Simultaneous FDP bounds for nested sequences of rejection sets

Eugene Katsevich

Department of Statistics and Data Science
Carnegie Mellon University

December 14, 2019

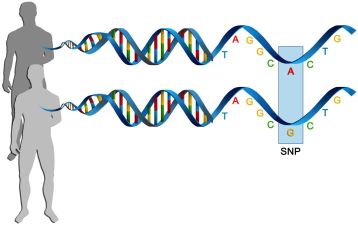
Joint work with Aaditya Ramdas

E. Katsevich and A. Ramdas. Simultaneous high-probability bounds on the FDP in structured, regression, and online settings. *Annals of Statistics*, to appear, 2020.

Genome-wide association studies

Genotypes $X_1, \ldots, X_p$ at p SNPs and trait Y measured for n individuals.

Goal: find a set of SNPs associated with the trait.



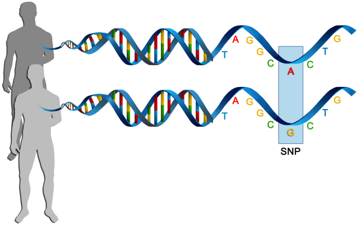
(Source: Google)

UK Biobank data: $p \approx n \approx 500{,}000$.

Genome-wide association studies

Genotypes $X_1, \ldots, X_p$ at p SNPs and trait Y measured for n individuals.

Goal: find a set of SNPs associated with the trait.



(Source: Google)

UK Biobank data: $p \approx n \approx 500{,}000$.

Knockoffs (Barber and Candès, 2015), a variable selection method with FDR control, recently applied to GWAS (Sesia et al., 2019).

Step 1: Compute knockoff statistic for each SNP

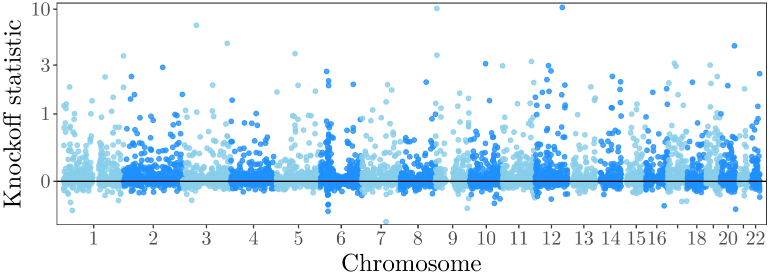
1. Generate synthetic negative control SNPs (knockoffs).
2. Apply lasso to all original and knockoff SNPs.
3. For SNP k, define knockoff statistic

$$W_k = |\hat{\beta}_k| \ - |\hat{\beta}_{k+p}|.$$

Step 1: Compute knockoff statistic for each SNP

1. Generate synthetic negative control SNPs (knockoffs).
2. Apply lasso to all original and knockoff SNPs.
3. For SNP k, define knockoff statistic

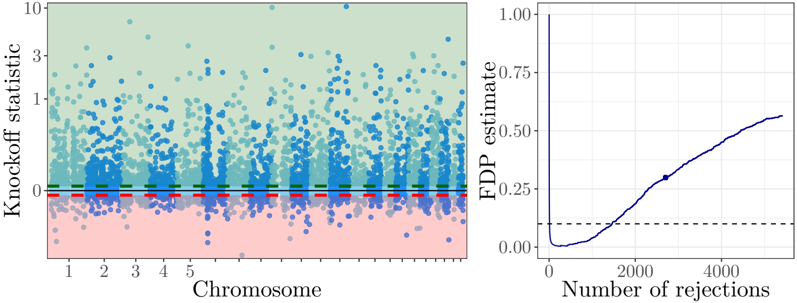
$$W_k = |\hat{\beta}_k| \ - |\hat{\beta}_{k+p}|.$$



Knockoff statistics for platelet count, UKBB data *(Sesia et al., 2019)*

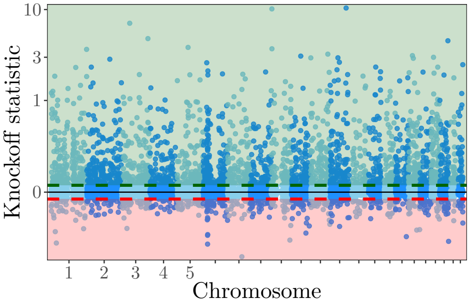
Step 2: Find the threshold for FDR control

