Quality Control of scRNAseq data

Åsa Björklund
asa.bjorklund@scilifelab.se

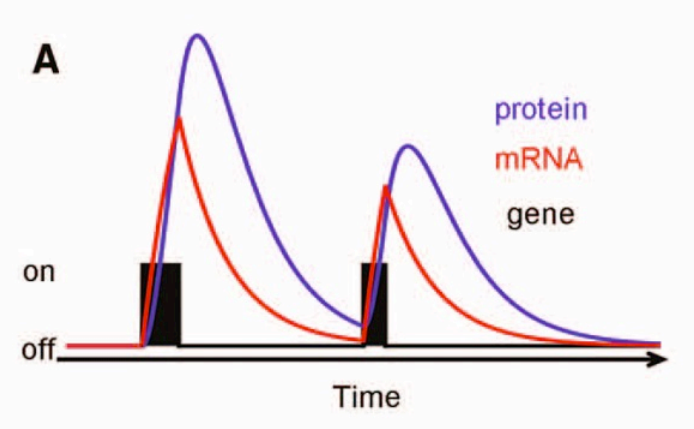
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

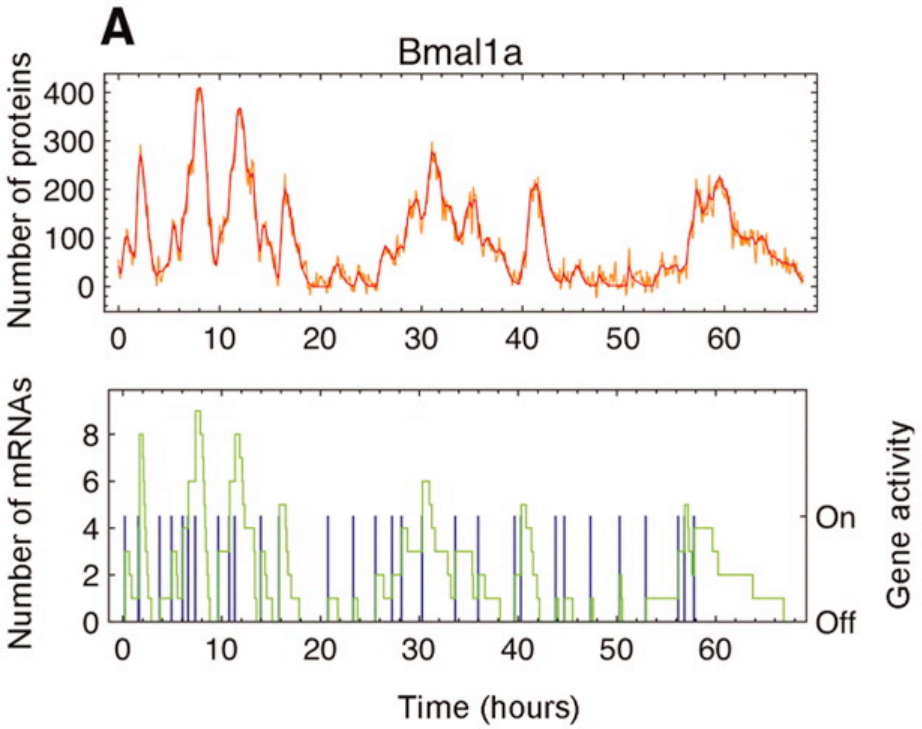
SciLifeLab

Outline

- Background on transcriptional bursting & drop-outs
- Experimental setup – what could go wrong?
- Spike-in RNAs
- Quality control metrics
- PCA for quality control

Transcriptional bursting

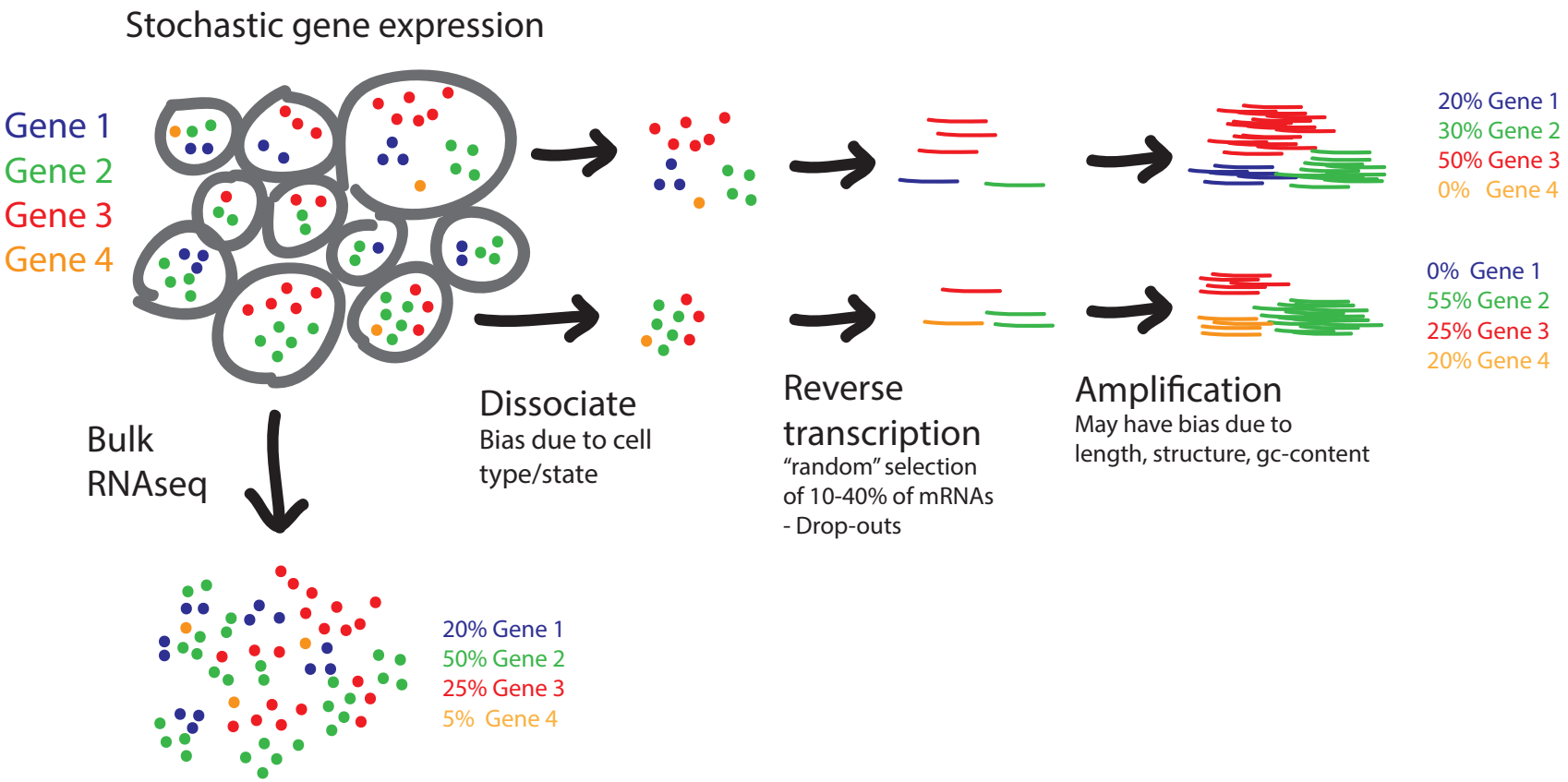




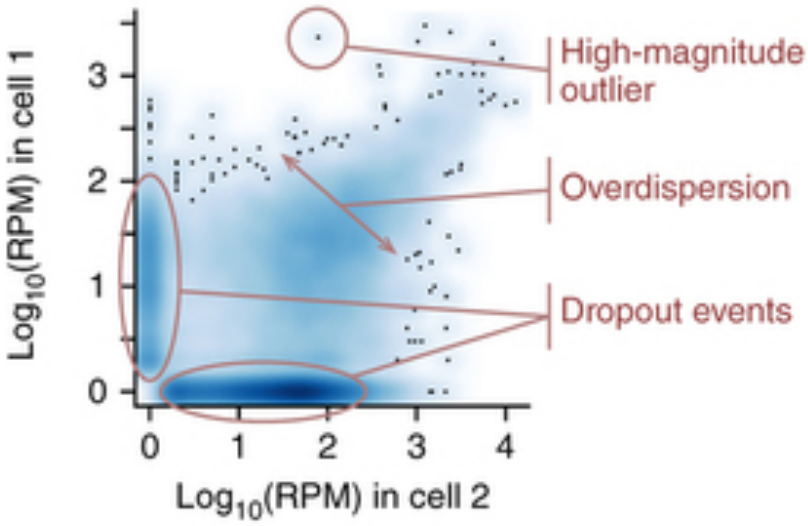
- Burst frequency and size is correlated with mRNA abundance
- Many TFs have low mean expression (and low burst frequency) and will only be detected in a fraction of the cells

