Differential Analysis for RNA-seq using Intensive Randomization.

D. Desaulle et Y. Rozenholc - UR7537 - BioSTM
en collaboration avec
C. Hoffmann et B. Hainque - UTCBS - CNRS UMR8258 - INSERM U1267

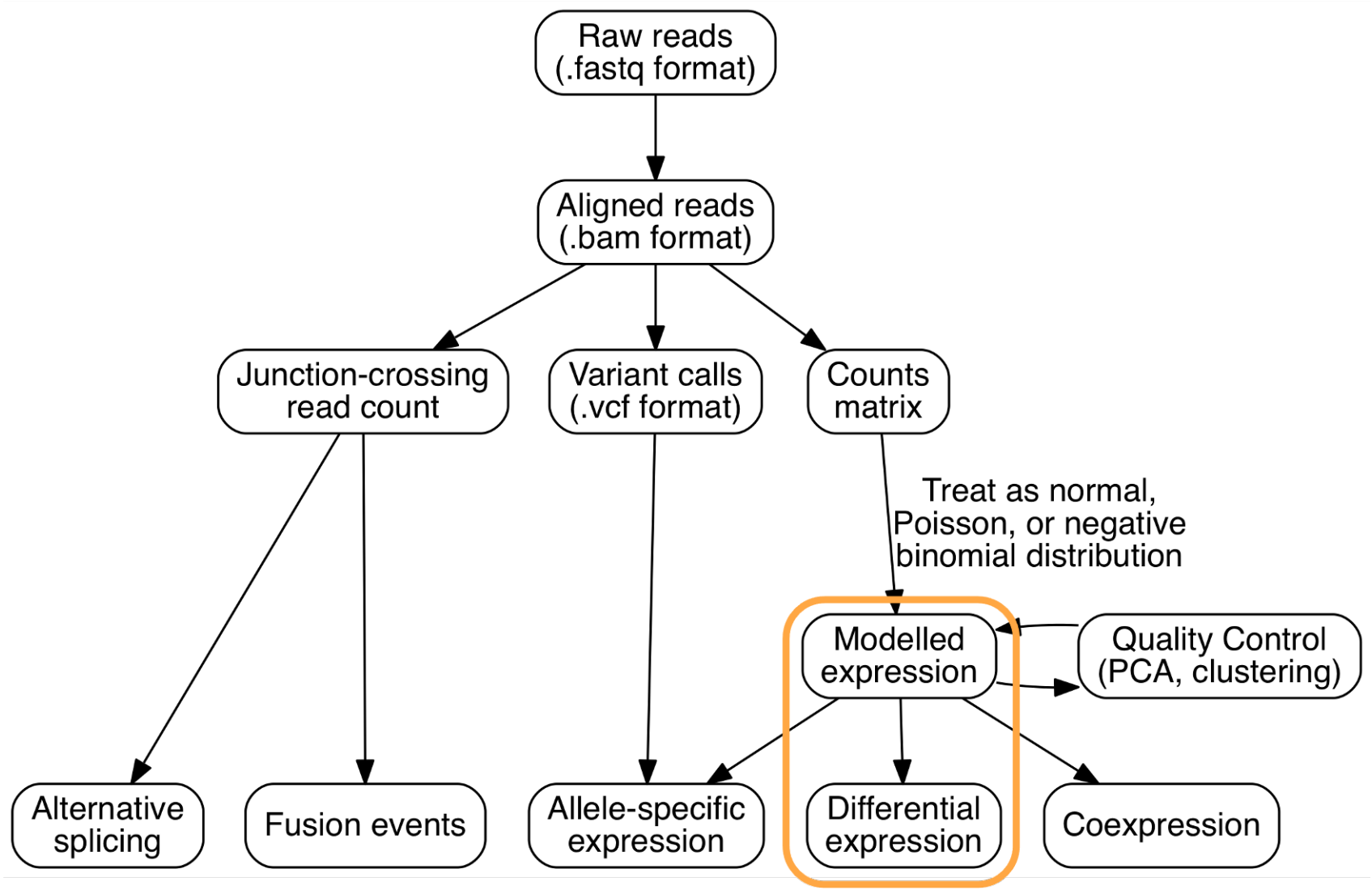
Faculté de Pharmacie de Paris
Université de Paris Cité

StatOmique — 25 Octobre 2022

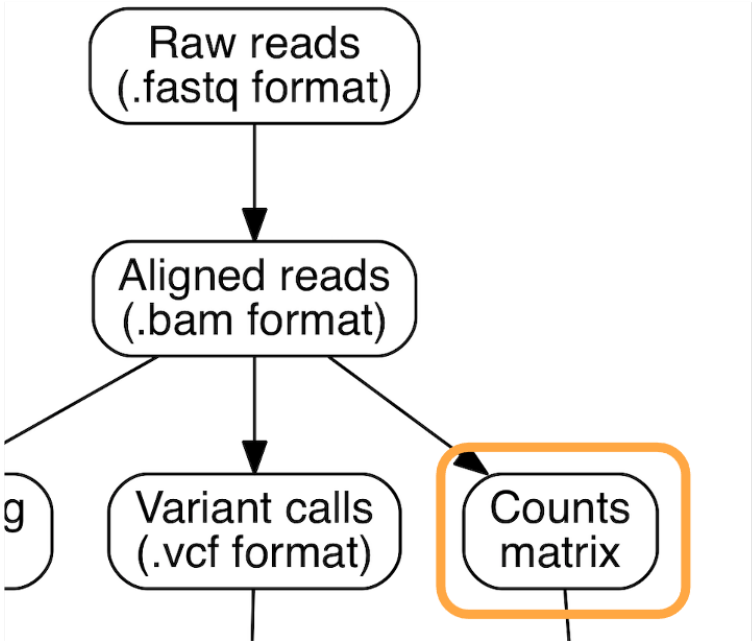
Table of Contents

Transcriptome data (RNA-seq)

RNA-Seq : From raw expressions to statistical analysis



RNA-Seq data : Counts data



	Gene 1	Gene j	...	Gene m
Sample 1	2	2	3	3
⋮				
Sample *i*	4	4	6	6
⋮				
Sample *n*	4	4	6	16

X_{ij} - Raw counts - number of reads

- in sample i, $i = 1, \ldots, n$
- for gene j, $j = 1, \ldots, m$

Modelization and normalization

Read counts modelization for the i-th sample

- Multinomial variable

$$(X_{i1}, ..., X_{im}) \sim \mathcal{M}(N_i, (\pi_1, ..., \pi_m))$$

- Independent Poisson variables

$$X_{ij} \sim \mathcal{P}(s_i \times \mu_j) \qquad \text{for} \quad j = 1, ..., m$$

- Independent negative binomials variables

$$X_{ij} \sim \mathcal{NB}(r_j, r_j/(r_j + \lambda_{ij})) \qquad \text{for} \quad j = 1, ..., m$$

with $\lambda_{ij} = s_i \times \mu_j$

In all cases :

- The expected number of reads is given by $\mathbb{E}(X_{ij}) = s_i \times \mu_j$
- The relative mean expression level μ_j is un unknown quantity of interest
- The scaling factors s_i are unknown and represent nuisance model parameters

Differential analysis in transcriptomic studies

Which are differential expressions between conditions A and B ?

- We assume that

$$\mathbb{E}(X_{ij}) = s_i \, \mu_j^A, \text{ for } i \in A \qquad \text{and} \qquad \mathbb{E}(X_{ij}) = s_i \, \mu_j^B \text{ for } i \in B.$$

- We aime at testing

$$H_0^j : \mu_j^A = \mu_j^B \quad \text{against} \quad H_1^j : \mu_j^A \neq \mu_j^B.$$

to identify those $j \in \{1, ..., m\}$ such that H_1^j is true.

UNFORTUNATELY the s_i are unknown !

- Must be taken into account in the analysis !

Scaling factors and normalization

Suppose that the scaling factors SF_i are known for each sample i :

				Scaling Factor
2	2	3	3	1
4	4	6	6	2
4	4	6	16	2

$$\Rightarrow \frac{X_{ij}}{SF_i} :$$

2	2	3	3
2	2	3	3
2	2	3	8

Scaling factors and normalization

Suppose that the scaling factors SF_i are known for each sample i :

				Scaling Factor
2	2	3	3	1
4	4	6	6	2
4	4	6	16	2

$$\Rightarrow \frac{X_{ij}}{SF_i} :$$

2	2	3	3
2	2	3	3
2	2	3	8

Scaling factor estimation in the context of Analyse Differentiel

- **Total counts normalization** (Marioni et al. 2008 ; Mortazavi et al. 2008)
- **Housekeeping genes normalization** (Vandesompele et al. 2002)
- **Upper quantile normalization** (Bullard et al. 2010)
- **Trimmed Mean of *M* values (TMM)** (Robinson and Oshlack 2010) in edgeR R package (Robinson, McCarthy, and Smyth 2010)
- **DESeq2 normalization** (Anders and Huber 2010) in DESeq2 R package (Love, Huber, and Anders 2014)

If most genes are not differentially expressed, both edgeR and DESeq2 methods are able to control the false positive rates while maintaning the power (Dillies et al. 2013)

The optimal approach has not reached a consensus to date (Abrams et al. 2019).

