Robustness of Controlling the False Discovery Rate

Weijie J. Su

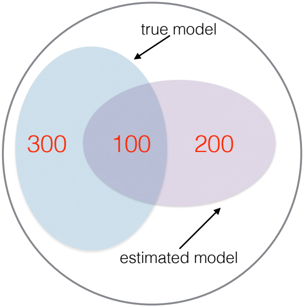
University of Pennsylvania

Robust and High-Dimensional Statistics, Simons Institute, October 31, 2018

False discovery rate (FDR)

$$\text{FDP} = \frac{\#\text{false discoveries}}{\#\text{discoveries}} = \frac{200}{100 + 200}$$

$$\text{FDR} = \mathbb{E}\,\text{FDP}$$



- FDP: false discovery proportion
- Want to control FDR $\le q$ (e.g. $q = 0.05, 0.1$)
- Proposed by Benjamini and Hochberg '95

The Benjamini–Hochberg (BH) procedure

Given p-values $p_1, \ldots, p_m$ corresponding to m hypotheses

BH procedure the "great"

- Let R be the largest such that at least R of $p_1, \ldots, p_m$ are $\le \frac{qR}{m}$
- Reject the R smallest p-values

The Benjamini–Hochberg (BH) procedure

Given p-values $p_1, \ldots, p_m$ corresponding to m hypotheses

BH procedure the "great"

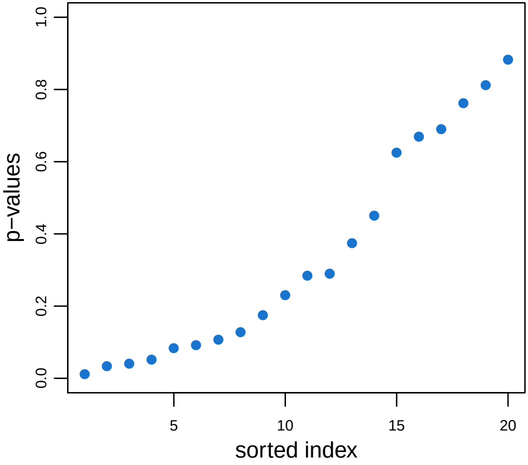
- Let R be the largest such that at least R of $p_1, \ldots, p_m$ are $\le \frac{qR}{m}$
- Reject the R smallest p-values

- A p-value is a measure of how extreme the observation is when the null hypothesis is true
- E.g., observe $y \sim \mathcal{N}(\mu, 1)$ and decide between $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$
- We call a p-value

$$\begin{cases} \text{null} & \text{if } H_0 \text{ is true} \\ \text{non-null} & \text{if } H_0 \text{ is false} \end{cases}$$

The BH procedure

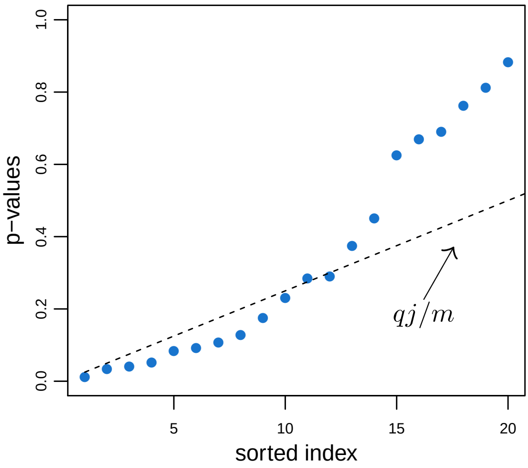
Let $p_1, p_2, \ldots, p_m$ be p-values of m hypotheses



- Sort $p_{(1)} \leq \cdots \leq p_{(m)}$

The BH procedure

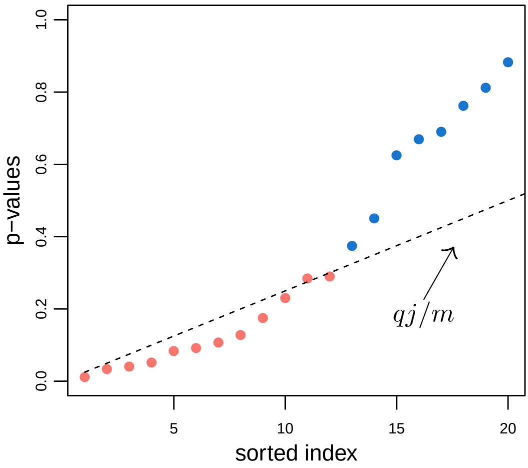
Let $p_1, p_2, \ldots, p_m$ be p-values of m hypotheses



