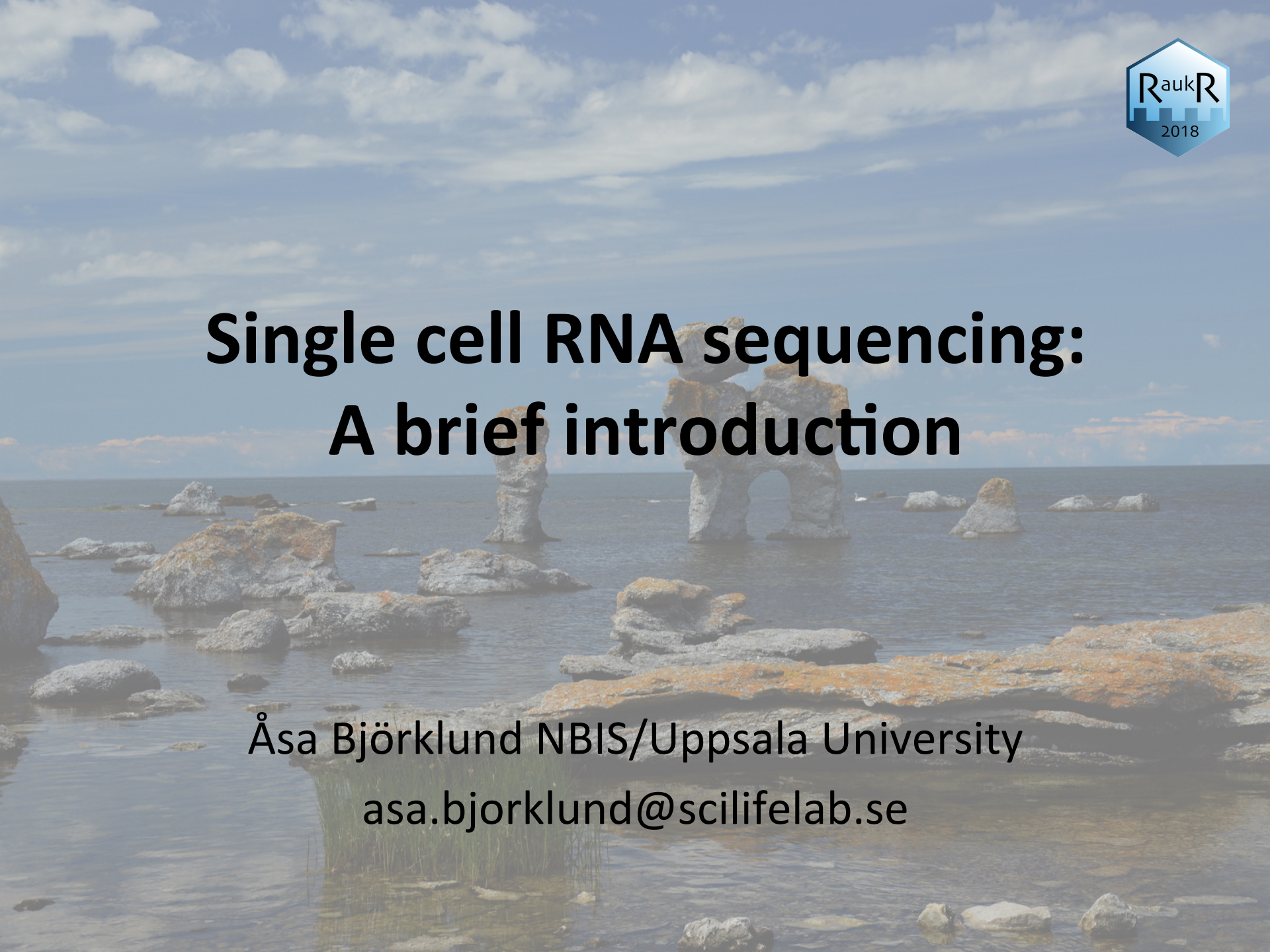


Single cell RNA sequencing: A brief introduction

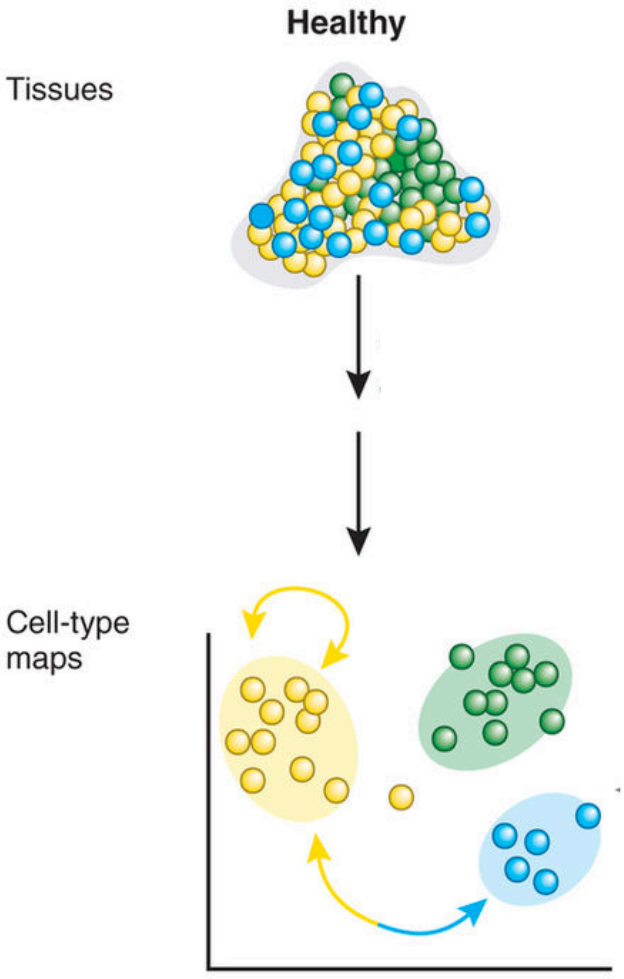
Åsa Björklund NBIS/Uppsala University

asa.bjorklund@scilifelab.se

Outline

- Overview of scRNA-seq:
  - Why scRNAseq?
  - How is it done?
  - Transcriptional bursting
  - Spike-in RNAs
- Data analysis:
  - Overview of main steps
  - Examples of tools
- Recent developments

Why single cell?



Between cell types

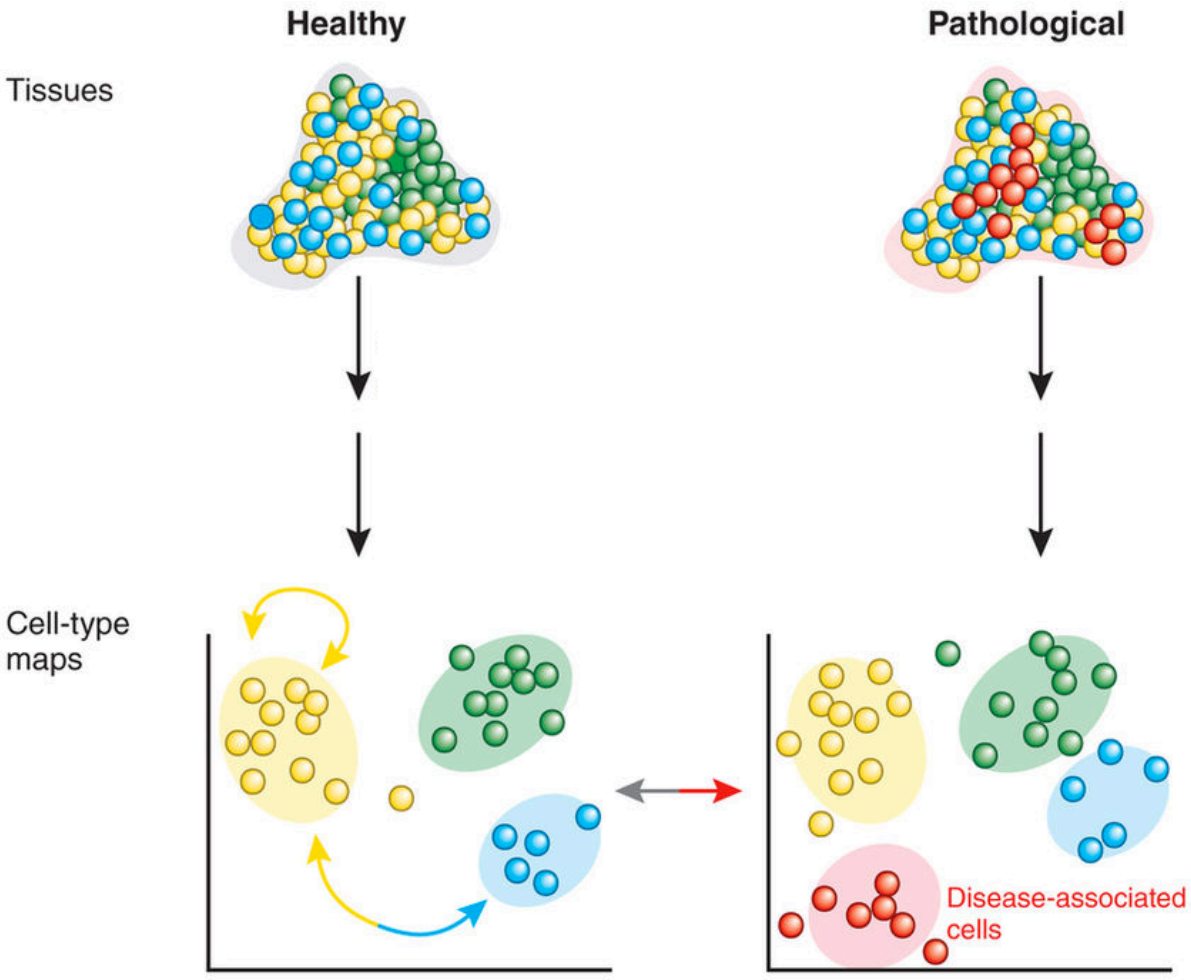
- Identify biomarkers
- Compare regulatory networks

Within cell type

- Transcriptional mechanisms
- Regulatory networks
- Allelic expression patterns
- Cell cycle progression

(Sandberg, *Nature Methods* 2014)

Why single cell?



Between tissues

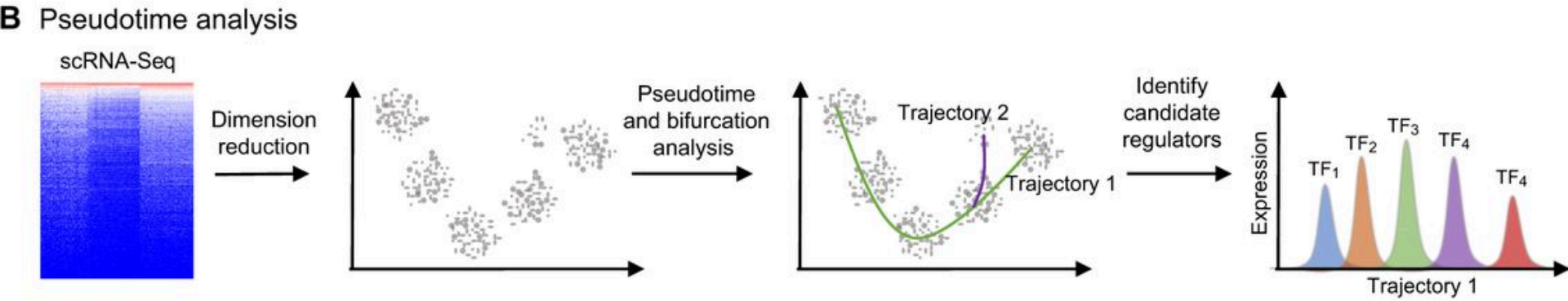
- Cell-type compositions
- Altered transcription in matched celltypes

(Sandberg, *Nature Methods* 2014)

Why single cell?

Developmental progression

- Transitions during differentiation
- Lineage decisions



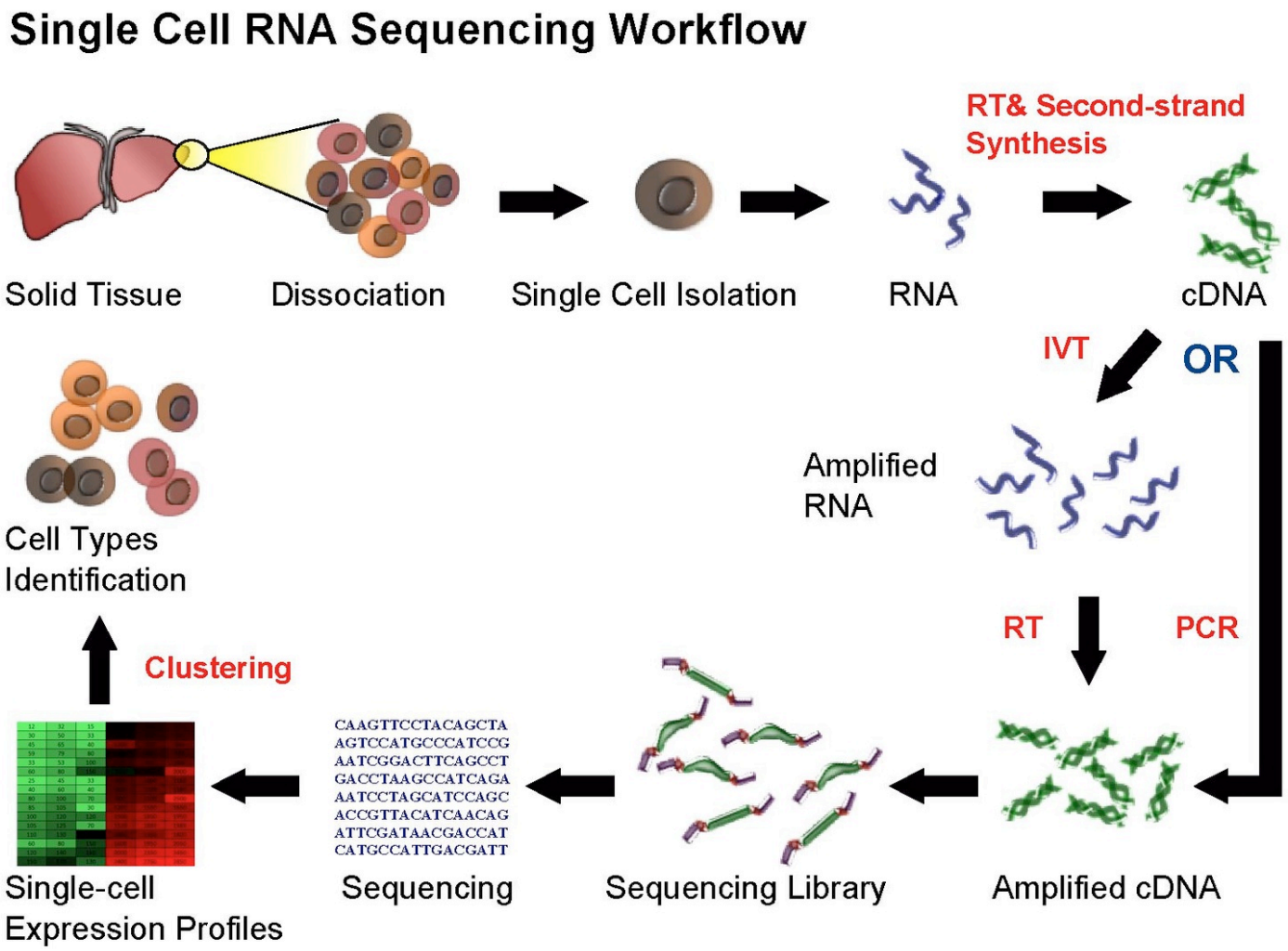
(Kuman *et al. Development*, 2017)

Why single cell?

Main applications:

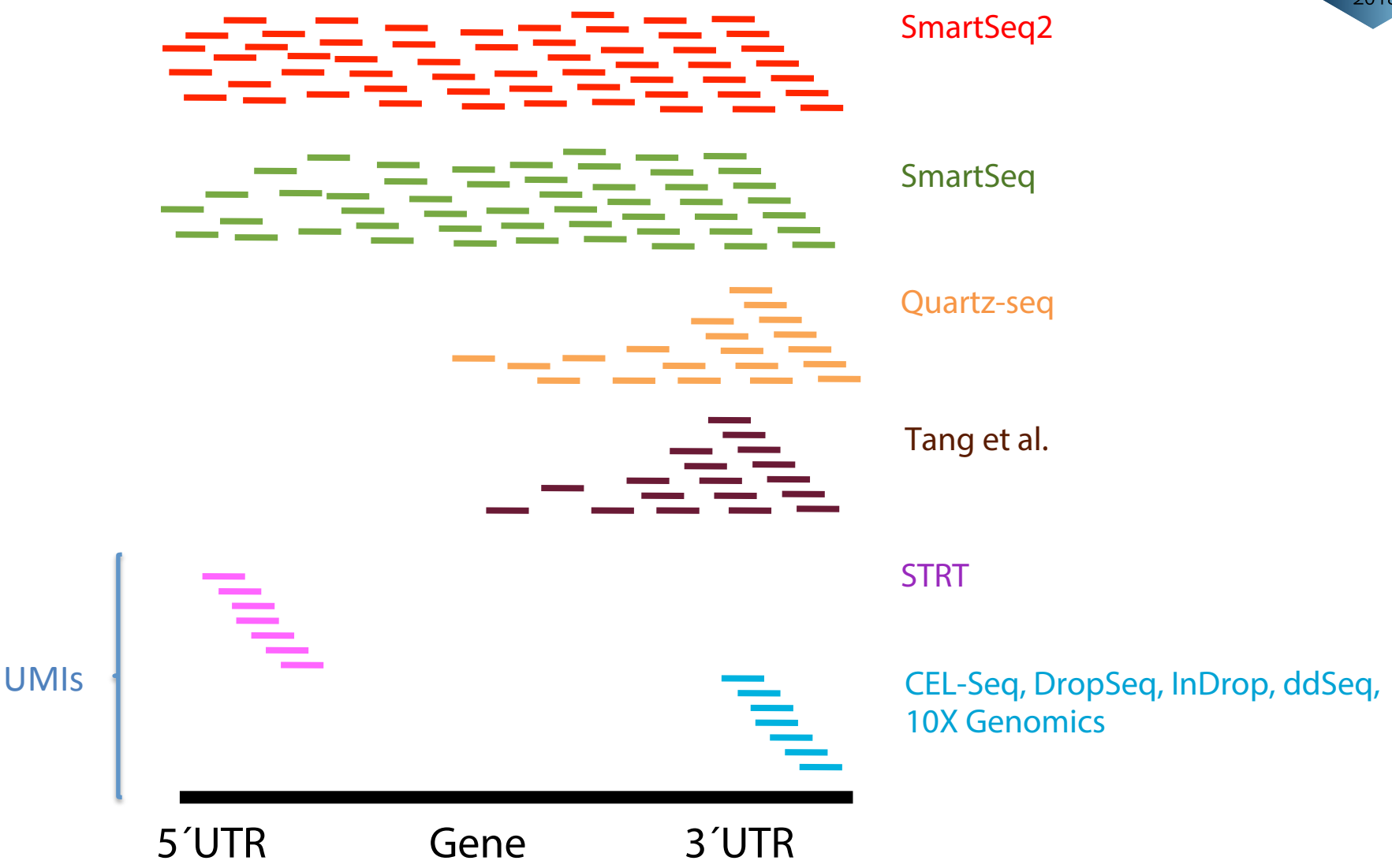
- Understanding heterogeneous tissues - Atlas
  - Identifying cell types in a tissue
- Changes in cellular composition and cellular states.
  - Upon a treatment
  - With time
  - Between healthy and disease
- Dissection of temporal changes − Trajectories
  - Understand the development of cell types

How is it done?

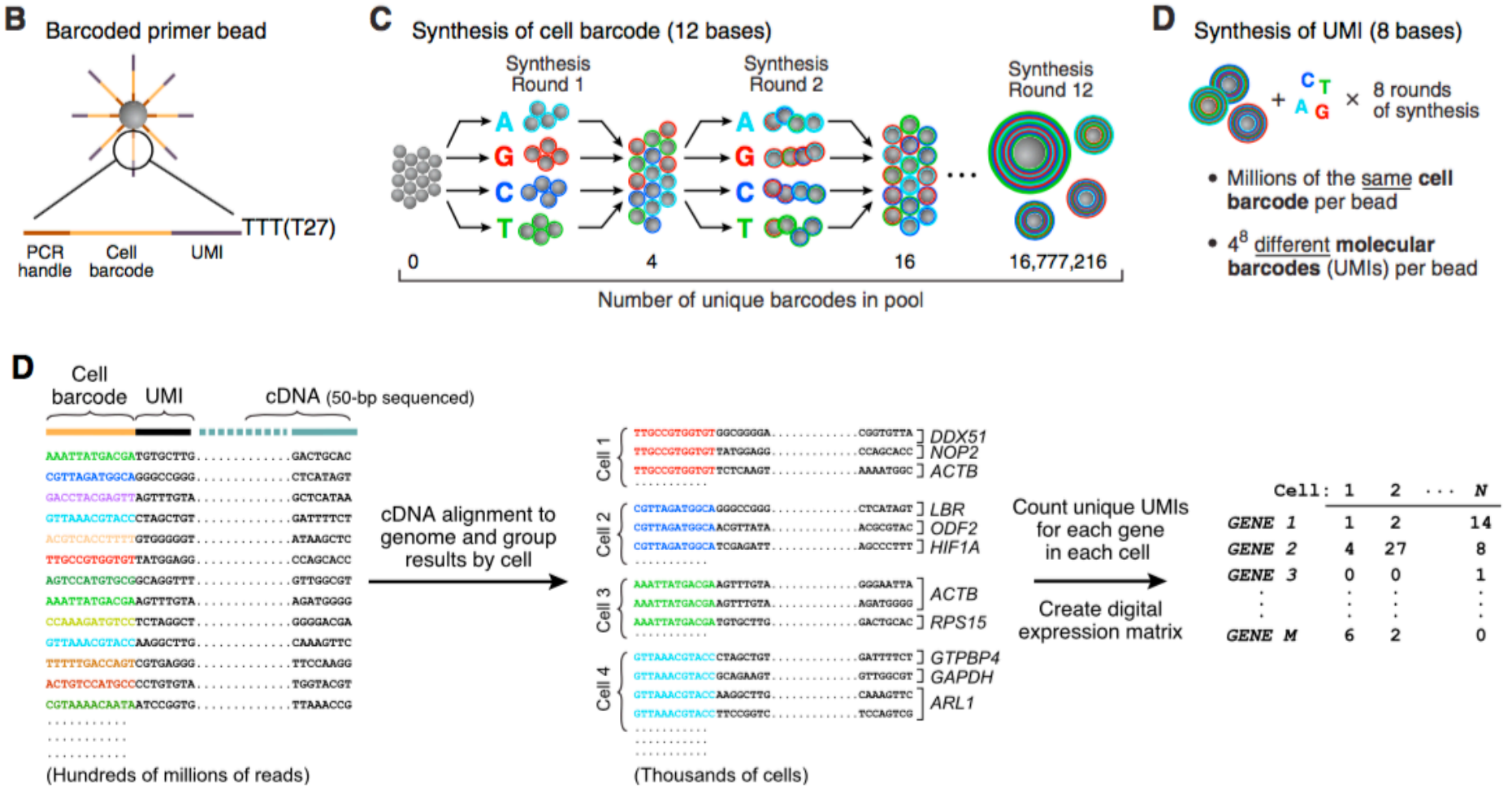


(Wikipedia)

scRNAseq methods

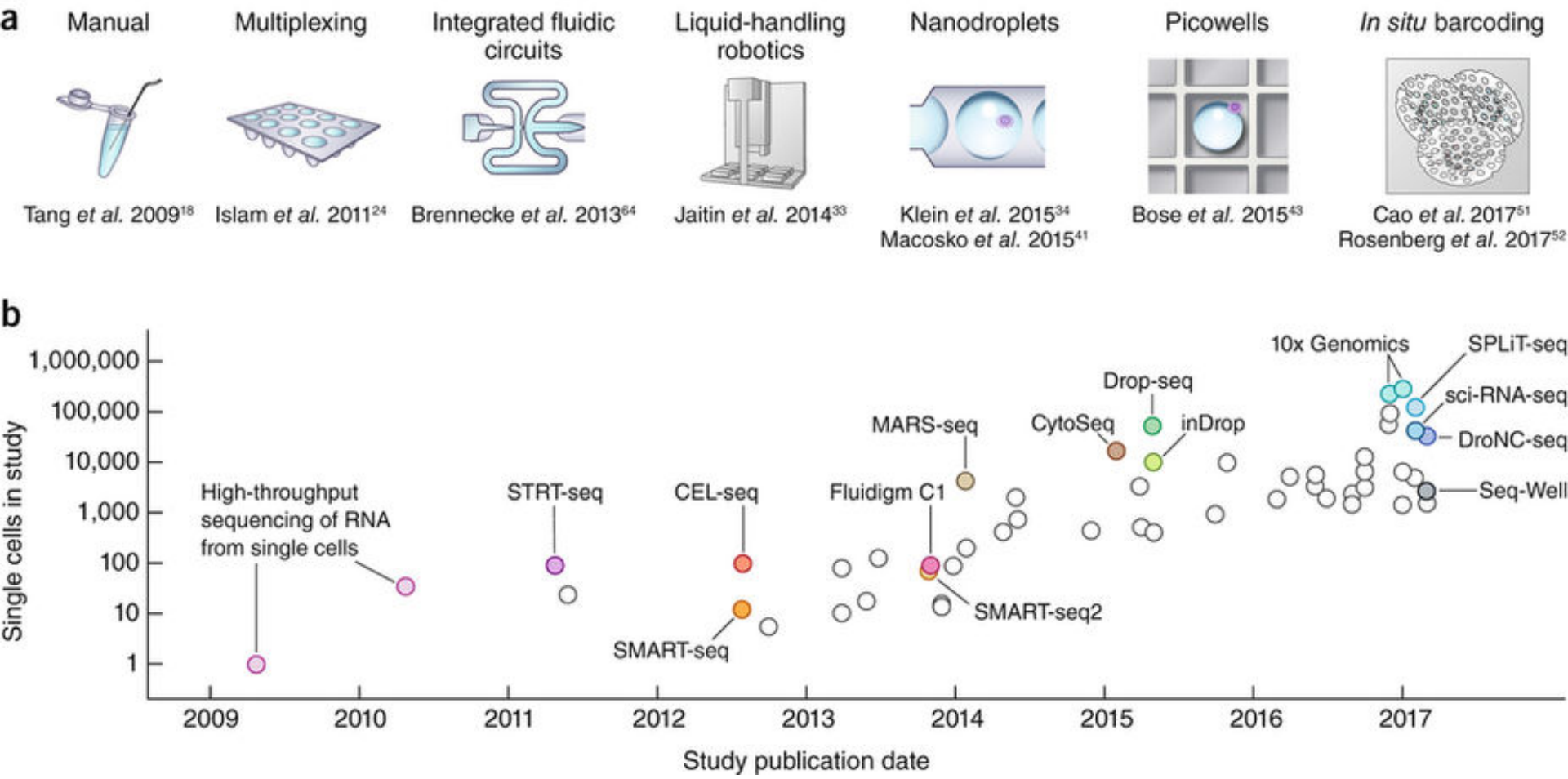


Unique molecular identifiers - UMIs



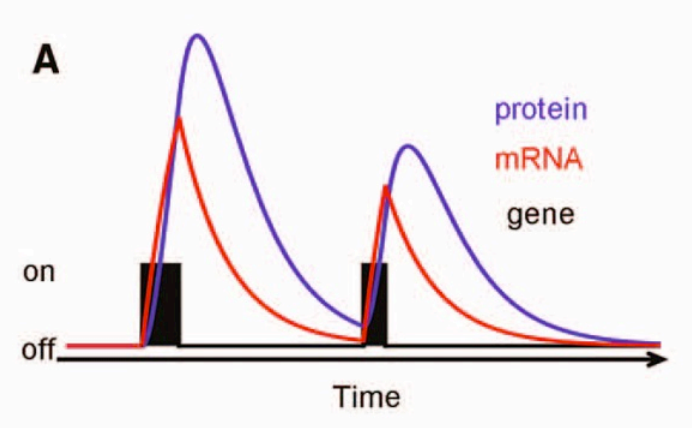
(http://mccarrolllab.com/dropseq/)

