Processing and Quality Control of scRNAseq data

Åsa Björklund
asa.bjorklund@scilifelab.se

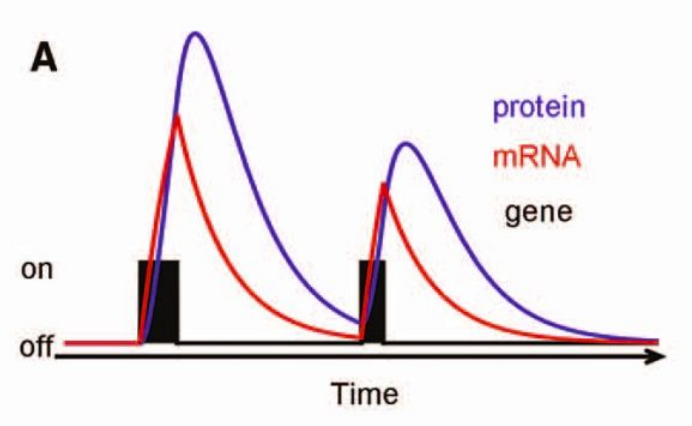
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

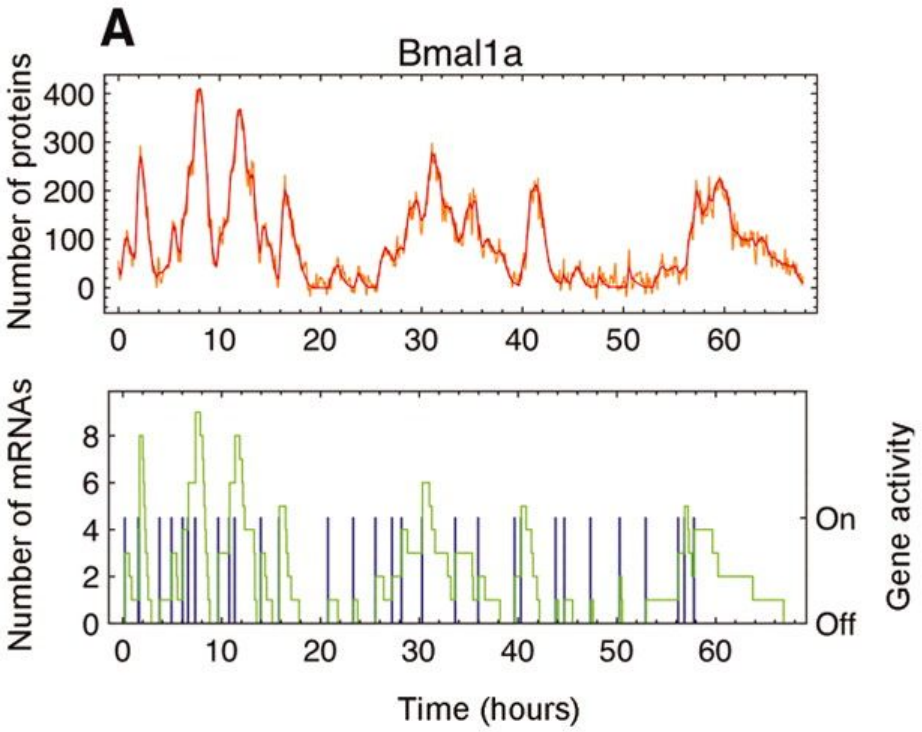
SciLifeLab

Outline

- Background on transcriptional bursting & drop-outs
- From reads to expression counts
- Experimental setup – what could go wrong?
- snRNAseq
- Spike-in RNAs
- Quality control:
  - Filtering low quality cells
  - Doublets
  - Ambient RNA effects
  - Filtering of genes

Transcriptional bursting

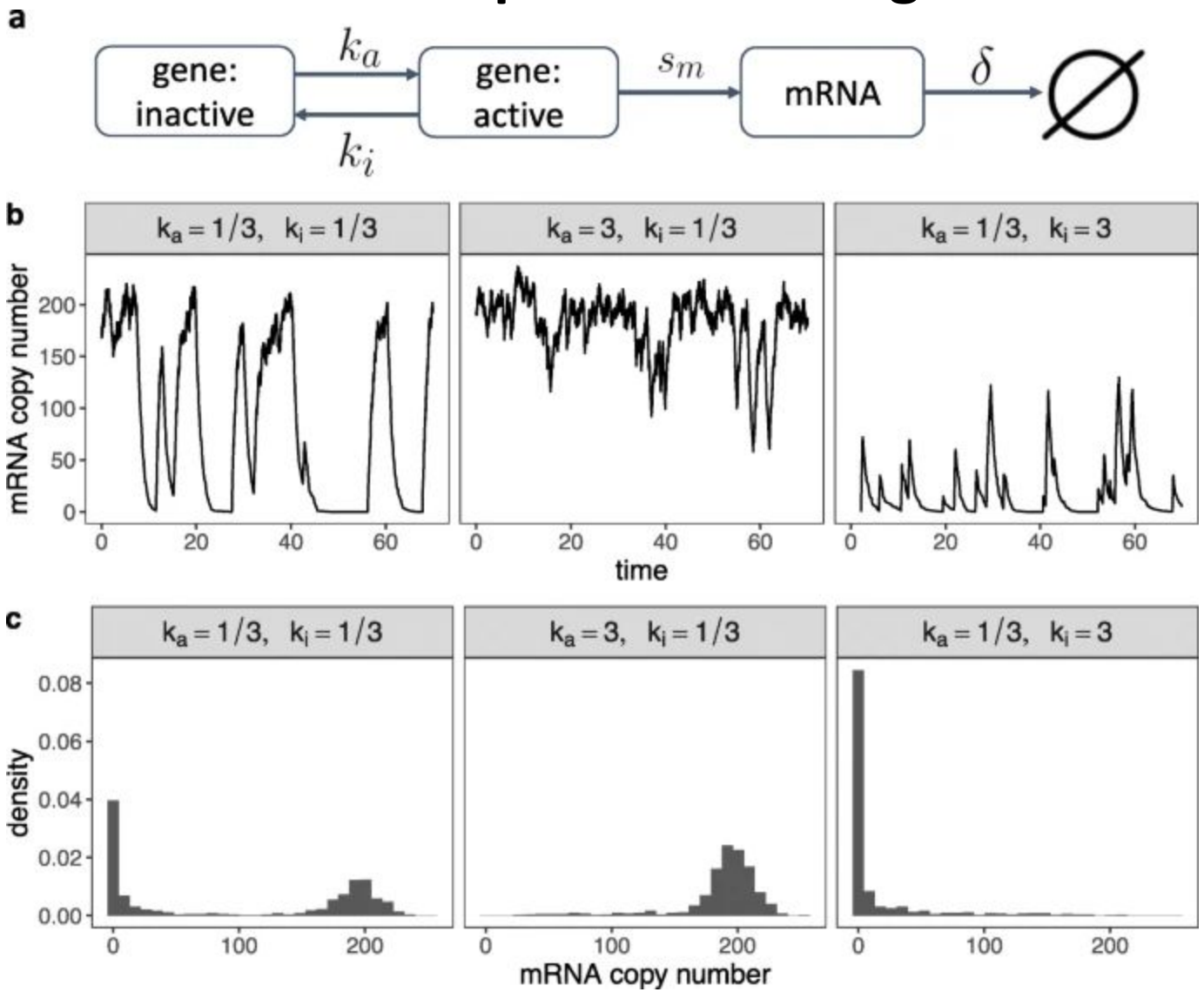




- Burst frequency and size is correlated with mRNA abundance
- Many TFs have low mean expression (and low burst frequency) and will only be detected in a fraction of the cells

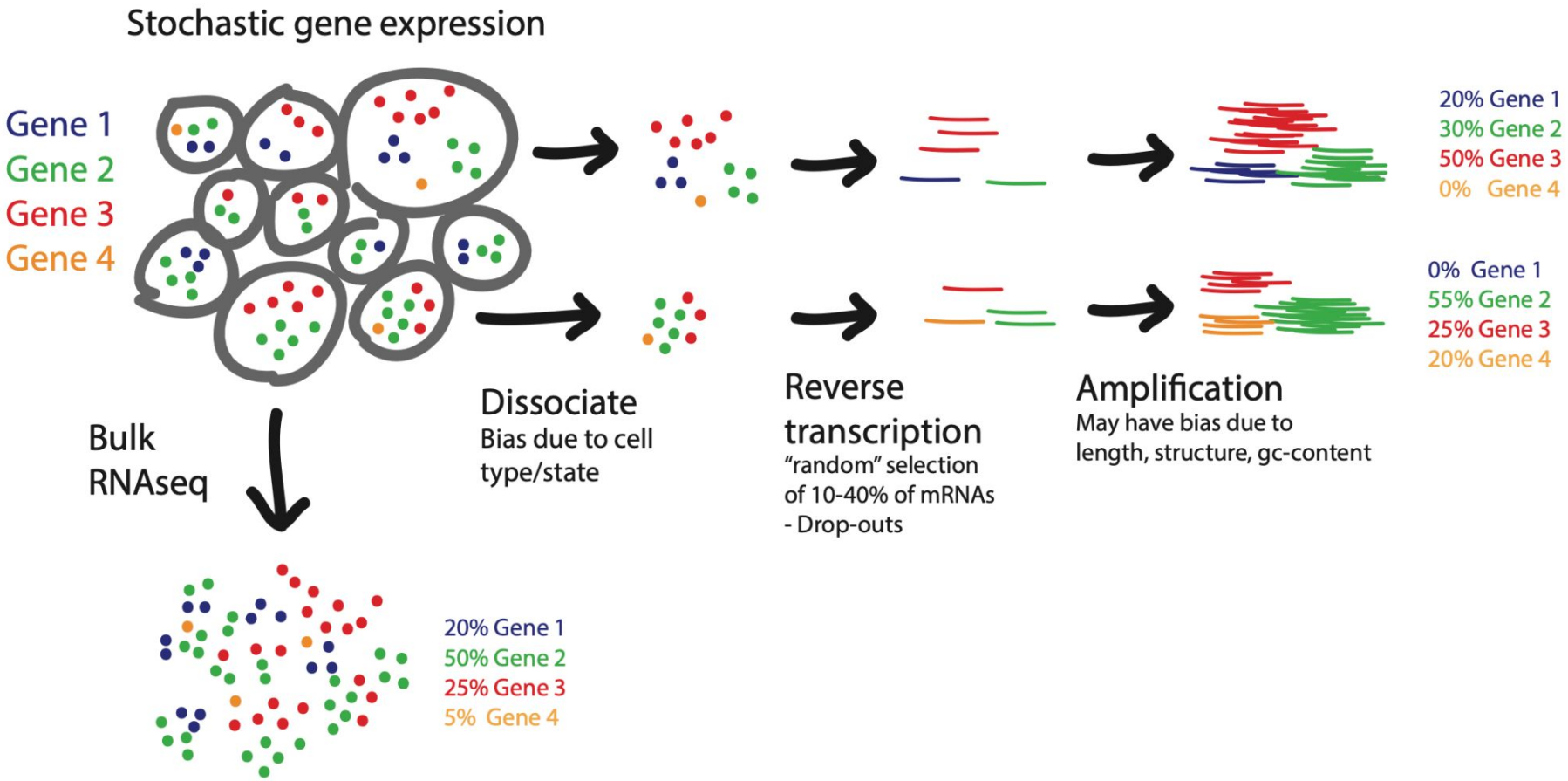
(Suter et al. *Science* 2011)

Transcriptional bursting



(Jiang et al. Genome Biol 2022)

Bursting, drop-outs and amplification bias



NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

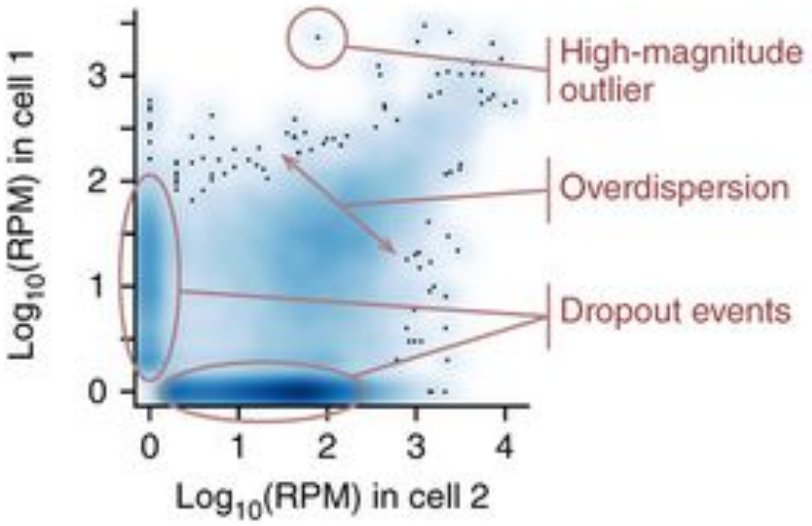
SciLifeLab

Zero inflation?

Many use the term zero inflation - partly true due to dropout rates.

Stochastic gene expression zeros - are biologically relevant!

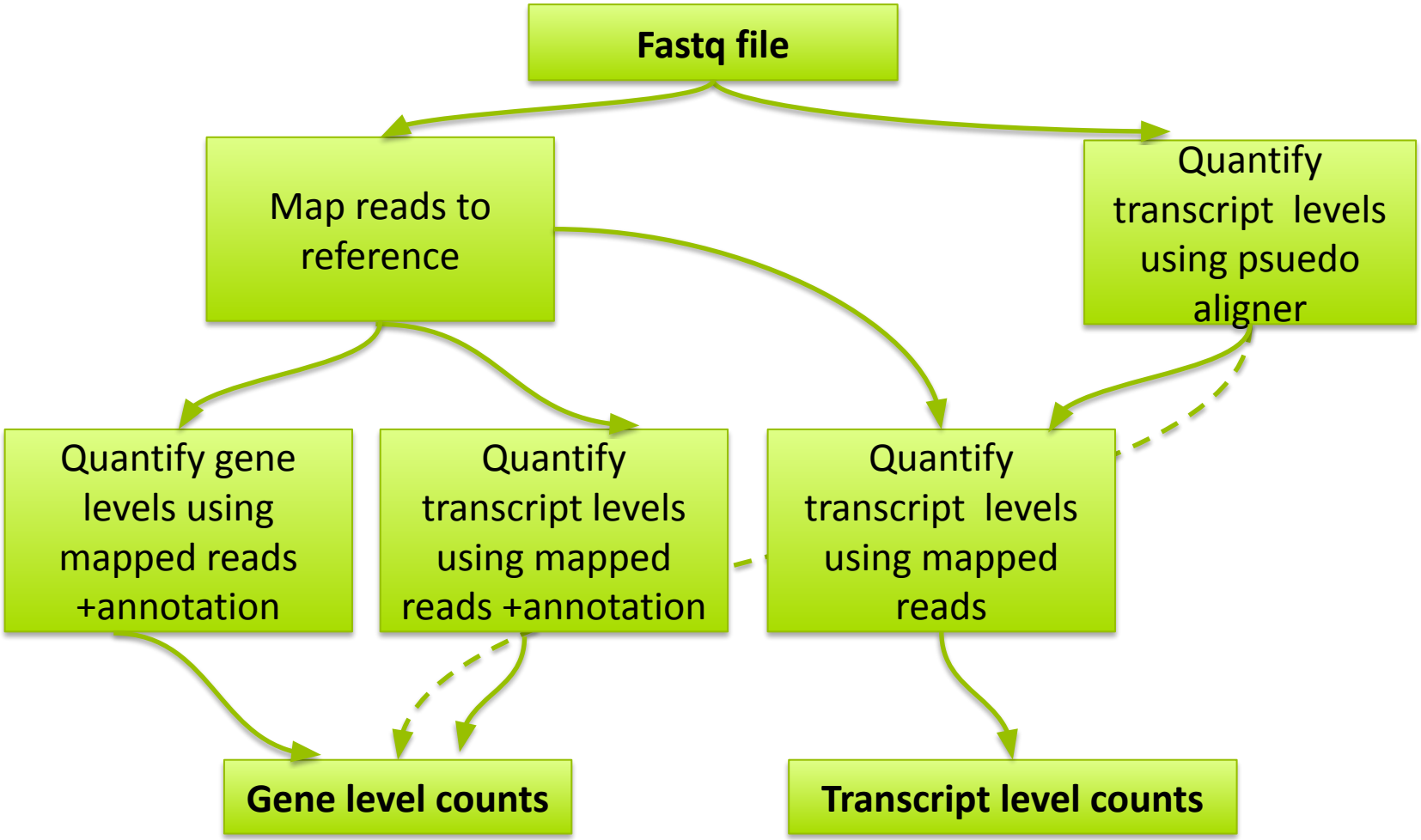
Problems compared to bulk RNA-seq



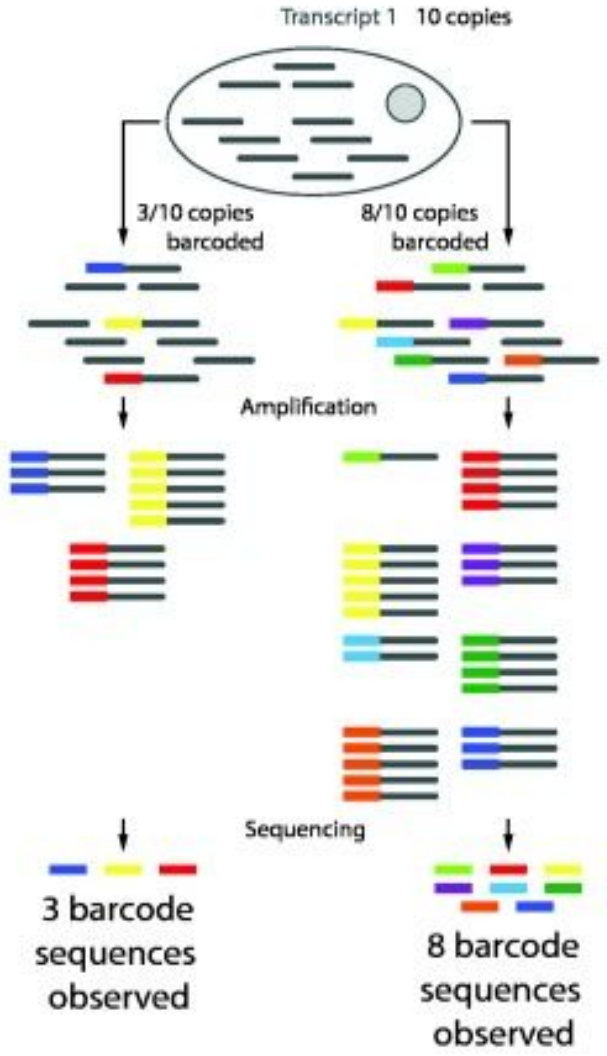
- Amplification bias
- Drop-out rates
- Transcriptional bursting
- Background noise
- Bias due to cell-cycle, cell size and other factors
- Often clear batch effects

