Meta-analysis of RNA-Seq data

Guillemette Marot

Meta-analysis of RNA-Seq data

Guillemette Marot
Univ. Lille, CHU Lille METRICS & Inria MODAL

28 June 2022

Overview

Meta-analysis

Particular context :

- Differential expression analysis
- Several studies available but direct comparison impossible
- Very few individuals in each individual study, a lot of genes or transcripts.

Objectives of meta-analysis

- Increase of sensitivity, the proportion of truly declared differentially expressed (DE) genes
- Eliminate false positives (genes declared DE but not DE in reality)

Overview

Statistical analysis of one study

Research of differentially expressed genes

- Normalisation
- Use of a test statistic appropriate for differential expression analysis
- Correction for multiple testing

	Microarrays	**RNA-Seq**
Information	Intensities	Counts of reads
Modelling	Normal distribution	Poisson, Negative binomial
Tests	Moderated t-tests	Likelihood ratio tests

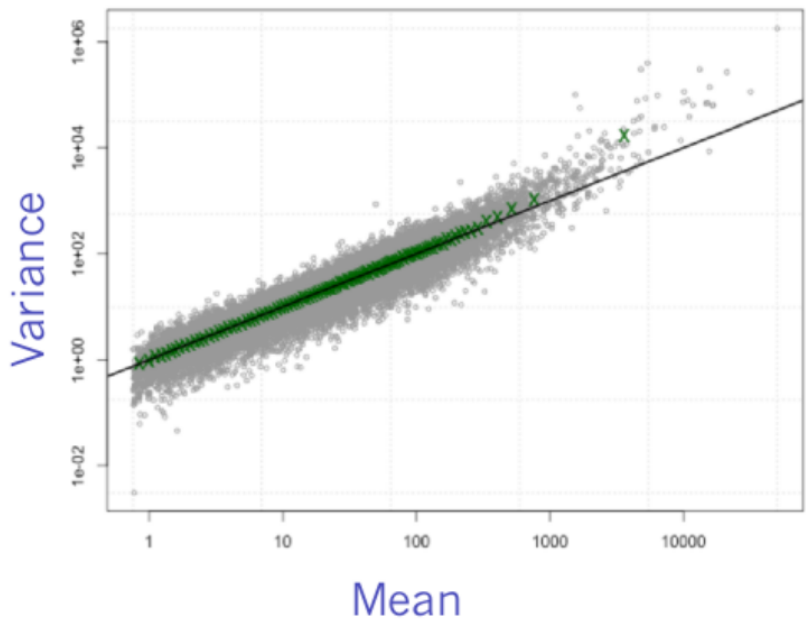
Adaptation of usual test statistics to find a compromise between a gene-by-gene analysis and a "common parameter" analysis

Mean-variance relationship in RNA-Seq data

The Poisson distribution to model counts

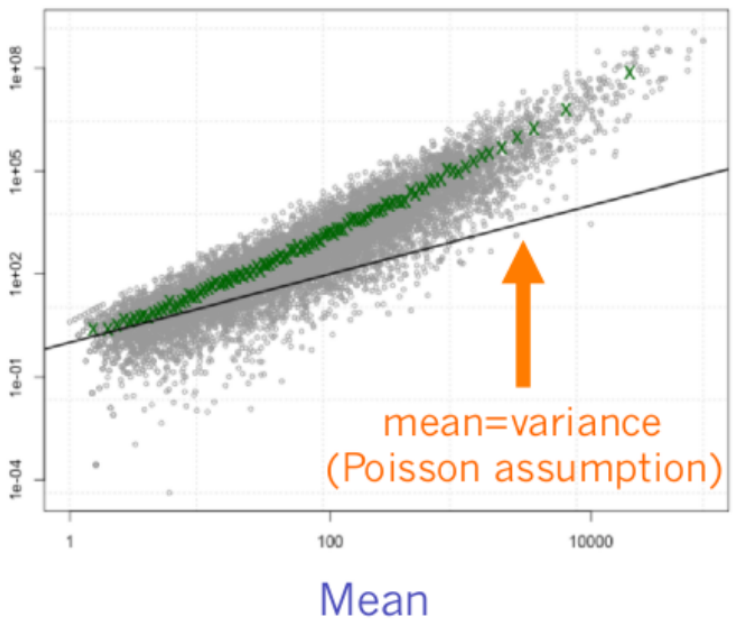
- Describes the number of occurences of rare events during a given time interval
- Property : Mean = Variance

Technical replicates



data from Marioni et al. Gen Res 2008

Biological replicates

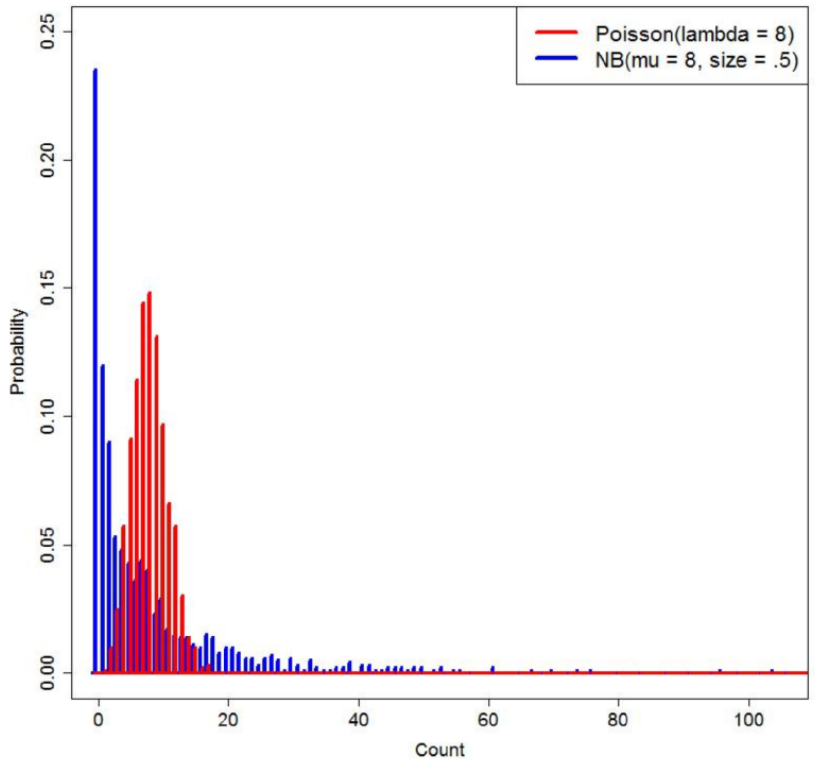


data from Parikh et al. *Genome Bio* 2010

Negative Binomial Models

A supplementary dispersion parameter ϕ to model the variance

Poisson vs Negative Binomial models



Technical variability is the main source of variability in low counts, whereas biological variability is dominant in high counts

Estimating the dispersion : the key question

Problem

Estimate a reliable dispersion from a very small number of replicates (sometimes less than 5)

Why using sophisticated approaches ?