Scaling factor estimation - Housekeeping genes normalization

Housekeeping genes normalization

- Assume a set of invariant genes is known
- compute *partial* library size using this gene set
- normalize using these partial library sizes

Partial Count

2	2	3	3	5
4	4	6	6	10
4	4	6	16	10

$\Rightarrow \frac{X_{ij}}{PC_i}$:

0.4	0.4	0.6	0.6
0.4	0.4	0.6	0.6
0.4	0.4	0.6	1.6

Seems to work well

Let us call it : **"good" normalization**

Scaling factor estimation - “Housekeeping” genes normalization

Some disadvantages

- Housekeeping genes are not necessary known under new conditions
- Housekeeping genes may vary more than expected
- Set size should be large to ensure good estimation

Partial Count

2	2	3	3	6
4	4	6	6	12
4	4	6	16	22

$\Rightarrow \frac{X_{ij}}{PC_i}$:

0.33	0.33	0.5	0.5
0.33	0.33	0.5	0.5
0.18	0.18	0.27	0.72

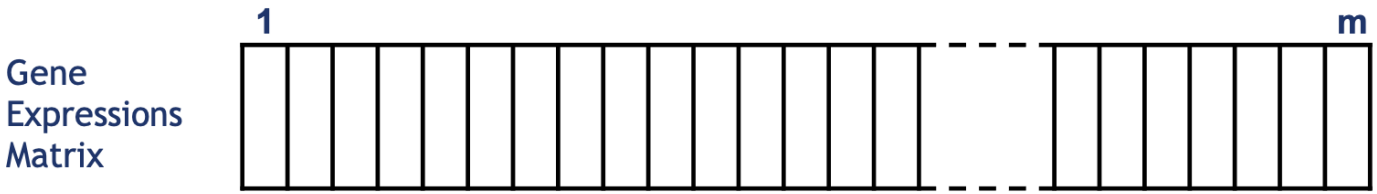
Does not work anymore

Let us call it : **"wrong" normalization**

Picking ‘reference’ genes at random

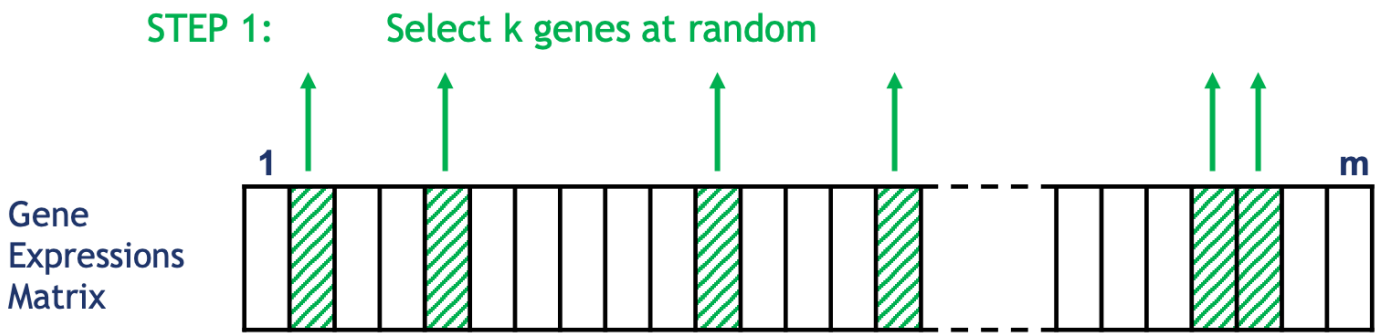
Housekeeping (reference) gene normalization

Single run of the procedure



Housekeeping (reference) gene normalization

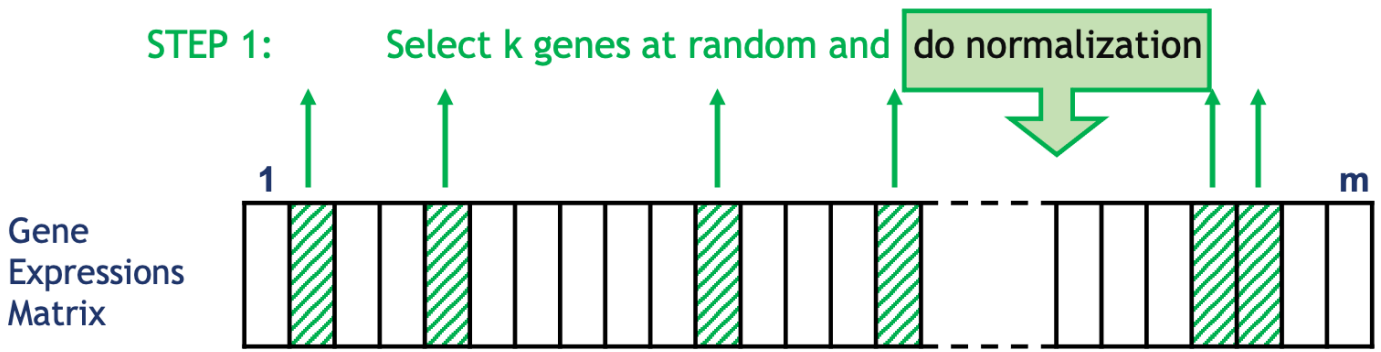
Single run of the procedure



- Step 1 : Chose a subset S of k randomly picked genes

Housekeeping (reference) gene normalization

Single run of the procedure

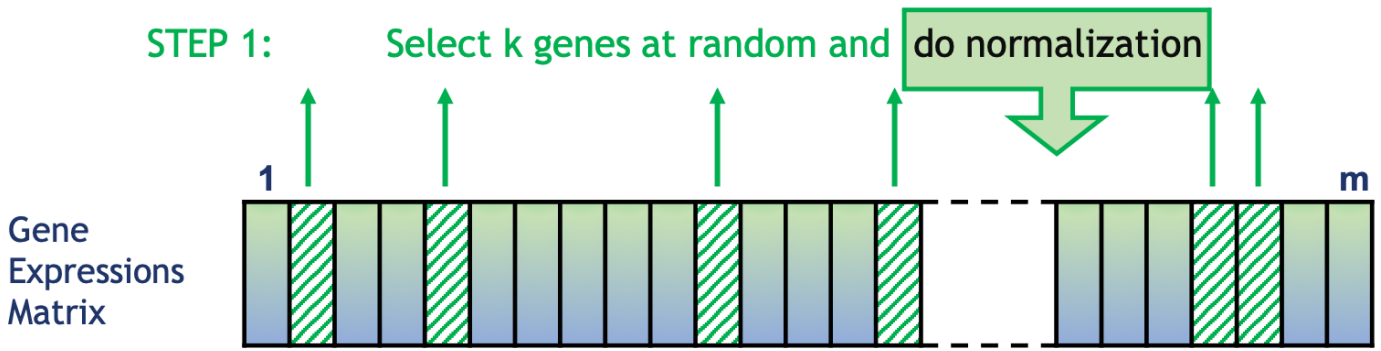


- Step 1 : Chose a subset S of k randomly picked genes

$\Rightarrow$ Estimate s_i

Housekeeping (reference) gene normalization

Single run of the procedure



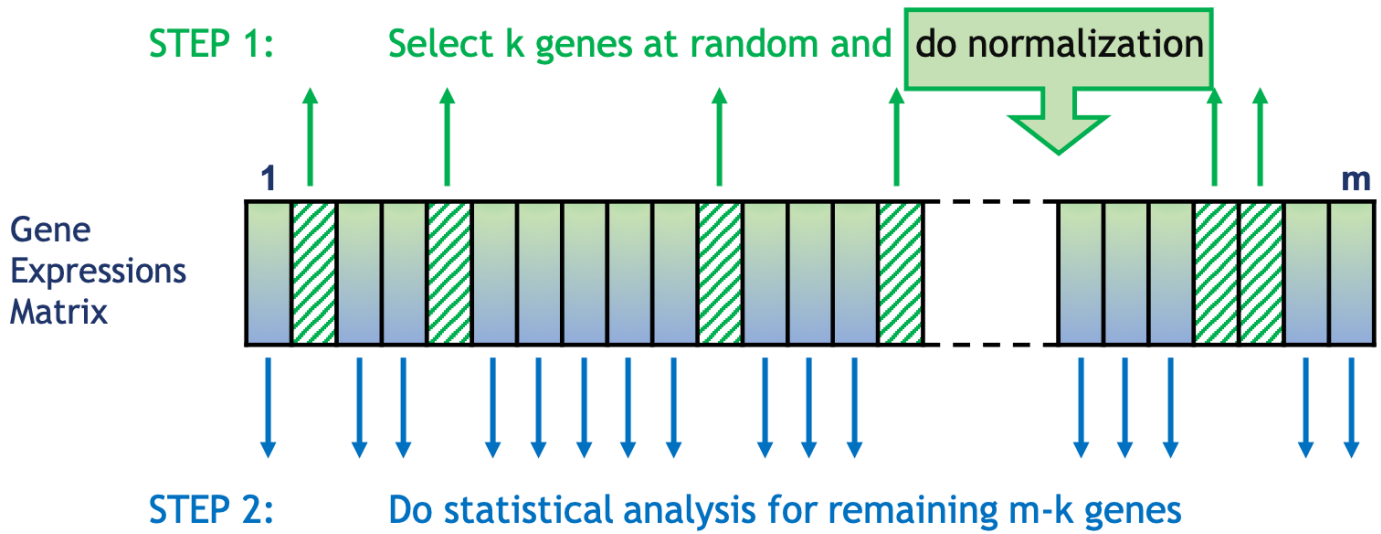
- Step 1 : Chose a subset S of k randomly picked genes

$\Rightarrow$ Estimate s_i

$\Rightarrow$ Normalize the data

Housekeeping (reference) gene normalization

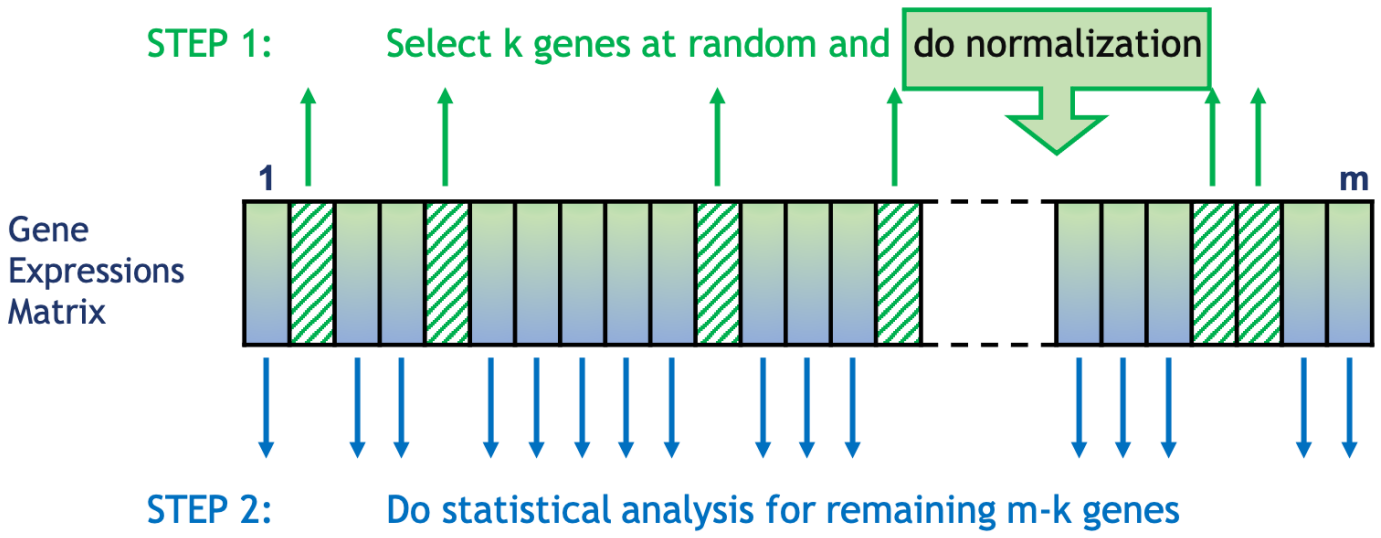
Single run of the procedure



- Step 2 : Do a differential analysis (test) for remaining genes

Housekeeping (reference) gene normalization

Single run of the procedure



- Step 2 : Do a differential analysis (test) for remaining genes

$\Rightarrow$ Get detection indicator

$$1_S(j) = \begin{cases} 1 & \text{if } j \text{ declared DE } (j\text{-th hypothesis is rejected}) \\ 0 & \text{otherwise} \end{cases}$$

Random normalization with a good normalization subset

In step 1, for one random subset $S \subset \{1, ..., m\}$ of size k used for normalization, we consider

$$\hat{s}_i^S = \frac{n \sum_{j \in S} X_{ij}}{\sum_{i=1}^n \sum_{j \in S} X_{ij}}$$

and in step 2 we test genes $j \in \{1, ..., m\} \setminus S$.

