HARVARD MEDICAL SCHOOL | BLAVATNIK INSTITUTE SYSTEMS BIOLOGY

The Single Cell RNA-seq Workflow:

A practical guideline to ensure experimental success

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

Single Cell Core at HMS Quad

We are here!

200 Longwood Ave., Armenise Room 517

Single Cell Core at HMS Quad

The Team

Dr. Mandovi Chatterjee
Director

Dr. Arpita Kulkarni
Associate Director

Dr. Pratyusha Bala
Associate Director
(Spatial Transcriptomics)

Theresa Torre
Research Technician

Dr. Ollie
Single Cell Pawlice,
Officer of the Floof

https://singlecellcore.hms.harvard.edu

First name_Last name @ hms.harvard.edu

Single Cell Core at HMS Quad

Faculty Advisors

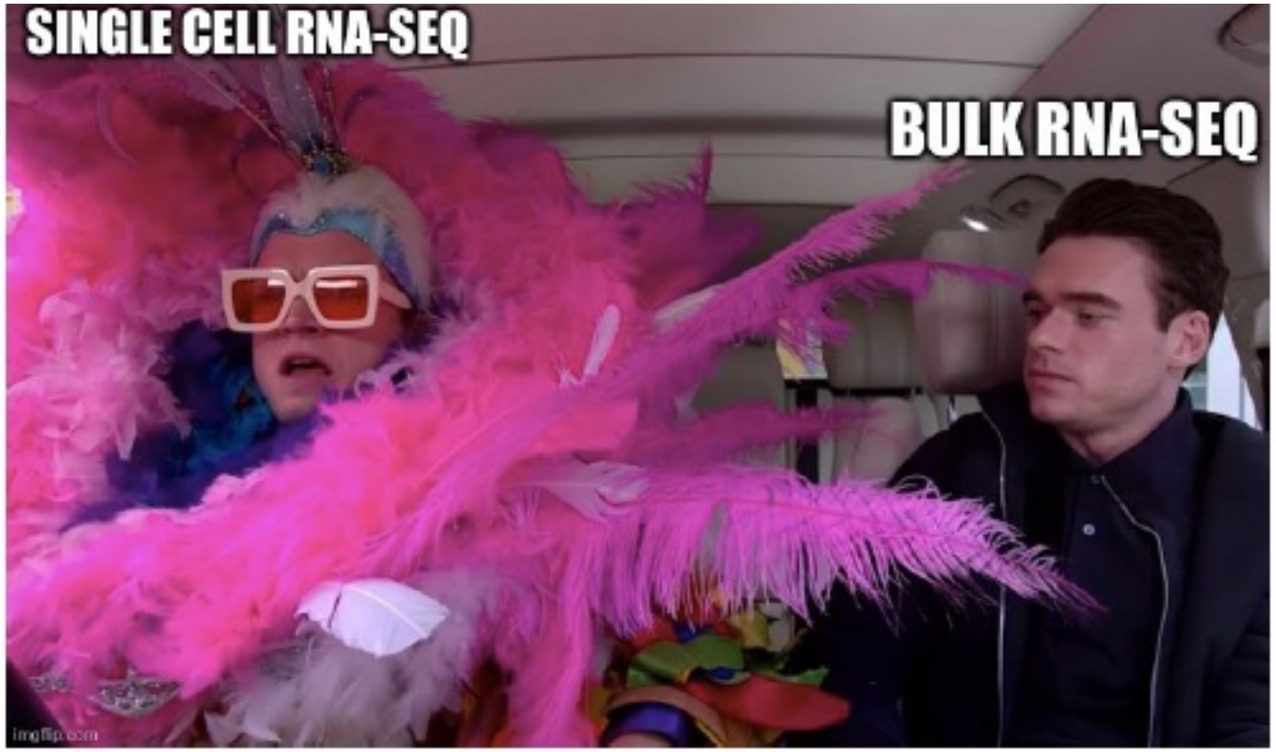
Dr. Allon Klein

Dr. Jeff Moffitt

Dr. Chris Benoist

Single Cell Core at HMS Quad

Mission: Enable novel discoveries by assisting in the design, execution and interpretation of single cell and spatial -omics assays using cutting edge technology



Single Cell Core at HMS Quad

We house different high-throughput platforms that allow encapsulation, barcoding and library preps from single cells for single cell/spatial -omics

10x GENOMICS

vizgen

BD

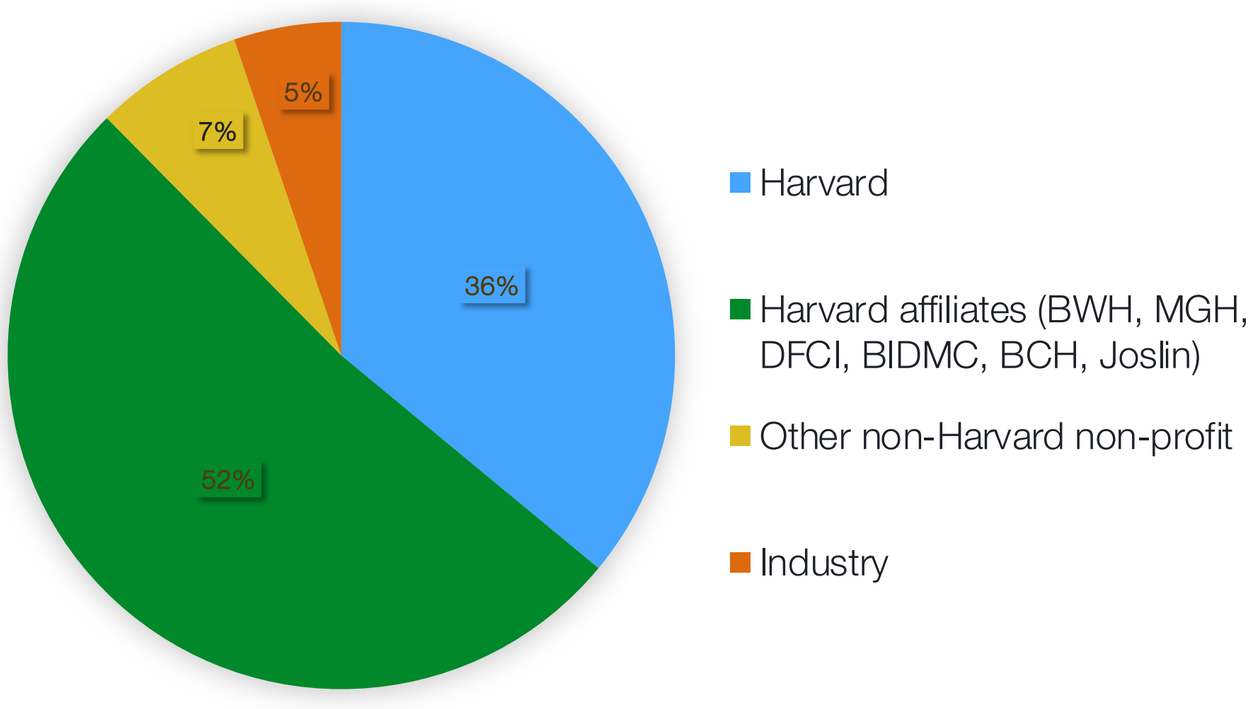
Parse BIOSCIENCES

AtlasXomics

Fluent BioSciences

curio BIOSCIENCE

STOmics



- We are oldest single cell core on campus!
- a fee-for-service core
- >500 PI's and 50,000+ samples

Key Services

❖ **Consultations**

❖ **Single Cell mRNA barcoding**
- inDrops (sunset FY2024)
- 10x Genomics
- BD Rhapsody
- Parse Biosciences (SPLiT-seq)
- Fluent Biosciences (PIP-seq)

❖ **Applications**
- Library Preparation (scRNA-seq for 3' and 5'GEX, scATAC-seq, Multiome, CITE-seq, Hashtagging, MULTI-seq, CellPlex)
- Sequencing Coordination

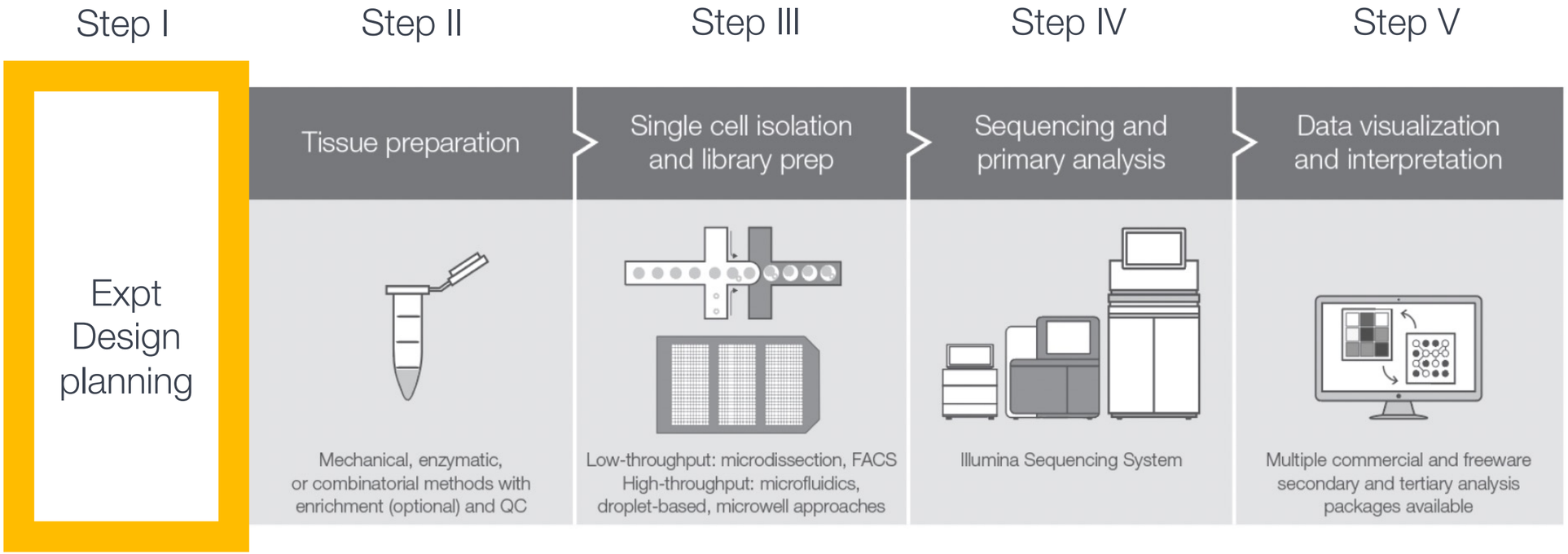
❖ **NEW Single cell spatial transcriptomics & epigenomics**
- Visium (10x Genomics)
- MERFISH (Vizgen)
- Stereo-seq and DBiTSeq (AtlasXenomics)

