GHGA THE GERMAN HUMAN GENOME-PHENOME ARCHIVE

A beginner's guide to scRNAseq analysis

Florian Heyl

In cooperation with

nfdi

(1) What is single cell sequencing?

(2) How can I use single cell sequencing data?

(3) What tools can I use?

GHGA THE GERMAN HUMAN GENOME-PHENOME ARCHIVE

Questions

(1) **Laboratory:**

- What is the difference between bulk and single cell sequencing?
- How do you perform a single cell experiment?
- What confounders do I have?

(2) **Bioinformatics:**

- What data do I get?
- What should I check for my single cell data?
- What can I do with my data?

(3) **Products:**

- Why should I automatize my data analysis?
- What is important for a workflow?
- What open challenges do we still have?

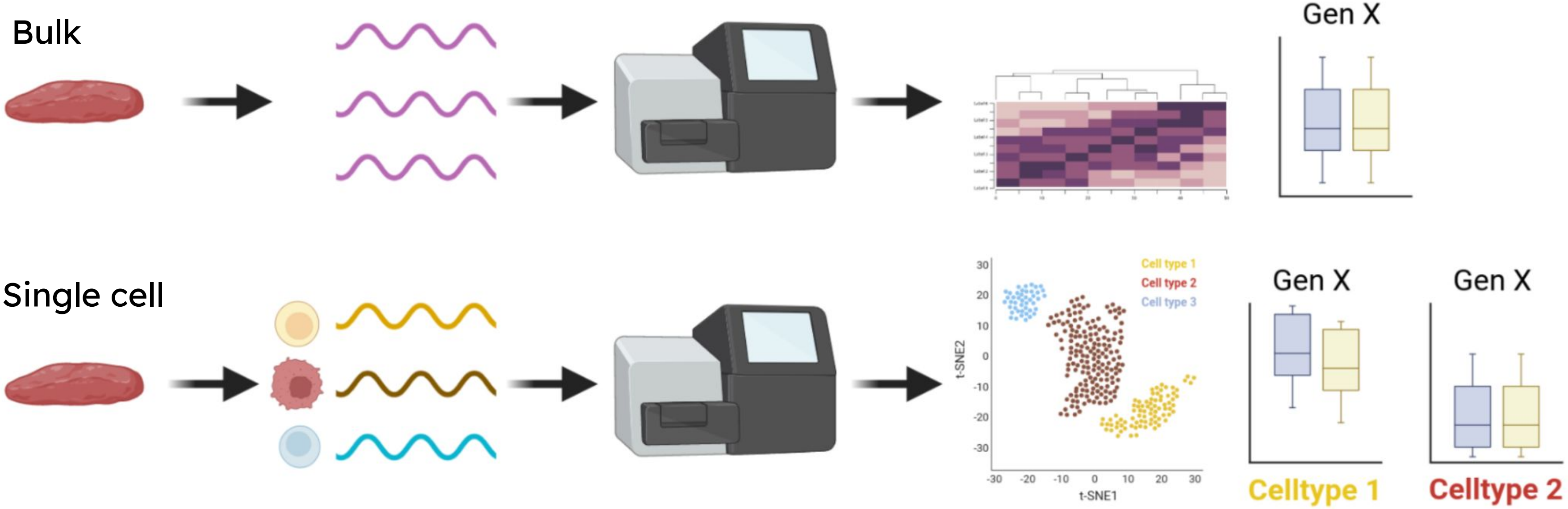
GHGA THE GERMAN HUMAN GENOME-PHENOME ARCHIVE

Slides will be available

Take home message

Literature

(1) What is single cell sequencing?

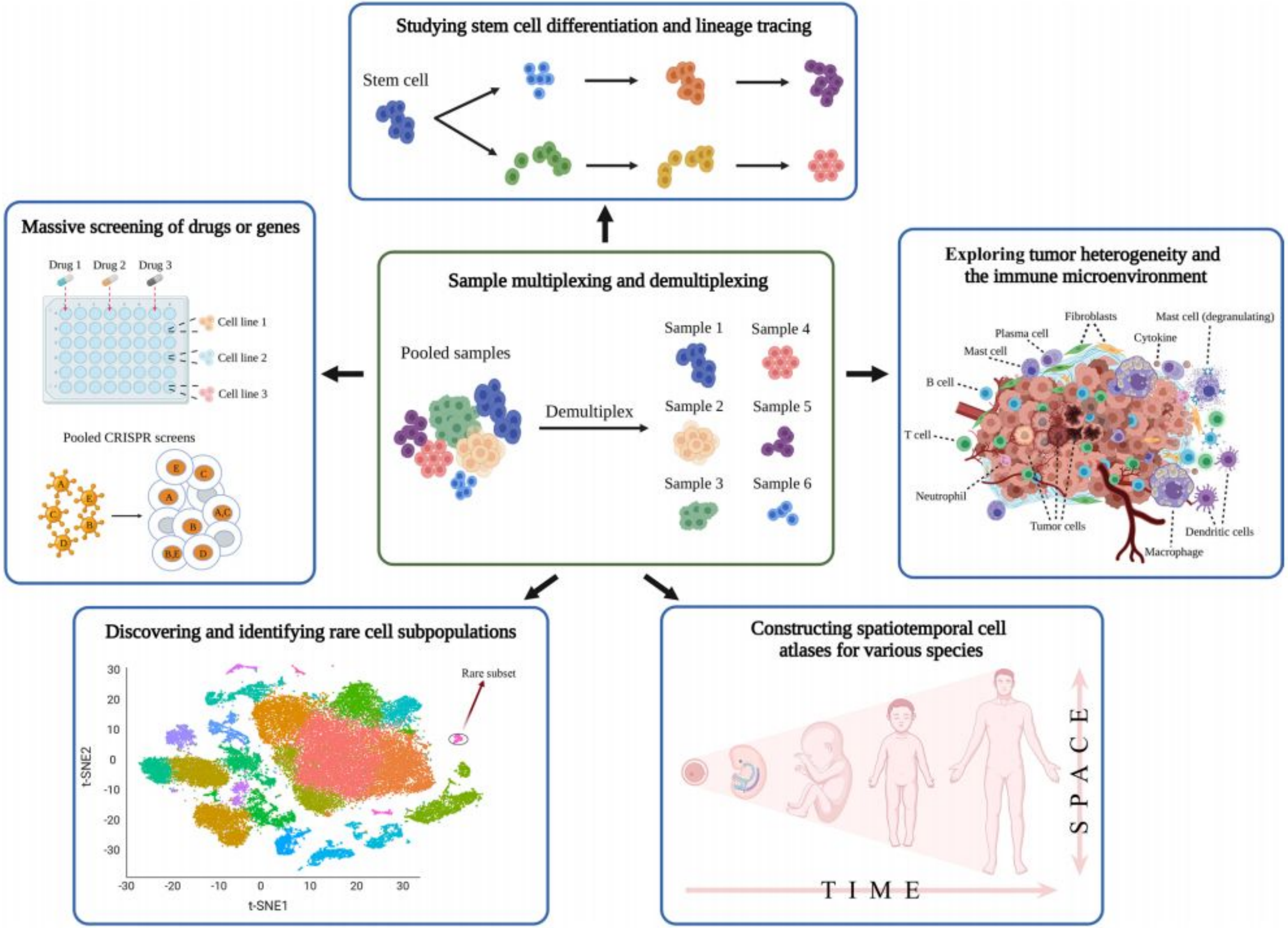


Bulk

Single cell

Bulk seq.: Average-based expression profile
Single cell seq.: Cell level expression profile

Created in **BioRender.com** bio



Studying stem cell differentiation and lineage tracing

Massive screening of drugs or genes

Sample multiplexing and demultiplexing

Exploring tumor heterogeneity and the immune microenvironment

Discovering and identifying rare cell subpopulations

Constructing spatiotemporal cell atlases for various species

Yulong Zhang et al. (2022)

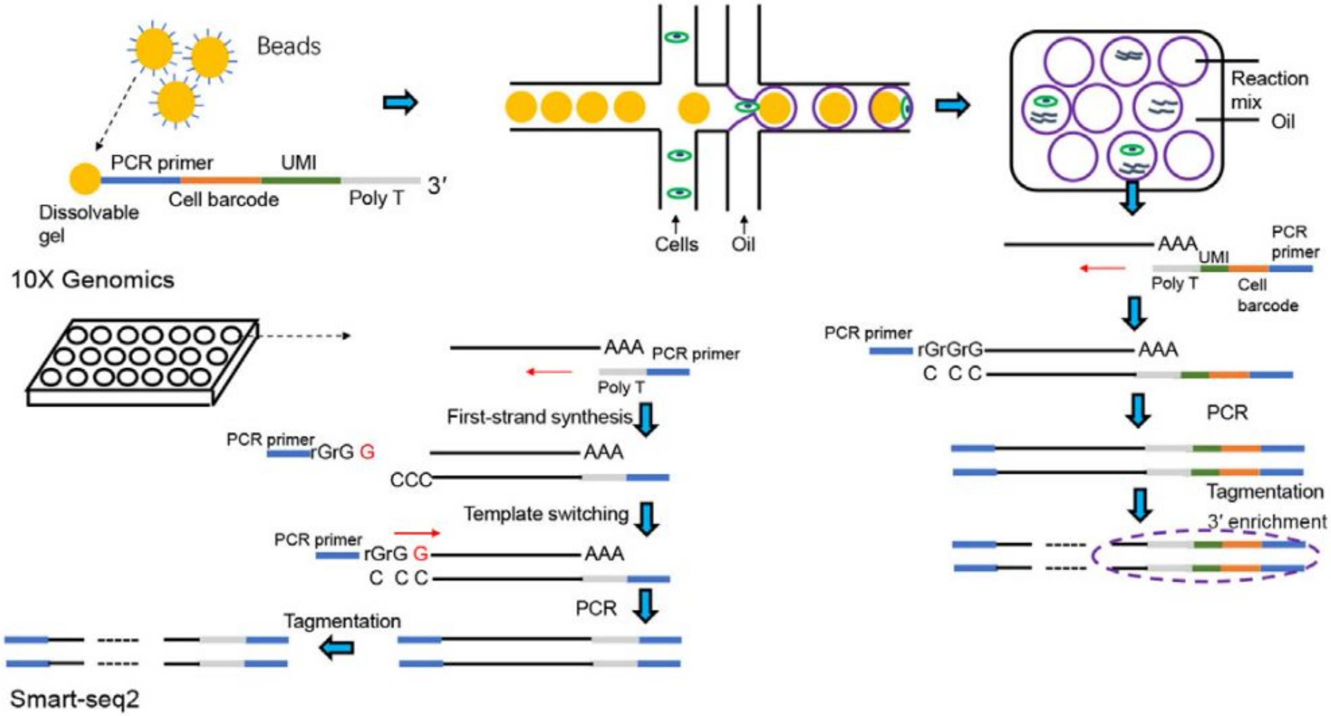
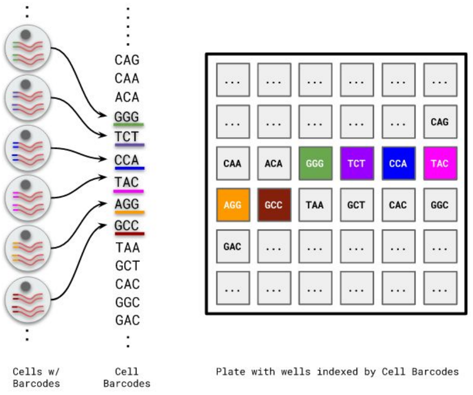
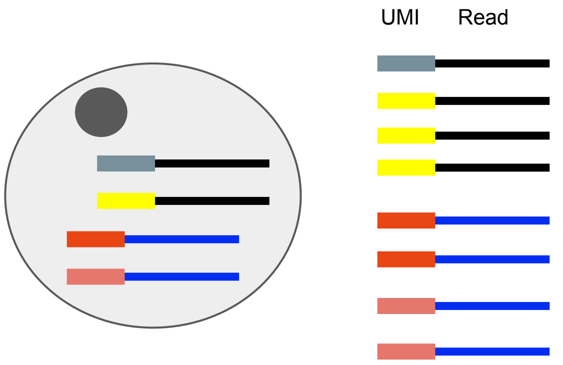
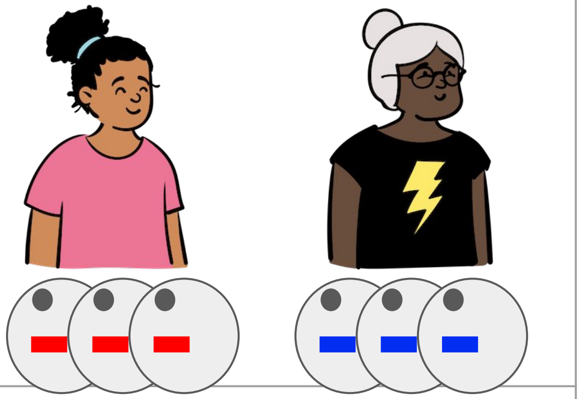