Remark :

For a "good" normalization subset S that is made of invariant genes, the rates of detection of gene j are assumed to be well controlled

$$\begin{cases} P(1_S(j) = 1 | S - \text{"good"}) \leq \eta & \text{if } j \text{ is invariant (under } H_0) \\ P(1_S(j) = 1 | S - \text{"good"}) \geq 1 - \beta & \text{if } j \text{ is DE (under } H_1) \end{cases}$$

Random normalization in general case

Assume that d genes are not invariant.

- π_d^0 : probability that a normalization subset is "wrong" when the tested gene is invariant,
- π_d^1 : same when the tested gene is DE.

Considering the events "S contains only invariant" or not, using total probability formula, when j is tested its detection rates satisfy

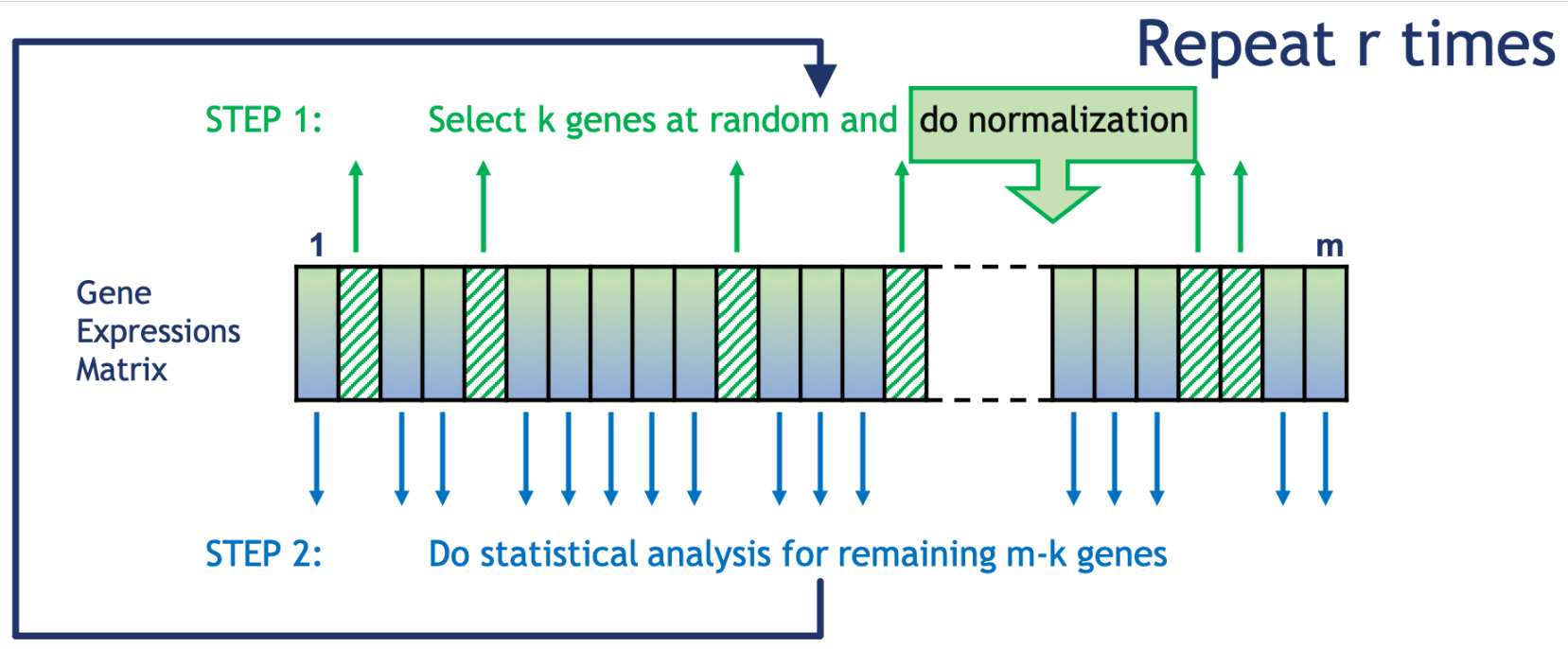
$$\begin{cases} P(1_S(j) = 1) \leq \eta(1 - \pi_d^0) + \pi_d^0 & \text{if } j \text{ is invariant (under } H_0^j), \\ P(1_S(j) = 1) \geq (1 - \beta)(1 - \pi_d^1) & \text{otherwise (under } H_1^j) \end{cases}$$

Remark : Since the normalization subset S may contain DE genes, single run of the procedure is not enough to declare gene j DE.

- The procedure shall be repeated !

Housekeeping (reference) gene normalization

Repeated procedure



- Do Step 1 (chose randomly reference genes and estimate s_i) and Step 2 (testing)
- Increment R_j it the j-th hypothese is rejected

For r drawn normalization subsets, R_j is the number of detections of gene j through the $r_j \leq r$ normalization subsets S such that $j \notin S$.

Final detection from the repeated procedure

- Consider R_j the number of detections of gene j through the $r_j \leq r$ normalization subsets S such that $j \notin S$
- Compute corresponding p-values under Binomial distribution under H_0^j

$$p_j^d(\eta) := 1 - B\left(R_j; r, [1 - k/m][\eta(1 - \pi_d^0) + \pi_d^0]\right).$$

- Account for multiplicity on α level : **Holm's rule**

$$p_{(1)}^d(\eta) \leq p_{(2)}^d(\eta) \leq \ldots \leq p_{(m)}^d(\eta)$$

  - $\delta = 0$ if $p_{(1)}^1 \geq \frac{\alpha}{m}$
  - $\delta = \arg\min_d \left[p_{(d)}^d < \frac{\alpha}{m-d+1} \text{ and } p_{(d+1)}^{d+1} \geq \frac{\alpha}{m-d}\right]$ otherwise

    with $p_{(1)}^d \leq p_{(2)}^d \leq \ldots \leq p_{(m)}^d$ being sorted p-values.
- If $\delta > 0$, declare as differentially expressed genes associated with $p_{(1)}^\delta, \ldots, p_{(\delta)}^\delta$.

Control of the FWER

Theorem 1 : Assuming the genes are independent and that, for any good normalization subset S, the detection rates satisfy

$$P(1_S(j) = 1) \leq \eta(1 - \pi_d^0) + \pi_d^0 \qquad \text{if } j \text{ is invariant,}$$
$$P(1_S(j) = 1) \geq (1 - \beta)(1 - \pi_d^1) \qquad \text{if } j \text{ is DE.}$$

then the FWER is bounded by $\alpha + o_r(1)$ as soon as

$$(1 - \beta)(1 - \pi_d^1) > \eta(1 - \pi_d^0) + \pi_d^0$$

where d is the unknown number of differential expressions.

A new test for the negative binomial model

Negative binomial model

$$X_{ij} \sim \mathcal{NB}(r_j, r_j/(r_j + \lambda_{ij})) \qquad \text{for} \quad j = 1, \ldots, m$$

with $\lambda_{ij} = s_i \times \mu_j^\bullet$ where $\bullet$ indicates the condition (A or B) to which i belongs to.

Poisson limiting distribution :

$$\lim_{r \to \infty} \mathcal{NB}(r, r/(r + \lambda)) = \mathcal{P}(\lambda).$$

We have

$$\mathbb{E}(X_{ij}) = \lambda_{ij} = s_i \times \mu_j^\bullet \quad \text{and} \quad \mathbb{V}(X_{ij}) = \lambda_{ij}(1 + \lambda_{ij}/r_j) = s_i \times \mu_j^\bullet (1 + s_i \times \rho_j^\bullet).$$

Gaussian approximation :

$$X_{ij} \sim \mathcal{NB}(r_j, r_j/(r_j + \lambda_{ij})) \approx \mathcal{N}(s_i \times \mu_j^\bullet, s_i \times \mu_j^\bullet (1 + s_i \times \rho_j^\bullet)).$$

Known scaling factor case

$$U_{ij} := 2\sqrt{X_{ij}/s_i} \approx \mathcal{N}(2\sqrt{\mu_j^\bullet}, \frac{1}{s_i} + \rho_j^\bullet) = 2\sqrt{\mu_j^\bullet} + V_{ij}.$$

with $V_{ij} := \sqrt{\rho_j^\bullet + 1/s_i} \times \varepsilon_{ij}$ and $\varepsilon_{ij} \sim \mathcal{N}(0,1)$

The empirical means

$$\bar{U}_j^A := \sum_{i \in A} U_{ij}/n_A \quad \text{and} \quad \bar{U}_j^B := \sum_{i \in B} U_{ij}/n_B.$$

satisfy

$$\frac{\bar{U}_l^A - \bar{U}_l^B}{\sum_s} \approx \mathcal{N}(2\sqrt{\mu_j^A} - 2\sqrt{\mu_j^B}, 1)$$

where

$$\sum\nolimits_S^2 := \frac{\rho_j^A}{n_A} + \frac{\rho_j^B}{n_B} + \frac{1}{n_A^2}\sum_{i \in A}\frac{1}{s_i} + \frac{1}{n_B^2}\sum_{i \in B}\frac{1}{s_i}.$$

UNFORTUNATELY : ρ_j^A, ρ_j^B and the s_i are unknown.