(Suter et al. *Science* 2011)

Bursting, drop-outs and amplification bias



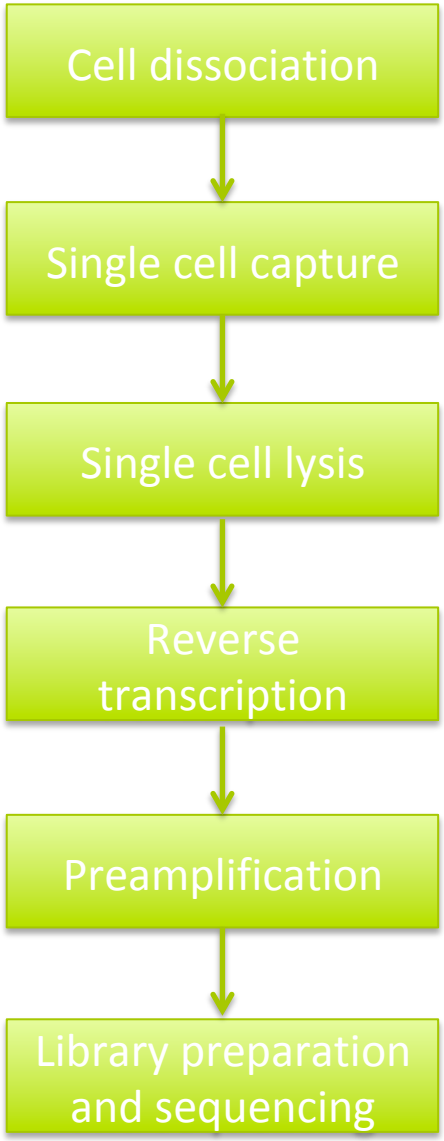
Problems compared to bulk RNA-seq



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

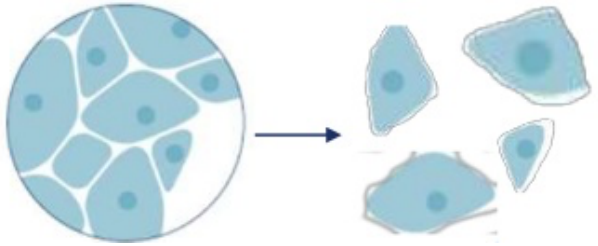
(Karchenko et al. *Nature Methods* 2014)

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. Tissues can be dissolved with mechanical methods, detergents or enzymatic digestion. Short time from dissociation to cell capture to reduce effect on transcriptional state.

Tissues that are hard to dissociate:

Laser capture microscopy (LCM)

Nuclei sorting

PROBLEMS:

- Incomplete dissociation can give multiple cells sticking together.
- To harsh lysis may damage the cells -> RNA degradation and RNA leakage, background RNA signal.
- Different lysis conditions may/may not give nuclear lysis.

(Kolodziejczyk et al. 2015)

Experimental setup

Cell dissociation

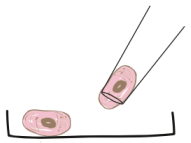
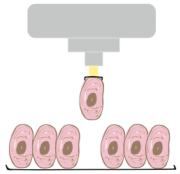
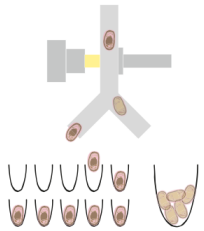
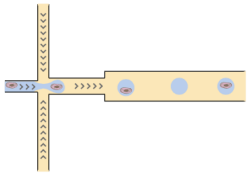
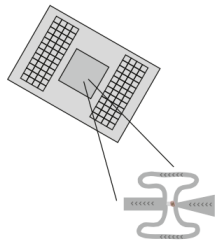
Single cell capture

Single cell lysis

Reverse transcription

Preamplification

Library preparation and sequencing

MICROPIPETTING MICROMANIPULATION	LASER CAPTURE MICRODISSECTION	FACS	MICRODROPLETS	MICROFLUIDICS e.g. FLUIDIGM C1
				
low number of cells	low number of cells	hundreds of cells	large number of cells	hundreds of cells
any tissue	any tissue	dissociated cells	dissociated cells	dissociated cells
enables selection of cells based on morphology or fluorescent markers	enables selection of cells based on morphology or fluorescent markers	enables selection of cells based on size or fluorescent markers	no selection of cells (can presort with FACS)	no selection of cells (can presort with FACS)
visualisation of cells	visualisation of cells	fluorescence and light scattering measurements	fluorescence detection	visualisation of cells
time consuming	time consuming	fast	fast	fast
reaction in microliter volumes	reaction in microliter volumes	reaction in microliter volumes	reaction in nanoliter volumes	reaction in nanoliter volumes

PROBMLEMS:

- All these 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.
- Long time for sorting may damage the cells

(Kolodziejczyk et al. 2015)

Experimental setup

Cell dissociation

Single cell capture

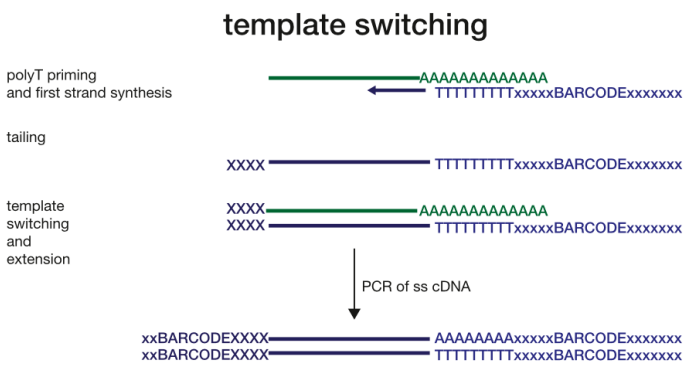
Single cell lysis

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 90-60% depending on the method used.

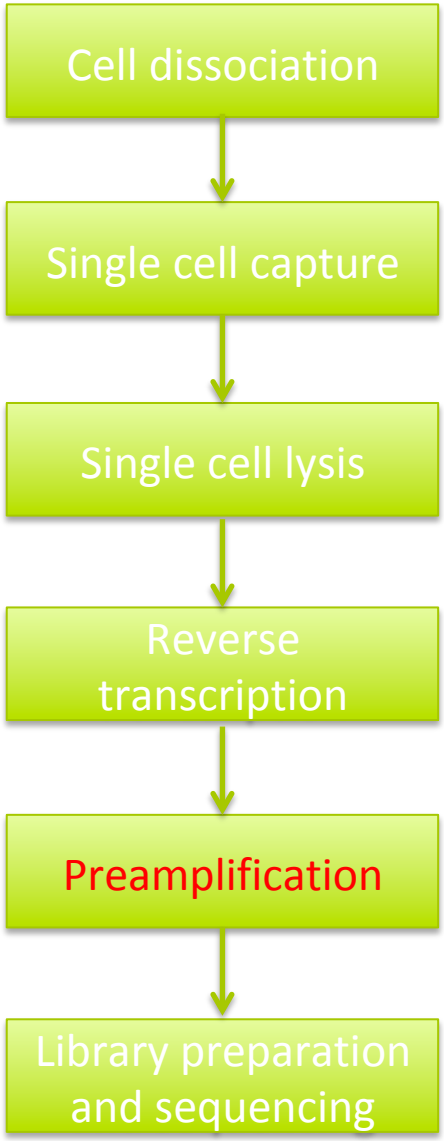
Two libraries with the same method using the same cell type may have very different drop-out rates.

(Kolodziejczyk et al. 2015)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Experimental setup



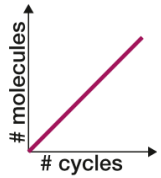
PCR

- 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.

