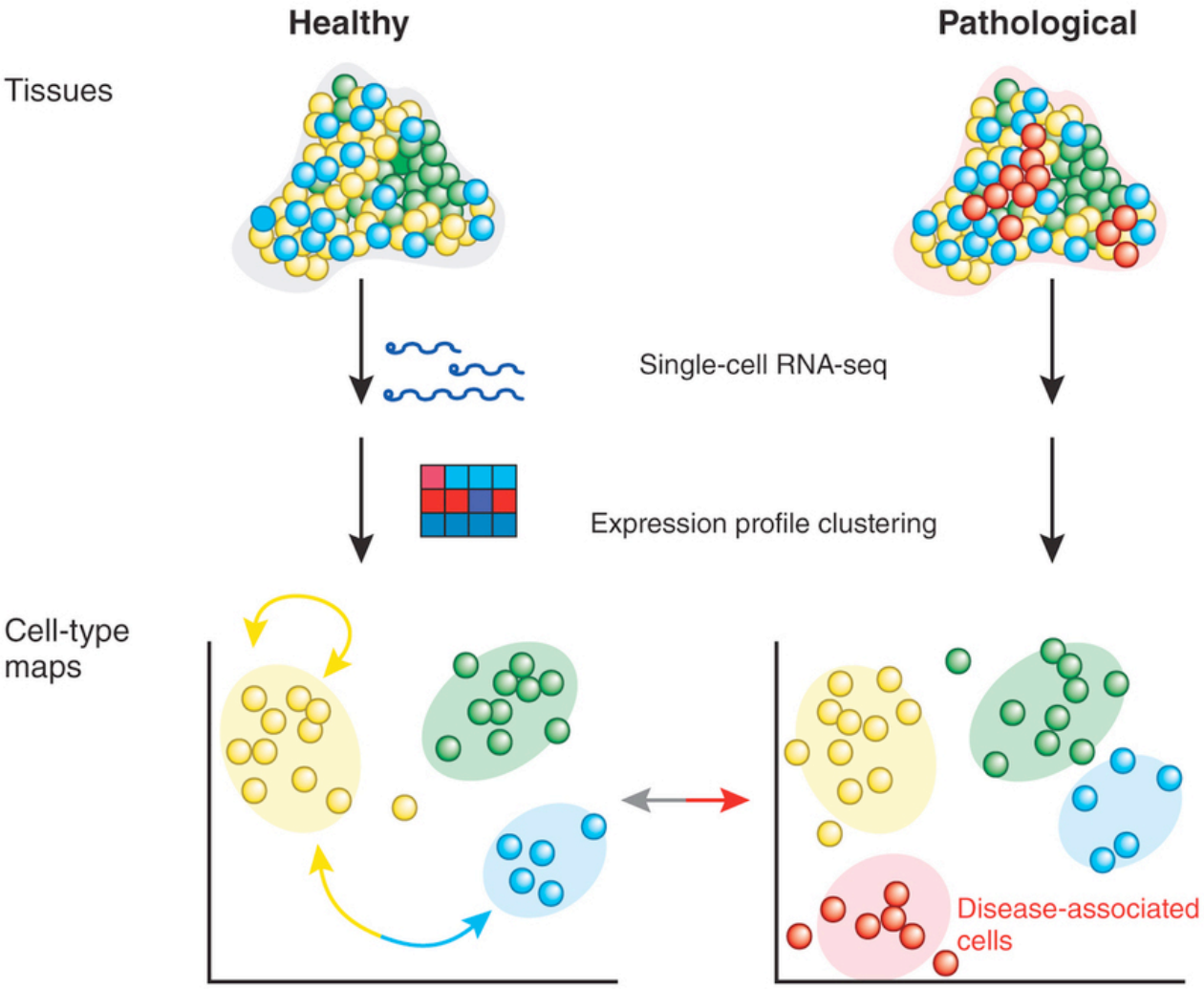
Single cell RNA sequencing

Johan Reimegård

johan.reimegard@scilifelab.se

Outline

- Why single cell gene expression?
- Library preparation methods
- Setting up experiments
- Defining cell types
- Identifying differentially expressed genes



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?

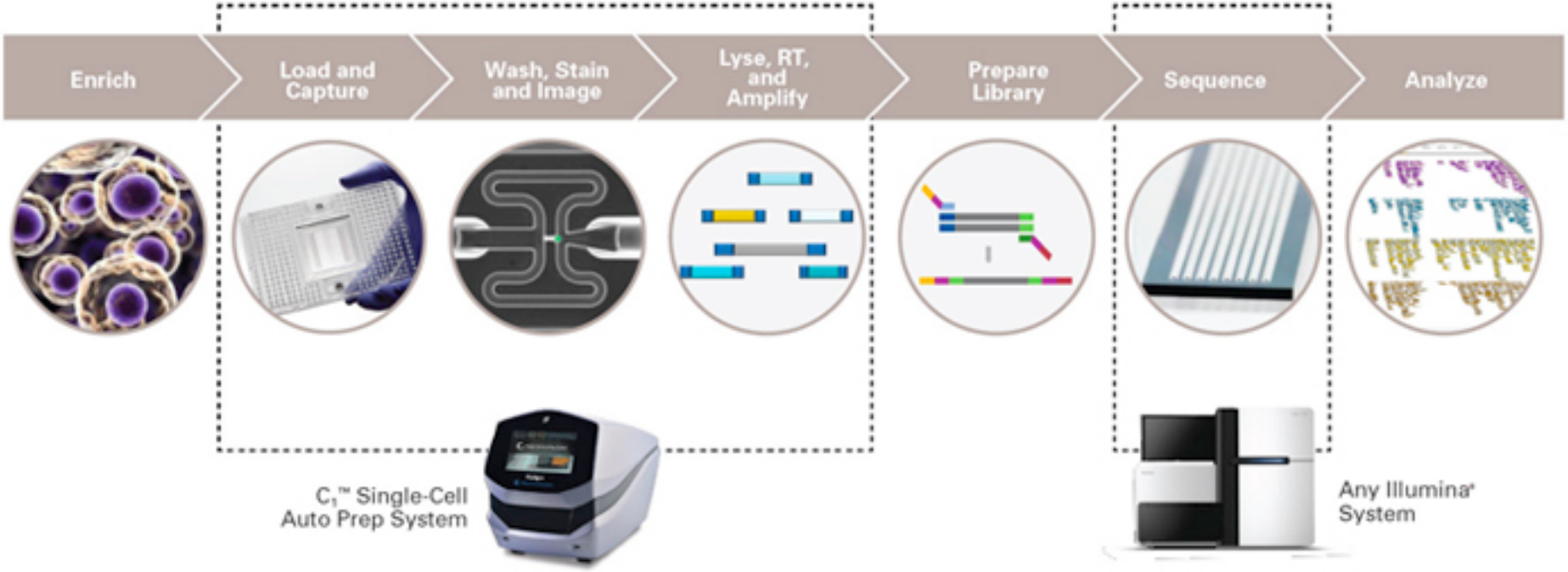
- Understanding heterogeneous tissues
- Identification and analysis of rare cell types
- Changes in cellular composition
- Dissection of temporal changes
- Example of applications:
  - Differentiation paths
  - Cancer heterogeneity
  - Neural cell definition
  - Embryonic development
  - Drug treatment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

In order to do single cell sequencing you first have to seperate the cells

- Depending on the cell types this can be more or less complicated.
  - Blood relatively easy since cells are already seperated
  - Brain much more complicated since the neurons are intertwined.
  - Also some tissues stick together and are not easy to separate

Different ways of isolating single cells

- FACS sorting
- Manual picking
- Fluidigm C1 system
- Dissociation of cells is a crucial step to minimize leakage and RNA degradation
- Celltypes that are hard to dissociate:
  - Laser capture microscopy (LCM)
  - Nuclei sequencing



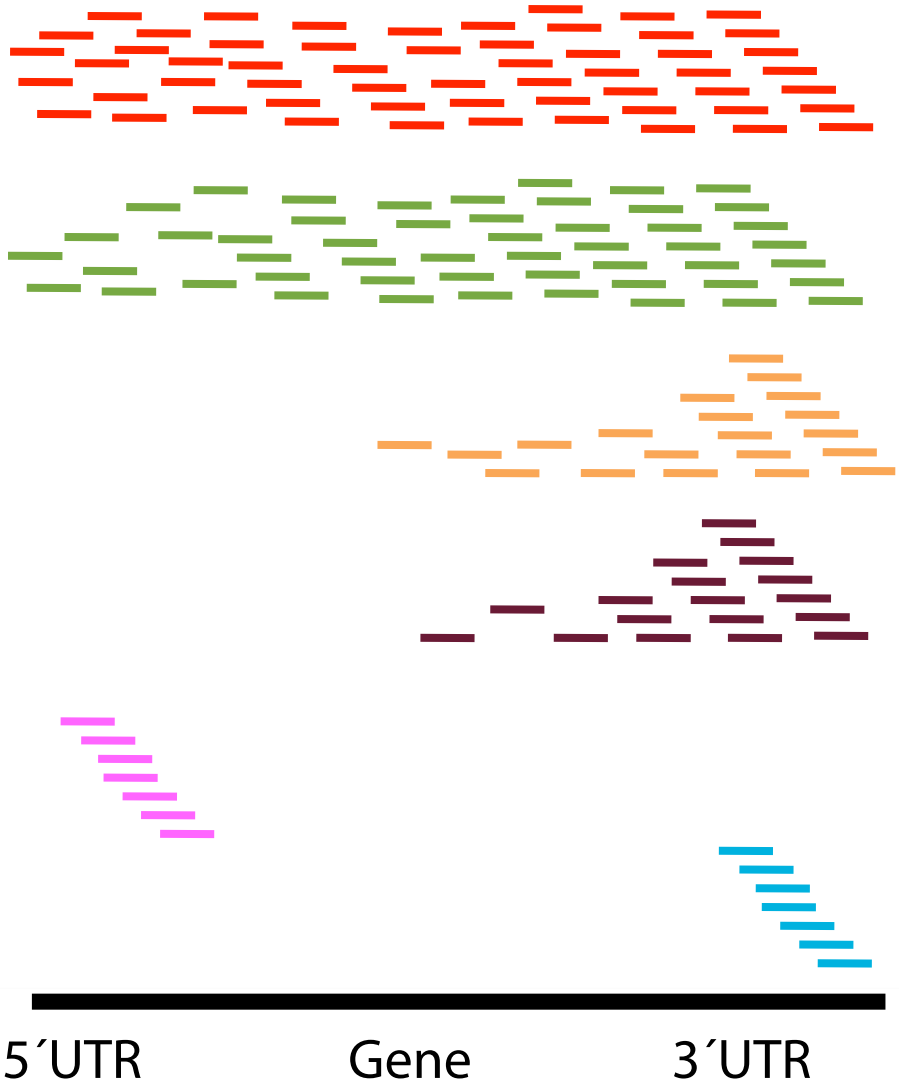
- Fluidigm C1 system
  - Limiting factor is size of capture chambers
  - Have protocols for running SMARTer, SmartSeq2, CEL-seq & STRT

(https://www.fluidigm.com)

Different ways of sequencing single cells.

- Lots of methods available for single cell sequencing
  - Methods for sequencing up to 1000 of cells
    - Different enrichment protocolls for the different methods on what they are sequencing
      - Full length RNAs (SmartSeq2)
      - 5’ end enrichment (STRT)
      - 3’ end enrichment (CEL-seq)
      - Smartseq2 and STRT developed at Karolinska Institute
      - All three protocols available at the Scilifelab
  - Methods available for sequencing up to 50 000 cells
    - Drop seq

scRNA seq methods



SmartSeq2
(Picelli et al. Nature Methods 2014)

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

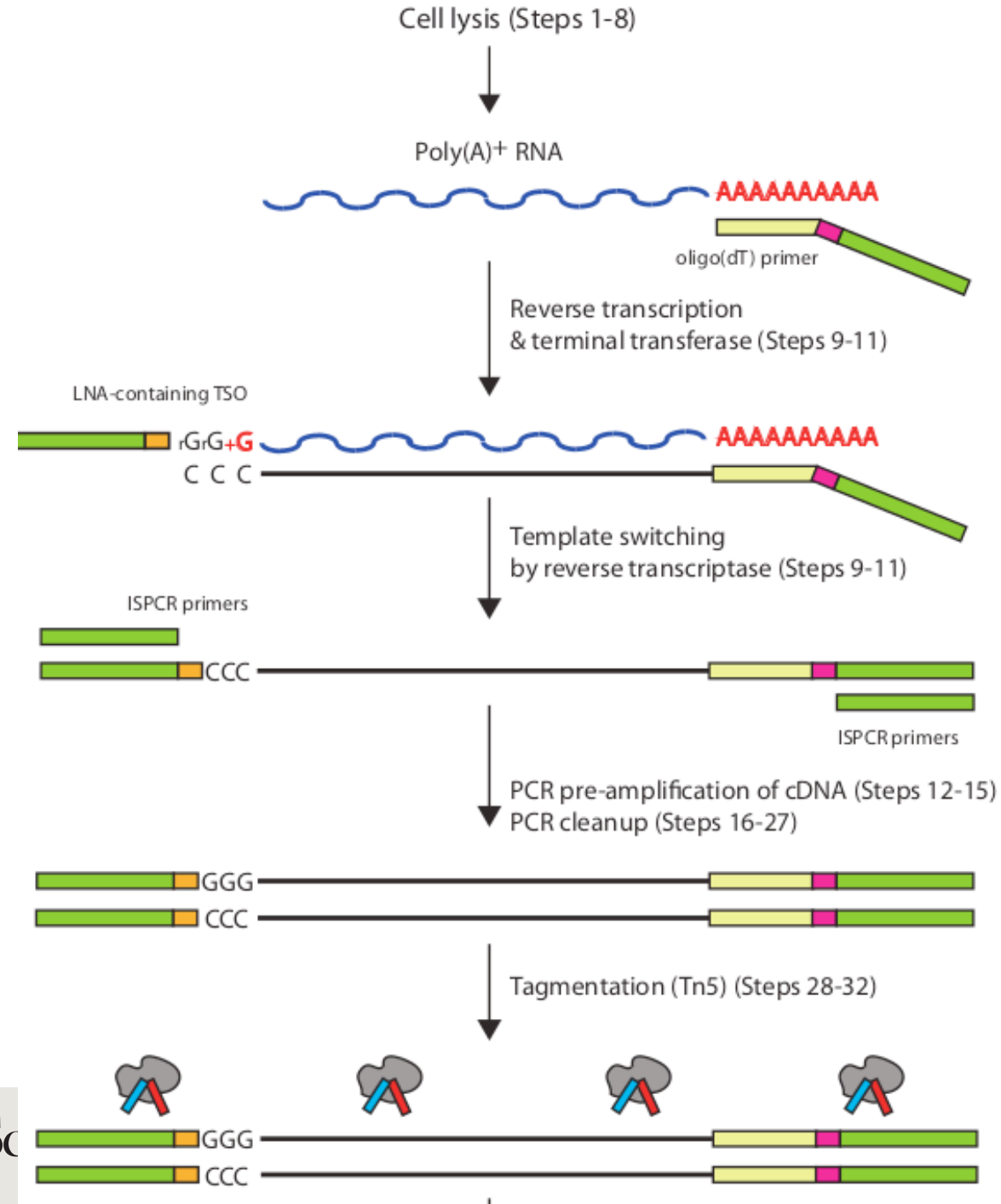
Quartz-seq
(Sasagawa et al. Genome Biology 2013)