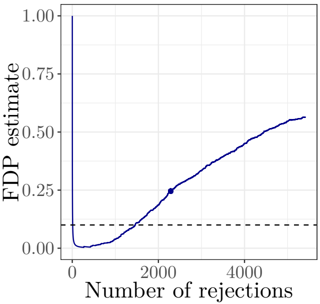
$$\mathcal{R}(t) = \{k : W_k \geq t\}; \quad \widehat{\text{FDP}}(t) = \frac{1 + |\{k : W_k \leq -t\}|}{|\mathcal{R}(t)|}$$



Step 2: Find the threshold for FDR control

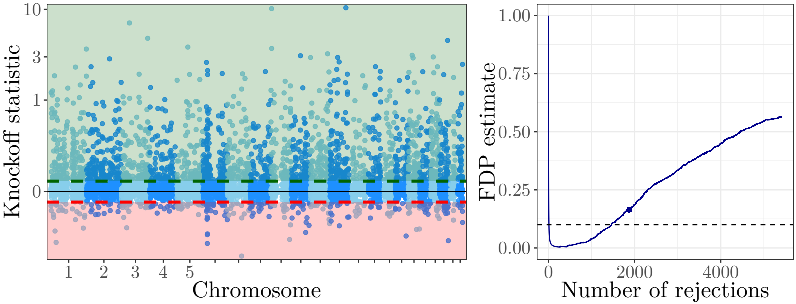
$$\mathcal{R}(t) = \{k : W_k \geq t\}; \quad \widehat{\mathrm{FDP}}(t) = \frac{1 + |\{k : W_k \leq -t\}|}{|\mathcal{R}(t)|}$$





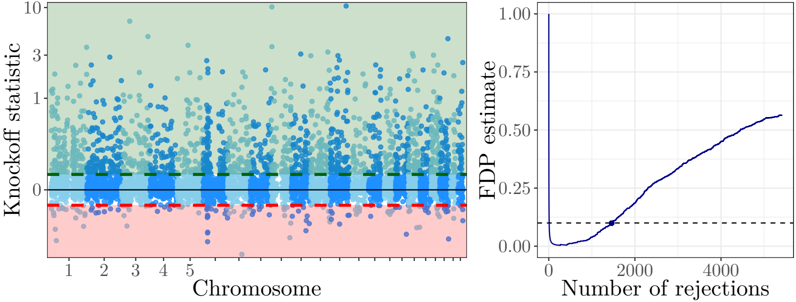
Step 2: Find the threshold for FDR control

$$\mathcal{R}(t) = \{k : W_k \geq t\}; \quad \widehat{\mathrm{FDP}}(t) = \frac{1 + |\{k : W_k \leq -t\}|}{|\mathcal{R}(t)|}$$



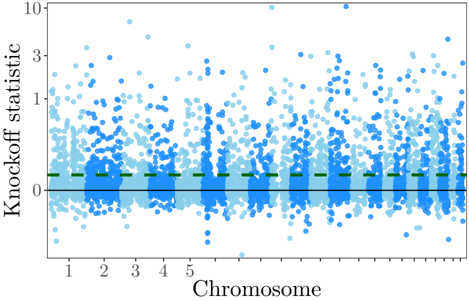
Step 2: Find the threshold for FDR control

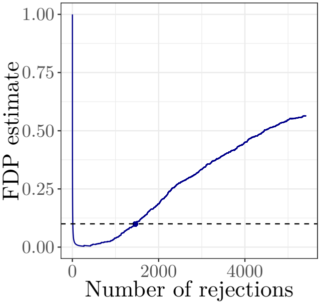
$$\mathcal{R}(t) = \{k : W_k \geq t\}; \quad \widehat{\text{FDP}}(t) = \frac{1 + |\{k : W_k \leq -t\}|}{|\mathcal{R}(t)|}$$



