# Gene expression clustering for single-cell RNA sequencing data

C. Réda
supervised by G. Ilsley & N. Luscombe

OIST, **Genomics and Regulatory Systems** Unit, Japan

March, $1^{st}$ - July, $31^{st}$

# My internship from March, 1$^{st}$ to July, 31$^{st}$


Figure: from **Google Maps & OIST website**

Okinawa Institute of Science and Technology (OIST)

# My internship from March, 1$^{st}$ to July, 31$^{st}$


Figure: from **Google Maps & OIST website**

Okinawa Institute of Science and Technology (OIST)
**Genomics and Regulatory Systems (Luscombe Unit)**

# Projects in the Luscombe Unit

- Study of *Ciona intestinalis*.

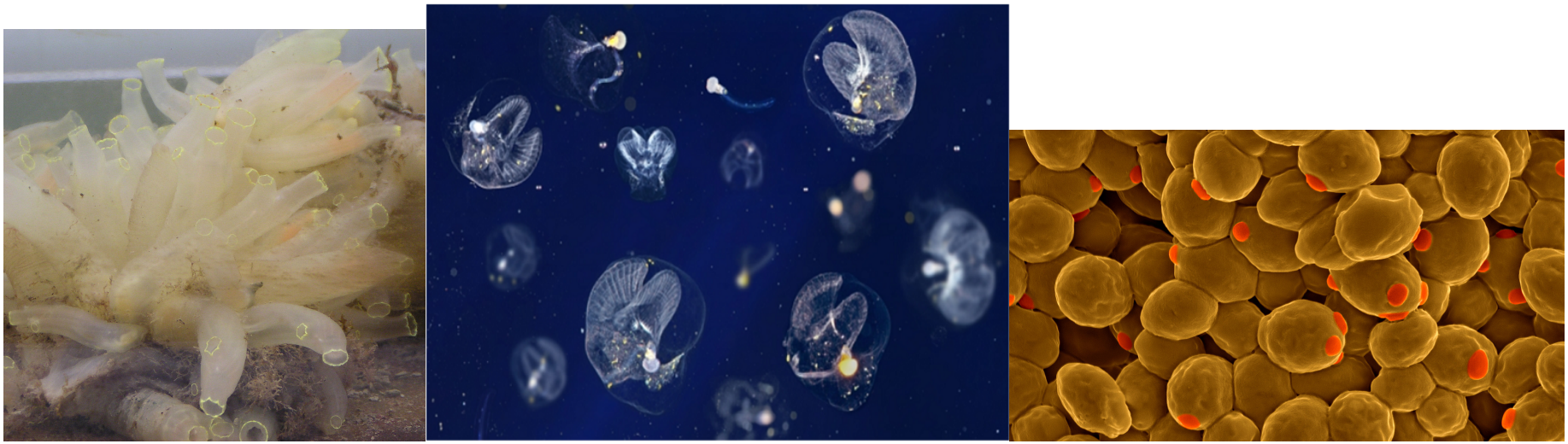


Figure: Ciona **(Wikipédia)** & Oikopleura **(OikoBase)** & Yeast **(NPR)**

# Projects in the Luscombe Unit

- Study of *Ciona intestinalis*.
- Culture of *Oikopleura dioica*.

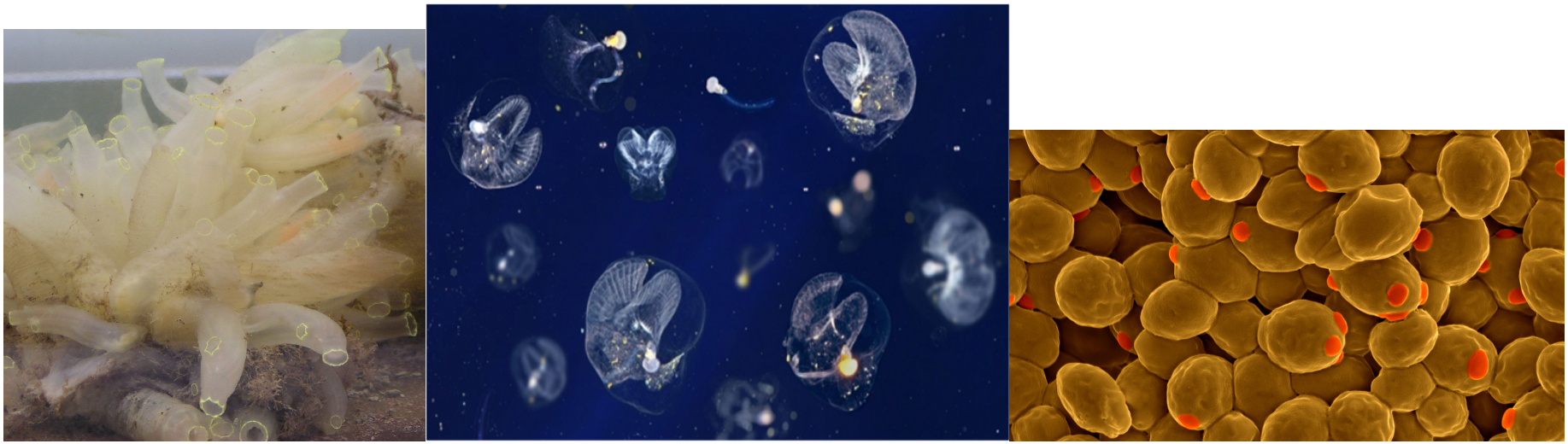


Figure: Ciona **(Wikipédia)** & Oikopleura **(OikoBase)** & Yeast **(NPR)**

# Projects in the Luscombe Unit

- Study of *Ciona intestinalis*.
- Culture of *Oikopleura dioica*.
- Research on the fly and the yeast, etc.

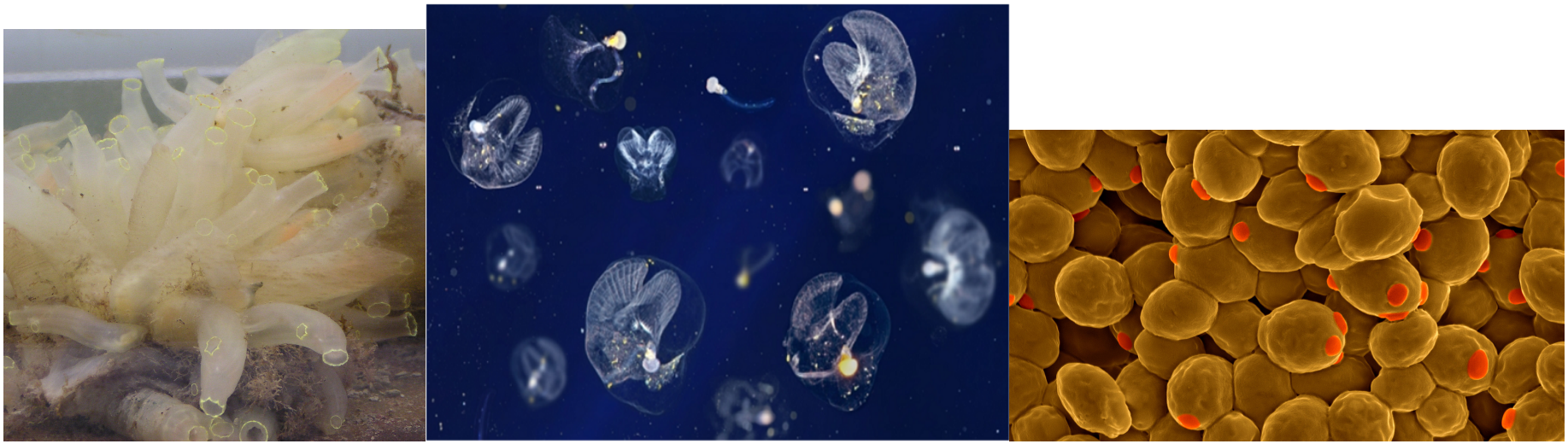
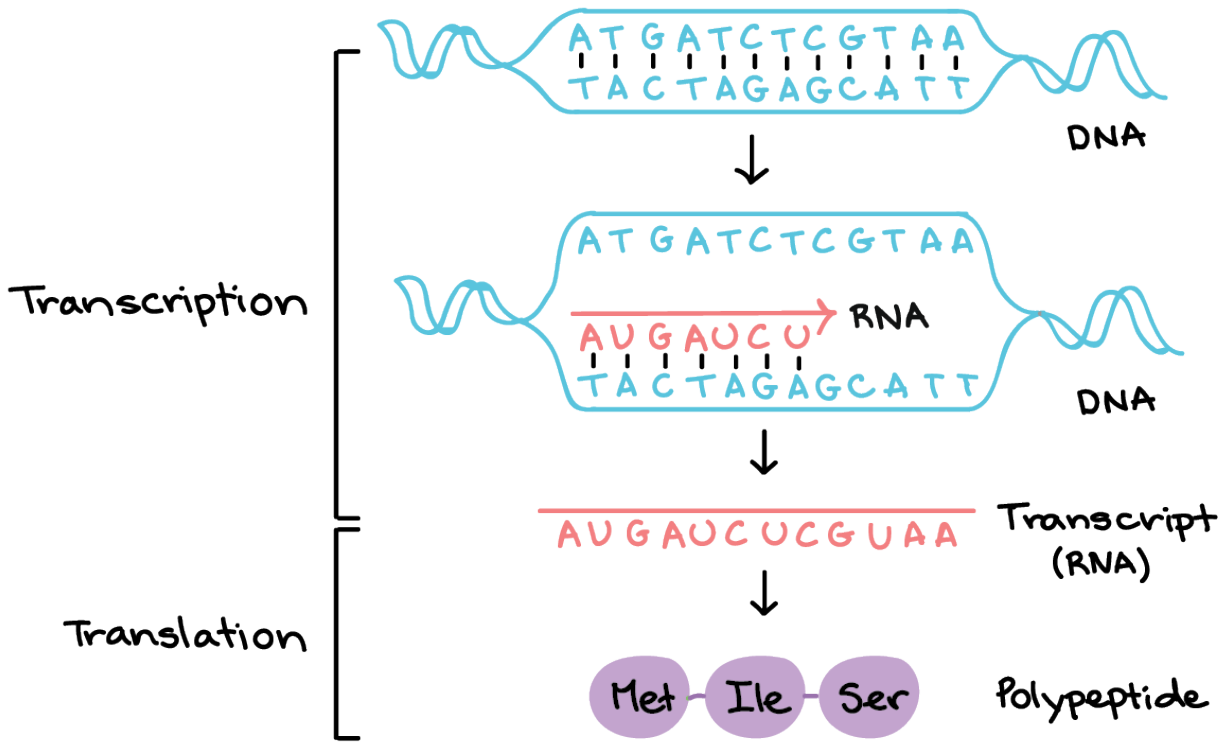


