

Imaging meets single-cell -omics: A brief introduction to the new era of Spatial Transcriptomics.

George Gavriilidis MSc, PhD

Post-doctoral researcher in Systems Pharmacology and single-cell omics

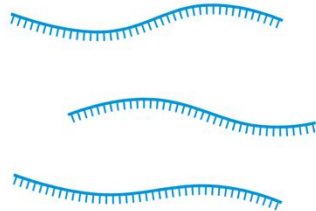
INAB, CERTH

ggeorav@certh.gr

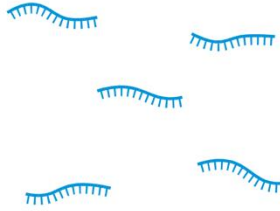
Sequencing 101

RNA Sequencing

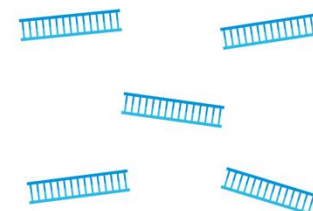
① Isolate RNA from samples



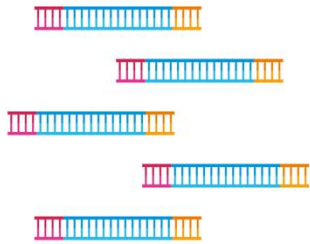
② Fragment RNA into short segments



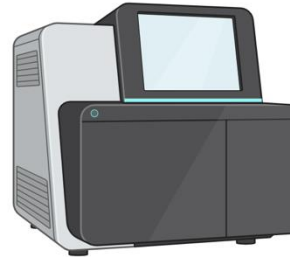
③ Convert RNA fragments into cDNA



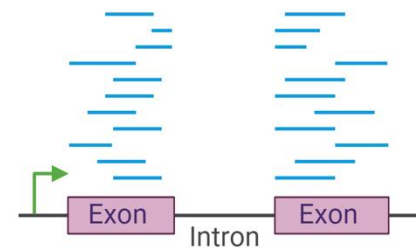
④ Ligate sequencing adapters and amplify



⑤ Perform NGS sequencing

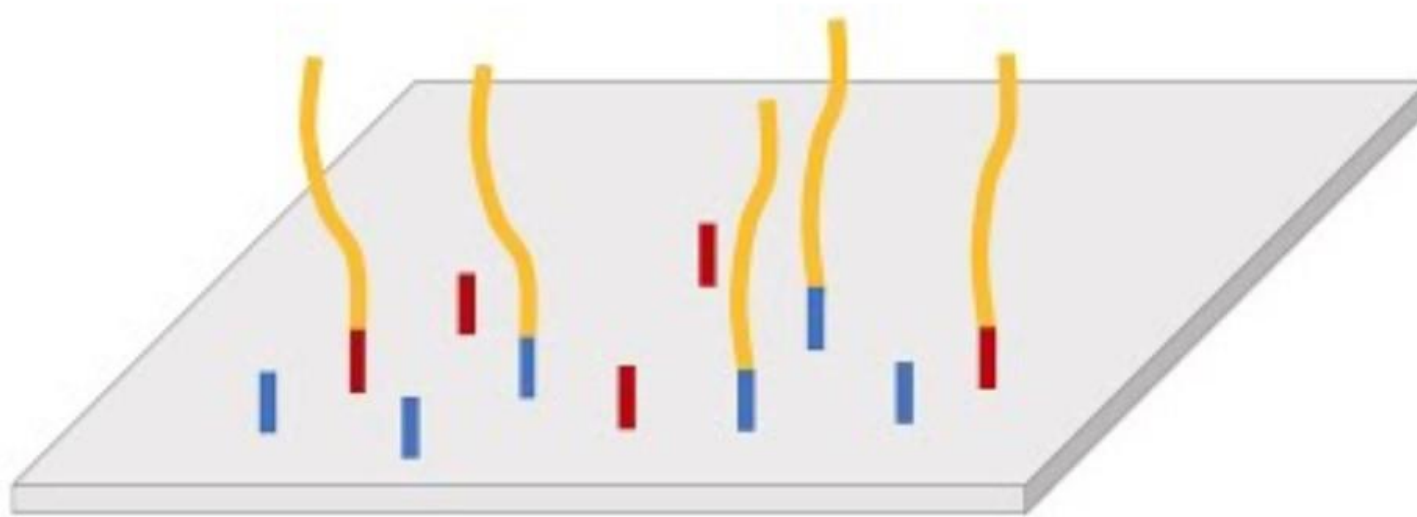


⑥ Map sequencing reads to the transcriptome/genome



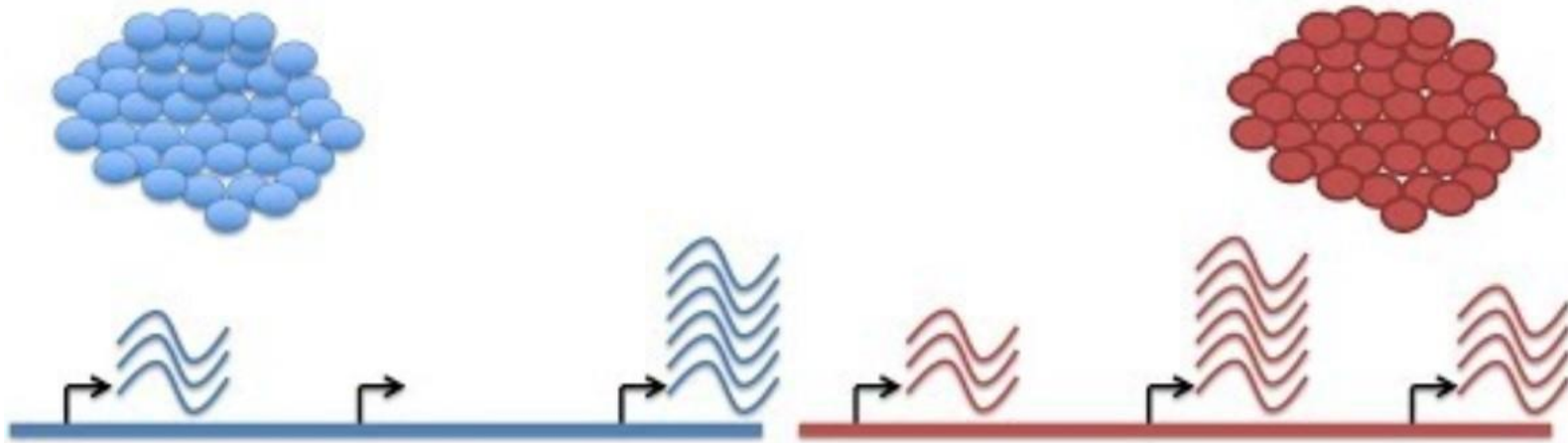
Sequencing 101

Next Gen Sequencing



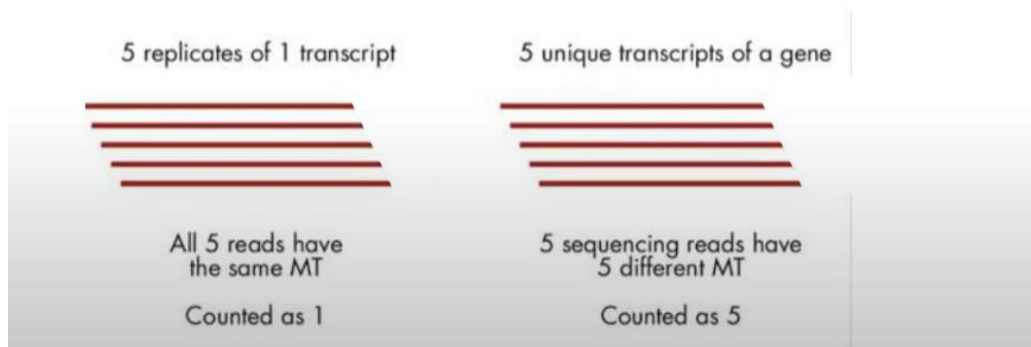
Sequencing 101

A Gentle Introduction To: RNA-Seq!!!!

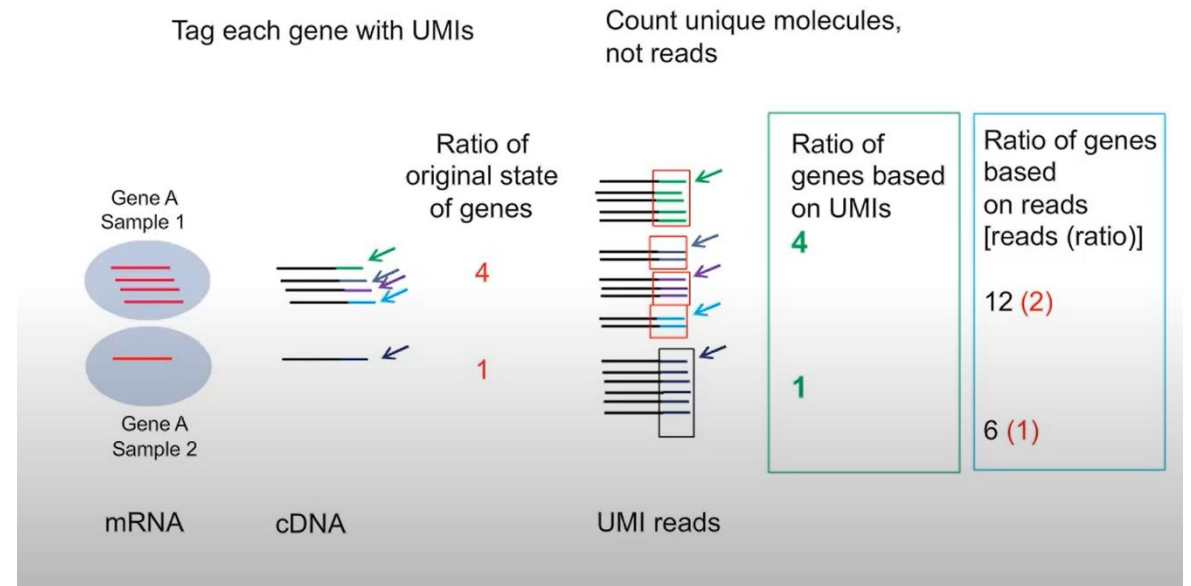


Unique Molecular Identifiers (UMIs)

How do we go from “Reads” to counting transcripts?



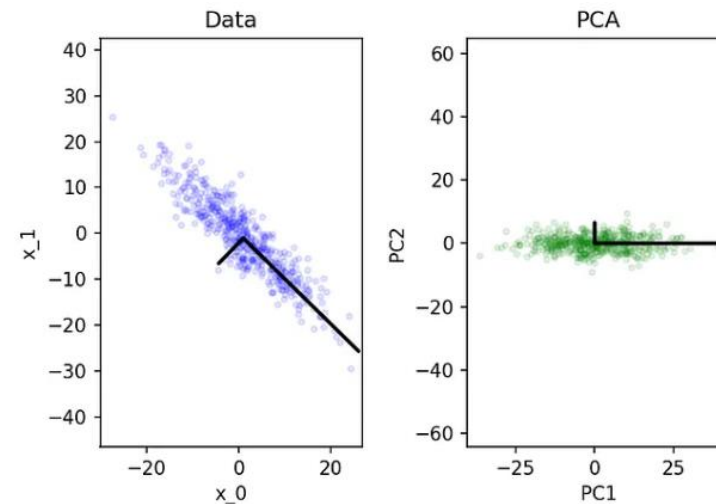
Molecular indexes allow the counting of original transcript levels instead of PCR duplicates, thereby enabling digital sequencing and resulting in unbiased and accurate gene expression profiles.



Statistical concepts for downstream analysis

- **Principal Component Analysis (PCA)**

- PCA transforms data into linearly, uncorrelated features or principal components (PCs). Importantly, these new features are meaningful — they indicate axes in the data where the variance is greatest
- The first principal component, or PC1, 'explains the most variance' in the original dataset — this also means that features that correlate with PC1 contribute to a large amount of variance in the data



Obligatory figure showing (left) random data in two dimensions and (right) PC1 and PC2 of the data. Black lines indicate the axes of along which the most variance exists. Note that these axes are orthogonal. In PCA space (right), the axes are vectors along PC1 and PC2.

