Single-cell data processing and analyses

How to get the most out of your data

CVRC Bioinformatics Team

May 2025

About us (CVRC Bioinformatics Team)

Florencia Schlamp, PhD
Assistant Director of Bioinformatics
- 2019 -

Mike Gildea, PhD
Senior Bioinformatician
- 2020 -

Alex Ferrena, PhD
Senior Bioinformatician
- 2024 -

Sofie Delbare, PhD
Senior Bioinformatician
- 2023 -

All CVRC Labs

PPG & Associated Labs
(Moore, Fisher, Schmidt)

PPG: “Macrophage dysfunction in diabetes, obesity, and atherosclerosis”

What kind of work do we do?

- Data analysis
  - sequencing data (single cell, bulk, spatial)
  - non-sequencing based omics data (proteomics, lipidomics, metabolomics)
- Data management
  - HPC BigPurple lab folders
  - GEO repositories
- Data mining
  - extracting data from published work and public resources
- Setting up collaborations
- Bioinformatics consultation (grants and proposals)
- Bioinformatics training (lectures, seminars, one-on-one, journal club)

Current types of data

- Single cell:
  - snRNA-seq
  - scRNA-seq (10x / PIP-seq)
  - scATAC-seq
  - CITE-seq
- Bulk:
  - RNA-seq, microRNA-seq
  - ATAC-seq, CHIP-seq, CUT&RUN
  - RIP-seq/ChIRP-seq
- Spatial:
  - Visium
  - GeoMx
- Non-sequencing based omics (lipids/proteins):
  - Targeted (olink)
  - Untargeted (mass spec)

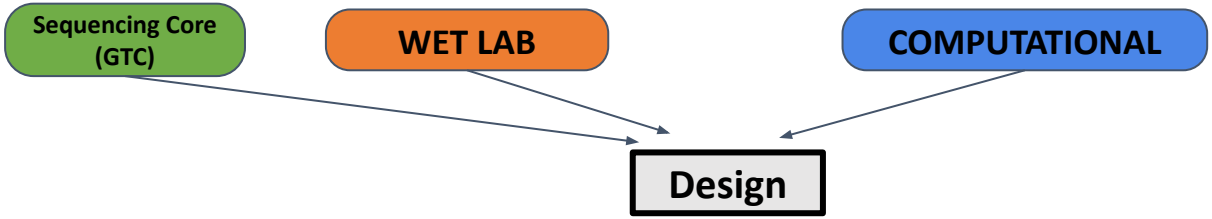
Single Cell data types

- **Transcriptome profiling: single-cell RNA-seq / single-nucleus RNA-seq**
  - Choice between cell vs. nucleus driven by tissue or cell types that need to be profiled
  - 10X / PIPseq

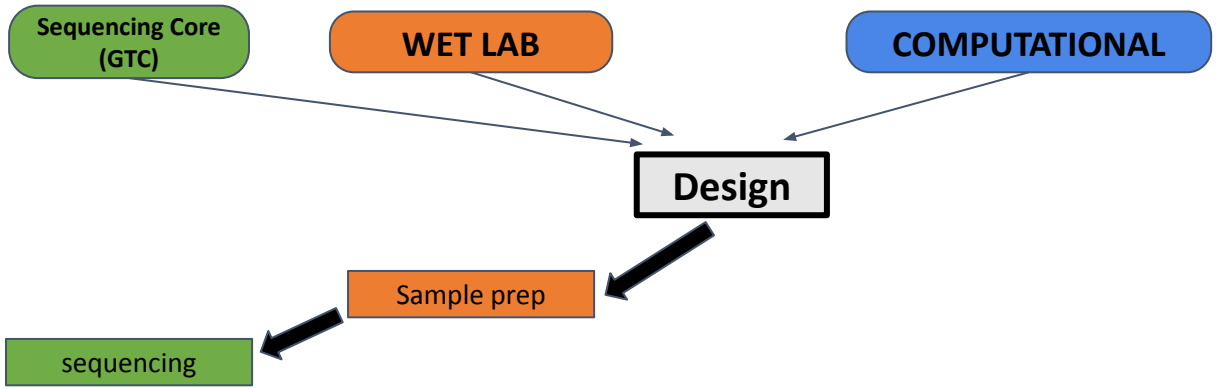
- **Chromatin accessibility: single-cell ATAC-seq**
  - Epigenetic information, mechanisms of gene regulation

- **Surface proteins: CITE-seq**
  - Defined panel of proteins can aid with cell type annotations, e.g. CD4 vs CD8

