Using the swfdr package to estimate false discovery rates conditional on covariates

Simina M. Boca[1], Tomasz Konopka[2], Leah Jager[3], Jeffrey T. Leek[3]

[1]Georgetown University Medical Center, [2]Queen Mary University of London, [3]Johns Hopkins Bloomberg School of Public Health

BioC2020

July 27, 2020

Problem

Multiple testing is a ubiquitous issue in modern science:

Need to test relationship between hundreds or thousands of variables/features and one outcome:

- Is the expression of each of these 20,000 genes associated with cancer survival?
- Is each of these 2.5 million SNPs associated with BMI?
- Is each of these 2,000 metabolites associated with disease status?

Why do we need more methods?

We already have the:

- Bonferroni approach to control the family-wise error rate (FWER),
- Benjamini-Hochberg (BH) approach to control the false discovery rate (FDR),
- Storey (q-value) approach to estimate the FDR.

Why do we need more methods?

We already have the:

- Bonferroni approach to control the family-wise error rate (FWER),
- Benjamini-Hochberg (BH) approach to control the false discovery rate (FDR),
- Storey (q-value) approach to estimate the FDR.

https://www.youtube.com/watch?v=j9Z_f3L56iY&t=1m19s
https://www.youtube.com/watch?v=j9Z_f3L56iY&t=2m58s

Why do we need more methods?

We already have the:

- Bonferroni approach to control the family-wise error rate (FWER),
- Benjamini-Hochberg (BH) approach to control the false discovery rate (FDR),
- Storey (q-value) approach to estimate the FDR.

MORE POWER!!!

Why do we need more methods?

- We already have the Bonferroni approach to control the family-wise error rate (FWER), the Benjamini-Hochberg approach to control the false discovery rate (FDR), and the Storey (q-value) approach to estimate the FDR.
- The FDR depends on the overall fraction of null hypotheses (variables/features not associated with the outcome), often denoted by π_0.
- Adaptive FDR procedures, such as q-values, can improve power by including an estimate of π_0 based on the distribution of the p-values (magic of Empirical Bayes!)

Why do we need more methods?

What if we have other data we can use in our estimates, besides the p-values themselves?

- Often have external covariates (meta-data, co-data, feature-level covariates), which can be incorporated into an adaptive procedure and help with the decision of whether to reject a hypothesis.
- Examples of these covariates:
  - Minor allele frequency (MAF) and sample size for SNPs in genome-wide association studies (GWAS)
  - Set size in gene set analyses
  - Mean nonzero gene expression and detection rate in single-cell RNA-seq

Korthauer et al, 2019, *Genome Biology*

Using a regression approach to incorporate covariates

- Our approach uses a regression framework for estimating π_0 as a function of the external covariates $\mathbf{x}$, so that we consider $\pi_0(\mathbf{x})$ and $\text{FDR}(\mathbf{x})$.
- After obtaining $\hat{\pi}_0(\mathbf{x})$, we simply use a plug-in estimator for $\widehat{\text{FDR}}(\mathbf{x})$, multiplying $\hat{\pi}_0(\mathbf{x})$ by the BH-transformed p-values.
  - This is essentially equivalent to the Storey q-values if there are no covariates.

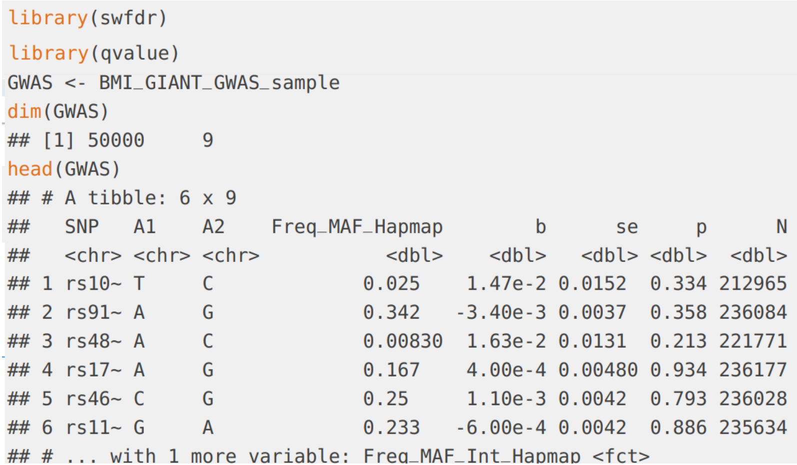
Boca SM, Leek JT. “A direct approach to estimating false discovery rates conditional on covariates.” *PeerJ*, 2018, 6:e6035. [link at *PeerJ*]
`https://www.bioconductor.org/packages/release/bioc/html/swfdr.html`
`https://github.com/leekgroup/swfdr`