- Sort $p_{(1)} \le \cdots \le p_{(m)}$
- Draw rank-dependent threshold qj/m

The BH procedure

Let $p_1, p_2, \ldots, p_m$ be p-values of m hypotheses



- Sort $p_{(1)} \le \cdots \le p_{(m)}$
- Draw rank-dependent threshold qj/m
- Reject hypotheses below cutoffs

FDR control

Theorem (Benjamini and Hochberg '95)

The BH procedure controls FDR if

- *the nulls are jointly independent,*
- *and the nulls are independent of the non-nulls*

- Recall that FDR controls means

$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{rejected null } p\text{-values}}{\#\text{rejected } p\text{-values}}\right] \leq q$$

- Replaced by "positive" dependence (Benjamini and Yekutieli '01)
- Arguably, conditions are very strigent for *provable* FDR control

Impact

- Perhaps the most popular error rate in genomics
- 49,443 citations as of October 29, 2018

Impact

- Perhaps the most popular error rate in genomics
- 49,443 citations as of October 29, 2018
- In summer 2014, two computer scientists became interested in FDR

Collaborators

Cynthia Dwork (Harvard)

Li Zhang (Google)

Summer 2014

I spent a wonderful summer at MSR Silicon Valley

What I was doing at MSR Silicon Valley

Prove FDR control of a differentially private version of BH

What I was doing at MSR Silicon Valley

Prove FDR control of a differentially private version of BH

Challenging because

smallest p-values may not be selected

- FDR proof techniques: martingale technique (Storey et al ’04) and “leave-one-out” technique (Benjamini and Yekutieli ’01)
- Existing approaches do not explore the *robustness*

Theory vs practice

- Provable FDR control rests on very stringent conditions
- In practice, works so well. Even very difficult to lose FDR control (Guo and Rao '08)

Theory vs practice

- Provable FDR control rests on very stringent conditions
- In practice, works so well. Even very difficult to lose FDR control (Guo and Rao '08)

Why?

This talk: it's the *robustness*, stxpid! (sorry ̈⌣)

- BH is a *robust* procedure
- FDR is a *robust* criterion

This talk: it's the *robustness*, stxpid! (sorry ☺)

- BH is a *robust* procedure
- FDR is a *robust* criterion

- Robust to even *adversary* dependence between nulls and non-nulls
- Null distribution matters most
- A new relaxed criterion: FDR consistency

Outline

An observation

Definition (Compliance)

A procedure is called compliant if any *rejected* p-value satisfies

$$p_i \le \frac{qR}{m}$$

- $R = $ #discoveries $=$ #rejected p-values

An observation

Definition (Compliance)

A procedure is called compliant if any *rejected* p-value satisfies

$$p_i \leq \frac{qR}{m}$$

- $R = \#\text{discoveries} = \#\text{rejected } p\text{-values}$

- Related to self-consistency condition (Blanchard and Roquain '08)
- BH is compliant
- So are the generalized step-up-step-down procedures (Tamhane, Liu, and Dunnett '98; Sarkar 02')

Compliances helps bound FDP

Compliance implies

$$\mathrm{FDP} \le \max_{1 \le j \le m_0} \frac{qj}{m p^0_{(j)}}$$

$p^0_{(1)} \le p^0_{(2)} \le \cdots \le p^0_{(m_0)}$ are the ordered m_0 null p-values

Compliances helps bound FDP

Compliance implies

$$\text{FDP} \le \max_{1 \le j \le m_0} \frac{qj}{m p^0_{(j)}}$$

$p^0_{(1)} \le p^0_{(2)} \le \cdots \le p^0_{(m_0)}$ are the ordered m_0 null p-values

Denote by V the number of false discoveries

- The largest rejected null p-value is at least $p^0_{(V)}$
- By compliance, $p^0_{(V)} \le \frac{qR}{m}$. Thus, $R \ge m p^0_{(V)}/q$
- Finally,

$$\text{FDP} = \frac{V}{R} \le \frac{V}{m p^0_{(V)}/q} \le \max_{1 \le j \le m_0} \frac{qj}{m p^0_{(j)}}$$

More comments

- Compliance implies

$$\text{FDP} \le \max_{1 \le j \le m_0} \frac{qj}{m p^0_{(j)}}$$

- Define $\text{FDR}_k = \mathbb{E}\left[\frac{V}{R}; V \ge k\right]$. Then

$$\text{FDP}_k \le \max_{k \le j \le m_0} \frac{qj}{m p^0_{(j)}}$$

- Hold *regardless* of the non-null p-values
- Non-null p-values can be *adversary* after looking at nulls!

