Post hoc bounds on false positives using reference families

Pierre Neuvial

CNRS and Institut de Mathématiques de Toulouse (France)

joint work with Gilles Blanchard, Guillermo Durand, Etienne Roquain, Marie Perrot-Dockès https://arxiv.org/abs/1910.11575

Funded by ANR SansSouci

Case study: differential expression in genomics

Example: Leukemia data set

Chiaretti et. al., *Clinical cancer research*, 11(20):7209–7219, 2005

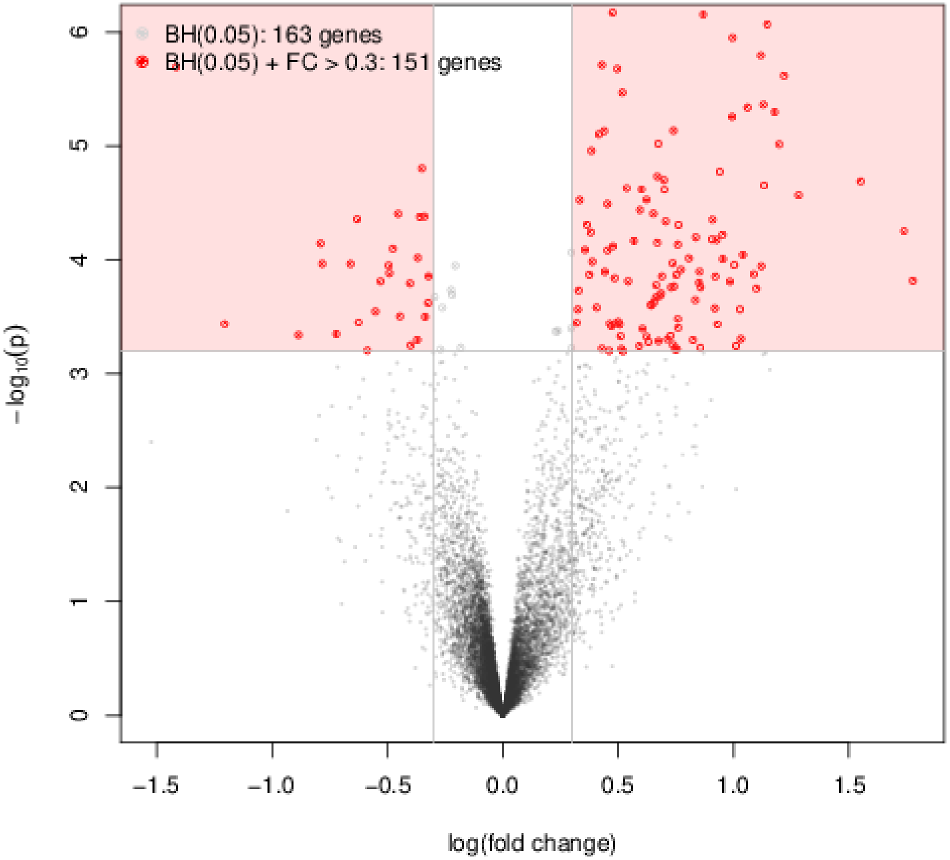
Data: gene expression measurements (mRNA)

- $m = 12625$ genes
- $n = 79$ cancer patients in two subgroups:
  - BCR/ABL: 37 patients
  - NEG: 42 patients

Question

Find genes whose average expression differs between the two groups

Leukemia data set: volcano plot



Notation

- $\mathcal{H} = \{1, \ldots m\}$ m null hypotheses to be tested
- $\mathcal{H}_0 \subset \mathcal{H}$: true null hypotheses, $\mathcal{H}_1 = \mathcal{H} \setminus \mathcal{H}_0$
- $m_0 = |\mathcal{H}_0|, \pi_0 = m_0/m$
- $(p_i)_{1 \leq i \leq m}$: p-values
- $R \subset \mathcal{H}$: a set of rejected hypotheses
- $|R \cap \mathcal{H}_0|$: number of "false positives" within R.

Goal: post hoc inference

Find a $(1-\alpha)$-level *post hoc upper bound* on $|S \cap \mathcal{H}_0|$, ie V_α such that

$$\mathbb{P}\left(\forall S \subset \{1 \ldots m\}, \quad |S \cap \mathcal{H}_0| \leq V_\alpha(S)\right) \geq 1 - \alpha$$

Some related works

- Genovese & Wasserman, *Ann. Stat.*, 2006; Goeman & Solari, *Stat. Sci.*, 2011
- Katsevich and Ramdas, ArXiv:1803.06790
- Meijer, Krebs, and Goeman *SAGMB*, 2015