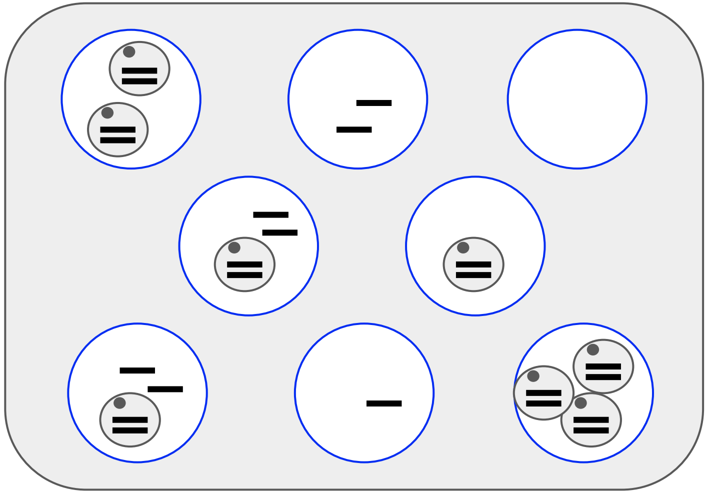
Plate-based: cells into wells on a plate
Droplet-based: each cell in its own microfluidic droplet
Each cell is a sample which cannot be replicated.

Jeanette Baran-Gale et al. (2018)
Malte D. Luecken & Fabian J. Theis (2019)
Xiliang Wang et al. (2021)

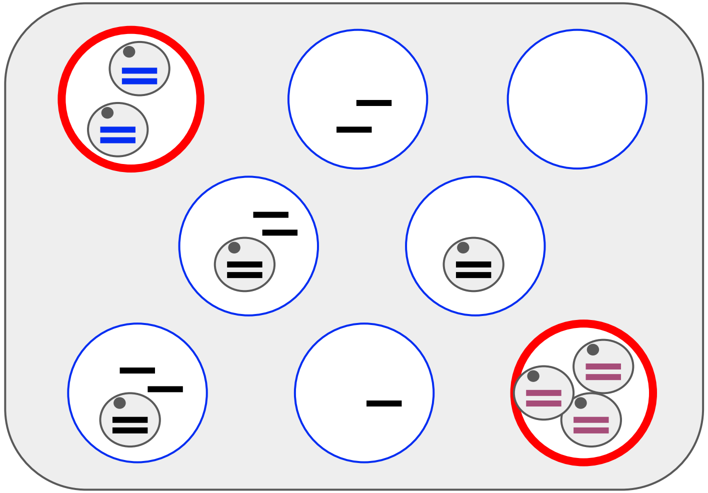
Cell Barcodes	Unique Molecular Identifier (UMI)	“Donor” (Multiplex) Barcode
 CAG CAA ACA GGG TCT CCA TAC AGG GCC TAA GCT CAC GGC GAC CAG CAA ACA GGG TCT CCA TAC AGG GCC TAA GCT CAC GGC GAC Cells w/ Barcodes Cell Barcodes Plate with wells indexed by Cell Barcodes		
To connect read (e.g., RNA) to a cell.	To reduce amplification bias (keep unique reads). In example: each gene (black & blue) has just two reads.	To connect read (e.g., RNA) to a donor (e.g., patient).

Mehmet Tekman / Galaxy Training material (An introduction to scRNA-seq data analysis)
Eric Vallabh Minikel (2012) How PCR duplicates arise
Yulong Zhang et al. (2022)

Confounders



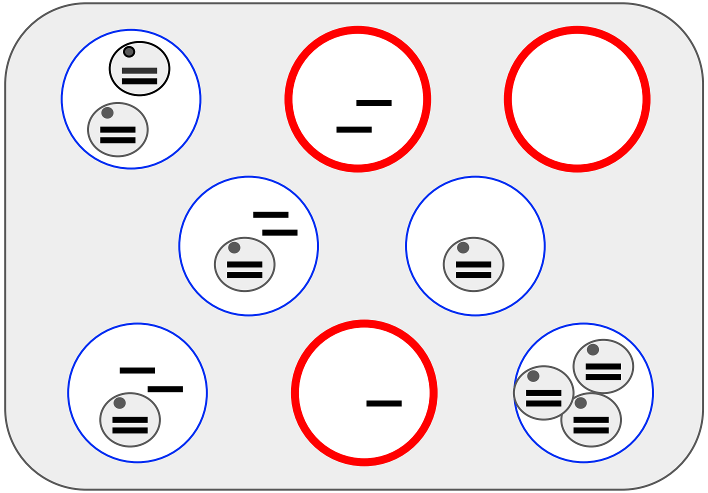
Problem 1: Doublets/Multiplets



Two or more cells are sequenced together
A high (e.g., RNA) count or number of detected regions is the result.

Malte D. Luecken & Fabian J. Theis (2019)
Samuel L. Wolock et al. (2019)
Tallulah S. Andrews et al. (2021)

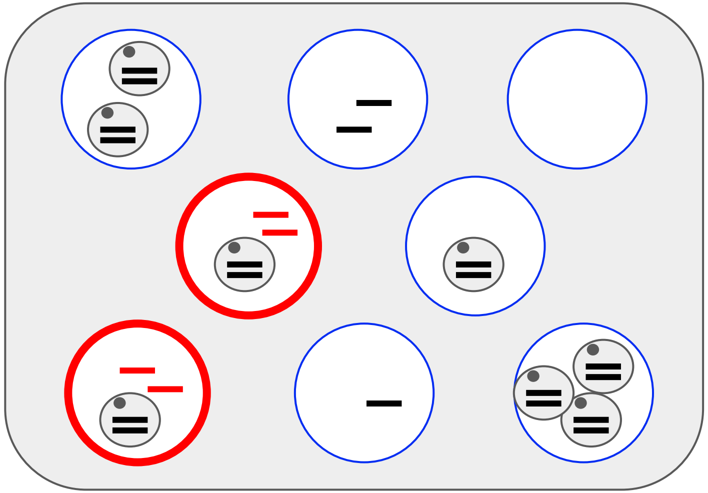
Problem 2: Empty droplets/wells



“Broken” cells or no cell can be collected
A low (RNA) count, few detected genes, and a high fraction of mitochondrial counts can be the result.

Malte D. Luecken & Fabian J. Theis (2019)
Aaron T. L. Lun et al. (2019)
Tallulah S. Andrews et al. (2021)

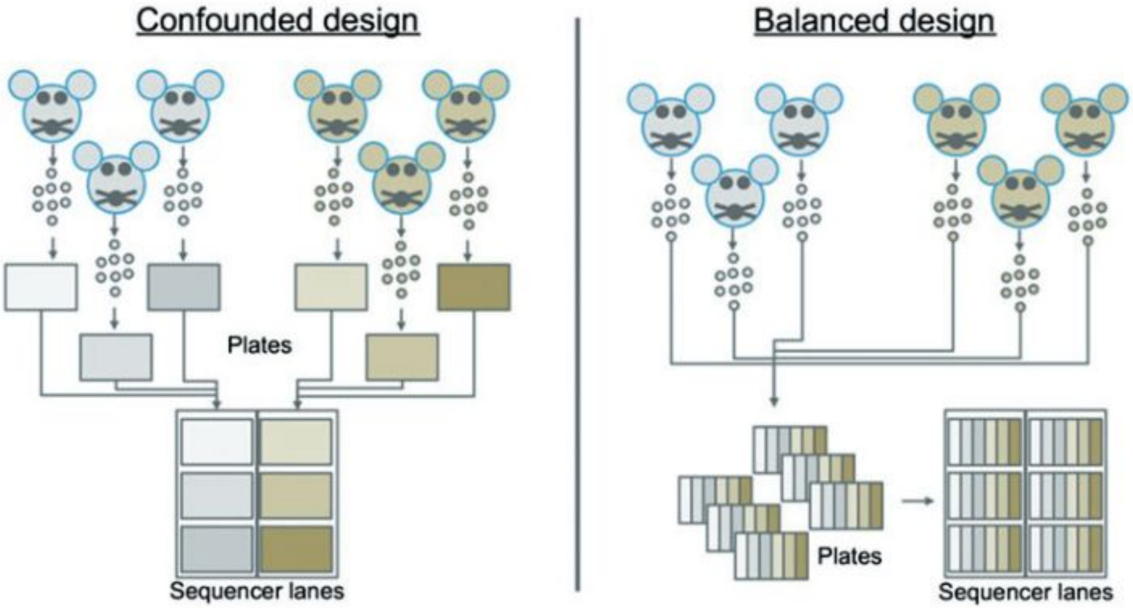
Problem 3: Ambient RNA



Counts that do not originate from the true barcoded cell, but from other lysed cells.
Can lead to overrepresentation of some cell clusters (spurious clusters), higher cluster overlaps, higher gene coverage.