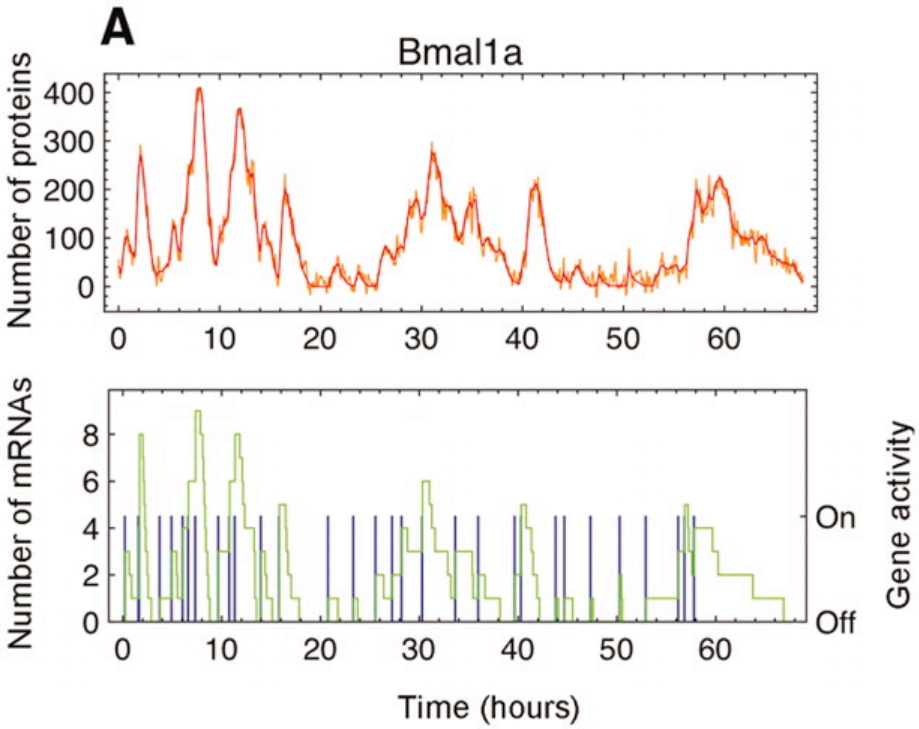
Method developments



(Svensson *et al. Nature Protocols* 2018)

Transcriptional bursting

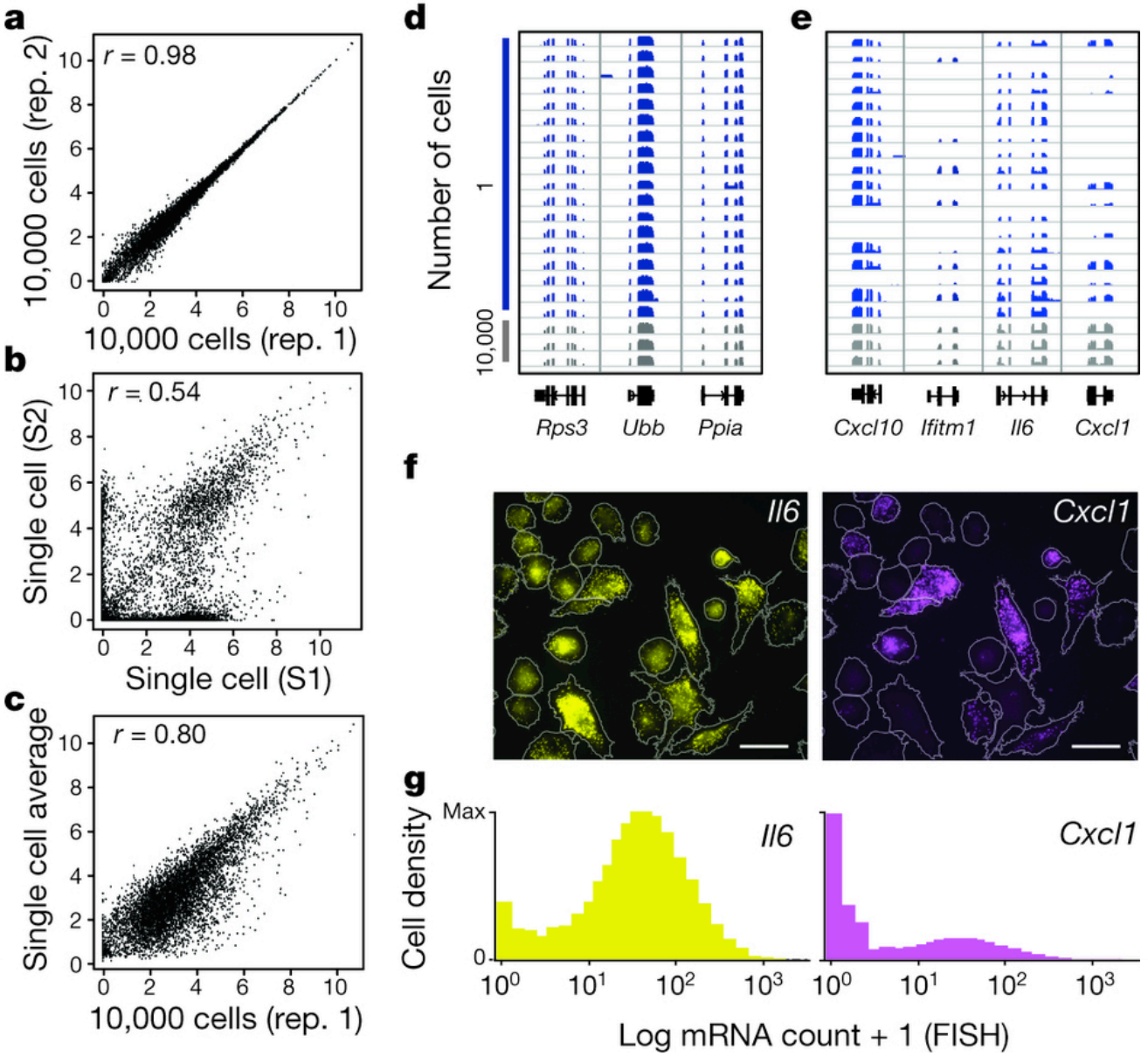




- Burst frequency and size is correlated with mRNA abundance but may differ per gene
- Many TFs have low expression and low burst frequency and will only be detected in a fraction of the cells

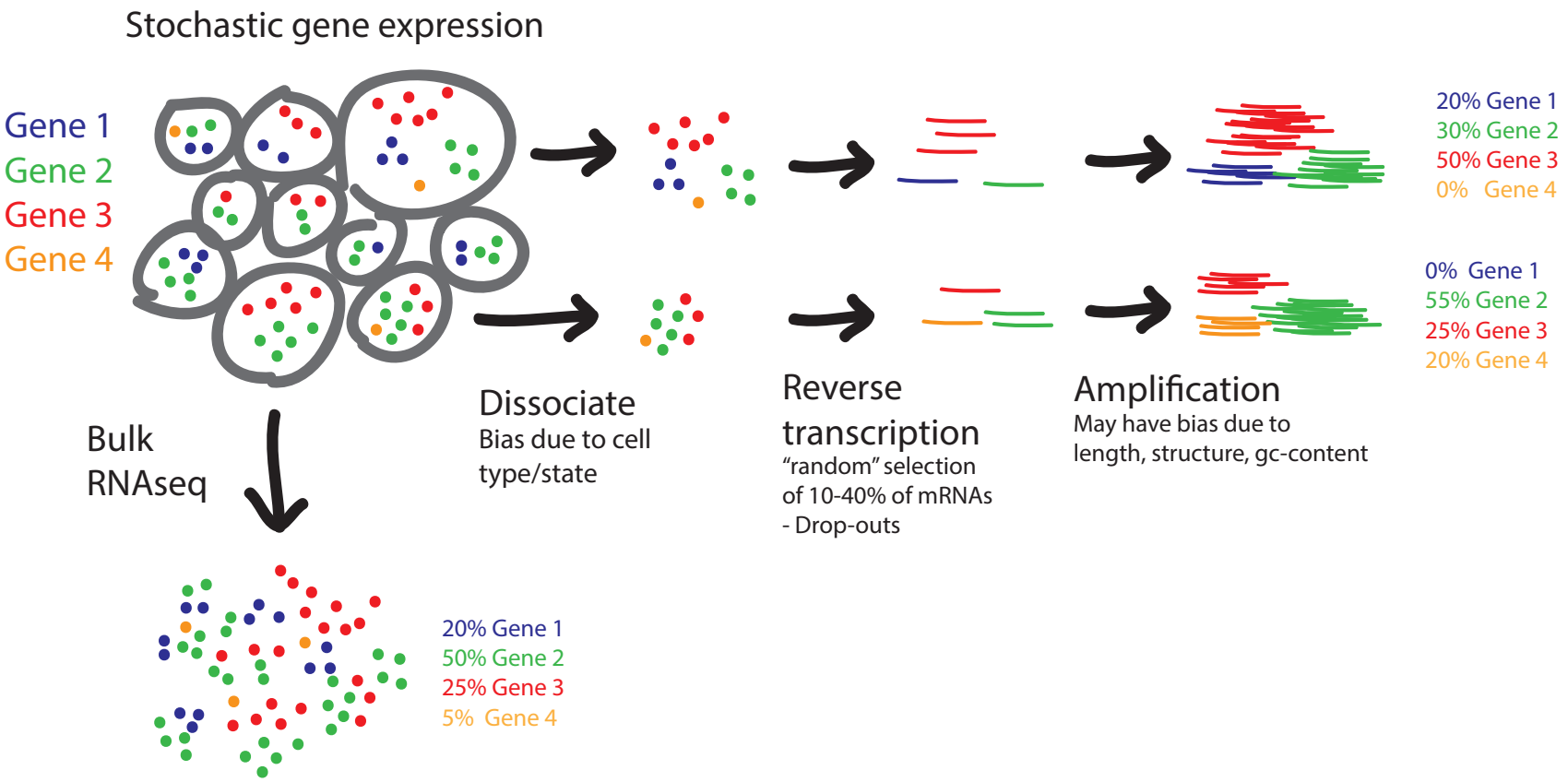
(Suter *et al. Science* 2011)

Example data - mouse bone-marrow-derived dendritic cells (BMDCs)

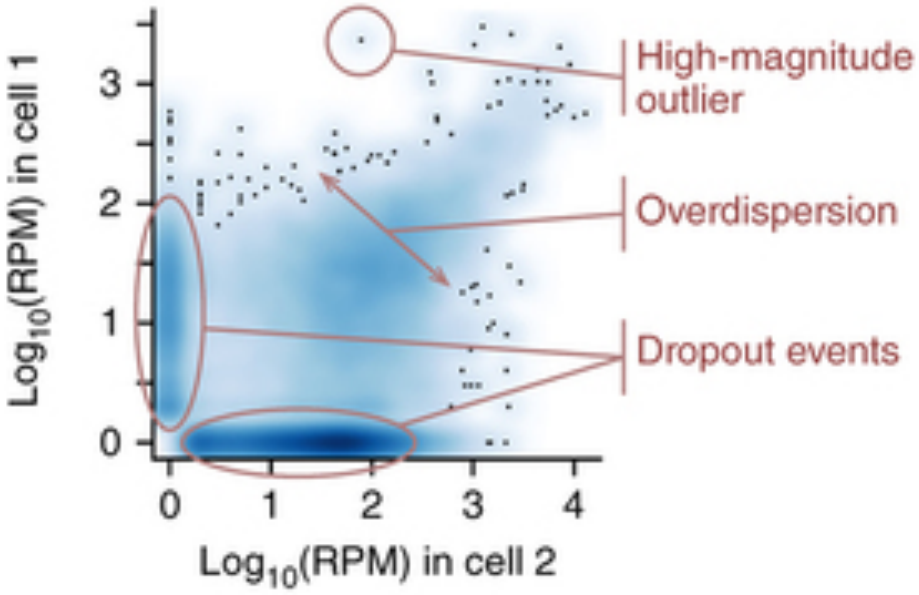


(Shalek *et al. Nature* 2014)

Bursting, drop-outs and amplification bias



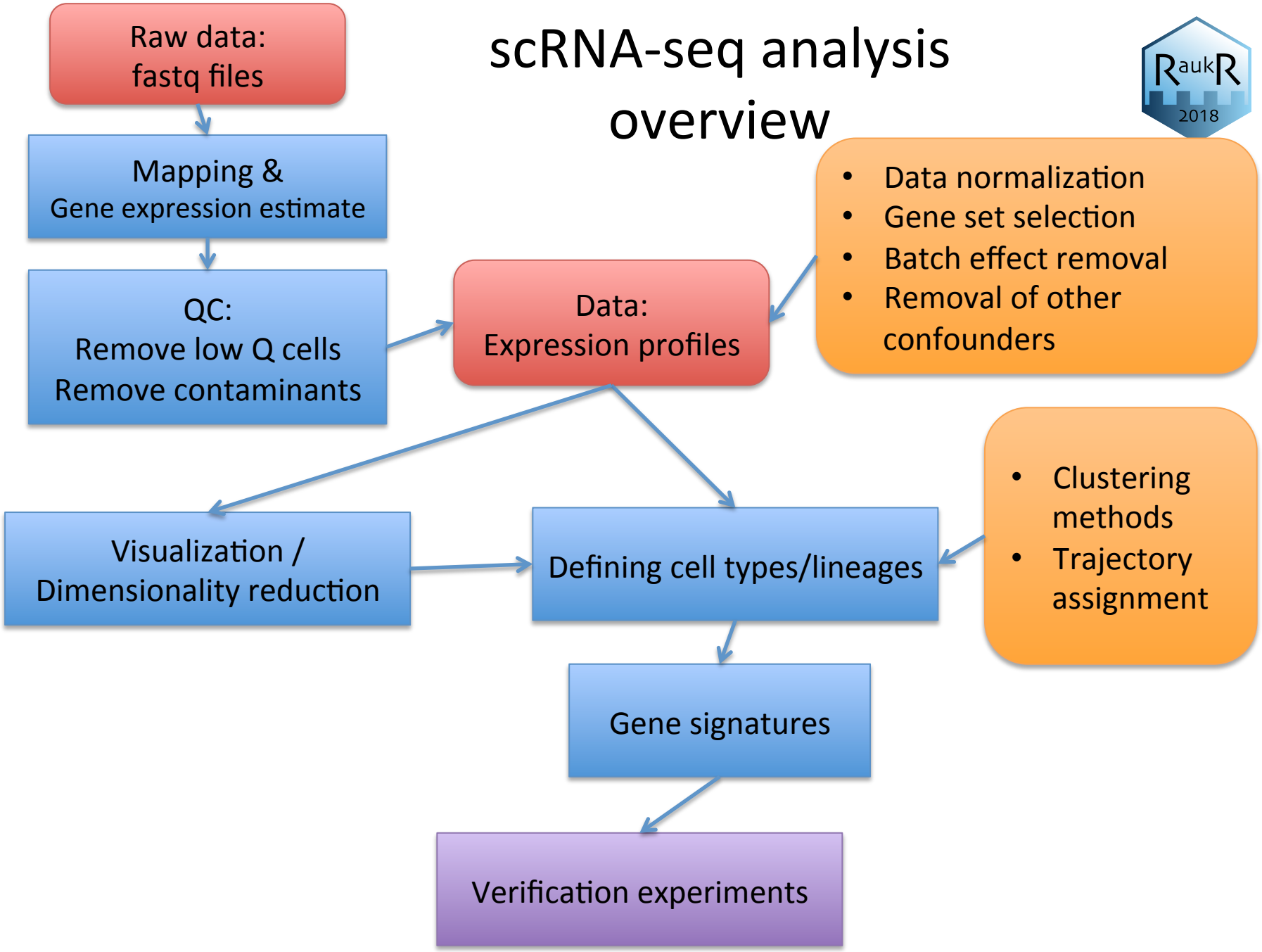
Problems compared to bulk RNA-seq



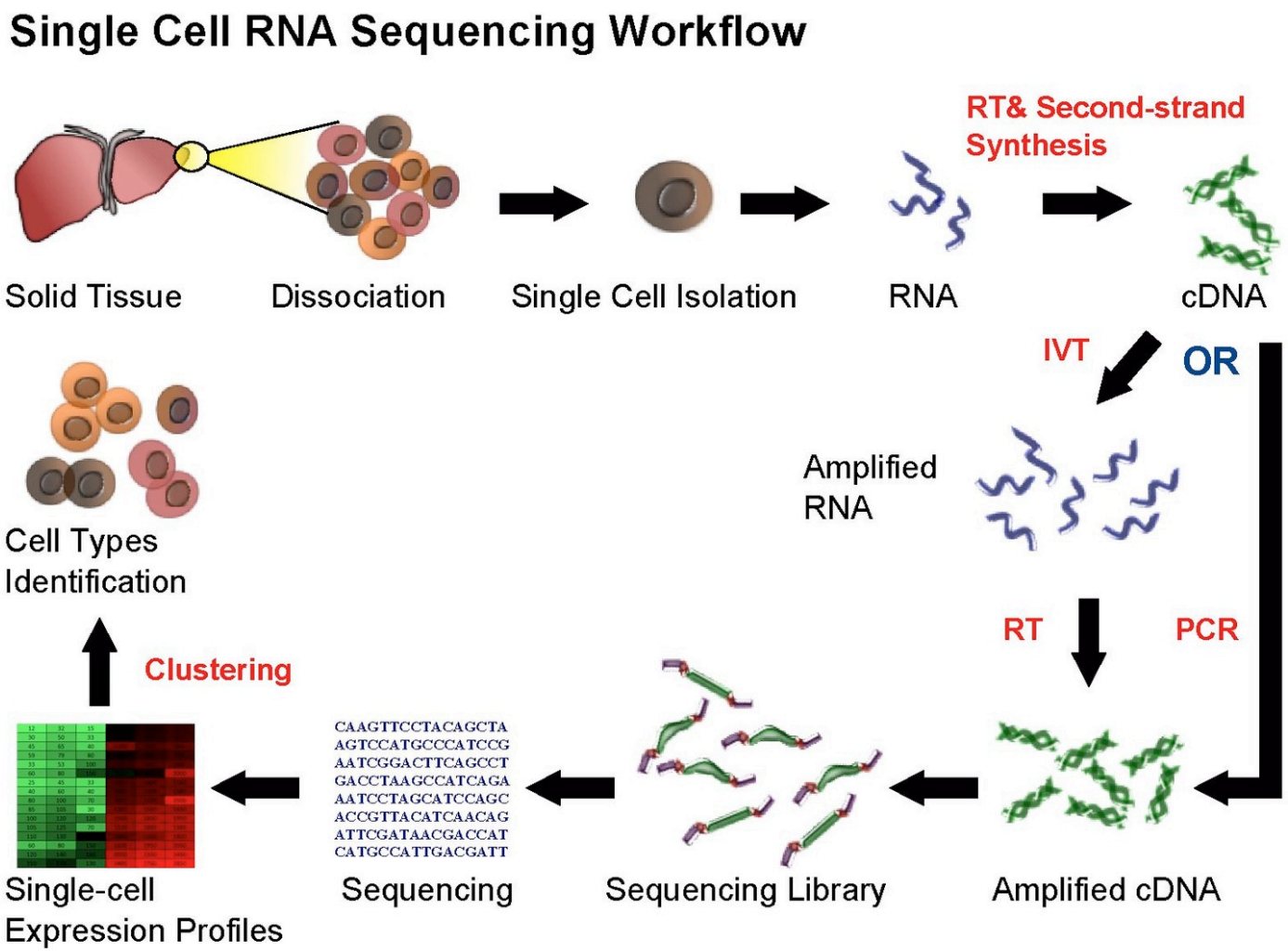
- Transcriptional bursting
- Drop-out rates
- Amplification bias
- Background noise
- Bias due to cell-cycle, cell size and other factors
- Often clear batch effects

- Zero-inflated data − different distribution compared to bulk RNAseq

(Karchenko *et al. Nature Methods* 2014)


scRNA-seq analysis overview
RaukR 2018
Raw data: fastq files
Mapping & Gene expression estimate
QC: Remove low Q cells Remove contaminants
Data: Expression profiles
Data normalization
Gene set selection
Batch effect removal
Removal of other confounders
Visualization / Dimensionality reduction
Defining cell types/lineages
Clustering methods
Trajectory assignment
Gene signatures
Verification experiments

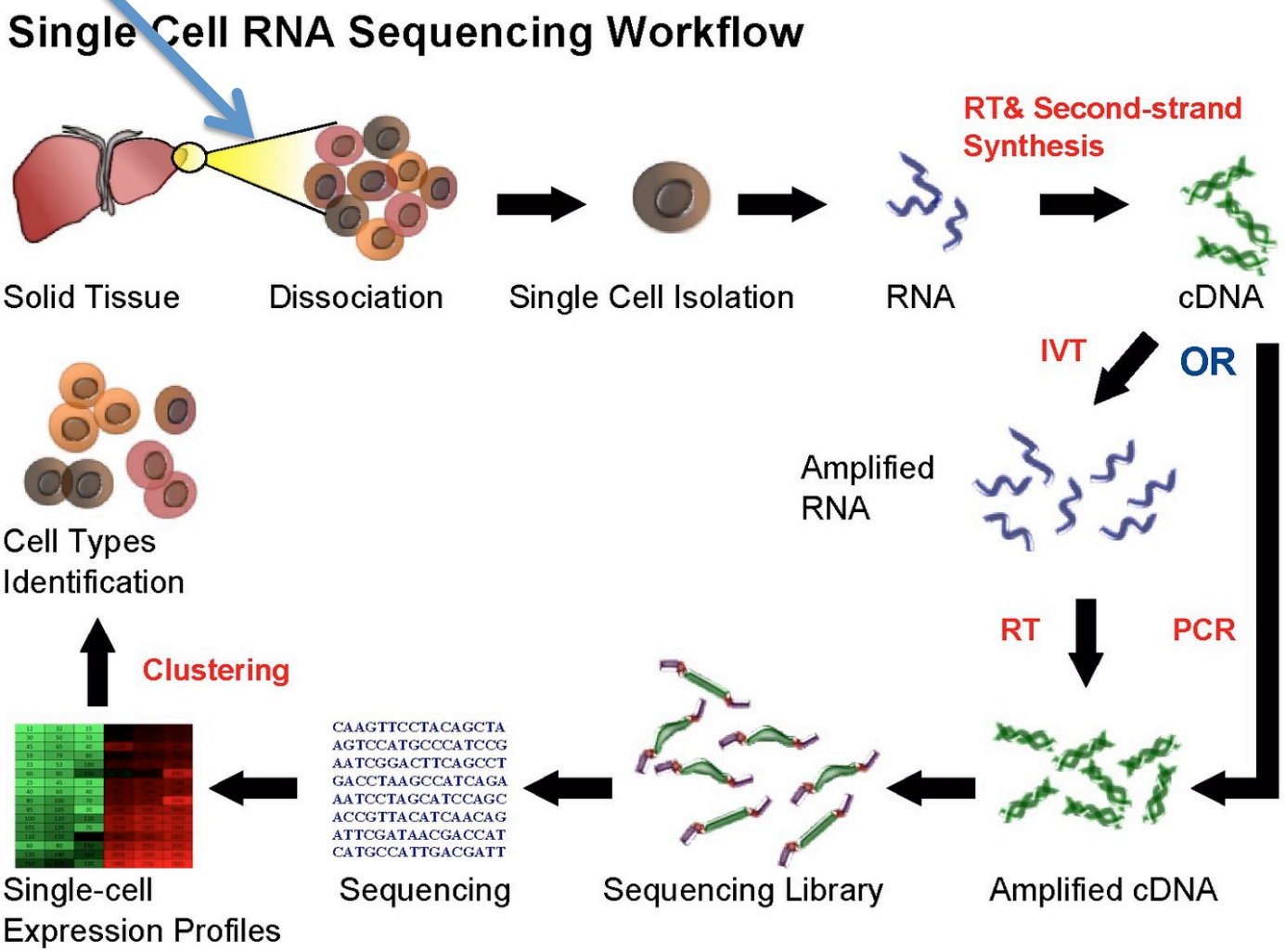
What can go wrong?