Figure: Ciona **(Wikipédia)** & Oikopleura **(OikoBase)** & Yeast **(NPR)**

# RiboNucleic Acid (RNA)

## Definition

DNA-like molecule that allows **gene expression**, and helps **producing proteins**.



# RNA sequencing

## Single-Cell RNA Sequencing (scRNAseq)

Gets the 4-letter (A, G, U, C) code that controls protein production in a given cell.

# RNA sequencing

## Single-Cell RNA Sequencing (scRNAseq)

Gets the 4-letter (A, G, U, C) code that controls protein production in a given cell.

## Gene expression level

In given sample and gene $g$, count of the **reads** which match with the coding sequence of $g$.

# Gene expression matrix

## Gene expression matrix

For a given set of samples, matrix that contains the gene expression **profiles** (for all genes) for each sample ∼ cell.

| | 1_P0_1 | 2_P0_1 | 4_P0_1 | 5_P0_1 | 6_P0_1 | 1_AB_2 |
|---|---|---|---|---|---|---|
| aap.1 | 45 | 0 | 13 | 0 | 0 | 0 |
| aat.1 | 21 | 98 | 0 | 0 | 8 | 0 |
| aat.2 | 112 | 144 | 260 | 8 | 1 | 258 |
| aat.3 | 0 | 0 | 0 | 0 | 0 | 0 |
| aat.4 | 0 | 0 | 0 | 0 | 0 | 0 |
| aat.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| aat.6 | 0 | 0 | 0 | 0 | 0 | 0 |
| aat.7 | 0 | 0 | 0 | 0 | 0 | 0 |
| aat.8 | 66 | 12 | 20 | 0 | 0 | 20 |
| aat.9 | 0 | 0 | 0 | 0 | 0 | 0 |
| abf.1 | 0 | 0 | 0 | 0 | 0 | 0 |

# Main focus

**Gene expression pattern**

Gene expression profile specific to a given **cell type** $\sim$ **cell functional family**.

a given **gene expression pattern** on a set of **important genes** $\equiv$ a cell function

# Why are gene expression profiles studied?

- To find cell sub-populations in a tumor **[Patel et al., 2014]**.

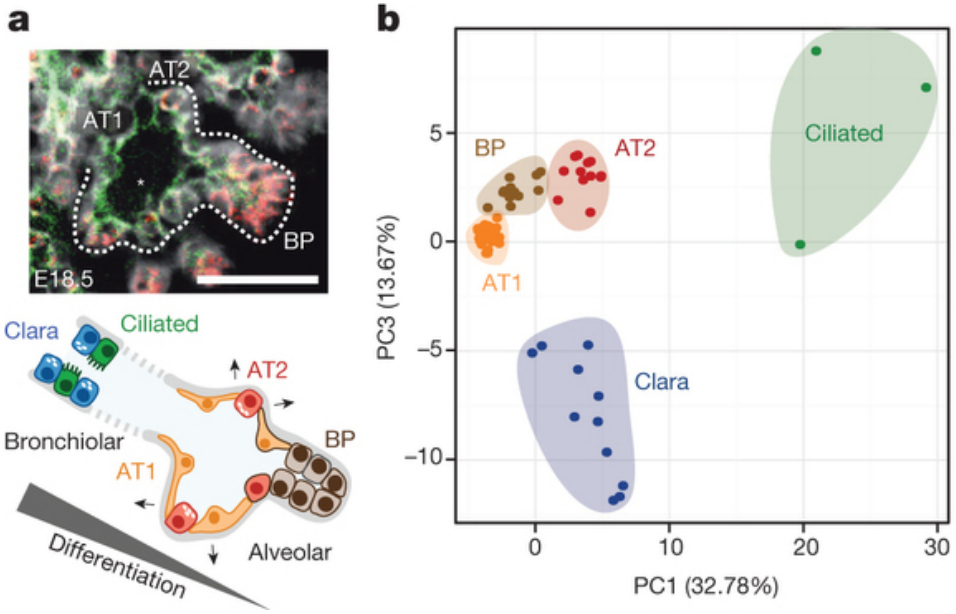



Figure: Mouse cell **clustering** (grouping) from **[Treutlein et al., 2014]**

# Why are gene expression profiles studied?

- To find cell sub-populations in a tumor [Patel et al., 2014].
- To study of the developmental stages of an organism **[Treutlein et al., 2014]**.

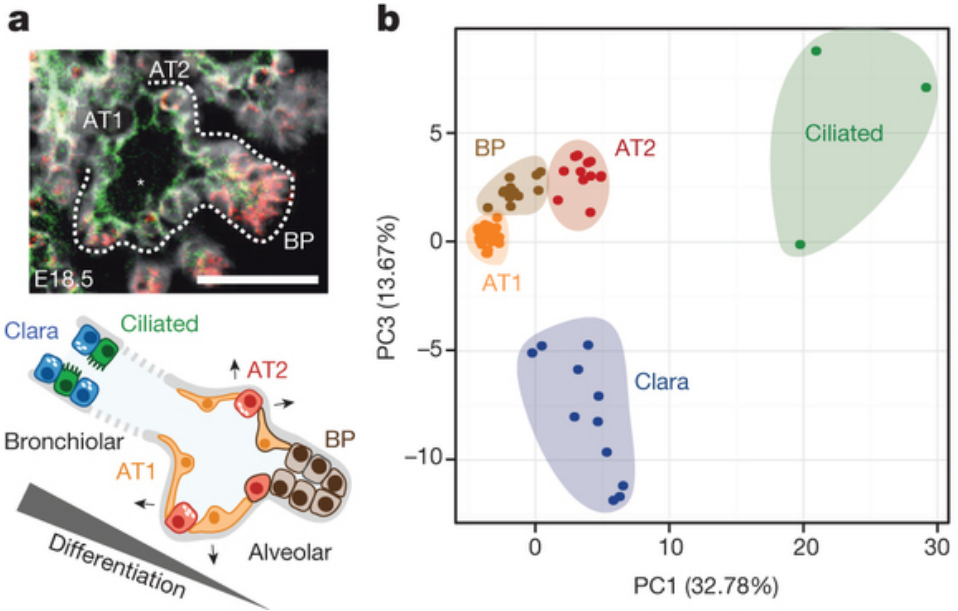



Figure: Mouse cell **clustering** (grouping) from **[Treutlein et al., 2014]**

# Why are gene expression profiles studied?

- To find cell sub-populations in a tumor [Patel et al., 2014].
- To study of the developmental stages of an organism [Treutlein et al., 2014].
- To discover new cell types **[Usoskin et al., 2015]**, etc.

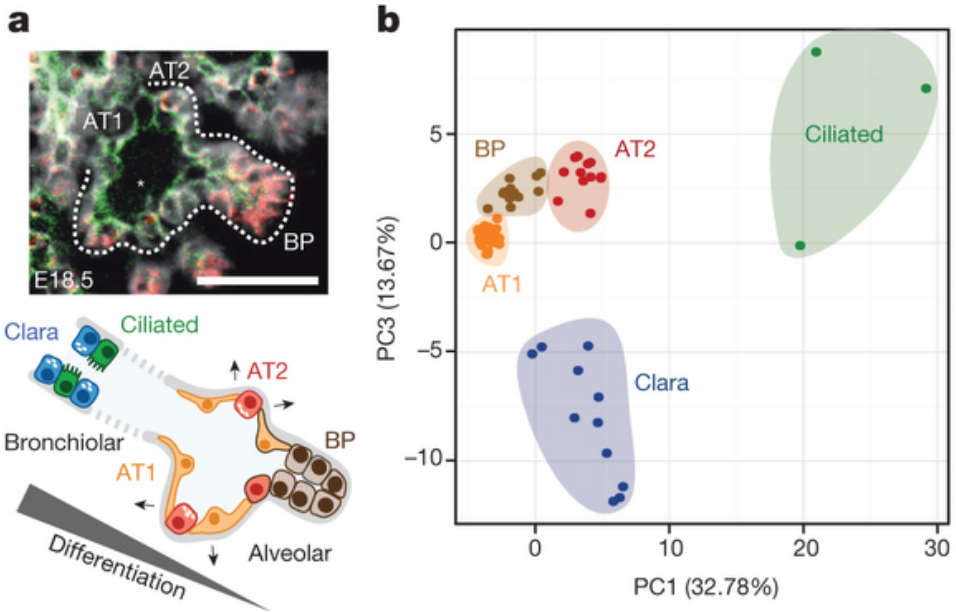



Figure: Mouse cell **clustering** (grouping) from **[Treutlein et al., 2014]**

# Motivation

## My contributions

1. To design a visualization tool for **gene expression profiles**.

# Motivation

## My contributions

1. To design a visualization tool for gene expression profiles.
2. To perform a benchmark on different **cell clustering algorithms**.