Starting point: post hoc bound via Simes' inequality

Under PRDS, Simes' inequality implies

$$\mathbb{P}\Big(\forall k, |R_k \cap \mathcal{H}_0| \leq k-1\Big) \geq 1-\alpha$$

where $R_k = \{i/p_i \leq \alpha k/m\}$

Corollary: $(1-\alpha)$ post hoc bound on $|S \cap \mathcal{H}_0|$

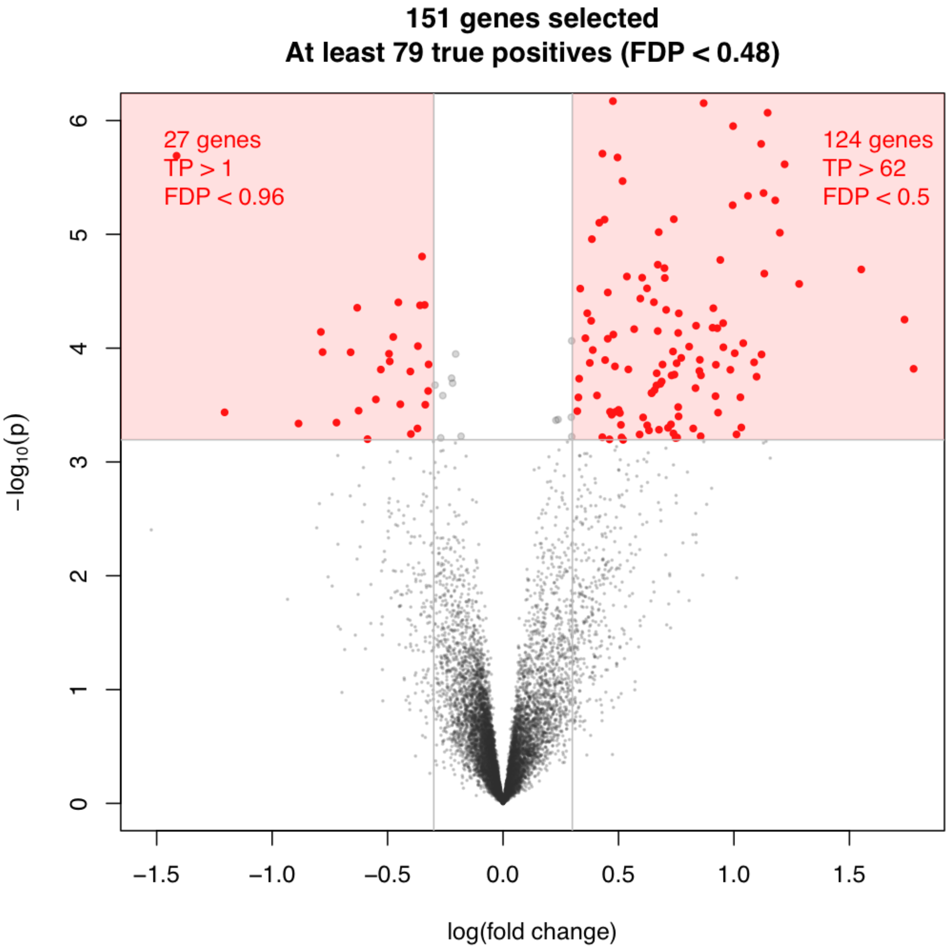
$$\overline{V}_\alpha(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} 1\{p_i > \alpha k/m\} + k - 1 \right\}$$

Recovers the bound of Goeman and Solari, *Stat. Science*, 2011.

Proof:

$$\begin{aligned} |S \cap \mathcal{H}_0| &= |S \cap R_k^c \cap \mathcal{H}_0| + |S \cap R_k \cap \mathcal{H}_0| \\ &\leq |S \cap R_k^c| + |R_k \cap \mathcal{H}_0| \end{aligned}$$

Leukemia data set: volcano plot (Simes-based bound)



Post hoc control via reference families

Joint Error Rate control implies post hoc bound

Definition: JER controlling family

$\mathfrak{R} = (R_k, \zeta_k)_k$ such that $\quad \mathbb{P}\big(\forall k, |R_k \cap \mathcal{H}_0| \leq \zeta_k\big) \geq 1 - \alpha$

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

Property: interpolation yields valid $(1 - \alpha)$ post hoc bounds

$$V_\alpha^*(S) = \max\{|S \cap A| : A \text{ s.t. } \forall k, |R_k \cap A| \leq \zeta_k\}$$

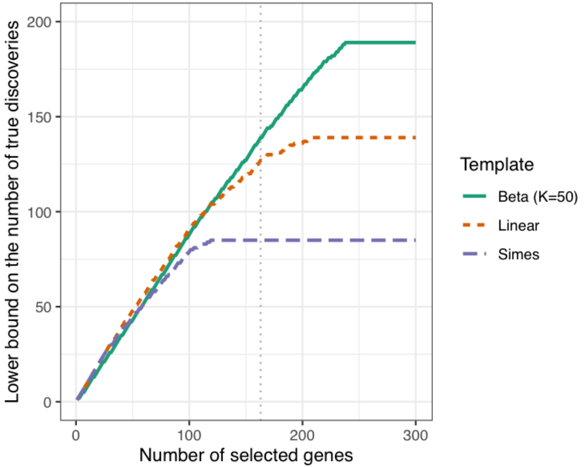
$$\overline{V}_\alpha(S) = \min_{1 \leq k \leq |S|} \{|S \cap R_k^c| + \zeta_k\}$$

Simes: $V_\alpha^*(S) = \overline{V}_\alpha(S) = \min_{1 \leq k \leq |S|} \left\{\sum_{i \in S} 1\{p_i > \alpha k / m\} + k - 1\right\}$

Main question: how to obtain JER control?