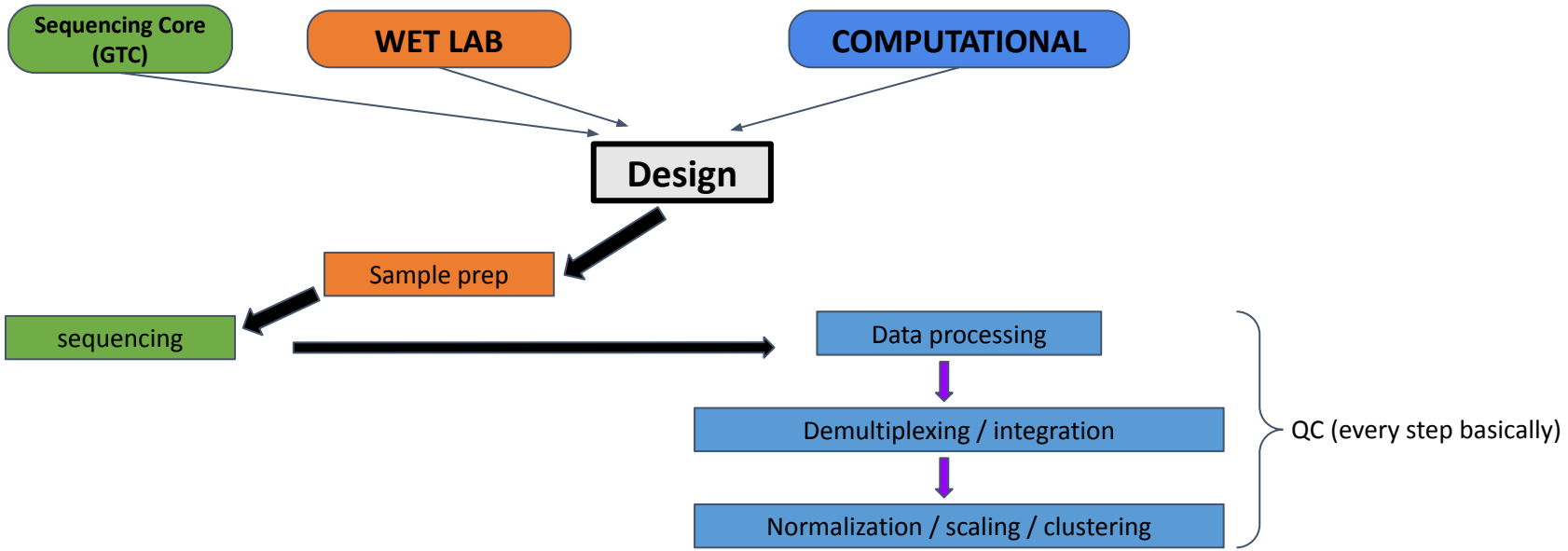
A bird's eye view of scRNA-seq analysis



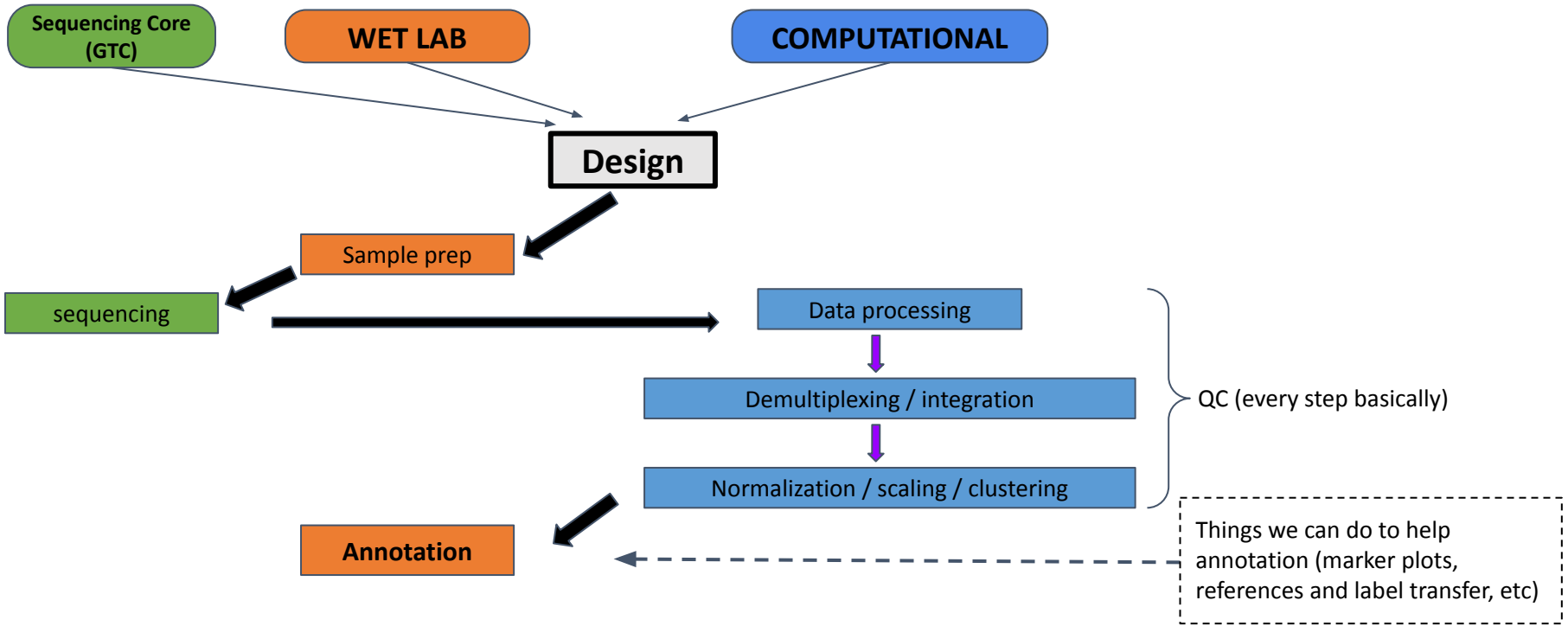
A bird's eye view of scRNA-seq analysis



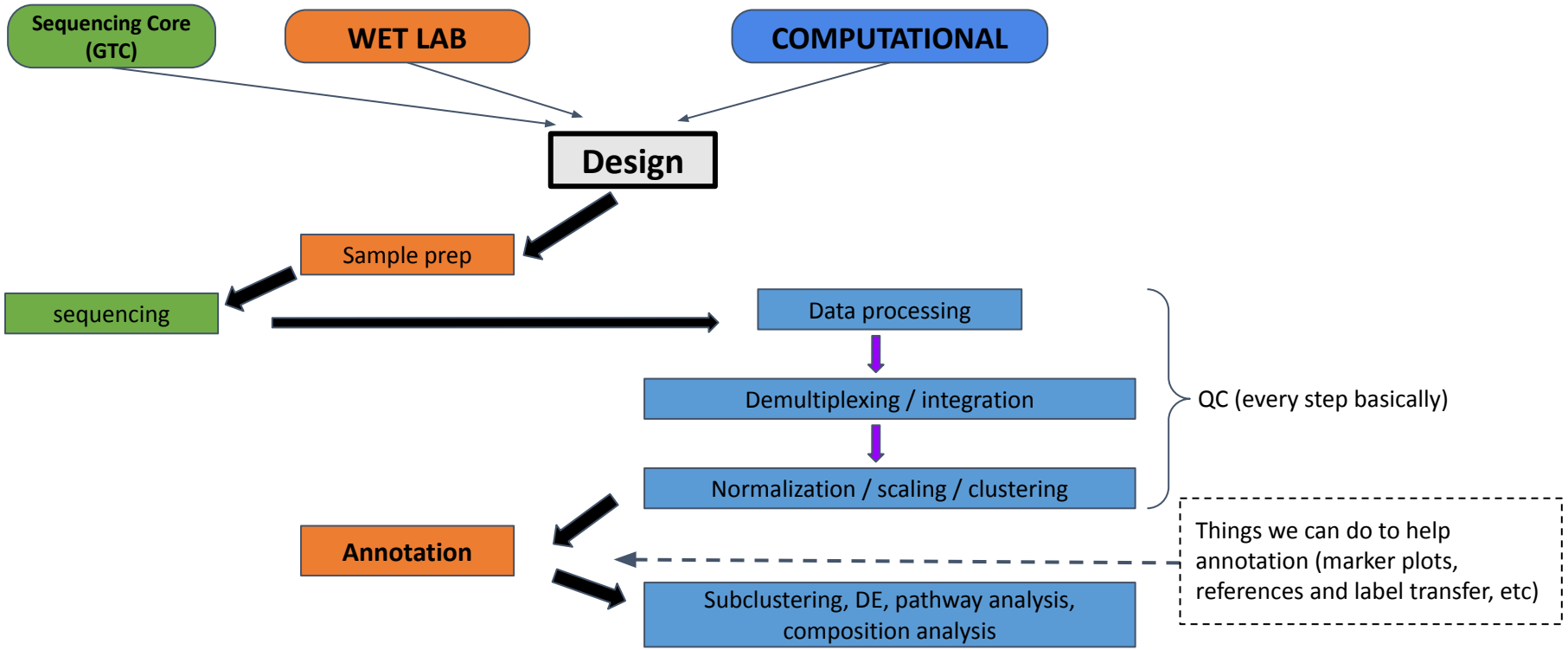
A bird's eye view of scRNA-seq analysis



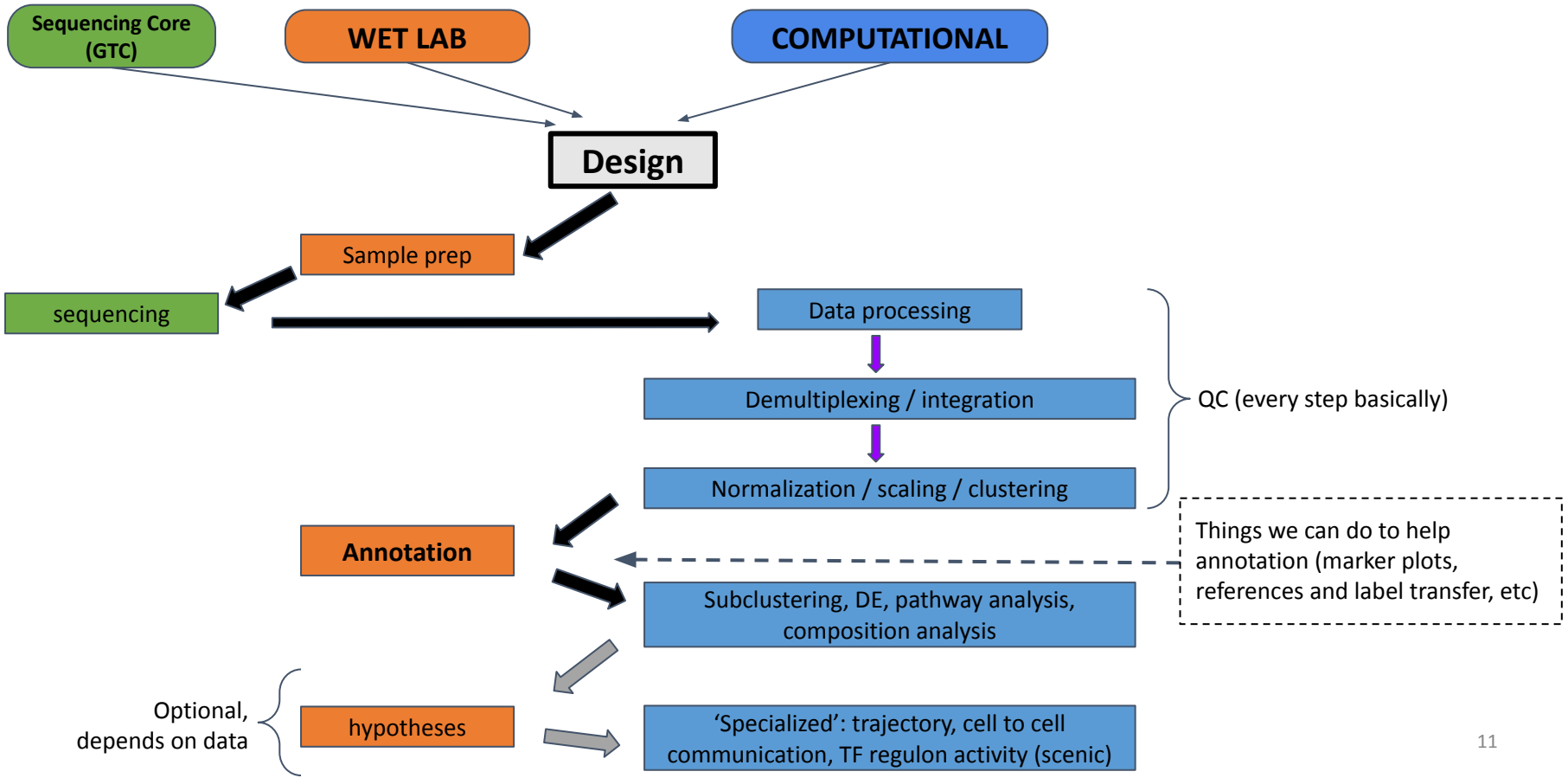
A bird's eye view of scRNA-seq analysis



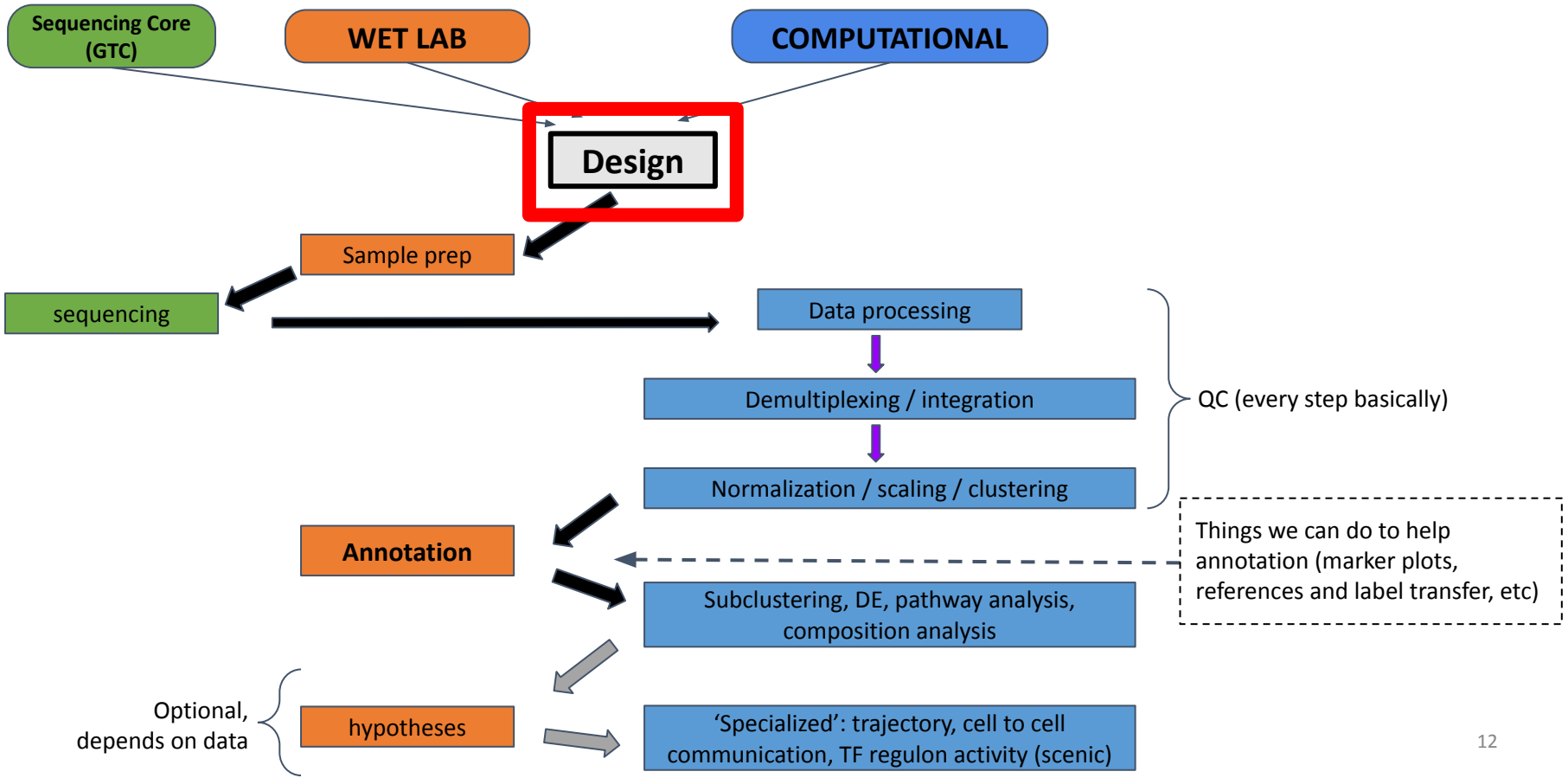
A bird's eye view of scRNA-seq analysis



A bird's eye view of scRNA-seq analysis



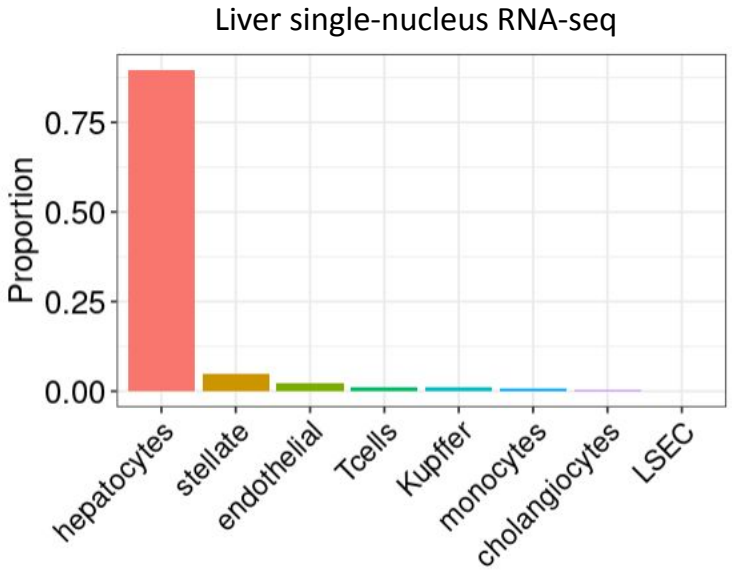
A bird's eye view of scRNA-seq analysis



Design:
What to have in mind before starting

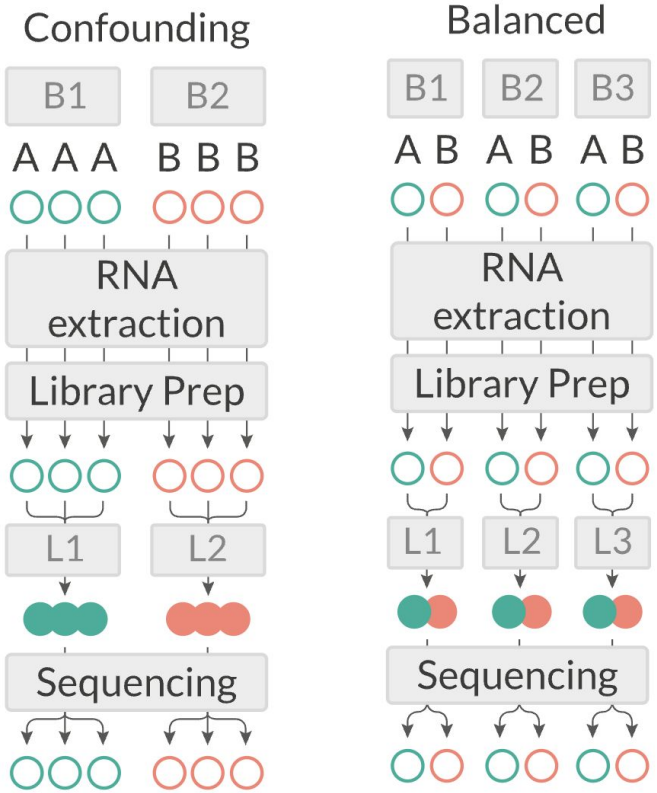
- **Single-cell vs. single-nucleus, with/without FACS?**
  - What are the cell types of interest?

