Neglecting normalization impact in semi-synthetic RNA-seq data simulations generates artificial false positives

Boris Hejblum

Inserm U1219 Bordeaux Population Health
Université de Bordeaux
Inria SISTM
Vaccine Research Institute

October 25th, 2022

RNA-seq differential expression

Objective:

Identify the genes whose expression is significantly associated with a factor or a group of factor in RNA-seq studies

Transcriptomics are **hypothesis generating** assays

- *Exploratory studies: (~20,000) univariate tests*
- influence downstream studies

⚠ False Positives (*science reproducibility crisis*)

RNA-seq data specificities:

⇒ Heteroscedasticity

⇒ High-dimensionality

State-of-the-art in bulk RNAseq differential analysis

Most popular methods:

`edgeR`: **Negative Binomial glm**

[Robinson *et al.*, *Genome Biology*, 2010]

(14,503 citations in PubMED)

`DESeq2`: **Negative Binomial glm**

[Love *et al.*, *Genome Biology*, 2014]

(22,016 citations in PubMED)

`limma-voom`: **weighted linear model**

Estimate mean-variance relationship
[Law *et al.*, *Genome Biology*, 2014]

(2,150 citations in PubMED)

Limitations:

- strong parametric assumptions
- tailored for small studies

Issues with Type-I error control

Rapaport *et al.* 2013
Reeb *et al.* 2013
Rocke et al. 2015
Germain *et al.* 2016
Yu *et al.* 2017
Assefa *et al.* 2018
Li *et al.* 2018

Introducing dearseq

0. r_{ij} aligned read counts from sample i for gene j

1. **log-counts per million** (normalize reading depth differences across samples i)

$$\boldsymbol{y}_{ij} = \log_2\left(10^6 \times \frac{0.5 + \boldsymbol{r}_{ij}}{1 + \mathscr{L}_i}\right) \quad \text{with } \mathscr{L}_i = \sum_{j=1}^{p} \boldsymbol{r}_{ij}$$

2. **Model global heteroscedasticity** (*not unlike voom*)

$$\text{Var}(\boldsymbol{y}_{ij}|X_i, \Phi_i) = \omega(E[\boldsymbol{y}_{ij}|X_i, \Phi_i]) + \boldsymbol{e}_{ij}$$

Local linear regression borrowing information across all genes $\Rightarrow \widehat{\omega}(\cdot)$

3. **Working model** for each gene j

$$\boldsymbol{y}_{ij} = X_i^T \boldsymbol{\alpha}_j + \Phi_i^T \boldsymbol{\beta}_j + \Phi_i^T \boldsymbol{\xi}_{ij} + \boldsymbol{\varepsilon}_{ij}$$

$$\boldsymbol{\xi}_j \sim N(0, \Sigma_{\boldsymbol{\xi}_j}), \quad \boldsymbol{\varepsilon}_j \sim N(0, \Sigma_{\boldsymbol{\varepsilon}_j})$$

Variance component score test

H_0: expression of gene j does not depend on covariates Φ

$\Rightarrow$ **Variance component score test** for $\left(\boldsymbol{\beta}_j + \boldsymbol{\xi}_{ij}\right) = \boldsymbol{\eta}_j \boldsymbol{\nu}_{ij}$

$$H_0: \ \boldsymbol{\eta}_j = 0$$

Test statistic: $Q_j = \boldsymbol{q}_j^T \boldsymbol{q}_j$ with $\boldsymbol{q}_j = n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{y}_{\mu_{ij}} \Sigma_{\boldsymbol{\varepsilon}_j}^{-1} \Phi_i$ $\Rightarrow$ $Q_j \underset{H_0}{\sim} \sum_{\ell} a_\ell \chi_1^2$

where $\boldsymbol{y}_{\mu_{ij}} = \boldsymbol{y}_{ij} - X_i^T \boldsymbol{\alpha}_j$

$\Rightarrow$ **robust to misspecification** (CLT)

$\Rightarrow$ **powerful**

[Commenges & Andersen, *Lifetime Data Analysis*, 1995]

[Lin, *Biometrika*, 1997]

[Wu *et al.*, *American Journal of Human Genetics*, 2011]

Numerical estimation

Score test $\Rightarrow$ model estimated only under H_0

- $\widehat{\alpha_j}$: OLS
- $\widehat{a_{\ell j}}$: eigen values of $cov(q_j)$
- p-value computed using [Kuonen D (1999) *Biometrika* 1999] saddlepoint approximation for distributions of quadratic forms implemented in `survey`
- permutation exact test: alternative for small sample sizes (<40)

Can you trust RNA-seq differential expression methods?

Difficulties to build convincing simulations:

- *parametric simulations*
  - 😁 easy, flexible, truth is certain
  - ☹ unrealistic, unfair
- *resampling*
  - 😁 realistic
  - ☹ not flexible (1 or 2 populations), requires large dataset
- *biological experiments*
  - 😁 ideallly realistic
  - ☹ expensive, truth is not fully known

[Adapted from Assefa]

