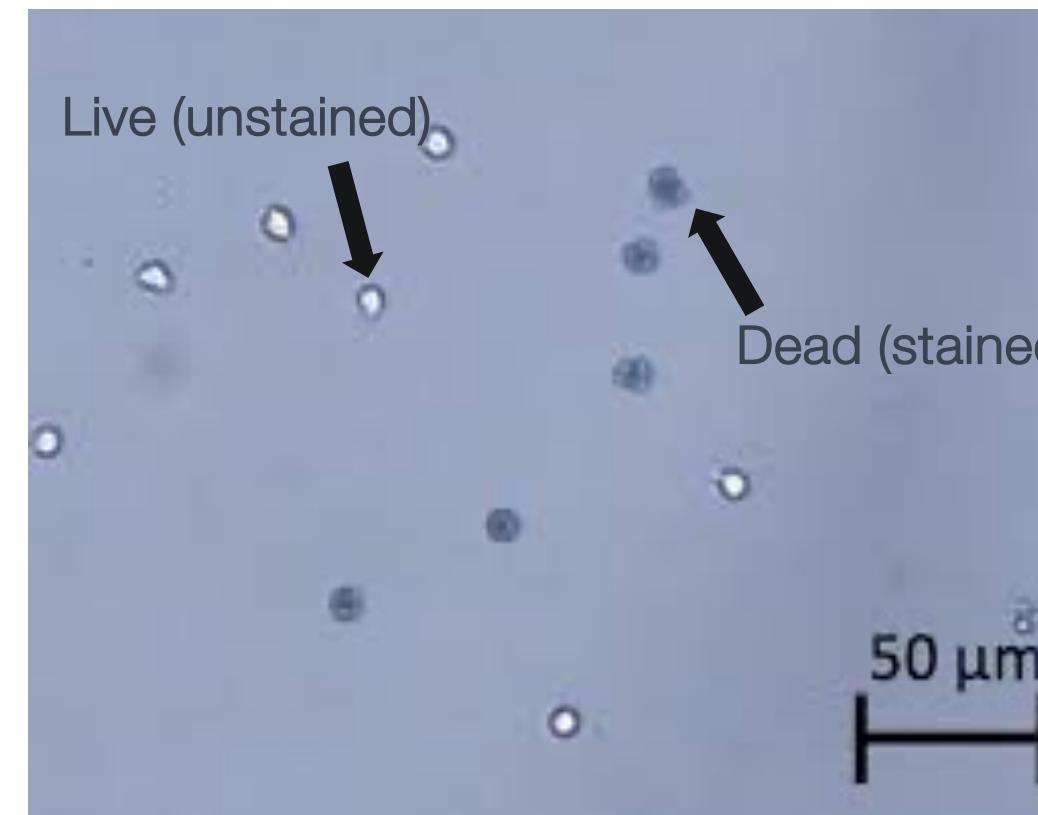
Factors affecting sample preparation: cell viability

1

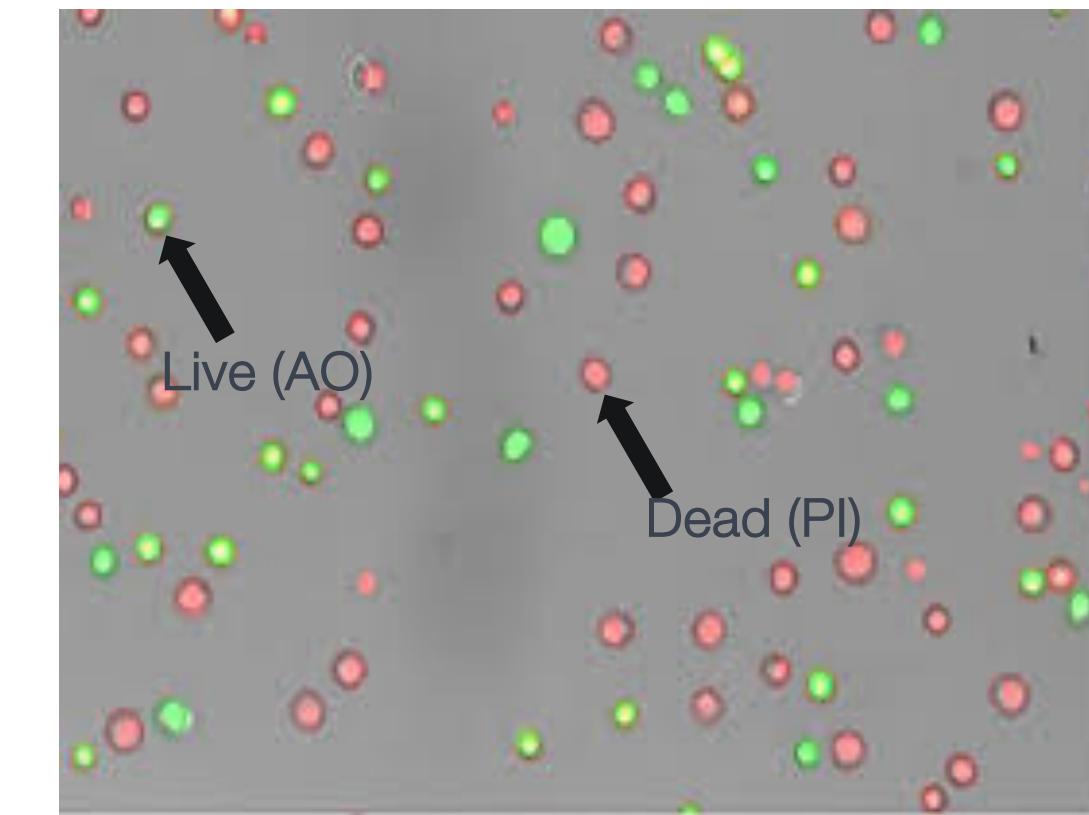
The higher the viability, the better (minimum 70-75%, ideally >90%)



Cell Viability
(high >70%)



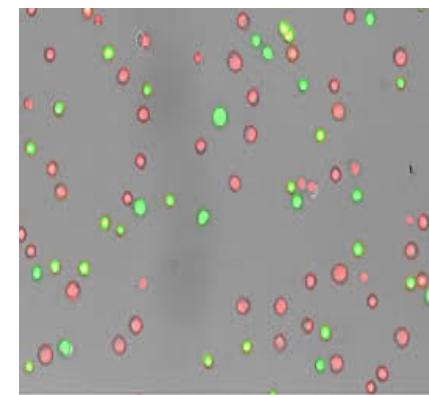
Trypan Blue (dead)



Acridine orange (live)/
Propidium iodide (dead)

Factors affecting sample preparation: cell viability

1

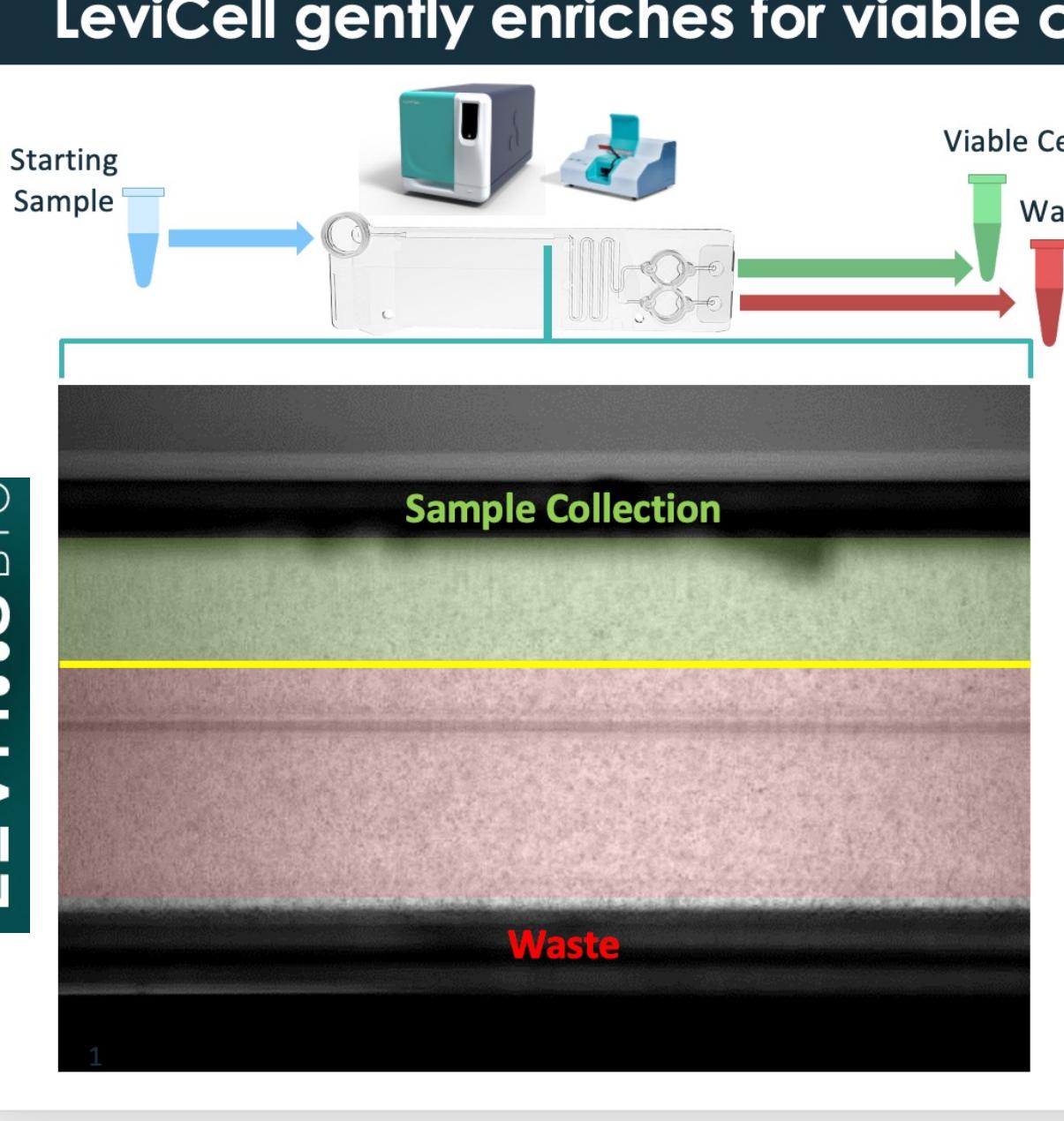


Cell Viability
(high >70%)



