Analysing single cell RNAseq

Laurent Modolo

March 16 2018

Table of Contents

1 single-cell RNA sequencing

2 Counting molecules

3 Cell quality control

4 Normalization

5 Dimension reduction

6 Differential expression analysis

7 Clustering

Table of Contents

1 single-cell RNA sequencing

2 Counting molecules

3 Cell quality control

4 Normalization

5 Dimension reduction

6 Differential expression analysis

7 Clustering

classical RNASeq (bulk RNASeq)

For each gene or transcript :

- We have a number of reads
- We can estimate the amount of RNA transcripts

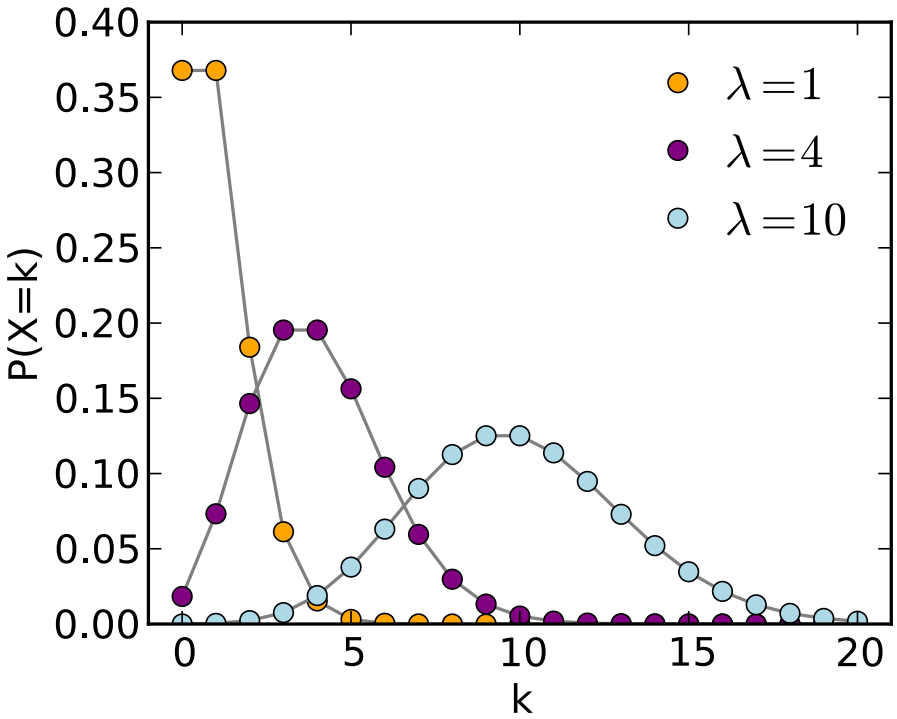
Compared to other RNA measurements we handle counts (Y positive integers)

The natural count distribution is the Poisson distribution.

$$Y_{ig} \sim \mathcal{P}(\lambda_{ig})$$

for the gene g in condition i.

- λ_{ig} RNA transcription rate



[1] **Gierlinski2015**

classical RNASeq (bulk RNASeq)

For each gene or transcript :

- We have a number of reads
- We can estimate the amount of RNA transcripts

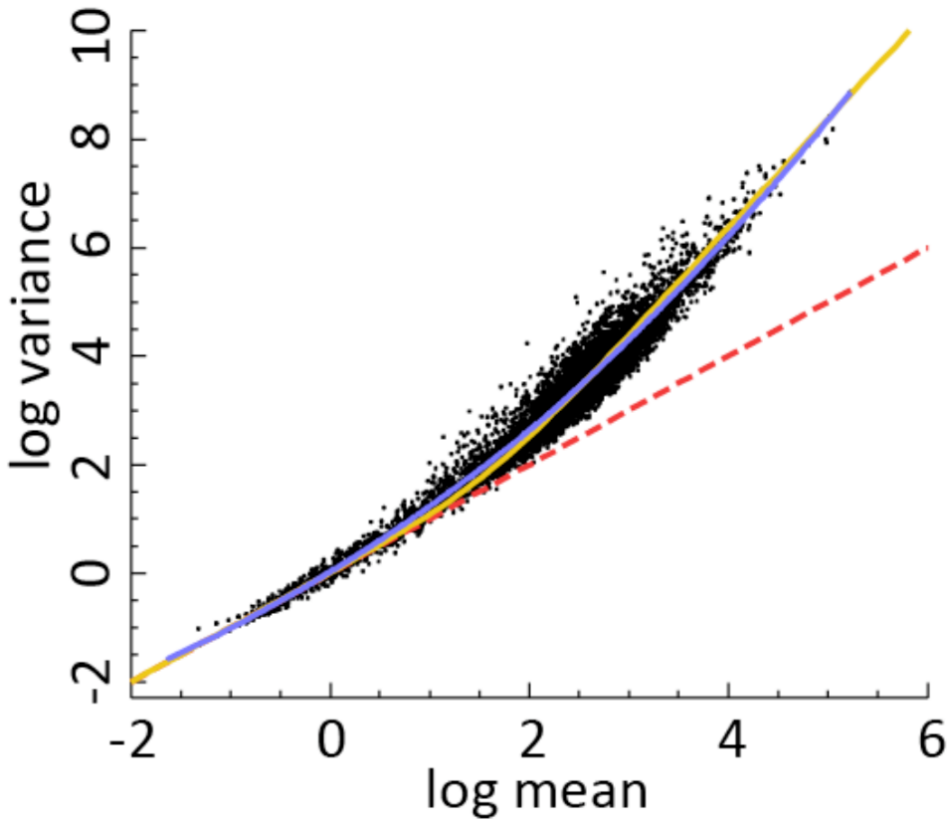
Compared to other RNA measurements we handle counts (Y positive integers)

The natural count distribution is the Poisson distribution.

$$Y_{ig} \sim \mathcal{P}(\lambda_{ig})$$

for the gene g in condition i.