Simulation study

- **Parametric simulations**
  1. Negative Binomial
  2. Nonlinear

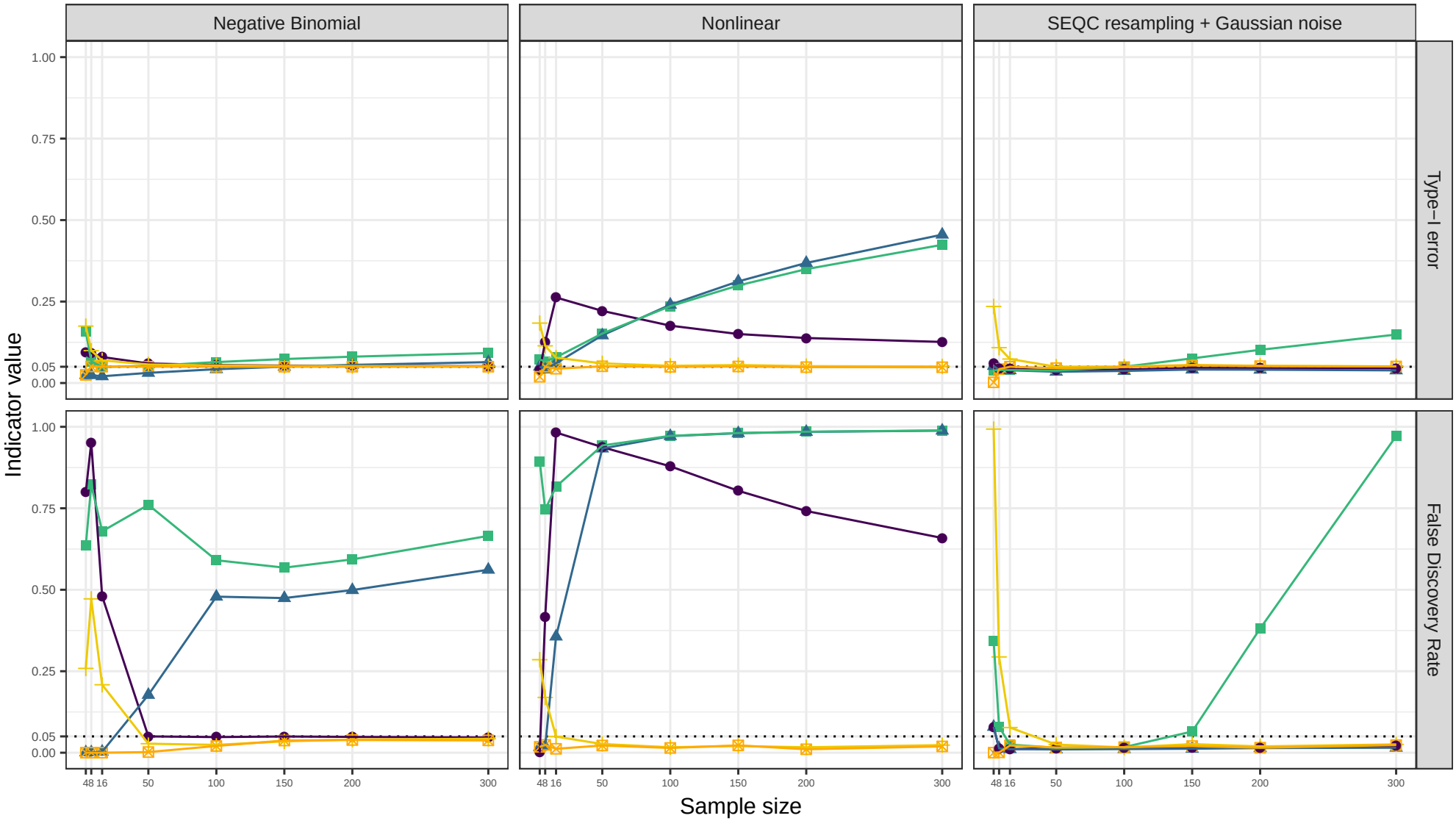
- **Resampling + parametric perturbation**
  3. SEQC data: 5 homogeneous RNA-seq samples

     $\Rightarrow$ *Gaussian noise with empirical covariance*

     [SEQC/MAQC-III Consortium. *Nature Biotechnology*, 2014]

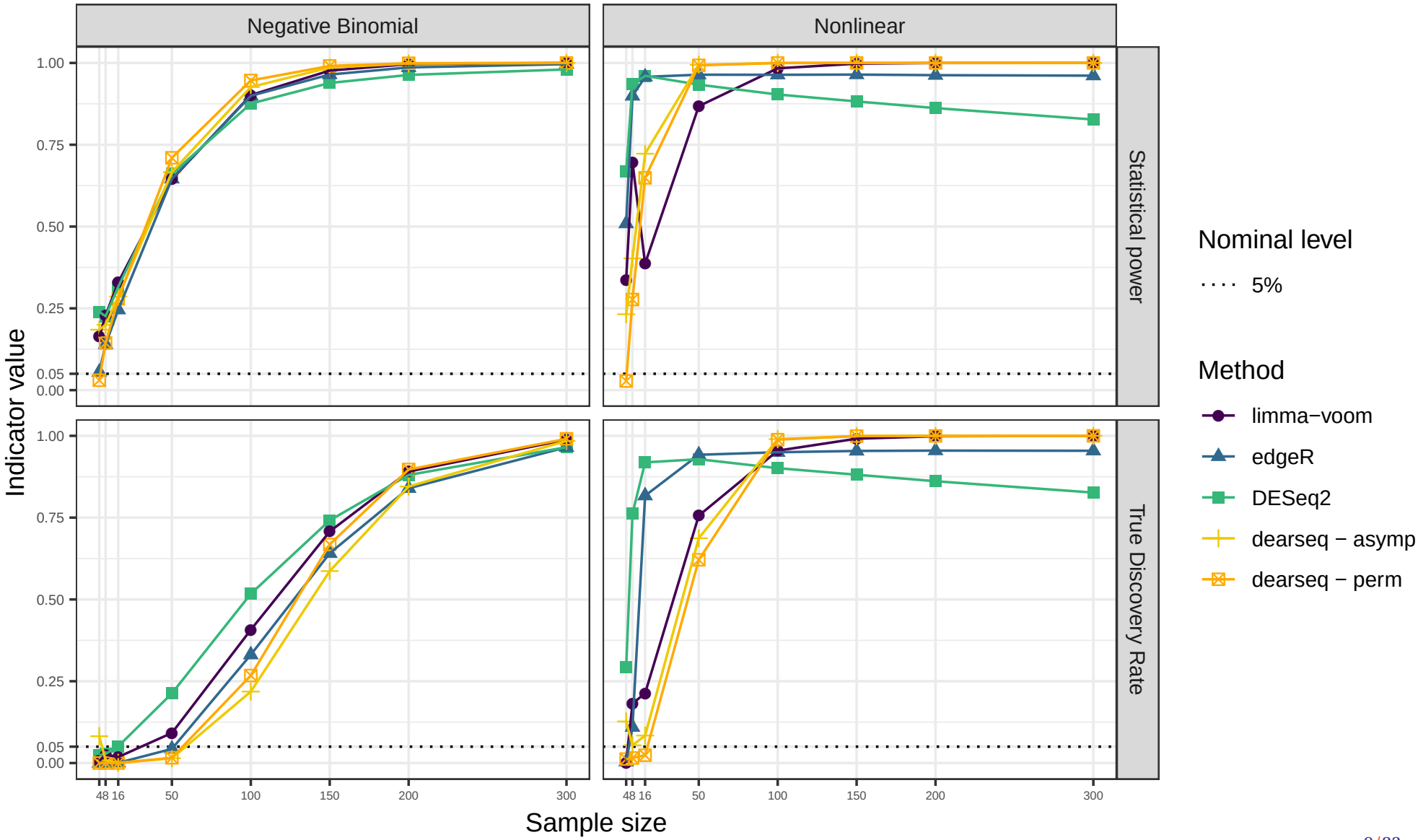
Monte−Carlo estimation over 1,000 simulation runs

(nominal testing level at 5%)