# Motivation

## My contributions

1. To design a visualization tool for **gene expression profiles**.
2. To perform a benchmark on different **cell clustering algorithms**.
3. To design a model for single-cell gene expression.

# Demonstration

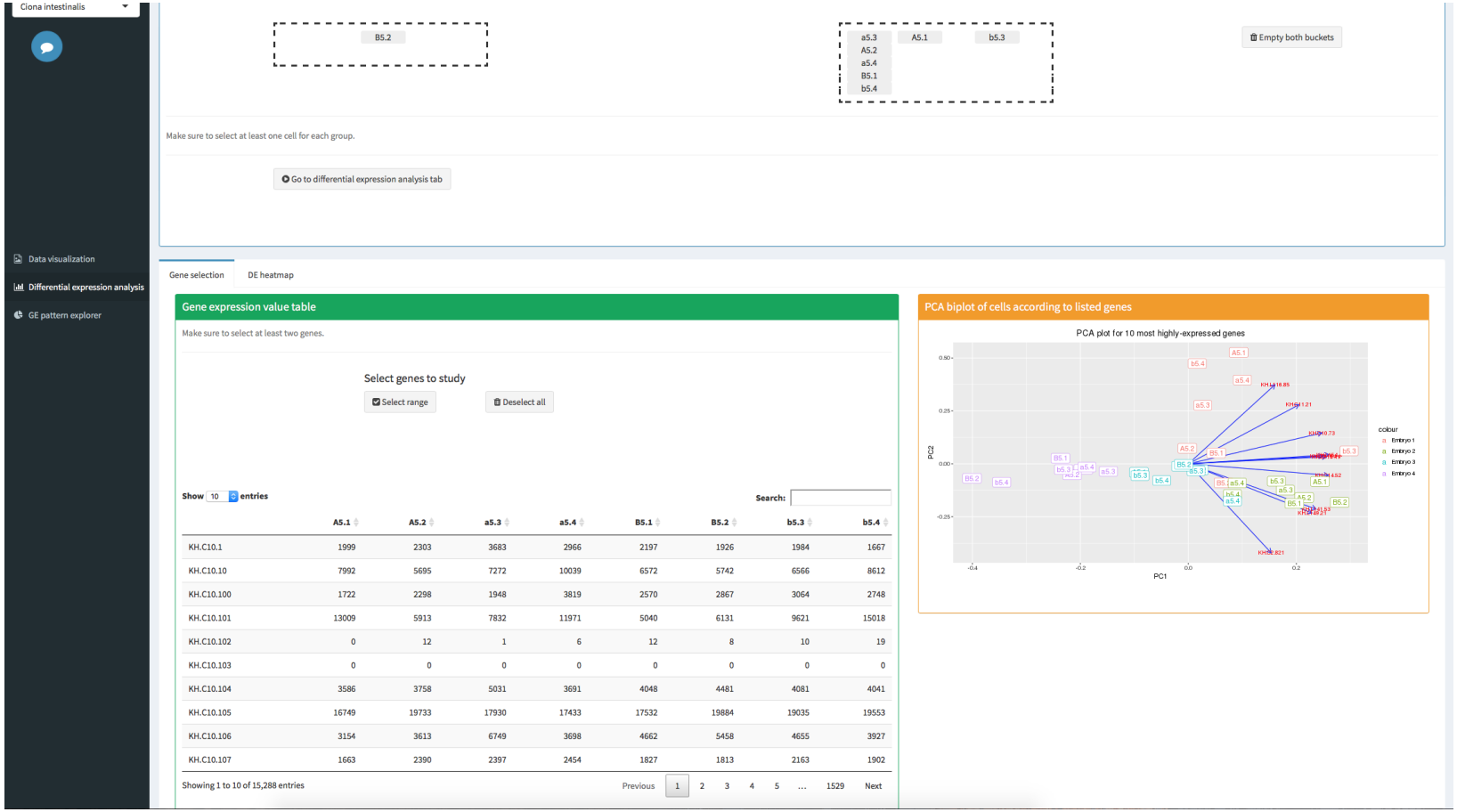



Figure: Screenshot of the application

# Benchmark method (1/2)

Each algorithm has been iterated 100 times on each dataset, with the best parameter values.

- **Accuracy** measure:

## Adjusted Rand Index

Compares a resulting clustering $\mathscr{C}$ to a reference labelling of points $\mathscr{R}$.

1. If $\text{ARI}(\mathscr{C}, \mathscr{R}) = 1$, then $\mathscr{C}$ and $\mathscr{R}$ are similar.

# Benchmark method (1/2)

Each algorithm has been iterated 100 times on each dataset, with the best parameter values.

- **Accuracy** measure:

## Adjusted Rand Index

Compares a resulting clustering $\mathscr{C}$ to a reference labelling of points $\mathscr{R}$.

1. If $\text{ARI}(\mathscr{C}, \mathscr{R}) = 1$, then $\mathscr{C}$ and $\mathscr{R}$ are similar.
2. If $\text{ARI}(\mathscr{C}, \mathscr{R}) = 0$, then the algorithm is not better than the random strategy.

# Benchmark method (1/2)

Each algorithm has been iterated 100 times on each dataset, with the best parameter values.

- **Accuracy** measure:

## Adjusted Rand Index

Compares a resulting clustering $\mathscr{C}$ to a reference labelling of points $\mathscr{R}$.

1. If $\text{ARI}(\mathscr{C}, \mathscr{R}) = 1$, then $\mathscr{C}$ and $\mathscr{R}$ are similar.
2. If $\text{ARI}(\mathscr{C}, \mathscr{R}) = 0$, then the algorithm is not better than the random strategy.
3. If $\text{ARI}(\mathscr{C}, \mathscr{R}) < 0$, then the two clusterings totally differ.

# Examples for the ARI measure (1/2)

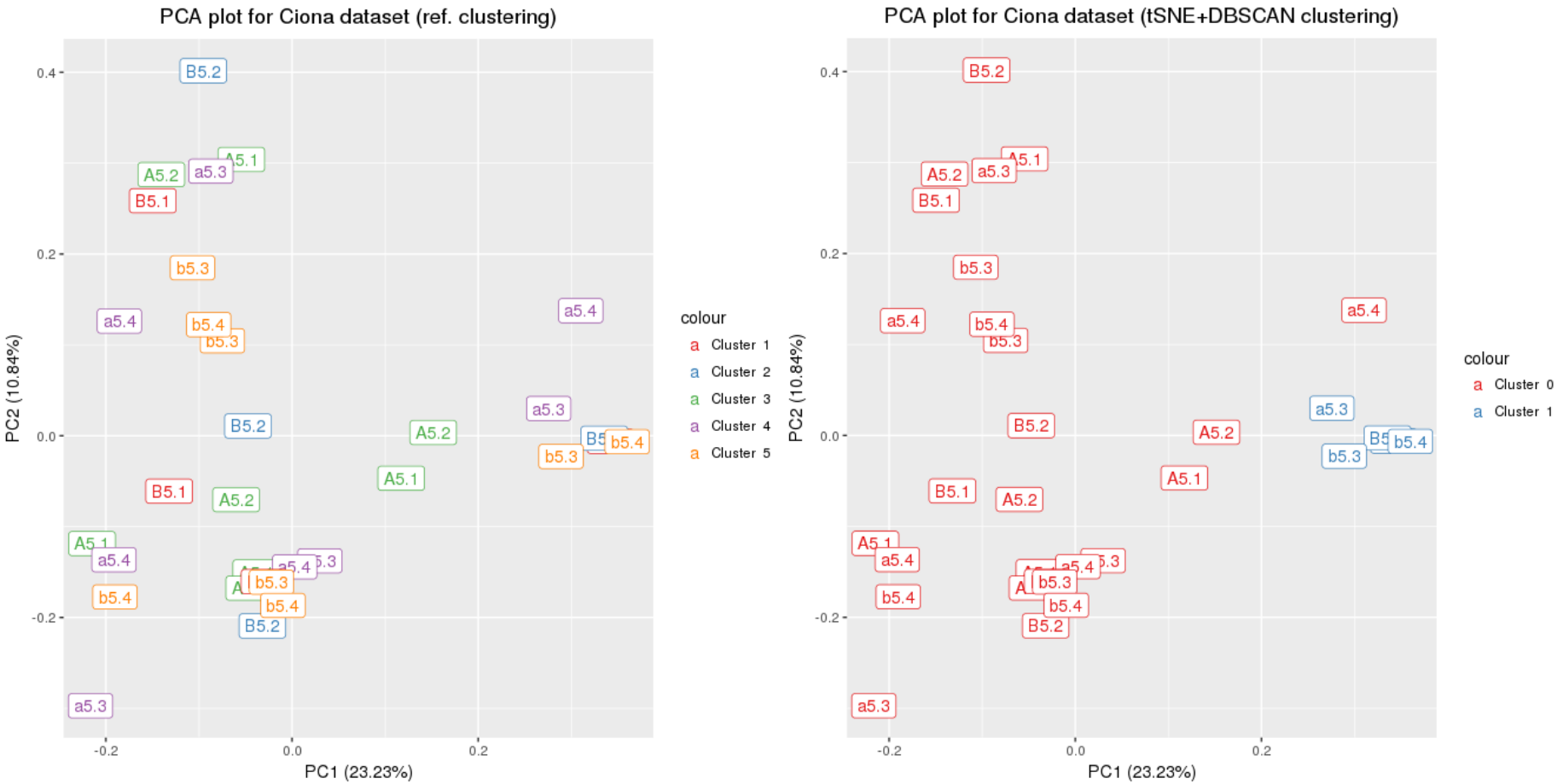



Figure: ARI $\approx$ 0.007 (**[Suyama et al., unp.]** dataset)

