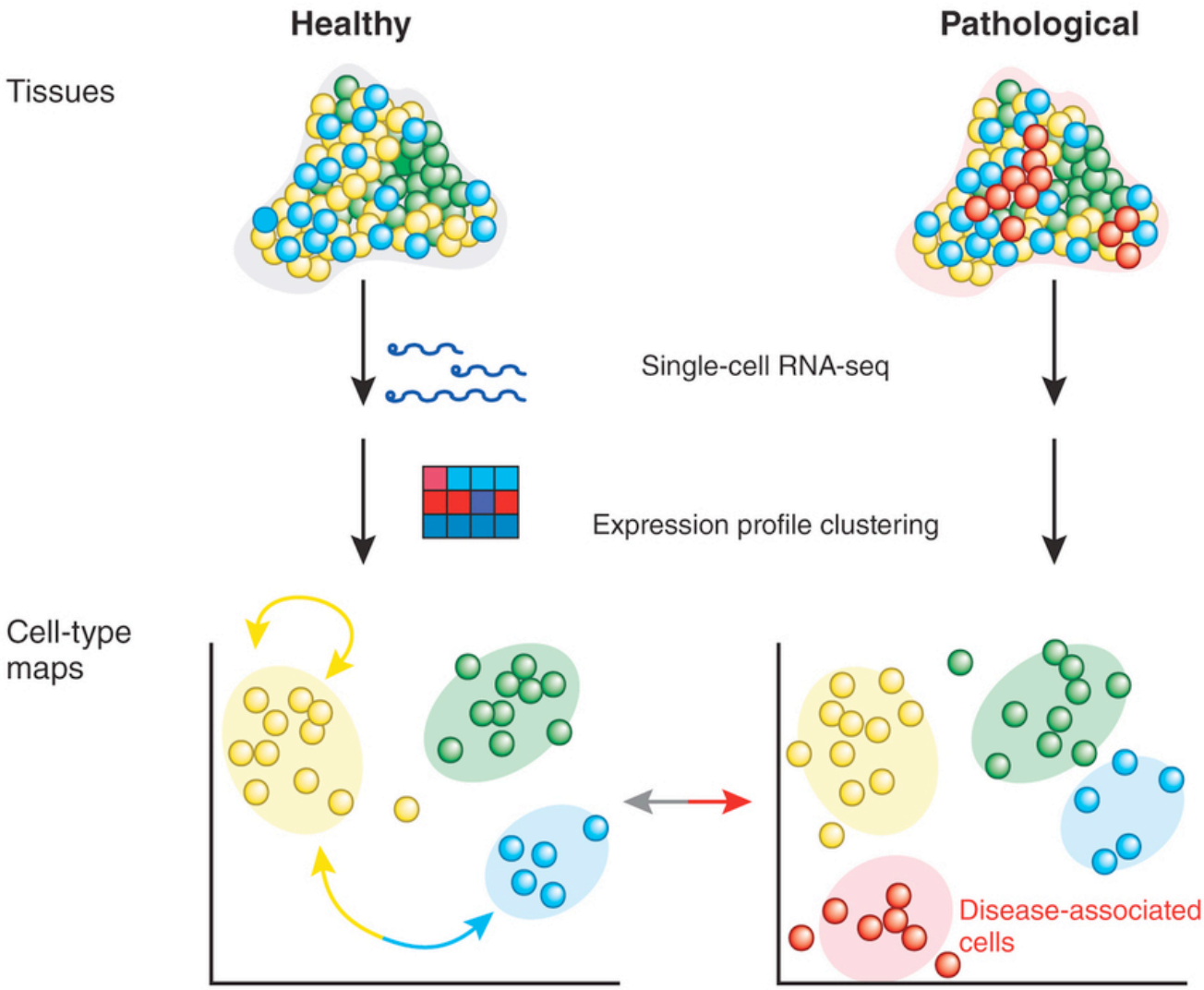
Single cell RNA sequencing

Åsa Björklund

asa.bjorklund@scilifelab.se

Outline

- Why single cell gene expression?
- Library preparation methods
- Setting up experiments
- Computational analysis
  - Defining cell types
  - Identifying differentially expressed genes
- Some recent papers



Types of analyses

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

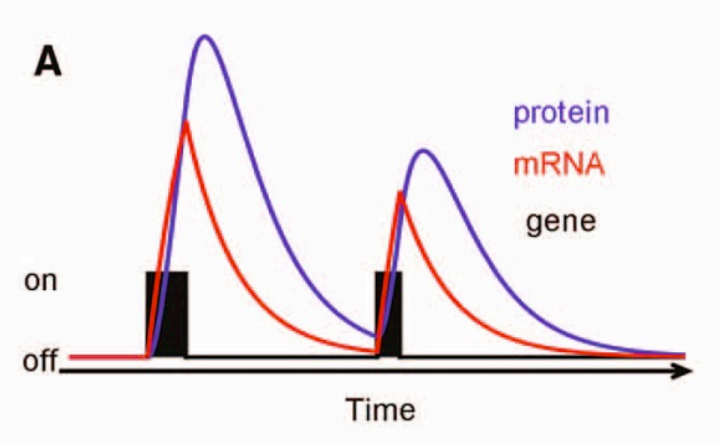
Between tissues
- Cell-type compositions
- Altered transcription in matched cell types

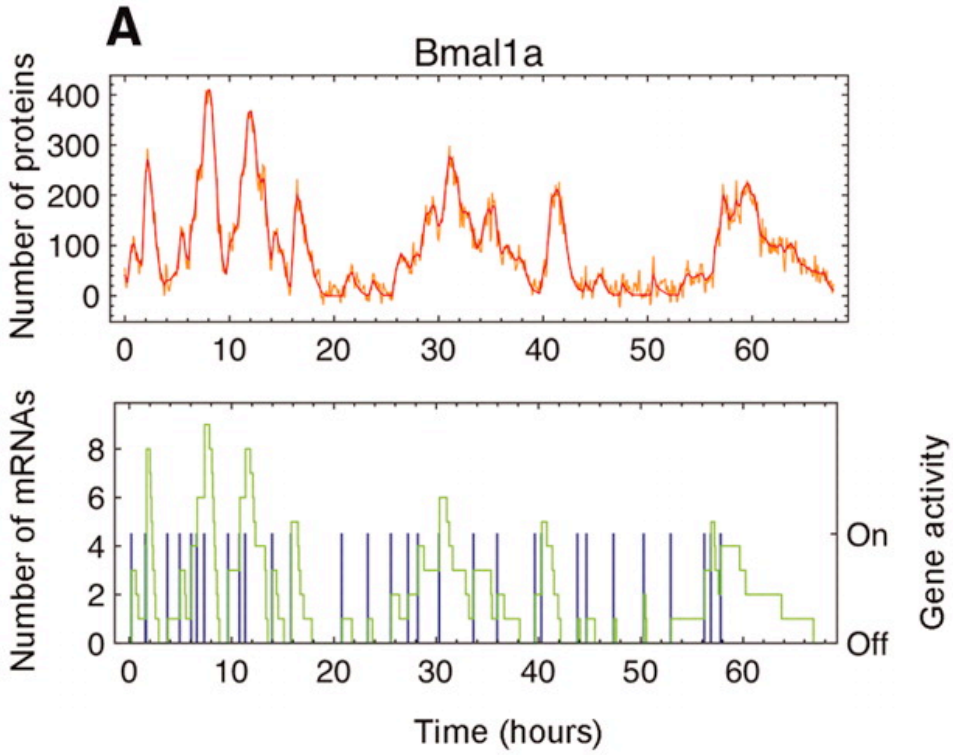
(Sandberg, *Nature Methods* 2014)

Why single-cell sequencing?

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

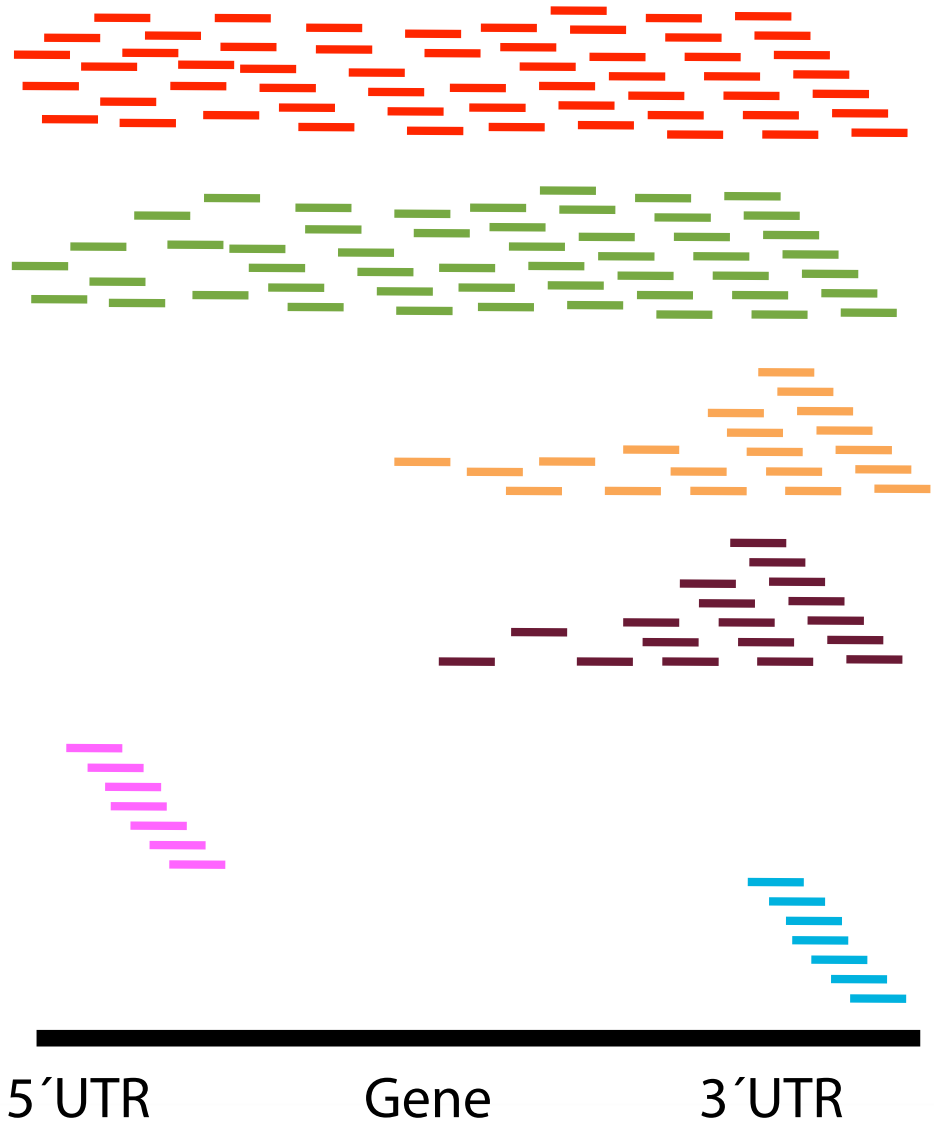
Transcriptional bursting





- 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

scRNA seq methods



SmartSeq2
(Picelli et al. Nature Methods 2014)

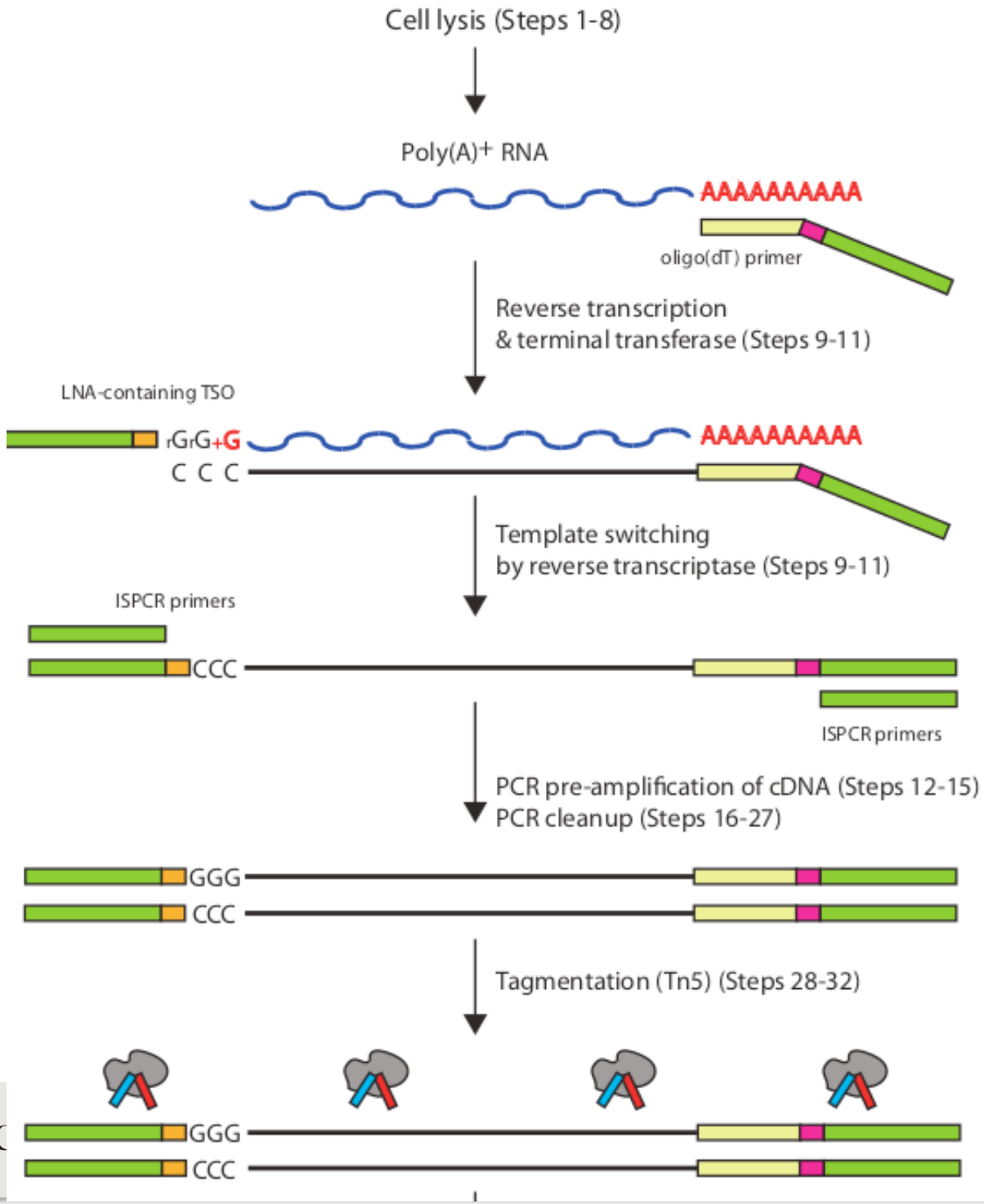
SmartSeq – SMARTer kit
(Ramsköld et al. Nature Biotech 2012)

