Single cell RNA sequencing

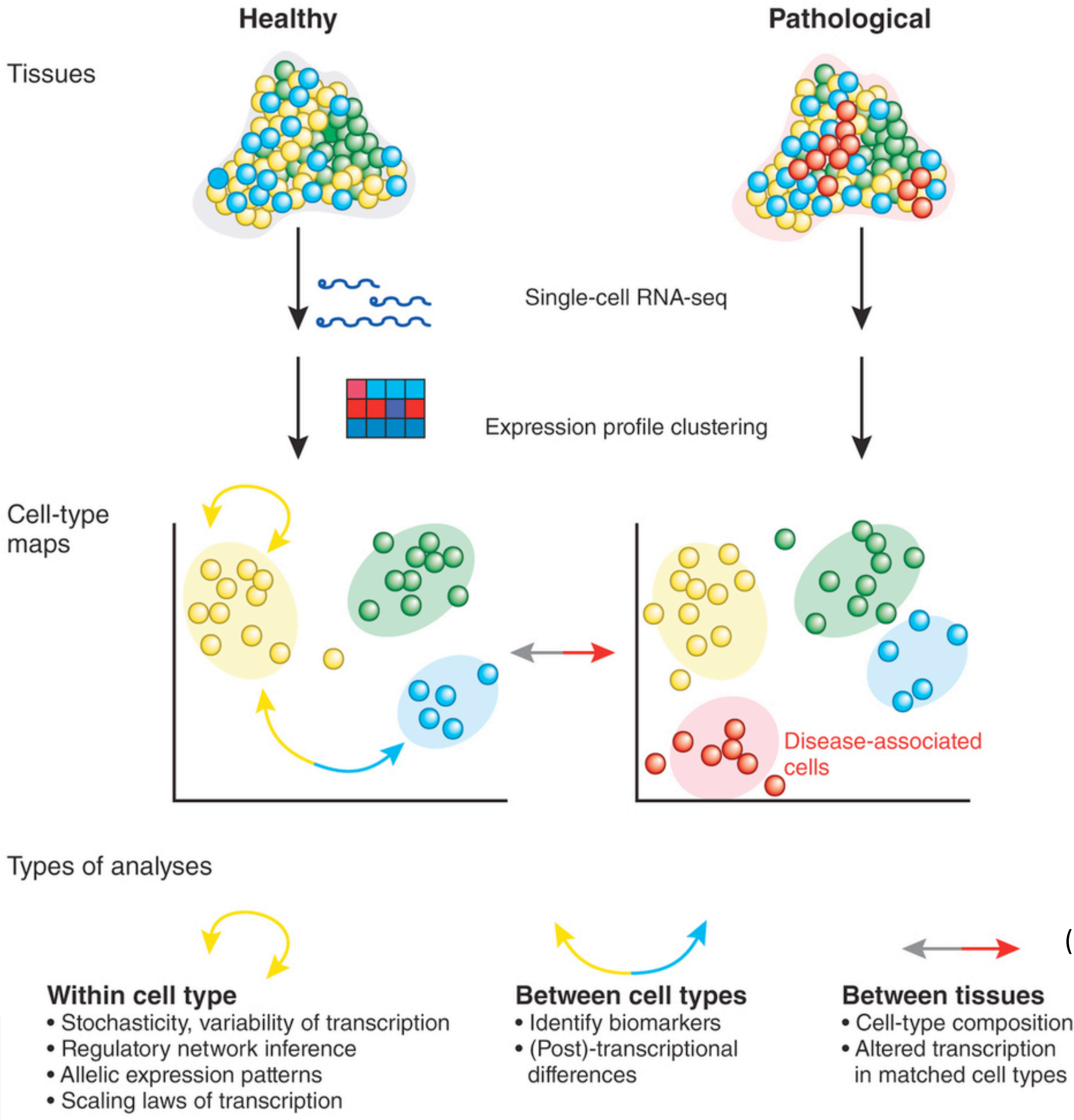
Åsa Björklund

asa.bjorklund@scilifelab.se

SciLifeLab Karolinska Institutet KTH VETENSKAP OCH KONST Stockholms universitet UPPSALA UNIVERSITET

Outline

- Why single cell transcriptomics?
- Experimental setup
- Computational analysis
- Examples of scRNA-seq experiments



Within cell type
- Stochasticity, variability of transcription
- Regulatory network inference
- Allelic expression patterns
- Scaling laws of transcription

Between cell types
- Identify biomarkers
- (Post)-transcriptional differences

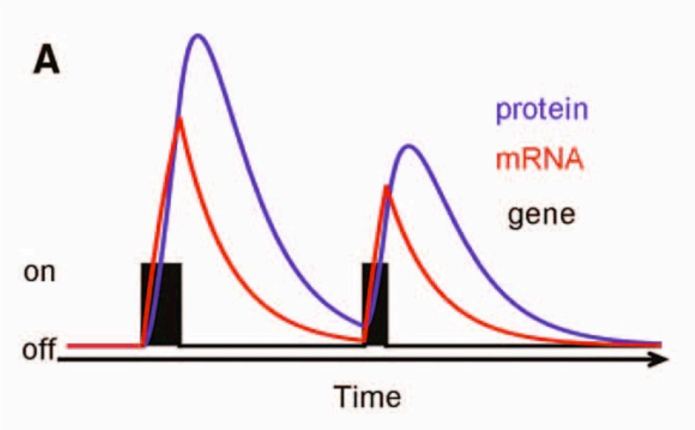
Between tissues
- Cell-type compositions
- Altered transcription in matched cell types

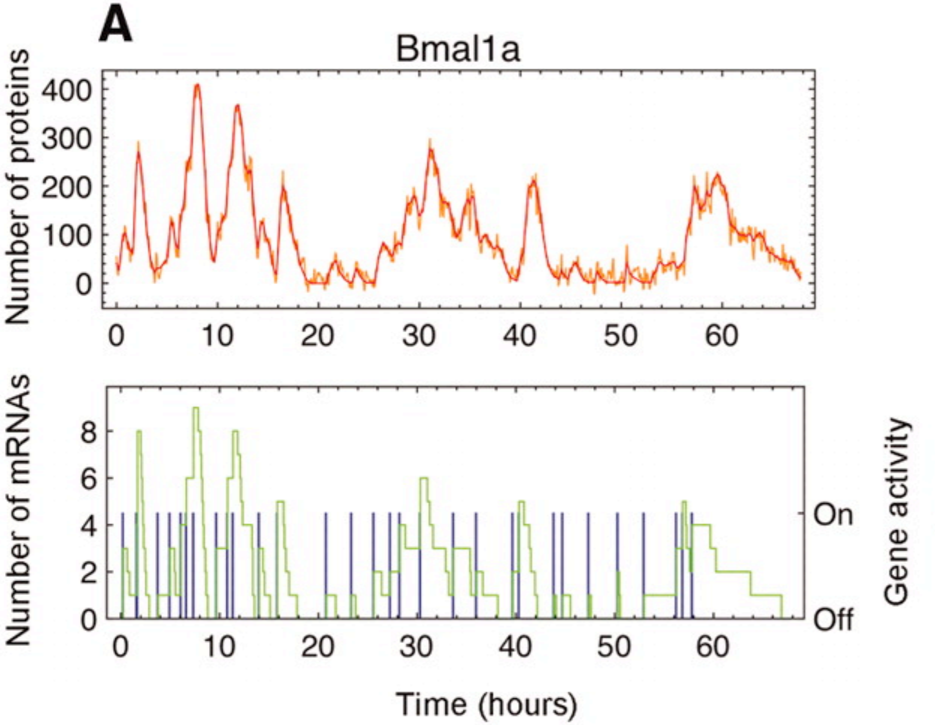
(Sandberg, *Nature Methods* 2014)

Why single-cell transcriptomics?

- Understanding heterogeneous tissues
- Identification and analysis of rare cell types
- Changes in cellular composition
- Dissection of temporal changes
- Example of applications:
  - Differentiation trajectories
  - Cancer heterogeneity
  - Neural cell classification
  - Embryonic development
  - Drug treatment response

Transcriptional bursting

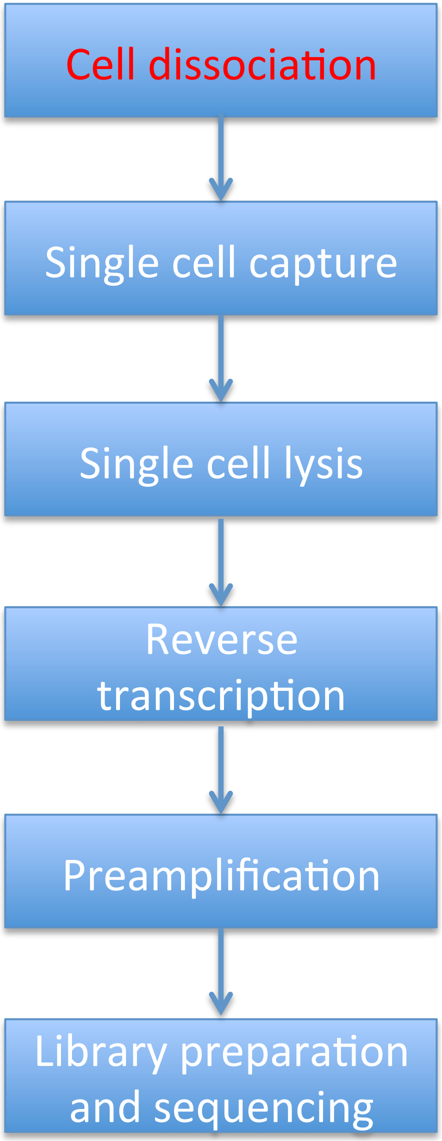




- Burst frequency and size is correlated with mRNA abundance
- Many TFs have low mean expression (and low burst frequency) and will only be detected in a fraction of the cells

(Suter et al. *Science* 2011)

Experimental setup


Cell dissociation

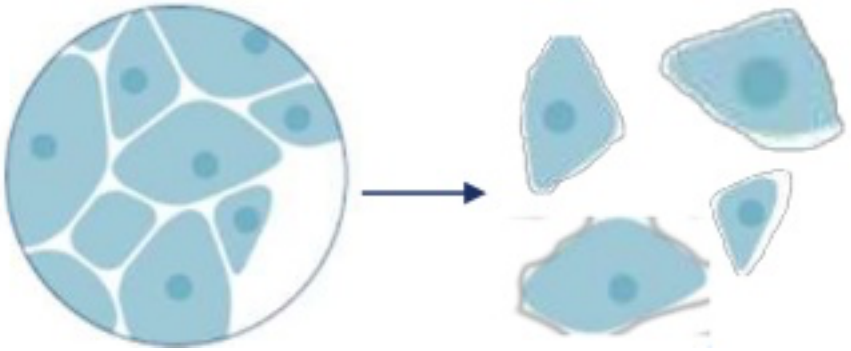
Single cell capture

Single cell lysis

Reverse transcription

Preamplification

Library preparation and sequencing



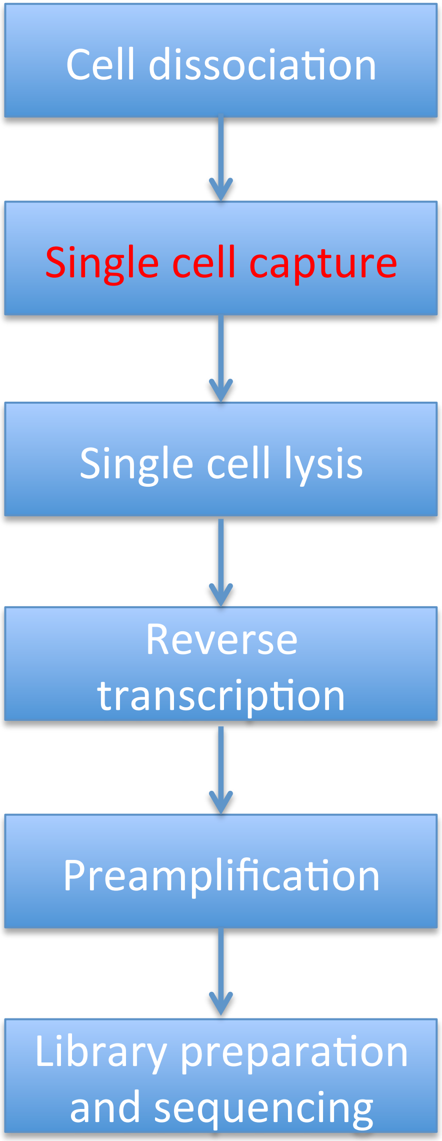
Tissues can be dissolved with mechanical methods, detergents or enzymatic digestion.

It is critical to have healthy whole cells with no RNA leakage.

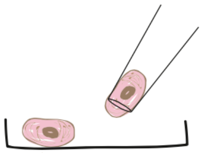
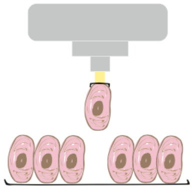
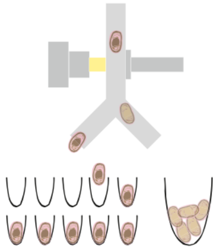
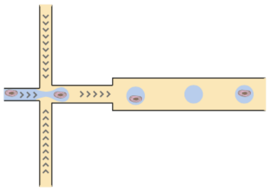
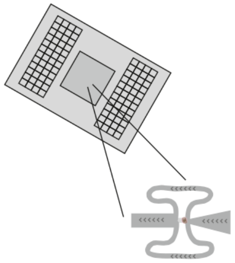
Should minimize time from dissociation to cell capture to reduce effect on transcriptional state.

Depending on your tissue type you may need to select different protocols.