# Examples for the ARI measure (2/2)

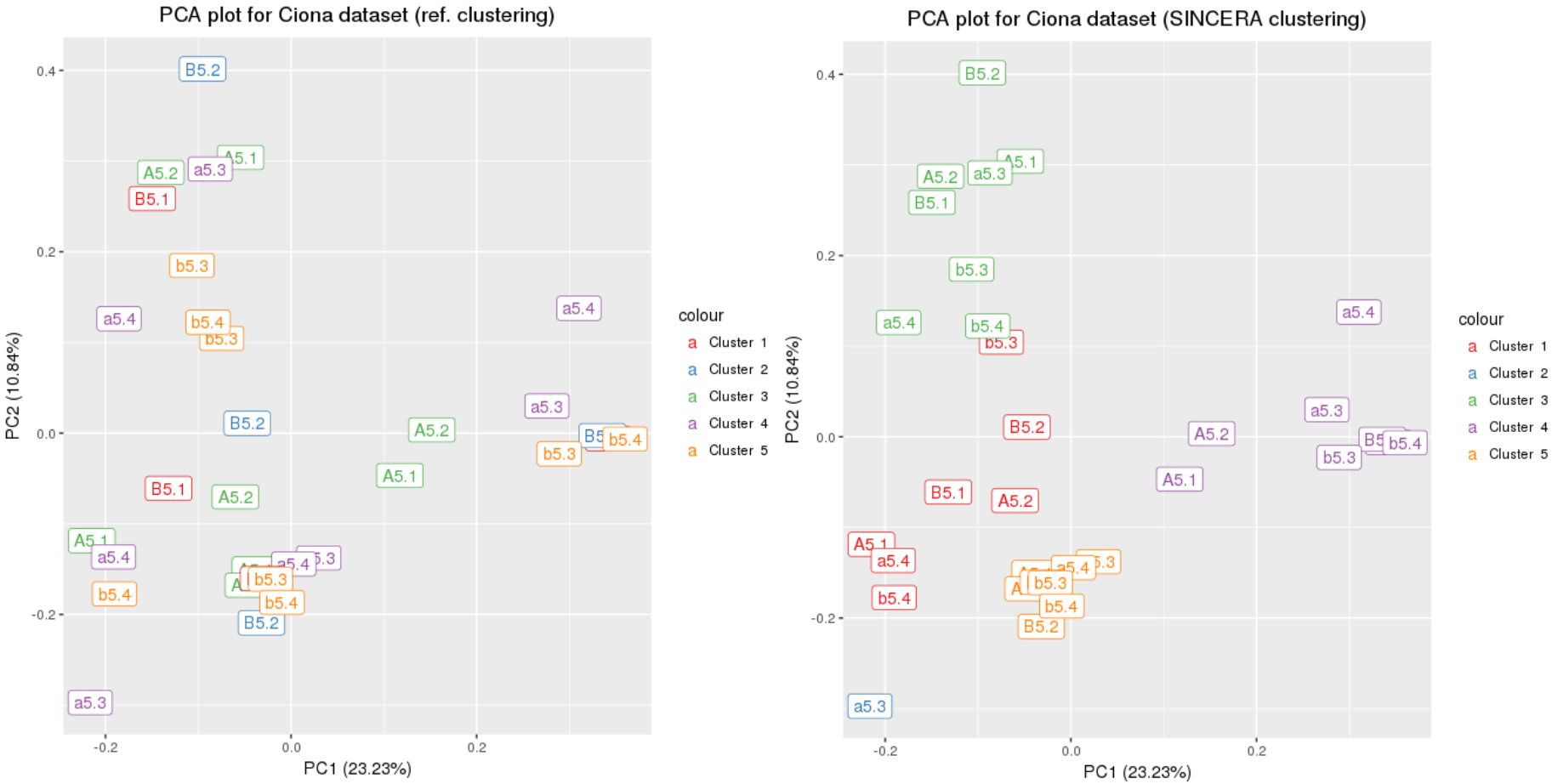


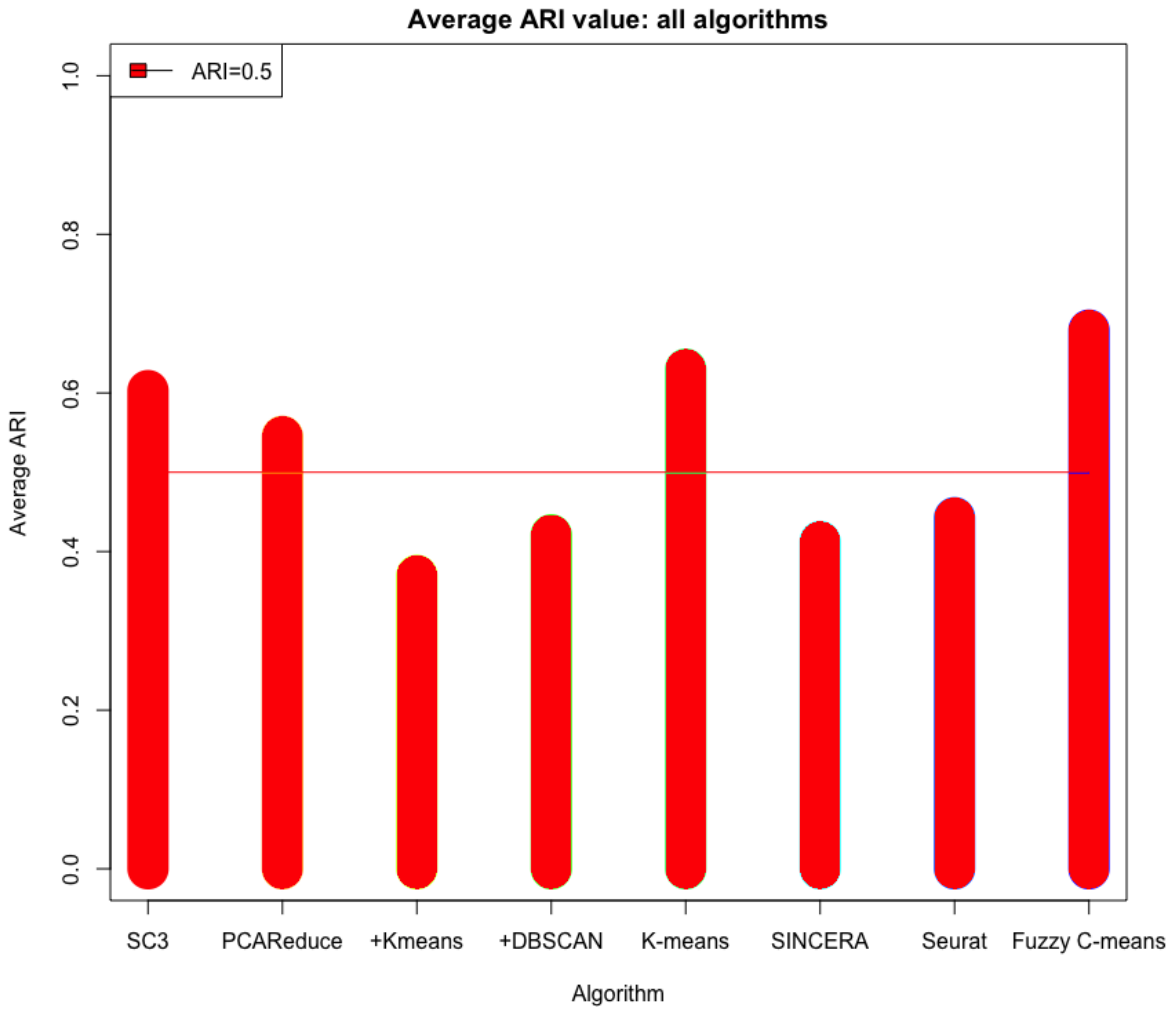

Figure: ARI ≈ -0.12 (**[Suyama et al., unp.]** dataset)