- λ_{ig} RNA transcription rate



[1] **Gierlinski2015**

classical RNASeq (bulk RNASeq)

For each gene or transcript :

- We have a number of reads
- We can estimate the amount of RNA transcripts

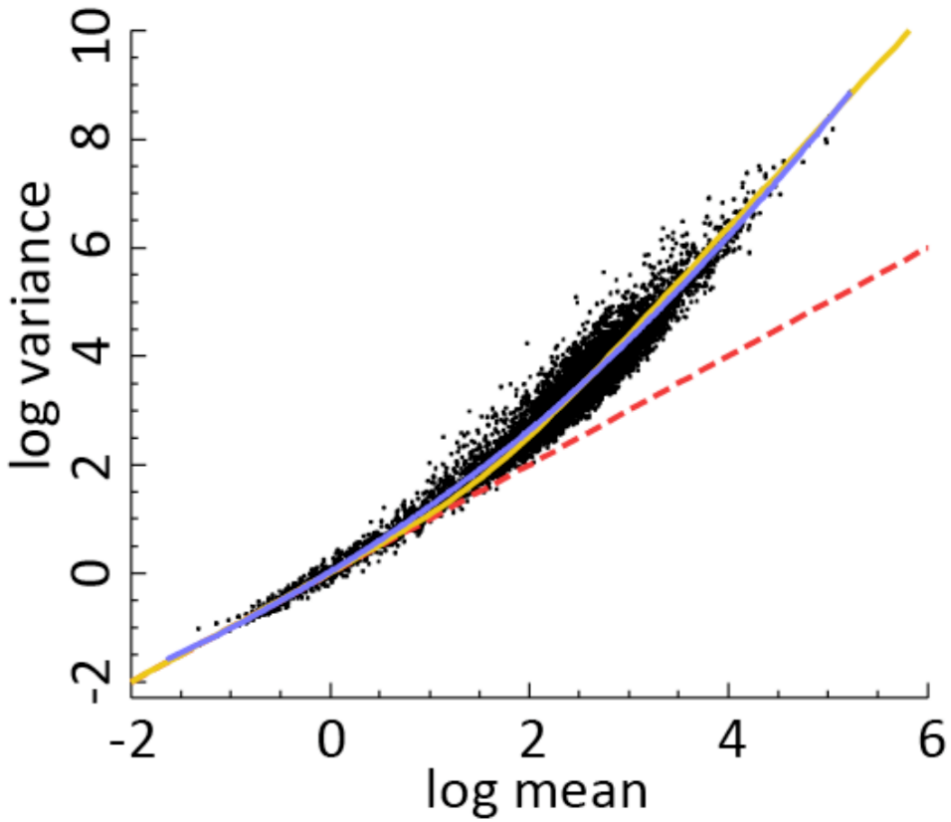
Compared to other RNA measurements we handle counts (Y positive integers)

The NegativeBinomial distribution allows for more or less variability.

$$Y_{ig} \sim NB(\mu_{ig}, \alpha_{ig})$$

for the gene g in condition i.

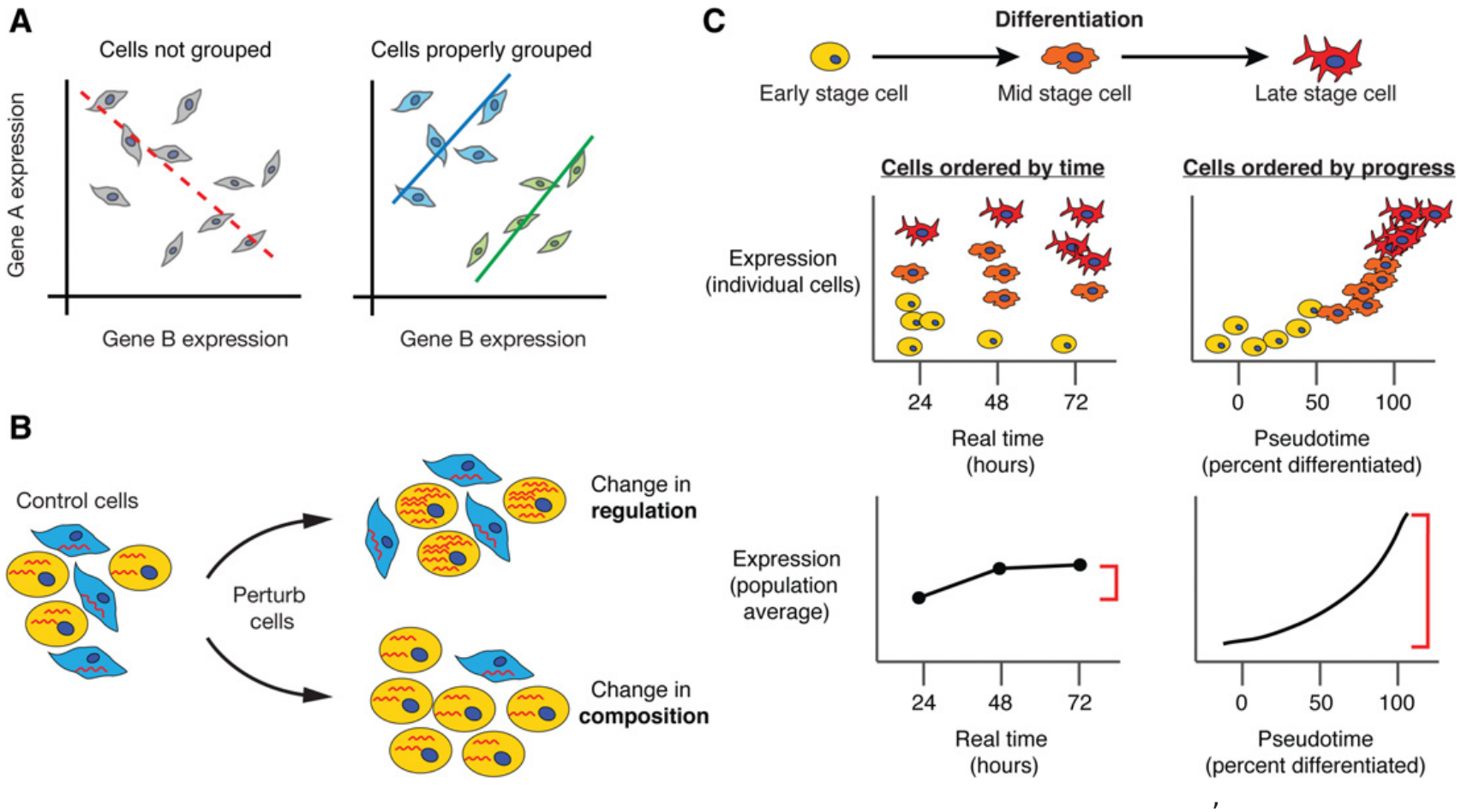
- μ_{ig} RNA transcription rate
- α_{ig} technical or biological variability



