A short introduction to single cell RNA-seq analyses

Nathalie Vialaneix

January 17th, 2019 - Biopuces

Unité MIAT, INRA Toulouse

Sources

These slides have been made using previous presentations from:

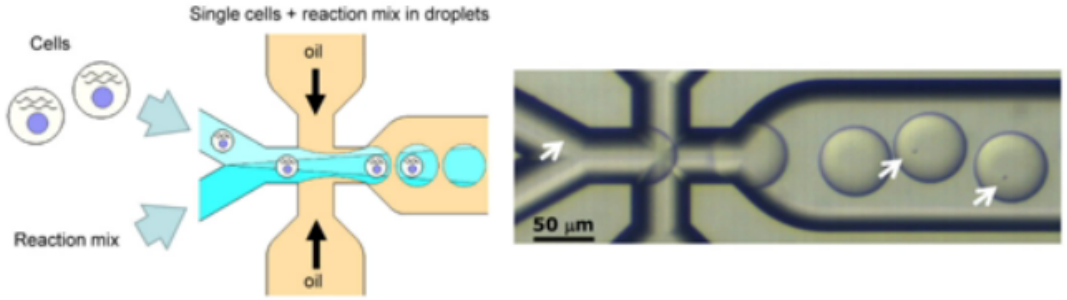
- Delphine Labourdette (LISBP) - diaporama
- Cathy Maugis (IMT) - diaporama
- Franck Picard (LBBE, Lyon) - diaporama

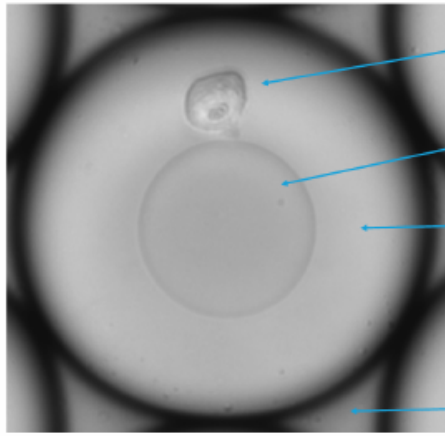
Simple description of single cell datasets

10x Genomics Chromium

Droplet biology

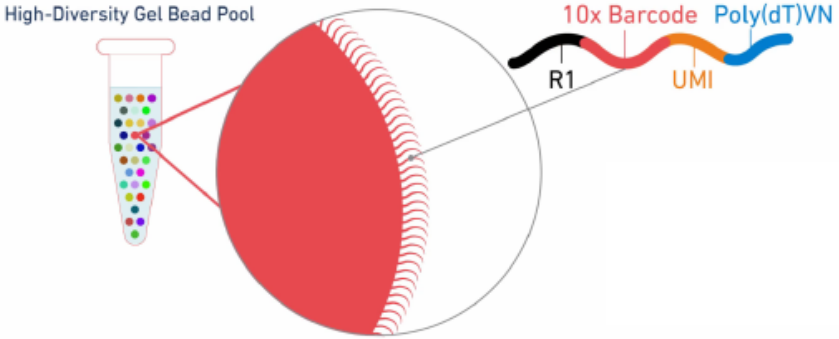
- Encapsulating millions of single cells in controlled, biocompatible, droplet micro-reactors





- Small droplet volumes lead to efficient & robust single cell reactions

10x Technology samples a pool of ~750,000 10x Barcodes to separately index each cell's transcriptome

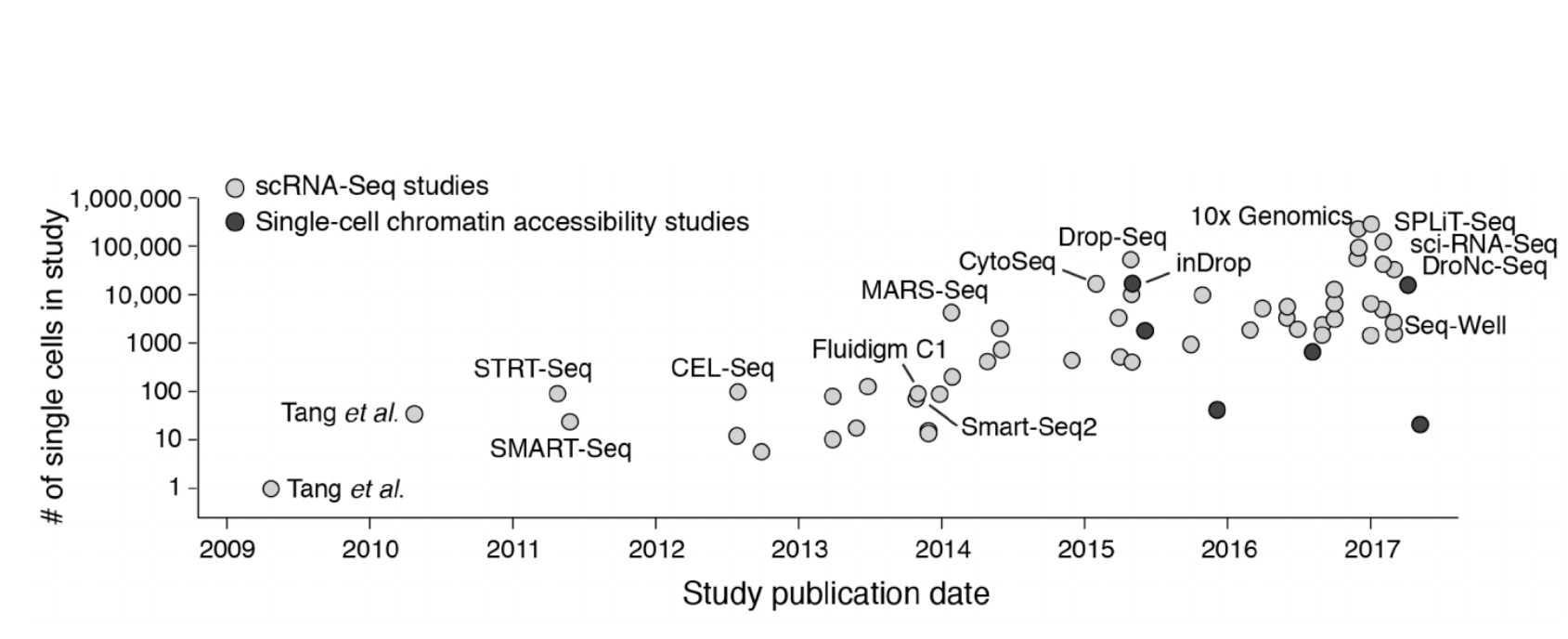


750 000 barcodes possibles

A few remarks

- barcoding is used to index each cell
- UMI are used to index each transcript and correct the amplification bias during library preparation
- droplet technology does not allow for spike-ins (which would be useful for normalization)
- droplets sometimes include duplicates or triplicates (more frequent in cancer cells; estimated at ~0-10% of the droplets, depending on the number of cells, it increases with the number of cells)
- many other sc technologies (check Delphine's slides)

