Post hoc inference via multiple testing

Pierre Neuvial

Institut de Mathématiques de Toulouse

Joint work with Gilles Blanchard and Etienne Roquain
Arxiv preprint: https://arxiv.org/abs/1703.02307

SPSML, Potsdam 2018-02-14

Outline

1. Introduction
   - Differential expression studies in cancer research
   - Post hoc inference
2. Post hoc bounds from JER control
   - JER control: definition and associated bounds
   - JER control based on Simes' inequality
   - Limitations of Simes-based JER control
3. Adaptive JER control
   - Calibration of a rejection template
   - Numerical experiments for Gaussian equi-correlation
   - Application: Leukemia data set

Example: Leukemia data set

- Expression measurements (mRNA) of $m = 12625$ genes in $n = 79$ cancer patients:
- Two groups of patients:
  - BCR/ABL: 37 patients
  - NEG: 42 patients

Question: find genes whose average expression differs between the two groups

Large-scale inference

- Setup: one statistical test for each gene g
  - e.g. Student's t test of $H_{0,g}$: no difference between group means
- Goal: select a subset S of genes with a "small" number $V(S)$ of false positives (genes in S but for which $H_{0,g}$ is true)

Step 1 (user): choose a (multiple testing) risk of interest

1. $\mathbb{P}(V(S) > 0)$: Family-Wise Error Rate
2. $\mathbb{E}(V(S)/(|S| \vee 1))$: False Discovery Rate

and an acceptable target level for this risk: α

Step 2 (statistician): select S satisfying the desired guarantee

1. Bonferroni, Bonferroni-Holm, Hommel, ...
2. Benjamini-Hochberg, Storey, ...

Example: FWER and FDR thresholding

State of the art answer

With $\alpha = 0.05$,

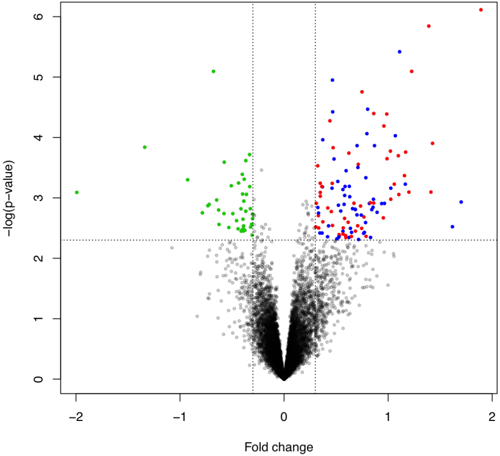
1. FWER control: $|S_1| = 20$: 1635_at, 1636_g_at, 1674_at...
   41815_at
2. FDR control: $|S_2| = 163$: 1000_at, 1001_at, 1002_f_at...
   1148_s_at

Post hoc questions

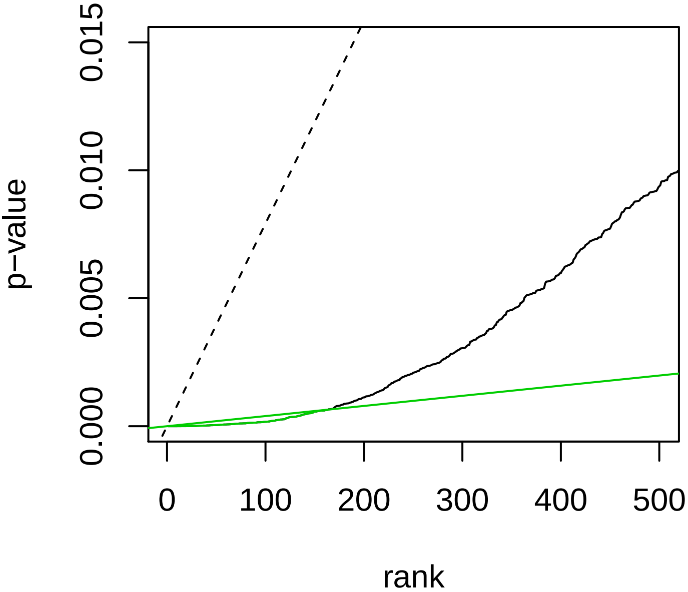
can we *incorporate prior biological knowledge*: fold change, gene pathways