Contributions: post hoc bounds in two dual cases

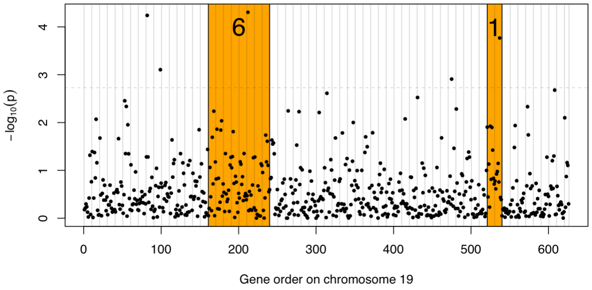
p-value level sets



- Fixed $\zeta_k (= k - 1)$
- $R_k = R_k(X)$

JER control = joint *control* of the k-FWER

structured hypotheses



- Fixed R_k given by prior knowledge
- Find $\zeta_k = \zeta_k(X)$

JER control = joint *estimation* of $|R_k \cap \mathcal{H}_0|$

Case 1: Fixed ζ_k, random R_k

- Blanchard, N., Roquain: Post Hoc Confidence Bounds on False Positives Using Reference Families *Annals of Statistics*, to appear.
- R package sansSouci

Setup: $\zeta_k = k - 1, R_k = \{i : p_i \leq t_k(\lambda)\}$

Properties

- The R_k are nested $\Rightarrow V^*_\alpha(S) = \overline{V}_\alpha(S)$
- For the reference family (R_k, ζ_k):

JER control holds for any λ such that

$$\mathbb{P}\left(\exists k, p_{(k:\mathcal{H}_0)} \leq t_k(\lambda)\right) \leq \alpha$$

Examples

- $\lambda = \alpha$ for $t_k(\lambda) = \lambda k/m$ under PRDS
- $\lambda = \alpha$ for $t_k(\lambda) = \lambda-$ quantile of $Beta(k+1, m-k+1)$ under independence
- adaptivity to dependence?