(Wikipedia)

Biased selection of cells.
Quality of cells

can go wrong?

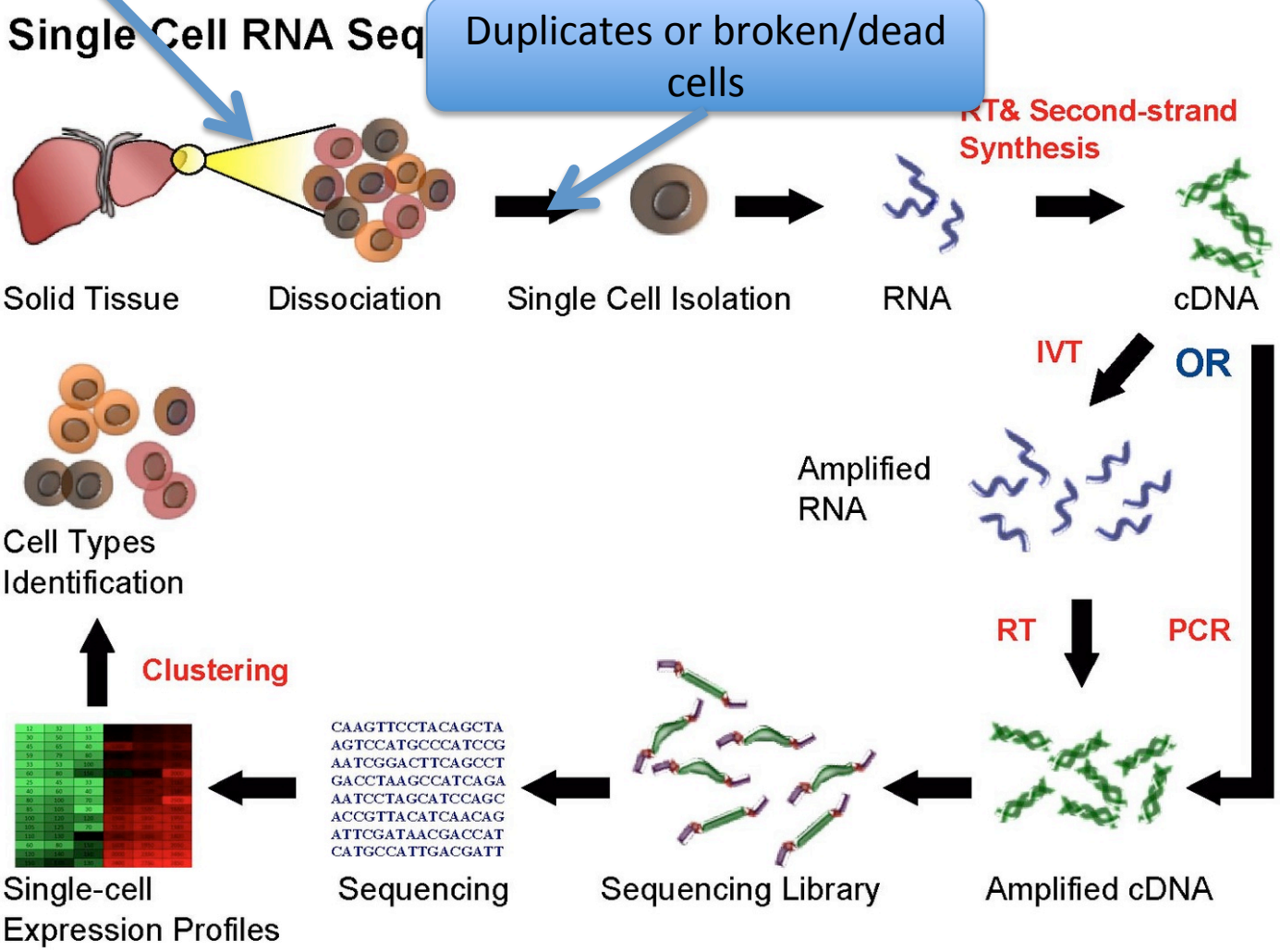


(Wikipedia)

Biased selection of cells.
Quality of cells

can go wrong?

Duplicates or broken/dead cells



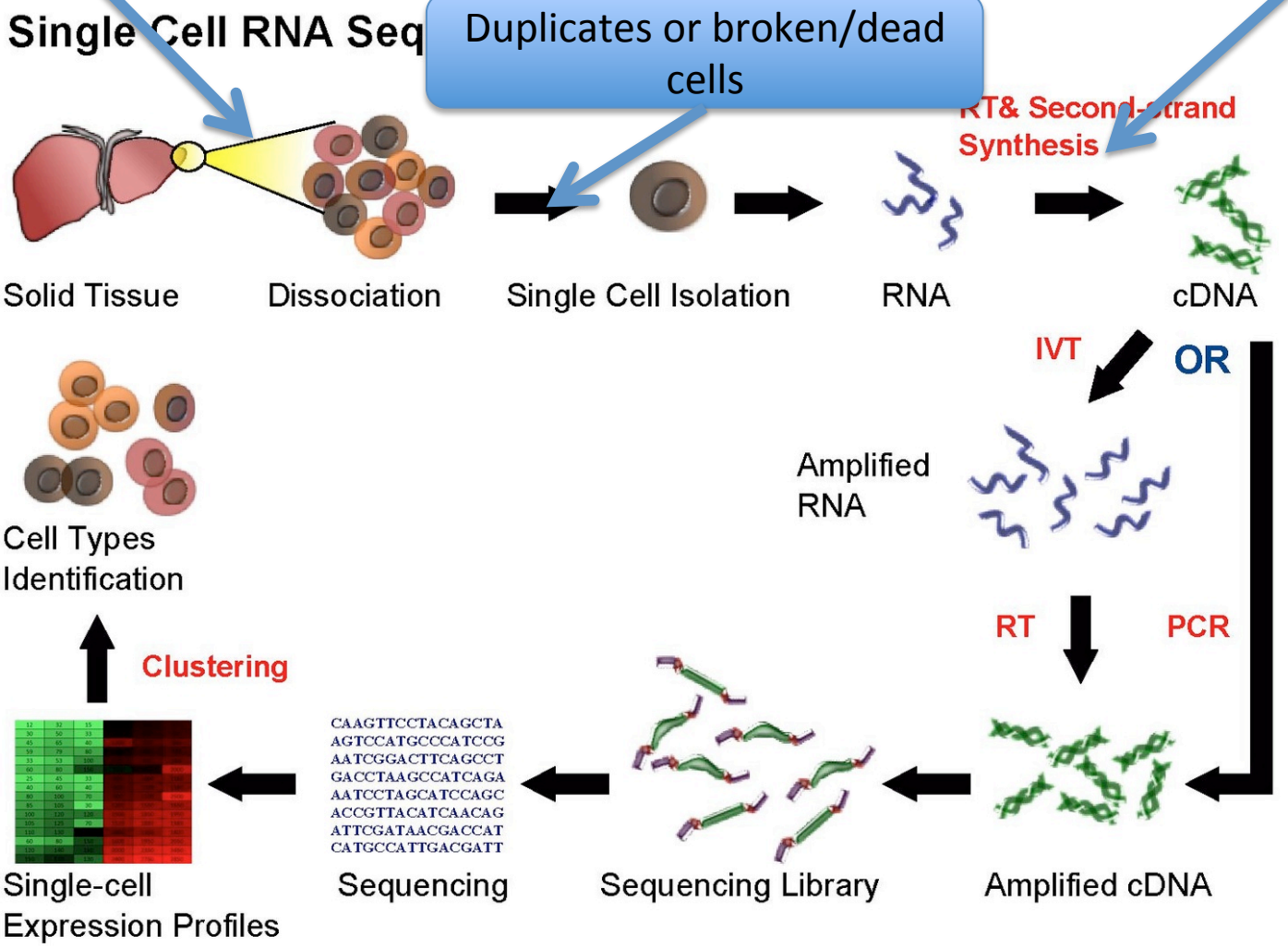
(Wikipedia)

can go wrong?

Biased selection of cells.
Quality of cells

Duplicates or broken/dead cells

Drop-outs
60-90%

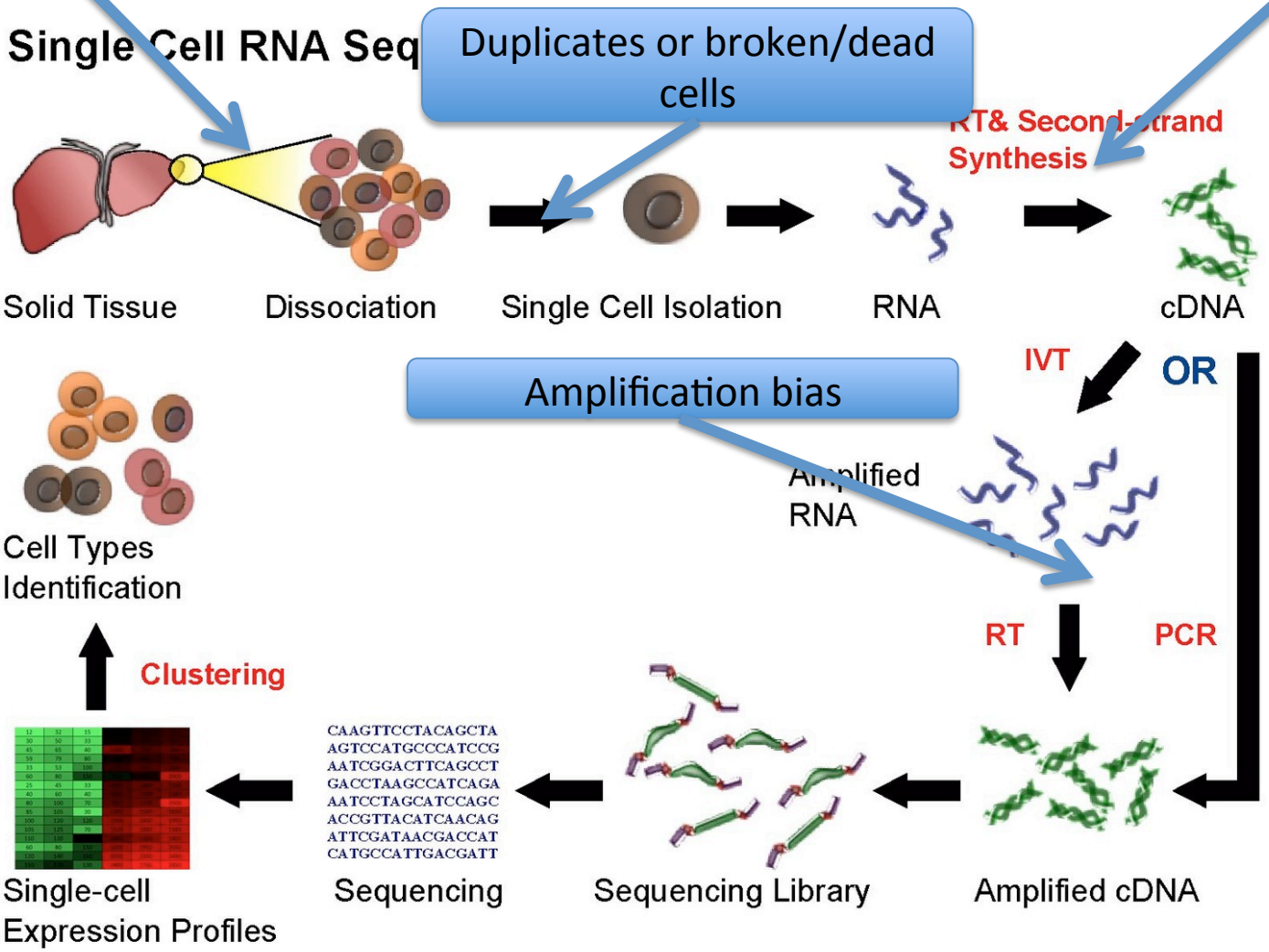


(Wikipedia)

can go wrong?

Biased selection of cells.
Quality of cells

Drop-outs
60-90%

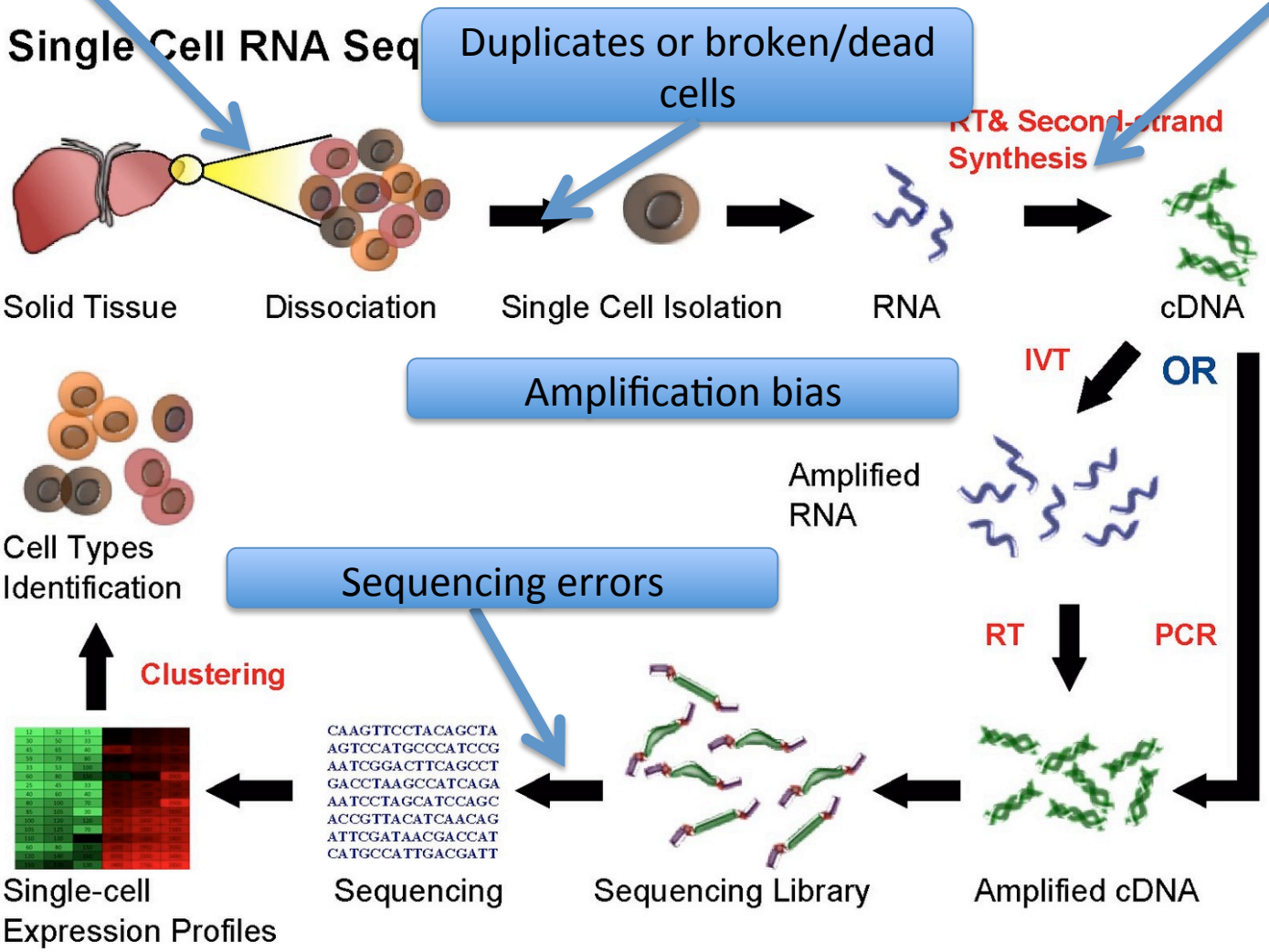


(Wikipedia)

can go wrong?

Biased selection of cells.
Quality of cells

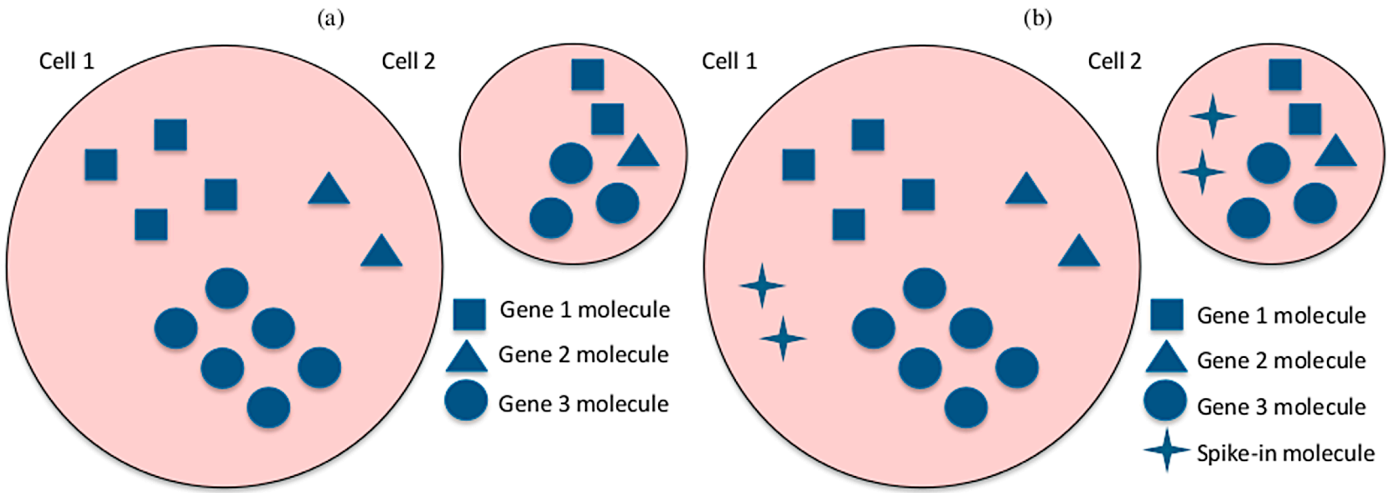
Drop-outs
60-90%



(Wikipedia)

Spike-in RNAs

- Addition of external controls
- ERCC spike-in most widely used, consists of 48 or 96 mRNAs at 17 different concentrations.
- Important to add equal amounts to each cell, preferably in the lysis buffer.