- inference on $S = S_1 \cup S_1'$?
- inference on $S = S_2 \setminus S_2'$

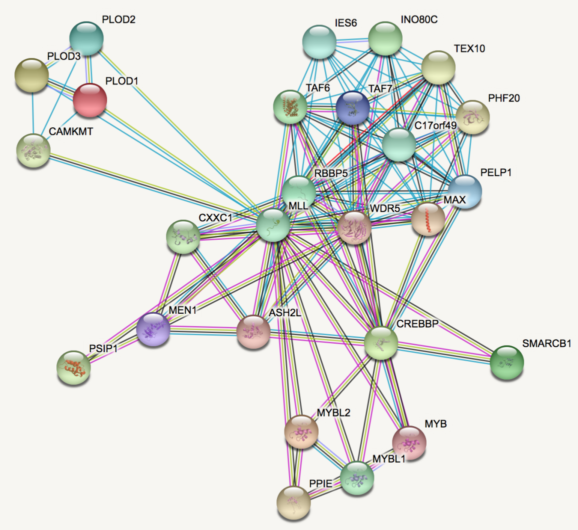
User-defined selection 1: volcano plot



User-defined selection 2: top k genes

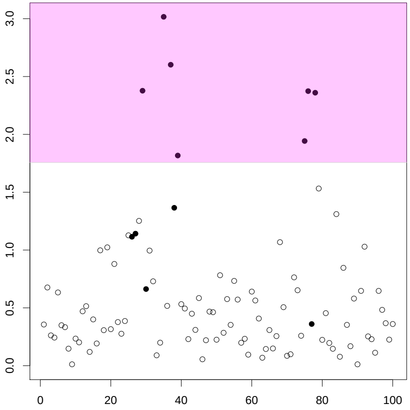


User-defined selection 3: gene pathways



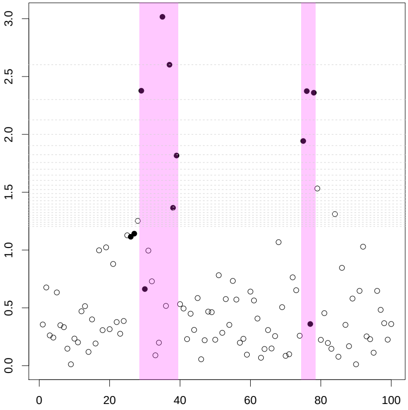
User-defined selection: toy example

Classical multiple testing



FDR $\leq 25\%$

Post hoc inference



With probability $\geq 75\%$

$|S \cap \mathcal{H}_1| \geq 2$ and $|S' \cap \mathcal{H}_1| \geq 1$

How can JER control be achieved?

The need for post hoc inference

Challenges

- large-scale multiple testing is exploratory in nature
- no formal statistical guarantee on such user-defined selections

Proposal: post hoc confidence bounds

- $\mathcal{H} = \{1, \dots m\}$: m null hypotheses to be tested
- $\mathcal{H}_0 \subset \mathcal{H}$: true null hypotheses, $m_0 = |\mathcal{H}_0|$
- $\mathcal{H}_1 = \mathcal{H} \setminus \mathcal{H}_0$
- $V(S) = |S \cap \mathcal{H}_0|$: number of false postives in $S \subset \mathcal{H}$

Goal: find $\overline{V}_\alpha$ such that

$$\mathbb{P}\left(\forall S \subset \{1 \dots m\},\ V(S) \leq \overline{V}_\alpha(S)\right) \geq 1 - \alpha$$

Related works: selective inference

for a specific selection rule

Inference for a specific selection rule S

- Lockhart et al. (2014), Fithian et al. (2014)

for an arbitrary, pre-decided selection rule

Inference for an arbitrary selection rule, to be chosen before looking at the data

- Benjamini and Yekutieli (2005)

Omnibus

Inference simultaneously over all $S \subset \{1, \dots, m\}$, possibly chosen after looging at the data

- Genovese and Wasserman (2006), Goeman and Solari (2011), Berk et al. (2013)

Basic idea: reference family

Reference set

Assume that we *know* an upper bound for $V(R) := |R \cap \mathcal{H}_0|$ for some $R \subset \mathcal{H}$

Then for any $S \subset \mathcal{H}$, we have $V(S) \leq |S \cap R^c| + V(R)$

Proof: simply note that $V(S) = |S \cap \mathcal{H}_0| = |S \cap \mathcal{H}_0 \cap R^c| + |S \cap \mathcal{H}_0 \cap R|$

Reference family

Idea: build a family of sets $(R_1, \ldots, R_K)$ for which we have an upper bound on $V(R_k)$ for each k.