(Karchenko et al. *Nature Methods* 2014)

RNA-seq - Different paths to get a count table



UMI (Unique molecular identifiers) will make sure that one fragment is counted as one read



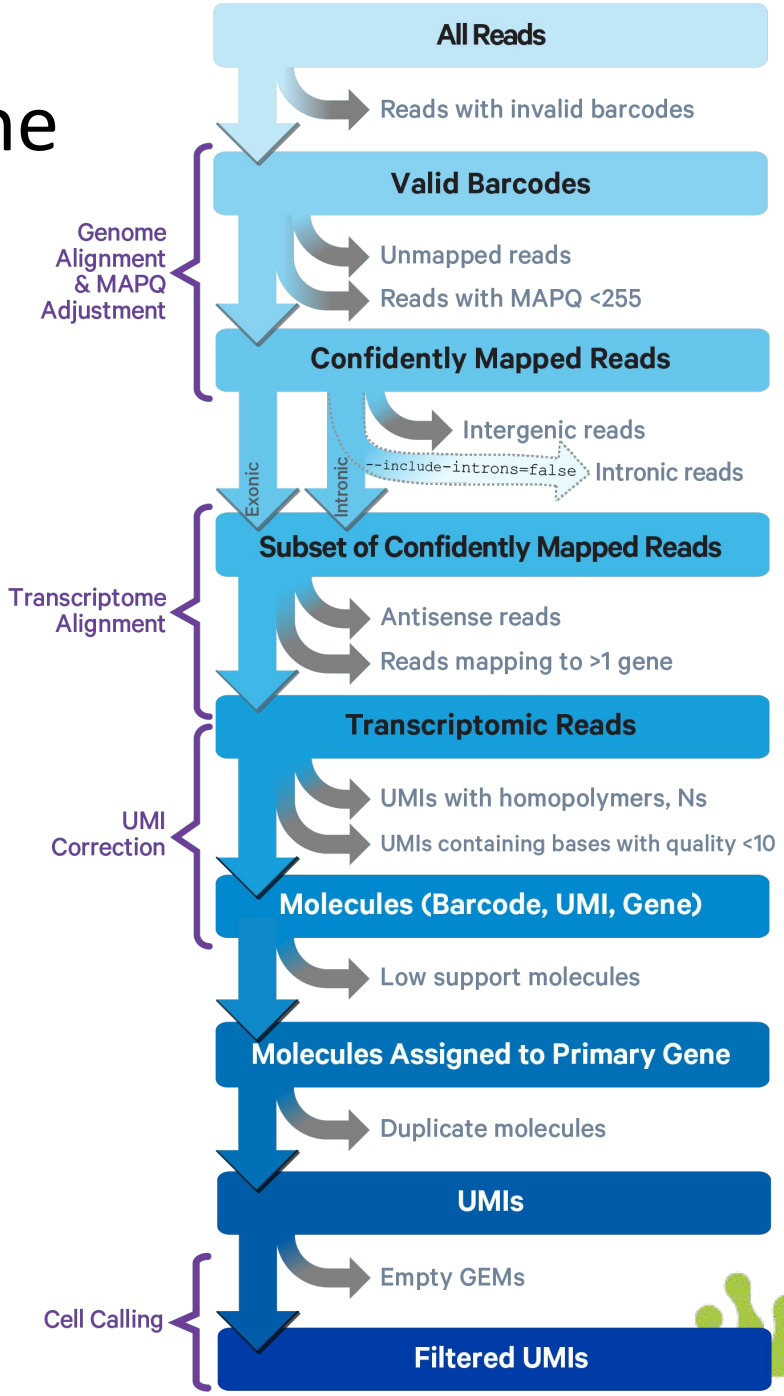
- Will remove errors that occur during the amplification step.
- Will not handle sampling bias

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Cellranger pipeline

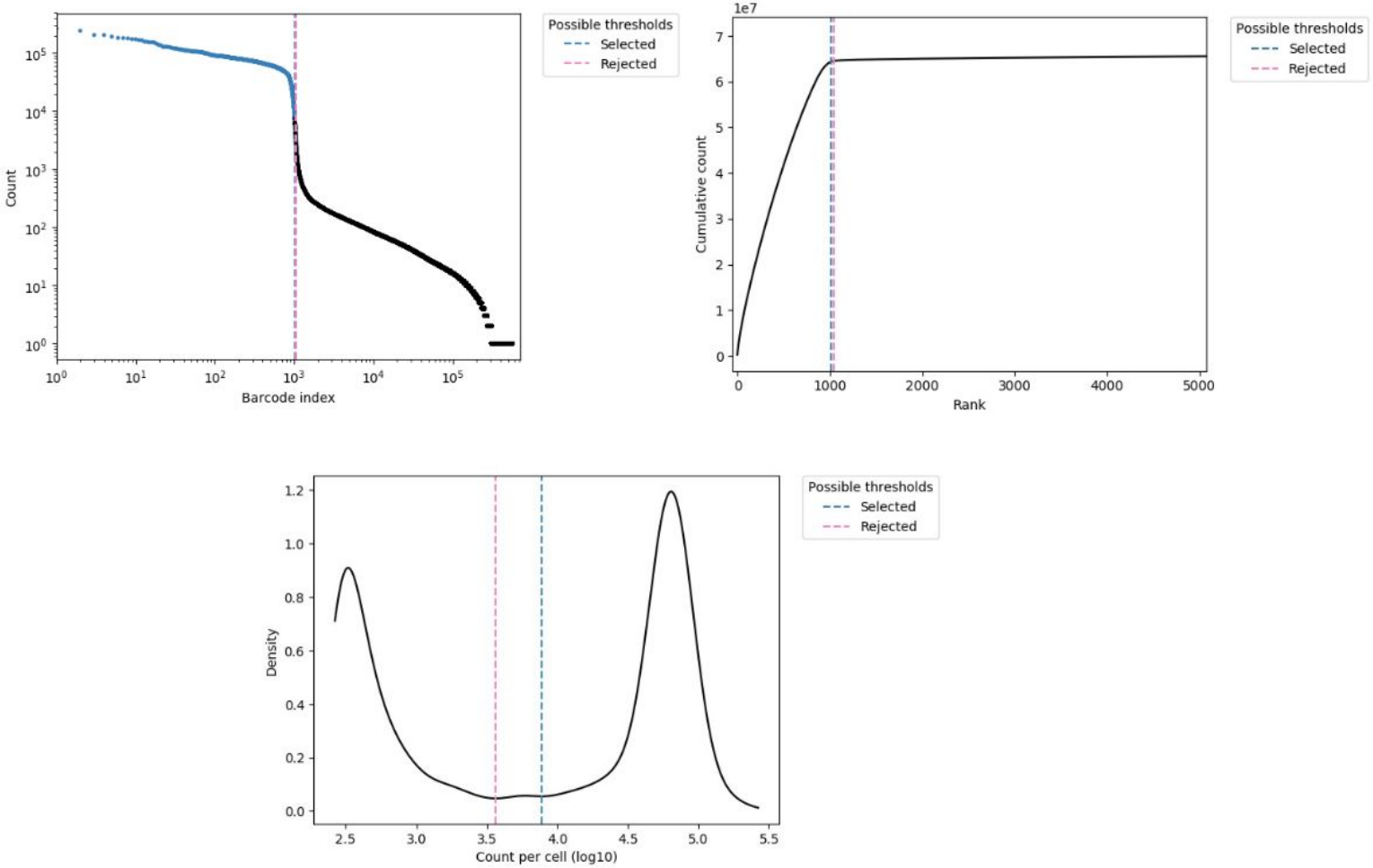
More detail on:
https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/algorithms/overview



NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Cell calling for droplet-based methods



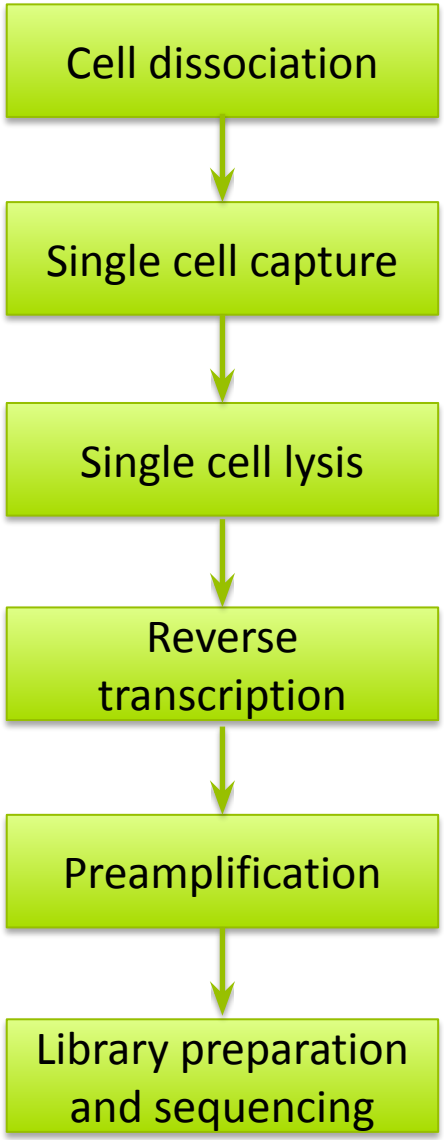
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Cellranger reports

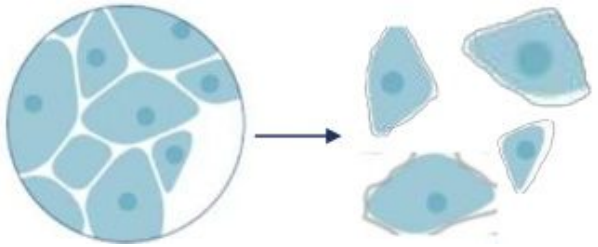
- https://cf.10xgenomics.com/samples/cell-exp/4.0.0/Parent_NGSC3_DI_HodgkinsLymphoma/Parent_NGSC3_DI_HodgkinsLymphoma_web_summary.html

Experimental setup



What could go wrong?

Experimental setup



Cell dissociation

Single cell capture

Single cell lysis

Reverse transcription

Preamplification

Library preparation and sequencing

It is critical to have healthy whole cells with no RNA leakage.
Short time from dissociation to prep!

PROBLEMS:

- Incomplete dissociation can give multiple cells sticking together.
- To harsh dissociation may damage cells -> RNA degradation and RNA leakage.
- Leakage of RNA –> background signal.
- Different celltypes are more/less sensitive to dissociation.

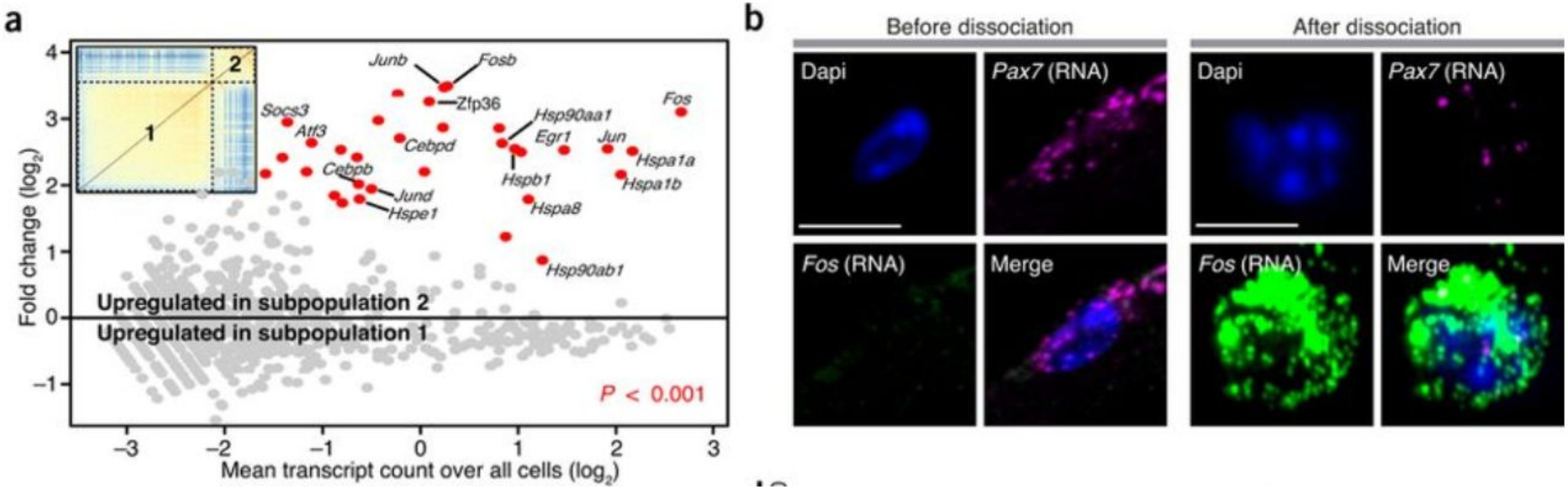
(Kolodziejczyk et al. 2015)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Dissociation artifacts

- Dissociation may bias your cell populations
- Dissociation protocols may introduce transcriptional changes.