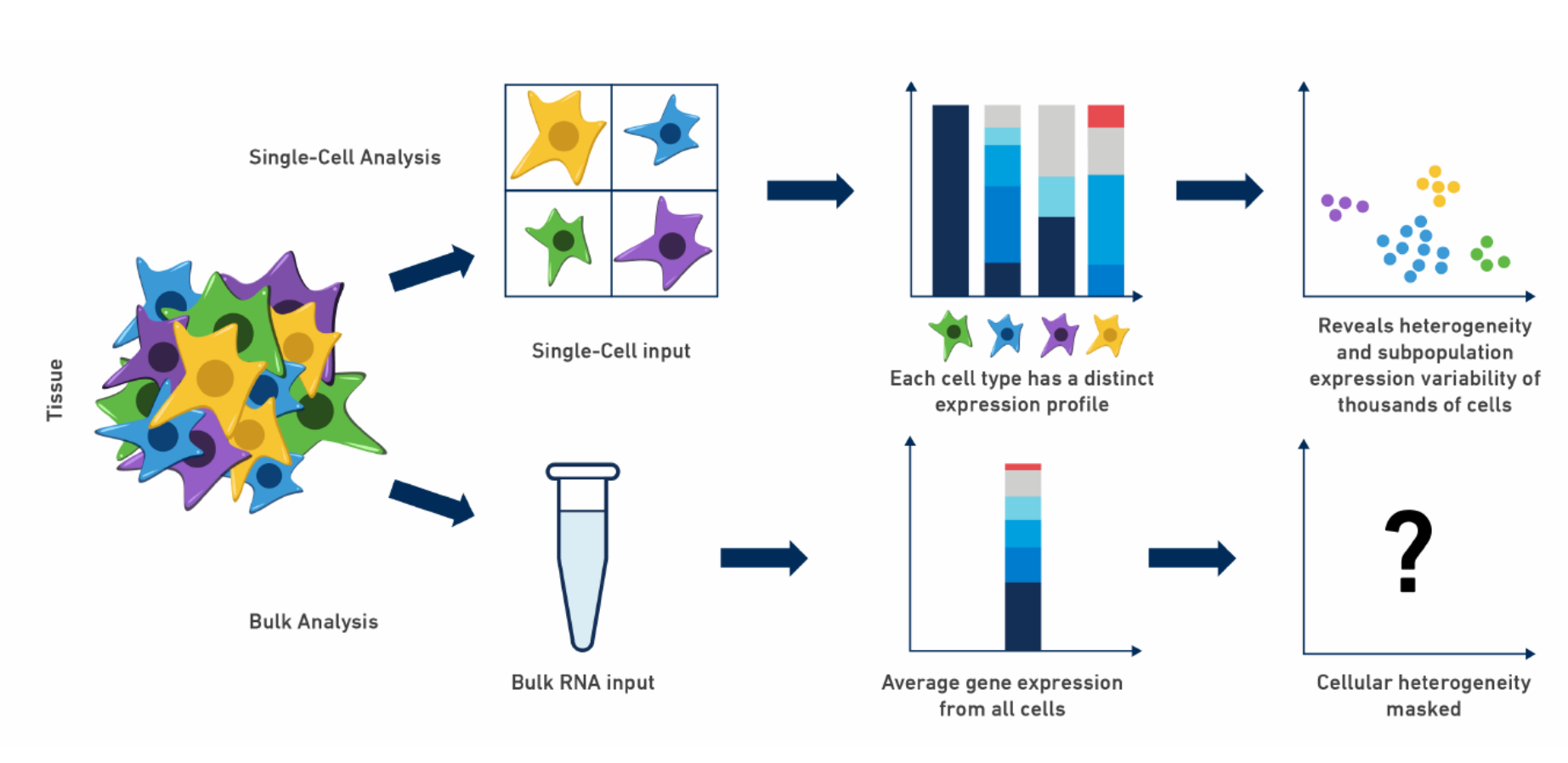
Single-cell technologies



[Regev et al., 2017]

Why single cell?

From a statistical perspective...



From 10x Genomics

Standard analyses and tools

- normalization and dimension reduction
- clustering
- differential expression

can be performed using:

- **the bioconductor workflow "single cell"** (that uses the packages `scatter` and `scran`)
- the all-in-one pipeline "seurat"

Description of datasets and requests from project TregDiab

Datasets

- Count dataset (as produced by Claire) with $n = 8,273$ cells and $p = 27,998$ genes (Unique Molecular Identifier)
- Metadata:
  - on cells: barcode (identifies the cell), group (IL15 or IL2) and genotype (WT or KO)
  - on genes: ENSEMBLE gene name and Gene name
- Frequency distribution of conditions over cells:

	WT	KO
IL15	2452	1609
IL2	2175	2037

The rest of the analysis will focus on cells coming from WT samples.

Questions

1. On the whole population of cells (not taking into account groups and genotypes), perform a typology of cells (unsupervised clustering).

2. Identify markers (genes) that are specific of each cell type.

Data cleaning and normalization

Different steps of the normalization

1. Quality control of the cells: library size distribution, number of expressed genes distribution, mitocondrial proportion distribution.

⇒ Atypical cells are removed from the analysis.

2. Cell cycle classification.

⇒ Only cells in G1 phase are used for the analysis.

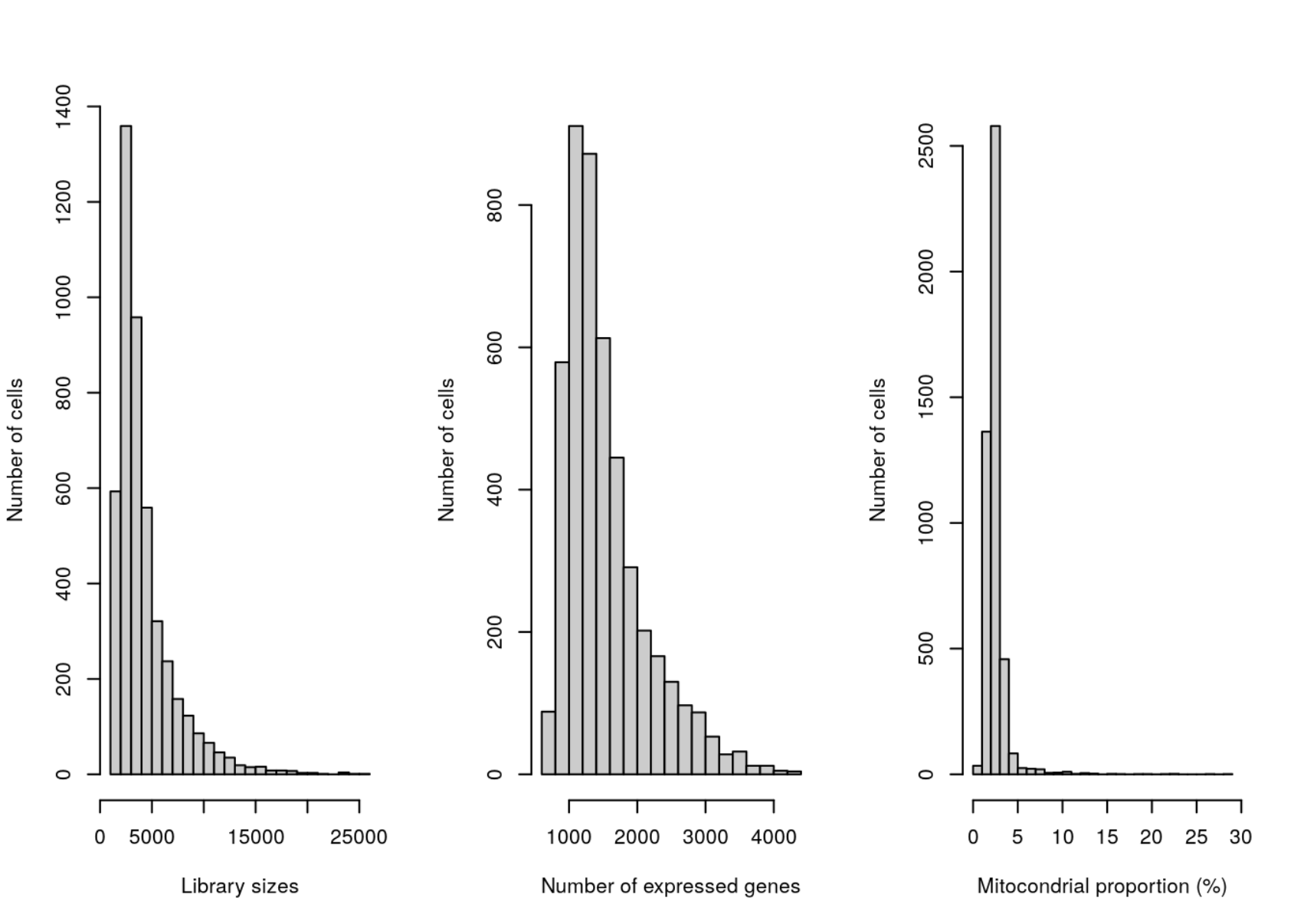
3. Quality control of genes: average count distribution, number of cells in which the gene is expressed.