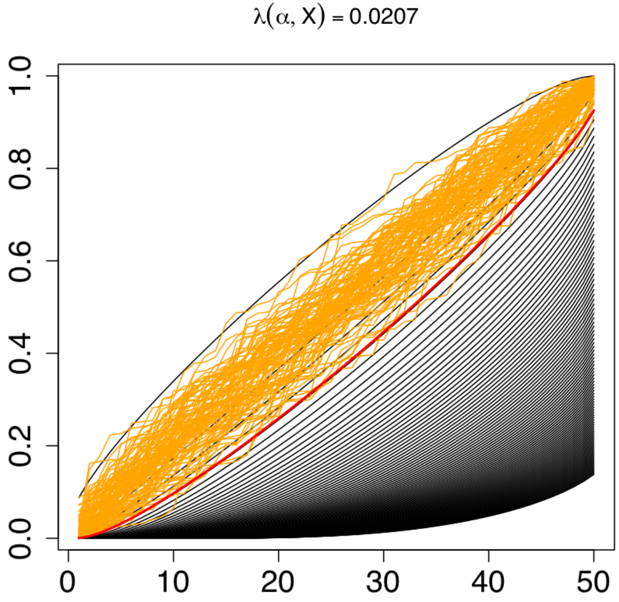
Adaptivity to dependence

Goal: estimate the largest λ such that

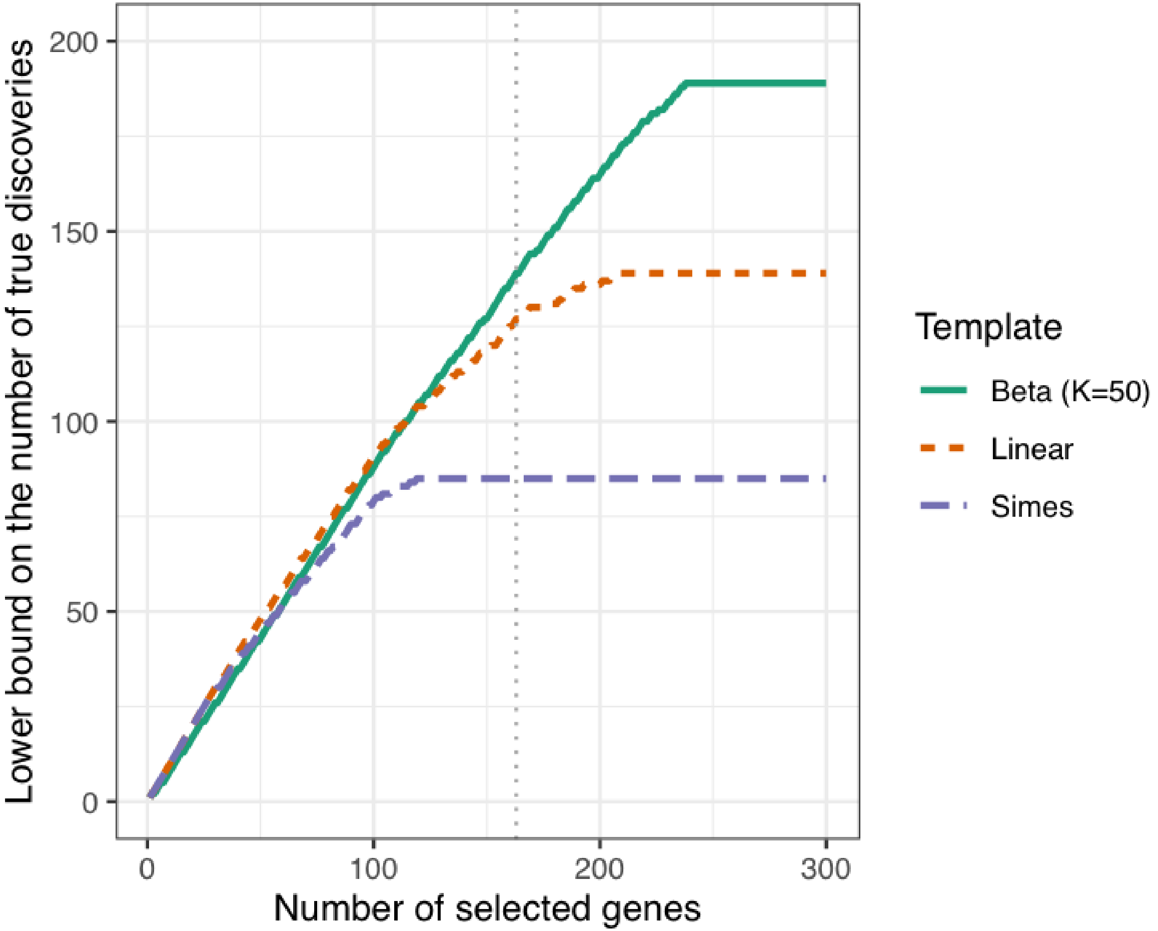
$$\mathbb{P}\left(\exists k, p_{(k:\mathcal{H}_0)} \leq t_k(\lambda)\right) \leq \alpha$$

Tool: *randomization*, e.g. class label permutation in multiple two-sample tests

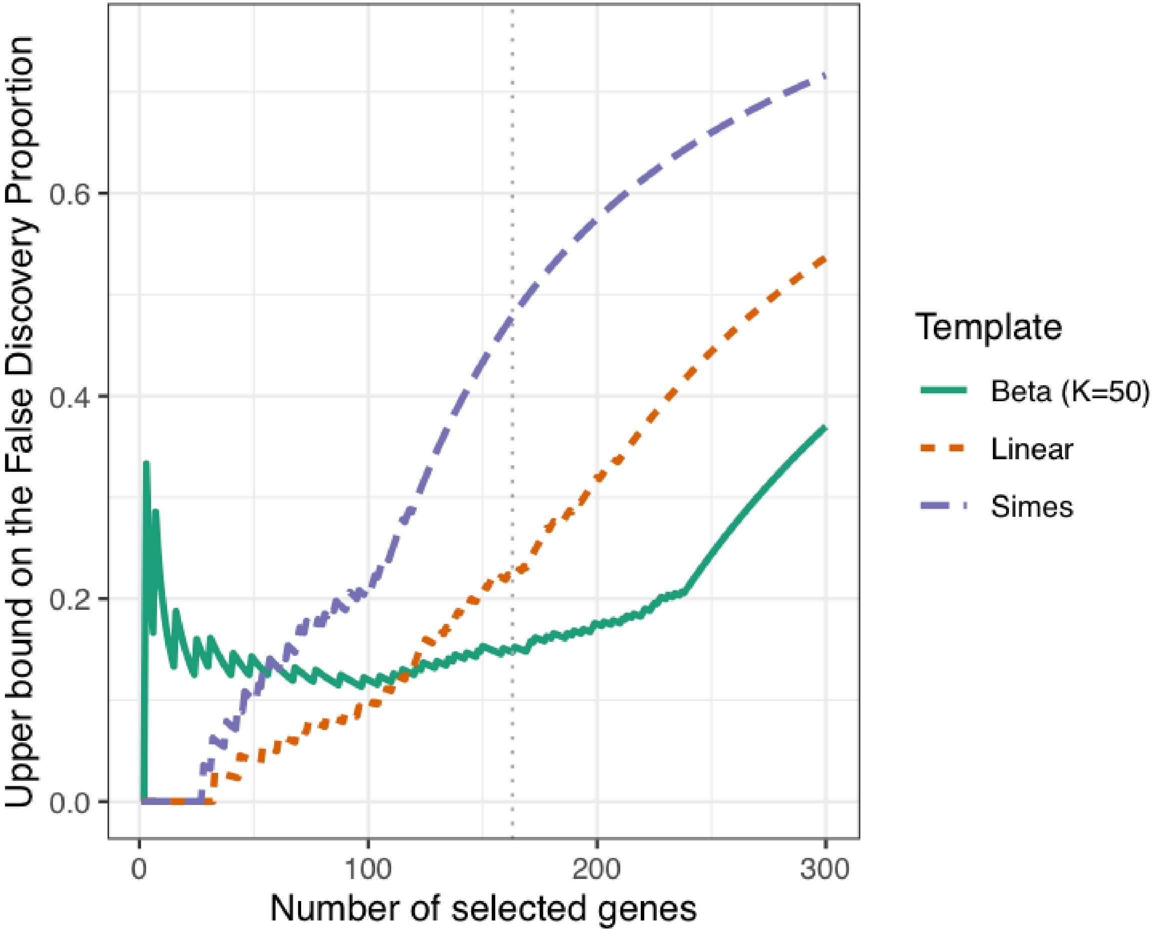
Example: $t_k(\lambda) = \lambda-$ quantile of $Beta(k+1, m-k+1)$