[1] **Gierlinski2015**

single-cell RNA sequencing (scRNASeq)

Examples where biological variability stay hidden in bulk RNASeq experiments



[1] **Trapnell2015**

Drawbacks

Compared to bulk RNASeq, we are not working with measurement on a population of cells:

- low starting amount of RNA (zeros)
- need to amplify (more errors)
- transcription status of each gene in each cell (more zeros)

Drawbacks

Compared to bulk RNASeq, we are not working with measurement on a population of cells:

- low starting amount of RNA (zeros)
- need to amplify (more errors)
- transcription status of each gene in each cell (more zeros)

$$\text{scRNASeq} = \text{bulk RNASeq} + \text{zeros} \times 2 + \text{noise}$$

single-cell RNA sequencing (scRNASeq)

The zero-inflated NegativeBinomial distribution accounts for the excess of zeros.

$$Y_{ig} \sim \pi_{ig}\delta_0 + (1 - \pi_{ig})\, NB(\mu_{ig}, \alpha_{ig})$$

for the gene g in condition i.

- μ_{ig} RNA transcription rate
- α_{ig} technical or biological variability
- π the proportion of additional zeros

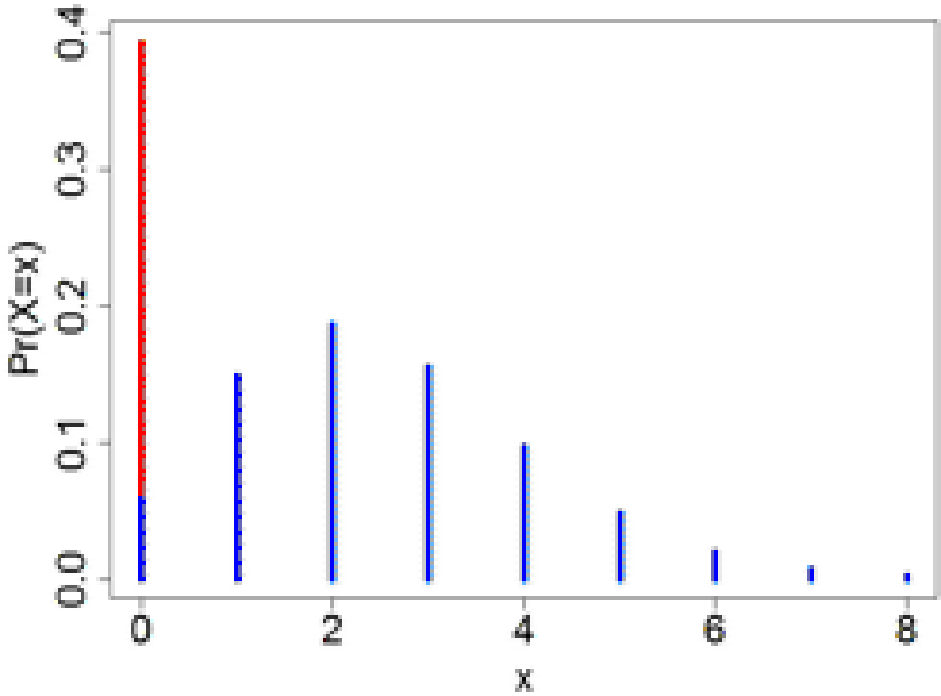


Table of Contents

1 single-cell RNA sequencing

2 Counting molecules

3 Cell quality control

4 Normalization

5 Dimension reduction

6 Differential expression analysis

7 Clustering

Bulk or scRNASeq reads are the same thing

You can use the same quality control and mapping tools

[1] **Smith2017**
[2] **Petukhov2017**
[3] **Bray2016**

Bulk or scRNASeq reads are the same thing

You can use the same quality control and mapping tools

Unique Molecular Identifiers UMI

- control of cDNA amplification
- we have access to the RNA transcripts counts (`dropEst`[1])
- risk of UMI collisions[2]
- alternative splicing?