⇒ Atypical genes (lowly expressed) are removed from the analysis.

4. Normalization of cell specific biases: size factor to correct library sizes are computed after a first (crude) clustering.

What has not been done: Doublet detection

Quality control of cells



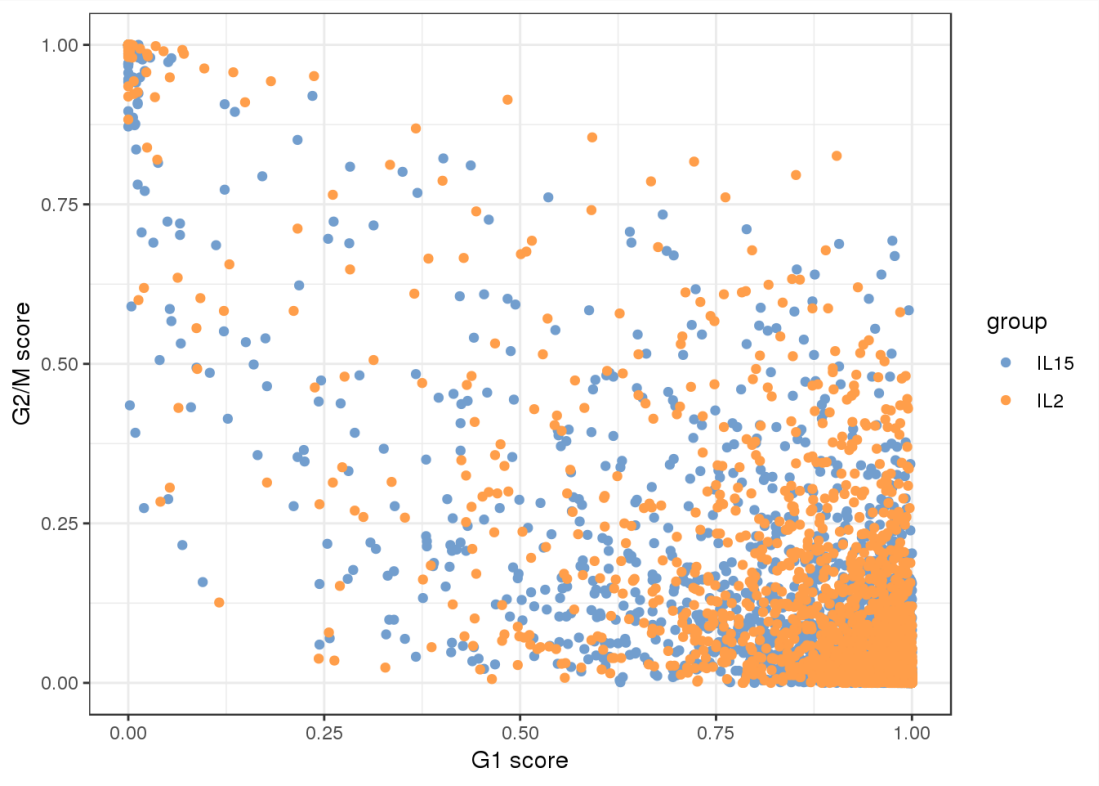
Filtering low quality cells

- remove cells with low library size
- remove cells with a low number of expressed genes
- remove cells with a too large number of mitocondrial genes

$\Rightarrow$ 4,282 remaining cells (out of 4,627 original cells)

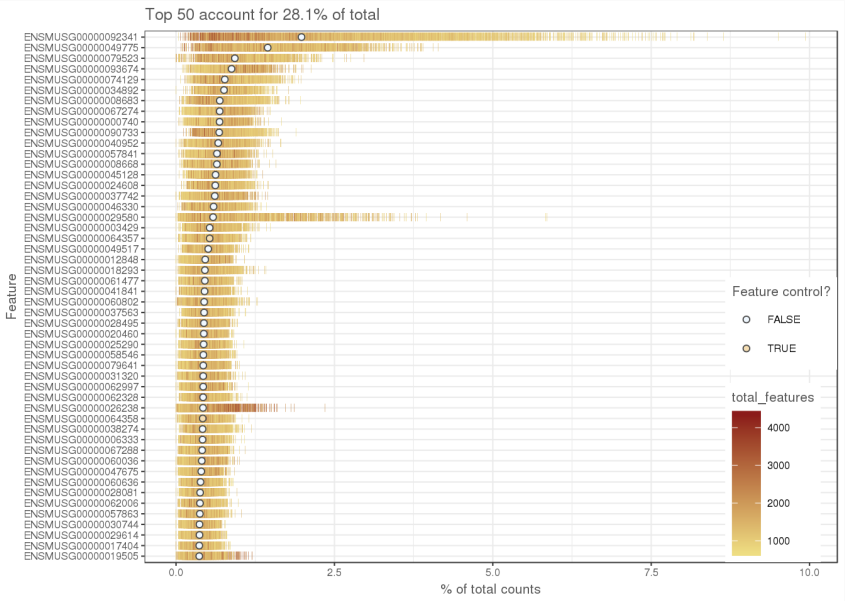
Cell cycle classification

Cell cycle classification is performed using `cyclone` (**R** package `scran`): based on a model that has been trained on specific markers of cell cycles (for mouse and human) ⇒ only cells in G1 phase are used in the analyses (to remove mitosis effects)