Dead cell removal,
Enrichment for live cells

LEVITAS BIO



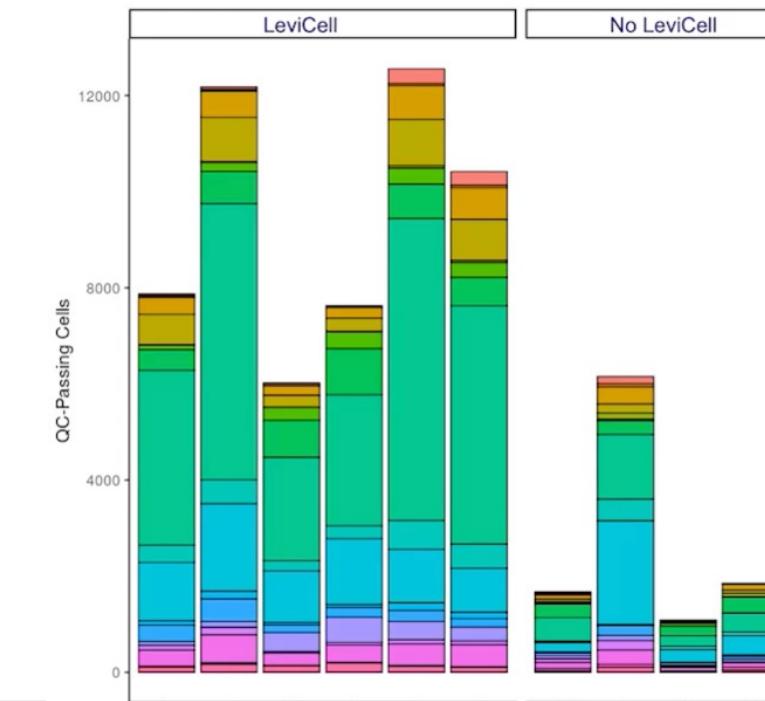
LeviCell gently enriches for viable cells improves SC data



Up to 5-fold increase in cell recoveries post-sequencing

Dr. Aane Antanaviciute

"With particularly bad-quality samples....the platform [LeviCell] becomes really applicable because it can rescue projects that would otherwise fail."

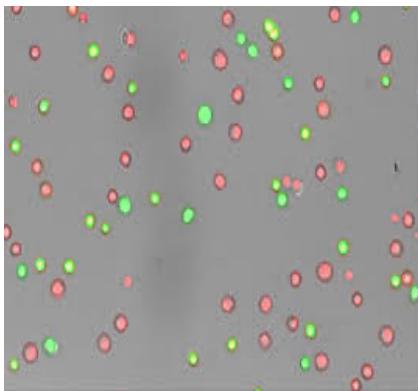


Factors affecting sample preparation: cell viability

1



Dead cell removal and any sort of enrichment comes at a cost!



1. How many dead cells are you removing?
2. What does this mean for the biology you are studying?
3. Are you recording this metadata?

Dead cells may increase the free-floating RNA in your prep
(check supernatant for RNA using e.g. Ribogreen Kit)

Cell Viability
(high >70%)



Dead cell removal,
Enrichment for live cells

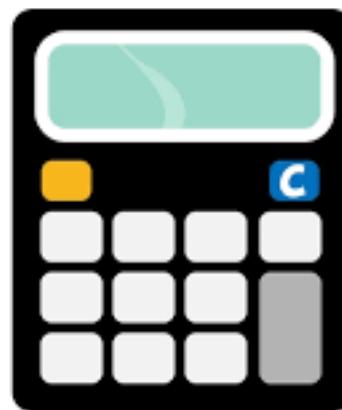
Low viability
Ambient RNA

Filter parameters:

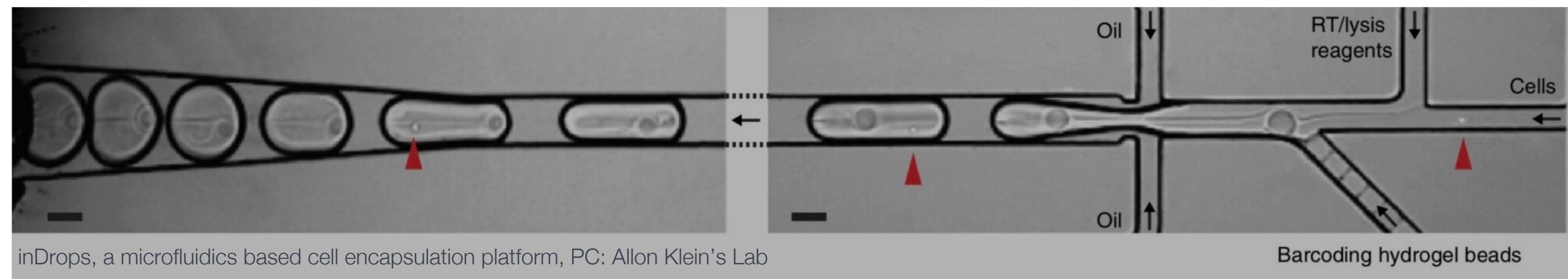
- nUMI > 500
- nGene > 250
- log10GenesPerUMI > 0.8
- mitoRatio < 0.2

Factors affecting sample preparation: cell count

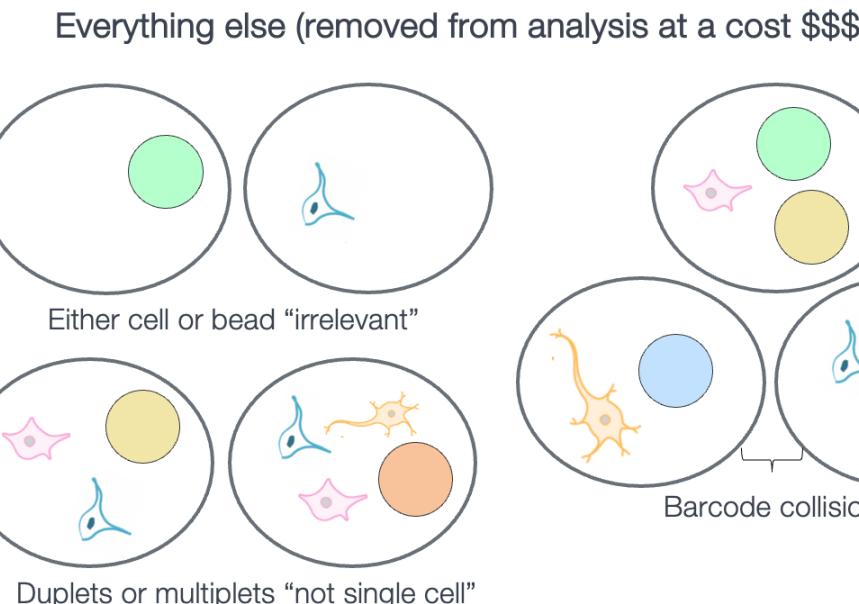
2



Single cell encapsulation follows Poisson's distribution (& tightly controlled math)



Cell Count
(accurate)



- nUMI > 500
- nGene > 250
- log₁₀GenesPerUMI > 0.8
- mitoRatio < 0.2

Factors affecting sample preparation: cell count

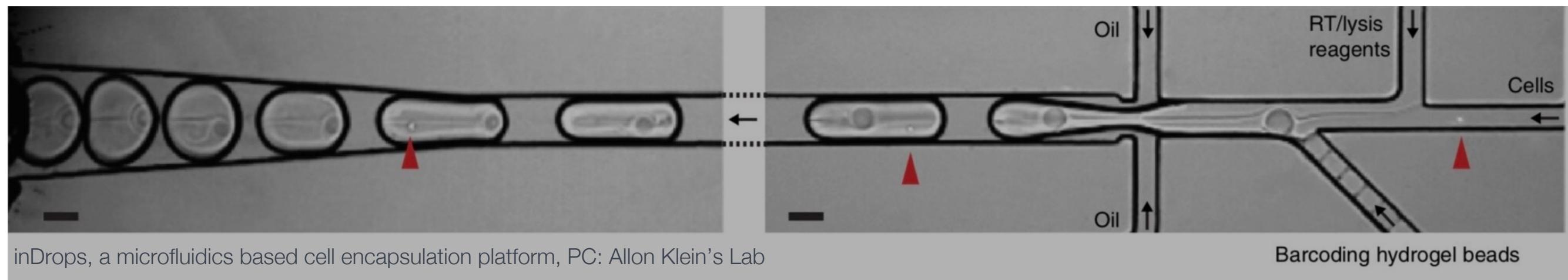
2



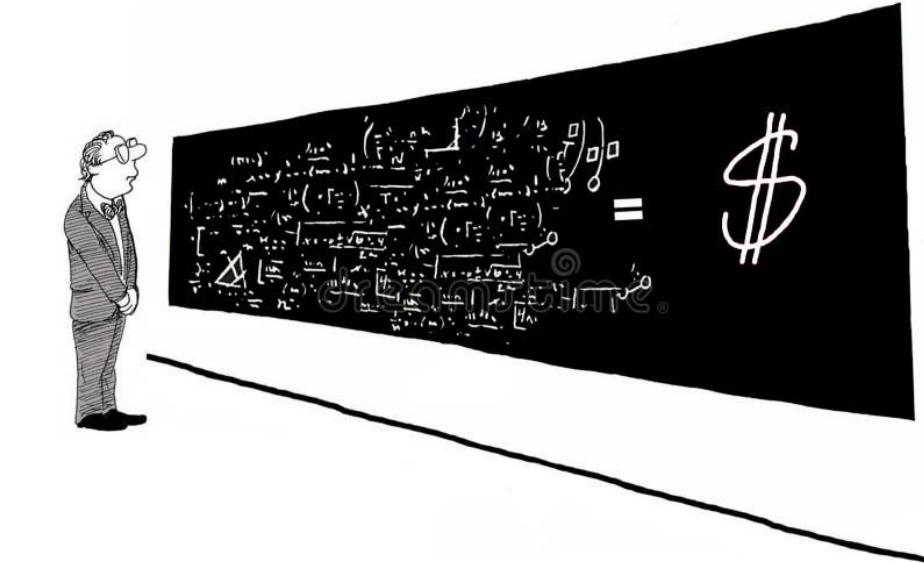
Cell Count
(accurate)

Manual count
at optimization

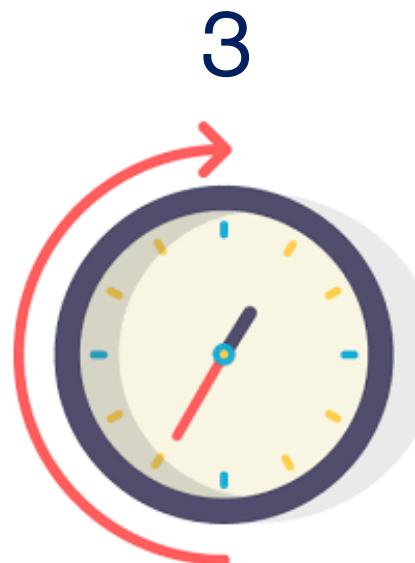
Single cell encapsulation follows Poisson's distribution



- Calibrate count by manual counting (hemocytometer)
- Cell sorters can overestimate by as much as 5-50%



Factors affecting sample preparation: time and temp



Time
(short)

Simple protocol,
minimal steps
1-3h

Less is more!