❖ **NEW Long read seq lib preps for PacBio & Nanopore**

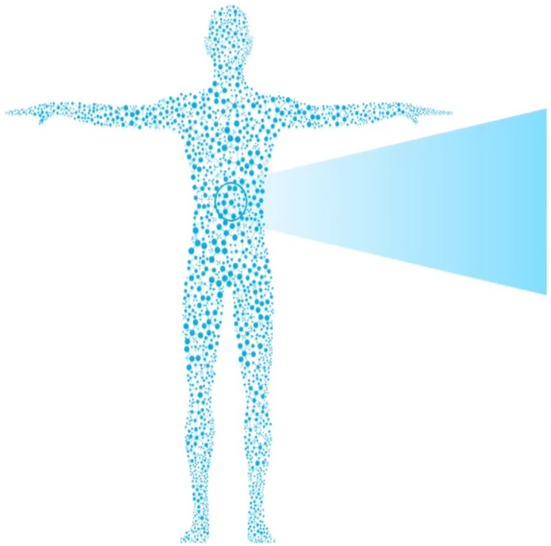
❖ Teaching HSPH Chan Bioinformatics Core

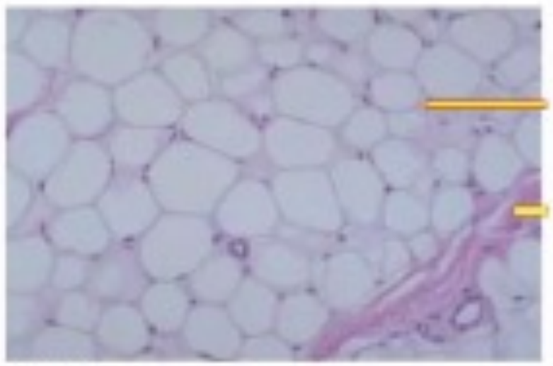
Outline for today's talk

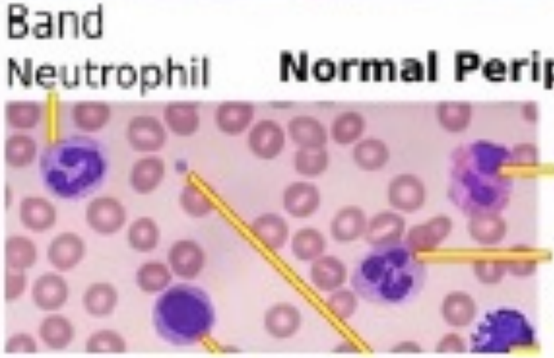
- scRNAseq background
- scRNAseq workflow

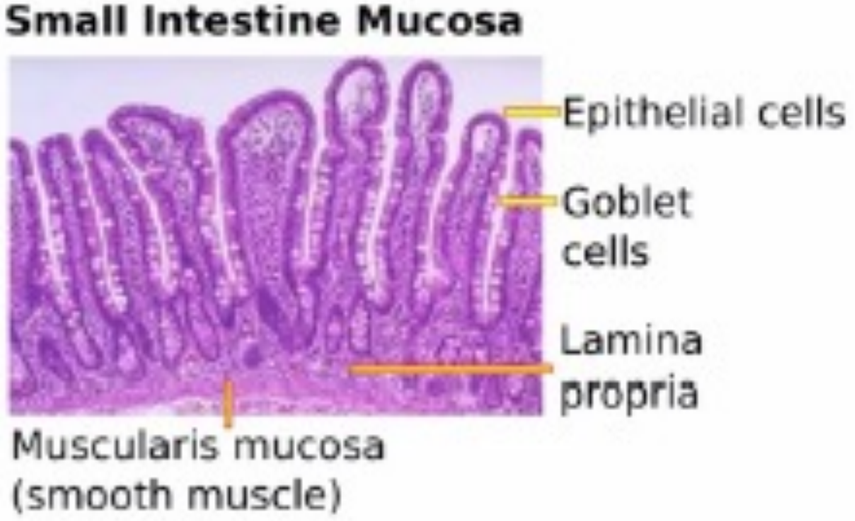


We know tissues are heterogeneous



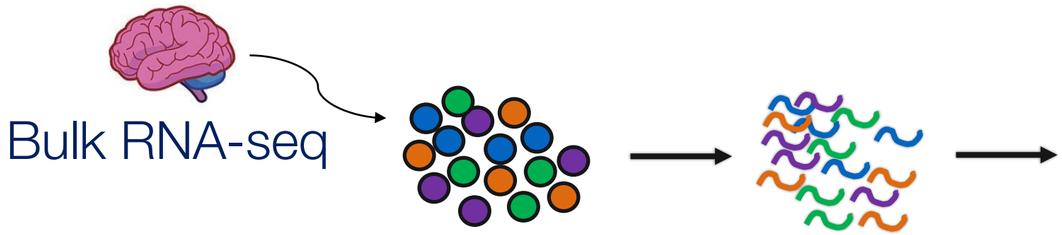






10x GENOMICS

Bulk RNA Sequencing (early ~2000s)

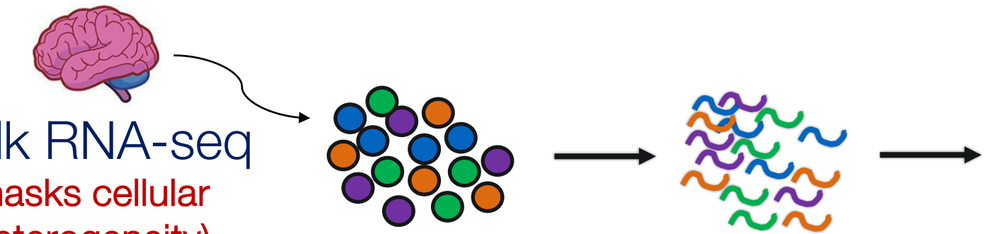


average expression level across a population

Bulk RNA-seq good for -

- comparative transcriptomics
- disease biomarker
- homogenous systems
- Great for studying broad level differences

Bulk RNA Sequencing (early ~2000s)

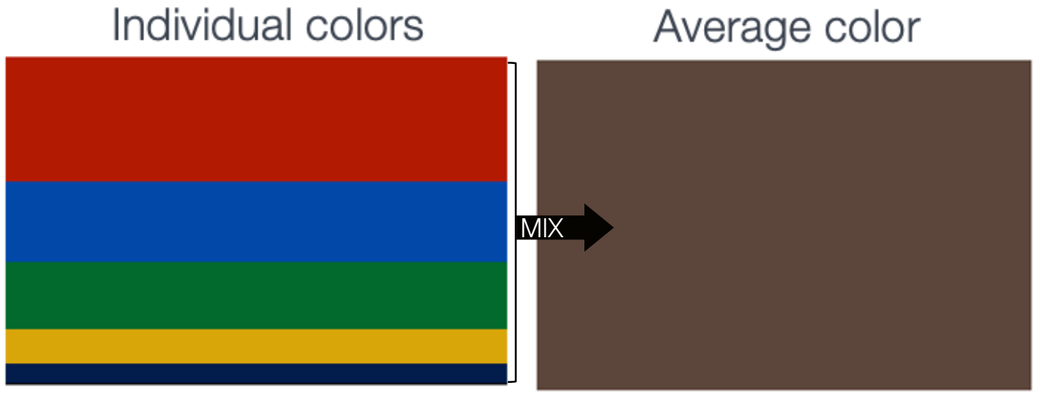


Bulk RNA-seq
(masks cellular heterogeneity)

average expression level across a population

Bulk RNA-seq good for -

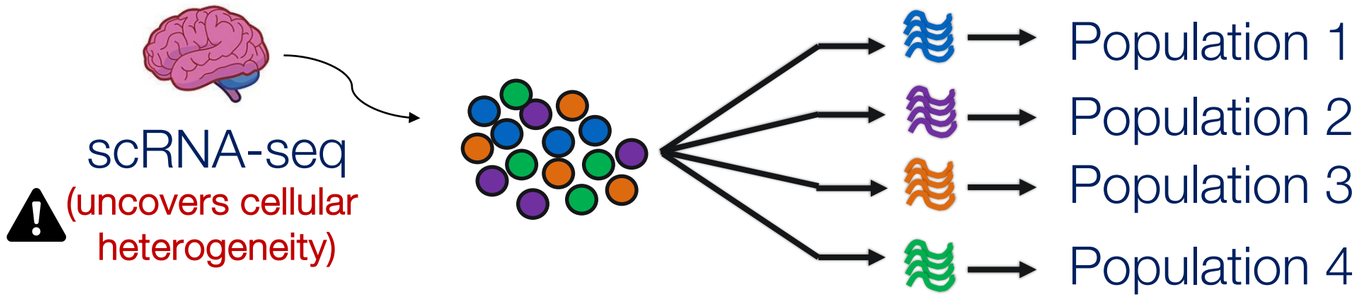
- comparative transcriptomics
- disease biomarker
- homogenous systems
- Great for studying broad level differences



The average does not represent any single color

Sometimes averages are not useful!

Single Cell RNA Sequencing (~2009)

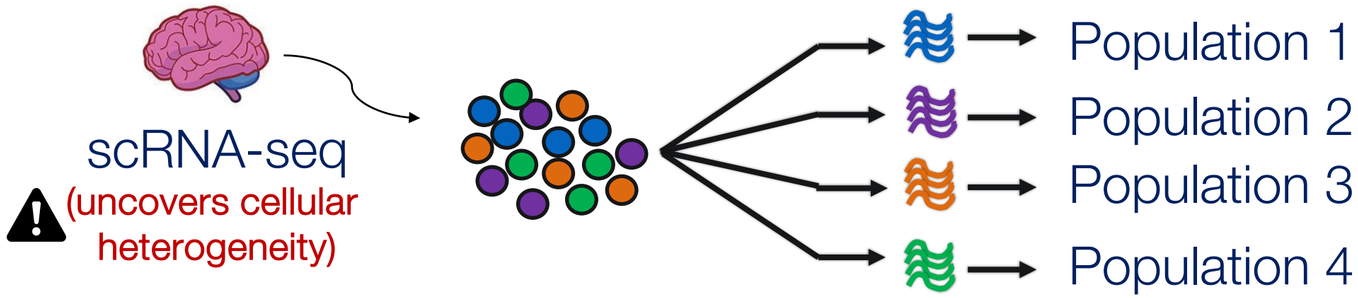


scRNA-seq good for -

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

Captures cell to cell variation in gene expression

Single Cell RNA Sequencing (~2009)



scRNA-seq good for -

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

Captures cell to cell variation in gene expression

The main difference between bulk and scRNA-seq is that in the latter each sequencing library represents a single cell, instead of a population of cells

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 an either/or situation

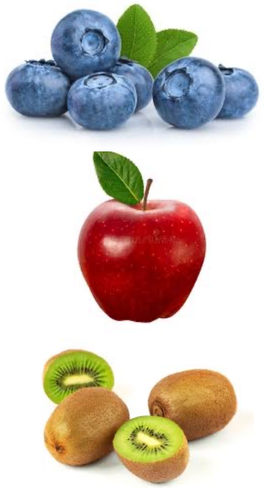
smoothie

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

Data Quality – Different Transcriptome Coverages (mRNA)

"Bulk RNAseq"