Post hoc bound via JER control

Definition (Joint Family-Wise Error Rate control)

Let $\mathfrak{R} = (R_k)_k$ be a *reference family* of subsets of $\mathcal{H}$.

$$\mathrm{JER}(\mathfrak{R}) := \mathbb{P}(\exists k, V(R_k) \geq k) \leq \alpha$$

That is, $\mathcal{E} = \{\forall k : V(R_k) \leq k - 1\}$ is of probability $\geq 1 - \alpha$

Proposition: post hoc upper bound on the number of false positives

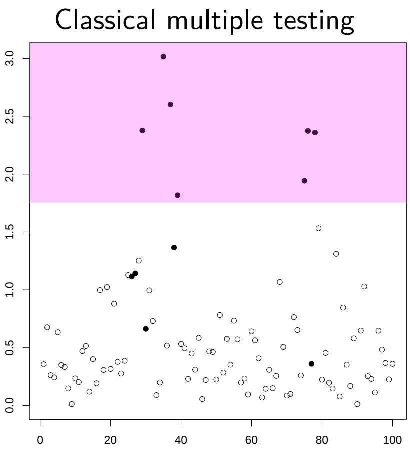
On the event $\mathcal{E}$, for any set $S \subset \{1, \dots m\}$,

$$V(S) \leq |S| \wedge \min_k \{|S \cap R_k^c| + k - 1\}$$

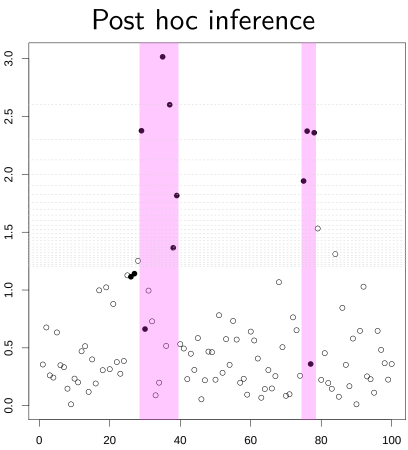
Recall: $V(S) \leq |S \cap R^c| + V(R)$

Applicable to any number of possibly data-driven sets!

Post hoc inference: toy example



FDR $\leq 25\%$



With probability $\geq 75\%$
$|S \cap \mathcal{H}_1| \geq 2$ and $|S' \cap \mathcal{H}_1| \geq 1$

How can JER control be achieved?

Simes-based[1] JER control and post hoc bound

Simes' inequality

- If the p-values (p_i), $1 \leq i \leq m$, are independent then

$$\mathbb{P}(\exists k \in \{1, \ldots, m_0\} \,:\, p_{(k:\mathcal{H}_0)} \leq \alpha k/m_0) = \alpha$$

- Under some forms of positive dependence (PRDS($\mathcal{H}_0$)): $\leq \alpha$

(*PRDS = Positive Regression Dependency on a Subset*)

Corollary: Simes-based JER control and post hoc bound

Under PRDS, the Simes reference family $(R_k)_k$, with

$$R_k = \{1 \leq i \leq m \,:\, p_i \leq \alpha k/m\}$$

achieves JER control at level α and thus provides a post hoc bound

[1] R. J. Simes. *Biometrika* 73.3 (1986), pp. 751–754.

Simes-based JER control and post hoc bound

Post hoc bound for the Simes family

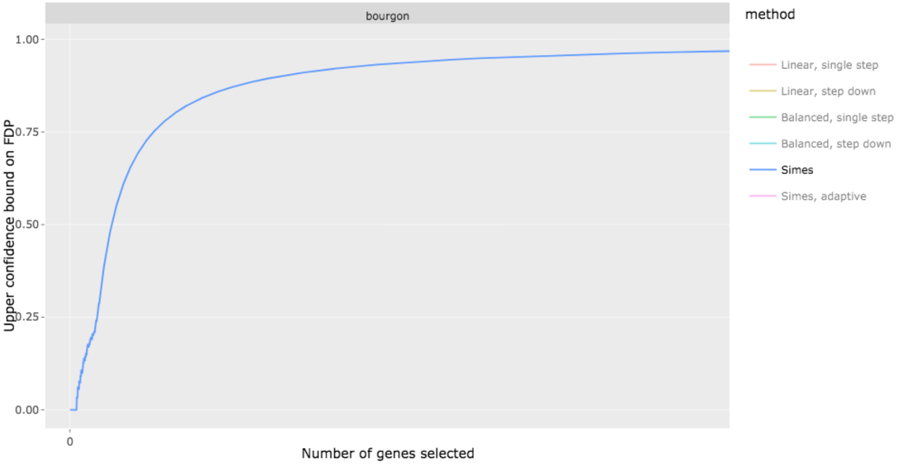
Under PRDS, with probability larger than $1-\alpha$, for any S,