- **Which cell types are expected to be present in the tissue? How can these be annotated?**
  - Marker genes < literature
  - Published reference datasets



Design:
What to have in mind before starting

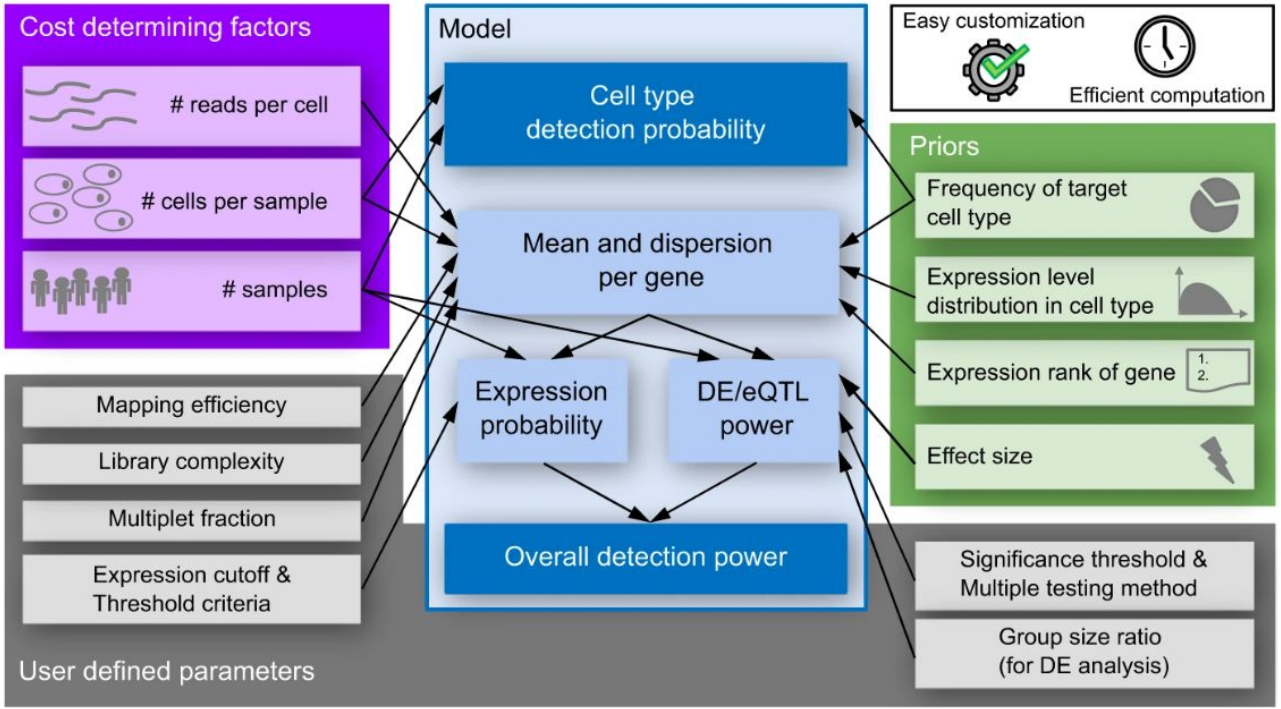
- **Batch setup and replicate multiplexing**
  - Balanced design
  - Optimize hashtag concentration



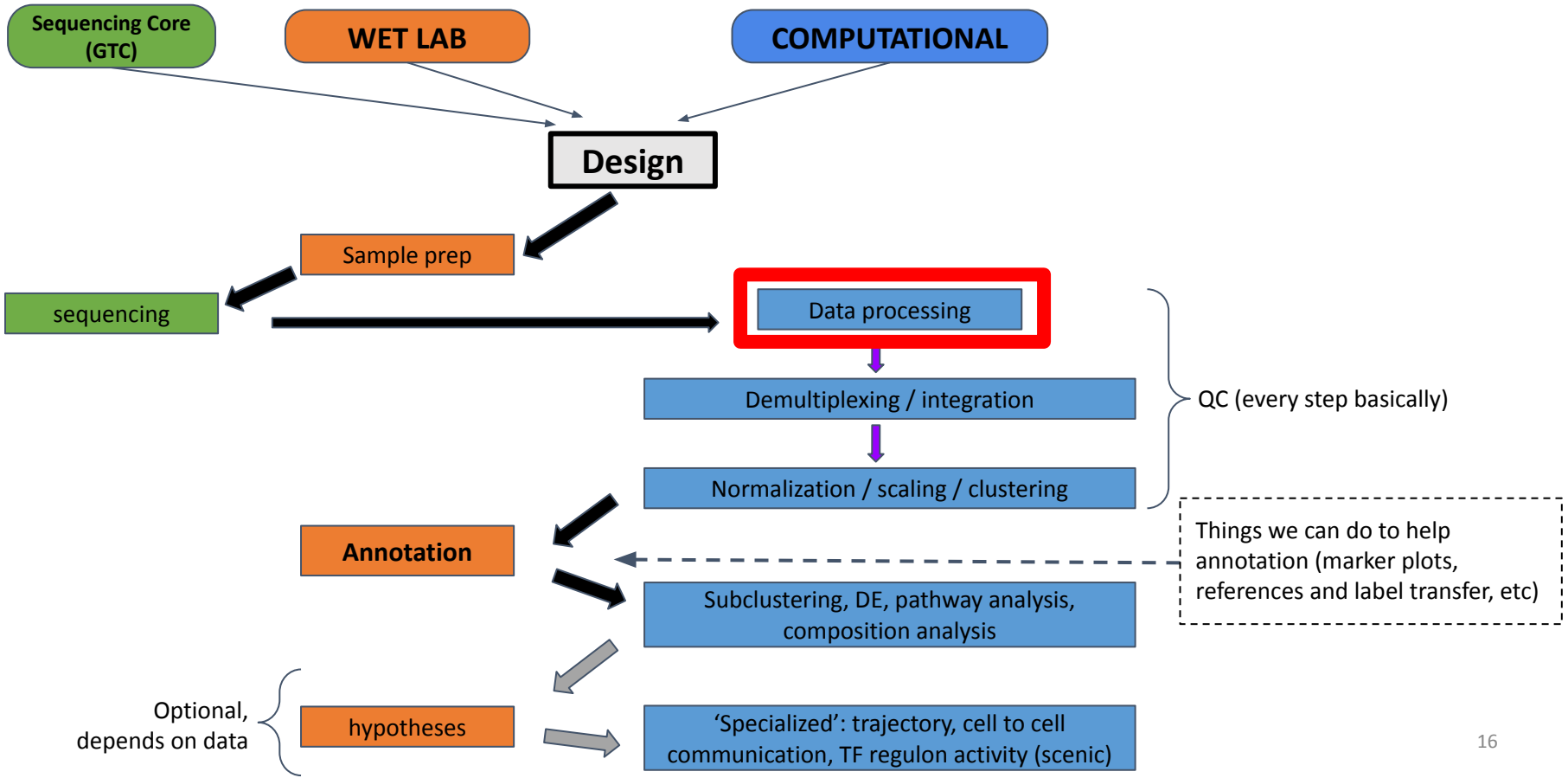
Ahren 2019 SciLifeLab

Design:
What to have in mind before starting

- **Power calculations:** number of replicates, number of cells, sequencing depth
  - Technical limitations: consult the sequencing core
  - scPower (*Schmid et al. 2021 Nat. Comm.*)

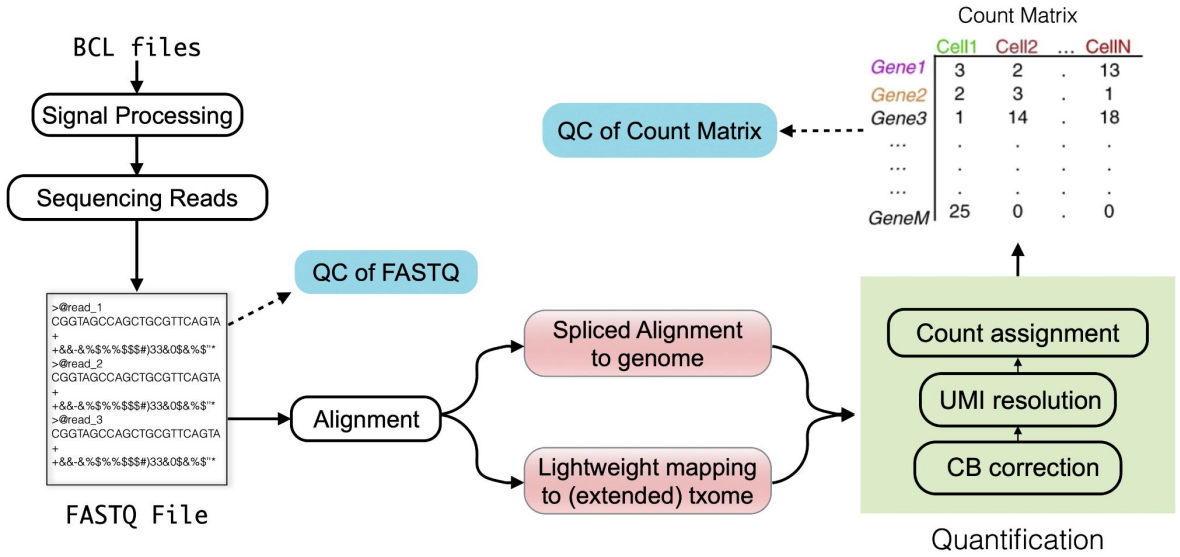


A bird's eye view of scRNA-seq analysis



Data processing:
from raw base calls to cell-level read quantification

Generation of a gene expression count matrix
(Cell Ranger, PIPseeker)



Heumos et al 2023; Chapter 3

Data processing:
Cell Ranger/PIPseeker output

10x GENOMICS Cell Ranger • count

Alerts

The analysis detected 1 informational notice.

	Alert	Value	Detail
ⓘ	Intron mode used		This data has been analyzed with intronic reads included in the count matrix. This behavior is different from previous Cell Ranger versions. If you would not like to count intronic reads, please rerun with the "include-introns" option set to "false". Please contact support@10xgenomics.com for any further questions.

Summary | Gene Expression | Antibody

8,319
Estimated Number of Cells

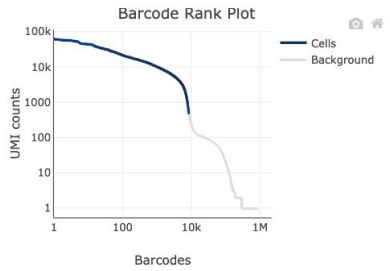
123,210
Mean Reads per Cell

1,641
Median Genes per Cell

Sequencing

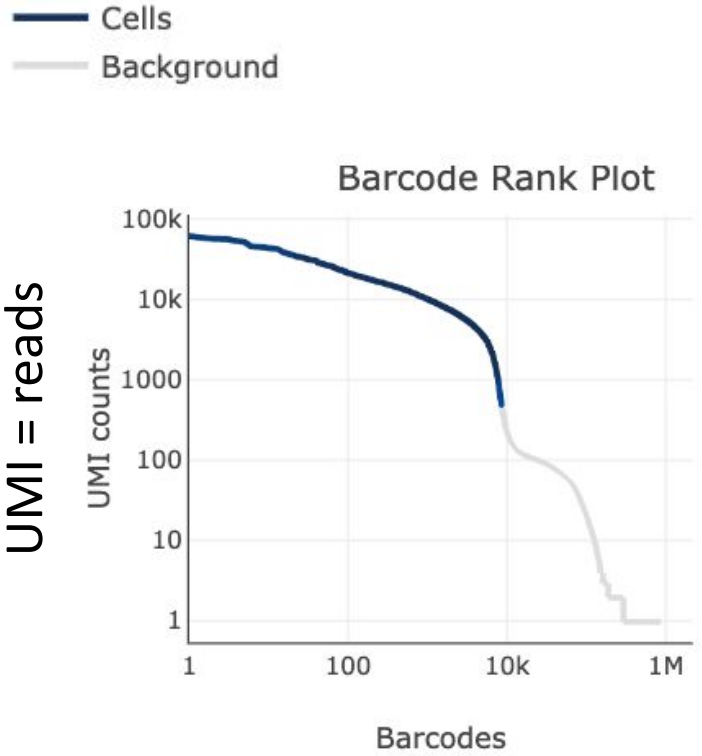
Number of Reads	1,024,981,512
Number of Short Reads Skipped	0
Valid Barcodes	96.6%
Valid UMIs	99.9%
Sequencing Saturation	92.0%

Cells



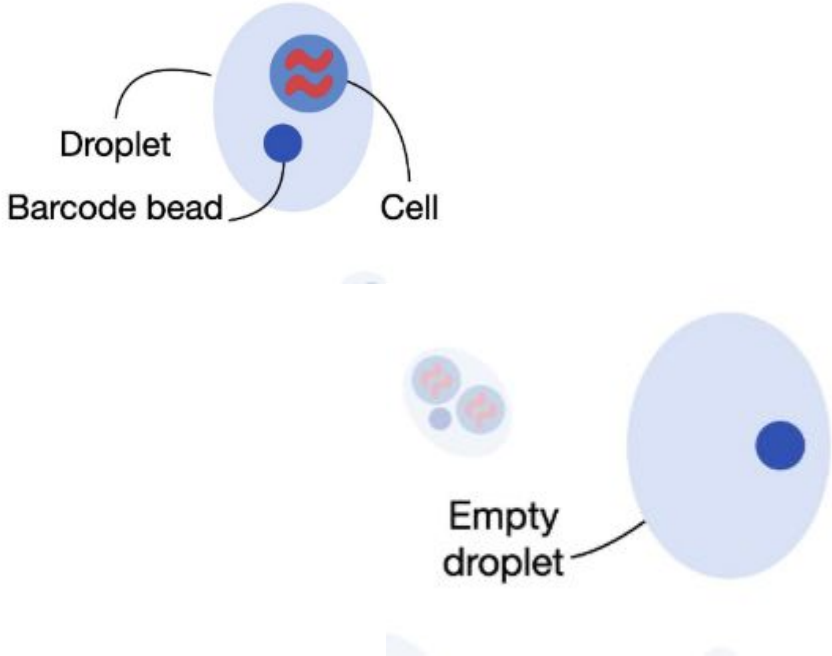
Estimated Number of Cells	8,319
Fraction Reads in Cells	87.5%
Mean Reads per Cell	123,210