- Higher starting RNA material (500ng-)
- **>~20,000 transcripts per cell** (more when consider splice variants / isoforms)
- Captures >**80-95% of transcriptome**

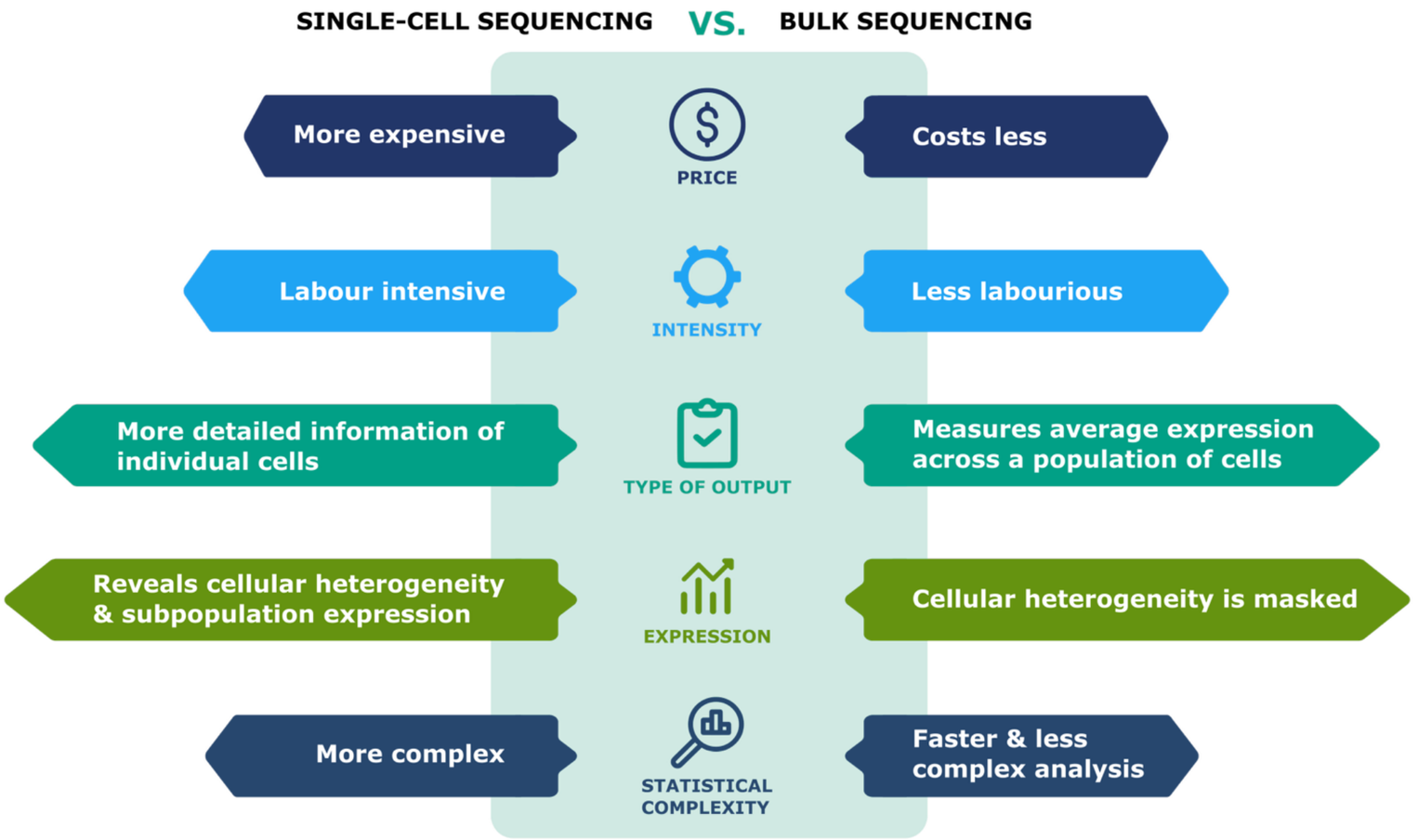
depending on sequencing depth

"Single Cell Methods"

- Lower starting RNA (noisier gene expression) 10^3-10^6 cells
- **200 -10,000 transcripts per cell**
- Capture <**10-40%** of the transcriptome

scRNAseq can be very powerful when done correctly,
but you want to be sure that it is the best method for your Q

Single cell vs Bulk RNA Sequencing: The face-off

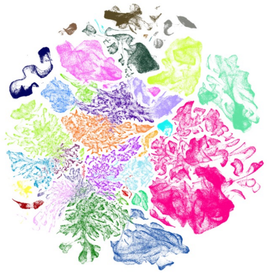


Common applications of scRNA-seq

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

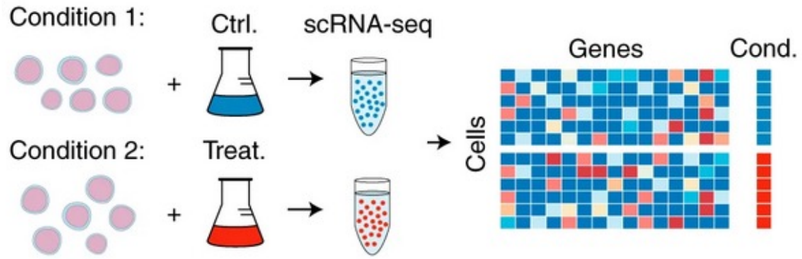
Bio. Trajectories/cell fate, Dev timelines, lineage tracing

e.g. embryogenesis

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

Uncover GEX diff on purturbation

e.g. CRISPR studies



“-omics” one can study at single cell level

Single cell _____

- Transcriptomics
- Epigenomics
- Genomics
- Proteomics
- Metabolomics
- Microbiomics
- Lipidomics
- Glycomics
- Multiomics

Each “–omics” produces large data

BUT

Integrating big data from multi “–omics” presents a considerable statistical challenge

Spatial transcriptomics at (sub)cellular resolution

High resolution spatial profiling of scRNA expression in their native context

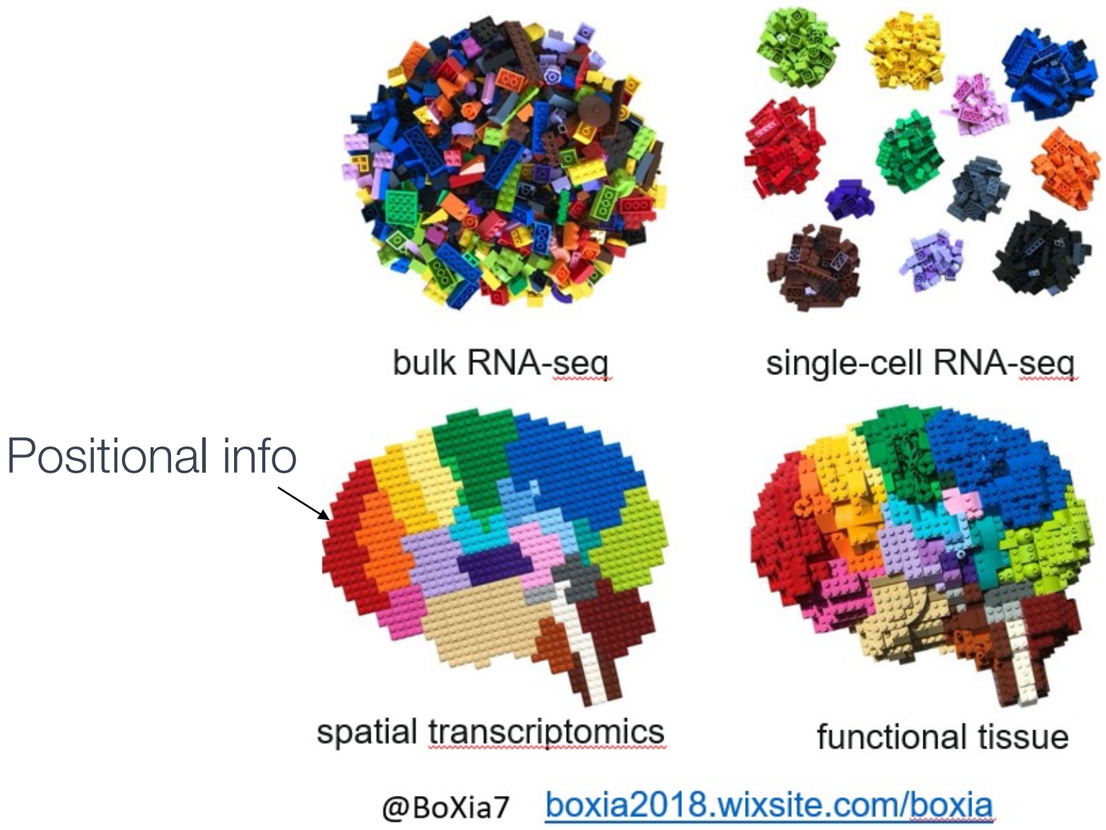
- Sequencing or Imaging based
- Require fresh-frozen tissue sections

Examples:

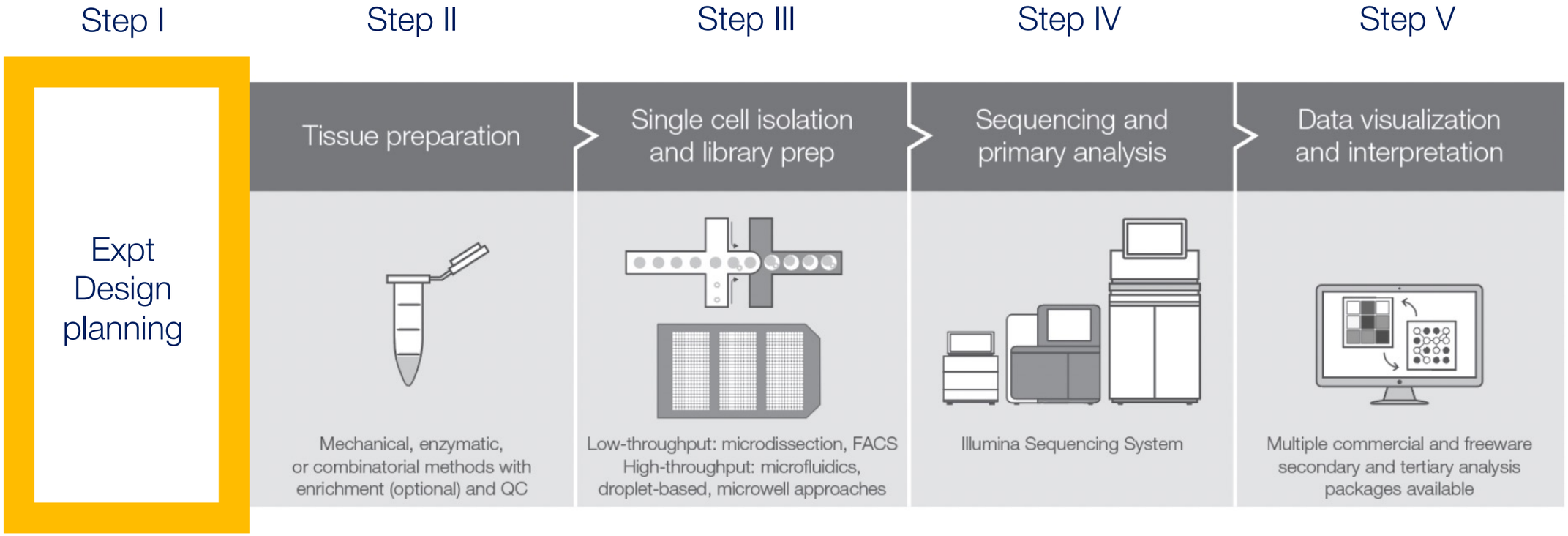
10x's Visium and Xenium

Vizgen's MERSCOPE

Nanostring's GeoMx and CosMx



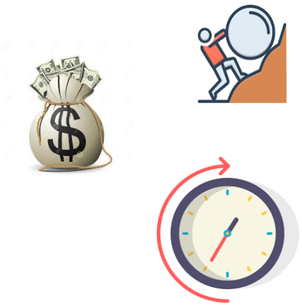
scRNA-seq workflow – STEP I



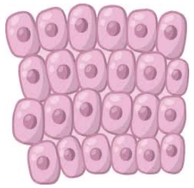
Goal: Put a well-thought-out, holistic plan in place!