(Kolodziejczyk et al. 2015)



Experimental setup

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

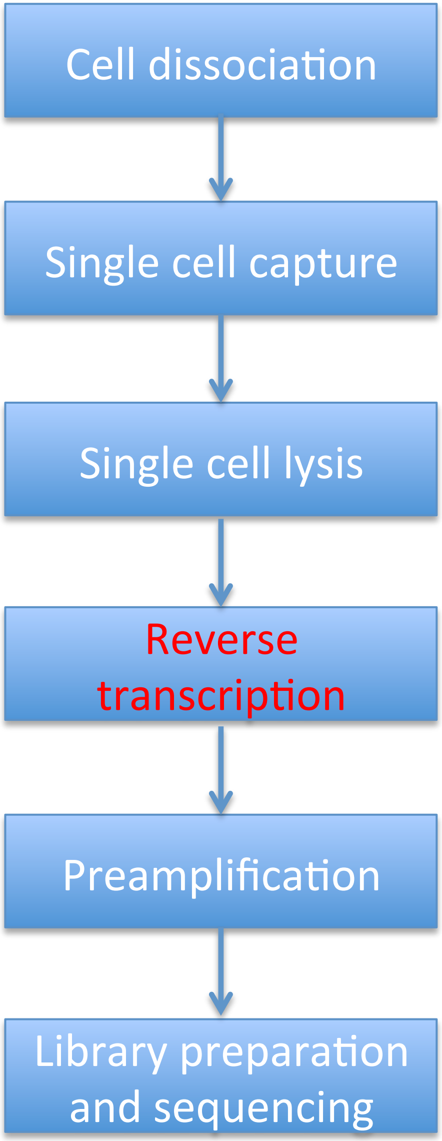
Tissues that are hard to dissociate:

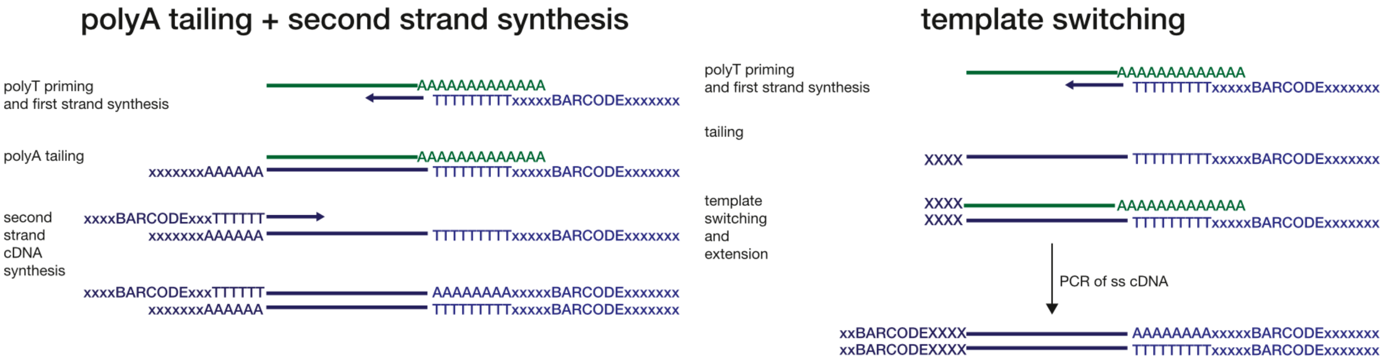
Laser capture microscopy (LCM)

Nuclei sorting

(Kolodziejczyk et al. 2015)

Experimental setup




polyA tailing + second strand synthesis
polyT priming and first strand synthesis
AAAAAAAAAAAAA
TTTTTTTTTxxxxxBARCODExxxxxxx
polyA tailing
xxxxxxxAAAAAA
AAAAAAAAAAAAA
TTTTTTTTTxxxxxBARCODExxxxxxx
second strand cDNA synthesis
xxxxBARCODExxxTTTTTT
xxxxxxxAAAAAA
TTTTTTTTTxxxxxBARCODExxxxxxx
xxxxBARCODExxxTTTTTT
AAAAAAAAxxxxxBARCODExxxxxxx
xxxxxxxAAAAAA
TTTTTTTTTxxxxxBARCODExxxxxxx
template switching
polyT priming and first strand synthesis
AAAAAAAAAAAAA
TTTTTTTTTxxxxxBARCODExxxxxxx
tailing
XXXX
TTTTTTTTTxxxxxBARCODExxxxxxx
template switching and extension
XXXX
AAAAAAAAAAAAA
XXXX
TTTTTTTTTxxxxxBARCODExxxxxxx
PCR of ss cDNA
xxBARCODEXXXX
AAAAAAAAxxxxxBARCODExxxxxxx
xxBARCODEXXXX
TTTTTTTTTxxxxxBARCODExxxxxxx

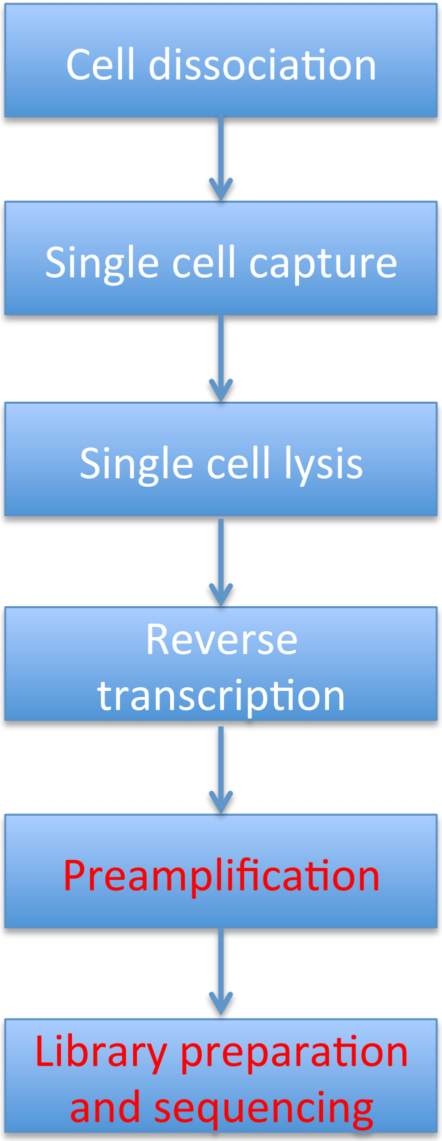
Tang protocol (Tang et al 2009)
CELseq/MARSseq (Hashimony et al. 2013, Jaitin et al. 2014)
QuartzSeq (Sasagawa et al. 2013)

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.

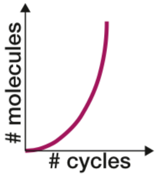
(Kolodziejczyk et al. 2015)

Experimental setup



PCR

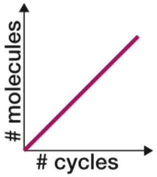
- exponential amplification
- PCR base specific biases



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)

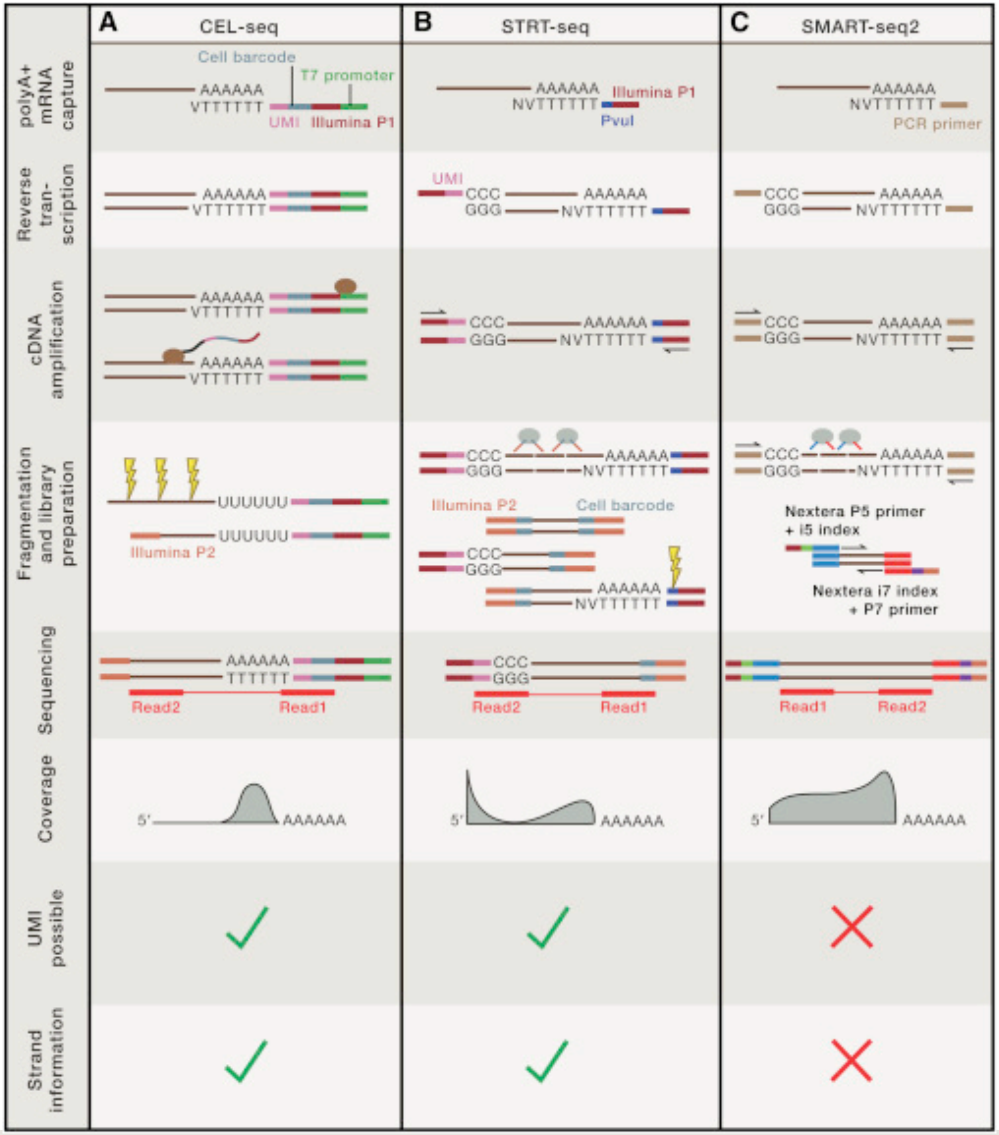
Illumina

AB SOLID

PacBio

(Kolodziejczyk et al. 2015)

Overview of the 3 common library preparation methods

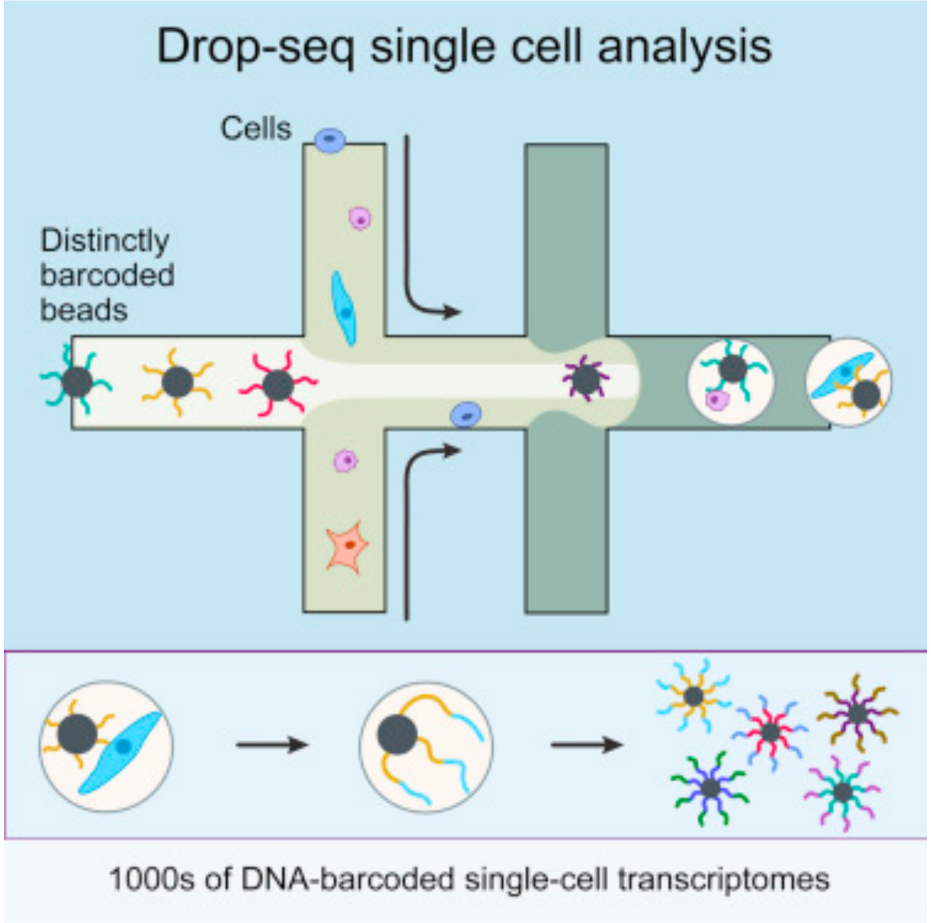


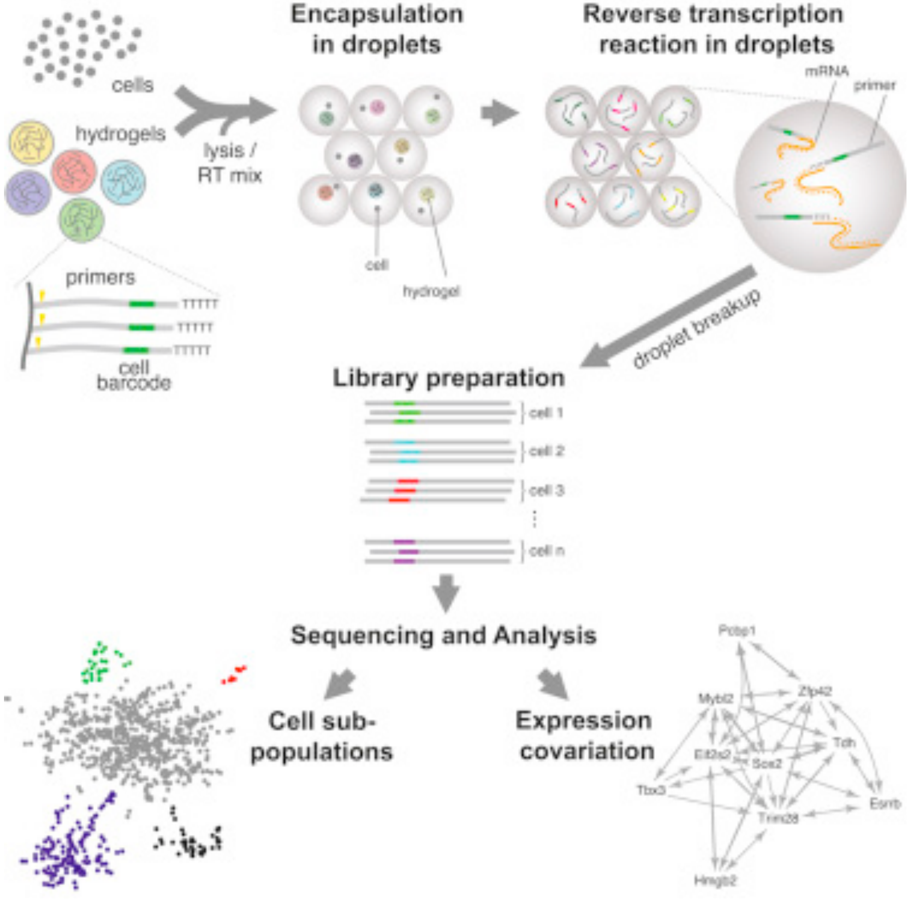
(Grun et al. *Cell* 2015)