Step 2: Find the threshold for FDR control

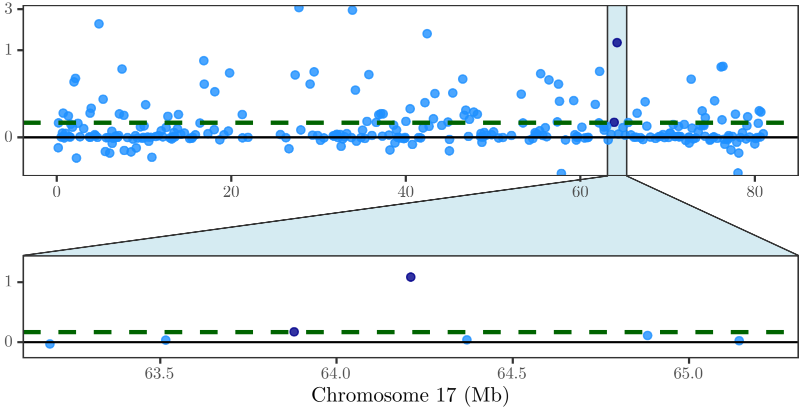
$$\mathcal{R}(t) = \{k : W_k \geq t\}; \quad \widehat{\mathrm{FDP}}(t) = \frac{1 + |\{k : W_k \leq -t\}|}{|\mathcal{R}(t)|}$$



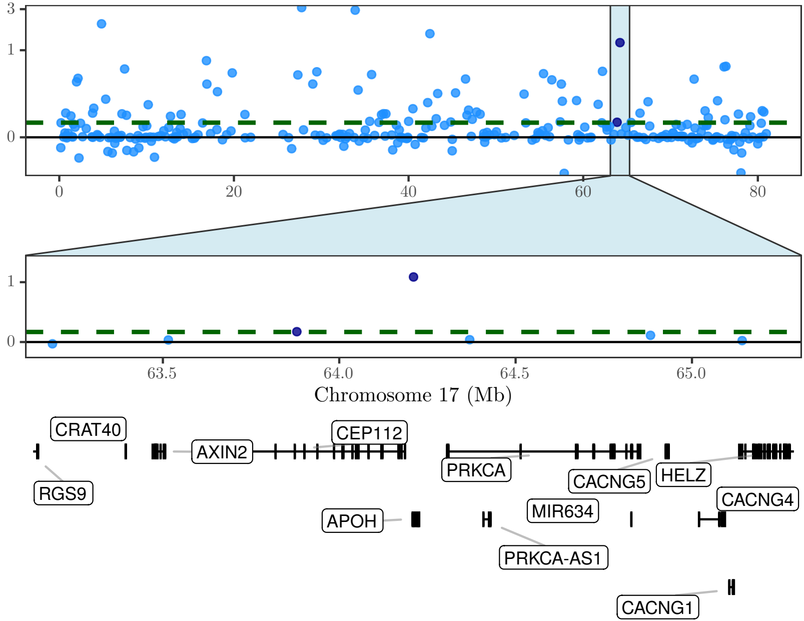


$|\mathcal{R}(t^*)| = 1460$ SNPs associated with platelet count at $q = 0.1$.

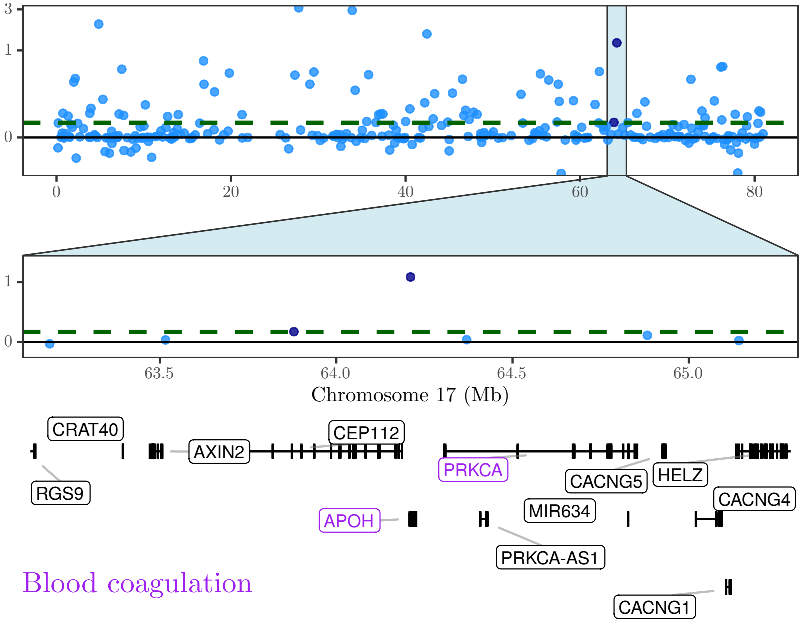
Next step: Biological interpretation of findings



