

IIDEA : Interactive Inference for Differential Expression Analyses

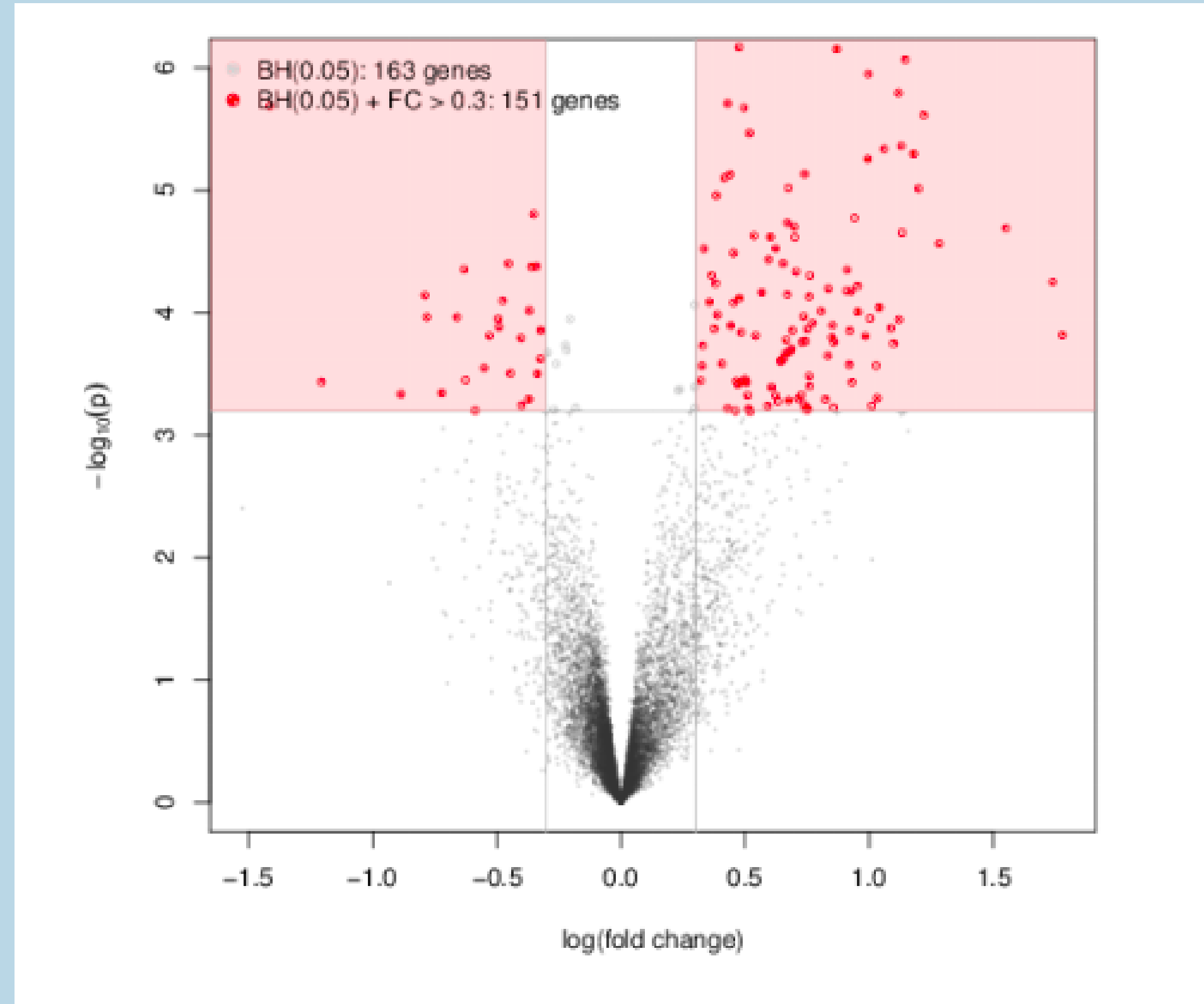
Nicolas Enjalbert-Courrech, Pierre Neuvial

nicolas.enjalbert-courrech@math.univ-tlse.fr, pierre.neuvial@math.cnrs.fr
<https://shiny-iidea-sanssouci.apps.math.cnrs.fr/>