Monte−Carlo estimation over 1,000 simulation runs

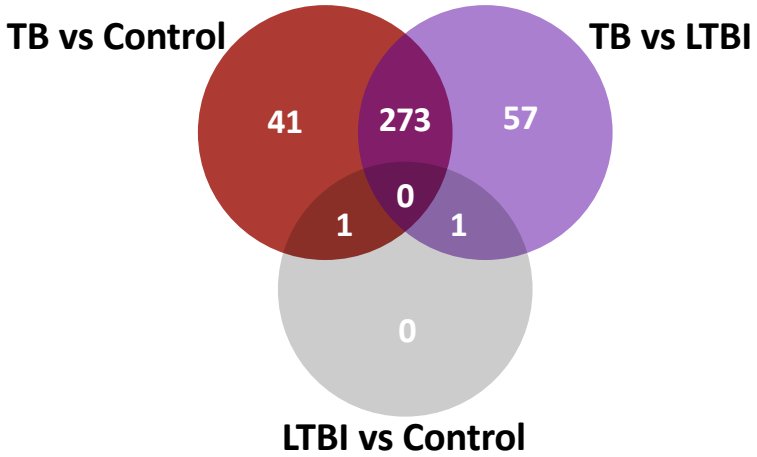
(nominal testing level at 5%)



Re-analysis of a tuberculosis infection study

▷ Singhania *et al.*, A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection, *Nature Communications*, 2018.

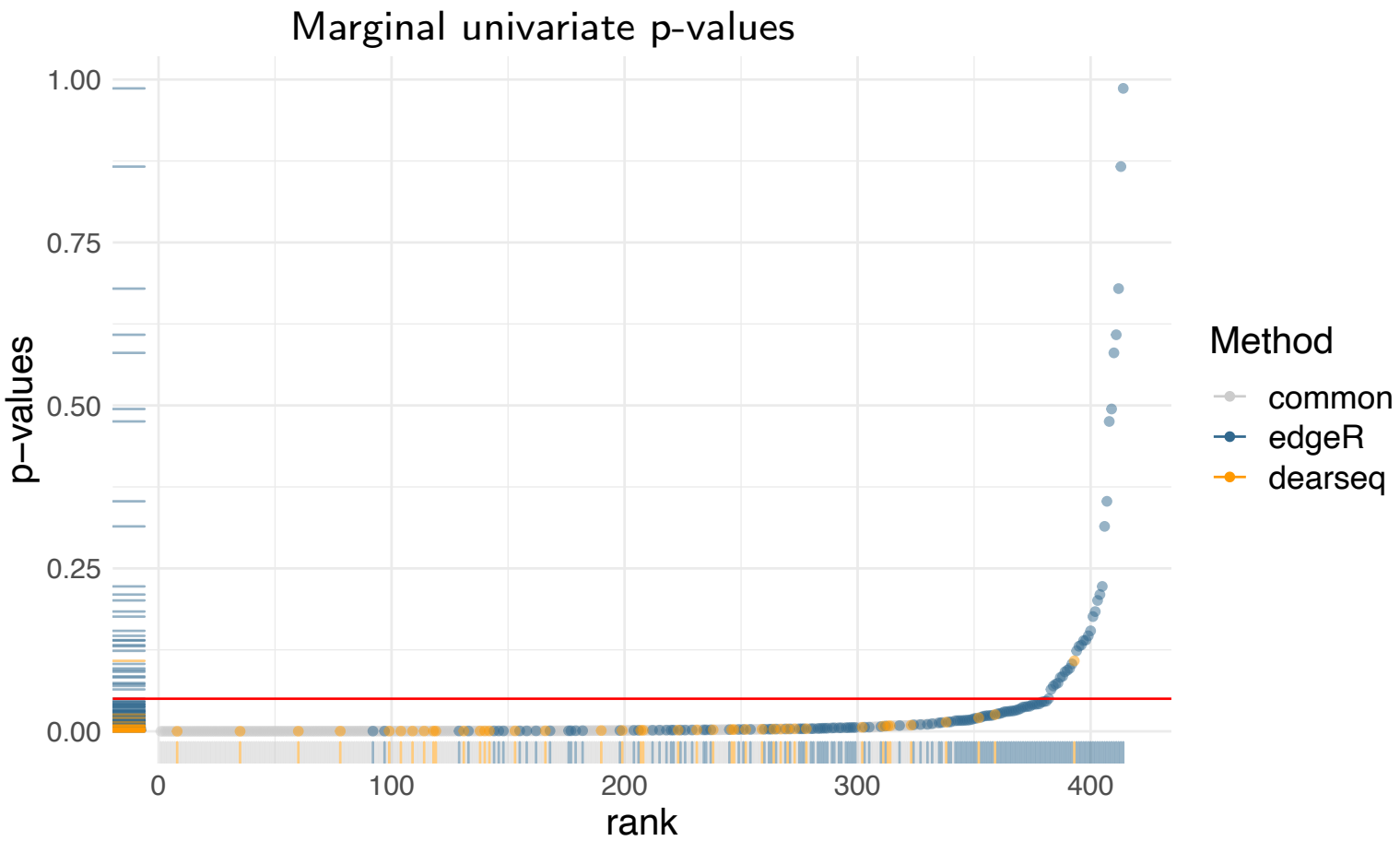
- 14,150 genes in 54 samples:
  - 21 ActiveTB
  - 21 LTBI
  - 12 Control
- edgeR ⇒ **373-gene** signature