Quartz-seq
(Sasagawa et al. Genome Biology 2013)

Tang et al.
(Nature methods 2009)

STRT
(Islam et al. Genome Res 2011)

CEL-Seq
(Hashimshony et al. Cell Reports 2012)



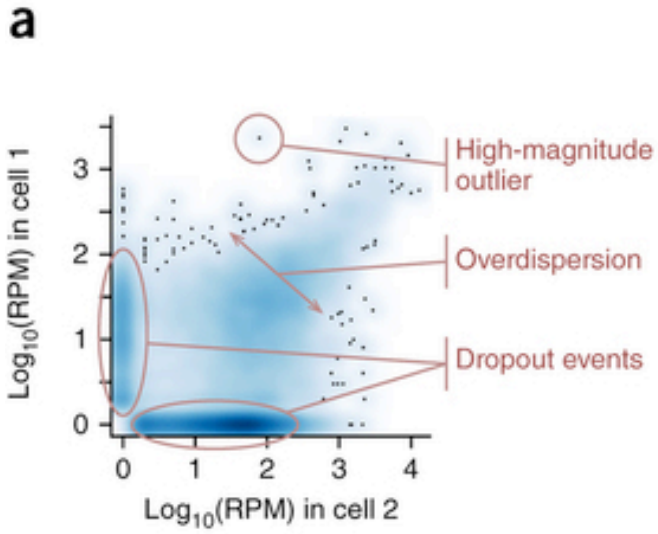
SmartSeq2 protocol

Reverse transcription efficiency limits the detection range
– Drop outs

(Picelli et al. *Nature Protocols*, 2014)

Problems compared to bulk RNA-seq

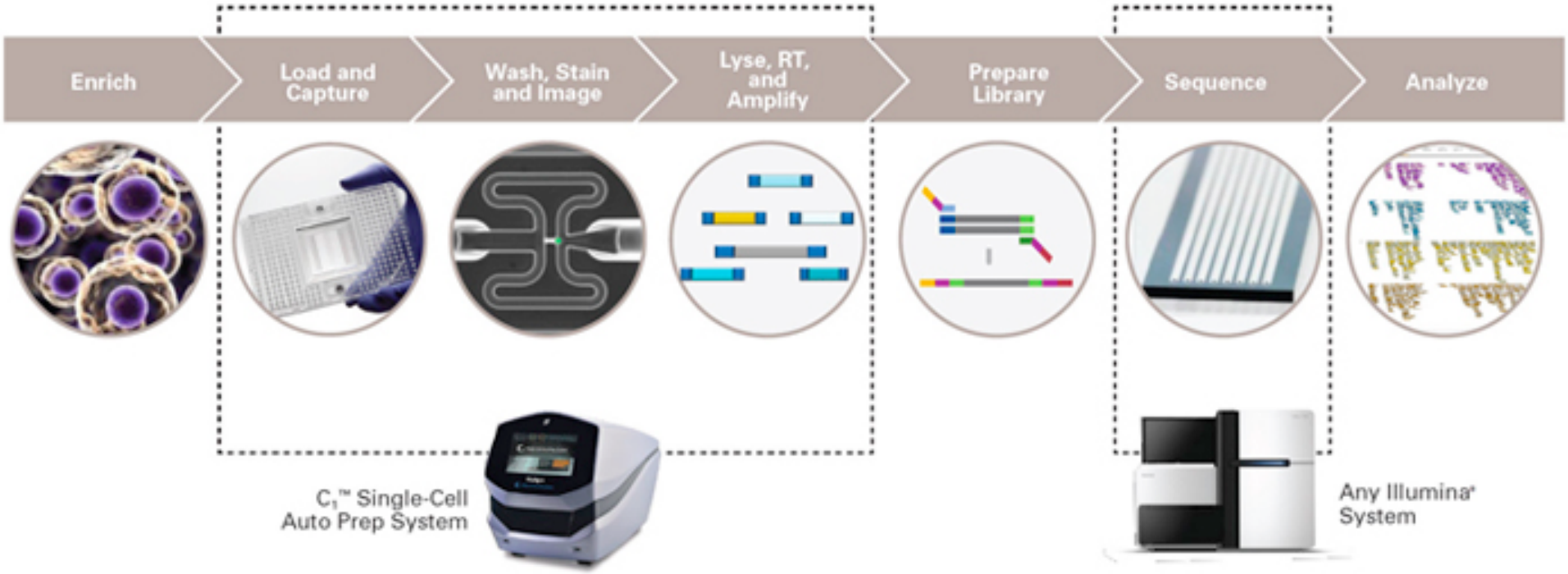
- Amplification bias
- Drop-out rates
- Stochastic gene expression
- Background noise
- Bias due to cell-cycle, cell size and other factors
- As of now, only polyA transcripts, no method for total RNA sequencing in single cells



(Karchenko et al. *Nature Methods* 2014)

Isolating single cells

- FACS sorting
- Manual picking
- Fluidigm C1 system
- Dissociation of tissue is a crucial step to minimize leakage and RNA degradation, different depending on tissue type.
- Cell types that are hard to dissociate:
  - Laser capture microscopy (LCM)
  - Nuclei sequencing

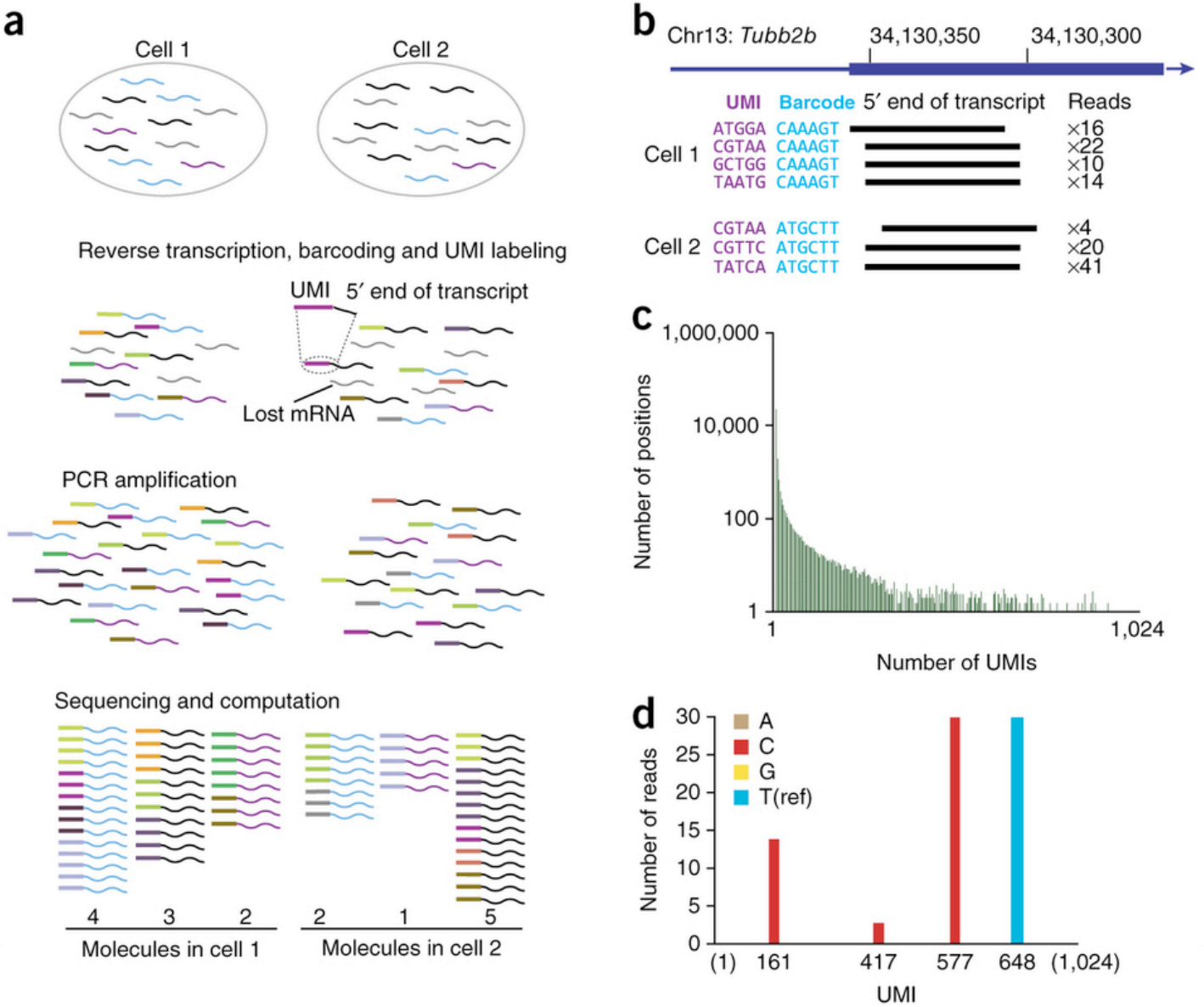


- Fluidigm C1 system
  - Limiting factor is size of capture chambers
  - Have protocols for running SMARTer, SmartSeq2, CEL-seq & STRT
  - Now with 800 cell chips

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

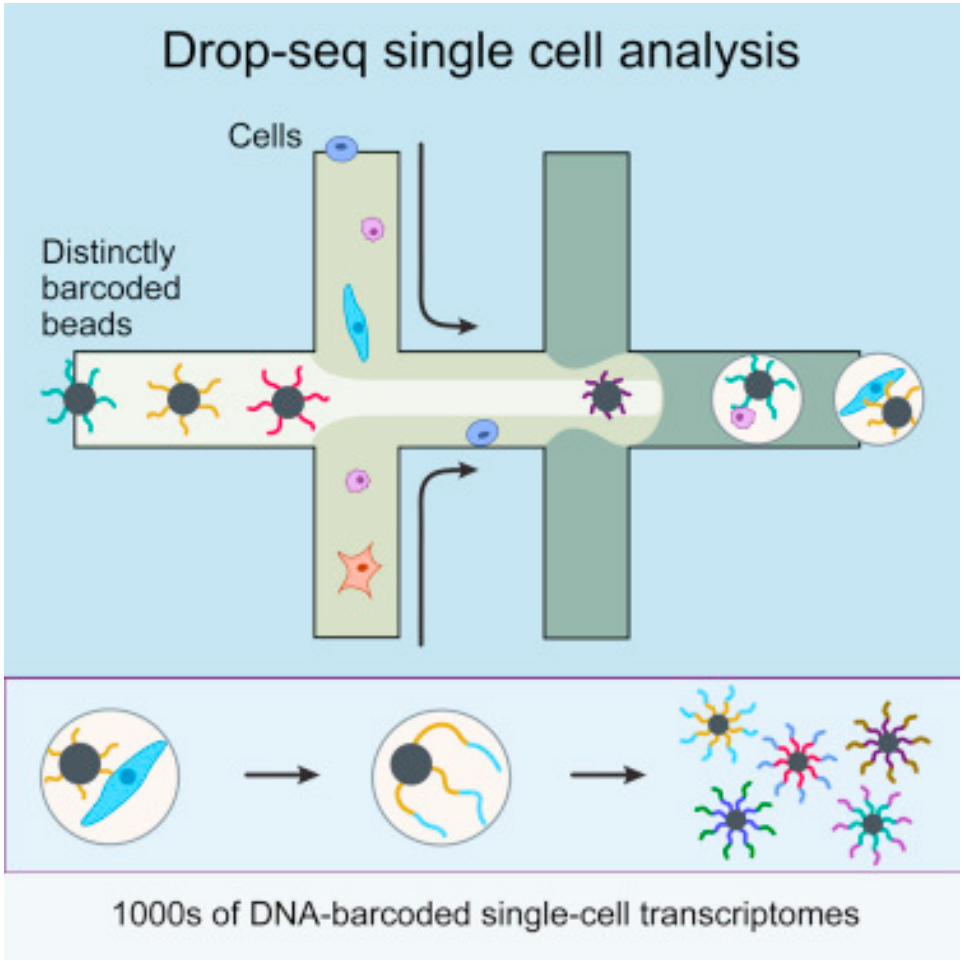


(Islam et al. *Nature Methods* 2014)