Motivation: volcano plots

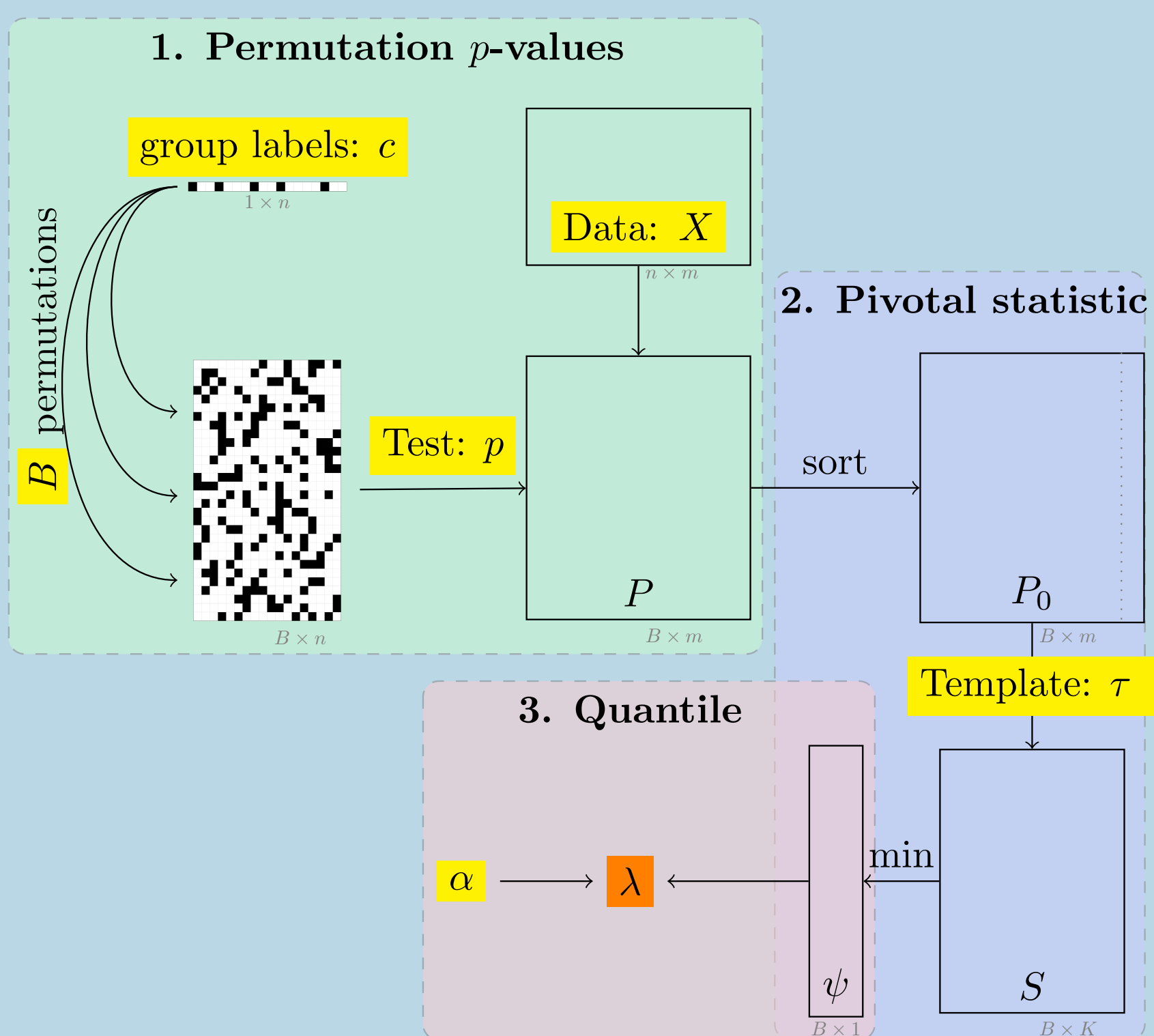
Differential gene expression studies aim at identifying genes whose mean expression level differs significantly between two known populations. **Post hoc inference** provides statistical guarantees on arbitrary gene selections, e.g. based on both statistical significance and effect size, as e.g. in **volcano plots** [1]:



To make post hoc inference more widely used, we combine (1) a **powerful permutation-based post hoc method** [4] with (2) an **R/shiny app for interactive user selections**.