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 effected by a number of factors:

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

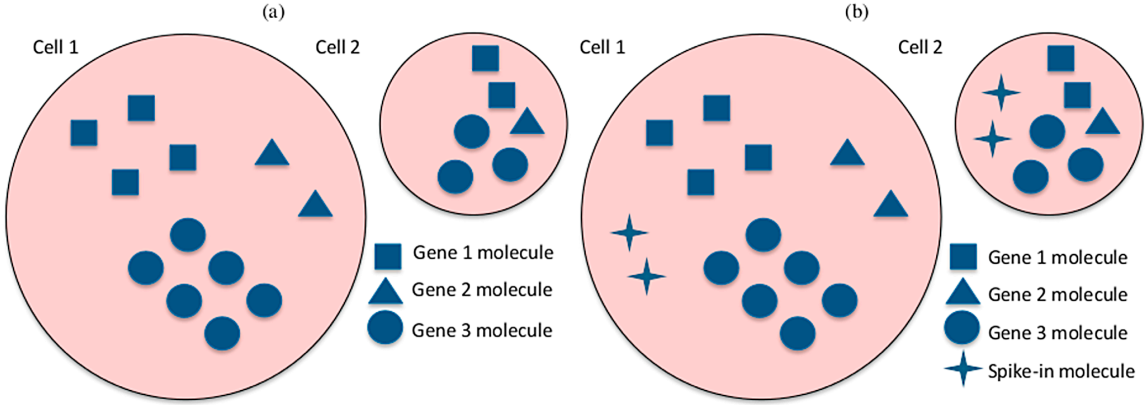
Index swapping

(Kolodziejczyk et al. 2015)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Spike-in RNAs

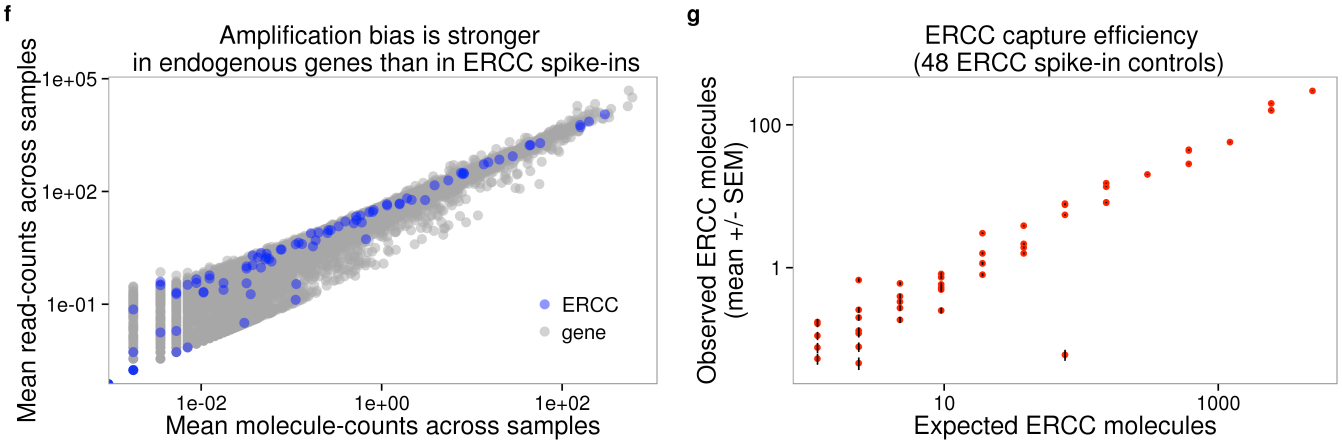


External molecules added in a known concentration.

- 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

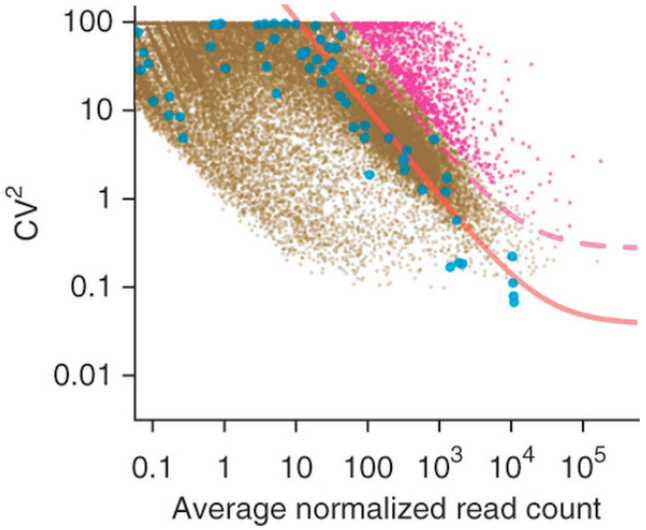


Spike-ins can be used to model:

- 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

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
- 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.)

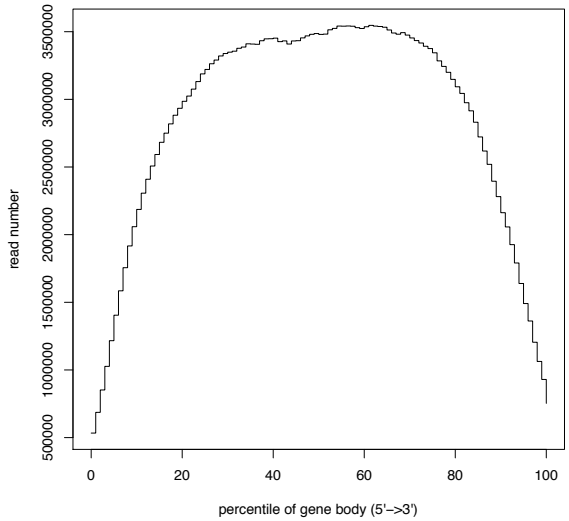
Low number of reads – may not have enough information for that cell.

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.

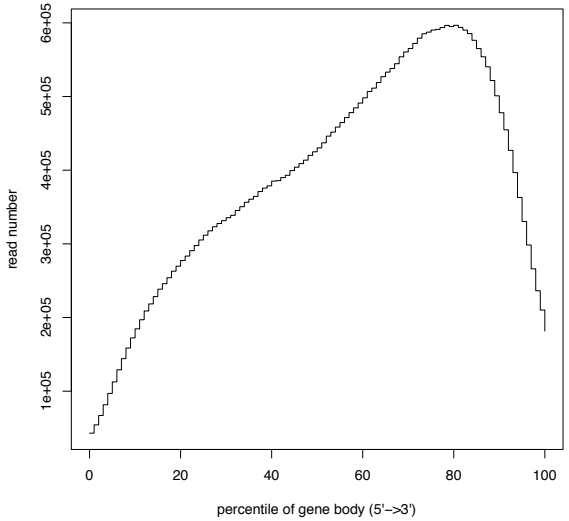
QC-metrics

– 3’ bias (degraded RNA) – for full length methods like SS2

Not degraded



Degraded



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

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

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.

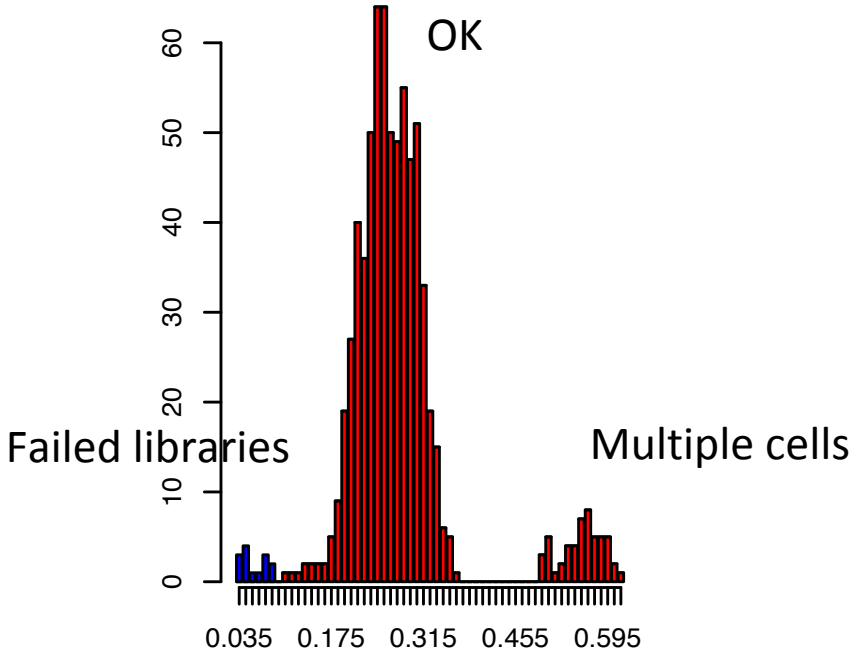
Proportion of cell to spike-in reads is an indication of the starting amount of RNA from the cell. Low amount of cell RNA can indicate breakage or just a smaller cell.