What can compliance do for us?

Compliance plus IWN implies FDR control

Definition (IWN)

A set of p-values are said to satisfy *independence within the null* (IWN) if the null p-values are jointly independent

Compliance plus IWN implies FDR control

Definition (IWN)

A set of p-values are said to satisfy *independence within the null* (IWN) if the null p-values are jointly independent

Theorem (Dwork, S., and Zhang)

For $k \geq 2$, any compliant procedure applied to IWN p-values satisfies

$$\mathrm{FDR}_k \leq C_k q$$

- Applies to BH and many variants
- $C_2 \approx 2.41, C_3 \approx 1.85, C_{10} \approx 1.32$
- Dependence between nulls and non-nulls can be adversarial!
- Explains partially why BH is so robust

Optimality of C_k

Theorem (Dwork, S., and Zhang)

For any $C < C_k$, if q is sufficiently small and m is sufficiently large, there exists a compliant procedure applied to IWN p-values such that

$$\mathrm{FDR}_k > Cq$$

Connection with the literature

- State-of-art FDR control requires certain positive dependence between nulls and non-nulls (Benjamini and Yekutieli '01)
- Arbitrary dependence, FDR is controlled at (Benjamini and Yekutieli '01)

$$\left(1 + \frac{1}{2} + \cdots + \frac{1}{m}\right) q \approx (\log m) q$$

- Robustness in uniform FDP bounds (Katsevich and Ramdas '18)

Let's prove it

Proof I

Let $p_{i_1}, \ldots, p_{i_R}$ be rejected p-values

Compliance implies

$$\mathsf{FDP}_k \le \max_{k \le j \le m_0} \frac{qj}{m p^0_{(j)}}$$

- Replacing the ordered null p-values by the uniform order statistics $U_{(1)} \le U_{(2)} \le \cdots \le U_{(m_0)}$
- Then

$$\mathsf{FDR}_k \le \mathbb{E}\left[\max_{k \le j \le m_0} \frac{qj}{mU_{(j)}}\right] = q\frac{m_0}{m}\mathbb{E}\left[\max_{k \le j \le m_0} \frac{j}{m_0 U_{(j)}}\right]$$

Proof II

Thus, it suffices to prove

$$\mathbb{E}\left[\max_{k\le j\le n}\frac{j}{nU_{(j)}}\right]\le C_k$$

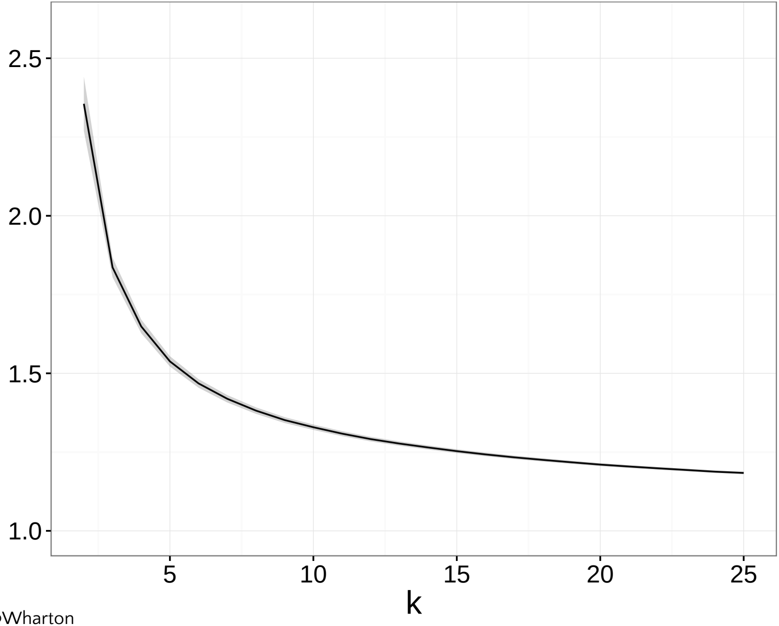
Lemma

Define for $n\ge k\ge 2$

$$C_k^{(n)}=\mathbb{E}\left[\max_{k\le j\le n}\frac{j}{nU_{(j)}}\right]$$

Then $C_k^{(n)}\le C_k^{(n+1)}$

The constant C_k



Controlling FDR^k

A variant of the FDR defined as

$$\mathrm{FDR}^k = \mathbb{E}\left[\frac{V}{R}; R \geq k\right]$$

Theorem (Dwork, S., and Zhang)