- gene-specific tests $\Rightarrow$ lack of sensitivity (proportion of true positives among positives) due to the lack of information
- common dispersion parameter for all tests $\Rightarrow$ many false positives

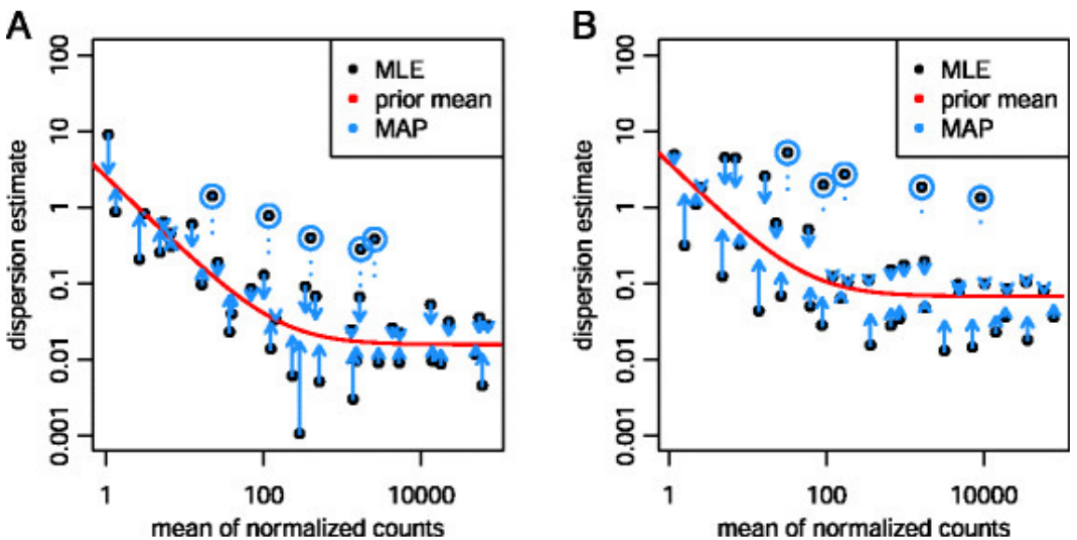
Empirical bayesian approaches = compromise between gene-specific and common dispersion parameter estimation

Exemple : edgeR, DESeq2

Dispersion estimation with DESeq2

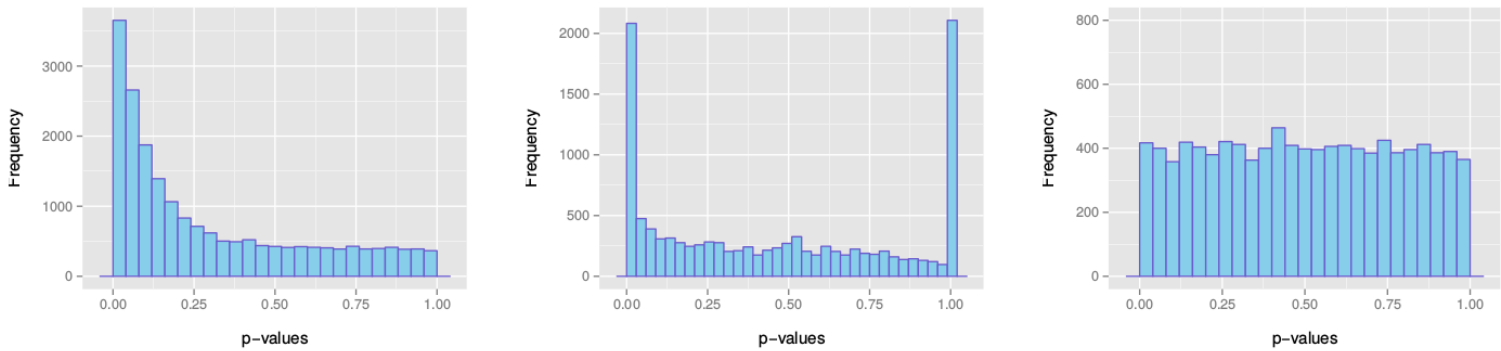
Hypothesis : genes of similar average expression strength have similar dispersion

1. Estimate gene-wise dispersion estimates using maximum likelihood (ML) (black dots)
2. Fit a smooth curve (red line)
3. Shrink the gene-wise dispersion estimates (empirical Bayes approach) toward the values predicted by the curve to obtain final dispersion values (blue arrow heads).



p-values histograms for diagnosis

Examples of expected overall distribution



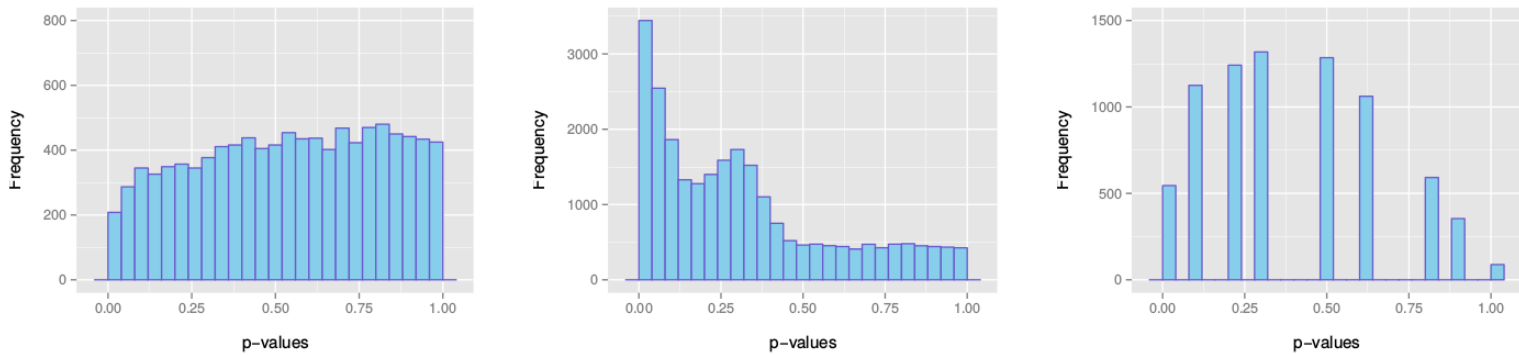
(a) : the most desirable shape

(b) : very low counts genes usually have large p-values

(c) : do not expect positive tests after correction

p-values histograms for diagnosis

Examples of not expected overall distribution



(a) : indicates a batch effect (confounding hidden variables)

(b) : the test statistics may be inappropriate (due to strong correlation structure for instance)

(c) : discrete distribution of p-values : unexpected

Overview

Back to microarray meta-analysis

(Marot et al., Bioinformatics, 2009)
$\Rightarrow$ p-value combination showed better performance in terms of sensitivity and AUC than effect size combination.

p-value combination was suggested with the inverse normal method (Liptak, 1958)