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

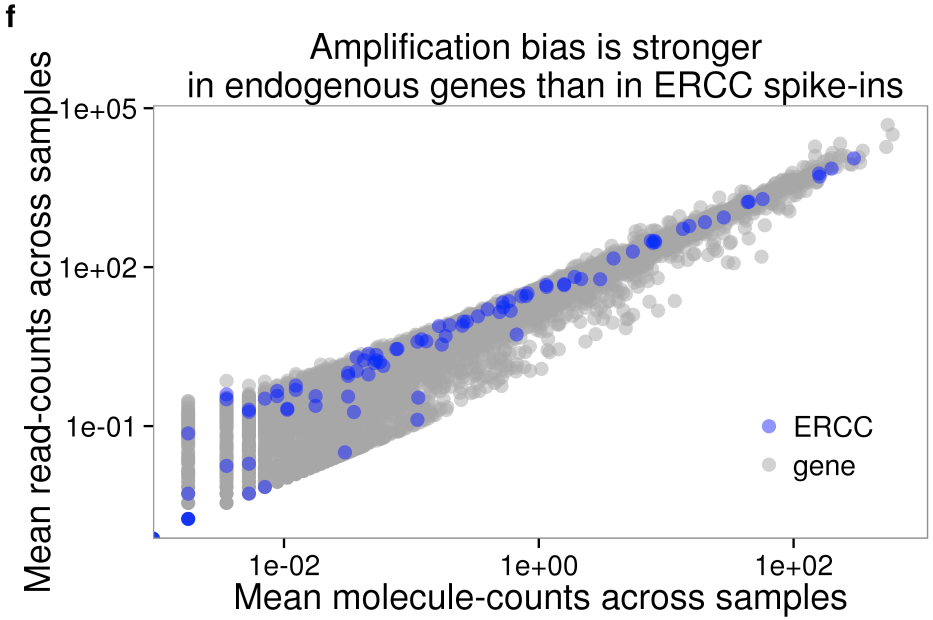
Spike-in RNAs

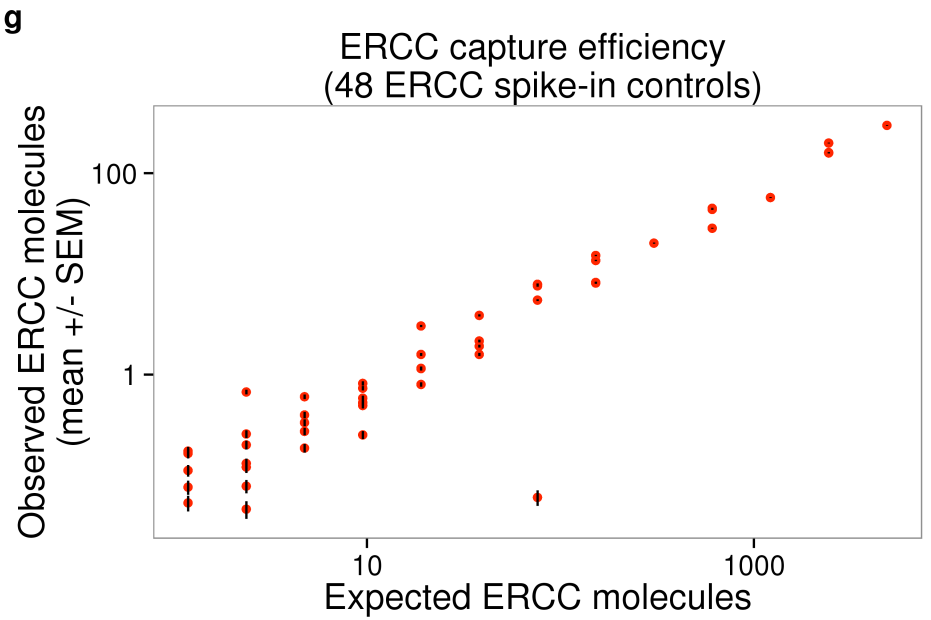
Spike-ins can be used to model:

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

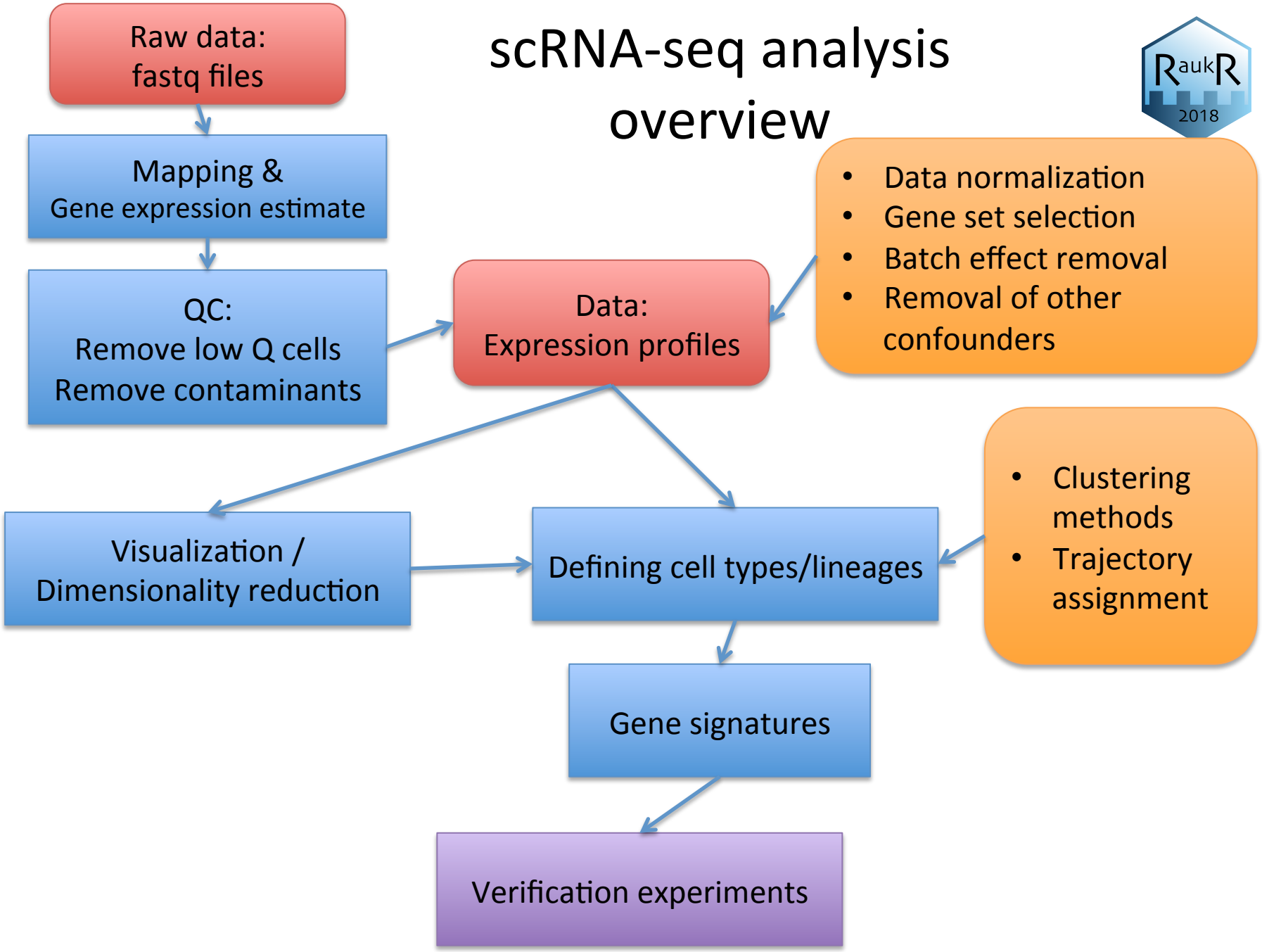
(Vallejos et al. *PLOS Comp Biol* 2015)

Spike-in RNAs





(Tung et al. *Scientific Reports* 2017)

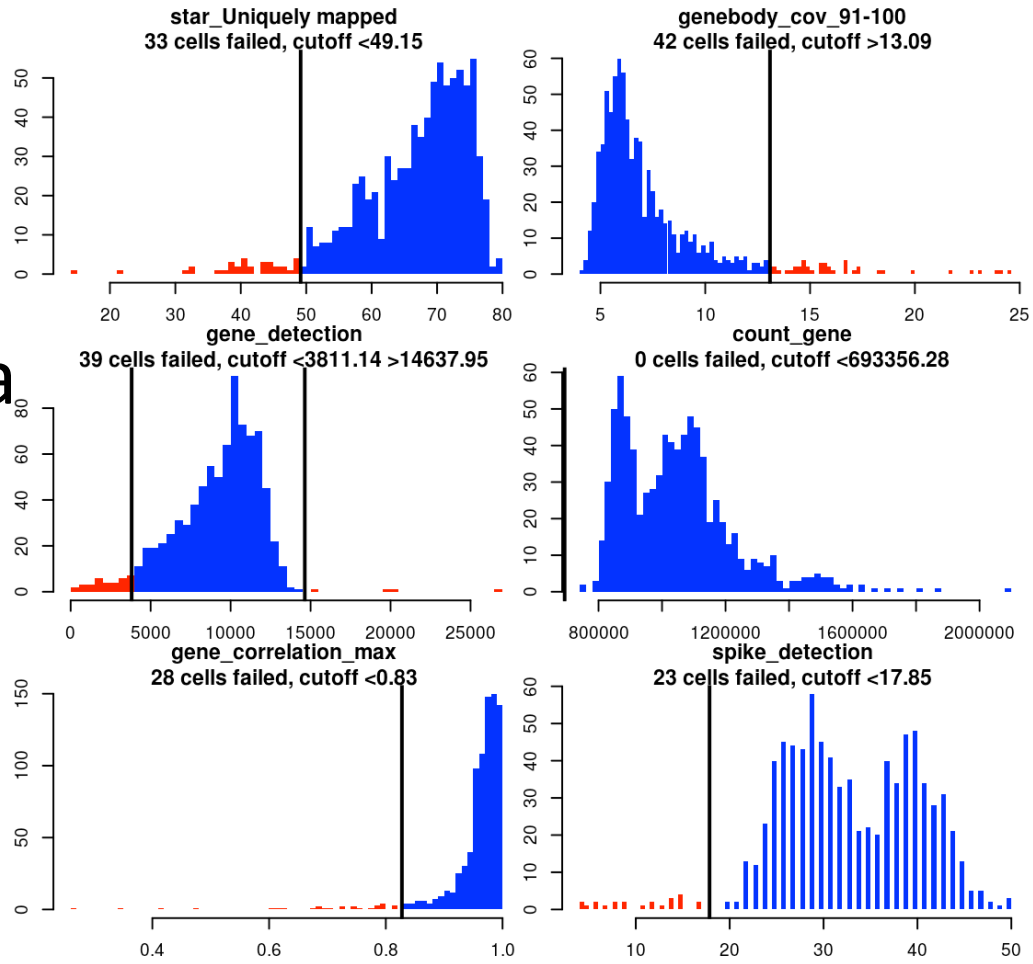

scRNA-seq analysis overview
RaukR 2018
Raw data: fastq files
Mapping & Gene expression estimate
QC: Remove low Q cells Remove contaminants
Data: Expression profiles
Data normalization
Gene set selection
Batch effect removal
Removal of other confounders
Visualization / Dimensionality reduction
Defining cell types/lineages
Clustering methods
Trajectory assignment
Gene signatures
Verification experiments

QC-metrics

- Starting amount of RNA
  - Number of reads
  - Mapping statistics (% uniquely mapping, fraction of exon mapping etc.)
  - Number of detected genes
  - Spike-in ratio
- RNA degradation:
  - 3’ bias – for full length methods like SmartSeq2
- Success of library prep:
  - Spike-in detection
- Quality of the cell
  - Mitochondrial read fraction
  - rRNA read fraction
  - Pairwise correlation to other cells
- Duplicate/multiple cells
  - High number of detected genes

How to filter cells

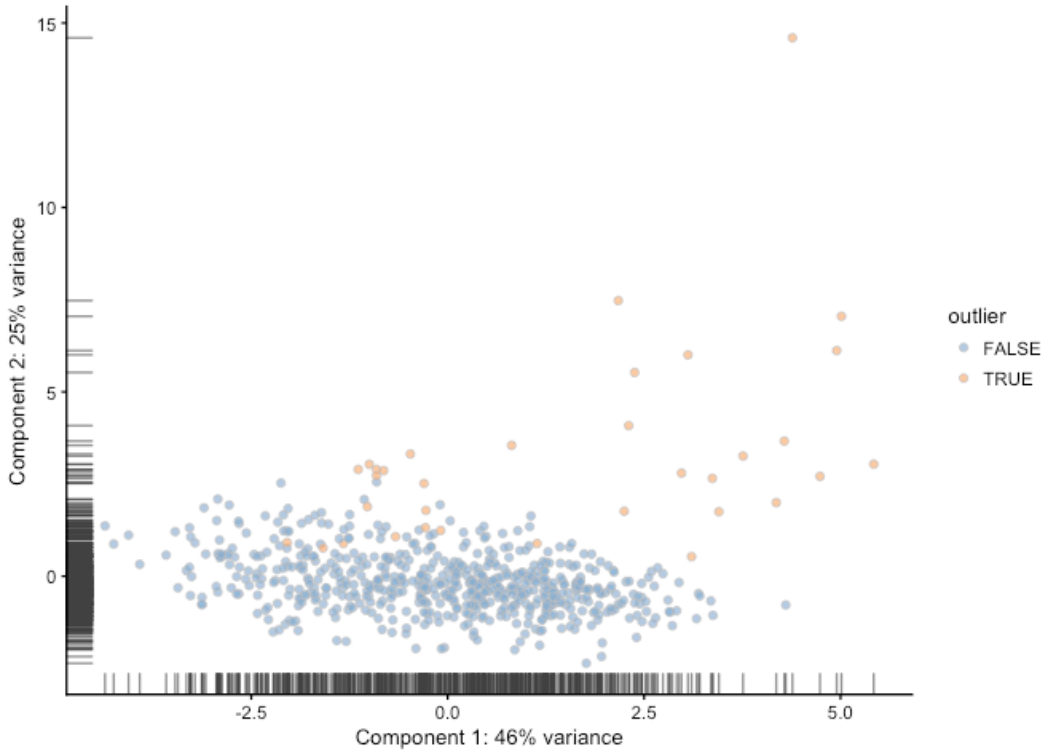
- Normally, most of these qc-metrics will show the same trends, so it could be sensible to use a combination of measures.
- Look at the distributions before deciding on cutoffs.



How to filter cells

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

  (Scater package)



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, t-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 – can be a sign of multiple cells
  - 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.

Data normalization

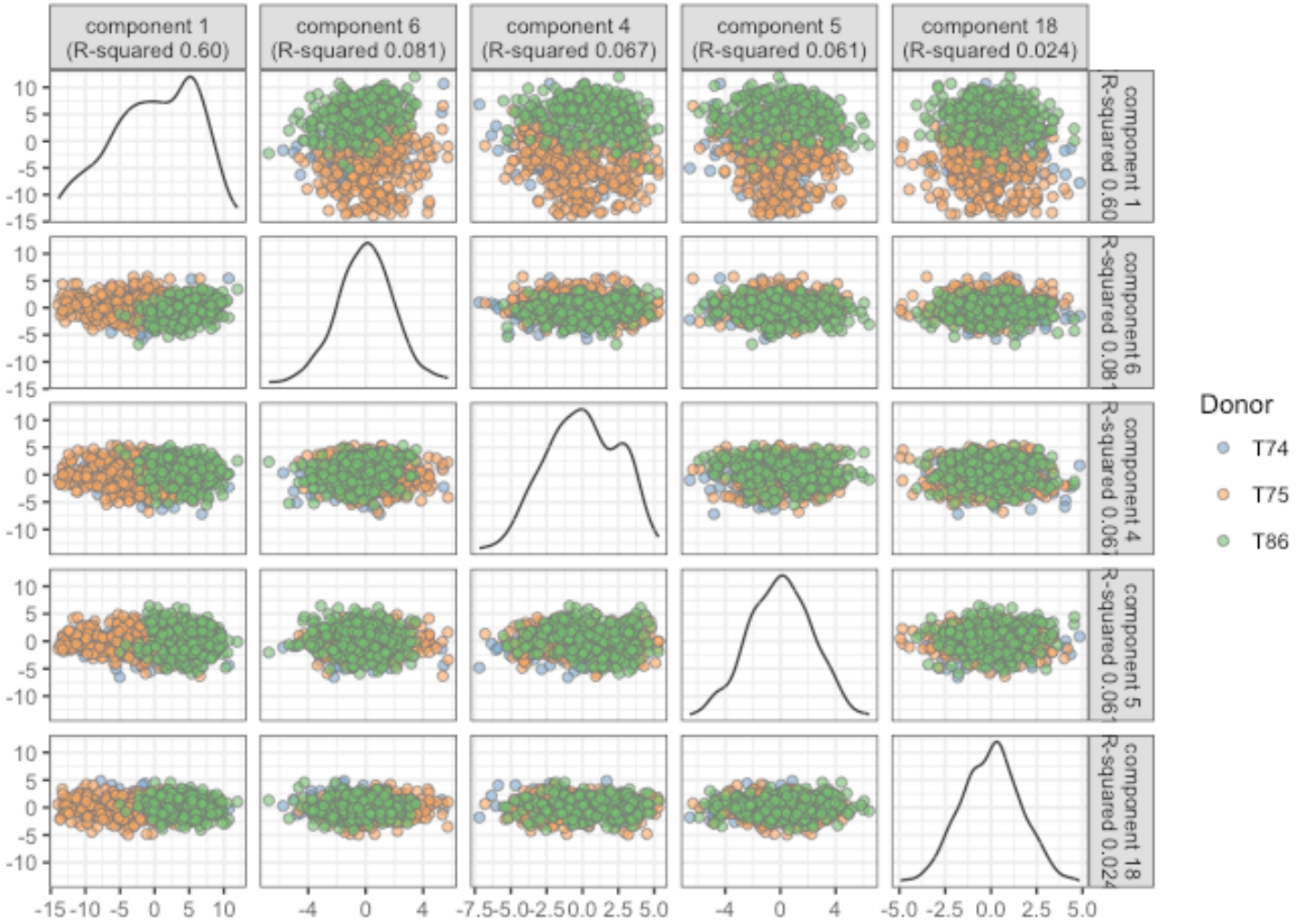
- Controlling for gene length - RPKMs, FPKM, TPMs
- Size factors from:
  - DESeq TMM, CPM
  - Deconvolution method (scran package) - Deals with the problem of vary large numbers of zeros by pooling cells together calculating a normalization for the sum of each pool.
- Error models in SCDE package − based on pairwise comparisons between cells, drop-out rates are estimated.
- Census algorithm to convert TPM to relative transcript counts.

Batch normalization

- Technical batches:
  - Dissociation batch
  - Library prep
  - Sequencing run
- Biological batches:
  - Cell culture
  - Animal
  - Patient

- Often clear biological variation between human samples.

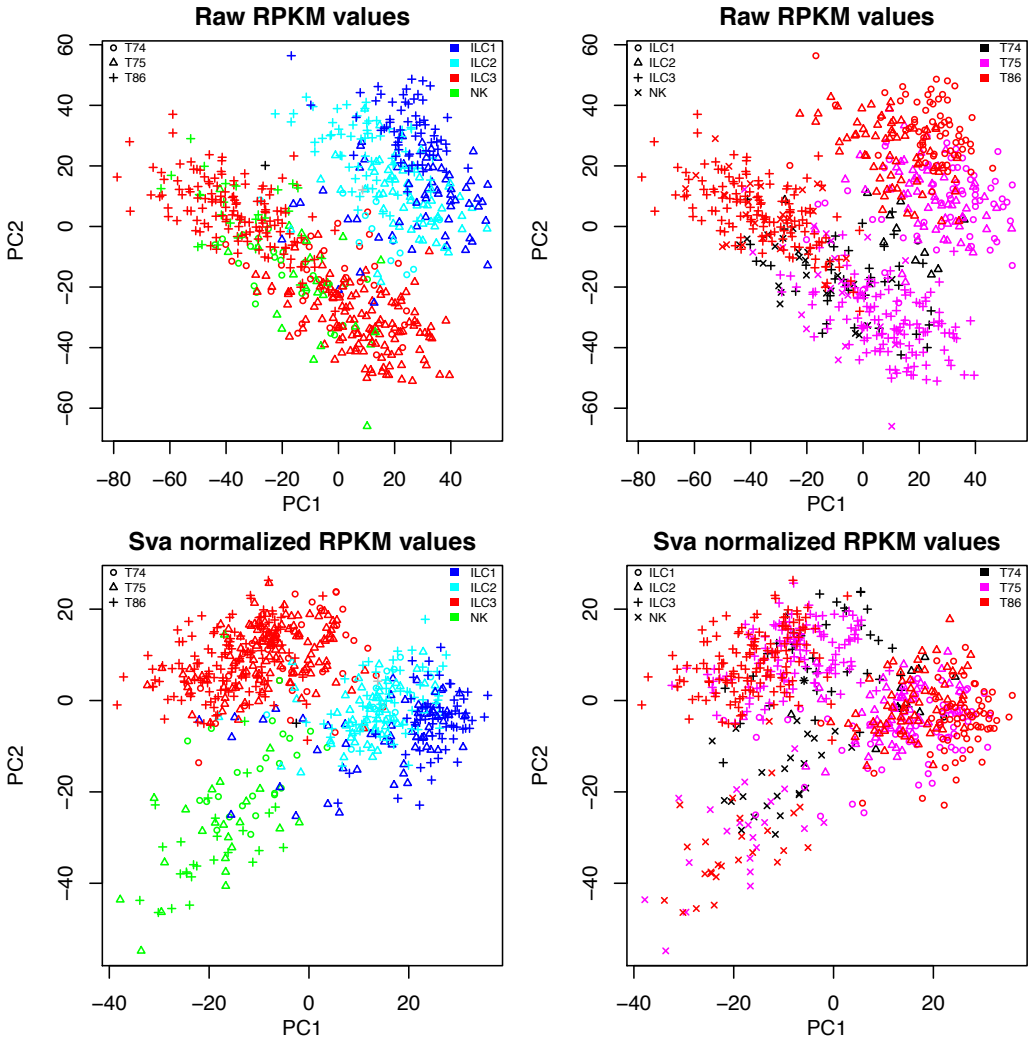
Check for batch effects in PCA



Scater package

Batch normalization with SVA function ComBat

Color by celltype



Color by donor

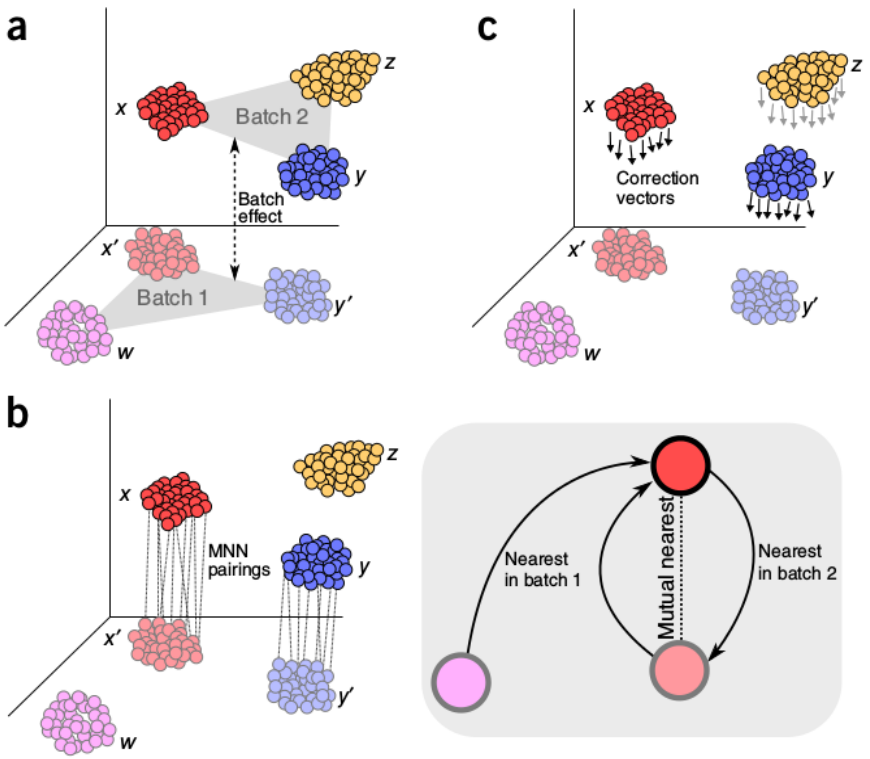
(Björklund *et al*. *Nature Immunology* 2016)

Combining different batches

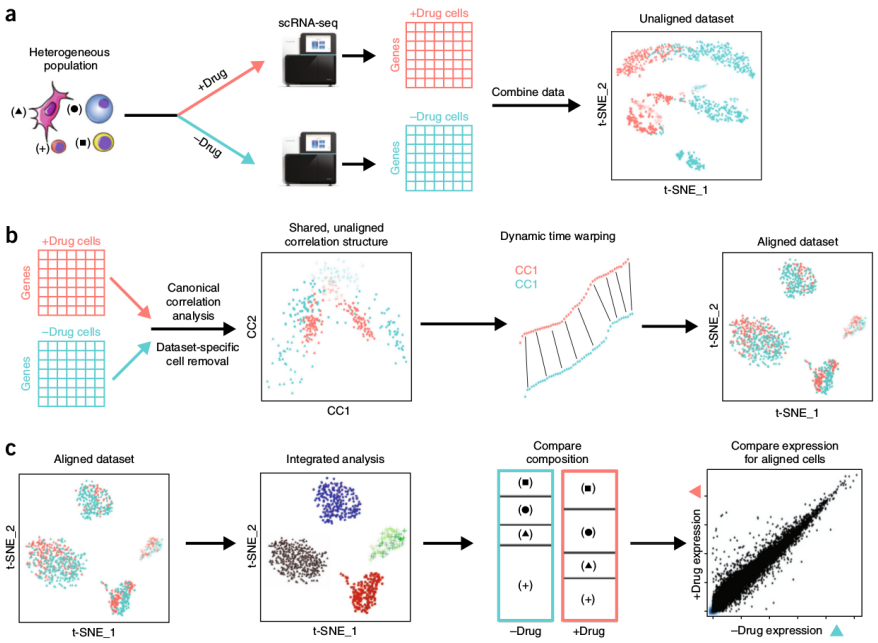
- SVA - ComBat function
- Seurat
  - ScaleData − linear model or generalized linear model (poisson, negative binomial)
  - Correlated Component Analysis (CCA)
- SCDE - pagoda.subtract.aspect − removes weighted projection of the expression matrix onto a specified aspect
- mnnCorrect - based on the detection of mutual nearest neighbors in the high-dimensional expression space

Other Methods for Batch Effects Corrections

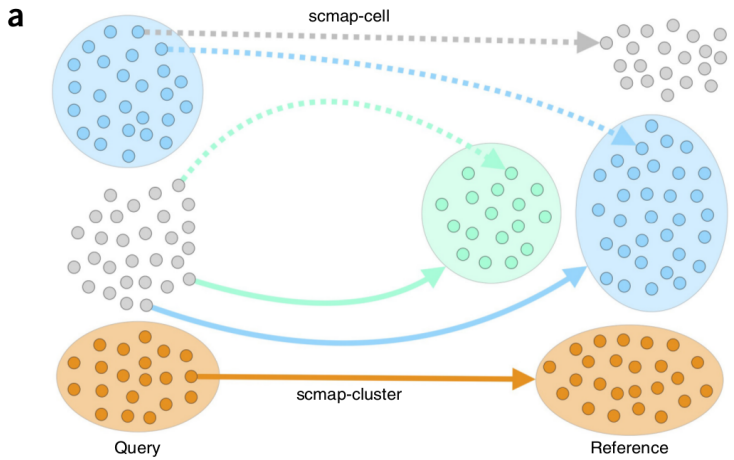
MNN



SEURAT

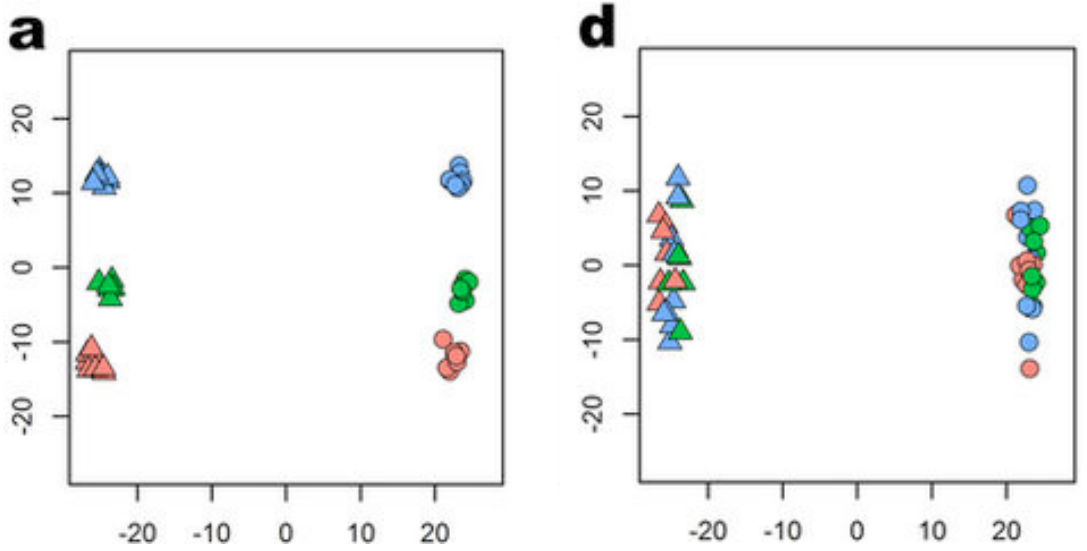


SCMAP



Remove other unwanted confounders

- Cell size / Number of detected genes
- Cell cycle
- Most methods use linear regression, implemented in SCDE, Seurat, scran etc.