Tang et al.
(Nature methods 2009)

STRT
(Islam et al. Genome Res 2011)

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

SmartSeq2 protocol



(Picelli et al. Nature Protocol, 2014)

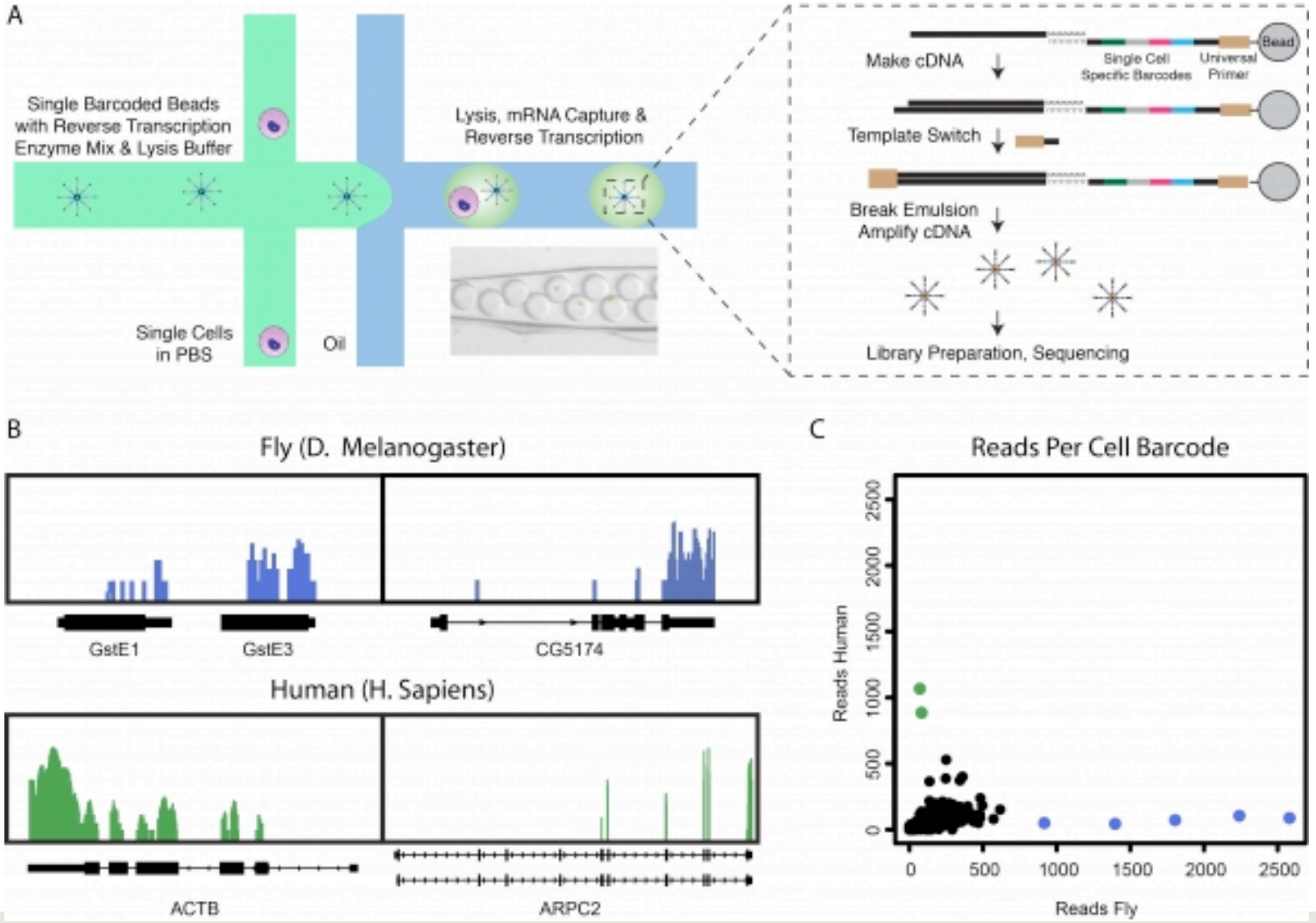
National single cell genomics platform at Scilifelab

- 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

Small volume approaches

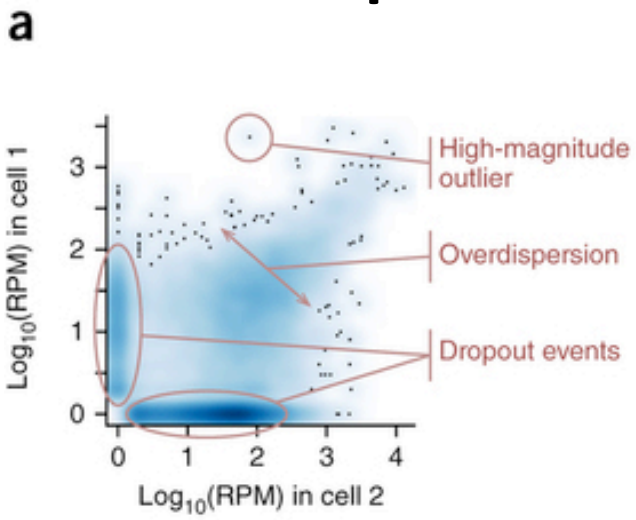
- Volume seem to be a key component in these reactions
  - Smaller volumes give higher detection and better reproducibility
- Smaller volumes = cheaper reagent costs
- Methods for high throughput (1000nds of cells)
- Sequencing cost becomes the bottleneck instead

Drop-seq – microfluidics approach that can produce thousands of single cell samples



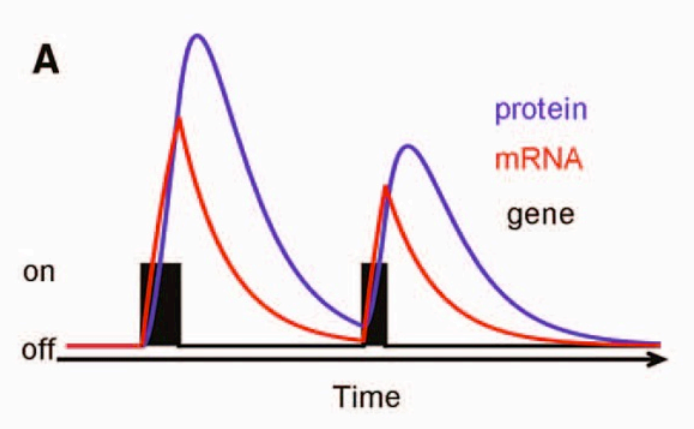
Problems compared to bulk RNA-seq

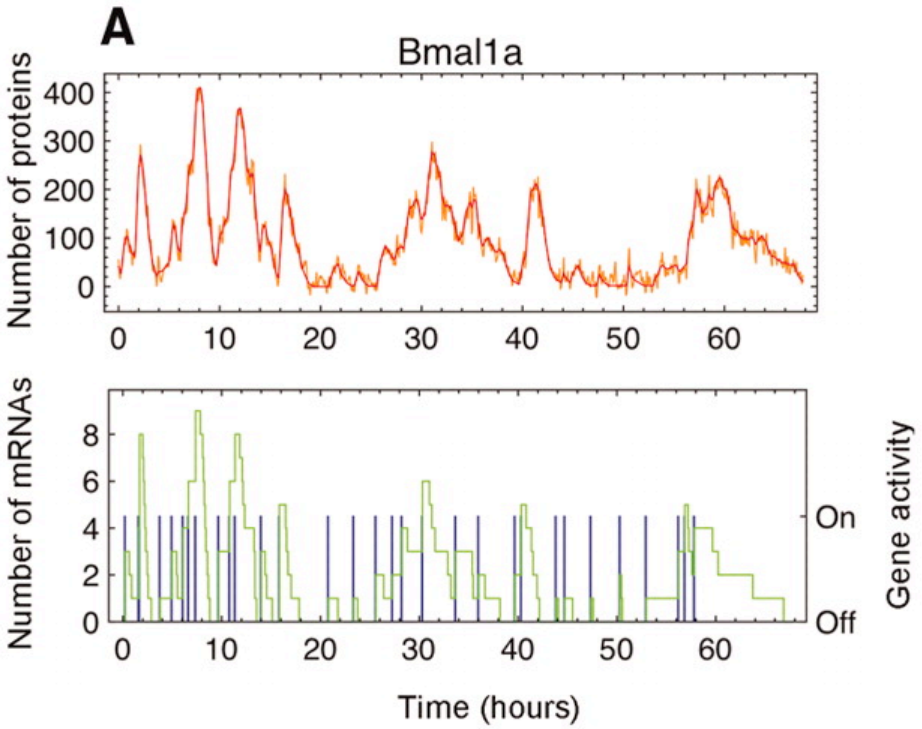
- Amplification bias
- Drop-out rates
- Stochastic gene expression
- Background noise
- Bias due to cell-cycle, cell size and other factors
- As of now, only polyA transcripts, no method for total RNA sequencing in single cells



(Karchenko et al. *Nature Methods* 2014)

Transcriptional bursting





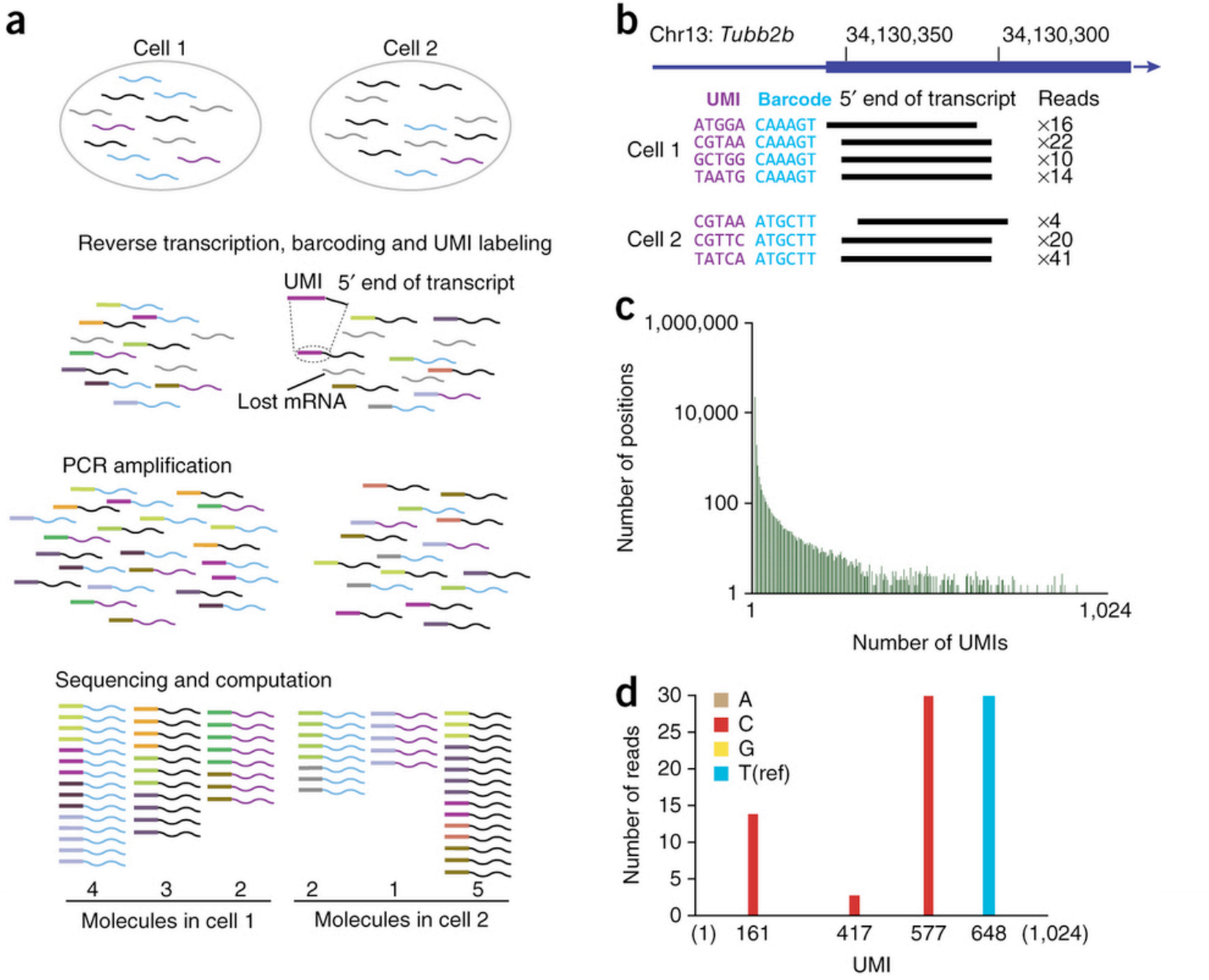
- Burst frequency and size seems to be correlated with expression level
- Many TFs have low mean expression (and low burst frequency) and will only be detected in a fraction of the cells