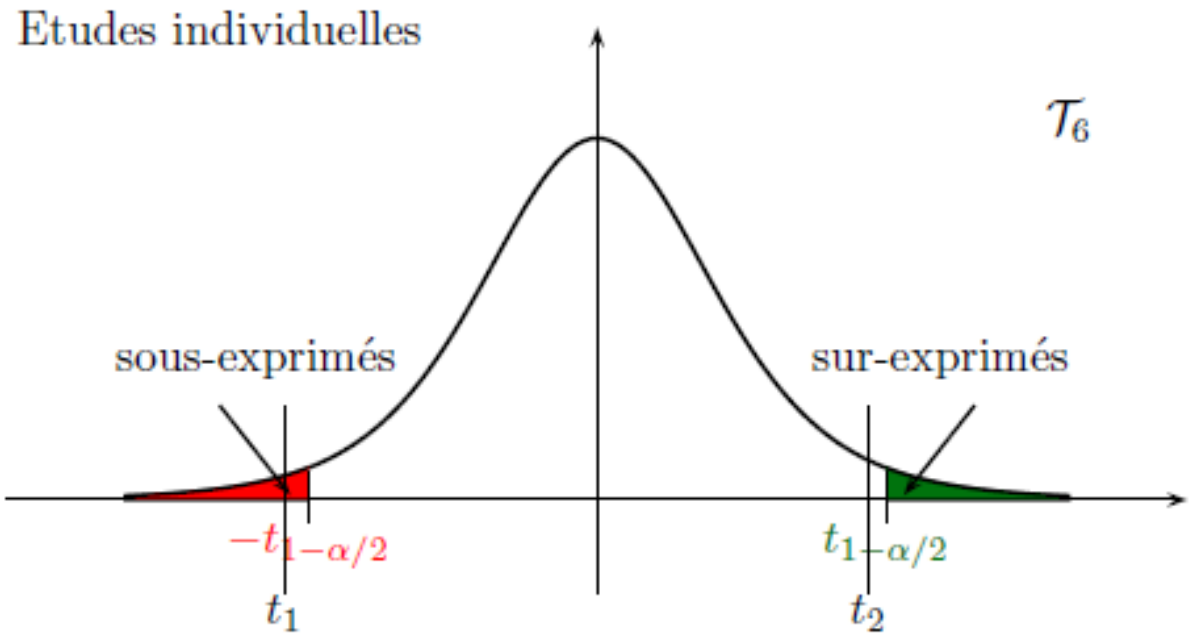
$$N_g = \sum_{s=1}^{S} w_s \Phi^{-1}(1 - p_{gs})$$

$$w_s = \sqrt{\frac{\sum_c R_{cs}}{\sum_\ell \sum_c R_{c\ell}}}$$

$$N_g \sim \mathcal{N}(0, 1)$$

Trick for microarrays : use unilateral p-values to avoid conflicts.

Back to microarray meta-analysis



Gene for which we observe a conflict of expression

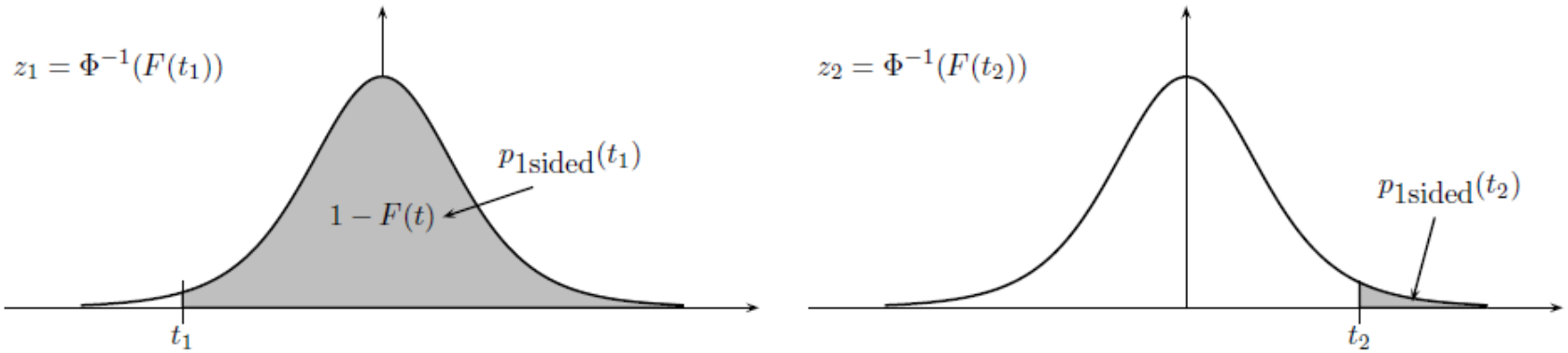
$t_1 < 0$ observed value of test statistic for study 1

$t_2 > 0$ observed value of test statistic for study 2

Back to microarray meta-analysis

scores from unilateral p-values

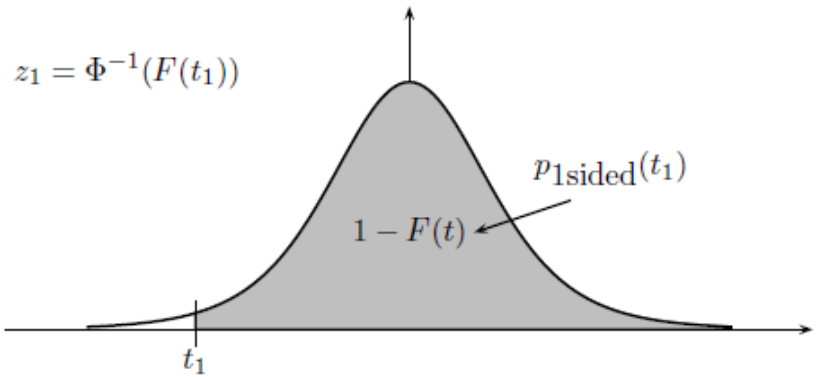
$z = \Phi^{-1}(1 - p_{1\text{sided}}) = \Phi^{-1}(F(t))$