Data processing:
Cell Ranger/PIPseeker output

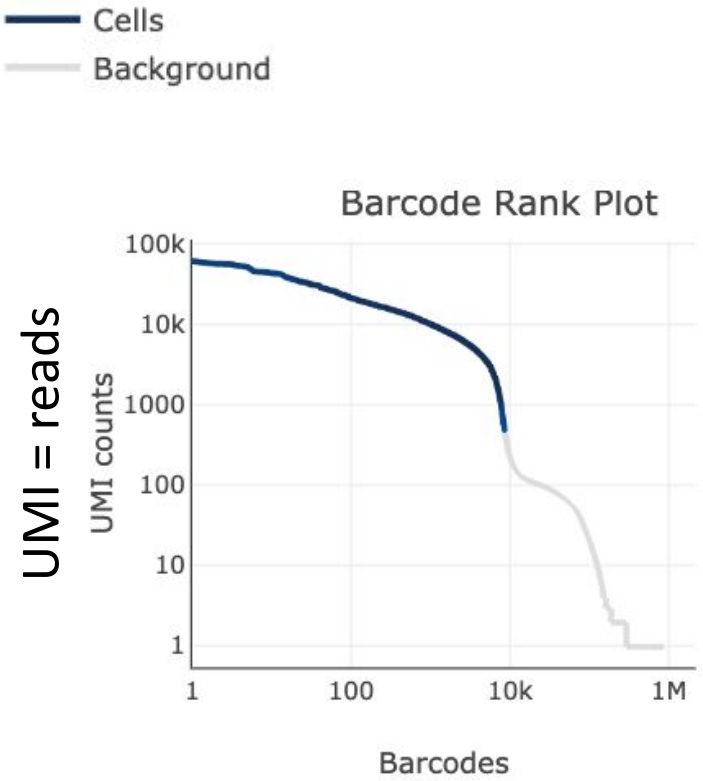


UMI = reads

Barcode ~ droplet

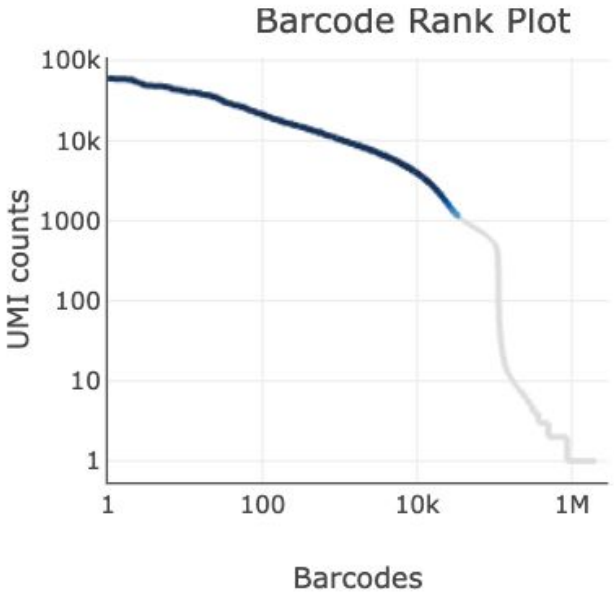


Data processing:
Cell Ranger/PIP-seq output



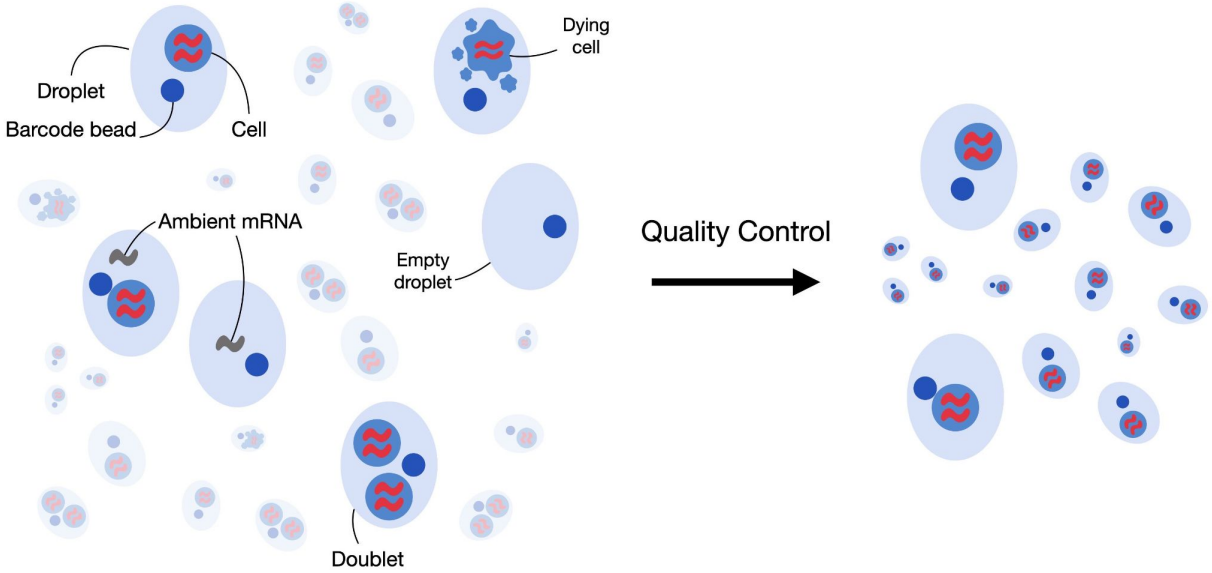
UMI = reads

Barcode = droplet



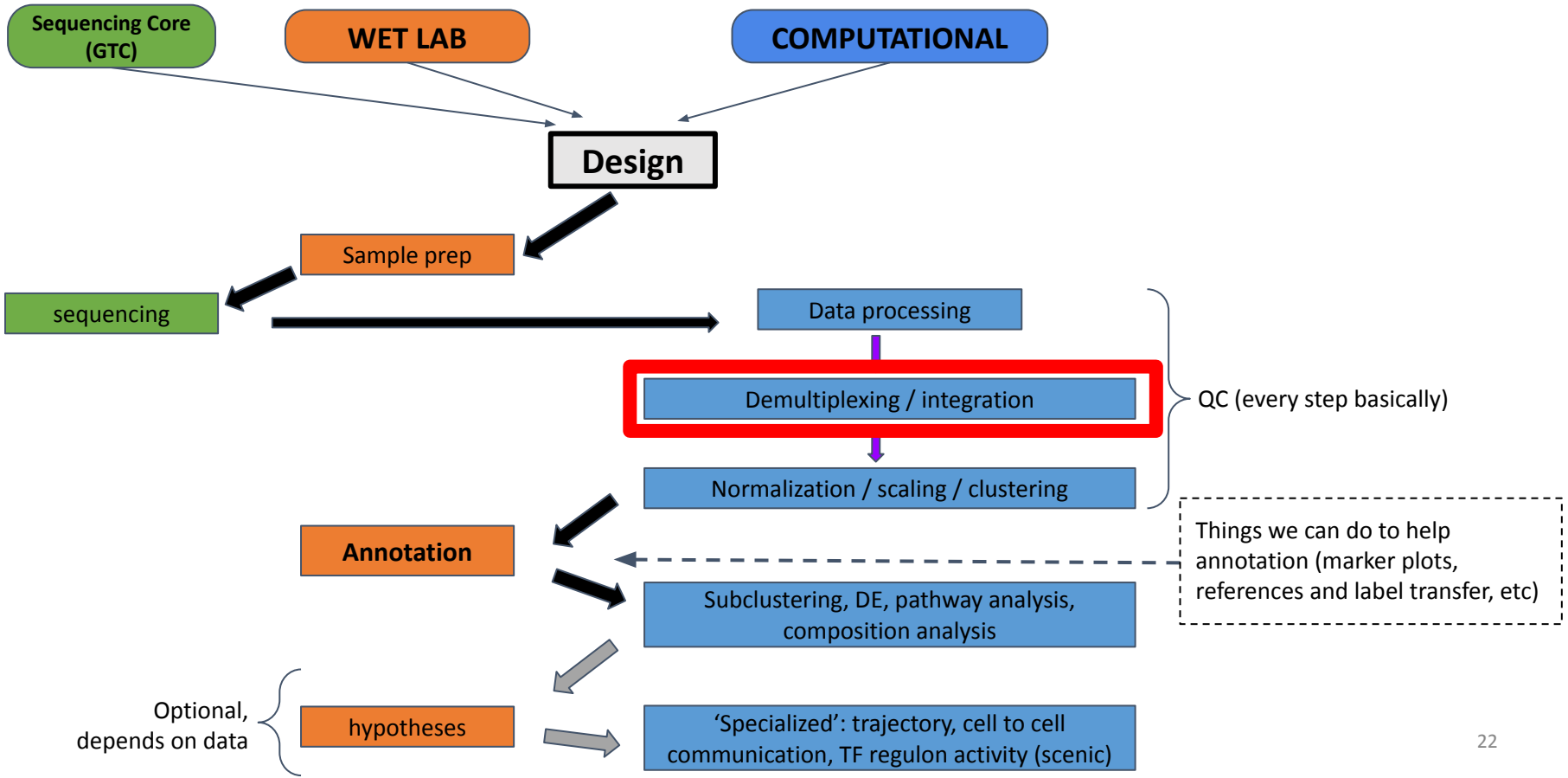
High levels of background noise

Data processing:
additional cell-level quality control

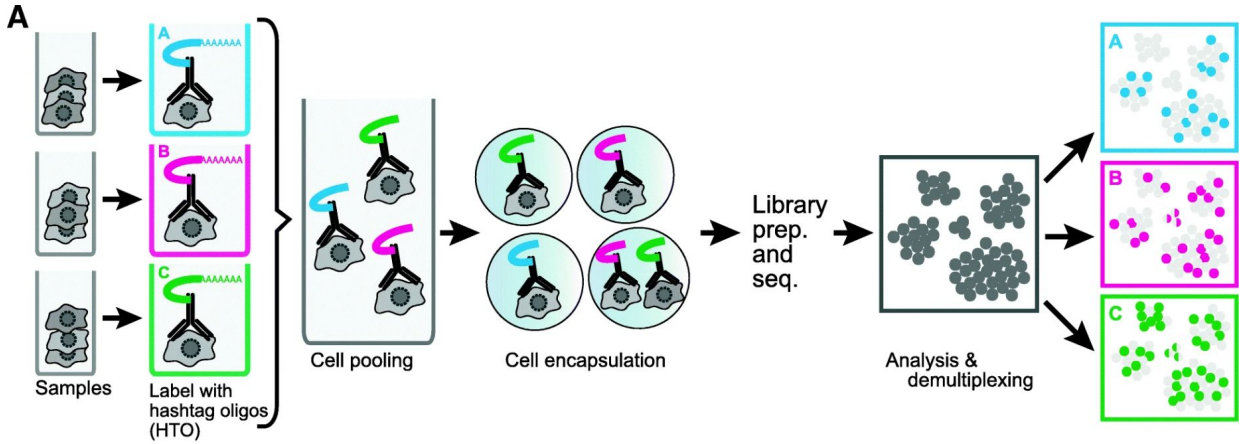


Heumos et al 2023; Chapter 6

A bird's eye view of scRNA-seq analysis



Scaling up sample sizes with sample-level “hashtag” multiplexing

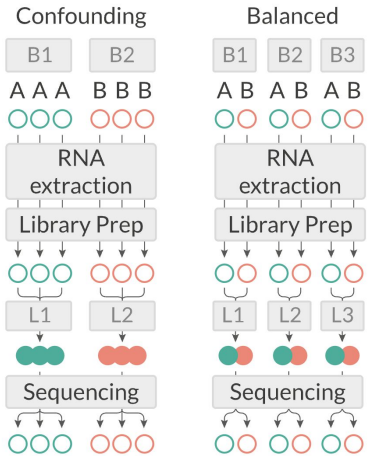


[Stoeckius et al Genom Biol 2018]

Multi-sample batch effects: Sequencing is a very sensitive protocol that can be strongly affected by technical biases. This can be mitigated by balanced experimental design

> **"Batch effects"**: Technical biases introduced during data generation
> **reagent lots**, personnel technique, ambient temperature, humidity, weather, the whims of the gods

Balanced design: If all **Genotype A** is Batch 1, and all **Genotype B** is Batch 2, then there is no way to distinguish real biology vs technical artifact

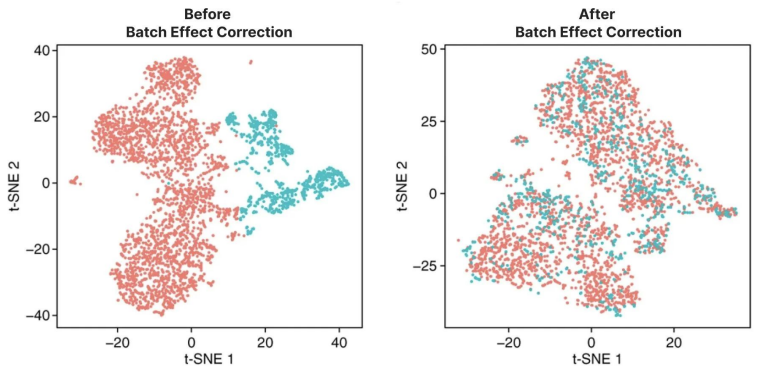


Ahren 2019

Integration: with computational tricks, we can "correct" for batch.

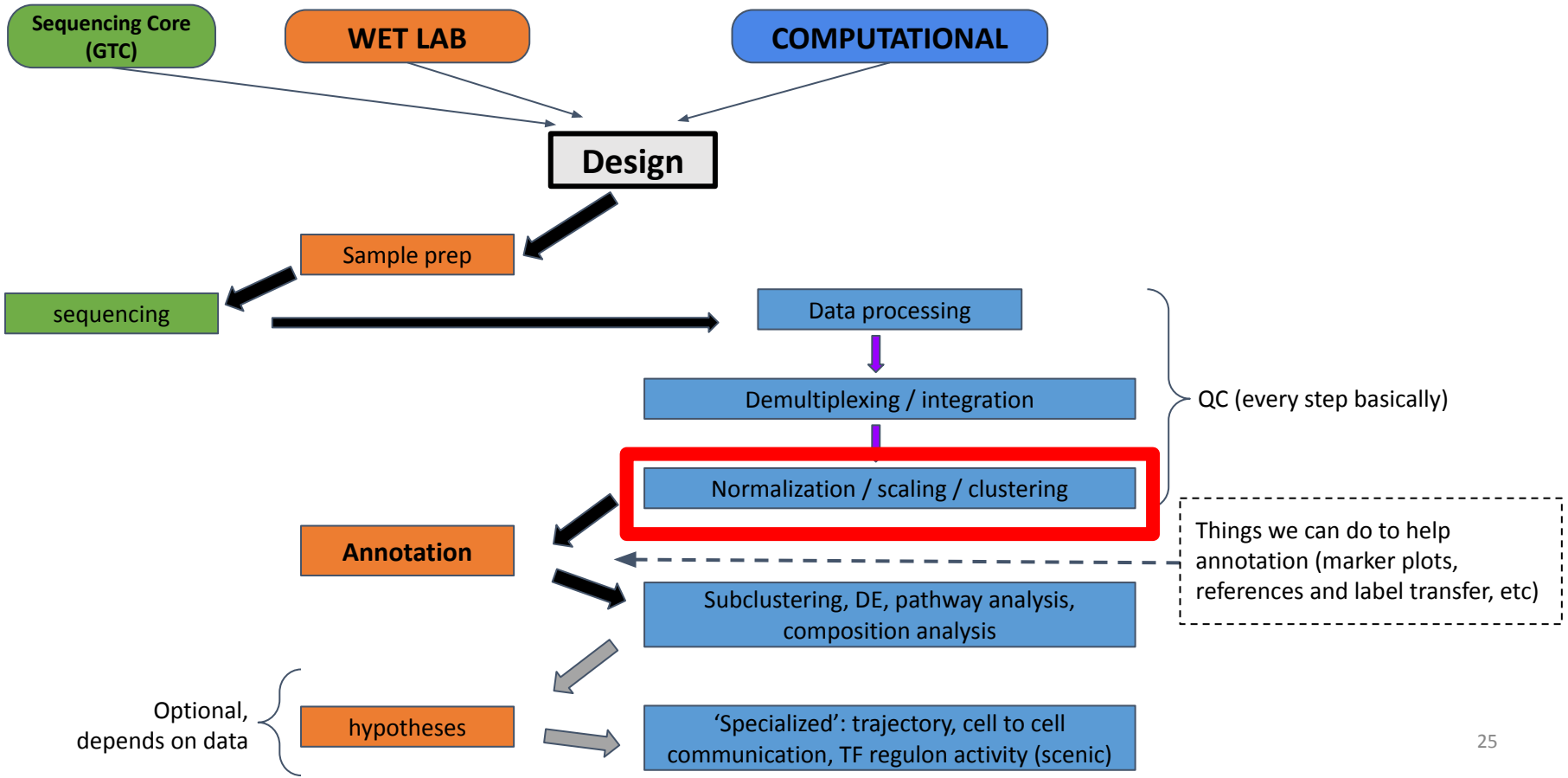
However, this can also eliminate true biology, so it cannot fix imbalanced design

Consultation: we are available to answer questions and help with experimental design!

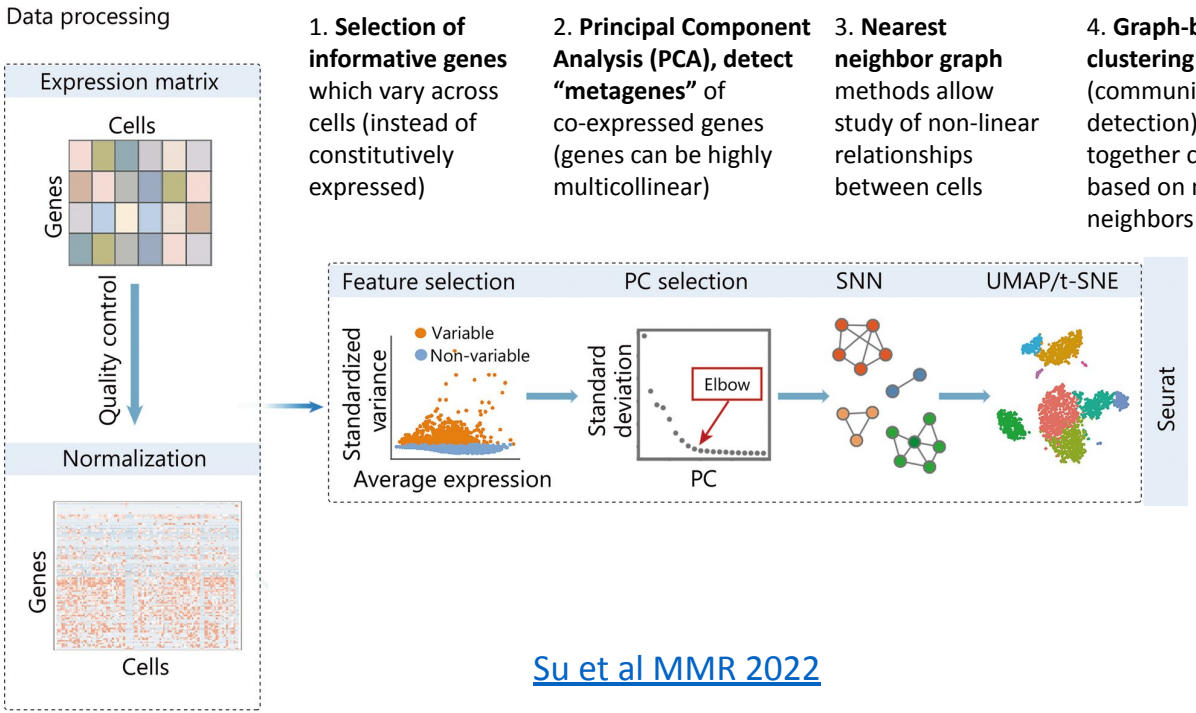


Srivastava 2023

A bird's eye view of scRNA-seq analysis



Clustering: classical machine learning methods are often used to group similar cells by shared transcriptomic patterns



1. **Selection of informative genes** which vary across cells (instead of constitutively expressed)

2. **Principal Component Analysis (PCA), detect "metagenes"** of co-expressed genes (genes can be highly multicollinear)

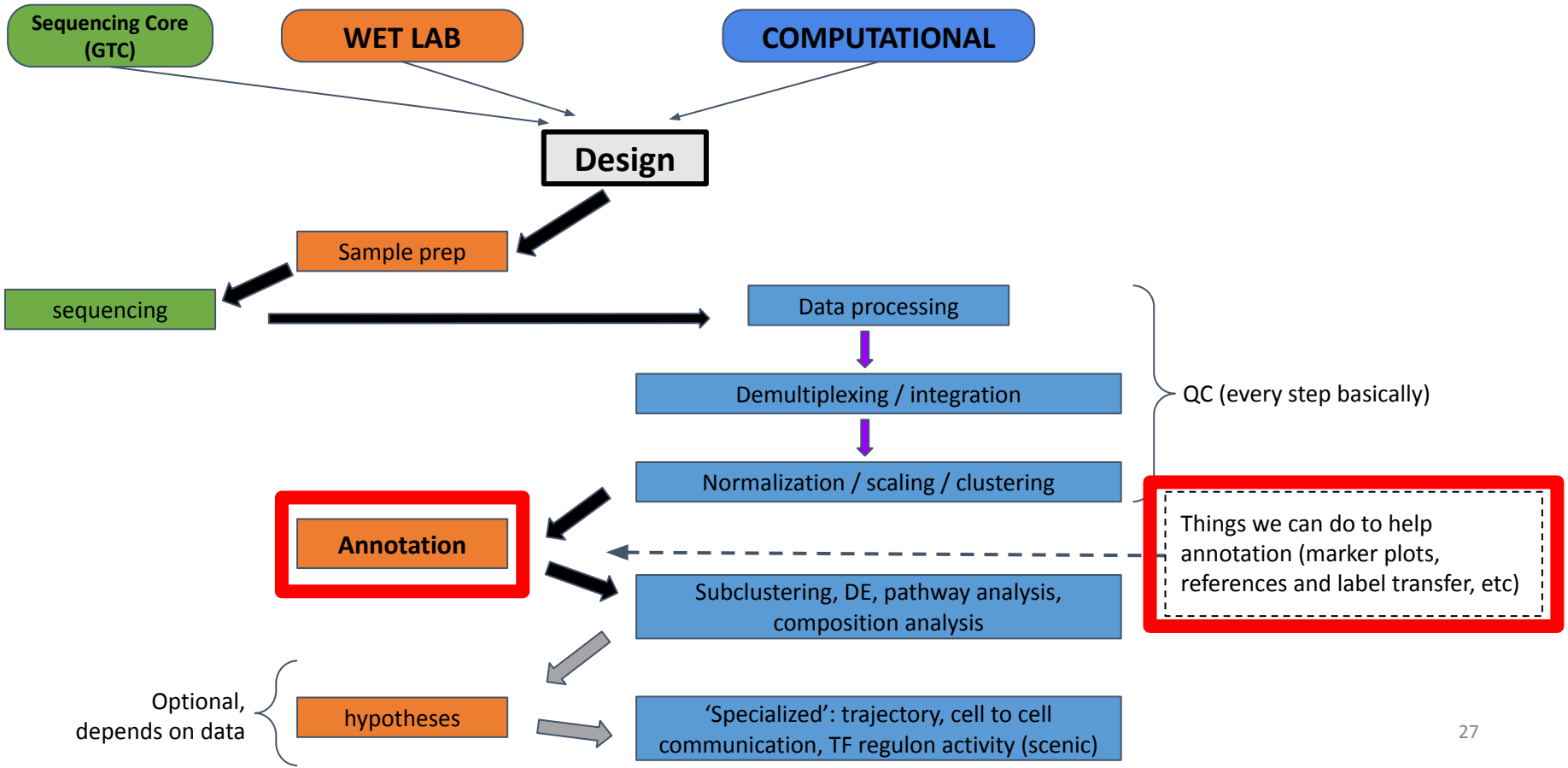
3. **Nearest neighbor graph** methods allow study of non-linear relationships between cells

4. **Graph-based clustering** (community detection) to group together cells based on nearest neighbors

Normalization: correct for sequencing bias, and place genes in similar numeric scale

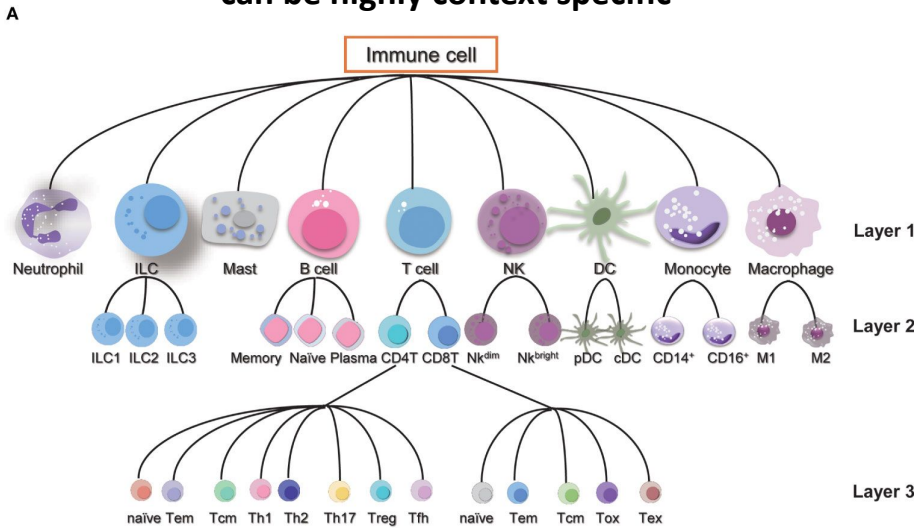
Su et al MMR 2022

A bird's eye view of scRNA-seq analysis



Annotation of clusters to celltypes: the most challenging aspect of single-cell analysis

Cell types are defined in various ways:
Lineage hierarchical or functional state;
marker based: binary (+/-) or relative (hi/lo);
can be highly context specific

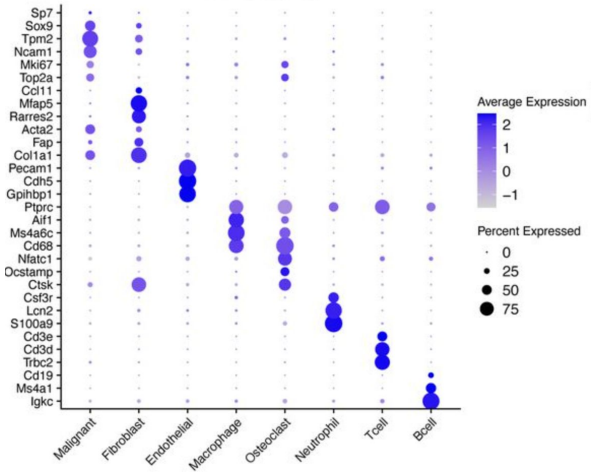


Jiang et al Front. Immunol. 2023

Annotation of clusters to celltypes: Marker- and reference-based annotation

Marker-based annotation:

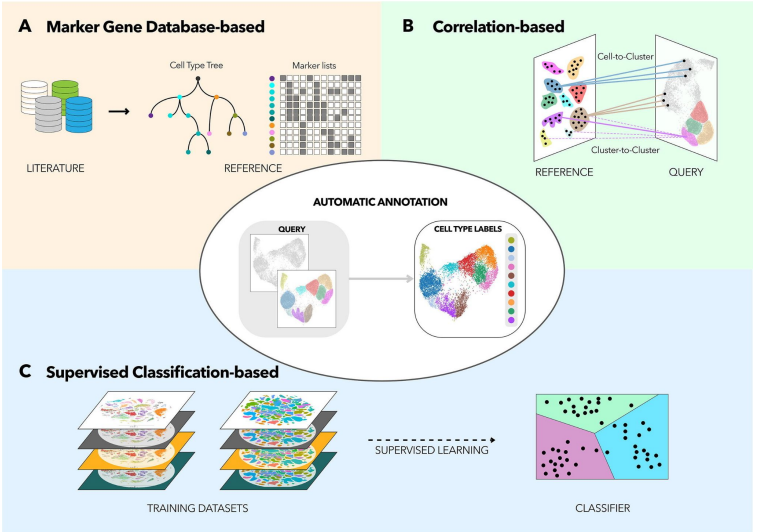
- ***top-down*** **(We checked the known T cell gene CD3 and saw it expressed in cluster 7)**
- ***bottom-up*** **(A test revealed cluster 5 overexpressed *CTSK* relative to other clusters, which Google says is an osteoclast gene)**



Ferrena et al bioRxiv 2024

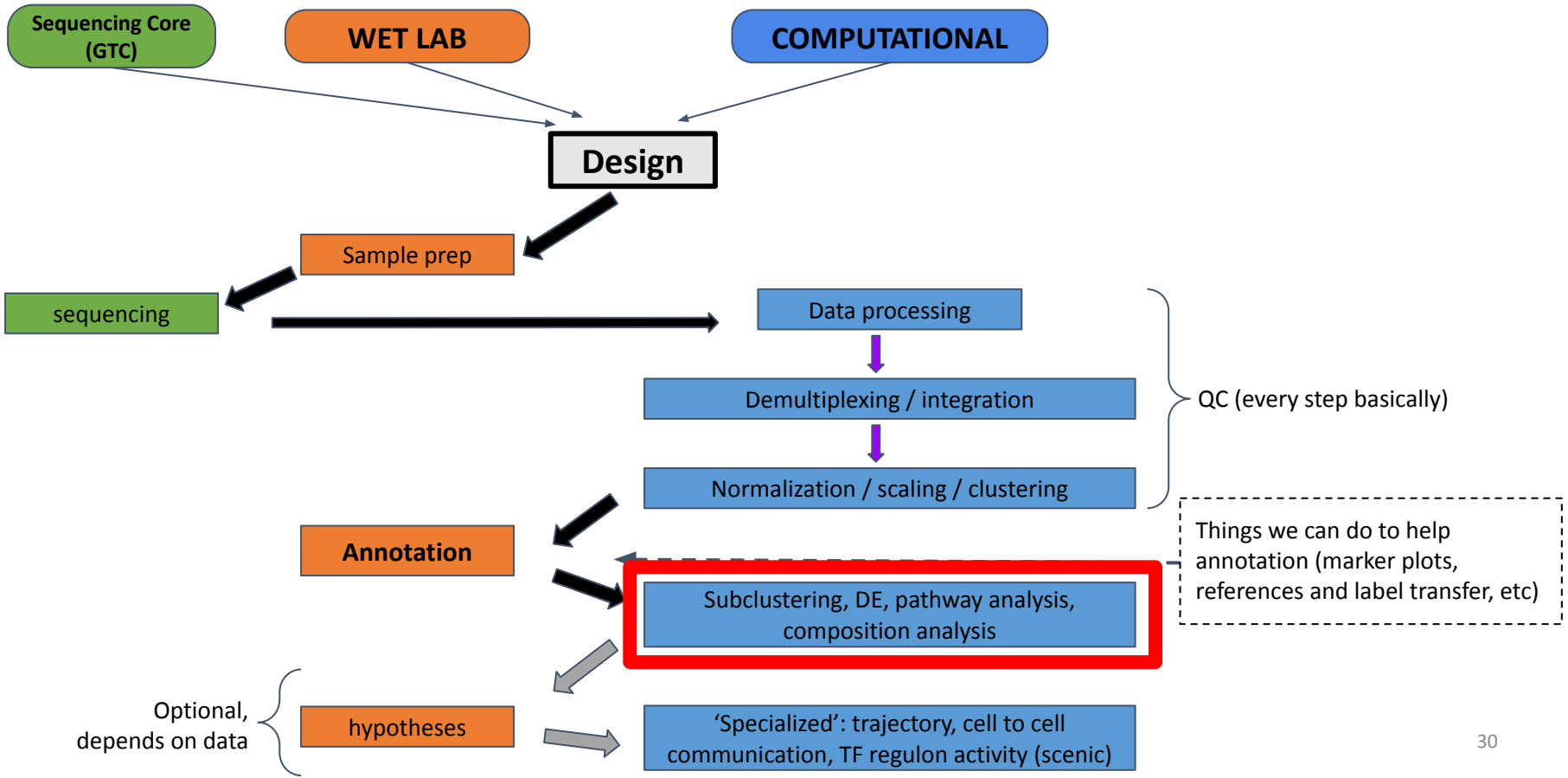
Reference-based suggestions: cell annotation can be aided by previously-annotated references via *Label Transfer*

In the near-future, supervised classification from large atlases will also be helpful

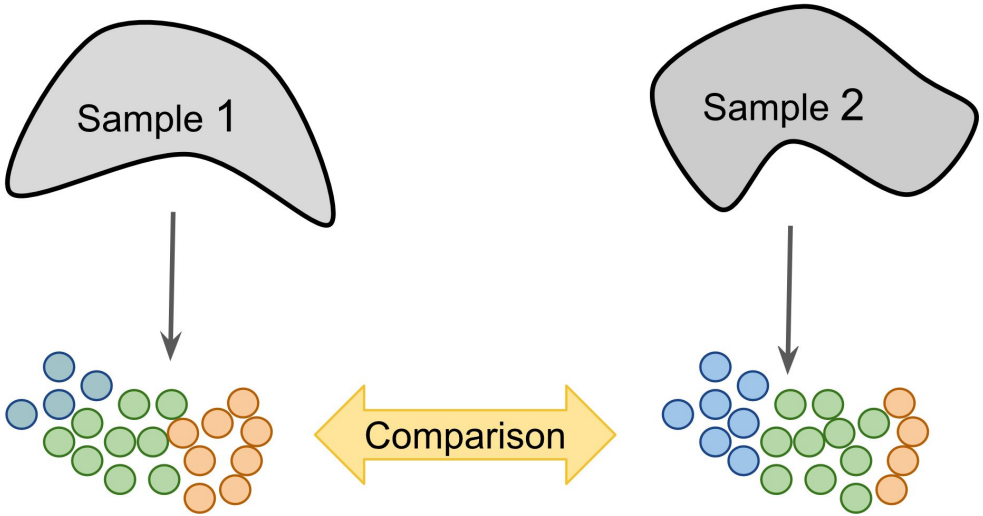


Pasquini et al CSBJ 2021

A bird's eye view of scRNA-seq analysis

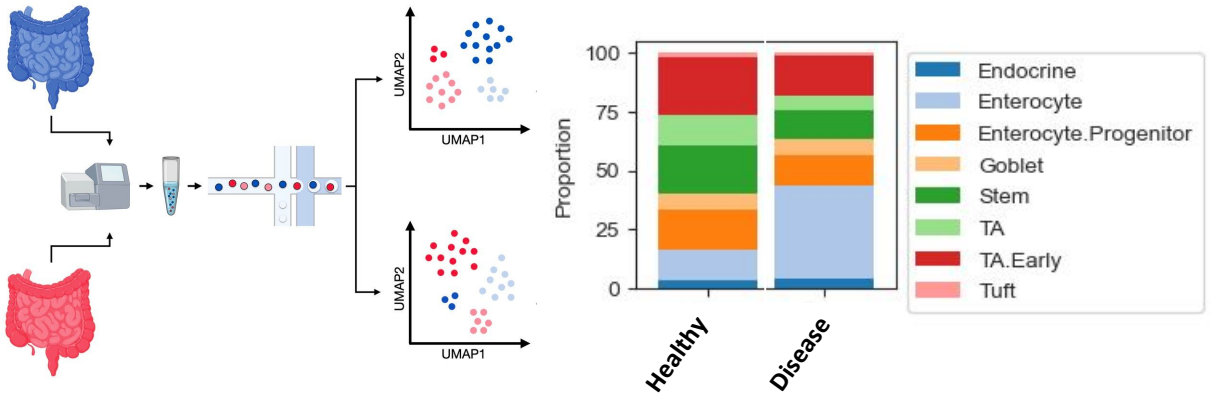


From **profiling heterogeneity** to **comparing biology**: scRNA-seq is increasingly used for comparative readout



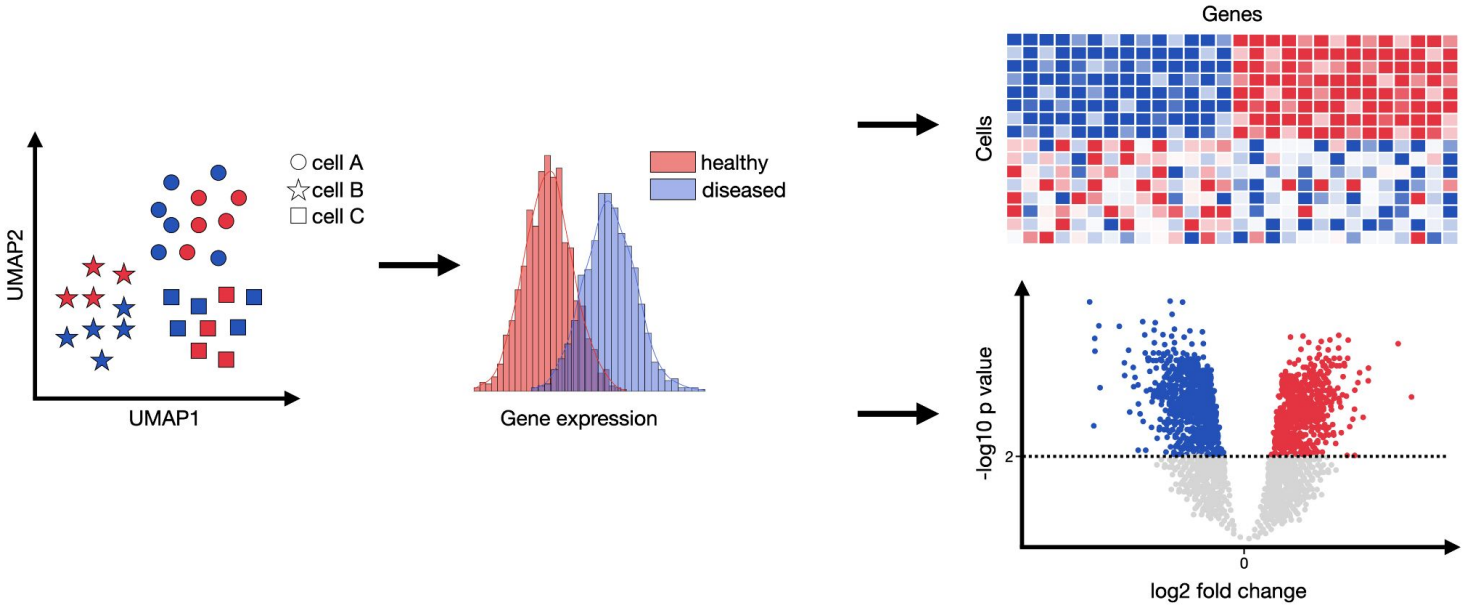
Heumos et al 2023; Chapter 16

Compositional analysis: differential cell abundance can also be compared, revealing celltypes which expand or decline in a tissue across across conditions



Adapted from [Heumos et al 2023](); Chapter 17

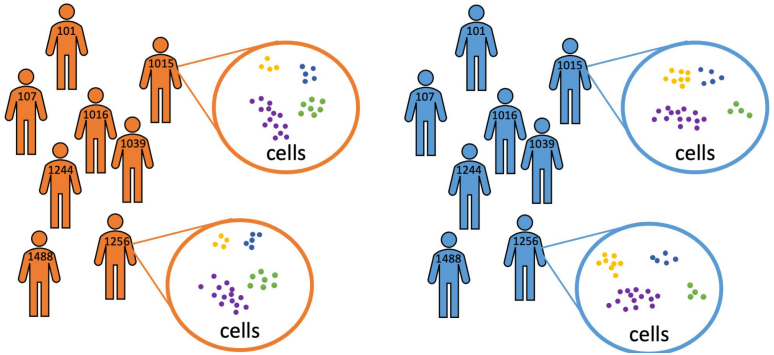
Differential expression: scRNAseq can be used to detect relative expression differences of genes and pathways



Heumos et al 2023; Chapter 16

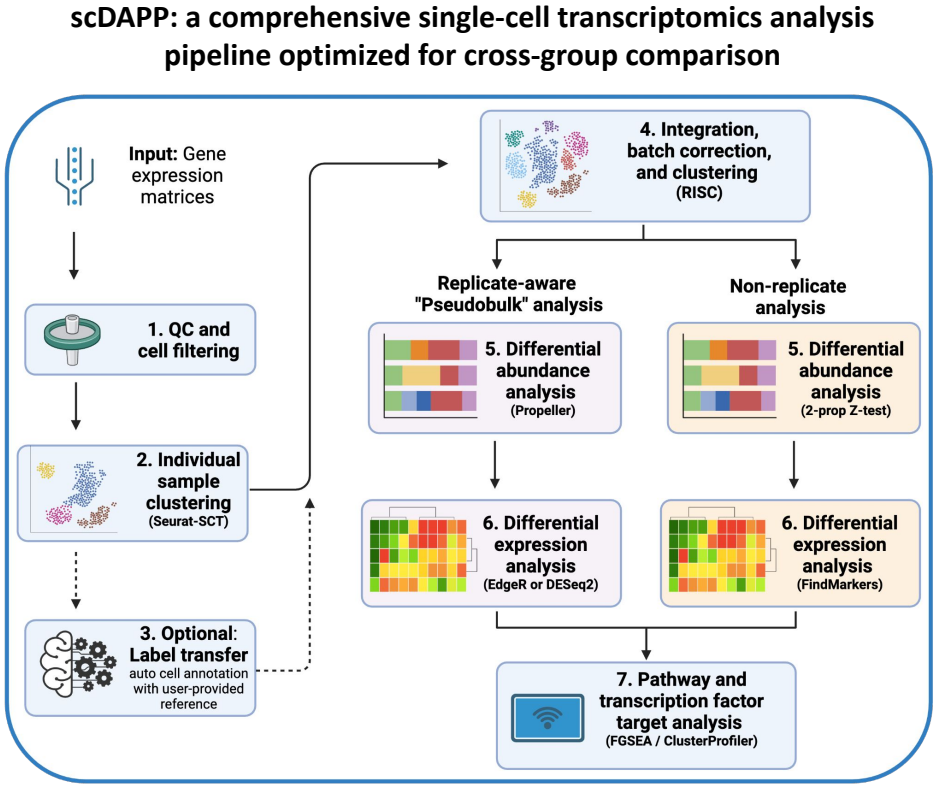
Replicates: scRNAseq is not just 1:1 anymore. The more replicates, the more generalizable the result

One recent bioinformatic development has been to make full use of replicates as the unit of analysis instead of cells via ***pseudobulking*** (ie summating gene expression per-sample and per-celltype). This is very important for external validity of studies



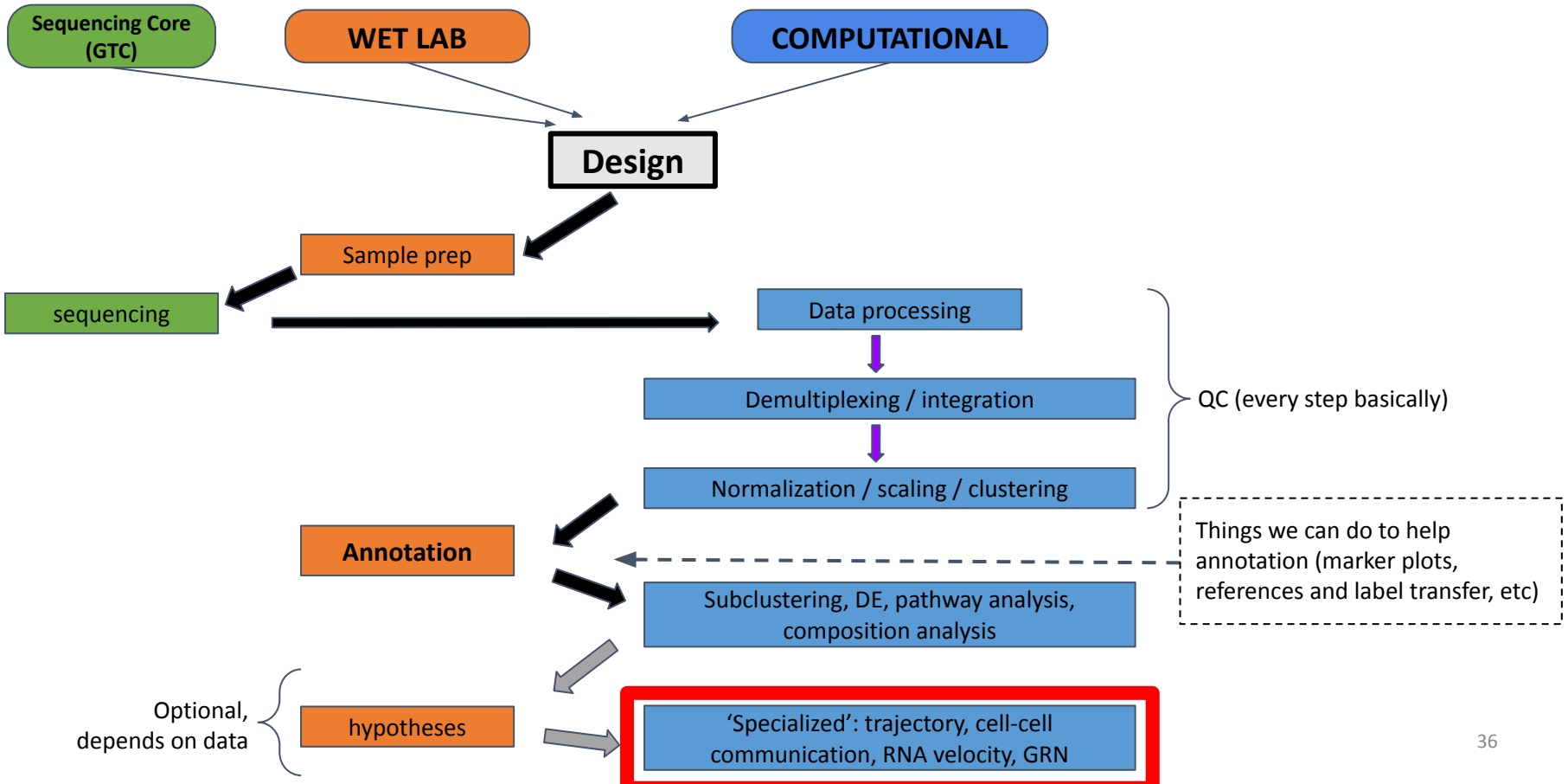
Ahlmann-Eltze 2025

Fast, automated, and reproducible comparative analysis



Ferrena et al NAR Genomics and Bioinformatics 2024

A bird's eye view of scRNA-seq analysis



Exploring single-cell data with Shiny

Shiny - framework for building R and python applications

- Efficiency
- Explore your data
- Aide in annotation process
- Save publication quality plots

1. **Create the app (We do this)**

2. **HPC access**
   We can help you set this up. It’s easy!

3. **Load the app**
   Detailed Tutorial on CVRC-Bioinformatics Github

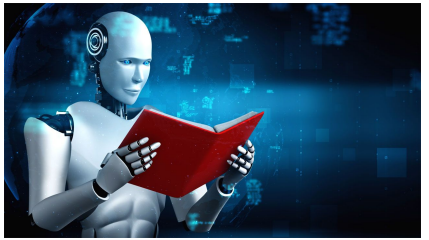
Email Address

CVRCBioinformatics <CVRCBioinformatics@nyulangone.org>

- For general CVRC Bioinformatics communications
- Requests for new projects will be sent here for management
- This address should be CC’d during project updates

- Goal: Centralized communication within bioinformaticians and continuity with any future turnovers

CVRC Bioinformatics Journal Club



~ usually last Friday of each month, hybrid format

Selected papers cover new bioinformatics methods, AI, benchmarks, etc

Upcoming sesion: *Genome modeling and design across all domains of life with Evo 2* (Friday May 16th 11am-1pm)

Past sessions:

- Cell2fate infers RNA velocity modules to improve cell fate prediction
- scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI
- High content of nuclei-free low-quality cells in reference single-cell atlases: a call for more stringent quality control using nuclear fraction
- Cell2location maps fine-grained cell types in spatial transcriptomics

If interested, Contact Alex (alexander.ferrena@nyulangone.org) or CVRCBioinformatics@nyulangone.org

Listserv Interest Form

Interest in Workshops / more talks?

Are there any LECTURE topics you would be interested in attending?

- [ ] Bulk RNA-seq - data processing and analyses
- [ ] Data Visualization and Interpretation
- [ ] Basic Stats
- [ ] Other: _____

Are there any WORKSHOP topics you would be interested in attending?

- [ ] Basics of R coding
- [ ] Basics of UNIX coding (terminal, HPC)
- [ ] Basics of Python
- [ ] Basics of data visualization with R
- [ ] bulk RNA-seq data processing and analyses
- [ ] single cell RNA-seq data processing and analyses
- [ ] ATAC-seq / ChIP-seq data processing and analyses
- [ ] Other: _____

https://forms.gle/3MTqgs9aL9qHShzW8

Resources

CVRC Bioinformatics Github

- Sample scripts: https://github.com/CVRC-Bioinformatics/sample_scripts

Previous CVRC Workshops and Lectures

- https://github.com/florschlamp/CVRC_NYU_Langone

Single Cell Best Practices (Theoretical Practices and Python application)

- https://www.sc-best-practices.org/preamble.html
-

Orchestrating single-cell analysis with Bioconductor (R based)

- https://bioconductor.org/books/release/OSCA/

Bioinformatics Training at the Harvard Chan Bioinformatics Core

- https://hbctraining.github.io/main/

Email us!

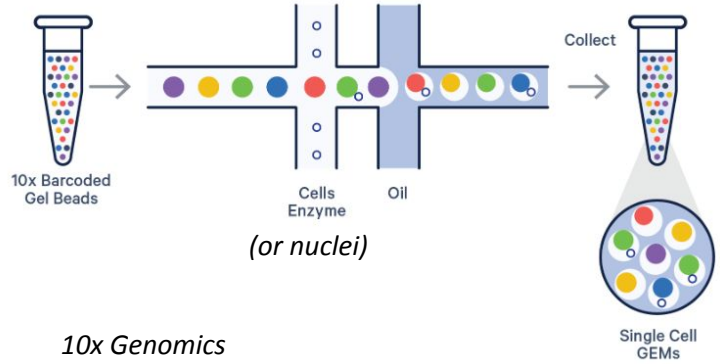
[include survey link]

CVRCBioinformatics <CVRCBioinformatics@nyulangone.org>

- Florencia: Florencia.schlamp@nyulangone.org
- Mike: Michael.gildea@nyulangone.org
- Sofie: Sofie.delbare@nyulangone.org
- Alex: Alexander.ferrena@nyulangone.org

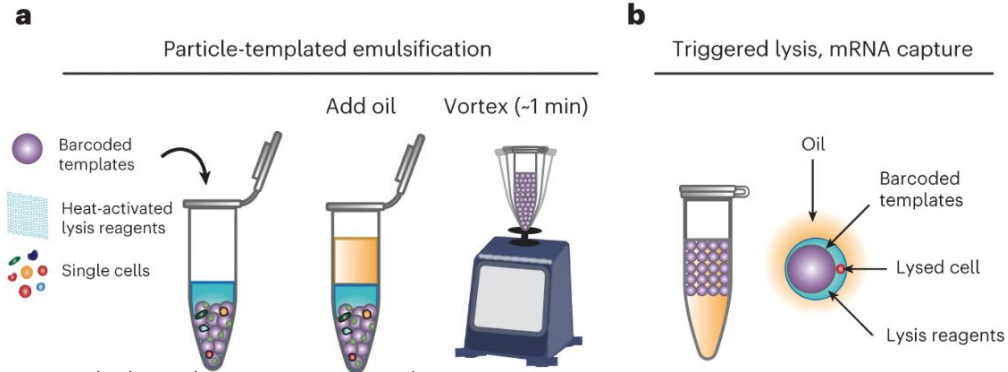
Extra slides

Single Cell data



(or nuclei)

10x Genomics

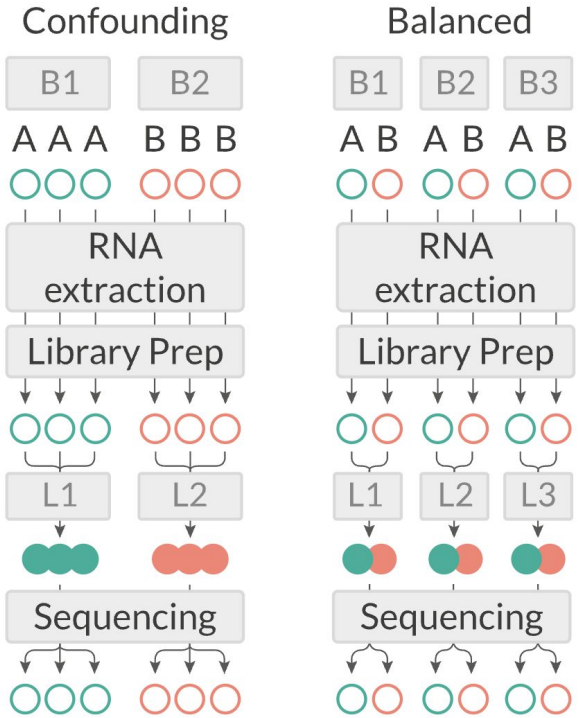


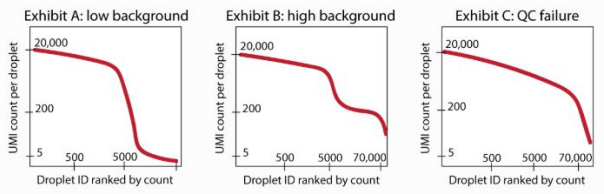
PIP-seq Clark et al. 2023 Nat. Biotechn.

Methods to assess single-cell transcriptome

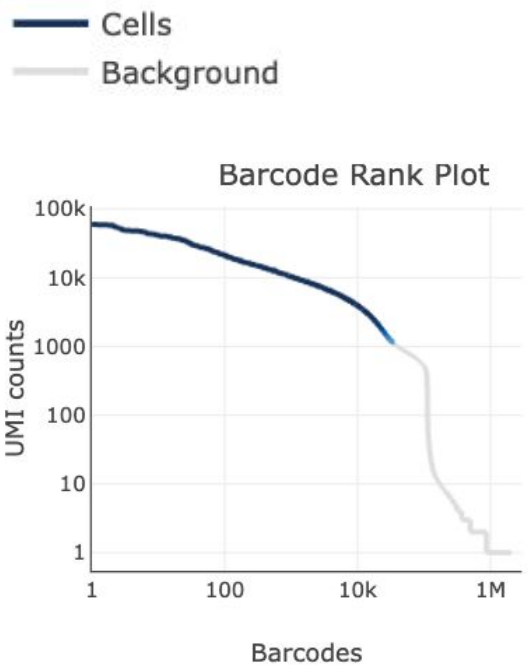
Sequenced output largely the same for all methods with some differences in upstream processing (e.g. ambient RNA removal for sn)

Choice of method depends on tissue/celltypes that need to be profiled (e.g. neurons or adipocytes don't do well with microfluidics, so nuclei isolation preferred; not sure how they work with pipseq?)


Confounding
B1
B2
A A A
B B B
RNA extraction
Library Prep
L1
L2
Sequencing
Balanced
B1
B2
B3
A B
A B
A B
RNA extraction
Library Prep
L1
L2
L3
Sequencing

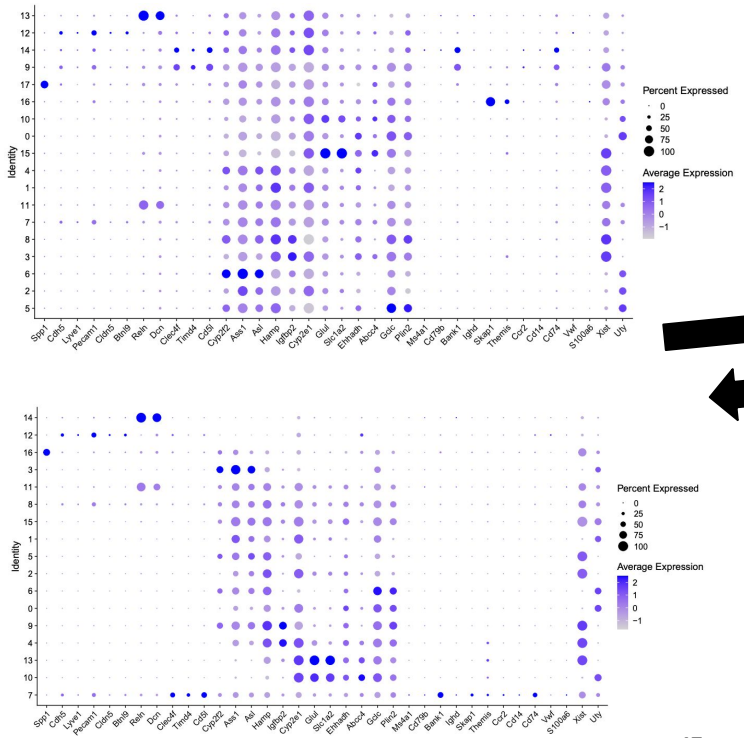


Data processing:
ell Ranger/PIP-seq output



Ambient RNA removal with cellbender

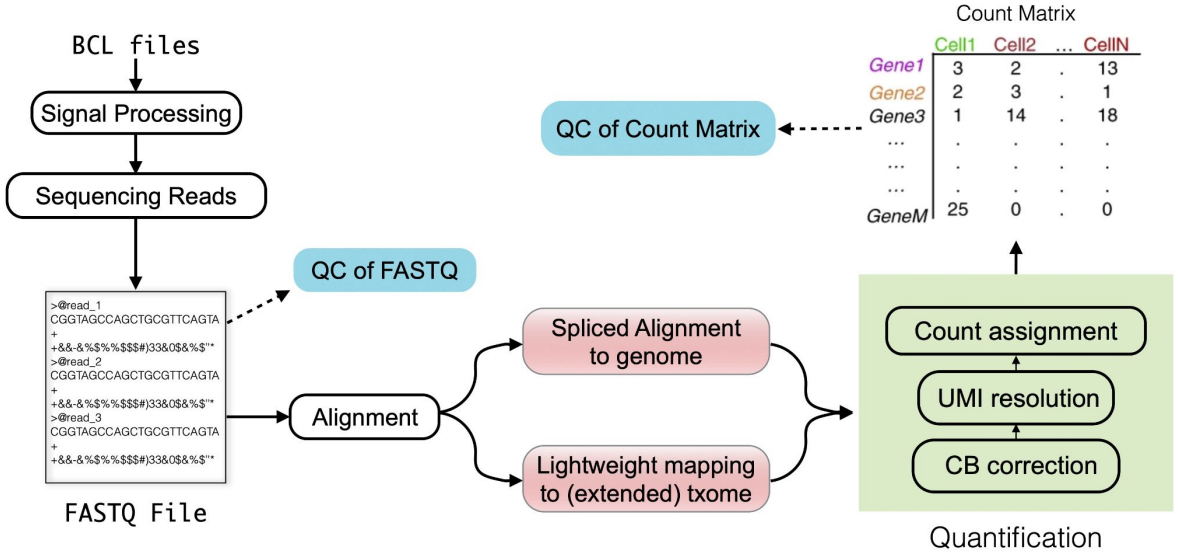
- Distinguishes empty droplets from droplets with cell/nucleus
- Removes ambient RNA from droplets with cell/nucleus



Fleming et al. 2023 Nat. Methods

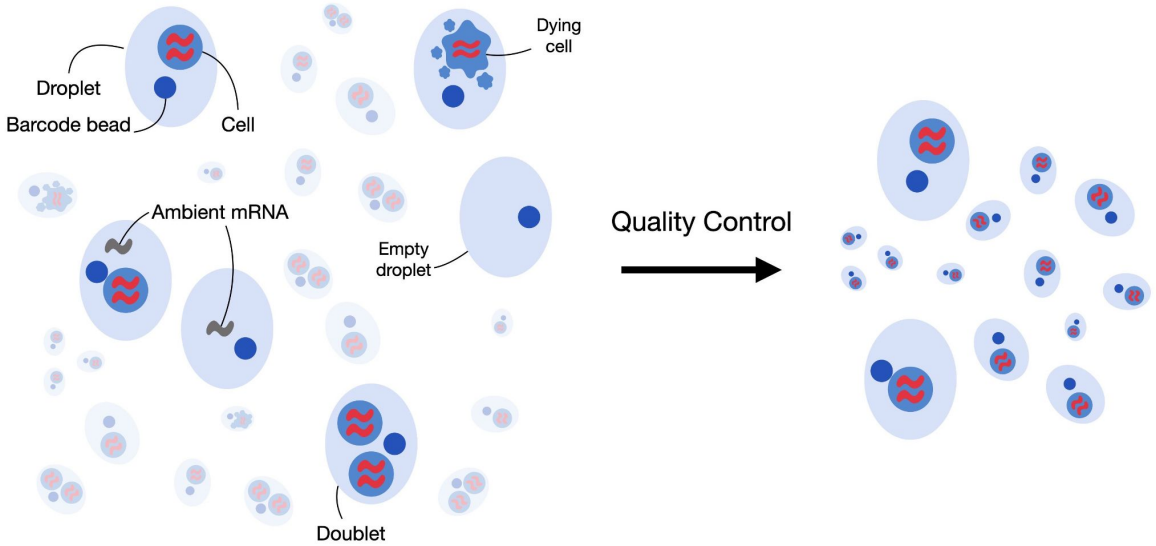
From raw base calls to cell-level read quantification going from sequencing reads to gene quantification involves some key bioinformatic steps including alignment and cell calling

We aim to generate a gene expression count matrix: table of genes by cells, where the values are number of transcripts



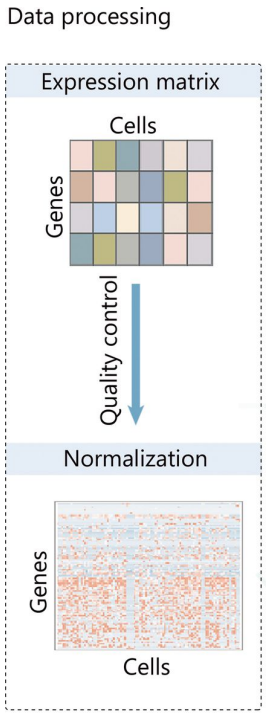
Heumos et al 2023; Chapter 3

Data processing: Cell-level Quality Control: identify and remove sources of noise, such as empty, leaky, dead, and multiplet “cells”, which can bias downstream results



Heumos et al 2023; Chapter 6

Clustering: classical machine learning methods are often used to group similar cells by shared transcriptomic patterns



Normalization: correct for sequencing bias, and place genes in similar numeric scale

1. **Selection of informative genes** which vary across cells (instead of constitutively expressed)

2. **Principal Component Analysis (PCA), detect "metagenes"** of co-expressed genes (genes can be highly multicollinear)

3. **Nearest neighbor graph** methods allow study of non-linear relationships between cells

4. **Graph-based clustering** (community detection) to group together cells based on nearest neighbors

[Su et al MMR 2022]