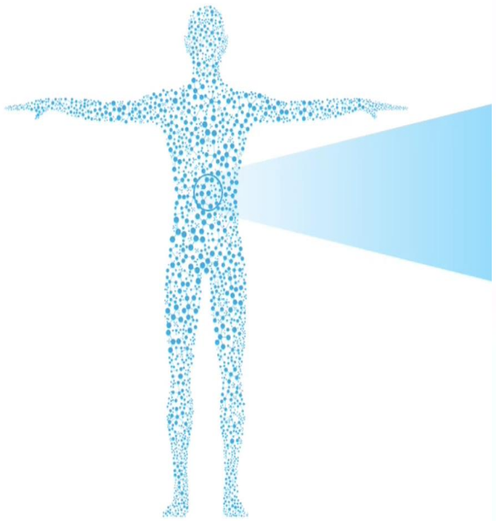
HARVARD MEDICAL SCHOOL | BLAVATNIK INSTITUTE SYSTEMS BIOLOGY | Beth Israel Lahey Health Beth Israel Deaconess Medical Center

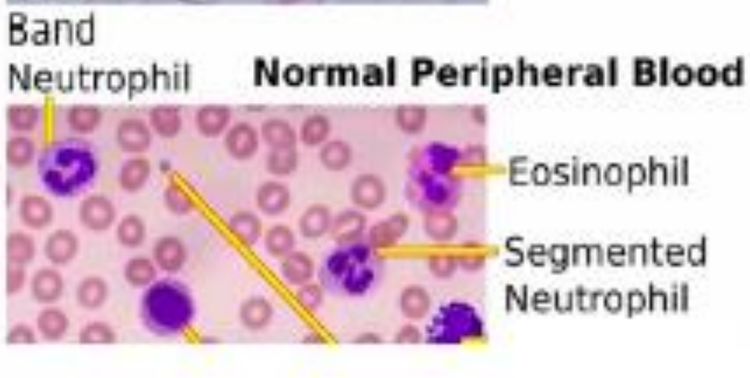
Bioinformatics begins at the bench:

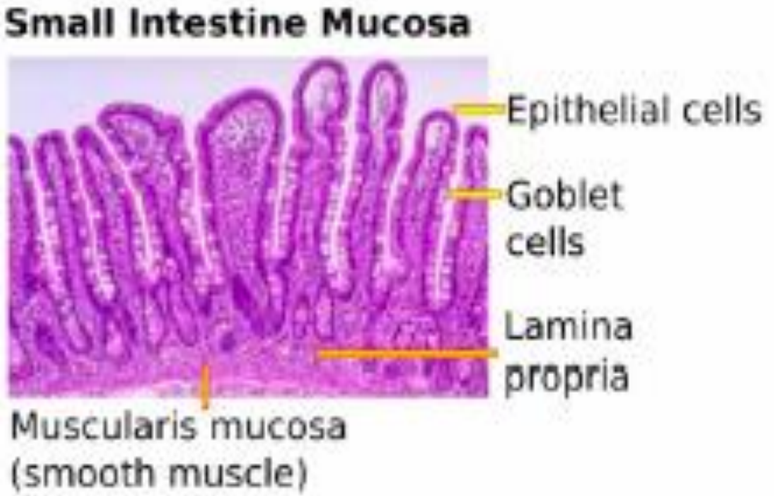
How experimental decisions shape single-cell RNA-seq analyses

Arpita Kulkarni, Ph.D.
Director, Functional Genomics & Bioinformatics Core, BIDMC/HMS
Former Associate Director, Single Cell Core, HMS

We know tissues are heterogeneous







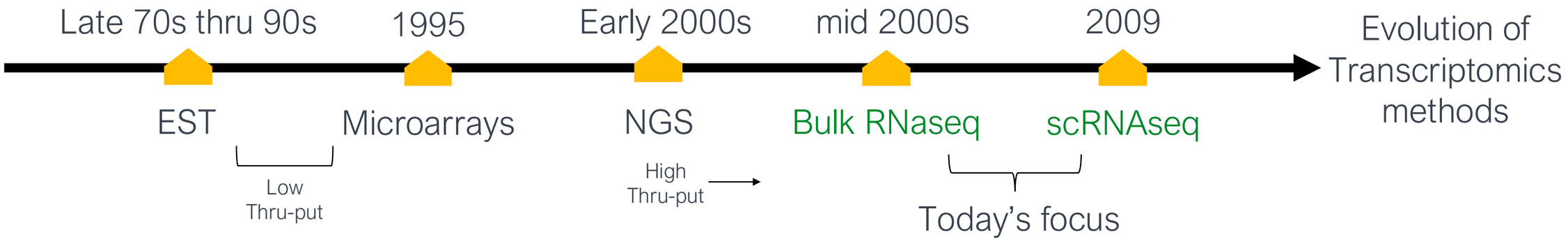
Q. What makes cells (& tissues) different from one another? How do we measure these differences?