$$V(S) \leq |S| \wedge \min_k \left\{ \sum_{i \in S} \mathbf{1}\left\{ p_i > \alpha k/m \right\} + k - 1 \right\}.$$

Comments

- Recovers the closed testing bound of Goeman and Solari (2011)
- JER: a generic device to build post hoc bounds
- Independence/PRDS assumption:
    - can we obtain dependence-free JER control?
    - how sharp is the Simes inequality under PRDS?

Application: Leukemia data set



Dependence-free JER control?

Under arbitrary dependence, with probability larger than $1-\alpha$, for any S,

$$V(S) \leq |S| \wedge \min_k \left\{ \sum_{i \in S} \mathbf{1}\left\{p_i > \alpha / C_m k/m\right\} + k - 1 \right\},$$

$C_m = \sum_{k=1}^m k^{-1} \sim \log(m)$: Hommel's correction factor for dependency[2]

Dependence-free adjustment is not a sensible objective

- implies adjusting to a worst case dependency
- very conservative (cf Benjamini-Yekutieli for FDR control)

We want to be adaptive to dependency

[2] G Hommel. "Tests of the overall hypothesis for arbitrary dependence structures". Biometrische Zeitschrift 25.5 (1983), pp. 423–430.

Sharpness and conservativeness of the Simes family

Simes' equality is sharp under independence, but conservative under positive dependence.

Conservativeness of JER control under PRDS

Example: Gaussian equi-correlation, white setting ($m_0 = m = 1,000$):
Test statistics $\sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$ for $i \neq j$.

Equi-correlation level: ρ	0	0.1	0.2	0.4	0.8
Achieved JER $\times\alpha^{-1}$	0.99	0.85	0.72	0.42	0.39

Can we build a family achieving sharper JER control?

We want to be adaptive to dependency

JER control with λ-calibration

Rejection template

Consider a reference family $\mathfrak{R}_\alpha = (R_k(\alpha))_k$, where

$$R_k(\alpha) = \{1 \leq i \leq m \ : \ p_i \leq t_k(\alpha)\}$$

where $t_k(0) = 0$ and $t_k(\cdot)$ is non-decreasing and left-continuous on $[0, 1]$

- Example (Simes family): $t_k(\alpha) = \alpha k/m$

Associated **rejection template**: collection $(t_k(\lambda))_k$ for all $0 \leq \lambda \leq 1$

Single-step λ-calibration

$$\lambda(\alpha) = \max\left\{\lambda \geq 0 \ : \ \mathbb{P}\left(\min_k \left\{t_k^{-1}\left(p_{(k:\mathcal{H}_0)}\right)\right\} \leq \lambda\right) \leq \alpha\right\}$$

The family $\mathfrak{R}_{\lambda(\alpha)}$ controls JER at level α.

Example: Gaussian location model

Setting: $X \sim \mathcal{N}(\mu, \Sigma)$, $p_i = 2\overline{\Phi}(|X_i|)$

$$\lambda(\alpha) = \max\left\{\lambda \geq 0 \,:\, \mathbb{P}_{Z\sim\mathcal{N}(0,\Sigma)}\left(\min_k \left\{t_k^{-1}\left(2\overline{\Phi}(|Z_{(k)}|)\right)\right\} \leq \lambda\right) \leq \alpha\right\}$$

yields $\mathrm{JER}(\mathfrak{R}_{\lambda(\alpha)}) \leq \alpha$

Choice of the template

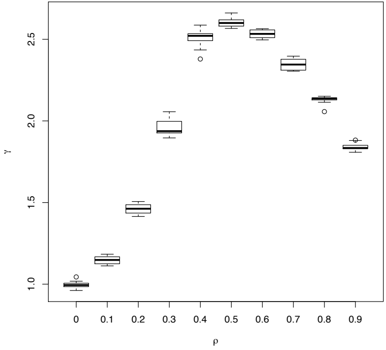
- Linear template: $t_k(\lambda) = \lambda k/m$ (generalizes Simes)
- Balanced template: $t_k(\lambda)$ such that $t_k^{-1}(2\overline{\Phi}(|X_{(k)}|)) \sim \mathcal{U}[0,1]$