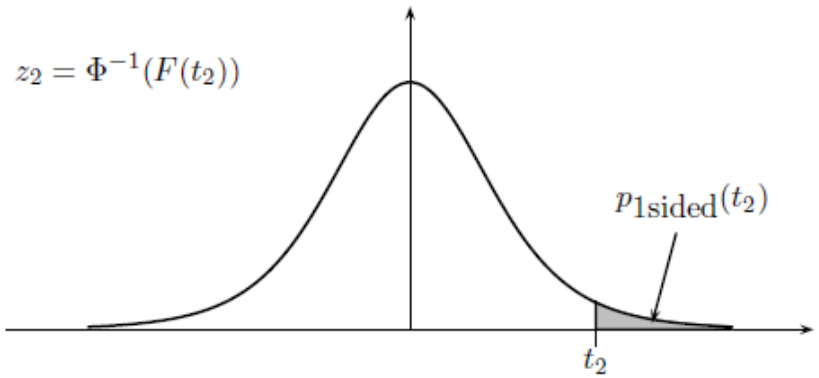


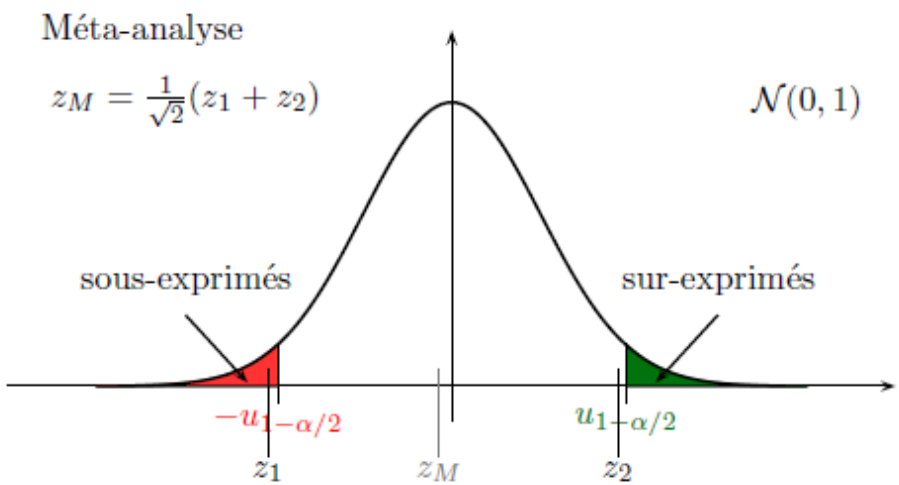
Back to microarray meta-analysis

scores from unilateral p-values

$z = \Phi^{-1}(1 - p_{\text{1sided}}) = \Phi^{-1}(F(t))$



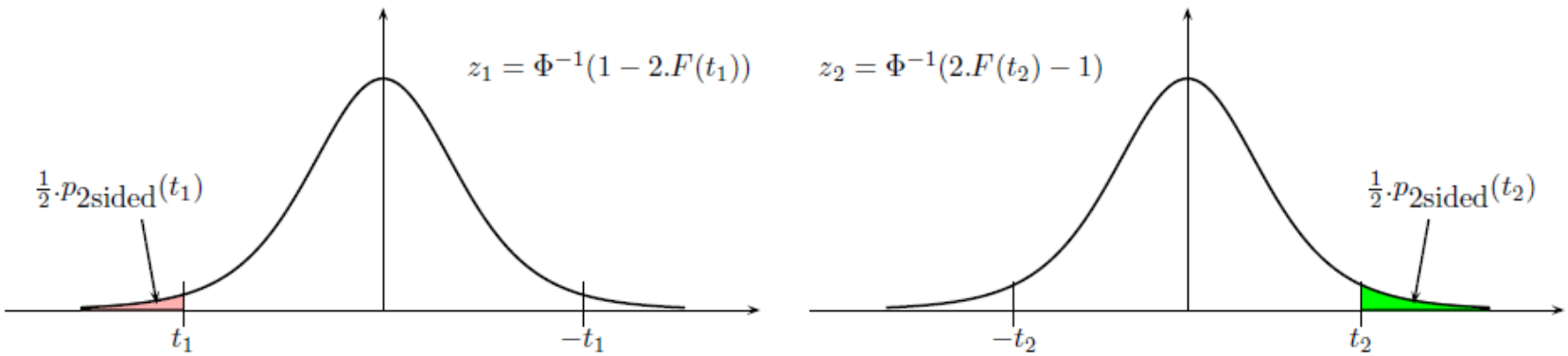




Back to microarray meta-analysis

scores from bilateral p-values

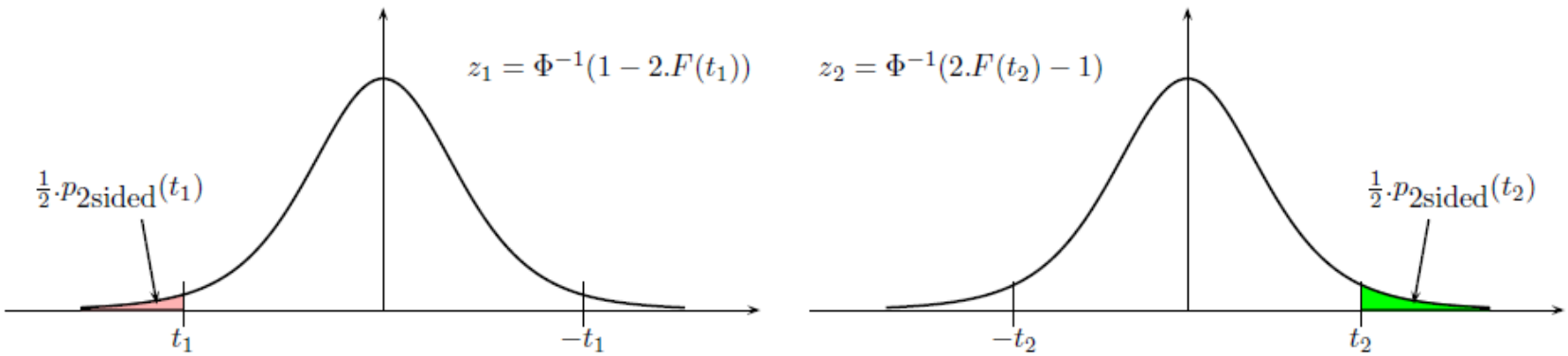
$z = \Phi^{-1}(1 - p_{2\text{sided}}) = \Phi^{-1}(1 - 2.(1 - F(|t|)))$