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.

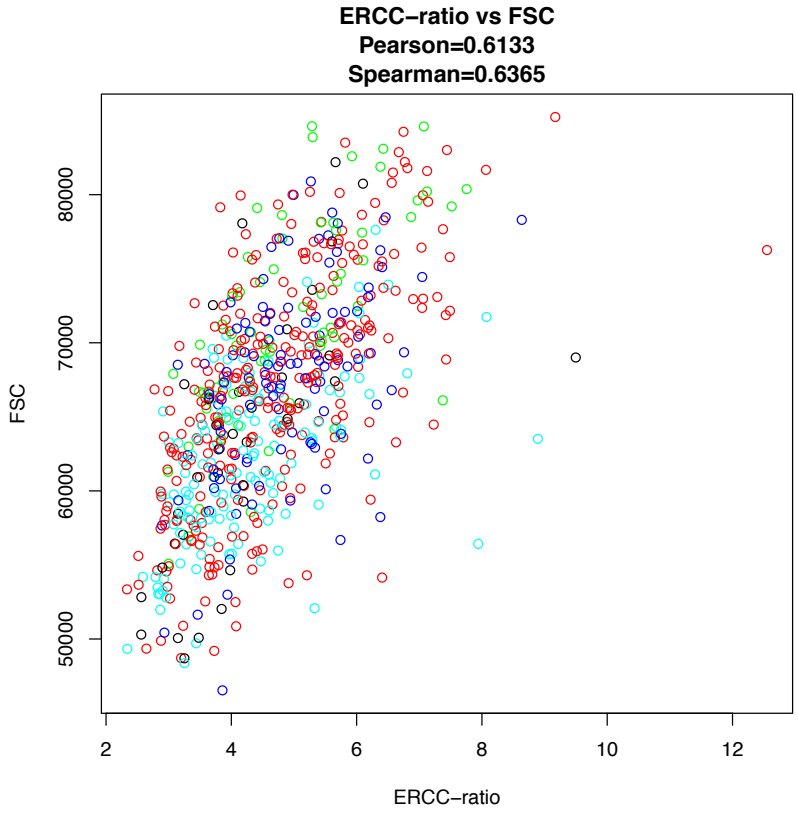


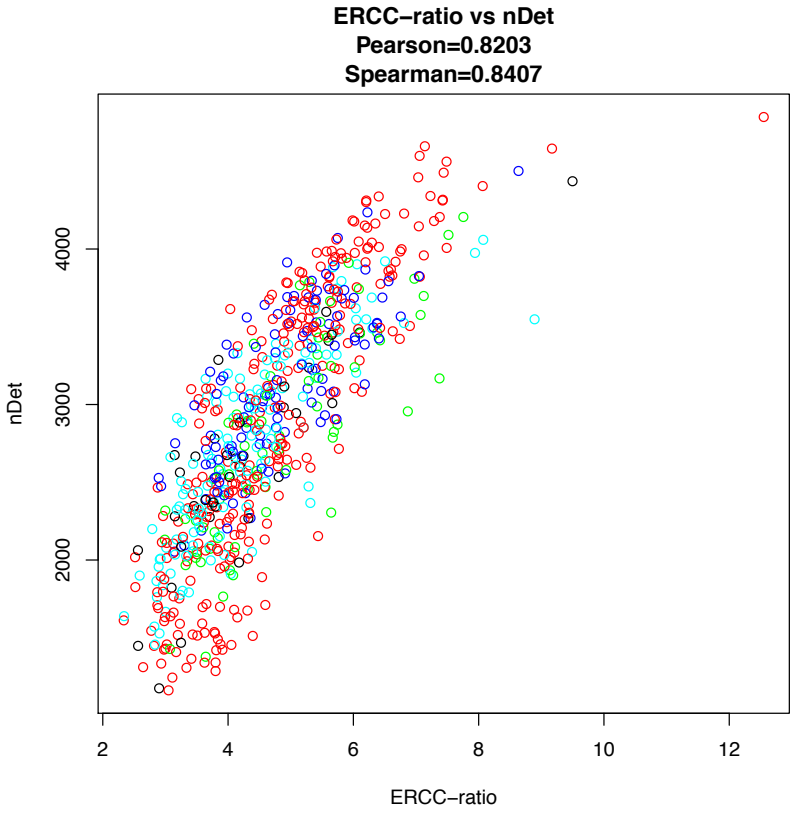
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

QC-metrics

- Cell size, spike-in ratio and number of detected genes are clearly correlated





FCS − Forward Scattering from FACS

QC-metrics

- Mitochondrial read fraction
- rRNA read fraction

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

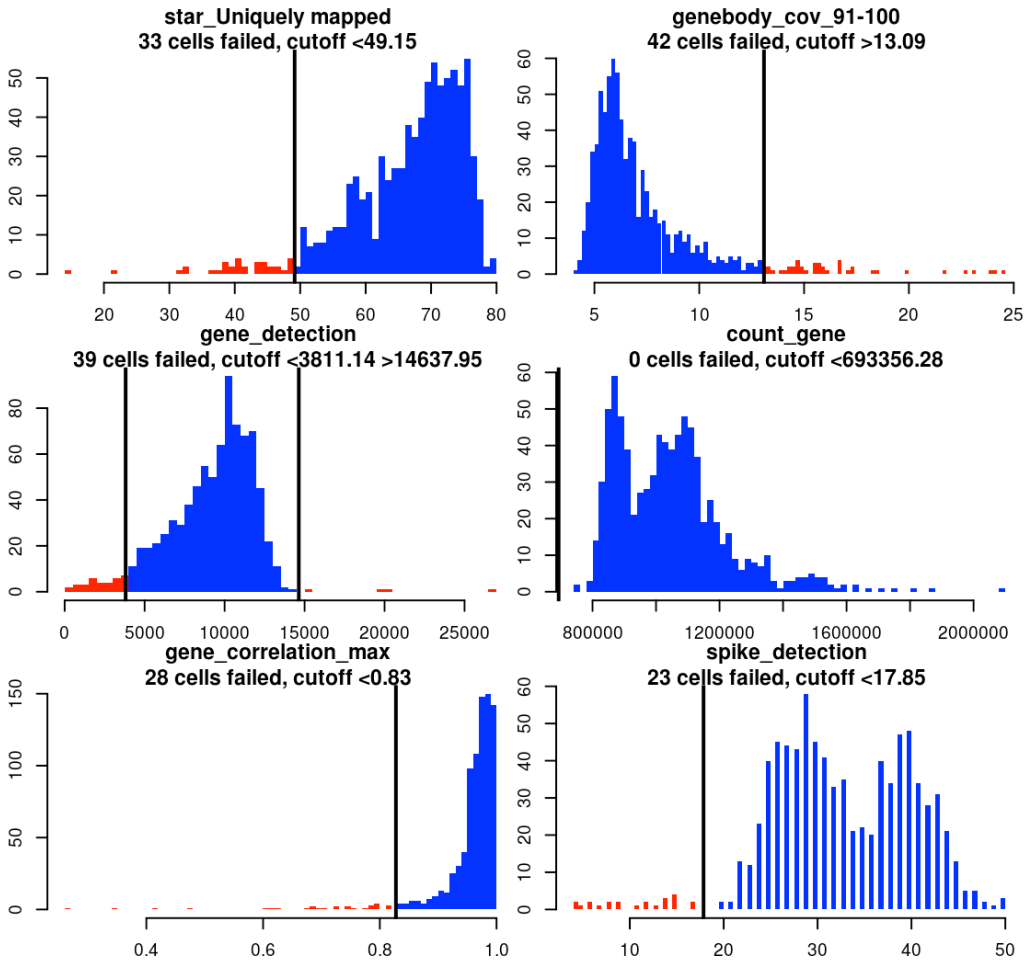
Possible that degradation of RNA leads to more templating of rRNA-fragments.