Test statistic

$$Y_{ij} := 2\sqrt{X_{ij}/\hat{s}_i}$$

Empirical expectations of the Y_{ij} into the subpopulation A and B :

$$\bar{Y}_j^A = \frac{1}{n_A}\sum_{i\in A} Y_{ij} \quad \text{and} \quad \bar{Y}_j^B = \frac{1}{n_B}\sum_{i\in B} Y_{ij}.$$

Estimate of $\sum_S^2$:

$$\hat{\sum}_S^2 = \frac{1}{n_A(n_A-1)}\sum_{i\in A}(Y_{ij}-\bar{Y}^A)^2 + \frac{1}{n_B(n_B-1)}\sum_{i\in B}(Y_{ij}-\bar{Y}^B)^2.$$

We build our testing procedure on the following statistic

$$T_j := \frac{\bar{Y}_j^A - \bar{Y}_j^B}{\hat{\sum}_S}.$$

Statistic decomposition

For any vector x of $\mathbb{R}^n$, $(\bar{x}^A - \bar{x}^B)\mathbb{1}_n = Hx$ where

$$\mathbb{1}_n = (\underbrace{1, \ldots, 1}_{n \text{ times}})^T \quad \text{and} \quad H = \begin{pmatrix} \frac{1}{n_A} J_{n_A} & -\frac{1}{n_B} J_{n_A, n_B} \\ \frac{1}{n_A} J_{n_B, n_A} & -\frac{1}{n_B} J_{n_B} \end{pmatrix}.$$

The following decomposition holds

$$T_j \mathbb{1}_n = \frac{\sum_S}{\hat{\sum}_S} \left(\frac{R_1 \varepsilon_{\bullet j}}{\sum_S} + 2\frac{R_2}{\sum_S} + \frac{2}{\sum_S}(\sqrt{\mu_j^A} - \sqrt{\mu_j^B}) + \frac{1}{\sum_S} H V_{\bullet j} \right).$$

with

$$\varepsilon_{\bullet j} = (\varepsilon_{1j}, \ldots, \varepsilon_{nj})^T$$

$$R_1 = H\, diag(\sqrt{s_i/\hat{s}_i} - 1)\, diag(\sqrt{\rho_j^{\bullet} + 1/s_i})$$

$$R_2 = H\, diag(\sqrt{s_i/\hat{s}_i} - 1)\, (\sqrt{\mu_j^A}\mathbb{1}_{n_A}^T, \sqrt{\mu_j^B}\mathbb{1}_{n_B}^T)^T$$

Distribution

$$T_j \mathbb{1}_n = \frac{\sum_S}{\hat{\sum}_S} \left(\frac{R_1 \varepsilon_{\bullet j}}{\sum_S} + 2\frac{R_2}{\sum_S} + \frac{2}{\sum_S}(\sqrt{\mu_j^A} - \sqrt{\mu_j^B}) + \frac{1}{\sum_S} HV_{\bullet j} \right).$$

Lemma : If $\max_i \left| \sqrt{s_i/\hat{s}_i} - 1 \right| \leq 1/2$ then, with probability larger than $1 - 5n^{-c}$,

$$\frac{\sum_S}{\hat{\sum}_S} \leq (1 + \sqrt{c \log n}) \left(1 + 2 \left[2(1+c)(1+o(1)) \frac{1 + s_{\max} \bar{\rho}_S}{\sum_{j \in S} \mu_j} \frac{\log n}{n} \right]^{1/2} \right),$$

$$\frac{\|R_1 \varepsilon\|^2}{\sum_S^2} \leq 2(1 + 2\sqrt{c \log n} + 2c \log n)(1+c)(1+o(1)) \frac{1 + s_{\max} \bar{\rho}_S}{\sum_{j \in S} \mu_j} \log n,$$

$$\frac{\|R_2\|^2}{\sum_S^2} \leq 2(1+c)(1+o(1)) \left(\sqrt{\mu_j^A} + \sqrt{\mu_j^B} \right)^2 \times \frac{1 + s_{\max} \bar{\rho}_S}{\sum_{j \in S} \mu_j} \left(\frac{n}{n_A} \vee \frac{n}{n_B} \right) \log n,$$

where $s_{\max} := \max_{i=1,\ldots,n} s_i$ and $\bar{\rho}_S := \sum_{j \in S} \mu_j \rho_j \big/ \sum_{j \in S} \mu_j$.

Remark : When $\sum_{j \in S} \mu_j$ is large, consequently $\sum_S / \hat{\sum}_S$ is of order $1 + \sqrt{c \log n}$ and the bias terms $R_1 \varepsilon / \sum_S$ and $R_2 / \sum_S$ are negligible.

From concentration inequalities for Gaussian (Hoeffding) and for quadratic form of Gaussian vector (Gendre 2014, Lemma 8.2).

Empirical study

Implementation - Part I

As we request

$$(1-\beta)(1-\pi_d^1) > \eta(1-\pi_d^0) + \pi_d^0,$$

there is a maximal number of discoveries $\delta_{\max}$:

$$\delta_{\max} = \max\{d \quad | \quad (1-\beta)(1-\pi_d^1) > \eta(1-\pi_d^0) + \pi_d^0\}$$

If $d > \delta_{\max}$, all the non invariant genes d cannot be discovered with this condition.

Approche

- apply an iterative procedure
  - starting with all genes and test at level $\alpha/2$ to get at most $\delta_{\max}^1$,
  - at iteration t, test genes at level $\alpha/2^t$ to get at most $\delta_{\max}^t$,
  - stop when the number of discoveries at step t is strictly less than $\delta_{\max}^t$.

Implementation - Part II

- The detection indicator is defined as

$$1_S(j) := (|T_j| > (1 + \sqrt{c \log n}) q(1 - \eta/2))$$

where $q(1 - \eta/2)$ is an upper quantile of a standard gaussian, $c > 0$

- In Lemma, non gaussian terms are all the smaller as $\sum_{j \in S} \mu_j$ is large.

In order to ensure that $\sum_{j \in S} \mu_j \geq K$

- If size of S is fixed, screen for genes s.t. $\mu_j \geq K/|S|$ which will be involved in normalization.
- If size of S is variable, at each iteration, increase size of S until $\sum_{j \in S} \mu_j > K$

Empirical setting

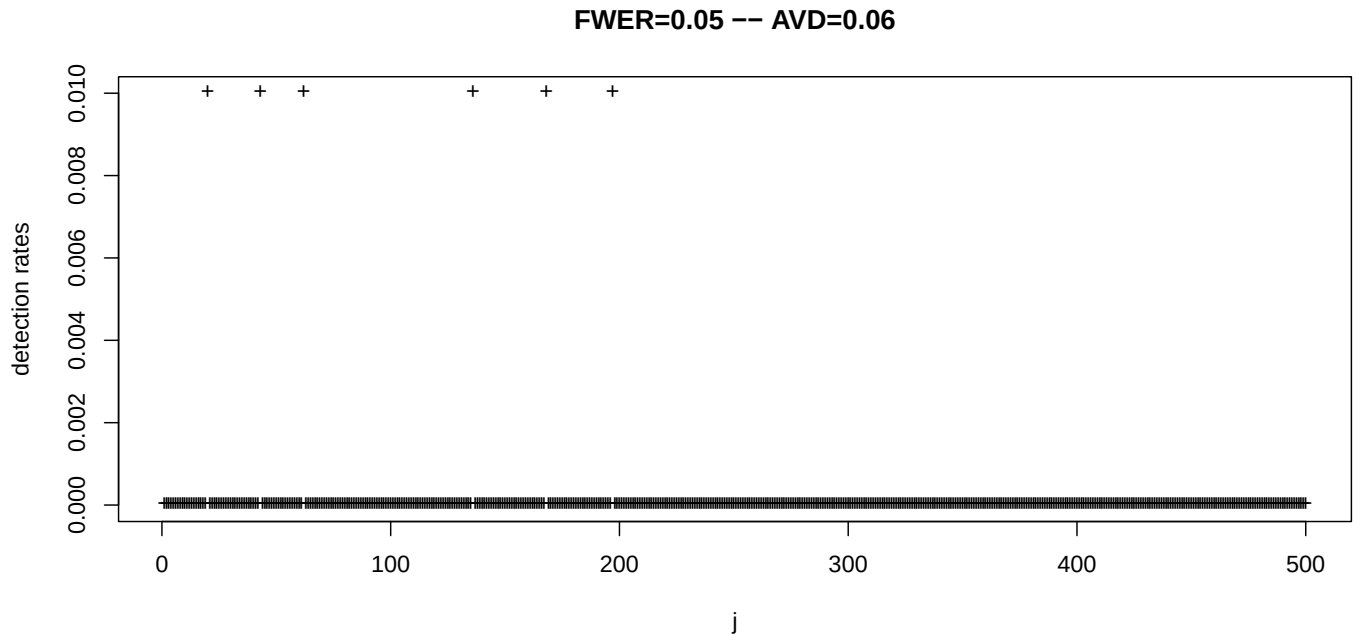
- $m = 500$, $n_1 = n_2 = 6$, $A = \{1, ..., n_1\}$ and $B = \{n_1 + 1, ..., n_1 + n_2\}$.
- $\mu_0 = 100$
- $\mu_j^A = \begin{cases} (1+\phi)\mu_0 & \text{for } j = 1, ..., d \\ \mu_0 & \text{for } j > d \end{cases}$,
- $\mu_j^B = \mu_0$ for all j
- $X_{ij} \sim \mathcal{P}(s_i \mu_j^\bullet)$ or $X_{ij} \sim \mathcal{NB}(r_j^\bullet, r_j^\bullet / (r_j^\bullet + s_i \mu_j^\bullet))$, with $r_j^\bullet = 100$ and $s_i = 1$.
- $|S| = k = 10$ and $r = 2500$.
- Use iterative procedure with $\eta = \alpha = 0.05$, $\beta = 0.1$ and $c = 2.5$.
- Run 100 simulations

Remark : Expected fold change is $2(\sqrt{\mu_j^B} - \sqrt{\mu_j^A})/\sum_S$. From $s_i = 1$, $\sum_S = 2/\sqrt{n}$, hence to get detection with probability greater than α the fold change should satisfy $\sqrt{n\mu_0}|\sqrt{1+\phi} - 1| > -q_{\alpha/2m}(1 + \sqrt{c \log n})$. From this relation we deduce ϕ_{min}.