Comparing dearseq to original edgeR results

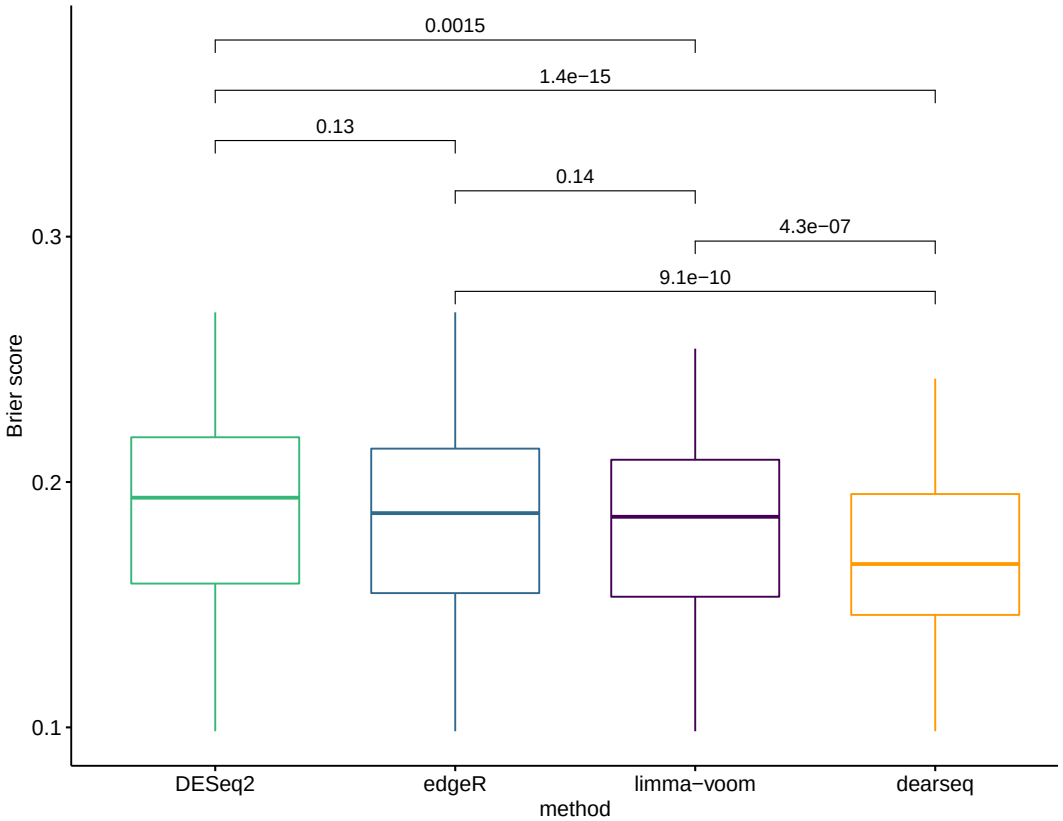
dearseq ⇒ **273-gene** signature

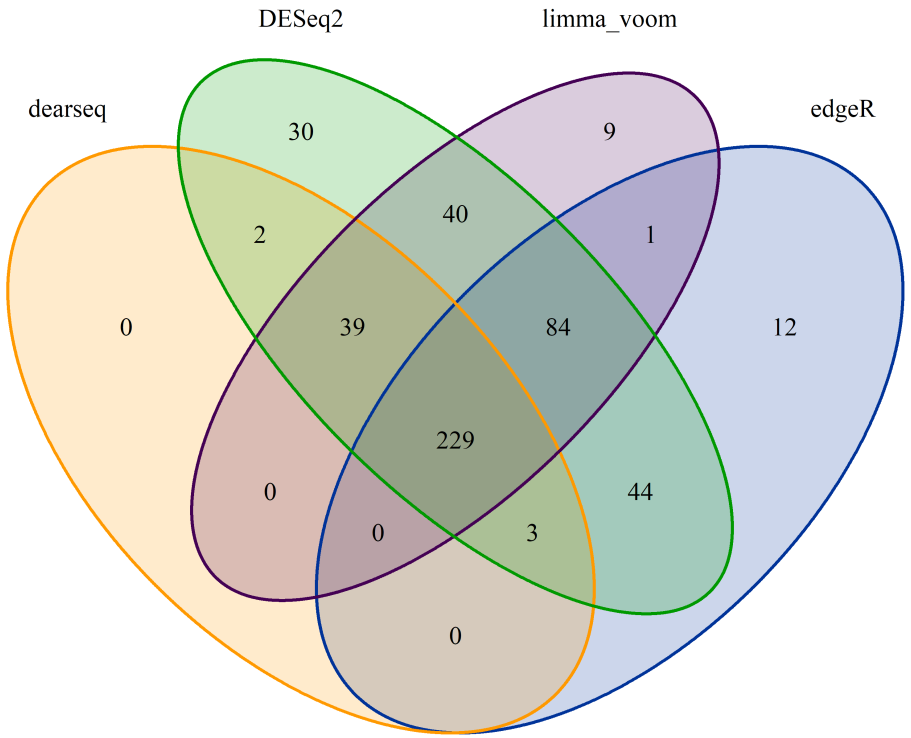
100 False Positives ? ⇒ univariate logistic regression



dearseq comparison with the state-of-the-art

Cross -validated Brier score for each gene (lower = better)





So, which method for differential analysis of RNA-seq data ?

Take home message

State-of-the-art methods may include **false positives** in the results, especially for **larger studies**

- `DESeq2` under-estimates low p-values ⇒ BH **X**
- `edgeR` also have Type-I error issues
  userguide only mention it once:
  "[latest test has] more rigorous control of Type-I error rate"
- `voom/limma` controls the type-I error and FDR **adequately**
  as long as **linear model is true**
- `dearseq` improves Type-I error & FDR control
  without losing statistical power