QC-metrics

- Number of reads
- Mapping statistics (**% uniquely mapping**)
- Fraction of exon mapping reads
- mRNA-mapping reads
- 3’ bias – for full length methods like SS2
- 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

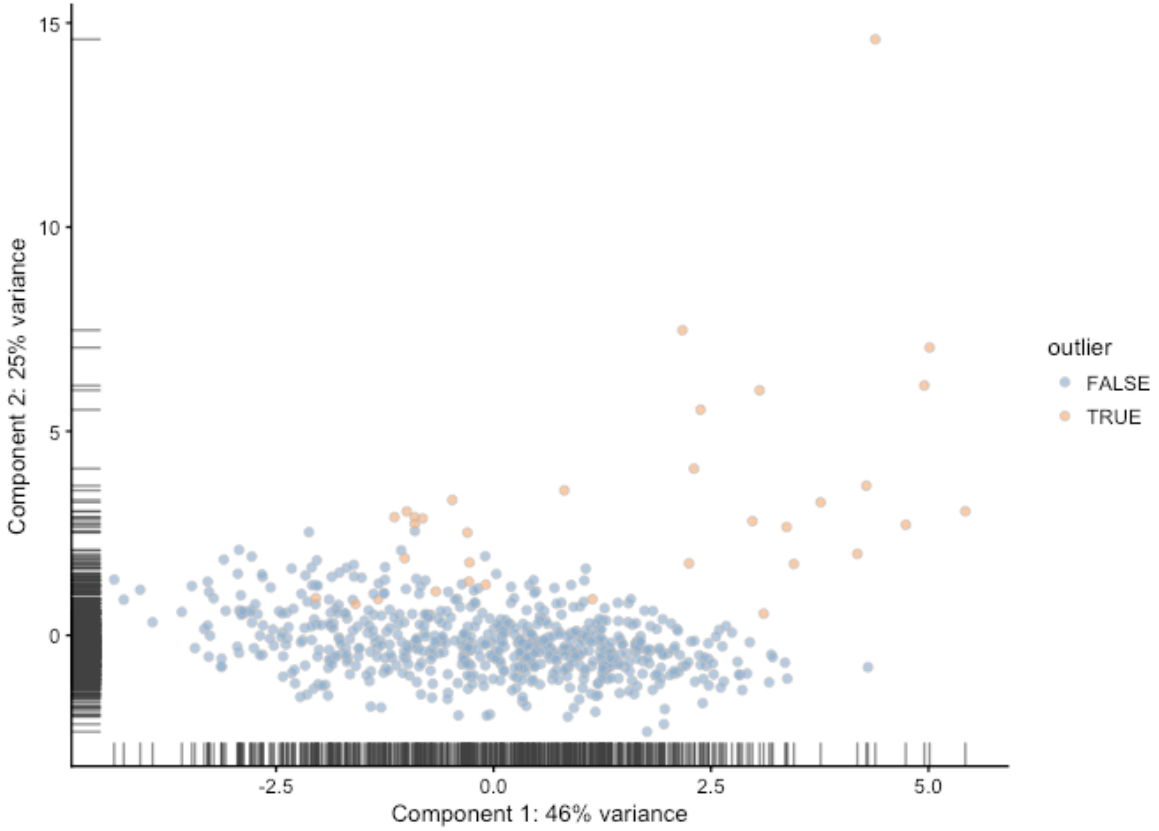
- Normally, most of these qc-metrics will show the same trends, so it could be sensible to use a combination of measures.
- Look at the distributions before deciding on cutoffs.



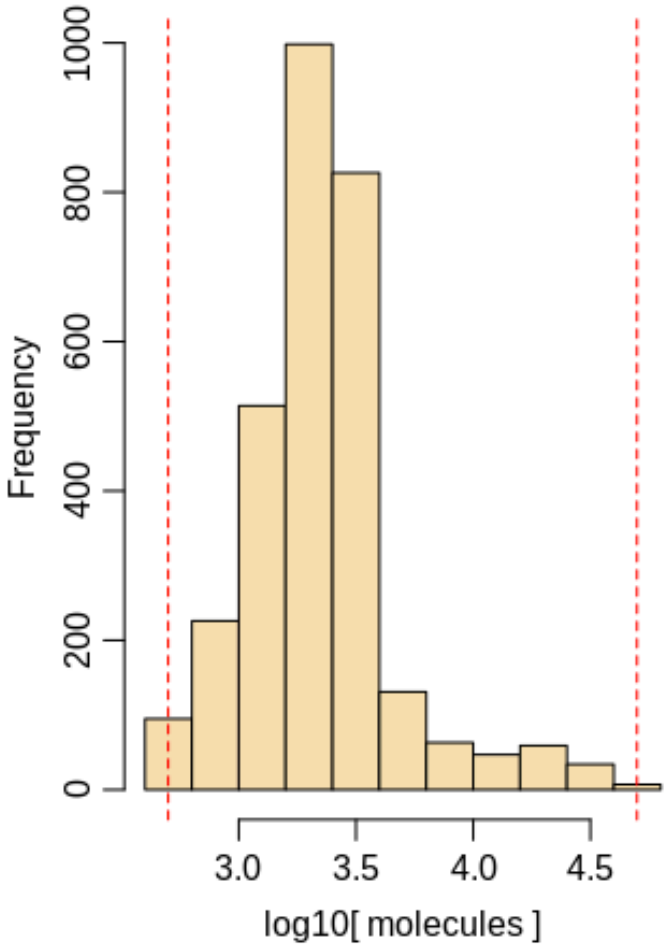
How to filter cells

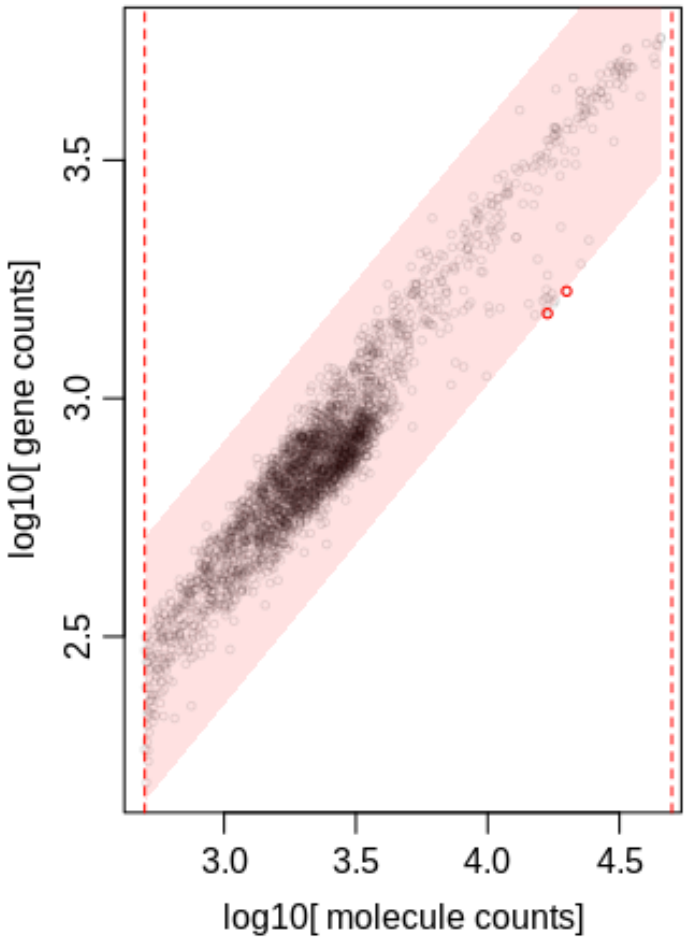
- Can use PCA based on QC-metrics to identify outlier cells.

(Scater package)



nUMI vs nGene





NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Deciding on cutoffs for filtering

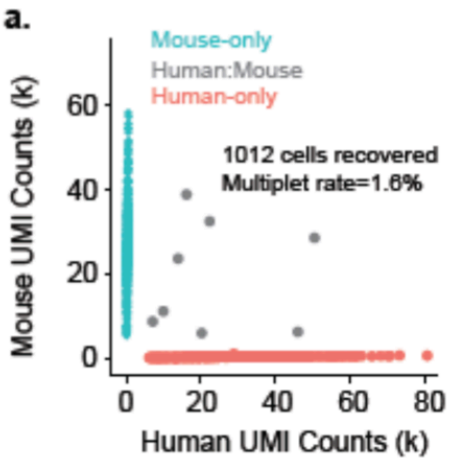
- Do you have a homogeneous population of cells with similar sizes?
- Is it possible that you will remove cells from a smaller celltype (e.g. red blood cells)
- Examine PCA/tSNE before/after filtering and make a judgment on whether to remove more/less cells.