Statistical concepts for downstream analysis

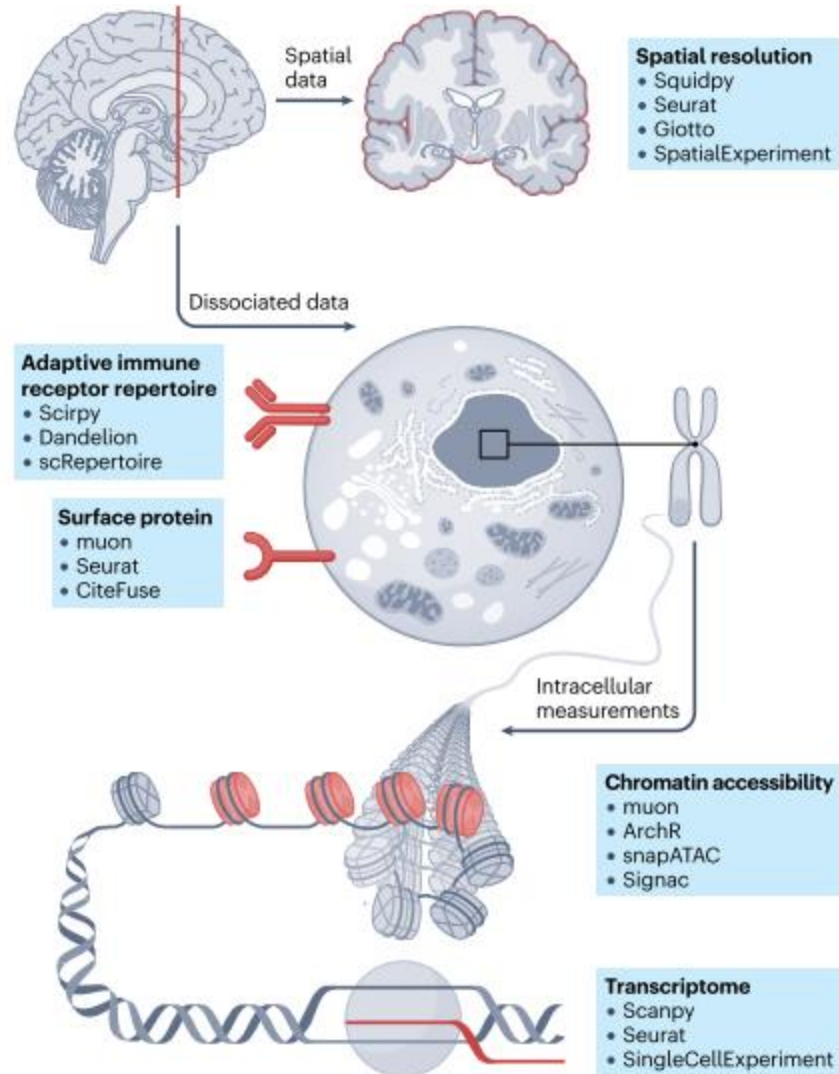
- **UMAP**

- Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE, but also for general non-linear dimension reduction. The algorithm is founded on three assumptions about the data:
 - The data is uniformly distributed on a Riemannian manifold;
 - The Riemannian metric is locally constant (or can be approximated as such);
 - The manifold is locally connected.

Hands-on practical #1- Understand PCA, UMAP, t-SNE interactively

- Before getting to analyze the scRNA-seq data, we need some basic understanding of dimensionality reduction
- PCA: <https://setosa.io/ev/principal-component-analysis/> , <https://www.youtube.com/watch?v=QSvpHo4p44M>
- UMAP/t-SNE: <https://pair-code.github.io/understanding-umap/#:~:text=The%20biggest%20difference%20between%20the,meaningful%20than%20in%20t%2DSNE>, https://appyters.maayanlab.cloud/dimensionality_reduction_visualization/

Single-cell analysis across modalities



The cellular state is characterized:

- RNA transcription,
- chromatin accessibility,
- surface proteins T cell receptors (TCRs), B cell receptors(BCRs)
- spatial location**

Core bioinformatic pipelines are cardinally:

- Seurat (R),
- Scanpy (Python),
- Signac (R)
- Squidpy (Python), Giotto (R)**

These frameworks are complemented with a myriad of additional tools for specific subsequent analysis tasks!

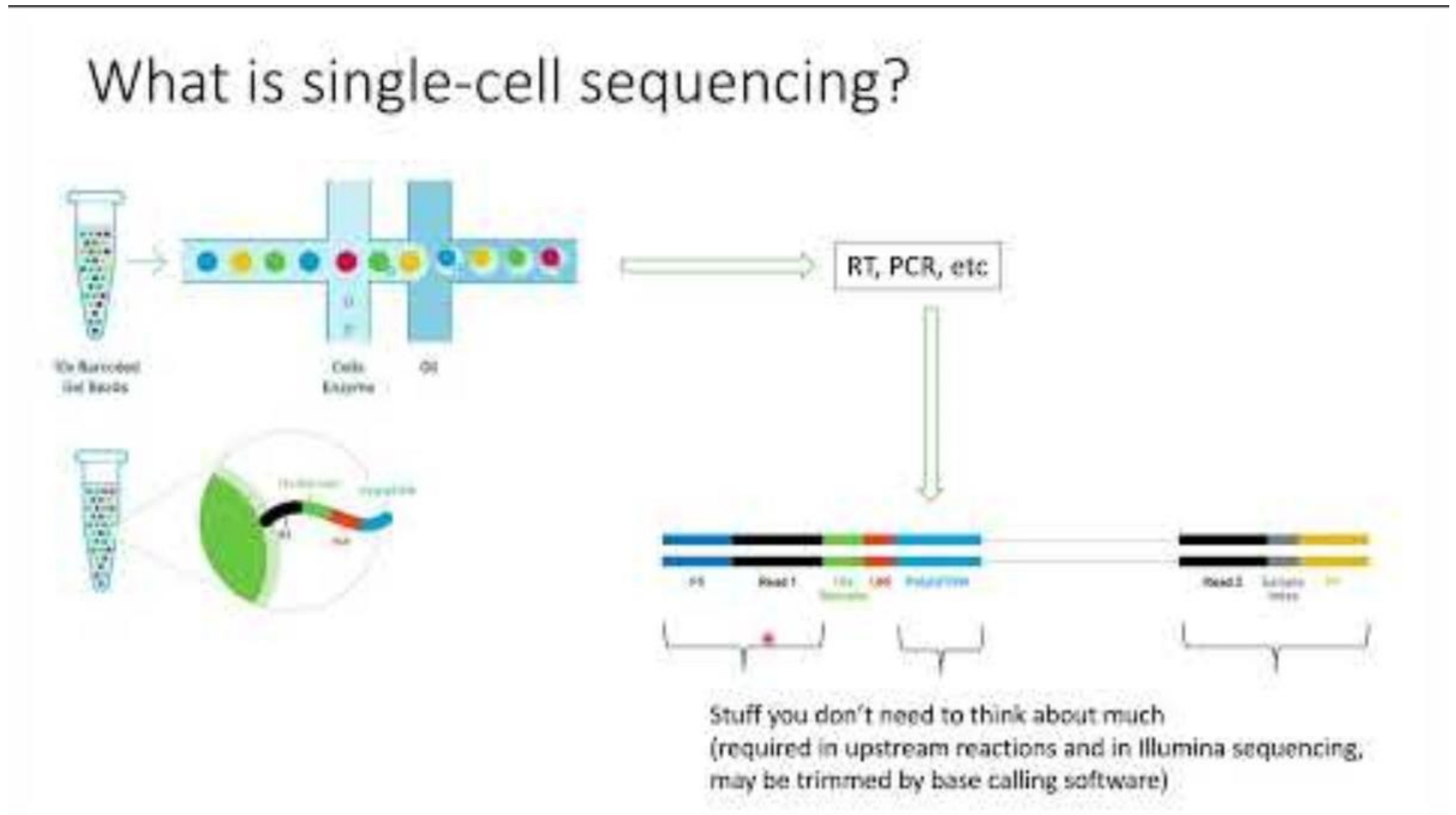
Single-cell sequencing (2013)



Single-cell multi-modal omics (2019)



What is single-cell sequencing?



What is scRNA-seq analysis?