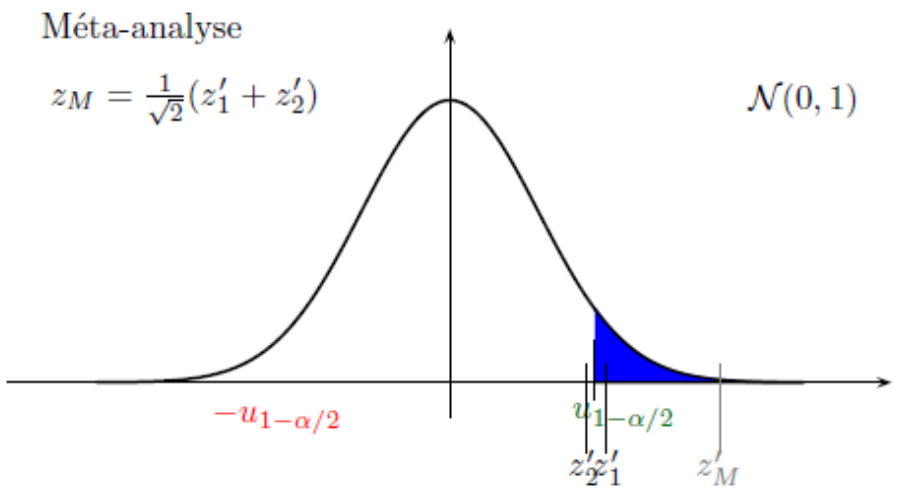


Back to microarray meta-analysis

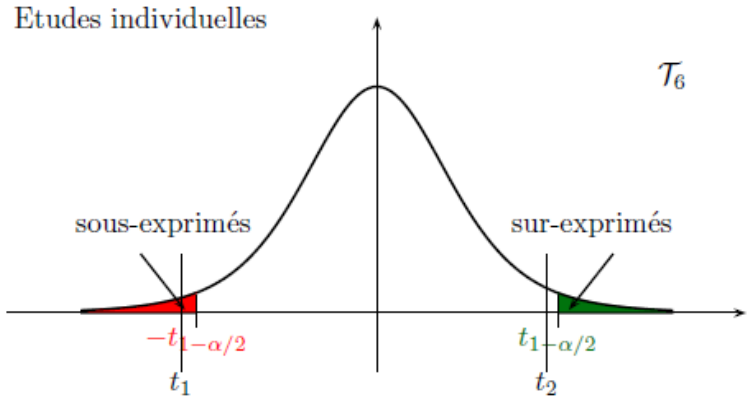
scores from bilateral p-values

$z = \Phi^{-1}(1 - p_{2\text{sided}}) = \Phi^{-1}(1 - 2.(1 - F(|t|)))$

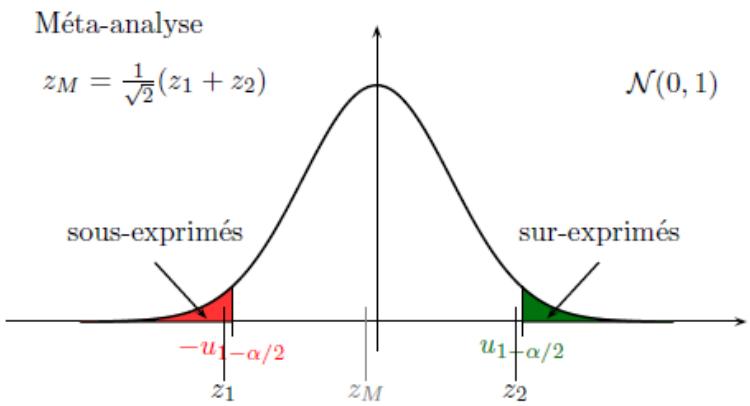




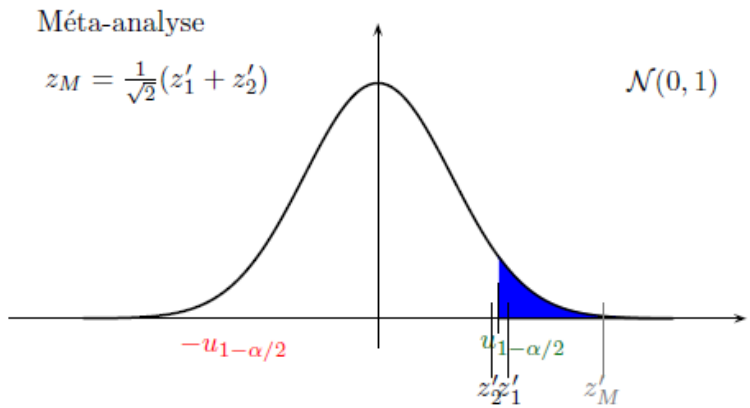
Back to microarray meta-analysis



Building of individual scores from p-values

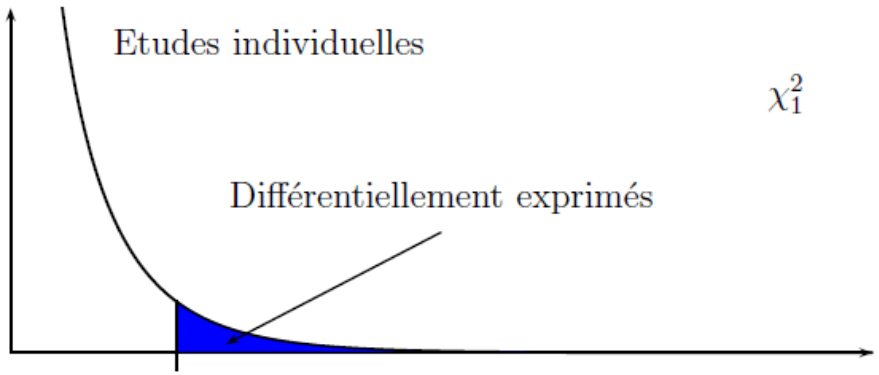


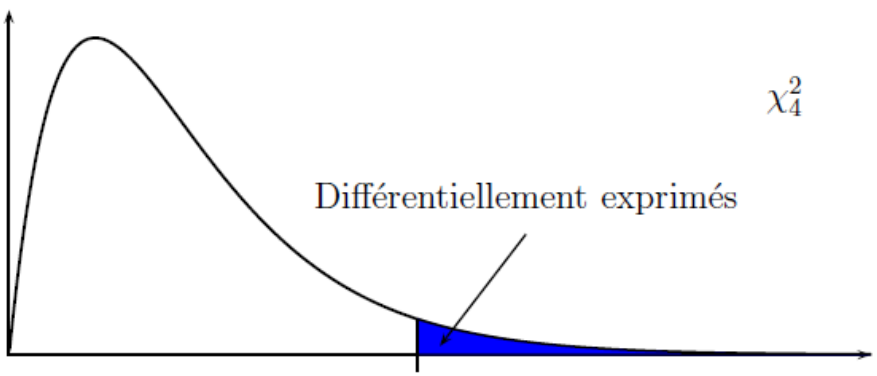
unilateral
then bilateral test

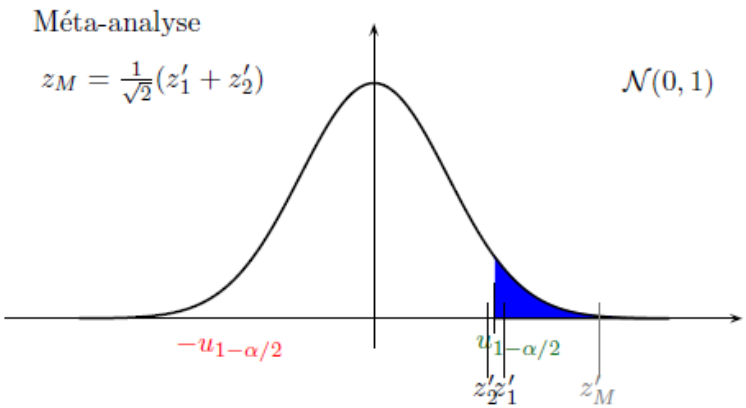


bilateral
then uni(bi)lateral test

RNASeq data meta-analysis







RNA-seq data meta-analysis

Inverse normal combination (Liptak, 1958)

$$N_g = \sum_{s=1}^{S} w_s \Phi^{-1}(1 - p_{gs})$$

$$N_g \sim \mathcal{N}(0, 1)$$

Fisher's method (1932)

$$F_g = -2 \sum_{s=1}^{S} \ln(p_{gs})$$

$$F_g \sim \chi^2_{2S}$$

Conflicts to be treated a posteriori

Global differential analysis (DESeq)

Gene counts : $Y_{gcrs} \sim \mathcal{NB}(\eta_{gcrs}, \phi_{gs})$

Full model $\log(\eta_{gcrs}) = \beta_g + \lambda_{gc} + \delta_{gs} + \log(\ell_{crs})$
where β_g is an intercept, λ_{gc} is a fixed condition effect, δ_{gs} a fixed study effect, ℓ_{crs} library size normalization factor

Reduced model $\log(\eta_{gcrs}) = \beta_g + \delta_{gs} + \log(\ell_{crs})$

Likelihood ratio test
$H_{0,g} : \forall c, \lambda_{gc} = 0$ vs $H_{1,g} : \exists c \mid \lambda_{gc} \neq 0$.
$\sim \chi^2$ with degrees of freedom equal to the number of conditions minus 1