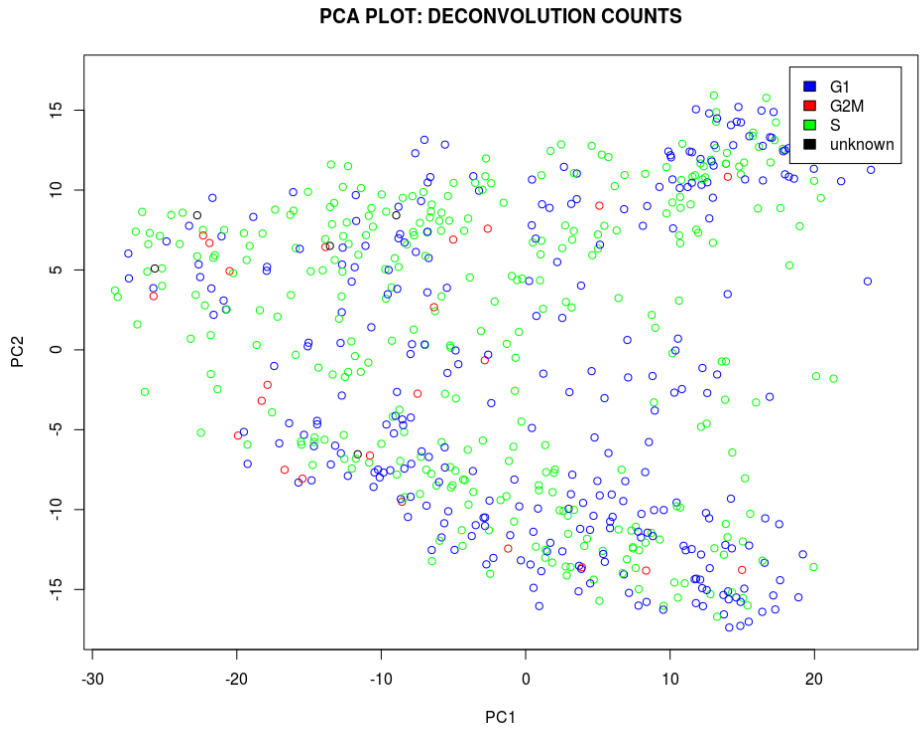


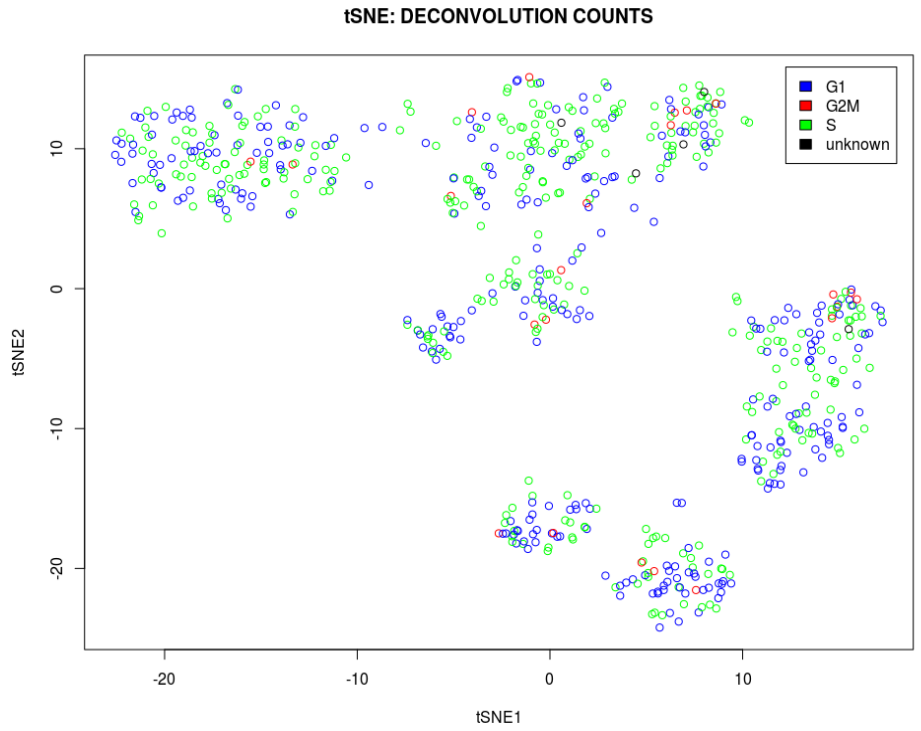
Color = CC phase
Shape = cell type

(ccRemover − Barron & Li, *Sci. Reports* 2016)

Cell Cycle Phase Assignment

Cyclone - Pre-trained classifier looks at pairs of genes having difference in expression that changes sign from phase to phase of cell cycle

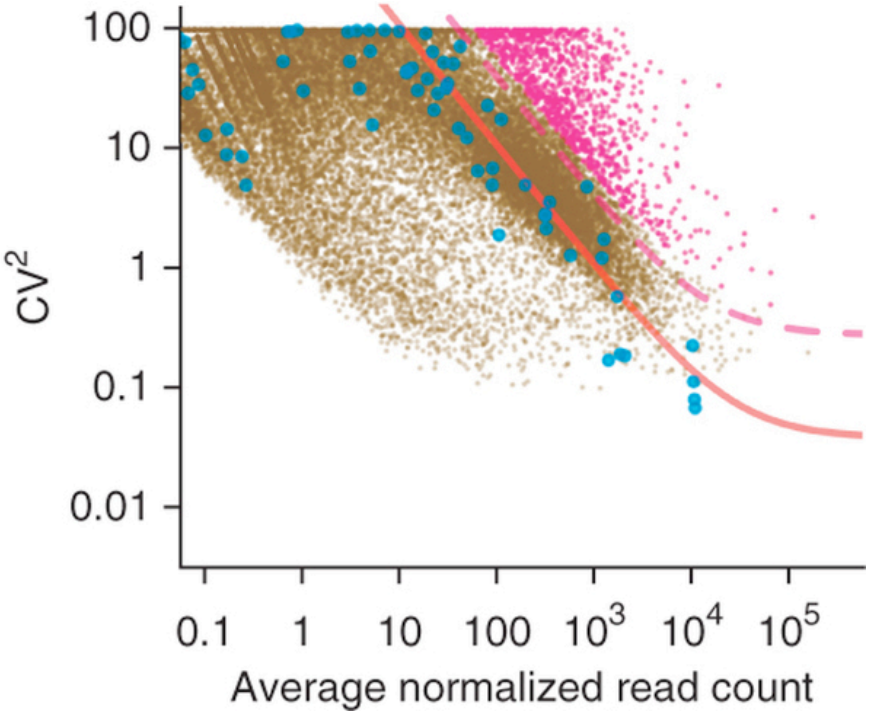




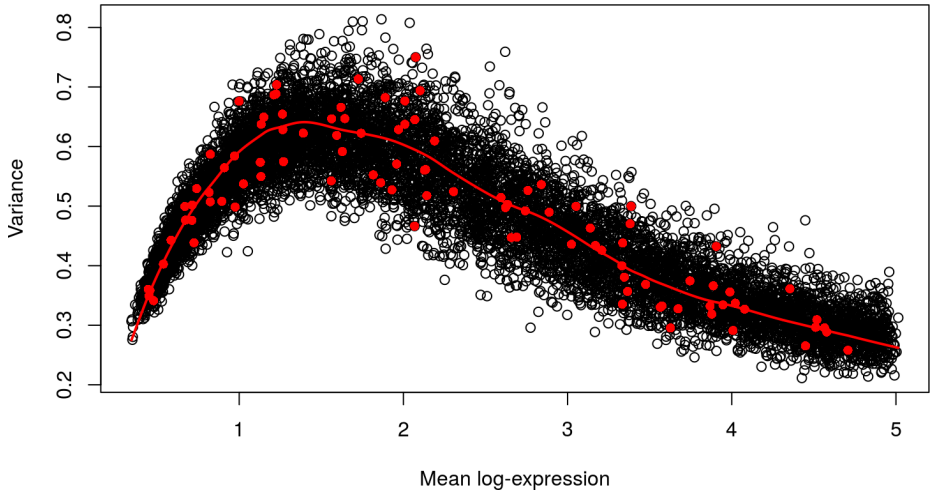
Feature (gene) selection

- In most cases, all genes are not used in analysis.
- Filtering based on:
  - Genes expressed over X in Y cells.
  - **Top variable genes** – implemented in most packages
  - Filter out genes with correlation to few other genes
  - Prior knowledge / annotation
  - DE genes from bulk experiments
  - Top PCA loadings
- Especially for trajectory analysis – important to build the trajectories on a relevant gene set

Spike-in RNAs
Finding biologically variable genes



Decompose variance of each gene into biological and technical component. Using spike-ins to model technical variation



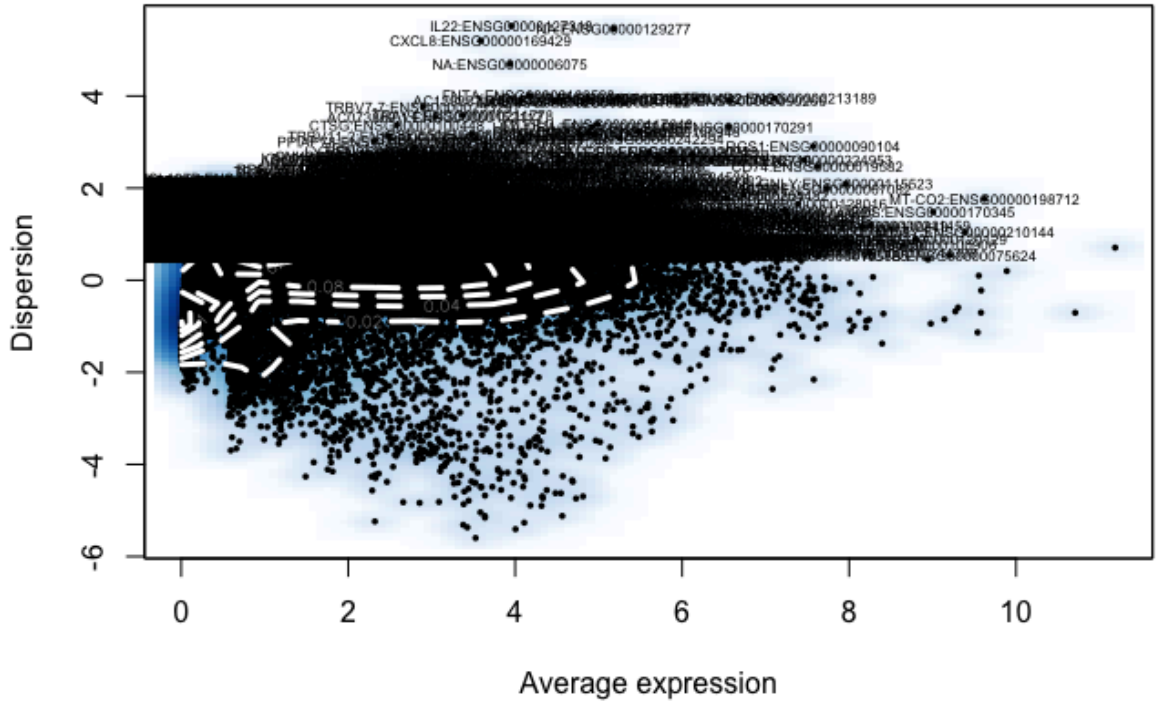
Coefficient of variation^2:

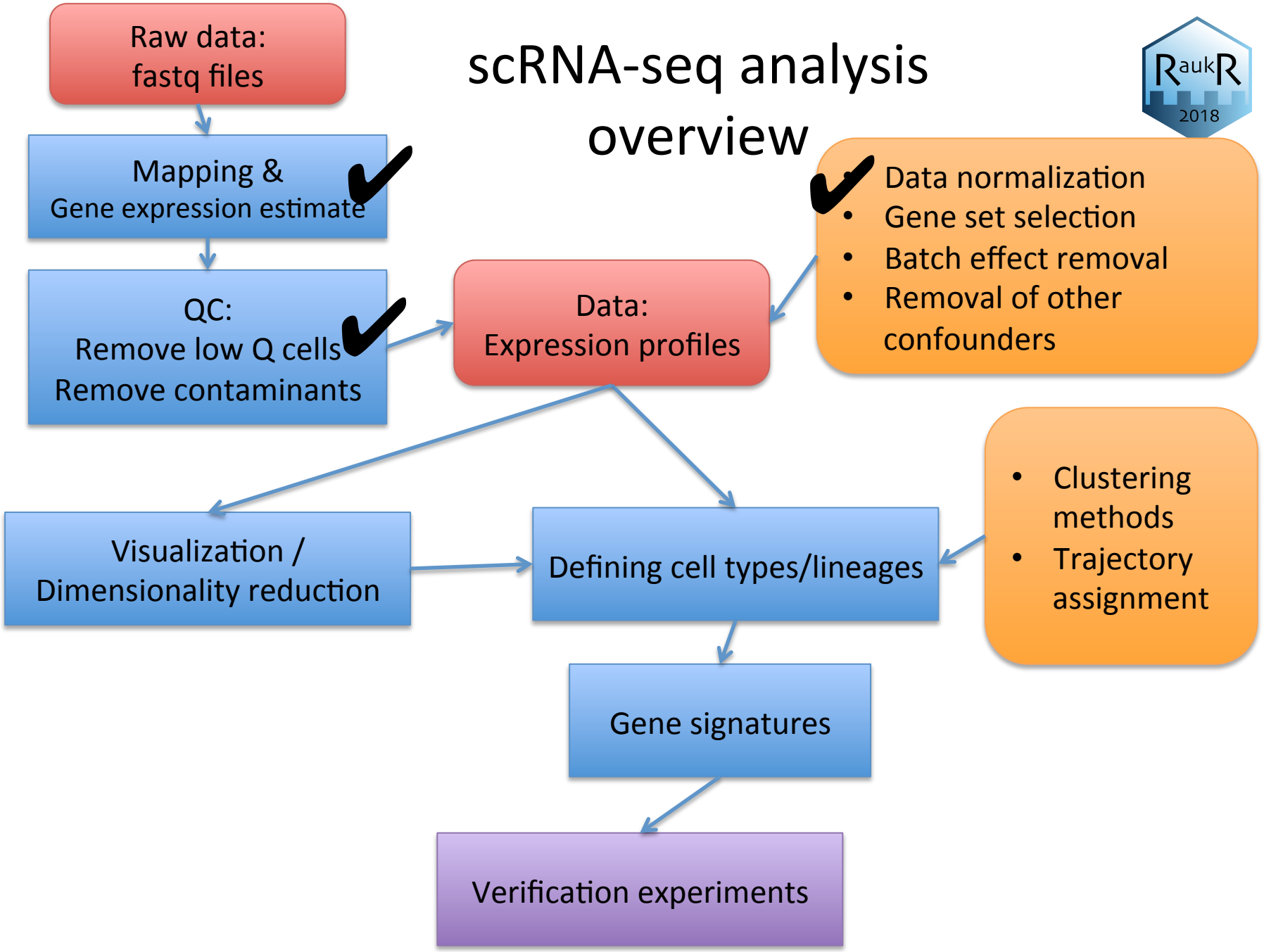
CV^2 = standard deviation / mean ^2

(Brennecke et al. *Nature Methods* 2013, Scran tutoral)

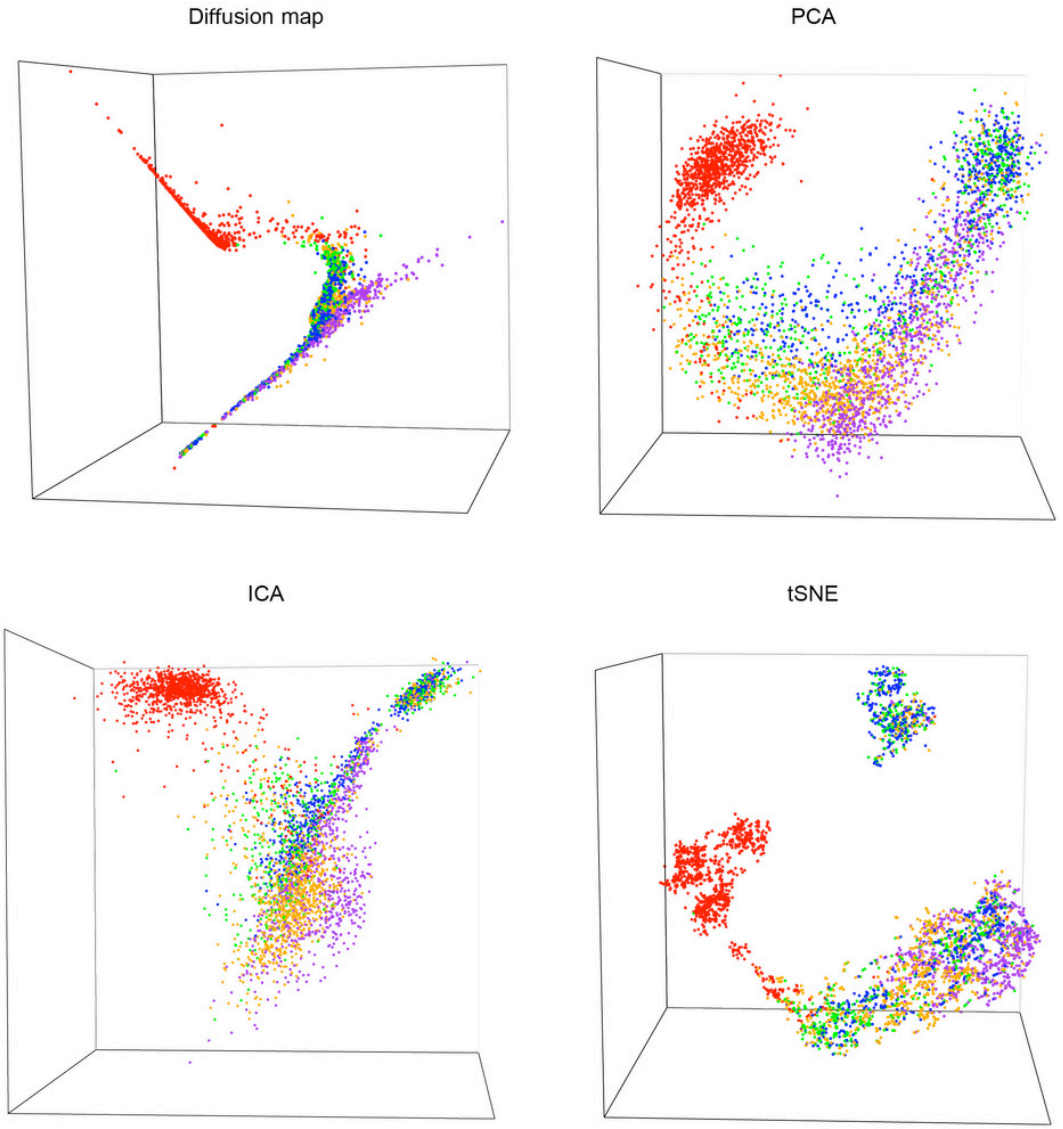
Top variable genes

Seurat uses Dispersion vs Average expression and you define cutoffs for x & y to select genes




scRNA-seq analysis overview
RaukR 2018
Raw data: fastq files
Mapping & Gene expression estimate
QC: Remove low Q cells Remove contaminants
Data: Expression profiles
Data normalization
Gene set selection
Batch effect removal
Removal of other confounders
Visualization / Dimensionality reduction
Defining cell types/lineages
Clustering methods
Trajectory assignment
Gene signatures
Verification experiments

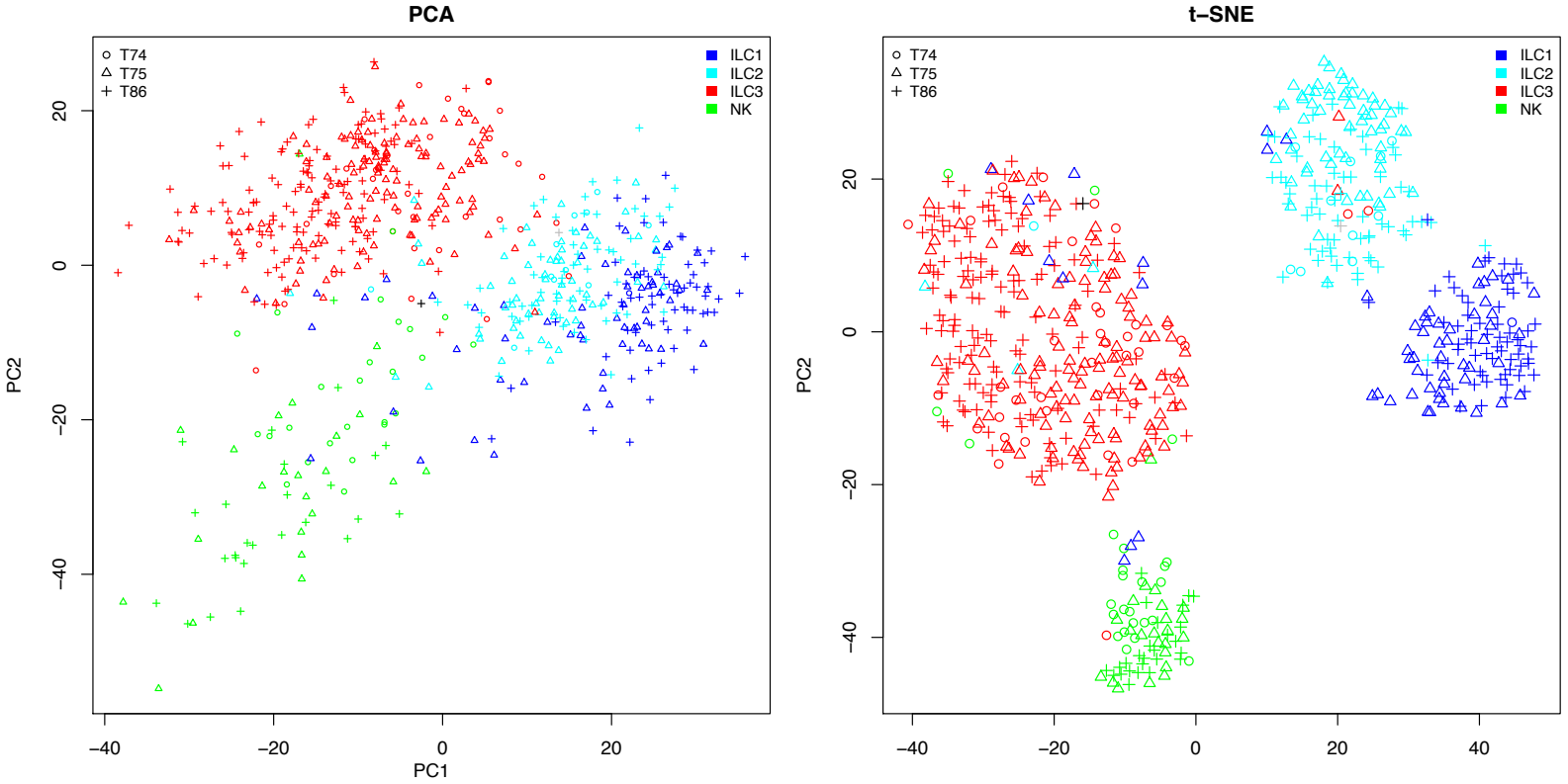
Visualization - Dimensionality



(Moignard *et al. Nature Biotech* 2015)

t-SNE – t-distributed stochastic neighbor embedding

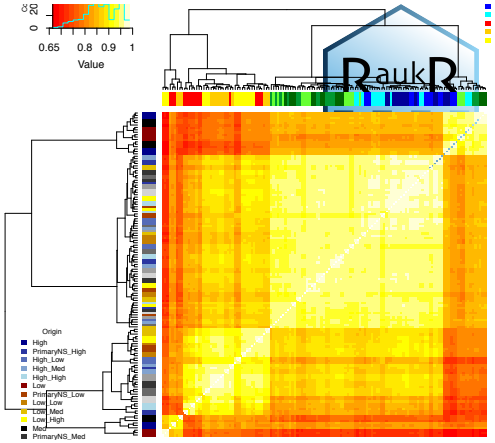
- Conserve pairwise distances in n-dimensional PCA space to 2D in local neighborhoods