For any $k \geq 1$, any compliant procedure applied to IWN p-values satisfies

$$\mathrm{FDR}^k \leq \left(1 + \frac{2}{\sqrt{qk}}\right) q$$

- Proof based on a backward martingale

Numerical examples of FDR control of BH

Multivariate normal

$$X \sim \mathcal{N}(\mu, \Sigma)$$

- Σ of size 1000×1000; m_1 the number of true effects; $m_0 = 1000 - m_1$
- Σ has ones on the diagonal, $\Sigma(ij) = -1/\sqrt{m_0 m_1}$ for $1 \le i \le m_0$ and $m_1 + 1 \le j \le m$, otherwise 0
- $\mu = 2$ for $1 \le i \le m_1$, otherwise 0
- $q = 0.1$

Multivariate normal

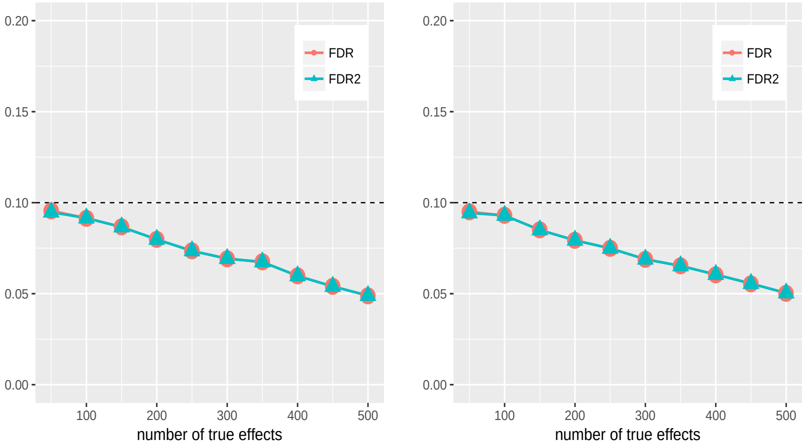
Multivariate t-distribution

$X^{(1)}, \ldots, X^{(n)} \sim \mathcal{N}(\mu, \Sigma)$. To test $\mu_i = 0$ vs $\mu_i > 0$, use

$$t_i = \frac{\sqrt{n}\bar{X}_i}{\sqrt{\frac{1}{n-1}\Sigma_{l=1}^{n}(X_i^{(l)} - \bar{X}_i)^2}}$$

- $n = 10$
- All the others the same as the previous example

Multivariate t-distribution



Outline

1. How does robustness arise?
2. From independence to PRDN
3. FDR consistency: the nulls matter

BH controls FDR under *IWN*

Positive regression dependence

Definition (Benjamini and Yekutieli ’01; Sarkar ’02)

$X = (X_1, \ldots, X_m)$ is said to satisfy the property of positive regression dependence on a subset I_0 (PRDS), if for any increasing set D and each $i \in I_0$

$$\mathbb{P}((X_1, \ldots, X_m) \in D | X_i = x)$$

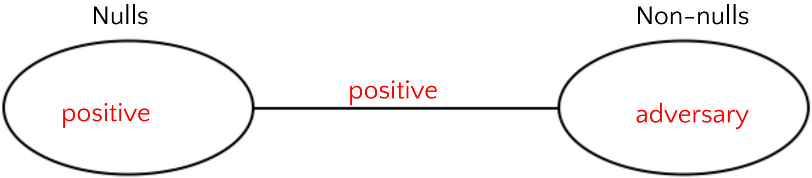
is increasing in x.