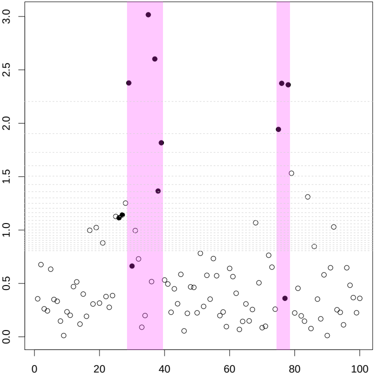
λ-calibration

- If Σ is known, $\lambda(\alpha)$ can be calibrated by Monte-Carlo
- If Σ is unknown, $\lambda(\alpha)$ can be calibrated by sign-flipping

JER control with λ-calibration for the linear template

Example under positive dependency (Gaussian equi-correlation)



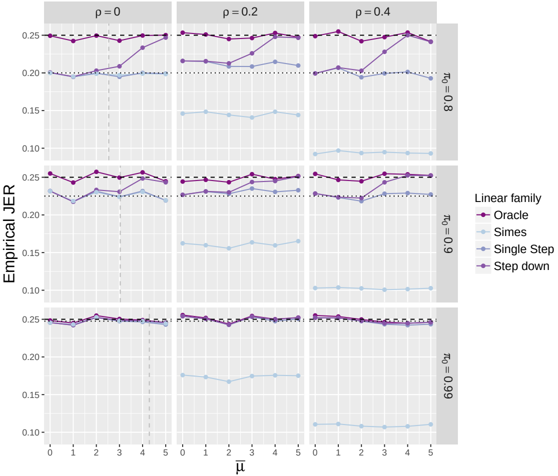


With probability $\geq 1 - \alpha = 75\%$:

$t_k(\alpha)$	Lower bound on $\lvert S \cap \mathcal{H}_1 \rvert$
$\alpha k/m$	$\lvert S \cap \mathcal{H}_1 \rvert \geq 2$ and $\lvert S' \cap \mathcal{H}_1 \rvert \geq 1$
$\lambda(\alpha) k/m$	$\lvert S \cap \mathcal{H}_1 \rvert \geq 3$ and $\lvert S' \cap \mathcal{H}_1 \rvert \geq 2$

JER control under Gaussian equi-correlation

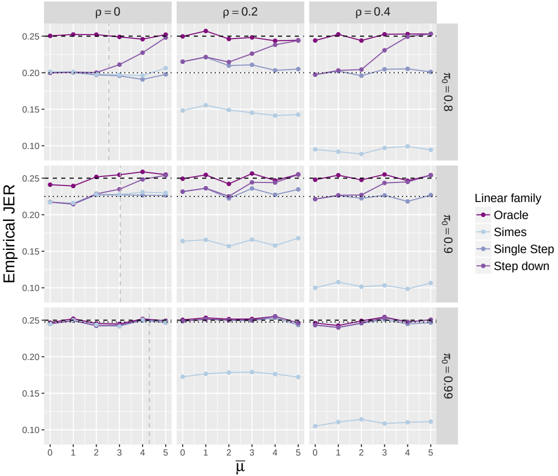
Linear template, known dependence (calibration by Monte-Carlo)



- $X_i \sim \mathcal{N}(0,1)$ under H_0
- $X_i \sim \mathcal{N}(\bar{\mu},1)$ under H_1
- $\mathrm{cor}(X_i, X_j) = \rho$ for $i \neq j$
- $\alpha = 0.25$

JER control under Gaussian equi-correlation

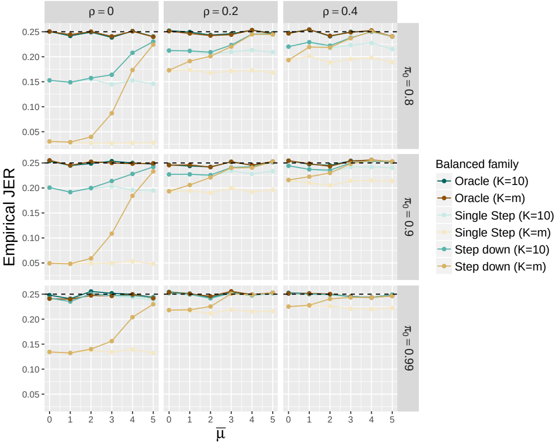
Linear template, unknown dependence (calibration by sign-flipping)