(Suter et al., Science 2011)

Unique molecular identifiers (UMIs) and cellular barcodes can address amplification biases

- 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

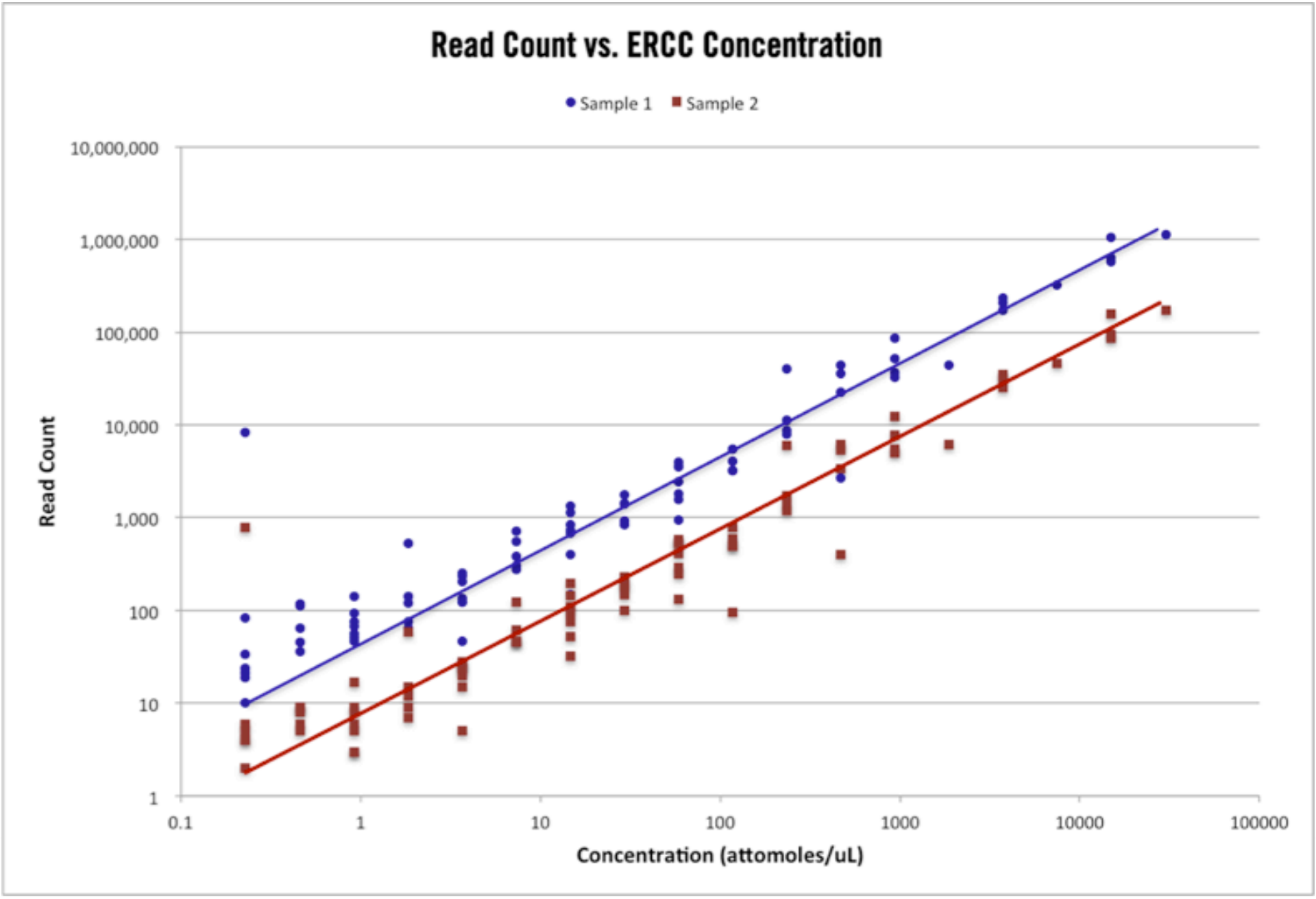
Unique molecular identifiers (UMIs) and cellular barcodes



(Islam et al. Nature Methods 2014)

Spike-in RNAs can address technical issues

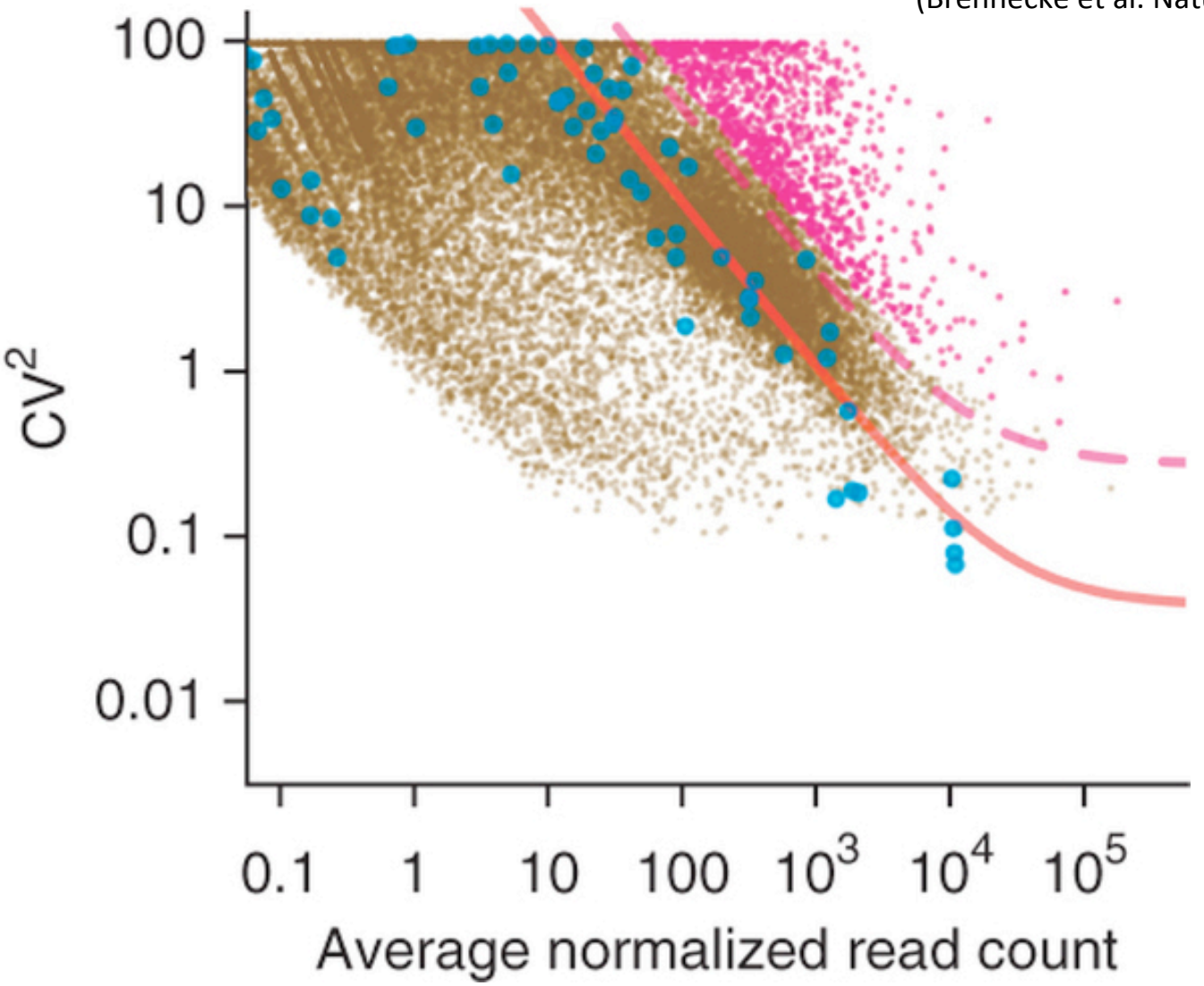
- Addition of external controls
- Used to model:
  - technical noise
  - drop-out rates
  - starting amount of RNA in the cell
- ERCC spike-in most widely used, consists of 48 or 96 mRNAs at 17 different concentrations.
- Add a ratio of about 1:10 to cell RNA.
- Important to add equal amounts to each cell, preferably in the lysis buffer.



(https://cofactorgenomics.com)

Finding biologically variable genes

(Brennecke et al. Nature Methods 2013)

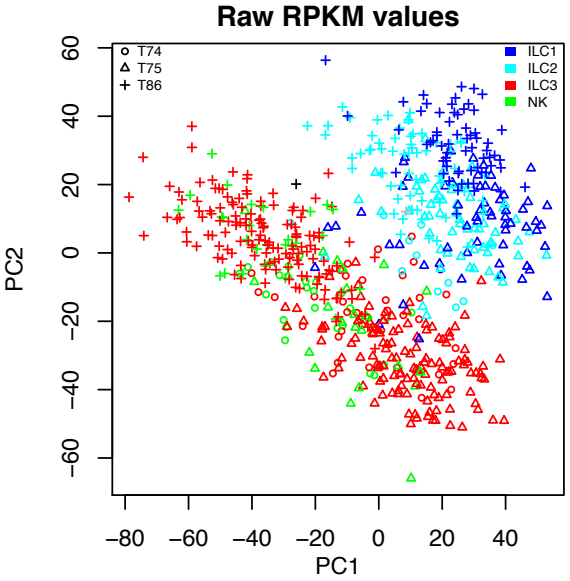


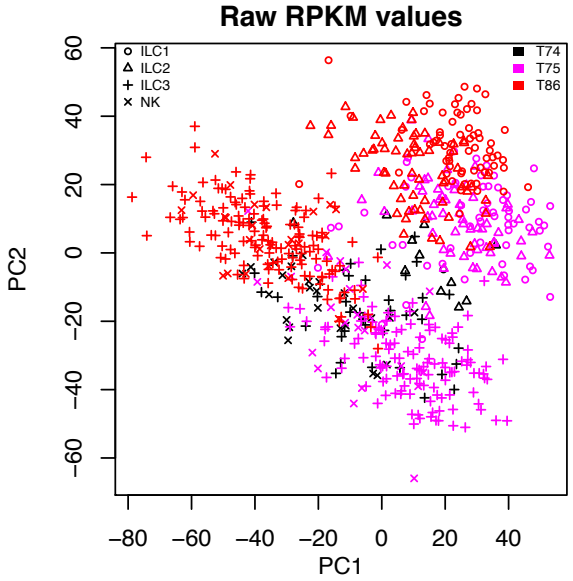
Downstream analysis can identify data bias

- May require normalization before clustering/PCA,
  - Batch effect removal (SVA ComBat function)
  - Remove cell-cycle effects or size bias (scLVM package)

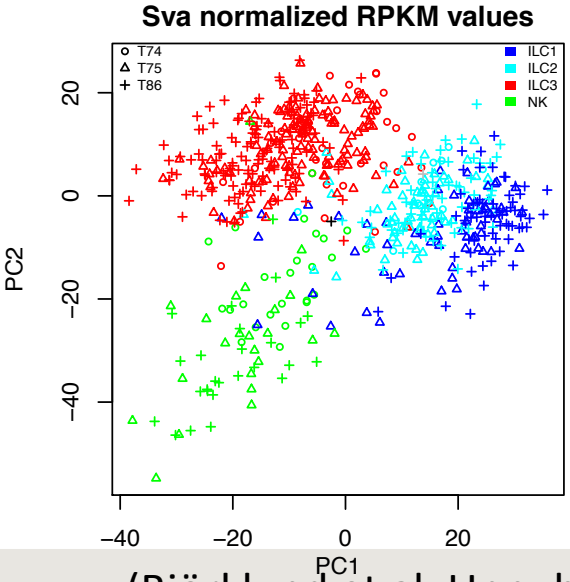
Batch normalization with SVA function ComBat