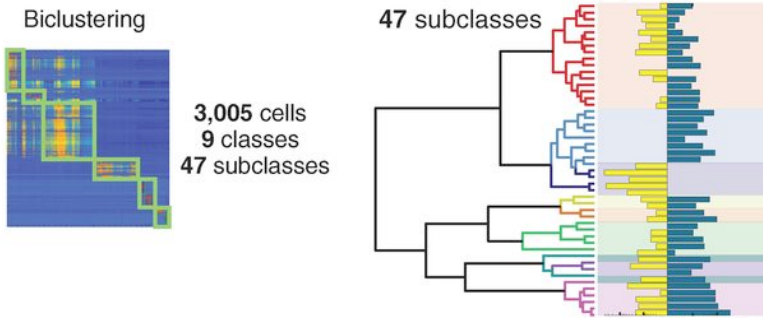


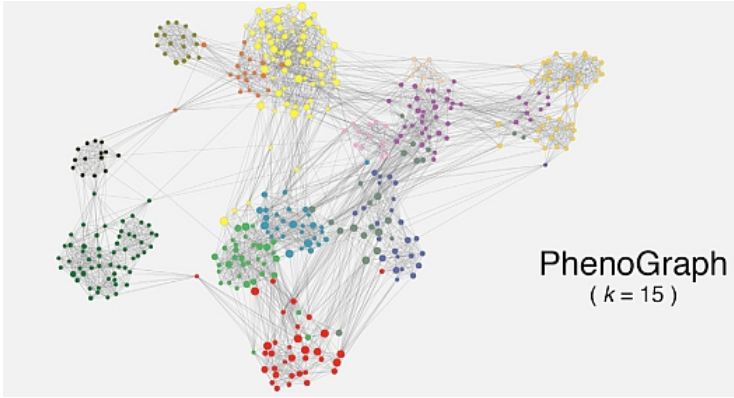
(Björklund *et al*. *Nature Immunology* 2016)

Clustering

- Classical clustering methods
  - Hierarchical
  - k-means clustering
- Biclustering
  - WGCNA
  - BackSPIN
- Graph based clustering
  - SNNCliq
  - Seurat
  - Pagoda2







Different distance measures

- Most commonly used in scRNA-seq:
  - Euclidean distance in multidimensional space (often with a reduced gene set)
  - Euclidean distance in PCA, tSNE or other reduced space
  - Inverted pairwise correlations (1-correlation)

Clustering

- About 1 paper/week with a new method for clustering single cells
- Consensus from several methods increases confidence in separation.
- Key challenges:
  - Deciding on the true number of clusters
  - Separation of true clusters from technical artifacts

Some toolkits for scRNAseq analysis

- Scater
- Scran
- Seurat
- SCDE
- Monocle

Scater

- Single Cell Experiment class for scRNAseq data:
  - Assays − counts, logcounts, tpms etc.
  - Reduced dimensions − PCA, tSNE, diffusion map
  - Spike in controls
  - Meta data
  - Feature data
- Contains functions for:
  - Visualization
  - Quality control, check for batch effects.
  - Normalization

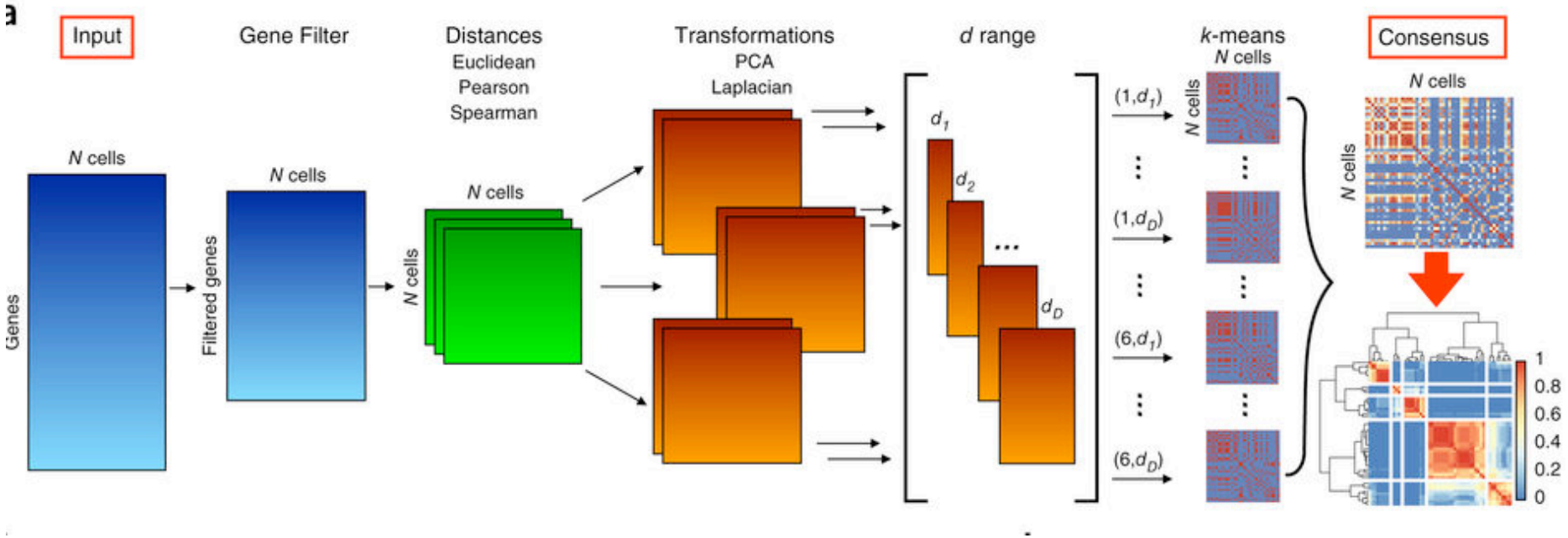
(McCarthy et al. Bionformatics 2017)

SCRAN – Single Cell RNA ANalisys

- Uses SingleCellExperiment class
- Includes cyclone method for predicting cell cycle phase.
- Includes Basics deconvolution strategy for size factors.
- Batch effect correction with mnnCorrect.
- Detection of variable genes by deconvolution of technical and biological variance.
- Also contains method for SNN graphs

http://bioconductor.org/packages/devel/bioc/vignettes/scran/inst/doc/scran.html

Single Cell Consensus Clustering – SC3



(Kiselev et al *Nat. Methods* 2017)

Single Cell Consensus Clustering – SC3

1. Gene filtering – rare and ubiquitous genes
2. Distance matrices (DM) – Euklidean, Spearman, Pearson
3. Transformation of DM with PCA or Laplacian
4. K-means clustering with first *d* eigenvectors
5. Consensus clustering – distance 1/0 for cells in same/ different clusters -> hierarchical clustering on average distances.

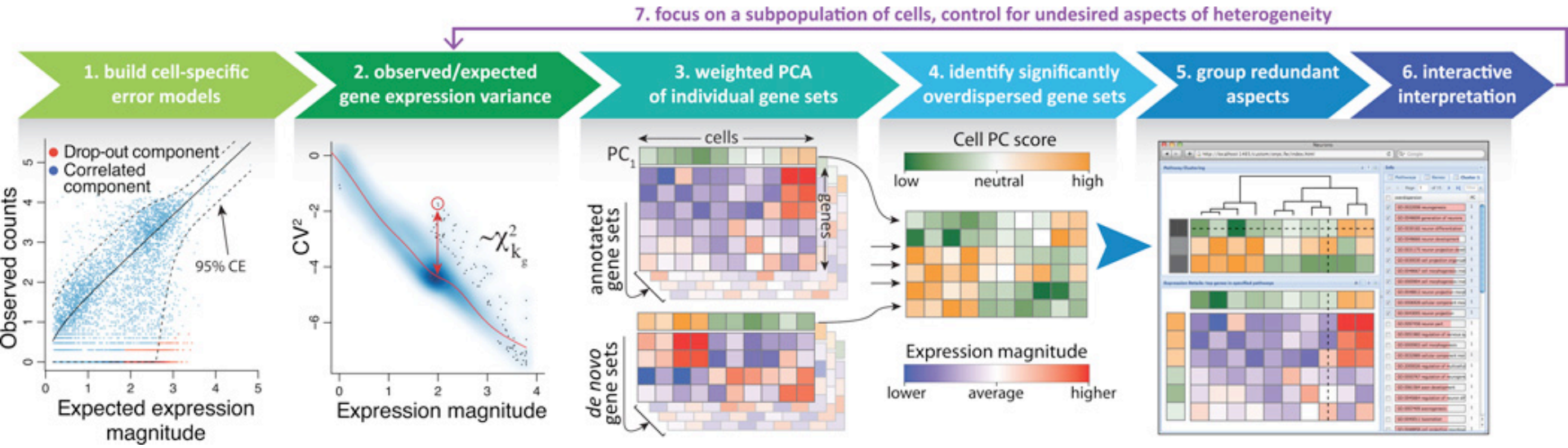
Differential expression with nonparametric Kruskal–Wallis test.

Marker genes with areas under the ROC curve (AUROC) from 100 permutations of cell cluster labels and P-values from Wilcoxon signed-rank test.

(Kiselev et al *Nat. Methods* 2017)

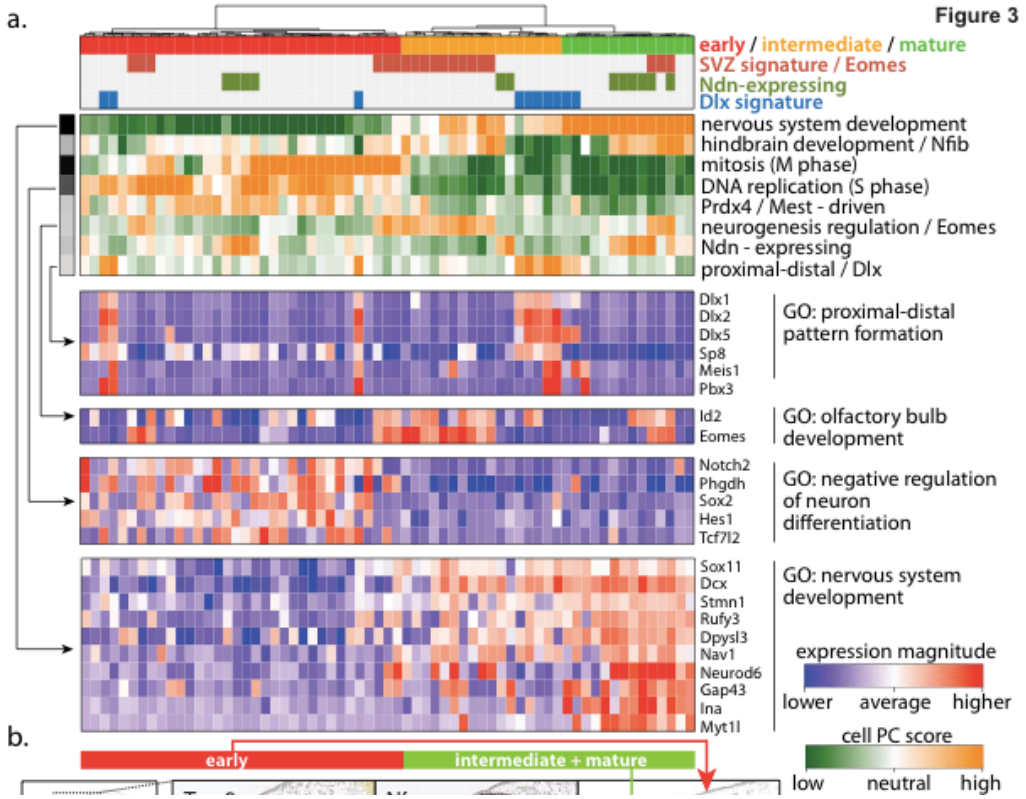
Pagoda – Pathway And Geneset OverDispersion Analysis

Implemented in the SCDE package



(Fan et al. *Nature Methods* 2016)

Pagoda – Pathway And Geneset OverDispersion Analysis



- Helps with biological interpretation of data
- Important to have good and relevant gene sets
- High memory consumption when running Pagoda
- Also has methods for removing batch effect, detected genes, cell cycle etc

(Fan et al. *Nature Methods* 2016)

Pagoda2

- Similar error modelling
- Now include KNN graph clustering
- Can visualize gene sets.
- https://github.com/hms-dbmi/pagoda2

Seurat

- Developed for drop-seq analysis − compatible with 10X output files. But works also for other types of data.
- Contains function for
  - Data normalization
  - Detection of variable genes
  - Regression of batch effects and other confounders
  - JackStraw to detect significant principal components
  - tSNE and other dimensionality reduction techniques
  - Clustering based on SNN graphs
  - Many different methods for Differential expression

(http://satijalab.org/seurat/)

Seurat - FindClusters

- First construct a KNN (k-nearest neighbor) graph based on the euclidean distance in PCA space.
  - Select which principal components to include
- Refine the edge weights between any two cells based on the shared overlap in their local neighborhoods (Jaccard distance).
- To cluster the cells, modularity optimization techniques to iteratively group cells together.
- **OBS!** Earlier versions of Seurat uses “spectral tSNE” and DBScan density clustering.

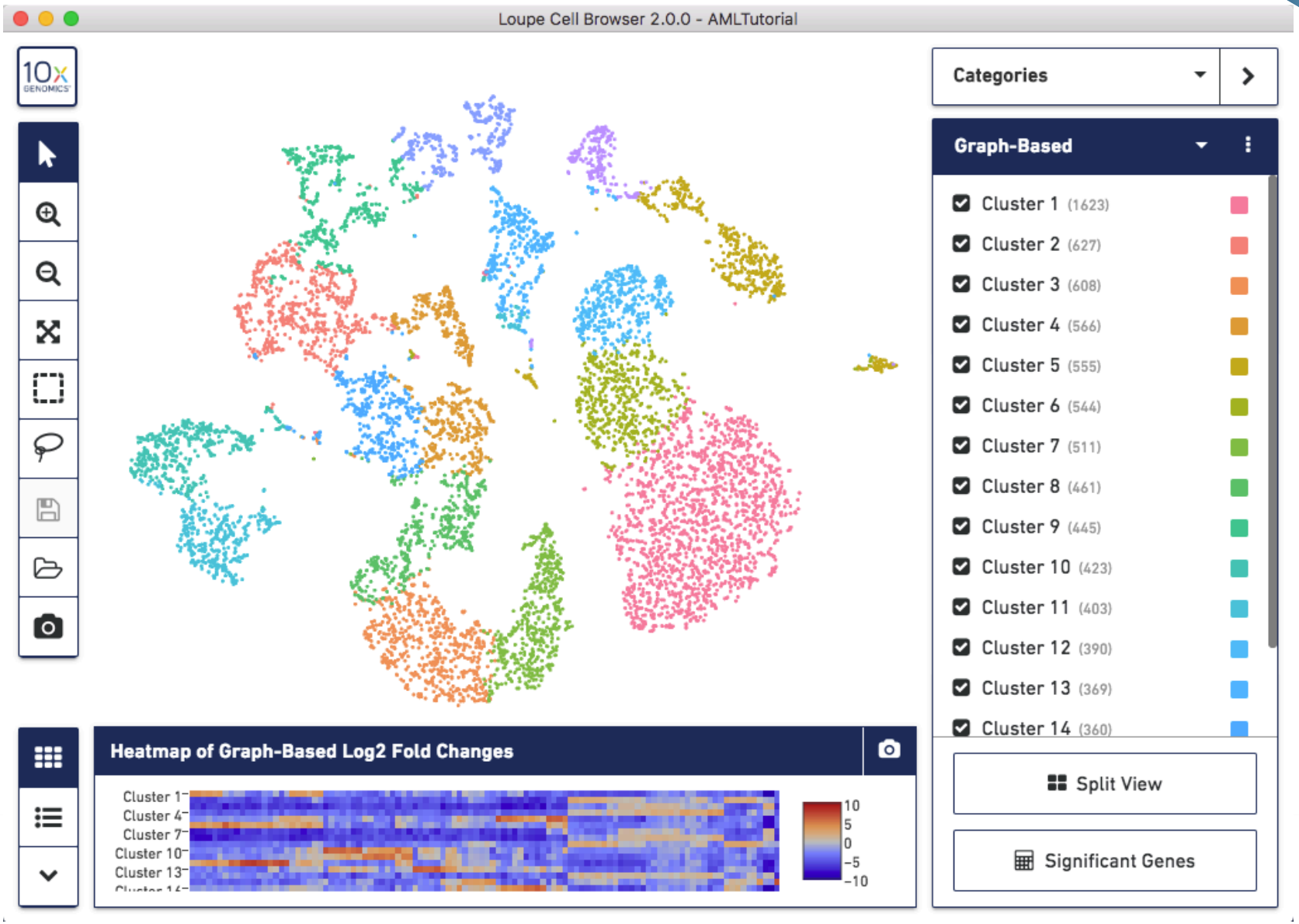
(http://satijalab.org/seurat/)

Seurat

- Also contains functions for:
  - Spatial reconstruction of single cell data using *in situ* references (Zebrafish embryos)
  - Integrated analysis across platforms with Correlated Component Analysis (CCA)
  - Analysis of multimodal datasets (e.g. RNA + protein)

(http://satijalab.org/seurat/)

Loupe – Cell Browser, from 10X Genomics



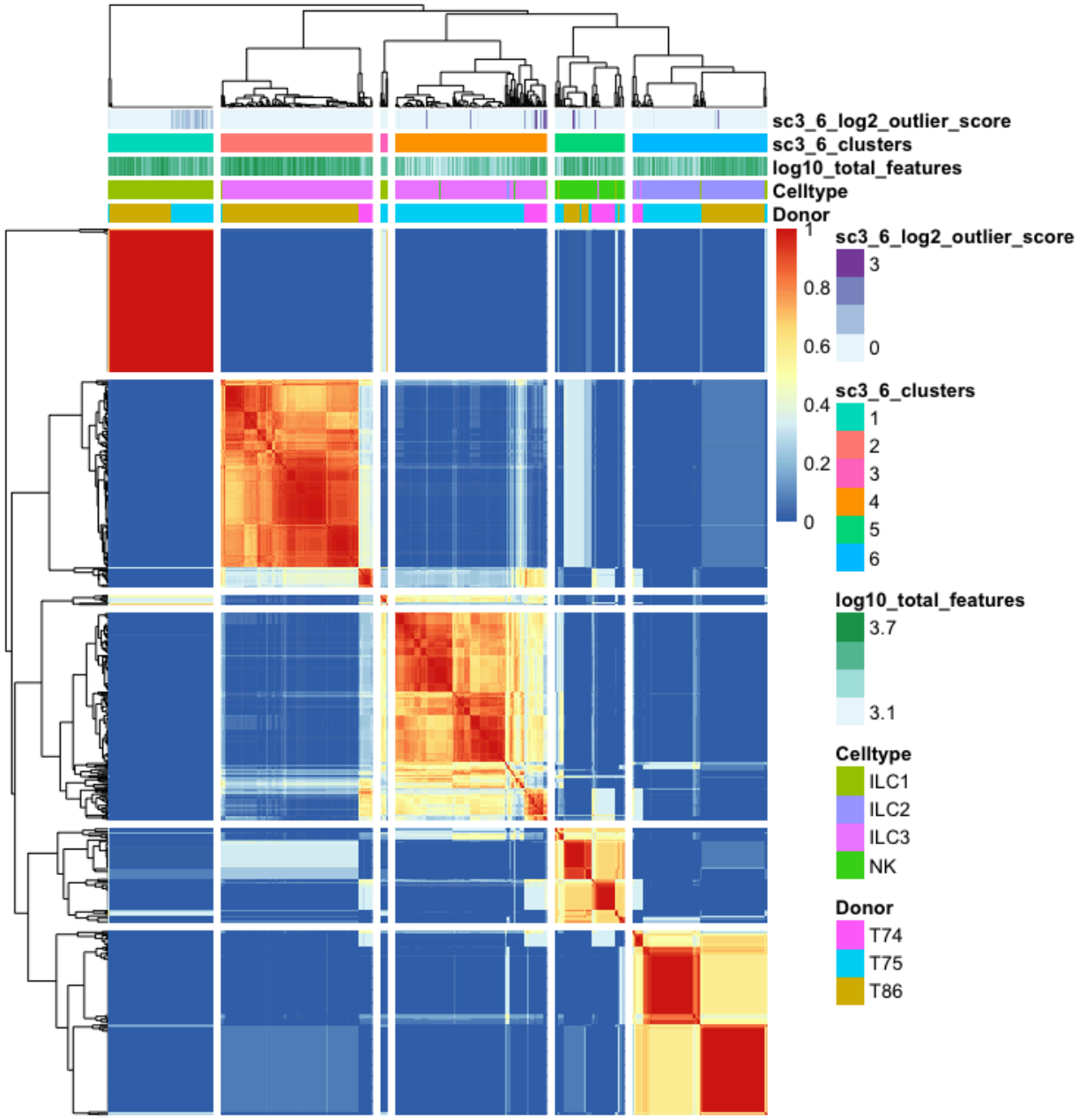
Which clustering method is best?

- Depends on the input data
- Consistency between several methods gives confidence that the clustering is robust
- The clustering method that is most consistent – best bootstrap values is not always best
- In a simple case where you have clearly distinct celltypes, simple hierarchical clustering based on euklidean or correlation distances will work fine.

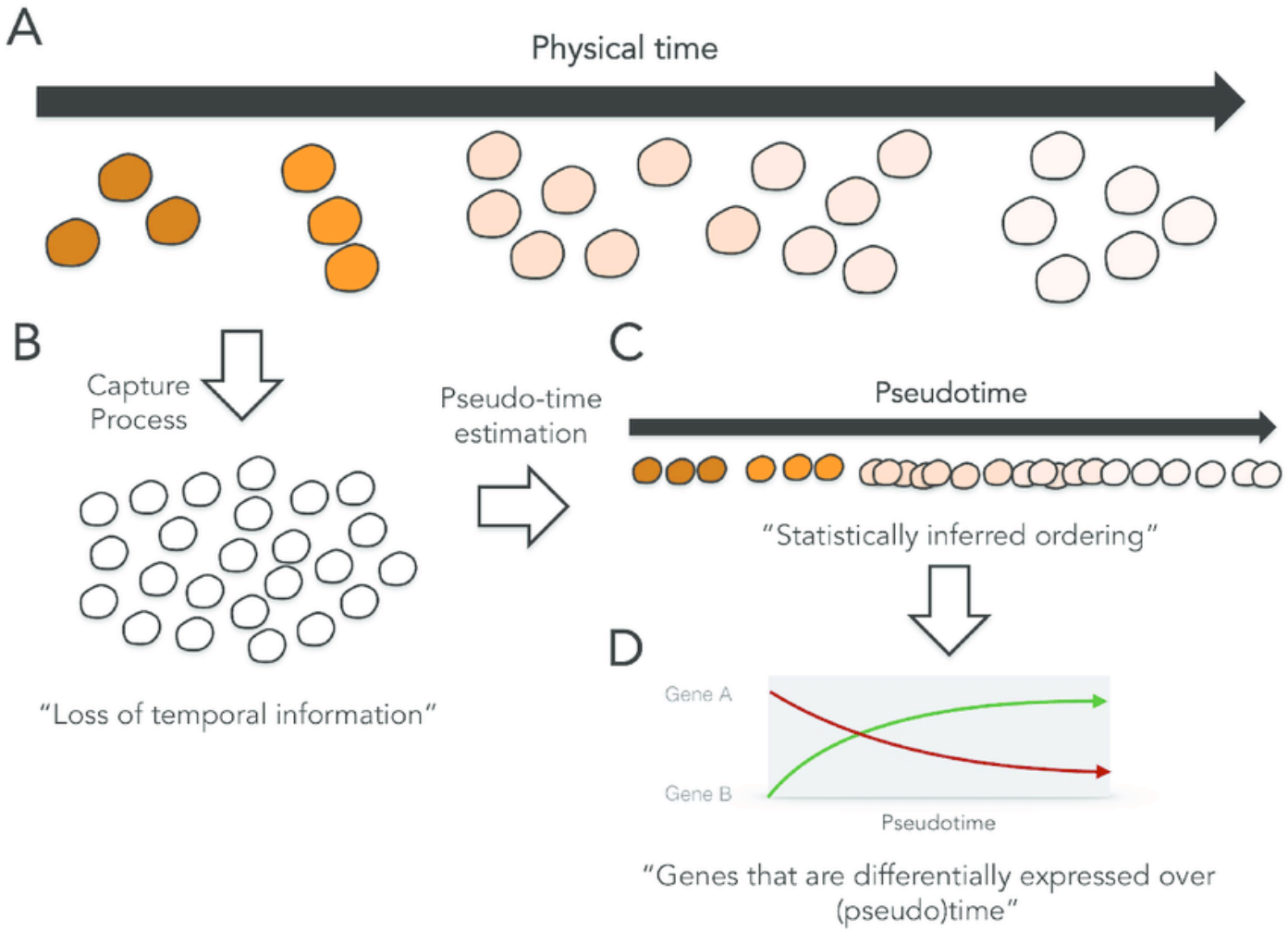
How many clusters do you really have?