[1] **Smith2017**
[2] **Petukhov2017**
[3] **Bray2016**

Bulk or scRNASeq reads are the same thing

You can use the same quality control and mapping tools

Unique Molecular Identifiers UMI

- control of cDNA amplification
- we have access to the RNA transcripts counts (`dropEst`[1])
- risk of UMI collisions[2]
- alternative splicing?

Classical RNASeq

- alternative splicing (`Kallisto`[3] with the `pseudo` mode)
- possible cDNA amplification bias
- we work with read counts

No tools to infer the transcript counts (yet)

[1] **Smith2017**
[2] **Petukhov2017**
[3] **Bray2016**

Table of Contents

1 single-cell RNA sequencing

2 Counting molecules

3 Cell quality control

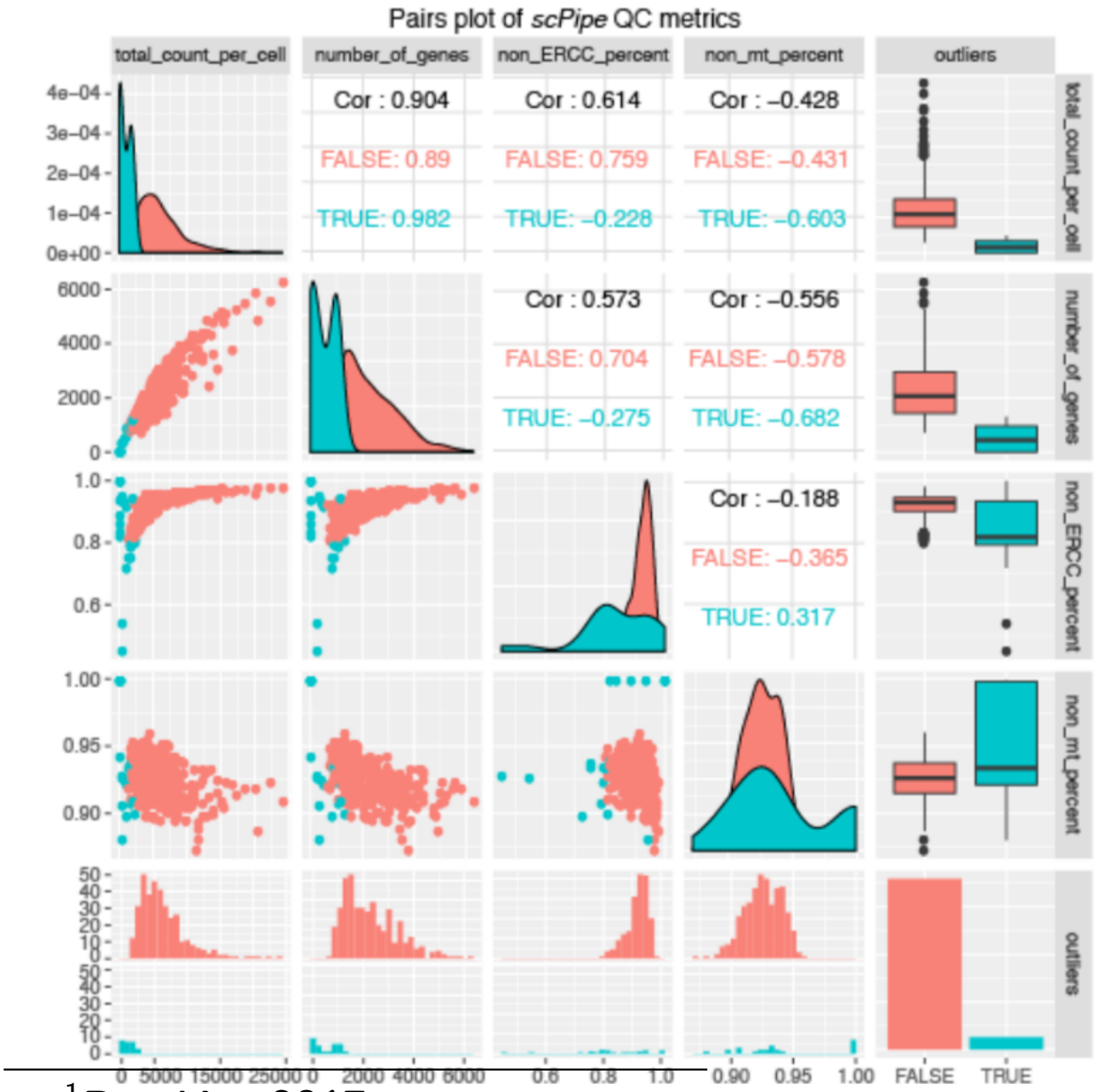
4 Normalization

5 Dimension reduction

6 Differential expression analysis

7 Clustering

Low quality cells must be removed



[1] **Petukhov2017**

Identifying low quality cells:

- detection of outliers
- sequencing of blanks
- ERCC counts

Gaussian mixture model (`mclust`)

Support vector machines (`e1071`)

Table of Contents

1 single-cell RNA sequencing

2 Counting molecules

3 Cell quality control

4 Normalization

5 Dimension reduction

6 Differential expression analysis

7 Clustering

Normalization

bulk RNASeq

- scaling between replicates
- as many replicate as we want
- (batch effects)

scRNASeq

- scaling between cells
- cells can be measured only once (for now)
- 10-60% mRNA capture efficiency
- large number of batches