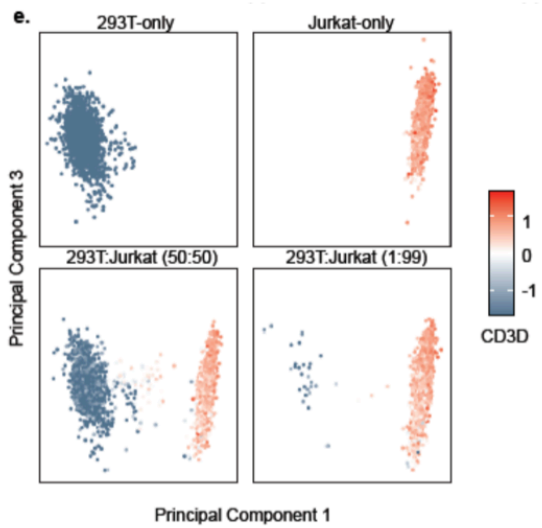
Detecting duplicate/multiple cells

- High number of detected genes or UMIs – can be a sign of multiples
  - But, beware so that you do not remove all cells from a larger celltype.
- After clustering – check if you have cells with signatures from multiple clusters.
- A combination of those 2 features would indicate duplicates.
- With 10X you should have a feeling for your doublet rate based on how many cells were loaded

Doublets in 10x

scRNA-seq is not always single-cell





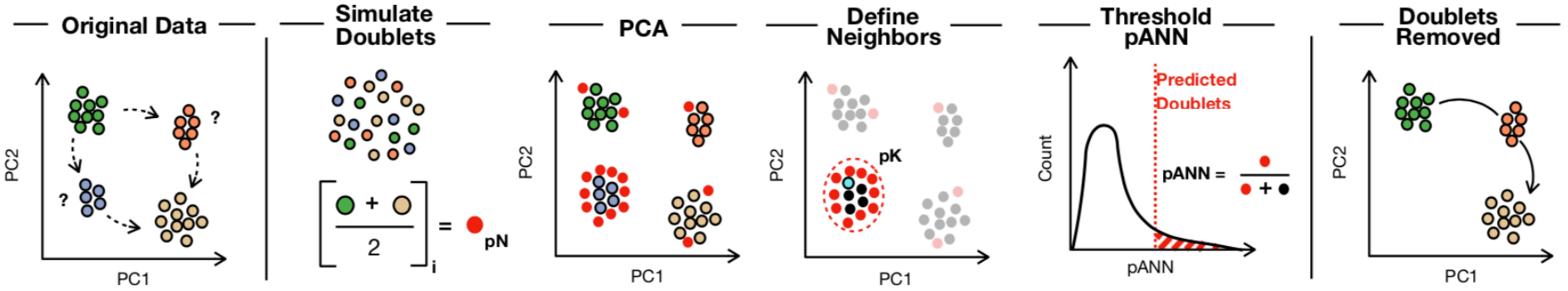
Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~870	~500
~0.8%	~1700	~1000
~2.3%	~5300	~3000
~3.9%	~8700	~5000
~7.6%	~17400	~10000

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Doublet detectors

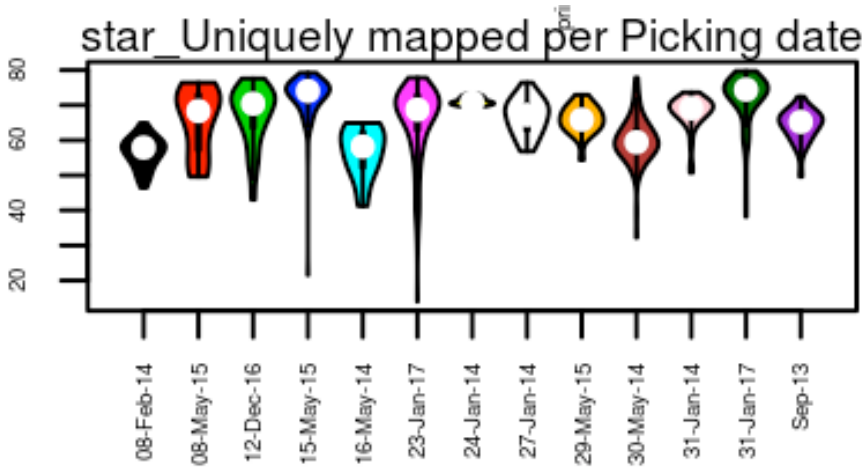
- DoubletFinder - https://github.com/chris-mcginnis-ucsf/DoubletFinder
- Scrublet - https://github.com/AllonKleinLab/scrublet
- DoubletDecon - https://github.com/EDePasquale/DoubletDecon

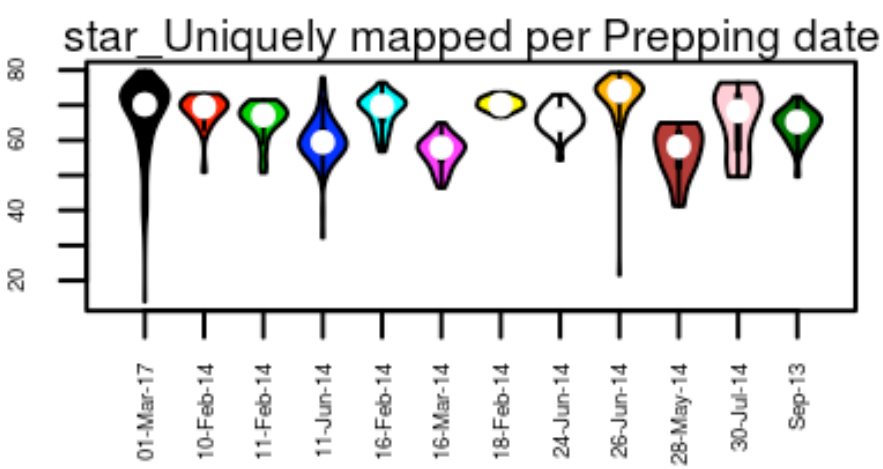


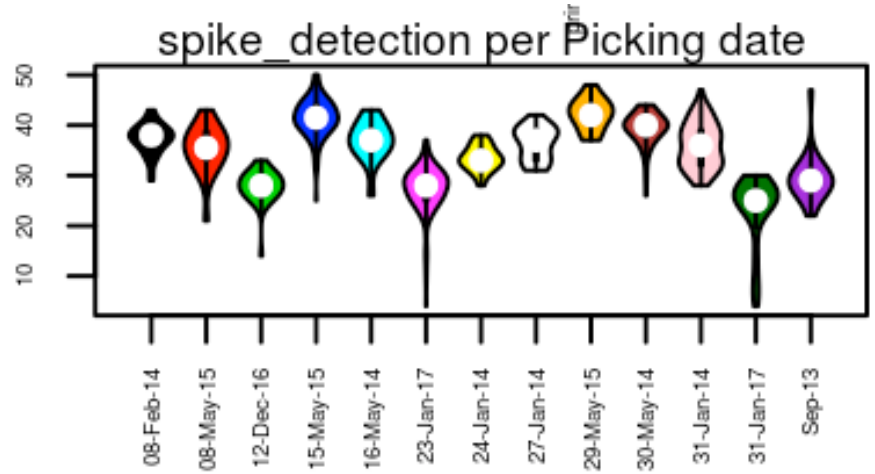
Batch effects

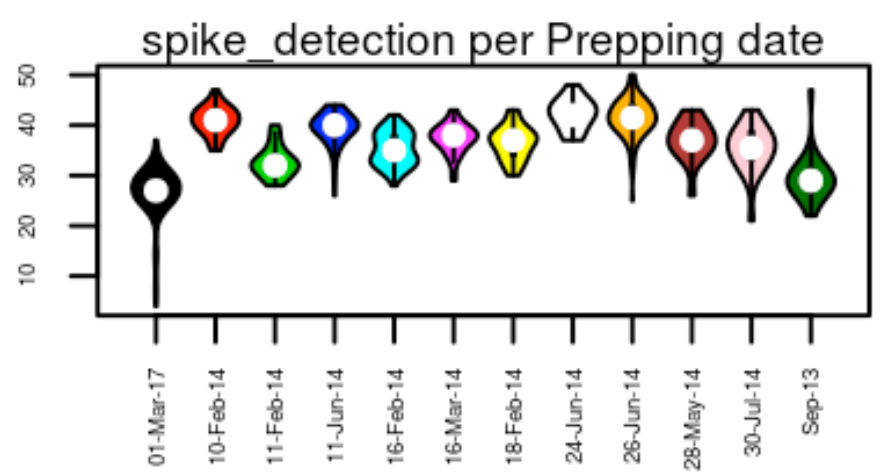
- Can be batch effects per
  - Experiment
  - Animal/Patient/Batch of cells
  - Sort plate
  - Sequencing lane
- Check if QC-measures deviates for any of those measures
- Check in PCA if any PC correlates to batches – Scater tutorial