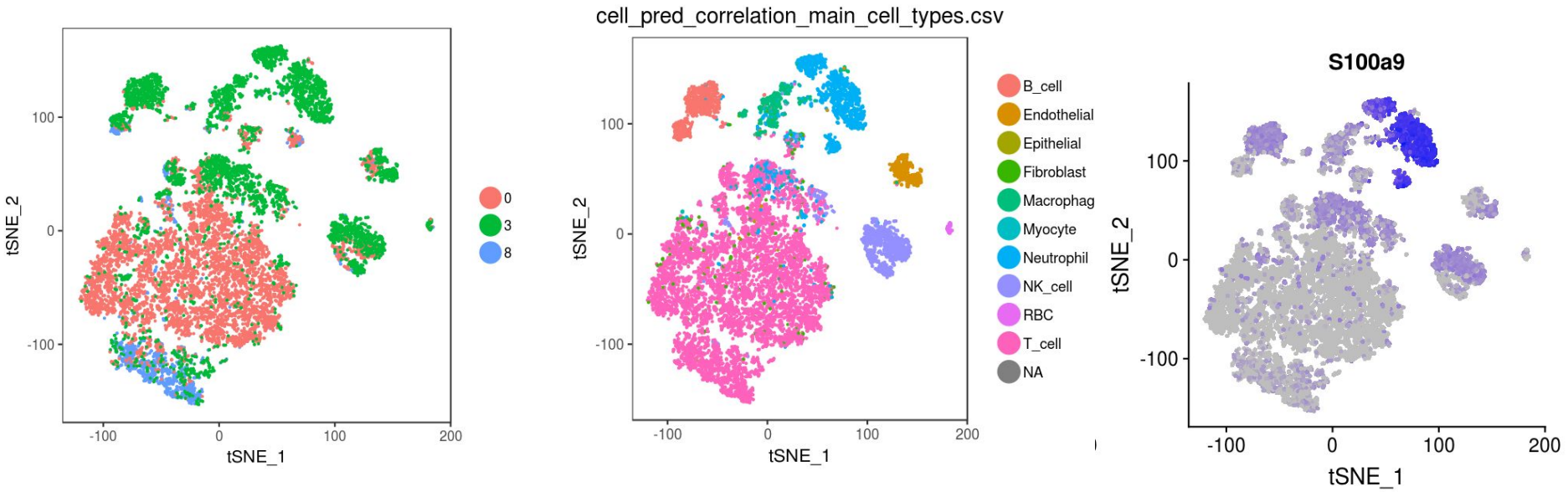


(van den Brink et al. Nature methods 2017)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

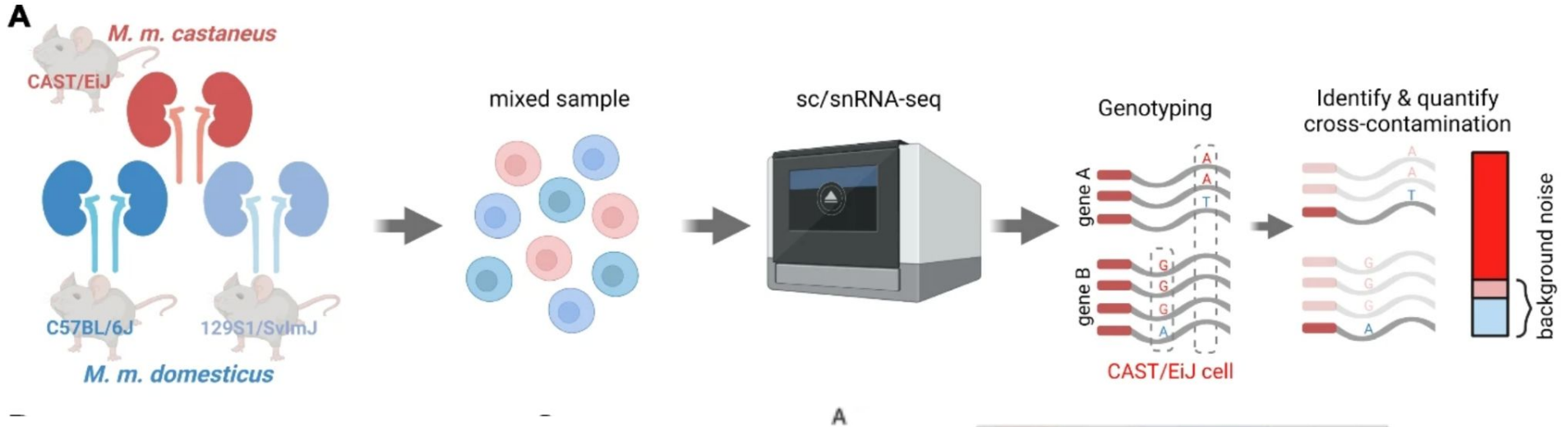
Ambient RNA

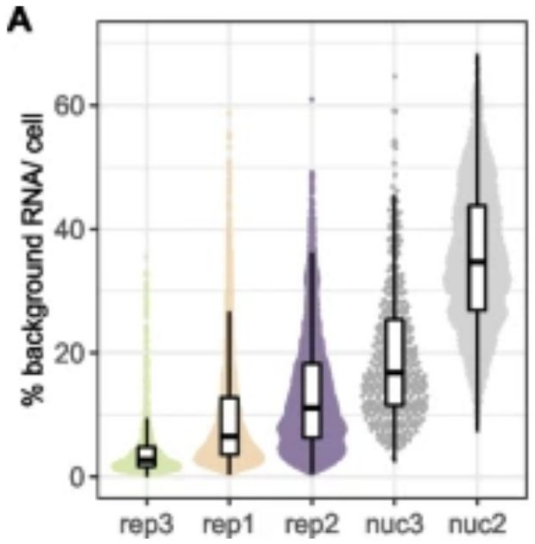


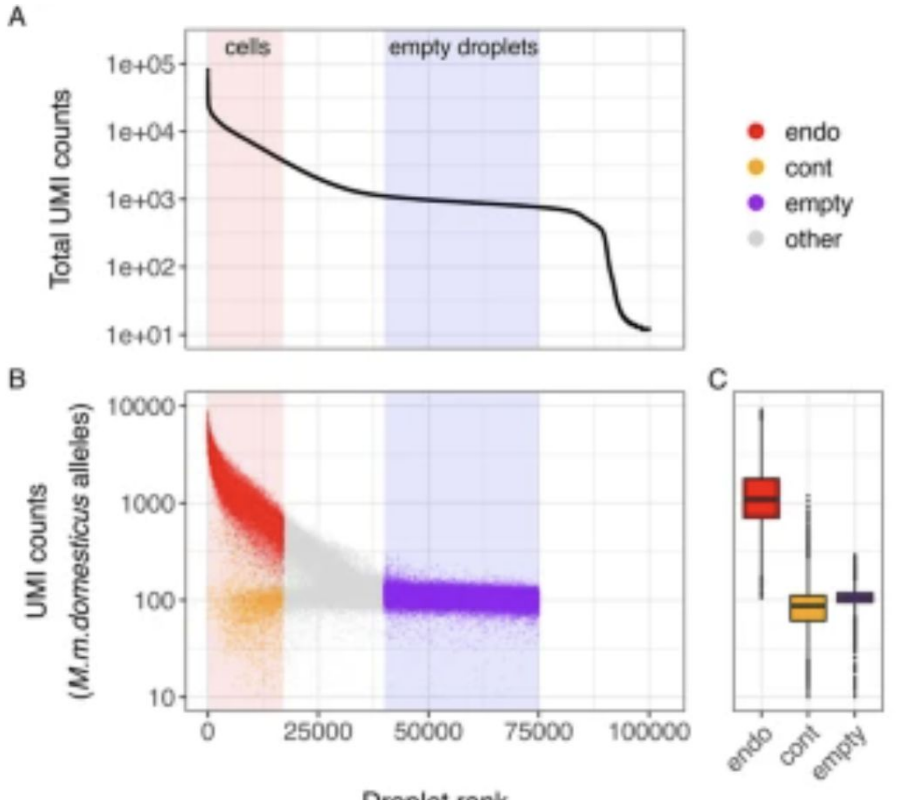
Sample from Day3 have detection of Neutrophil markers in all cells.

Probably contamination from ambient RNA.

Ambient RNA







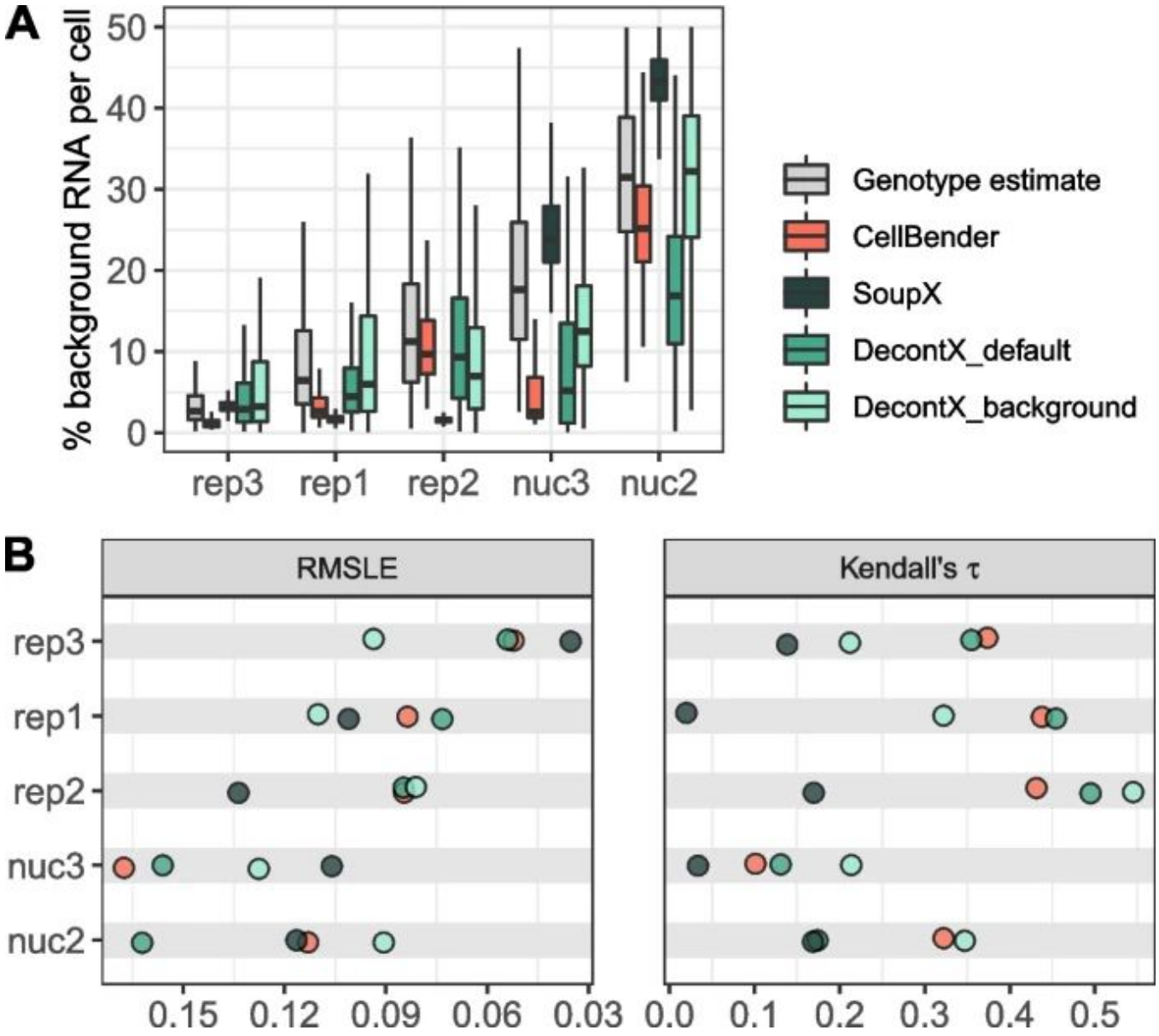
(Janssen et al. Genome Biology 2023)

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Ambient RNA

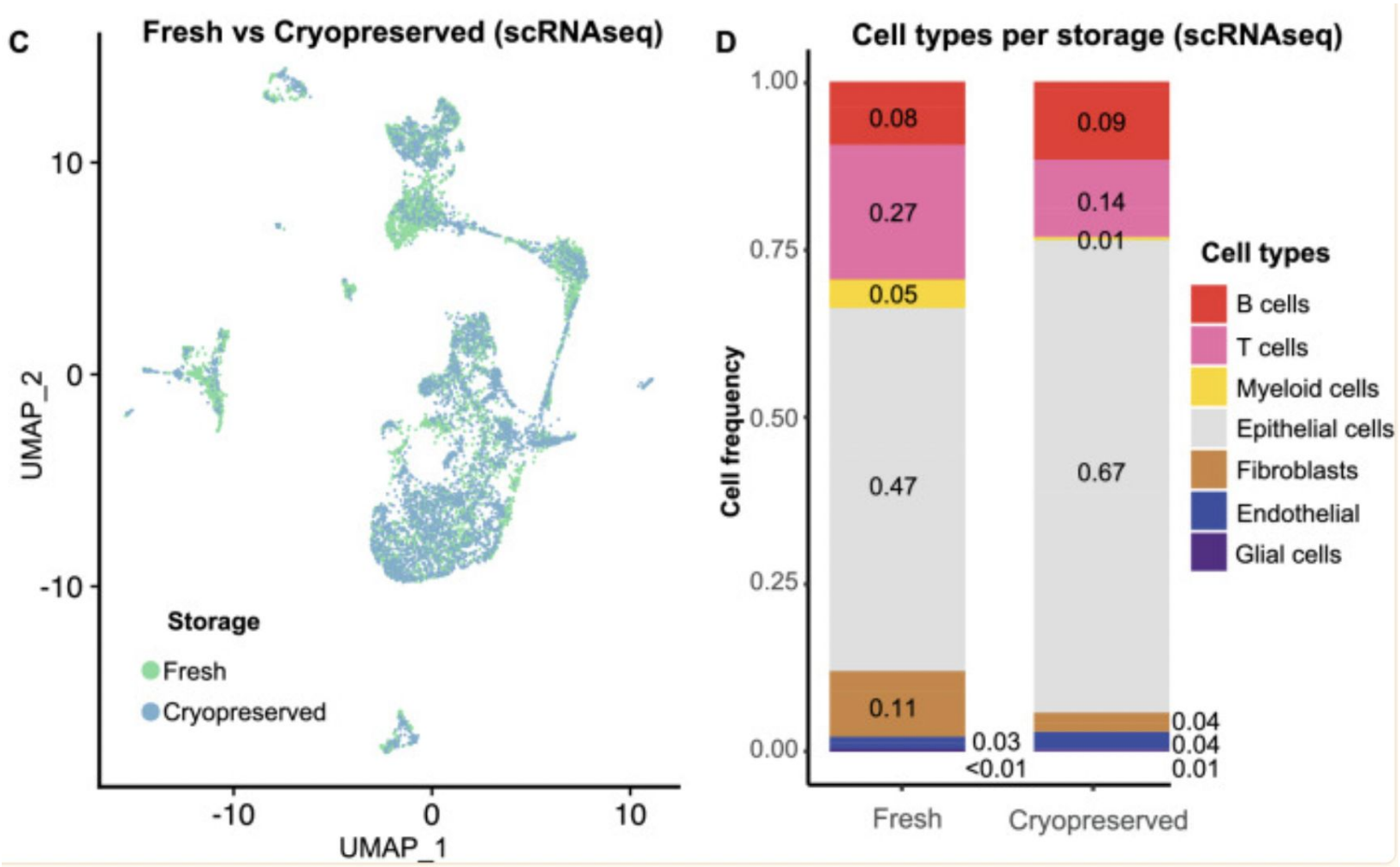
- Using empty droplets, estimate background signal – gives warning in cellranger report.
- Some methods are:
  - SoupX (Young MD, GigaScience 2020)
  - Cellbender (Flemming et al. Nature Methods 2023)
  - DecontX (Yang et al. Genome Biology 2020)

SoupX and DecontX depends on defining celltype clusters in the data. Cellbender uses only the empty droplets.



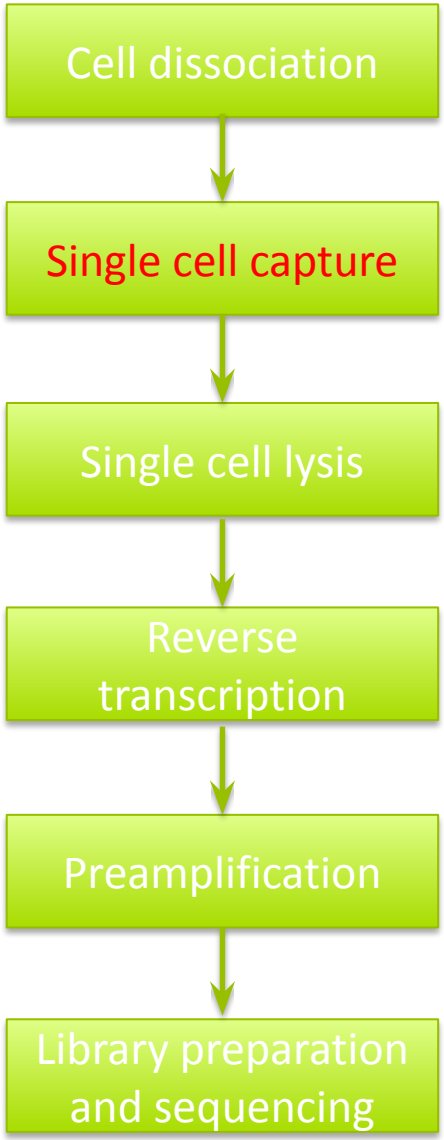
(Janssen et al. Genome Biology 2023)

Biased celltype distribution



(Uniken Venema *et al.* Sci. Rep. 2022)

Experimental setup



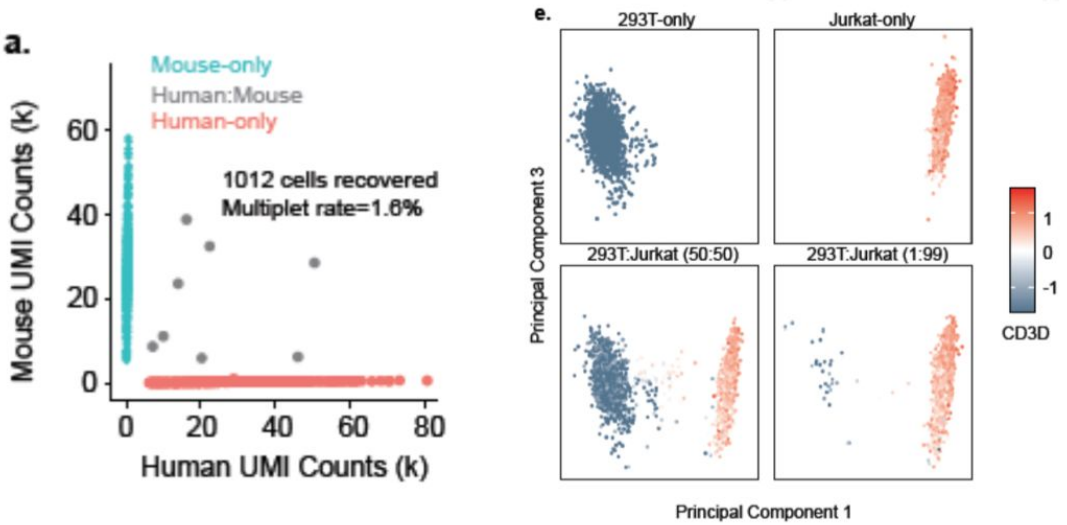
PROBMLEMS:

- All methods may give rise to **empty** wells/droplets, and also **duplicates** or multiples of cells.
- Size selection bias for many of the methods – dropseq has upper limit for **cell size**.
- Biased selection of certain celltype(s)
- Long **time** for sorting may damage the cells

(Kolodziejczyk et al. 2015)

Doublets in scRNAseq

scRNA-seq is not always single-cell

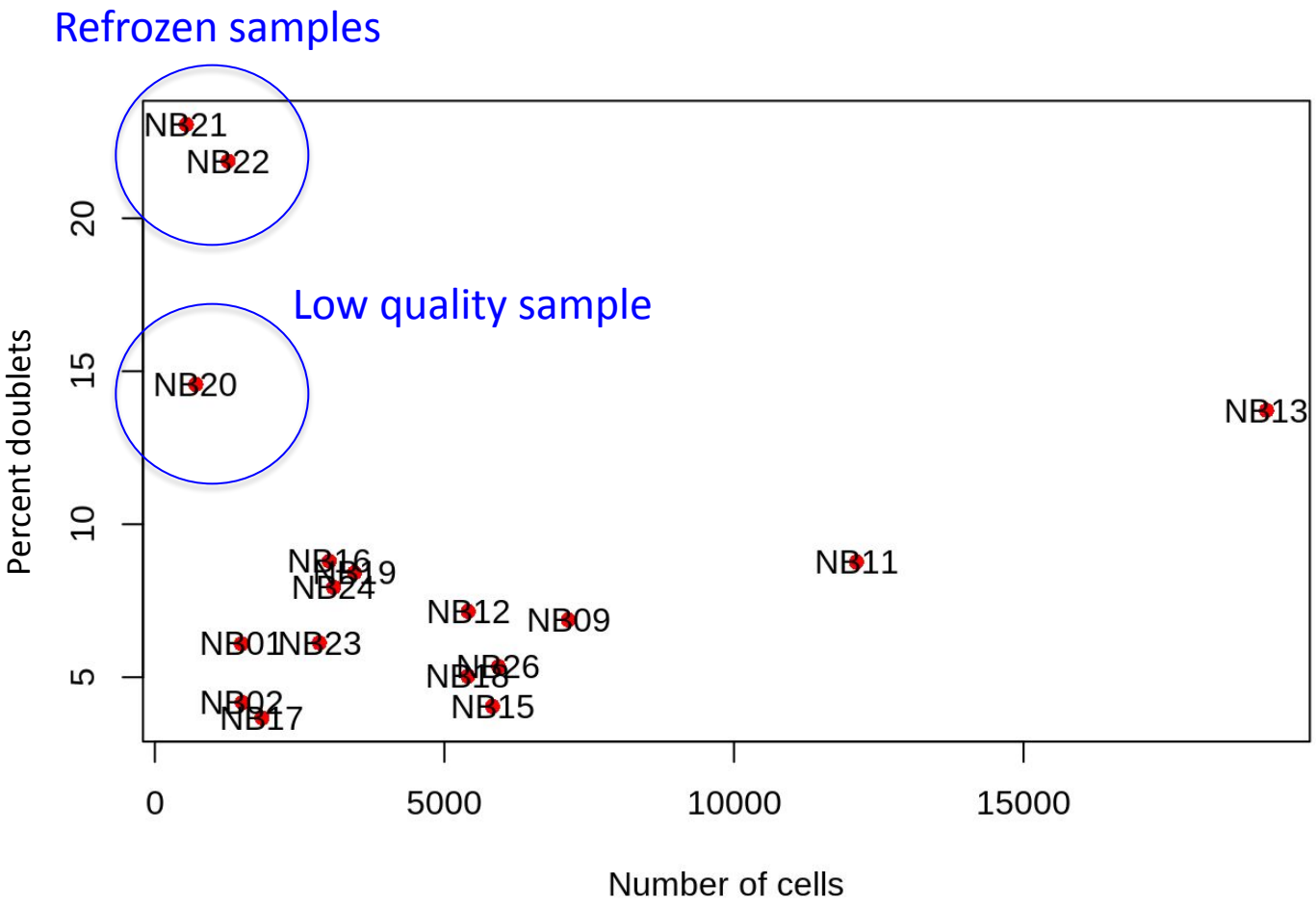


Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~870	~500
~0.8%	~1700	~1000
~2.3%	~5300	~3000
~3.9%	~8700	~5000
~7.6%	~17400	~10000

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Cell debris may cause doublet signatures



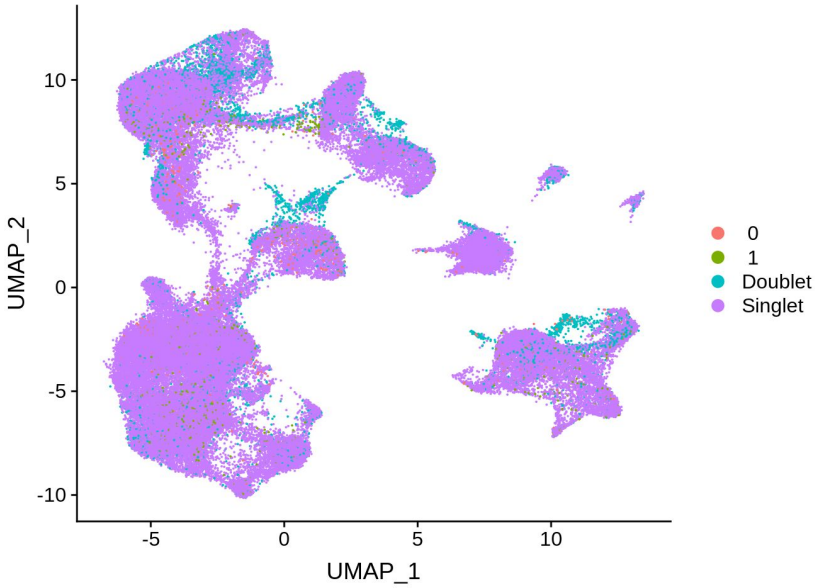
Predicted doublets based on celltype signatures

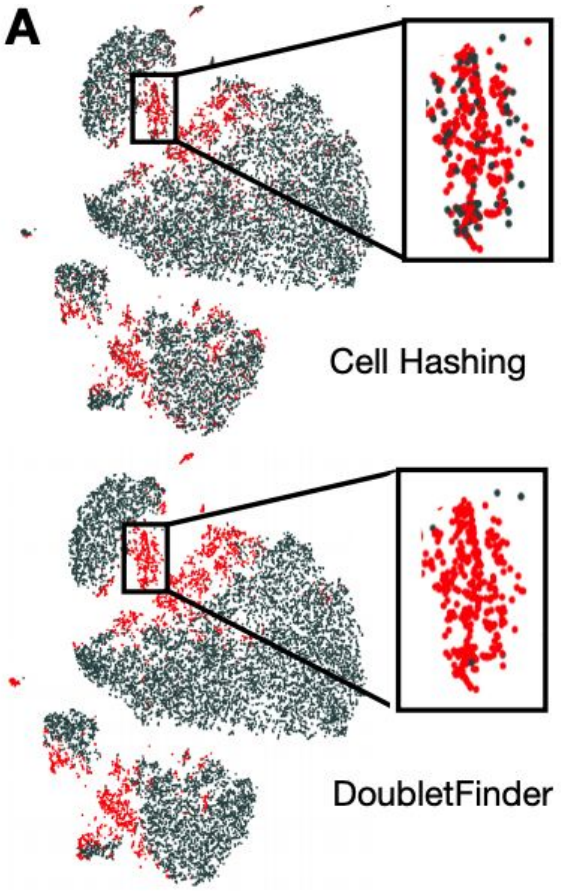
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Doublets in scRNAseq

- Can be distinct cluster
- Can be a streak between clusters.





(McGinnis et al. BioRxiv 2018)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

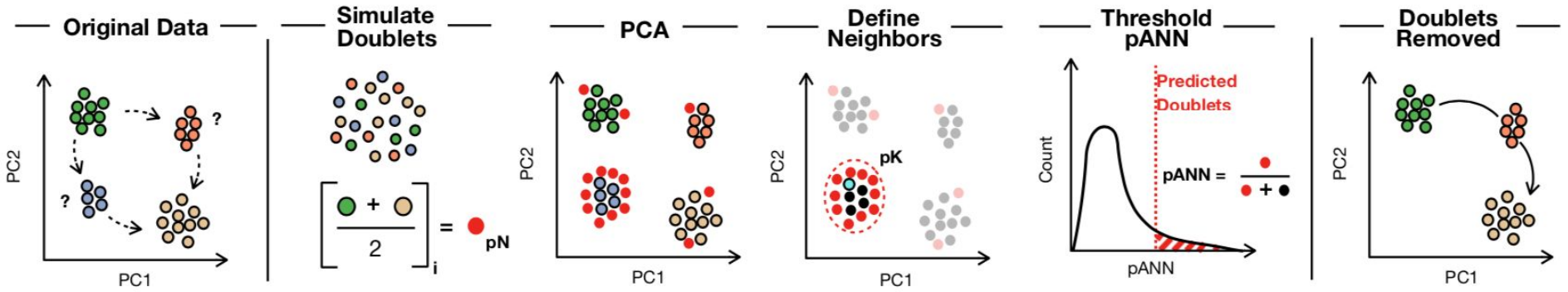
Detecting duplicate/multiple cells

- High number of detected genes or UMIs – can be a sign of multiples
  - But, be aware so that you do not remove all cells from a larger celltype.
- After clustering – check if you have cells with signatures from multiple clusters.

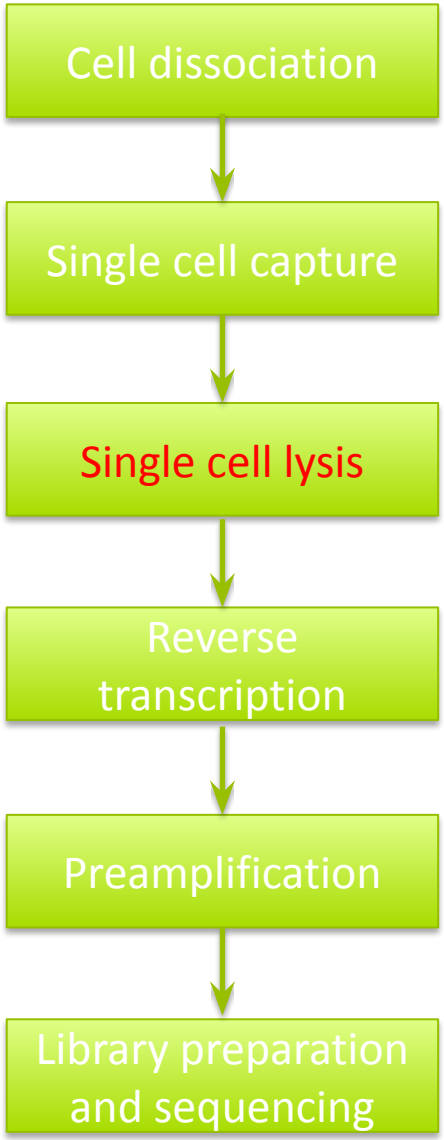
- With 10X you should have a feeling for your doublet rate based on how many cells were loaded. May be more than in their benchmark.

Doublet detectors

- DoubletFinder - https://github.com/chris-mcginnis-ucsf/DoubletFinder
- Scrublet - https://github.com/AllonKleinLab/scrublet
- DoubletDecon - https://github.com/EDePasquale/DoubletDecon
- DoubletCluster / DoubletCell in Scran



Experimental setup



Optimal lysis conditions may vary from celltype to celltype and for nuclei vs cells.

PROBLEMS:

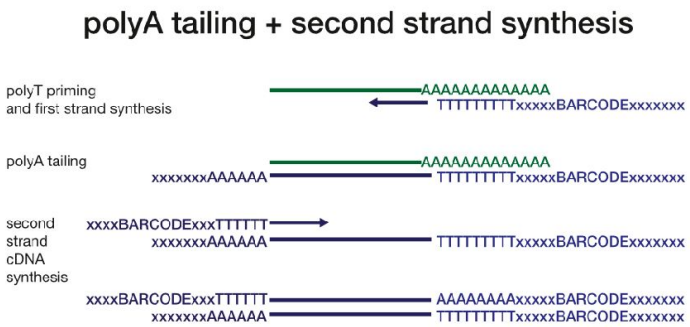
- To harsh lysis conditions may interfere with library prep.
- Different lysis conditions may/may not give nuclear lysis.

Can give biased cell populations.

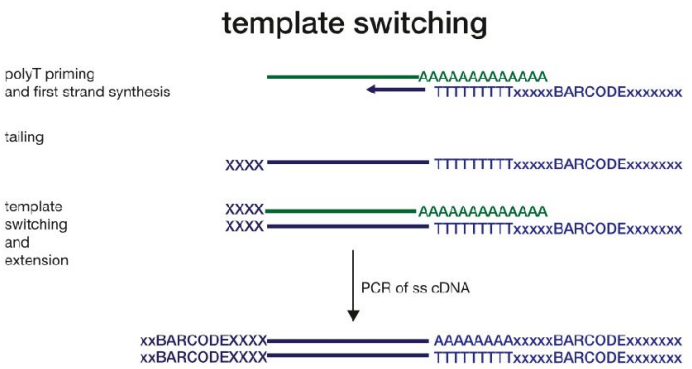
(Kolodziejczyk et al. 2015)

Experimental setup

- Cell dissociation
- Single cell capture
- Single cell lysis
- Reverse transcription
- Preamplification
- Library preparation and sequencing



Tang protocol (Tang et al 2009)
CELseq/MARSseq (Hashimony et al. 2013, Jaitin et al. 2014)
QuartzSeq (Sasagawa et al. 2013)



SmartSeq/SmartSeq2 (Ramskold et al. 2012, Deng et al. 2014)
STRT (Islam et al. 2011)

Efficiency of reverse transcription is the key to high sensitivity. Drop-out rate is around 90-40% depending on the method used.

Two libraries with the same method using the same cell type may have very different drop-out rates.

(Kolodziejczyk et al. 2015)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Experimental setup

Cell dissociation

Single cell capture

Single cell lysis

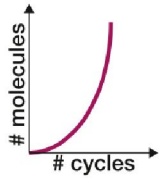
Reverse transcription

Preamplification

Library preparation and sequencing

PCR

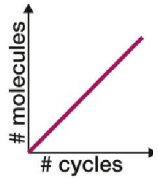
- exponential amplification
- PCR base specific biases



Tang protocol (Tang et al. 2009)
STRT (Islam et al. 2011)
SmartSeq/SmartSeq2 (Ramskold et al. 2012, Deng et al. 2014)

IVT

- linear amplification
- 3' bias due to two rounds of reverse transcription



CELseq/MARSseq (Hashimony et al. 2013, Jaitin et al. 2014)

Any amplification step will introduce a bias in the data.

Methods that uses **UMIs** will control for this to a large extent, but the chance of detecting a transcript that is amplified more is higher. Sequencing saturation matters.

Some full length methods like SmartSeq2 has no UMIs, so we cannot control for amplification bias.

(Kolodziejczyk et al. 2015)

Experimental setup

Cell dissociation

Single cell capture

Single cell lysis

Reverse transcription

Preamplification

Library preparation and sequencing

Multiplexing of samples will not always be perfect, so the number of reads per cell may vary quite a lot.