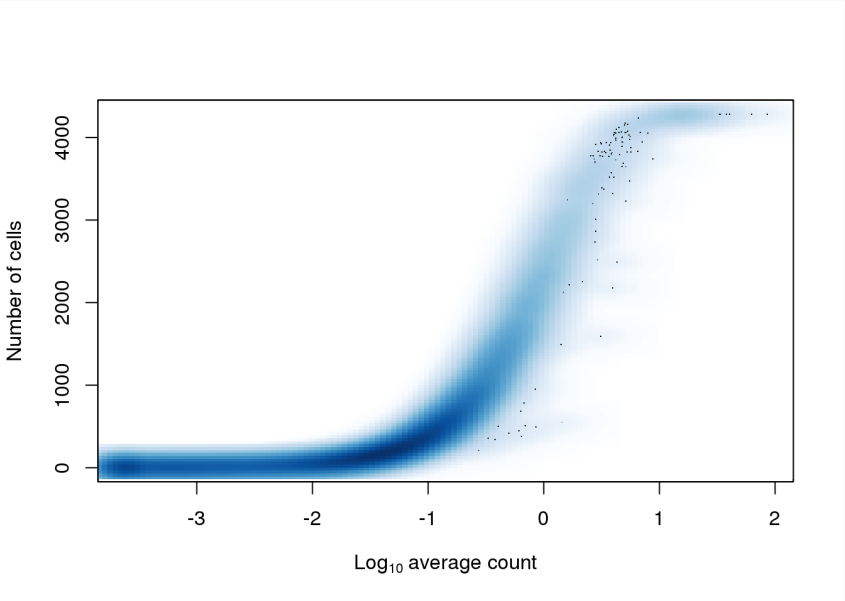


	IL15	IL2
G1	2027	1871
G2	148	99
S	84	53

Gene quality



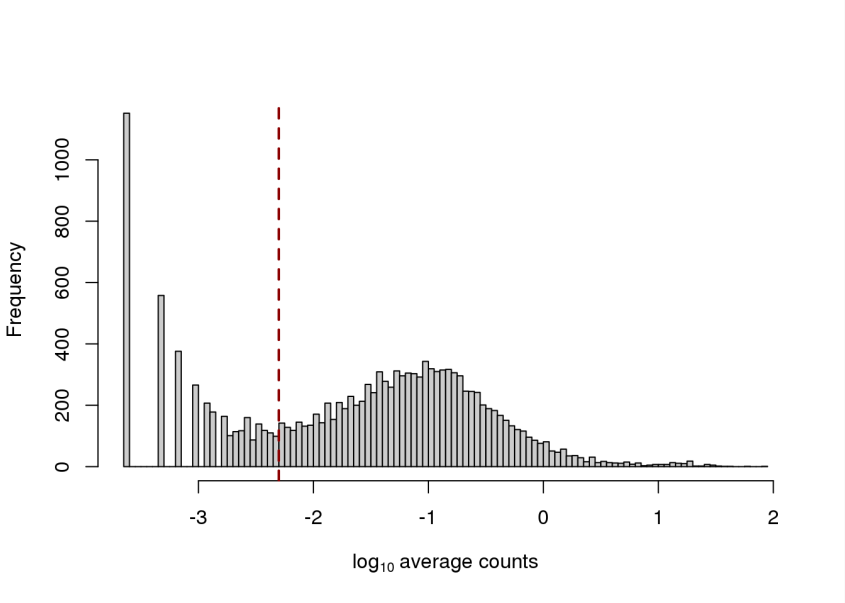
distribution of high expressed genes



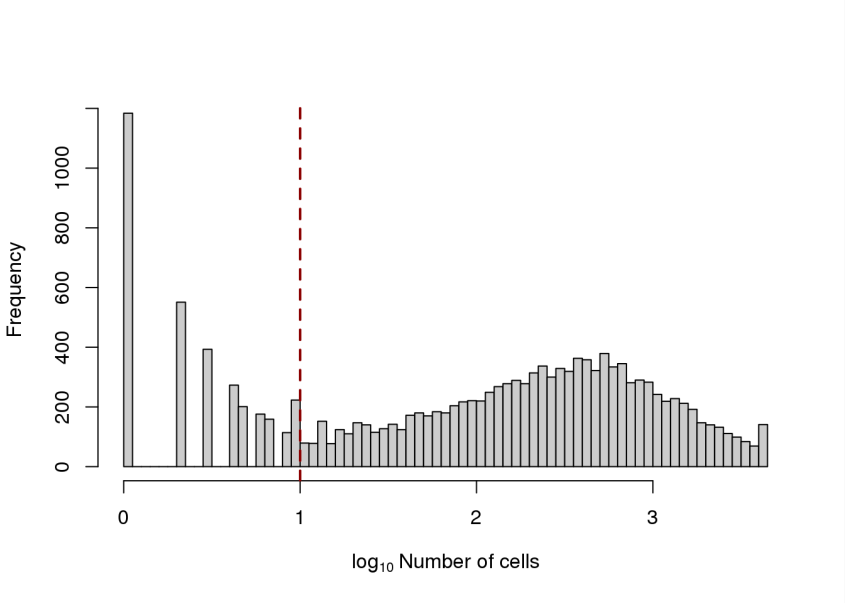
average log expression distribution

Filtering atypical genes

Removed non variable genes: $13,629$ with a variance equal to 0 (48.7%).



low expressed genes



genes expressed in few cells

$\Rightarrow$ 10,418 remaining genes (out of 27,998 initial genes)

Normalization

Normalization is performed after similar cells have been clustered together (based on the most expressed genes; **R** package `scater`).

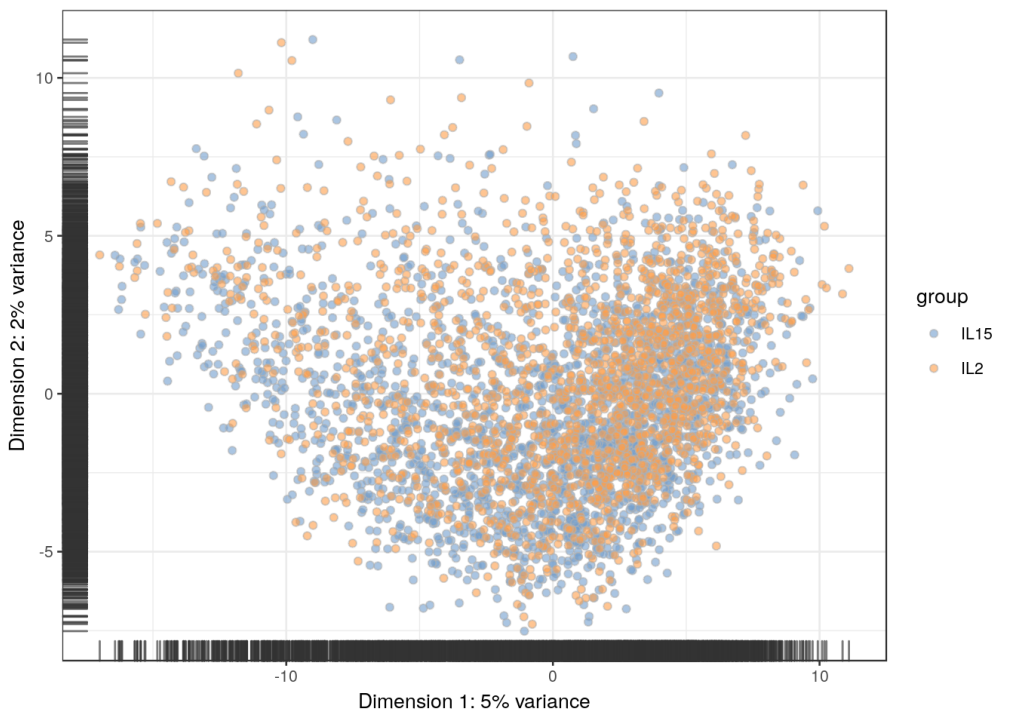
$\Rightarrow$ Scaling factors of library size are obtained (similar to RNA-seq, one can even normalize the library size as in `edgeR`).