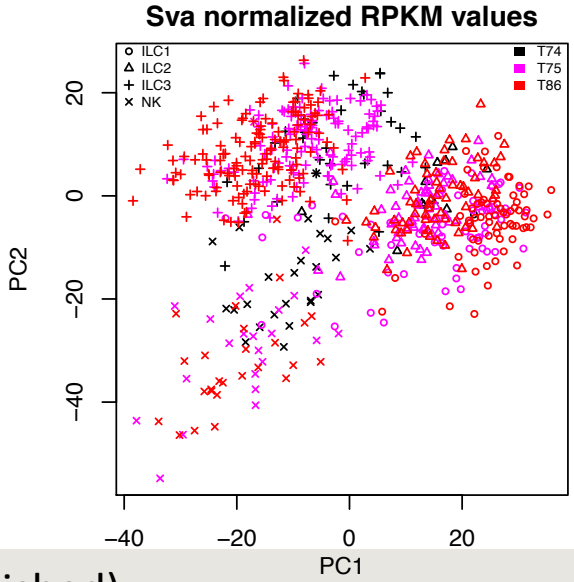
Color by celltype





Color by donor

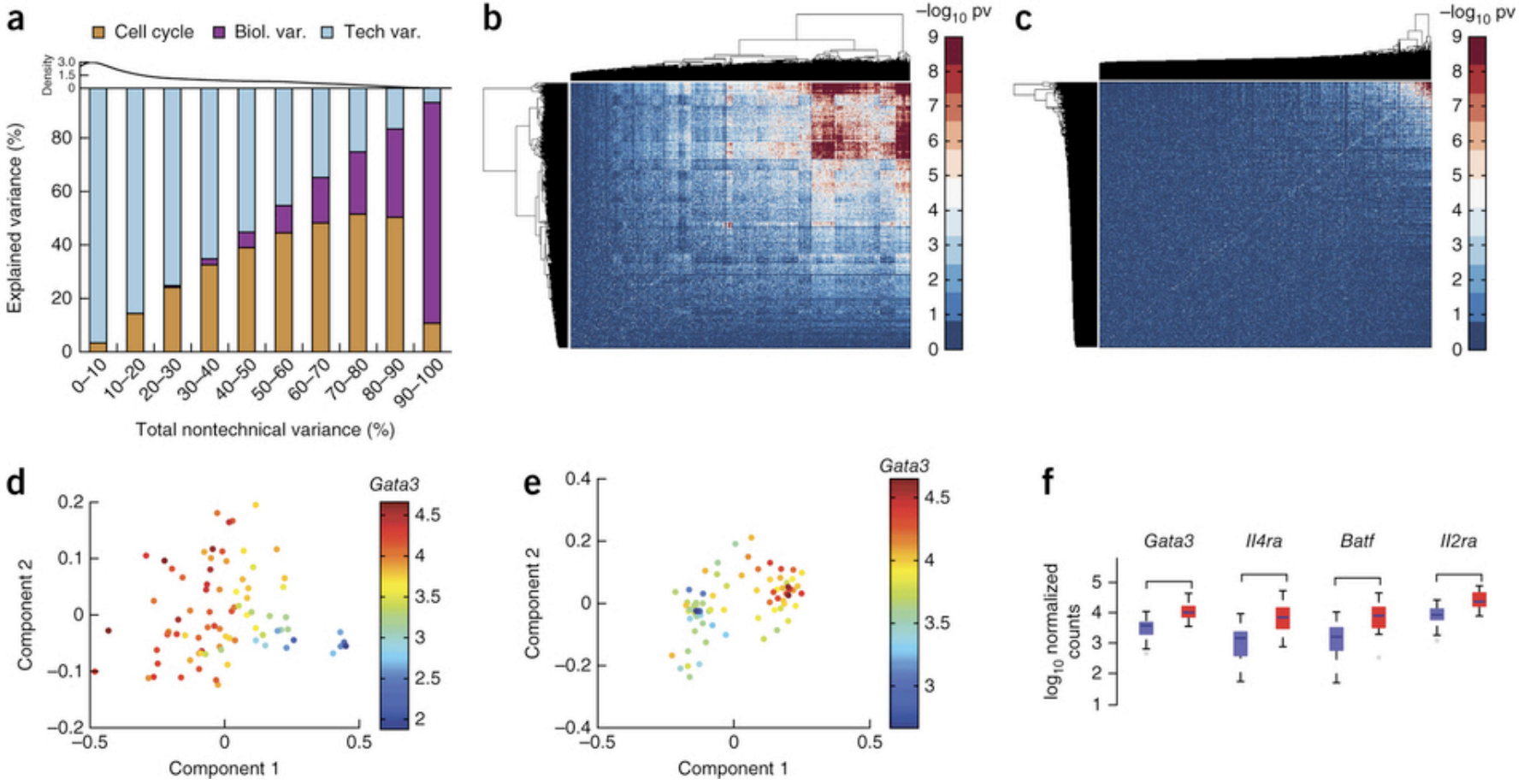




(Björklund et al. Unpublished)

scLVM - Marioni lab

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

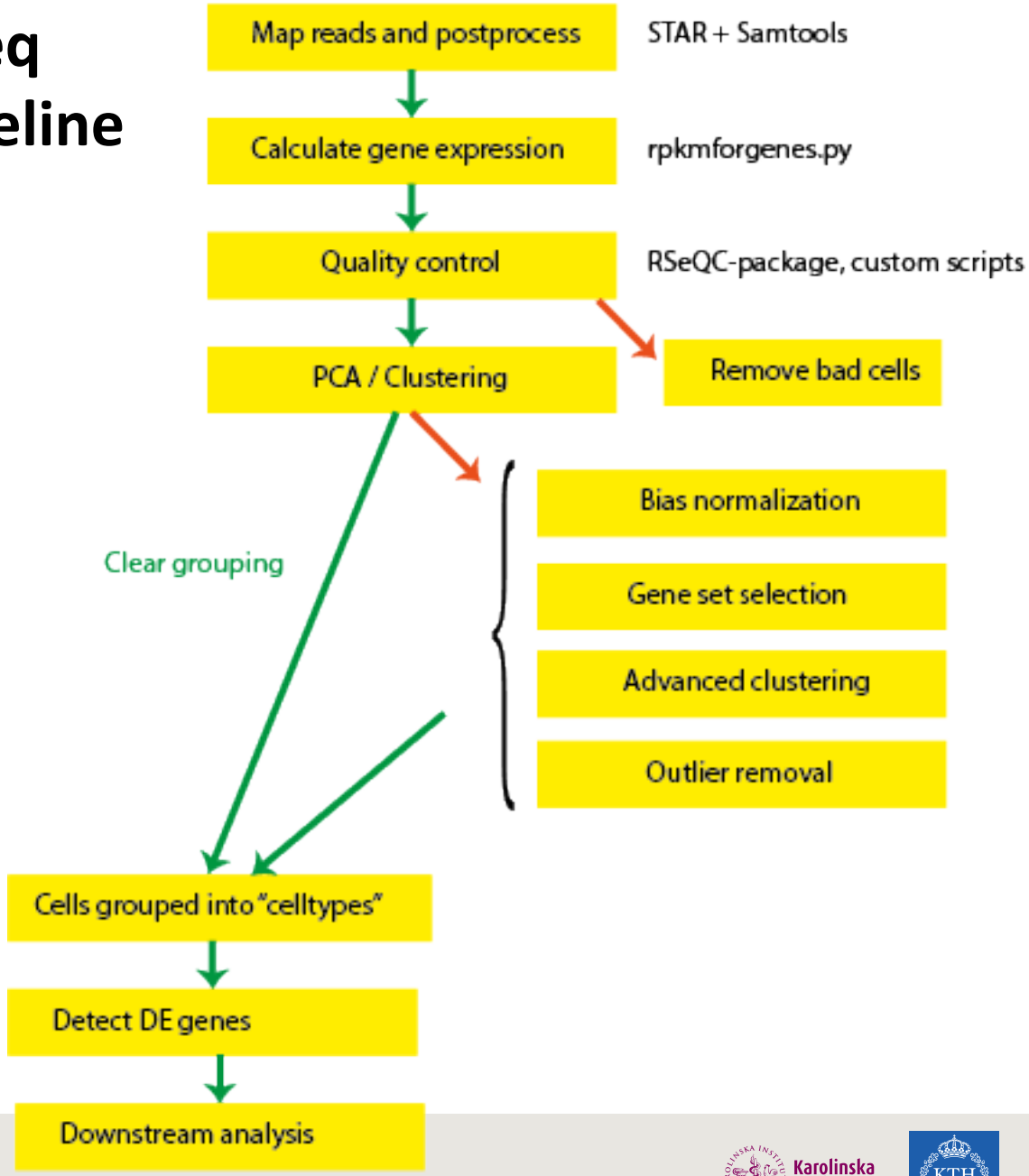


(Buettner et al. Nature Biotech 2015)

Replicates – how many cells do you have to sequence?

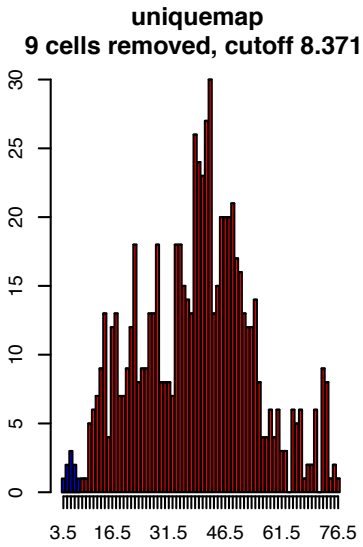
- Recommended to have around 20-30 cells from each cell type
  - A sample with a minor cell type at 5% requires sequencing of 400 cells.
  - Preselecting cells may be necessary, but unbiased cell picking is preferred.
- To study gene expression only, sequencing depth does not have to be deep.
  - Multiplexing of 96 samples on one lane is common. If mapping quality is good – you should get at least 1M reads per sample.
  - For tag-based methods sequencing is often more shallow.

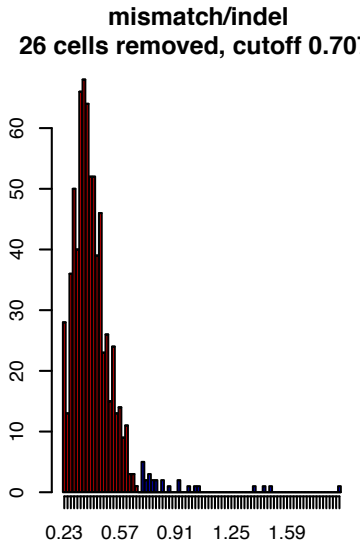
scRNA-seq analysis pipeline

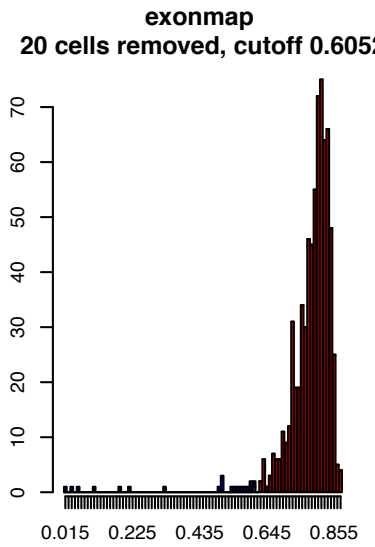


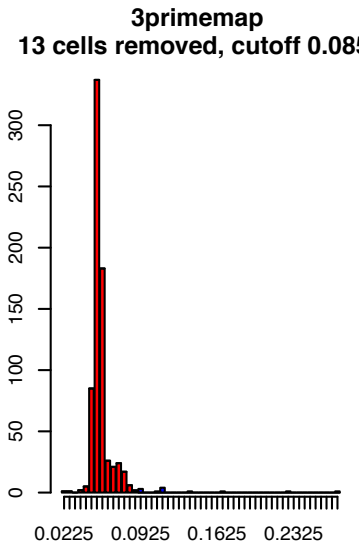
Quality Control (QC)

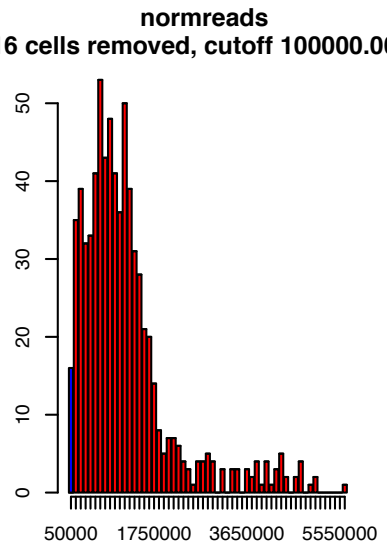
- QC is a crucial step in scRNA-seq
- Any experiment will have a number of failed libraries – important to remove these before analysis.
- OBS! Smaller celltypes gives lower mapping rates and more primer dimers.
- Look at:
  - Mismatch rate
  - 3’ bias (degraded RNA)
  - Mapping statistics (% uniquely mapping)
  - Fraction of exon mapping reads
  - Number of detected genes
- Depending on cell type, around 500K exon mapping reads saturates the gene detection. Can be deduced from subsampling.