Theorem (Benjamini and Yekutieli ’01)

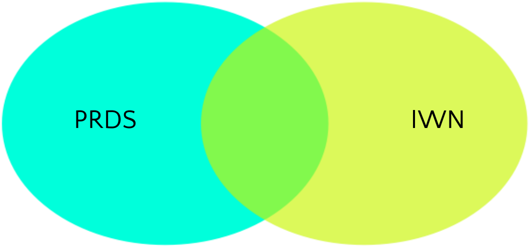
If the the test statistics are PRDS on the set of nulls, then BH gives

$$\text{FDR} \le \frac{q m_0}{m} \le q$$

BH controls FDR under *PRDS*



The current *provable* FDR control world



Can we find a new continent?

Recall compliance

Compliance implies

$$\mathrm{FDP} \le \min\left\{\max_{1\le j\le m_0} \frac{qj}{mp^0_{(j)}}, 1\right\}$$

$$\le \min\left\{\frac{qm_0/m}{\min_{1\le j\le m_0} \frac{m_0 p^0_{(j)}}{j}}, 1\right\}$$

$$\le \min\left\{\frac{q}{\min_{1\le j\le m_0} \frac{m_0 p^0_{(j)}}{j}}, 1\right\}$$

Recall compliance

Compliance implies

$$\mathsf{FDP} \le \min\left\{\max_{1\le j\le m_0} \frac{qj}{mp^0_{(j)}}, 1\right\}$$
$$\le \min\left\{\frac{qm_0/m}{\min_{1\le j\le m_0} \frac{m_0 p^0_{(j)}}{j}}, 1\right\}$$
$$\le \min\left\{\frac{q}{\min_{1\le j\le m_0} \frac{m_0 p^0_{(j)}}{j}}, 1\right\}$$

What's the distribution of

$$\min_{1\le j\le m_0} \frac{m_0 p^0_{(j)}}{j}?$$

A new dependence structure: PRDN

Definition (S.)

A set of p-values are said to satisfy the *positive regression dependence within nulls* (PRDN) if the nulls satisfy PRDS

A new dependence structure: PRDN

Definition (S.)

A set of p-values are said to satisfy the *positive regression dependence within nulls* (PRDN) if the nulls satisfy PRDS

- Includes PRDS and IWN as special cases
- No assumption regarding the non-nulls
- Under PRDN, one can show that

$$\min_{1 \le j \le m_0} \frac{m_0 p^0_{(j)}}{j}$$

  is stochastically larger than or equal to $U[0,1]$
- Connection with the Simes method

FDR control under PRDN

Theorem (S.)

Any compliant procedure applied to PRDN p-values satisfies

$$\text{FDR} \le q + q \log \frac{1}{q}$$

FDR control under PRDN

Theorem (S.)

Any compliant procedure applied to PRDN p-values satisfies

$$\text{FDR} \le q + q \log \frac{1}{q}$$

$$\begin{aligned}
\text{FDR} &\le \mathbb{E}\left[\min\left\{\frac{q}{\min_{1 \le j \le m_0} \frac{m_0 p^0_{(j)}}{j}}, 1\right\}\right] \\
&\le \mathbb{E}\left[\min\left\{\frac{q}{U[0,1]}, 1\right\}\right] \\
&= \mathbb{P}(U[0,1] \le q) + \int_q^1 \frac{q}{x} \, \mathrm{d}x \\
&= q + q \log \frac{1}{q}
\end{aligned}$$

Optimality

Theorem (S.)

Let $c < q + q \log \frac{1}{q}$ for sufficiently small q. If m is sufficiently large, BH applied to certain PRDN p-values gives

$$\text{FDR} > c$$

Possible to get rid of the logarithmic factor $\log(1/q)$?

Bounded adversariness

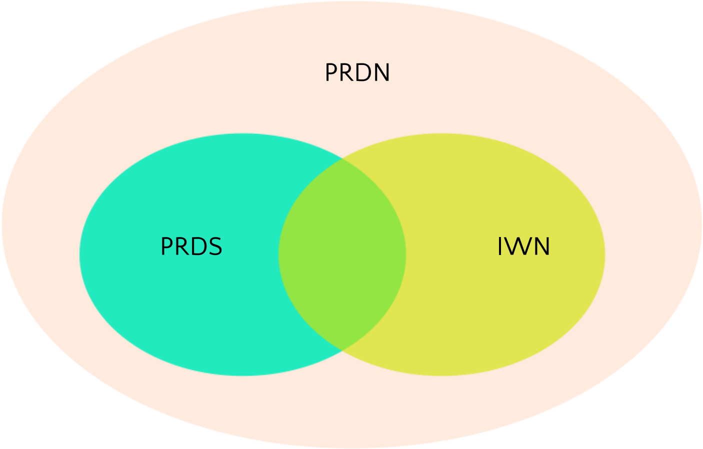
Theorem (S.)