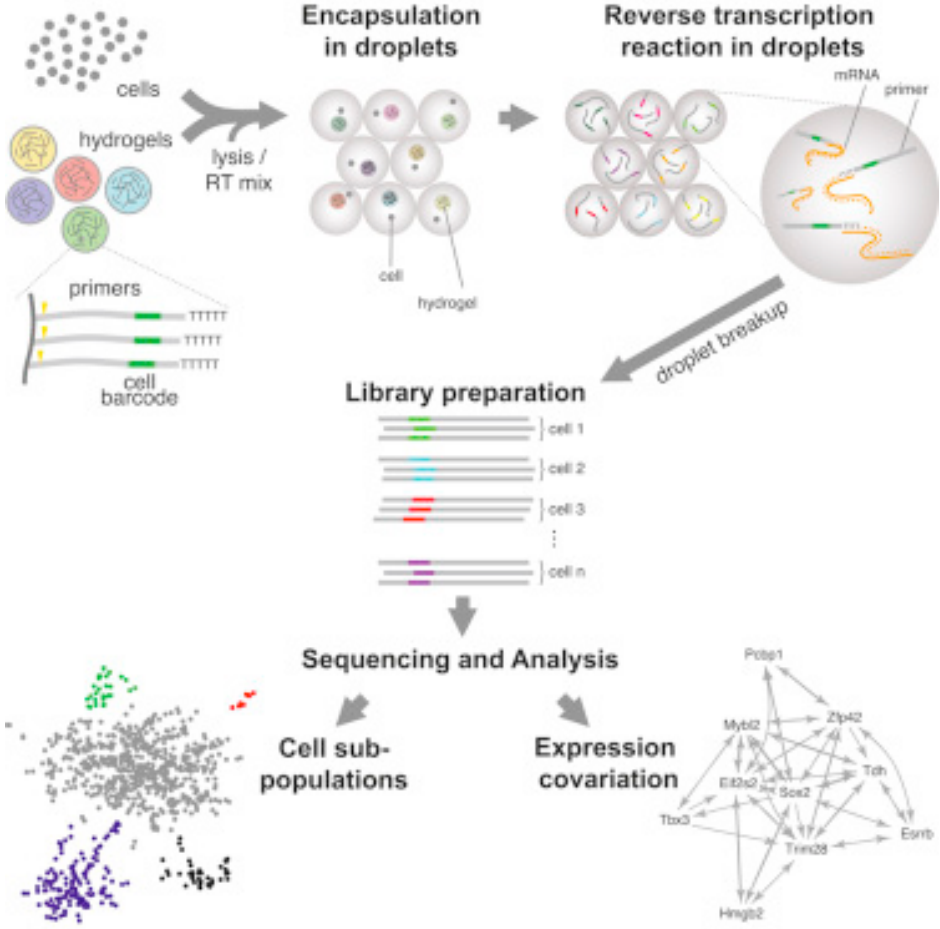
Small volume approaches

- Volume seem to be a key component in these reactions
  - Smaller volumes give higher detection and better 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



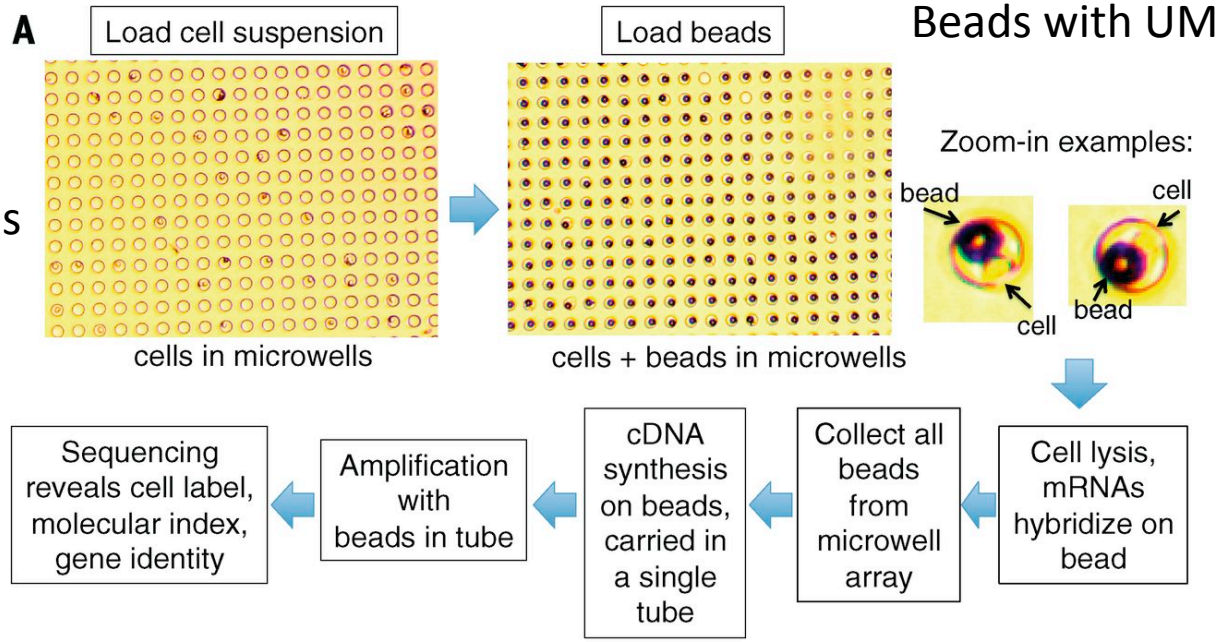
Macosko et al. *Cell* 2015
McCarrol, Regev etc. Broad/Harvard

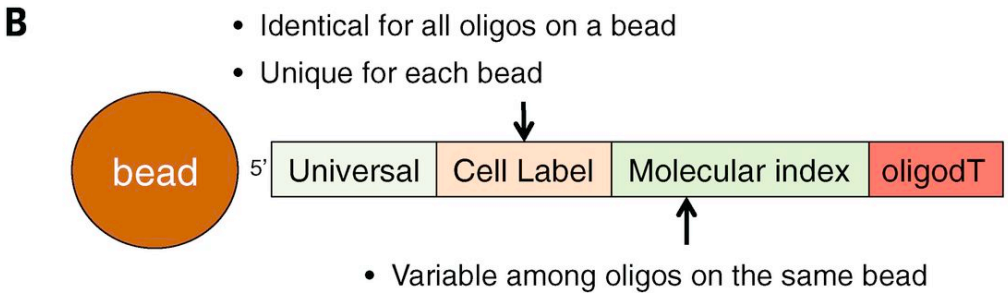


Klein et al. *Cell* 2015
Kirschner, Weitz etc. Harvard

Cyto-seq method

Surface with
100,000 microwells



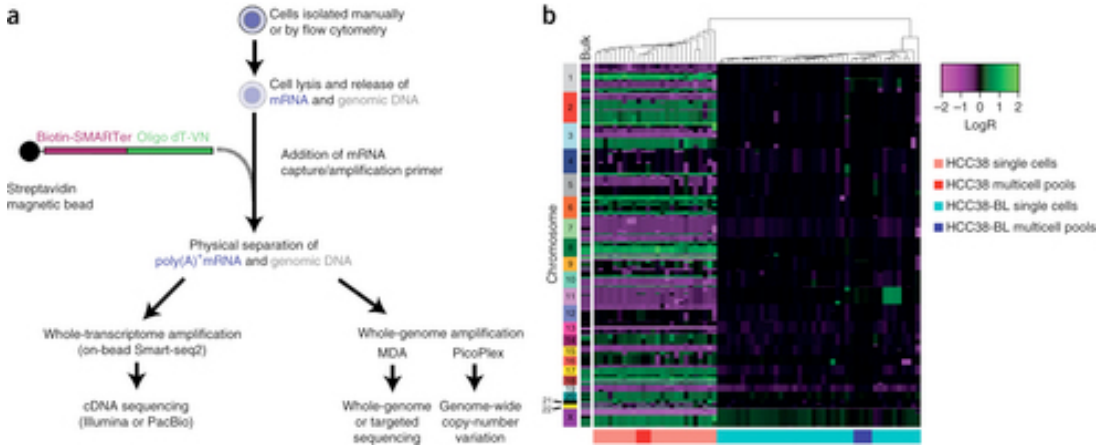


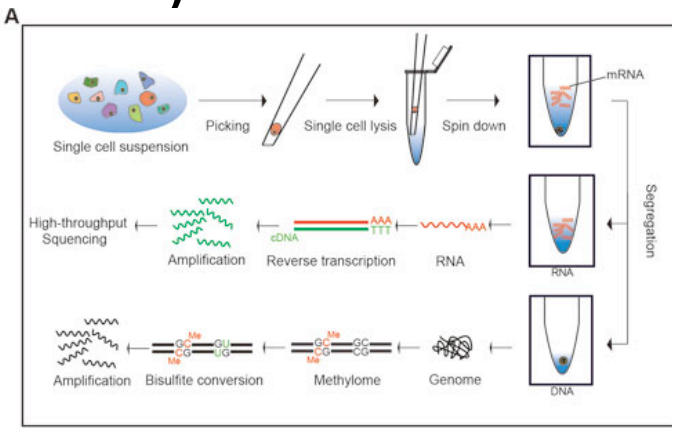
Presented as targeted sequencing method, but could be used with universal primers.

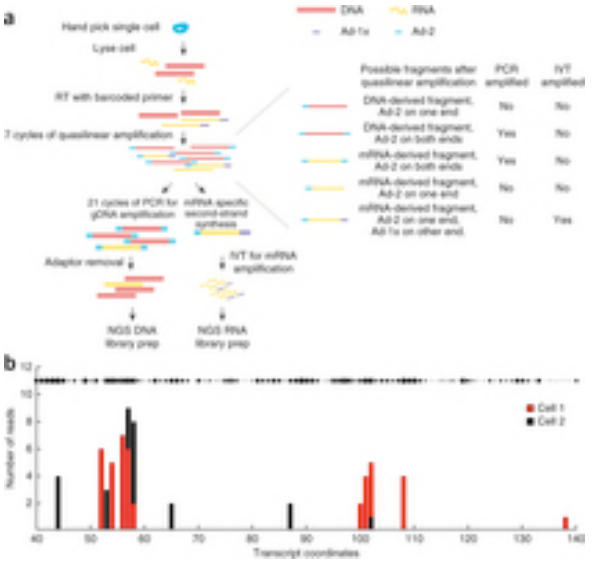
(Fan et al. *Science* 2015)

Combination with single cell genome sequencing

- G&T-seq (Macaulay et al. *Nature Methods* 2015)
- DR-seq (Dey et al. *Nature Biotech* 2015)
- Triple omics (Hou et al. *Cell Research* 2016)



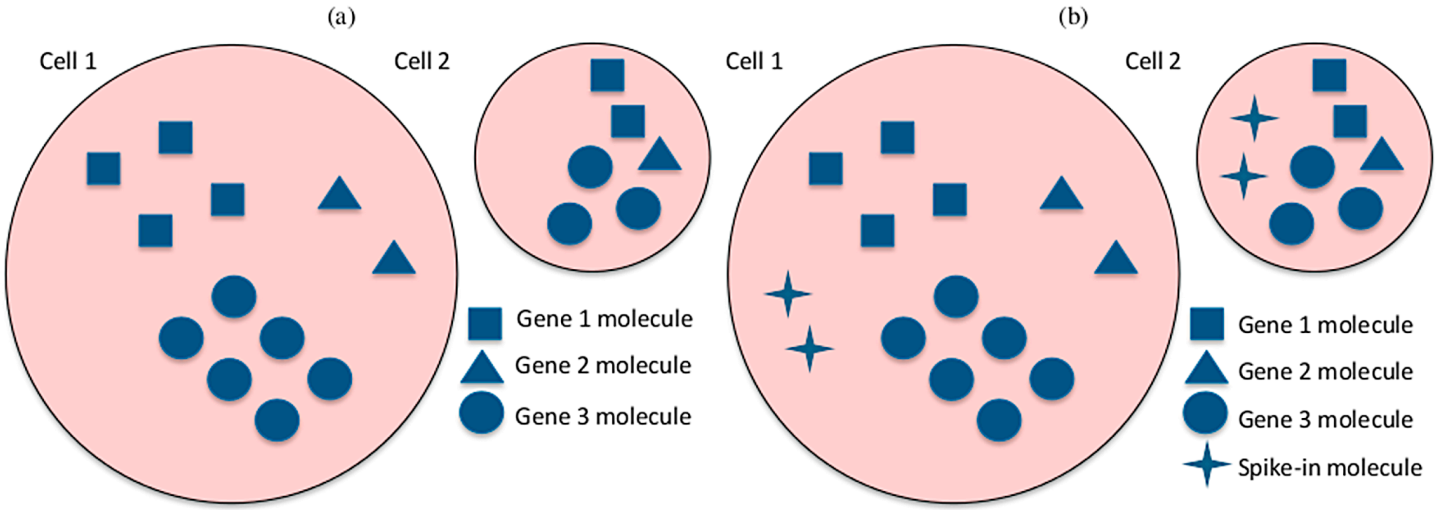




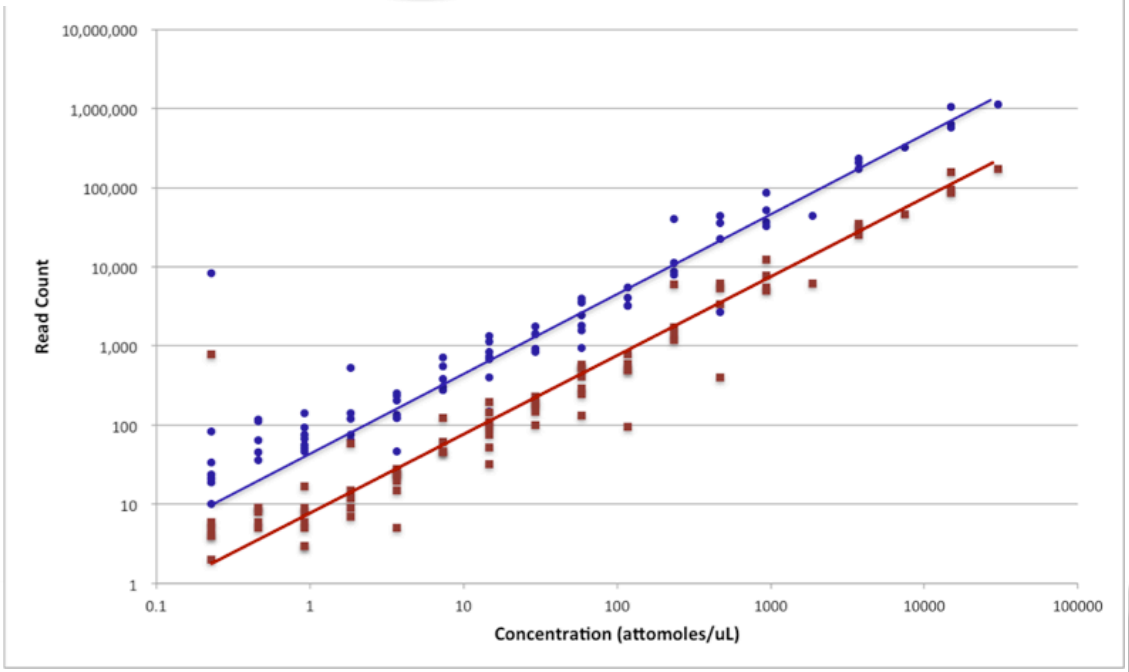
Spike-in RNAs

- 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.
- Add a ratio of about 1:10 to cell RNA.
- Important to add equal amounts to each cell, preferably in the lysis buffer.