- $\phi = a/\sqrt{j}$, with detection up to index $j \leq (a/\phi_{min})^2$

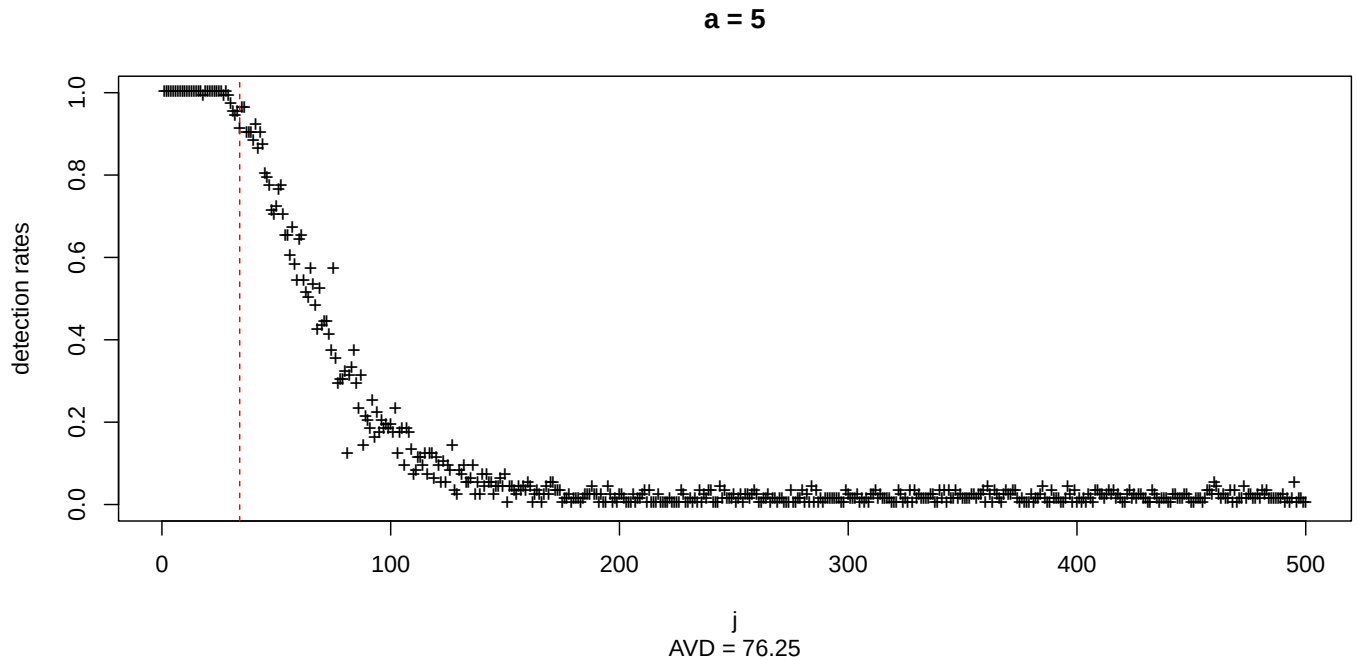
Empirical setting : Poisson $\phi = 0$ (global null)

Using random picking for scale estimation (our procedure)



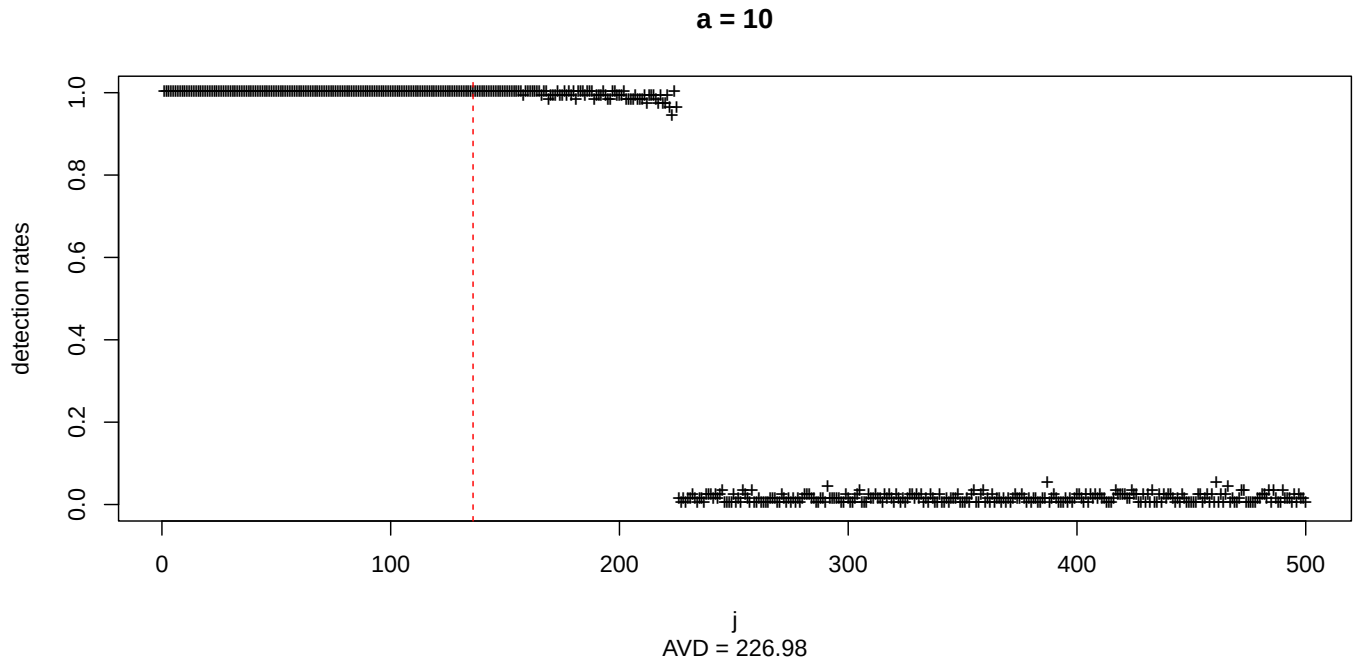
Empirical setting : Poisson $\phi = 5/\sqrt{j}$ for $j = 1, ..., d$ with $d = 225$

Using random picking for scale estimation (our procedure)

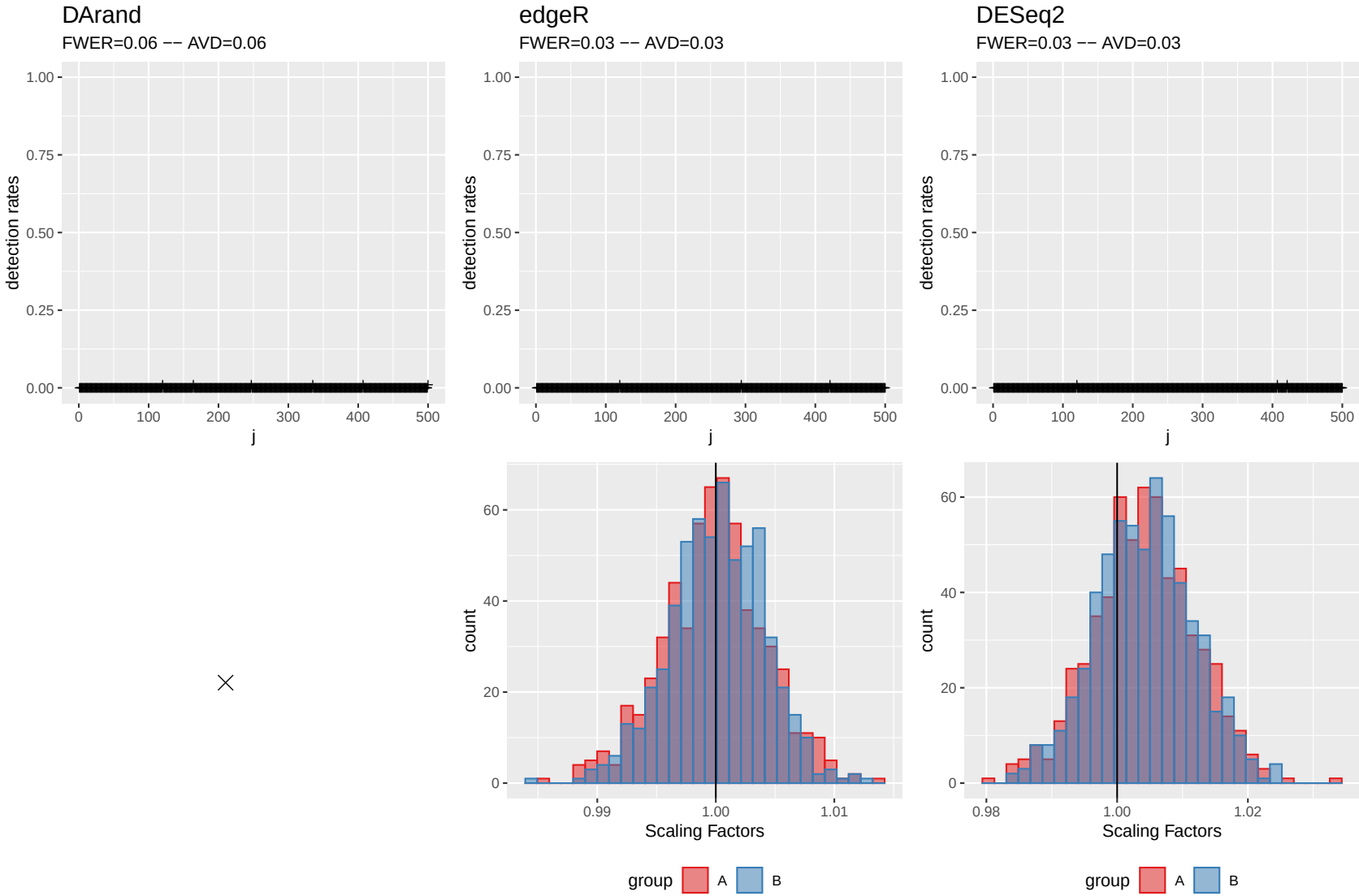


Empirical setting : Poisson $\phi = 10/\sqrt{j}$ for $j = 1, \ldots, d$ with $d = 225$