- It is hard to know when to stop clustering – you can always split the cells more times.
- Can use:
  - Do you get any/many significant DE genes from the next split?
  - Some tools have automated predictions for number of clusters – may not always be biologically relevant
- Always check back to QC-data – is what your splitting mainly related to batches, qc-measures (especially detected genes)

Check QC data

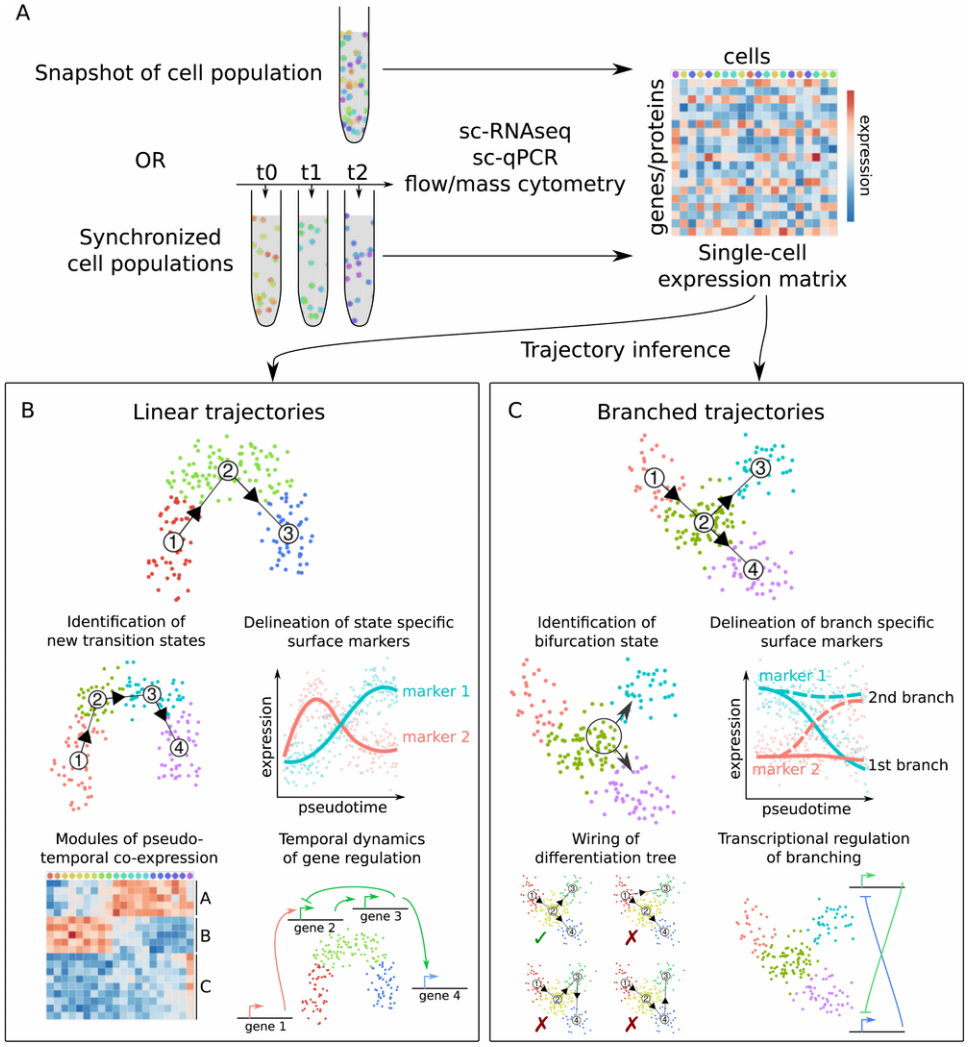


Pseudotime/trajectory analysis



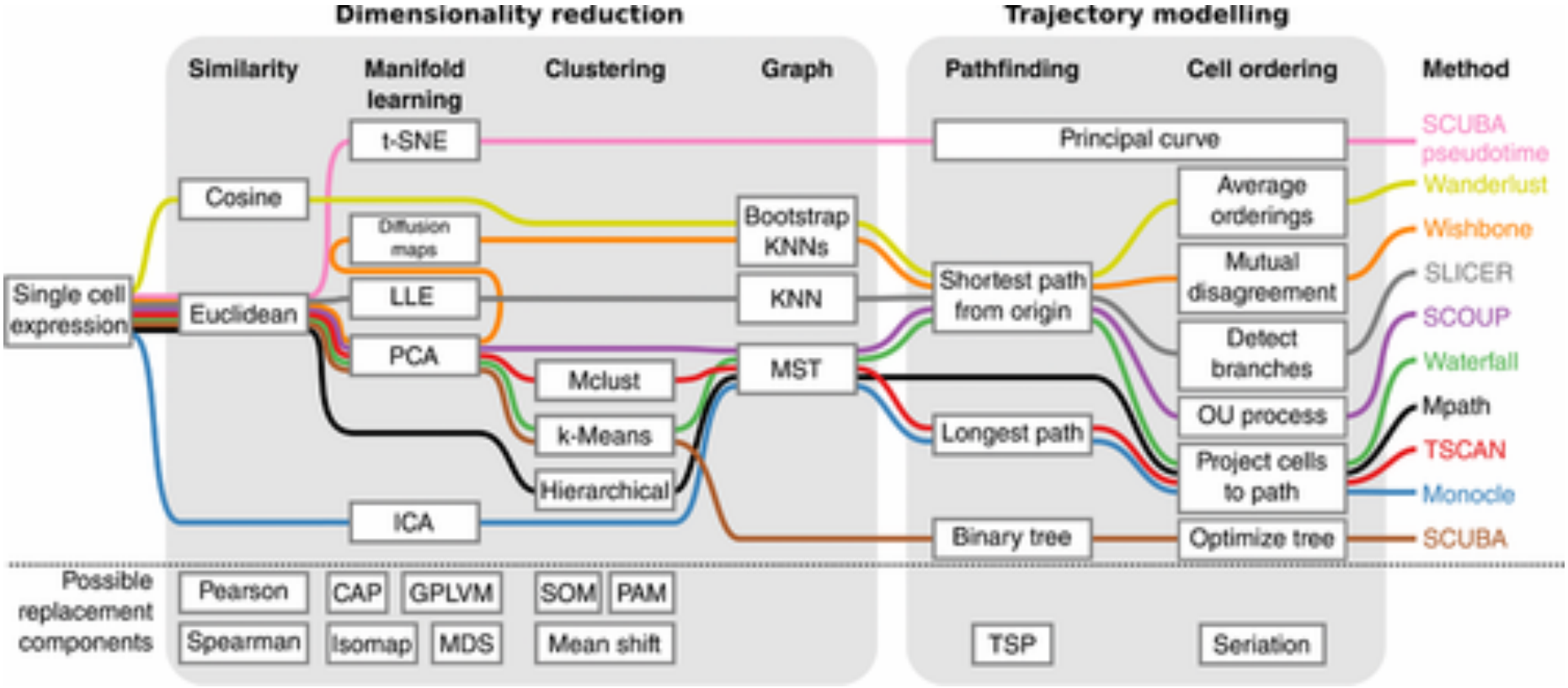
(Kieran *et al*. *Plos Comp Biol*. 2017)

Pseudotime/trajectory analysis



(Cannoodt *et al*. *EJI* 2016)

Trajectory tools



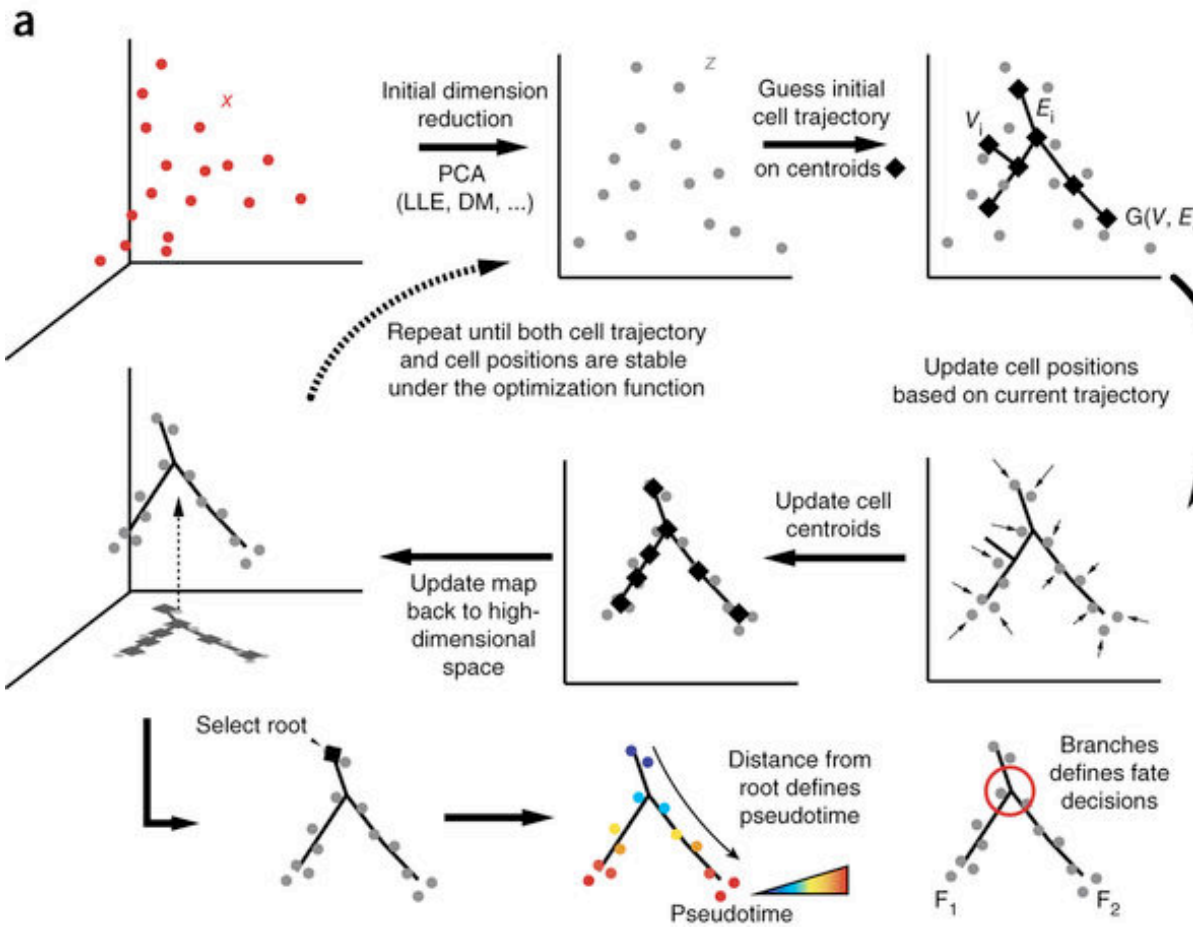
(Cannoodt *et al*. *EJI* 2016)

Should you run trajectory analysis?

- Are you sure that you have a developmental trajectory?
- Do you believe that you have branching in your trajectory?
- Do you have enough cells along the trajectory?

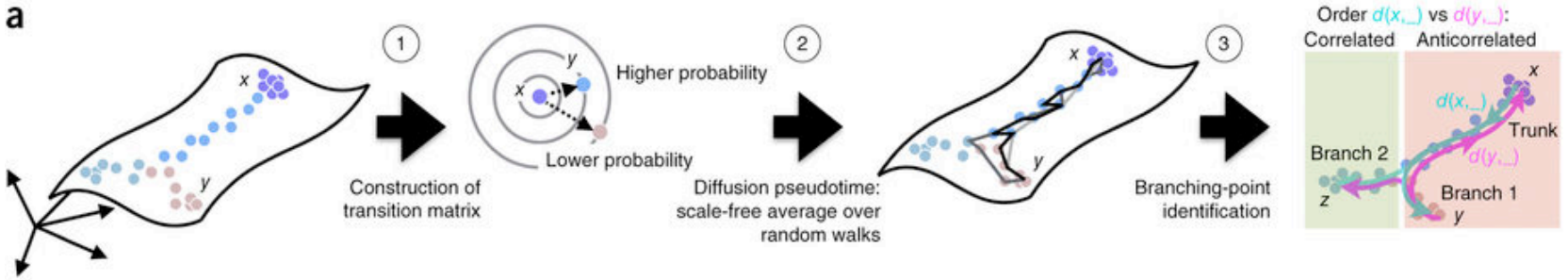
- Be aware, any dataset can be forced into a trajectory without any biological meaning!
- First make sure that gene set and dimensionality reduction captures what you expect.

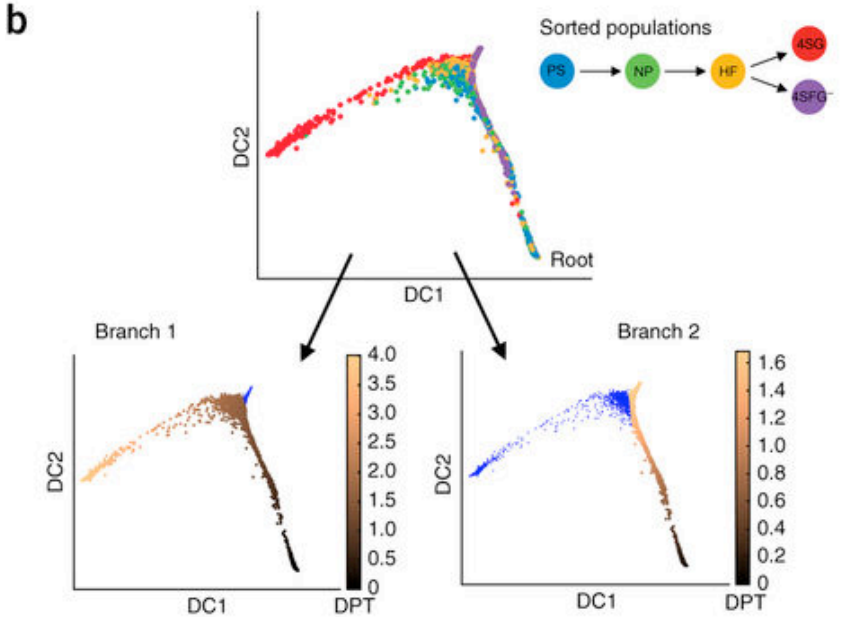
Monocle2 – reversed graph enbedding

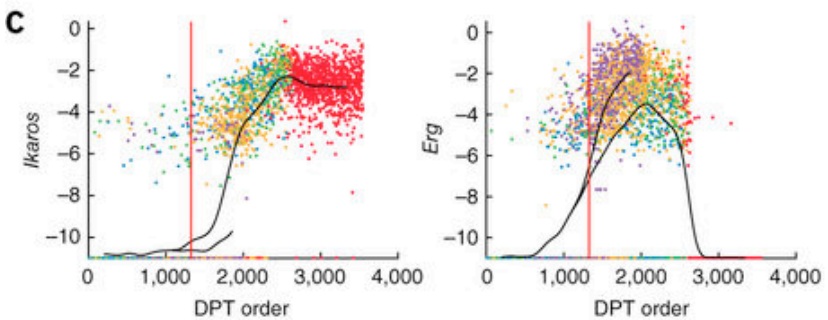


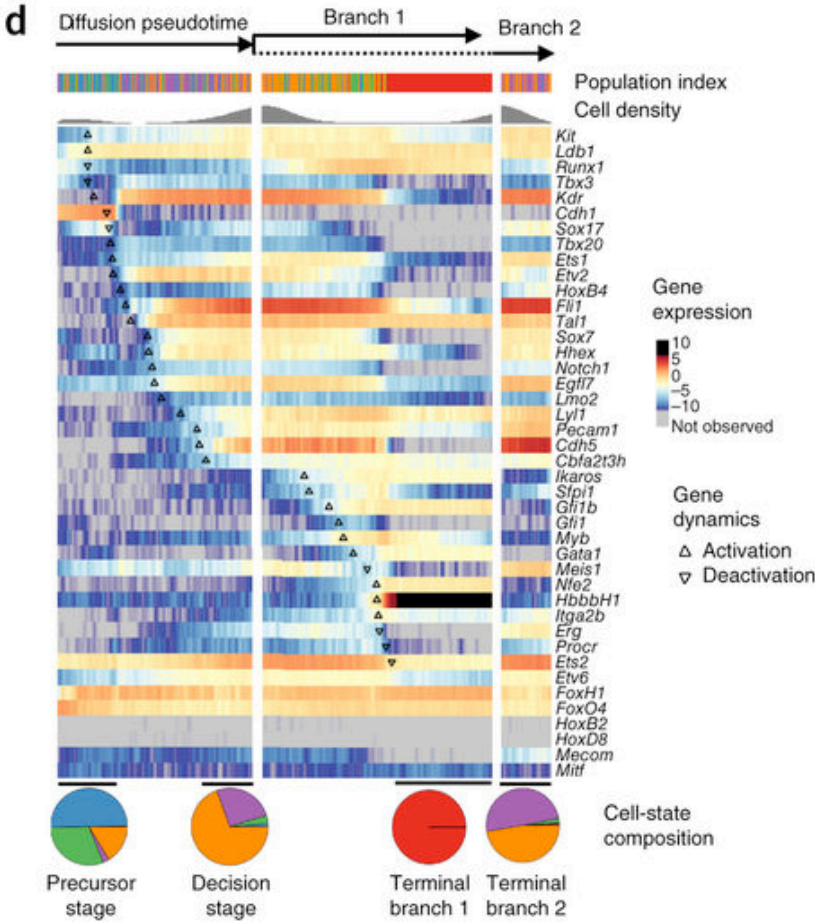
(Qiu et al. *Nat Methods* 2017)

Diffusion pseudotime (DPT) (Haghvedi *et al.* *Nature Methods* 2016)





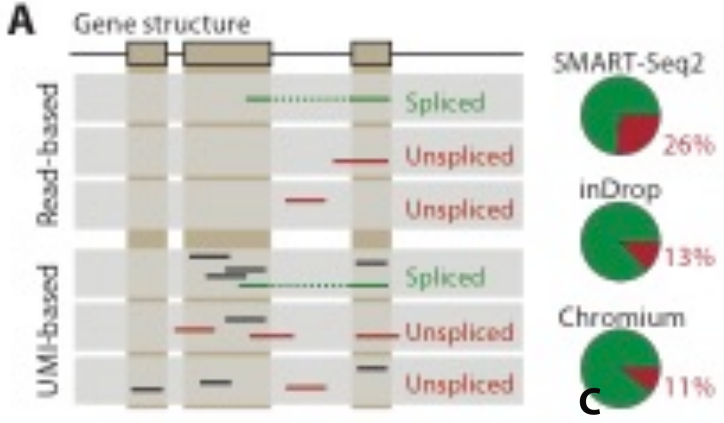


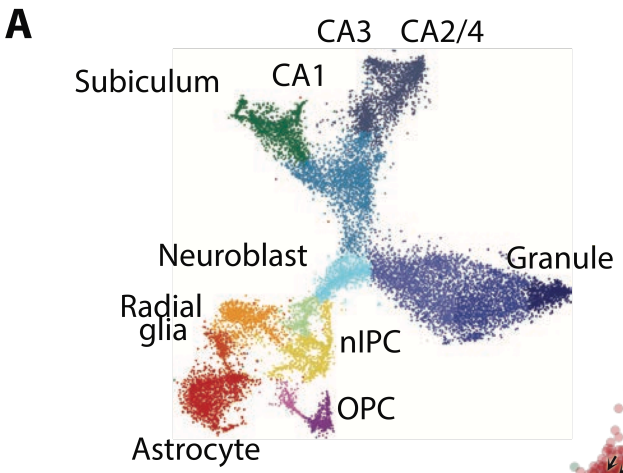


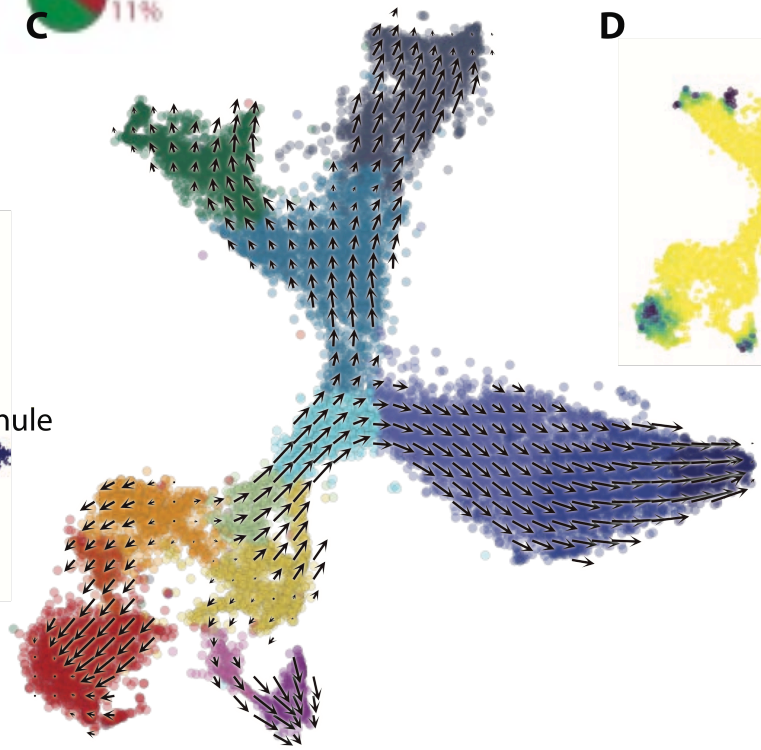
RNA velocity

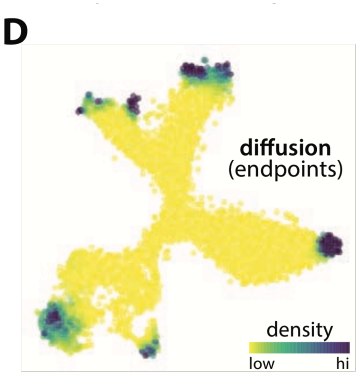
Uses proportion of unspliced reads to identify recently activated genes.

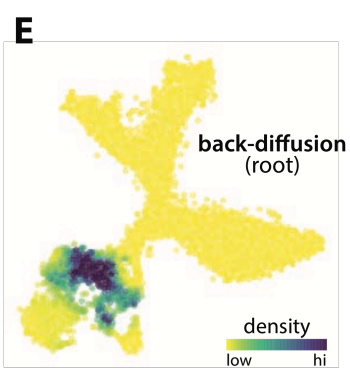
Predicts future fate of the cells.