Base calls in the sequencing may be affected by a number of factors:

- Low complexity of library – may be an issue when there are many primer dimers
- Base call quality scores may be affected if there are contaminations in the flow cell

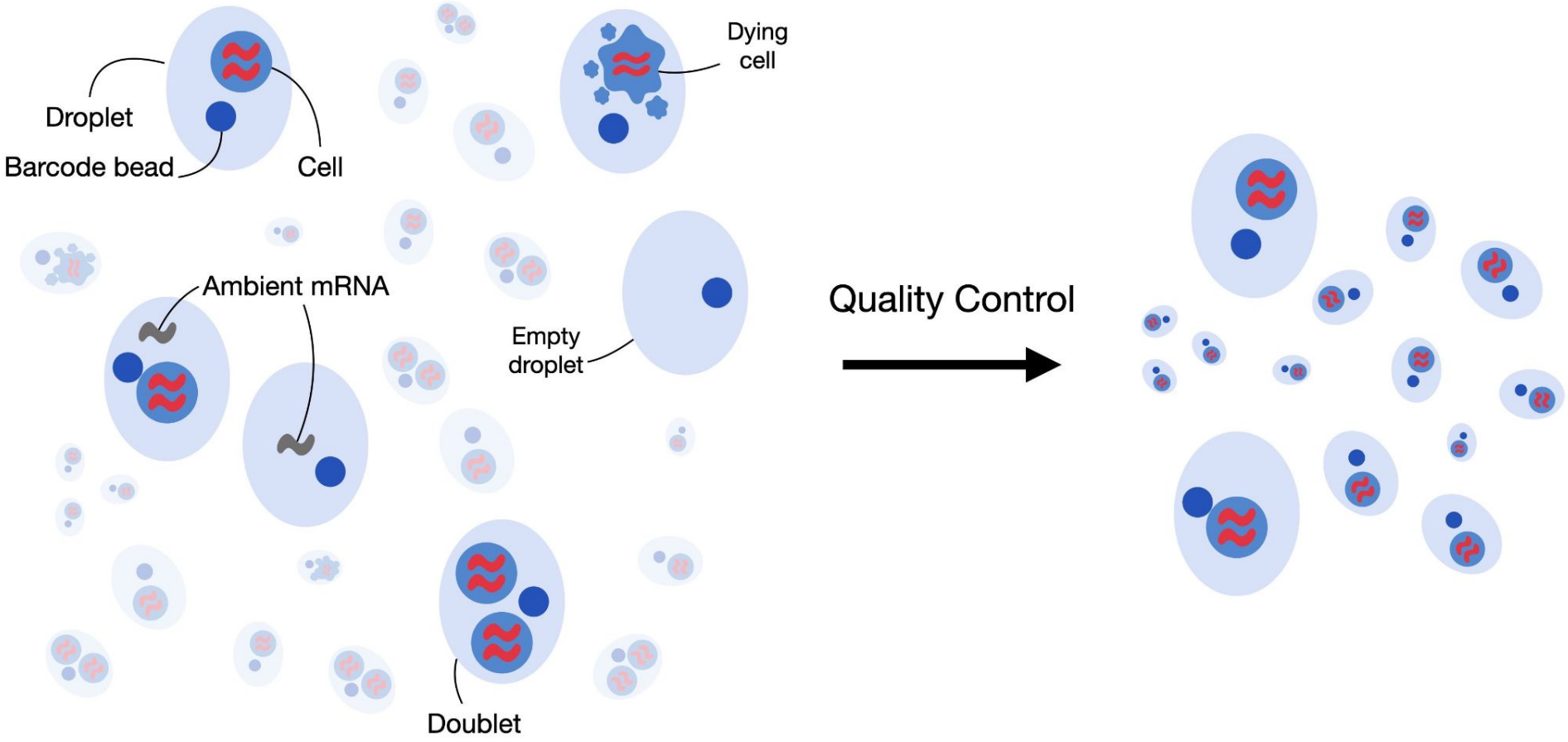
Index swapping

(Kolodziejczyk et al. 2015)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

QC summary



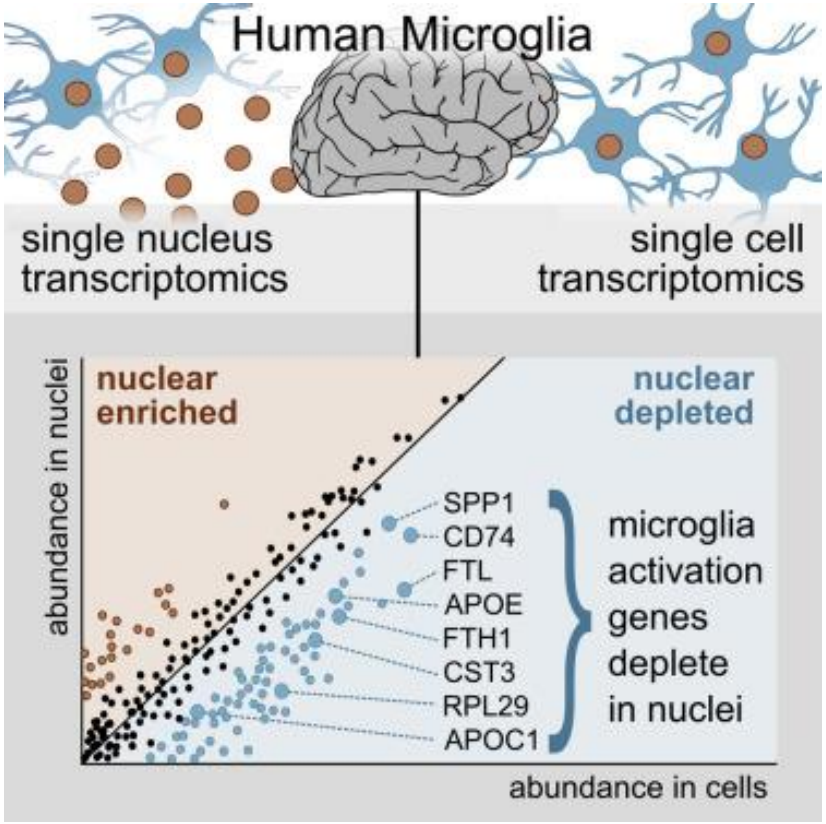
(https://www.sc-best-practices.org/)

Single cell or single nuclei?

- snRNAseq pros:
  - Can avoid some biases due to dissociation.
  - Hard to dissociate celltypes (e.g. neurons, muscle fibres, adipocytes)
  - Frozen tissues
- snRNAseq cons:
  - Less mRNA per nuclei
  - More dominated by nuclear lincRNAs
  - Internal priming of polyA stretches in introns
  - More ambient RNA

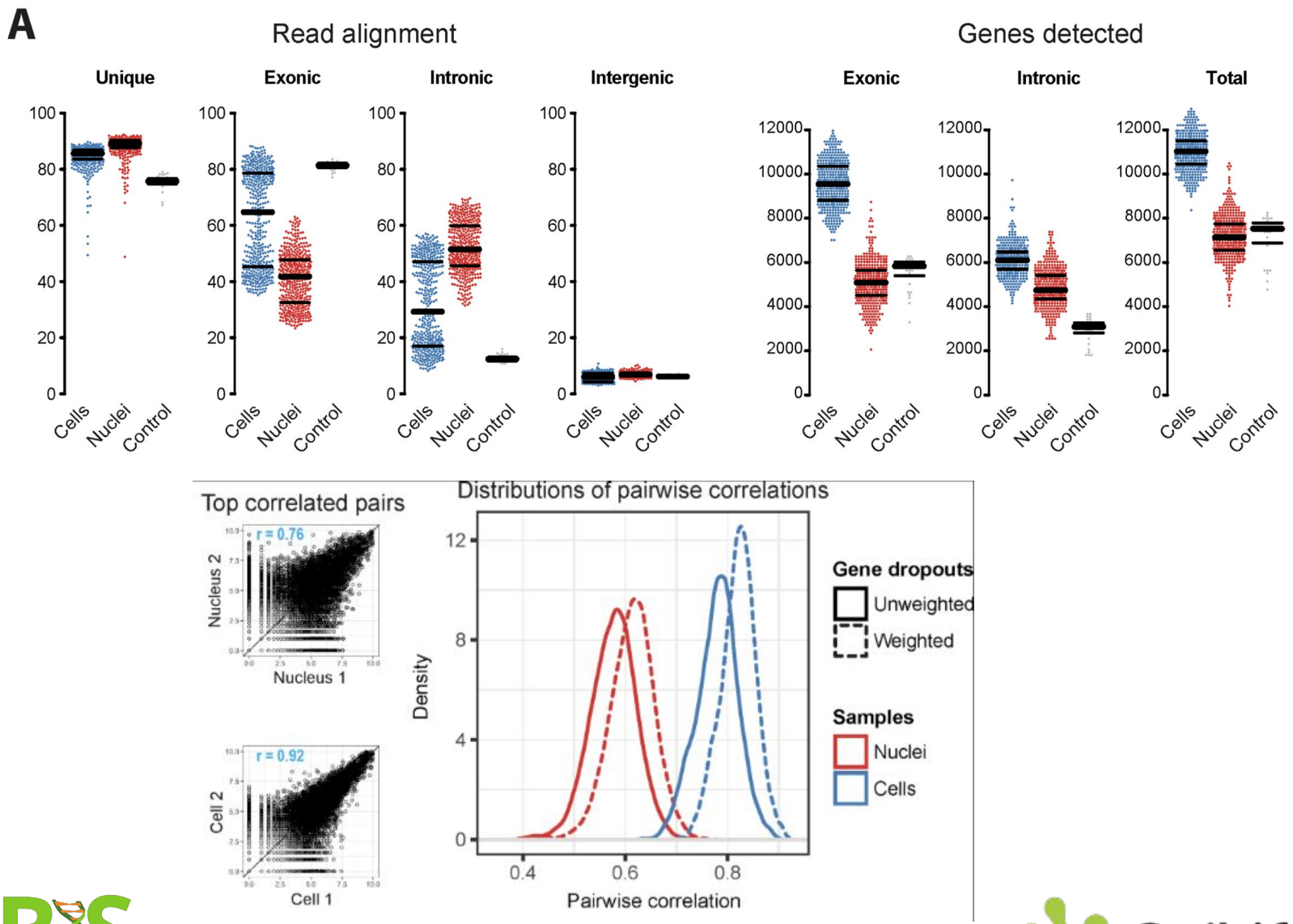
Single cell or single nuclei

- For some celltypes there may be biased detection of genes in nuclei vs cell.



(Thrupp et al. Cell Rep. 2020)

Single cell or single nuclei



(Bakken et al. Plos One 2018)

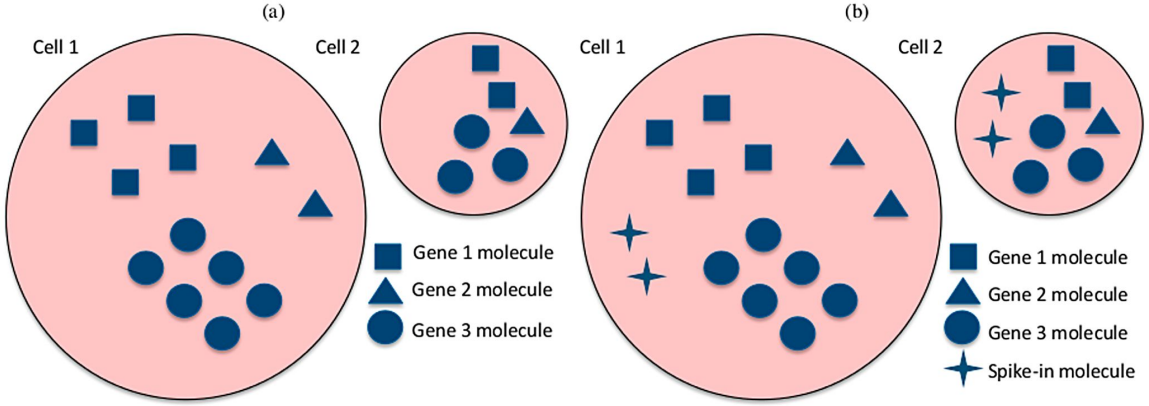
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Single cell or single nuclei

- Usually need to include intronic counts to increase transcript detection in snRNAseq
- Is default in Cellranger v7 (July 2022) also for scRNAseq

Spike-in RNAs

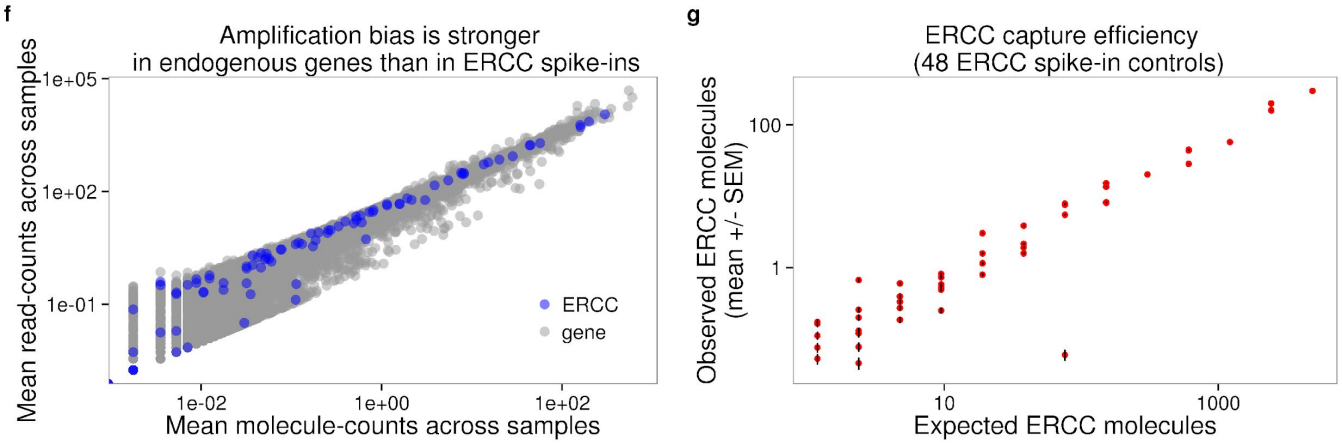


External molecules added in a known concentration.

- ERCC:
  – 92 bacterial RNA species, different lengths, GC contents
  – 22 abundance levels, 2 mixes for fold-change accuracy assessment
- SIRV:
  – 69 artificial transcripts
  – Mimic human genes
  – Used for isoforms detection

(Vallejos et al. *PLOS Comp Biol* 2015)

Spike-in RNAs

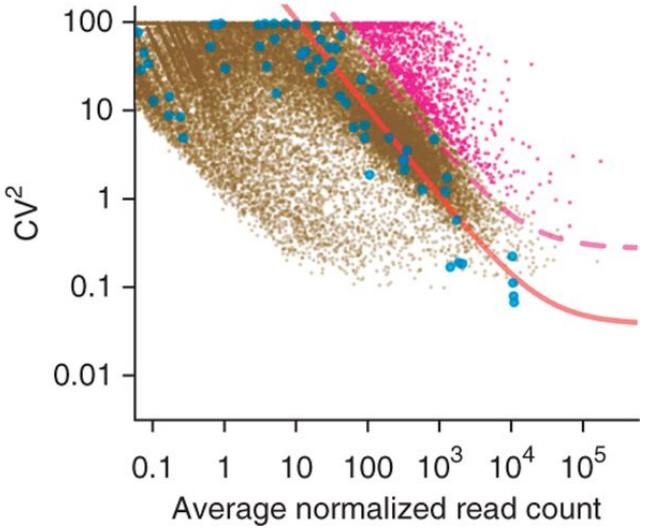


Spike-ins can be used to model:

- Technical noise
- Drop-out rates / capture efficiency
- Starting amount of RNA in the cell
- Data normalization

Problems:

- Spike-ins behave differently to endogenous genes
- Cannot be used in drop-seq methods

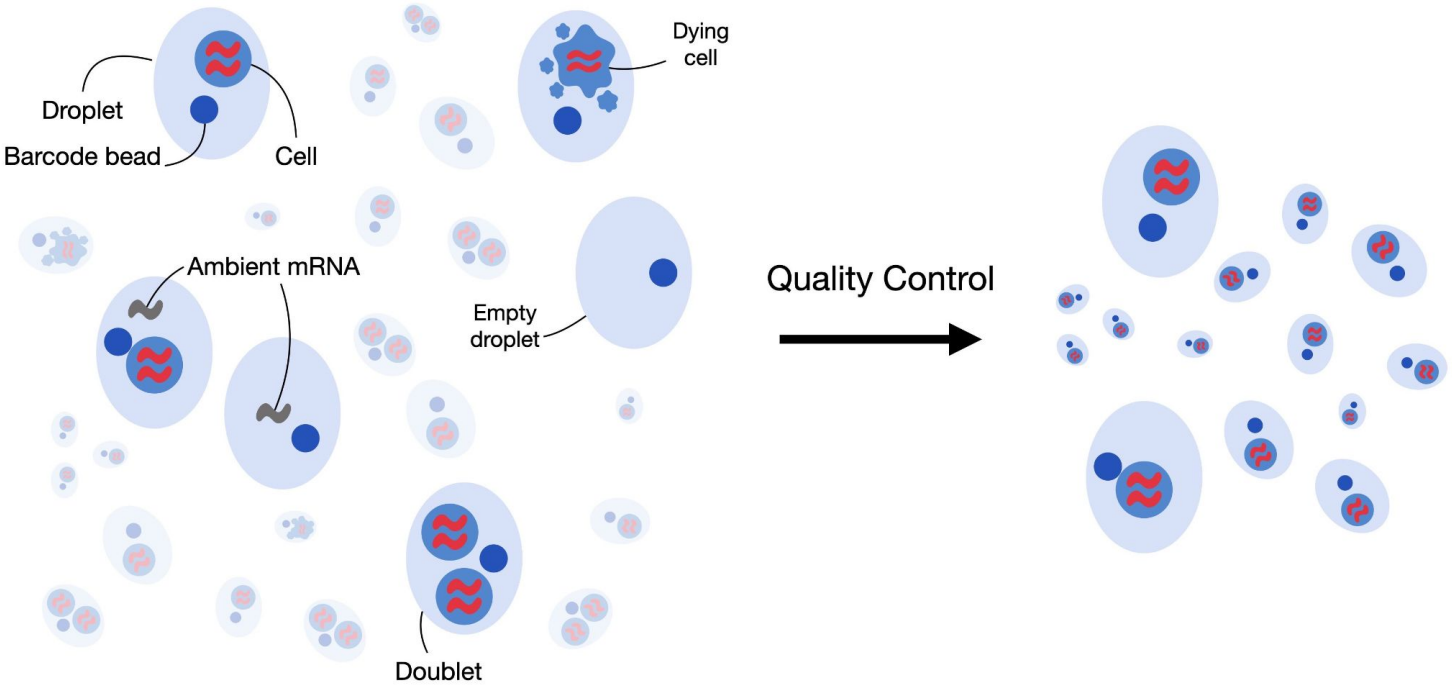


(Brennecke et al. *Nature Methods* 2013)
(Tung et al. *Scientific Reports* 2017)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

How do we define a failed vs successful cell capture and library prep?