PC: 10x Genomics

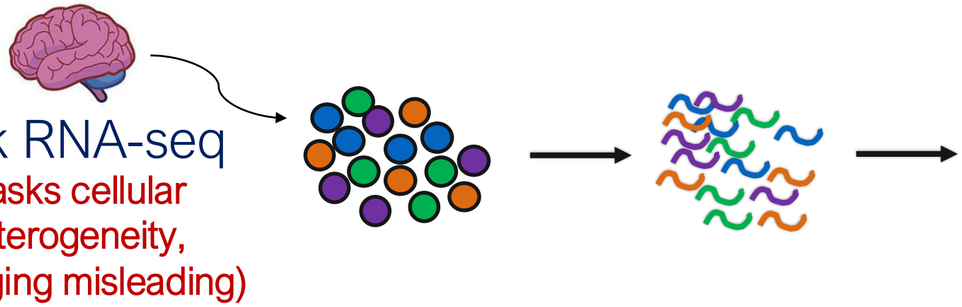
Transcriptomics: Unraveling the nature of cellular identity

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

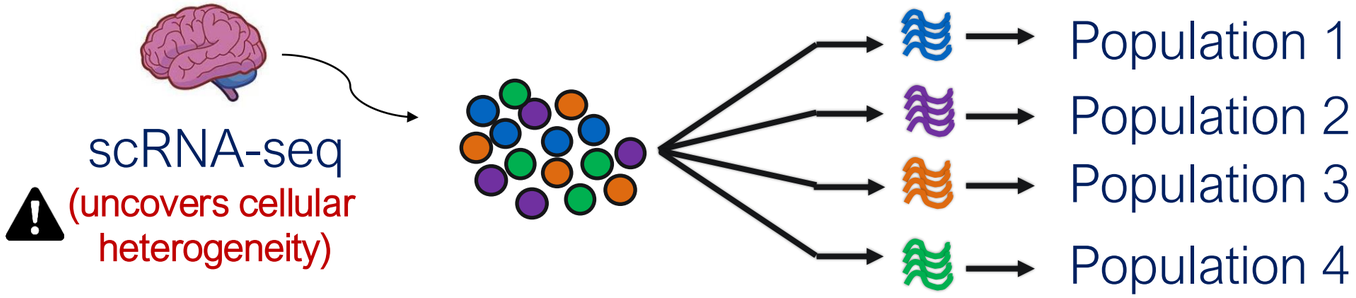
(masks cellular heterogeneity, averaging misleading)

average expression level across a population

Bulk RNA-seq good for -

- comparative transcriptomics
- ID disease biomarkers
- homogenous systems

Allows identification of broad level differences in gene expression



scRNA-seq

(uncovers cellular heterogeneity)

scRNA-seq good for -

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

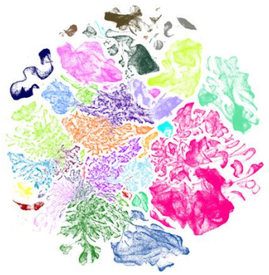
Allows identification of cell to cell variation in gene expression

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

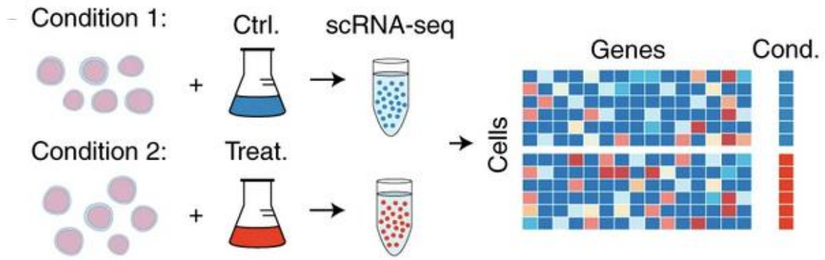
Bio. trajectories/cell fate, Dev timelines, lineage tracing

e.g. embryogenesis

c) "screening"-type studies
- Single cells as individual expt.