Simulations

Data simulated according to a negative binomial distribution,

$$Y_{gcrs} \sim \mathcal{NB}(\mu_{gcs}, \phi_{gs})$$

where μ_{gcs} and ϕ_{gs} represent the mean and dispersion, respectively

Mean-variance relationship defined by

$$\mathrm{Var}(Y_{gcrs}) = \mu_{gcs} + \frac{\mu_{gcs}^2}{\phi_{gs}}.$$

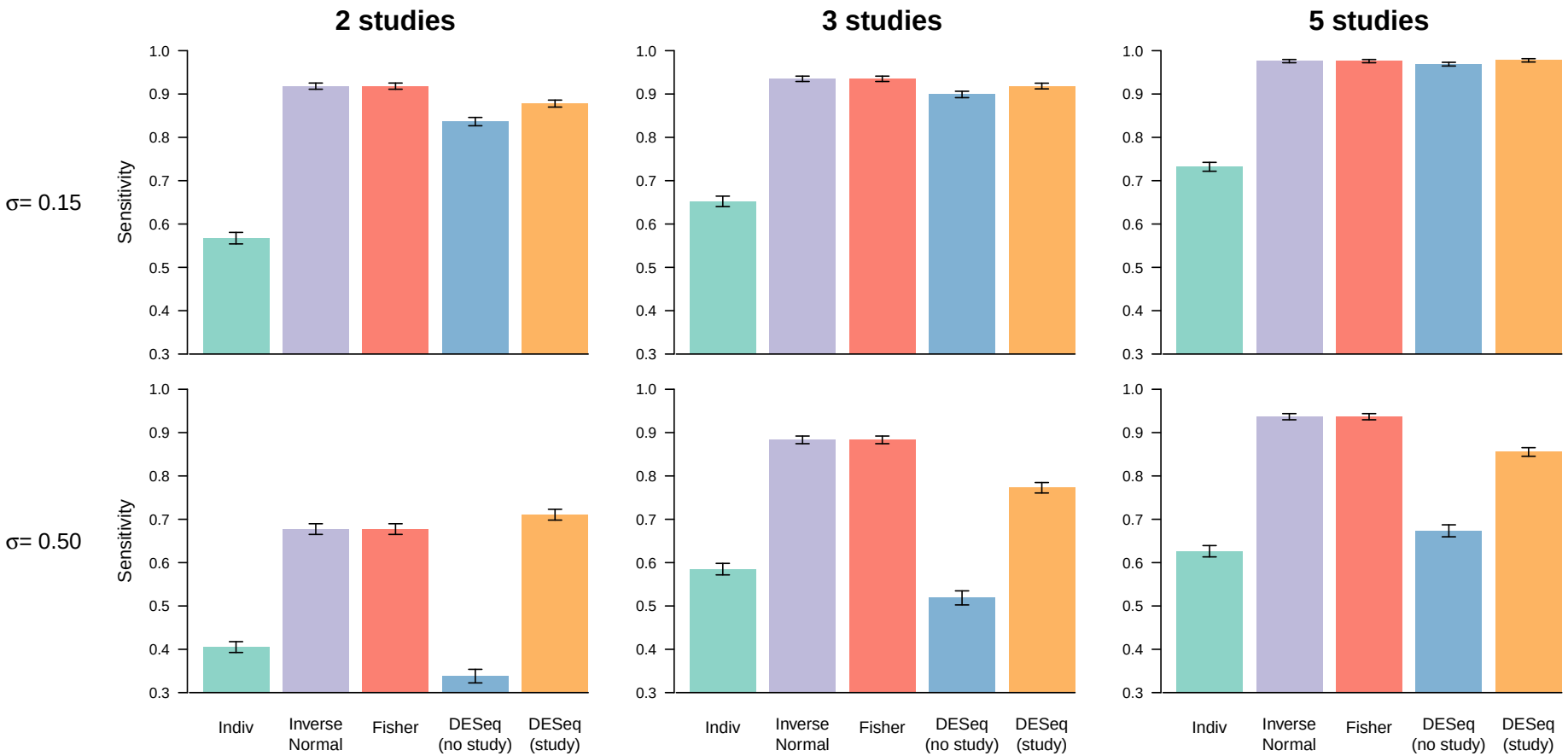
To incorporate inter-study variability :

$$\log(\mu_{gcs}) = \theta_{gc} + \varepsilon_{gcs}, \text{and } \varepsilon_{gcs} \sim \mathcal{N}(0, \sigma^2),$$

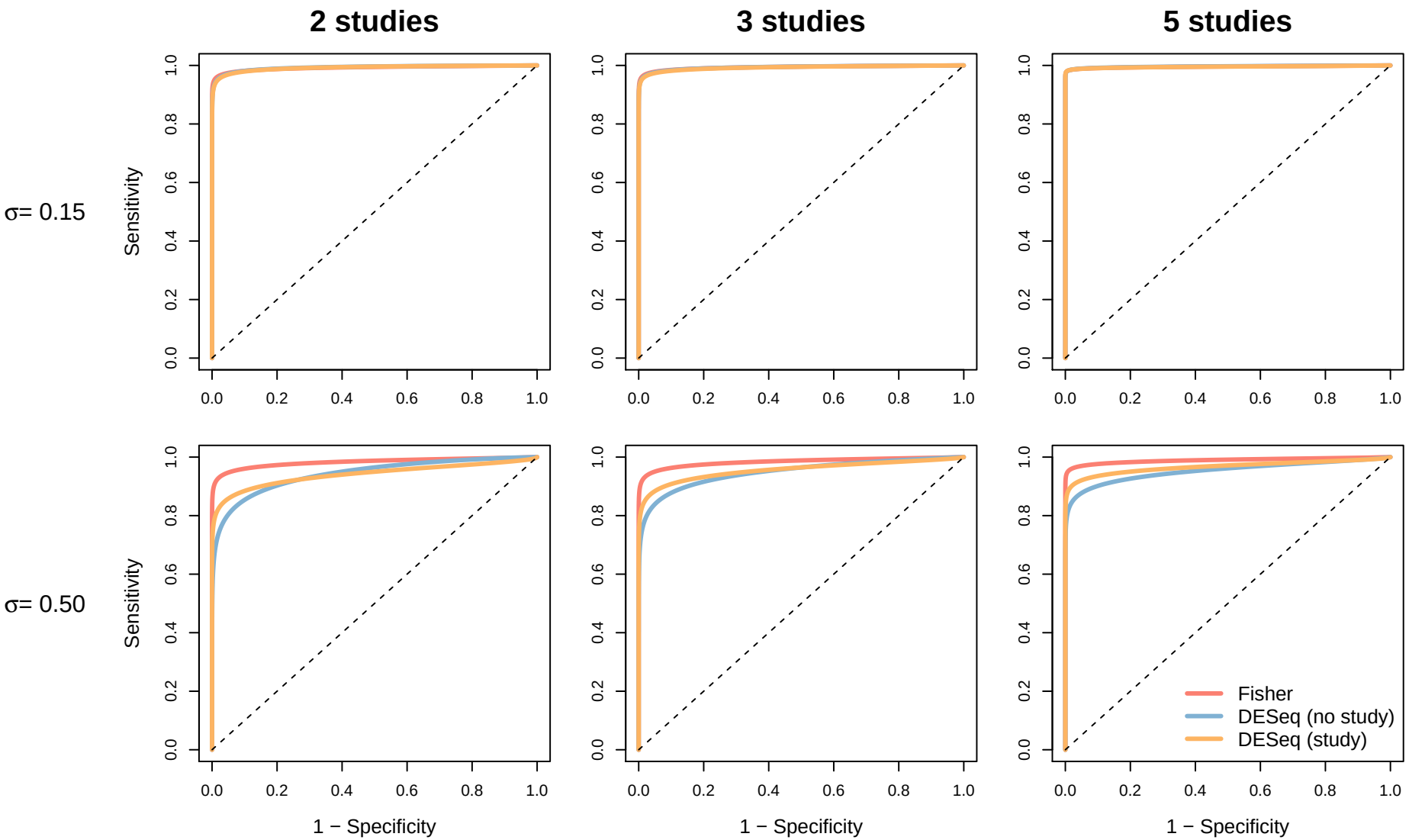
where σ^2 is the size of the inter-study variability.