Next step: Biological interpretation of findings



Next step: Biological interpretation of findings

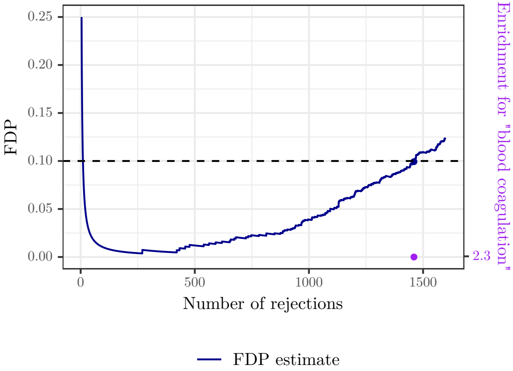


Blood coagulation

Gene Ontology enrichment analysis (McLean et al., 2010)

Enrichment: freq. of annotation near all discovered SNPs.

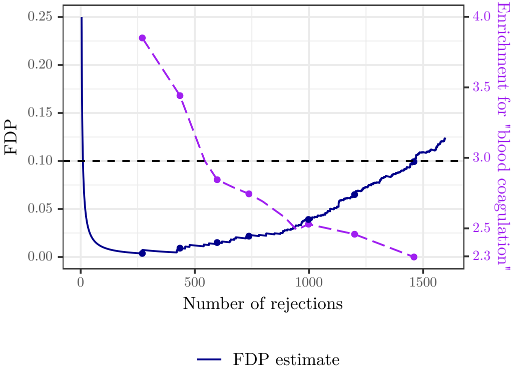
Knockoffs SNPs have enrichment 2.3 for blood coagulation.



Gene Ontology enrichment analysis (McLean et al., 2010)

Enrichment decreases with rejection set size.

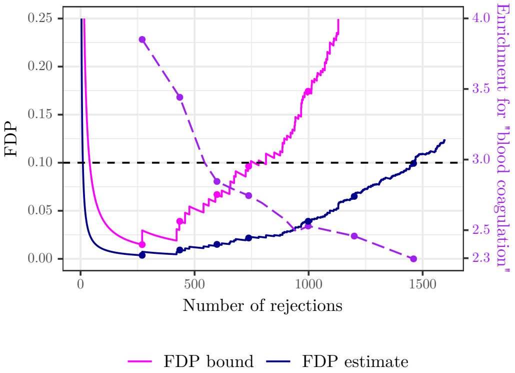
Desirable to explore along knockoffs path.



Simultaneous FDP upper bound permits exploration

$$\overline{\mathrm{FDP}}(t) = \frac{-\log(\alpha)}{\log(2-\alpha)} \cdot \widehat{\mathrm{FDP}}(t)$$

Theorem (KR '19). With prob. $1-\alpha$, $\mathrm{FDP}(t) \leq \overline{\mathrm{FDP}}(t)\ \forall t$.

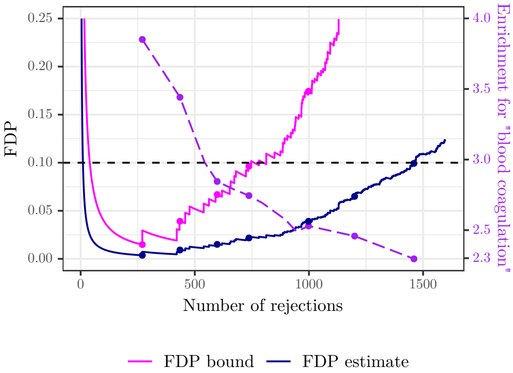


Simultaneous FDP upper bound permits exploration

$$\overline{\mathrm{FDP}}(t) = \frac{-\log(\alpha)}{\log(2-\alpha)} \cdot \widehat{\mathrm{FDP}}(t)$$

Theorem (KR '19). With prob. $1-\alpha$, $\mathrm{FDP}(t) \leq \overline{\mathrm{FDP}}(t)\ \forall t$.