- Minimal handling
- Gentle protocol
- Reduce/arrest metabolic activity of cells
- Not induce extra stress response in cells (high mito)

- nUMI > 500
- nGene > 250
- log₁₀GenesPerUMI > 0.8
- mitoRatio < 0.2



Temp.
(cold, 4C)

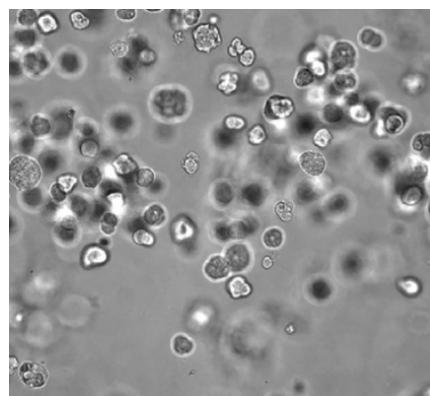
On ice,
RNase-free

- RT accelerates cell death, clumping, results in “ambient RNA” or degraded RNA
- Ambient RNA creates noisy, unusable data, at high costs, and wasted \$

- nUMI > 500
- nGene > 250
- log₁₀GenesPerUMI > 0.8
- mitoRatio < 0.2

Factors affecting sample preparation: quality

5



Quality
(no clumps/debris)

 Use micron-filters,
Gentle pipette-mixing
or centrifugation (<400-500g, 4C),
Use Dnase,
Be quick

Your expt is not single-cell if there are clumps! (removed at filtering)



Cell aggregates

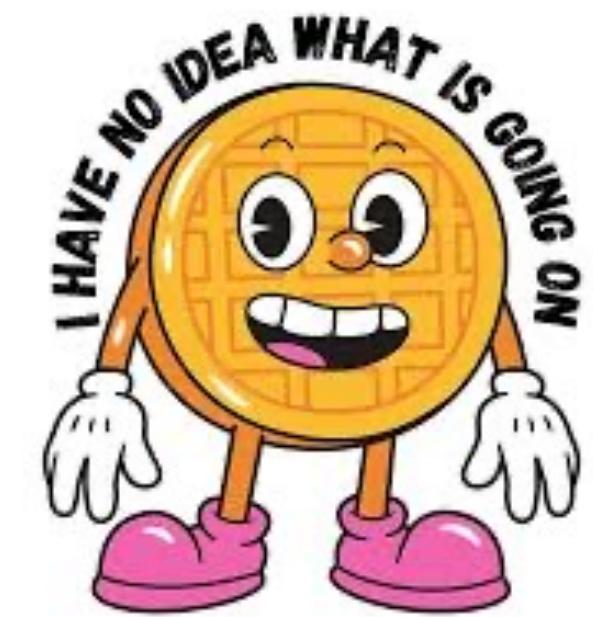
< 10% doublets

Single cell platforms do not distinguish between live or dead cells, debris or clumps and will encapsulate everything, making your data noisy/unusable

- nUMI > 500
- nGene > 250
- log10GenesPerUMI > 0.8
- mitoRatio < 0.2

Common causes of a bad prep: too many!

- Long prep times (>3-4h)
- Harsh dissociation conditions or harsh handling
- Too many dead cells
- Debris
- Using wrong buffer/media
- Cell/Nuclear membrane damage



PRACTICE PRACTICE PRACTICE!

This is why the actual “scRNA-seq run day” should not be the 1st time you attempt the protocol

Wide variety of input samples compatible for scRNAseq- decouple sample collection from sample processing



✓ Cryopreserved samples

- Success dependent on cell type
- Beware of data bias: some cells more prone to death upon thaw

DMSO cryopreservation is the method of choice to preserve cells for droplet-based single-cell RNA sequencing

Christian T. Wohnhäas, Germán G. Leparc, Francesc Fernandez-Albert, David Kind, Florian Gantner, Coralie Violet, Tobias Hildebrandt & Patrick Baum

Cryopreservation with BAMBANKER™



✓ Snap-Frozen samples

- Segue into next slide
- Offers greater flexibility



'Frankenstein' protocol for nuclei isolation from fresh and frozen tissue for snRNAseq ↗ ▾

Luciano Martelotto¹



✓ Fixed samples

- Opens up a treasure trove!
- Offers greater flexibility



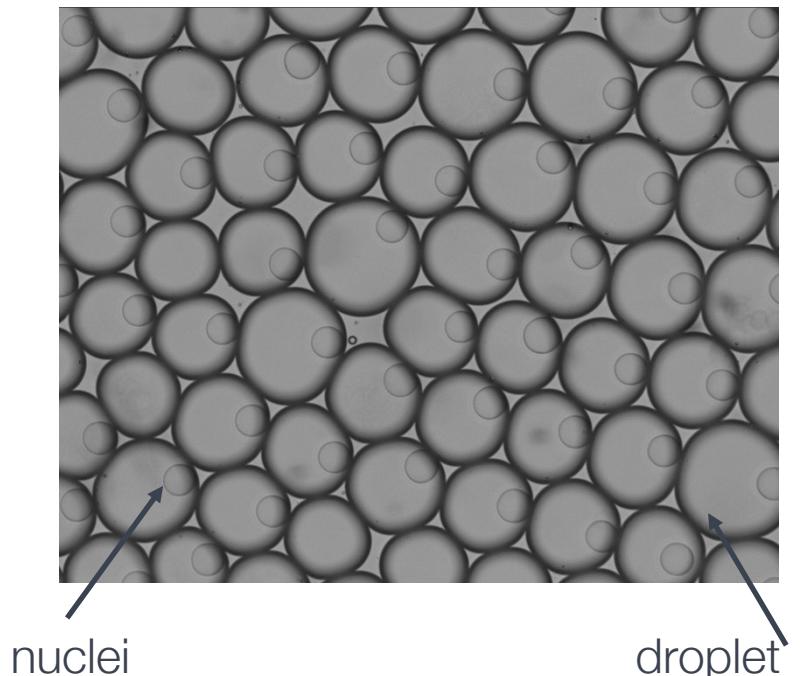
FixNCut: A Practical Guide to Sample Preservation by Reversible Fixation for Single Cell Assays

Method | Open access | Published: 08 April 2021

ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics

single nuclei RNA-seq (snRNAseq): data comparable to scRNAseq

- Removes transcriptional noise from dead/dying cells
- snRNAseq most often used for
 - ✓ difficult to isolate/dissociate samples e.g. neuronal samples
 - ✓ low viability samples e.g. good for flash frozen clinical samples
 - ✓ tissues problematic for sc-processing e.g. adipose tissue, where fat inhibits RT enz. or pancreatic tissue (high in RNases)



- ✓ Cell types hard to get from single cell preparations or cells too big to encapsulate on microfluidic platforms
- ✓ ATAC (to study the epigenome)/Multiome studies (to study the epigenome along w/ transcriptome)

Same QC metrics to evaluate sample prep

Garbage in, Garbage out, every single time



Poor quality input (cells) contributes to poor quality output (data) in scRNAseq!



=



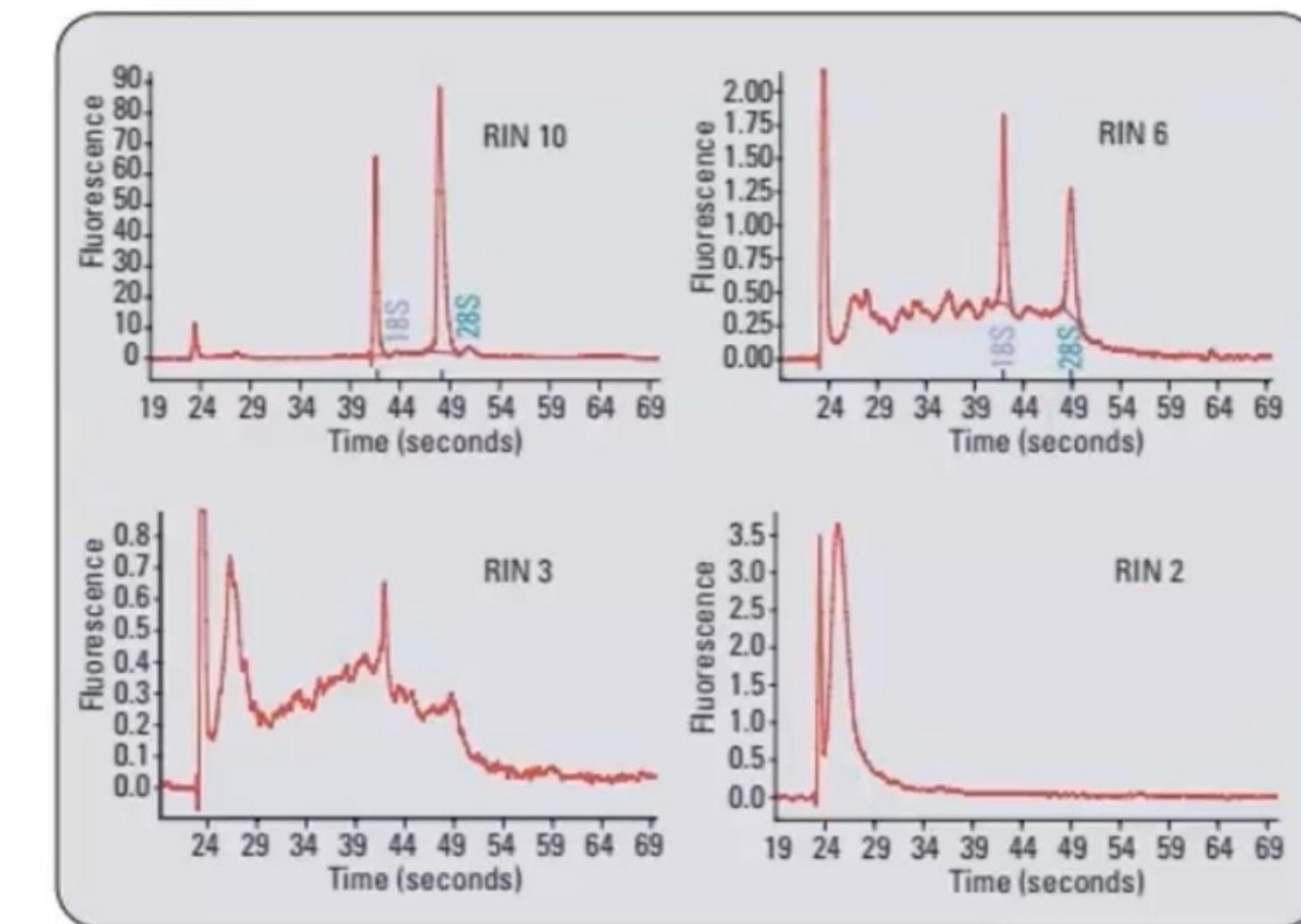
The RIN score: RNA integrity as final check at optimization