Meta-analysis

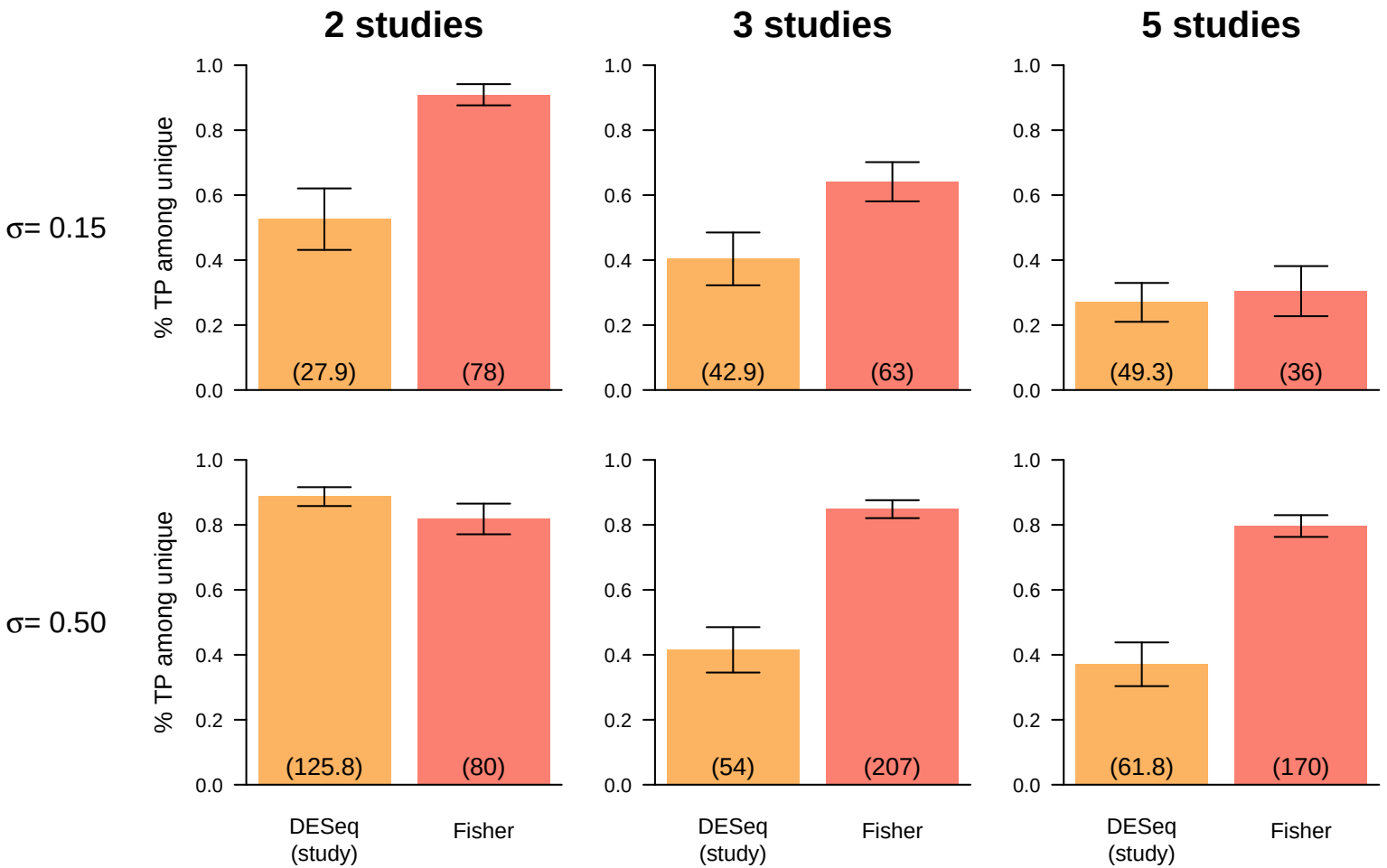
Sensitivity : $E(\frac{VP}{VP+FN})$



Meta-analysis



Meta-analysis



Proportion of true positives among unique discoveries.

Overview

R package

R package metaRNASeq available on CRAN
Publication : Rau, Marot and Jaffrézic, BMC Bioinformatics (2014)

```
library(metaRNASeq)
vignette("metaRNASeq")
```

- data simulation with `sim.function`
- single individual analyses with DESeq2
- use of HTSFilter (if needed) to validate the p-value uniform distribution assumption
- p-value combination with `fishercomb` and `invnorm`
- treatment of conflicts from the extraction of fold changes