Spike-in RNAs

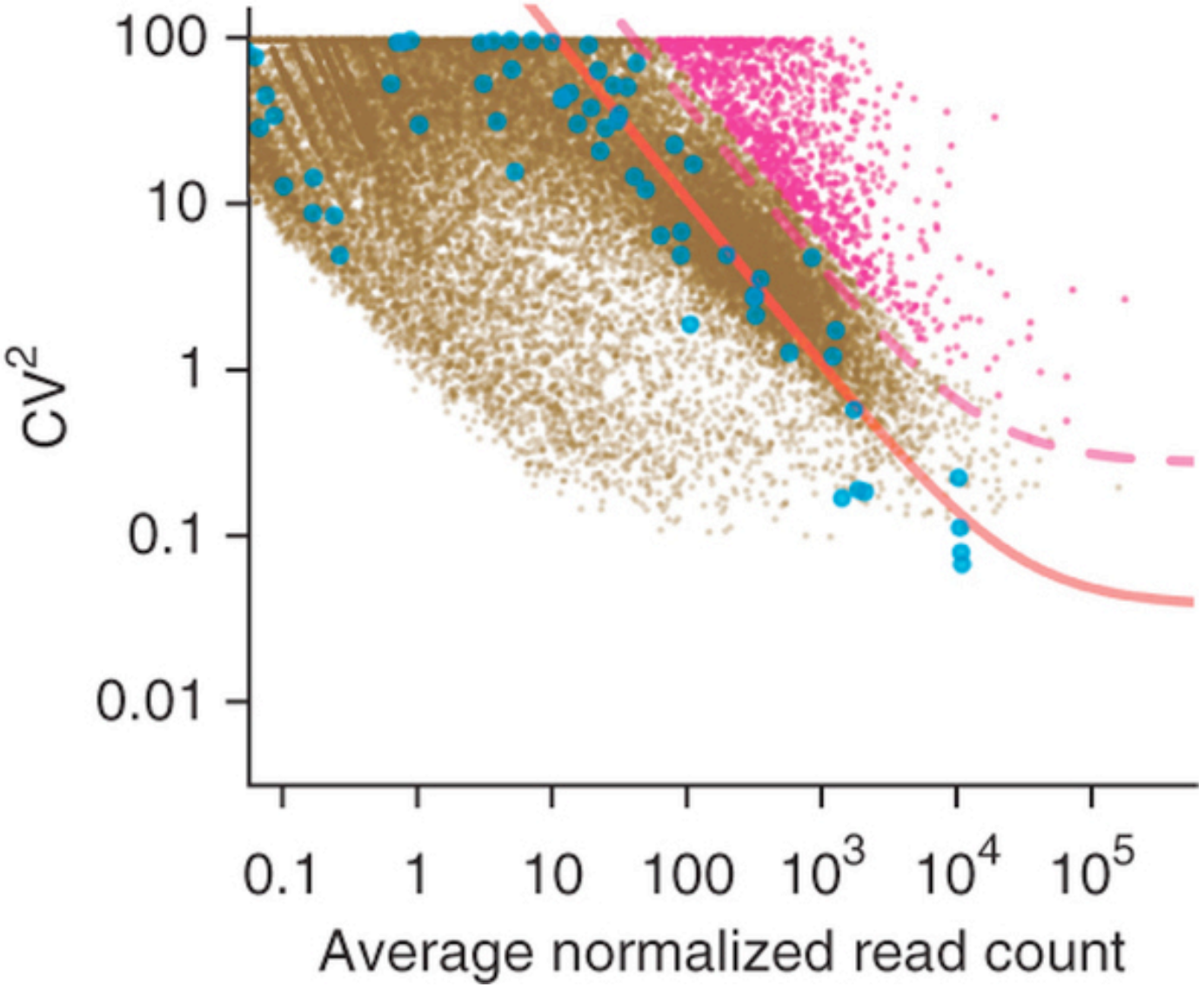


(Vallejos et al. *PLOS Comp Biol* 2015)



(https://cofactorgenomics.com)

Finding biologically variable genes



(Brennecke et al. *Nature Methods* 2013)

Replicates – how many cells do you have to sequence?

- Recommended to have around 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.
- 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.

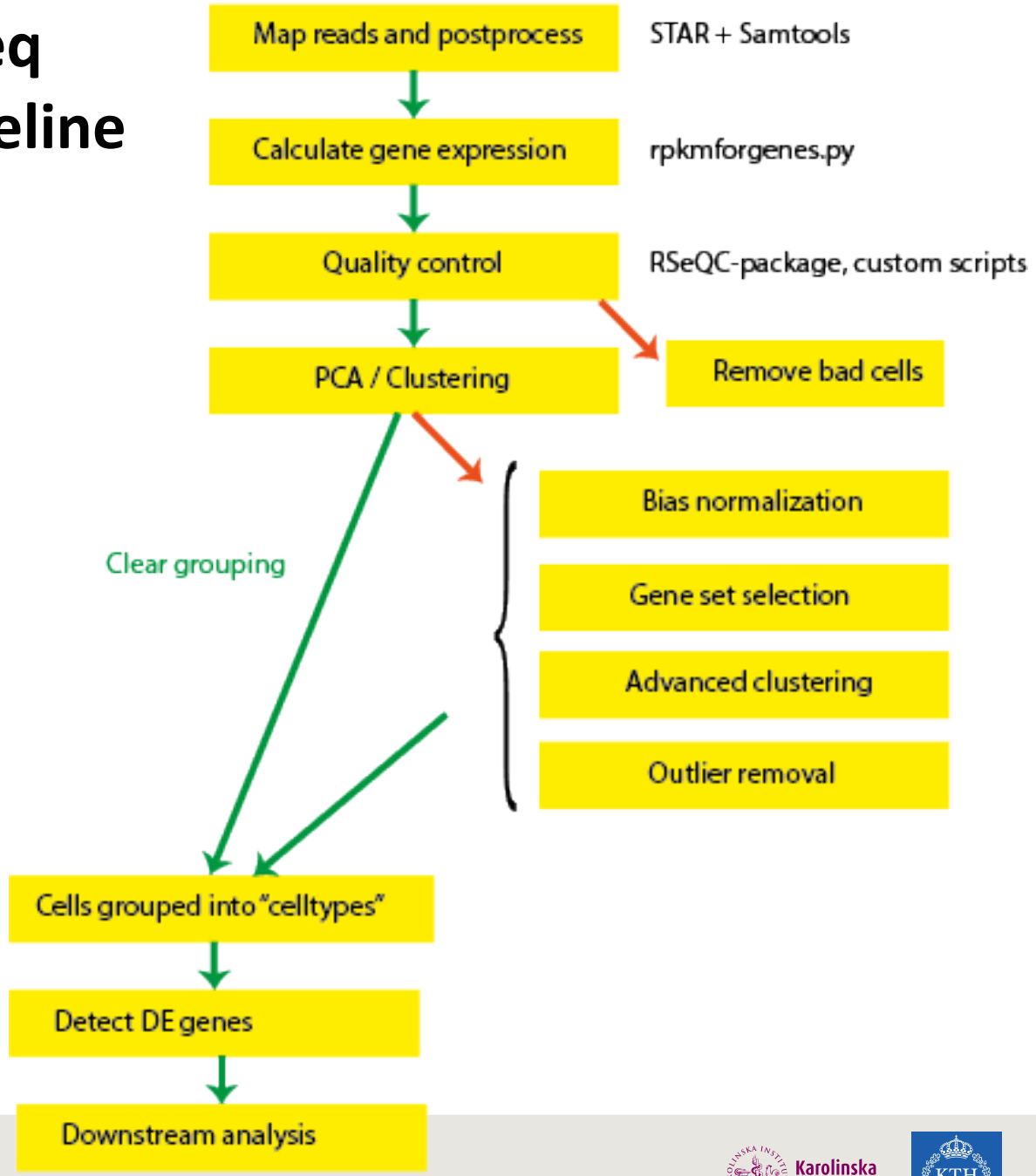
Which method should I use?

- 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
- The Single-cell platform offers SmartSeq2 in 384 well format or STRT on Fluidigm C1.

National single cell genomics platform at Scilifelab

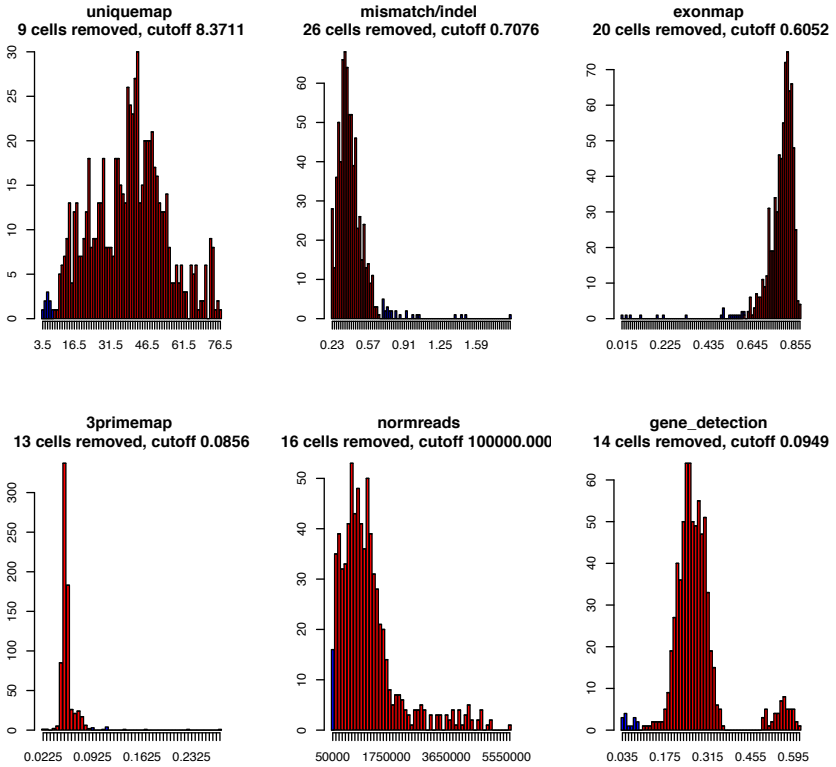
- Uppsala node – microbial single cell genome sequencing
  - http://www.scilifelab.se/facilities/single-cell/
  - MDA of whole genomes
  - qPCR of selected target genes
- Stockholm node – eukaryotic single cell RNA / genome sequencing
  - http://www.scilifelab.se/facilities/eukaryotic-single-cell-genomics/
  - STRT and cell isolation on Fluidigm C1 system
  - SmartSeq2 on isolated cells on plates
  - MDA whole genome sequencing

scRNA-seq analysis pipeline



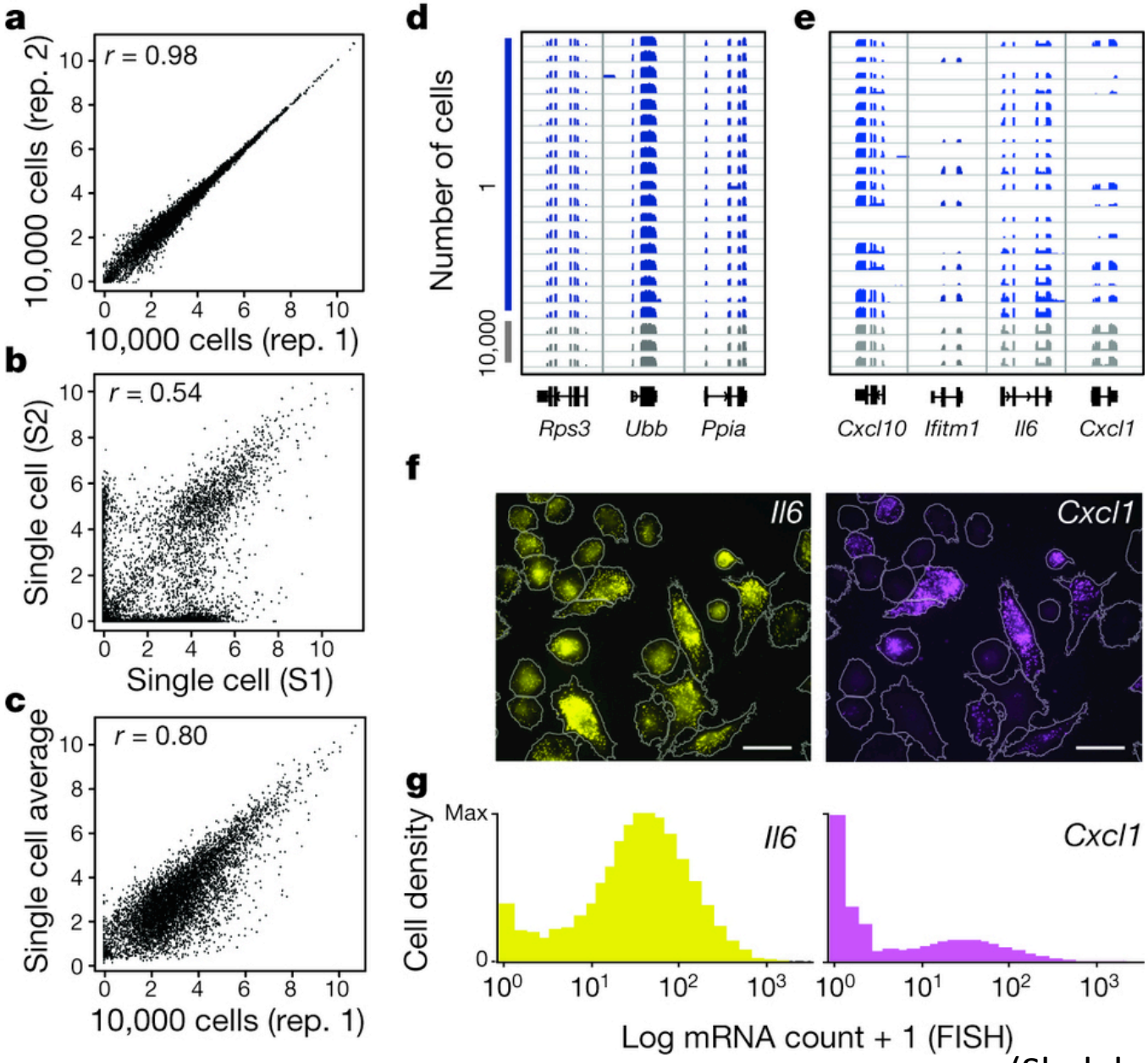
Quality Control (QC)