RIN stands for “RNA Integrity Number”,
i.e. how degraded is the RNA in your sample(s)



RIN 7-10 (Proceed), RIN < 3 (no go)

Informative for:
Assessing prep quality
Predicting data quality



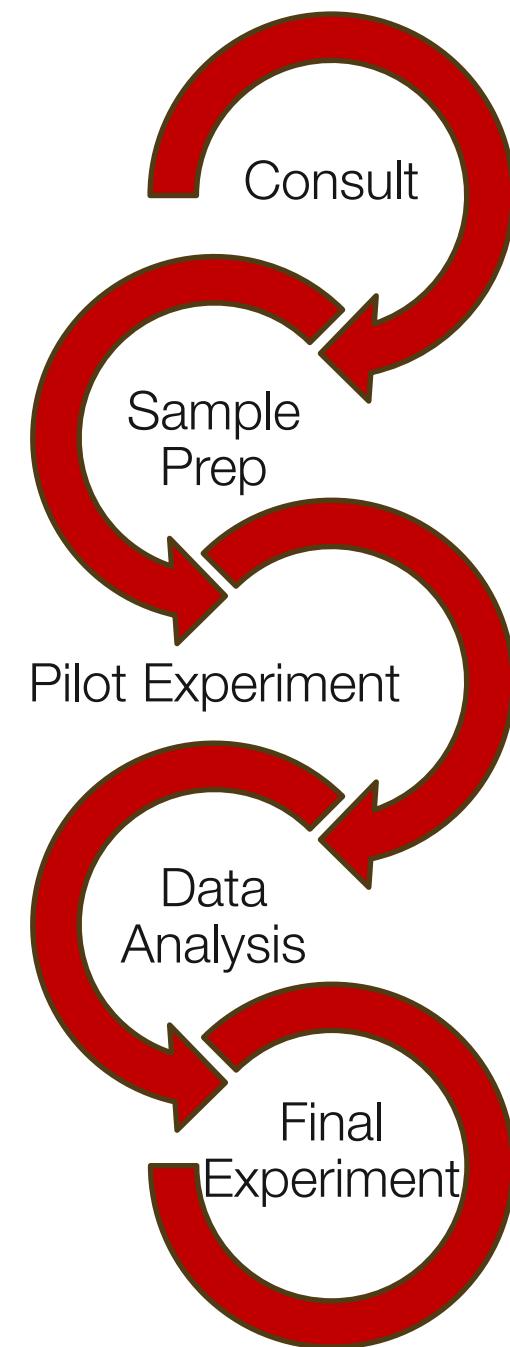
PC: blog.genohub.com

Do a (small scale) pilot experiment



pilot is key to success (closes gaps between theory & practice)

- Do not rush to the final experiment
- A well-planned pilot experiment is essential for
 - ✓ tweaking or redefining bio. objectives
 - ✓ providing rational expt design/optimal approach for research Q
 - ✓ evaluating sample preparation
 - ✓ figuring out the required number of cells needed statistically to answer your biological question



A pilot will warn you about the sources of technical noise (before you've spent big money)

“Technical Noise”: When non-biological, technical factors cause changes in the data produced by the expt. leading to wrong conclusions

 Single cell expts prone to technical noise

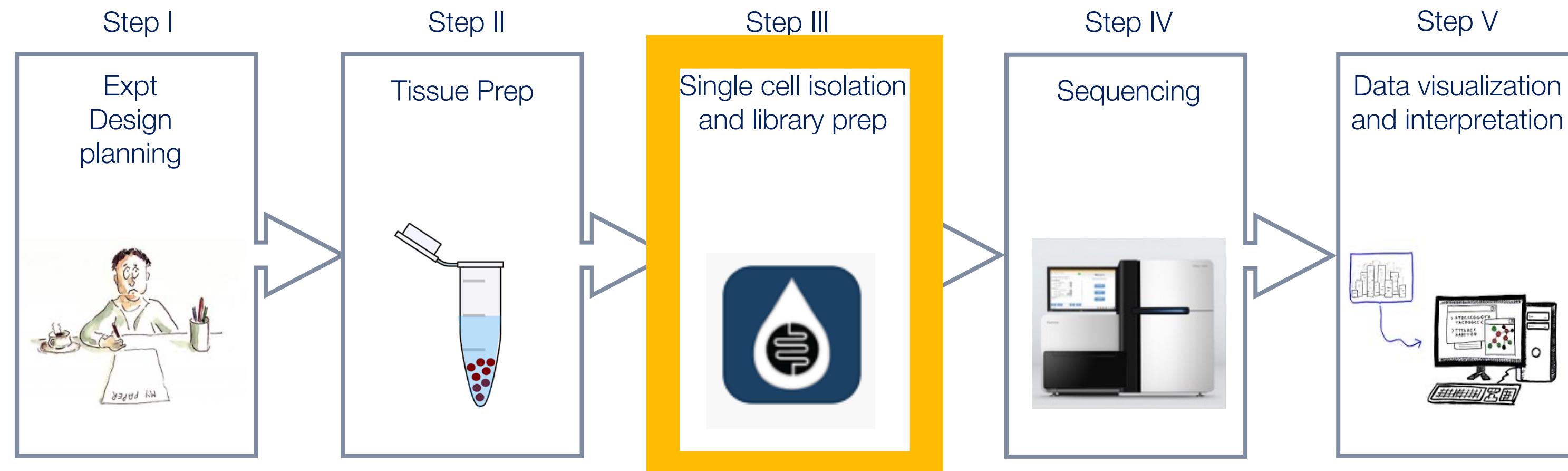
2 kinds of technical noise -

- from experimental designs and handling (e.g. different personnel, reagent lots, PCR amp cycles, equipment, protocols etc) -> **“Batch effect correction”**
- from sequencing (e.g. library prep, GC content, amp bias etc) -> **“Normalization”**



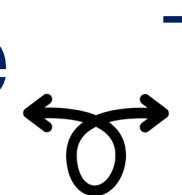
- nUMI > 500
- nGene > 250
- log10GenesPerUMI > 0.8
- mitoRatio < 0.2

Steps in a scRNAseq work flow



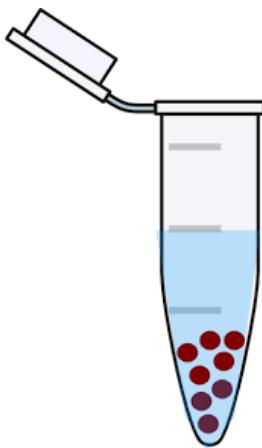
Goal: Capture and isolate single cells on a suitable platform, & prep libraries