Experimental design considerations

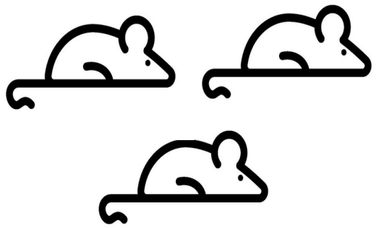
End goal:
Hypothesis testing
Publication?
Grant-writing?
New study or conti.?



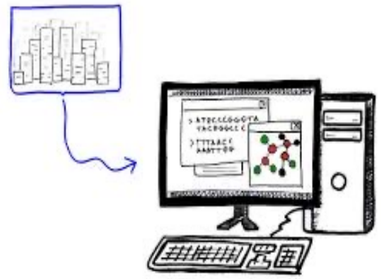
Resources



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

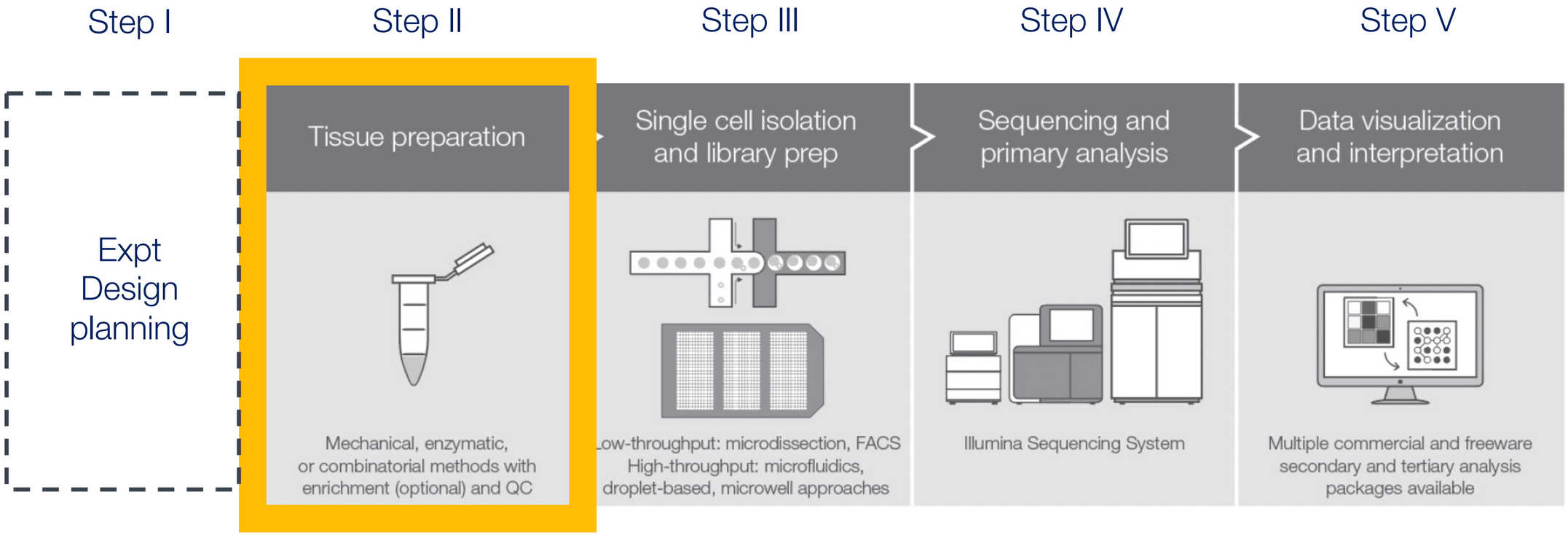


Technical &
biological replicates



Bioinformatics &
analyses capabilities,
Cloud storage,
Computing power

scRNA-seq workflow – STEP II

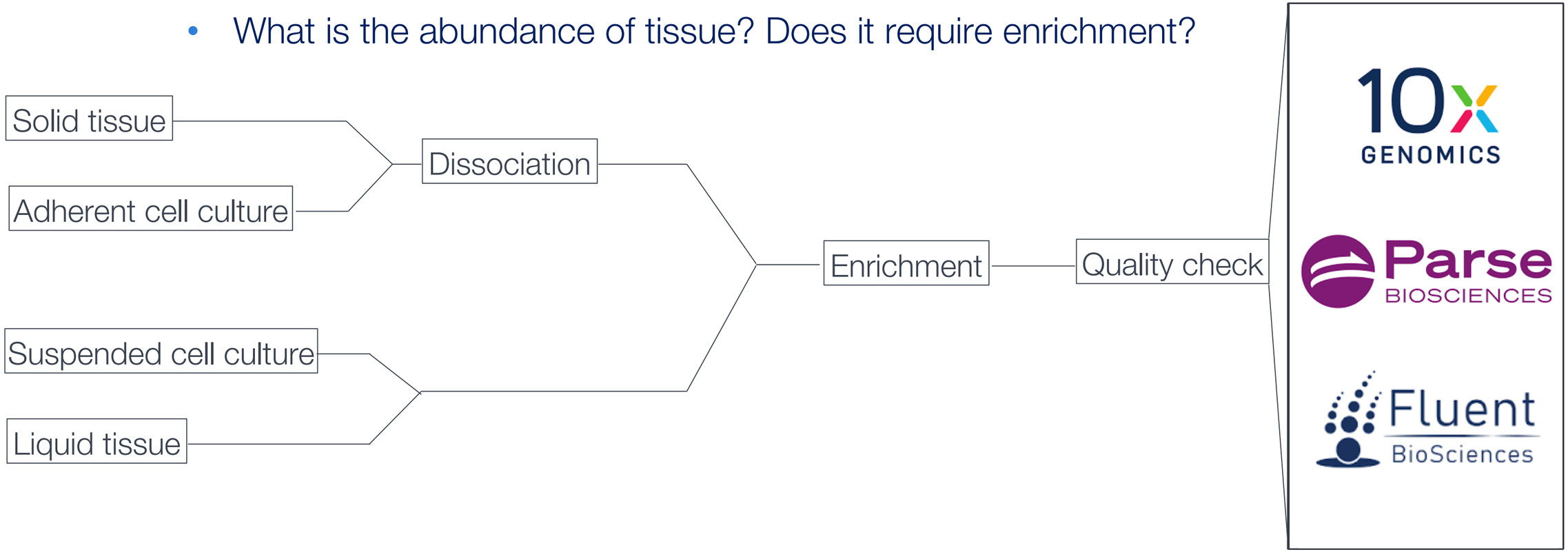


Goal: Get high quality, viable, single cell suspension from tissue, assess prep and do sample QC

STEP II – Tissue preparation

Logistics: What is your sample of interest? How would you obtain that?

- Which population in a tissue should be examined?
- What is the abundance of tissue? Does it require enrichment?



Making a single cell suspension: Dissociation

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

Centrifugation

Vortexing

Pipette mixing

2. Enzymatic dissociation -

Enzymes (trypsin, collagenase)

3. Combination

Dissociator

No universal protocol

Requires optimization of protocol for every tissue/cell type

Making a single cell suspension: Dissociation

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

Centrifugation

Vortexing

Pipette mixing

2. Enzymatic dissociation -

Enzymes (trypsin, collagenase)

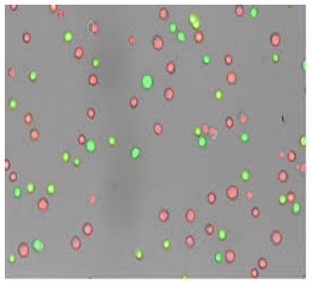
3. Combination

Dissociator

How to find a protocol-

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

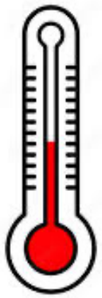
Factors affecting sample preparation and quality



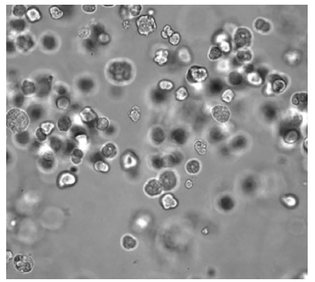
Cell Viability

Cell Count

Time



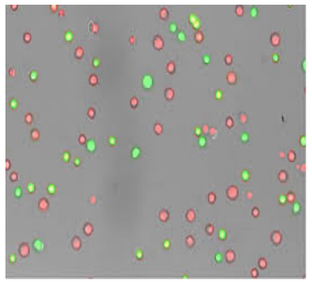
Temp



Quality

Factors affecting sample preparation and quality

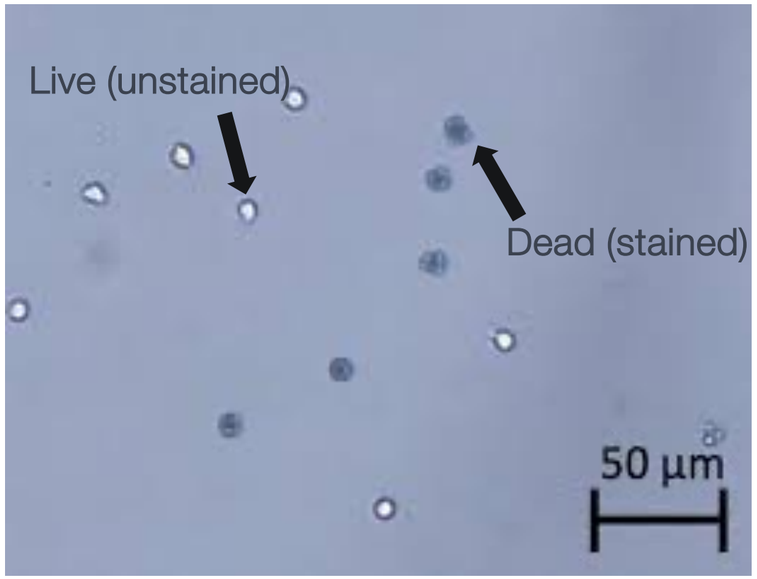
1



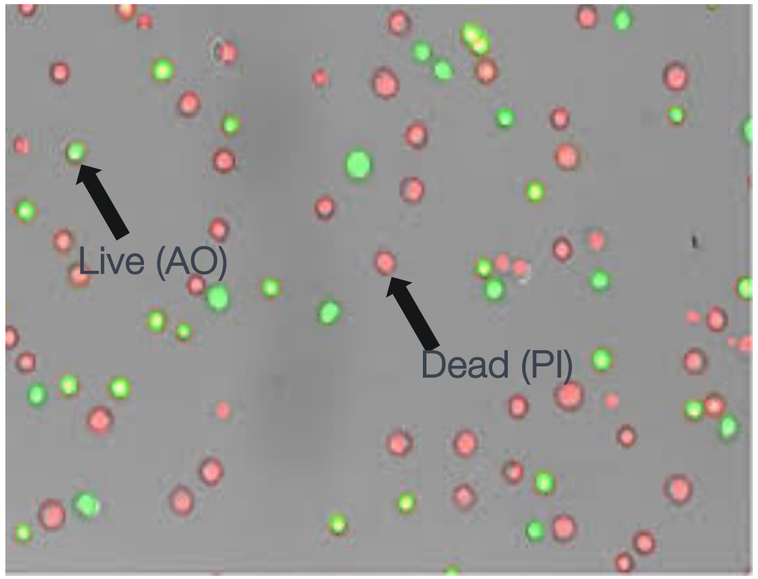
Cell Viability
(high >70%)