Small volume approaches

- Volume seem to be a key component in these reactions
  - Smaller volumes give better detection and reproducibility
- Smaller volumes = cheaper reagent costs
- Methods for high throughput (1000nds of cells)
- Sequencing cost becomes the bottleneck instead – often shallow sequencing

Droplet / microfluidics approaches


Drop-seq single cell analysis
Cells
Distinctly barcoded beads
1000s of DNA-barcoded single-cell transcriptomes

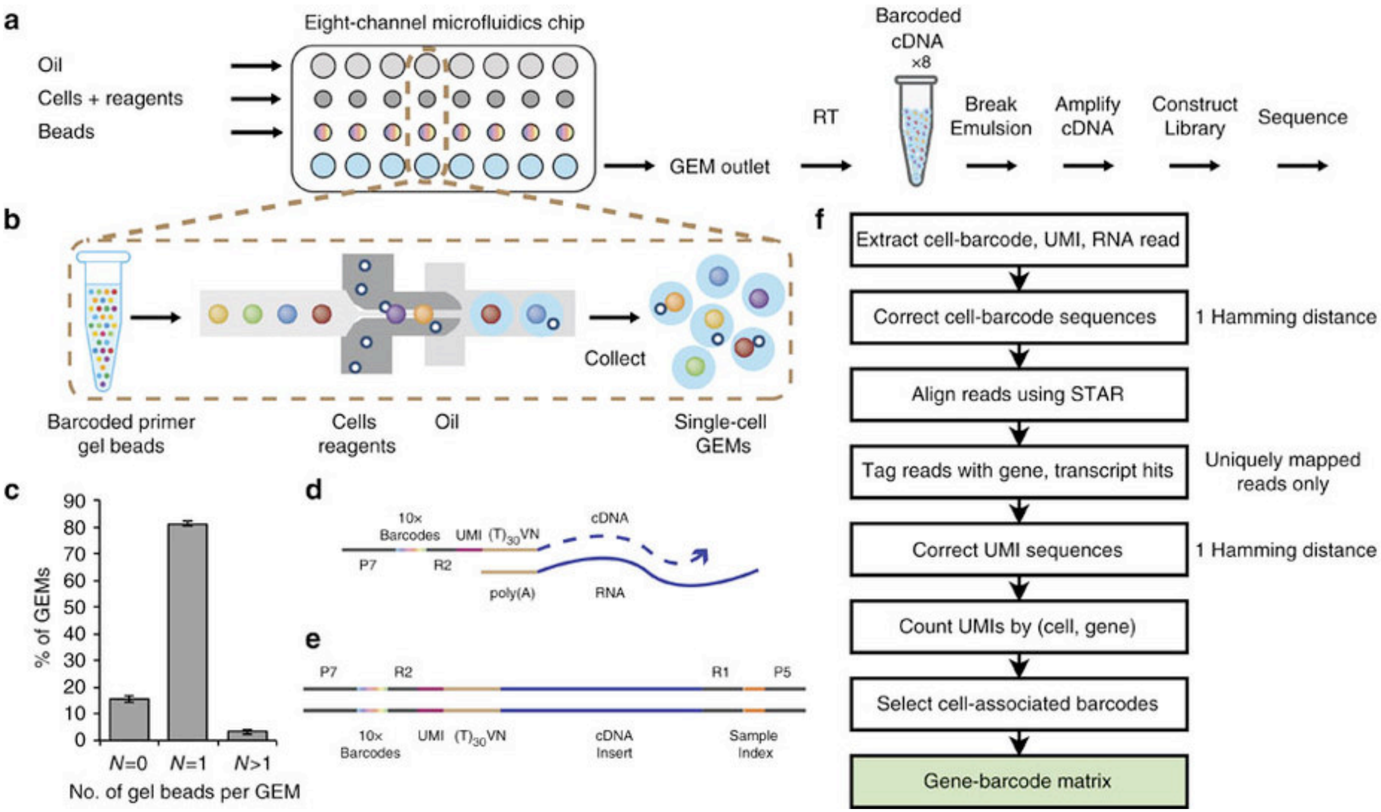


Macosko et al. *Cell* 2015
McCarrol, Regev etc. Broad/Harvard

Klein et al. *Cell* 2015
Kirschner, Weitz etc. Harvard

Chromium 10X Genomics

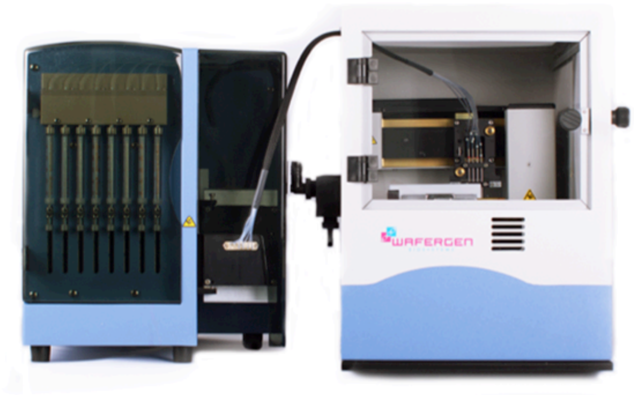
- Droplet based system for scRNAseq and genome sequencing
- 500-10,000 single cells
- CellRanger sofware for analysis of results



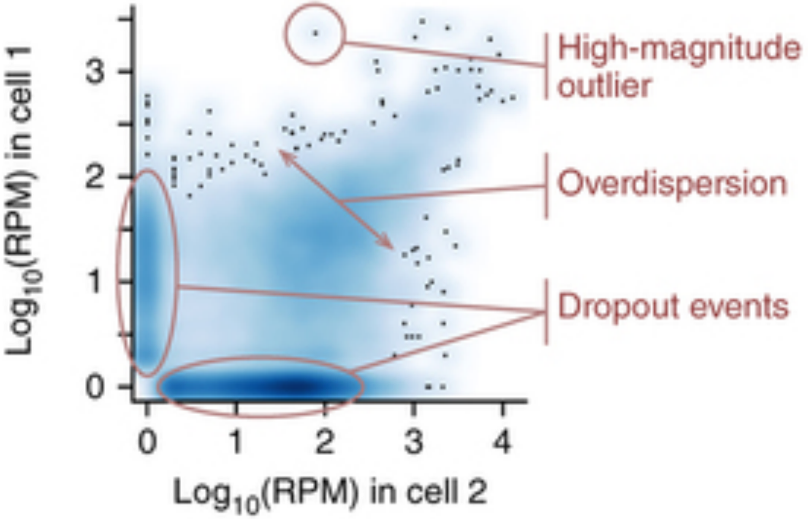
(GXY Zhang et al *Nat Communications* 2017)

Wafergen ICELL8 microwell system

- ICELL8 chip with 9600 wells
- Multi sample nano dispenser
- Can use FACS sorting
- Imaging station
- Software to select cells -> minimize number of doublets
- Possible to save images of cells with a few different colors
- Implemented with STRT at the ESCG platform



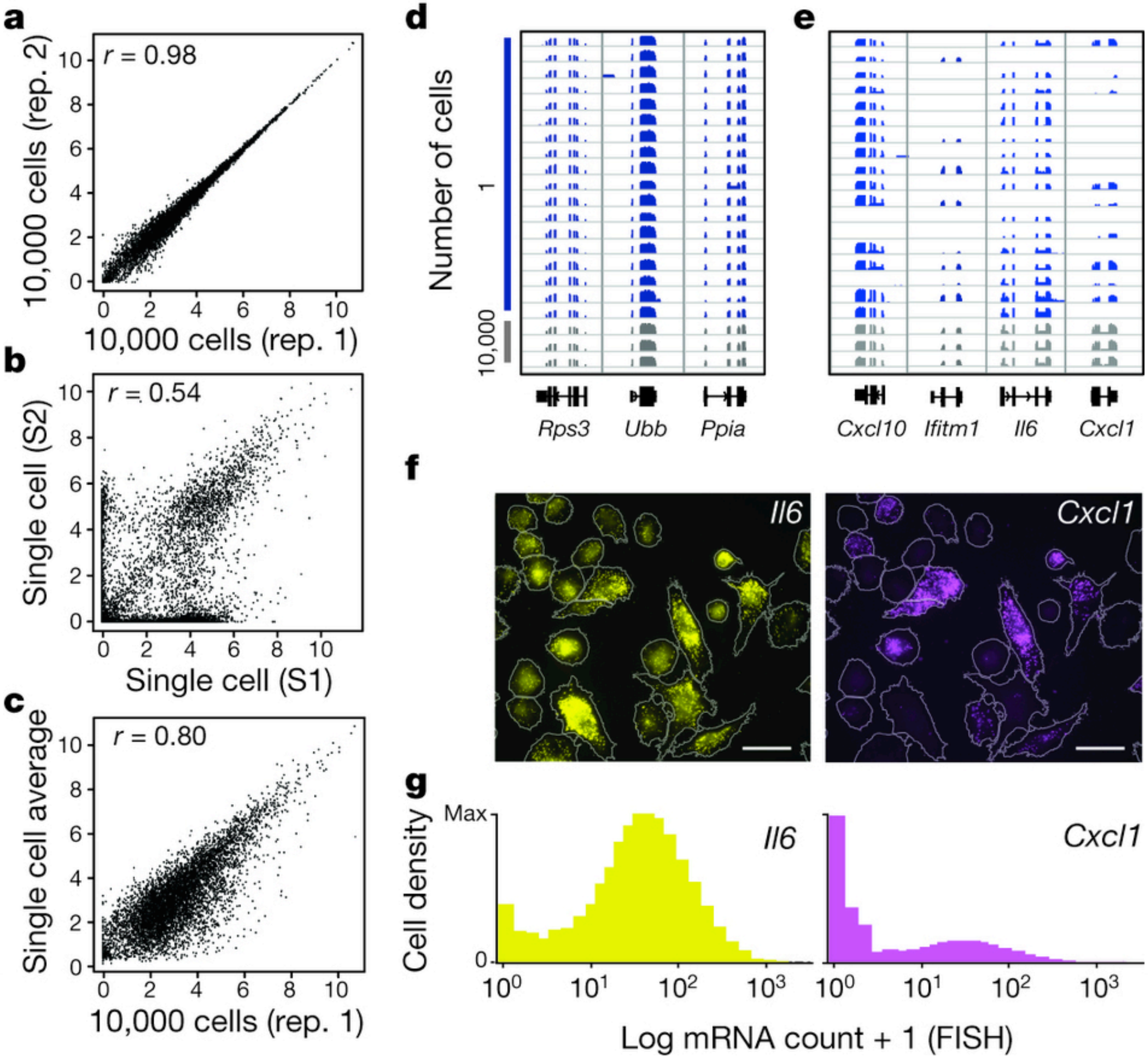
Problems compared to bulk RNA-seq



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

(Karchenko et al. *Nature Methods* 2014)

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

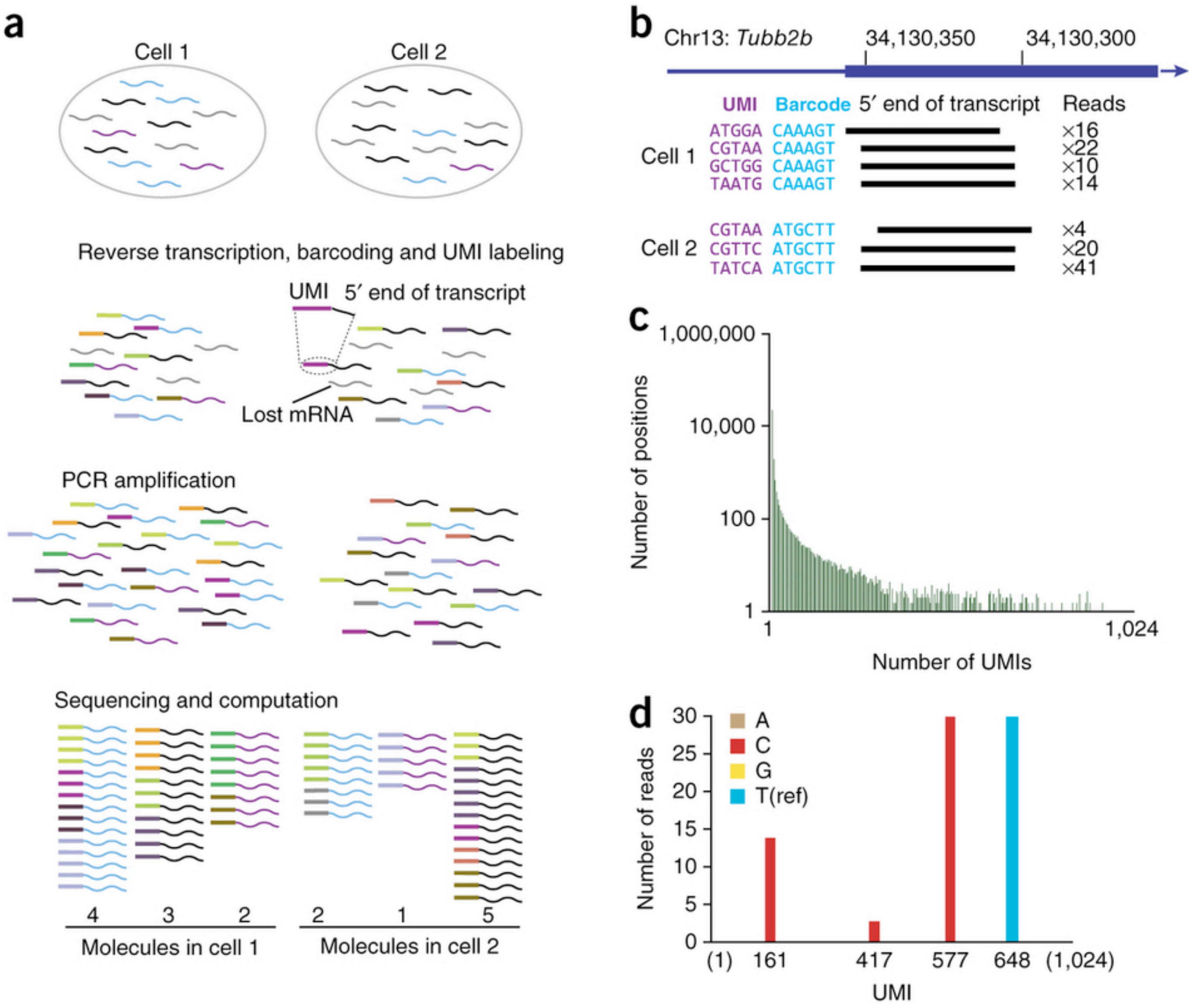


(Shalek et al. *Nature* 2013)