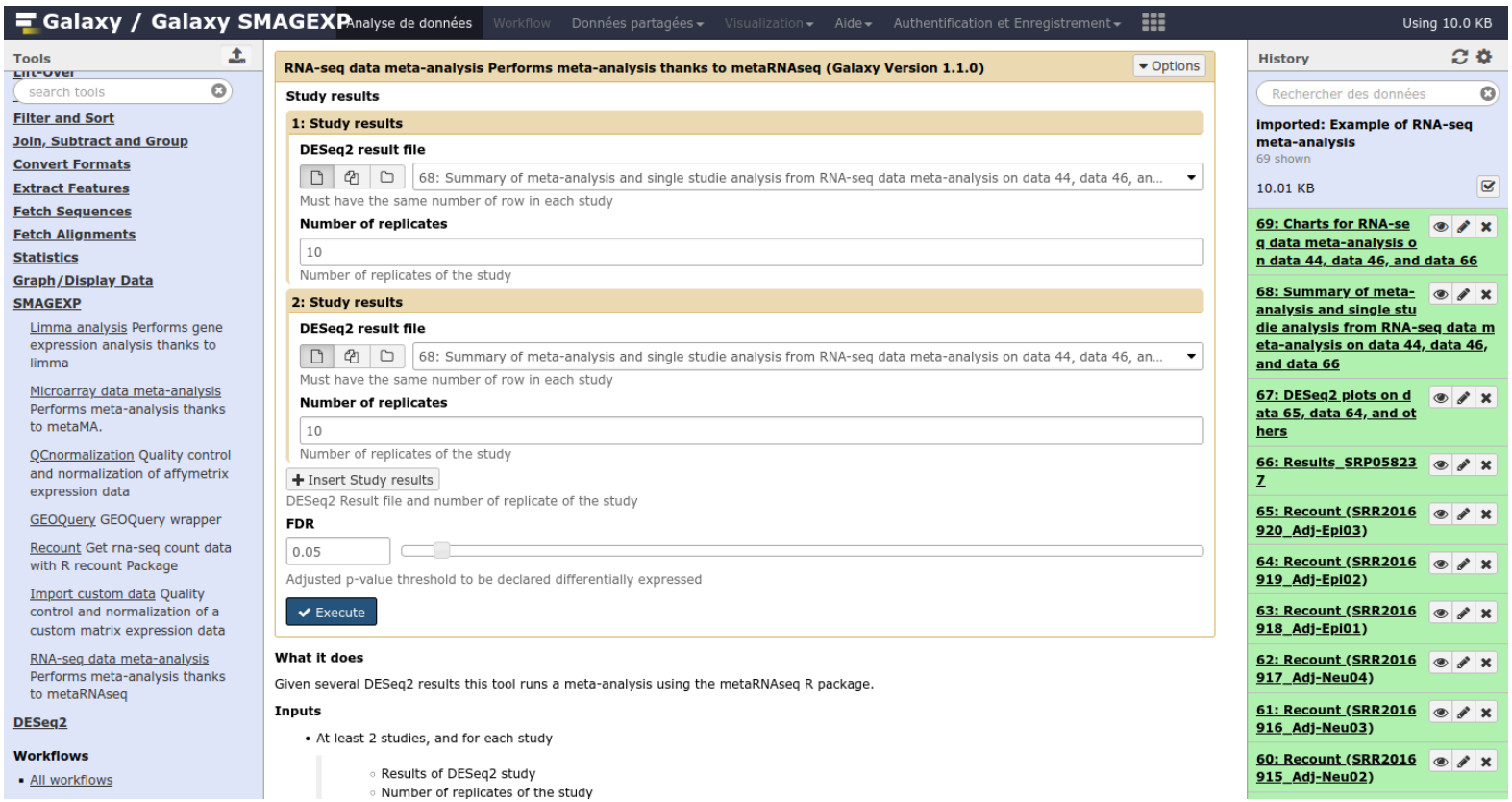
Key figures

- DE (differentially expressed) : number of DE genes
- IDD (integration-driven discoveries) : number of genes that are declared DE in the meta-analysis that were not identified in any of the single studies alone
- Loss : number of genes that are identified DE in single studies but not in meta-analysis
- IDR (integration-driven discovery rate) : corresponding proportion of IDD
- IRR (integration-driven revision) : corresponding proportion of loss

SMAGEXP

SMAGEXP available on Galaxy main tool shed or in a dockerised instance

Publication : Blanck and Marot, Gigascience, (2019)



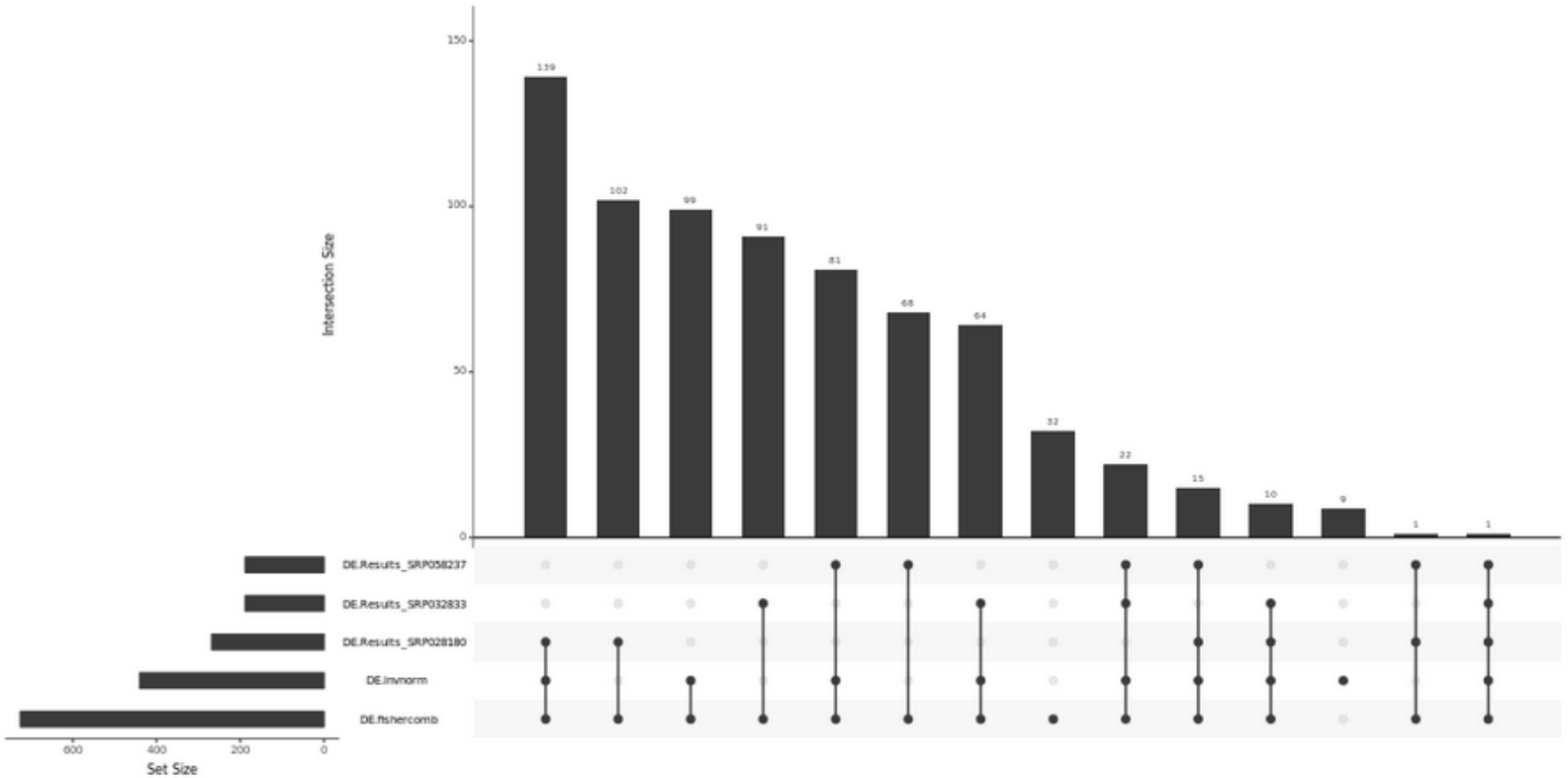
SMAGEXP

Source code, help, and installation instructions available on Github : https ://github.com/sblanck/smagexp

- Step by step example of a RNA-seq meta-analysis
  - Data used in this example
  - First Analysis
    - Run the recount tool
    - Run a DESeq2 analysis
  - Second Analysis
    - Run the recount tool
    - Run a DESeq2 analysis
  - Third Analysis
    - Run the recount tool
    - Run a DESeq2 analysis
  - Run the Meta-analysis with metaRNASeq

SMAGEXP

UPSETR DIAGRAM



Fisher combination summary

DE	IDD	Loss	IDR	IRR
725	131	0	18.07	0

Overview

Conclusion - Discussion

- Meta-analysis useful when strong inter-study effect and more than 3 studies
- p-value combinations enable to take advantage of empirical bayesian approaches - especially appropriate when few replicates
- with RNA-Seq data, necessity to treat conflicts a posteriori
- p-values histograms and PCA graphs enable to decide whether using or not metaRNASeq.
- metaRNASeq available on CRAN
- SMAGEXP available on Galaxy tool shed, Docker, Github.

Acknowledgements

Andrea Rau

Florence Jaffrézic

Samuel Blanck

Claus-Dieter Mayer

Jean-Louis Foulley