Scale impacts technology choice



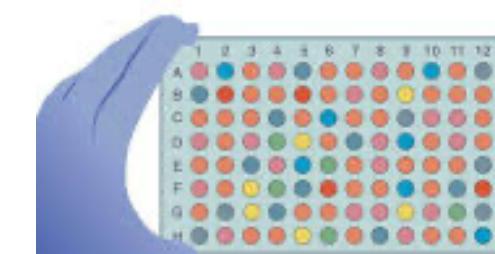
Technology choice impacts scale

Low Through-put approaches



Manual

1st Gen



Plate/Wells

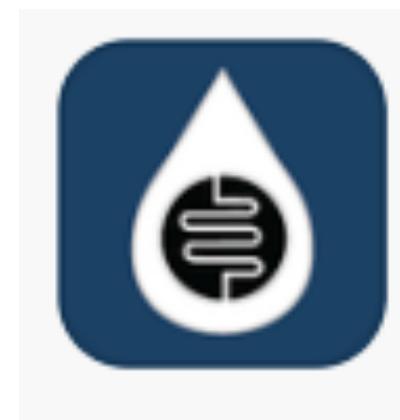


Liquid handlers

1-few cells

Minimal equipment
Cost effective
Scale prohibitive

High Through-put approaches



Microfluidics/Droplet

2nd Gen

3rd Gen

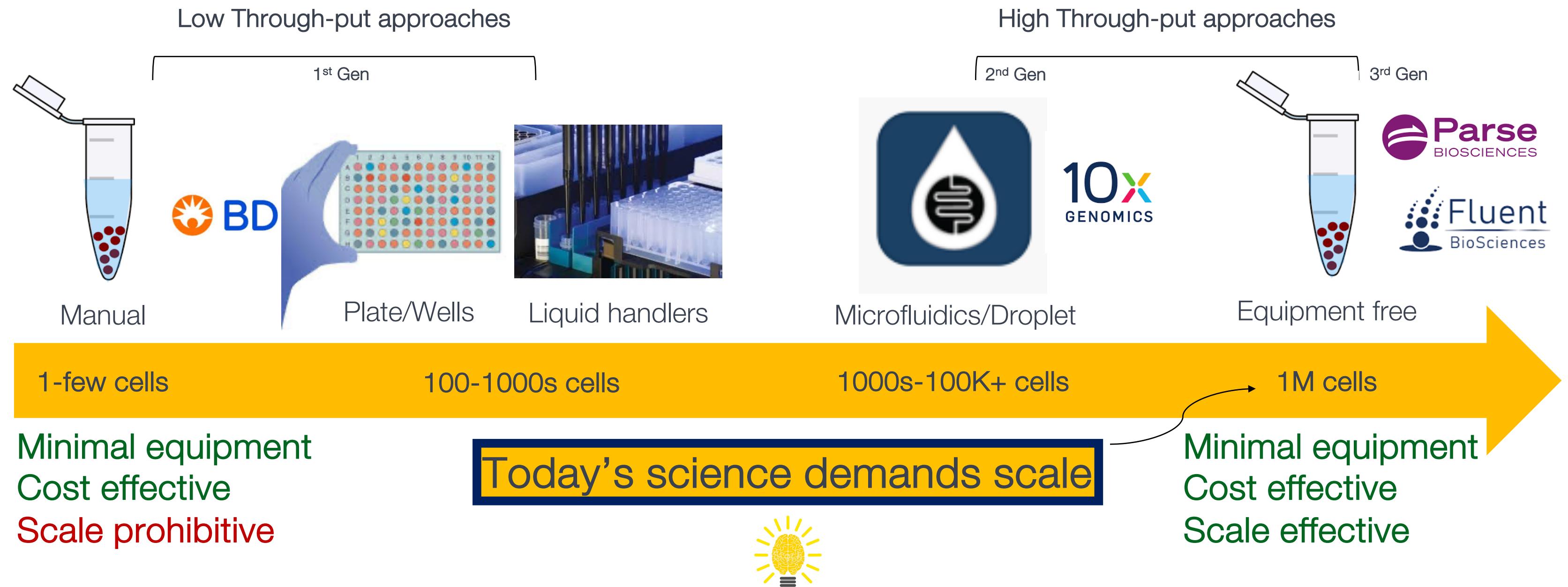
100-1000s cells

1000s-100K+ cells

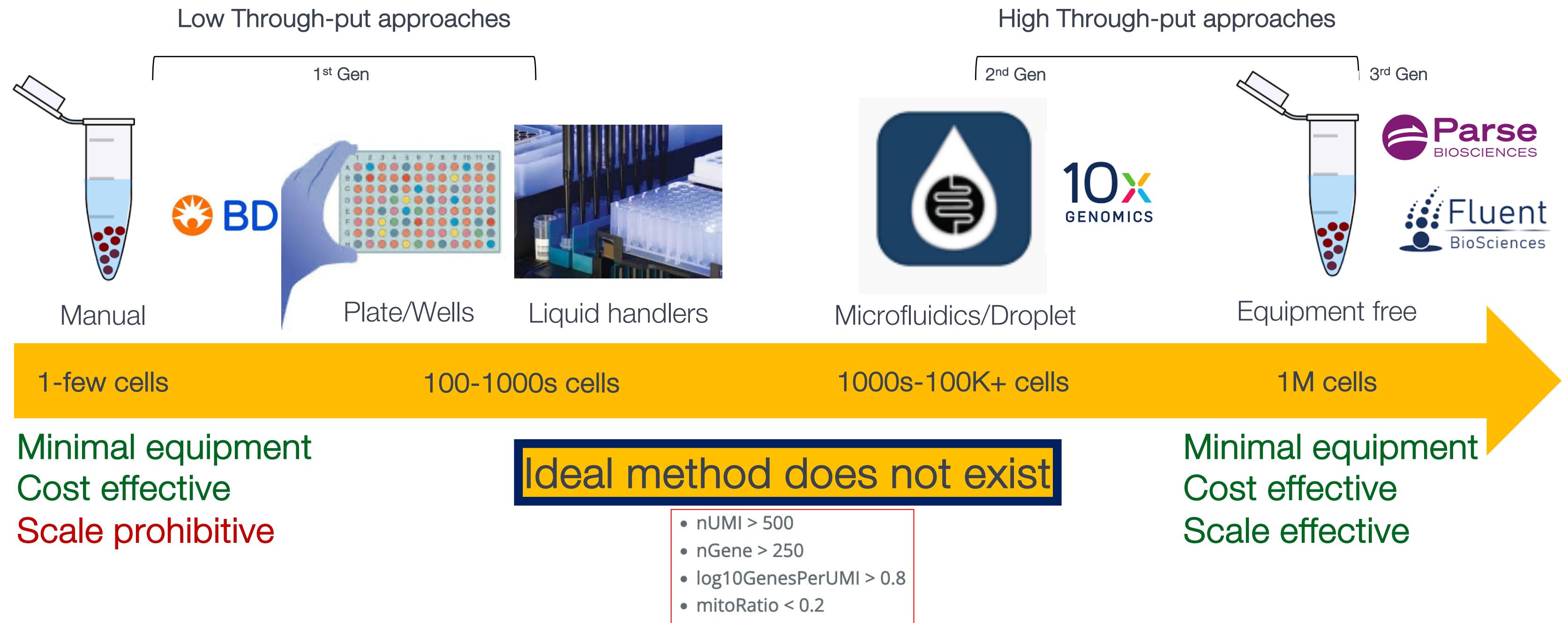
1M cells

Sophisticated equipment
Cost prohibitive
Scale effective

Scale impacts technology choice \leftrightarrow Technology choice impacts scale

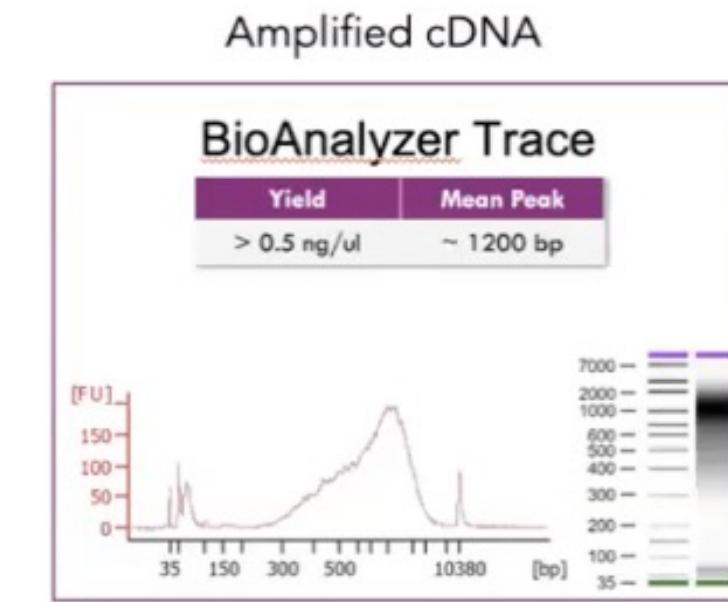
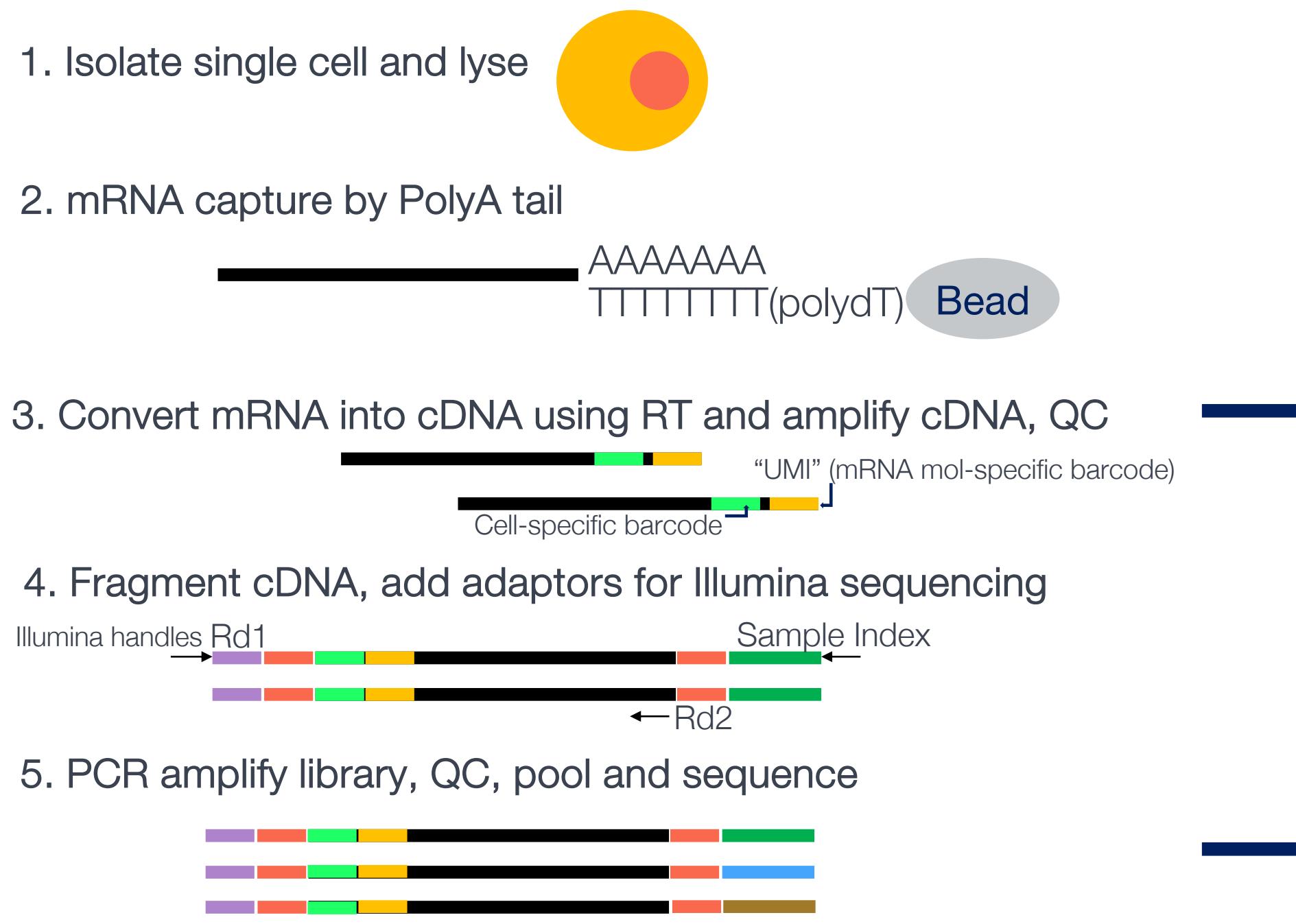


Scale impacts technology choice ↪ Technology choice impacts scale

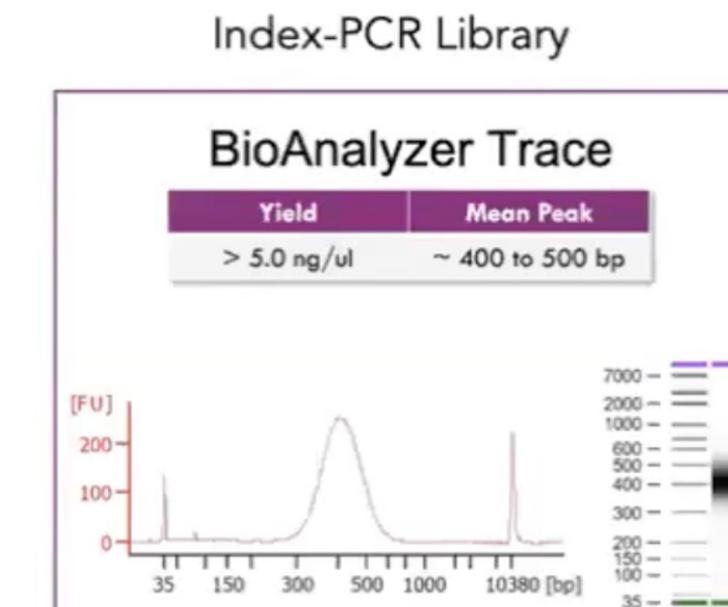


Structure and scRNA-seq library preparation

(just an illustrative example)