[1] **Andrews2018**

Normalization

bulk RNASeq

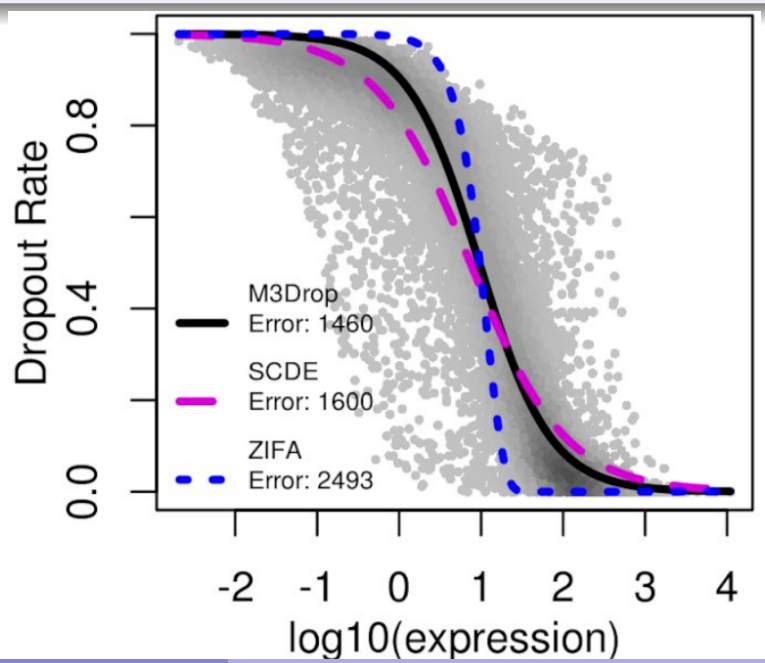
- scaling between replicates
- as many replicate as we want
- (batch effects)

scRNASeq

- scaling between cells
- cells can be measured only once (for now)
- 10-60% mRNA capture efficiency
- large number of batches

Need to normalize for:

- differences between cells
- differences between batches
- differences between genes



[1] **Andrews2018**

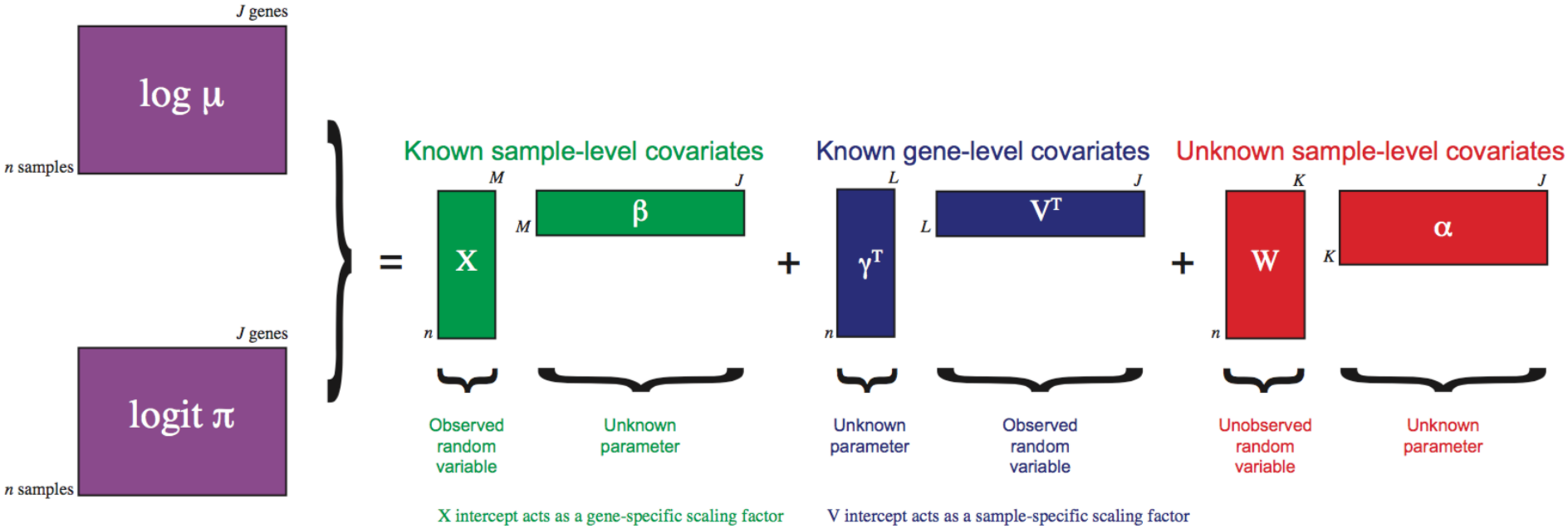
Normalization

- single-cell scaling factor with `SCnorm`[1]
- batch effect `DASC`[2]