GWAS example

- We consider a meta-analysis from a GWAS for BMI (Locke et al, 2015, *Nature*).
  - Meta-analysis of 339,224 individuals (322,154 of European origin) measuring 2,555,510 SNPs.
  - Different SNPs are genotyped in different individuals, leading to a different sample size per SNP.
  - Minor allele frequencies (MAF) — frequencies of least common allele for each SNP — also vary per SNP.
- The `swfdr` package includes a subset of results for the individuals of European origin for a subset of 50,000 random SNPs.

GWAS example in `swfdr` package

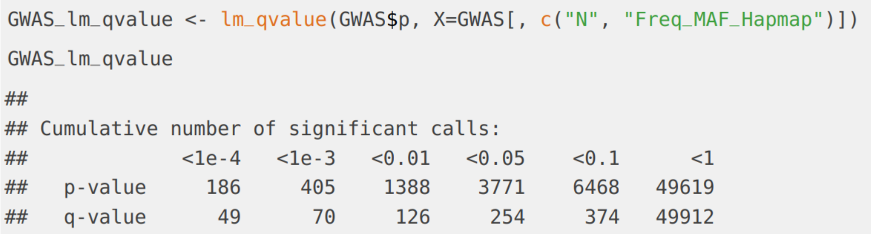
- First load and explore the dataset:



```
library(swfdr)
library(qvalue)
GWAS <- BMI_GIANT_GWAS_sample
dim(GWAS)
## [1] 50000     9
head(GWAS)
## # A tibble: 6 x 9
##   SNP   A1    A2    Freq_MAF_Hapmap        b     se     p      N
##   <chr> <chr> <chr>           <dbl>    <dbl>  <dbl> <dbl>  <dbl>
## 1 rs10~ T     C             0.025    1.47e-2 0.0152  0.334 212965
## 2 rs91~ A     G             0.342   -3.40e-3 0.0037  0.358 236084
## 3 rs48~ A     C             0.00830  1.63e-2 0.0131  0.213 221771
## 4 rs17~ A     G             0.167    4.00e-4 0.00480 0.934 236177
## 5 rs46~ C     G             0.25     1.10e-3 0.0042  0.793 236028
## 6 rs11~ G     A             0.233   -6.00e-4 0.0042  0.886 235634
## # ... with 1 more variable: Freq_MAF_Int_Hapmap <fct>
```

GWAS example in `swfdr` package

- After loading the dataset, use the `lm_qvalue` function, based on the `qvalue` function in the `qvalue` package:



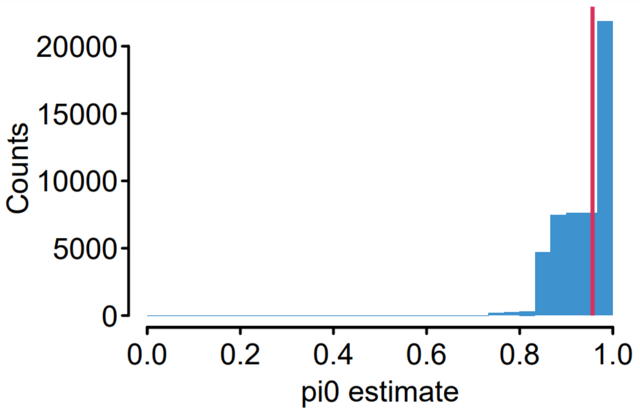
```
GWAS_lm_qvalue <- lm_qvalue(GWAS$p, X=GWAS[, c("N", "Freq_MAF_Hapmap")])

GWAS_lm_qvalue

##
## Cumulative number of significant calls:
##               <1e-4   <1e-3   <0.01   <0.05    <0.1      <1
##    p-value      186     405    1388    3771    6468   49619
##    q-value       49      70     126     254     374   49912
```