- Quantify
- Size
- contam

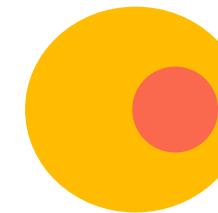


- Quantify
- Size
- contam

Structure and scRNA-seq library preparation

(just an illustrative example)

1. Isolate single cell and lyse



2. mRNA capture by PolyA tail



3. Convert mRNA into cDNA using RT and amplify cDNA, QC

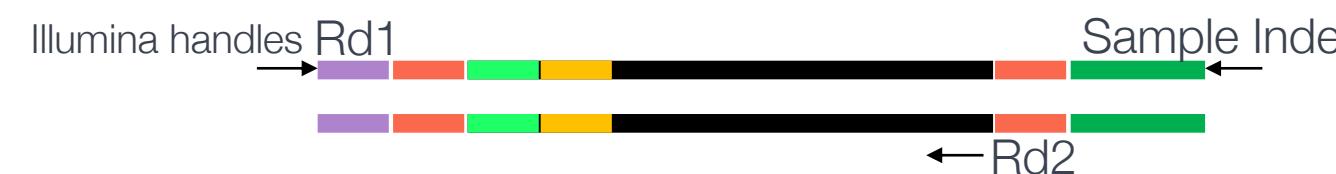


Library structure is sc-platform, kit & sequencing platform dependent

Lib prep is the biggest source of batch effect

- nUMI > 500
- nGene > 250
- log₁₀GenesPerUMI > 0.8
- mitoRatio < 0.2

4. Fragment cDNA, add adaptors for Illumina sequencing



- When possible, prep all libs together
- Prep control and treatment together, per run, so you can normalize

5. PCR amplify library, QC, pool and sequence



Parallel, multimodal assays to add layered info to scRNAseq data

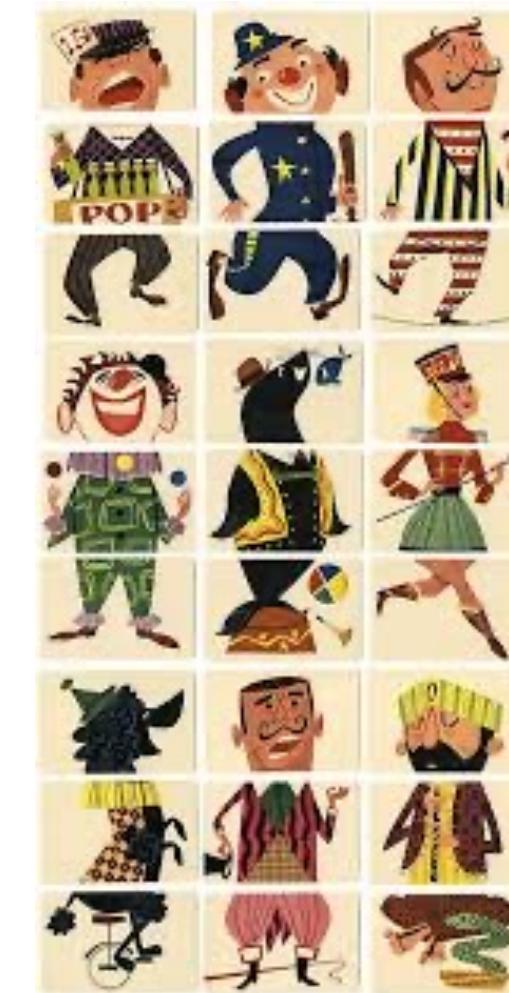
Multiple libraries from same sample for multimodal sc-analysis:

- scRNAseq ([transcriptome](#)) + scATACseq ([epigenome](#))
- scRNAseq ([transcriptome](#)) + CITEseq/Hashing/VDJ ([surface proteins](#))
- scRNAseq ([transcriptome](#)) + DNA ([genome](#))



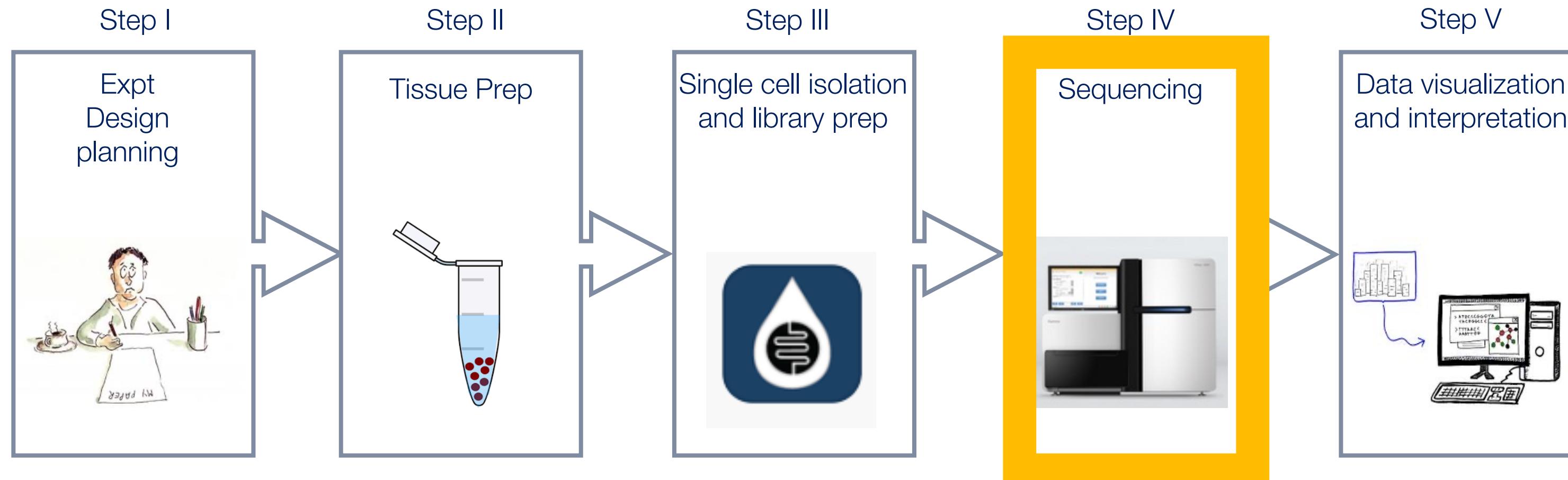
More informative data at same or lower cost!

But expt has to be designed as such at the beginning



Mix-n-match

Steps in a scRNAseq work flow



Goal: sequence your libraries on the appropriate platform

Step IV: Sequencing platforms for scRNAseq

(current) common compatible single cell sequencing platforms – NextSeq and NovaSeq



More output

Simple benchtop

Affordable & low cost

Fast data turnaround



Advantages	Power of high-throughput sequencing with the simplicity and affordability of a benchtop system	Unprecedented output and throughput
Ideal for	Mid- to high-throughput sequencing applications and average scale single-cell sequencing studies, such as studies to profile cell function in both development and disease.	Extensive screening studies, such as pharmaceutical screens and cell atlas studies.

Step IV: how much should one sequence?



Sequencing depth dependent on sample type and experimental objective

Table 8: Recommended reads for different single-cell sequencing applications

Method	Recommended no. of reads ^a
3' gene expression	15K–50K reads per cell
5' gene expression	50K reads per cell
Antibody sequencing	100 reads per antibody/cell
scATAC-Seq	50K reads per nuclei
5' TCR/BCR	5K reads per cell
Takara SMARTer	1M–2M reads per cell (> 300,000 reads per cell)

The recommended number of reads is based upon manufacturer recommendations

Step IV: choosing a platform

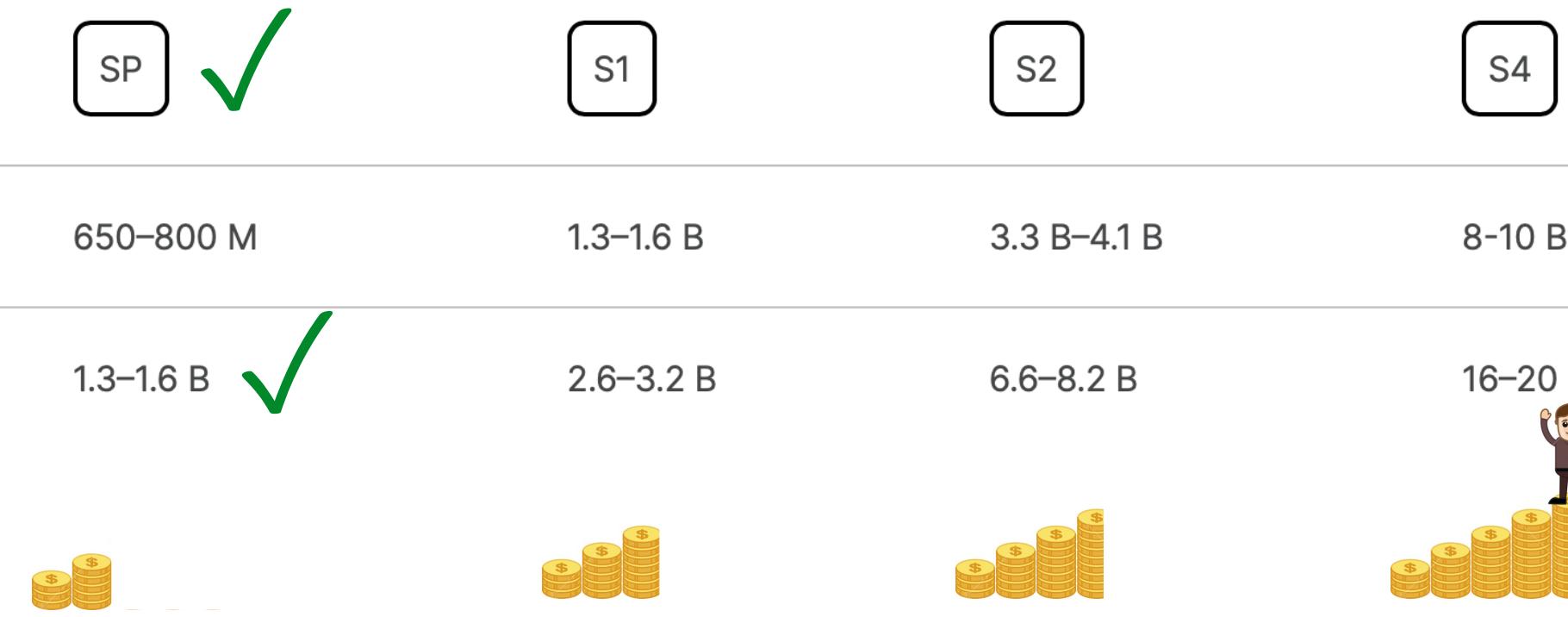
Example: You have 4 samples & you have barcoded 10K cells each for 5'GEX = 40K barcoded cells
40K x 50,000 reads/cell = 1 Billion total reads needed

NovaSeq 6000 System

Flow Cell Type	SP	✓	S1	S2	S4
Single-end Reads		650–800 M	1.3–1.6 B	3.3 B–4.1 B	8–10 B
Paired-end Reads		1.3–1.6 B	✓	2.6–3.2 B	6.6–8.2 B

Shallow seq {

- nUMI > 500
- nGene > 250
- log10GenesPerUMI > 0.8
- mitoRatio < 0.2



Step IV: choosing a platform

Example: You have 4 samples & you have barcoded 10K cells each for 5'GEX = 40K barcoded cells
 $40K \times 50,000 \text{ reads/cell} = 1 \text{ Billion total reads needed}$



1 slice of bread does not need an entire jar of peanut butter!
Similarly, you don't need sequencing-overkill on your sample

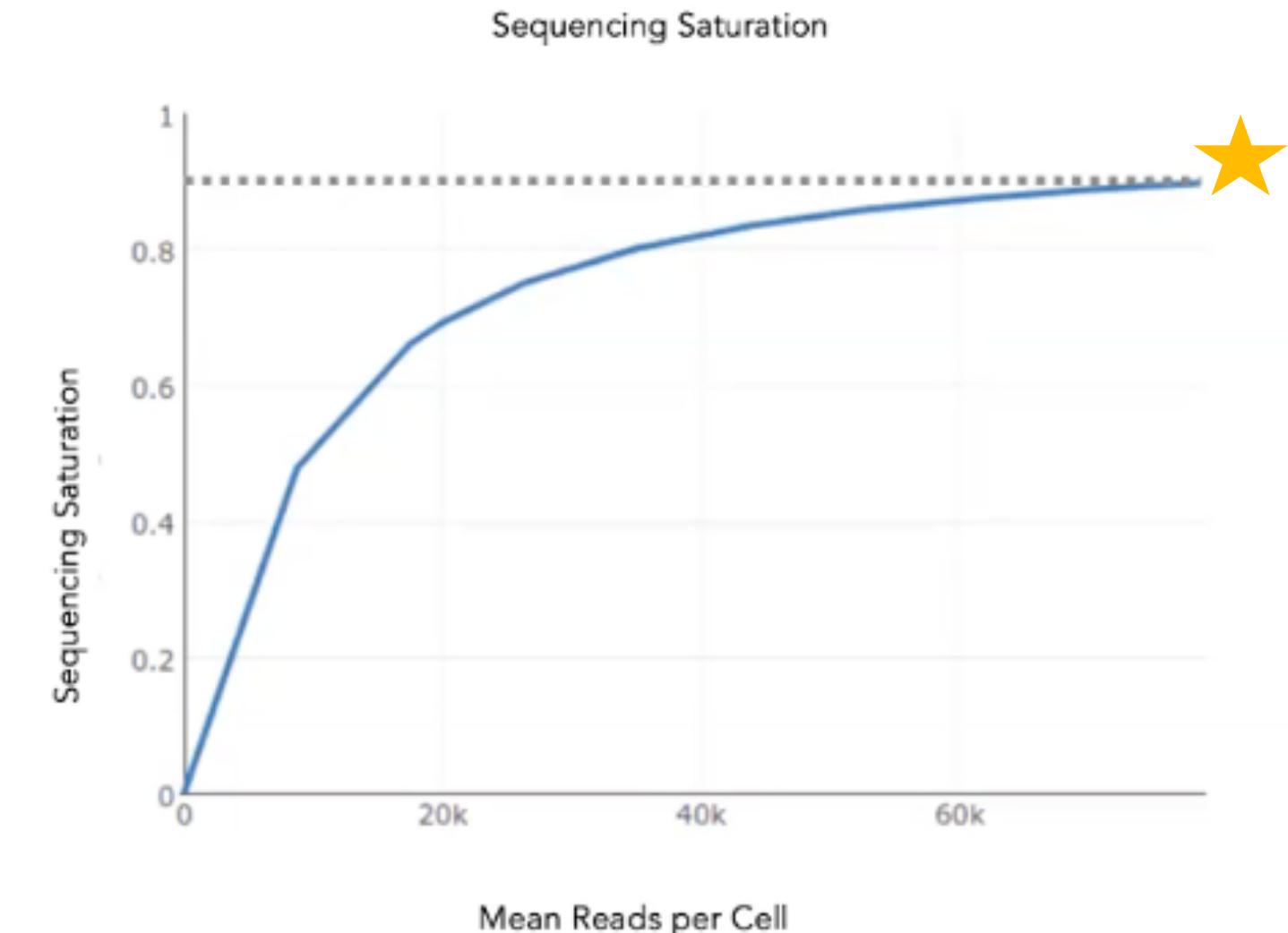


Step IV: dialing in on sequencing saturation

How to know what's sequencing over-kill?

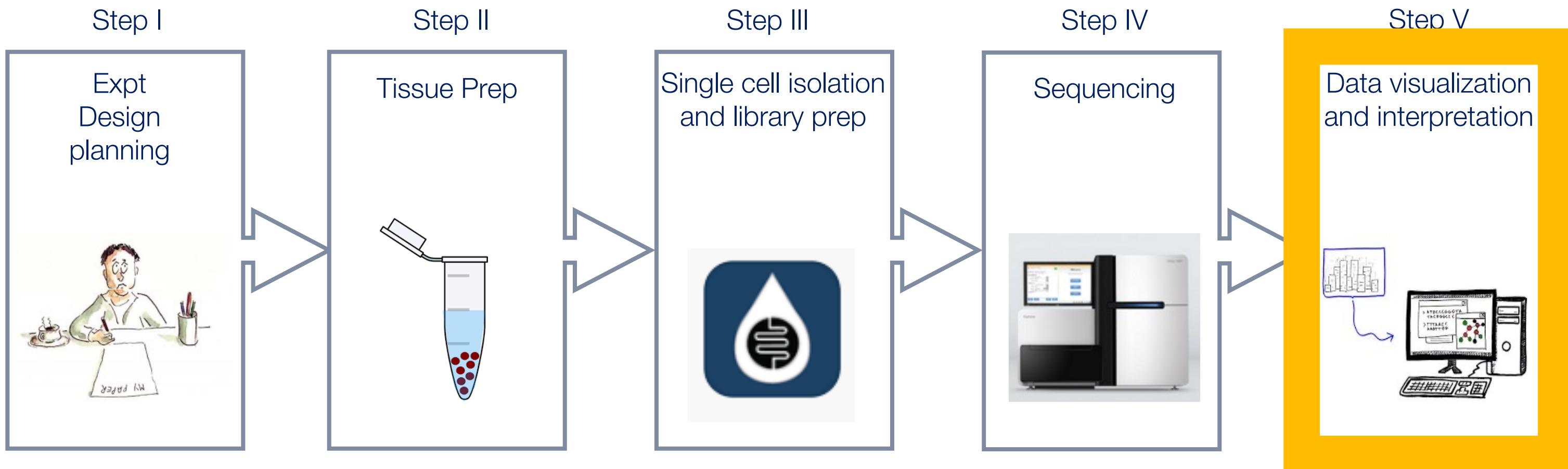
★ Seq saturation = $\frac{\text{# of unique mRNA detected}}{\text{# of total reads}}$

- Differs by RNA amount per cell type (cell type dependent)
- Depends on sample metrics – how many cells barcoded, what is rarest cell population of interest?
- Rarer the cell type (or transcript), more sequencing needed



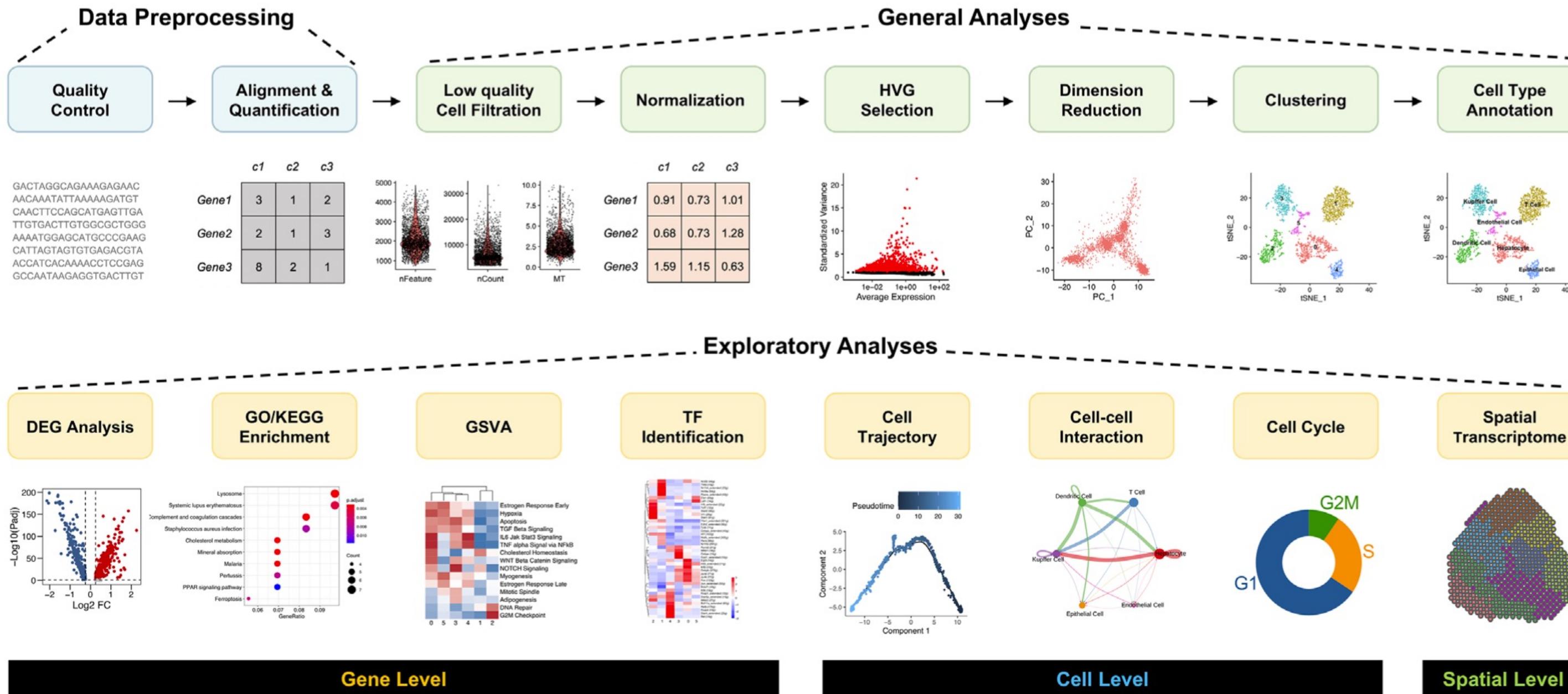
You can always re-seq your libraries!