If the null p-values are iid uniform and the adversary only has access to all (sorted) p-values but the smallest one. Then any compliant procedure satisfies

$$\text{FDR} \le 3.41q$$

FDR control under *PRDN*

The *new* provable FDR control world



Outline

This rate is “consistent”

An observation

$$\lim_{q \to 0} \; q + q \log \frac{1}{q} = 0$$

independent of the dimension m

This rate is "consistent"

An observation

$$\lim_{q \to 0} \quad q + q \log \frac{1}{q} = 0$$

independent of the dimension m

- But the rate

$$\left(1 + \frac{1}{2} + \cdots + \frac{1}{m}\right) q \approx (\log m) q$$

does *not* tend to zero uniformly

A weak version of FDR control

Definition (FDR consistency)

A dependence structure (indexed by the dimension m) of p-values is said to be FDR-consistent if the FDR of BH satisfies

$$\text{FDR} \le f(q),$$

where $f(q) \to 0$ as $q \to 0$ uniformly over all m

A weak version of FDR control

Definition (FDR consistency)

A dependence structure (indexed by the dimension m) of p-values is said to be FDR-consistent if the FDR of BH satisfies

$$\text{FDR} \le f(q),$$

where $f(q) \to 0$ as $q \to 0$ uniformly over all m

- If dependence of nulls is "positive," then $f(q) = q + q\log(1/q)$ is FDR-consistent
- For the most *adversary* dependence, $f(q) = (1 + 1/2 + \cdots + 1/m)q$. FDR consistency not satisfied (Benjamini and Yekutieli '01)!

It's the nulls that matter for FDR consistency

Theorem (S.)

If the null dependence structure is FDR-consistent, then the (full) dependence structure is FDR-consistent

It's the nulls that matter for FDR consistency

Theorem (S.)

If the null dependence structure is FDR-consistent, then the (full) dependence structure is FDR-consistent

- FDR consistency is robust to adversary non-nulls
- Future theoretical FDR research: focus on the nulls!

Proof

Lemma (S.)

Let a compliant procedure applied to the nulls control the FDR at $\mathsf{FDR}_0(q)$. *Then, the procedure applied to all* p*-values satisfies*

$$\mathsf{FDR} \le q + q \int_q^1 \frac{\mathsf{FDR}_0(x)}{x^2} \mathrm{d}x$$

Proof

Lemma (S.)

Let a compliant procedure applied to the nulls control the FDR at $\mathrm{FDR}_0(q)$*. Then, the procedure applied to all* p*-values satisfies*

$$\mathrm{FDR} \le q + q \int_q^1 \frac{\mathrm{FDR}_0(x)}{x^2} \mathrm{d}x$$

- Step 1:

$$\mathrm{FDP} \le \min \left\{ \frac{q}{\min_{1 \le j \le m_0} \frac{m_0 p^0_{(j)}}{j}}, 1 \right\}$$

- Step 2: the CDF of $\min_{1 \le j \le m_0} \frac{m_0 p^0_{(j)}}{j}$ is $\le \mathrm{FDR}_0(q)$
- Step 3: $q + q \int_q^1 \frac{f(x)}{x^2} \mathrm{d}x \to 0$ if $f(x) \to 0$ as $x \to 0$

Extending the provable FDR *consistent* world?

Summary

Take-home messages

- Both FDR and BH are robust to adversary dependence between nulls and non-nulls

Take-home messages

- Both FDR and BH are robust to adversary dependence between nulls and non-nulls
- The joint distribution of nulls matters most

Take-home messages

- Both FDR and BH are robust to adversary dependence between nulls and non-nulls
- The joint distribution of nulls matters most
- If proving FDR control is too difficult, let's consider FDR consistency under global null!

Thank you!

1. *Private False Discovery Rate Control*
   Cynthia Dwork, Weijie J. Su, and Li Zhang, arXiv:1511.03803 (subsumed)

2. *Differentially Private False Discovery Rate Control*
   Cynthia Dwork, Weijie J. Su, and Li Zhang, arXiv:1807.04209

3. *The FDR-Linking Theorem*
   Weijie J. Su, in preparation