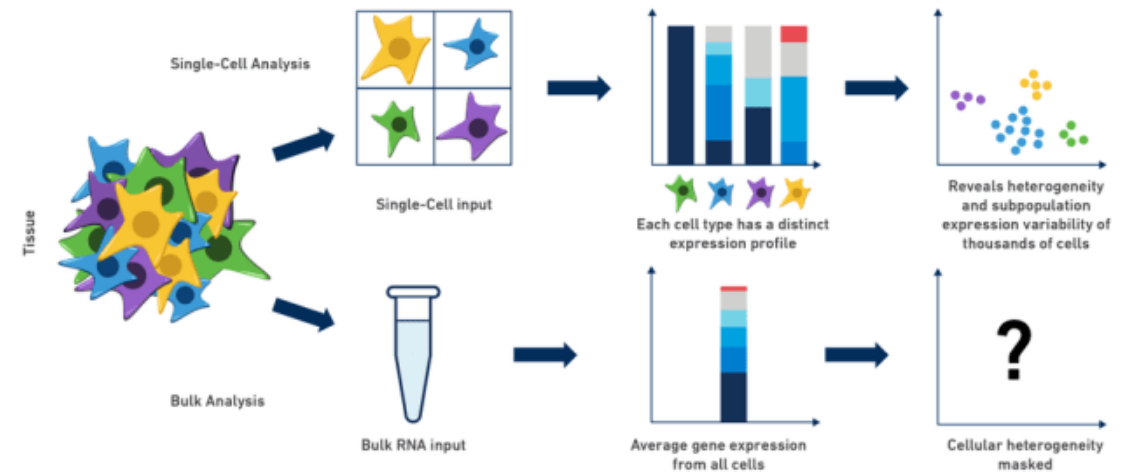
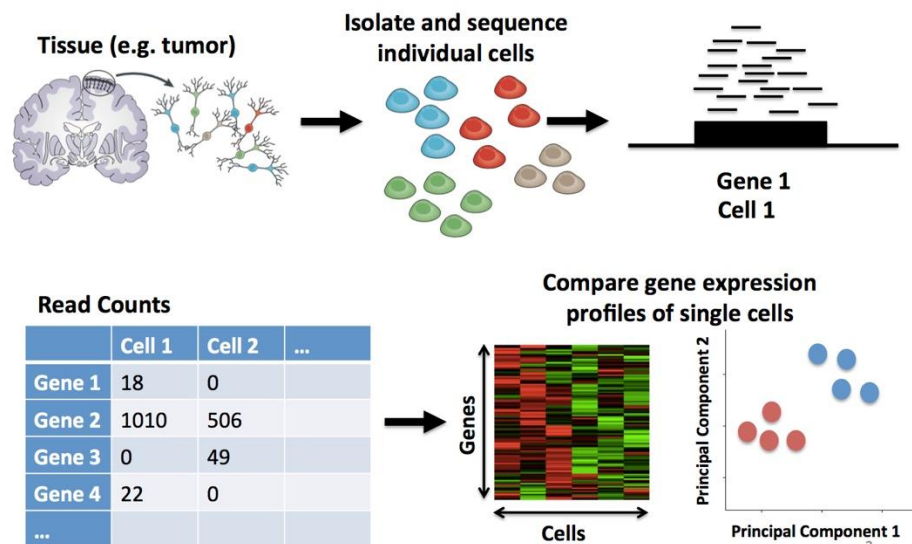
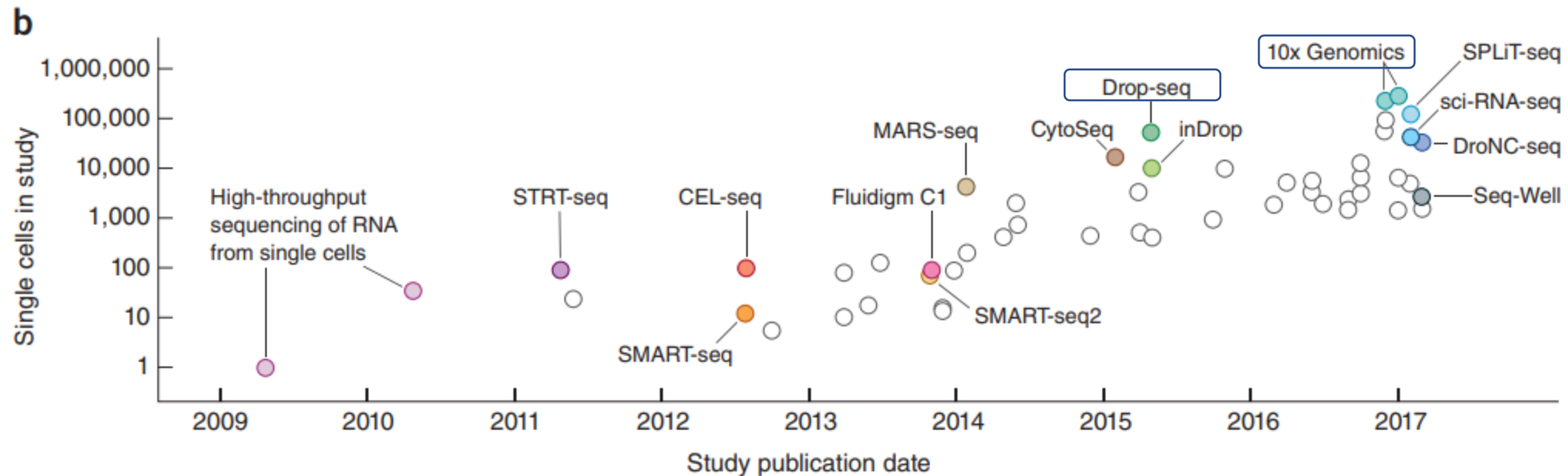
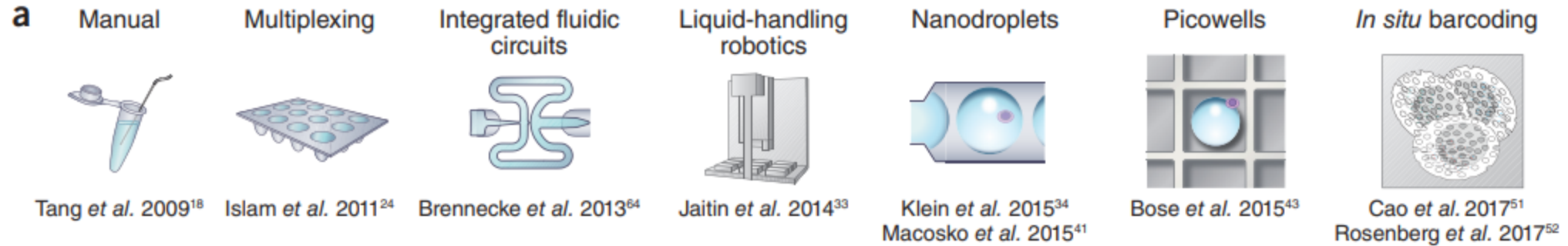
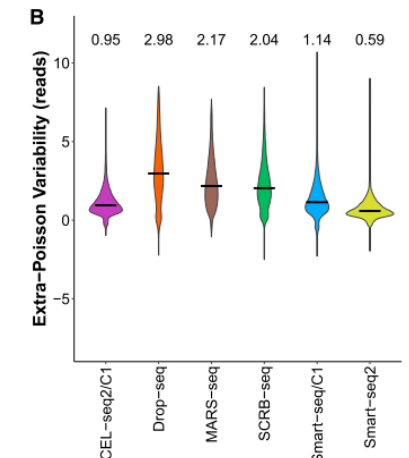
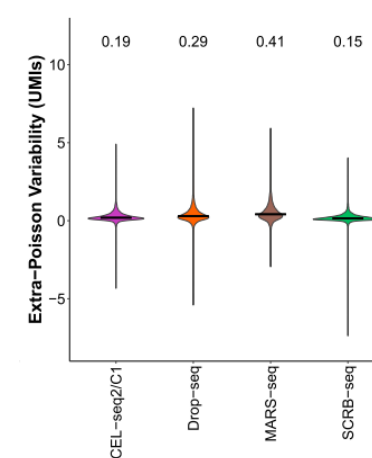
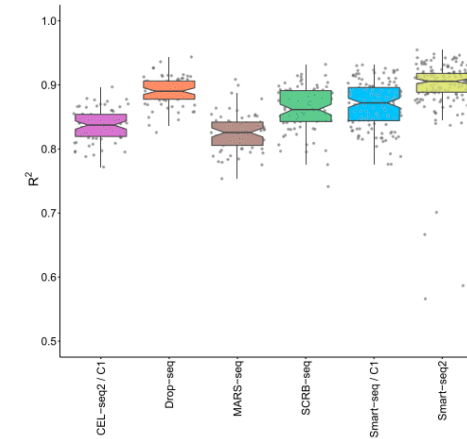
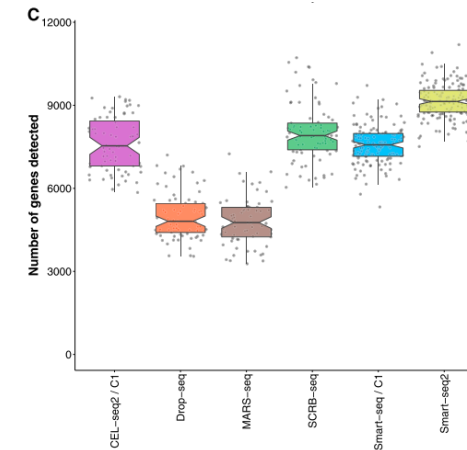
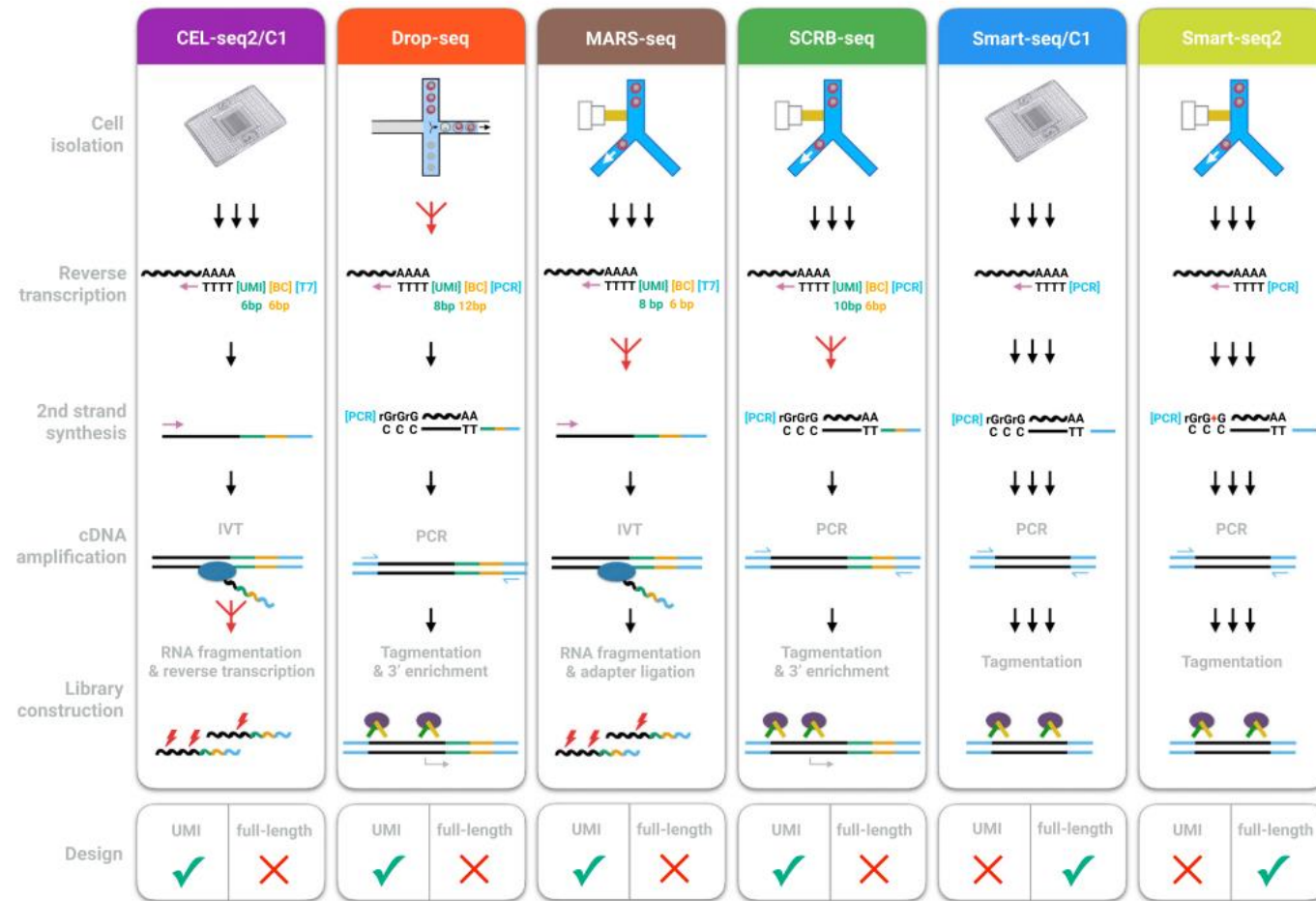


Figure 1. Single-cell RNA-seq reveals cellular heterogeneity that is masked by bulk RNA-seq methods.