(1) Permutation method

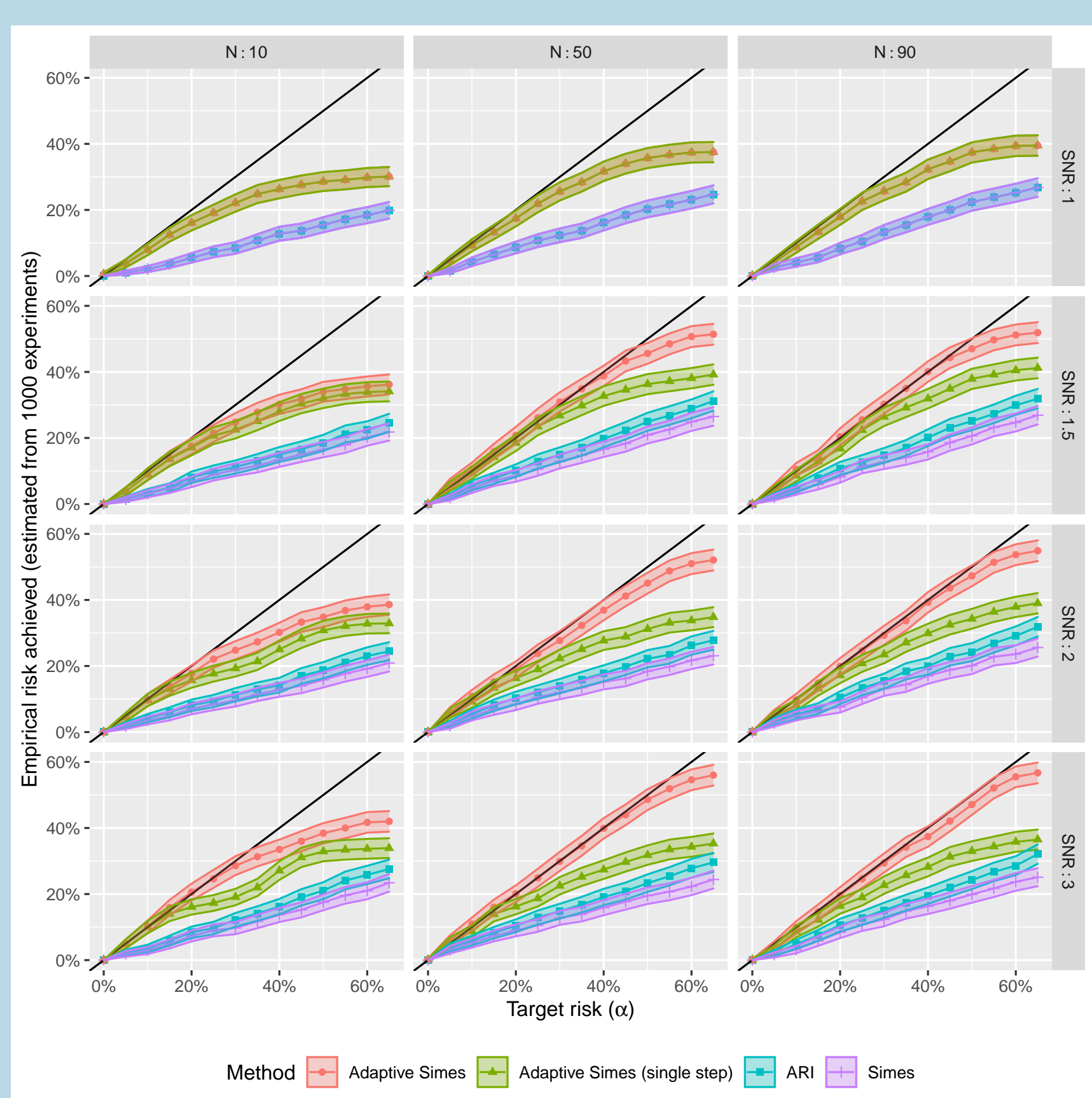
Adaptive Simes, our calibration method based on permutation for two-sample tests, adapts to the dependency structure of data [3, 4].



Complexity: $O(mB(n + \log(m)))$ in time and $O(m(B + n))$ in space.

Statistical performance

Adaptive Simes yields **valid** and **sharp** bounds:



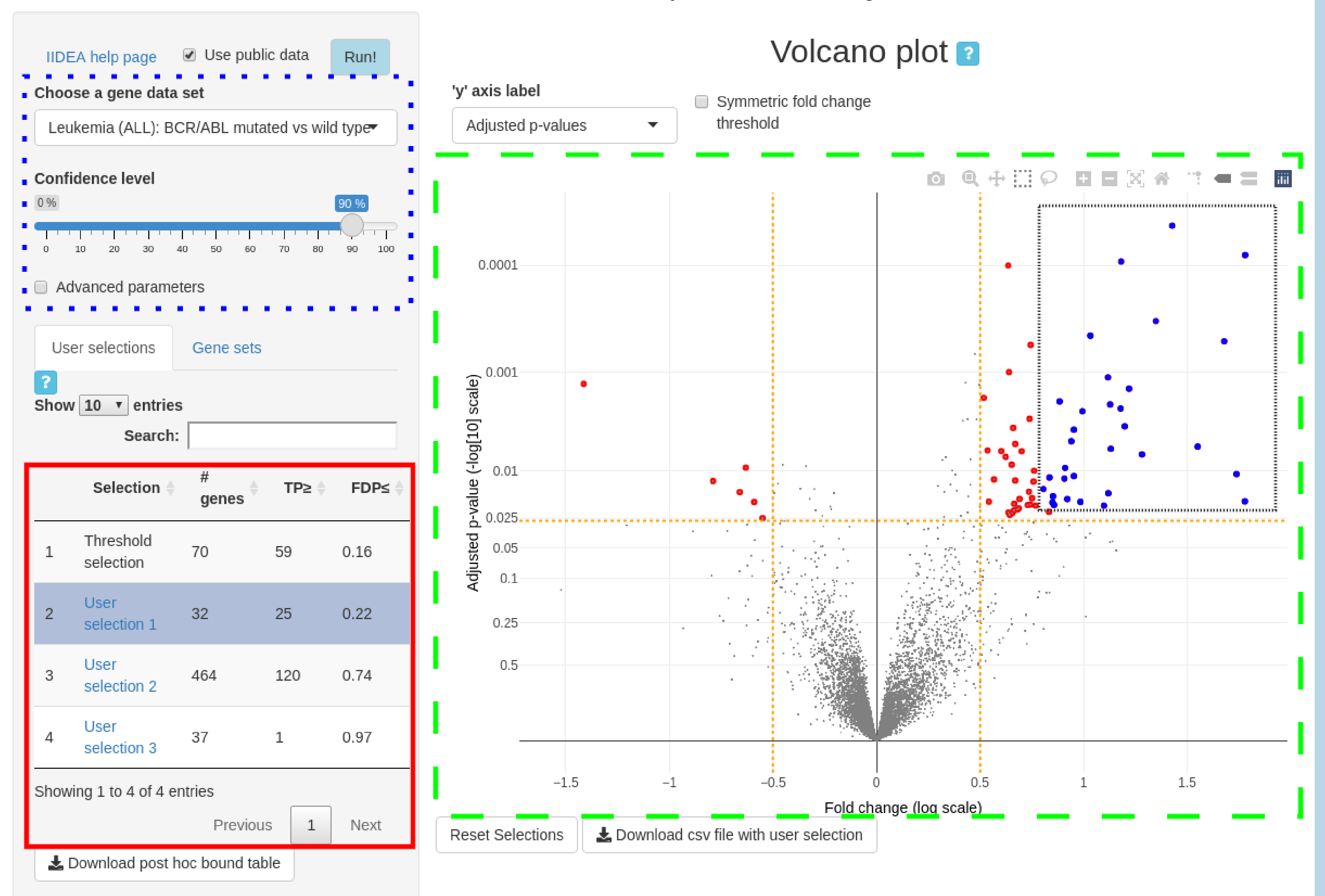
Review of existing post hoc inference methods

Method	R Package	Adaptivity to:		Time complexity(**)
		dependence	π_0 (*)	
Simes [1, 3]	cherry	NO	NO	quadratic
ARI [6]	ARI	NO	YES	linear
permutation ARI [5]	pARI	YES	YES	quadratic
Adaptive Simes [4, 3]	sanssouci	YES	YES	linear