- Simultaneous
- Closed form
- Finite sample

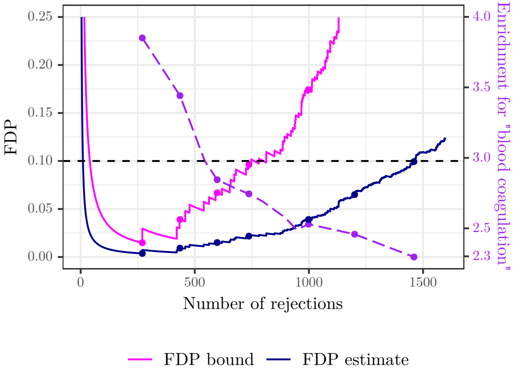


Simultaneous FDP upper bound permits exploration

$$\overline{\mathrm{FDP}}(t) = \frac{-\log(\alpha)}{\log(2-\alpha)} \cdot \widehat{\mathrm{FDP}}(t)$$

Theorem (KR '19). With prob. $1-\alpha$, $\mathrm{FDP}(t) \leq \overline{\mathrm{FDP}}(t)\ \forall t$.

- Simultaneous
- Closed form
- Finite sample



For a factor of 4.5, can move from bounding FDP on average at one point to bounding it with 95% confidence at all points.

A glimpse of the proof

We have

$$\frac{\mathrm{FDP}(t)}{\widehat{\mathrm{FDP}}(t)} \leq \frac{|\{\text{null } k : |W_k| \geq t,\ \text{sign}(W_k) = \text{``} + \text{''}\}|}{1 + |\{\text{null } k : |W_k| \geq t,\ \text{sign}(W_k) = \text{``} - \text{''}\}|}.$$

A glimpse of the proof

We have

$$\frac{\mathrm{FDP}(t)}{\widehat{\mathrm{FDP}}(t)} \le \frac{|\{\text{null } k : |W_k| \ge t,\ \text{sign}(W_k) = \text{``}+\text{''}\}|}{1 + |\{\text{null } k : |W_k| \ge t,\ \text{sign}(W_k) = \text{``}-\text{''}\}|}.$$

Knockoffs FDR proof uses backward martingale to show

$$\mathbb{E}\left[\frac{\mathrm{FDP}(t^*)}{\widehat{\mathrm{FDP}}(t^*)}\right] \le 1.$$

A glimpse of the proof

We have

$$\frac{\mathrm{FDP}(t)}{\widehat{\mathrm{FDP}}(t)} \leq \frac{|\{\text{null } k : |W_k| \geq t,\ \mathrm{sign}(W_k) = \text{``}+\text{''}\}|}{1 + |\{\text{null } k : |W_k| \geq t,\ \mathrm{sign}(W_k) = \text{``}-\text{''}\}|}.$$

Knockoffs FDR proof uses backward martingale to show

$$\mathbb{E}\left[\frac{\mathrm{FDP}(t^*)}{\widehat{\mathrm{FDP}}(t^*)}\right] \leq 1.$$

Our proof uses forward martingale to show

$$\mathbb{P}\left[\sup_{t\geq 0} \frac{\mathrm{FDP}(t)}{\widehat{\mathrm{FDP}}(t)} \geq x\right] \leq \exp(-x\theta_x); \quad \theta_x \approx \log(2).$$

Other FDP bounds of this type

General idea: Repurpose path constructions and FDP estimates from existing FDR procedures.

Other FDP bounds of this type

General idea: Repurpose path constructions and FDP estimates from existing FDR procedures.

We prove similar bounds in the following settings:

Other FDP bounds of this type

General idea: Repurpose path constructions and FDP estimates from existing FDR procedures.

We prove similar bounds in the following settings:

- hypotheses ordered by p-value ($\mathcal{R}(t) = \{k : p_k \leq t\}$);

$$\overline{\mathrm{FDP}}(t) = \frac{-\log(\alpha)}{\log(1 - \log(\alpha))} \cdot \frac{1 + m \cdot t}{|\mathcal{R}(t)|};$$

Other FDP bounds of this type

General idea: Repurpose path constructions and FDP estimates from existing FDR procedures.