- 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.
- Look at:
  - Mapping statistics (% uniquely mapping)
  - Mismatch rate
  - Fraction of exon mapping reads
  - 3’ bias (degraded RNA)
  - mRNA-mapping reads
  - Number of detected genes



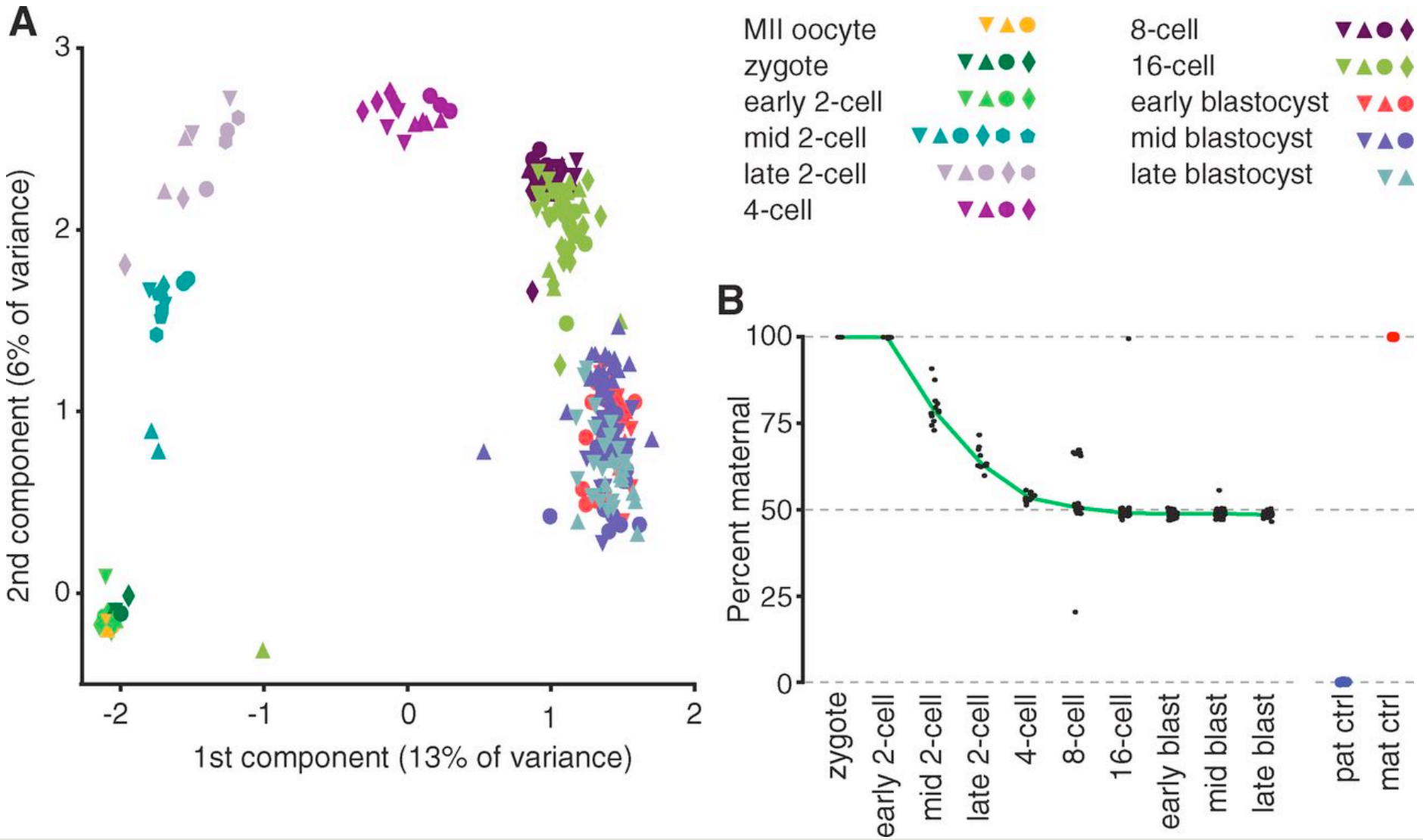
- Depending on cell type, around 500K exon mapping reads saturates the gene detection. Can be deduced from subsampling.

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



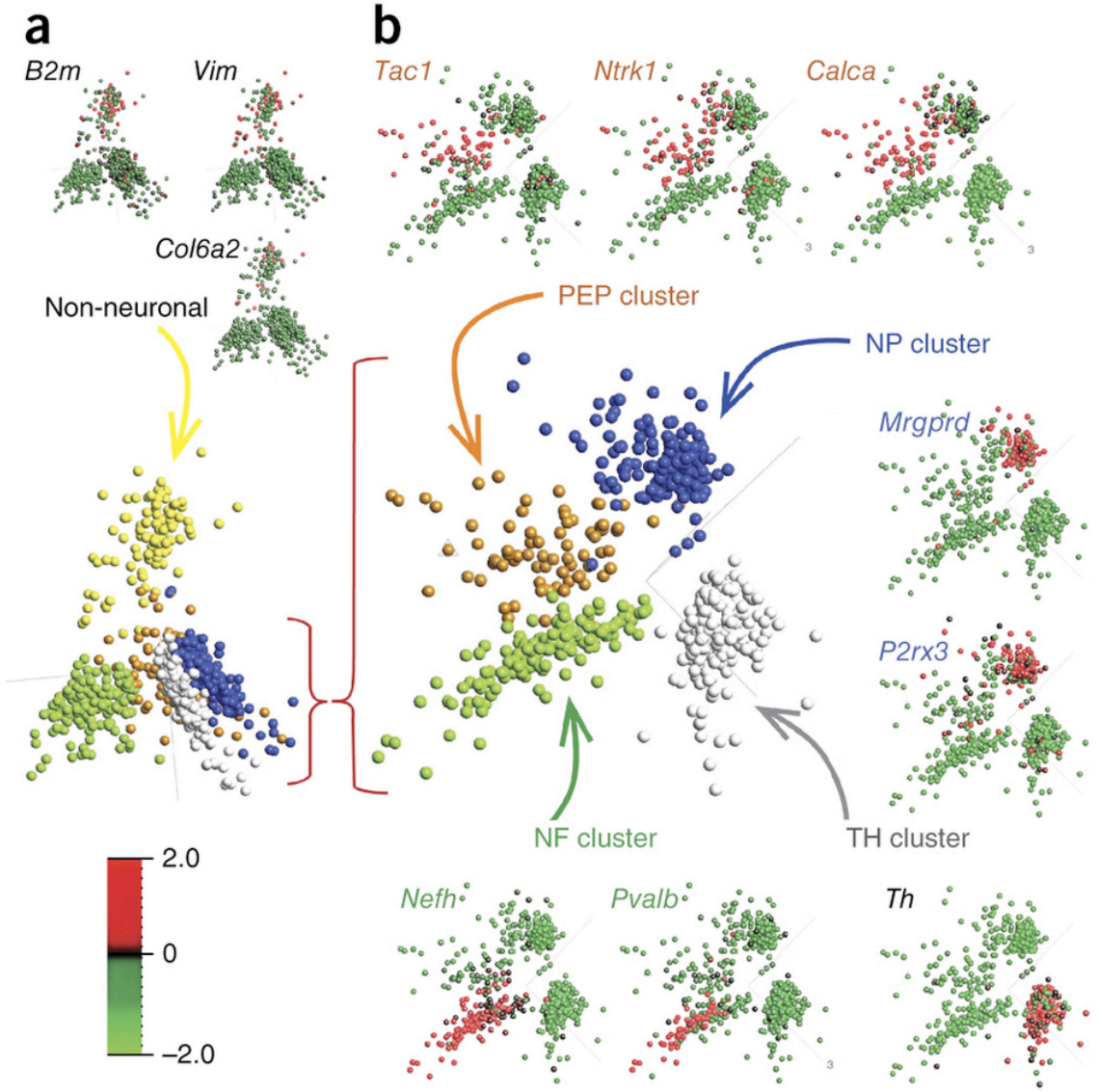
(Shalek et al. Nature 2014)

Identifying celltypes



(Deng et al. Science 2014)

Identifying celltypes



(Usoskin et al. Nat. neuroscience 2015)

Identifying celltypes – Dimensionality reduction

- Many different methods are used:
  - PCA (principal component analysis)
  - ICA (independent component analysis)
  - MDS (multidimensional scaling)
  - Non-linear PCA
  - t-SNE (t-distributed stochastic neighbor embedding)
  - Diffusion maps
  - Network based methods

Identifying celltypes - Clustering

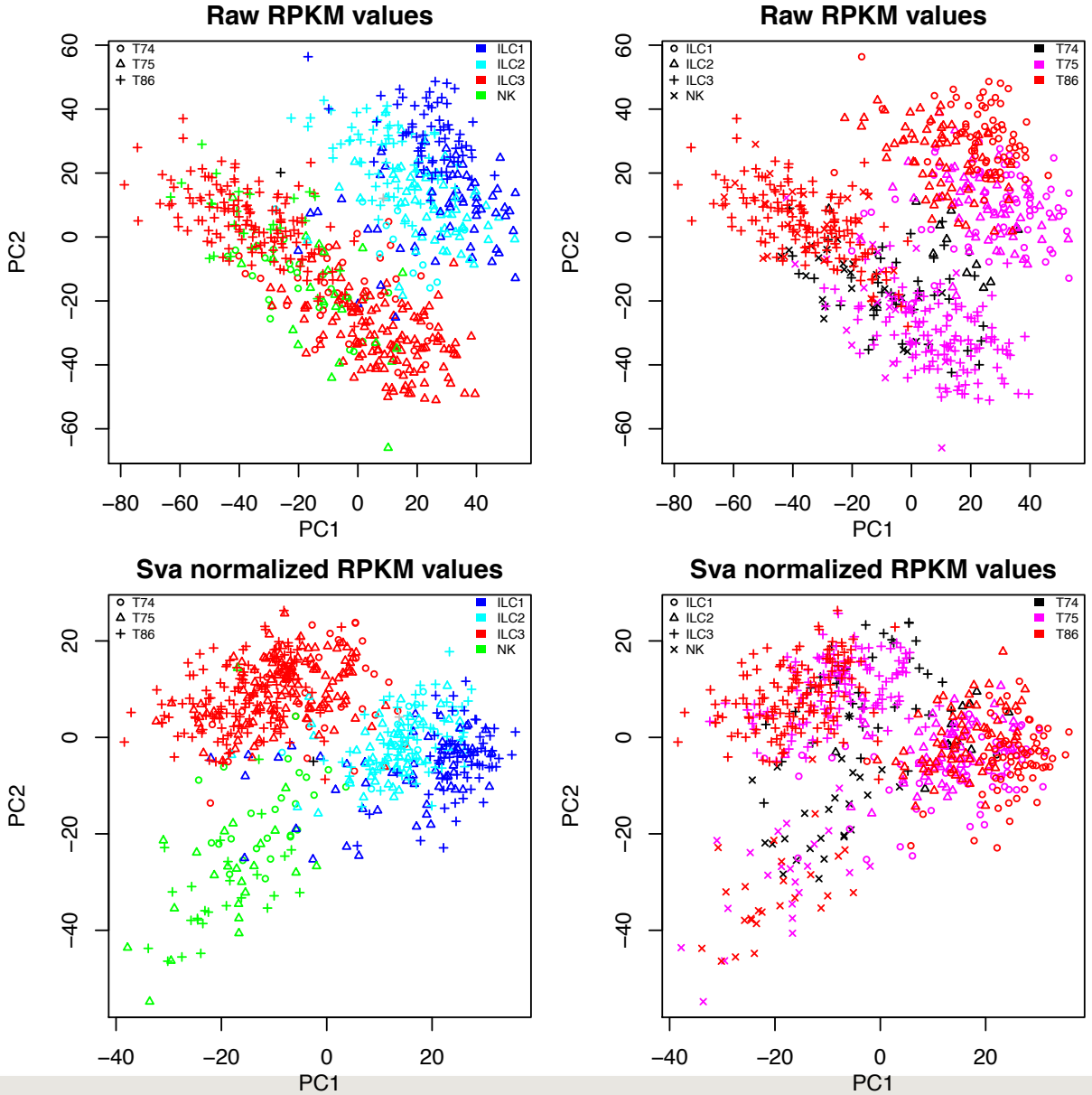
- Clustering based on
  - rpkms/counts
  - Correlations
  - PCA or other dimensionality reduction method
- Method of choice: hierarchical, k-means, biclustering
- Some programs:
  - WGCNA
  - BackSPIN
  - Pagoda
  - DBscan
- OBS! Outlier removal as an initial step may be necessary, especially with PCA-based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