[1] **Batcher2017**
[1] **Yi2017**
[3] **Risso2017**
[4] **Li2018**

Normalization

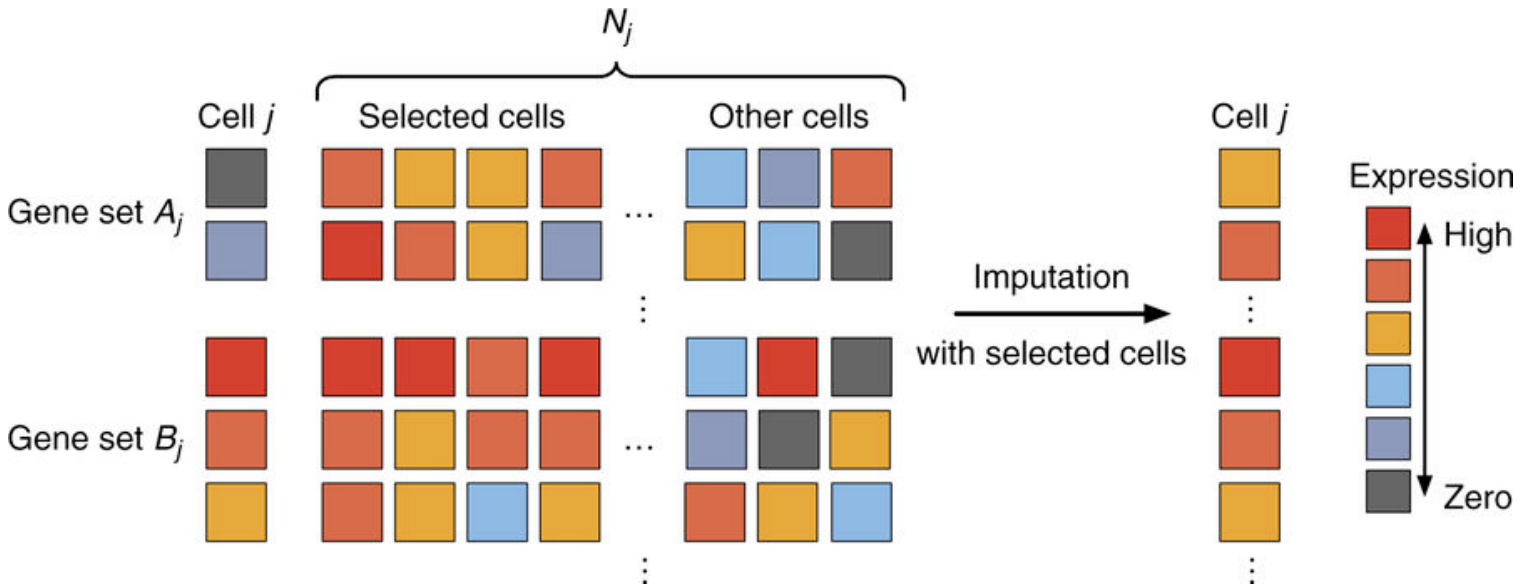
- single-cell scaling factor with `SCnorm`[1]
- batch effect `DASC`[2]
- both `ZINB-WaVe`[3]



[1] **Batcher2017**
[1] **Yi2017**
[3] **Risso2017**
[4] **Li2018**

Normalization

- single-cell scaling factor with `SCnorm`[1]
- batch effect `DASC`[2]
- both `ZINB-WaVe`[3]
- correct dropout `scImpute`[4]



[1] **Batcher2017**
[1] **Yi2017**
[3] **Risso2017**
[4] **Li2018**

Scaling

scaled counts

$$\frac{Y_g}{\exp(\widehat{\alpha_g})} \times (1 - \widehat{\pi_g})$$

- zeros stay zeros
- the more zeros the less the gene will contribute
- we use the empirical dispersion