Using random picking for scale estimation (our procedure)

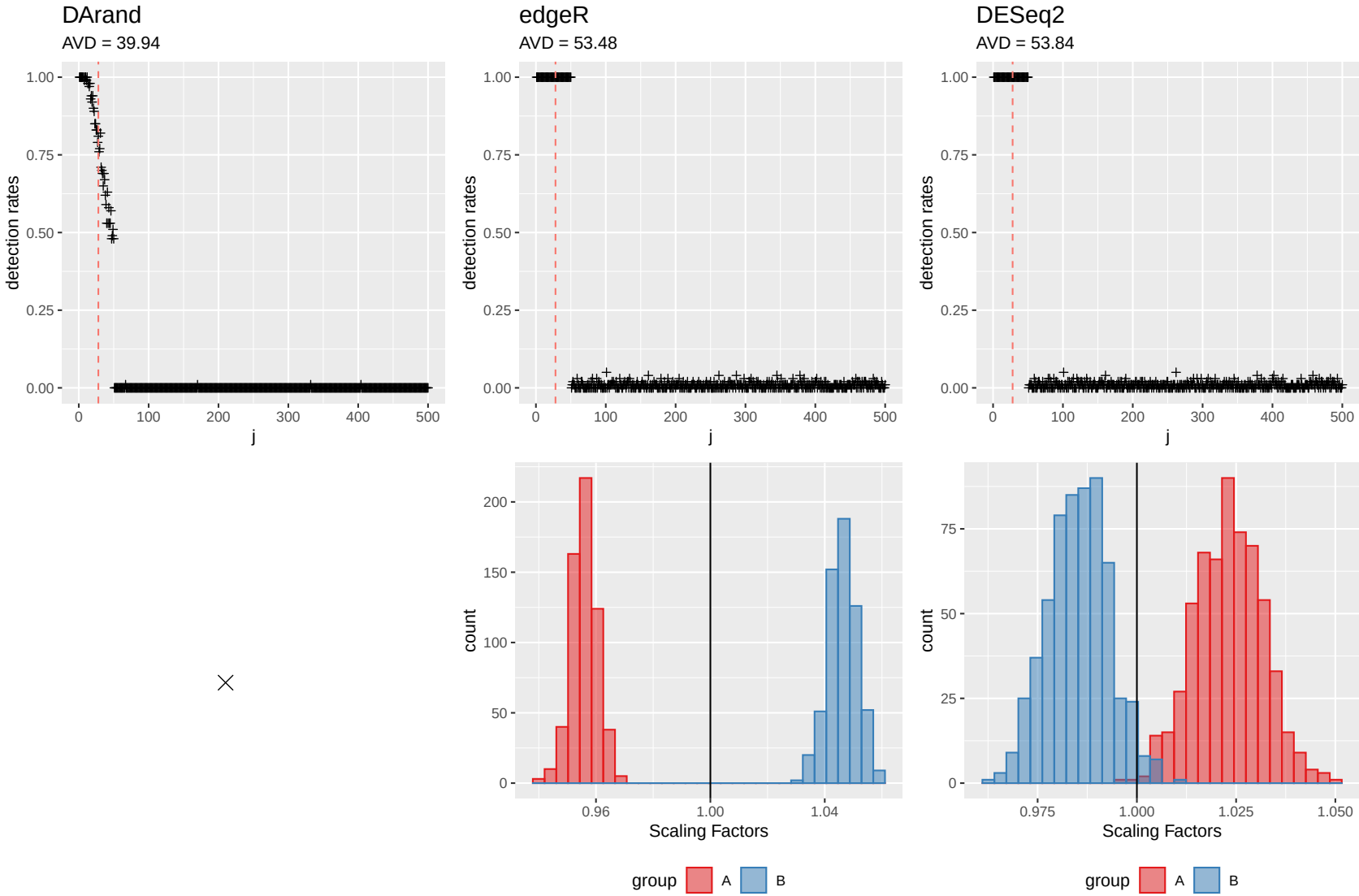


Comparison of methods : Negative Binomial $\phi = 0$ (global null)