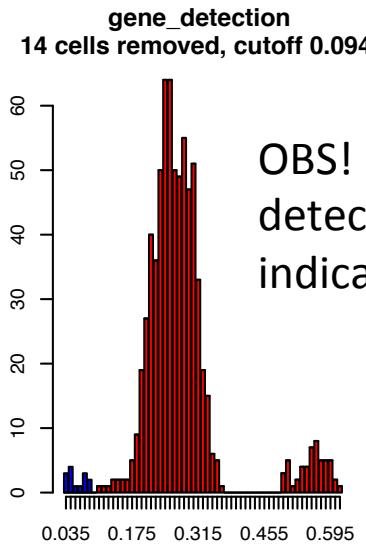








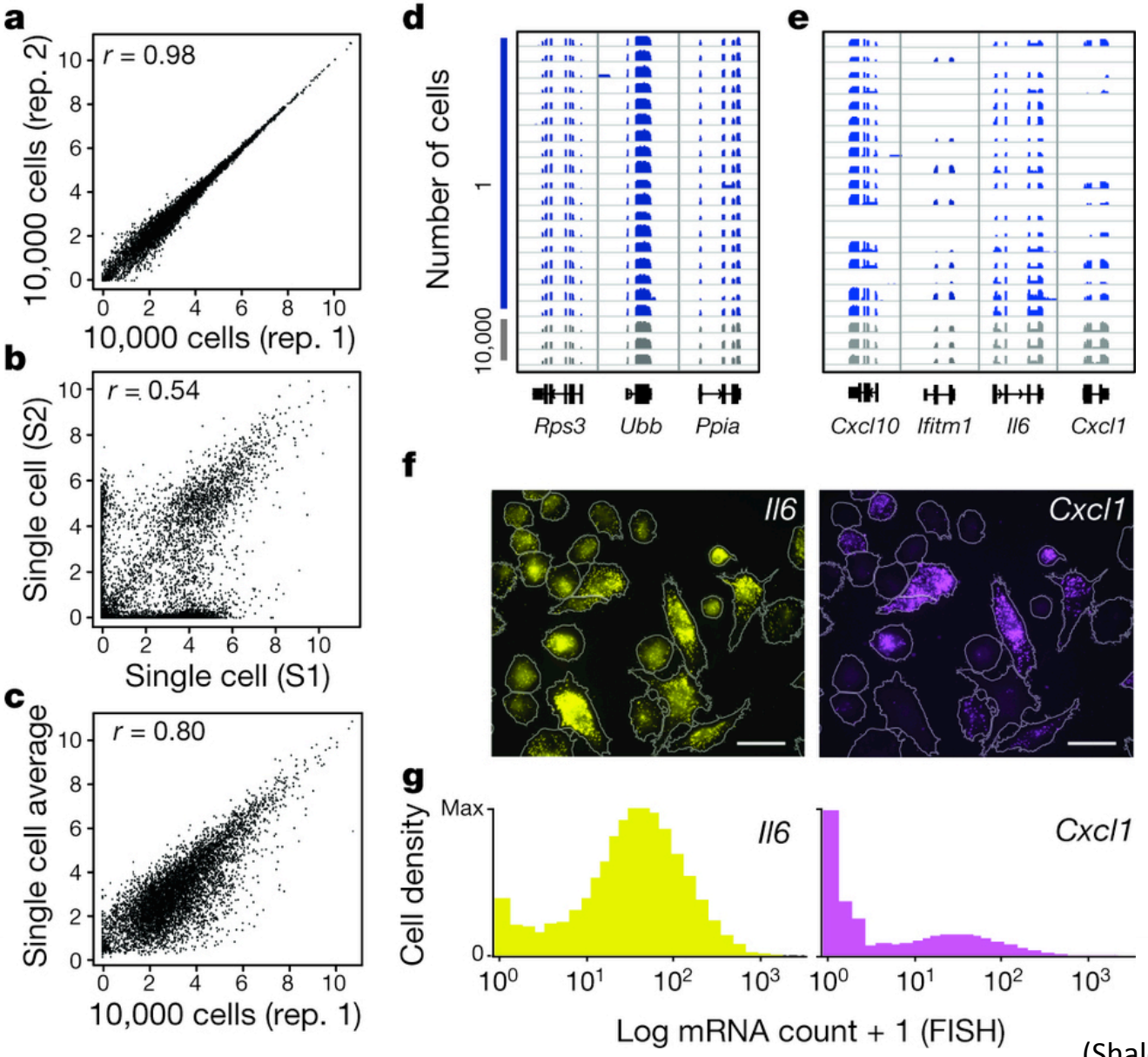




OBS! High number of detected genes could indicate multiple cells

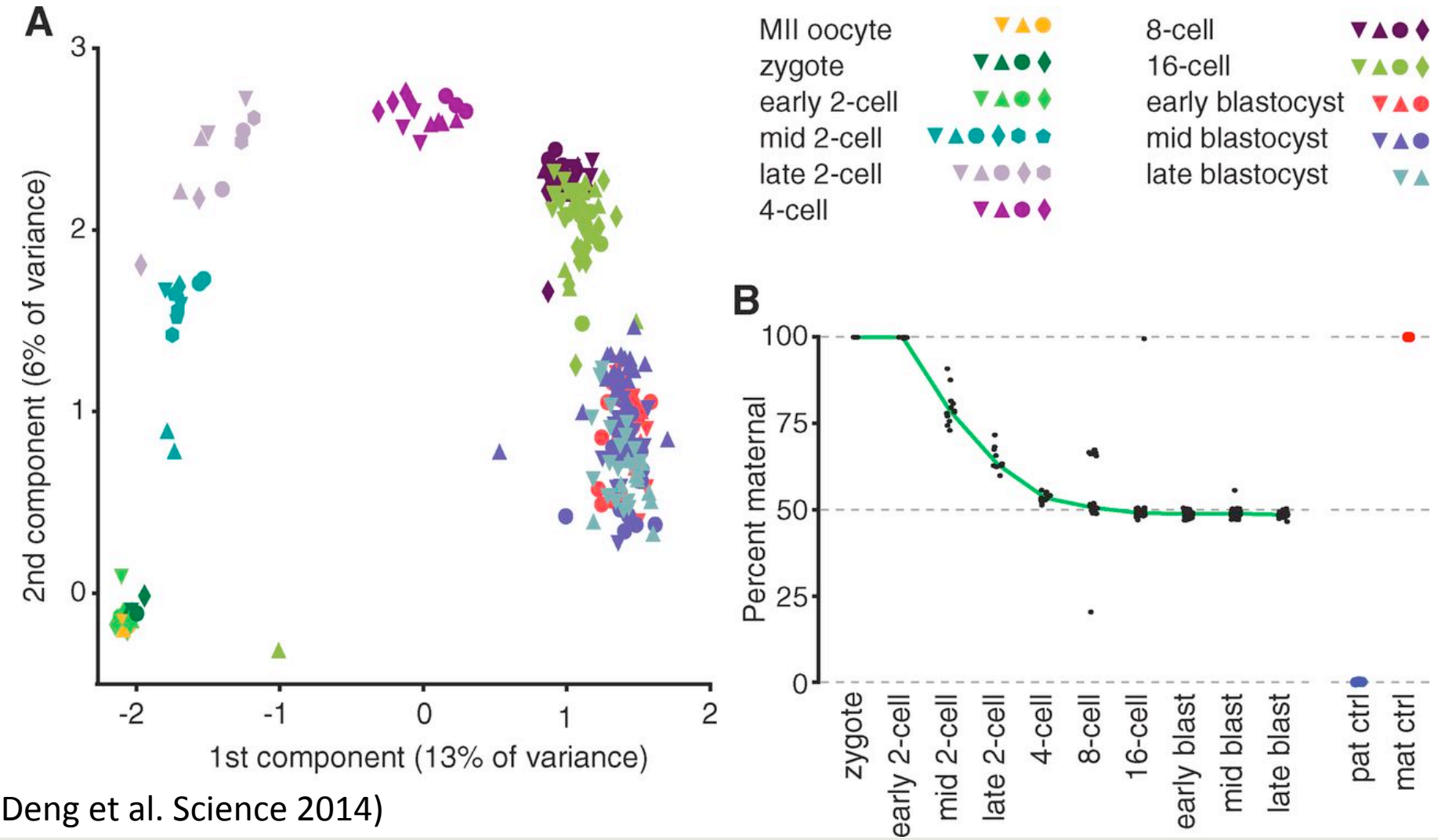
total 48 samples failed QC
Total samples: 722

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



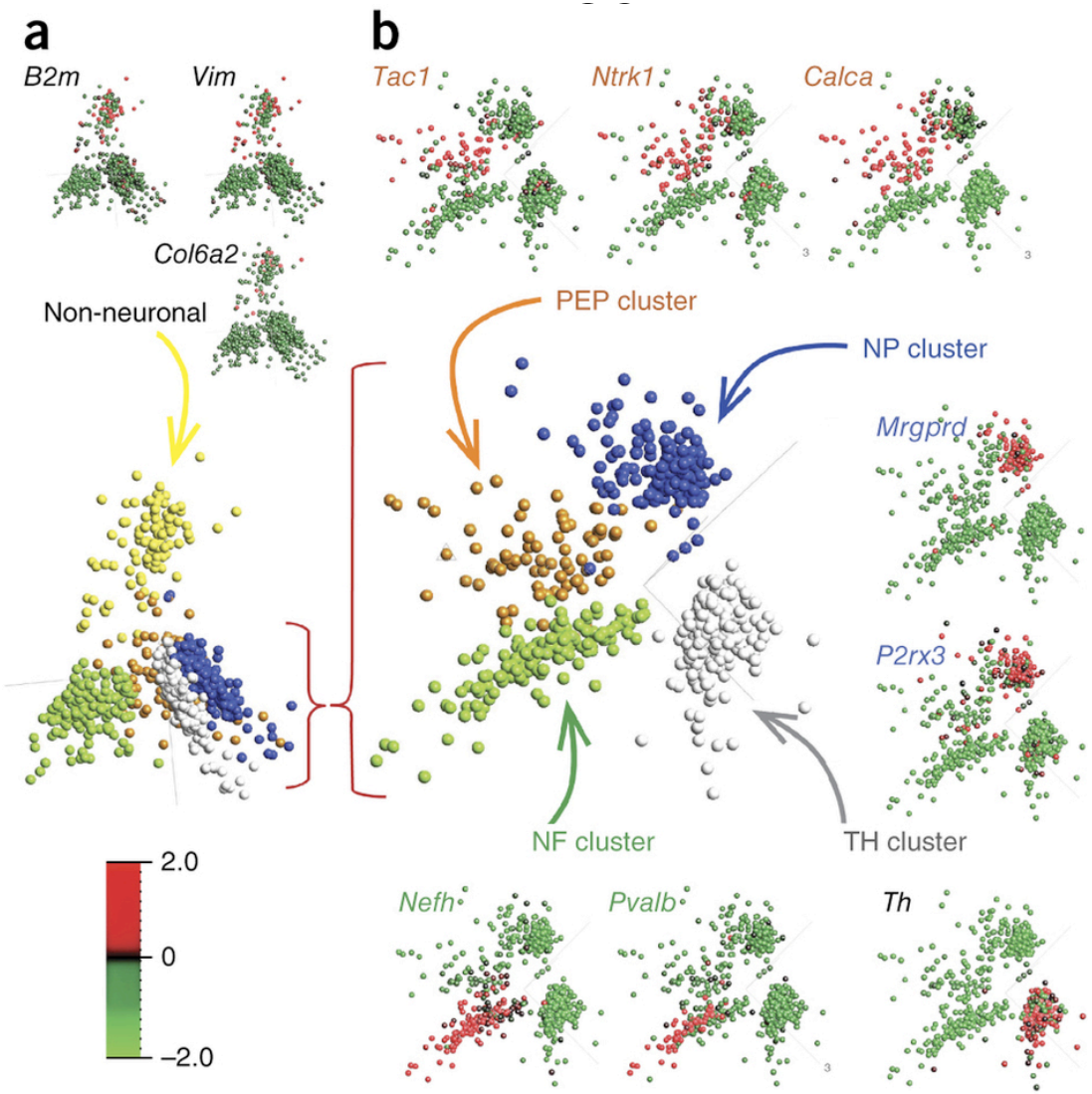
(Shalek et al. Nature 2014)

Identifying celltypes using dimensional reduction 1



(Deng et al. Science 2014)

Identifying celltypes using dimensional reduction 2



(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)