Malte D. Luecken & Fabian J. Theis (2019)
Shiyi Yang et al. (2020)
Tallulah S. Andrews et al. (2021)
Emre Caglayan et al. (2022)
Stephen J. Fleming et al. (2022)

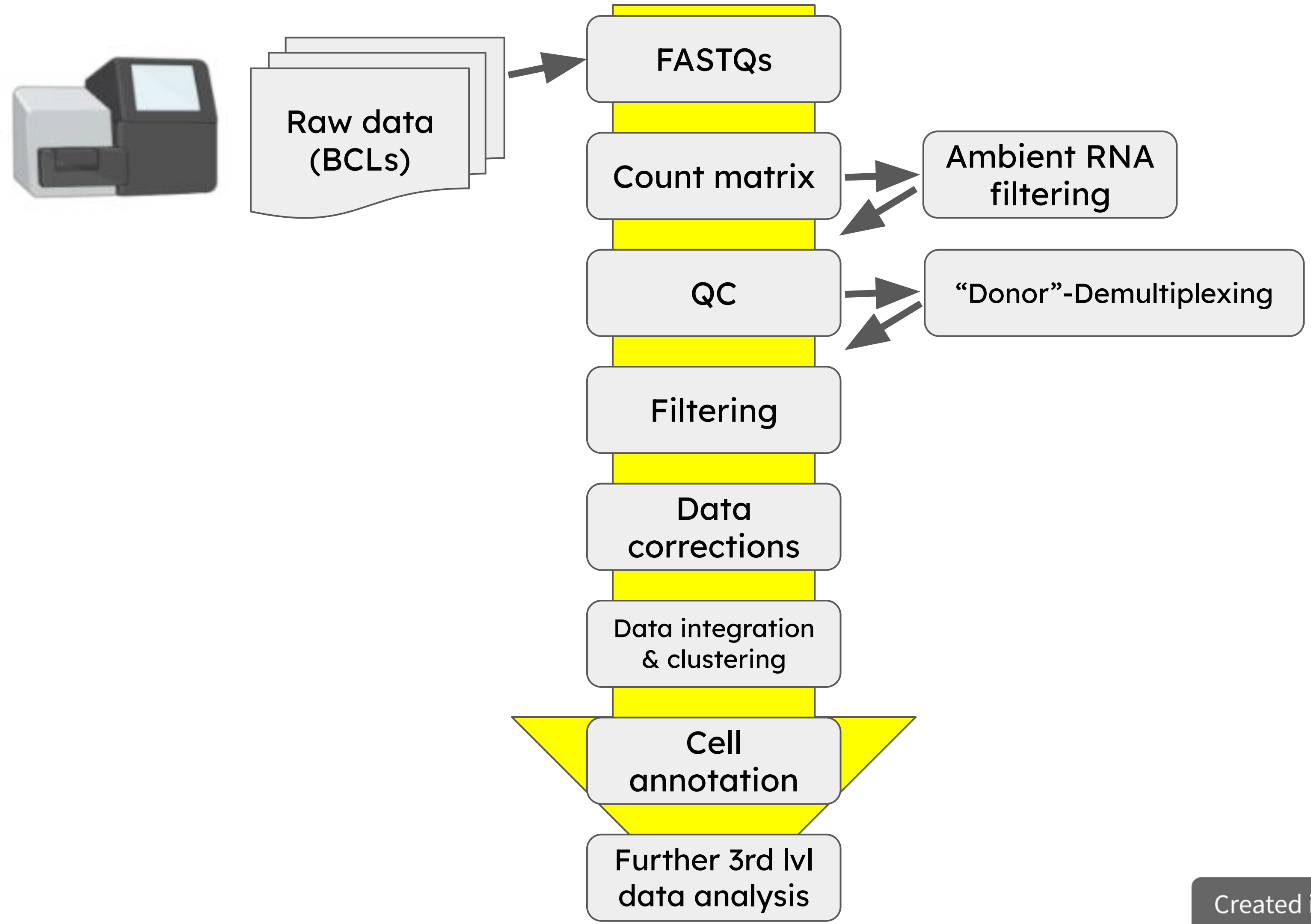
Problem 4: Batch effects



Sequencing your experiments in batches
Might lead to spurious results (e.g., clusters or correlations)

Jeanette Baran-Gale et al. (2018)
Malte D. Luecken & Fabian J. Theis (2019)
Tallulah S. Andrews et al. (2021)

(2) How can I use single cell sequencing data?



Raw data (BCLs)

FASTQs

Count matrix

Ambient RNA filtering

QC

“Donor”-Demultiplexing

Filtering

Data corrections

Data integration & clustering

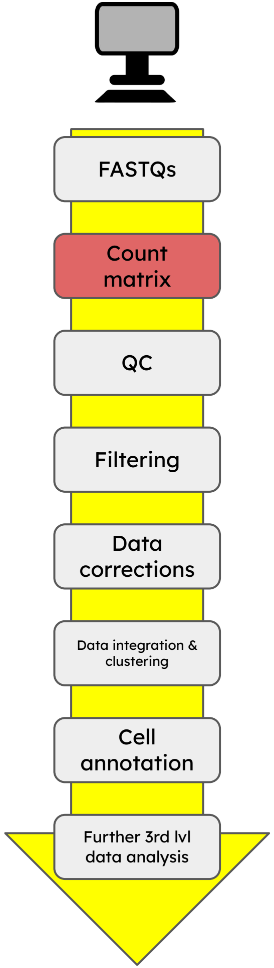
Cell annotation

Further 3rd lvl data analysis

Created in BioRender.com bio

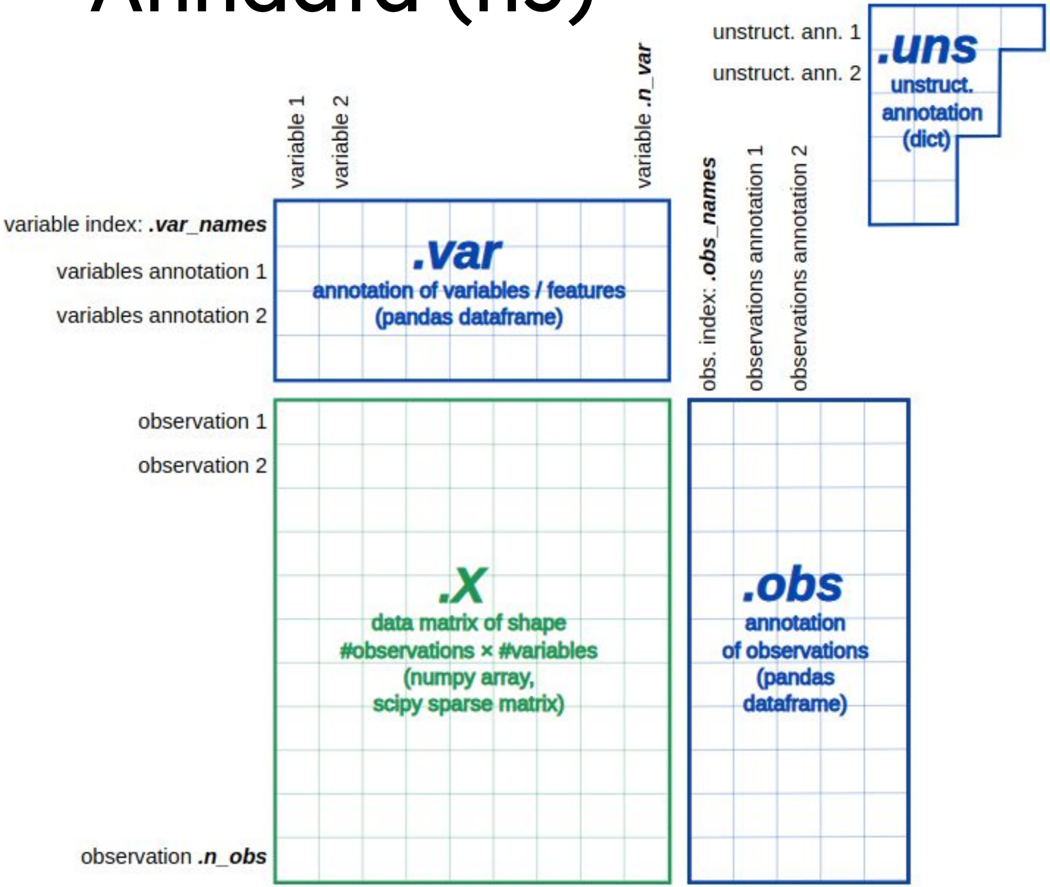
Cellranger

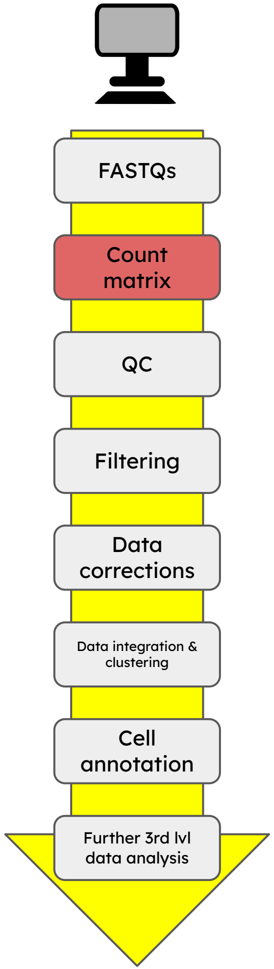
1. Read trimming
2. Genome/Transcriptome alignment
3. MAPQ adjustment
4. 10x barcode correction
5. UMI counting
6. Calling cell barcodes



10x Genomics (Gene Expression Algorithms Overview)
Ralf Schulze Brüning et al. (2022)

Anndata (h5)


unstruct. ann. 1
unstruct. ann. 2
.uns
unstruct. annotation (dict)
variable 1
variable 2
variable .n_var
obs. index: .obs_names
observations annotation 1
observations annotation 2
variable index: .var_names
variables annotation 1
variables annotation 2
.var
annotation of variables / features (pandas dataframe)
observation 1
observation 2
observation .n_obs
.X
data matrix of shape #observations × #variables (numpy array, scipy sparse matrix)
.obs
annotation of observations (pandas dataframe)



Bérénice Batut et al. / Galaxy Training material (Clustering 3K PBMCs with Scanpy)

Anndata (h5)