Dead cell removal,
Enrichment for live cells

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



Trypan Blue (dead)



Acridine orange (live)/
Propidium iodide (dead)

Factors affecting sample preparation and quality

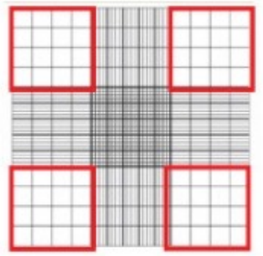
2

Cell Count
(accurate)

Manual counting

Wrong counts can lead to -

- wrong interpretation of the biology or uninformative expts
- high duplets/multiplets (tight maths involved)



- Calibrate using manual counting (hemocytometers)
- Cell sorter counts off by as much as 5-50%
- Wrong counts require higher seq depth,

  = wasted $$

Factors affecting sample preparation and quality

3

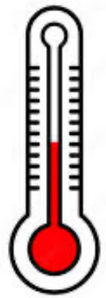
Time
(short)

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
- Mis-interpretation of biology

4



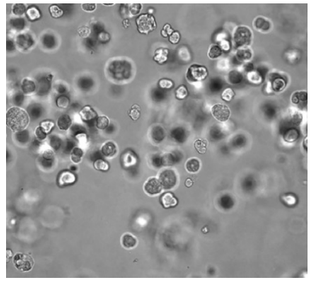
Temp.
(cold, 4C)

On ice,
RNAse-free

- RT accelerates cell death, clumping, results in “ambient RNA”
- Ambient RNA creates noisy, unusable data, at higher $
- Mis-interpretations of biology

Factors affecting sample preparation and quality

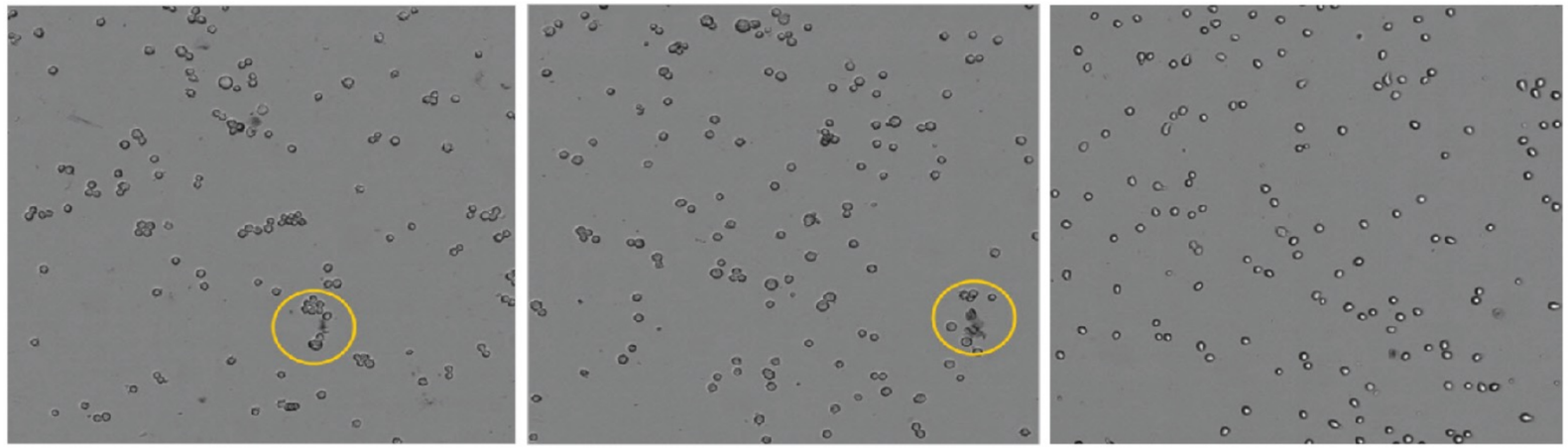
5



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!



Cell aggregates

< 10% doublets

Single cell platforms do not distinguish between
live or dead cells, debris or clumps and will encapsulate
everything

Common causes of cell clumping or poor viability

- Long prep times (>3-4h)
- Harsh dissociation conditions or harsh handling (cell pelleting, centrifugation, pipetting, FACS sorting)
- Too many dead cells
- Debris
- Using wrong buffer/media: cations like Ca^{++} and Mg^{++} [also EDTA, heparin in final media (inhibits RT)]
- Cell/Nuclear membrane damage: using DNAse to reduce clumping

Common causes of cell clumping or poor viability

- Long prep times (>3-4h)
- [illegible]
- [illegible]
- [illegible]
- [illegible]
  media (inhibits RT)]
- Cell/Nuclear membrane damage: using DNAse to reduce clumping

PRACTICE PRACTICE PRACTICE!

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

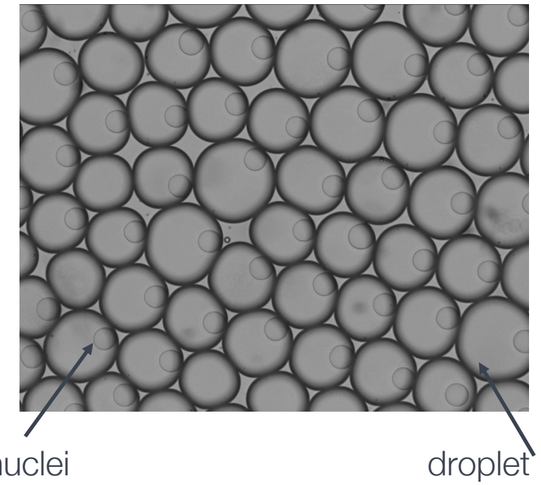
Tissue Preparation: cryopreservation/cryopreserved samples

- Several sc-papers on various cryopreservation techniques
- Success of cryopreservation is dependent on the sample/cell type (e.g. blood and immune cells do great!)
- Cell viability upon thaw is key to success
- The quality of the tissue at the time of freezing is a major factor in the quality of data downstream
- **Disadv**: you don't know ahead of time if one of your cell types in more sensitive to thawing/death at thaw, meaning you could heavily bias your sc data if you are not careful!

Use Std growth media+FBS/DMSO for best results

Tissue Preparation: single nuclei RNA-seq (snRNAseq)

- 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

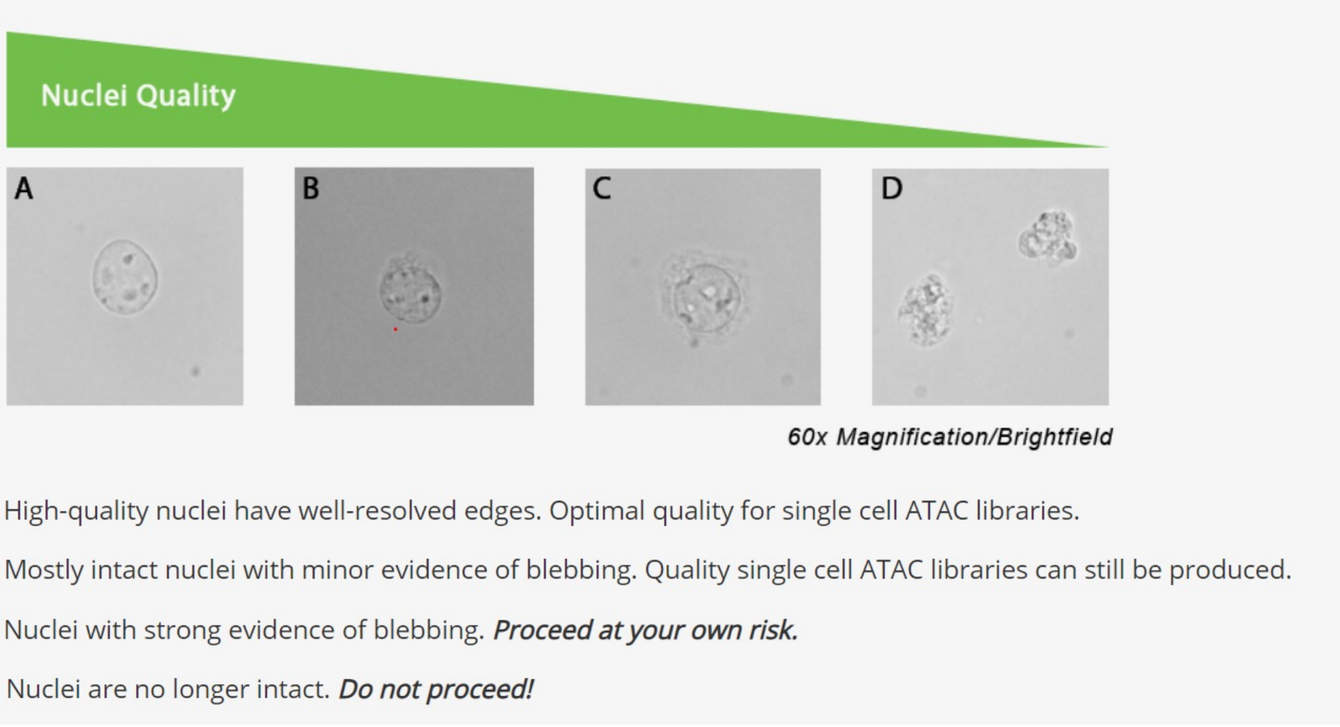
✓ ATAC (to study the epigenome)/Multiome studies (to study the epigenome along w/ transcriptome)

Data from scRNAseq is comparable to data from snRNAseq

✓ Analysis for snRNAseq different due to presence of introns

Tissue Preparation: single nuclei RNA-seq

- Good single nuclei suspension. No clumps and minimal debris
- Nuclear membrane integrity is required until nuclei are encapsulated



A: High-quality nuclei have well-resolved edges. Optimal quality for single cell ATAC libraries.