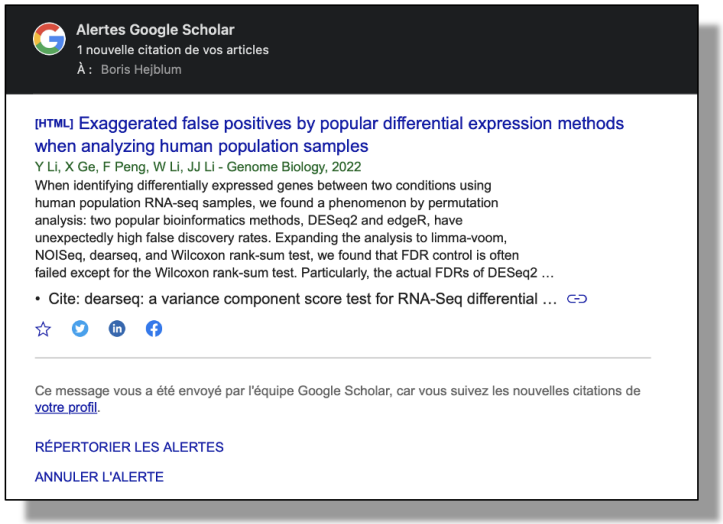
dearseq summary

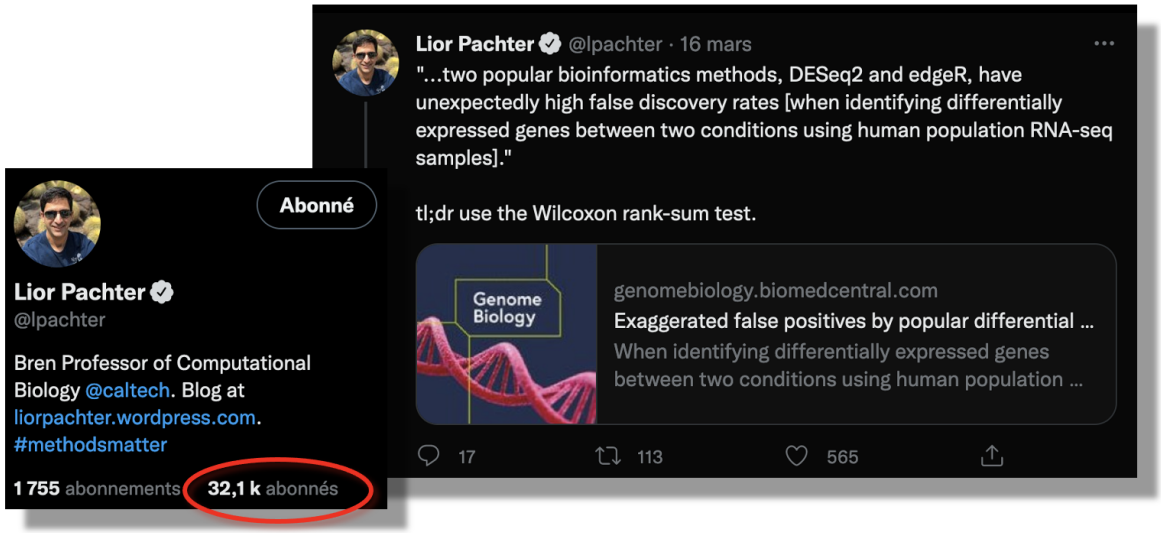
- **Robust** and **powerful** variance component score test accounting for the **mean-variance** relationship in RNA-seq data
- **Complex designs**
  ⇒ longitudinal data . . .
- **Multiple outcomes**
  ⇒ immune response, complex phenotypes . . .
- **R package** `dearseq` available on Bioconductor

- Versatile **DEA** ⇒ *NAR G & B* `DOI: 10.1093/nargab/lqaa093`
  [Gauthier, Agniel, Thiébaut & Hejblum, dearseq: a variance component score test for RNA-seq differential analysis that effectively controls the false discovery rate, *NAR Genomics & Bioinformatics* 2020]
- **Gene Set Analysis** ⇒ *Biostatistics* `DOI: 10.1093/biostatistics/kxx005`
  [Agniel & Hejblum, Variance component score test for time-course gene set analysis of longitudinal RNA-seq data, *Biostatistics*, 2017]

Serendipity, Literature watch & Twitter

One quiet Friday morning last March...





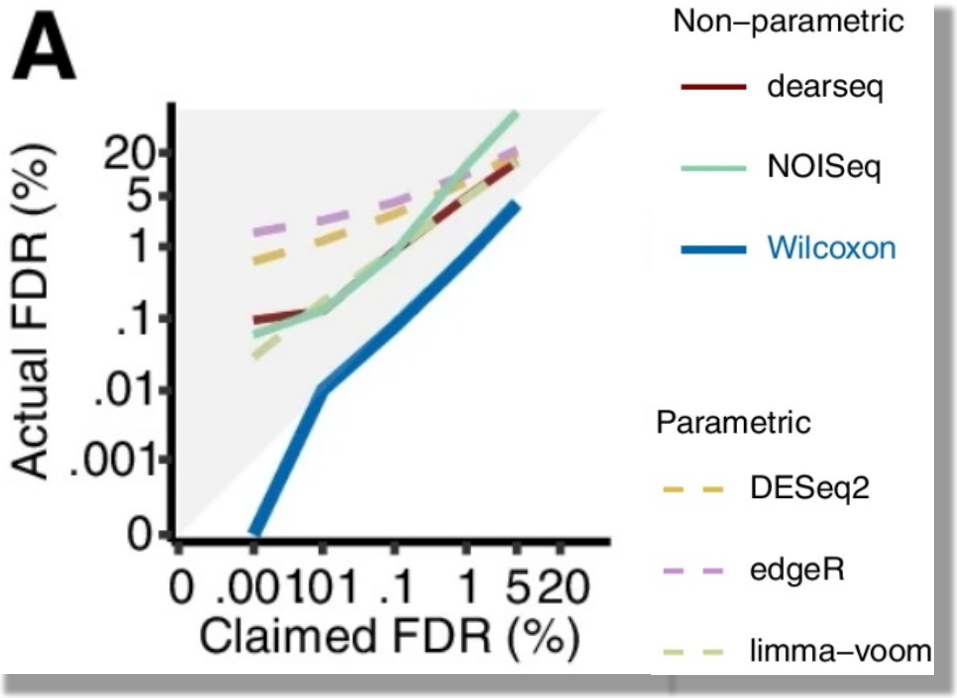
Too Long ; Didn't Read

Genome Biology

Short Report | Open Access | Published: 15 March 2022

Exaggerated false positives by popular differential expression methods when analyzing human population samples

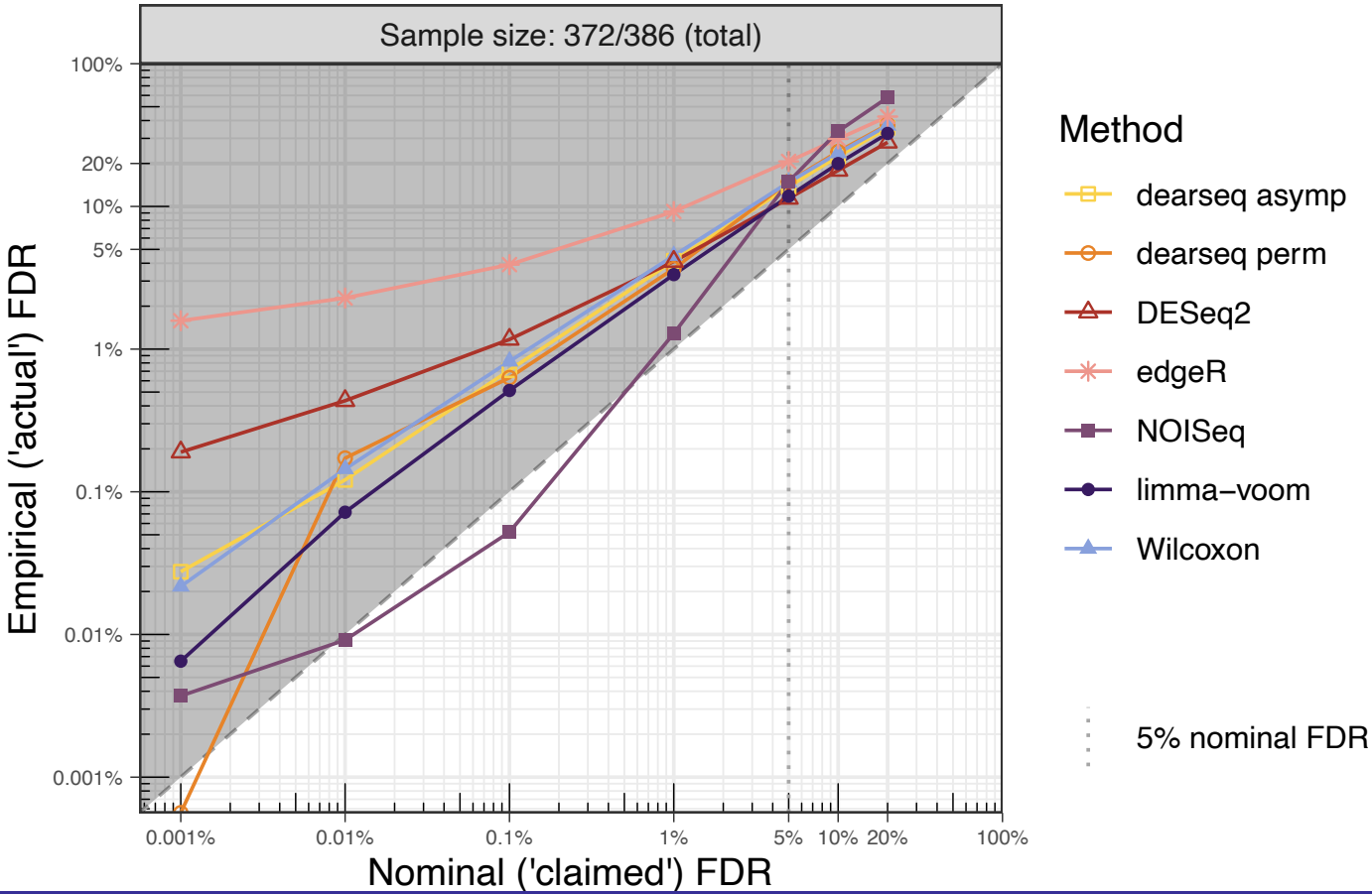
Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li & Jingyi Jessica Li