# Benchmark method (2/2)

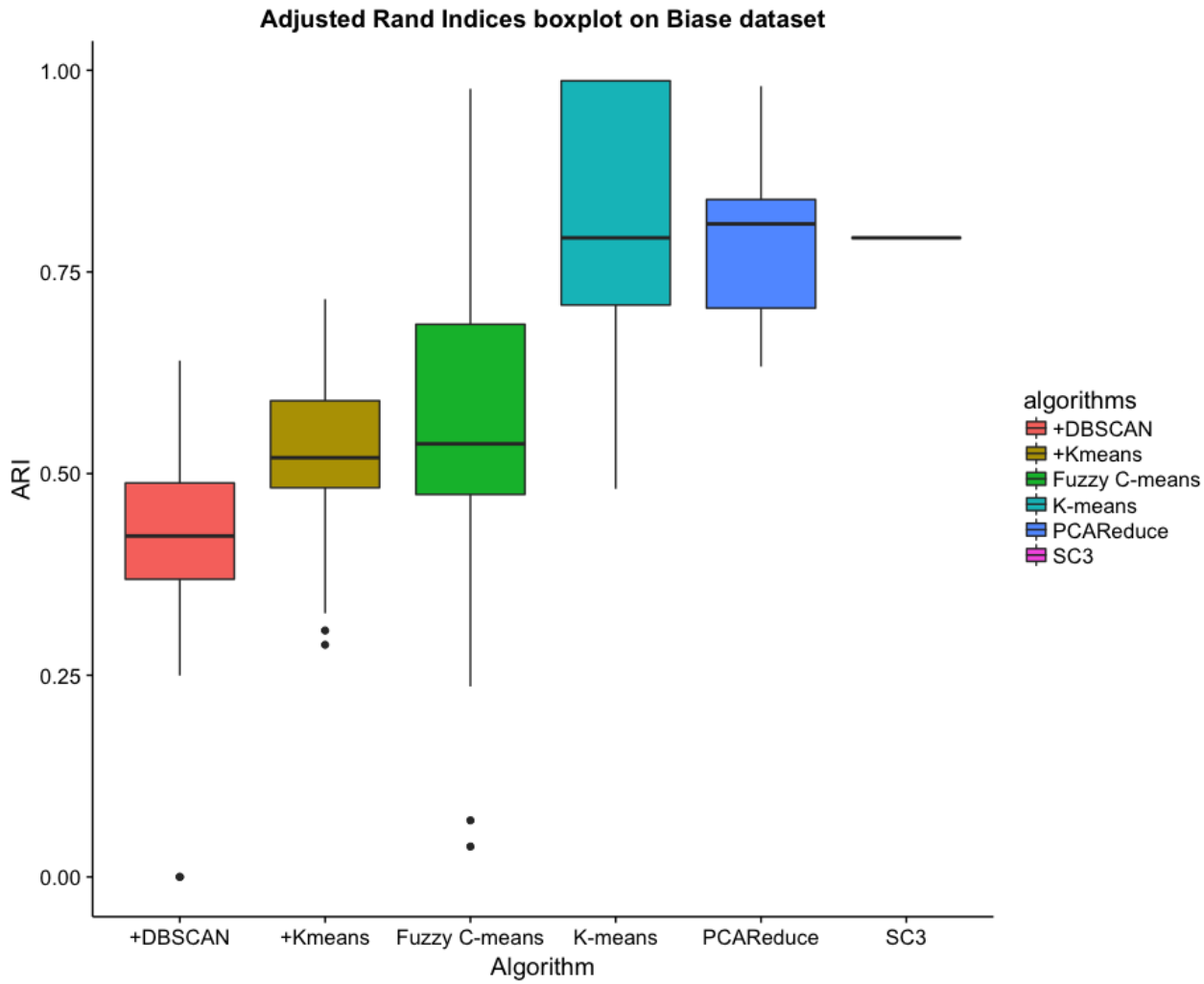
Each algorithm has been iterated 100 times on each dataset, with the best parameter values.

- **Time complexity**: depending on the number of cells and number of genes.
- **Stability**: study of the ARI index variation across all 100 iterations.

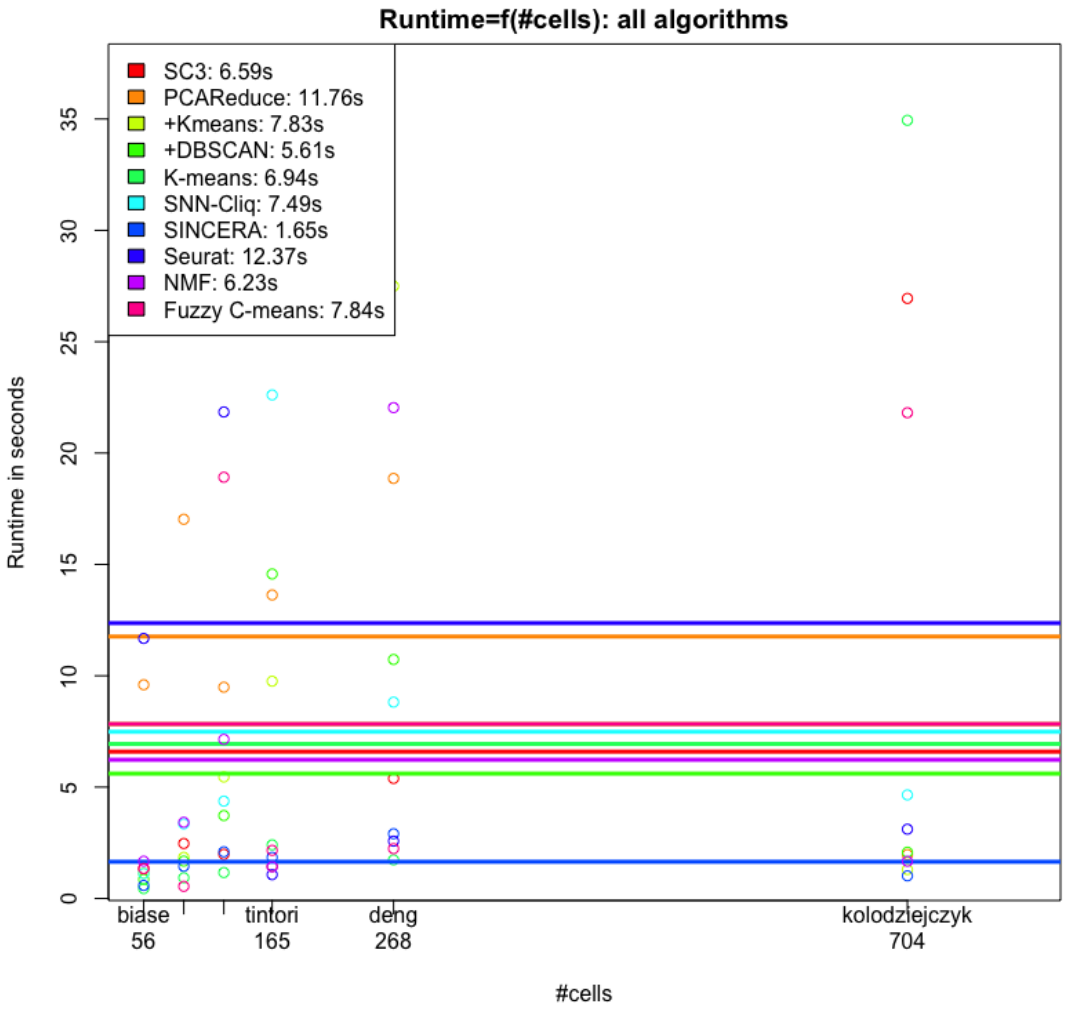
# Benchmark results (1/3)



# Benchmark results (2/3)



# Benchmark results (3/3)



# Characteristics of ***single-cell*** *RNA sequencing* data

1. Lots of **dropout** events **[Lun et al., 2016]**.

# Characteristics of ***single-cell*** *RNA sequencing* data

1. Lots of **dropout** events [Lun et al., 2016].
2. High cell-to-cell variability **[Martinez et al., 2017]**.

# Characteristics of ***single-cell*** *RNA sequencing* data

1. Lots of **dropout** events **[Lun et al., 2016]**.
2. High cell-to-cell variability **[Martinez et al., 2017]**.
3. Log-"normal" model for gene expression **[Finak et al., 2015]**.

# Characteristics of ***single-cell*** *RNA sequencing* data

1. Lots of **dropout** events **[Lun et al., 2016]**.
2. High cell-to-cell variability **[Martinez et al., 2017]**.
3. Log-"normal" model for gene expression **[Finak et al., 2015]**.
4. Bimodal gene expression distribution **[McDavid et al., 2014]**, etc.

# A single cell gene expression model

## Goal

Design a model to pseudo-randomly generate *single cell gene expression data* from a small sample.

# A single cell gene expression model

## Goal

Design a model to pseudo-randomly generate *single cell gene expression data* from a small sample.

## Core idea

- **Gene correlation**.

# A single cell gene expression model

## Goal

Design a model to pseudo-randomly generate *single cell gene expression data* from a small sample.

## Core idea

- Gene correlation.
- A **model for single gene expression** is known.

# A single cell gene expression model

## Goal

Design a model to pseudo-randomly generate *single cell gene expression data* from a small sample.

## Core idea

- Gene correlation.
- A model for single gene expression is known.
- The **whole gene regulation system** is not known.