B: Mostly intact nuclei with minor evidence of blebbing. Quality single cell ATAC libraries can still be produced.

C: Nuclei with strong evidence of blebbing. ***Proceed at your own risk.***

D: Nuclei are no longer intact. ***Do not proceed!***

https://support.10xgenomics.com/single-cell-atac/sample-prep/doc/demonstrated-protocol-nuclei-isolation-for-single-cell-atac-sequencing

Garbage in, Garbage out

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

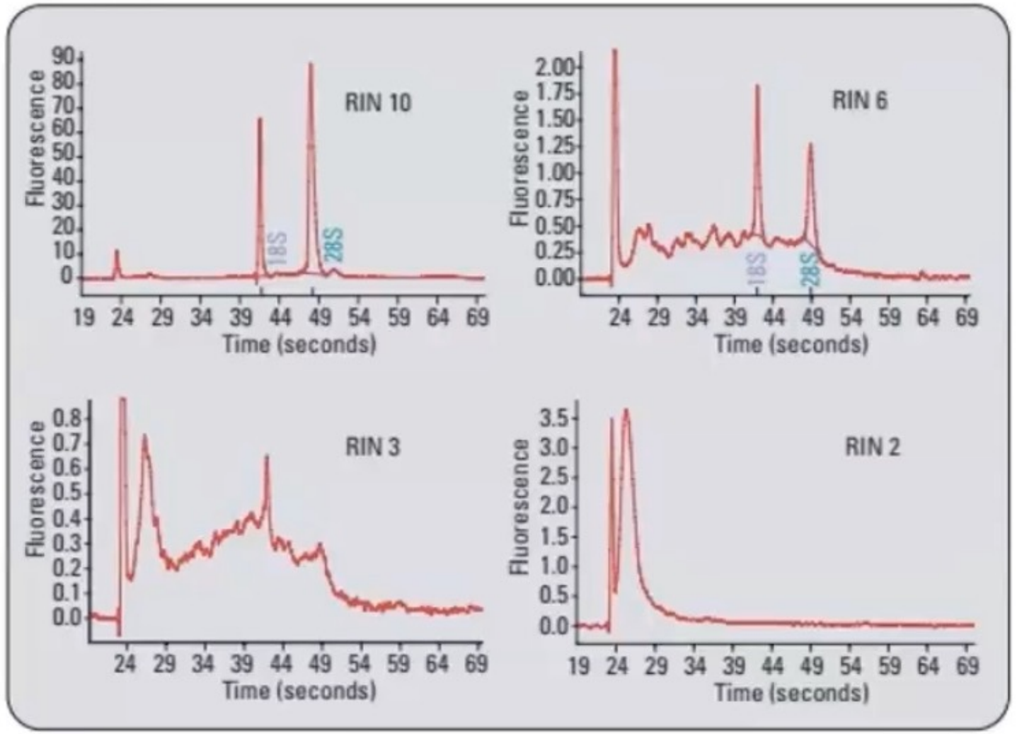
The RIN score: RNA integrity and quality check

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

1. RIN score ranges from 1-10

2. Higher the RIN score = better data

Informative for –
Prep quality
Data quality



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

Do a (small scale) pilot experiment

- Do not rush to the final experiment
- A well-planned pilot experiment is essential for

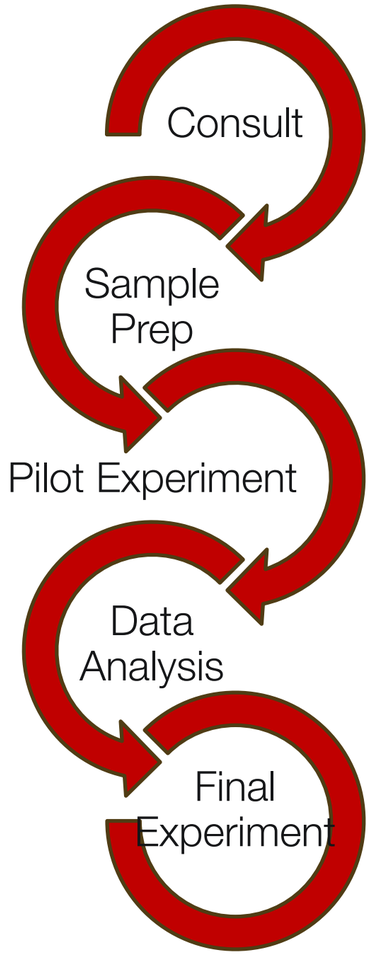
✓ coming up w/ well defined bio. objectives

✓ rational expt design/optimal approach for research Q

✓ evaluating sample preparation

✓ figuring our the required number of cells needed to (statistically) answer your biological question

- Good sample prep is the key to success



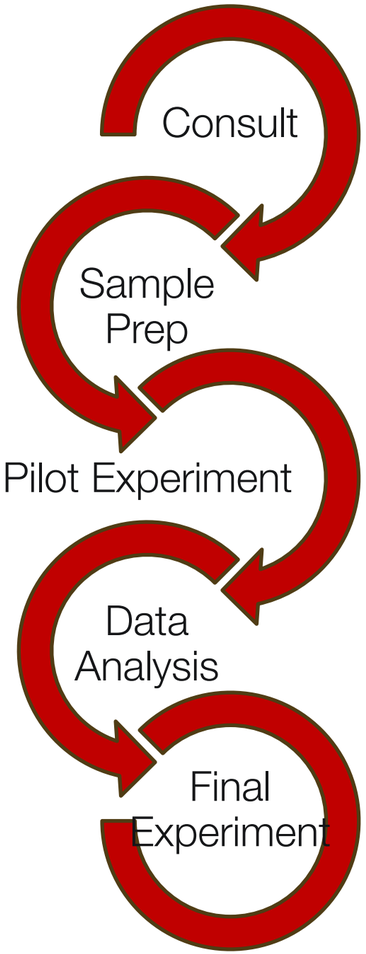
Do a (small scale) pilot experiment

What causes technical noise in single cell expts?

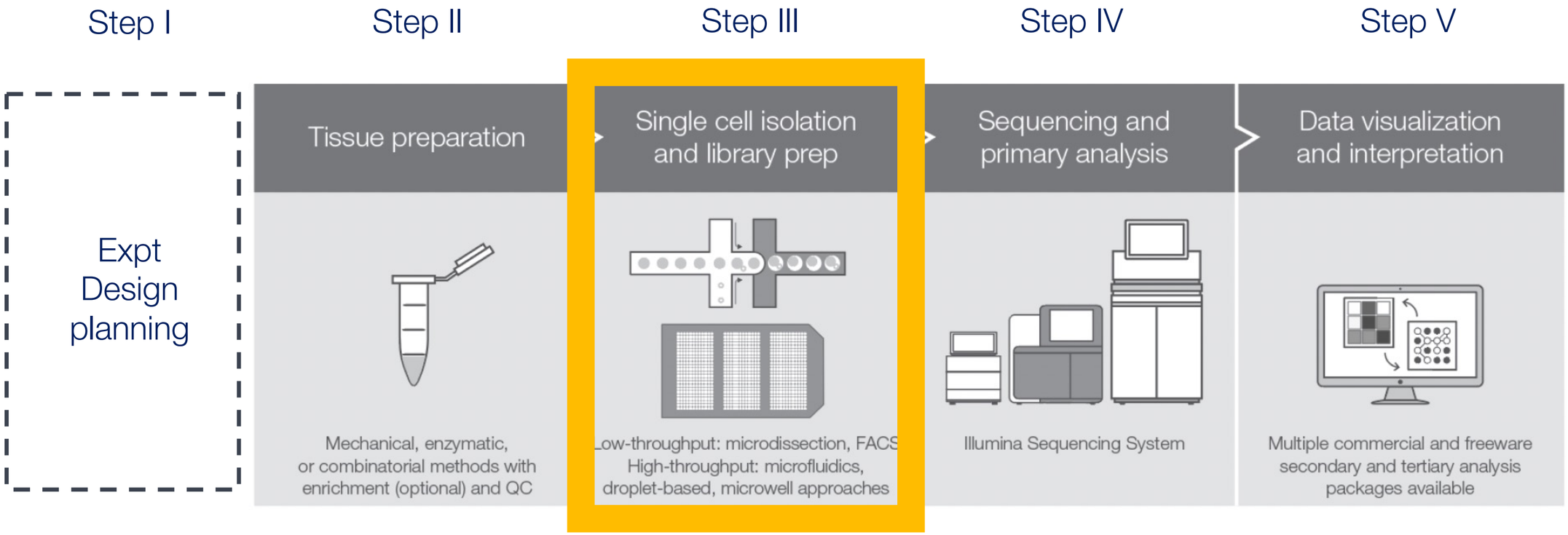
"Technical Noise": When non-biological, technical factors cause changes in the data produced by the expt. leading to wrong conclusions

2 kinds of technical noise -

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



scRNA-seq workflow – STEP III

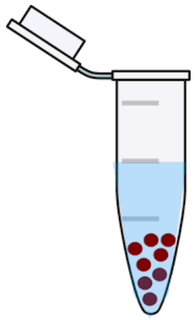


Goal: Capture and isolate single cells on identified platform, & prep libraries

Scale impacts technology choice/Technology choice impacts scale

Low Through-put approaches

High Through-put approaches



Manual

Wells

Liquid handlers

Microfluidics/Droplet

1-few cells

100-1000s cells

10K-100K cells

1M cells

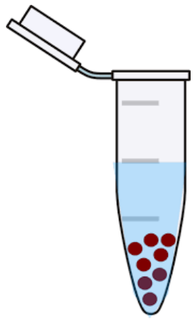
Minimal equipment
Cost effective
Scale prohibitive

Sophisticated equipment
Cost prohibitive
Scale effective

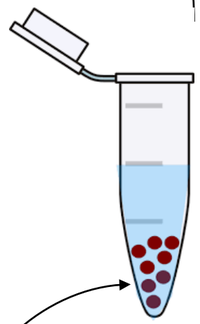
Scale impacts technology choice/Technology choice impacts scale

Low Through-put approaches

High Through-put approaches



10x GENOMICS



Parse BIOSCIENCES

Fluent BioSciences

Manual

Wells

Liquid handlers

Microfluidics/Droplet

Equipment free

1-few cells

100-1000s cells

10K-100K cells

1M cells

Minimal equipment
Cost effective
Scale prohibitive

Today's science demands scale

Minimal equipment
Cost effective
Scale effective

Parallel assays to add layered info to scRNAseq data

Transcriptome (mRNA), Genome (DNA), Epigenome (ATAC) and Protein Capture (CITEseq/HASH)

Multiple libraries from same sample for multimodal sc-analysis:

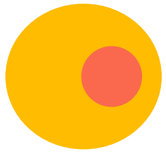
- scRNAseq (3’ or 5’ transcriptome) + scATACseq (epigenome)
- scRNAseq (3’ or 5’ transcriptome) + CITEseq (surface proteins)
- scRNAseq (3’ or 5’ transcriptome) + cell hashing (surface proteins)

More informative data at same or lower cost!