- $X_i \sim \mathcal{N}(0, 1)$ under H_0
- $X_i \sim \mathcal{N}(\bar{\mu}, 1)$ under H_1
- $\mathrm{cor}(X_i, X_j) = \rho$ for $i \neq j$
- $\alpha = 0.25$

JER control under Gaussian equi-correlation

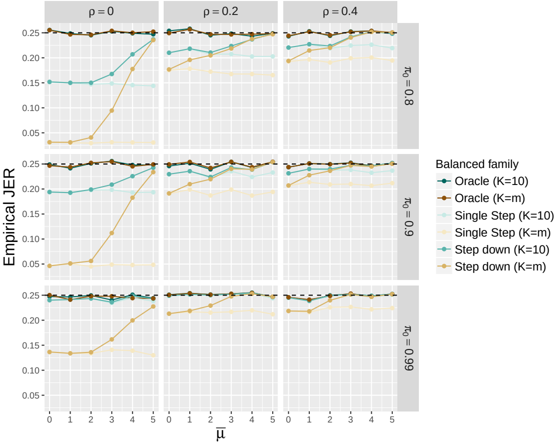
Balanced template, known dependence (calibration by Monte-Carlo)



- $X_i \sim \mathcal{N}(0,1)$ under H_0
- $X_i \sim \mathcal{N}(\bar{\mu},1)$ under H_1
- $\mathrm{cor}(X_i, X_j) = \rho$ for $i \neq j$
- $\alpha = 0.25$

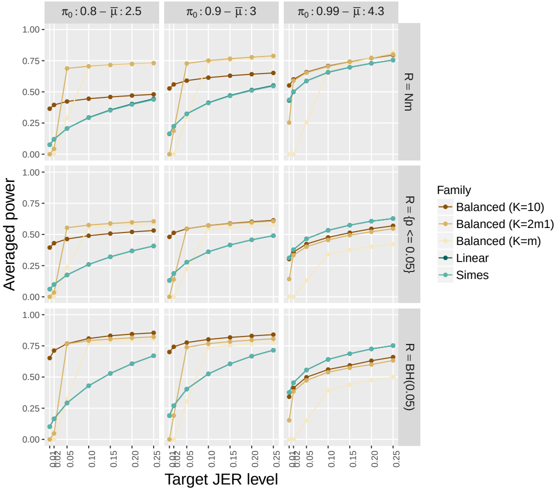
JER control under Gaussian equi-correlation

Balanced template, unknown dependence (calibration by sign-flipping)



- $X_i \sim \mathcal{N}(0,1)$ under H_0
- $X_i \sim \mathcal{N}(\bar{\mu},1)$ under H_1
- $\mathrm{cor}(X_i, X_j) = \rho$ for $i \neq j$
- $\alpha = 0.25$

Estimation power for under independence



- $X_i \sim \mathcal{N}(0,1)$ under H_0
- $X_i \sim \mathcal{N}(\bar{\mu},1)$ under H_1
- $\mathrm{cor}(X_i, X_j) = 0$ for $i \neq j$
- $\bar{\mu} = 2$
- Estimation power: $E(\overline{S}(\mathcal{H}_1))/m_1$