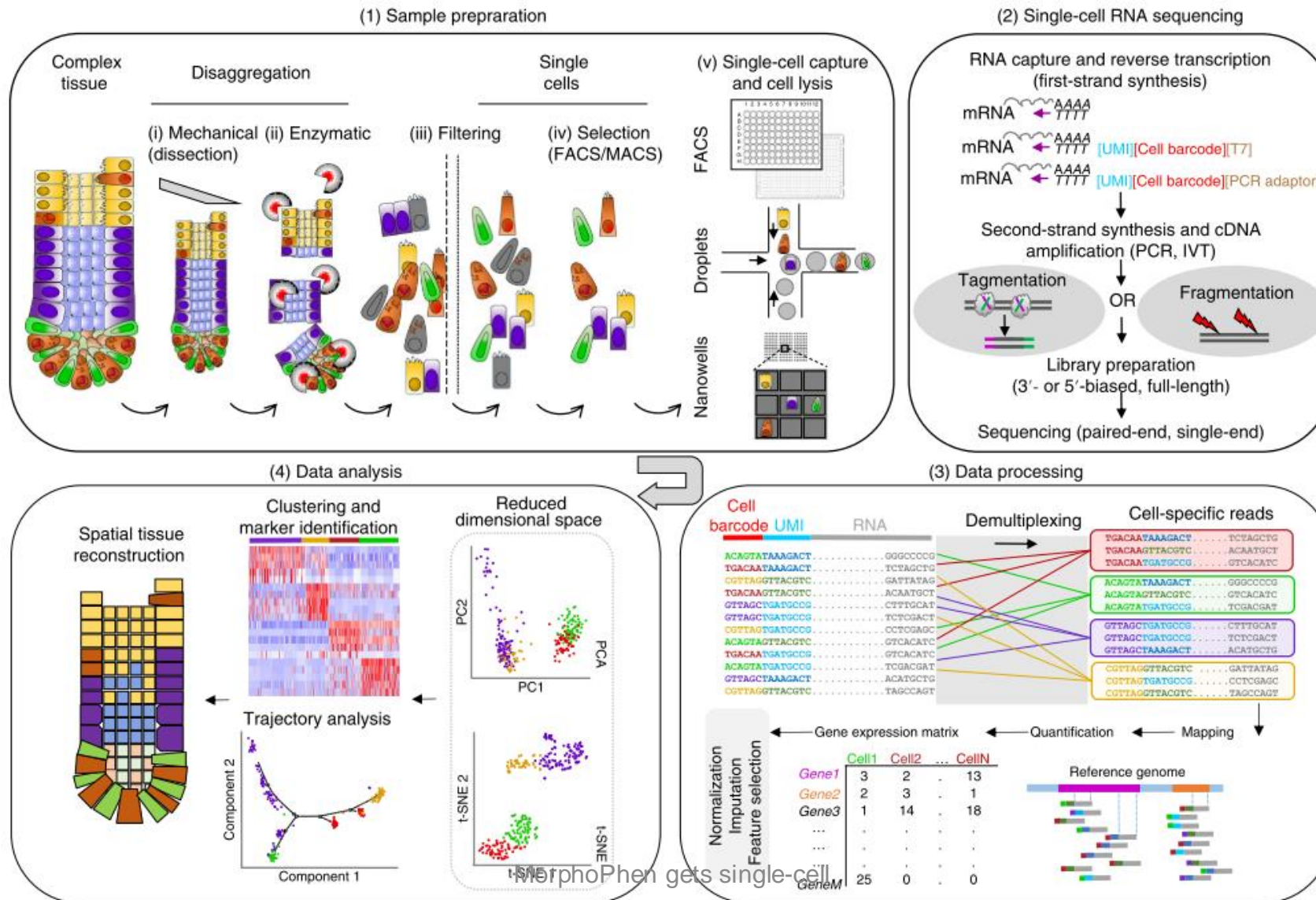
Diverse scRNA-seq technologies



Distinct features and comparison of scRNA-seq technologies

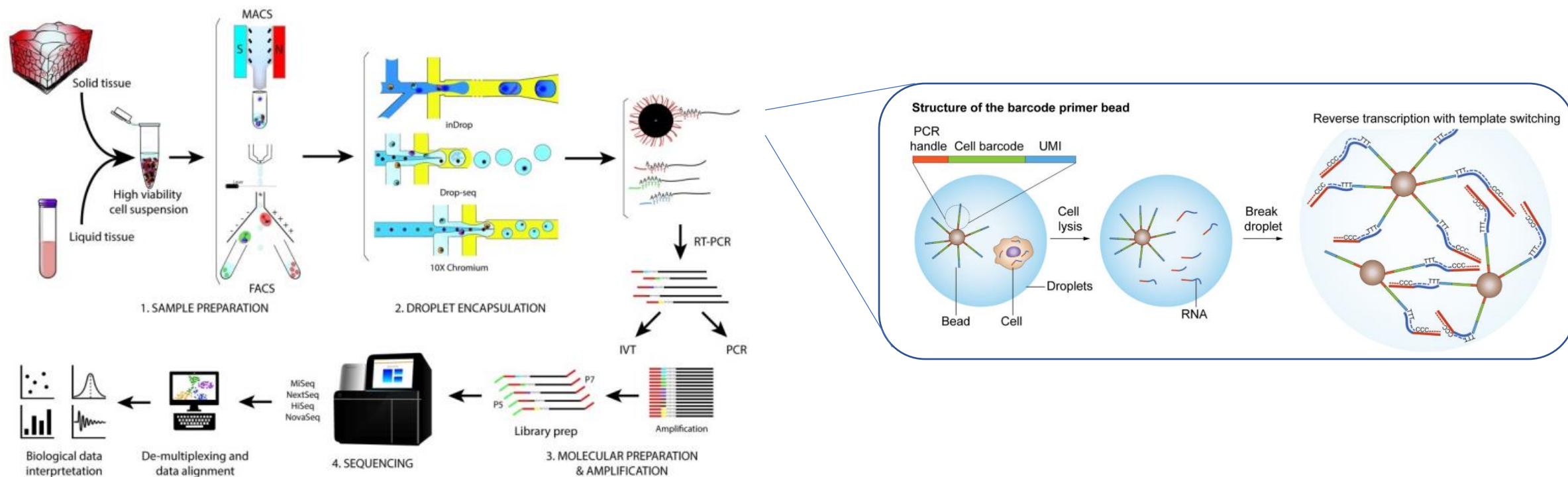


Overview of the scRNA-seq process



Lafzi et al. 2018 REVIEW ARTICLE
<https://doi.org/10.1038/s41596-018-0073-y>
 Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies

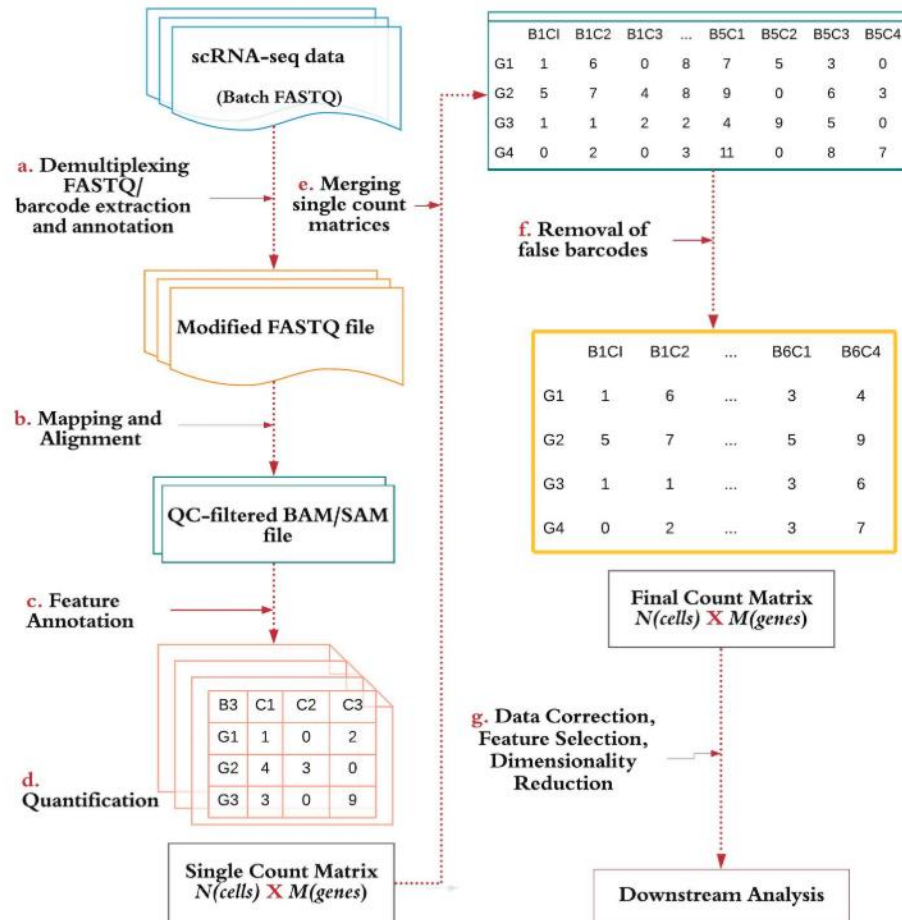
Droplet-based sc-RNAseq



Salomon et al. 2019. Lab Chip. 2019 May 14;19(10):1706-1727.
Hwang et al. Experimental & Molecular Medicine (2018) 50:96

Raw data to gene-cell expression matrices

Cell barcode handling
UMI deduplication
Transcriptome/genome mapping
Counting



Count matrices and Quality Control

Count matrices of cells by genes are obtained from raw data processing pipelines, after correction for:

(a) cell-free ambient RNA – (b) doublets – (c) dying cells

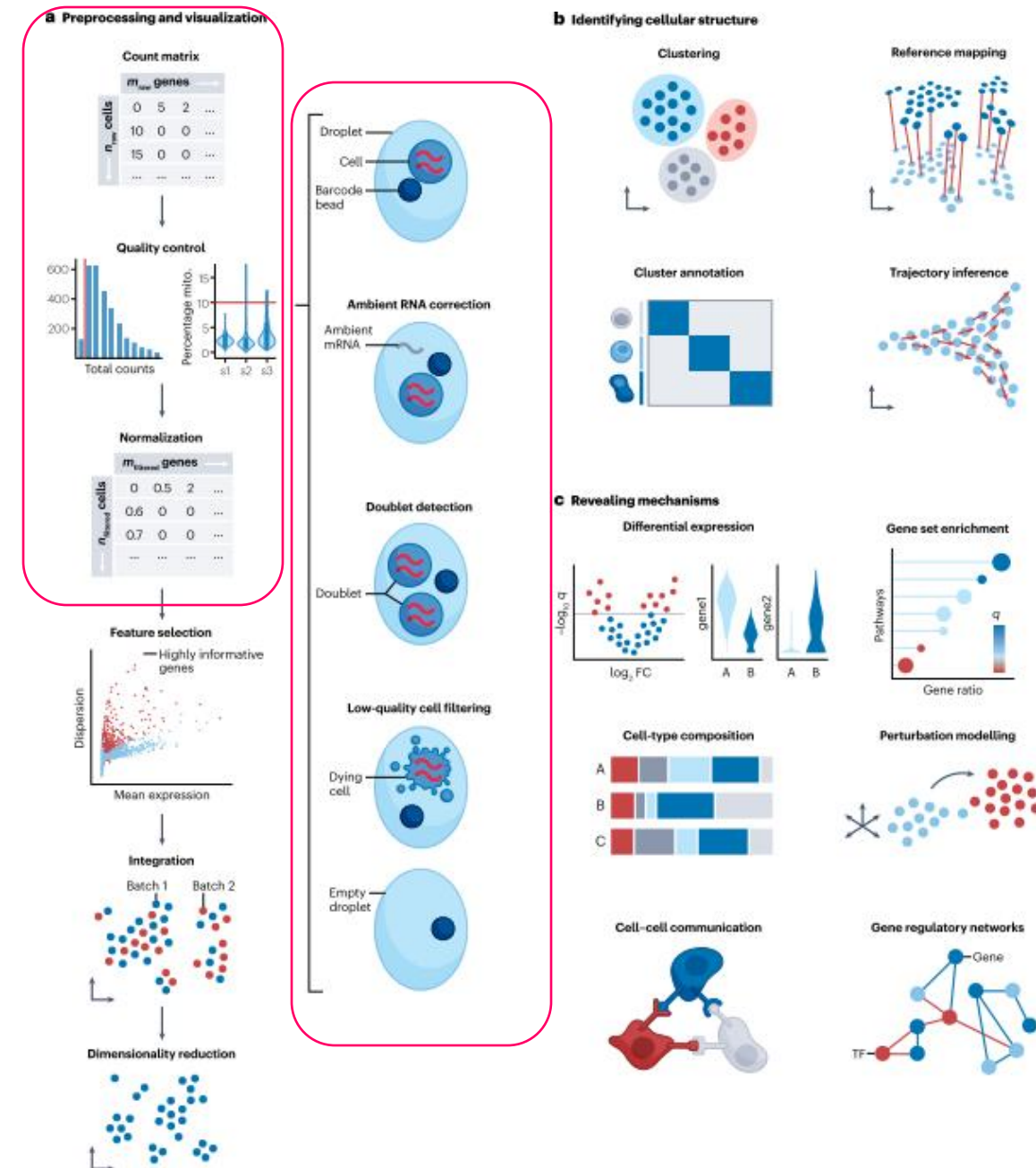
!!!Quality Control metrics!!!:

- the number of counts per barcode (library size)
- the number of genes per barcode
- the fraction of counts from mitochondrial genes per barcode (percent. mito)

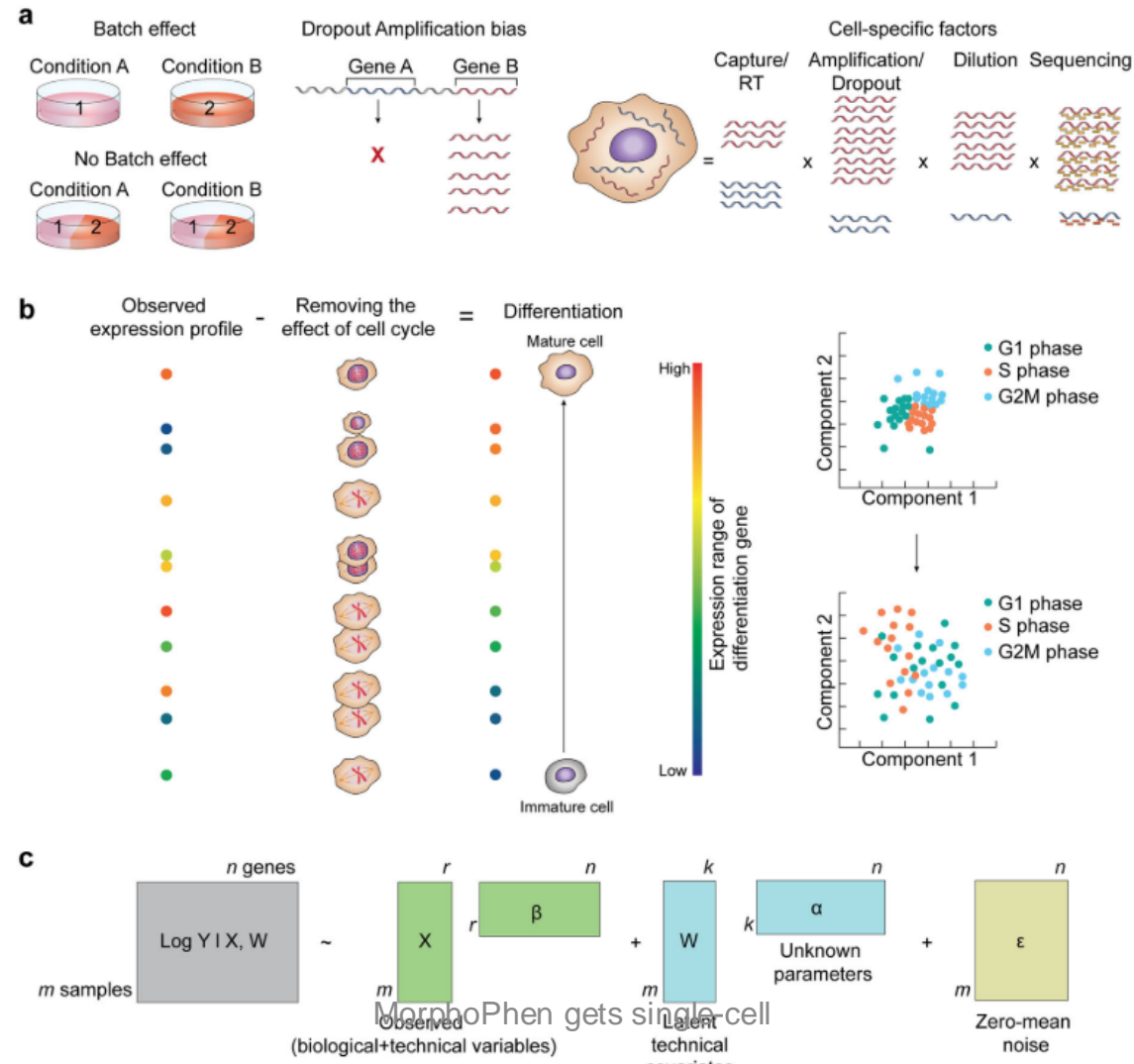
But what is a count???