Identifying celltypes – Data bias

- May require normalization before clustering/PCA,
  - Batch effect removal (SVA ComBat function)
  - Remove cell-cycle effects or size bias (scLVM package)
  - RT efficiency / drop-out rate (SCDE package)
  - Technical noise (BASiCS package, GRM, Brenneke method)

Batch normalization with SVA function ComBat

Color by celltype

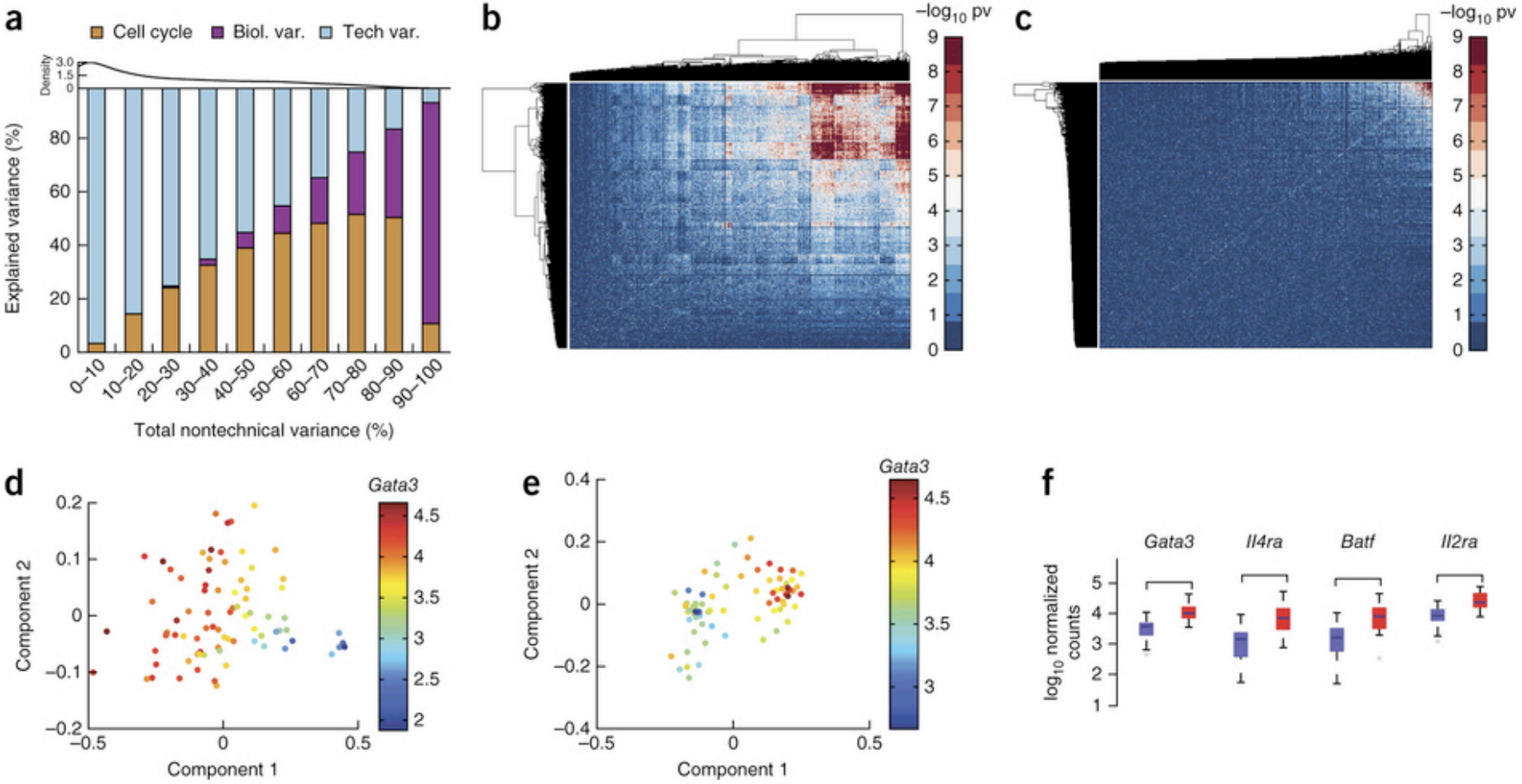


Color by donor

(Björklund et al. Nature Immunology 2016)

scLVM - Marioni lab

https://github.com/PMBio/scLVM)

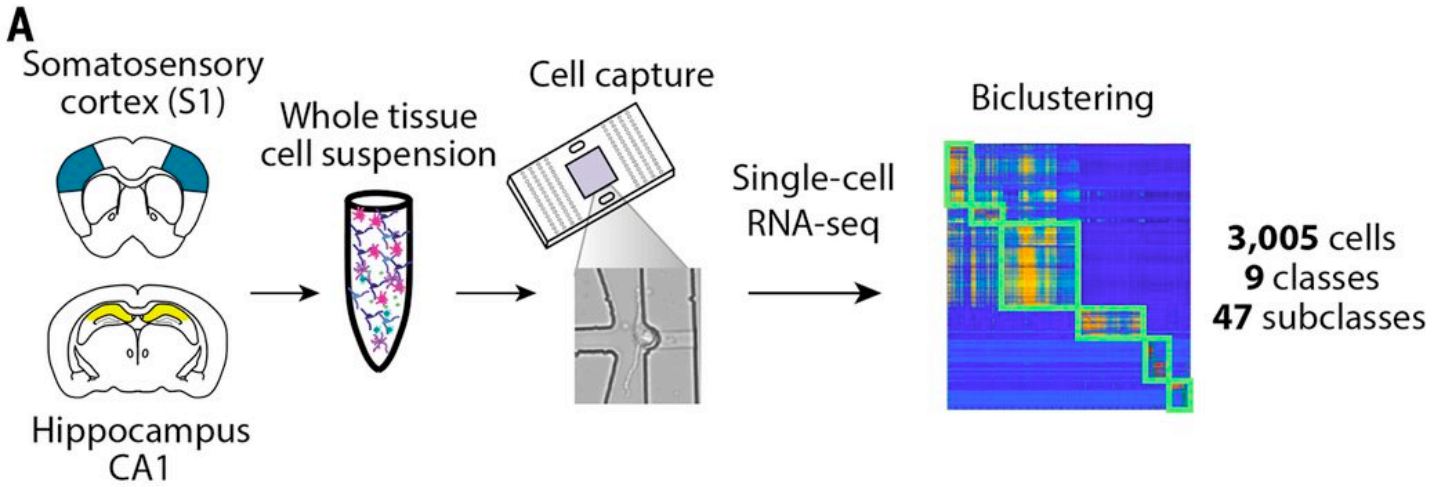


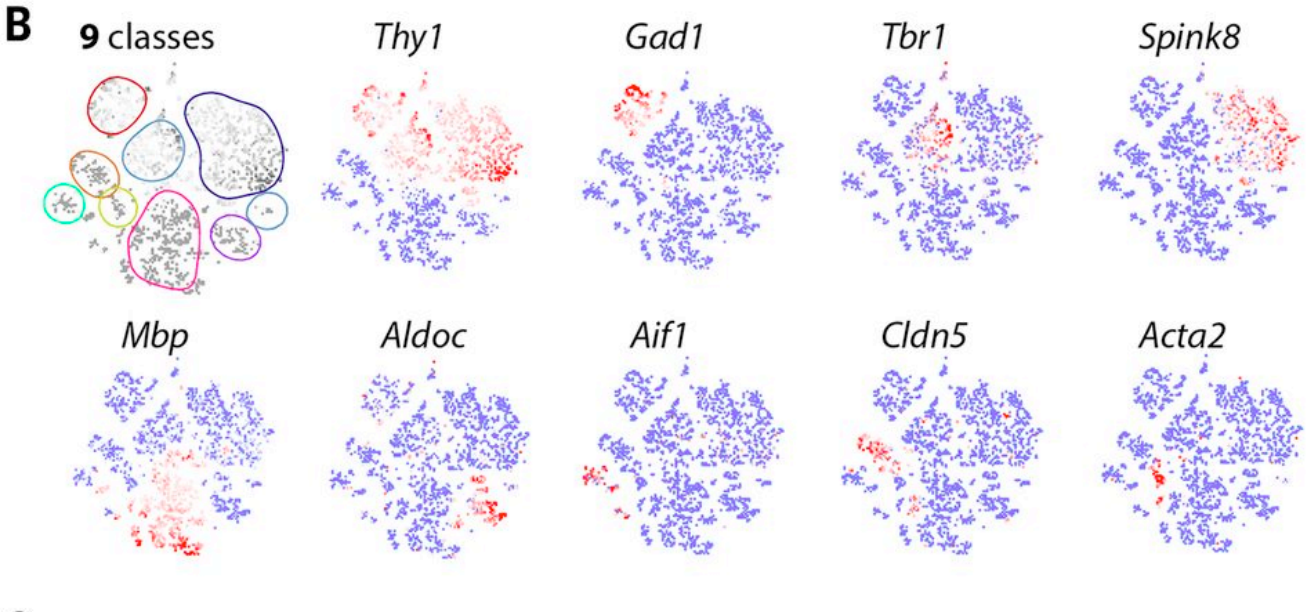
(Buettner et al. Nature Biotech 2015)

t-SNE – t-distributed stochastic neighbor embedding

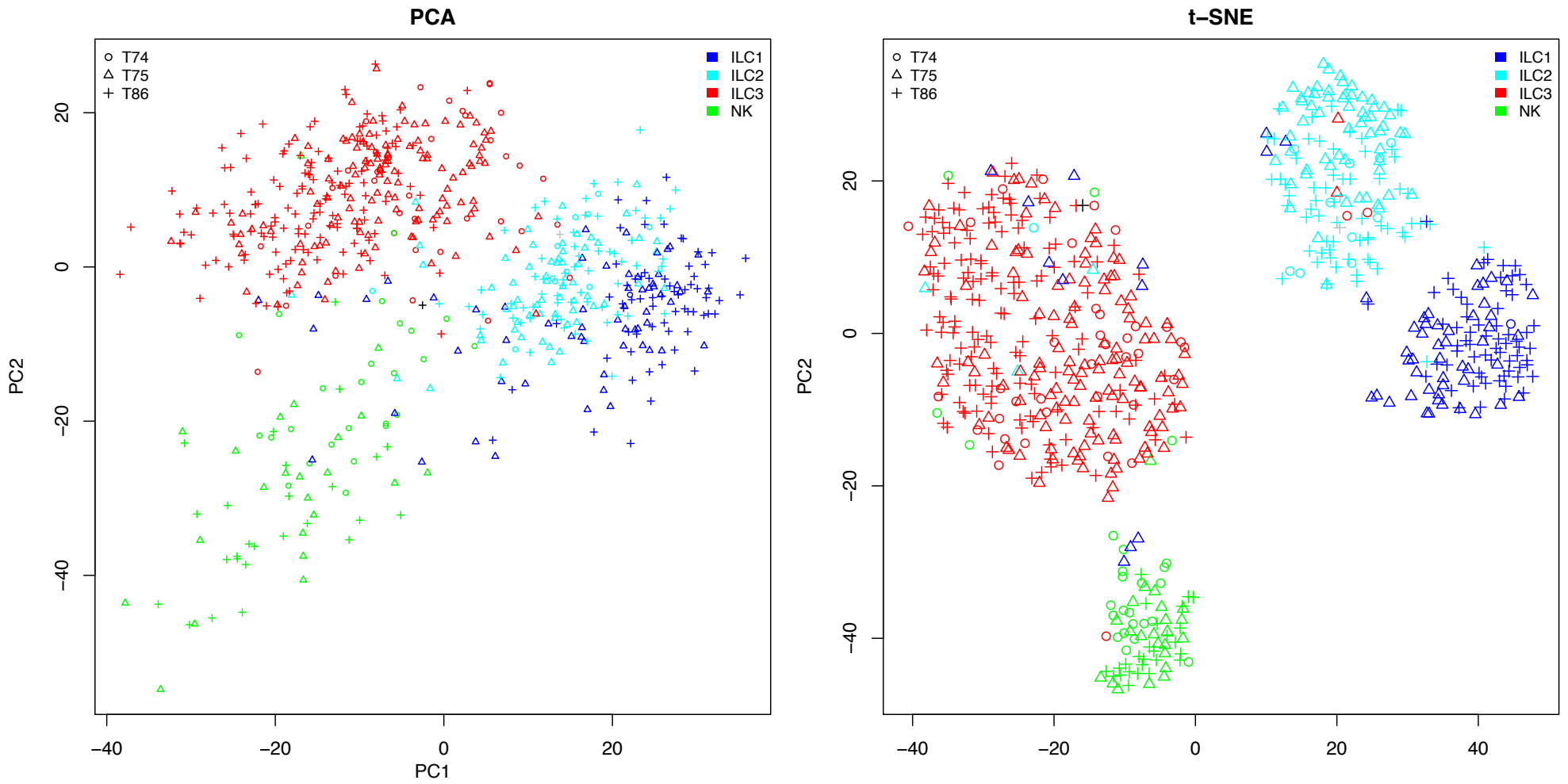
- Method often used in single cell proteomics.
- **Step 1** – probability distribution for all pairs in PCA space with N principal components
- **Step 2** – dimensionality reduction with similar probability distribution and minimization of divergence between distributions
- Implementations in R:
  - tsne
  - Rtsne (Barnes-Hut t-SNE)
- For other languages (python, java, matlab, C++ etc.):
  - http://lvdmaaten.github.io/tsne/

t-SNE – t-distributed stochastic neighbor embedding





t-SNE vs PCA dimensionality reduction



(Björklund et al. Nature Immunology 2016)