(*) proportion of null genes; (**) for calculating the post hoc bound

(2) R/shiny app for interactive differential expression analysis

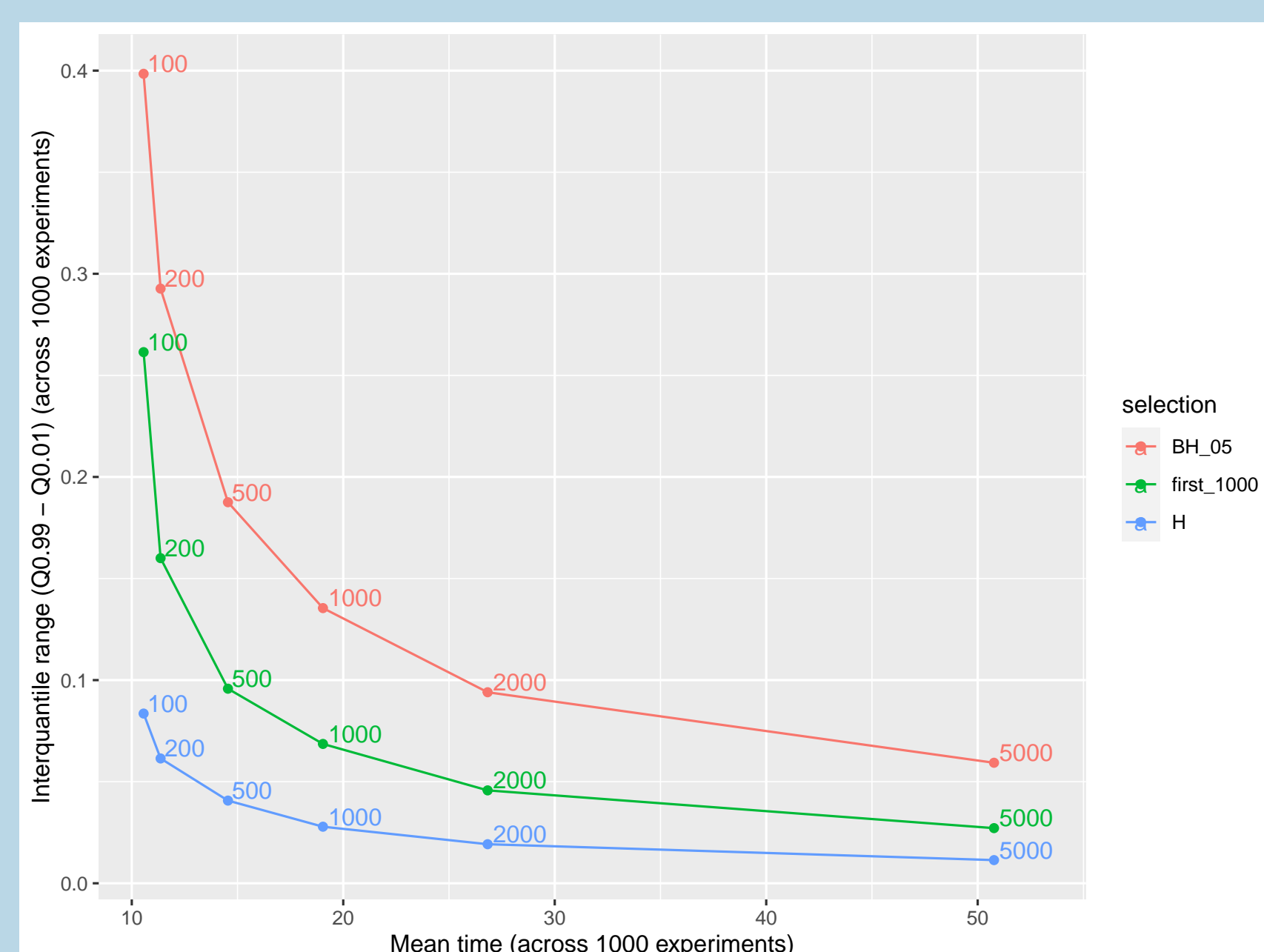
IIDEA: Interactive Inference for Differential Expression Analyses



This interactive shiny application is developed by Nicolas Enjalbert-Courrech and Pierre Neuvial for the R package sansSouci. It implements permutation-based post hoc inference bounds for differential gene expression analysis, see dedicated vignette. The source code for this app is freely available. For any question, please file an issue.

How many permutations?

The number B of permutations at the calibration step induces a tradeoff between the precision of post hoc bounds, and time complexity:



The default in IIDEA is set to $B = 1000$.

References

- [1] Goeman & Solari, *Stat. Sci.* 2011
- [2] Ebrahimipour & Goeman. *Brief. Bioinf.* 2021
- [3] Blanchard, Neuvial, Roquain. *Ann. Stat.* 2020
- [4] Enjalbert-Courrech & Neuvial. *Bioinformatics* (minor revision), 2022. doi:10.1101/2022.03.08.483449
- [5] Andreella et al. arXiv:2012.00368
- [6] Rosenblatt et al. *Neuroimage* 2018
- [7] Blain, Thirion, Neuvial. *Neuroimage* 2022
- [8] Davenport, Thirion, Neuvial. FDP control in multivariate linear models using the bootstrap. *In preparation*, 2022.

Acknowledgements

- ANR-16-CE40-0019 (SansSouci)
- Fondation Catalyses at Université Paul Sabatier
- Mission for Transversal and Interdisciplinary Initiatives (MITI) at CNRS (DDisc project).
- the SansSouci team!

Conclusion

Currently implemented in IIDEA:

- interactive volcano plot
- powerful **post hoc** guarantees
- **linear** computation of post hoc bounds
- **no programming skills** required
- **input data**: gene expression matrix or a matrix of p -values and fold changes (no adaptivity)

Perspective

Features yet to be implemented in IIDEA:

- RNA-seq data analysis (counting data)
- non-parametric calibration using **notip** [7]
- **multiple contrasts in multivariate linear models** using the bootstrap [8]
- parallelization of calibration
- use of promises for concurrent users