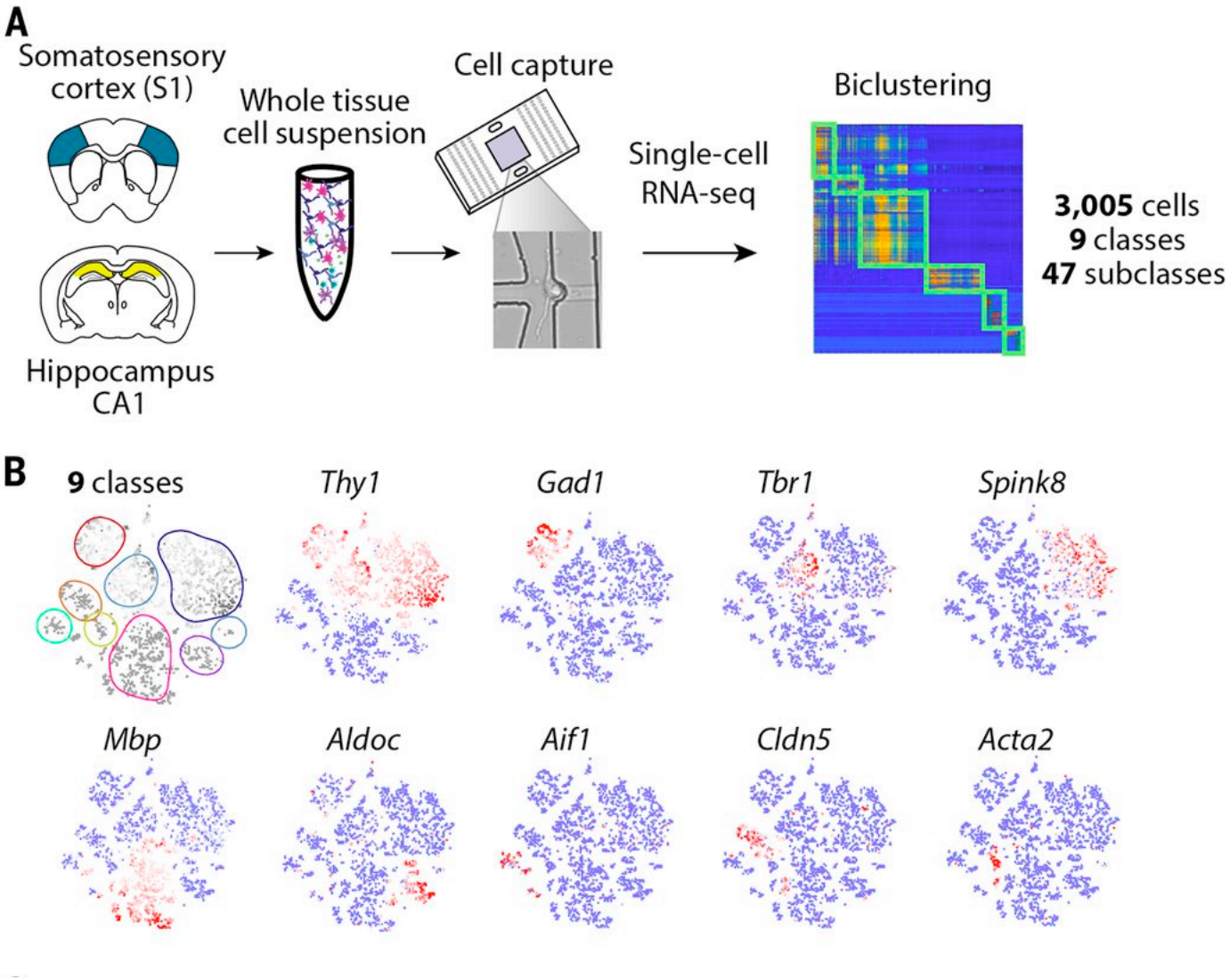
Identifying celltypes - Clustering

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

t-SNE – t-distributed stochastic neighbor embedding

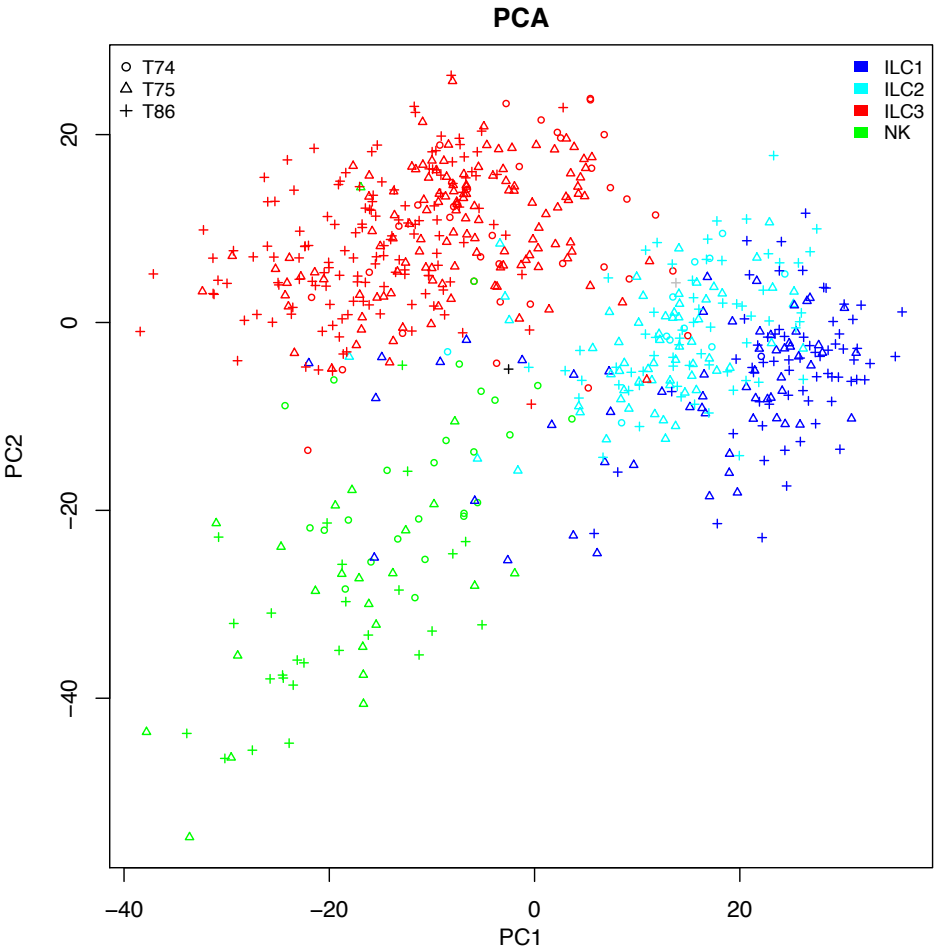
- Method often used in single cell proteomics.
- **Step 1** – probability distribution for all pairs in PCA space with N principal components
- **Step 2** – dimensionality reduction with similar probability distribution and minimization of divergence between distributions
- Implementations in R:
  -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

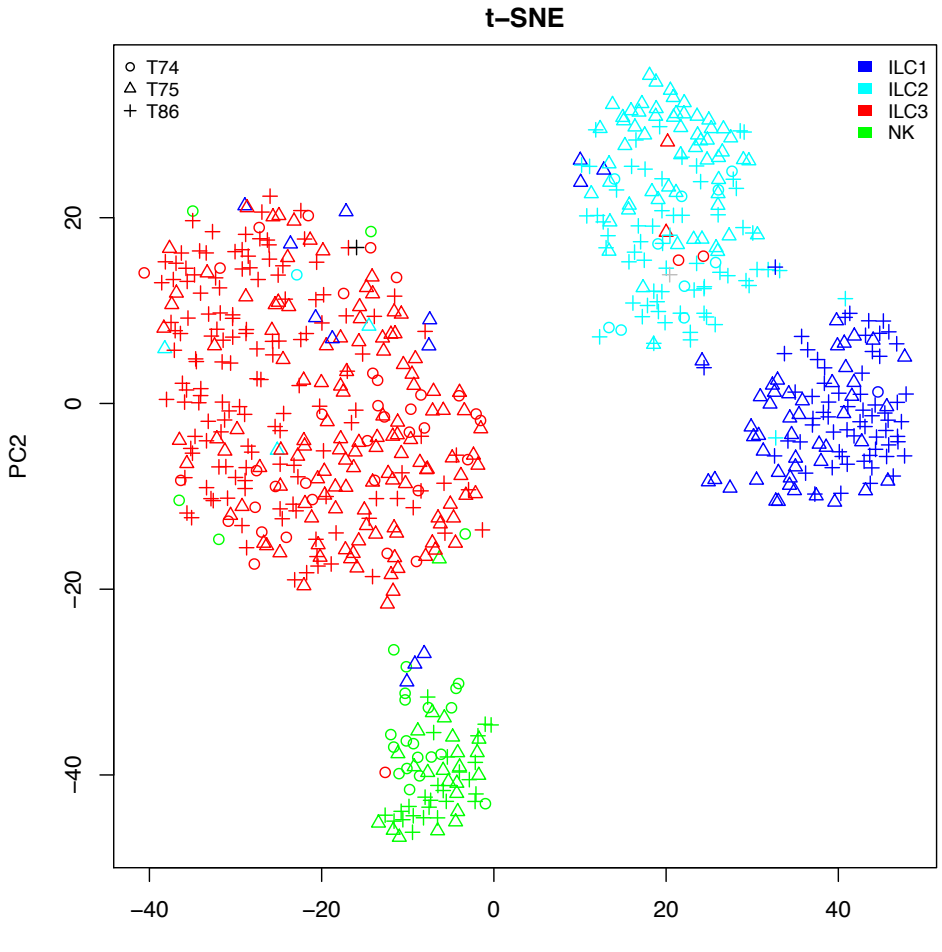
t-SNE – t-distributed stochastic neighbor embedding



(Zeizel et al. Science 2015)

t-SNE vs PCA dimensionality reduction





(Björklund et al. Unpublished)

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

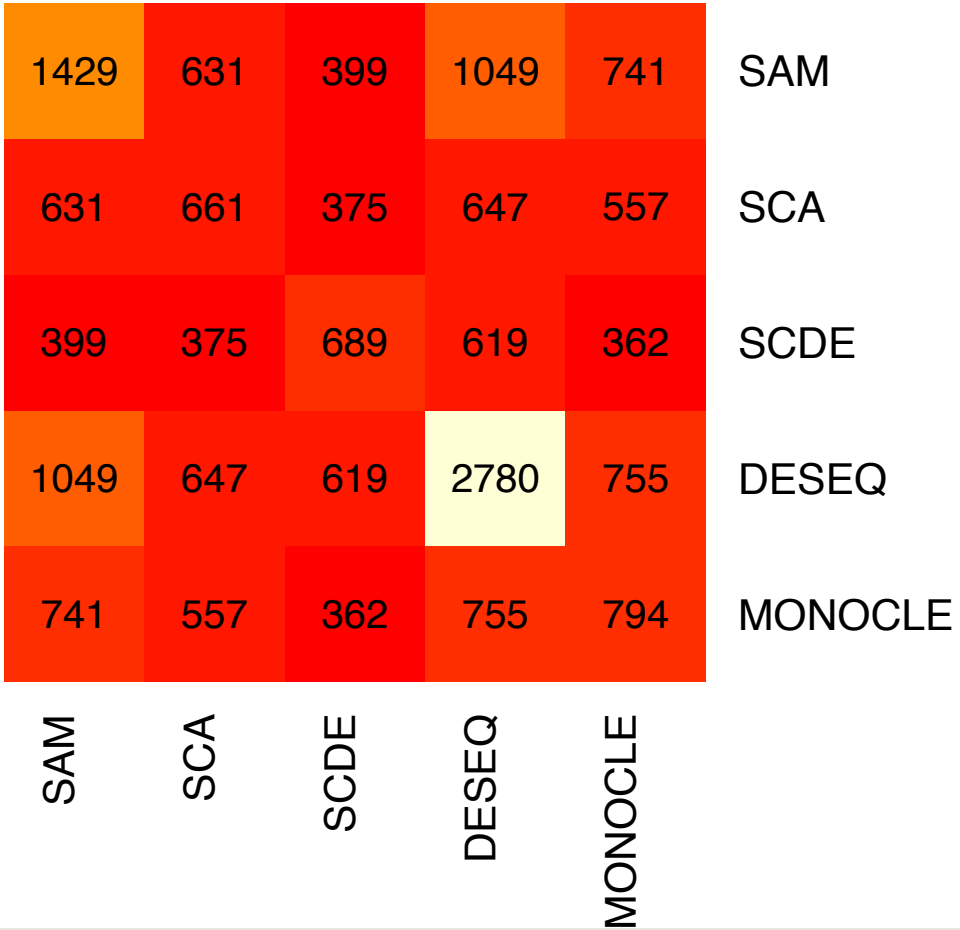
- Parametric methods like EdgeR & DESeq not suitable for scRNA seq since the parameter assumptions in those methods does not apply here.
- Simple fisher or chi square test works better in most cases
- Or non-parametric methods like SAMseq

Detecting differentially expressed genes

- Non-parametric methods like SAMseq can be used.
  - SAMstrt – extention to SAMseq with spike-in normalization
- Available single cell DE methods:
  - SingleCellAssay – developed for qPCR experiments
  - Monocle package
  - Single Cell Differential Expression - SCDE
- Some papers use PCA contribution (loadings) to define celltype specific genes

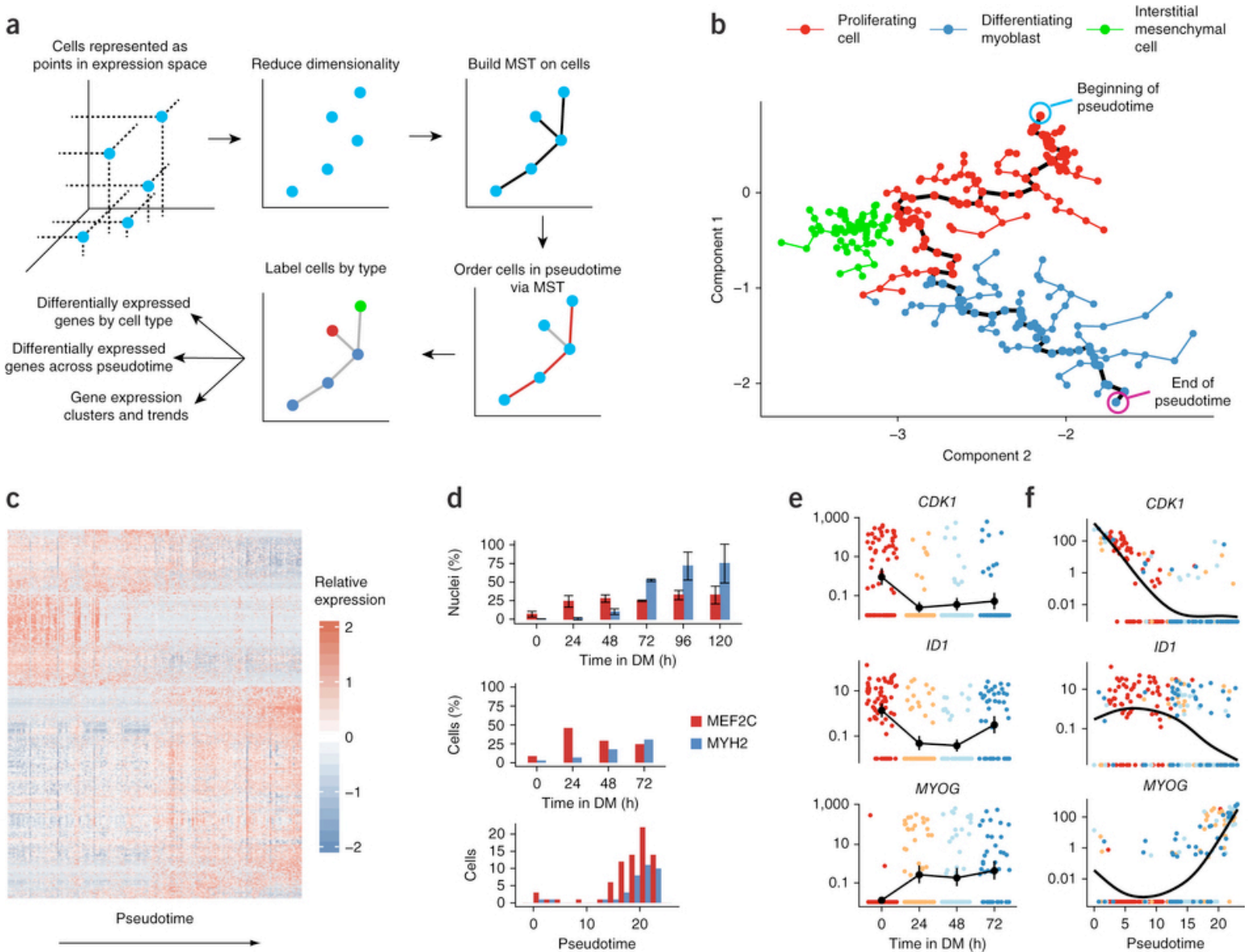
Comparison of some methods

ILC3 overlap of DE genes



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)



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.