All counts represent successful capture, reverse transcription and sequencing of an mRNA molecule.

Count depths for identical cells can differ- when comparing gene expression between cells, **differences may originate solely from sampling effects** => Normalization is needed to obtain correct relative gene abundances between cells!

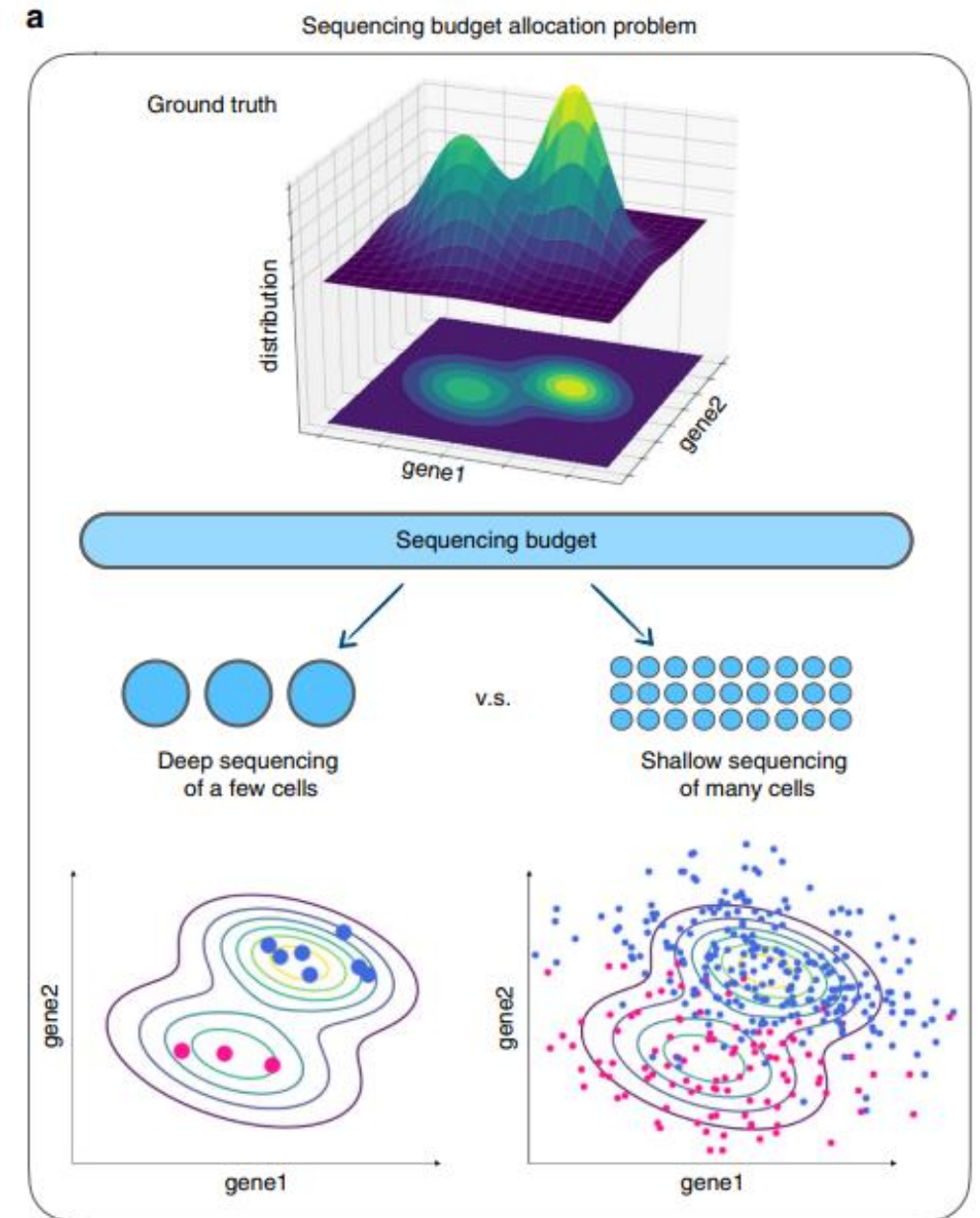


Addressing confounding factors in scRNA-seq



Out-of-our-depth?

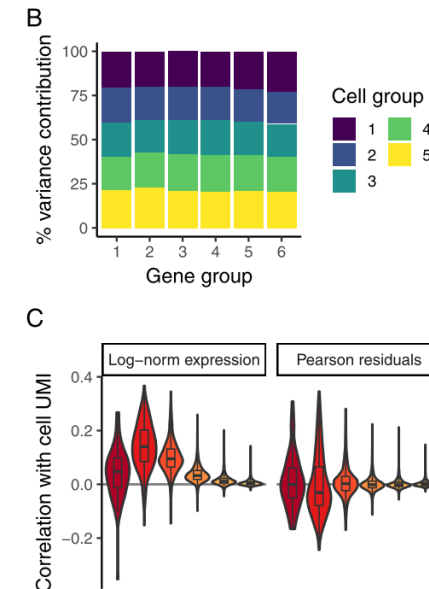
- Which technology has more genes analyzed single-cell or bulk?
- The sequence budget allocation problem...
- How do you make a good trade-off between depth of sequencing and ground truth?



Removing sequencing depth as a technical confounder

- In general, the normalized expression level of a gene should not be correlated with the total sequencing depth of a cell. Downstream analytical tasks (dimensional reduction, differential expression) should also not be influenced by variation in sequencing depth.
- The variance of a normalized gene (across cells) should primarily reflect biological heterogeneity, independent of gene abundance or sequencing depth.
- For example, genes with high variance after normalization should be differentially expressed across cell types, while housekeeping genes should exhibit low variance. Additionally, the variance of a gene should be similar when considering either deeply sequenced cells, or shallowly sequenced cells.

SCTransform: A statistical model which relates cellular sequencing depth to gene molecule counts.



Highly Variably Genes, batch integration, cellular structure

~30,000 genes for humans.

HVG! They drive downstream analysis— **less noise!**

Batch integration and correction! We usually integrate multiple scRNA-seq studies (e.g., data from each patient), **but the technical issue of integrating multiple studies (“batch”) can mask biological variability!**

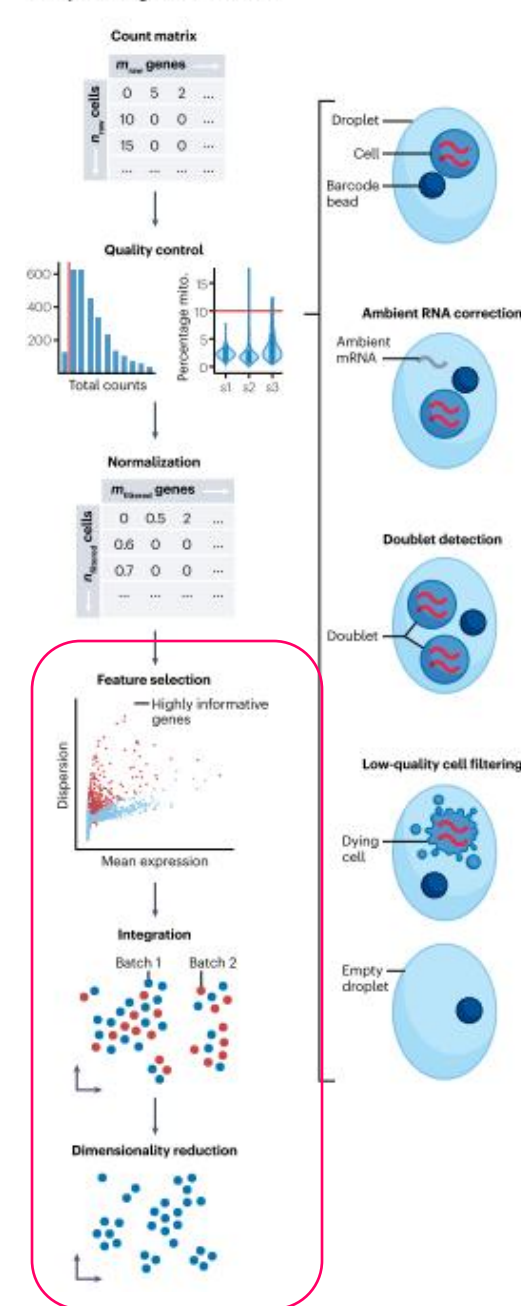
Dimensionality reduction! **Less computational burden**, easier to navigate and **interpret through visualization...**

Clustering! The low-dimensional space with cell embeddings can be clustered into **distinct cell types but annotation needed!**

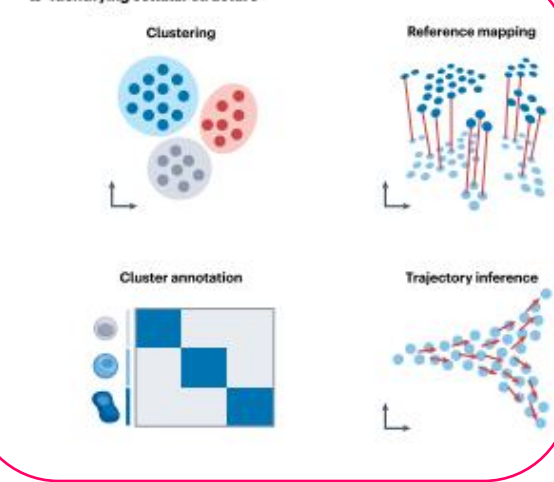
Cell annotation! **Manual, semi-manual, or automated through ML/DL**, usually with biological priors

Cell states! Differentiation trajectories: **Stem cell->Precursor->fully developed cell**

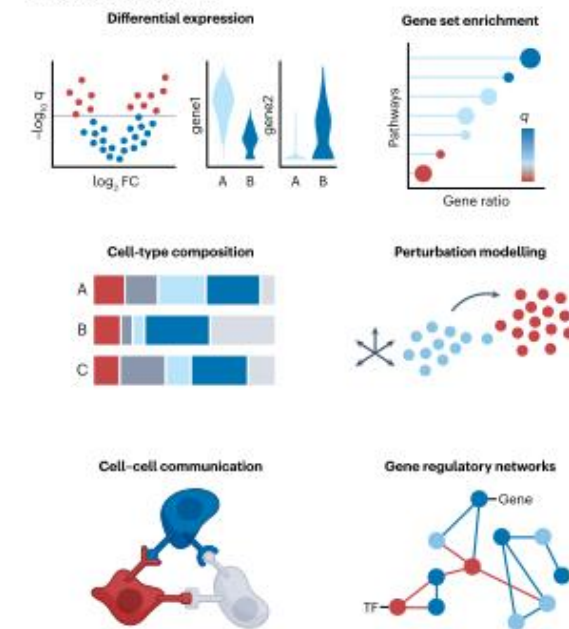
a Preprocessing and visualization



b Identifying cellular structure



c Revealing mechanisms



Revealing mechanisms

Upregulated or downregulated genes! – how does this compare with bulk omics differential expression???