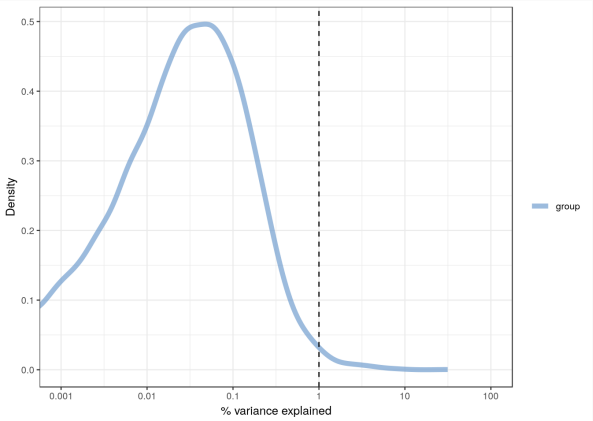
Dimension reduction and clustering

Standard approach for exploratory analysis

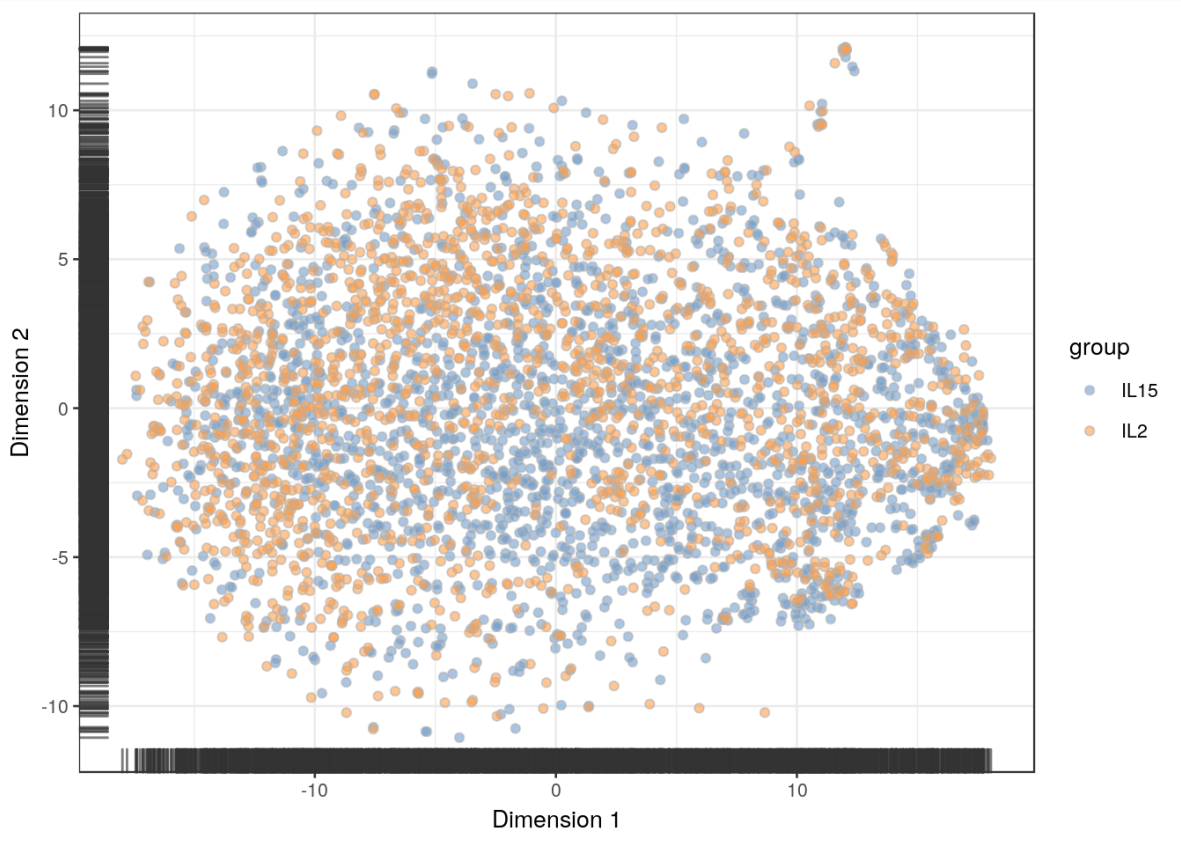
- dimension reduction (PCA, nearest neighbors graphs...)
- visualization (PCA, or t-SNE based on PCA or on any other dimension reduction)
- clustering

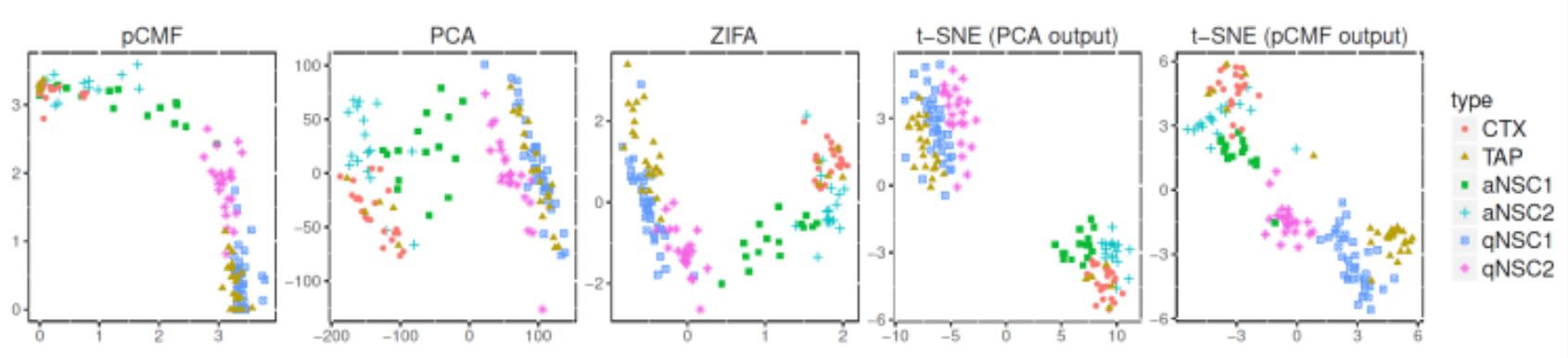
PCA (all genes)





t-SNE (perplexity: 50, R package `scater`)





What does t-SNE?

If cell expressions are noted x_1, ..., x_n (n cells, x_i is in $\mathbb{R}^p$), then

- compute a similaritly between samples with:

$$p_{i|j} = \frac{\exp(-\gamma^2 \|x_i - x_j\|^2)}{\sum_{k \neq j} \exp(-\gamma^2 \|x_k - x_j\|^2)}$$

- search for representation in $\mathbb{R}^2$, y_1, ..., y_n with a similarity between points in the new representation based on:

$$q_{i|j} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq j} \exp(-\|y_k - y_j\|^2)}$$

- based on the minimization KL divergence between p and q

But: the objective function is not convex and the results are very sensitive to γ (perplexity) and to the initialization

t-SNE: remarks

- t-SNE is good at representing local distances but not global ones (non linear dimension reduction)
- the perplexity can change a lot the representation (no good values found for this dataset)
- the population of cells seem **very homogeneous** and **not related to the genotype**
  (the same is observed on PCA projection)

How could we improve that? Use log / raw expression, base the algorithm on PCA results, try a wider range of perplexity values...?

Clustering

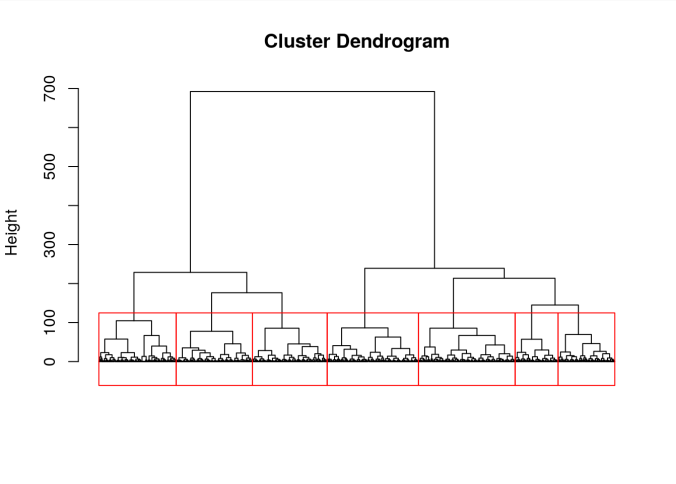
- extract t-SNE coordinates
- use HAC on those