(https://www.sc-best-practices.org/)

QC-metrics

- Mapping statistics (% uniquely mapping)
- Fraction of exon mapping reads
- 3’ bias – for full length methods like SS2
- mRNA-mapping reads
- Number of UMIs/reads
- Number of detected genes
- Spike-in detection
- Mitochondrial read fraction, ribosomal read fraction
- rRNA read fraction
- Pairwise correlation to other cells

QC-metrics

- Number of reads
- Mapping statistics (% uniquely mapping)
- Fraction of exon mapping reads
- mRNA-mapping reads (vs other types of genes like rRNA, sRNA, non coding, pseudogenes etc.)

Low number of reads – may not have enough information for that cell.

Bad mapping may be an indication of a failed library prep. Low content of mRNAs will lead to more primer dimers and more spurious mapping and fewer mapping reads.

QC-metrics

- Spike-in detection
- Spike-in ratio

If the number of spike-in molecules that are detected is low, this is a clearly failed library prep.

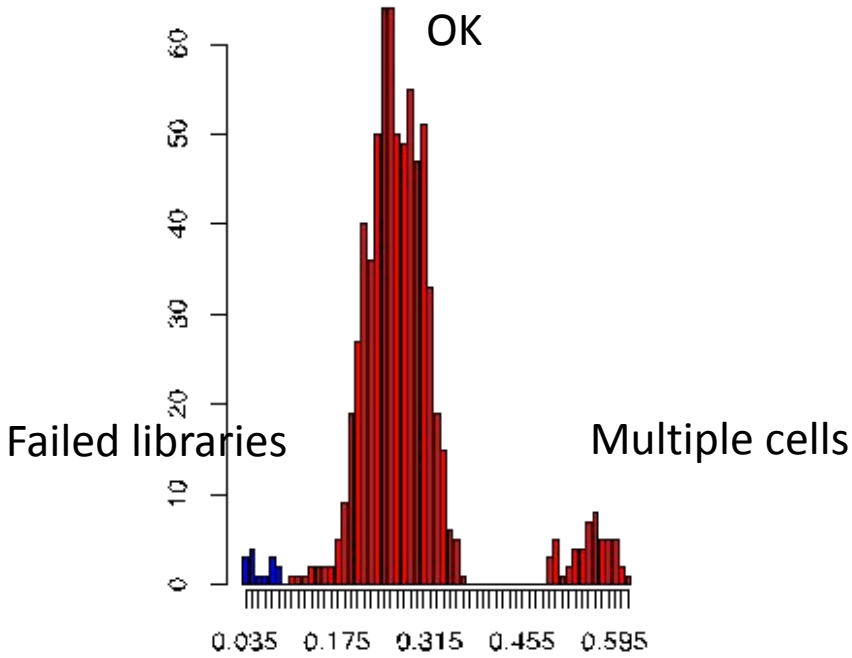
Proportion of cell to spike-in reads is an indication of the starting amount of RNA from the cell. Low amount of cell RNA can indicate breakage or just a smaller cell.

QC-metrics

– Number of detected genes

Number of detected genes clearly correlates to the size of the cells, so be careful if you are working with cells with very varying sizes.

High number of detected genes may be an indication of duplicate/multiple cells.
But can also be a larger celltype.

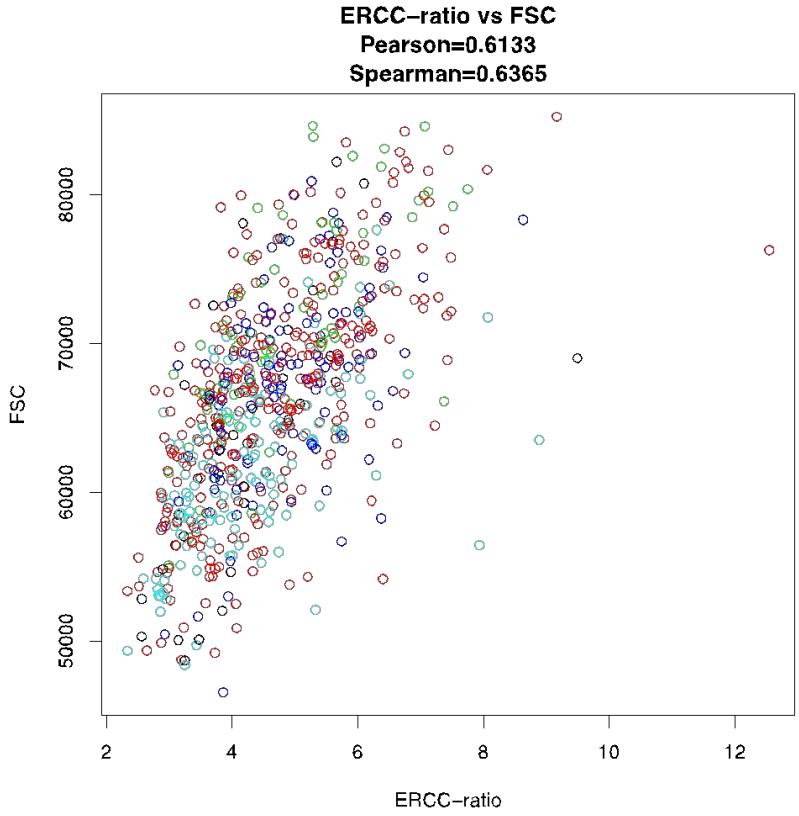


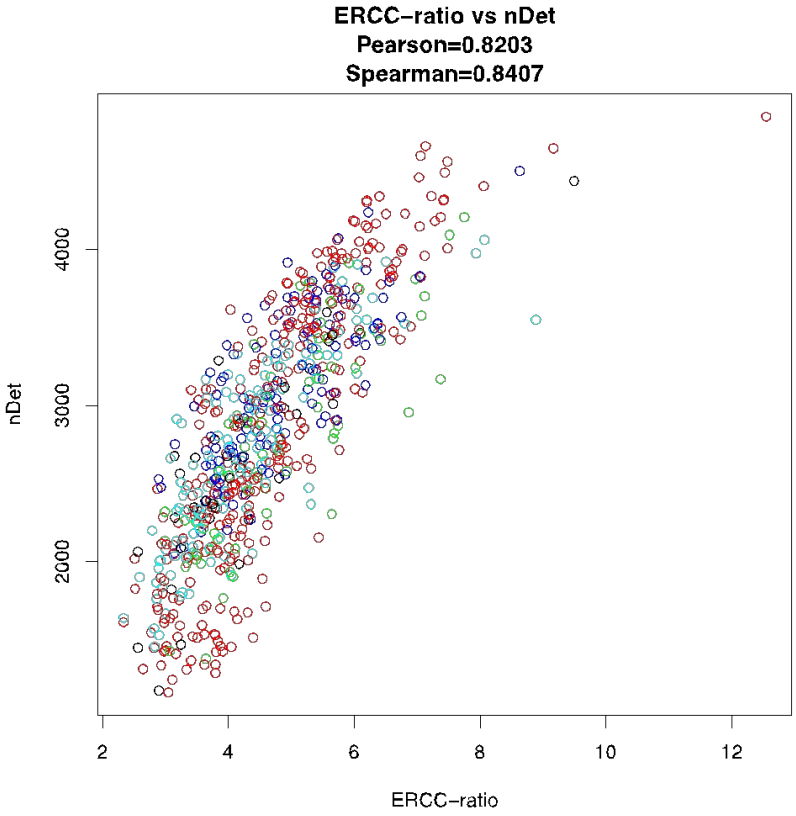
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

QC-metrics

- Cell size (by FCS), spike-in ratio and number of detected genes are clearly correlated





FCS – Forward Scattering from FACS

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

QC-metrics

- Mitochondrial read fraction

Suggested that when the cell membrane is broken, cytoplasmic RNA will be lost, but not RNAs inside the mitochondria.

High content of mitochondrial RNA may indicate apoptosis.

OBS! With snRNAseq mitochondrial reads are due to contamination, not related to quality.

QC-metrics

- Ribosomal RNA read fraction
- Ribosomal protein read fraction

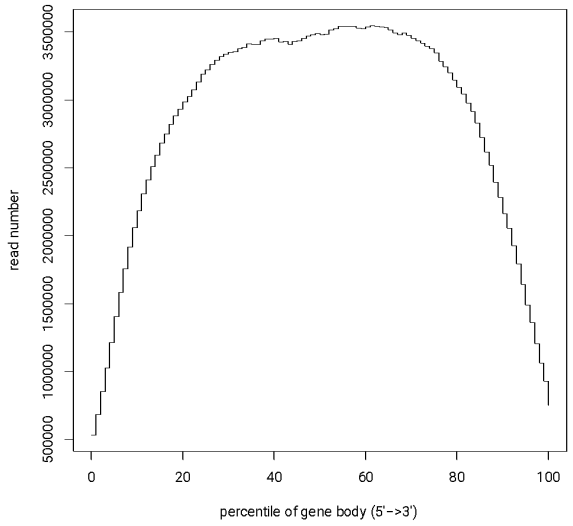
Possible that degradation of RNA leads to more templating of rRNA-fragments.

Proportion ribosomal proteins may be an artifact from handling of samples.

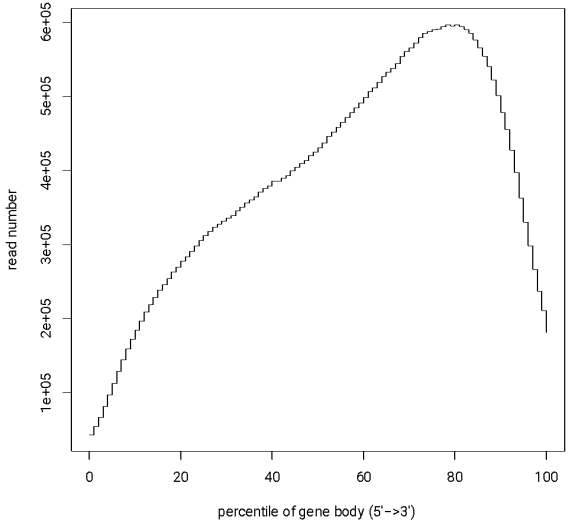
QC-metrics

– 3’ bias (degraded RNA) – for full length methods like Smartseq

Not degraded



Degraded



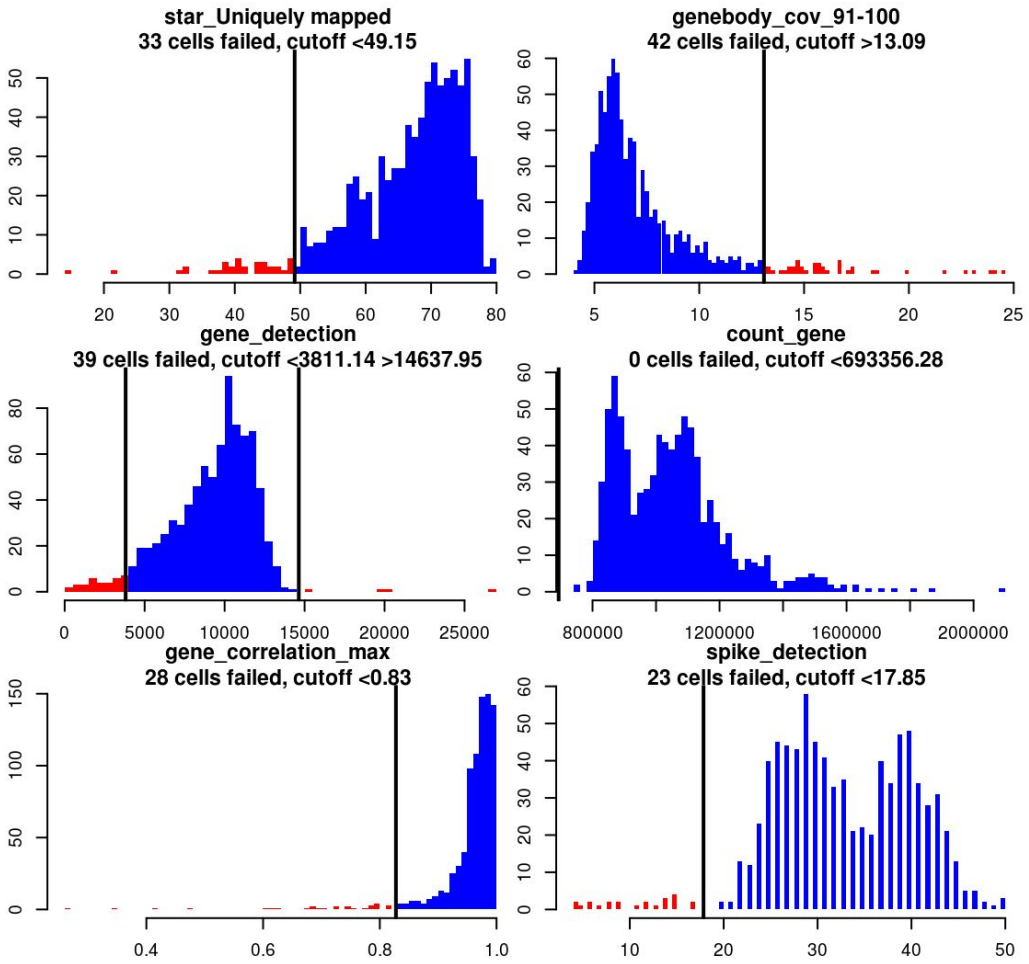
Look at proportion of reads that maps to the 10-20% most 3’ end of the transcript

QC-metrics

- Number of reads
- Mapping statistics (**% uniquely mapping**)
- Fraction of exon mapping reads
- mRNA-mapping reads
- 3’ bias – for full length methods like SS
- mRNA-mapping reads
- **Number of detected genes**
- **Spike-in detection**
- **Mitochondrial read fraction**
- rRNA read fraction
- Pairwise correlation to other cells

How to filter cells

- Normally, most of these qc-metrics will show the same trends, so it could be sensible to use a combination of measures.
- Look at the distributions before deciding on cutoffs.