Effects on pathways (gene set enrichment)! – how does this compare with bulk omics pathway enrichment?

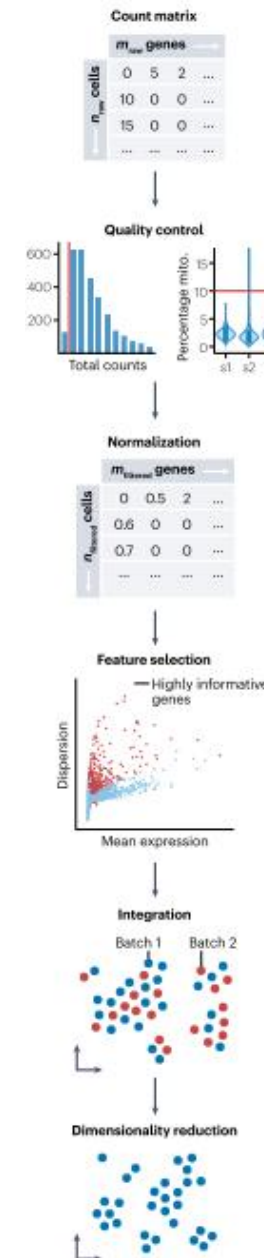
Changes in cell composition!

Perturbation modeling! Enabling the assessment of the effect of induced perturbations and the prediction of unmeasured perturbations (that's my thing!! 😊).

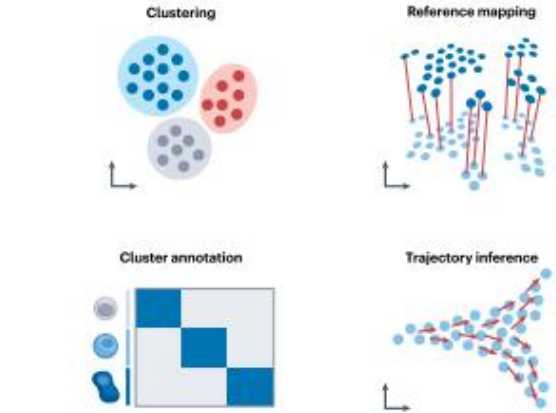
Cell-Cell communication! Networks (!) of ligand-receptors

Gene Regulatory Networks! Remember some methodologies from bulk?

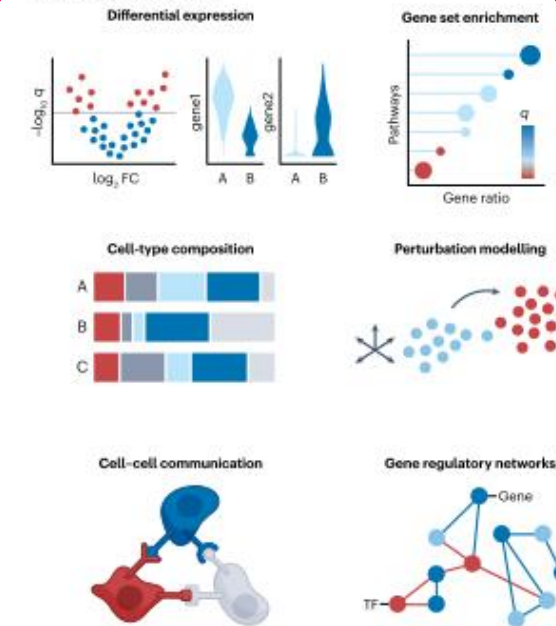
a Preprocessing and visualization



b Identifying cellular structure



c Revealing mechanisms



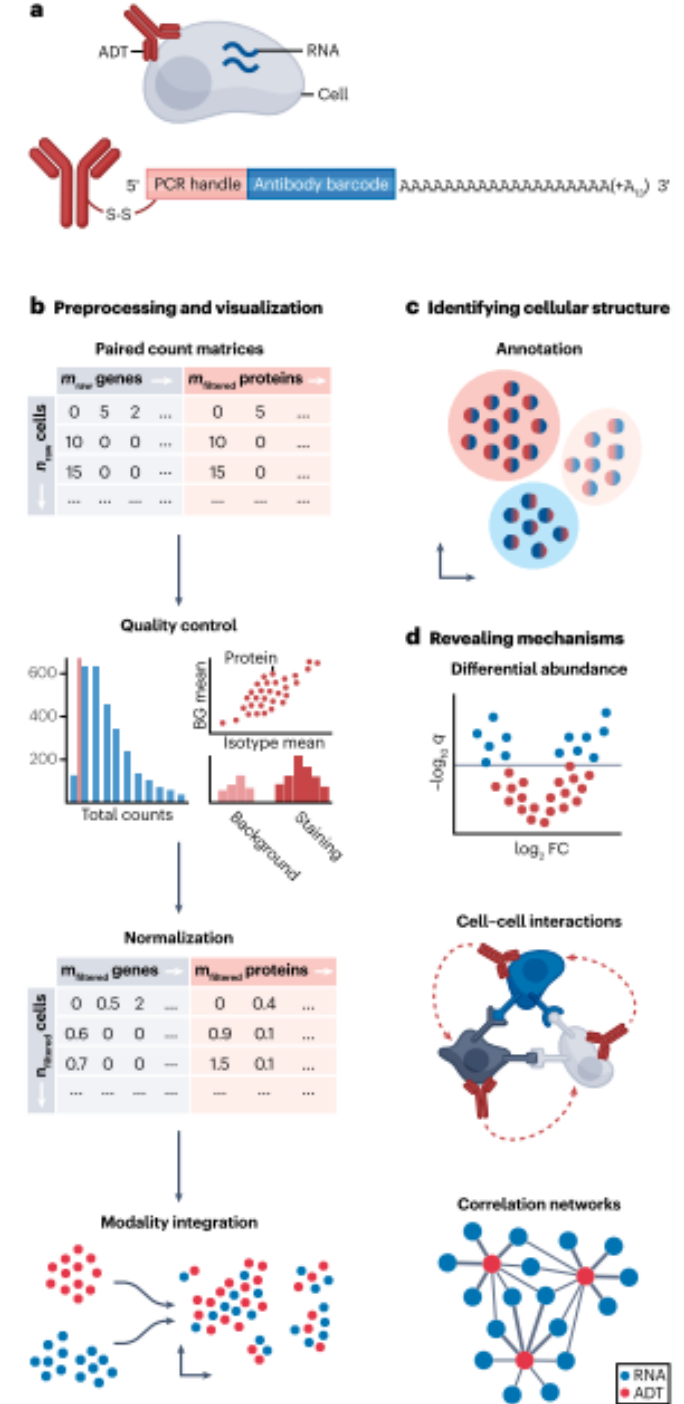
CITE-seq: Some proteins and many genes!

CITE-Seq (cellular indexing of transcriptomes and epitopes) is a sequencing-based method that simultaneously quantifies cell surface protein and transcriptomic data within a single cell readout

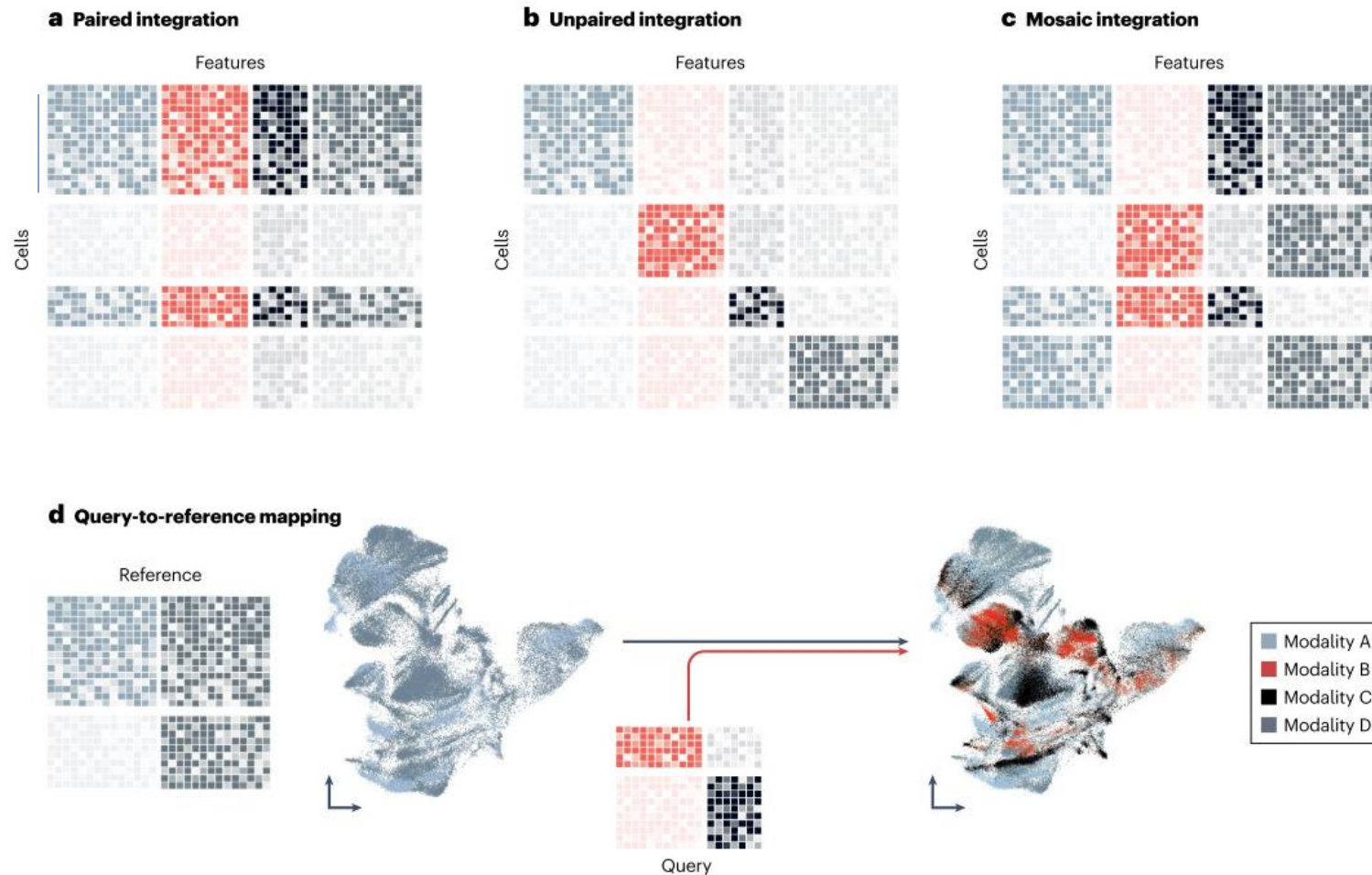
Antibody-derived tags (ADTs) are antibody clones with unique barcodes attached to poly(A) sequences and a PCR handle that is specifically amplified in subsequent library processing steps.

The antibody binds to surface proteins, and the sequenced ADT counts represent the expression level of those proteins.