Unique molecular identifiers (UMIs) and cellular barcodes

- Cellular barcodes
  - Introduced at RT step with one unique sequence per cell
  - Enables pooling of many libraries into one tube for subsequent steps
- UMIs
  - Introduce random sequences at the beginning of each sequence
  - Reduces effect of amplification bias by removing PCR duplicates
- Implemented with tag-based methods such as STRT and CEL-seq

Unique molecular identifiers (UMIs) and cellular barcodes



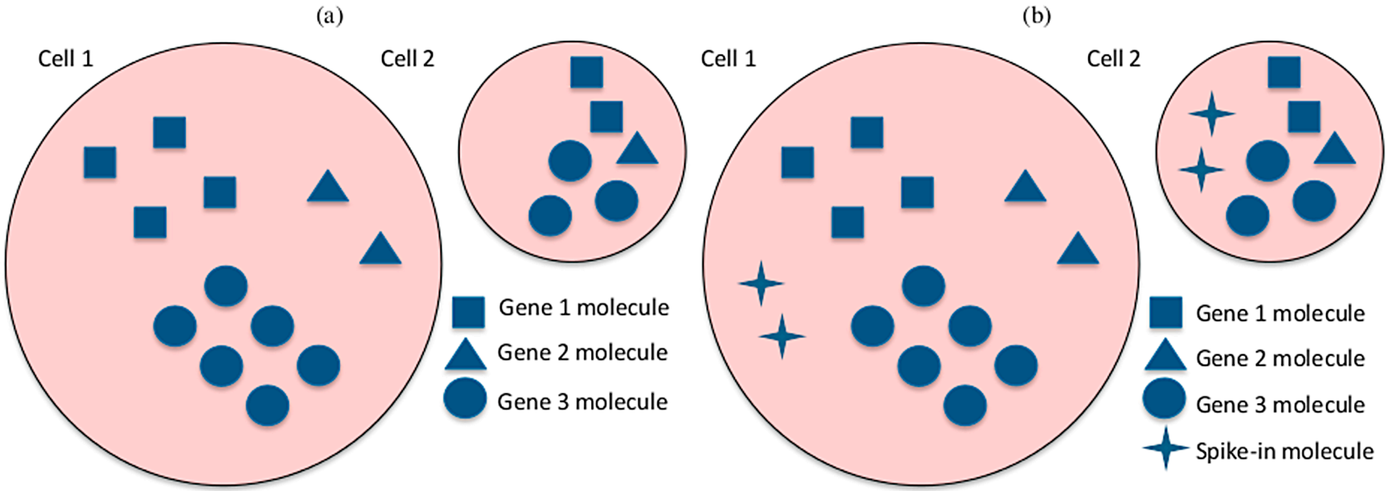
	UMI	Barcode	5′ end of transcript	Reads
Cell 1	ATGGA	CAAAGT		×16
	CGTAA	CAAAGT		×22
	GCTGG	CAAAGT		×10
	TAATG	CAAAGT		×14
Cell 2	CGTAA	ATGCTT		×4
	CGTTC	ATGCTT		×20
	TATCA	ATGCTT		×41

(Islam et al. *Nature Methods* 2014)

Spike-in RNAs

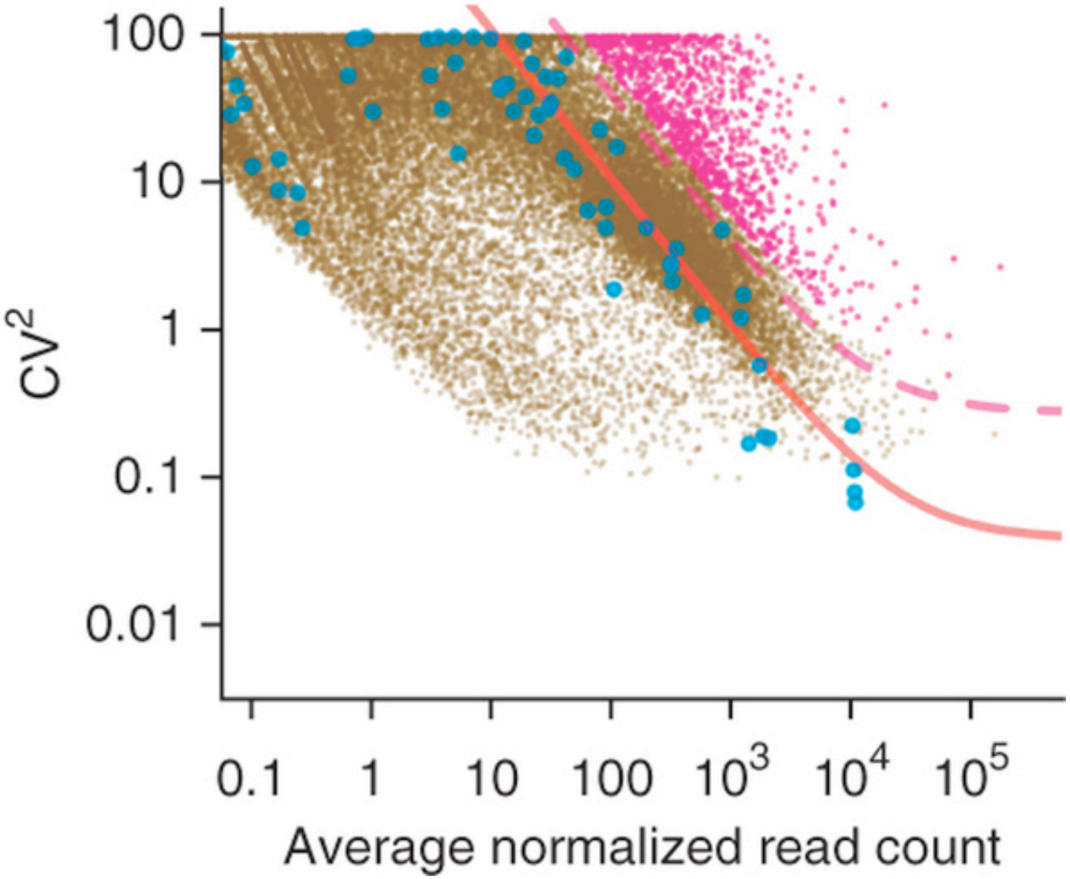
- Addition of external controls
- Used to model:
  - technical noise / drop-out rates
  - starting amount of RNA in the cell
- ERCC spike-in most widely used, consists of 48 or 96 mRNAs at 17 different concentrations.
- Important to add equal amounts to each cell, preferably in the lysis buffer.

Spike-in RNAs



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

Spike-in RNAs
Finding biologically variable genes



(Brennecke et al. *Nature Methods* 2013)

Replicates – how many cells do you have to sequence?

- Recommended to have at least 20-30 cells from each cell type
  - A sample with a minor cell type at 5% requires sequencing of 400 cells.
  - Preselecting cells may be necessary, but unbiased cell picking is preferred.
  - Depending on the sensitivity of your method you may need more/less cells
- To study gene expression only, sequencing depth does not have to be deep.
  - Multiplexing of hundreds of samples on one lane is common.
  - For tag-based methods sequencing is often more shallow.
- Possible to have a consultancy session with someone at NBIS for experimental design.

Which method should I use?

- Full length (SmartSeq2) vs tag-based (CELseq/STRT) methods:
  - Trade-off between throughput and sensitivity
  - Unique molecular identifiers (UMI) implementation with the tag-based methods
- Practical issues such as sorting of cells

National single cell genomics platform at Scilifelab

- Uppsala node – Microbial single cell genomics
  - http://www.scilifelab.se/facilities/single-cell/
  - MDA of whole genomes
  - qPCR of selected target genes
- Stockholm node – Eukaryotic single cell genomics (ESCG)
  - http://www.escg.se
  - Several technologies for scRNA-seq
  - MDA whole genome sequencing

RNA-seq services at ESCG

	Full-length	Quantitative		
	Smart-seq2	**STRT-C1**	**STRT-Wafergen**	**10xGenomics**
Format	Eppendorf Twin-tek	C1 microfluidics chip (Fluidigm)	Microwell chip	Chromium microfluidics chip
Cell number	384	3 x 96	9,600 (~2,500)	8 x 500-10,000
Input	FACS-sorted cells	Cell suspension	Cell suspension	Cell suspension
Transcript coverage	Full-length	5’	5’	3’
Advantage	• Flexible delivery • SNPs, mutations • Nuclei	• Imaging • Cell selection	• Unbiased • Cell selection • 8 samples parallel • Nuclei	• High throughput • 8 samples parallel • Sample pooling
Limitation	• No UMI (ERCC)	• Low throughput • Cell size bias	• Limiting dilution • Challenging FACS sort	• Potential clogs

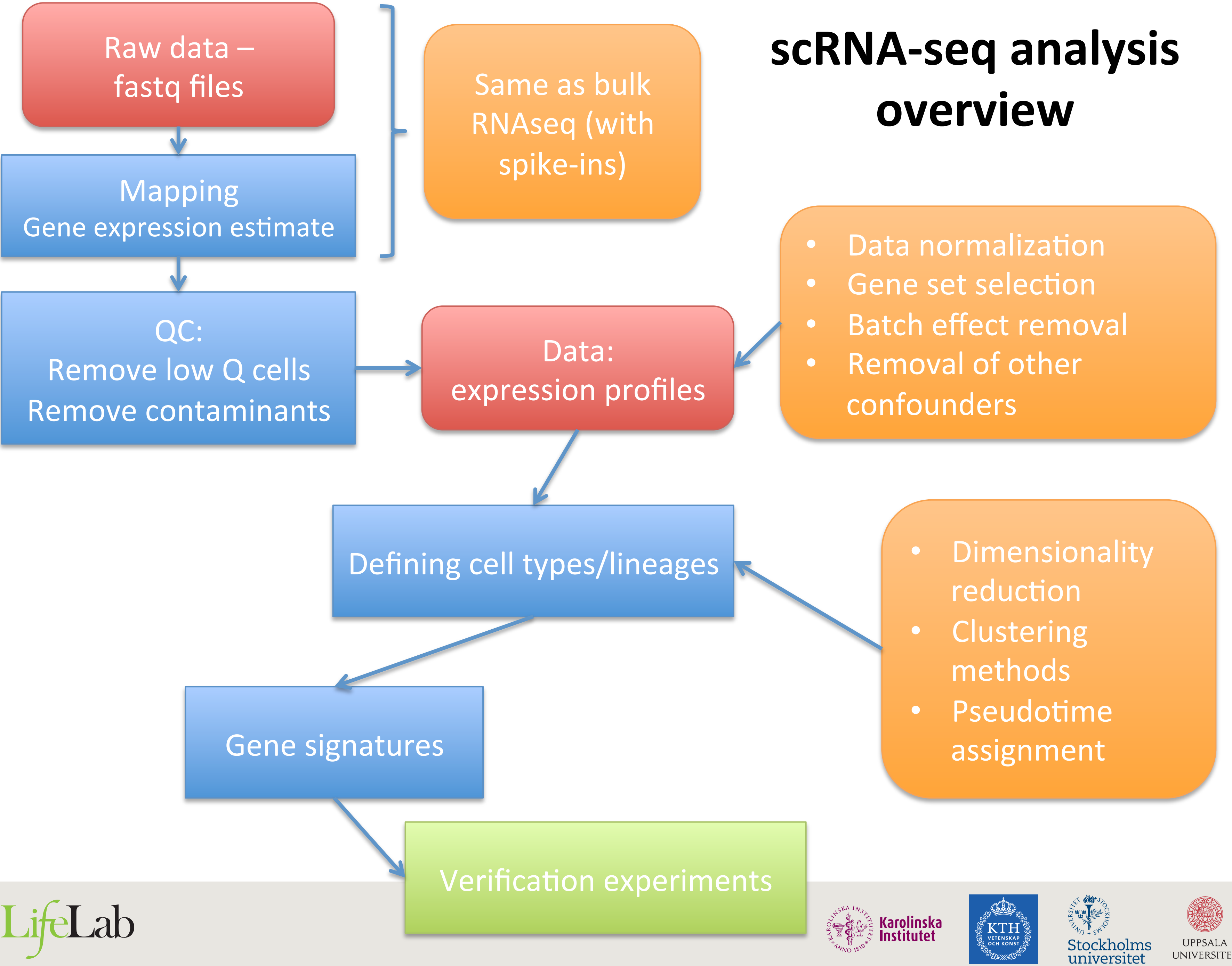
(From Karolina Wallenborg at ESCG)

User fees

Smart-seq2 384 well plate	STRT-C1 96 cell chip (50-96 cells)	STRT-Wafergen 9600 wells chip (~2,500 cells)	10XGenomics 1 sample (~3,000 cells)
• Validation • Smart-seq2 library • Sequencing • (50 bp, single-read	• Validation • STRT-C1 library • Sequencing (50bp single-read)	• Validation • STRT library (dual index) • Sequencing (50 bp single-read)	• Validation • Illumina library • Sequencing (paired-end, dual index)
~40,500 SEK	~22,500 SEK	~45,000-50,000 SEK	~43,000 SEK

Costs include: Reagents, consumables, instrument depreciation, instrument service, personnel. Overhead is not included.

scRNA-seq analysis overview



Raw data – fastq files

Mapping
Gene expression estimate

Same as bulk RNAseq (with spike-ins)

QC:
Remove low Q cells
Remove contaminants

Data:
expression profiles

- Data normalization
- Gene set selection
- Batch effect removal
- Removal of other confounders

Defining cell types/lineages

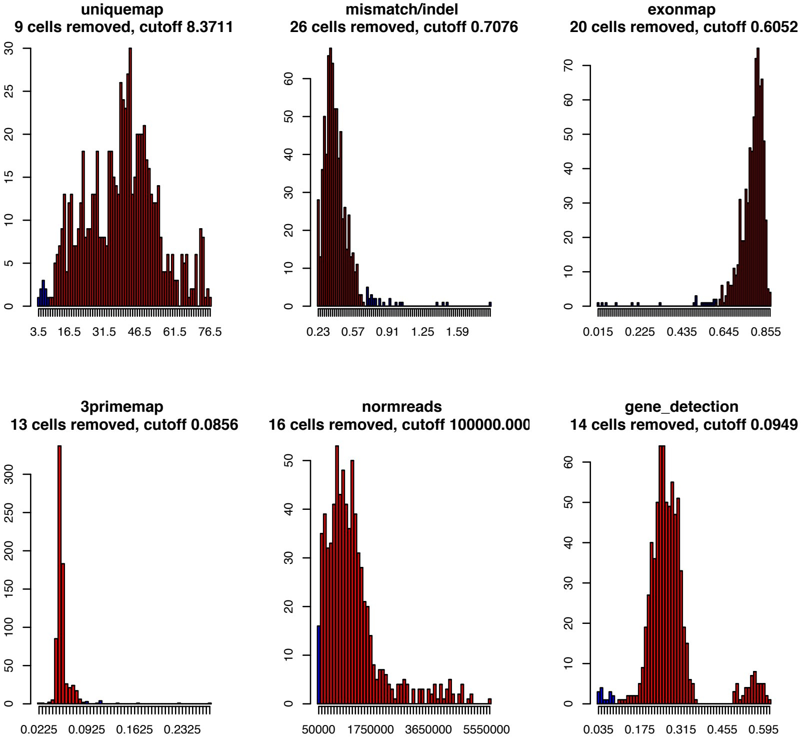
- Dimensionality reduction
- Clustering methods
- Pseudotime assignment