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









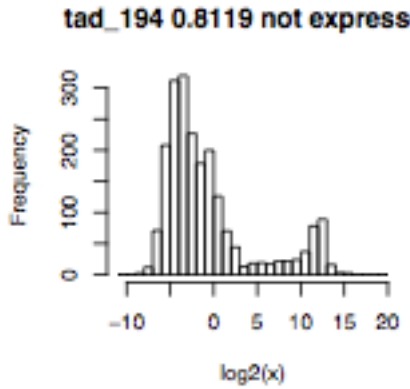
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

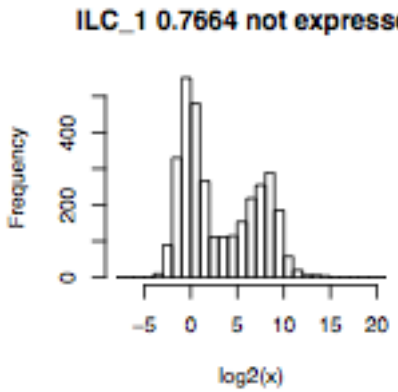
SciLifeLab

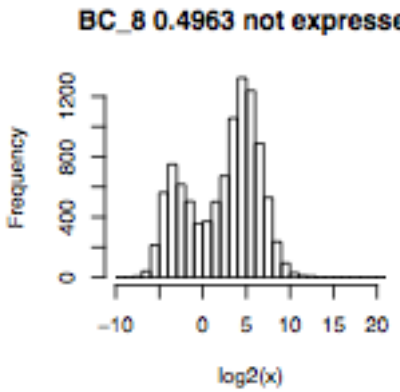
How to filter genes

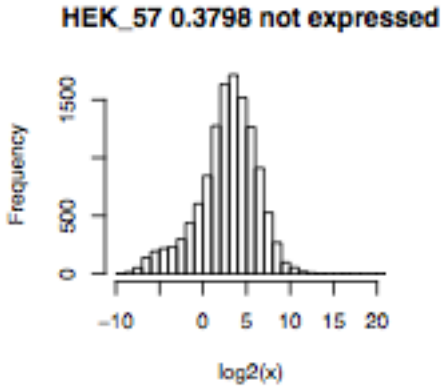
- In most cases, all genes are not used in dimensionality reduction and clustering.
- Gene set selection based on:
  - Genes expressed in X cells over cutoff Y.
  - Variable genes – using spike-ins or whole distribution.
  - Filter out genes with correlation to few other genes
  - Prior knowledge / annotation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

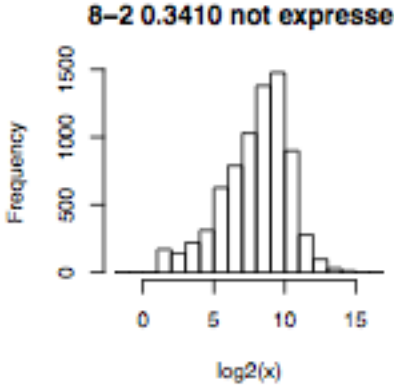
Defining cutoffs for gene expression – bimodal gene expression or background expression?









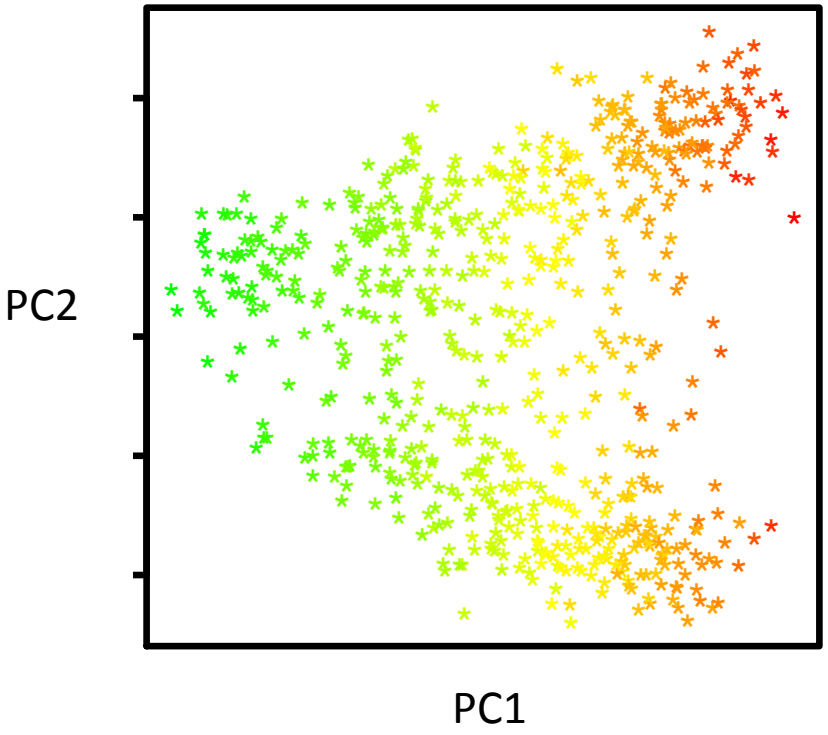


Cells ordered by size (approximate)
tad − *Trixoplax adherens* cell
ILC − Innate lymphoid cell
BC − B-cell
HEK − human embryonic kidney cell
8-2 − Mouse embryonic cell (8-cell stage)