Tools for single cell analysis

- Tutorial from Harvard WS:
  - http://pklab.med.harvard.edu/scw2014/
- For differential expression:
  - SCDE: http://pklab.med.harvard.edu/scde/index.html
  - SCA: https://github.com/RGLab/SingleCellAssay
  - 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
  - Pagoda: http://pklab.med.harvard.edu/scde

Recommended reading

- **Single-cell sequencing-based technologies will revolutionize whole-organism science.** Shapiro et al. Nature Reviews Genetics, 14, 618–630 (2013)
- **Computational and analytical challenges in single-cell transcriptomics.** Stegle et al. . Nature Reviews Genetics, 16(3) (2015)
- **Entering the era of single-cell transcriptomics in biology and medicine.** Sandberg, Nature Methods, 11(1): 22-4 (2014)
- **Accounting for technical noise in single-cell RNA-seq experiments.** Brennecke et al. *Nature Methods* 10, 1093–1095 (2013)
- **Bayesian approach to single-cell differential expression analysis.** Karchenko et al. *Nature Methods* (2014)
- **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)
- **Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments.** McDavid et al. *Bioinformatics* 29.4 (2013): 461-467.
- **Quantitative single-cell RNA-seq with unique molecular identifiers**. Islam et al. *Nature Methods* 11, 163–166 (2014)
- **Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells.** Ramsköld et al. *Nature Biotechnology* 30, 777–782 (2012)
- **Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.** Shalek et al. *Nature* 498, 236–240 (13 June 2013)
- **Validation of noise models for single-cell transcriptomics.** Grun et al. *Nature Methods* 11, 637–640 (2014)
- **Smart-seq2 for sensitive full-length transcriptome profiling in single cells.** Picelli et al. *Nature Methods* (2013).
- **SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization.** Katayama et al. *Bioinformatics*. 2013 Nov 15;29(22):2943-5

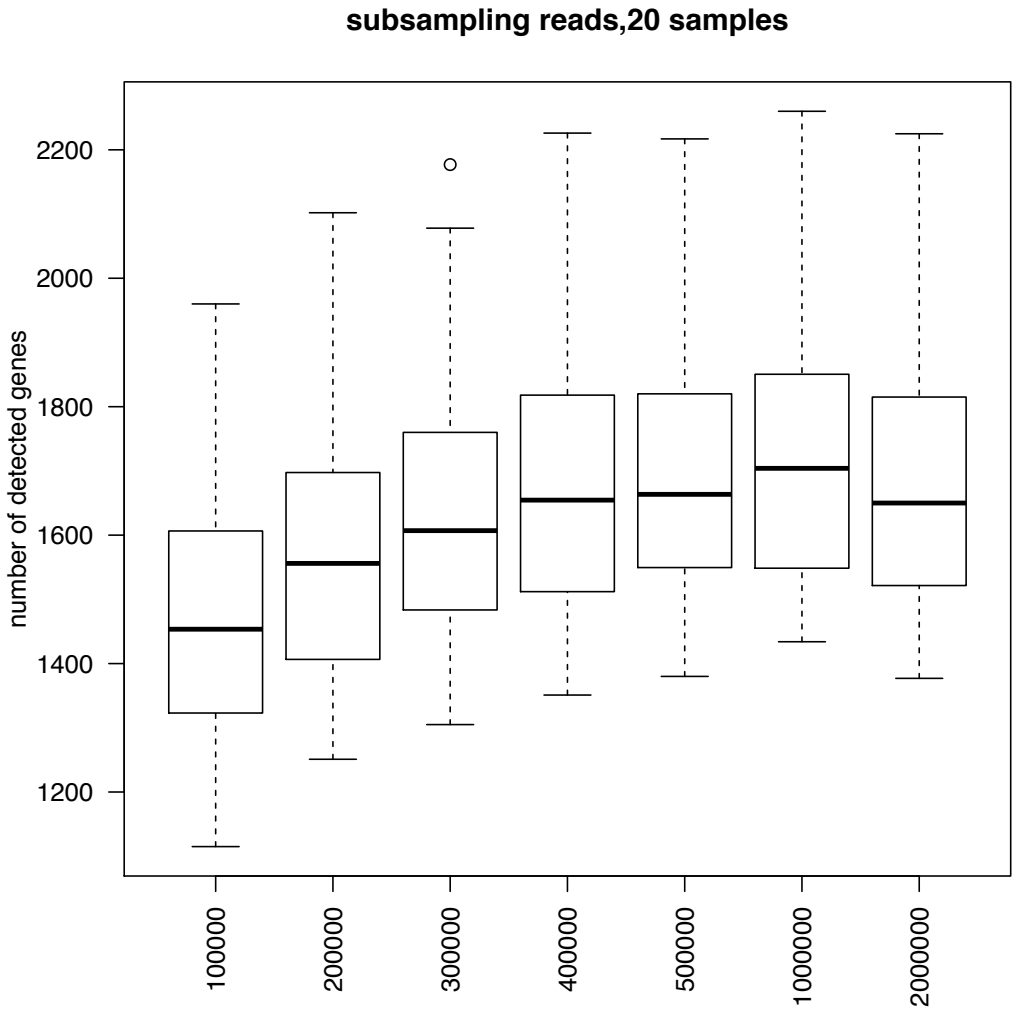
Questions?

Some concepts

- RNA-seq expression values (RPKM, FPKM, counts) close to log normal distribution – always use logged values (log2 or log10)
- Coefficient of variation (CV) – ratio between standard deviation and mean

$$c_v = \frac{\sigma}{\mu}$$

Gene detection – subsampling of reads

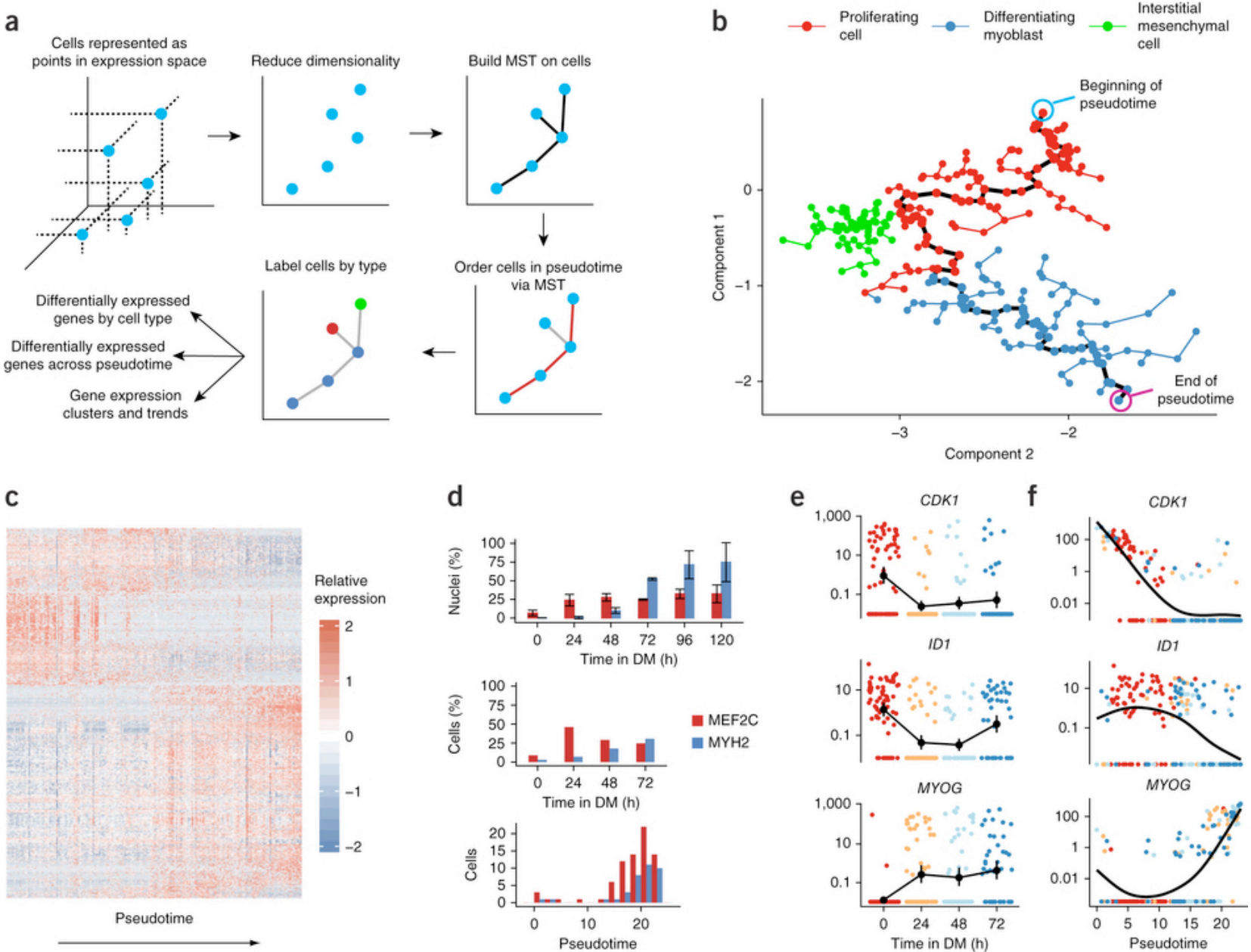


Monocle

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)

- Tool for finding differentiation paths
- Test for differential expression between groups
- Test for differential temporal expression pattern – could use any continuous ordering of cells
- Generalized additive models (GAMs) to model the expression of each gene with a normally distributed error term
- Approximate χ^2 likelihood ratio test
- OBS! Quite slow to run

Monocle

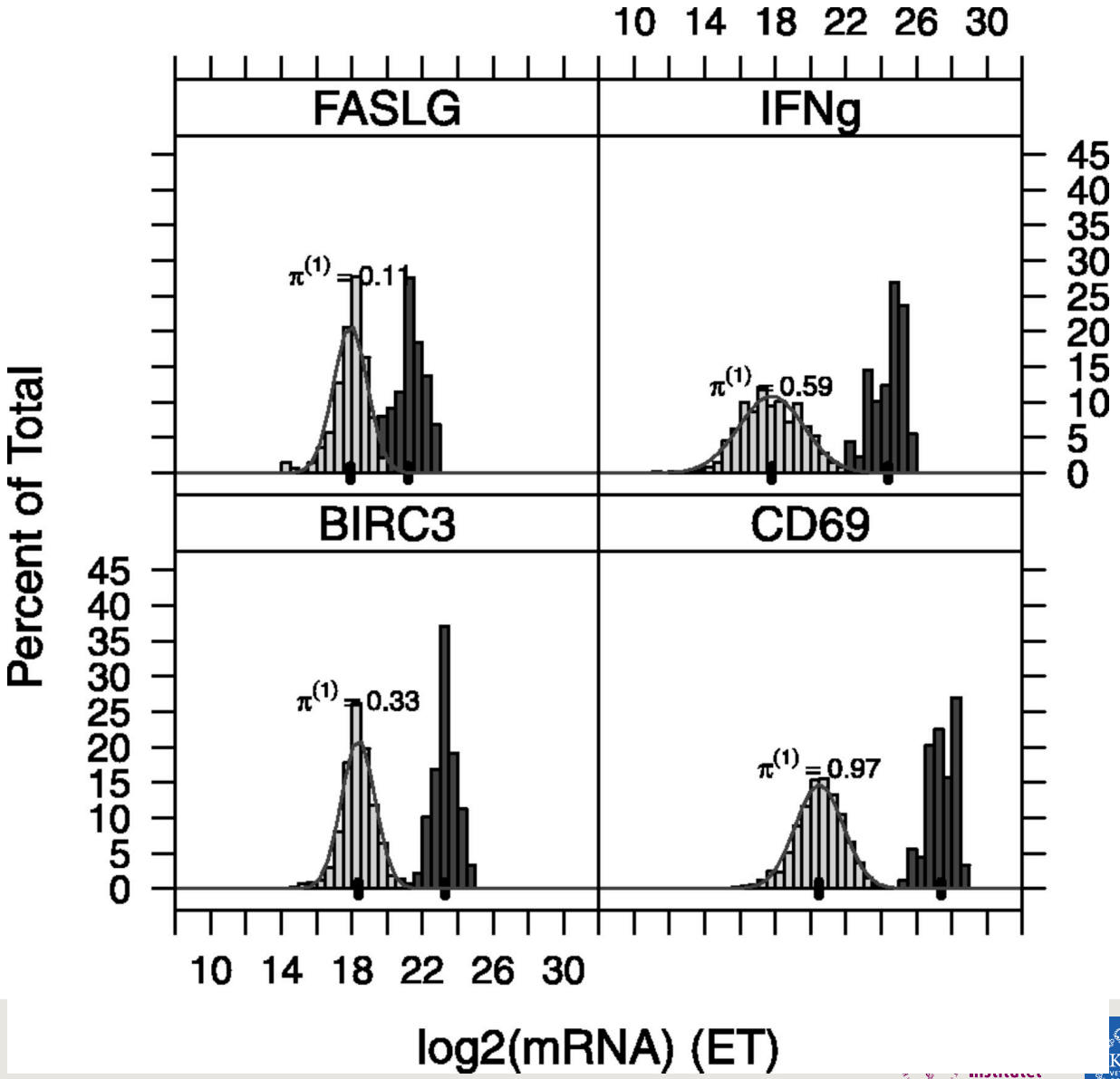

a
Cells represented as points in expression space
Reduce dimensionality
Build MST on cells
Order cells in pseudotime via MST
Label cells by type
Differentially expressed genes by cell type
Differentially expressed genes across pseudotime
Gene expression clusters and trends
b
Proliferating cell
Differentiating myoblast
Interstitial mesenchymal cell
Beginning of pseudotime
End of pseudotime
Component 1
Component 2
c
Relative expression
Pseudotime
d
Nuclei (%)
Time in DM (h)
Cells (%)
MEF2C
MYH2
Cells
Pseudotime
e
CDK1
ID1
MYOG
Time in DM (h)
f
CDK1
ID1
MYOG
Pseudotime

Single Cell Assay

Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. McDavid et al. *Bioinformatics* 29.4 (2013): 461-467.