Steps in a scRNAseq work flow



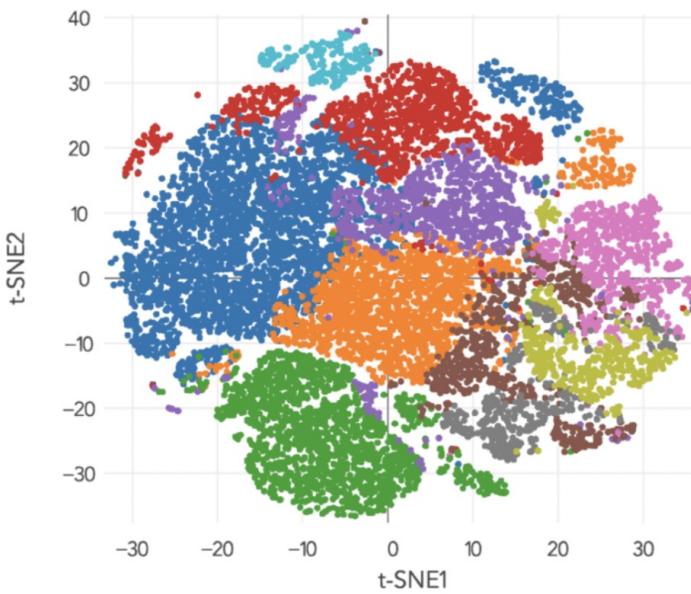
Goal: analysis, interpretation and visualization of data!

Key analysis steps in scRNAseq analysis



Pipelines for data visualization and interpretation

A wealth of bioinformatic tools are available for scRNA-seq analysis:



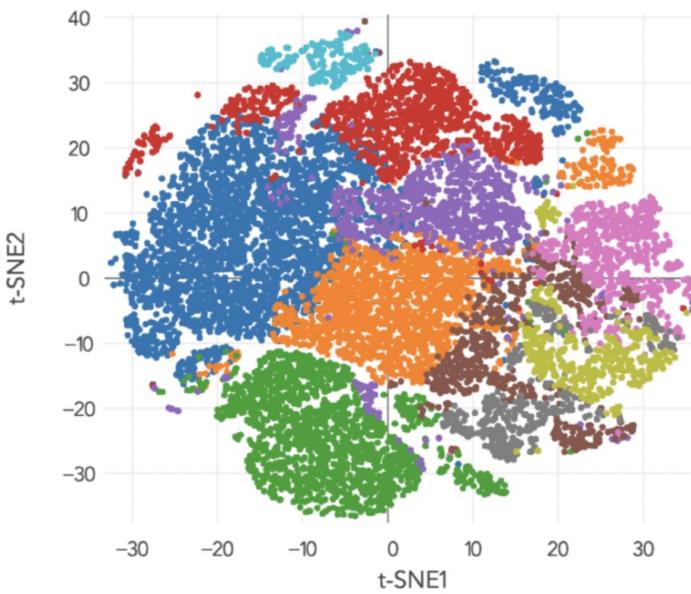
Publication-worthy figs



Interpretation

Pipelines for data visualization and interpretation

A wealth of bioinformatic tools are available for scRNA-seq analysis:



Publication-worthy figs

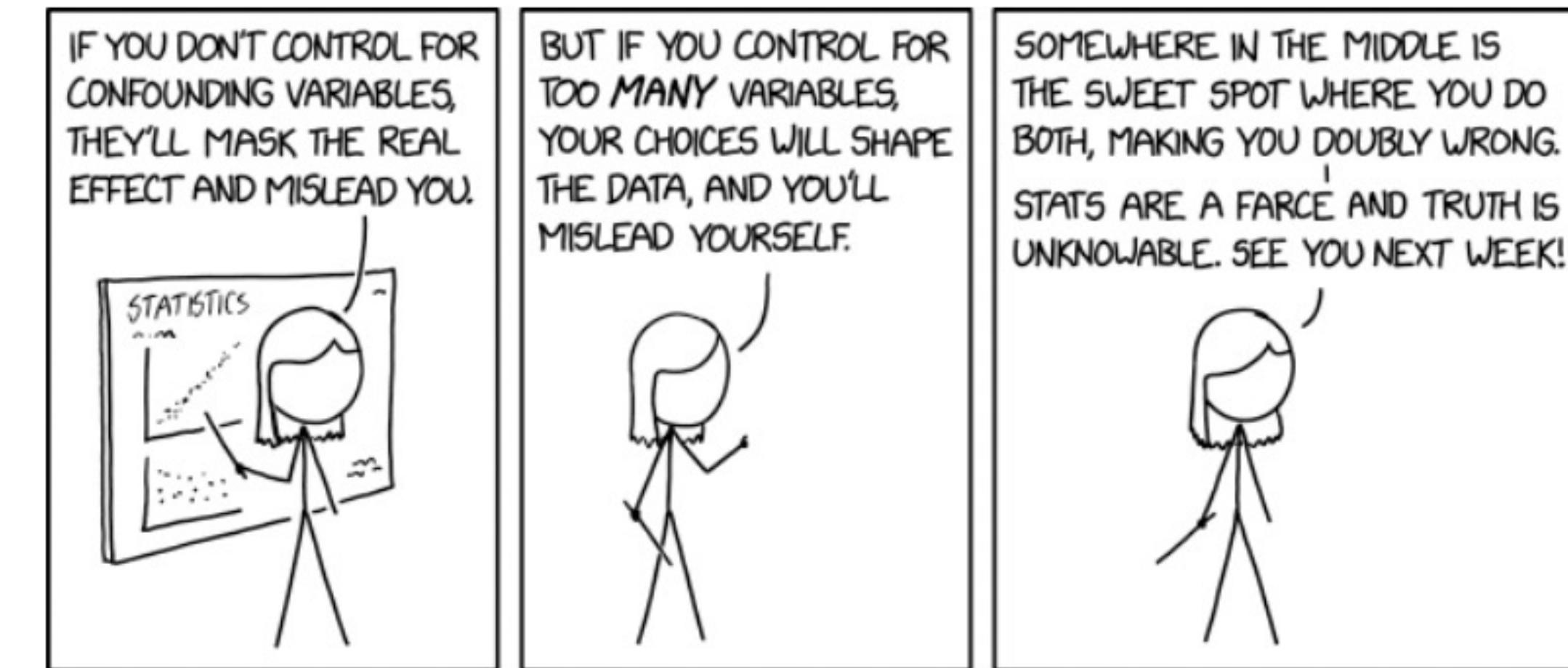
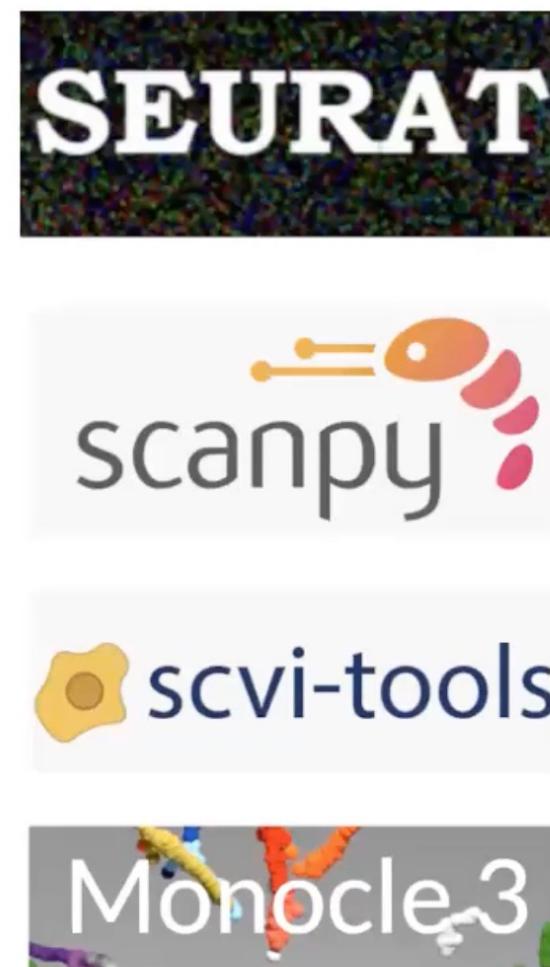


Interpretation

Leading pipelines for scRNAseq analysis-

- Seurat (Satija Lab, 2015): written in the programming language R
- Scanpy (Theis Lab, 2017): written in the programming language Python

More tools do not translate to more understanding of biology



Choice boils down to the user's programming preference

Implicit assumption: choice among packages and versions has little to no impact on the interpretation of results (**proven wrong**)



Key takeaways!

Summary –

- (I know it seems daunting) You can do this!
- Plan your experiments, put in effort into your sample prep. It pays off at data analysis

Just remember: **Garbage in, Garbage out**

- Pay attention to doing informed, responsible, transparent and reproducible analyses
- Talk to experts – this is a fast evolving field (technology, methodology and computationally)



Thank you! Got Questions?

Get (stay) in touch!

For consultations, trainings and questions:

arpita_kulkarni@hms.harvard.edu,
singlecell@hms.harvard.edu



@HMS_SCC; @ArpitaBKulkarni



Citation and use:

hbctraining/scRNA-seq_online: scRNA-seq Lessons from HCBC (first release). Zenodo. <https://doi.org/10.5281/zenodo.5826256>

These materials have been developed as part of a teaching course at the Harvard Chan Bioinfo Core. They are open access materials distributed under the terms of the *Creative Commons Attribution license (CC BY 4.0)*, which permits unrestricted use, distribution and reproduction in any medium, provided the original author & source are credited