Analyzing proteins and transcripts at the same time offers key advantages! **Why?????**



Data integration across omic modalities



Paired Integration:

RNA, protein, ATAC... all from the same dataset – Mostly linear approaches (MOFA+, WNN)

Unpaired Integration:

RNA, protein, ATAC... NOT from the same dataset or experiment – Mostly non-linear approaches (Machine Learning, Deep Learning)

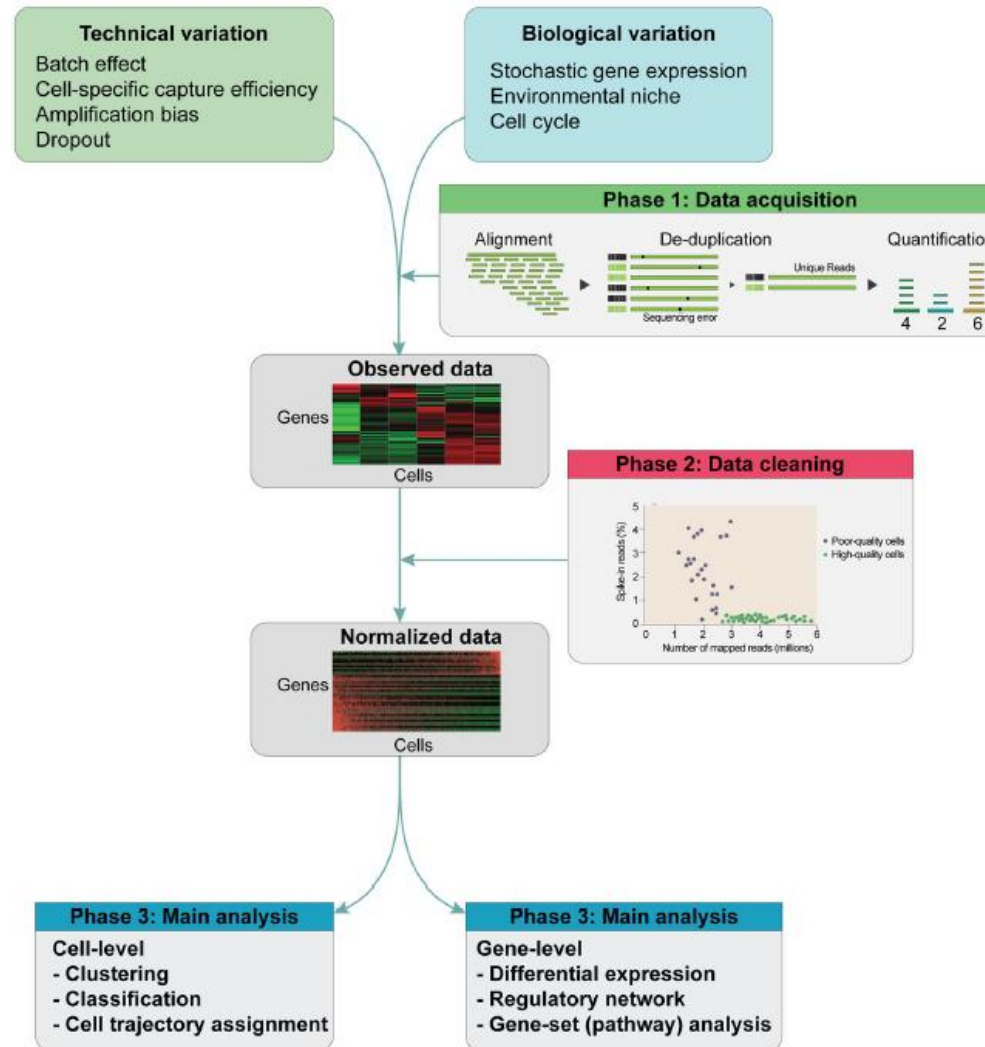
Mosaic integration:

RNA, protein, ATAC... Profiling individual modalities on different populations of cells from the same biological sample is more common, leading to completely missing data matrices - Non-linear approaches (Machine Learning, Deep Learning)

Query-to-reference mapping:

Prior knowledge from many integrated studies is leveraged!

scRNA-seq preprocessing overview



Hwang et al. Experimental & Molecular Medicine (2018) 50:96

Recap no1!

- Single-cell technologies provide a glimpse into cell biology at an unprecedented resolution, revealing biological mechanisms and cell states that would be impossible to discover with traditional bulk –omic techniques
- Single-cell sequencing is the first version of these technologies – more have come into the fold (single-cell proteomics, scATAC-seq etc..) as well spatial omics
- Bioinformatic analysis of single-cell technologies requires sound knowledge of biostatistics, Machine Learning and Deep Learning
- QC, Highly Variable Genes, Scaling, Normalization, Dimensionality Reduction, Clustering, Visualization are the cornerstones of single-cell analysis
- Many downstream analysis can reveal detailed information about cell fate, pathways, cell-cell interaction, perturbations and many more...
- Single-cell integration of multi-omics is a very active field of research and various strategies are being explored – not having multi-omic measurements from the same exact cells is a huge challenge!!

Hands-on practical #2- Run Seurat tutorial!

- We are running the basic tutorial for PBMCs from Seurat
 - What will happen if I change the PCs in the analysis? How many PCs?
 - What will happen if I change the resolution parameter during clustering?
- We are running cell-cycle regression tutorial to understand how this biological regulation can be removed from our data for downstream analysis
 - Should we always regress cell-cycle from downstream analysis?
- We are running Mixscape to see how a cell-line that has experienced perturbations through CRISPR is clustered depending on the target of the perturbation
 - What is the value of this type of analysis on single-cell?

Spatial transcriptomics

- Cellular organization bridges the gap between tissue biology and pathology, which enables the discovery of new cellular functionalities and creates new computational challenges for which distinct analysis methods are required
- Spatial omics resolves features and cellular identities by adding two additional modalities to single-cell genomics: histological imaging and spatial profiling measurements
- Adding information extracted from the imaging data can enhance, for example, cell identification or the resolution of the molecular features can help to identify spatial patterns of variation
- **Array-based methods** and **Image-based methods**

Spatial Transcriptomics

(A) **Array-based spatial transcriptomics**: quantify gene expression *in predefined barcoded (BC) regions* ($10\ \mu\text{m}$ and $200\ \mu\text{m}$) BC = multiple cells \rightarrow count matrices and spatial coordinates where each observation is a BC region.

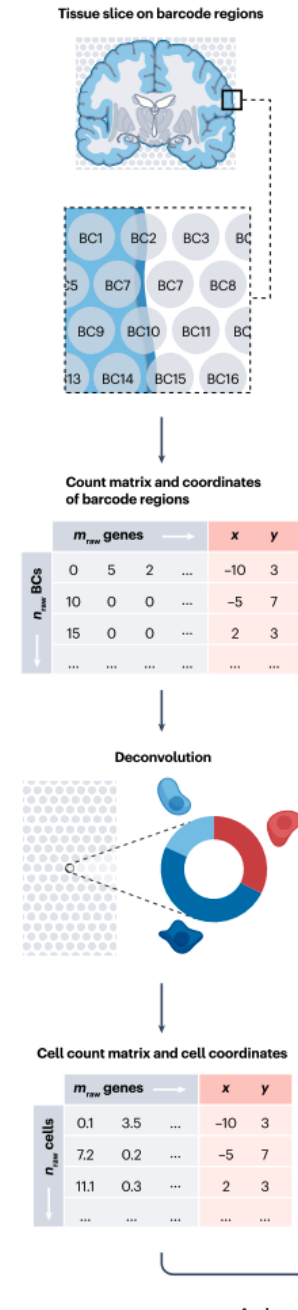
Cell-type deconvolution: From BC \rightarrow count matrices and spatial coordinates where each observation is a single cell

(B) **Image-based spatial transcriptomics**: capture individual locations of transcripts in multiple sequential hybridization rounds

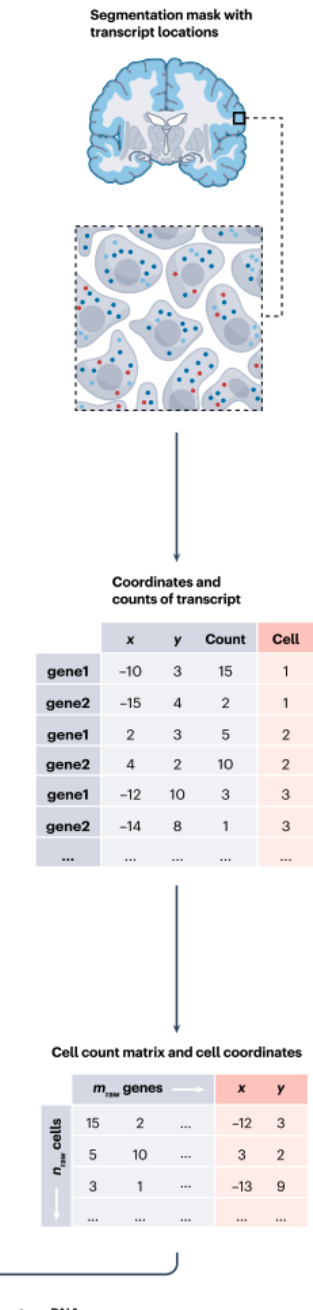
Transcript locations can be aggregated to obtain count matrices and spatial coordinates at single-cell level

Cell segmentation: understanding where the cell perimeter is, we can aggregate its content and finally arrive at gene x cell matrix with coordinates