Uncover GEX diff on perturbation

e.g. CRISPR studies, Therapeutics discovery

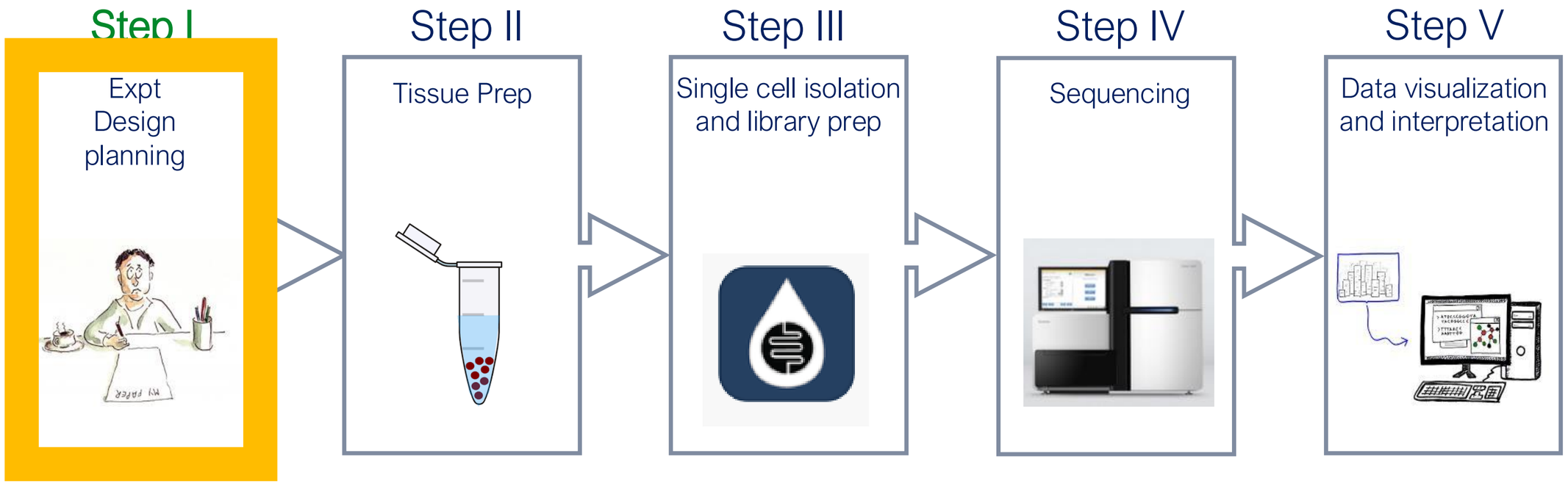


The Central Idea

Every computational result has an experimental history.

We'll follow the scRNA-seq workflow and ask:
How does each experimental decision shape what the bioinformatician sees?

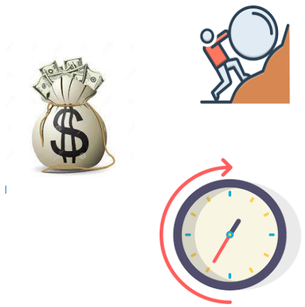
The scRNAseq workflow



Goal: Put a well-thought-out, holistic plan in place

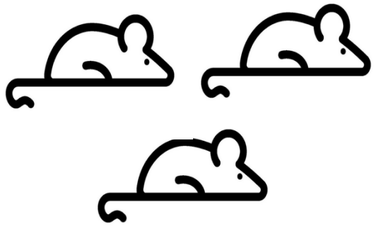
Practical planning

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

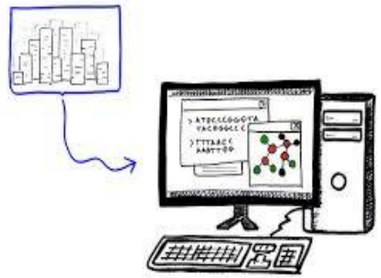


Resources

Sample type:
cells vs nuclei/
fresh vs frozen
abundance,
scale



Technical &
biological replicates



Bioinformatics &
analyses capabilities,
Cloud storage,
Computing power

Your Intellectual Framework

1. **Biological question:** What am I actually trying to resolve at single-cell resolution?

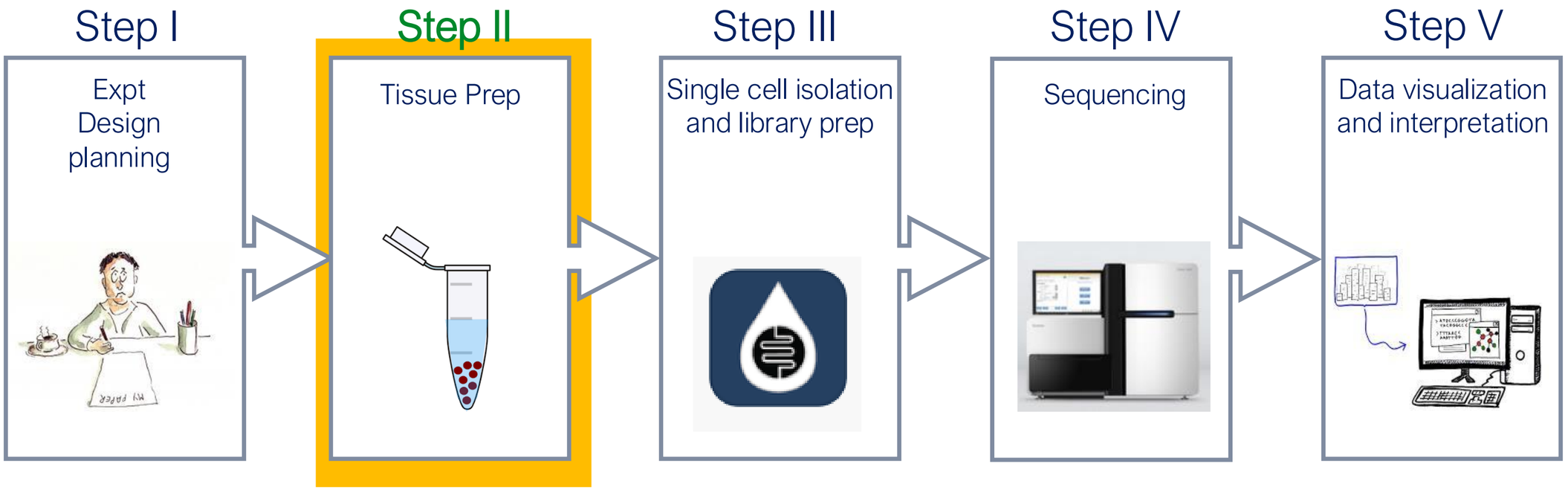
2. **Tissue/sample handling:** How much biology could I have changed before the cells ever entered the workflow?