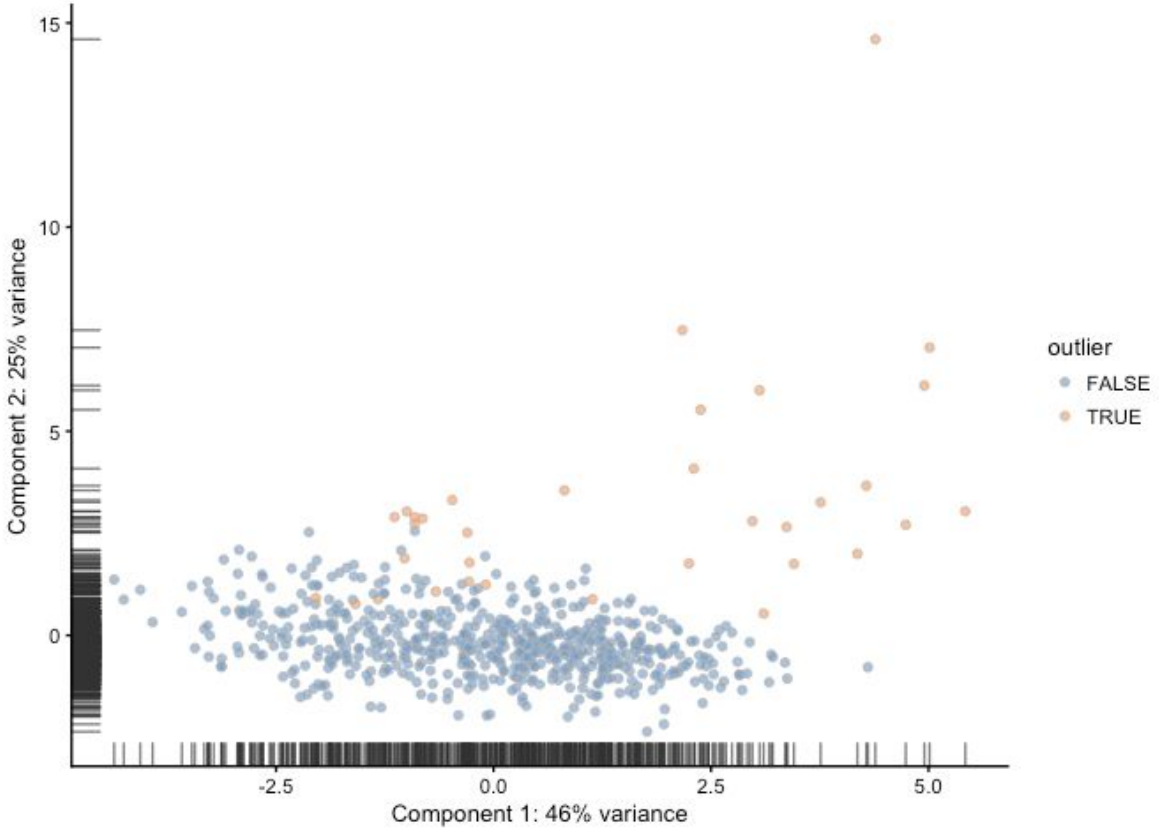
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

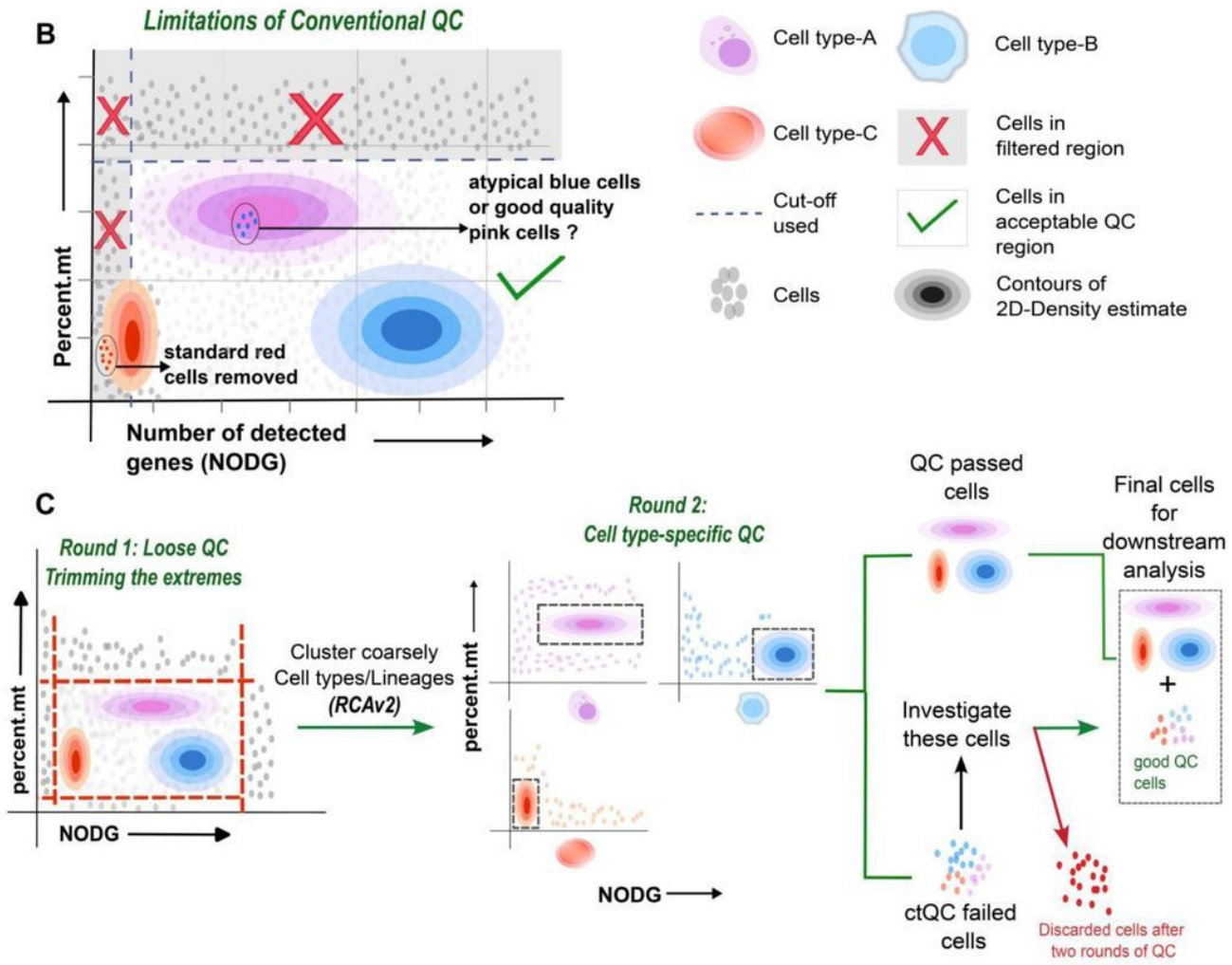
How to filter cells

- Can use PCA based on QC-metrics to identify outlier cells.

(Scater package)



Cluster based QC - ctQC

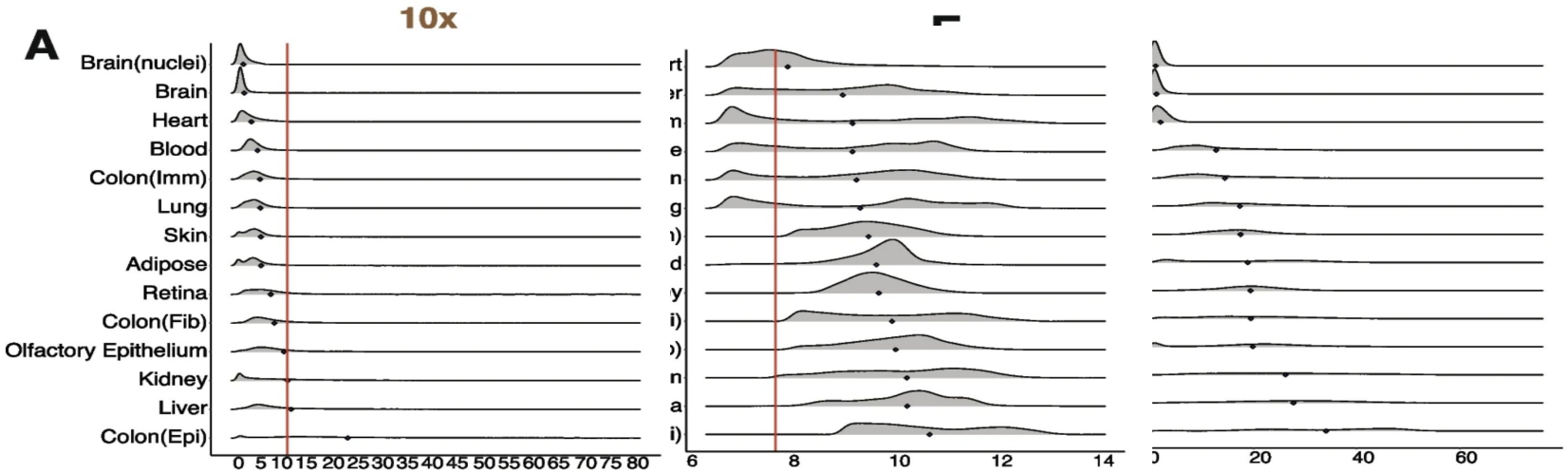


(Lakshmanan et al. BioRxiv 2024)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

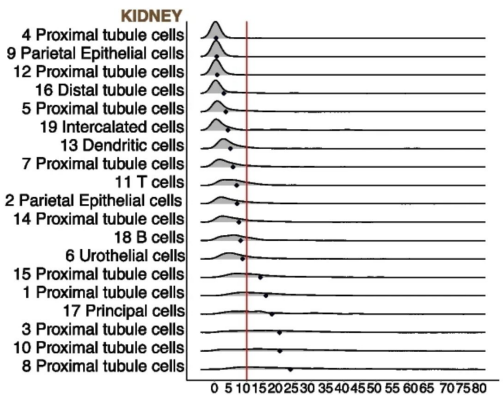
Cluster based QC - ddQC

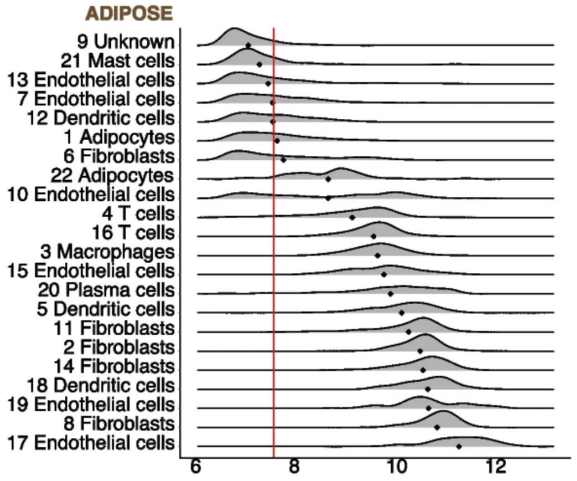


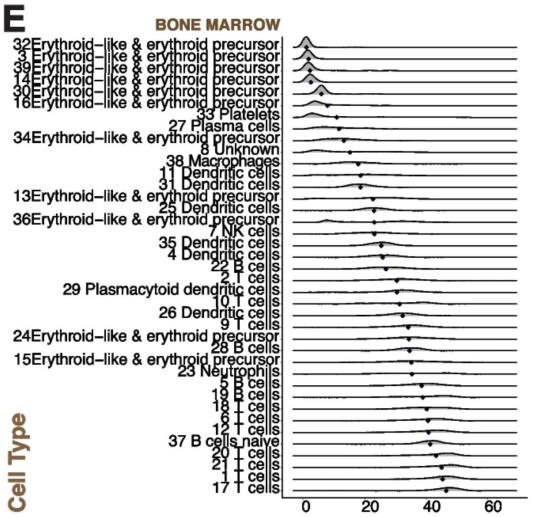
Fraction mitochondrial

Number of genes

Fraction ribosomal







(Subramanian et al. Genome Biology 2022)

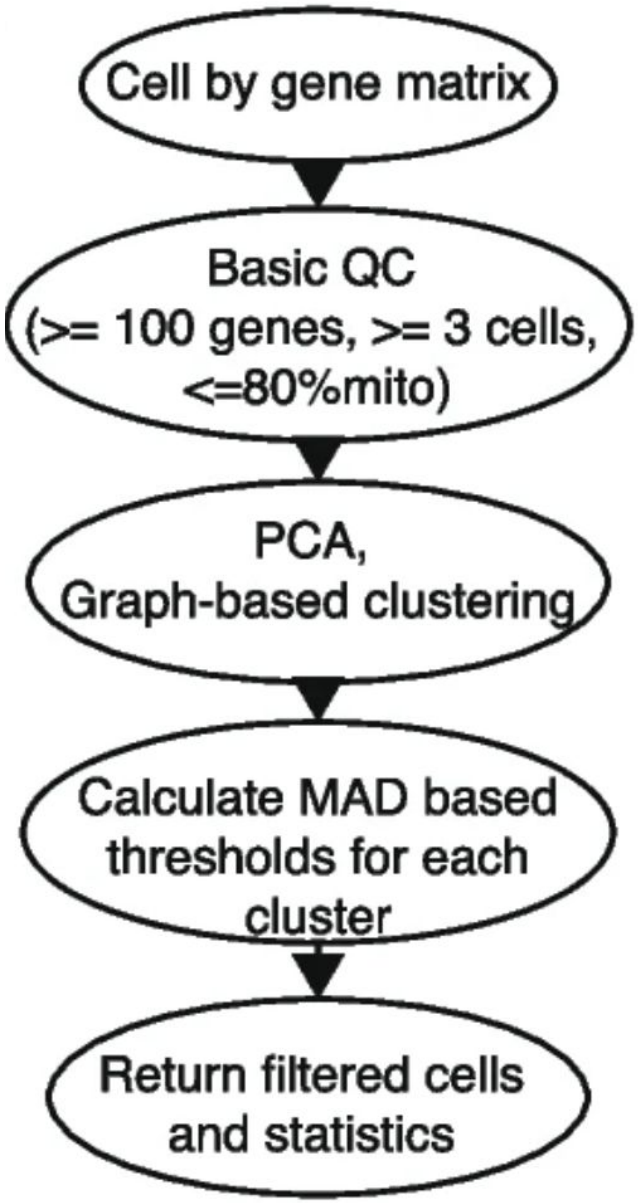
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Cluster based QC - ddQC

- First initial clustering
- Then define automatic cutoffs for each cluster

OBS! Depends strongly on first clustering being correct.



(Subramanian et al. Genome Biology 2022)

Deciding on cutoffs for filtering

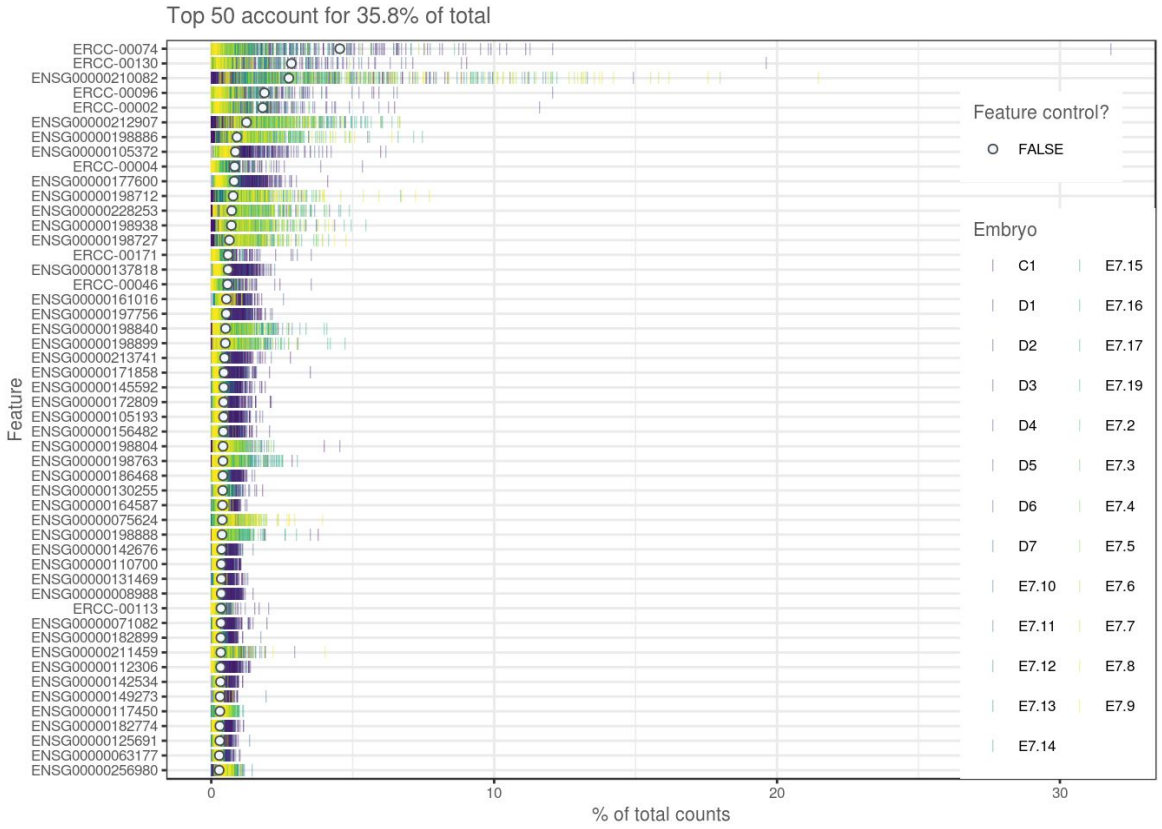
- Do you have a homogeneous population of cells with similar sizes?
- Is it possible that you will remove cells from a smaller celltype (e.g. red blood cells, immune cells) or a larger celltype (e.g. tumor cells)
- Examine PCA/tSNE before/after filtering and make a judgment on whether to remove more/less cells.
- Better to include too much than to filter out interesting celltypes in the first step.

Filter genes

- In most cases, all genes are not used in dimensionality reduction and clustering.
- Gene set selection based on:
  - Genes expressed in X cells over cutoff Y.
  - Variable genes – using spike-ins or whole distribution.
  - Filter out genes with correlation to few other genes
  - Prior knowledge / annotation
  - DE genes from bulk experiments
  - Top PCA loadings

Look at total contribution to expression

- Sometimes individual genes may have very high expression – may be problematic for normalization.

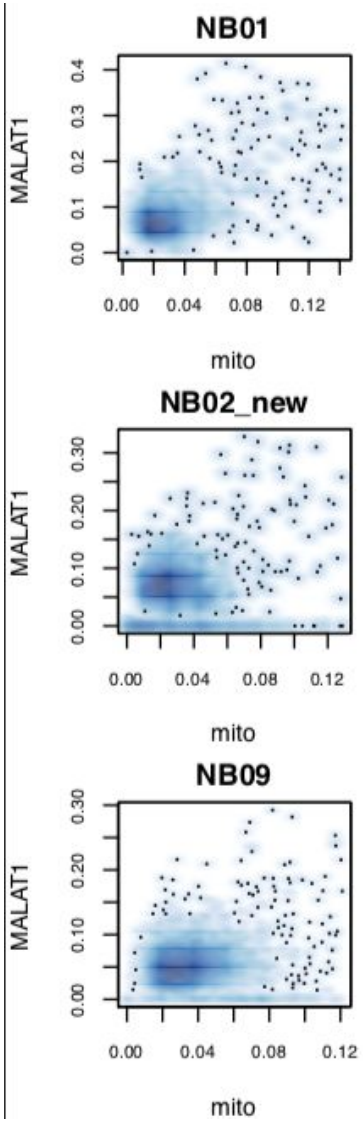


Look out for MALAT1 and other nuclear lincRNAs.
Mitochonrial or ribosomal genes, actin and hemoglobin.