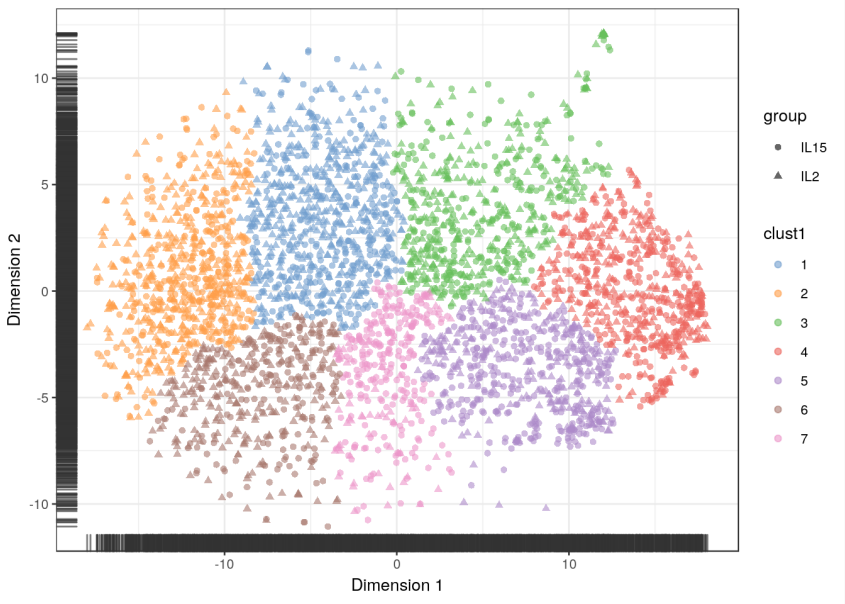
Other approaches for clustering

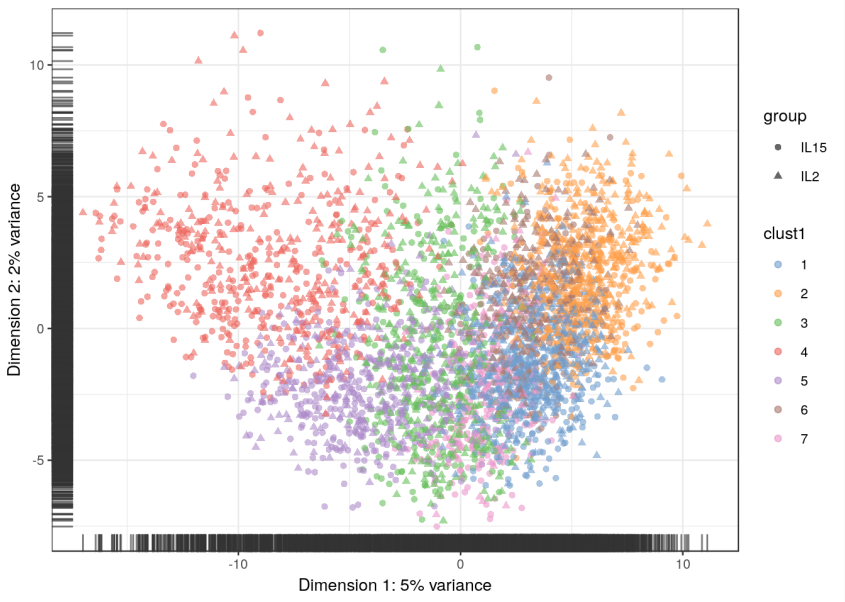
- use a NN network + clustering of graph (Louvain algorithm that optimizes the modularity)
- use other dimension reduction methods and perform any clustering algorithm

$\Rightarrow$ results are different (visualization can even be extremely different)