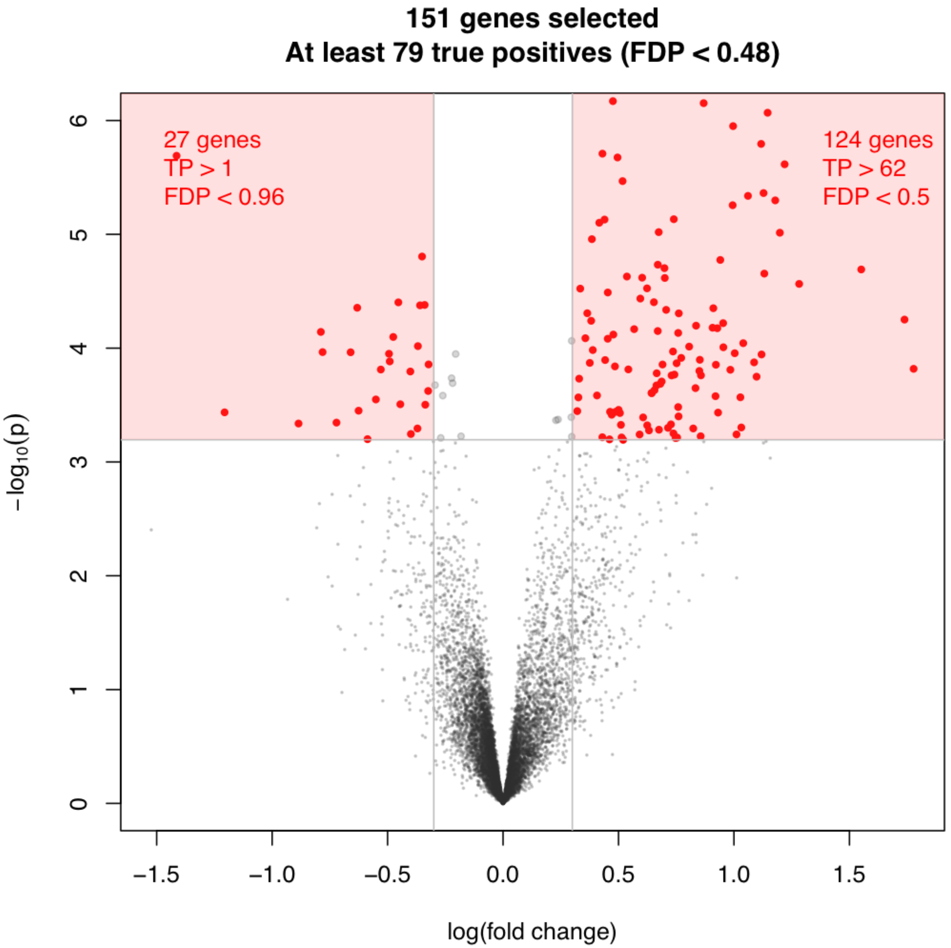
Leukemia data: confidence bounds on $|S \cap \mathcal{H}_1|$