# A single cell gene expression model

## Goal

Design a model to pseudo-randomly generate *single cell gene expression data* from a small sample.

## Core idea

- Gene correlation.
- A model for single gene expression is known.
- The whole gene regulation system is not known.

# A single cell gene expression model

## Goal

Design a model to pseudo-randomly generate *single cell gene expression data* from a small sample.

## Core idea

- Gene correlation.
- A model for single gene expression is known.
- The whole gene regulation system is not known.

$\Rightarrow$ **Idea:** design separately **single gene expression** and **single cell gene expression**.

# How to link gene-level expression and gene regulation?

## Copula (**Sklar, 1959**)

A multivariate joint CDF $\mathscr{C}$ which margins follow standard uniform distributions.

# How to link gene-level expression and gene regulation?

## Copula (**Sklar, 1959**)

A multivariate joint CDF $\mathscr{C}$ which margins follow standard uniform distributions.

Let $U_1, U_2, U_3, ..., U_p \sim \mathscr{U}_{0,1}$ r. v.:

$$\forall 0 \leq x_1 \leq 1, ..., 0 \leq x_p \leq 1,$$
$$\mathscr{C} : x \to \mathbb{P}(U_1 \leq x_1 \wedge ... \wedge U_p \leq x_p)$$

# Use of copula in modelling

- $(X_i)_{i\in\{1,2,...,p\}}$ is a set of $\mathbb{R}$-valued r. v. with CDF $(\mathscr{F}_i)_{i\in\{1,2,...,p\}}$.
- $\mathscr{F}$ is a joint CDF of $(X_i)_{i\in\{1,2,...,p\}}$, i.e.:

$$\forall x \in \mathbb{R}^p, \mathscr{F}(x_1, ..., x_p) = \mathbb{P}(X_1 \leq x_1, ..., X_p \leq x_p)$$

# Use of copula in modelling

- $(X_i)_{i\in\{1,2,...,p\}}$ is a set of $\mathbb{R}$-valued r. v. with CDF $(\mathscr{F}_i)_{i\in\{1,2,...,p\}}$.
- $\mathscr{F}$ is a joint CDF of $(X_i)_{i\in\{1,2,...,p\}}$, i.e.:

$$\forall x \in \mathbb{R}^p, \mathscr{F}(x_1, ..., x_p) = \mathbb{P}(X_1 \leq x_1, ..., X_p \leq x_p)$$

**Theorem [Sklar, 1959]**

1. Then there is a copula $\mathscr{C}$ such as: $\forall x \in \mathbb{R}$, $\mathscr{F}(x_1, x_2, ..., x_p) = \mathscr{C}(\mathscr{F}_1(x_1), \mathscr{F}_2(x_2), ..., \mathscr{F}_p(x_p))$.

# Use of copula in modelling

- $(X_i)_{i\in\{1,2,...,p\}}$ is a set of $\mathbb{R}$-valued r. v. with CDF $(\mathscr{F}_i)_{i\in\{1,2,...,p\}}$.
- $\mathscr{F}$ is a joint CDF of $(X_i)_{i\in\{1,2,...,p\}}$, i.e.:

$$\forall x \in \mathbb{R}^p, \mathscr{F}(x_1, ..., x_p) = \mathbb{P}(X_1 \leq x_1, ..., X_p \leq x_p)$$

**Theorem [Sklar, 1959]**

1. Then there is a copula $\mathscr{C}$ such as: $\forall x \in \mathbb{R}, \mathscr{F}(x_1, x_2, ..., x_p) = \mathscr{C}(\mathscr{F}_1(x_1), \mathscr{F}_2(x_2), ..., \mathscr{F}_p(x_p))$.
2. Given a copula $\mathscr{D}$, $\mathscr{H} : (\mathbb{R}^+)^p \to [0,1]$, $x \to \mathscr{D}(\mathscr{F}_1(x_1), ..., \mathscr{F}_p(x_p))$ is a CDF.