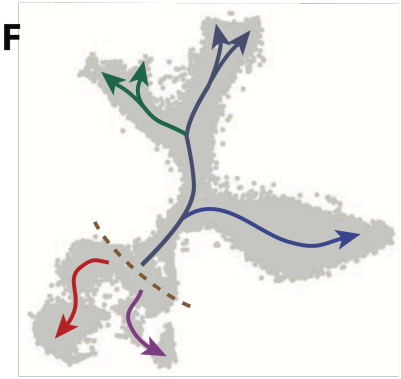




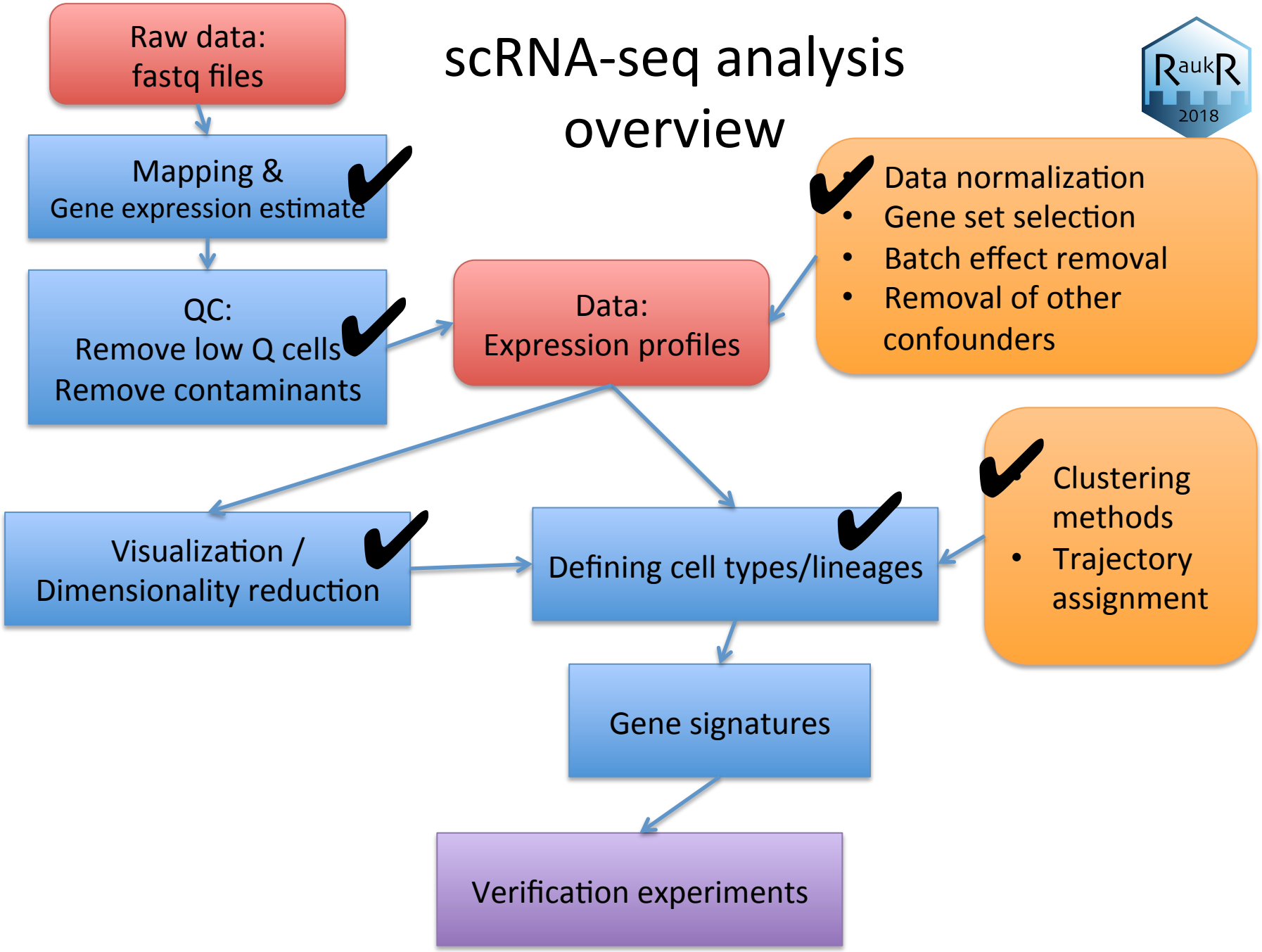








(LaManno *et al*. BioRxiv 2017)

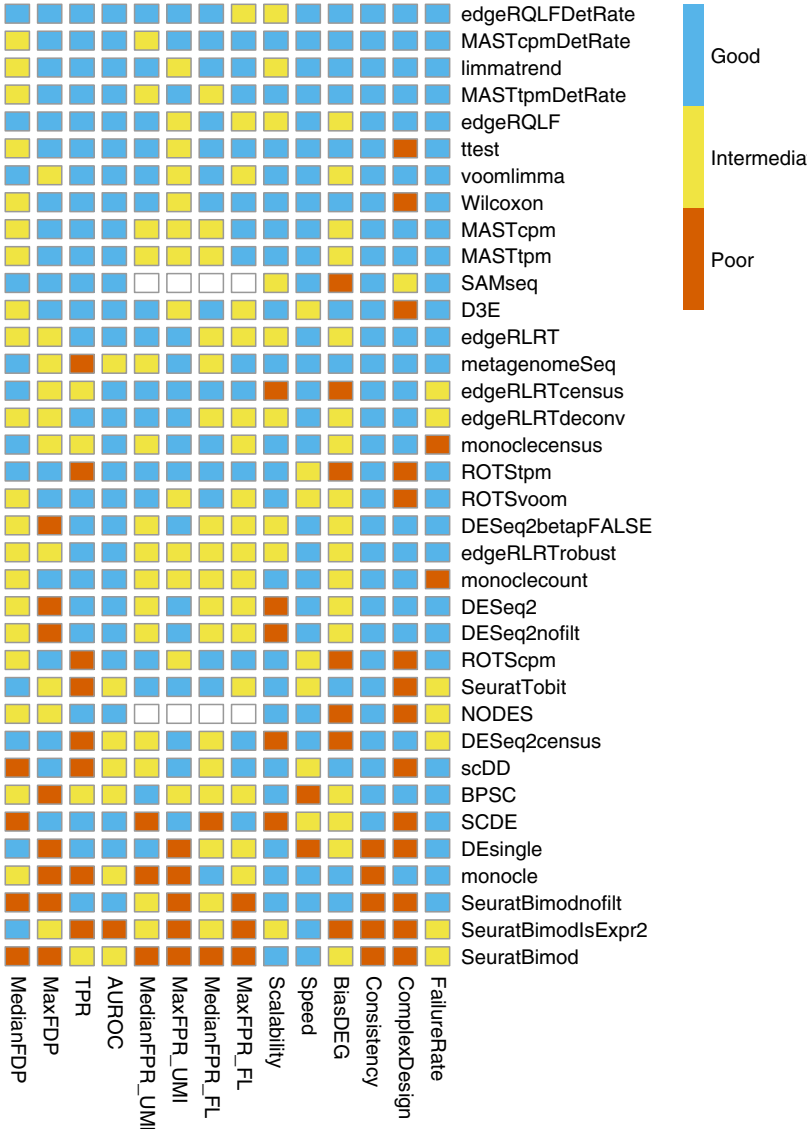

scRNA-seq analysis overview
RaukR 2018
Raw data: fastq files
Mapping & Gene expression estimate
QC: Remove low Q cells Remove contaminants
Data: Expression profiles
Data normalization
Gene set selection
Batch effect removal
Removal of other confounders
Visualization / Dimensionality reduction
Defining cell types/lineages
Clustering methods
Trajectory assignment
Gene signatures
Verification experiments

Differential expression in scRNAseq data

- scRNAseq is affected by higher noise (technical and biological)
- Usually scRNAseq data have many more replicates, can make use of the distributions for each gene, not only the mean + dispersion.
- What is used:
  - Bulk methods: edgeR, limma, DESeq, SAMseq
  - SC methods: MAST, SCDE, Monocle, D3E
  - Classical tests: Kruskal-Wallis, ROTS, t-test, RandomForest classifier, LRT, Kolmogorov-Smirnov test, Wilcox-rank-sum test, etc.

	Short name	Method	Software version	Input	Available from	Reference
	BPSC	BPSC	BPSC 0.99.0/1	CPM	GitHub	[11]
	D3E	D3E	D3E 1.0	raw counts	GitHub	[12]
	DESeq2	DESeq2	DESeq2 1.14.1	raw counts	Bioconductor	[13]
	DESeq2betapFALSE	DESeq2 without beta prior	DESeq2 1.14.1	raw counts	Bioconductor	[13]
	DESeq2census	DESeq2	DESeq2 1.14.1	Census counts	Bioconductor	[13]
	DESeq2nofilt	DESeq2 without the built-in independent filtering	DESeq2 1.14.1	raw counts	Bioconductor	[13]
	DEsingle	DEsingle	DEsingle 0.1.0	raw counts	GitHub	[14]
	edgeRLRT	edgeR/LRT	edgeR 3.19.1	raw counts	Bioconductor	[15–17]
	edgeRLRTcensus	edgeR/LRT	edgeR 3.19.1	Census counts	Bioconductor	[15–17]
	edgeRLRTdeconv	edgeR/LRT with deconvolution normalization	edgeR 3.19.1, scran 1.2.0	raw counts	Bioconductor	[15, 17, 18]
	edgeRLRTrobust	edgeR/LRT with robust dispersion estimation	edgeR 3.19.1	raw counts	Bioconductor	[15–17, 19]
	edgeRQLF	edgeR/QLF	edgeR 3.19.1	raw counts	Bioconductor	[15, 16, 20]
	edgeRQLFDetRate	edgeR/QLF with cellular detection rate as covariate	edgeR 3.19.1	raw counts	Bioconductor	[15, 16, 20]
	limmatrend	limma-trend	limma 3.30.13	log_2(CPM)	Bioconductor	[21, 22]
	MASTcpm	MAST	MAST 1.0.5	log_2(CPM+1)	Bioconductor	[23]
	MASTcpmDetRate	MAST with cellular detection rate as covariate	MAST 1.0.5	log_2(CPM+1)	Bioconductor	[23]
	MASTtpm	MAST	MAST 1.0.5	log_2(TPM+1)	Bioconductor	[23]
	MASTtpmDetRate	MAST with cellular detection rate as covariate	MAST 1.0.5	log_2(TPM+1)	Bioconductor	[23]
	metagenomeSeq	metagenomeSeq	metagenomeSeq 1.16.0	raw counts	Bioconductor	[24]
	monocle	monocle (tobit)	monocle 2.2.0	TPM	Bioconductor	[25]
	monoclecensus	monocle (Negative Binomial)	monocle 2.2.0	Census counts	Bioconductor	[25, 26]
	monoclecount	monocle (Negative Binomial)	monocle 2.2.0	raw counts	Bioconductor	[25]
	NODES	NODES	NODES 0.0.0.9010	raw counts	Author-provided link	[27]
	ROTScpm	ROTS	ROTS 1.2.0	CPM	Bioconductor	[28, 29]
	ROTStpm	ROTS	ROTS 1.2.0	TPM	Bioconductor	[28, 29]
	ROTSvoom	ROTS	ROTS 1.2.0	voom-transformed raw counts	Bioconductor	[28, 29]
	SAMseq	SAMseq	samr 2.0	raw counts	CRAN	[30]
	scDD	scDD	scDD 1.0.0	raw counts	Bioconductor	[31]
	SCDE	SCDE	scde 2.2.0	raw counts	Bioconductor	[32]
	SeuratBimod	Seurat (bimod test)	Seurat 1.4.0.7	raw counts	GitHub	[33, 34]
	SeuratBimodnofilt	Seurat (bimod test) without the internal filtering	Seurat 1.4.0.7	raw counts	GitHub	[33, 34]
	SeuratBimodIsExpr2	Seurat (bimod test) with internal expression threshold set to 2	Seurat 1.4.0.7	raw counts	GitHub	[33, 34]
	SeuratTobit	Seurat (tobit test)	Seurat 1.4.0.7	TPM	GitHub	[25, 33]
	ttest	t-test	stats (R v 3.3)	TMM-normalized TPM	CRAN	[16, 35]
	voomlimma	voom-limma	limma 3.30.13	raw counts	Bioconductor	[21, 22]
	Wilcoxon	Wilcoxon test	stats (R v 3.3)	TMM-normalized TPM	CRAN	[16, 36]

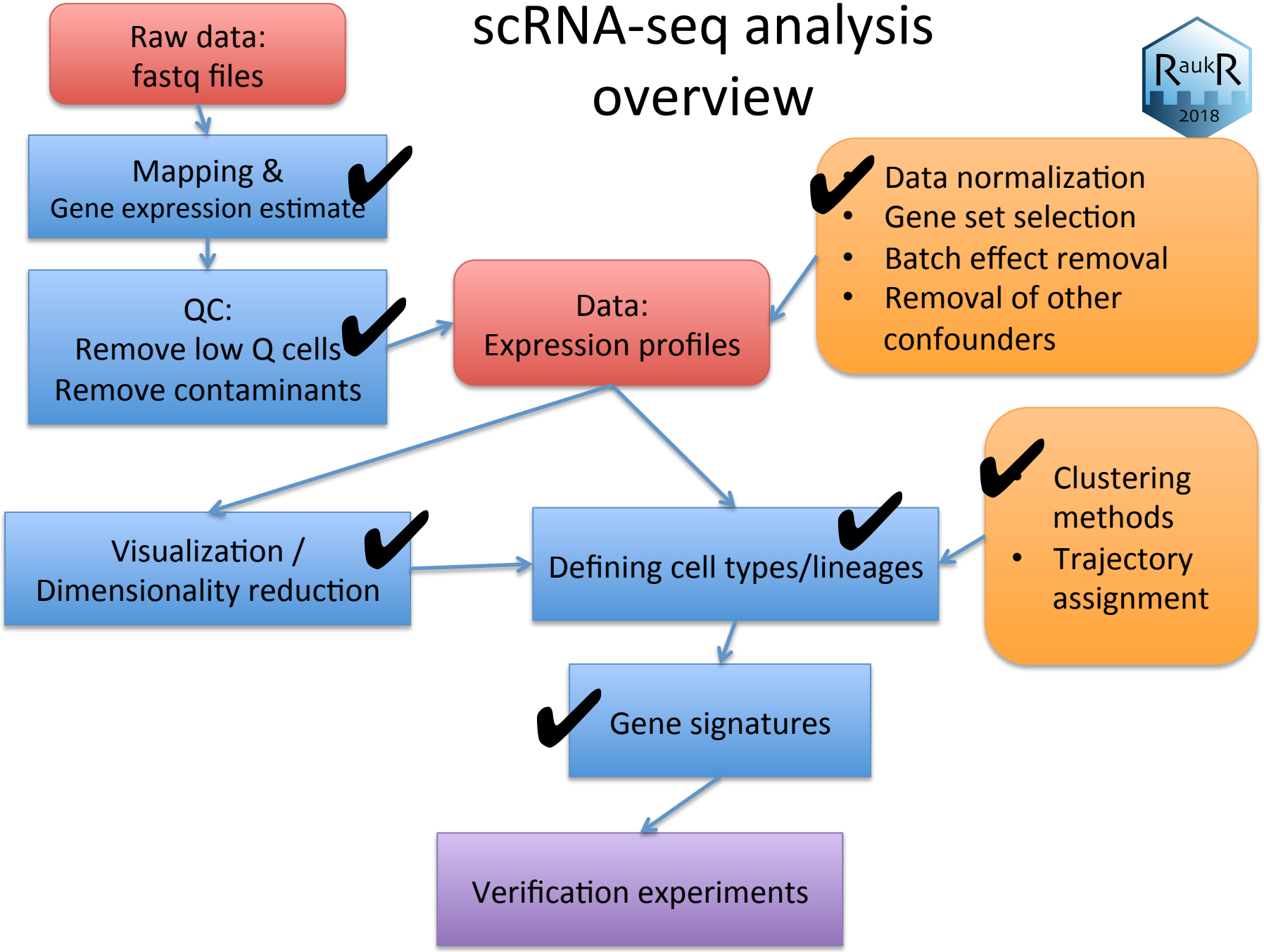
(Soneson & Robinson, *Nature Methods* 2018)



- EdgeR & Limma seem to work well
- MAST ranked high among scRNA-seq methods

(Soneson & Robinson, *Nature Methods* 2018)

scRNA-seq analysis overview



What more can you do?

- Allelic expression
- X chromosome inactivation
- Variant calling
- Copy-number variation
- Alternative splicing

Combination methods

- CITE-seq & REAP-seq − barcoded antibodies to measure protein markers + RNA-seq
- Proximity ligation assay (PLA) + RNA-seq − internal proteins
- T-cell receptor VD(J) profiling + RNA-seq
- Genome + RNA (WGS, ATAC-seq, Bis-seq)
- CRISPR-mediated perturbations + RNA-seq
- Lineage tracing with CRISPR barcodes
- ncRNA methods − single cell small RNA libraries and total RNA libraries
- Combination with spatial methos (*in situs*, smRNA fish, Spatial transcriptomics)

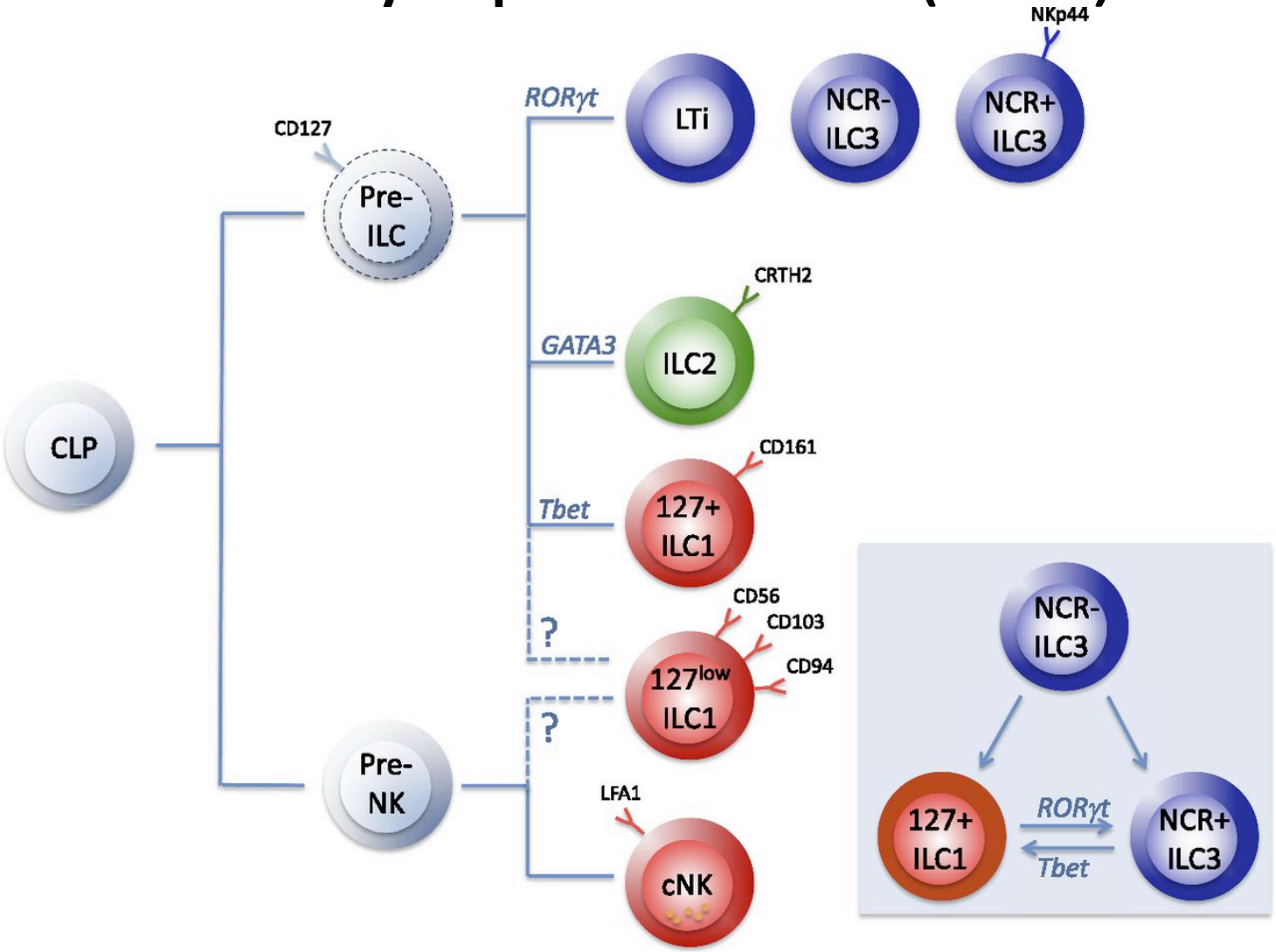
Resources

- Tutorials and lectures from our NBIS course: https://nbisweden.github.io/workshop-scRNAseq/
- Other good course at: https://hemberg-lab.github.io/scRNA.seq.course/
- Many of the packages have very thorough tutorials on their websites
- Repo with scRNA-seq tools: https://github.com/seandavi/awesome-single-cell
- Single cell experiment objects for many datasets: https://hemberg-lab.github.io/scRNA.seq.datasets/
- Conquer datasets - salmon pipeline for many different datasets: http://imlspenticton.uzh.ch:3838/conquer/

Presentation of tutorials

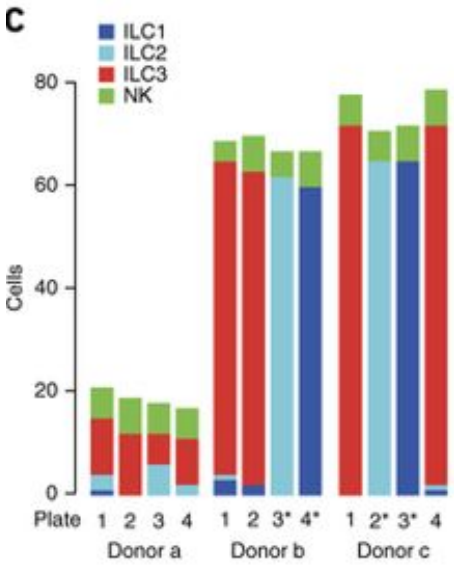
- **Seurat -** Single-cell RNA-seq data workflow with Seurat including filtering, QC and clustering.
- **Scater -** Quality Control with Scater package and batch correction with ComBat
- **SC3 -** Single-cell RNA-seq concensus clustering and with the SC3 package.

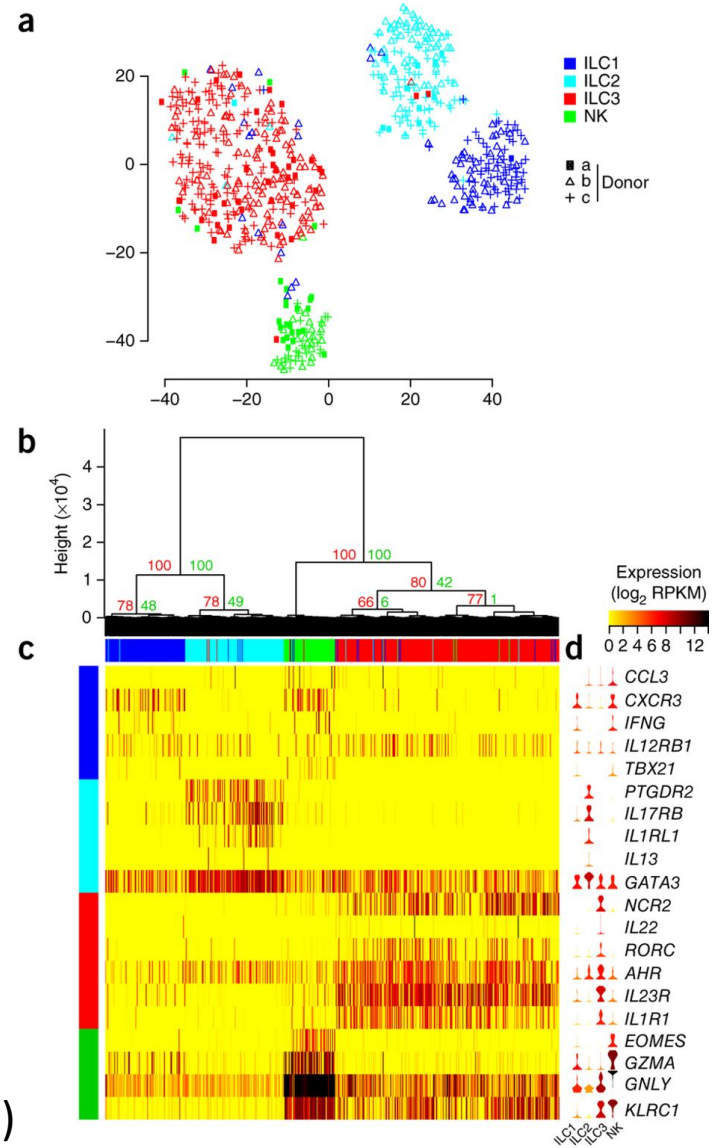
Innate lymphoid cells (ILCs)



(Hazenberg & Spitz, *Blood* 2014)

Innate lymphoid cells (ILCs)





(Björklund *et al. Nature Immunology* 2016)