We prove similar bounds in the following settings:

- hypotheses ordered by p-value ($\mathcal{R}(t) = \{k : p_k \leq t\}$);

$$\overline{\text{FDP}}(t) = \frac{-\log(\alpha)}{\log(1-\log(\alpha))} \cdot \frac{1 + m \cdot t}{|\mathcal{R}(t)|};$$

- hypotheses have a priori ordering;

Other FDP bounds of this type

General idea: Repurpose path constructions and FDP estimates from existing FDR procedures.

We prove similar bounds in the following settings:

- hypotheses ordered by p-value ($\mathcal{R}(t) = \{k : p_k \leq t\}$);

$$\overline{\mathrm{FDP}}(t) = \frac{-\log(\alpha)}{\log(1 - \log(\alpha))} \cdot \frac{1 + m \cdot t}{|\mathcal{R}(t)|};$$

- hypotheses have a priori ordering;
- hypothesis order determined interactively;

Other FDP bounds of this type

General idea: Repurpose path constructions and FDP estimates from existing FDR procedures.

We prove similar bounds in the following settings:

- hypotheses ordered by p-value ($\mathcal{R}(t) = \{k : p_k \leq t\}$);

$$\overline{\mathrm{FDP}}(t) = \frac{-\log(\alpha)}{\log(1 - \log(\alpha))} \cdot \frac{1 + m \cdot t}{|\mathcal{R}(t)|};$$

- hypotheses have a priori ordering;
- hypothesis order determined interactively;
- hypotheses arrive in an online fashion.

Other FDP bounds of this type

General idea: Repurpose path constructions and FDP estimates from existing FDR procedures.

We prove similar bounds in the following settings:

- hypotheses ordered by p-value ($\mathcal{R}(t) = \{k : p_k \leq t\}$);

$$\overline{\mathrm{FDP}}(t) = \frac{-\log(\alpha)}{\log(1 - \log(\alpha))} \cdot \frac{1 + m \cdot t}{|\mathcal{R}(t)|};$$

- hypotheses have a priori ordering;
- hypothesis order determined interactively;
- hypotheses arrive in an online fashion.

Results require p-value independence, but some robustness to correlation observed in simulations.

Prior work on simultaneous inference and exploration

Multiple testing setting:

- Goeman and Solari (2011)
- Blanchard, Neuvial and Roquain (2017)
- Rosenblatt, Finos, Weeda, Solari, and Goeman (2018)

Regression setting:

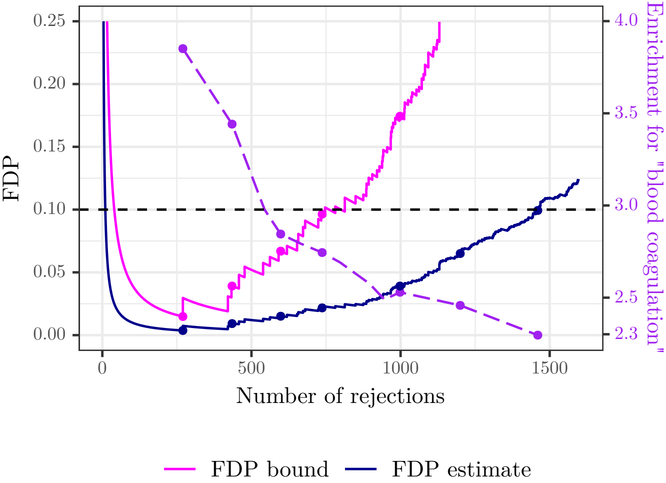
- Berk, Brown, Buja, Zhang, and Zhao (2013).
- Bachoc, Preinerstorfer, and Steinberger (2016)
- Kuchibhotla, Brown, Buja, George, and Zhao (2018)

Conclusion

Simultaneous high-probability FDP bounds for nested sequences of rejection sets.

- Our bounds are finite sample and closed form.
- We add to growing literature on simultaneous inference, broadening its scope to include variety of testing settings.
- Link between simultaneous inference and FDR literature.

Thank you!


FDP
0.25
0.20
0.15
0.10
0.05
0.00
0
500
1000
1500
Number of rejections
4.0
3.5
3.0
2.5
2.3
Enrichment for "blood coagulation"
FDP bound
FDP estimate