# Which distributions should be chosen? (1/4)

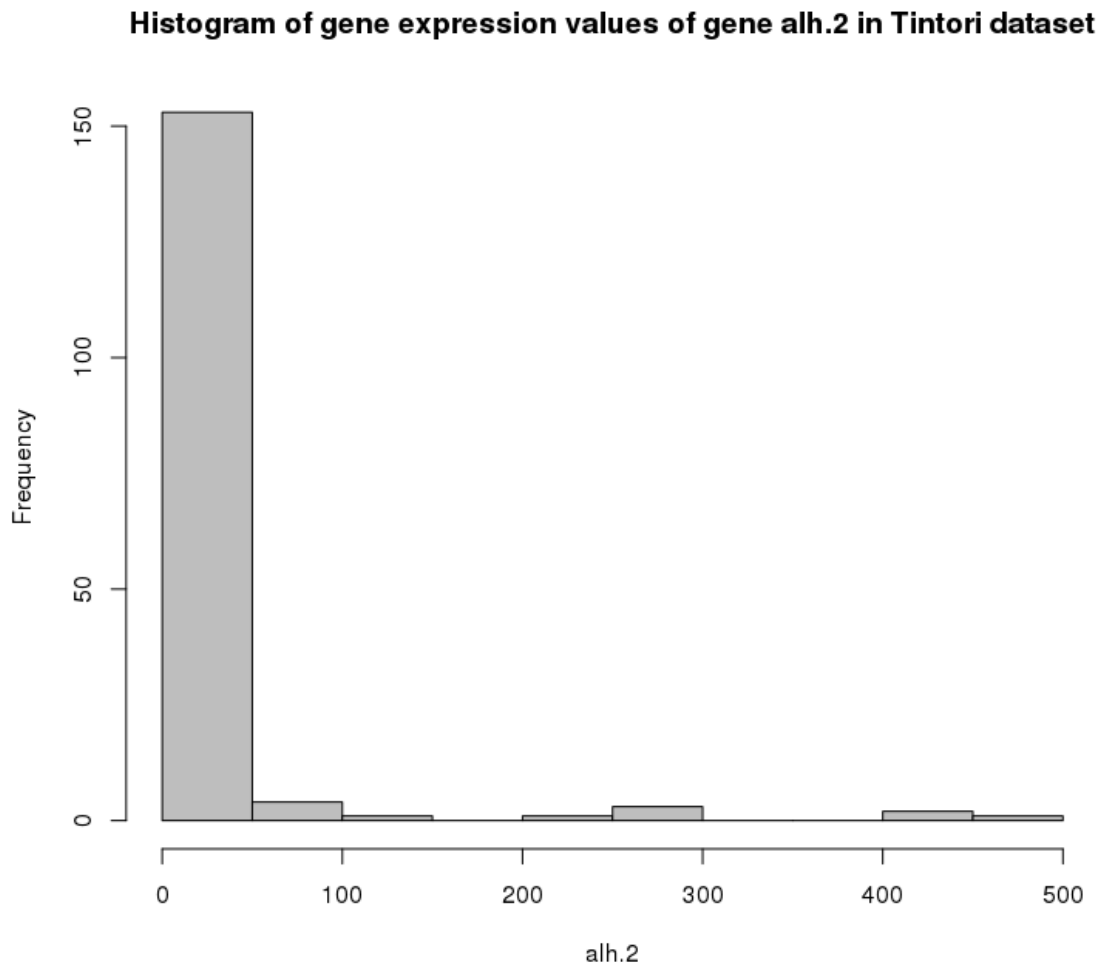



Figure: Dataset from **[Tintori et al., 2016]**

# Which distributions should be chosen? (2/4)

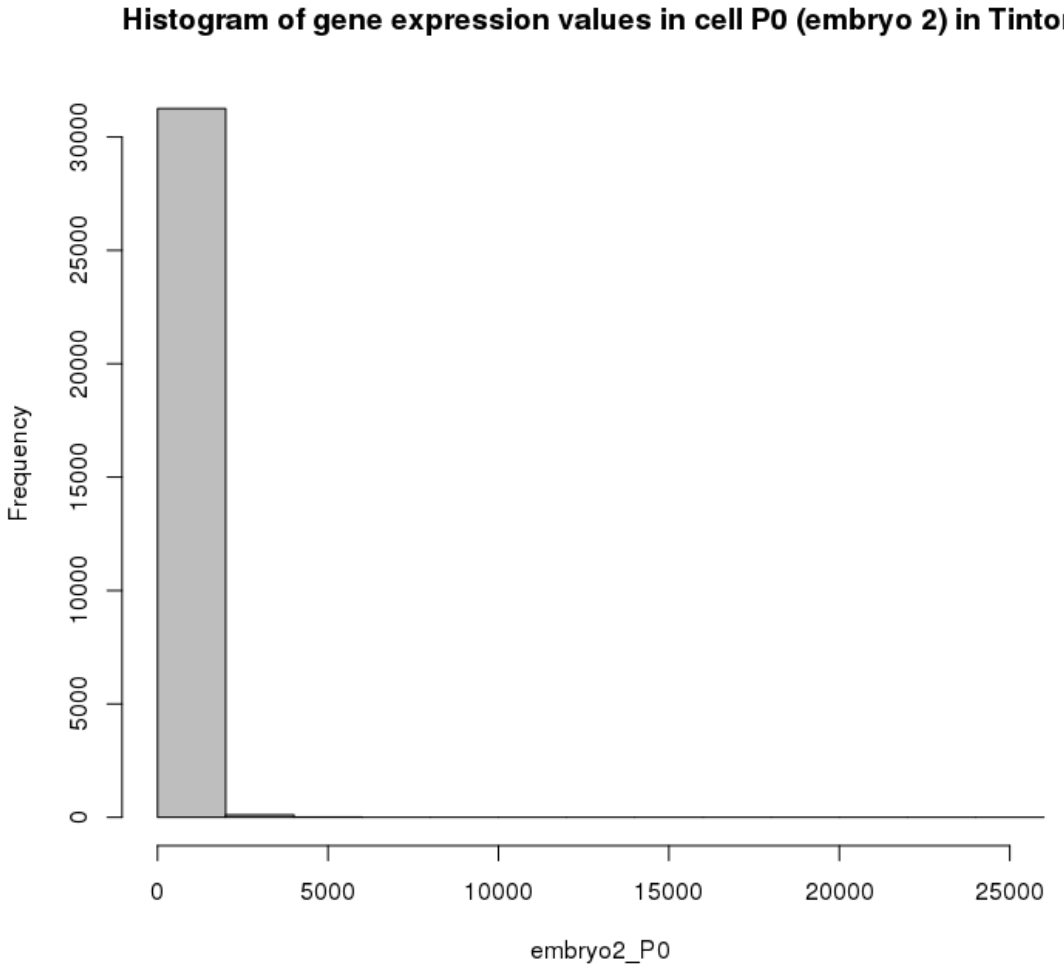



Figure: Dataset from **[Tintori et al., 2016]**

# Which distributions should be chosen? (3/4)

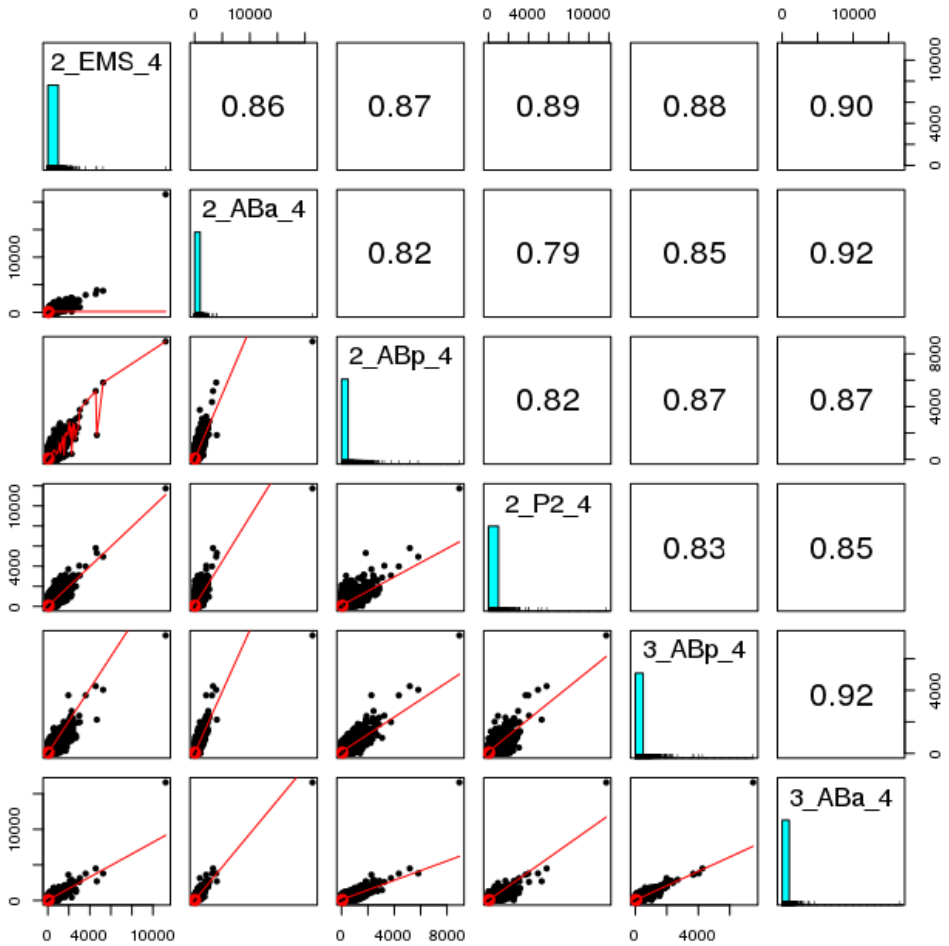



Figure: Dataset from **[Tintori et al., 2016]** (all genes)

# Which distributions should be chosen? (4/4)

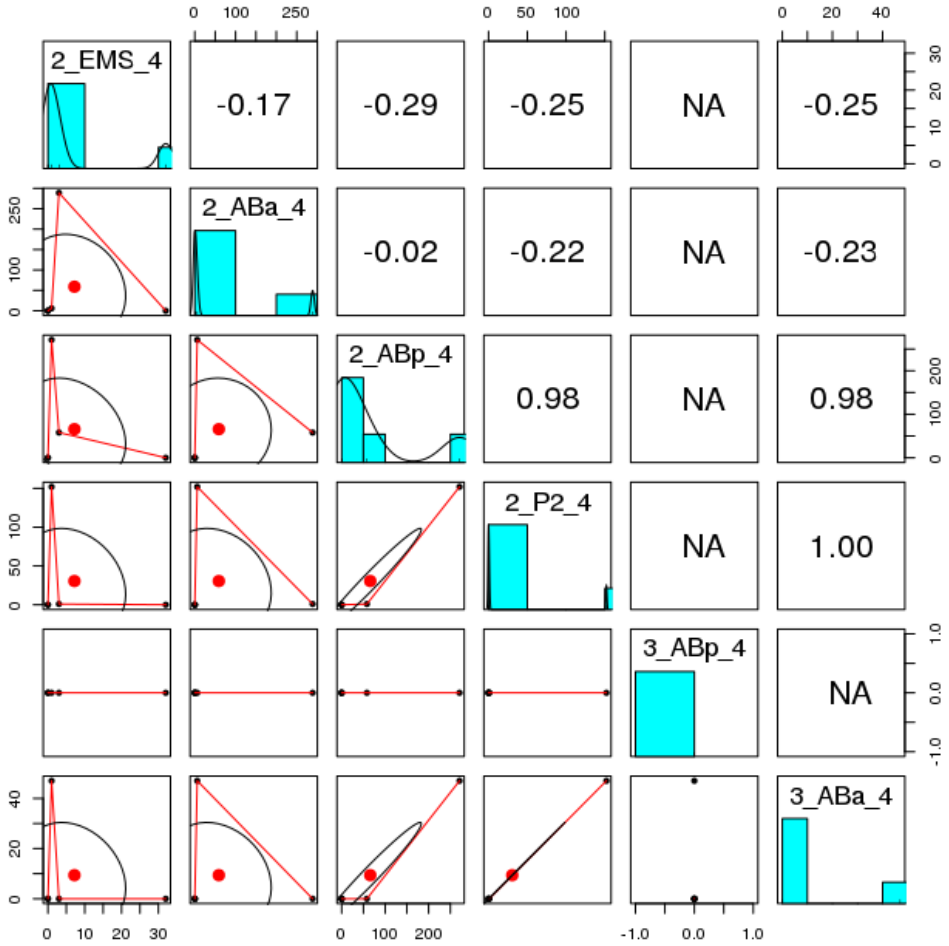



Figure: **[Tintori et al., 2016]** (informative genes: thres=0.75)

# Model for single-cell gene expression

After **feature selection**
and computation of the **two modes** for each cell:

# Model for single-cell gene expression

After **feature selection**
and computation of the **two modes** for each cell:

**For a given cell $j$ and $p$ genes (log-normalized values)**

- $\mathscr{C}_{j\mu_{1j},\mu_{2j},\alpha_j,\Sigma}$ is a bimodal Gaussian copula of means: $\mu_{1j}$, $\mu_{2j}$, covariance: $\Sigma$, rate: $\alpha_j$.

# Model for single-cell gene expression

After **feature selection**
and computation of the **two modes** for each cell:

**For a given cell $j$ and $p$ genes (log-normalized values)**

- $\mathscr{C}_{j\mu_{1j},\mu_{2j},\alpha_j,\Sigma}$ is a bimodal Gaussian copula of means: $\mu_{1j}$, $\mu_{2j}$, covariance: $\Sigma$, rate: $\alpha_j$.
- $\forall i \leq p, X_{i,j}$ is the r.v. associated with expression of gene $i$.
- $X_j = (X_{1,j}, ..., X_{p,j})$ is the r.**v**. associated with expression in cell $j$.

# Model for single-cell gene expression

After **feature selection**
and computation of the **two modes** for each cell:

**For a given cell $j$ and $p$ genes (log-normalized values)**

- $\mathscr{C}_{j\mu_{1j},\mu_{2j},\alpha_j,\Sigma}$ is a bimodal Gaussian copula of means: $\mu_{1j}$, $\mu_{2j}$, covariance: $\Sigma$, rate: $\alpha_j$.
- $\forall i \leq p, X_{i,j}$ is the r.v. associated with expression of gene $i$.
- $X_j = (X_{1,j}, ..., X_{p,j})$ is the r.**v**. associated with expression in cell $j$.
- For all $i \leq p$, $\mathscr{F}_{X_{i,j}}$ (CDF of $X_{i,j}$) is the known model for single gene $i$ expression.