GWAS example in `swfdr` package

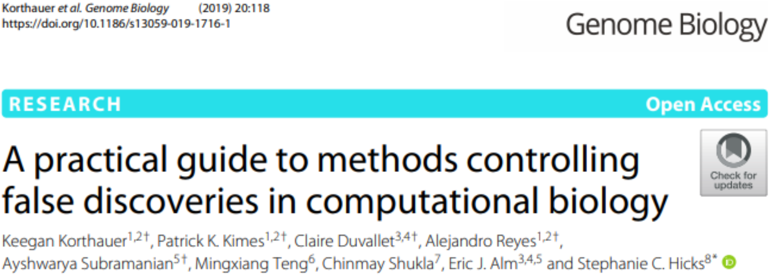
- Can compare the estimates $\hat{\pi}_0(\mathbf{x})$ to the estimate one would obtain without conditioning (vertical line):



Other new developments, including plots, can be found at `https://github.com/leekgroup/swfdr/tree/dev`.

The combined powers of open-access and open-source

- Korthauer et al wrote a paper where they compared methods that controlled false discovery rates adjusting for covariates:



Korthauer *et al. Genome Biology* (2019) 20:118
https://doi.org/10.1186/s13059-019-1716-1

Genome Biology

RESEARCH Open Access

A practical guide to methods controlling false discoveries in computational biology

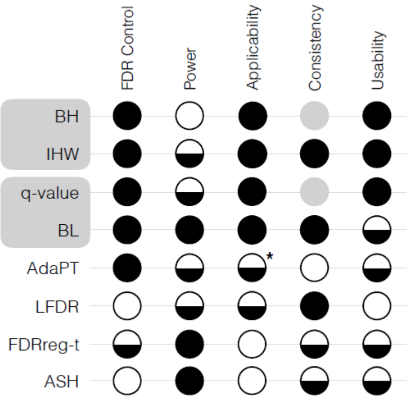
Keegan Korthauer[1,2†], Patrick K. Kimes[1,2†], Claire Duvallet[3,4†], Alejandro Reyes[1,2†], Ayshwarya Subramanian[5†], Mingxiang Teng[6], Chinmay Shukla[7], Eric J. Alm[3,4,5] and Stephanie C. Hicks[8*]

The combined powers of open-access and open-source

- Korthauer et al wrote a paper where they compared methods that controlled false discovery rates adjusting for covariates.
- They were able to compare 8 (eight!) methods in terms of FDR control, power, and software usability, on a number of simulated and real examples, using a univariate covariate.
- Our paper was in the process of “flunking out” of a series of journals, but we had a preprint out (initial version from December 2015!) and had already added the method to the `swfdr` package on Bioconductor.

The combined powers of open-access and open-source

- We got a good “score” in Korthauer et al, despite our paper not being yet published:



	FDR Control	Power	Applicability	Consistency	Usability
BH	●	○	●	◌	●
IHW	●	◒	●	●	●
q-value	●	◒	●	◌	●
BL	●	●	●	●	◒
AdaPT	●	◒	◒*	○	◒
LFDR	○	◒	◒	●	○
FDRreg-t	◒	●	○	◒	◒
ASH	○	●	○	◒	◒

The combined powers of open-access and open-source

- Tomasz Konopka, from the UK, read Korthauer et al, then read our preprint, and offered to help us out with improving the usability aspect.
- He's now one of the main developers for the `swfdr` package, having written the `lm_qvalue` function, among other developments.

Questions?

Email: smb310@georgetown.edu

Twitter: @siminaboca