3. **Dissociation:** Am I preserving the populations and states that existed in the tissue—or selecting for the cells that survive my protocol?

4. **Cell quality and representation:** Do I have enough viable, representative cells to answer the question?

5. **Experimental design / replication:** Can this experiment distinguish biological variation from technical variation?

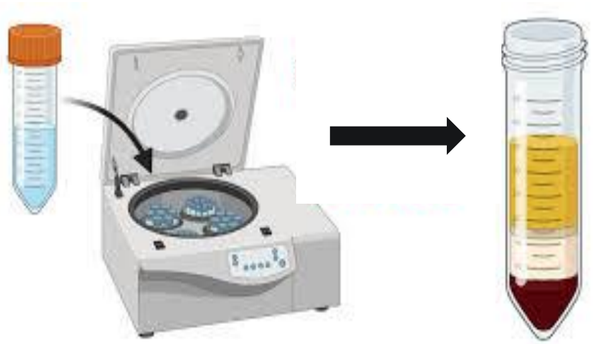
The scRNAseq workflow



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

Dissociation: making a single cell suspension

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



Centrifugation

Vortexing

Pipette mixing

2. Enzymatic dissociation

3. Combination

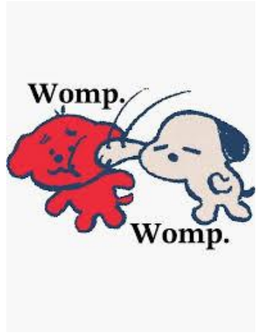
4. Automated

Enzymes (trypsin, collagenase liberase, accutase)