Preselection of a gene set

- In most cases, all genes are not used in PCA/ clustering.
- Filtering 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

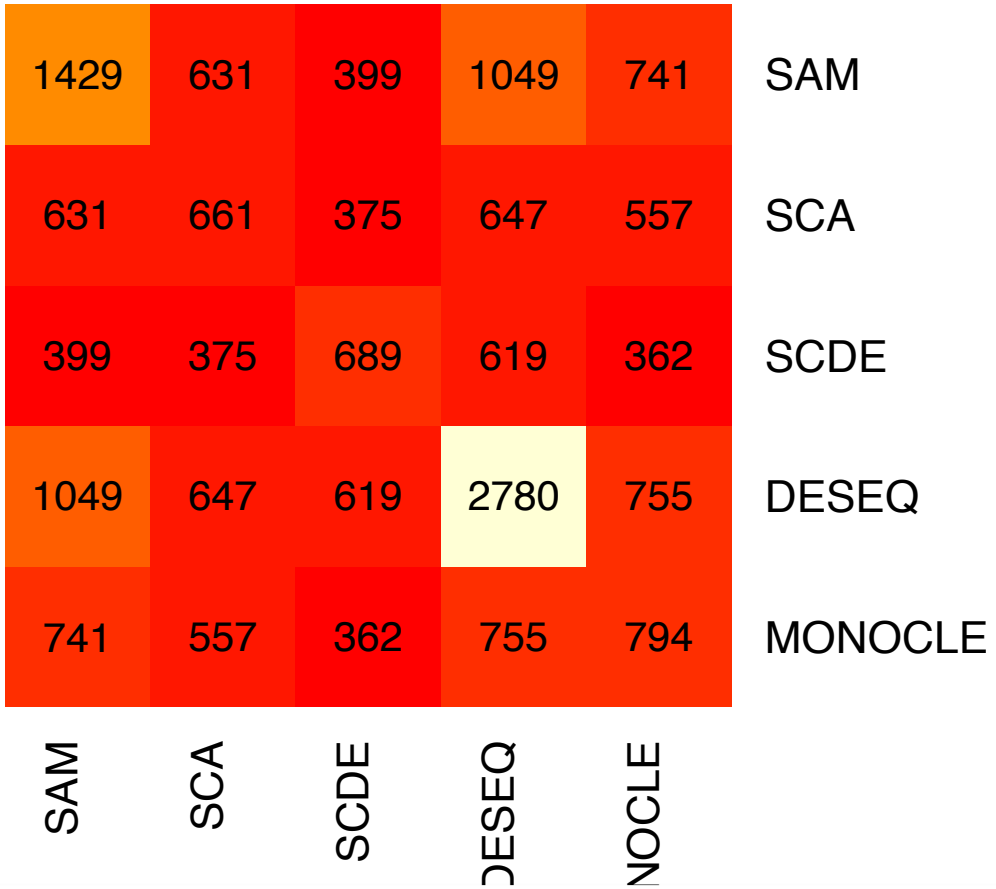
Detecting differentially expressed genes

- Parametric methods like EdgeR & DESeq not suitable for scRNAseq since the parameter assumptions in those methods does not apply here.
- Simple fisher or chi square test works better in most cases
- Or non-parametric methods like SAMseq

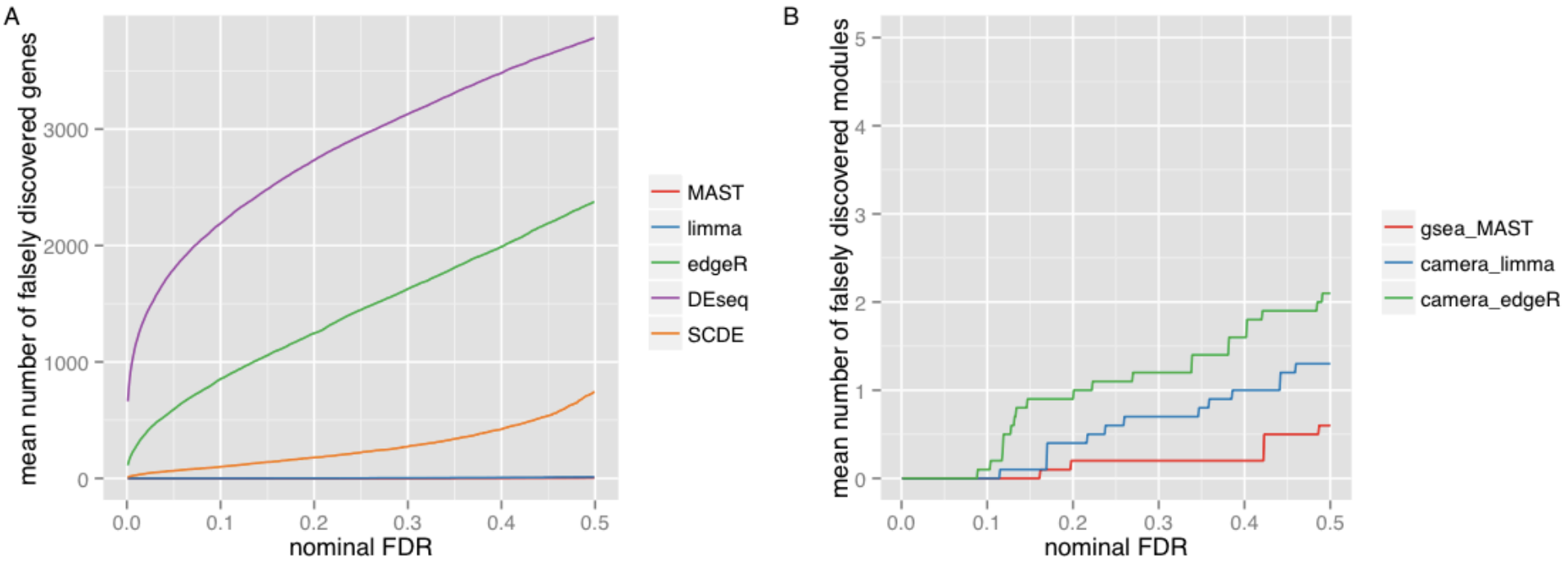
Detecting differentially expressed genes

- 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



High false discovery rate For DESeq and EdgeR

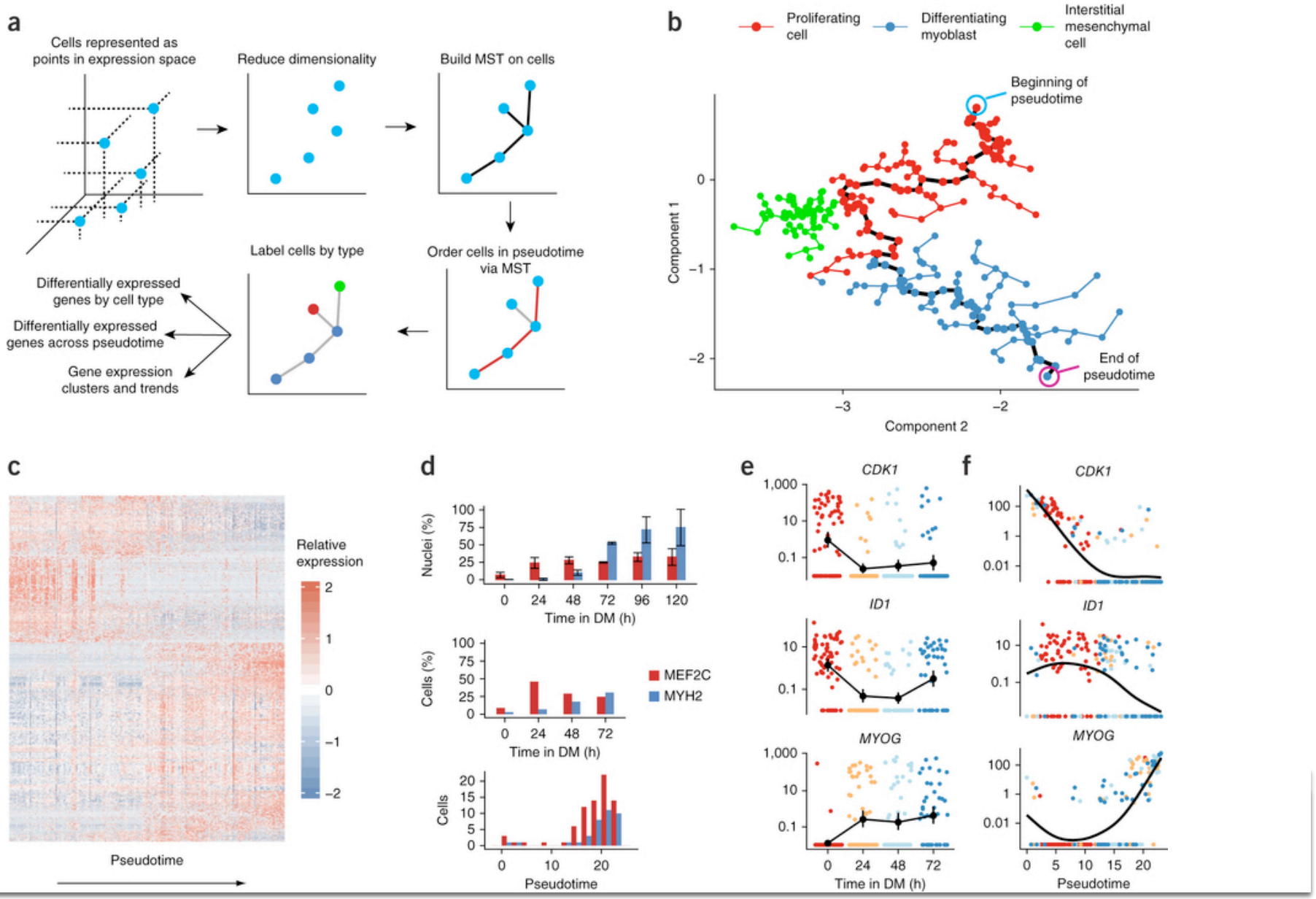


Detected DE genes and gene sets using randomly permuted cells from unstimulated MAIT cells

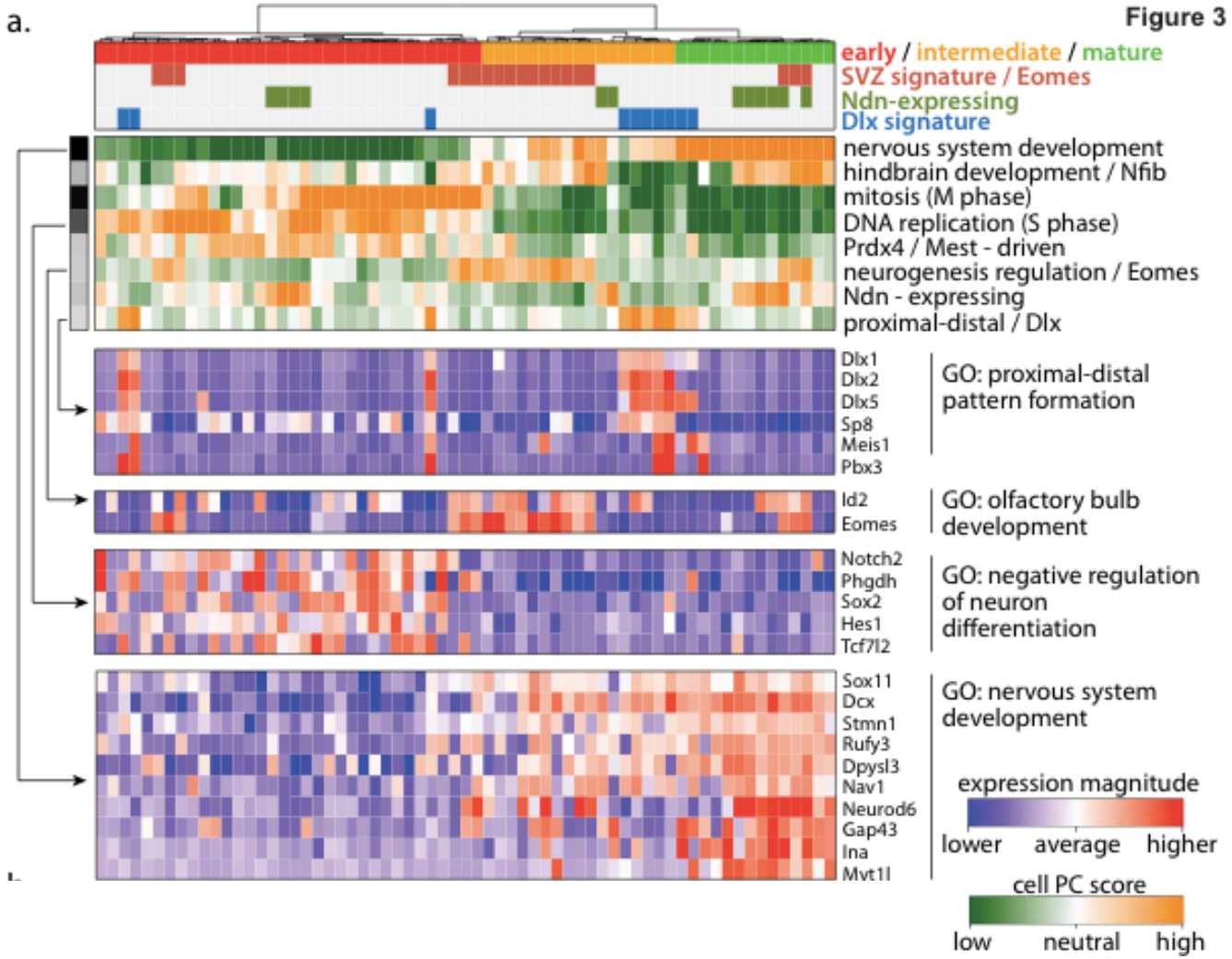
Pseudotime ordering - Monocle

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

Trapnell et al. *Nature Biotechnology* 32, 381–386 (2014)