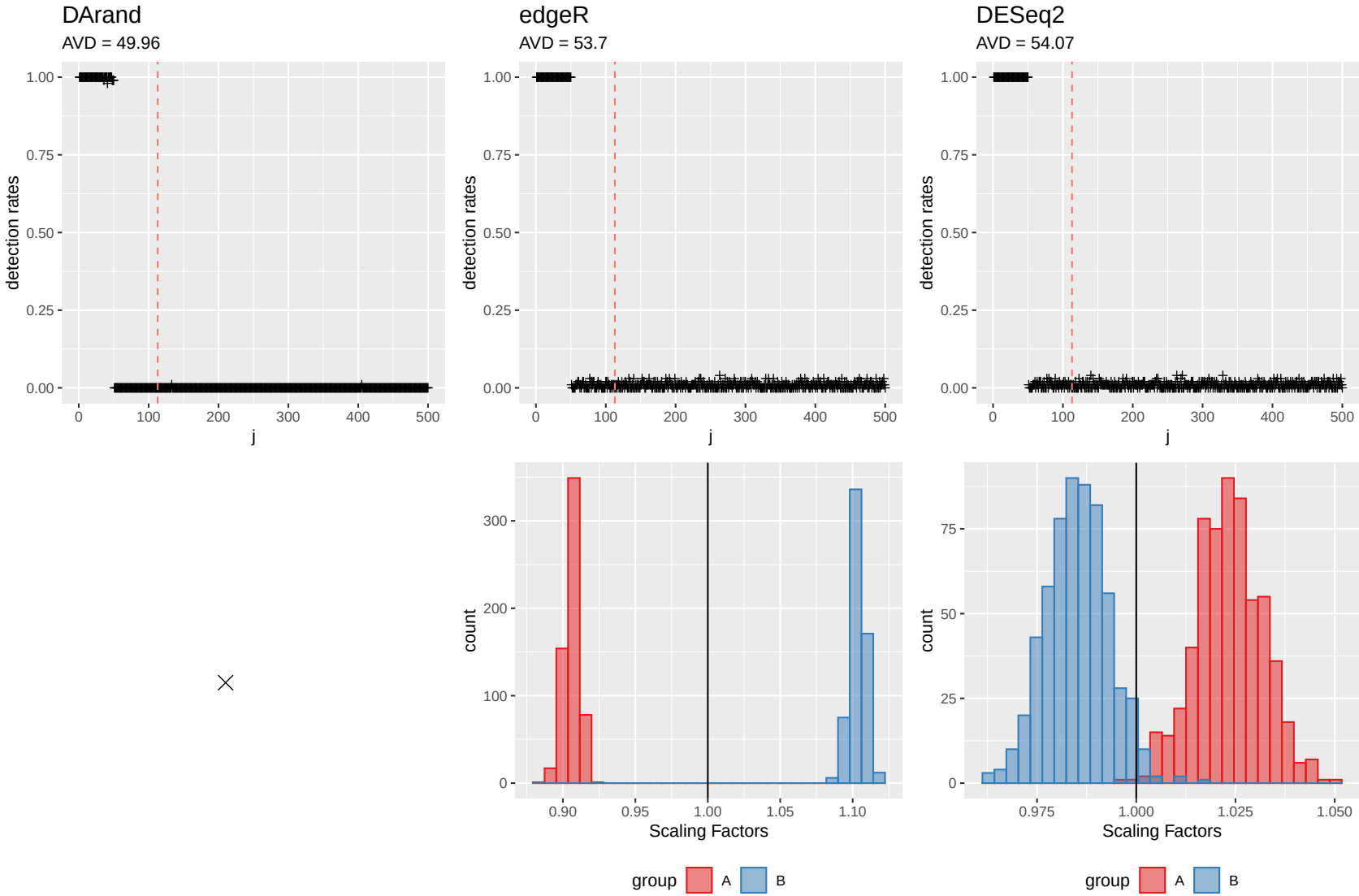
100% upregulated in A

Comparison of methods : Negative Binomial $\phi = 5/\sqrt{j}$ with $d = 50$



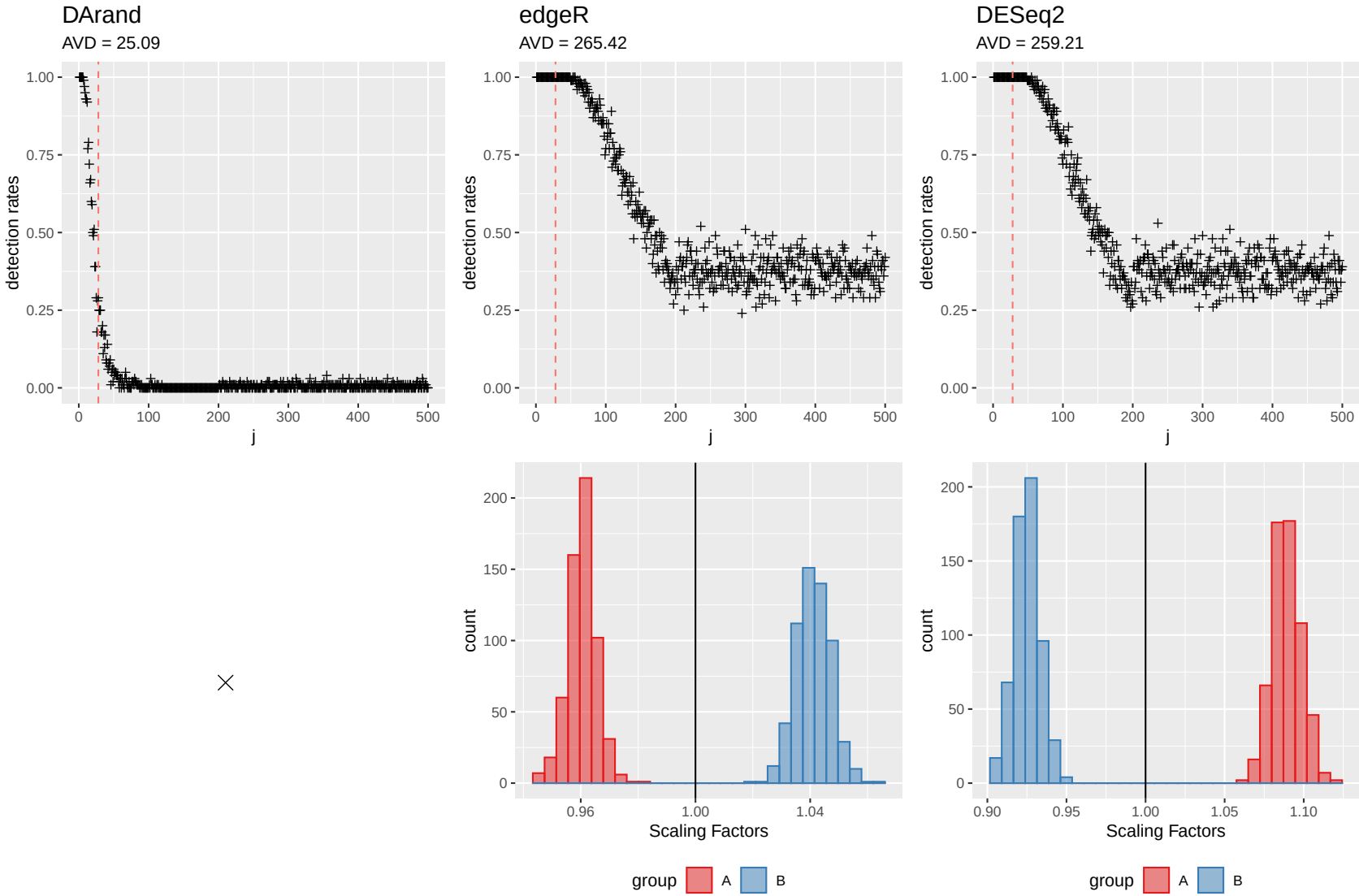
100% upregulated in A

Comparison of methods : Negative Binomial $\phi = 10/\sqrt{j}$ with $d = 50$



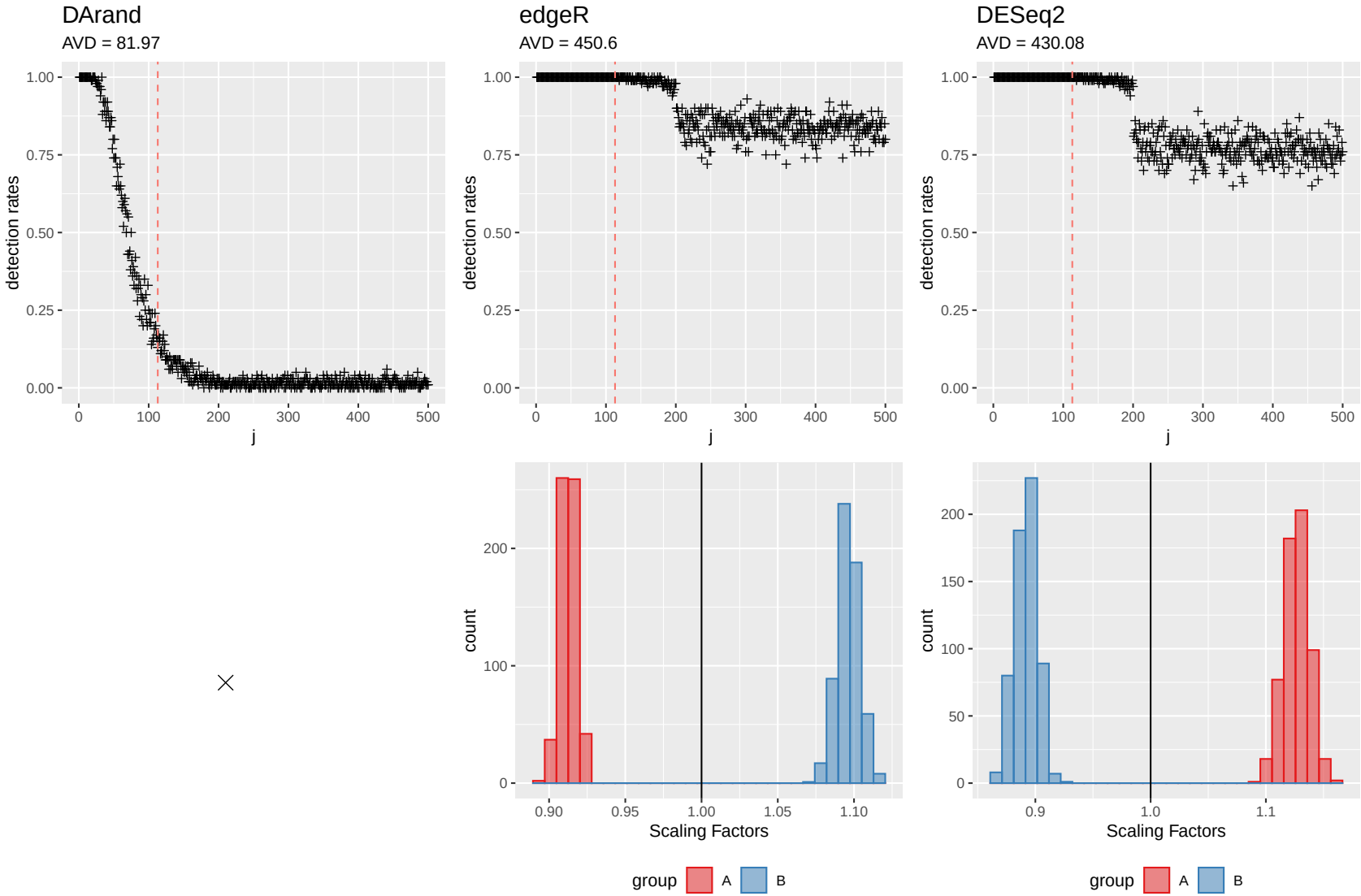
100% upregulated in A

Comparison of methods : Negative Binomial $\phi = 5/\sqrt{j}$ with $d = 200$



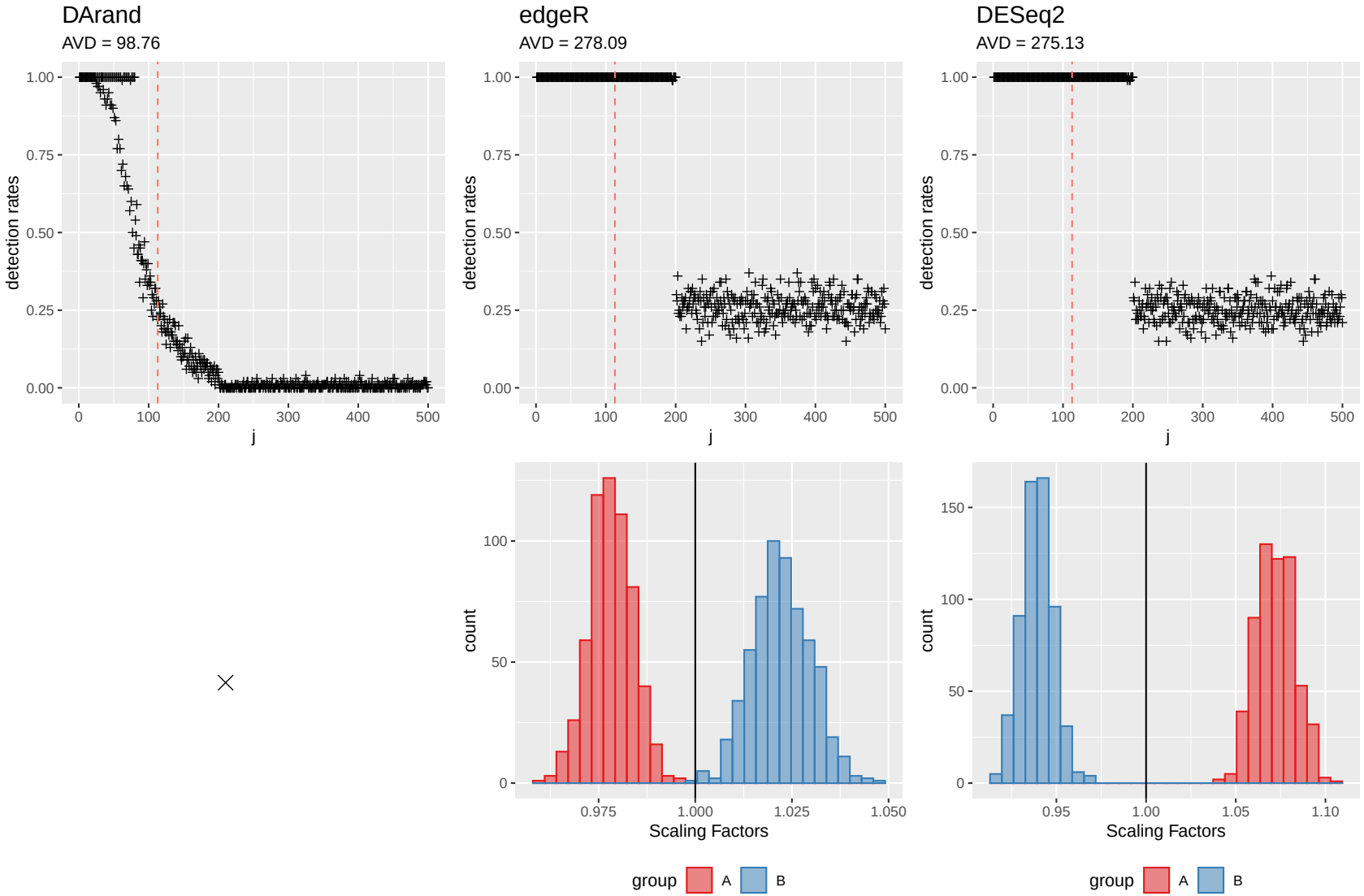
100% upregulated in A

Comparison of methods : Negative Binomial $\phi = 10/\sqrt{j}$ with $d = 200$



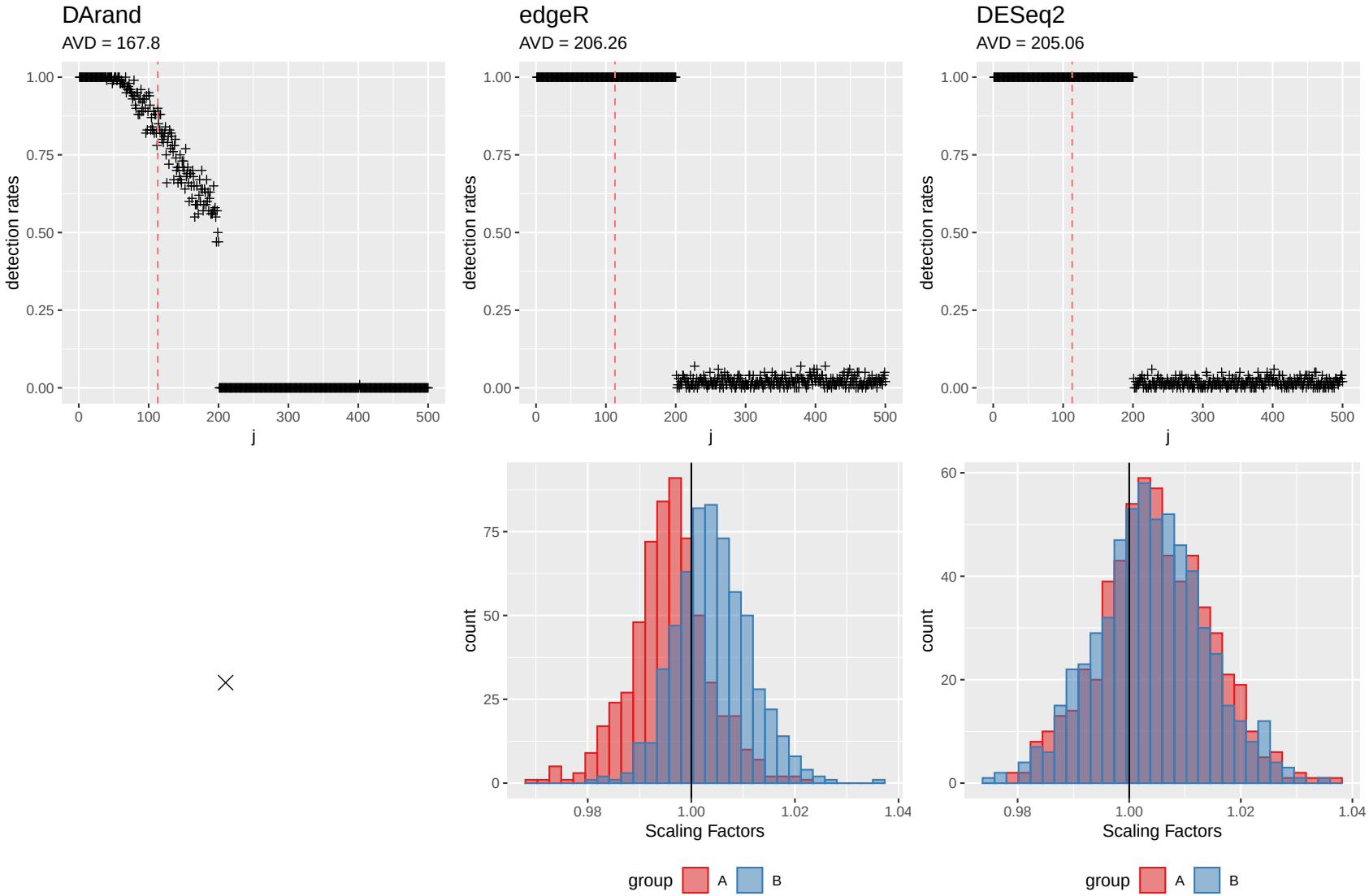
100% upregulated in A

Comparison of methods : Negative Binomial $\phi = 10/\sqrt{j}$ with $d = 200$



80% upregulated in A

Comparison of methods : Negative Binomial $\phi = 10/\sqrt{j}$ with $d = 200$



50% upregulated in A

Real data - Mice model of NASH

Differential analysis of miRNA in mice between 4 NASH models and 4 controls.

Test using miRNA showing more than 20 total reads

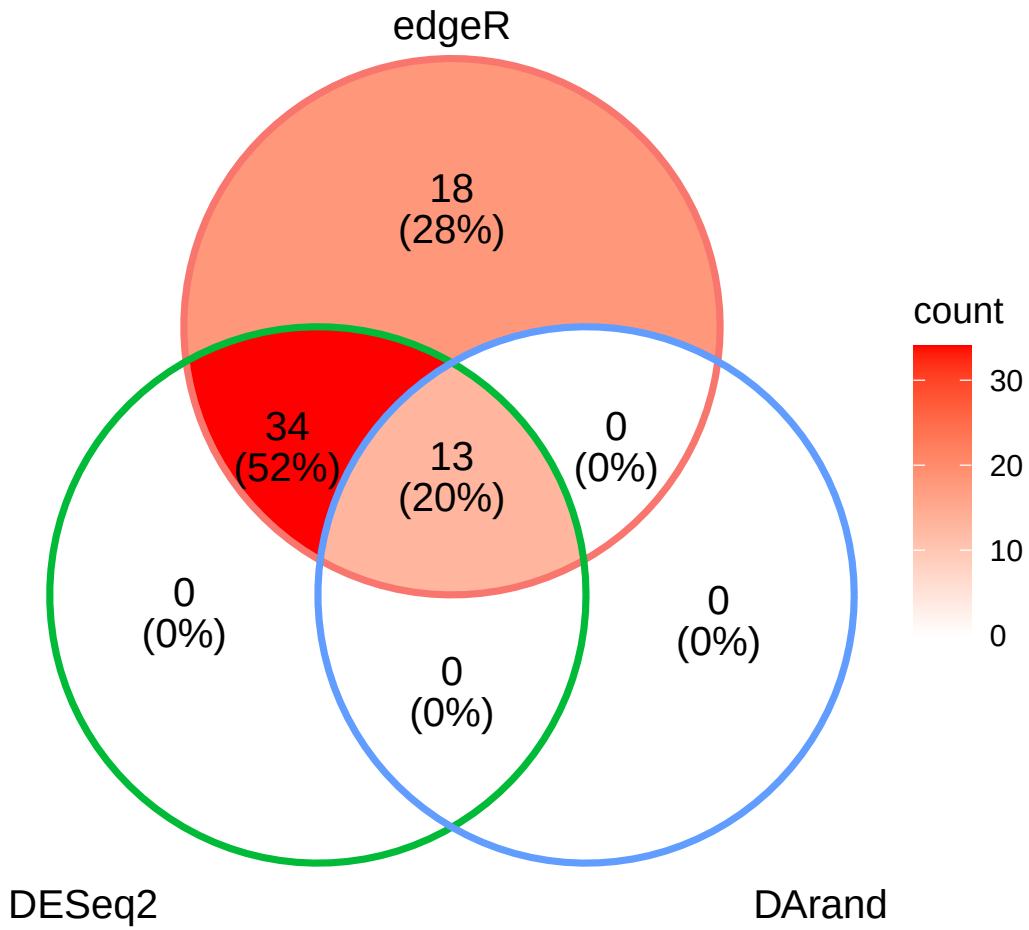
with $|S| = 10$ random reference genes.

The procedure detectes the total of 100 miRNA differentially expressed :

mmu-miR-31-3p, mmu-miR-31-5p, mmu-miR-34a-5p, mmu-miR-141-3p, mmu-miR-141-5p, mmu-miR-200a-3p, mmu-miR-200b-3p, mmu-miR-200b-5p, mmu-miR-200c-3p, mmu-miR-429-3p, mmu-miR-582-5p, mmu-miR-802-3p, mmu-miR-802-5p.

- First 11 miRNAs are related to fibrosis.
- The activity of the last three miRNAs is not specific to fibrosis : **mmu-miR-582-5p,mmu-miR-802-3p, mmu-miR-802-5p**

Real data - Mice model of NASH : DArand, edgeR, DESeq2



Conclusions

- New framework for differential analysis in transcriptomics with good control of the statistical errors (Desaulle et al. 2021)
- Implemented in the R package DArand (Desaulle and Rozenholc 2021)
- Lower false discovery rates when the number of differentially expressed genes increases comparing to DESeq2 and edgeR

Extentions

- Intensive Randomisation framework is extensible to other type of analysis e.g. in PCA methods for genetic features detection : Multiple Factor Analysis fits the iteratively normalized data stucture.
- Optimisation of the choice of the normalization subset size k and scaling factors iterative estimation.

References