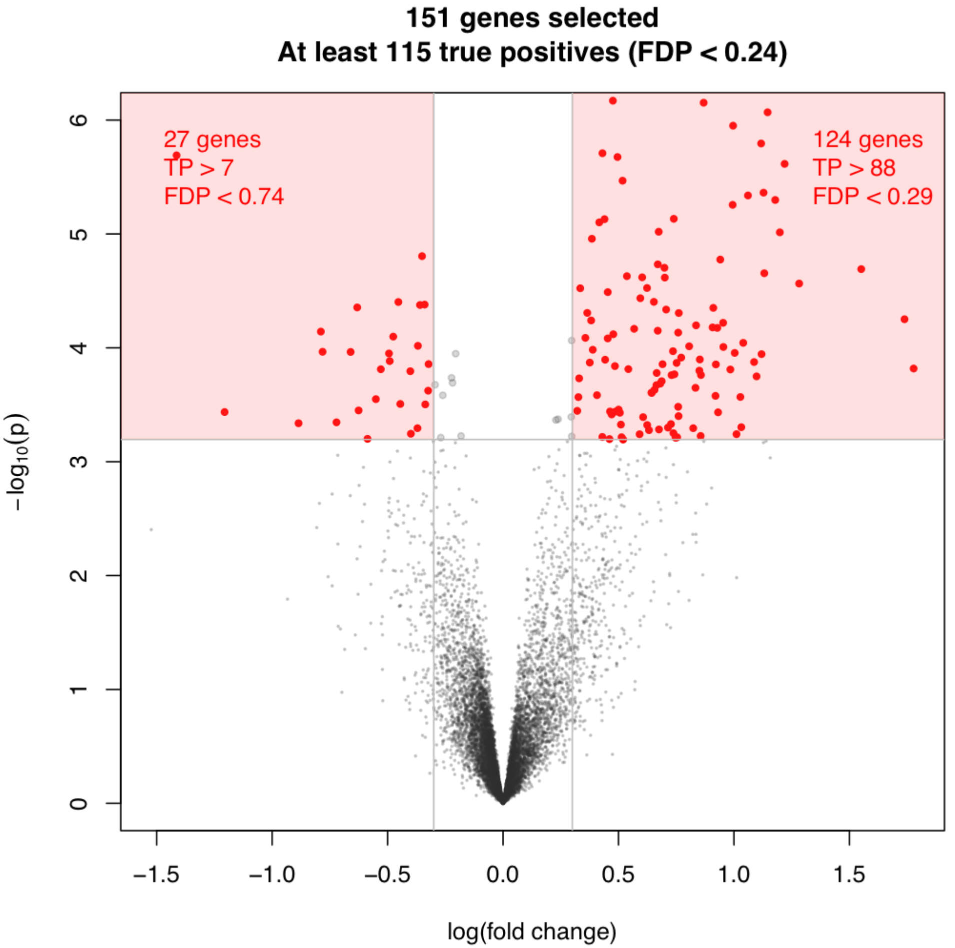
Leukemia data: confidence bounds on FDP = $\frac{|S \cap \mathcal{H}_0|}{|S| \vee 1}$



Leukemia data set: volcano plot (Simes-based bound)



Leukemia data set: volcano plot (after λ-calibration)

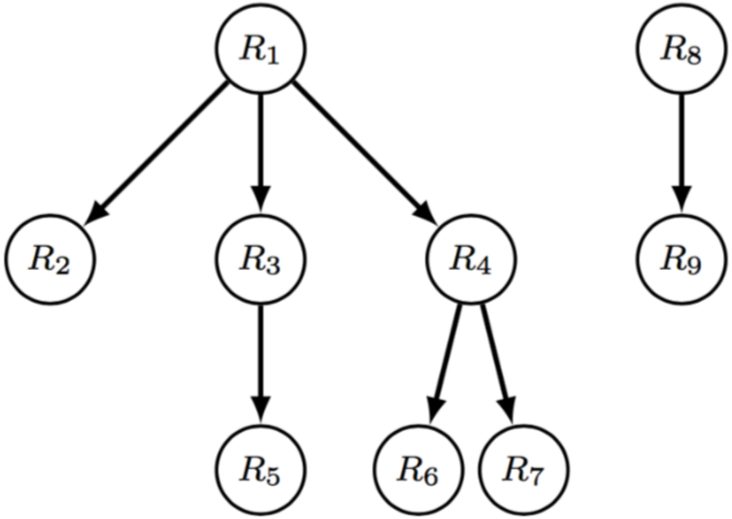


Case 2: Fixed R_k, random ζ_k

- Durand, Blanchard, N., Roquain: Post hoc false positive control for structured hypotheses, Scandinavian Journal of Statistics (2020). arxiv:1807.01470
- R package sansSouci

Setup: Fixed R_k, random ζ_k

Forest assumption: the $(R_k)_{k=1...K}$ are either nested or disjoint



Questions:

1. How to chose $\zeta_k(X)$ yielding JER control?
2. How to estimate the associated post hoc bound V_α^*

1. JER control

Device: DKWM inequality

- Dvoretzky, Kiefer, and Wolfowitz (1956) *Ann. Math. Stat.*
- Massart (1990) *Ann. Prob.*

