

Data-Driven Hypothesis Weighting Increases Detection Power In Genome-Scale Multiple Testing



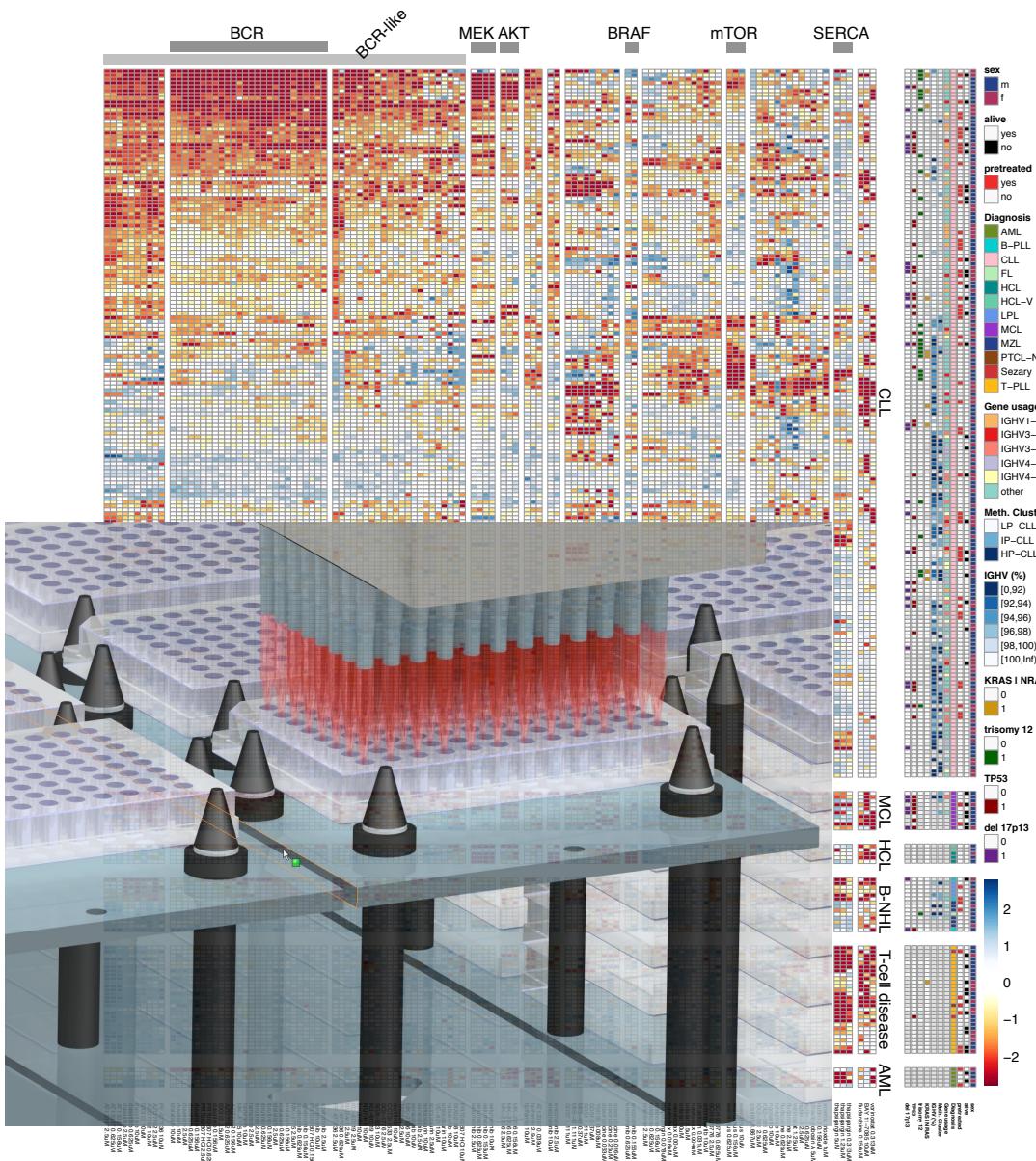
Wolfgang Huber

Research Group Leader, Senior Scientist
European Molecular Biology Laboratory (EMBL)

Multiple Testing

Many data analysis approaches in genomics employ item-by-item testing:

- Expression profiling
- ChIP-Seq
- Genetic or chemical compound screens
- Genome-wide association studies
- Proteomics
- Variant calling



The Multiple Testing Burden

When performing several tests, type I error goes up: for $\alpha = 0.05$ and n independent tests, probability of no false positive result is

$$\underbrace{0.95 \cdot 0.95 \cdot \dots \cdot 0.95}_{n\text{-times}} \lll 0.95$$



The Multiple Testing Opportunity

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)

THIS NEUTRINO DETECTOR MEASURES
WHETHER THE SUN HAS GONE NOVA.