Pagoda – Pathway And Geneset OverDispersion Analysis

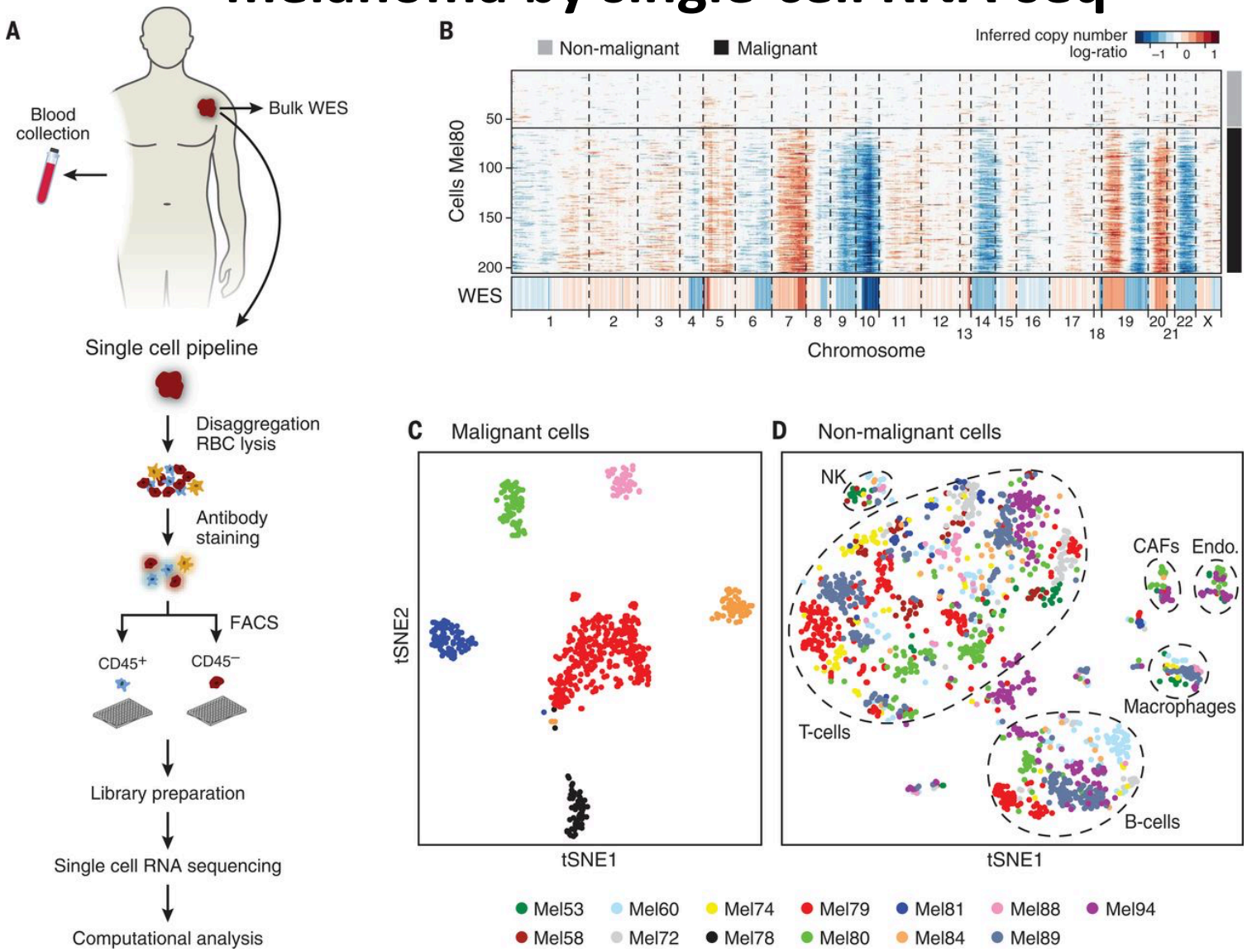


(Fan et al. Nature Methods 2016)

Additional analyses

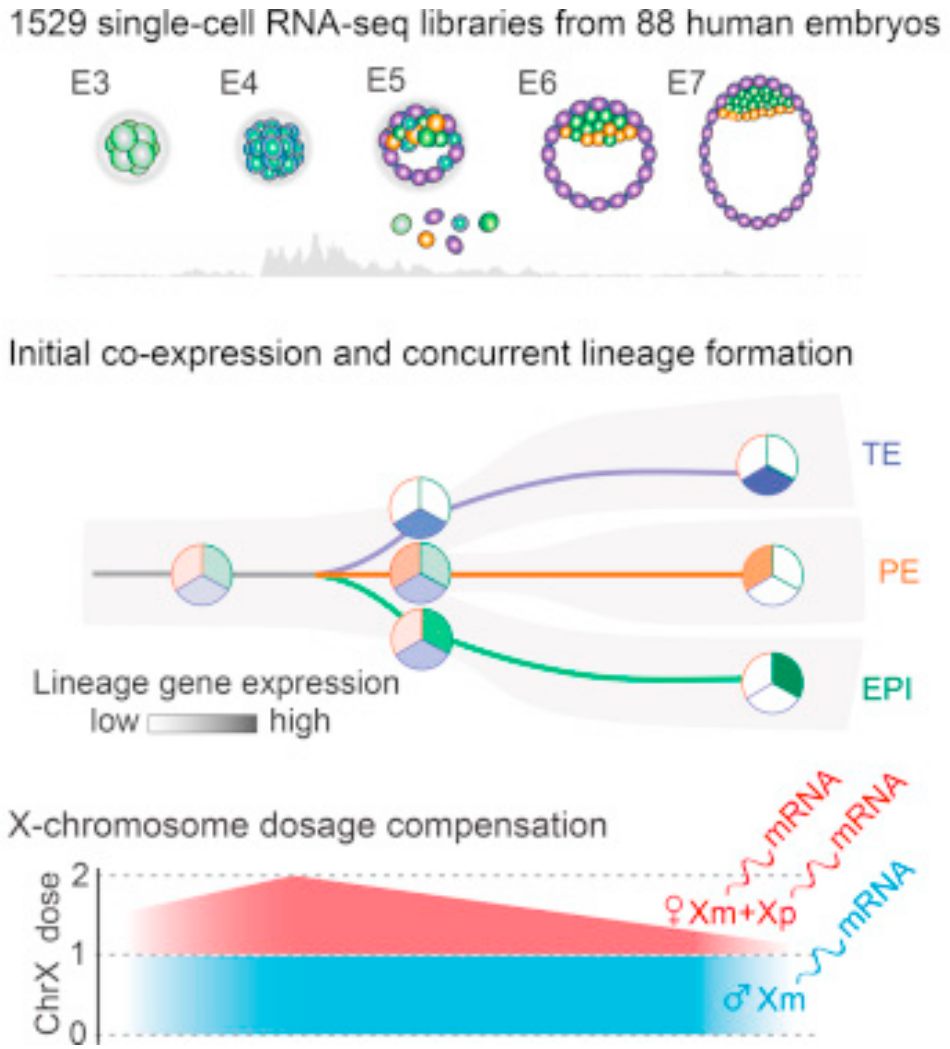
- Alternative splicing
- Allelic expression
- Copy-number variation
- Alternative splicing and allelic expression requires full length methods.
  - But only for highly expressed genes with good read coverage
  - Must be careful to take into consideration the drop-out rate, a unique splice form/allele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

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



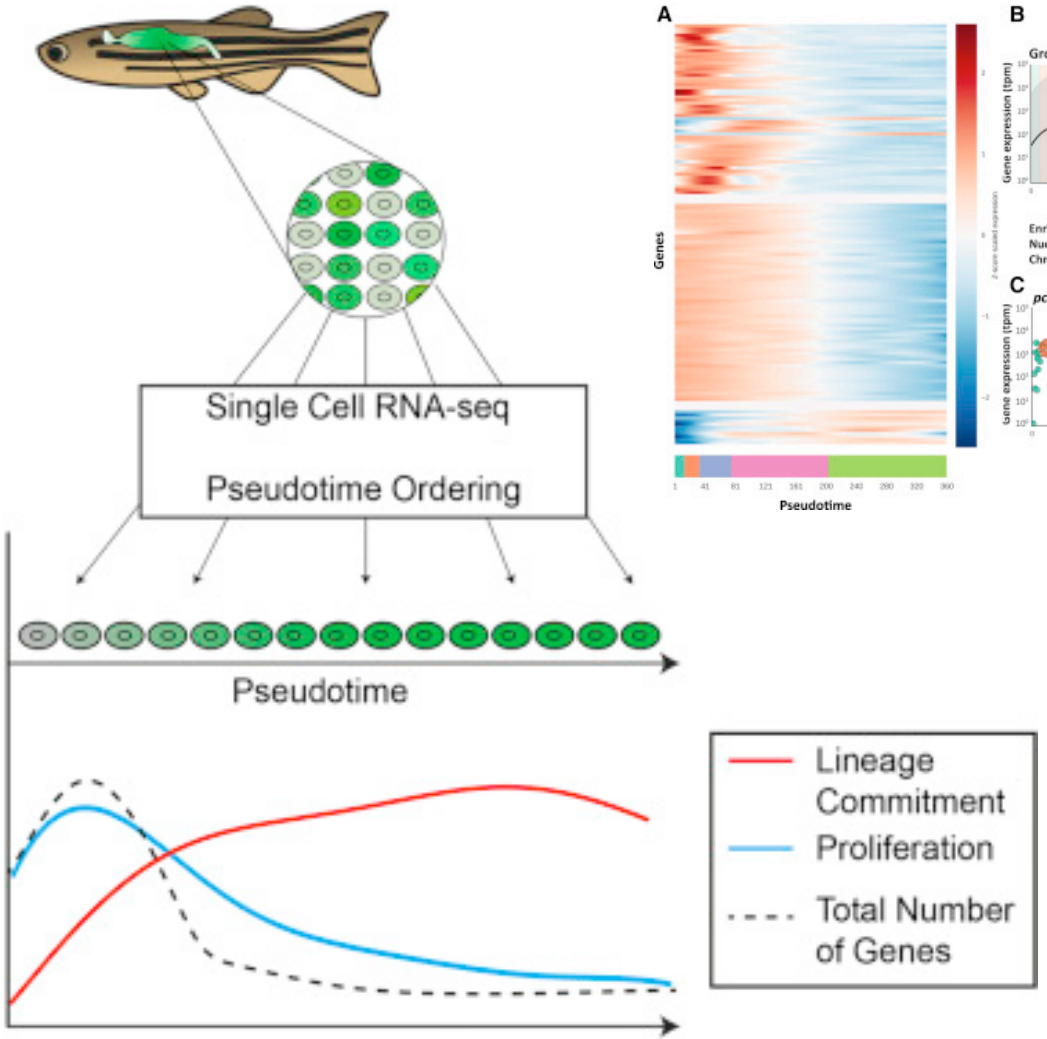
(Tirosh et al. Science 2016)

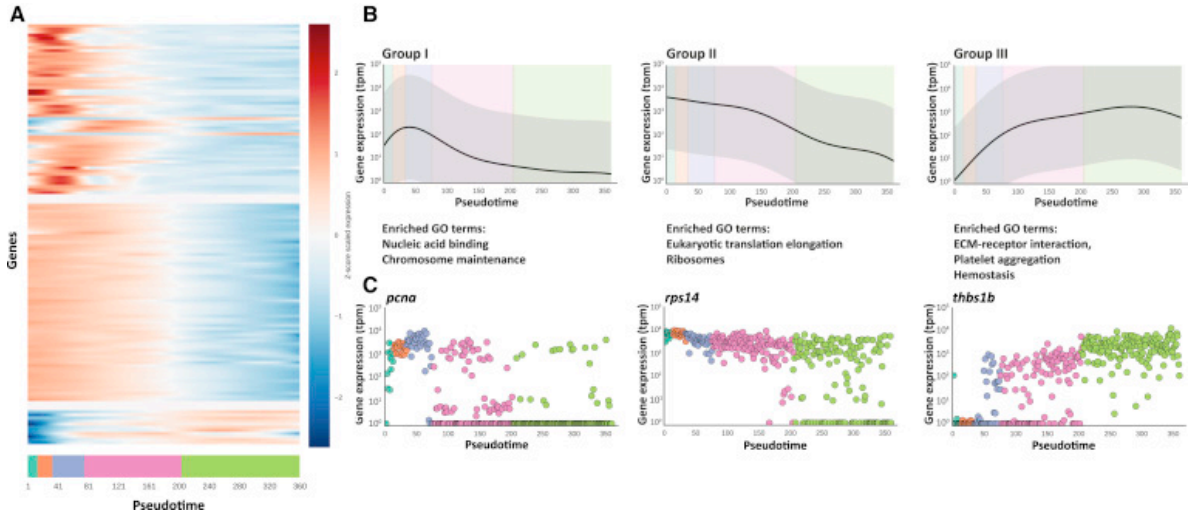
Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos



(Petropolus et al. Cell 2016)

Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells





(Macaulay et al. Cell Reports 2016)

Summary

- For diverse cell-types often straight forward to group cells into clusters and detect differentially expressed genes.
- For highly similar cells, or with subtle changes in cellular status – feature selection and different clustering methods may be required.

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 etc.:
  - Monocle: 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