Gene signatures

Verification experiments

Quality Control (QC)

- QC is a crucial step in scRNA-seq - Any experiment will have a number of failed libraries!
- OBS! Smaller celltypes gives lower mapping rates and more primer dimers.
- Can look at:
  - Mapping statistics (**% uniquely mapping**)
  - Mismatch rate
  - Fraction of exon mapping reads
  - 3’ bias (degraded RNA)
  - mRNA-mapping reads
  - **Number of detected genes**
  - **Spike-in detection**
  - Mitochondrial read fraction
  - Pairwise correlation to other cells



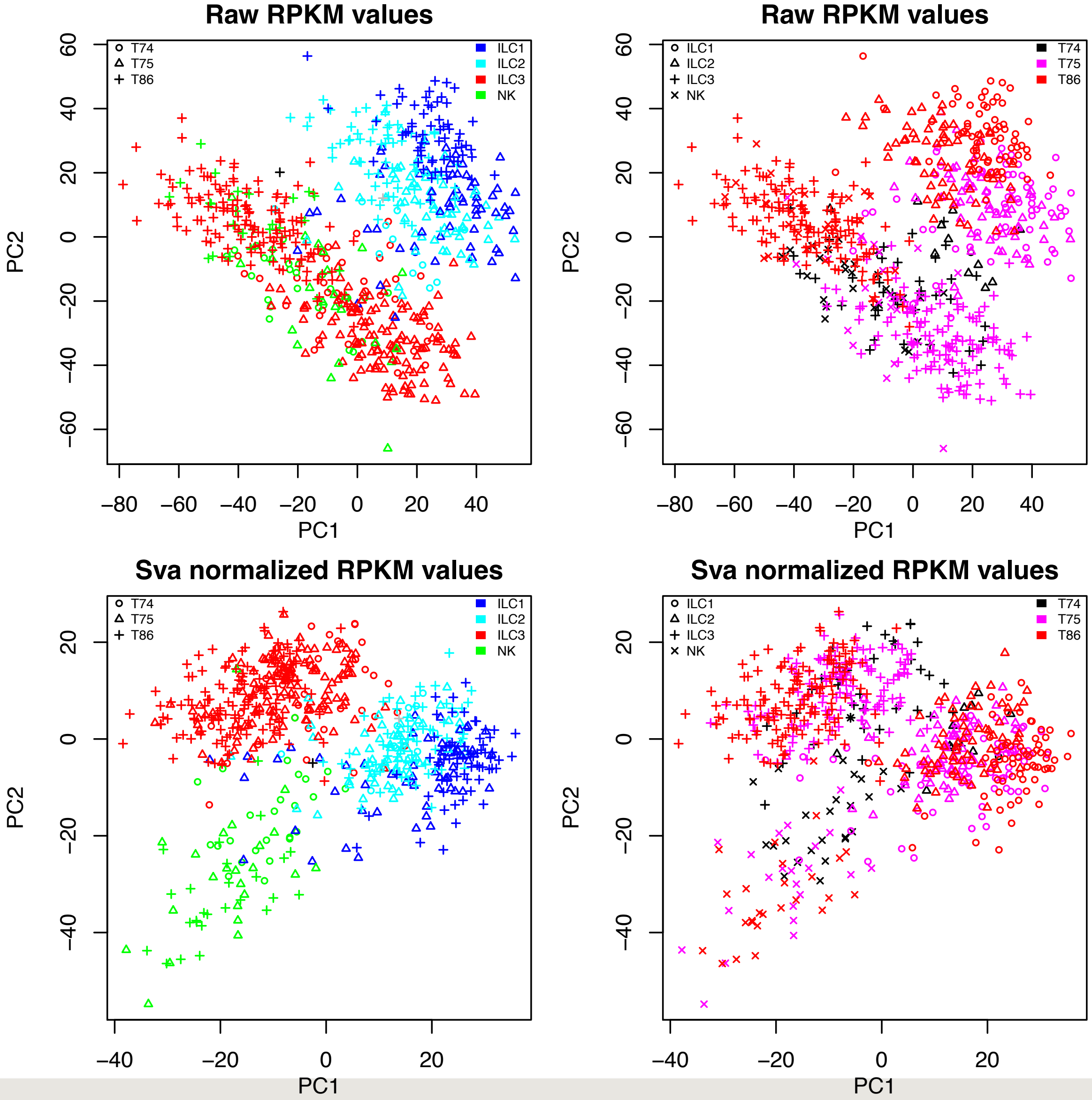
- Depending on cell type, around 500K exon mapping reads saturates the gene detection (deduced from subsampling in SS2 data).

Data bias

- Often need to do data transformation before clustering/PCA
  - Normalize by spike-in RNAs
  - Normalize by total counts
  - Length normalized RPKM/FPKM
  - Remove cell-cycle effects, size bias or similar (scLVM package, SCDE package)
  - RT efficiency / drop-out rate (SCDE package, scran package)
  - Technical noise (BASiCS package, GRM)
  - Batch effect removal (SVA ComBat function, SCDE package)

Batch normalization with SVA function ComBat

Color by celltype


Raw RPKM values

Raw RPKM values

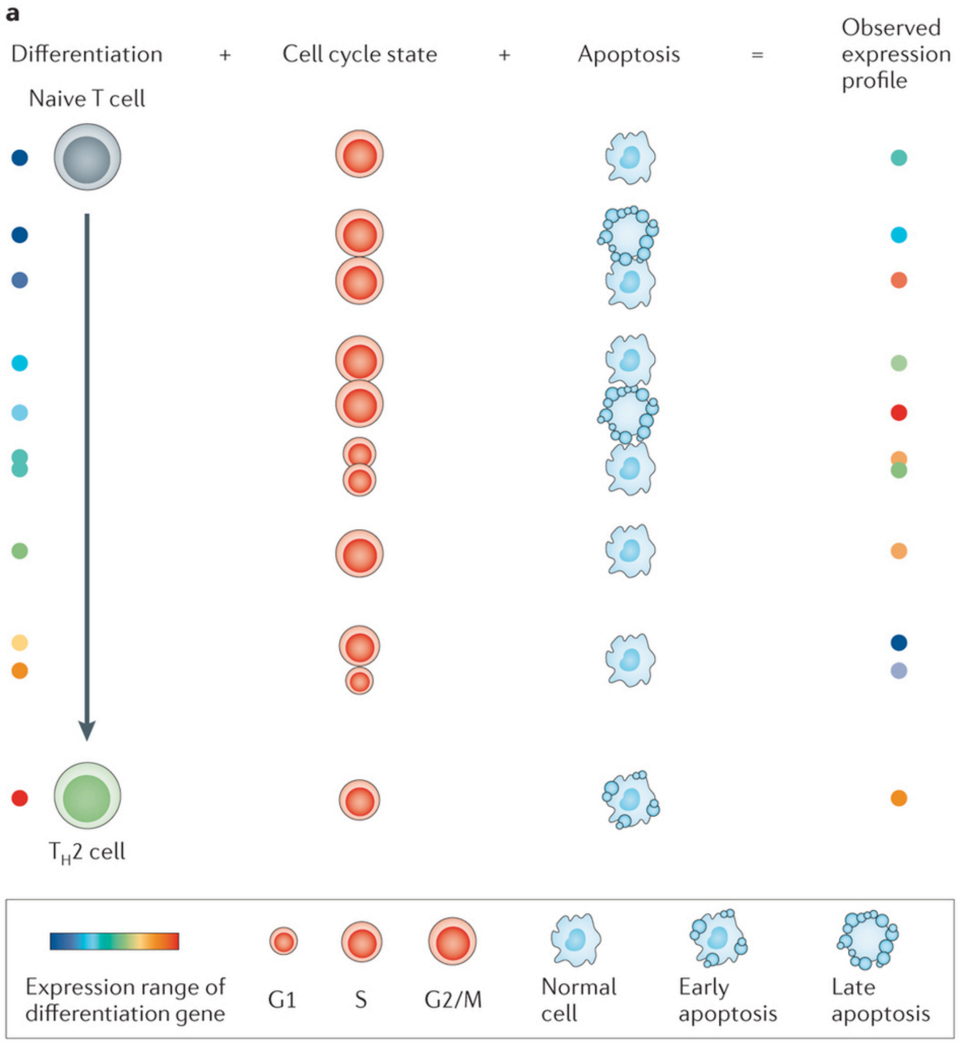
Sva normalized RPKM values

Sva normalized RPKM values

Color by donor

(Björklund et al. *Nature Immunology* 2016)

scLVM - Marioni lab
https://github.com/PMBio/scLVM)



(Stegle et al. *Nat Rev Genetics* 2015)

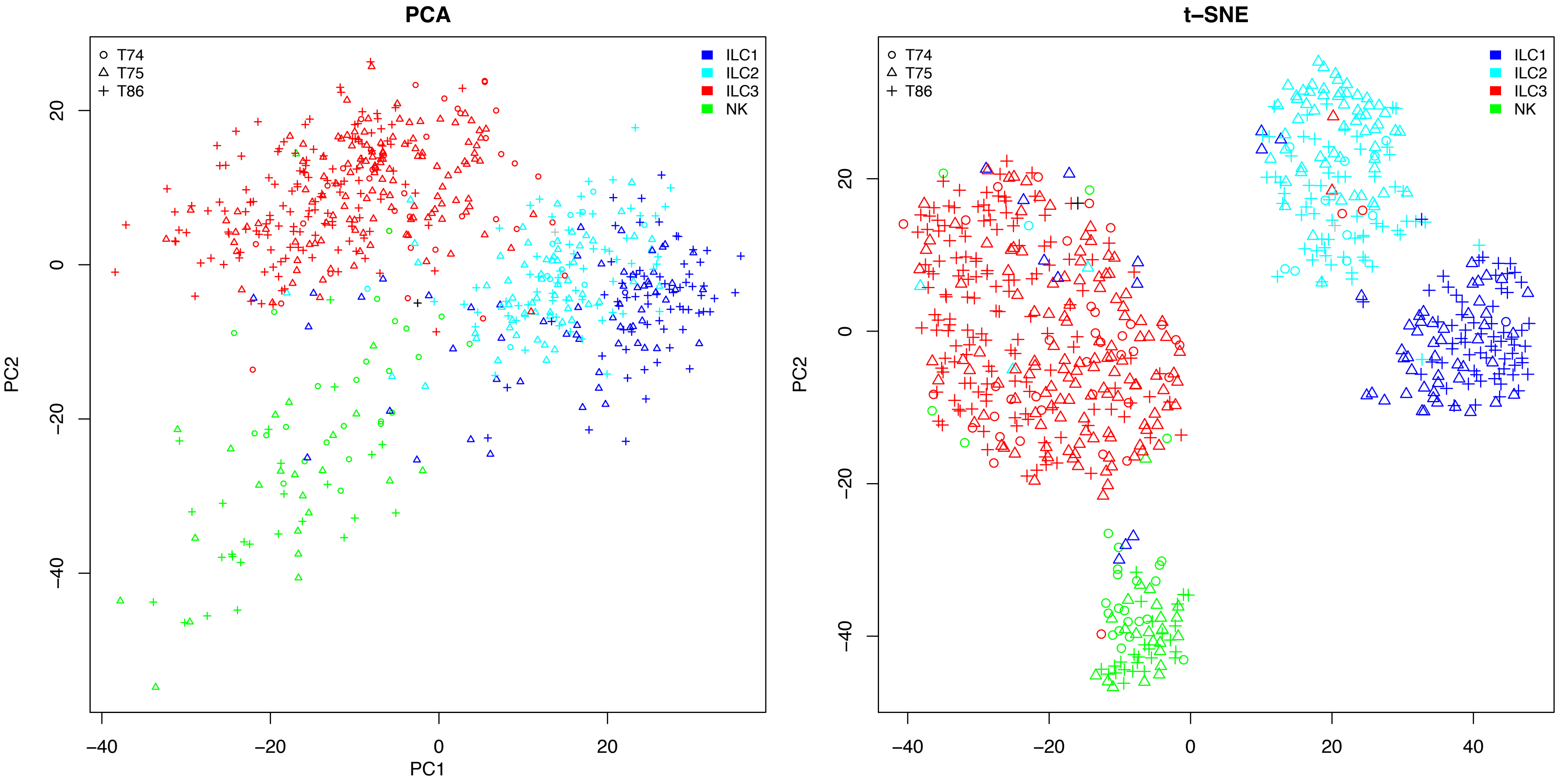
Feature selection – subset of genes

- In most cases, all genes are not used in PCA/ clustering.
- Gene set selection based on:
  - Biologically variable genes (Brenneke method based on spike-in data) or top variable genes if no spike-in data.
  - Genes expressed in X cells.
  - Filter out genes with correlation to few other genes
  - Prior knowledge / annotation
  - DE genes from bulk experiments
  - Top PCA loadings