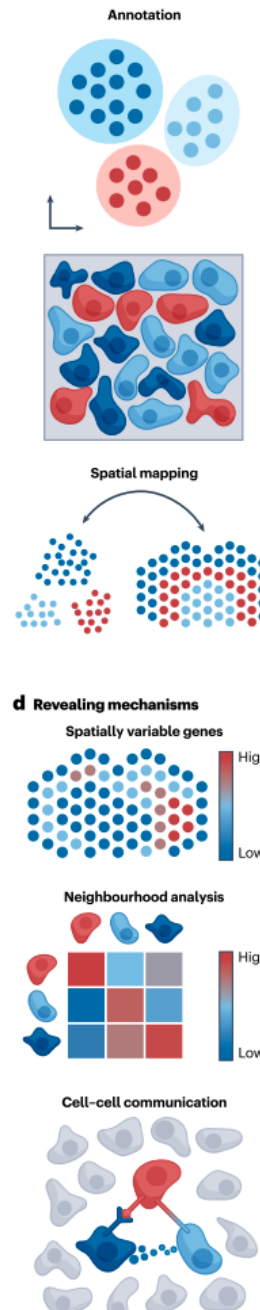
a Array-based spatial transcriptomics



b Image-based spatial transcriptomics



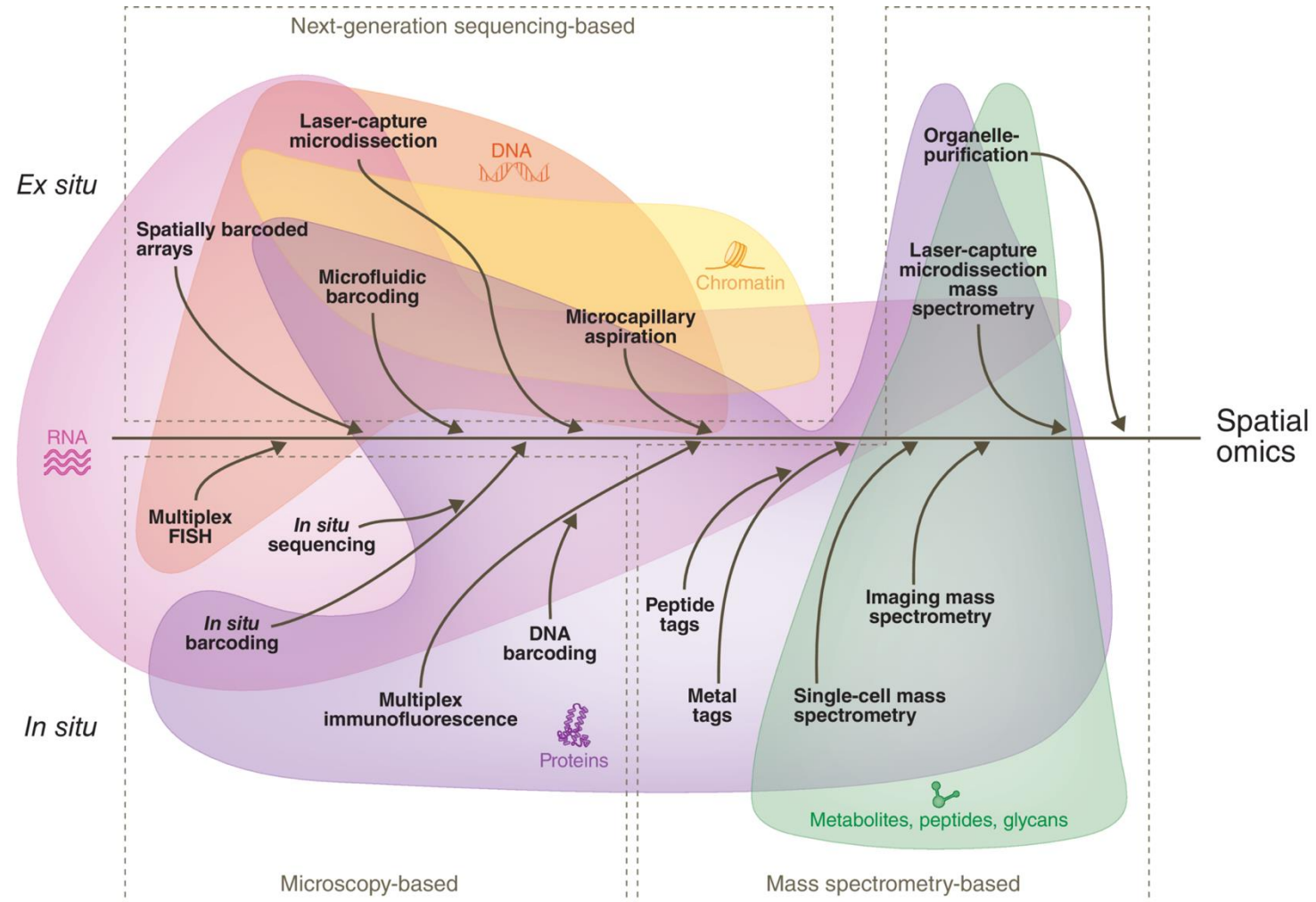
c Identifying cellular structure



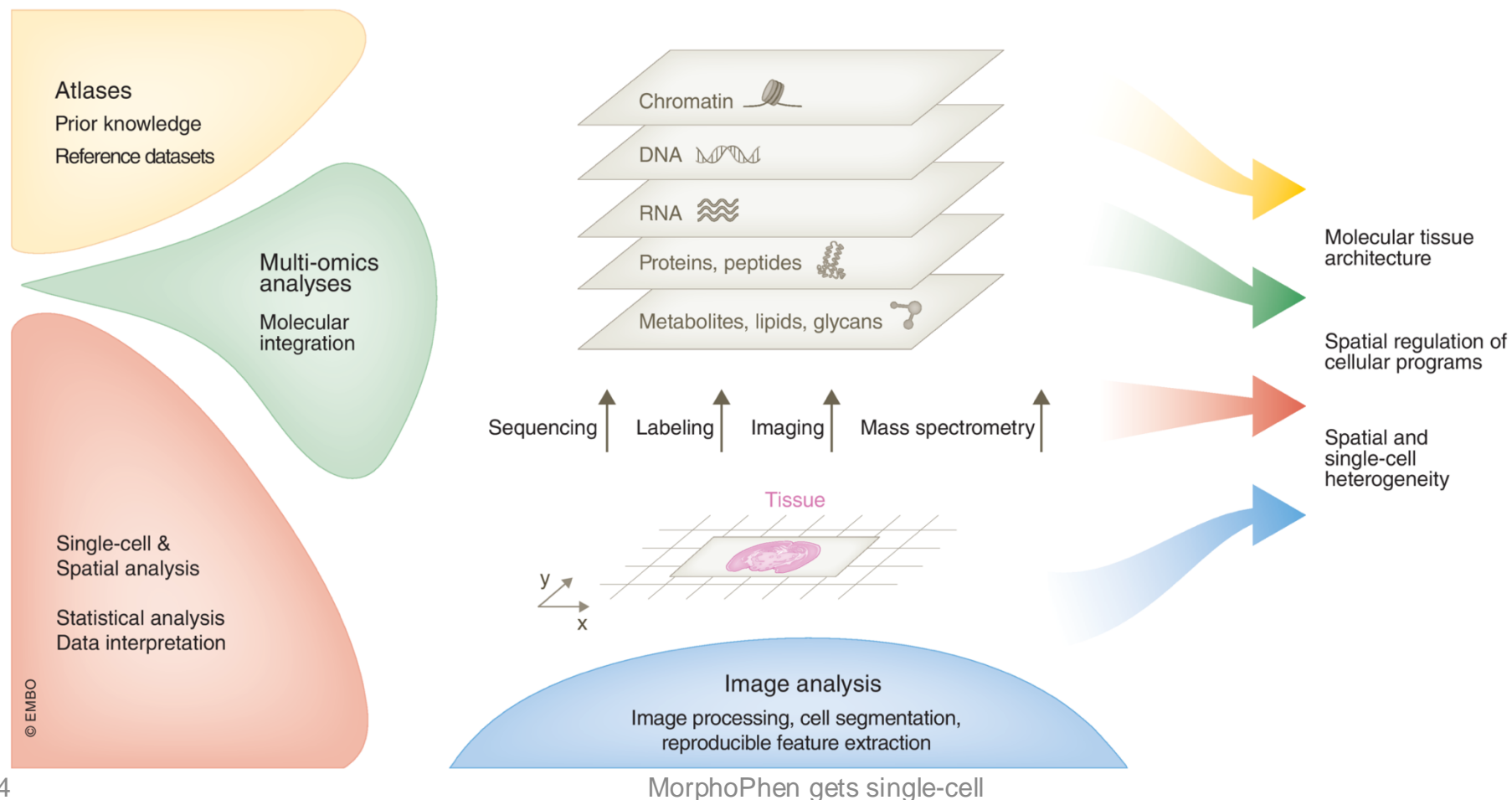
Granularity of spatial technologies

| Technology | Spot Size (μm) | Resolution (Pixels per Spot) | Array Dimensions |
|--|--------------------------------|------------------------------|-------------------------------|
| 10x Genomics Visium | ~55 μm | ~100-150 pixels per spot | 6.5 mm × 6.5 mm (4992 spots) |
| 10x Genomics Visium HD | ~2 μm | ~2500-3000 pixels per spot | ~100K spots per slide |
| ST (Spatial Transcriptomics by SciLifeLab) | ~100 μm | ~50-100 pixels per spot | 6 mm × 6 mm (1007-2000 spots) |
| GeoMx DSP (NanoString) | Region of Interest (ROI) based | Not fixed (varies per ROI) | Custom selection of areas |
| Slide-seq v2 | ~10 μm | ~500 pixels per spot | ~100K spots per array |
| DBiT-seq | ~10 μm | ~500 pixels per spot | Customizable grid |
| MERFISH | Subcellular (~0.3 μm) | Single-molecule resolution | Whole tissue |
| SeqFISH+ | Subcellular (~0.3 μm) | Single-molecule resolution | Whole tissue |
| HDST (High-Definition Spatial Transcriptomics) | ~2 μm | ~2500 pixels per spot | High-density grid |

Spatial Multi-omics



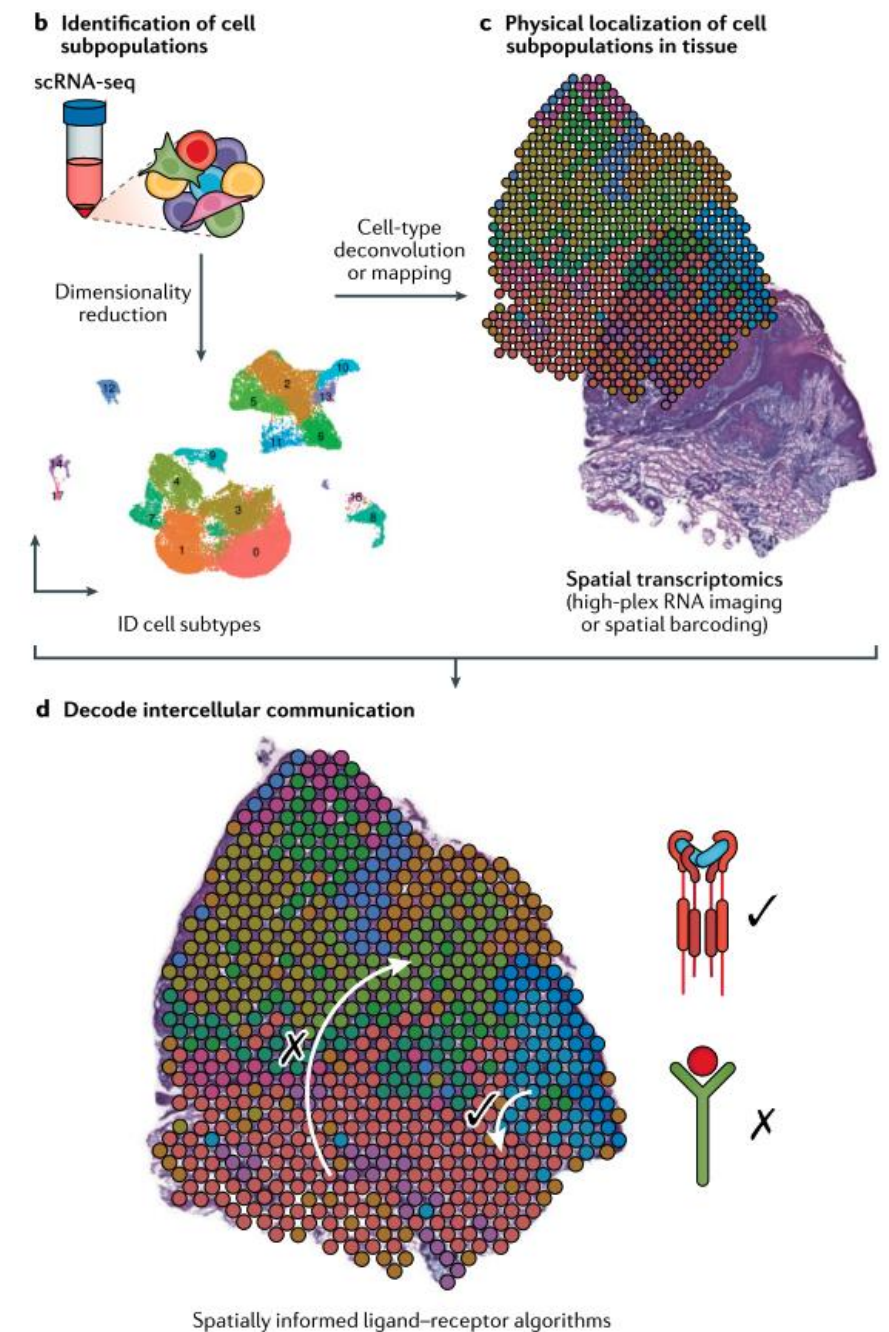
Spatial Multi-omics and new insights



Integrating scRNA-seq with Spatial datasets

Several spatial methods are still unable to create single-cell resolution spatial transcriptomics maps in which the transcriptome of each cell is captured at a depth akin to scRNA-seq.


This limitation underscores the need to integrate current spatial transcriptomics platforms with scRNA-seq to maximize resolution in tissue.



Visium

The Visium Spatial Gene Expression Solution

Unravel biological architecture in normal and diseased tissue and discover new biomarkers



The image shows two vertical rectangular slides, likely representing the Visium Spatial Gene Expression solution. Each slide has a white header with a QR code and some text, and a blue body with a grid of small white squares, possibly representing a tissue section or gene expression data.

MERFISH

The screenshot displays the Vizgen MERFISH website. At the top left is the Vizgen logo. The navigation bar includes links for Products, Technology, Applications, Resources, Support, Company, and Contact. A search bar with the placeholder text "Enter Search Term" and a magnifying glass icon is located at the top right. The main heading "MERFISH" is centered, with the subtitle "Multiplexed Error-Robust Fluorescence in situ Hybridization" below it. A descriptive paragraph states: "MERFISH is a massively multiplexed single-molecule imaging technology for spatially resolved transcriptomics capable of simultaneously measuring the copy number and spatial distribution of hundreds to tens of thousands of RNA species in individual cells." Below this, three key features are listed: "COMBINATORIAL LABELING", "SEQUENTIAL IMAGING", and "ERROR ROBUST BARCODING". The bottom section illustrates the workflow, starting with "Optical Barcodes" showing a cell with fluorescent spots, followed by "Identified RNA Transcripts" showing a cell with labeled transcripts (A, B, C, D) and a final image of a tissue section with many cells.

XENIUM



Recap no2!

- Spatial technologies revolutionize how we analyze and study cellular ecosystems – Multi-omics + Spatial, micro-anatomical regions combine to provide a holistic view
- 2 types of methods dominate the field: array or sequence-based and image-based!
- Cell deconvolution and cell segmentation are two very important new parameters for analysis in spatial omics compared to classic single-cell analysis like in scRNA-seq
- Often, combining scRNA-seq with spatial omics can be very informative in understanding cellular heterogeneity

Closing remarks

- A 3-day clustering is coming to an end!
 - Network biology and single-cell omics with an emphasis on Spatial!
 - **Cytoscape introduction, a bit of Drug Repurposing and Seurat pipelines!**
 - **Systems approach are MANDATORY for any modern biomedical scientist!**
 - Bioinformatics is the most democratic form of science-you can learn anything!
 - Biomedical scientists need to infuse this rapidly expanding field of Systems to improve our models, simulations, and computational analyses
 - Ideally, whatever you learned in MorphoPhen, can be applied for Network Analysis and single-cell studies (Machine Learning/Deep Learning, R or Python, Phenotypic analysis, biological knowledge etc....)
 - Stay enthusiastic and often travel outside your comfort zone!
-
- Let's keep a Slack channel active to stay in touch! <https://slack.com/intl/en-gb/>

Hands-on practical #3- Run sequence-based Seurat tutorial

- https://satijalab.org/seurat/articles/spatial_vignette#overview

Let's stay “Networked”

- Thank you!! Don't be a stranger!