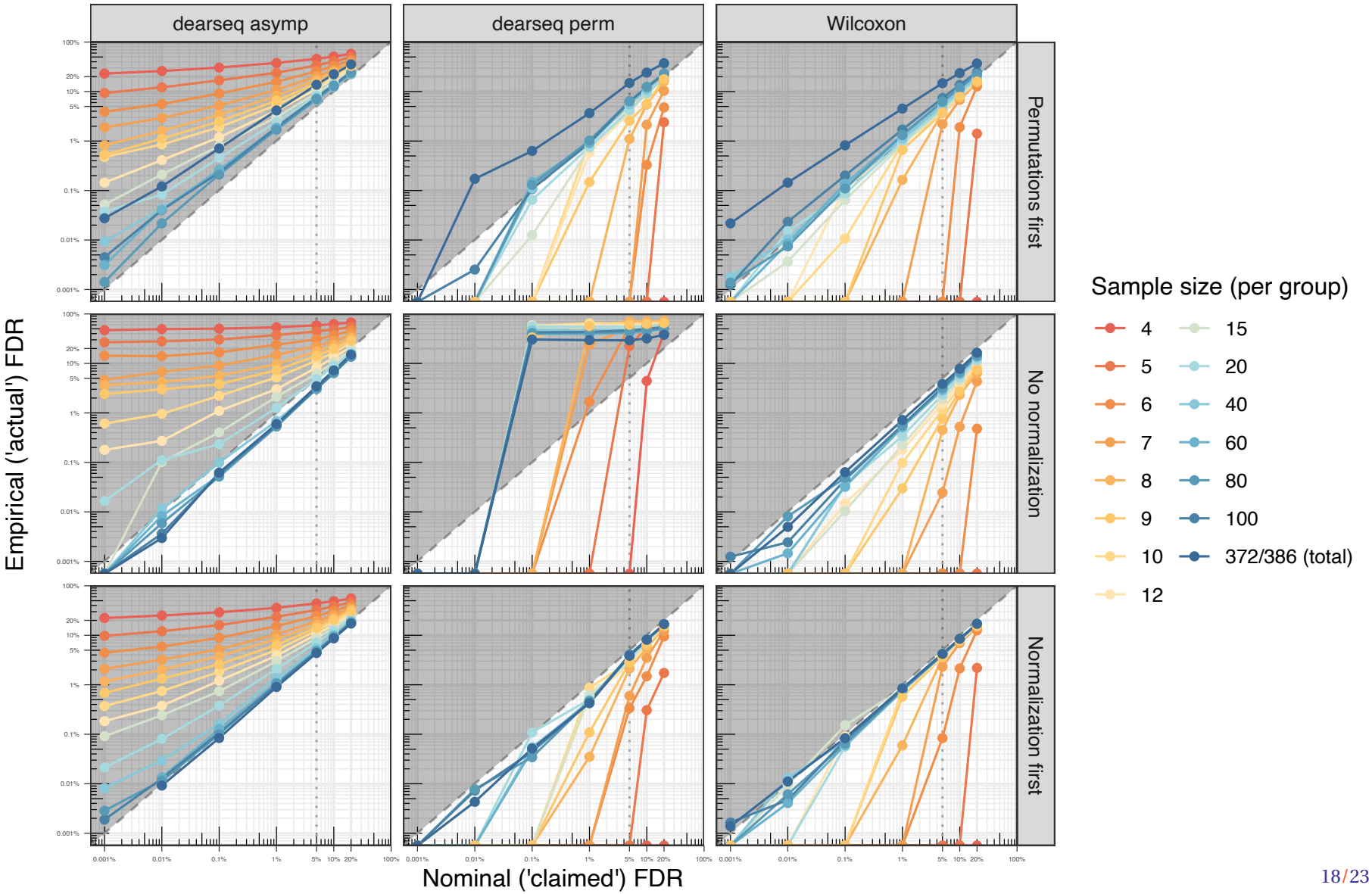
- Discrepancy with our simulation results
- Explained with parametric assumptions
  => But irrelevant for *dearseq* !
- + Ignore our permutation method
  while advocating for permutations to solve this issue