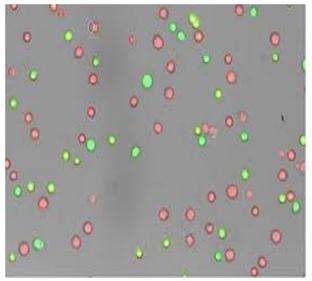
Dissociator

No universal protocol

Requires optimization of protocol for every tissue/cell type



The Top 5: Decisions that shape the dataset



Cell Viability

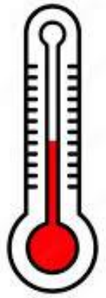
High

Cell Count

Accurate

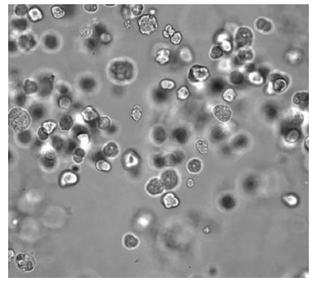
Time

Short



Temp

4°C



Quality

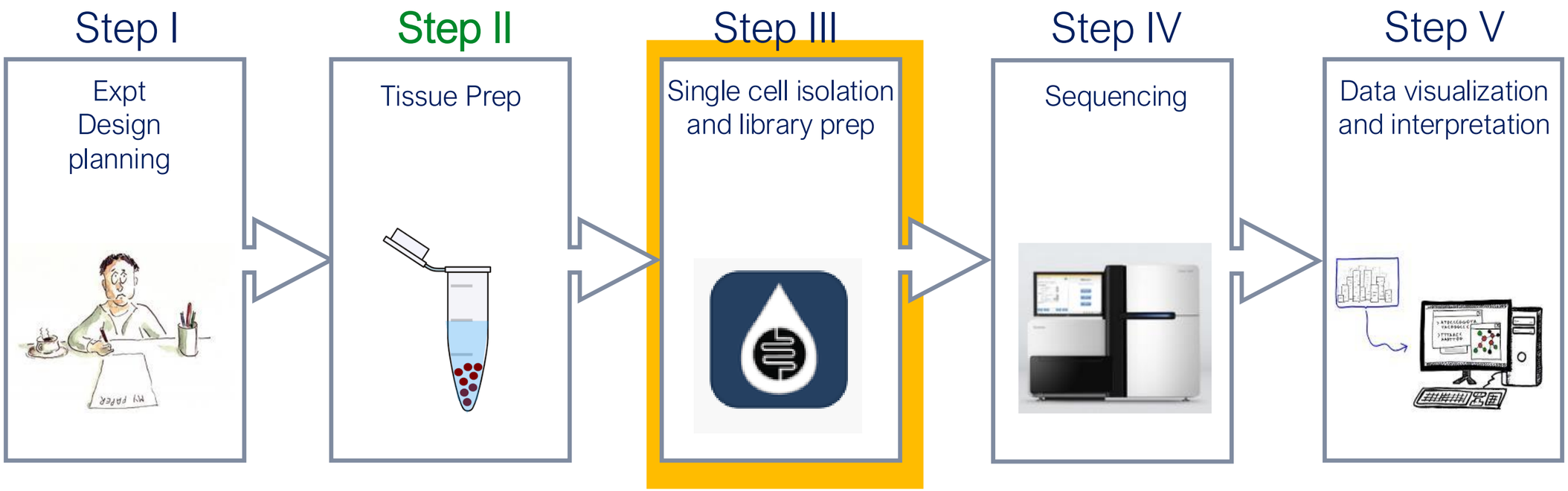
Debris,
Clump free

Consequences: Bioinformatics won’t fix sloppy bench work

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

Garbage in, Garbage out

The scRNAseq workflow



Goal: Generate high quality, sequencable libraries

Core principles of single-cell library preparation

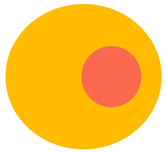
Every platform differs in chemistry, but they all aim to:

- preserve cell identity
- barcode molecules
- generate sequencing-ready libraries (exact structure varies by platform)

The implementation differs between technologies; the underlying principles remain the same.

Illustrative example: 10x genomics scRNAseq library

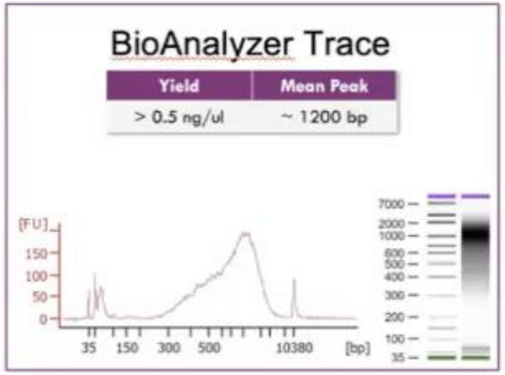
1. Isolate single cell and lyse



2. mRNA capture & barcoding molecules

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

Amplified cDNA

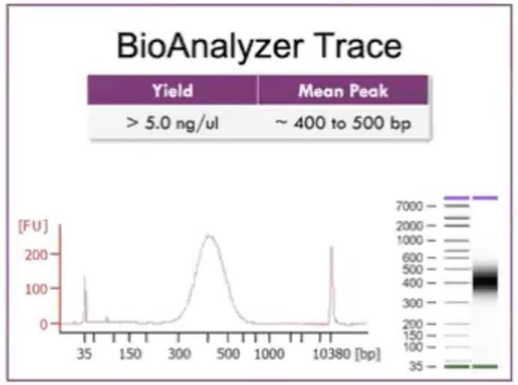


- Quantify
- Size
- contam

4. Fragment cDNA, add adaptors for Illumina sequencing

5. PCR amplify library, QC, pool and sequence

Index-PCR Library



- Quantify
- Size
- contam

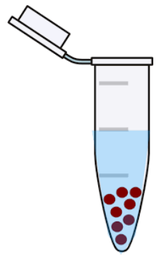
scRNA-seq workflow

Step I

Expt Design planning

Step II

Tissue Prep



Step III

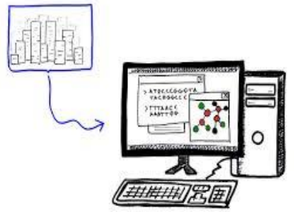
Single cell isolation and library prep

Step IV

Sequencing

Step V

Data visualization and interpretation



Goal: sequence your libraries on the appropriate sequencer

STEP IV: Sequencing platforms for scRNAseq

Common compatible sequencing systems -

More power/output
Simple benchtop
Affordable & low cost
Fast 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: coverage aka sequencing read depth

Choose sequencing capacity based on the biological question—not simply on what's available

Example: You have barcoded 10K cells from 4 samples for 5'GEX = 40K barcoded cells
40K x 50,000 reads/cell = 1Billion 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 $$	6.6–8.2 B $$$	16–20 B $$$$

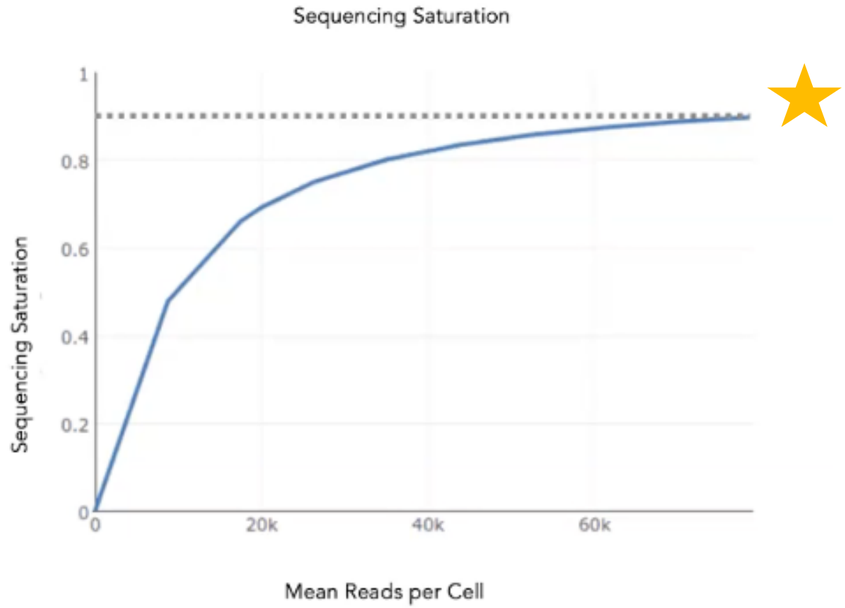
STEP IV: dialing in on sequencing saturation

Beyond a point, more reads ≠ better biology

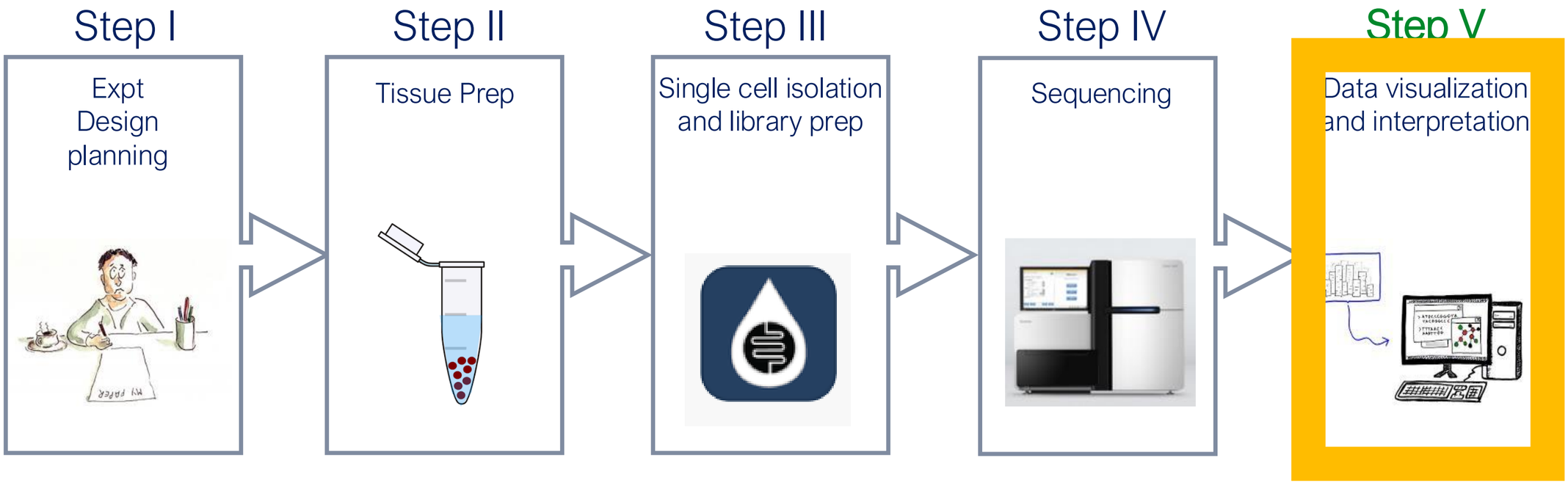
$$\text{Seq saturation} = \frac{\text{\# of unique mRNA detected}}{\text{\# of total reads}}$$

- 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 = $$$

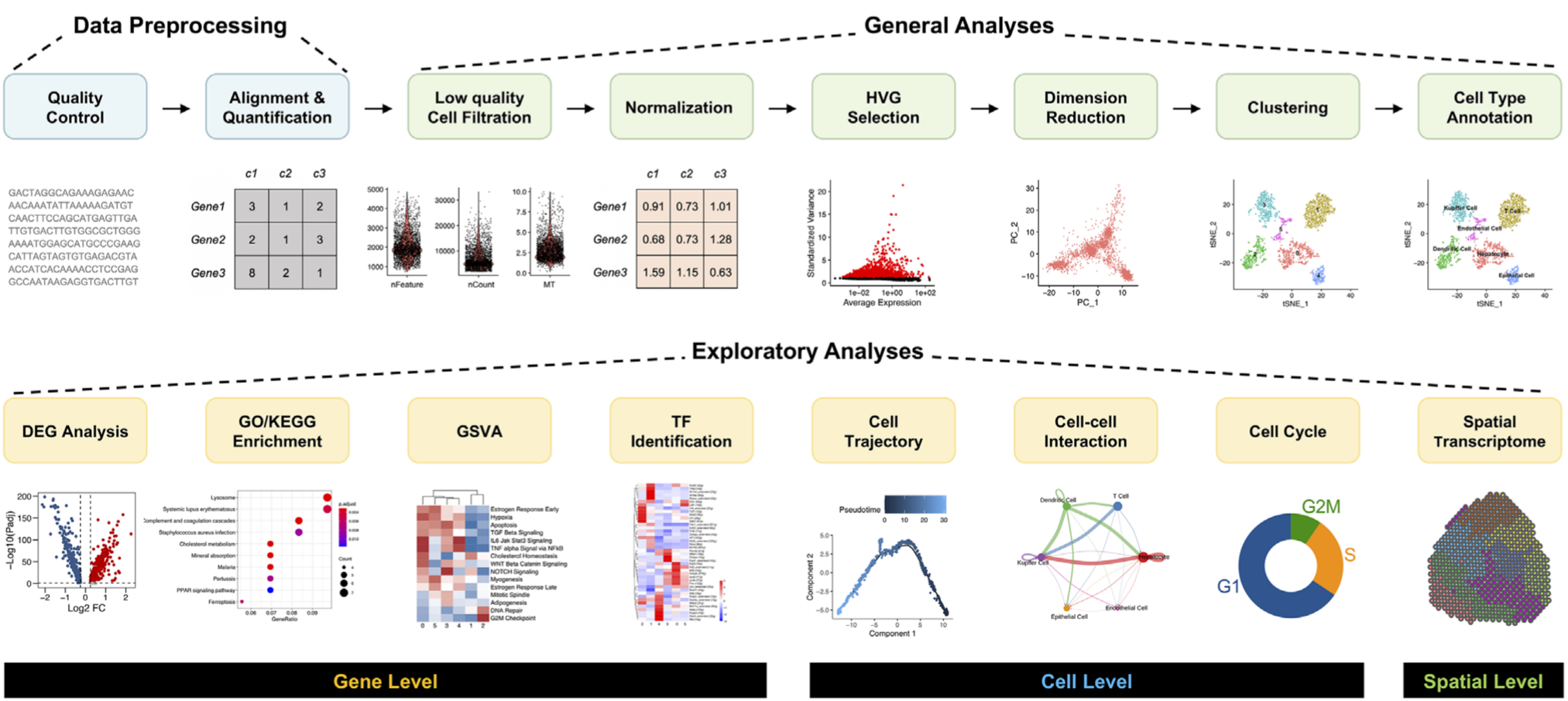


scRNA-seq workflow



Goal: analysis, interpretation and visualization of data

Key analysis steps in scRNAseq analysis



Kharchenko lab., 2021

Defining metrics for good quality scRNAseq data

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

Quality metrics -

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

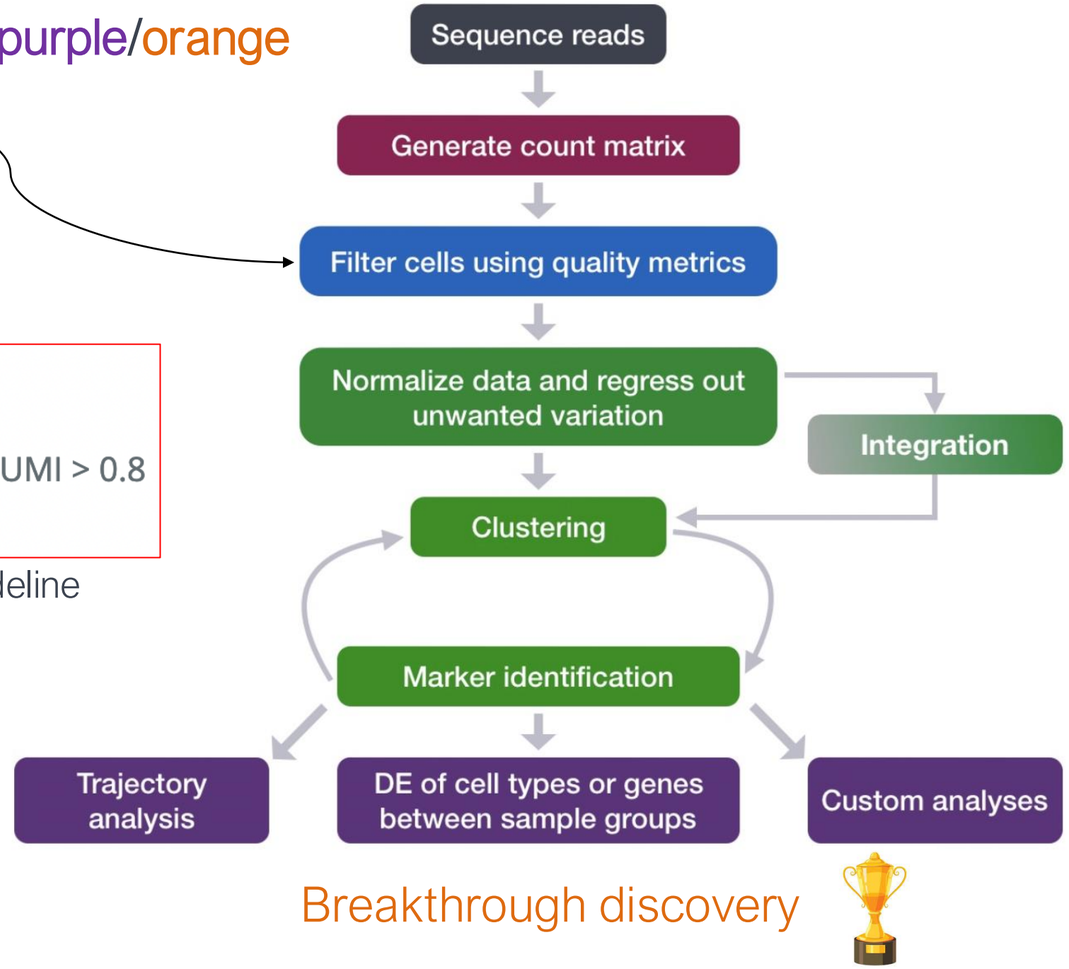
(and many others)

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

General guideline

- no metric in isolation
- setting thresholds is an iterative process

(starting from lenient to stringent)



QC: What are we actually asking about the biology?

QC Metric	Biological question
Cell count	Have I preserved the biological diversity that was present in the original tissue?
UMIs	Did we capture enough molecules to accurately represent this cell?
Genes detected	Does this cell contain enough biological information to be informative?
Mitochondrial % RNA	Was this cell healthy when we captured it?
Novelty / Complexity Score	How much biology did we successfully capture from each cell?
Doublets	Am I actually looking at one cell?
Ambient RNA	Am I measuring this cell, or RNA from another cell?
Batch effects	Is this real biology or technical artefacts?
Sequencing saturation	Have we learned everything this library can realistically tell us?

Do a (small scale) pilot experiment

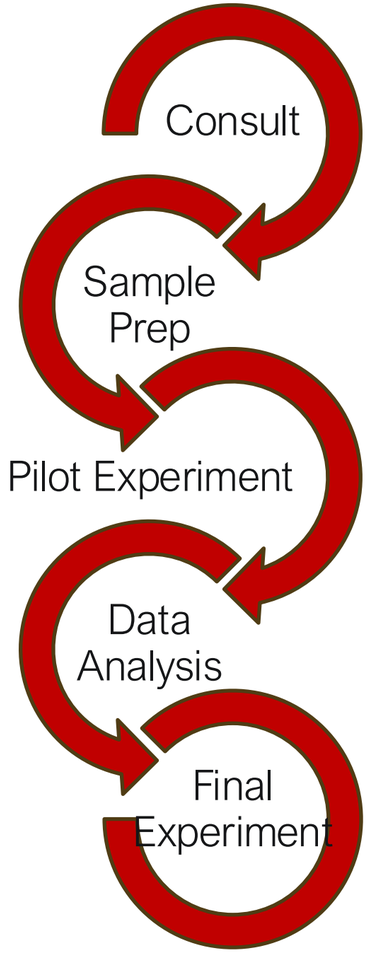
Do not rush to the final experiment

A well-planned pilot experiment is essential for:

- ✓ coming up w/ well defined bio. objectives
- ✓ rational expt design/optimal approach for research Q
- ✓ evaluating sample preparation
- ✓ Identifying sources of technical noise (handling vs sequencing bias)
- ✓ figuring out the required number of cells needed to

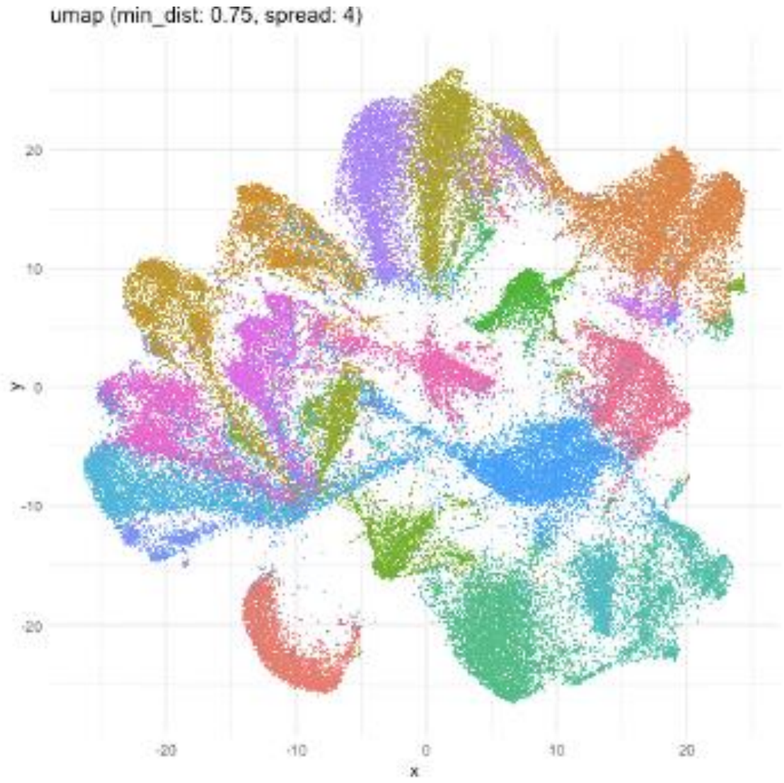
  (statistically) answer your biological question

A Pilot brings the talk back from bioinformatics to expt decision-making



Biological payoff: One beautiful UMAP and a stunning discovery

This is the the type of biological insight all those upstream decisions enable



PC: Pachter Lab

Key takeaways!

1. Biology begins before sequencing.
2. Every computational result has an experimental history.
3. Cells are observations; biological replicates are your experimental unit.
4. QC metrics are biological questions—not just thresholds.
5. Statistics cannot compensate for poor experimental design.

Bioinformatics begins at the bench