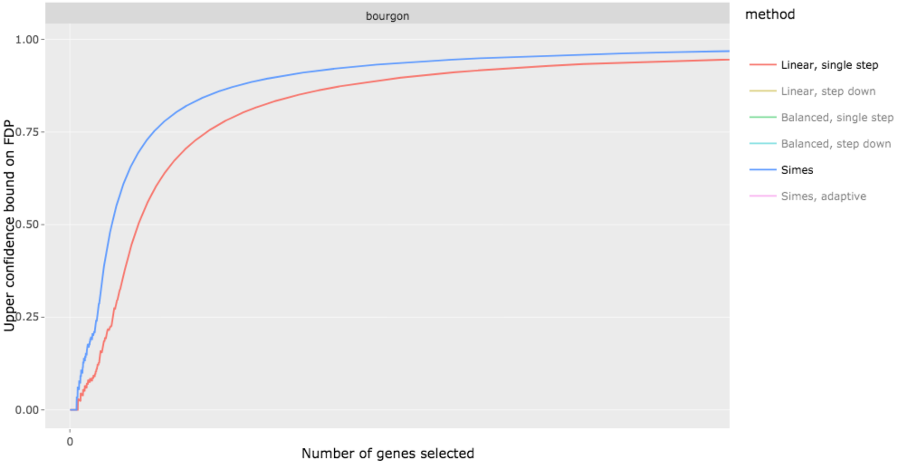
λ-Calibration by permutations

For two sample tests, the distribution of

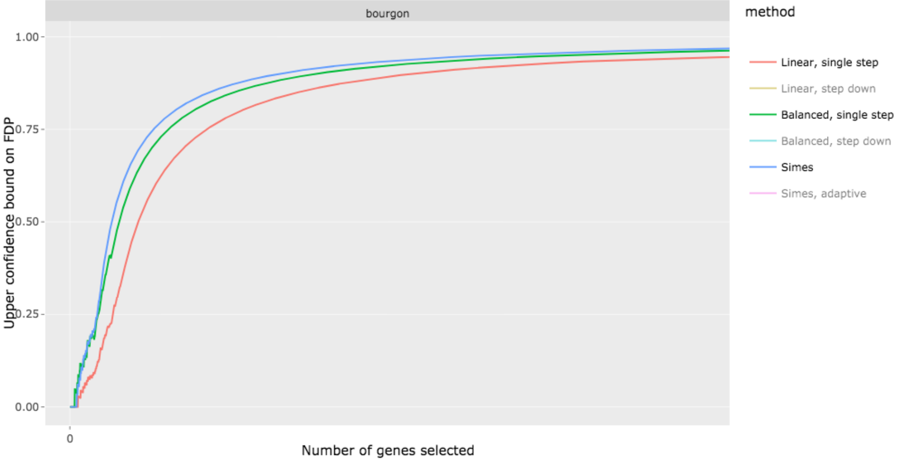
$$\min_k \left\{ t_k^{-1} \left(p_{(k:\mathcal{H}_0)} \right) \right\}$$

can be sampled from using *permutations* of the group labels

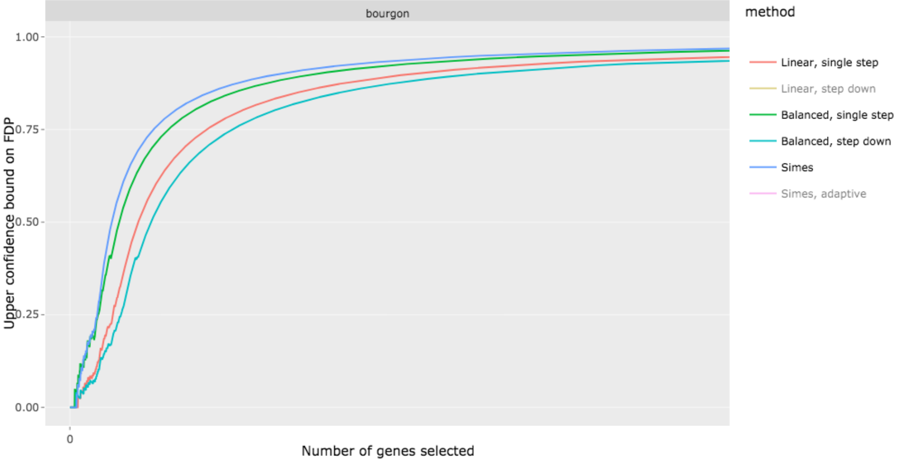
Improved confidence envelope using permutations



Improved confidence envelope using permutations



Improved confidence envelope using permutations



Conclusions

Summary

- JER control induces post hoc bounds
- Existing bounds recovered from probabilistic inequalities (Simes)
- Framework to build adaptive JER control

Results not discussed here

- Step-down procedures (adaptation to $|\mathcal{H}_0|$)
- Detection power: connection to "higher criticism" in a sparse setting

Ongoing/future works

- Choice of the template and its size
- Applications (GWAS, differential expression, neuro-imaging)
- Structured rejection sets: algorithms and statistical results
- Software (R package sansSouci) and visualization tools