spread counts

- log-transform: $\log(Y + 1)$
- Anscomb transform: $\sqrt{Y + 3/8}$

Table of Contents

1 single-cell RNA sequencing

2 Counting molecules

3 Cell quality control

4 Normalization

5 Dimension reduction

6 Differential expression analysis

7 Clustering

Dimension reduction

We have a large number of cells and a large number of genes

PCA

Classical PCA on scaled and spread counts

[1] **Risso2017**

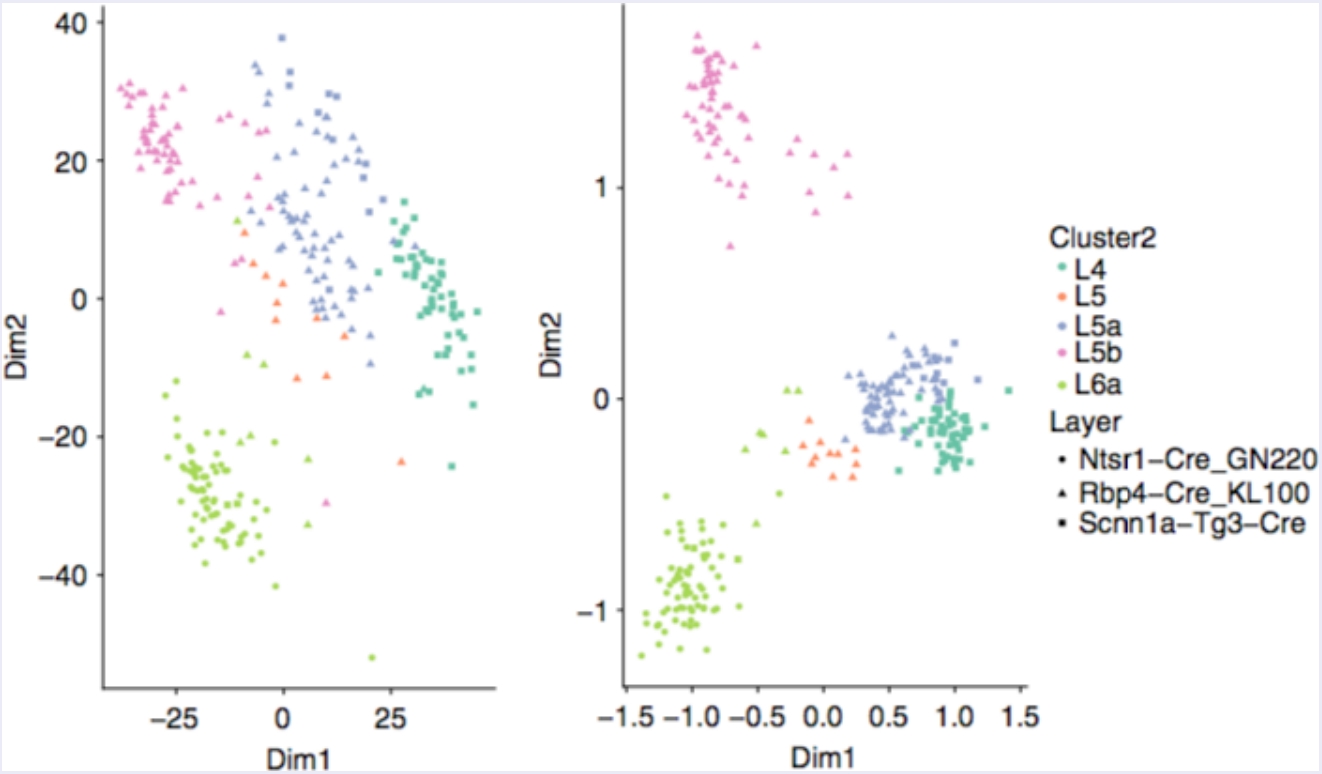
Dimension reduction

We have a large number of cells and a large number of genes

PCA

Classical PCA on scaled and spread counts

ziNB-WaVe



Dimension reduction

We have a large number of cells and a large number of genes

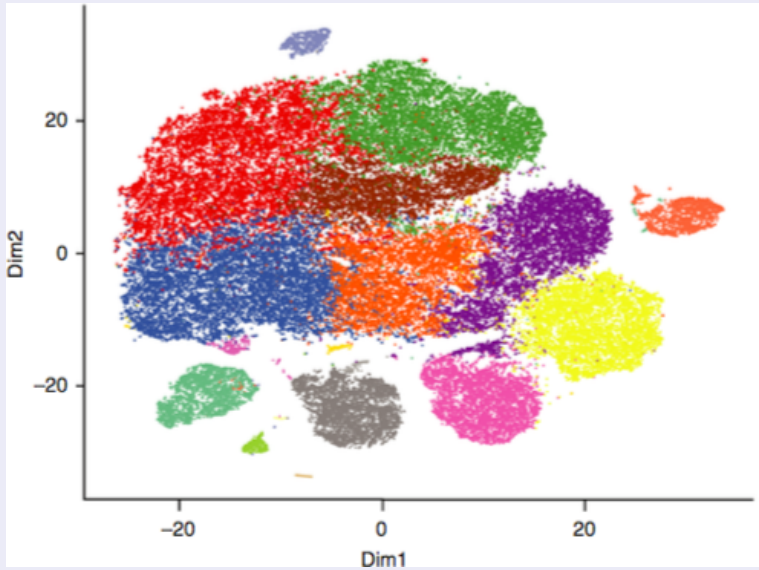
PCA

Classical PCA on scaled and spread counts

ziNB-WaVe

Using W

t-SNE on the above



[1] **Risso2017**

Table of Contents

1 single-cell RNA sequencing

2 Counting molecules

3 Cell quality control

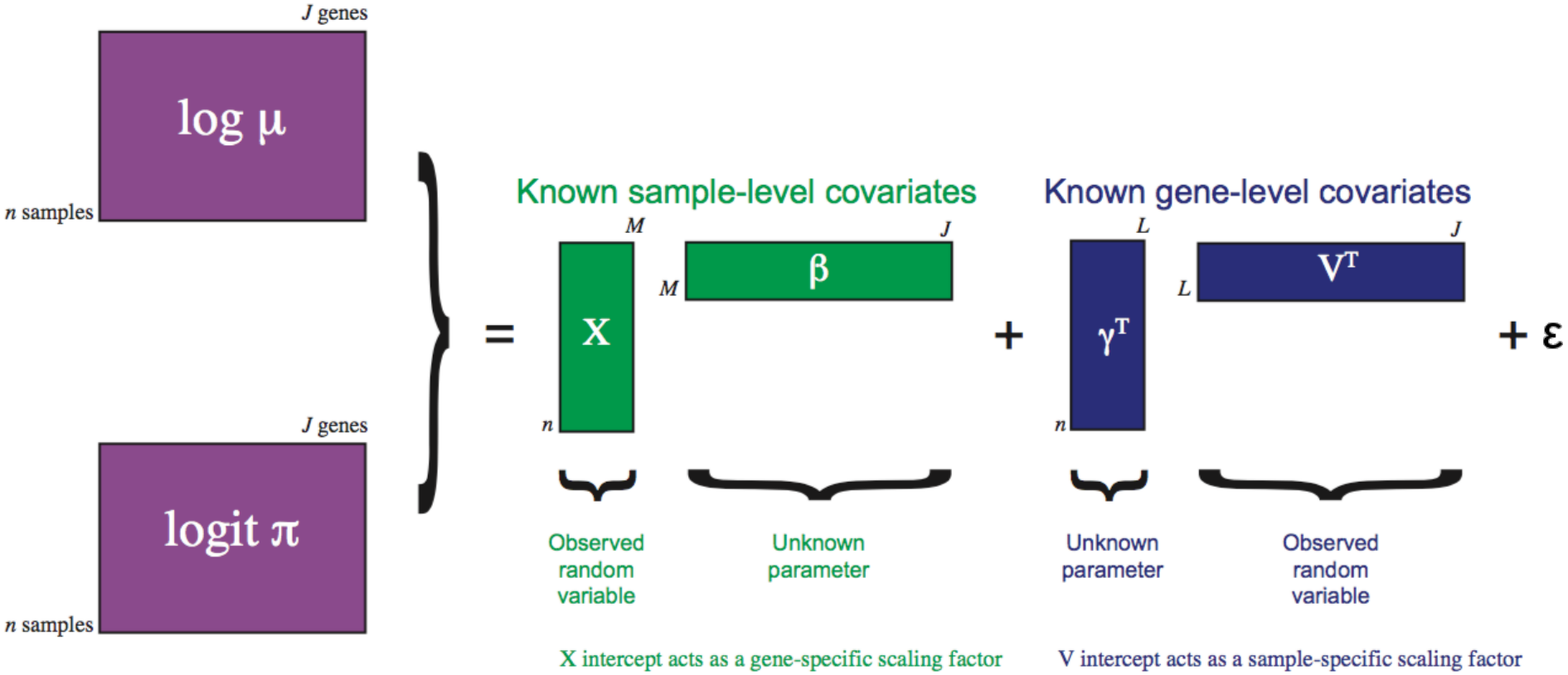