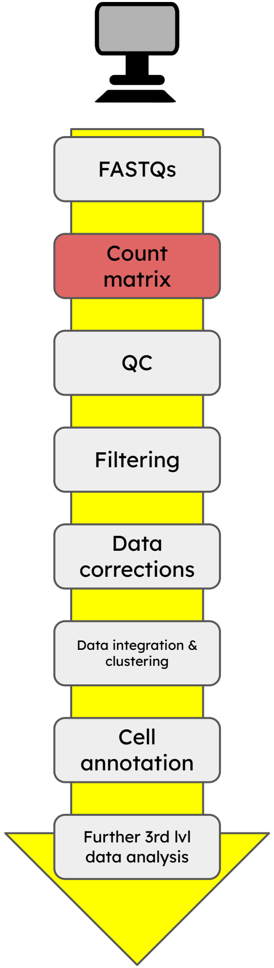
- Cell labels
- QC
- Sample IDs

Gene

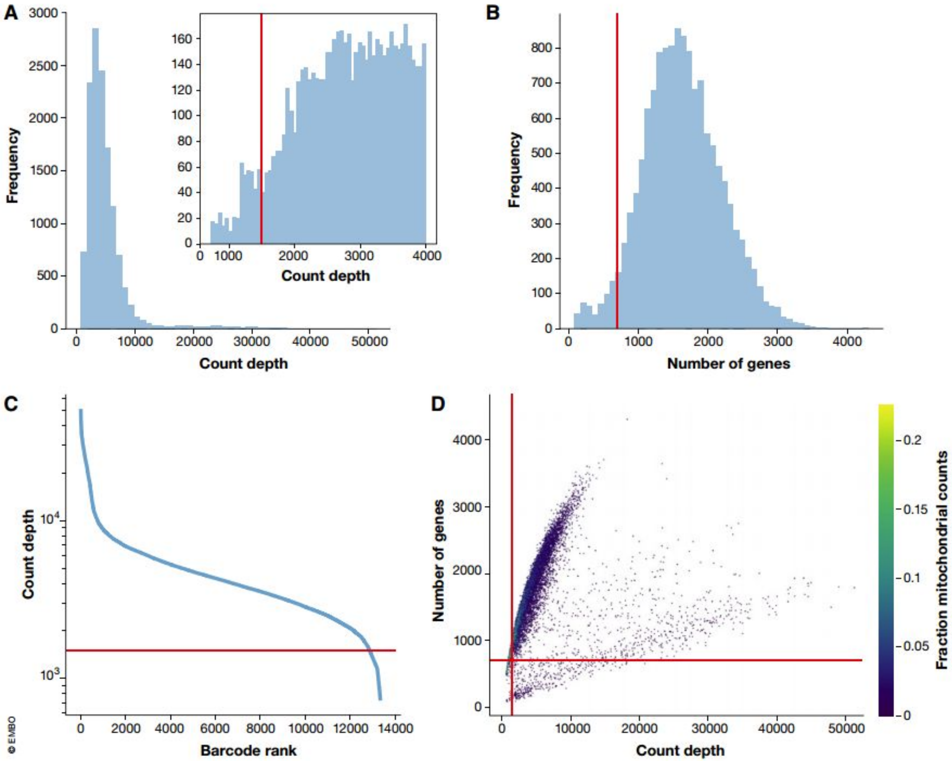
A	B	C	D	E
1	0	0	5	0
0	2	0	4	4
1	1	0	0	0
0	0	3	0	3
0	0	0	2	0

Cell

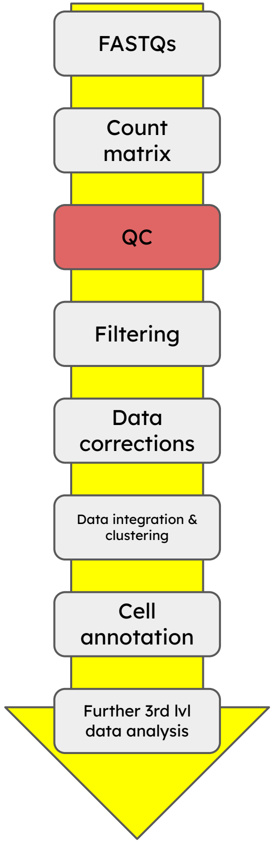
Bérénice Batut et al. / Galaxy Training material (Clustering 3K PBMCs with Scanpy)



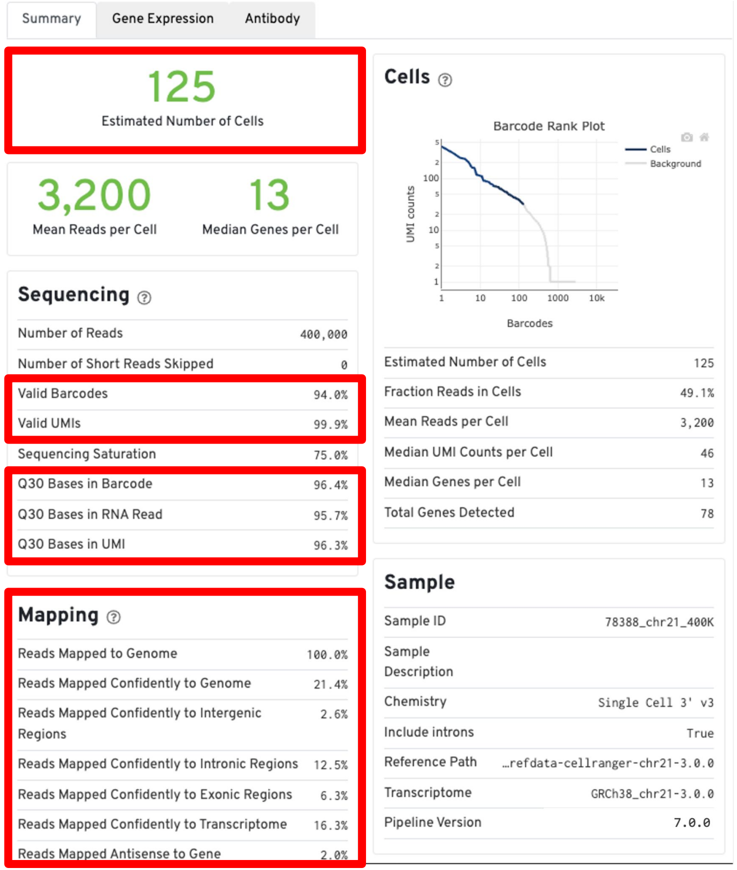
Quality control of cells / barcodes



Malte D. Luecken & Fabian J. Theis (2019)
Tallulah S. Andrews et al. (2021)
Rui Hong et al. (2022)



Quality control of experiment



Summary | Gene Expression | Antibody

125
Estimated Number of Cells

3,200
Mean Reads per Cell

13
Median Genes per Cell

Sequencing

Number of Reads	400,000
Number of Short Reads Skipped	0
Valid Barcodes	94.0%
Valid UMIs	99.9%
Sequencing Saturation	75.0%
Q30 Bases in Barcode	96.4%
Q30 Bases in RNA Read	95.7%
Q30 Bases in UMI	96.3%

Mapping

Reads Mapped to Genome	100.0%
Reads Mapped Confidently to Genome	21.4%
Reads Mapped Confidently to Intergenic Regions	2.6%
Reads Mapped Confidently to Intronic Regions	12.5%
Reads Mapped Confidently to Exonic Regions	6.3%
Reads Mapped Confidently to Transcriptome	16.3%
Reads Mapped Antisense to Gene	2.0%

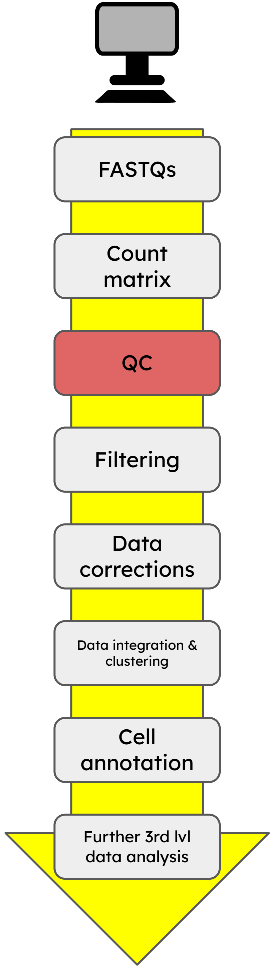
Cells

Estimated Number of Cells	125
Fraction Reads in Cells	49.1%
Mean Reads per Cell	3,200
Median UMI Counts per Cell	46
Median Genes per Cell	13
Total Genes Detected	78

Sample

Sample ID	78388_chr21_400K
Sample Description	
Chemistry	Single Cell 3' v3
Include introns	True
Reference Path	…refdata-cellranger-chr21-3.0.0
Transcriptome	GRCh38_chr21-3.0.0
Pipeline Version	7.0.0

10x Genomics (Web summary)