Small cells tend to have fewer detected genes and more background detection

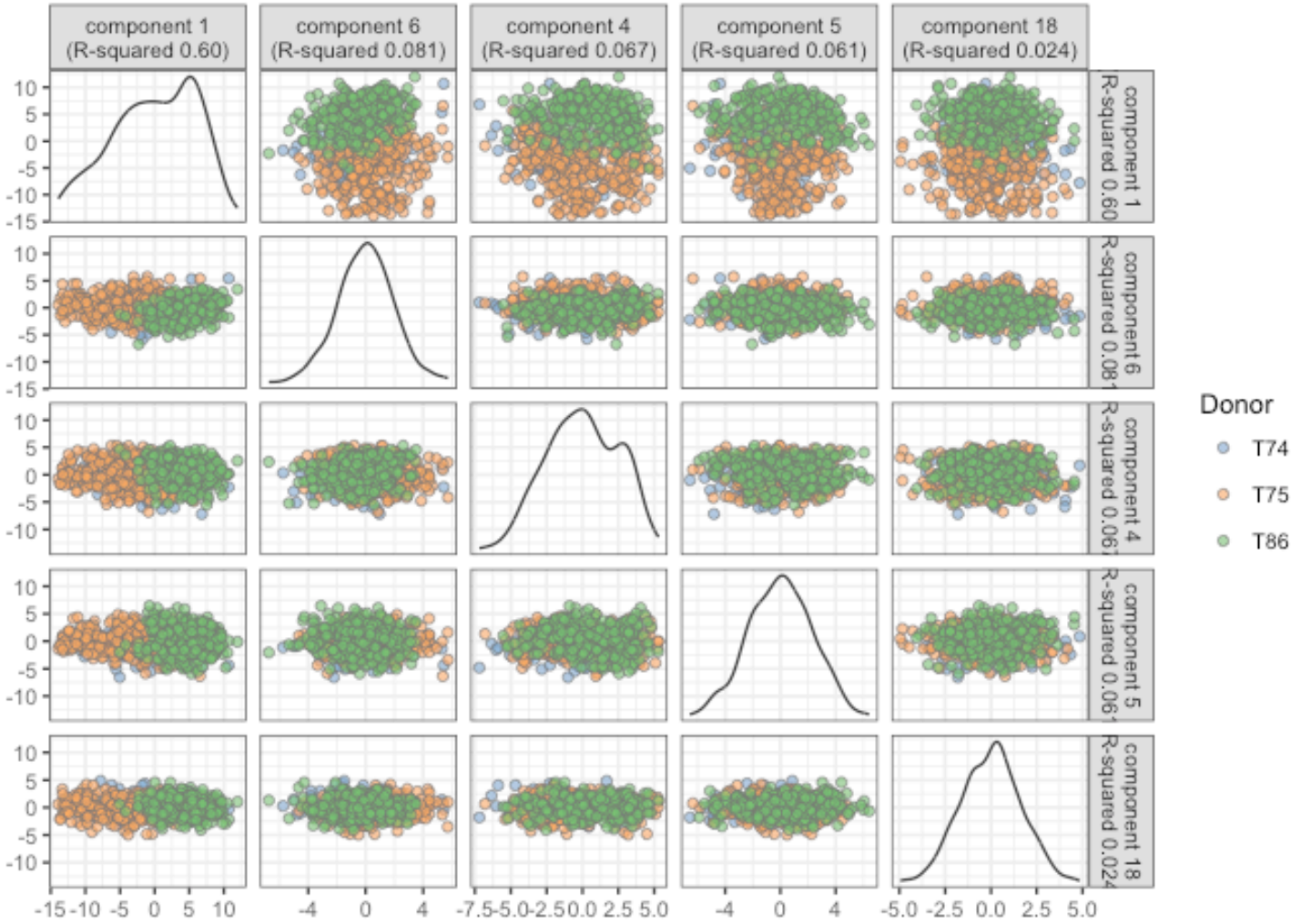
PCA for QC

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



Red – high number of detected genes
Green - low

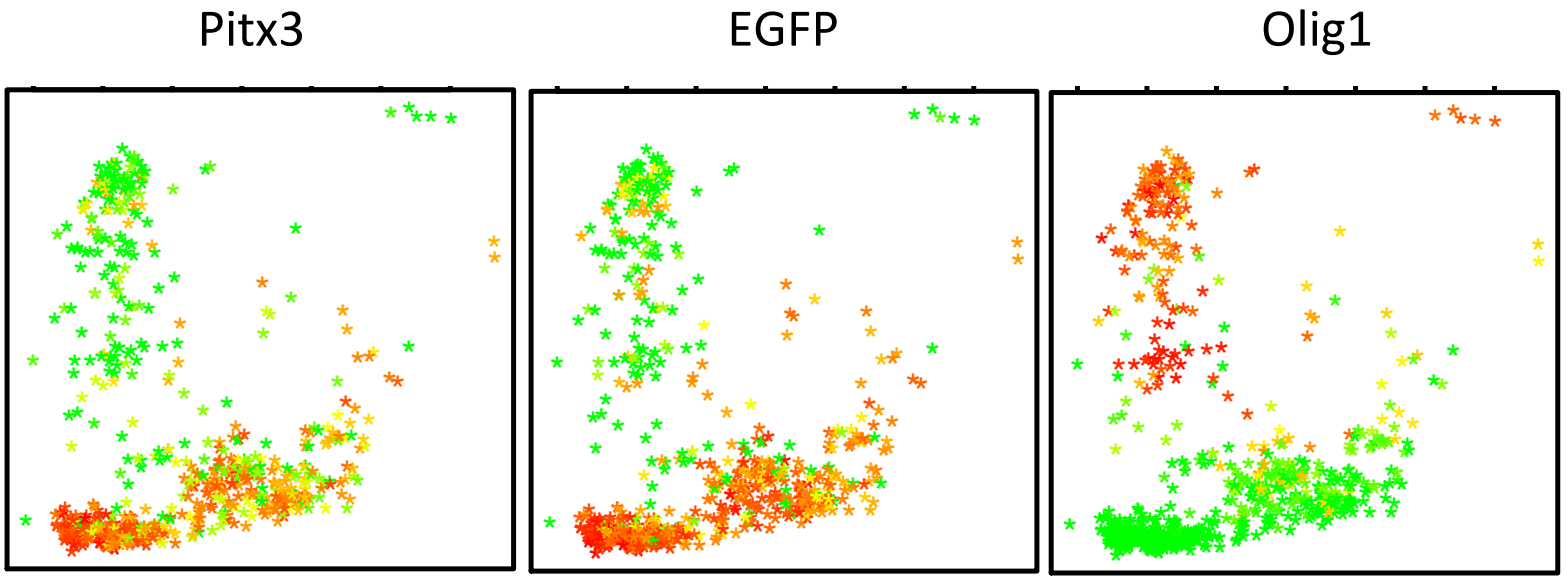
Check for batch effects in PCA



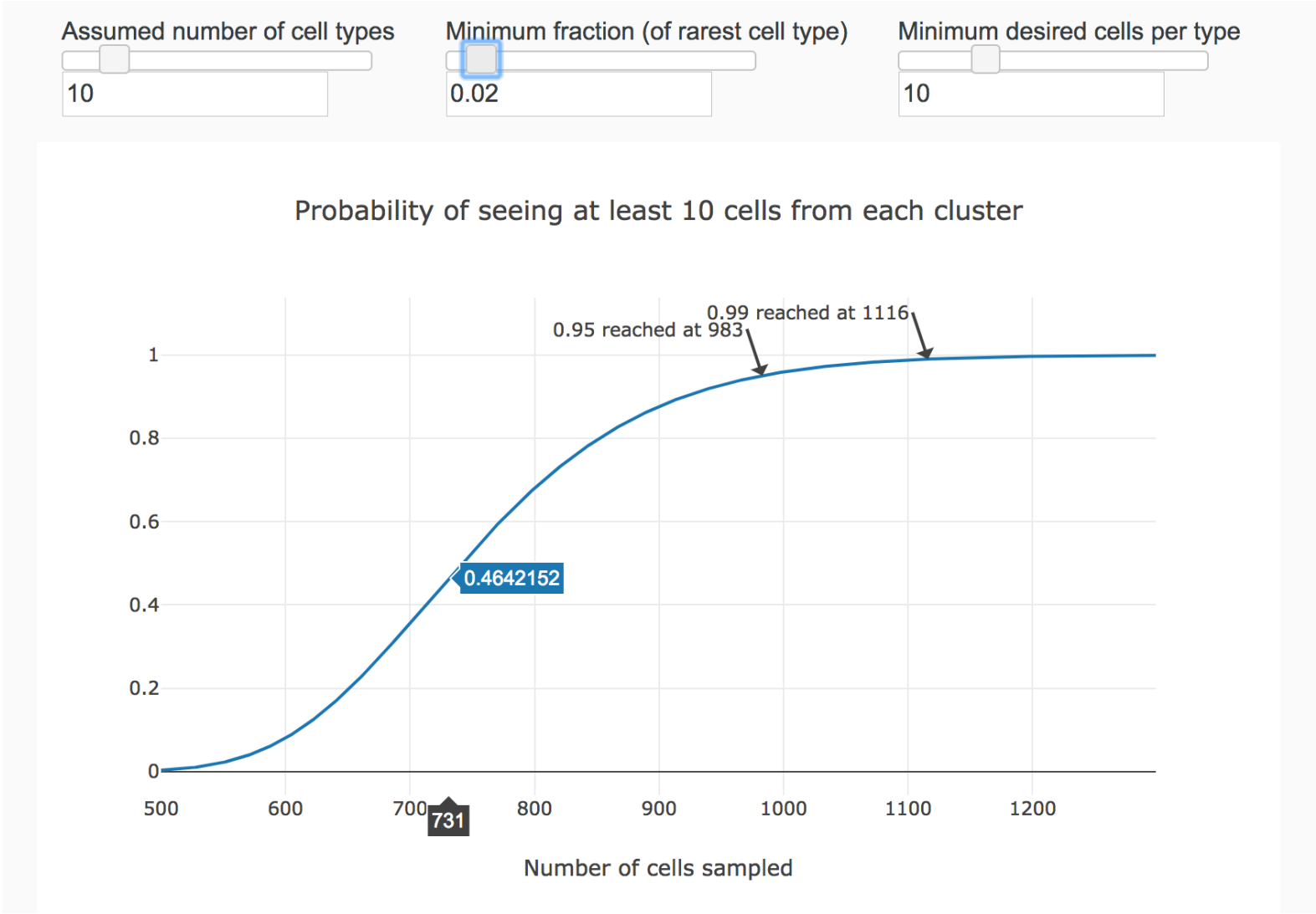
Scater package

PCA for QC

- PCA can be used to identify contaminant cells when you are sorting for a specific cell type.



How many cells do you need to sequence?



NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

http://satijalab.org/howmanycells

SciLifeLab

Conclusions

- Try to plan your experiment in a way so that the biological signal you are looking for is not confounded by batch effects
- 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

QC-summary for scRNAseq data

- Scripts for creating a QC-summary report from 2 or 3 files:
  - A file with all QC-stats
  - A metadata table with batch info etc
  - A file with all expression values (rpkms,counts or similar) – optional, only needed if you also want PCA plots.
- There are also some scripts for converting all SS2 data delivered from the ESCG to the correct format for making the qc-report.

https://bitbucket.org/asbj/qc-summary_scrnaseq