# Model for single-cell gene expression

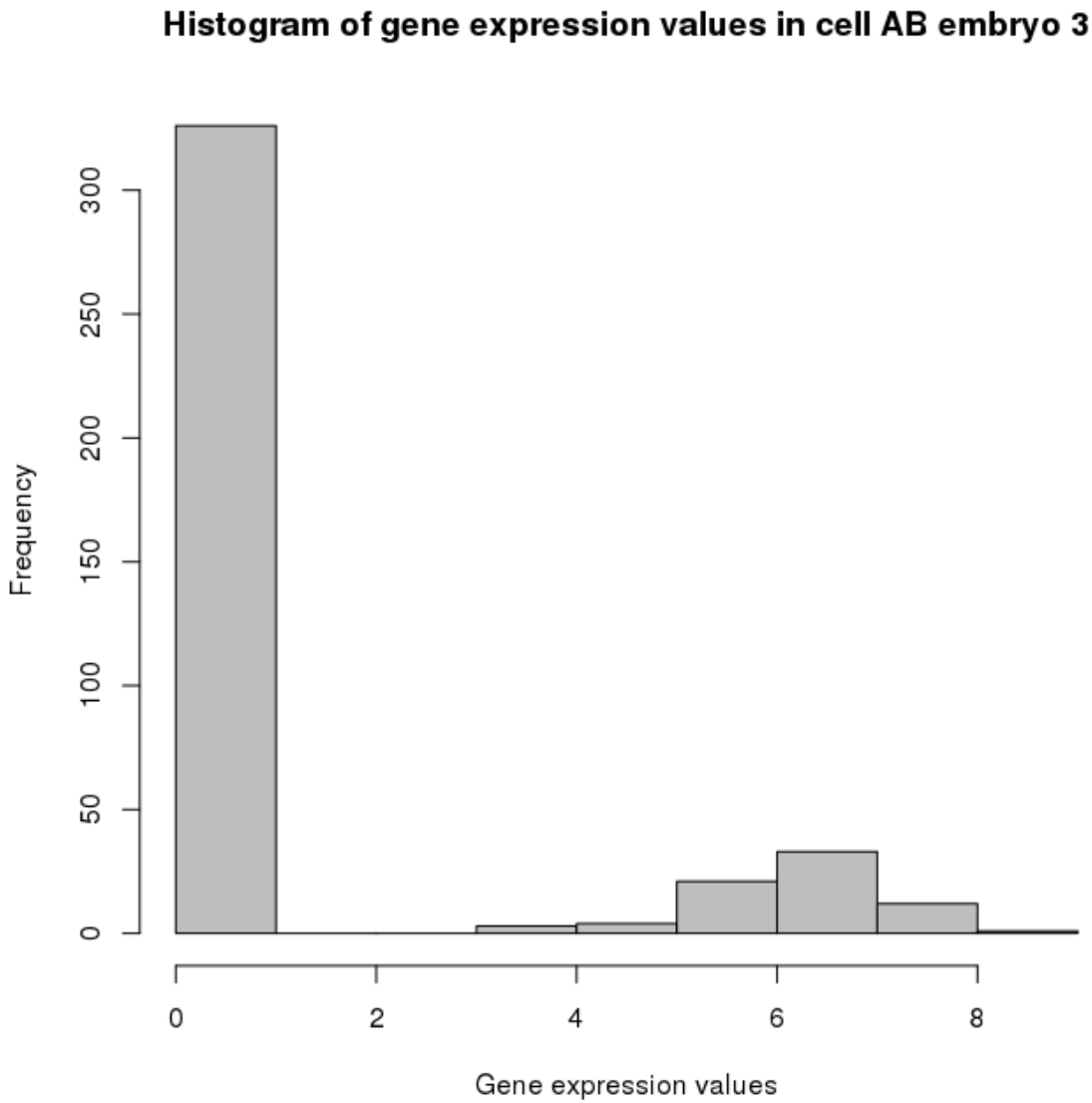
After **feature selection**
and computation of the **two modes** for each cell:

**For a given cell $j$ and $p$ genes (log-normalized values)**

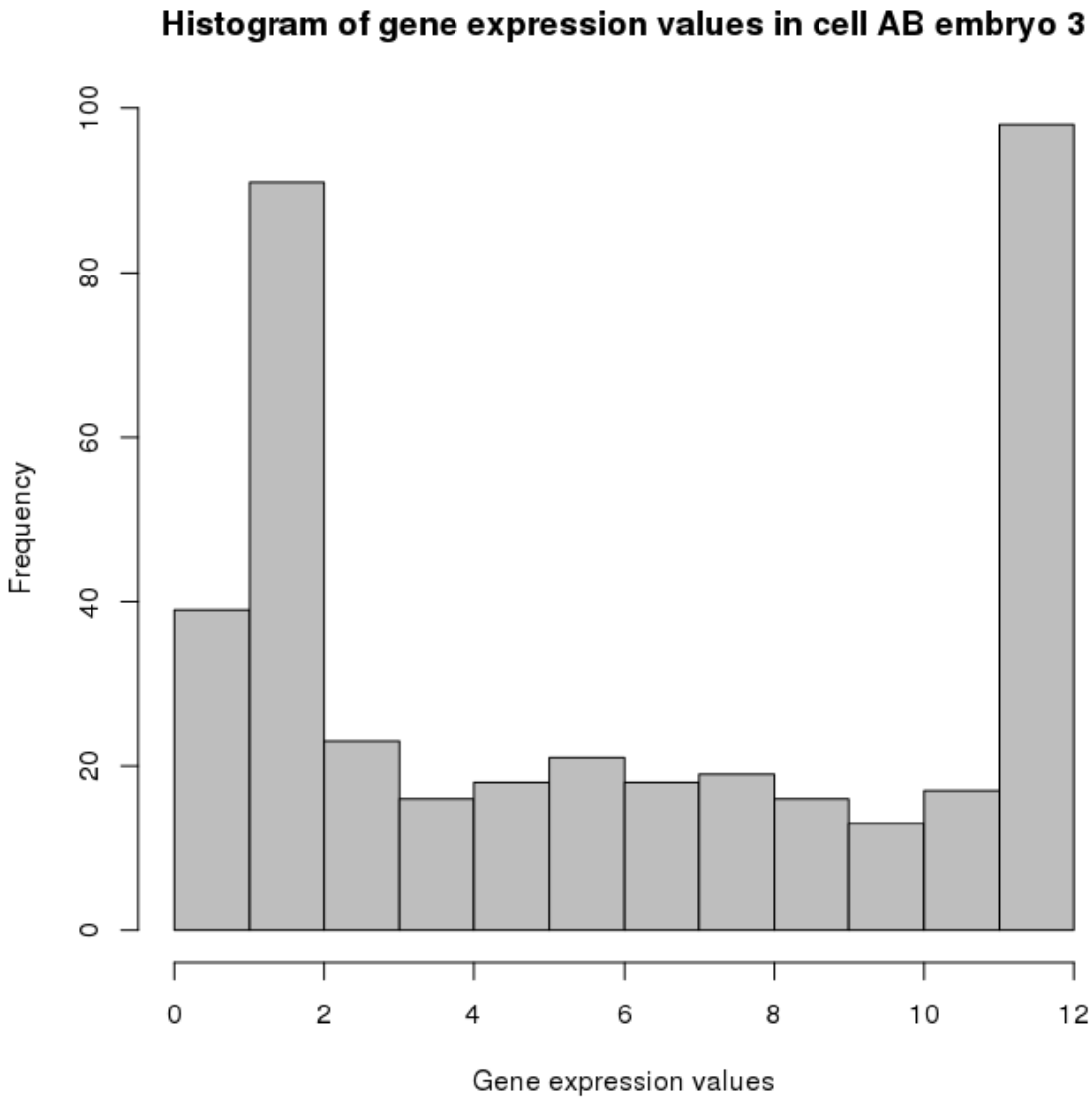
- $\mathscr{C}_{j\mu_{1j},\mu_{2j},\alpha_j,\Sigma}$ is a bimodal Gaussian copula of means: $\mu_{1j}$, $\mu_{2j}$, covariance: $\Sigma$, rate: $\alpha_j$.
- $\forall i \leq p, X_{i,j}$ is the r.v. associated with expression of gene $i$.
- $X_j = (X_{1,j}, ..., X_{p,j})$ is the r.**v**. associated with expression in cell $j$.
- For all $i \leq p$, $\mathscr{F}_{X_{i,j}}$ (CDF of $X_{i,j}$) is the known model for single gene $i$ expression.

$$\mathrm{P}(X_j \leq q) = \mathscr{C}_{j\mu_{1j},\mu_{2j},\alpha_j,\Sigma}(\phi^{-1} \circ \mathscr{F}_{X_{1,j}}(q_1), ..., \phi^{-1} \circ \mathscr{F}_{X_{p,j}}(q_p))$$

# Results (1/2): values from dataset [Tintori et al.]



# Results (2/2): generation from the model



# Conclusion

## Contribution

1. An online application for data analysis.

# Conclusion

## Contribution

1. An online application for data analysis.
2. A benchmark performed on clustering algorithms.

# Conclusion

## Contribution

1. An online application for data analysis.
2. A benchmark performed on clustering algorithms.
3. A model for *single-cell* gene expression, that needs improvement.