But expt has to be designed at the beginning for multimodal analysis

Structure and scRNA-seq (transcriptome) library preparation

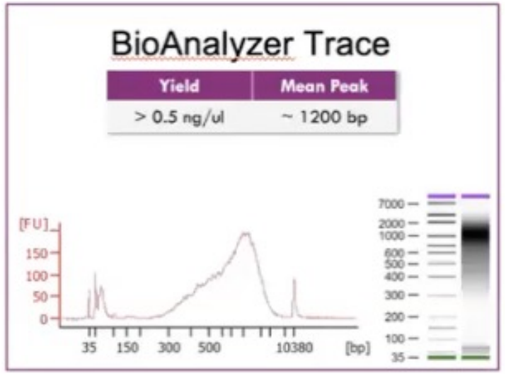
1. Isolate single cell and lyse



2. mRNA capture by PolyA tail

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

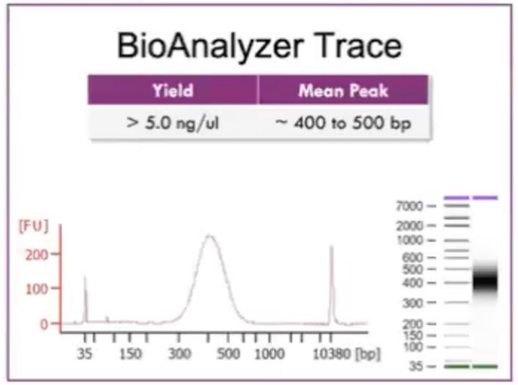
Amplified cDNA



- Quantify
- Size
- contam

4. Fragment cDNA, add adaptors for Illumina sequencing

Index-PCR Library

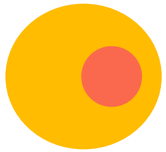


- Quantify
- Size
- contam

5. PCR amplify library, QC, pool and sequence

Structure and scRNA-seq (transcriptome) library preparation

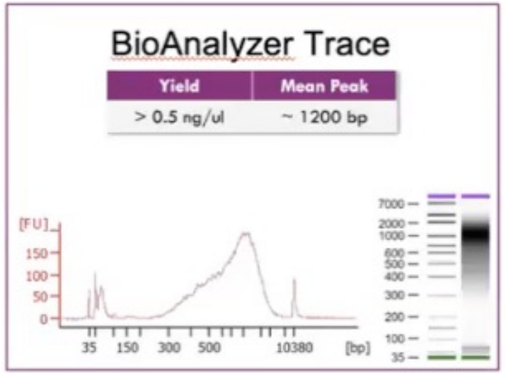
1. Isolate single cell and lyse



2. mRNA capture by PolyA tail

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

Amplified cDNA

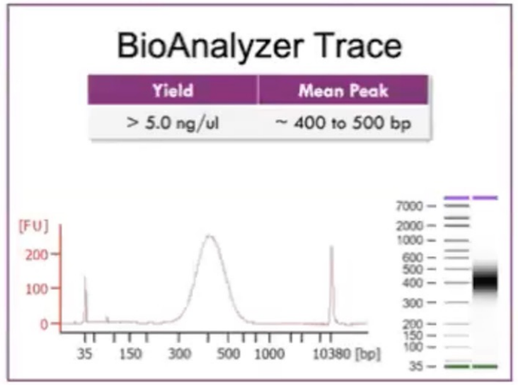


- Quantify
- Size
- contam

4. Fragment cDNA, add adaptors for Illumina sequencing

5. PCR amplify library, QC, pool and sequence

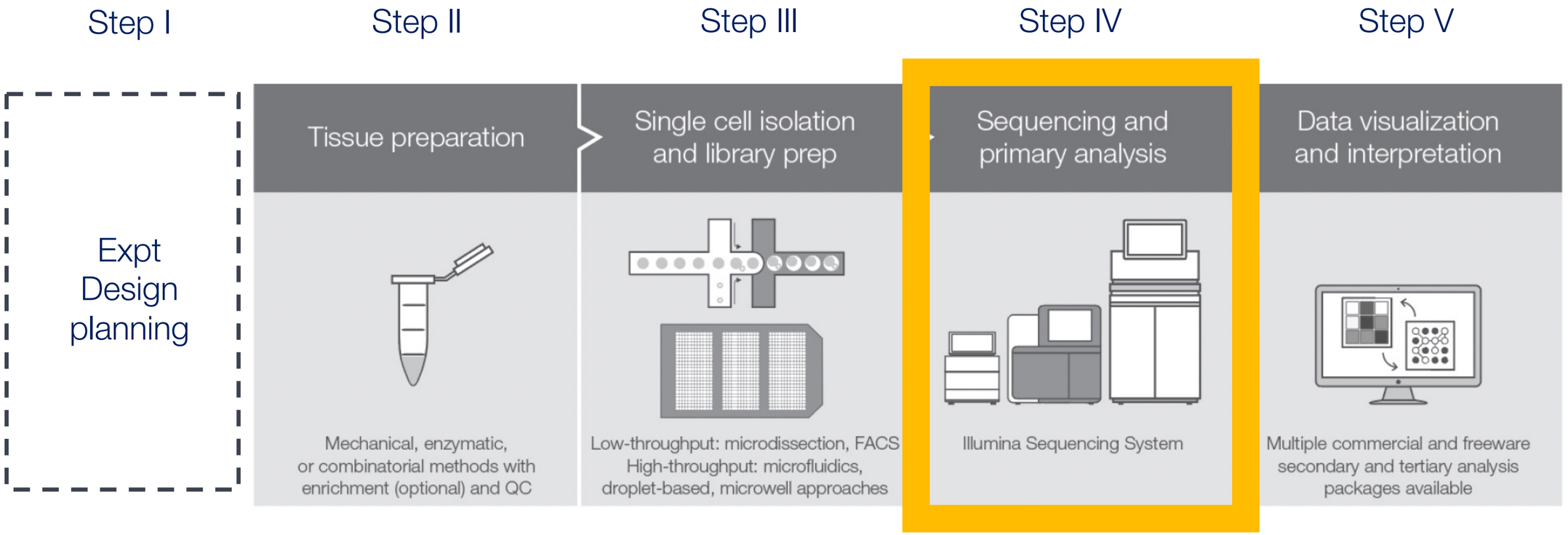
Index-PCR Library



- Quantify
- Size
- contam

Library structure platform and kit dependent

scRNA-seq workflow – STEP IV



Goal: sequence your libraries on the appropriate platform

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 or sequencing read depth

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: coverage or sequencing read depth

Experimental planning - Read depth or ‘coverage’

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: coverage or sequencing read depth

Experimental planning - Read depth or ‘coverage’

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

1 slice of bread does not need an entire jar of peanut butter!

Similarly, you don’t need sequencing-overkill on your sample

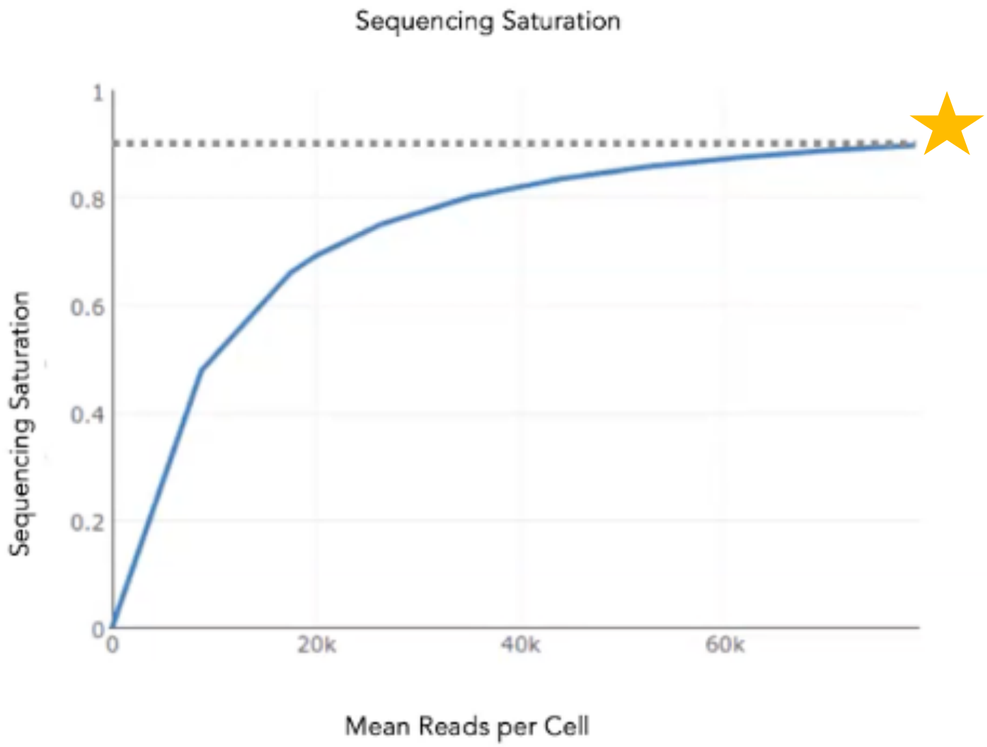
STEP IV: dialing in on sequencing saturation

How to know what's 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?

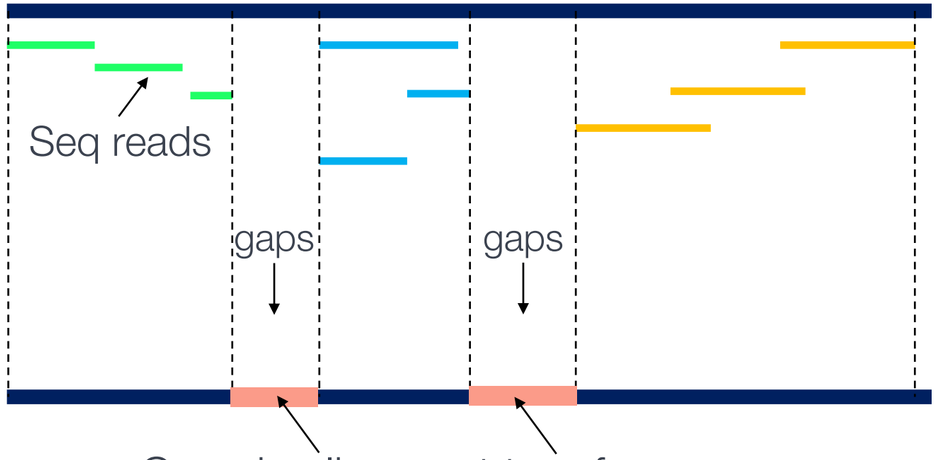
Rarer the cell type (or transcript), more sequencing needed = $$$



Choosing a sequencing platform: short or long-read?