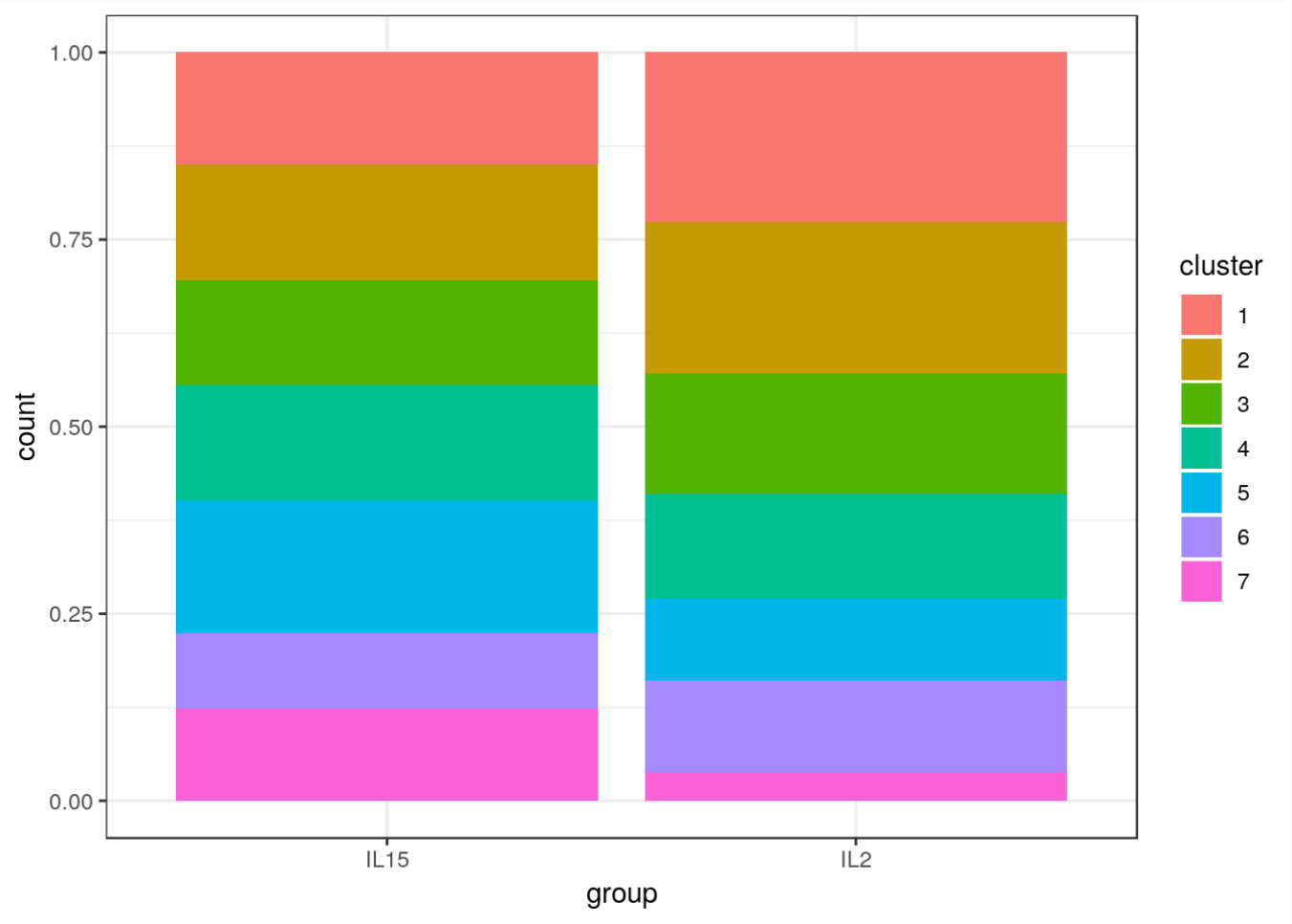
Clustering results



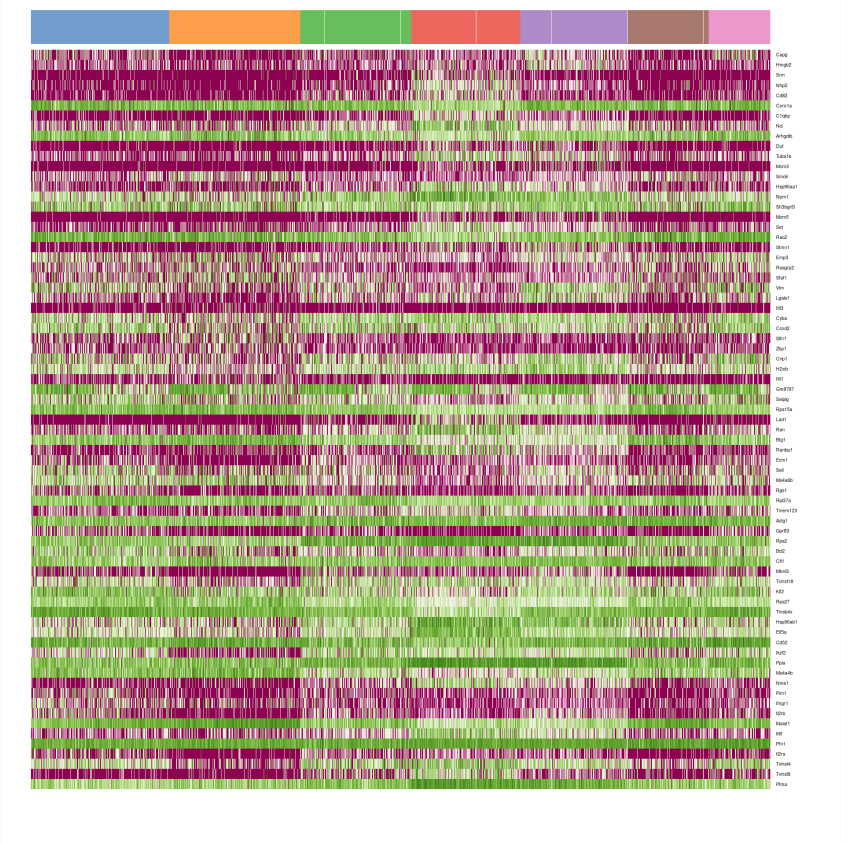




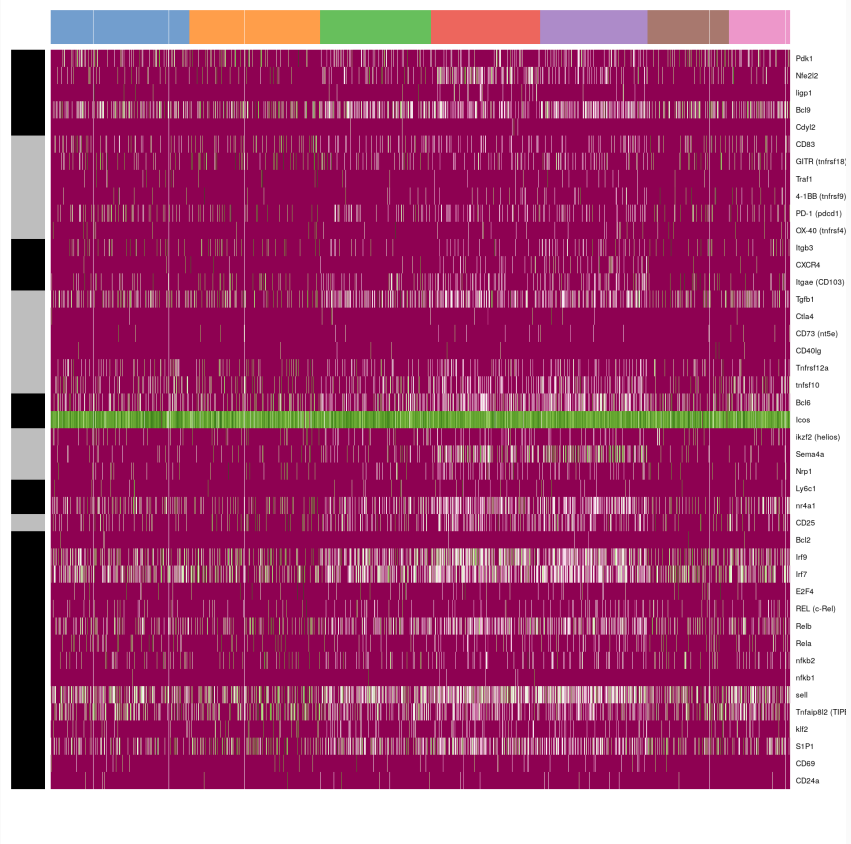
Conditions in clusters



Exploratory analysis of markers

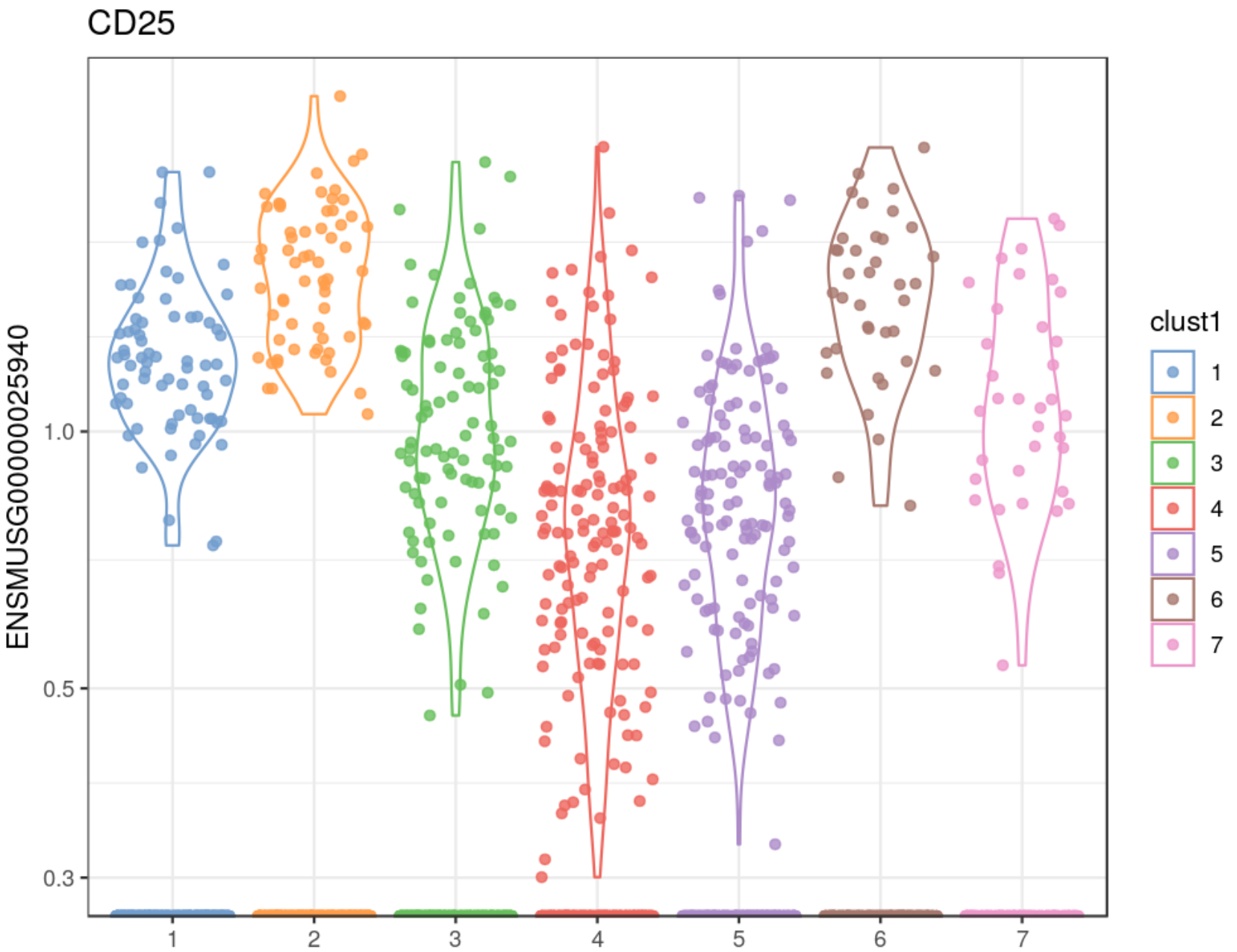


automatic detection



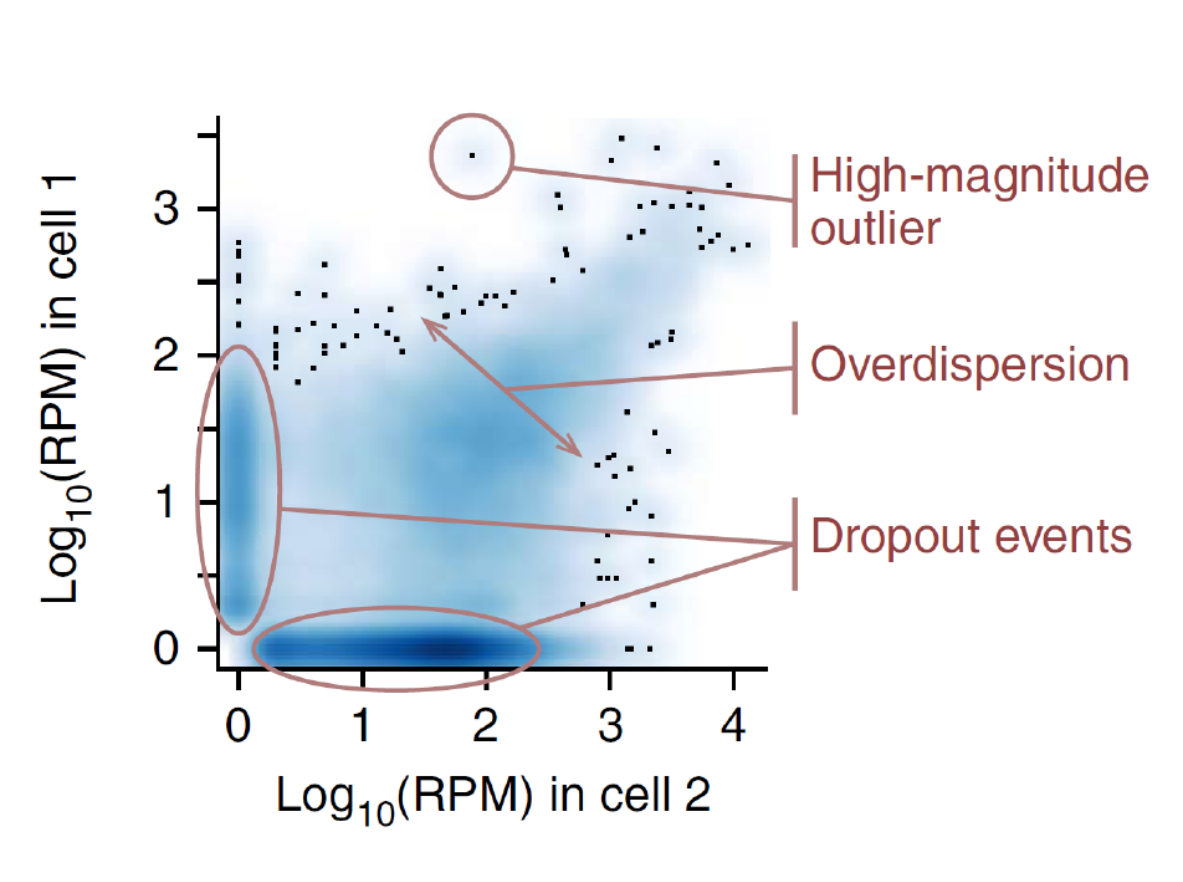
prior knowledge

Not too bad for some known markers...



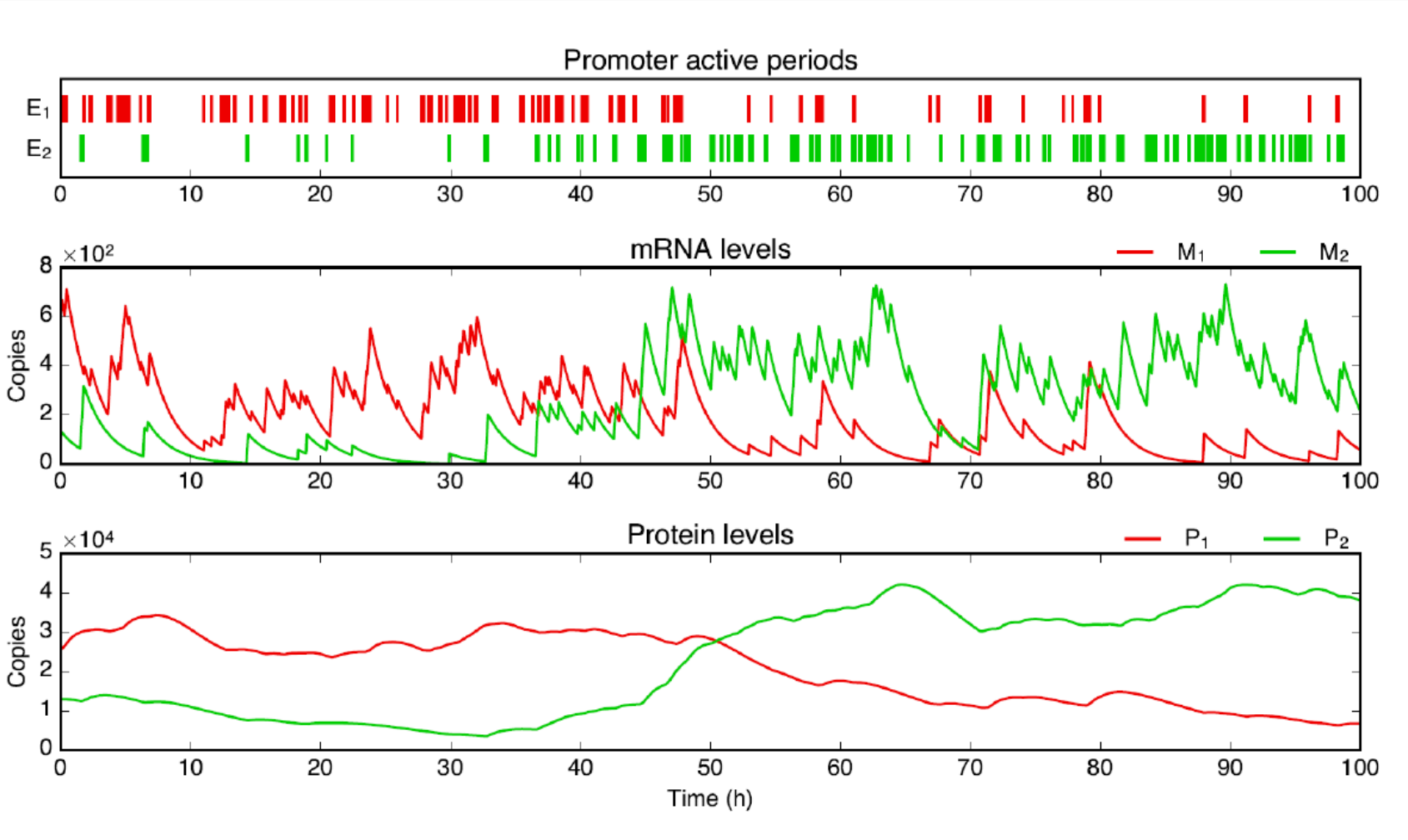
General overview of sc models in statistics

How bad is the situation in single cell data?



Overdispersion is mainly biological because diversity is high between cells

Expression is a bursty process: zeros are biological



sc Differential Expression Analysis with ZINB

[Risso et al., 2018] - package `zinbwave`

For cell i, gene j in condition r, gene expression is modeled by:

$$X_{ijr} \sim \pi_{ijr}\delta_0 + (1 - \pi_{ijr})NB(\mu_{ijr})$$

Remaining problems:

- We are not really able to discriminate low expression from no expression
- Estimation is hard (use of a Bayesian framework to address this issue)
- a similar method exists for PCA [Durif et al., 2018]

References

Durif, G., Modolo, L., Mold, J., Lambert-Lacroix, S., and Picard, F. (2018).
Probabilistic count matrix factorization for single cell expression data analysis.
In Raphael, B. J., editor, *Proceedings of Research in Computational Biology (RECOMB 2018)*, volume 10812 of *Lecture Notes in Computer Science*, pages 254–255, Paris, France. Springer.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein