# A direct approach to estimating false discovery rates conditional on covariates

Simina Boca[1], Jeff Leek[2]

[1]Georgetown University Medical Center, [2]Johns Hopkins Bloomberg School of Public Health

Joint Statistical Meetings, Baltimore, MD

July 31, 2017

# Goal

Extend false discovery rate framework to easily incorporate external covariates (meta-data) in decision of whether to reject a hypothesis when testing many hypotheses at once.

# Background

Multiple testing is a ubiquitous issue in modern science:

Need to test relationship between hundreds or thousands of variables/features and one outcome (genomics, metabolomics etc.)

| | Fail to reject null | Reject null | Total |
|---|---|---|---|
| Null true | $U$ | $V$ | $m_0$ |
| Null false | $T$ | $S$ | $m - m_0$ |
| | $m - R$ | $R$ | $m$ |

$R$ = number of discoveries
$V$ = number of false discoveries

Benjamini and Hochberg, 1995, *JRSSB*.

# Background

False discovery rate (FDR) often used as framework to control for multiple testing.

Natural definition:

$$FDR = E\left[\frac{V}{R}\right].$$

Since $R$ can be equal to 0, usually defined as:

$$FDR = E\left[\frac{V}{R}\middle| R > 0\right] Pr(R > 0).$$

Benjamini and Hochberg, 1995, *JRSSB*.

# Background

FDR control and estimation approaches rely on an estimate of the proportion of null hypotheses, $\pi_0$:

$$\pi_0 = P(\text{hypothesis } i \text{ is null}).$$

- Original Benjamini-Hochberg (BH) approach assumes $\pi_0 \equiv 1$
- Storey, 2002, *JRSSB* estimates $\pi_0$ as a fixed value and multiplies the BH FDR estimates by $\hat{\pi}_0$.
- We expand this framework to estimate $\pi_0$ as a function of external covariates.

# Motivating case study

- Genome-wide association (GWAS) study looking at associations between millions of genetic loci and BMI (Locke et al, 2015, *Nature*).
- Loci are single nucleotide polymorphisms (SNPs) that usually have 2 possible variants (alleles).
  - Major allele - more common allele, minor allele - less common allele.
- Each SNP has a different population-level frequency (coded as MAF = minor allele frequency).
- Not all SNPs have the same sample size (N), since they may be genotyped in different individuals.

Plan to use MAF and N as external covariates!

# Building on prior work

Our work builds on the work of Benjamini and Hochberg, Efron, Storey, Scott et al, 2015, *JASA*, who framed the concept of FDR regression, extending FDR and $\pi_0$ to include covariates.

We focus on estimating $\pi_0$ as a function of covariates, then using it as a plug-in estimator to estimate FDR as a function of covariates, à la Storey.

We also use ideas from Ignatiadis et al, 2016, *Nat. Methods* that adjusting for covariates independent of the data - conditional on the null being true - can improve power.

# Approach: Extend definitions of $\pi_0$ and FDR to include covariates

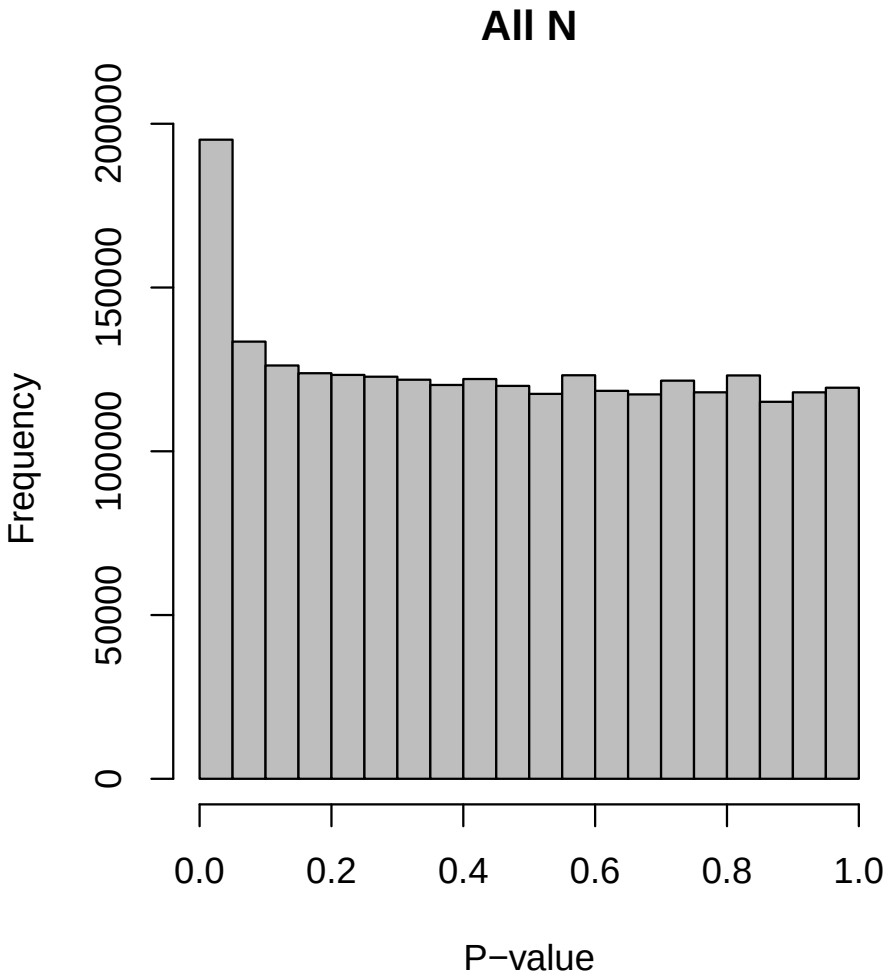
Assume a set of covariates in a column vector $\mathbf{X}_i$ of length $c$, possibly with $c = 1$ and extend definitions:

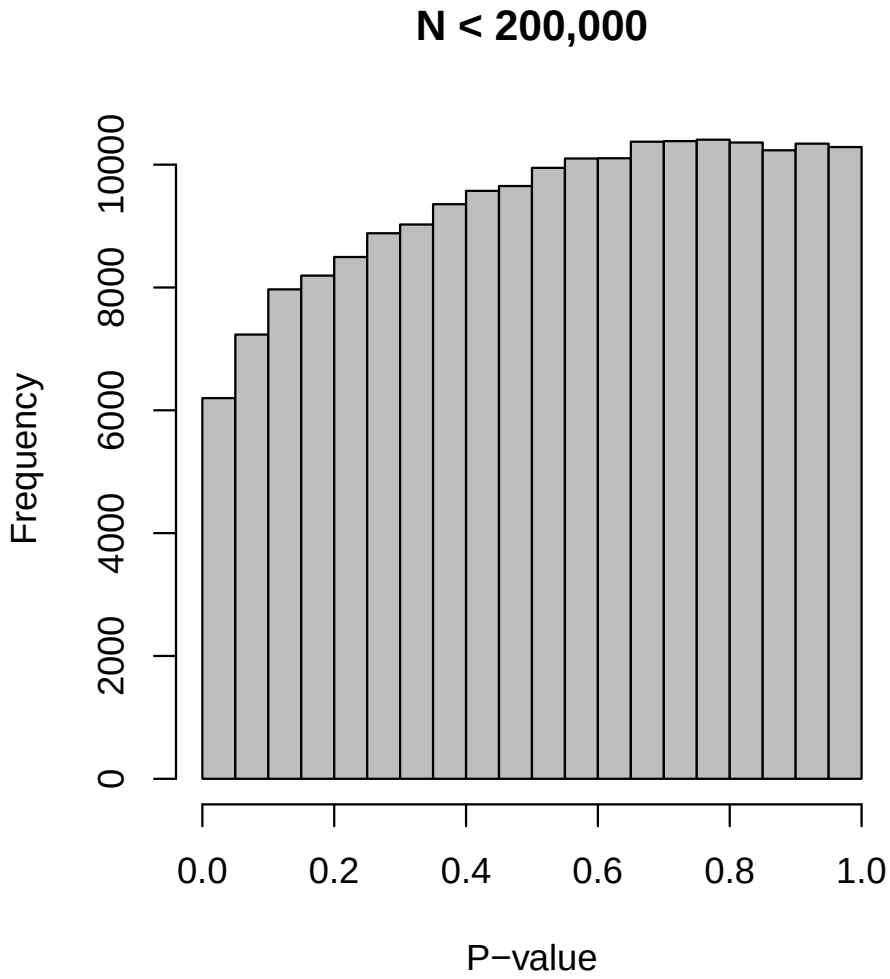
$$\begin{aligned} \pi_0(\mathbf{x}_i) &= Pr(\theta_i = 1 | \mathbf{X}_i = \mathbf{x}_i), \\ FDR(\mathbf{x}_i) &= E\left[\frac{V}{R}\middle| R > 0, \mathbf{X}_i = \mathbf{x}_i\right] Pr(R > 0 | \mathbf{X}_i = \mathbf{x}_i). \end{aligned}$$

# Motivating case study: GWAS meta-analysis for BMI

- Looked at $\sim$2.5 million SNPs in $\sim$340,000 individuals, considering their association with BMI.
- $\sim$320,000 were of European descent
  - Used HapMap CEU population as reference for MAF
- Sample size (N) varies between SNPs (50,002 - 339,224, median = 235,717)

# Dependence of p-values on sample sizes



# Approach: Estimate $\pi_0(\mathbf{x}_i)$ using logistic regression

Assume that null p-values come from Uniform(0,1) and the alternative p-values from a distribution with cdf G, so that for a large enough $\lambda \in (0,1)$, $G(\lambda) \approx 1$.

Define:
$Y_i = 1(\text{P-value from test } i > \lambda) = P(\text{test } i \text{ not rejected at level } \lambda)$.

For a fixed threshold $\lambda$:

$$\pi_0(\mathbf{x}_i) \approx \frac{E[Y_i|\mathbf{X}_i = \mathbf{x}_i]}{1-\lambda}.$$

Use a regression framework to estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}_i]$, then estimate $\pi_0(\mathbf{x})$ by:

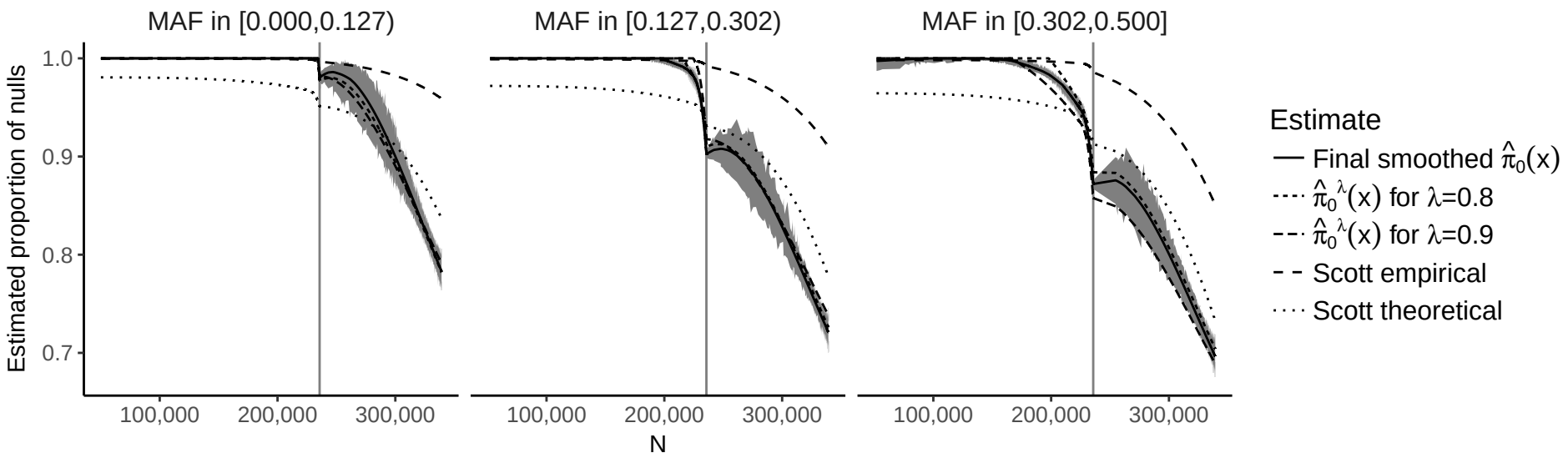
$$\hat{\pi}_0(\mathbf{x}_i) = \frac{\hat{E}[Y_i|\mathbf{X}_i = \mathbf{x}_i]}{1-\lambda}.$$

# Approach: Estimate $\pi_0(\mathbf{x}_i)$ using logistic regression

Additional details:

- May need to threshold at 1, given division by $1 - \lambda$.
- Can fix $\lambda$ at e.g. 0.8 or 0.9 *or* smooth over a series of thresholds $\lambda \in (0, 1)$.
- Can use a bootstrap approach for confidence intervals.
- For BMI GWAS, split MAFs into 3 categories and used cubic splines with 5 degrees of freedom for N.

# Estimates of $\pi_0(\mathbf{x}_i)$ for BMI GWAS




Grey = 90% bootstrap CI
Vertical line = median sample size

# Approach: Use plug-in estimate of $\pi_0(\mathbf{x}_i)$ for FDR$(\mathbf{x}_i)$

- Multiply estimate of FDR obtained via BH approach by our estimate of $\pi_0(\mathbf{x}_i)$
- Note that is the same as approach of Storey, but considers extension to covariates
- Major assumption: Conditional on the null and the alternative, the p-values do not depend on the covariates

  i.e. the probability of a feature being from one of the two distributions depends on the covariates but the actual test statistic and p-value does not depend on the covariates further.

# Results for estimated FDR($\mathbf{x}_i$) for BMI GWAS

Number of SNPs with an estimated FDR $\leq$ 5% for various approaches.

| | BL | Scott T | Scott E | Storey | BH |
|---|---|---|---|---|---|
| Number with $\widehat{FDR} \leq 5\%$ | 13,384 | 16,697 | 7,636 | 12,771 | 12,500 |

BL = our approach
Scott T = Scott et al approach with theoretical null
Scott E = Scott et al approach with empirical null

All the discoveries from the BH approach are also present in BL.
Overlap with Storey = 12,740.

# Simulations to check FDR control and power

Extensive simulations are presented in our paper/Github page.

Power = TPR (true positive rate) = fraction of truly alternative discoveries out of the total number of truly alternative features.

**The good:**

- If there is no or low correlation between test statistics, generally shows good control of FDR
- Always leads to an improvement in power over Benjamini-Hochberg, which increases with lower $\pi_0(\mathbf{x}_i)$ (max 6%-11% in absolute terms)
- Improved interpretability compared to Storey's approach
- FDR control is much better compared to Scott FDR regression approach when test statistics are not from a normal distribution

# Simulations to check FDR control and power

Extensive simulations are presented in our paper/Github page.

Power = TPR (true positive rate) = fraction of truly alternative discoveries out of the total number of truly alternative features.

**The caveats:**

- If test statistics are highly correlated, does not appropriately control the FDR
- If $\pi_0(\mathbf{x}_i)$ is high, not much gain in power over Benjamini-Hochberg
- Gain in power over Storey's approach is usually minimal (0-2%)
- Power is better for Scott FDR regression approach when test statistics are from a normal distribution

# Conclusions

We developed a direct approach to estimating FDR conditional on covariates.

Why should you use it?

- Improved power compared to BH
- Improved interpretability compared to Storey
- Improved robustness compared to Scott

We hope to extend/apply this approach to a number of other scenarios.

# Questions?

Email: smb310@georgetown.edu

Twitter: @siminaboca

Preprint:
`http://www.biorxiv.org/content/early/2017/07/25/035675`

Code for all analyses/simulations in paper:
`https://github.com/SiminaB/Fdr-regression`

Package which includes FDR regression approach: `https://bioconductor.org/packages/release/bioc/html/swfdr.html`
(uses linear, not logistic regression)