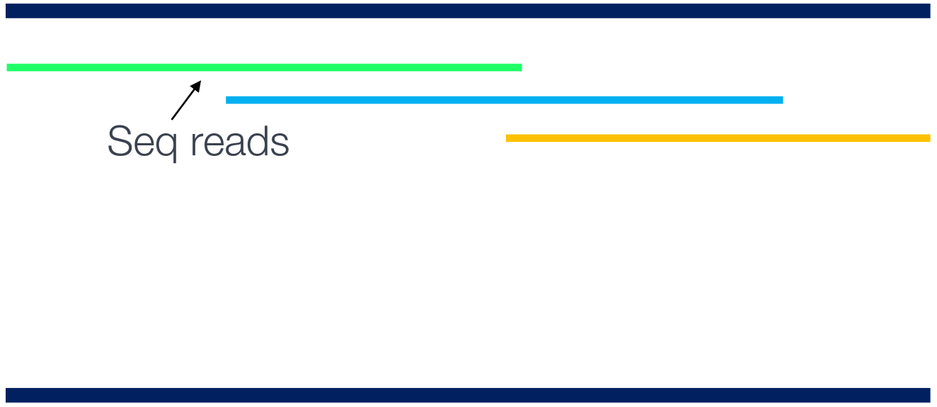
Short read (NGS/Illumina)

Reference genome



Long read (PacBio/ONT)

Reference genome

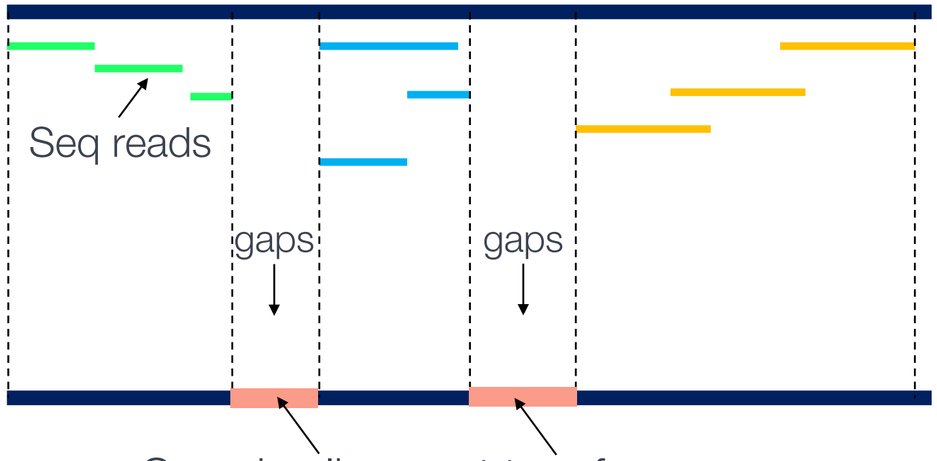


No/Less gaps in alignment

Choosing a sequencing platform: short or long-read?

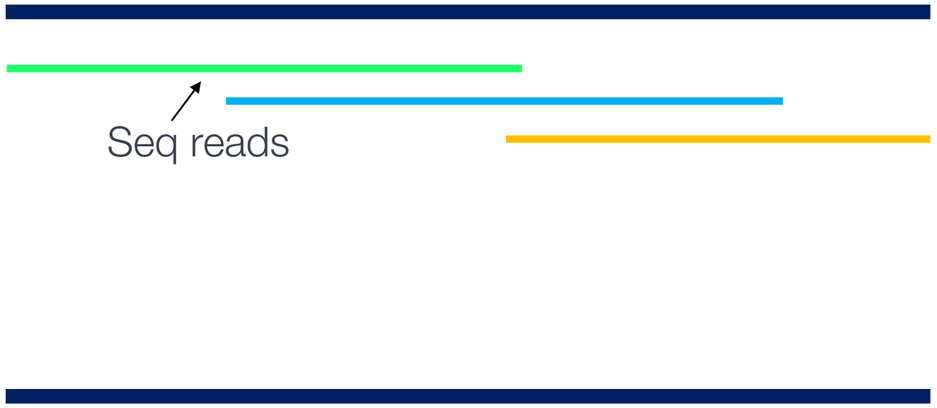
Short read (NGS/Illumina)

Reference genome



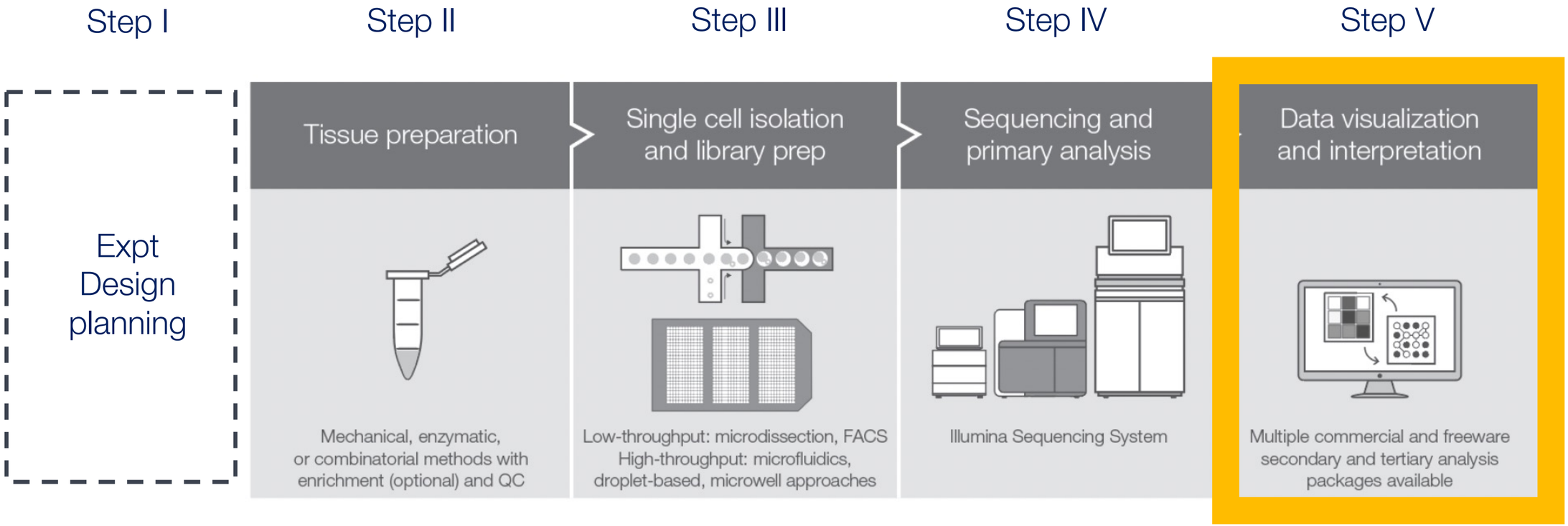
Long read (PacBio/ONT)

Reference genome



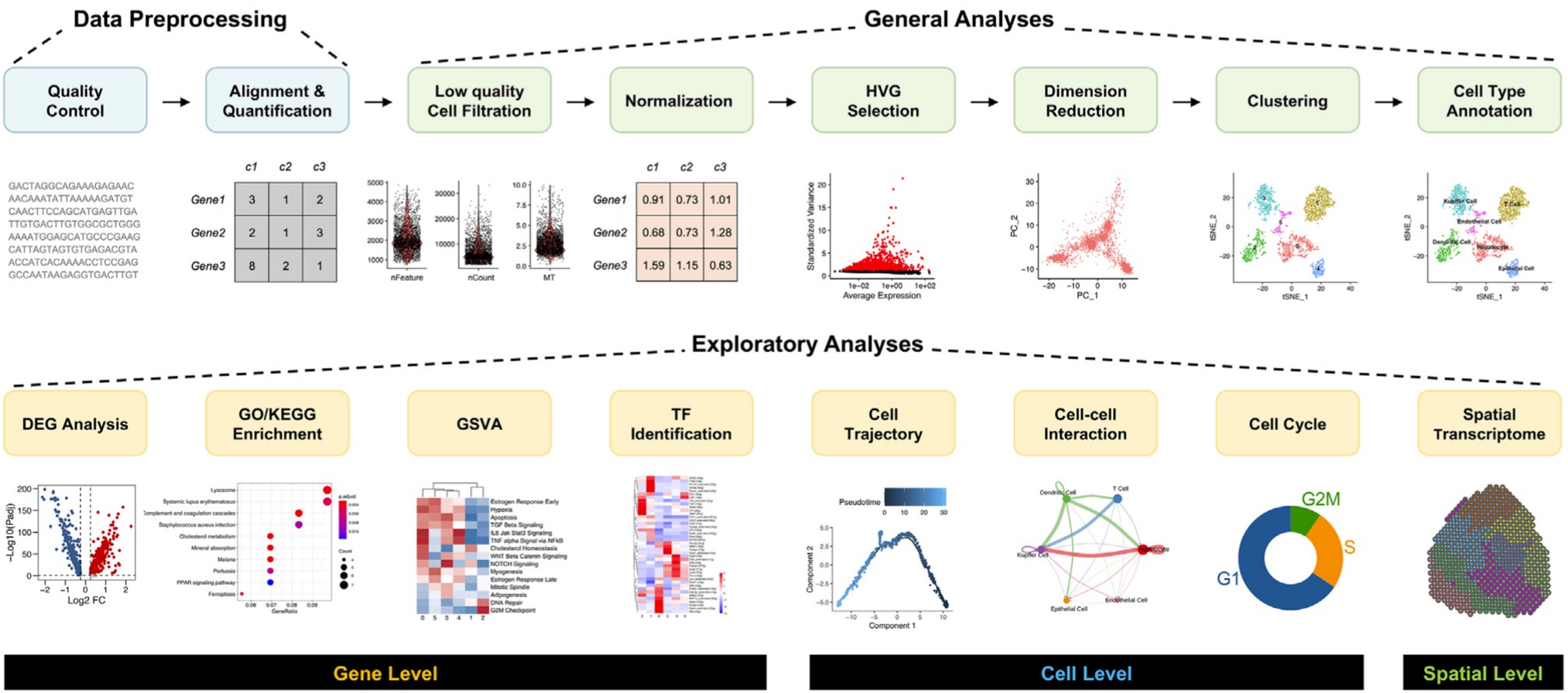
Decision of short vs long read, corresp. platform, lib prep made at Expt designing

scRNA-seq workflow – STEP V



Goal: analysis, interpretation and visualization of data for publication!

Key analysis steps in scRNAseq analysis



Kharchenko lab., 2021

Pipelines for data visualization and interpretation

A wealth of bioinformatic tools are available for scRNA-seq analysis: **pipelines give publication-worthy figures + help in-depth interpretation of the biology of your dataset**

- **Seurat** (Satija Lab): leading pipeline for the R language
- **Scanpy** (Theis Lab): leading pipeline for the Python language

Other pipelines that can incorporate statistical and machine learning models, but require more computational expertise -

- **Scvi-tools** (Yosef lab)
- **Monocle** (Trapnell Lab)

Key takeaways!

Summary –

- (I know it seems daunting) You can do this!
- Plan your experiments, put in effort into your sample prep!

  Mantra: Garbage in, Garbage out!
- 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