Benefits of open reproducible research

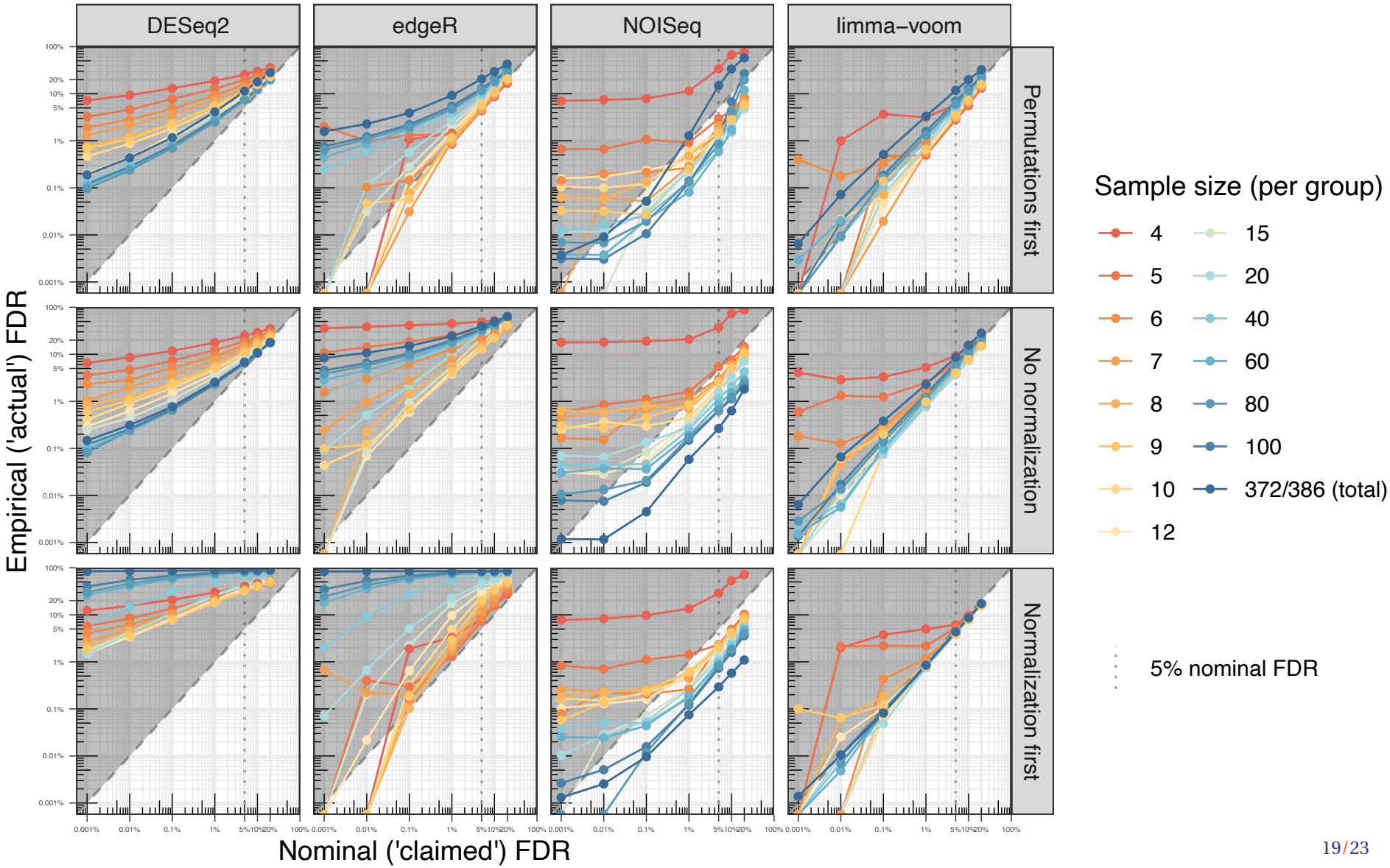
1. Able to redproduce their Fig 2
   *(after 2 days of searching and examining their GitHub **AND** Zenodo repositories)*
2. ⚠ They use **DIFFERENT** data for i) Wilcoxon and for ii) the other compared state-of-the-art methods !!



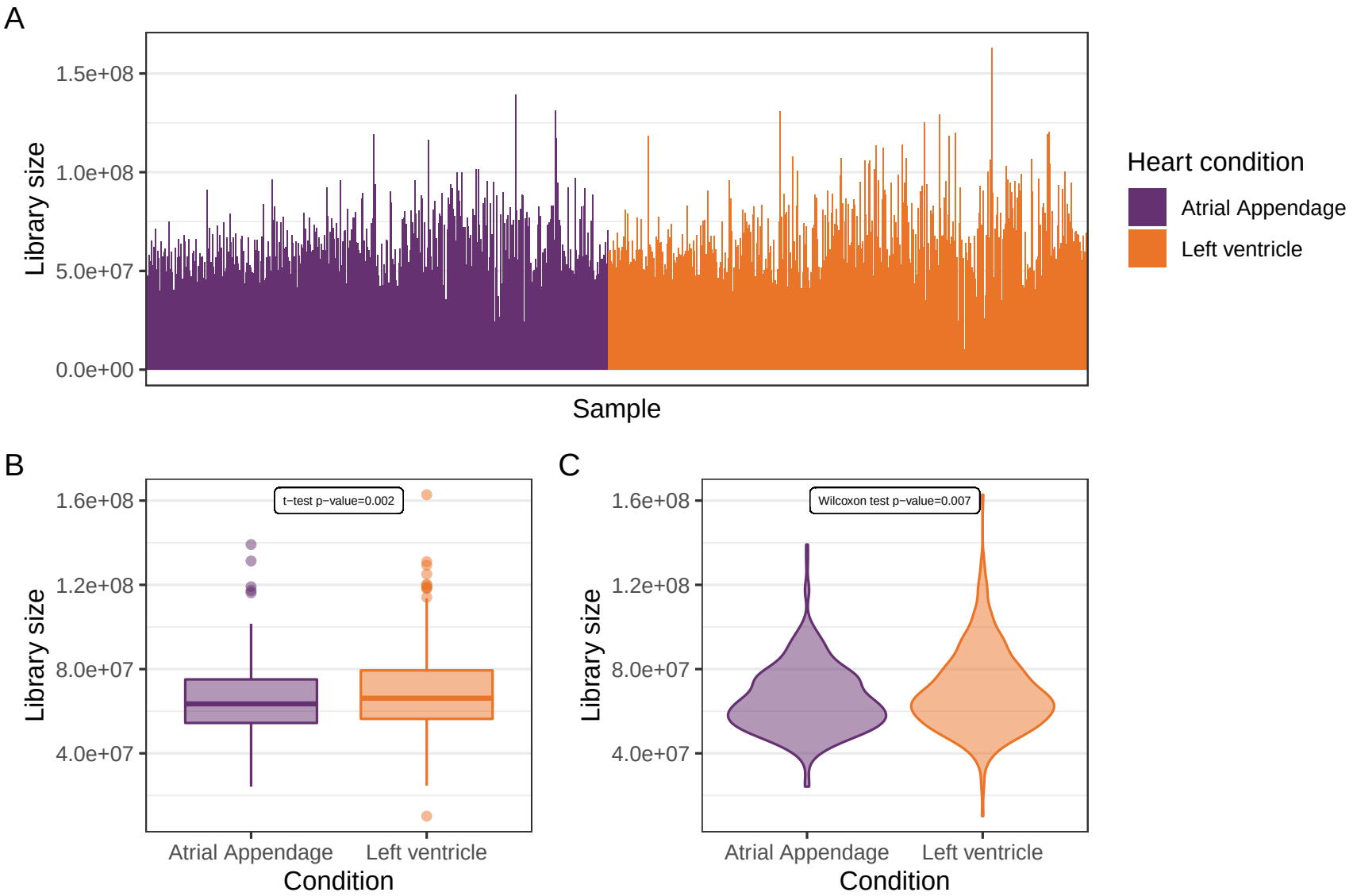
A fairer comparison: **same data for ALL methods** (I)



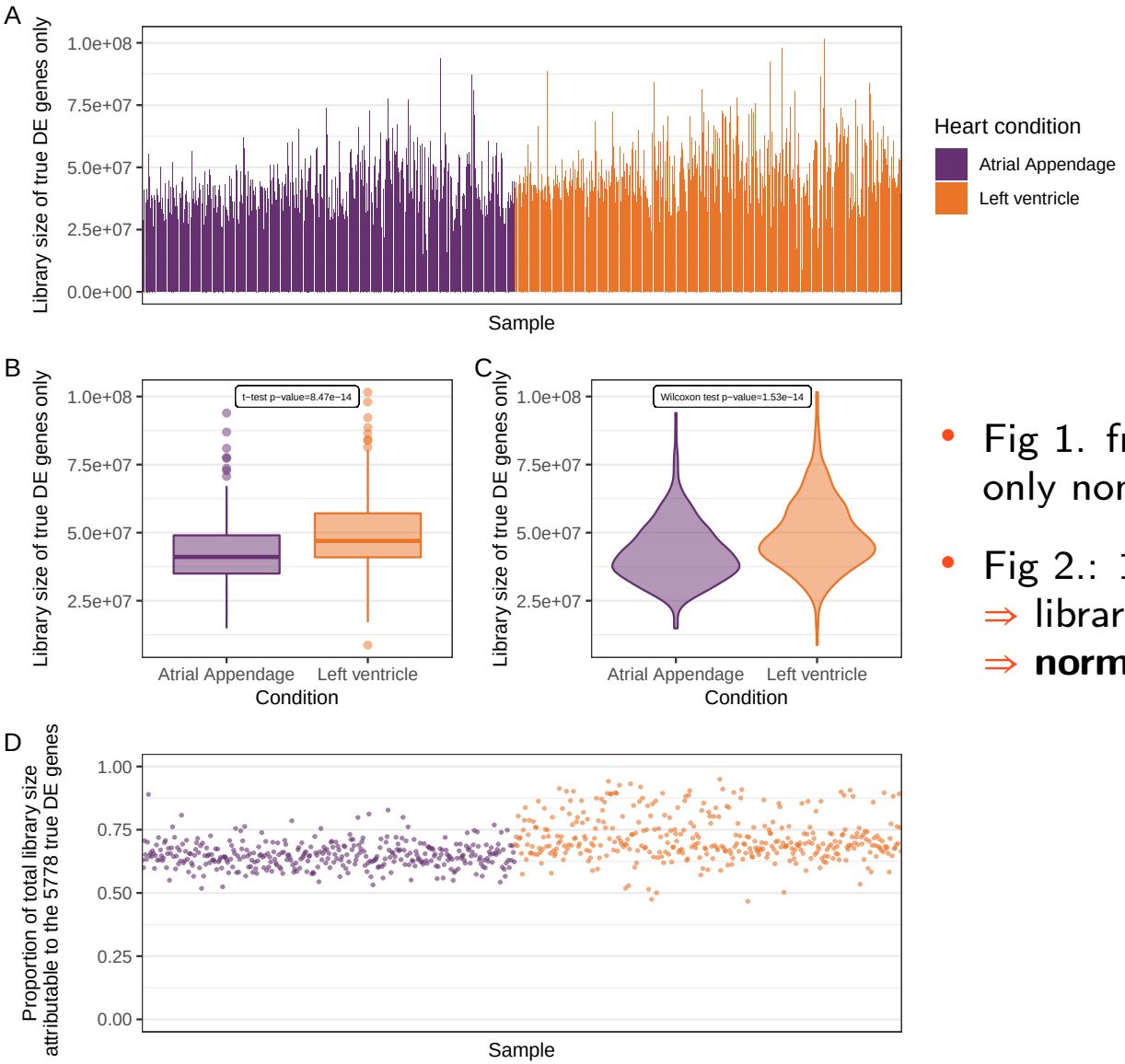
A fairer comparison: **same data for ALL methods** (I)



Confounding of library size differences with groups



Confounding of library size differences with groups



- Fig 1. from Li *et al.* only non-DE genes
- Fig 2.: 10% most DE genes kept unpermuted
  ⇒ library size confounding persists
  ⇒ **normalization induces artificial FP !**

We agree to disagree

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

Contradictory Results

Neglecting normalization impact in semi-synthetic RNA-seq data simulation generates artificial false positives

Boris P Hejblum, Kalidou Ba, Rodolphe Thiébaut, Denis Agniel

doi: https://doi.org/10.1101/2022.05.10.490529

Confirmatory Results

Wilcoxon rank-sum test still outperforms dearseq after accounting for the normalization impact in semi-synthetic RNA-seq data simulation

Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li, Jingyi Jessica Li

doi: https://doi.org/10.1101/2022.06.07.494963

Submitted as a comment in *Genome Biology* in May
⇒ "under review" since . . .

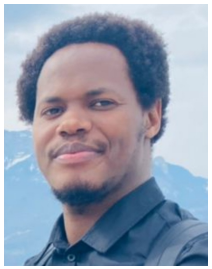
Discussion

- Bottom line: **simulations are hard** (and unfair)
- Disregarding standard practices in **data processing can bite you**
- **Wilcoxon test**: great for large 2 group comparisons
  False short for more complex designs
- `dearseq` **does control Type-I error !**

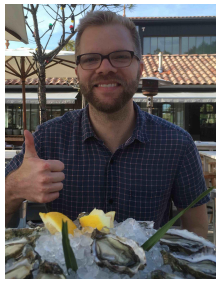
Thank you for your attention ! – **Questions ?**

Marine Gauthier

Kalidou Ba



Denis Agniel



Rodolphe Thiébaut

🚨🚨 **Internship, PhD & postdoc openings**

👉 *https://borishejblum.science*

boris.hejblum@u-bordeaux.fr