Proposition

Under independence, JER control is obtained for

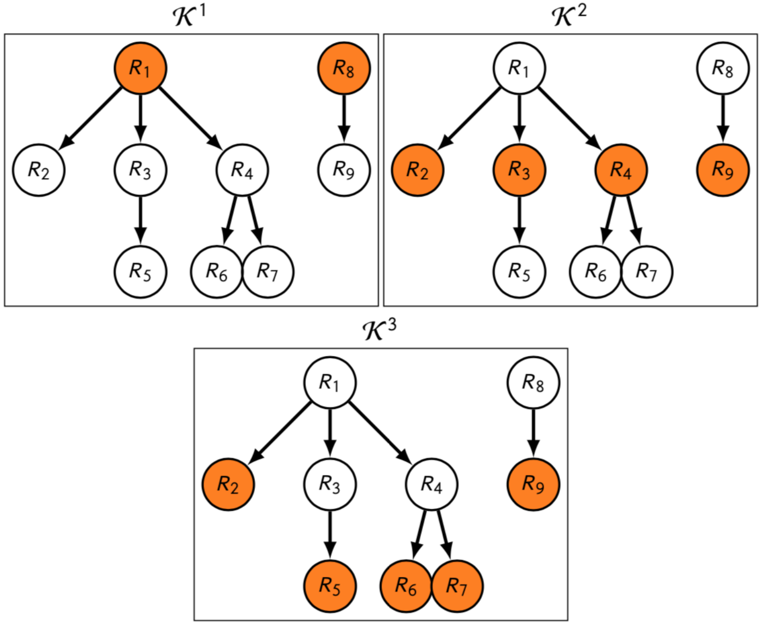
$$\zeta_k(X) = |R_k| \wedge \min_{t\in[0,1)} \left\lfloor \frac{C}{2(1-t)} + \left(\frac{C^2}{4(1-t)^2} + \frac{\sum_{i\in R_1} \mathbf{1}\{p_i(X) > t\}}{1-t} \right)^{1/2} \right\rfloor^2,$$