Identifying celltypes – Dimensionality reduction

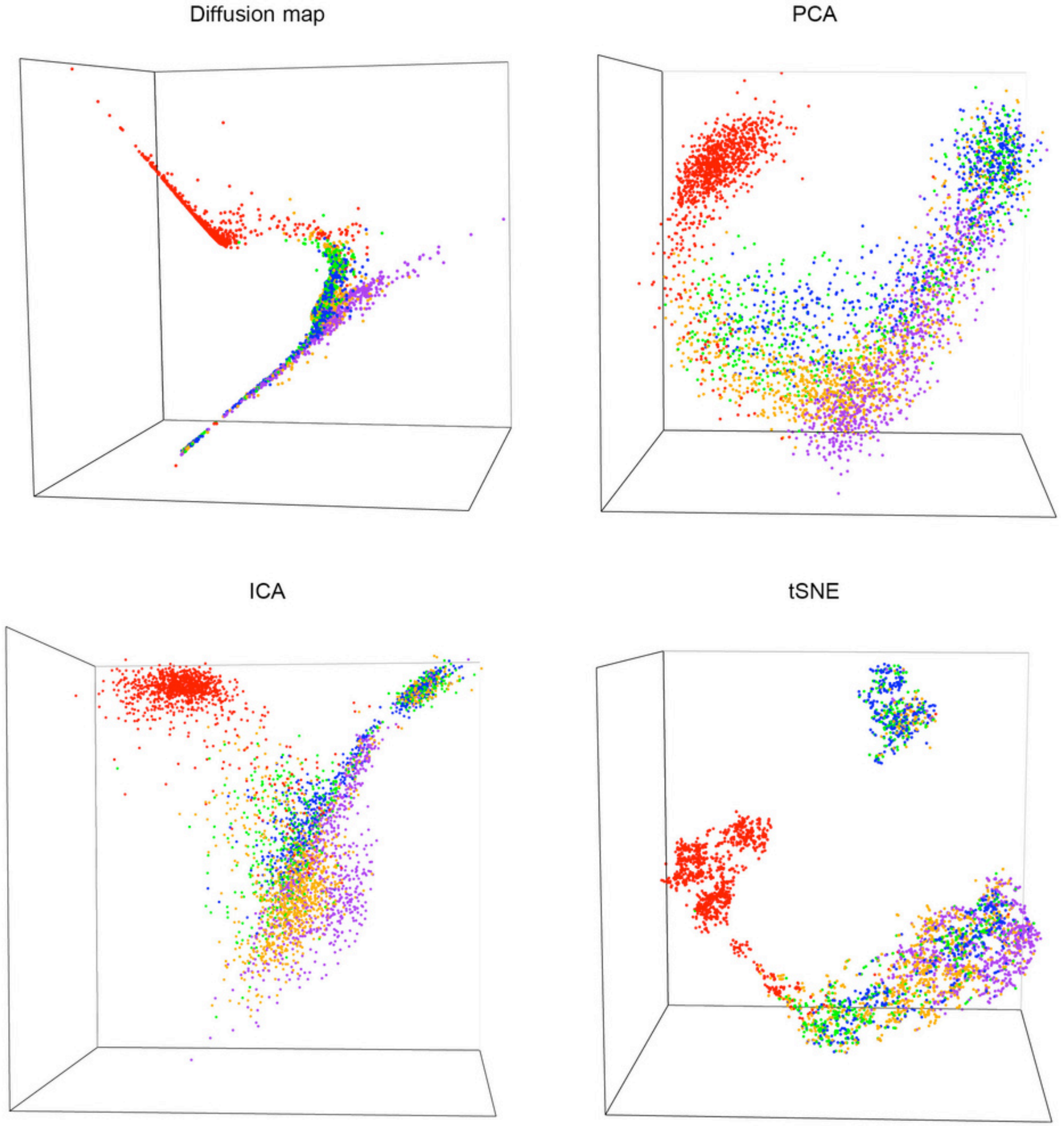
- Linear methods:
  - PCA (principal component analysis)
  - ICA (independent component analysis)
  - MDS (multidimensional scaling)
- Non-linear methods:
  - Non-linear PCA
  - t-SNE (t-distributed stochastic neighbor embedding)
  - Diffusion maps
  - Network based methods
- A PCA is a very good start in getting to know your data and understanding biases, batch effects etc.

t-SNE vs PCA dimensionality reduction



(Björklund et al. *Nature Immunology* 2016)

More dimensionality reductions

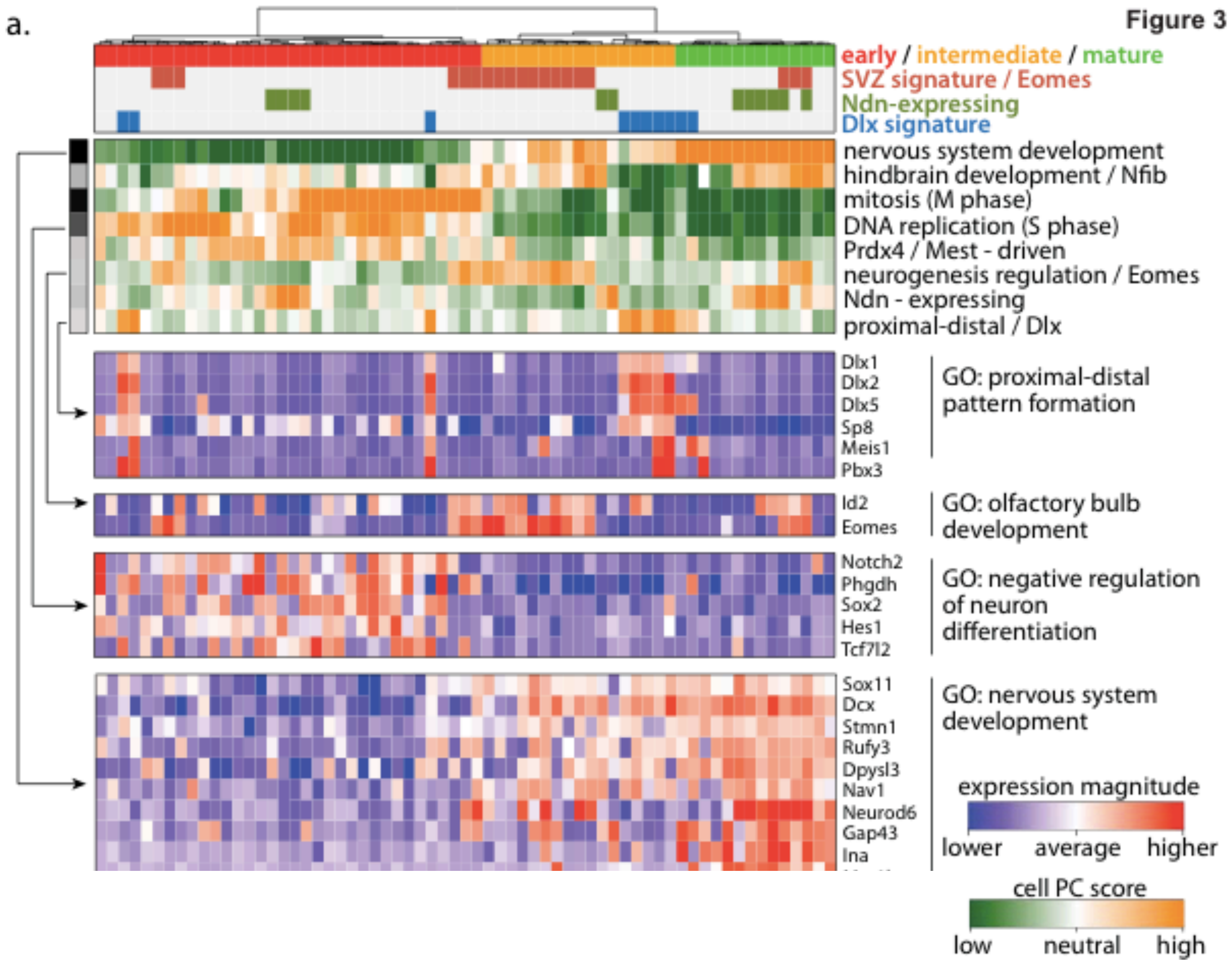


(Moignard et al. *Nature Biotech* 2015)

Identifying celltypes - Clustering

- Clustering based on
  - rpkms/counts – Euklidean distances
  - Pairwise correlations
  - PCA or other dimensionality reduction method
- Method of choice: hierarchical, k-means, biclustering
- Some programs:
  - WGCNA
  - BackSPIN
  - Pagoda
  - DBscan
- OBS! Outlier removal as an initial step may be necessary, especially with PCA-based clustering or similar.

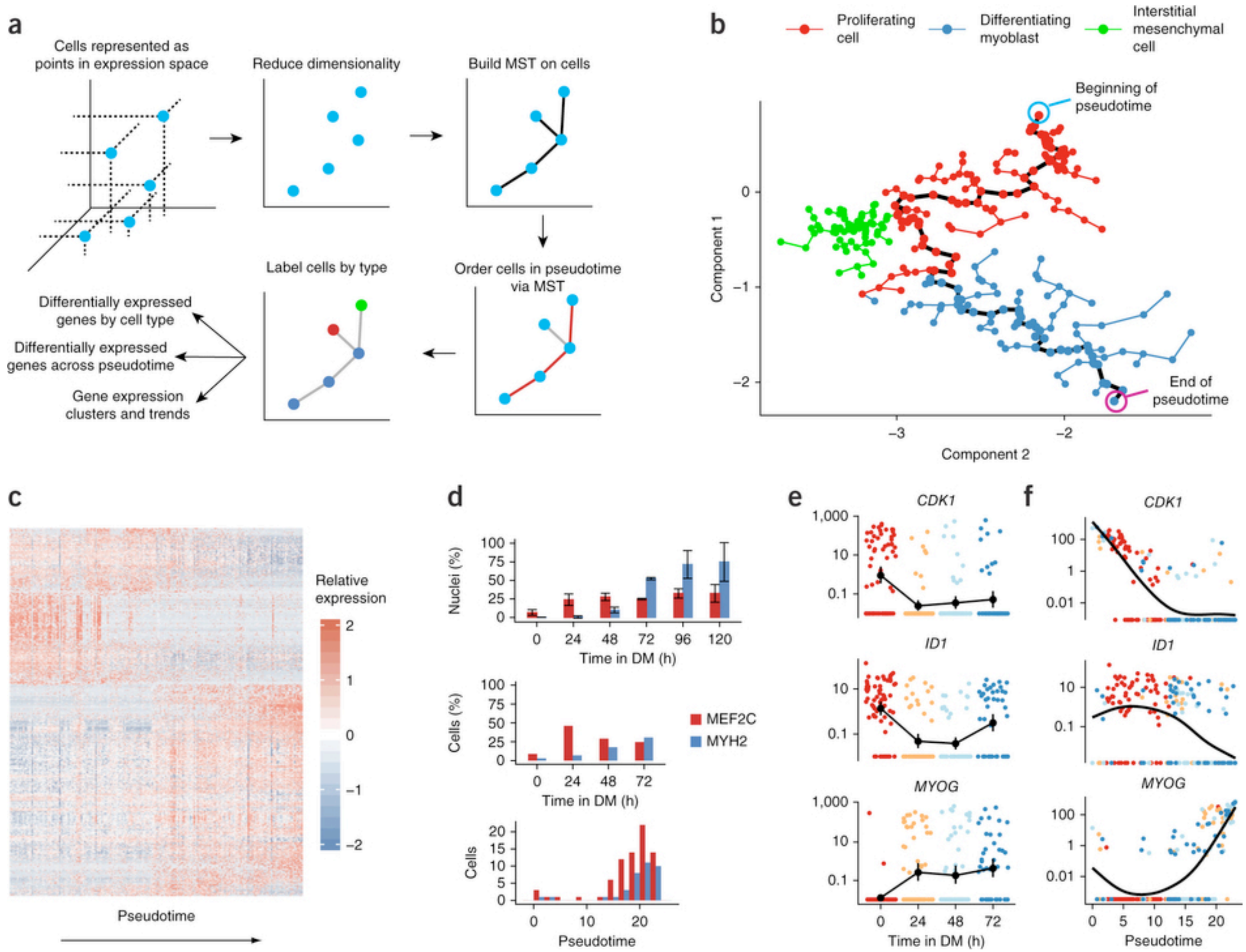
Pagoda – Pathway And Geneset OverDispersion Analysis



(Fan et al. *Nature Methods* 2016)

Pseudotime ordering - Monocle

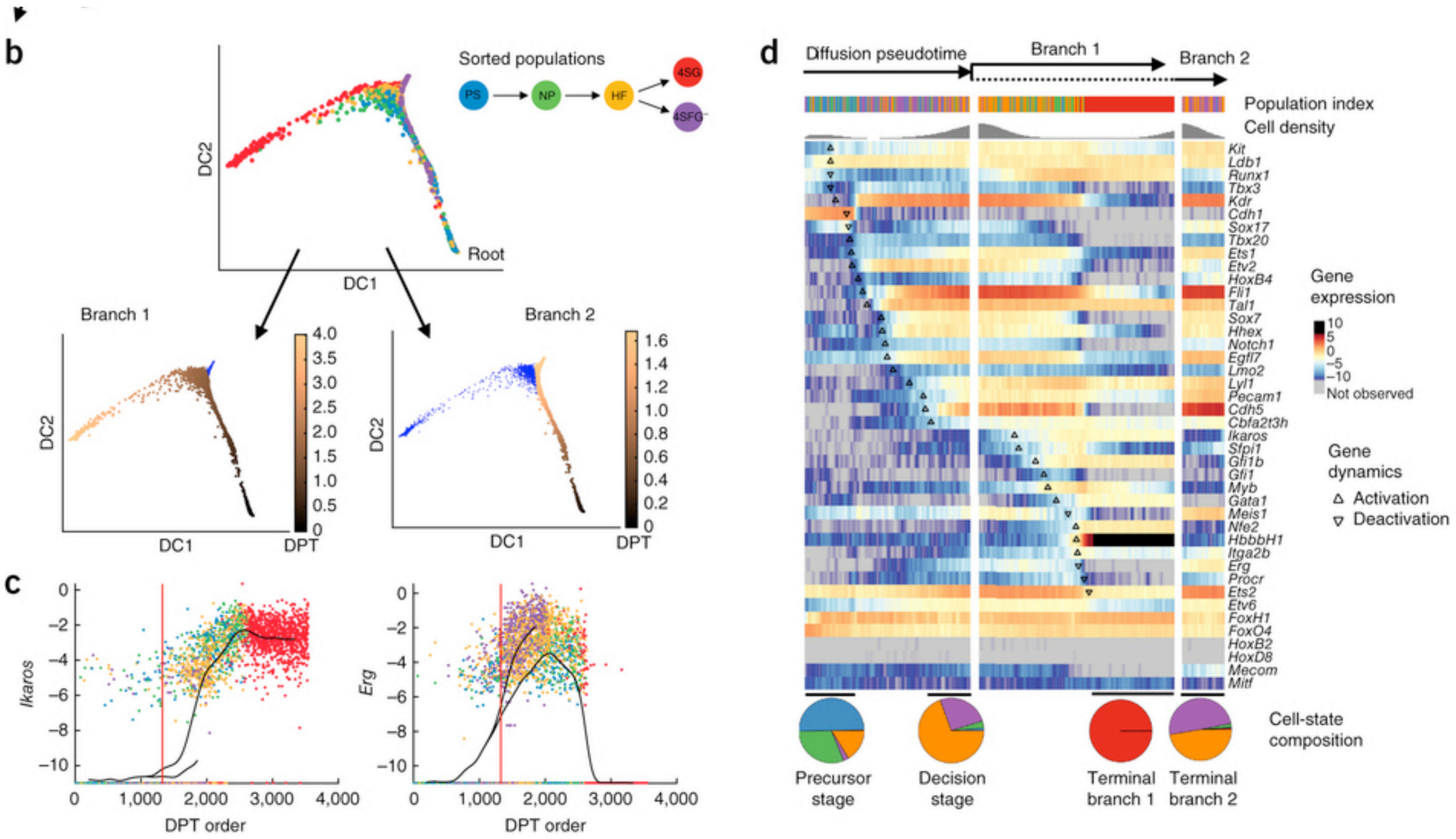
The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.