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Look at total contribution to expression



- MALAT1 clearly correlates with percent mitochondrial genes in some samples.

Look out for MALAT1 and other nuclear lincRNAs.
Mitochonrial, ribosomal genes, actin and hemoglobin.

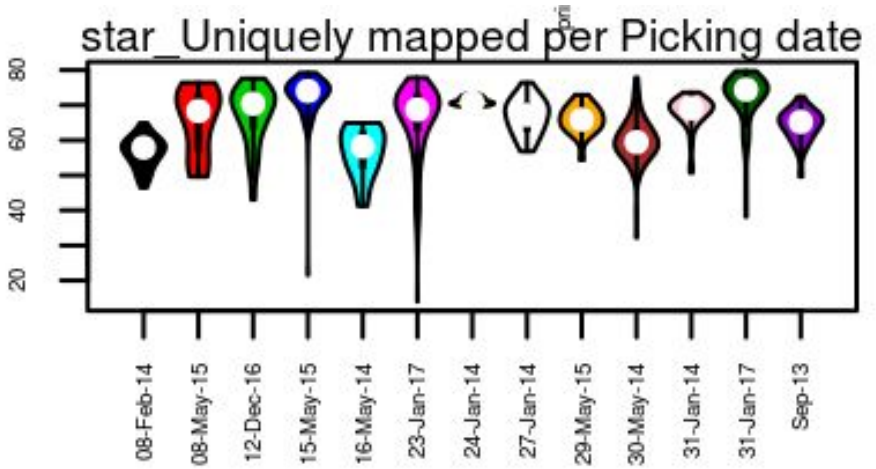
Removal of genes before analysis

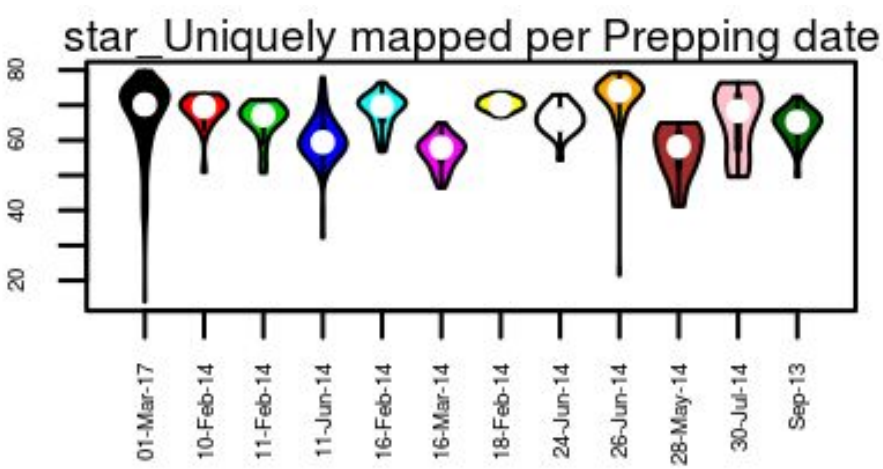
- Mitochondrial encoded genes – often mainly technical bias.
- Other genes suspected to be technical bias - often nuclear lincRNAs
- Genes that are too highly expressed
- Genes that may not contribute to celltype variation (e.g. ribosomal genes)

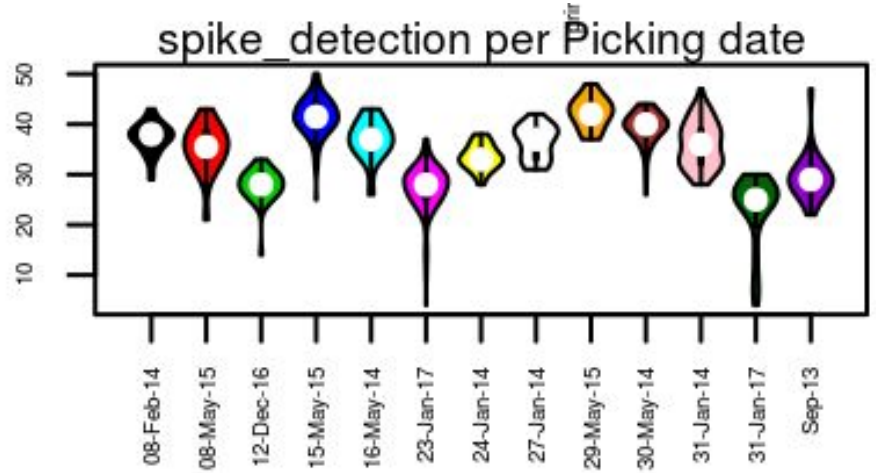
Batch effects

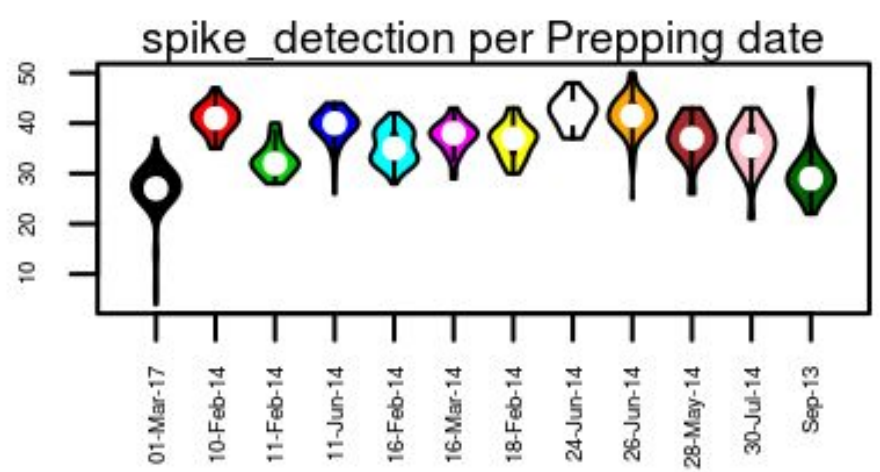
- Can be batch effects per
  - Experiment
  - Animal/Patient/Batch of cells
  - Sort plate
  - Sequencing lane
- Check if QC-measures deviates for any of those categories
- Check in PCA if any PC correlates to batches

Also check if your different qc-measures are different between batches.







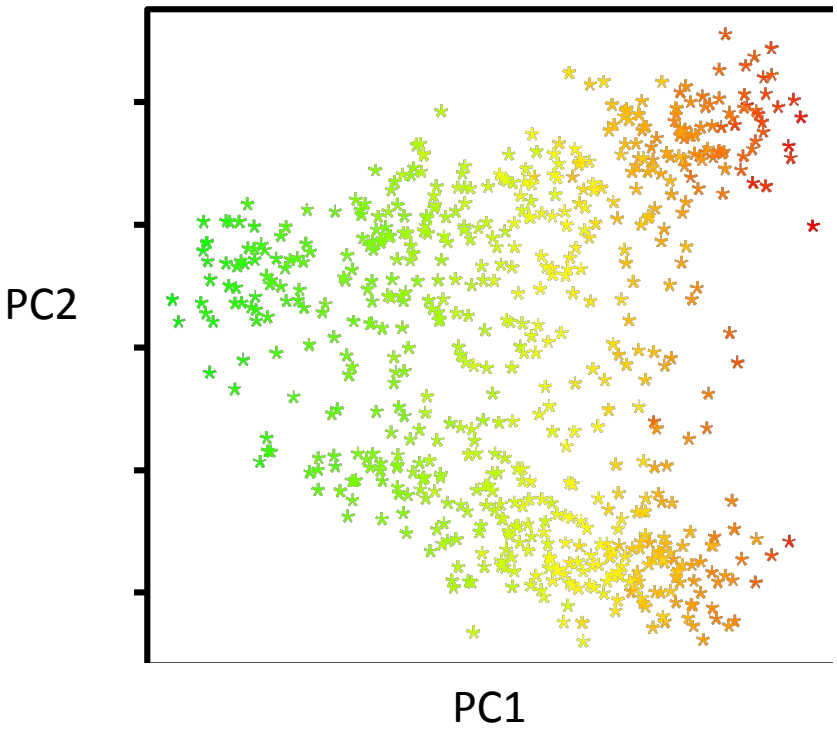


NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

PCA for QC

- One of the first PCs will (always?) correlate with number of detected genes

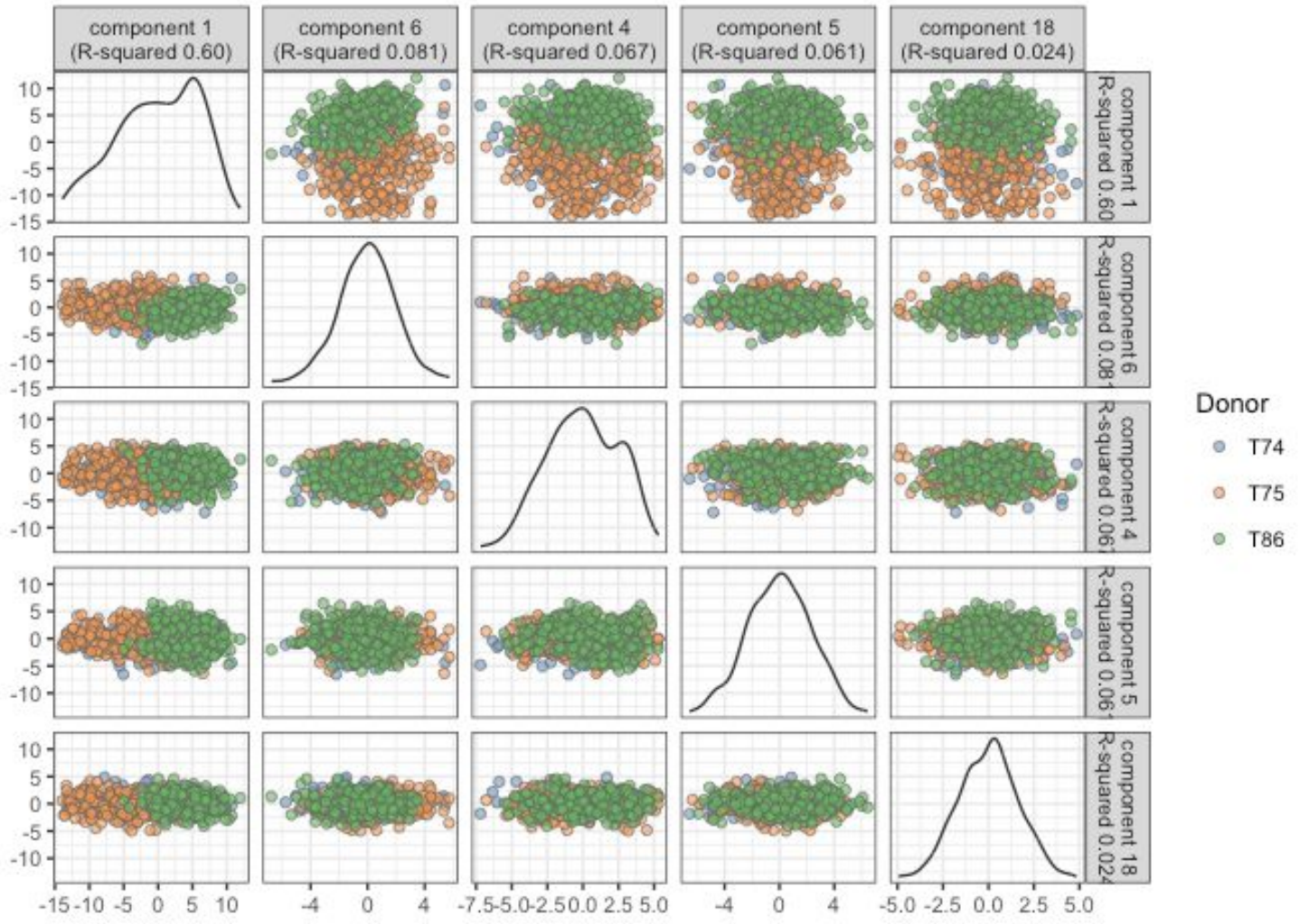


Red – high number of detected genes
Green - low number of detected genes

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Check for batch effects in PCA



Scater package

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

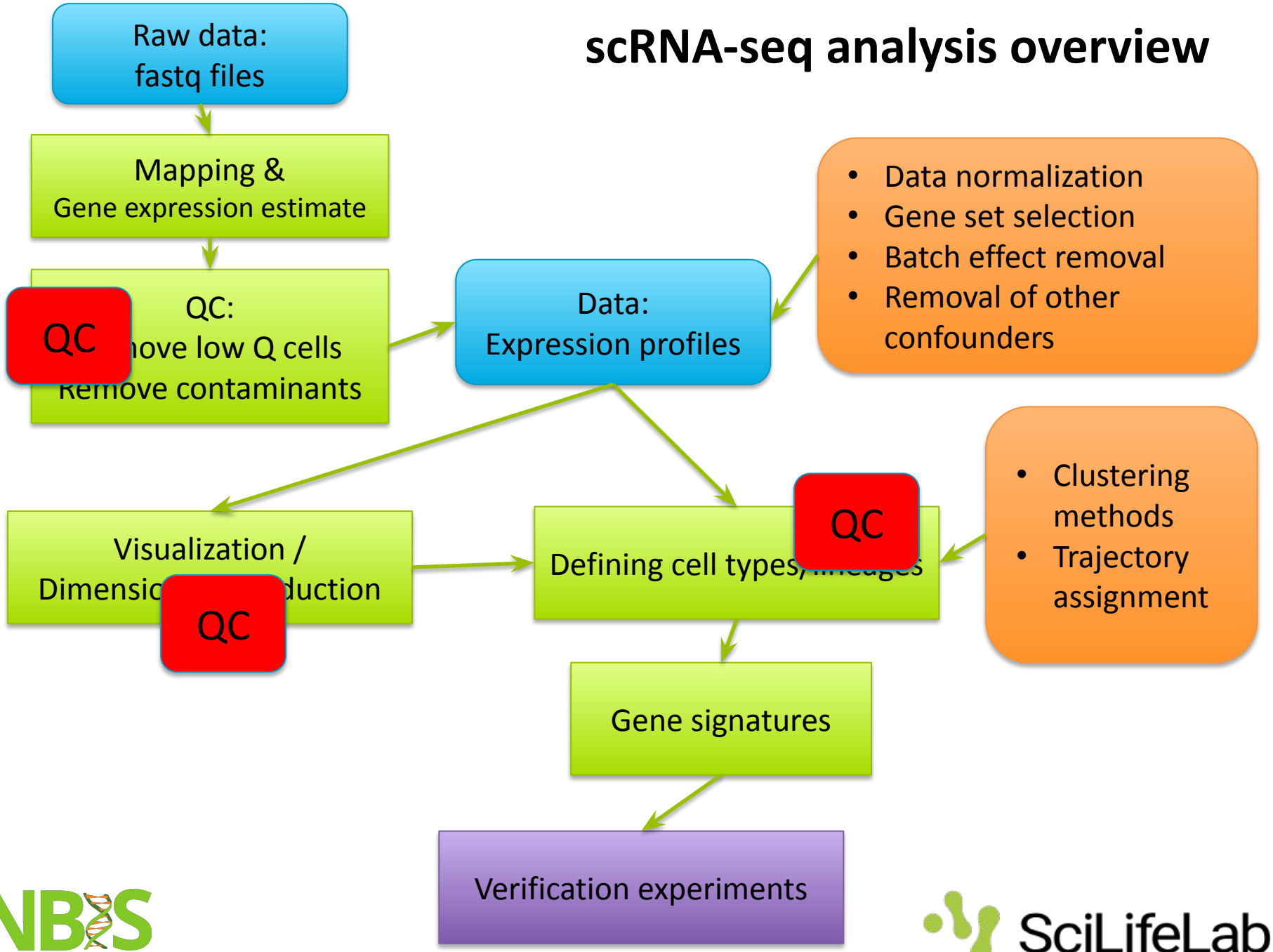
SciLifeLab

QC overview

1. a) Filter clearly failed libraries
   - low counts/detected genes,
   - high mito content

   b) Filter genes
2. Dimensionality reduction and clustering – check again for QC stats. Go back to 1. Possibly include more QC measures, filter genes more.
3. Iterate over 1 and 2 until results look good!

scRNA-seq analysis overview



NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Conclusions

- Try to plan your experiment in a way so that the biological signal you are looking for is not confounded by technical artifacts.
- Think about what distribution of cells you are expecting in your dataset when looking at the qc-measures. When you have homogeneous cells – deviant cells will be failed library. Otherwise be careful what you remove.
- Distinguishing duplicate cells is very hard, sometimes it will take some clustering first.