where $C = \sqrt{\frac{1}{2}\log\left(\frac{K}{\alpha}\right)}$

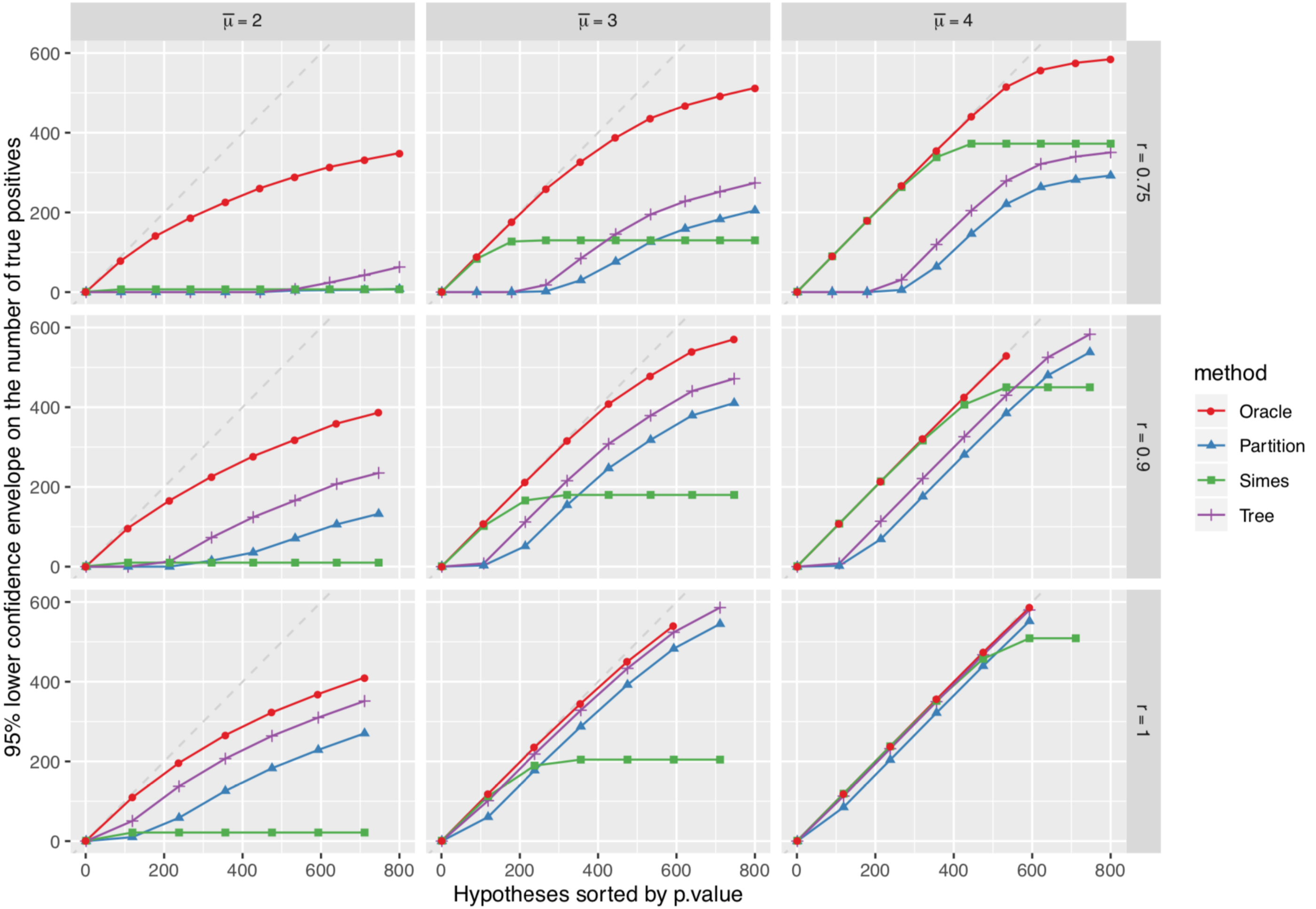
2. Algorithm to compute V_α^*

Proposition

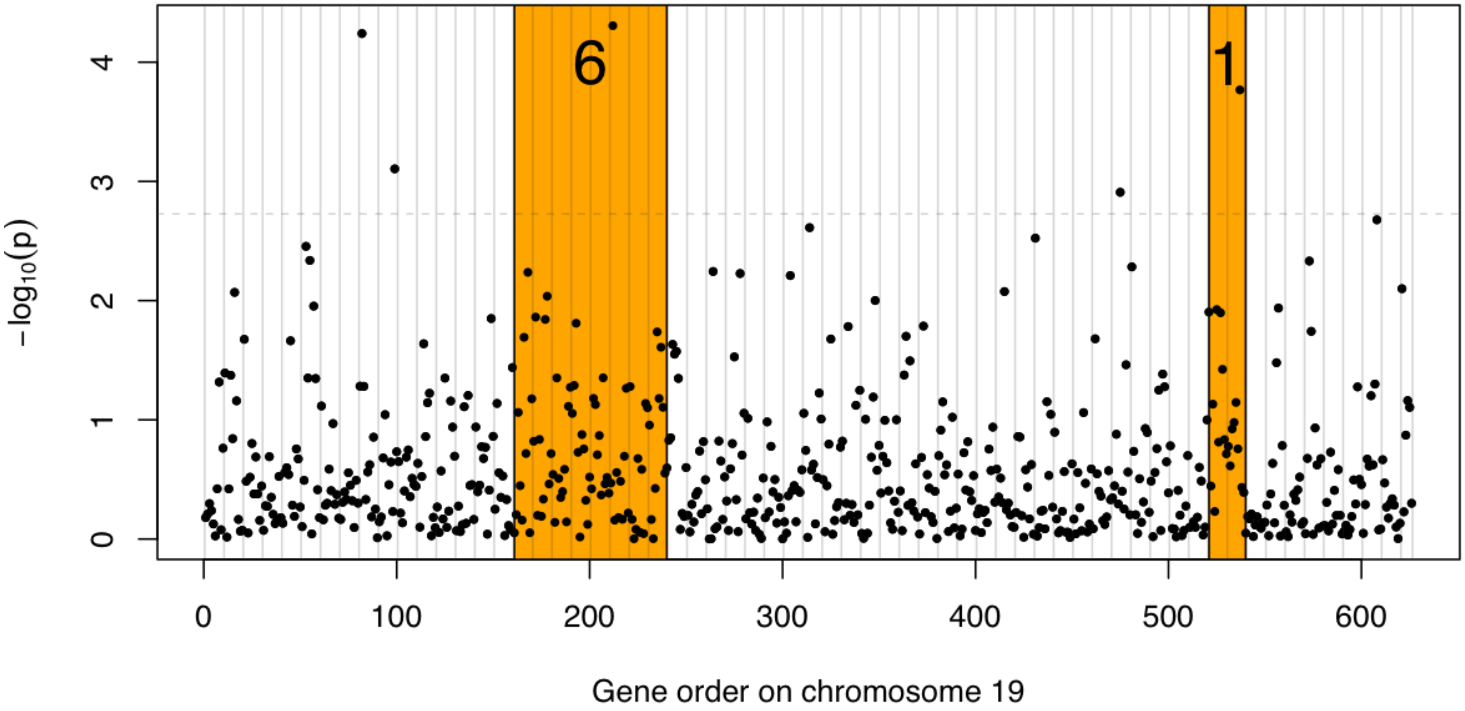
The bound V_α^* is obtained recursively by examining partitions at each possible depth in the forest.



Numerical experiments: Simes vs tree-based methods



Leukemia data set: regional association plot



The selection can be done interactively:

https://pneuvial.shinyapps.io/posthoc-bounds_ordered-hypotheses/

Conclusions

- Versatile approach to post hoc inference
  - JER control $\Rightarrow$ post hoc bounds
- JER control can be obtained from classical probabilistic inequalities
  - Fixed ζ_k, random R_k: Simes' inequality under PRDS
  - Fixed R_k, random ζ_k: DKWM inequality under independence
- adaptation to dependence: sharper JER control can be obtained by randomization

Extensions

- Applications to genomic data analysis
  - e.g. differential analysis along the genome
- Fixed R_k, random ζ_k: extension to specific dependence settings

See poster of Marie Perrot-Dockès:

"Improving structured post hoc inference via a Hidden Markov Model"