- Combines discrete and continuous models with a mixture of a point mass at zero and a log-normal distribution.
- Likelihood ratio test (LRT) that can simultaneously test for changes in mean expression (conditional on the gene being expressed) and in the percentage of expressed cells.
- OBS! Requires quite a lot of reordering of data and definition of cutoff for where a gene is defined as non-expressed.

Single Cell Assay

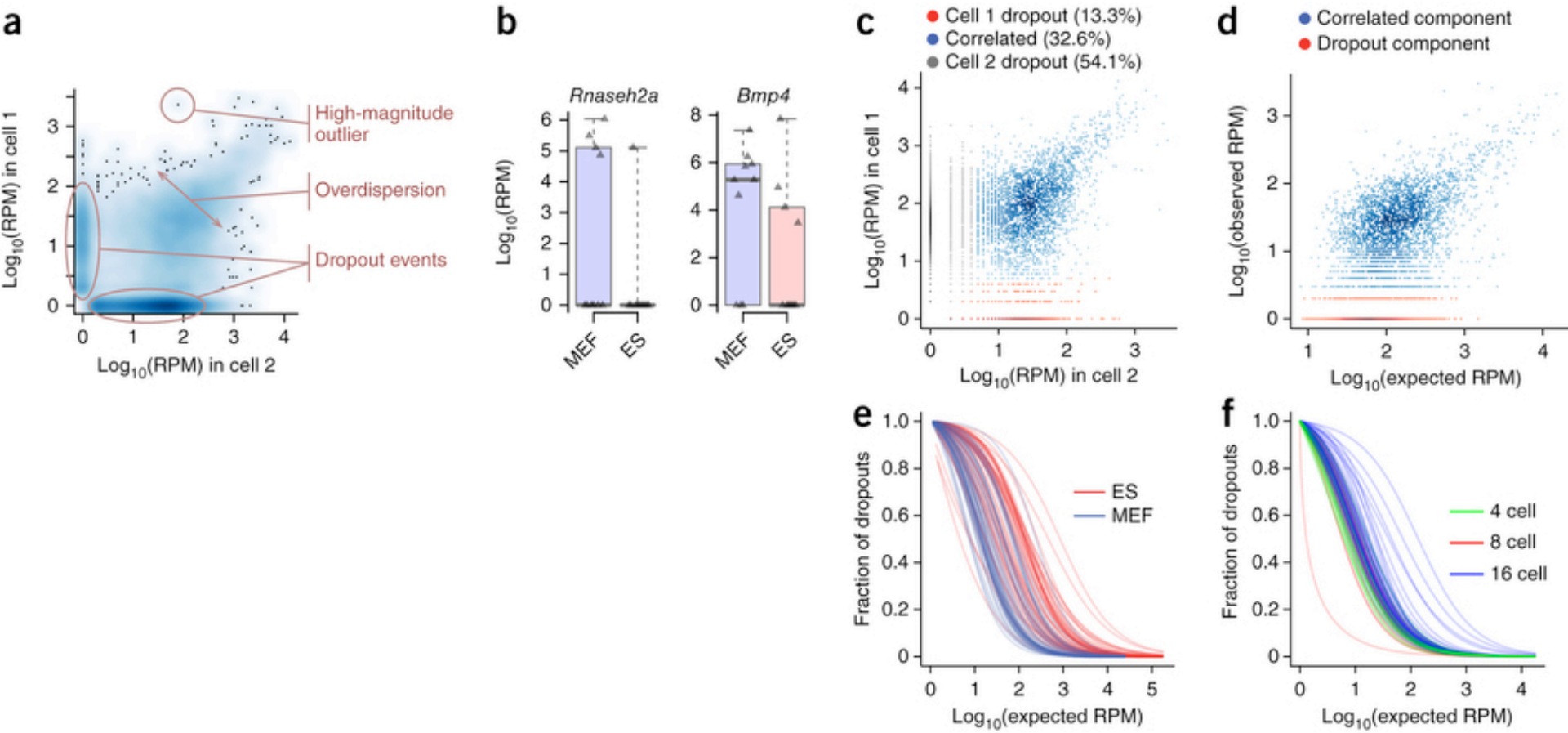


SCDE – Single cell differential expression

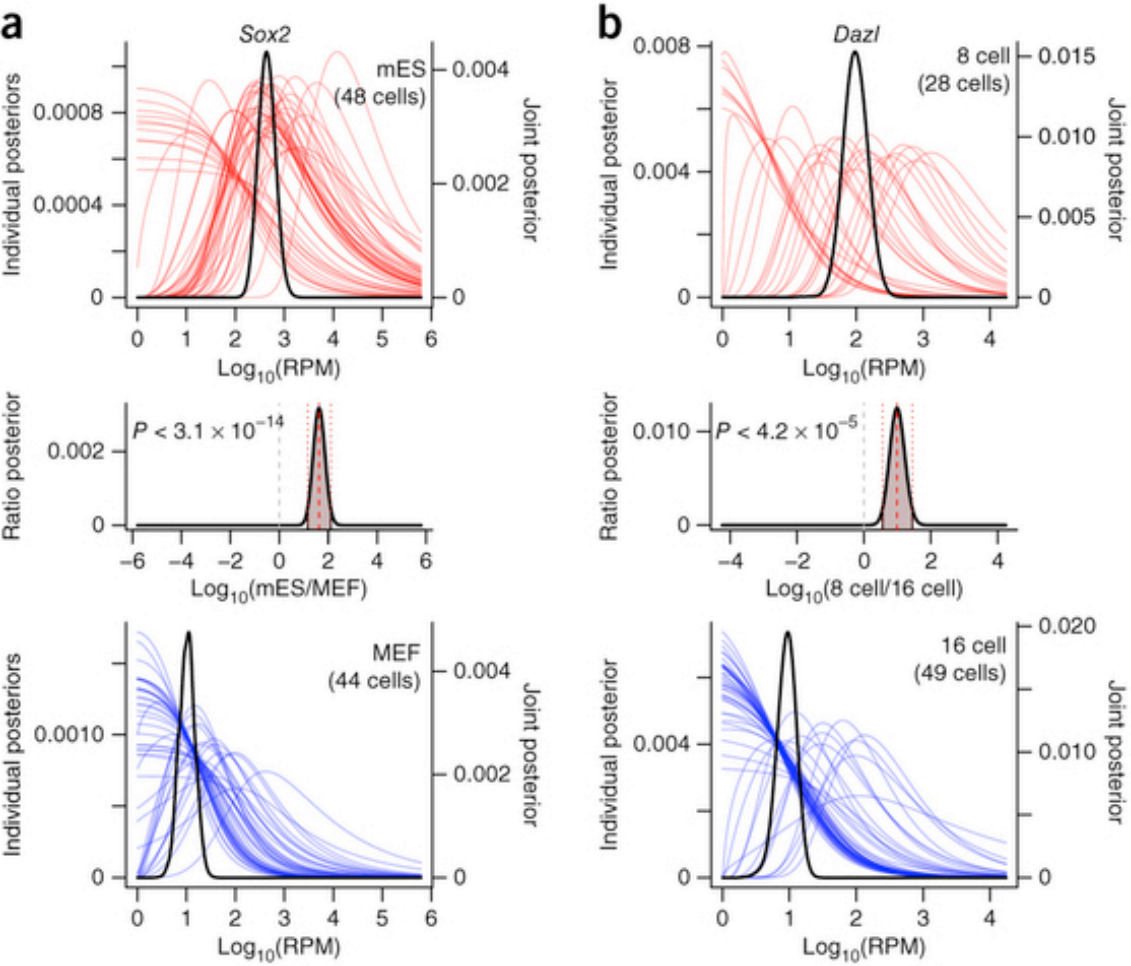
Bayesian approach to single-cell differential expression analysis.
Karchenko et al. *Nature Methods* (2014)

- Models drop out rates for each cell with different rates depending on expression magnitude.
- Fits a posterior distribution for each gene in each cell based on model with a negative binomial distribution for "true" expression and drop out as a low-magnitude Poisson process
- Parameters of the negative binomial distribution fitted based on pairwise comparisons within cell types.

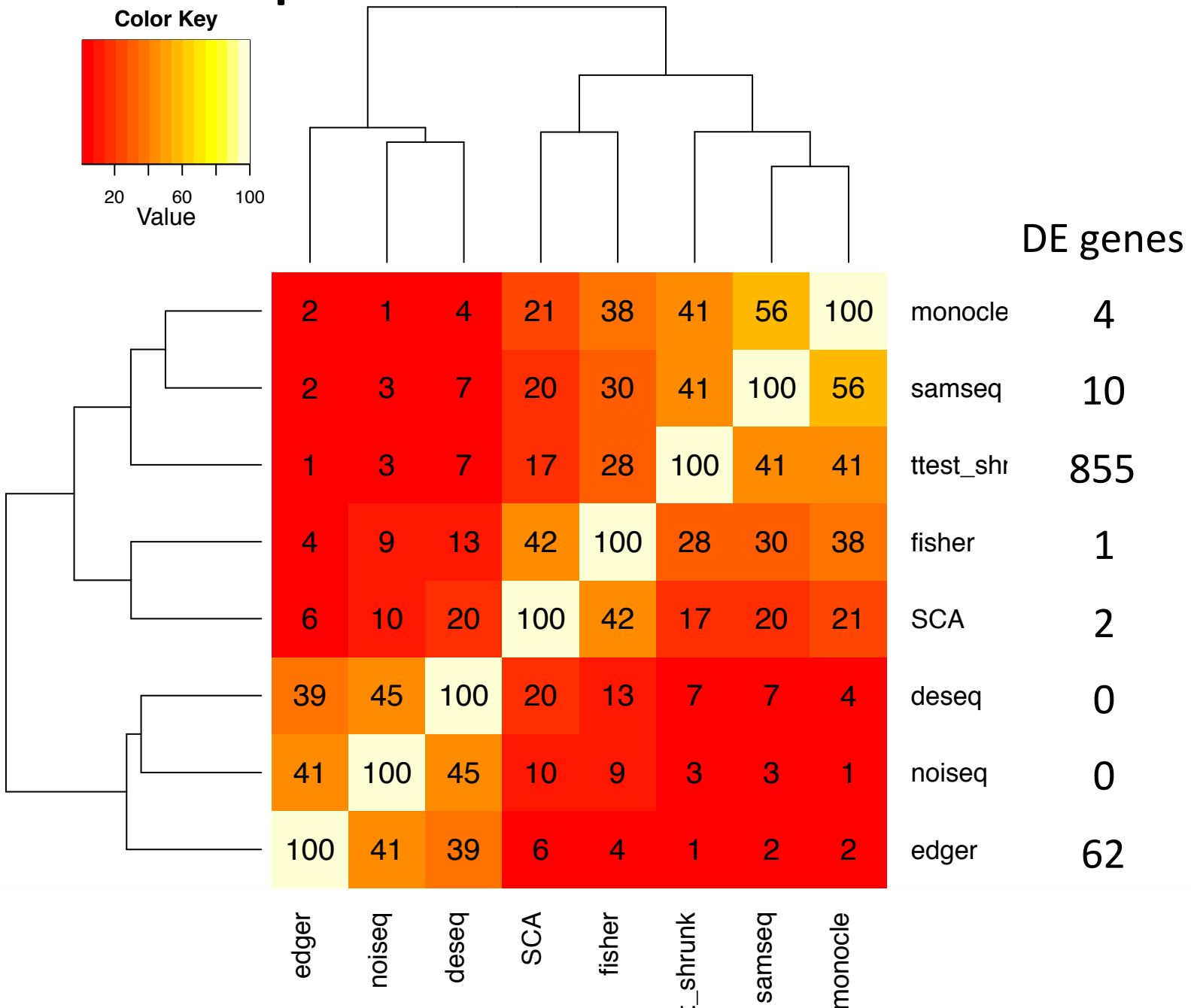
SCDE


a
High-magnitude outlier
Overdispersion
Dropout events
Log10(RPM) in cell 1
Log10(RPM) in cell 2
b
Rnaseh2a
Bmp4
Log10(RPM)
MEF
ES
c
Cell 1 dropout (13.3%)
Correlated (32.6%)
Cell 2 dropout (54.1%)
Log10(RPM) in cell 1
Log10(RPM) in cell 2
d
Correlated component
Dropout component
Log10(observed RPM)
Log10(expected RPM)
e
Fraction of dropouts
ES
MEF
Log10(expected RPM)
f
Fraction of dropouts
4 cell
8 cell
16 cell
Log10(expected RPM)

SCDE


a
Sox2
mES
(48 cells)
Individual posteriors
Joint posterior
Log10(RPM)
Ratio posterior
P < 3.1 × 10−14
Log10(mES/MEF)
MEF
(44 cells)
b
Dazl
8 cell
(28 cells)
P < 4.2 × 10−5
Log10(8 cell/16 cell)
16 cell
(49 cells)
c
SCDE
DESeq
CuffDiff2
SCA
SCDE: AUC = 0.76
DESeq: AUC = 0.67
CuffDiff2: AUC = 0.59
SCA: AUC = 0.60
True-positive rate
False-positive rate
d
Reciprocal
Pearson
Direct
Bray-Curtis
Fraction correctly classified
Top differential genes omitted

Comparison of some methods



Bimodal gene expression or background expression