Abrams, Zachary B., Travis S. Johnson, Kun Huang, Philip R. O. Payne, and Kevin Coombes. 2019. “A Protocol to Evaluate RNA Sequencing Normalization Methods.” BMC Bioinformatics 20 (24) : 679. https://doi.org/10.1186/s12859-019-3247-x.

Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” Genome Biology 11 (October) : R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Bullard, James H, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. 2010. “Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments.” BMC Bioinformatics 11 (February) : 94. https://doi.org/10.1186/1471-2105-11-94.

Desaulle, Dorota, Céline Hoffmann, Bernard Hainque, and Yves Rozenholc. 2021. “Differential Analysis in Transcriptomic : The Strength of Randomly Picking ’Reference’ Genes.” https://arxiv.org/abs/2103.09872.

Desaulle, Dorota, and Yves Rozenholc. 2021. DArand : Differential Analysis with Random Reference Genes. https://CRAN.R-project.org/package=DArand.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2013. “A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis.” Brief Bioinform 14 (6) : 671–83. https://doi.org/10.1093/bib/bbs046.

Gendre, Xavier. 2014. “Model Selection and Estimation of a Component in Additive Regression.” ESAIM : Probability and Statistics 18 : 77–116. https://doi.org/10.1051/ps/2012028.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” Genome Biology 15 : 550. https://doi.org/10.1186/s13059-014-0550-8.

Marioni, John C., Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. 2008. “RNA-Seq : An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays.” Genome Research 18 (9) : 1509–17. https://doi.org/10.1101/gr.079558.108.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” Nature Methods 5 (7) : 621–28. https://doi.org/10.1038/nmeth.1226.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “edgeR : A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” Bioinformatics 26 (1) : 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Robinson, Mark D., and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” Genome Biology 11 (March) : R25. https://doi.org/10.1186/gb-2010-11-3-r25.

Vandesompele, Jo, Katleen De Preter, Filip Pattyn, Bruce Poppe, Nadine Van Roy, Anne De Paepe, and Frank Speleman. 2002. “Accurate Normalization of Real-Time Quantitative RT-PCR Data by Geometric Averaging of Multiple Internal Control Genes.” Genome Biology 3 (7) : research0034.1–11. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC126239/.