4 Normalization

5 Dimension reduction

6 Differential expression analysis

7 Clustering

Differential expression analysis



- with classical RNASeq tools `zinbwaveZinger`[1]
- with zero-inflated NegativeBinomial GLM `pscl`, `glmmADMB`
- with dropout modelization `M3Drop`[2]

[1] **Risso2018**

[2] **Tallulah2018**

Table of Contents

1 single-cell RNA sequencing

2 Counting molecules

3 Cell quality control

4 Normalization

5 Dimension reduction

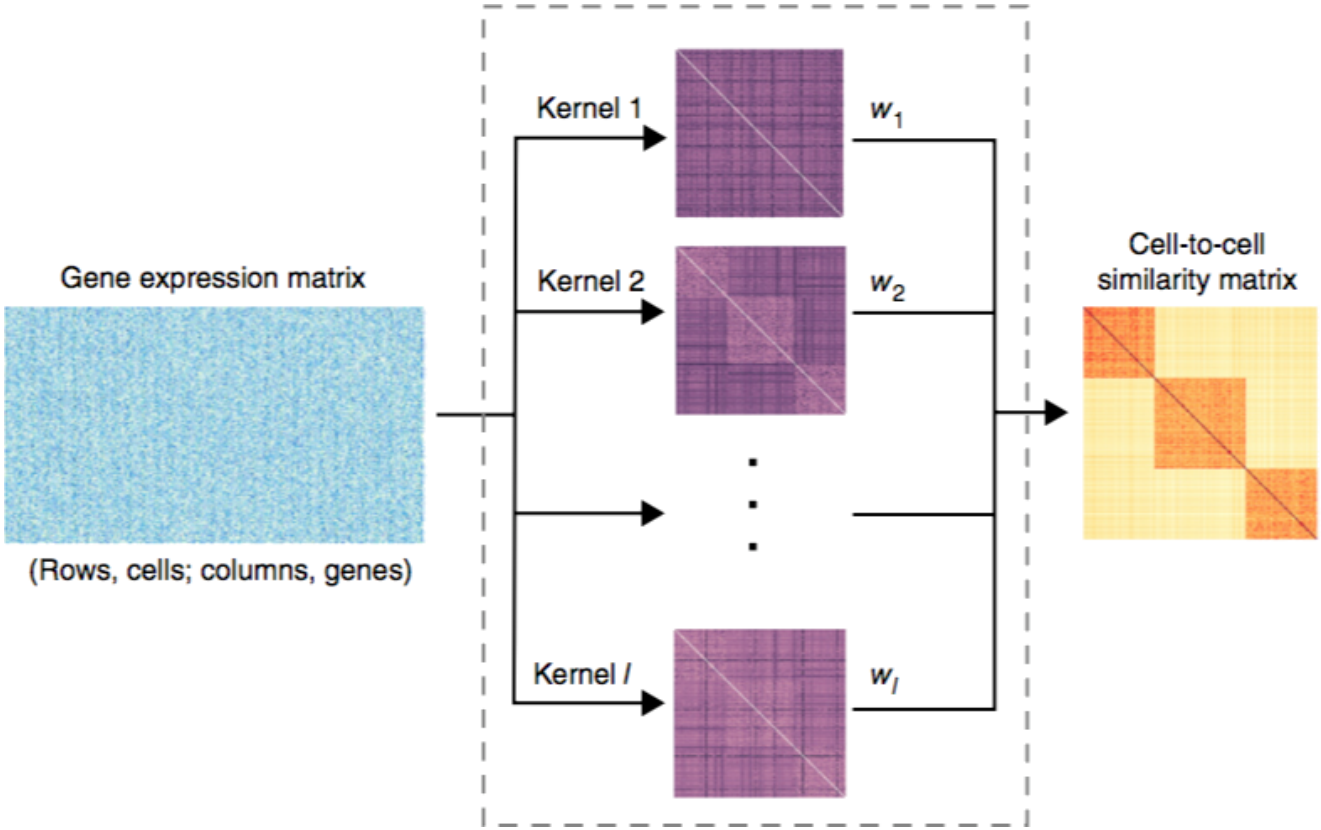
6 Differential expression analysis

7 Clustering

Clustering

Euclidian distance don't work with more than 80% of zeros

- multiple kernel method `SIMLR`[1]
- dropout imputation `scImpute`[2]



[1] **Wang2017**

[2] **Li2018**

Thank you

$$Y_{ig} \sim \pi_{ig}\delta_0 + (1 - \pi_{ig})\, NB(\mu_{ig}, \alpha_{ig})$$

for the gene g in condition i.

More than 200 tools at:
`http://www.scrna-tools.org/`

Tutorials and tools at:
`https://github.com/seandavi/awesome-single-cell`