FASTQs

Count matrix

QC

Filtering

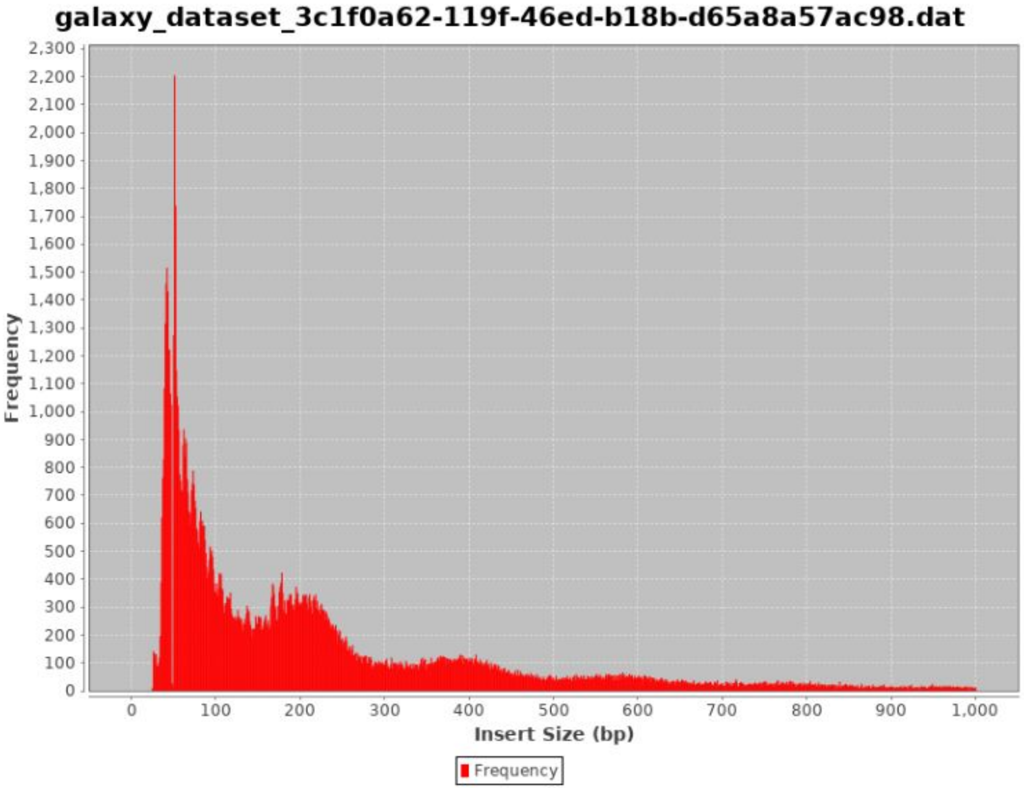
Data corrections

Data integration & clustering

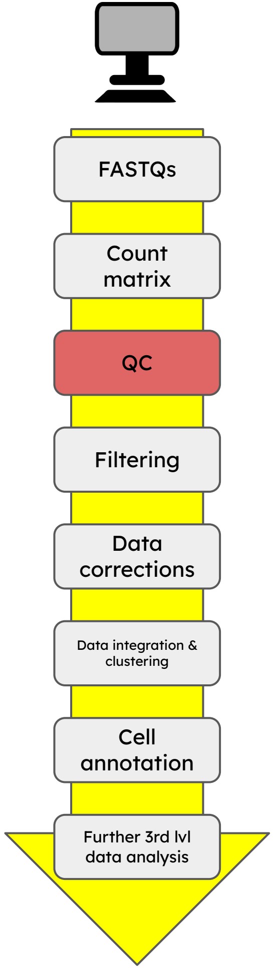
Cell annotation

Further 3rd lvl data analysis

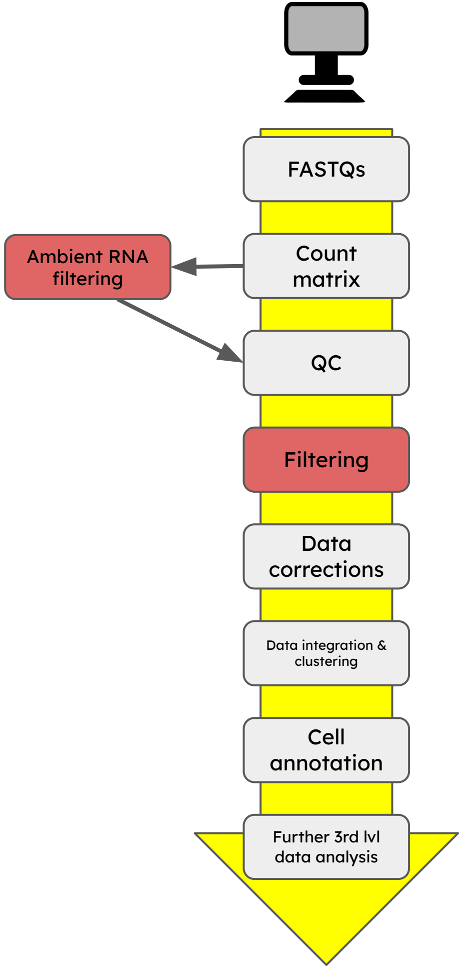
Quality control of experiment



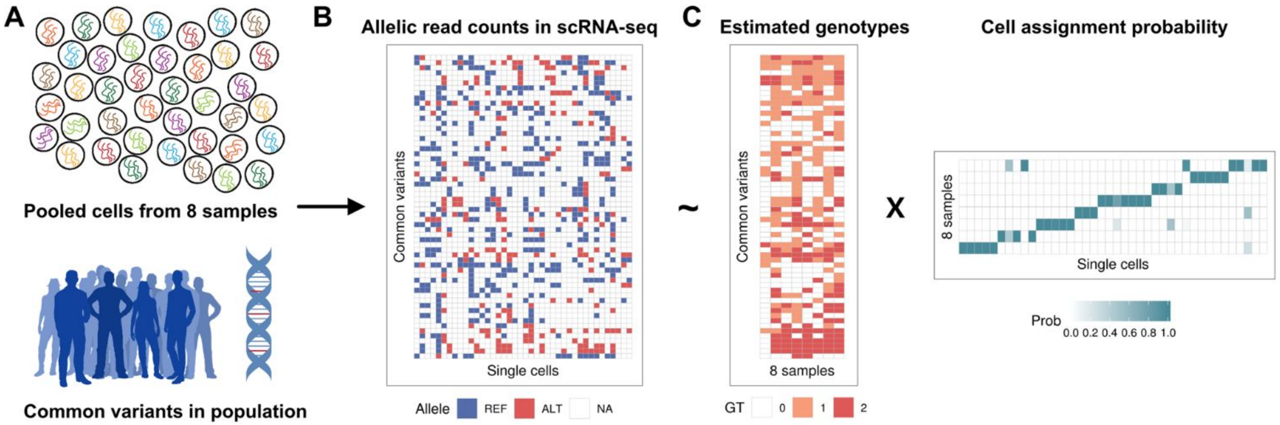
Lucille Delisle et al. / Galaxy (ATAC-Seq data analysis)

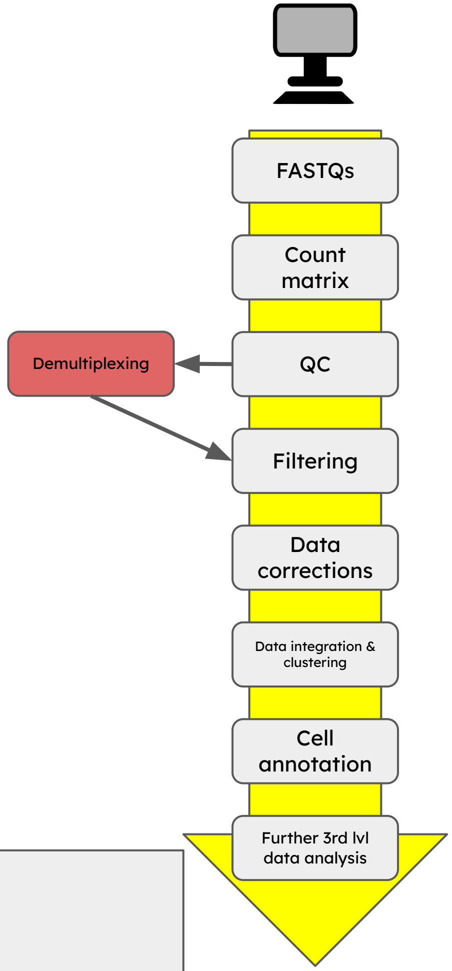


1. Emptyplets/Douplets/Multiplets
2. Ambient RNA
3. Experimental important QC measures (e.g., fragment sizes scATAC)



“Donor”-Demultiplexing

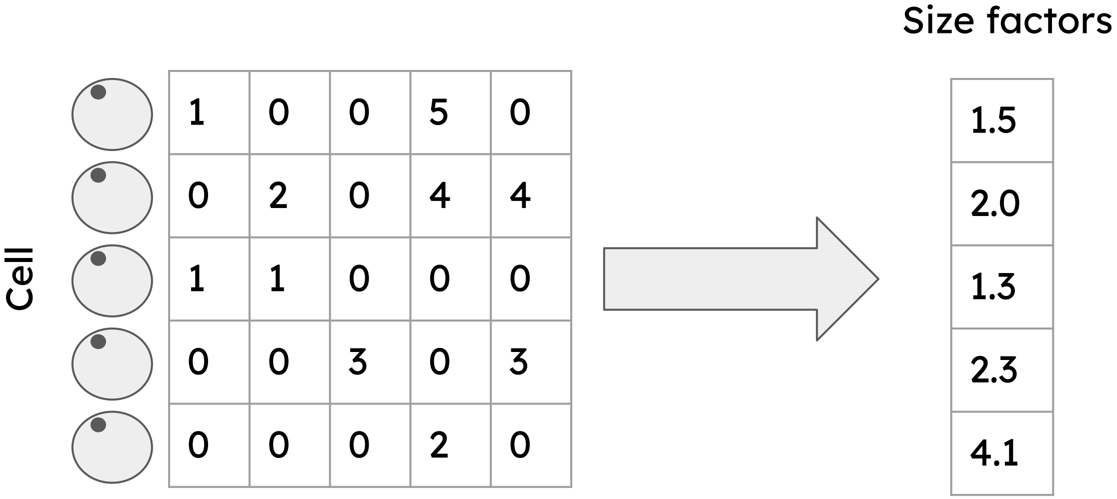




Demultiplexing approach depends on experimental design.

Yuanhua Huang et al. (2019)
Xianjie Huang and Yuanhua Huang (2021)
Yulong Zhang et al. (2022)
Joseph F. Cardiello et al. (2022)

Normalization

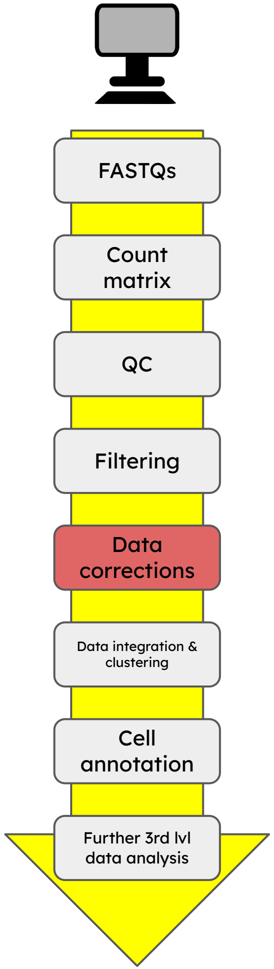


1	0	0	5	0
0	2	0	4	4
1	1	0	0	0
0	0	3	0	3
0	0	0	2	0

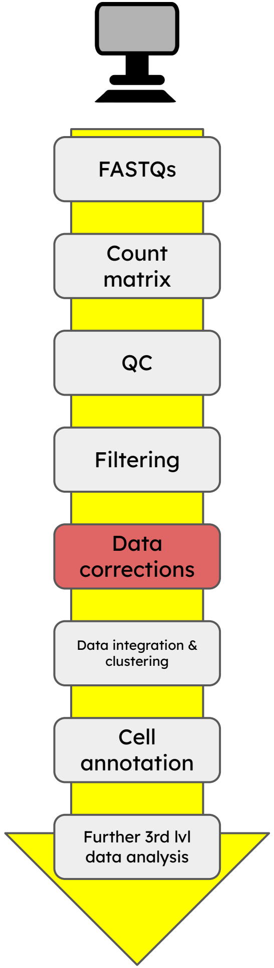
Size factors
1.5
2.0
1.3
2.3
4.1

Malte D. Luecken & Fabian J. Theis (2019)
Nicholas Lytal et al. (2020)
Tallulah S. Andrews et al. (2021)

Brenda Marquina-Sanchez et al. (2020)
Xin Wang et al. (2021)



Spike-In Normalization


FASTQs

Count matrix

QC

Filtering

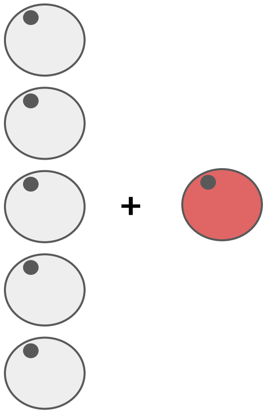
Data corrections

Data integration & clustering

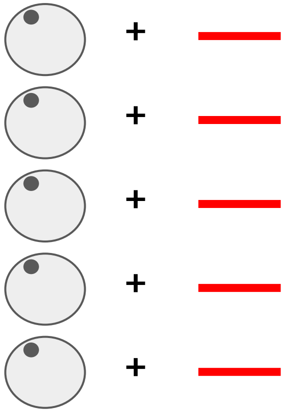
Cell annotation

Further 3rd lvl data analysis

Reference Cell



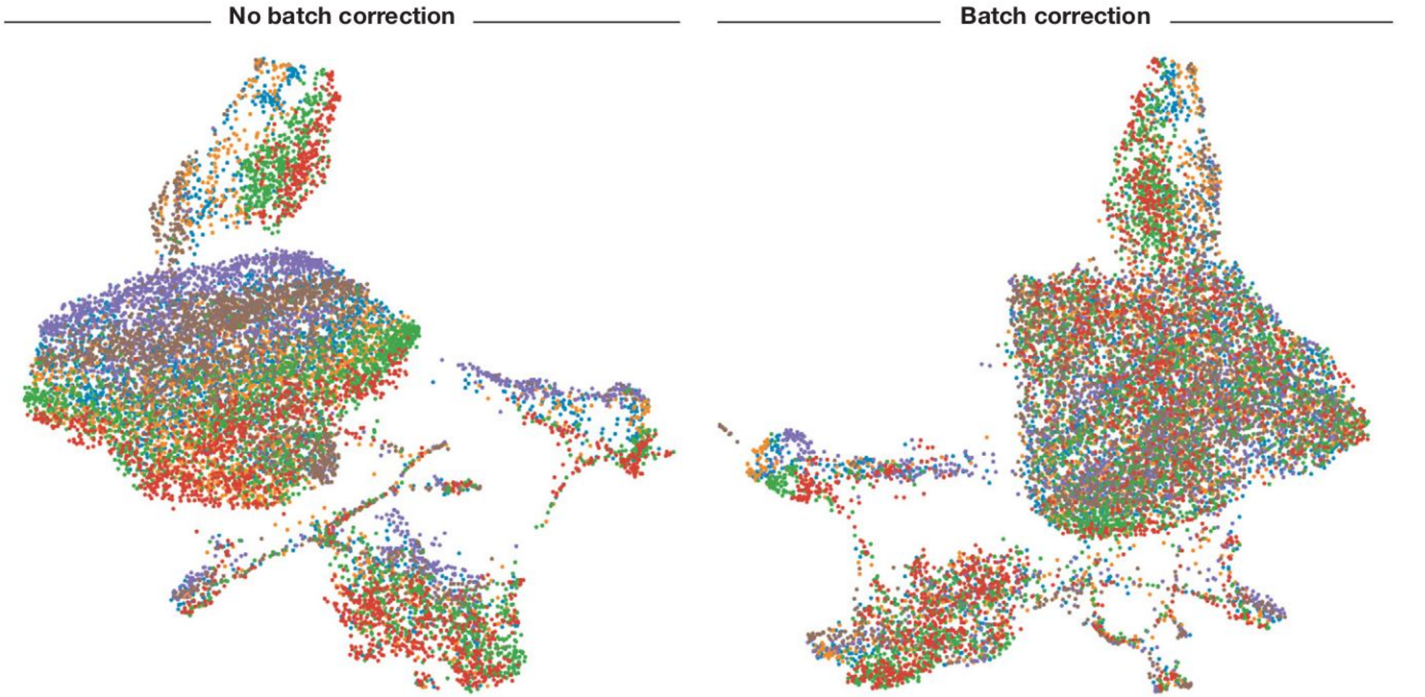
Reference RNA

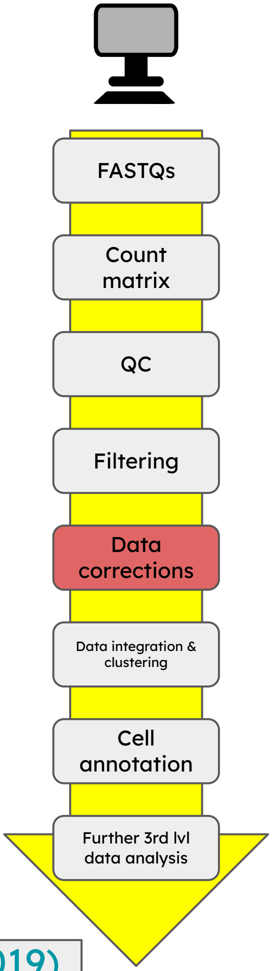


Malte D. Luecken & Fabian J. Theis (2019)
Nicholas Lytal et al. (2020)
Tallulah S. Andrews et al. (2021)

Brenda Marquina-Sanchez et al. (2020)
Xin Wang et al. (2021)

Batch correction




FASTQs

Count matrix

QC

Filtering

Data corrections

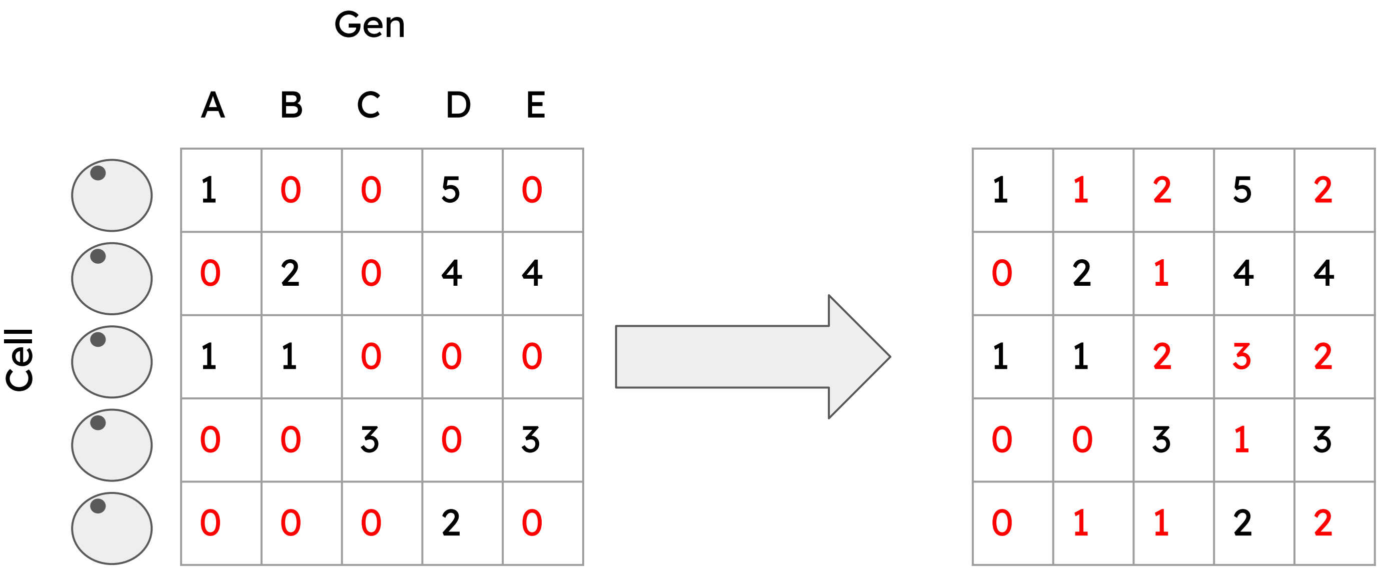
Data integration & clustering

Cell annotation

Further 3rd lvl data analysis

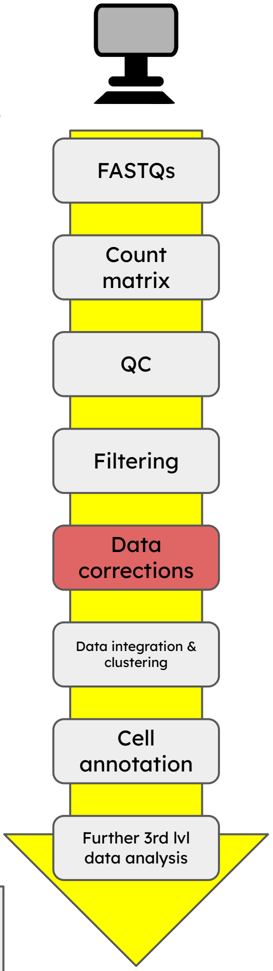
Malte D. Luecken & Fabian J. Theis (2019)
Tallulah S. Andrews et al. (2021)

Imputation and smoothing ! Caution !



A	B	C	D	E
1	0	0	5	0
0	2	0	4	4
1	1	0	0	0
0	0	3	0	3
0	0	0	2	0

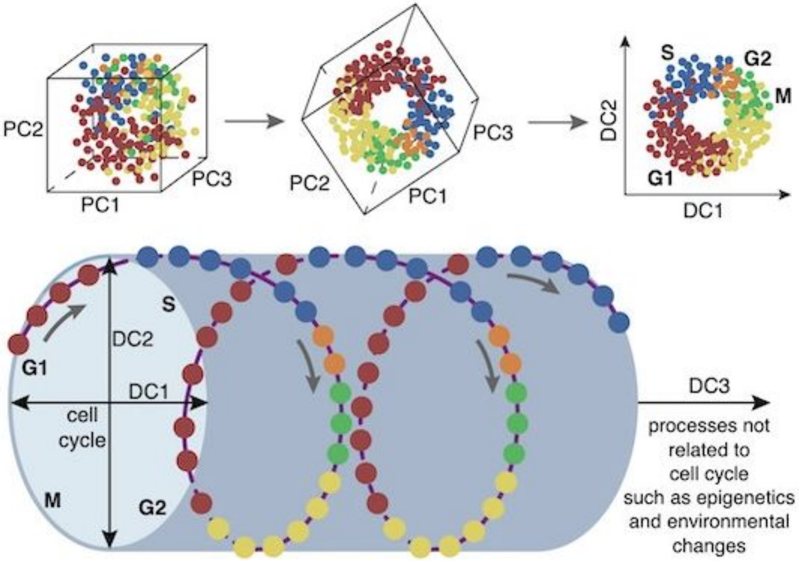
1	1	2	5	2
0	2	1	4	4
1	1	2	3	2
0	0	3	1	3
0	1	1	2	2



Zeros result from either (a) true expression or (b) technical variance.

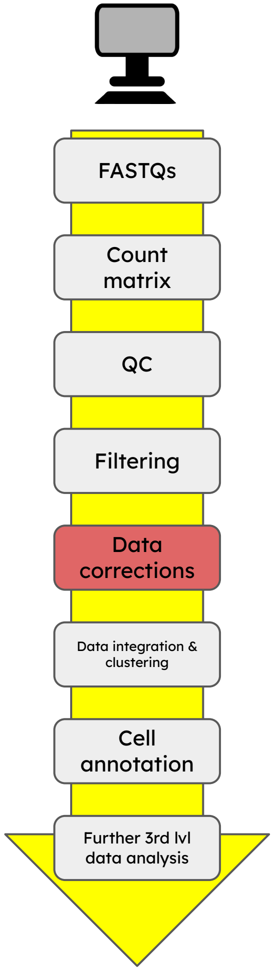
Malte D. Luecken & Fabian J. Theis (2019)
Wenpin Hou et al. (2020)
Tallulah S. Andrews et al. (2021)

Cell cycle removal ! Caution !



processes not related to cell cycle such as epigenetics and environmental changes

Malte D. Luecken & Fabian J. Theis (2019)
Daniel Schwabe et al. (2020)
Tallulah S. Andrews et al. (2021)
Jiajia Liu et al. (2021)



FASTQs

Count matrix

QC

Filtering

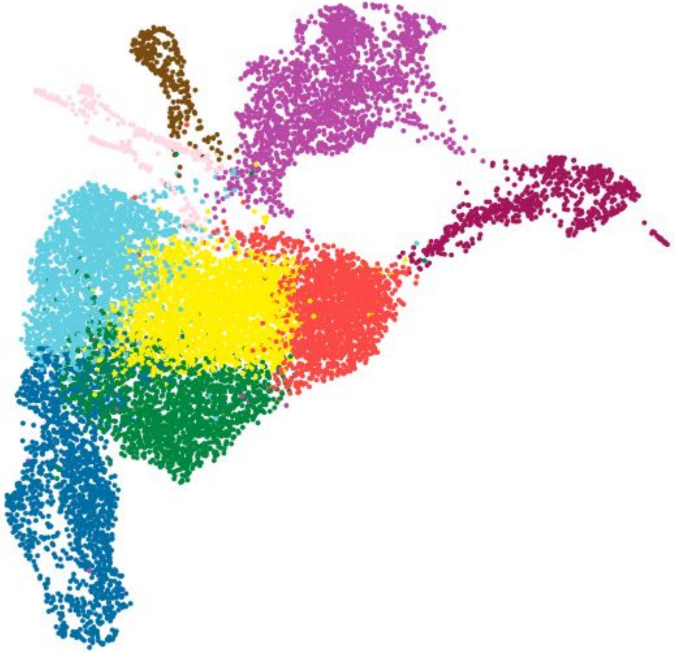
Data corrections

Data integration & clustering

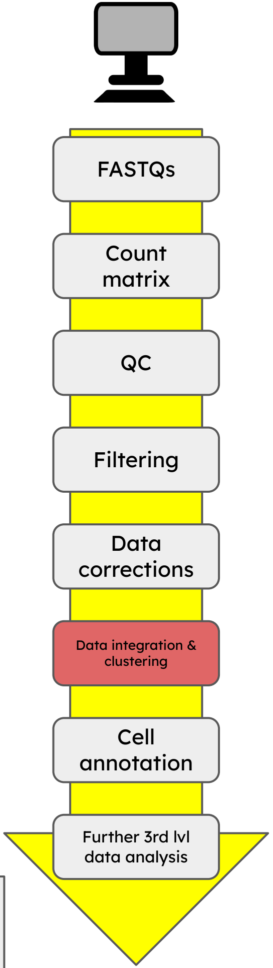
Cell annotation

Further 3rd lvl data analysis

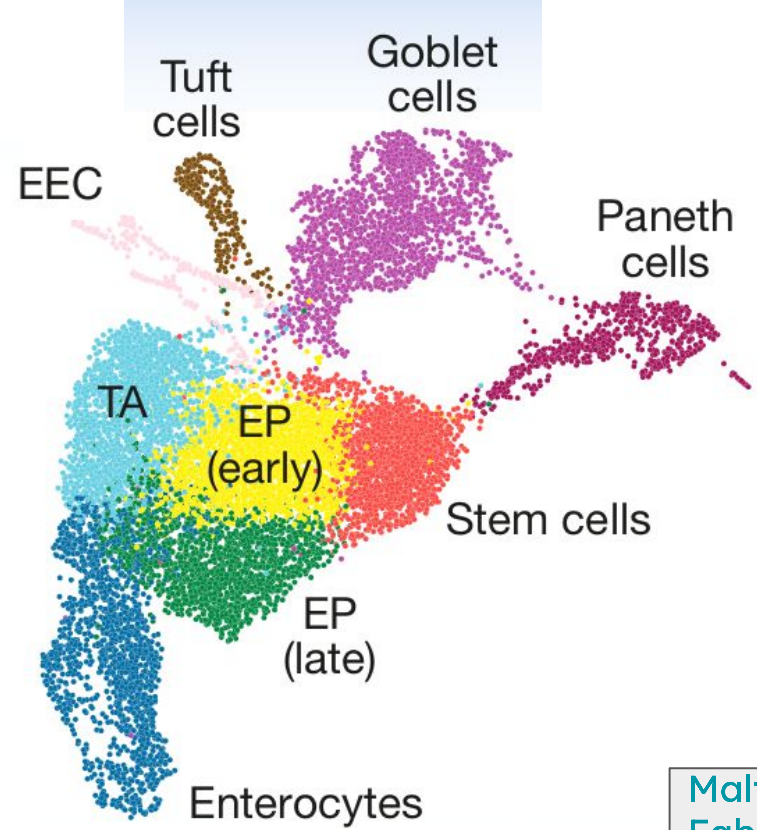
Clustering



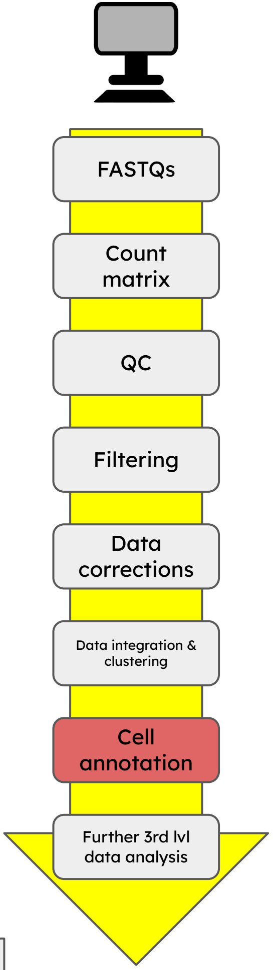
Malte D. Luecken & Fabian J. Theis (2019)
Ren Qi et al. (2020)
Tallulah S. Andrews et al. (2021)



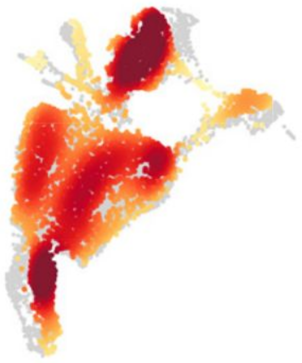
Cell annotation



Malte D. Luecken & Fabian J. Theis (2019)

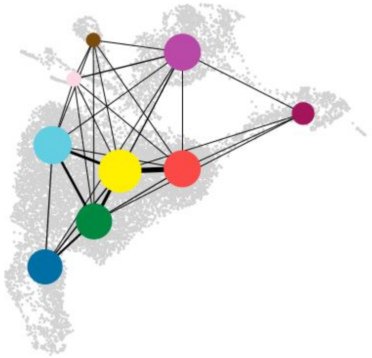


Compositional Analysis



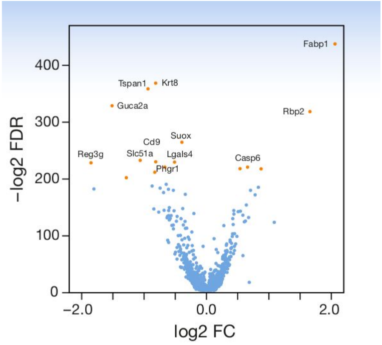
Sean Simmons (2022)

Trajectory Analysis



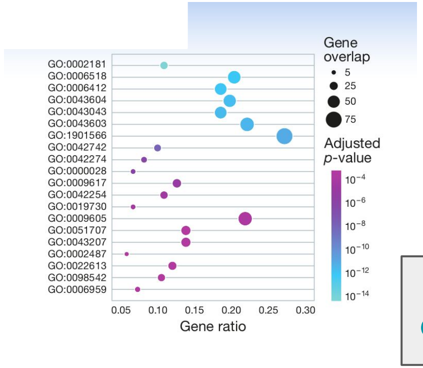
Shijie C. Zheng et al. (2022)

Differential Expression Analysis

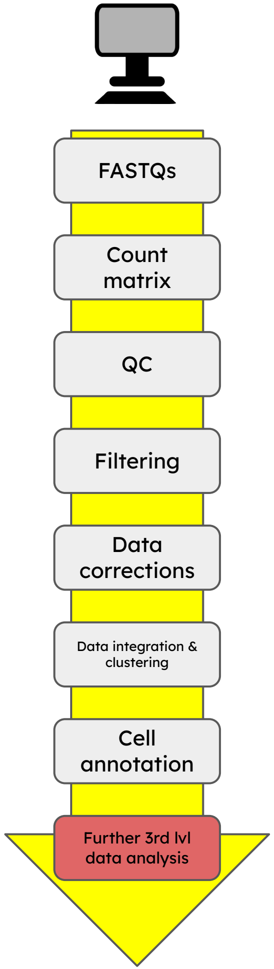


Samarendra Das et al. (2022)

Gene Set Analysis



Farhad Maleki et al. (2020)



FASTQs

Count matrix

QC

Filtering

Data corrections

Data integration & clustering

Cell annotation

Further 3rd lvl data analysis

(3) What tools can I use?

We need a **standard** for single cell data to make it FAIR

WHY?

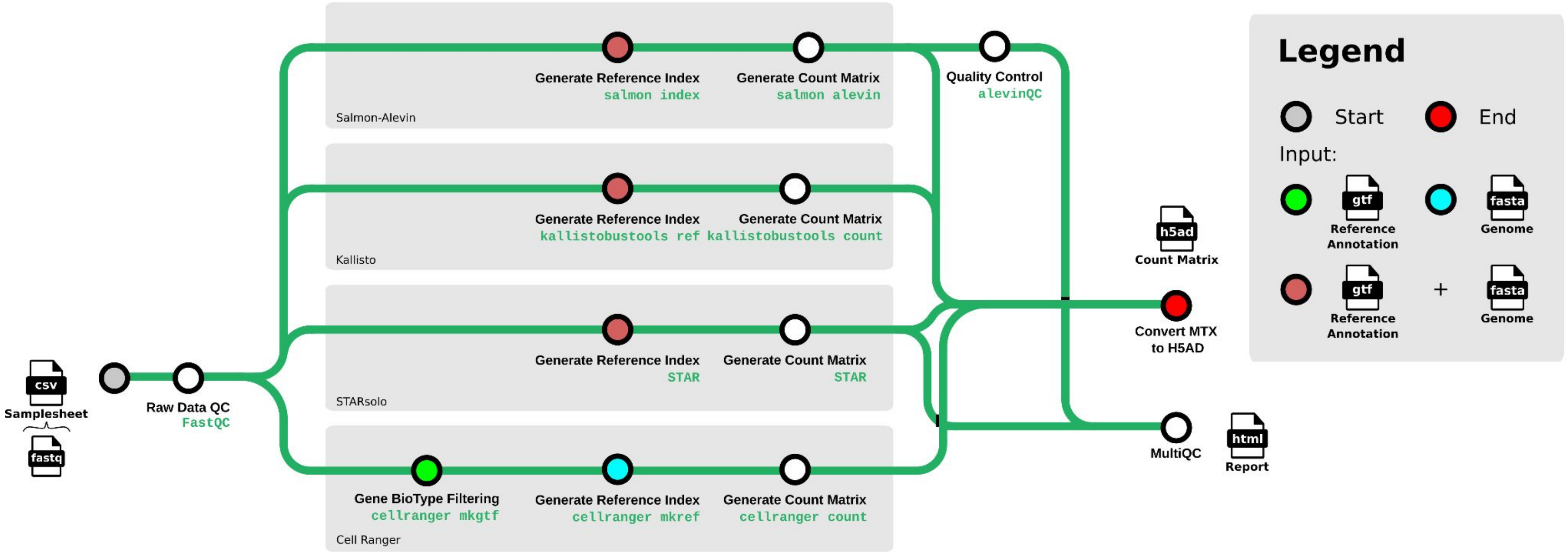
https://www.freepikcompany.com

nf-core/scrnaseq

- nf-core based
- scrnaseq version 2.0 released in June (DSL2)

- Protocols: SmartSeq2, 10xChromium, Drop-Seq

- 4 tools generating count matrix

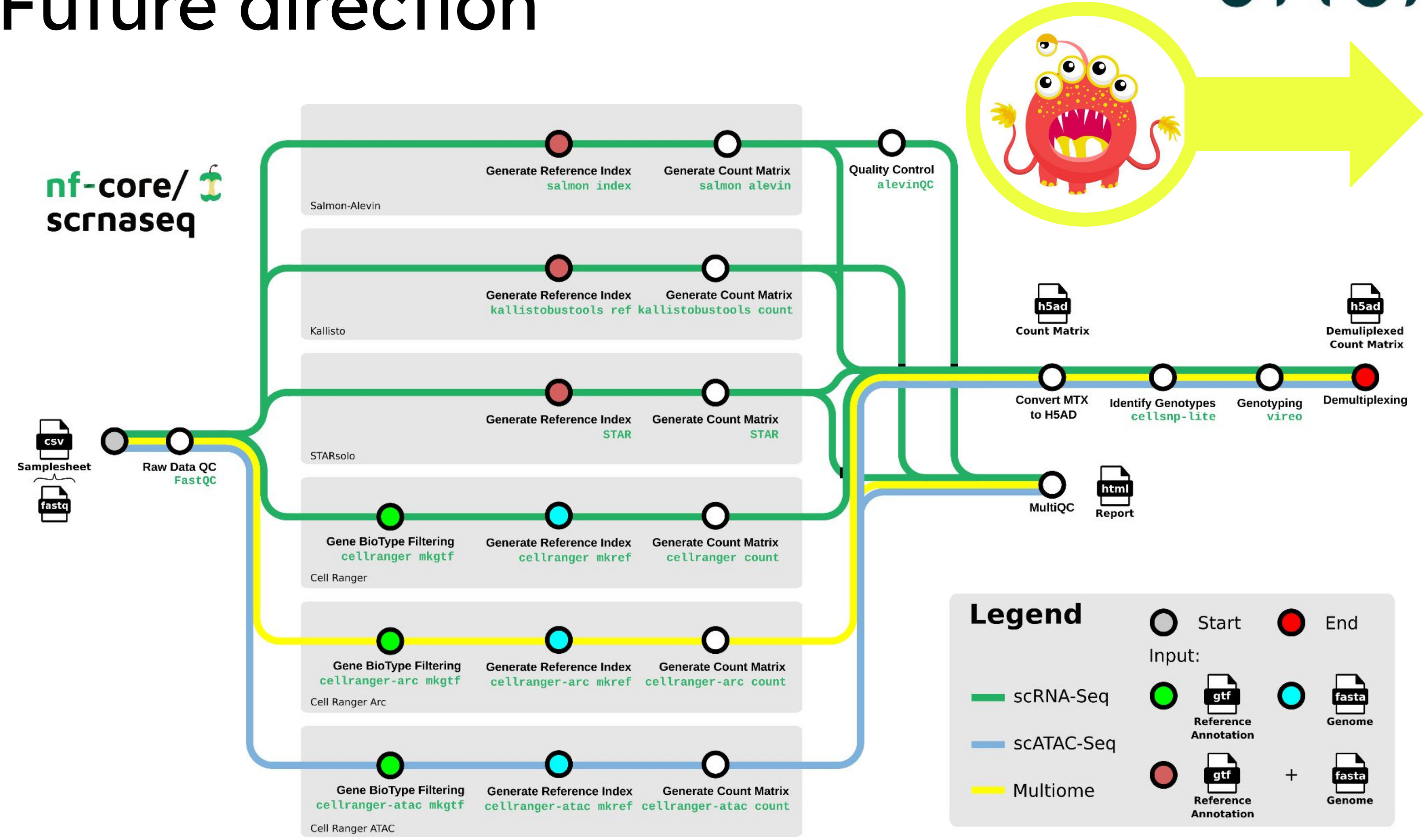


What does nf-core provide?

- Documentation
- CI Testing
- Stable Releases
- Packaged software
- Portable and reproducible
- Cloud-ready

Future direction

GHGA THE GERMAN HUMAN GENOME-PHENOME ARCHIVE



nf-core/scrnaseq

csv Samplesheet

fastq

Raw Data QC FastQC

Salmon-Alevin

Generate Reference Index salmon index

Generate Count Matrix salmon alevin

Quality Control alevinQC

Kallisto

Generate Reference Index kallistobustools ref

Generate Count Matrix kallistobustools count

STARsolo

Generate Reference Index STAR

Generate Count Matrix STAR

Cell Ranger

Gene BioType Filtering cellranger mkgtf

Generate Reference Index cellranger mkref

Generate Count Matrix cellranger count

Cell Ranger Arc

Gene BioType Filtering cellranger-arc mkgtf

Generate Reference Index cellranger-arc mkref

Generate Count Matrix cellranger-arc count

Cell Ranger ATAC

Gene BioType Filtering cellranger-atac mkgtf

Generate Reference Index cellranger-atac mkref

Generate Count Matrix cellranger-atac count

h5ad Count Matrix

Convert MTX to H5AD

Identify Genotypes cellsnp-lite

Genotyping vireo

h5ad Demuliplexed Count Matrix

Demultiplexing

MultiQC

html Report

Legend

Start End

Input:

scRNA-Seq

scATAC-Seq

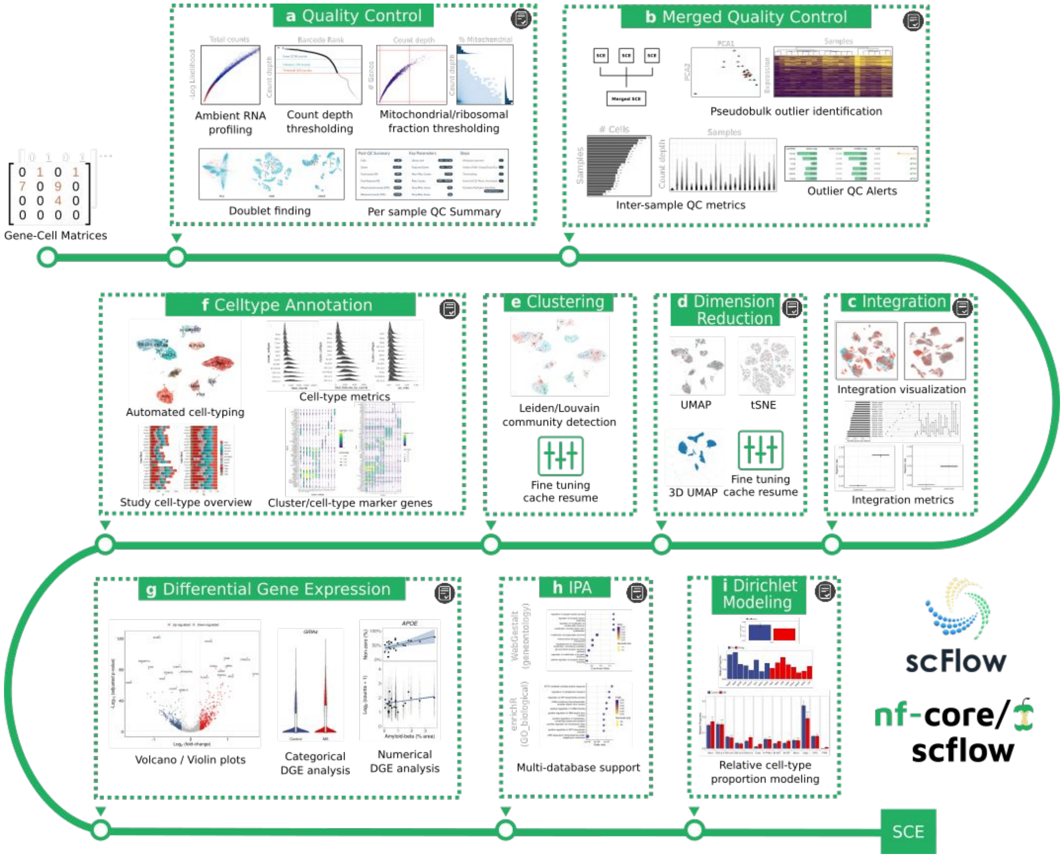
Multiome

gtf Reference Annotation

fasta Genome

gtf Reference Annotation + fasta Genome

scflow for further analysis



A benchmark “set” helps to cement a standard for GHGA

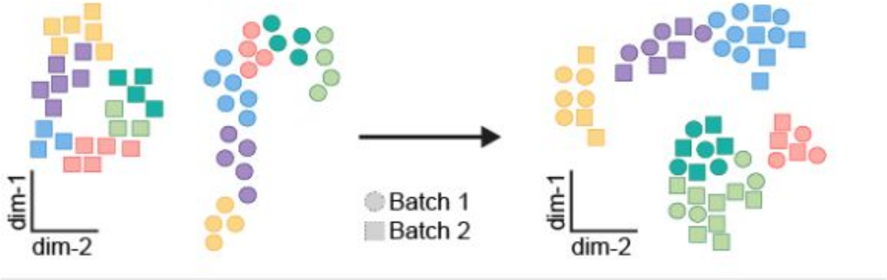
WHY?

https://www.freepikcompany.com

- Clear Tasks
- Easily accessible
- Quantitative metrics
- Runtime
- CI/CD + Continuous ranking (CR) = CI/CD/CR

Open Problems in Single-Cell Analysis

https://openproblems.bio/#about



Batch integration graph

Removing batch effects while preserving biological variation (graph output)

Christopher Lance et al. (2021)

Tools – Literature

Guidelines	• Jiajia Liu et al. (2021) • Malte D. Luecken & Fabian J. Theis (2019) • Galaxy Training Material (Single cell) • Tallulah S. Andrews et al. (2021) • Christopher Lance et al. (2021) • Rui Hong et al. (2022) • Sean Davis (https://github.com/seandavi/awesome-single-cell) • Ren Qi et al. (2020) • Yulong Zhang et al. (2022)	• Samarendra Das et al. (2022) • Farhad Maleki et al. (2020)
Software	• Cellranger: QC, count matrix, etc. • scverse (scanpy, muon, scvi-tools, …): QC, normalization, data integration, clustering, cell annotation, etc. • Scrublet: doublet/multiplet filtering • Cellsnp-lite: read pileup & genotyping • Vireo: genotyping & demultiplexing • ArchR: scATAC QC, etc.	• CellBender: emptyplet filtering • MultiQC: QC • Seurat: QC, filtering, etc. • nf-core/scrnaseq: QC, count matrix, etc. • nf-core/scflow: QC, clustering, etc.
Benchmarks	• Luyi Tian et al. (2019) • Malte D. Luecken et al. (2022) • Cody N. Heiser et al. (2021) • Ralf Schulze Brüning et al. (2022) • Huidong Chen et al. (2019) • Wenpin Hou et al. (2020)	• Sean Simmons (2022)