(Trapnell et al. *Nature Biotech* 2014)

Diffusion pseudotime

Diffusion pseudotime robustly reconstructs lineage branching



(Haghvedi et al *Nature Methods* 2016)

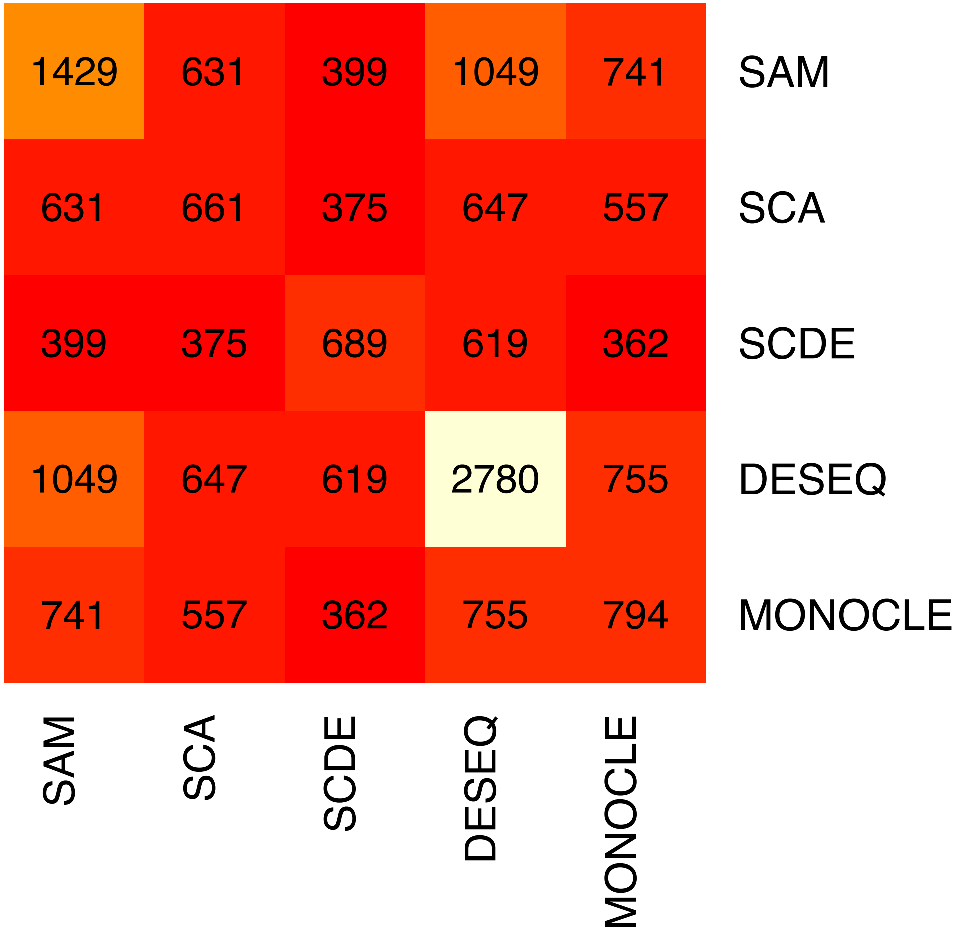
Detecting differentially expressed genes

- Parametric methods like EdgeR & DESeq not suitable for scRNAseq since the parameter assumptions in those methods does not apply here.
- Can use non-parametric methods like SAMseq

Detecting differentially expressed genes

- Available single cell DE methods:
  - SingleCellAssay – developed for qPCR experiments
  - Monocle package
  - Single Cell Differential Expression - SCDE
  - Model-based Analysis of Single-cell Transcriptomics – MAST
  - SAMstrt – extention to SAMseq with spike-in normalization
  - Many other recent publications…….
- Some studies use PCA contribution (loadings) or gene clustering to define celltype specific genes with no statistical DE test at all.

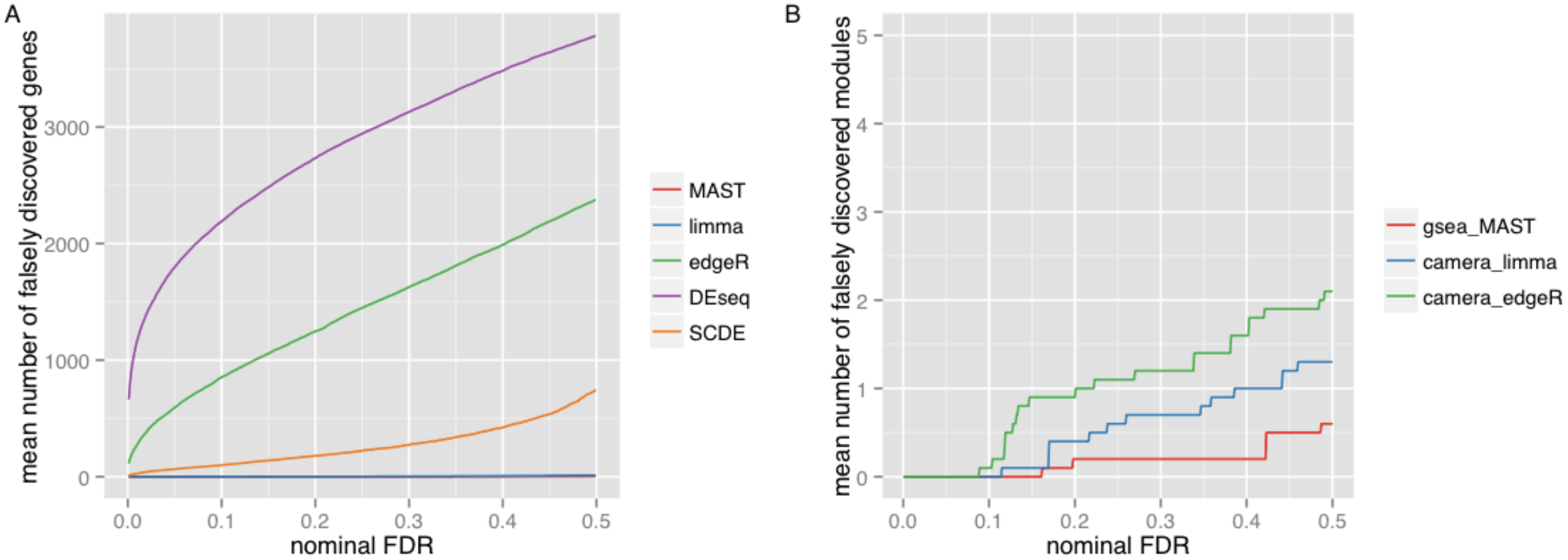
Comparison of DE detection methods



	SAM	SCA	SCDE	DESEQ	MONOCLE
SAM	1429	631	399	1049	741
SCA	631	661	375	647	557
SCDE	399	375	689	619	362
DESEQ	1049	647	619	2780	755
MONOCLE	741	557	362	755	794

Number of DE genes between 2 celltypes using SAMseq, Single Cell Assay, SCDE, DESeq2 and Monocle. Numbers along diagonal are total number of genes, boxes on each side shows overlap between methods.

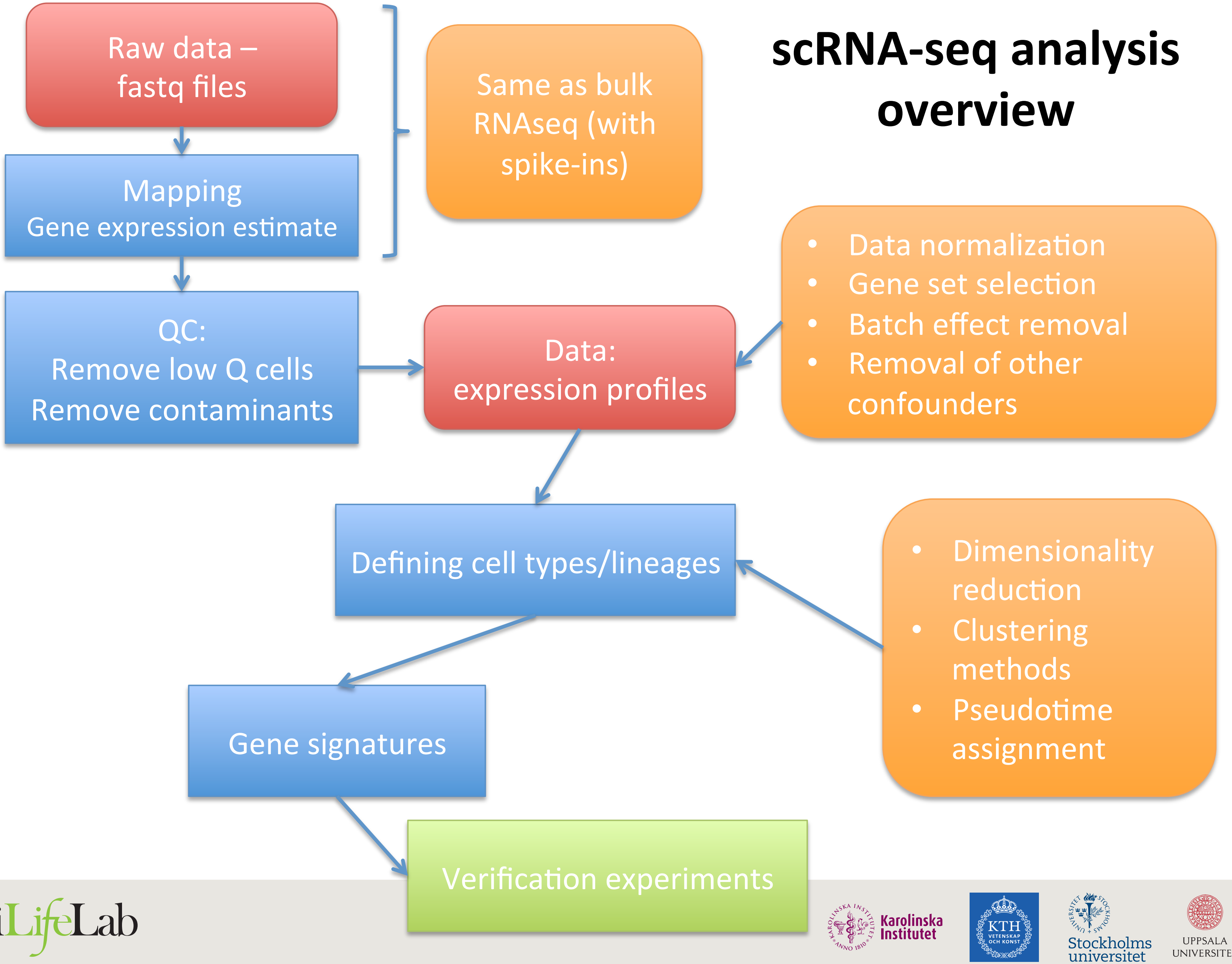
High false discovery rate for DESeq and EdgeR



Detected DE genes and gene sets using randomly permuted cells from unstimulated MAIT cells

(Finak et al. *Genome Biology* 2015, Supplementary material)

scRNA-seq analysis overview



Raw data – fastq files

Mapping
Gene expression estimate

Same as bulk RNAseq (with spike-ins)

QC:
Remove low Q cells
Remove contaminants

Data:
expression profiles

- Data normalization
- Gene set selection
- Batch effect removal
- Removal of other confounders

Defining cell types/lineages

- Dimensionality reduction
- Clustering methods
- Pseudotime assignment

Gene signatures

Verification experiments

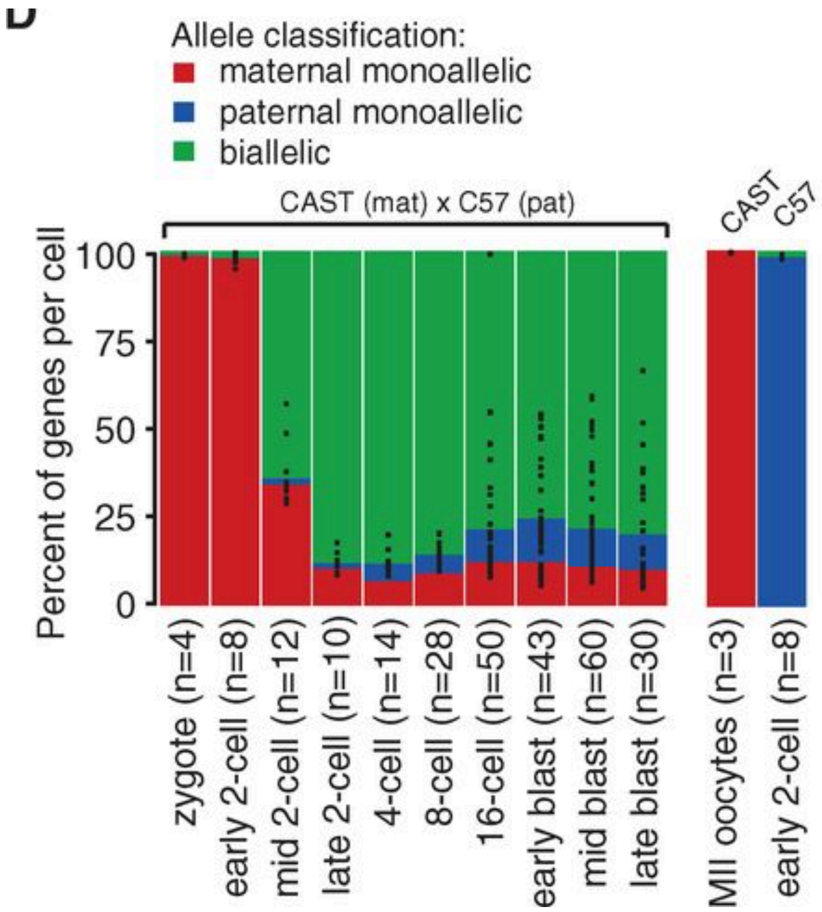
Additional analyses

- Allelic expression
- Variant calling
- Copy-number variation
- Alternative splicing

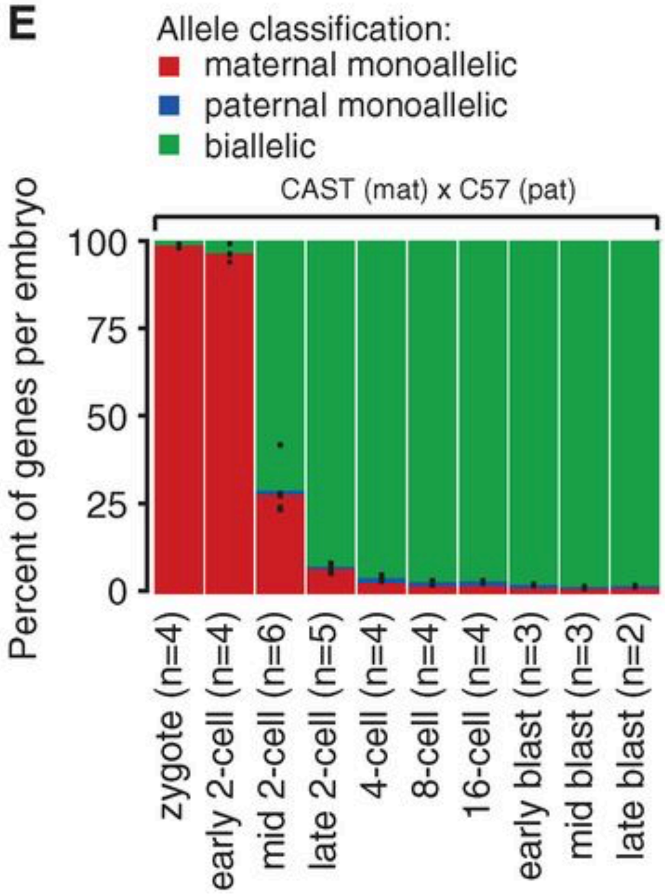
- Alternative splicing and allelic expression requires full length methods.
  - But only works for highly expressed genes with good read coverage
  - Must be careful to take into consideration the drop-out rate, a unique splice form/allele in a single cell may actually be a detection issue.

Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells

Single cells

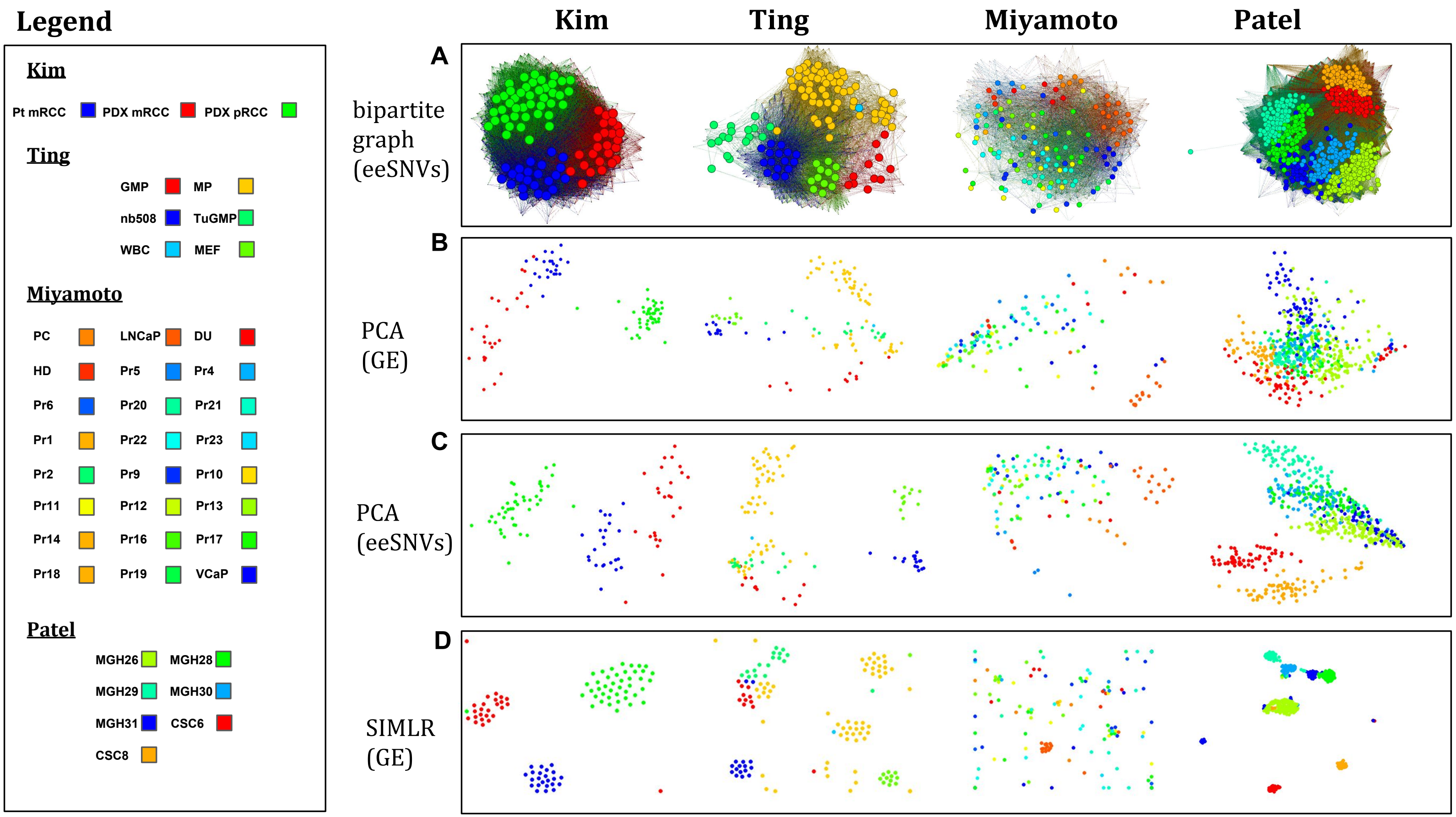


Pooled embryos



(Deng et al. *Science* 2014)

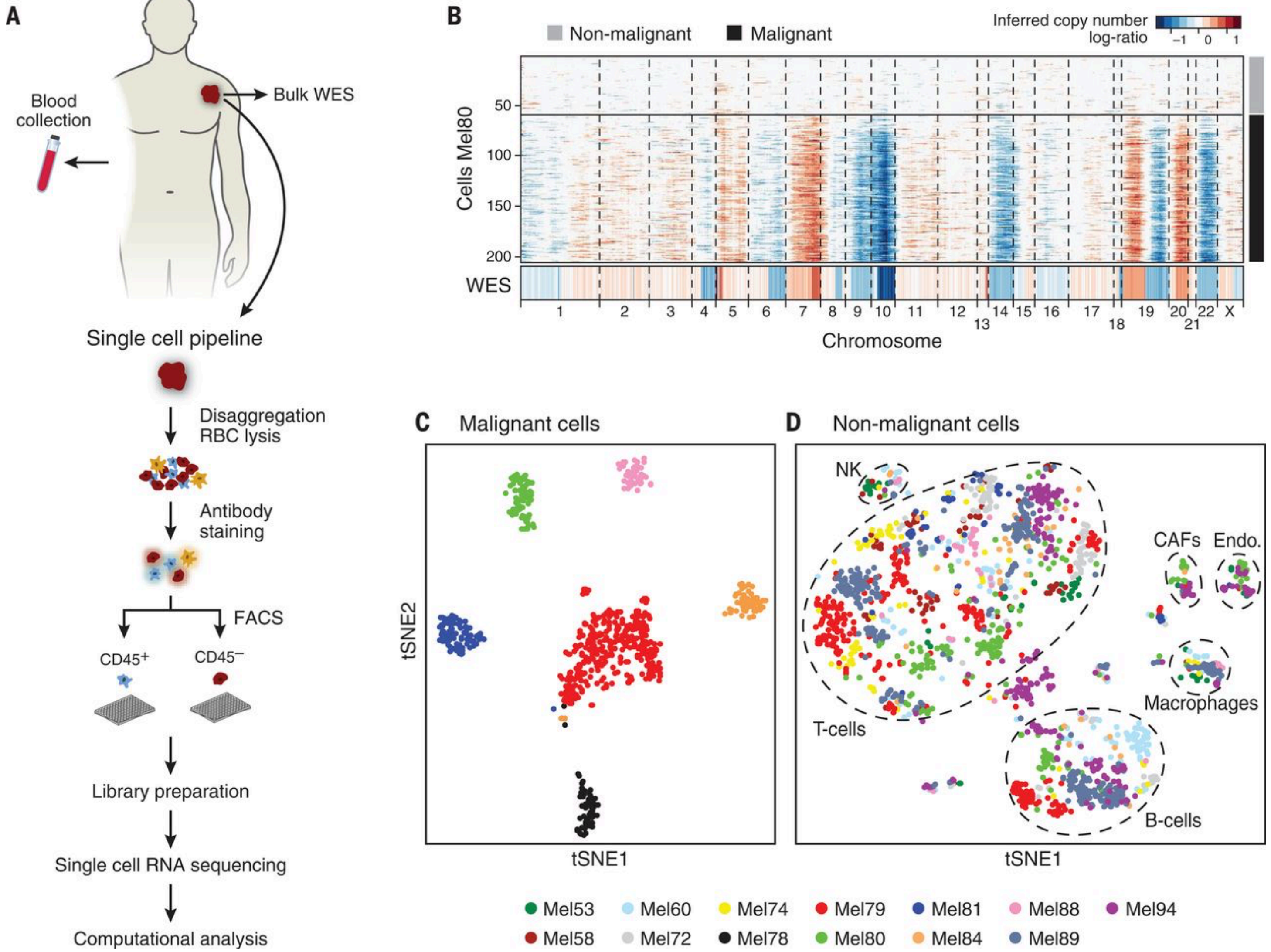
Using Single Nucleotide Variations in Cancer Single-Cell RNA-Seq Data for Subpopulation Identification and Genotype-phenotype Linkage Analysis



(Poiron et al. *BioRxiv* 2016)

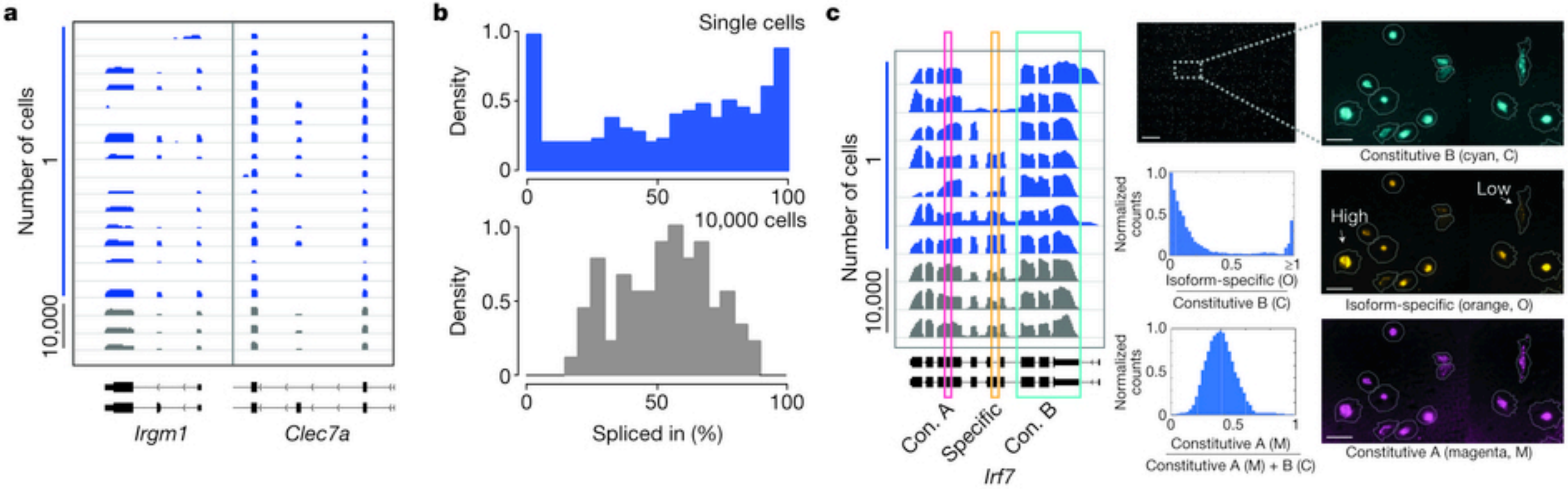
SciLifeLab Karolinska Institutet KTH VETENSKAP OCH KONST Stockholms universitet UPPSALA UNIVERSITET

Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq



(Tirosh et al. *Science* 2016)

Cell specific alternative splicing



(Shalek et al. *Nature* 2013)

Combination with single cell genome sequencing

- G&T-seq (Macaulay et al. *Nature Methods* 2015) – DNA + RNA
- DR-seq (Dey et al. *Nature Biotech* 2015) – DNA + RNA
- scTrio-seq (Hou et al. *Cell Research* 2016) - RNA + DNA & DNA methylome
- scM&T-seq (Angermueller *Nature Methods* 2016) – RNA + DNA methylome

Non coding RNA single cell library

- SUPeR-seq – random hexamer primer instead of polyA-priming (Fan et al. *Genome Biology* 2015)
  - Detect circular RNAs and other non-coding RNAs as well as mRNAs
- small RNA library prep from single cells (Faridani et al. *Nature Biotech* 2016)
  - Ligate adapters to 5′ phosphate and 3′ hydroxyl groups
  - Detect miRNA, snoRNA etc.

Conclusions

- For diverse cell-types often straight forward to group cells into clusters and detect differentially expressed genes.
- For highly similar subtypes or with subtle changes in cellular states – feature selection and different clustering methods may be required.
- PCA or other dimensionality reduction technique is a good start to get to know your data.

Some tools for single cell analysis

- Tutorial from Harvard WS:
  - http://pklab.med.harvard.edu/scw2015/
- For differential expression:
  - SCDE: http://pklab.med.harvard.edu/scde/index.html
  - SCA: https://github.com/RGLab/SingleCellAssay
  - MAST: https://github.com/RGLab/MAST
  - SAMseq: http://cran.r-project.org/web/packages/samr
- For clustering, normalization etc.:
  - Monocle2/Census: https://github.com/cole-trapnell-lab/monocle-release
  - Rtsne: http://cran.r-project.org/web/packages/Rtsne
  - Sincell: http://master.bioconductor.org/packages/devel/bioc/html/sincell.html
  - scLVM: https://github.com/PMBio/scLVM
  - BASiCS: https://github.com/catavallejos/BASiCS
  - Pagoda: http://pklab.med.harvard.edu/scde
  - Seurat toolkit: http://www.satijalab.org/seurat.html
  - Sincera pipeline: https://research.cchmc.org/pbge/sincera.html
  - SimpleSingleCell pipeline: https://www.bioconductor.org/help/workflows/simpleSingleCell/