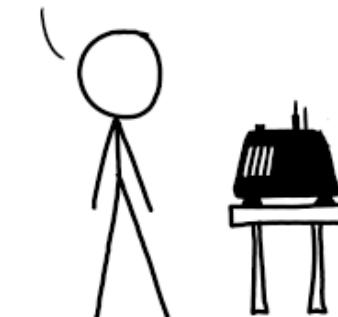
LET'S TRY.
DETECTOR! HAS THE
SUN GONE NOVA?

THEN, IT ROLLS TWO DICE. IF THEY
BOTH COME UP SIX, IT LIES TO US.
OTHERWISE, IT TELLS THE TRUTH.



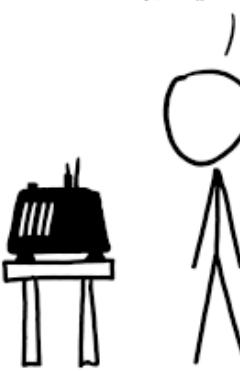
FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT
HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE
THAT THE SUN HAS EXPLODED.

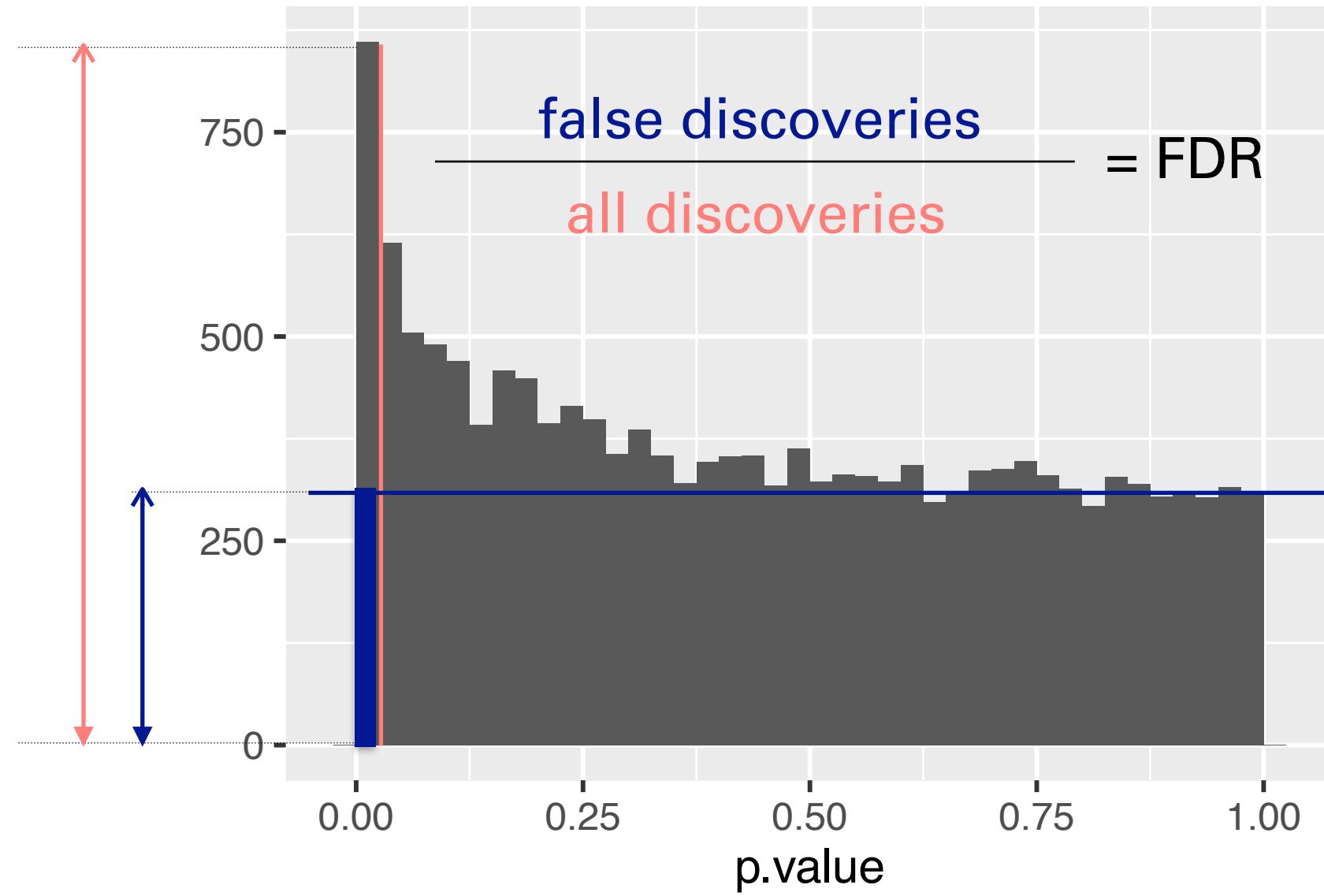


BAYESIAN STATISTICIAN:

BET YOU \$50
IT HASN'T.

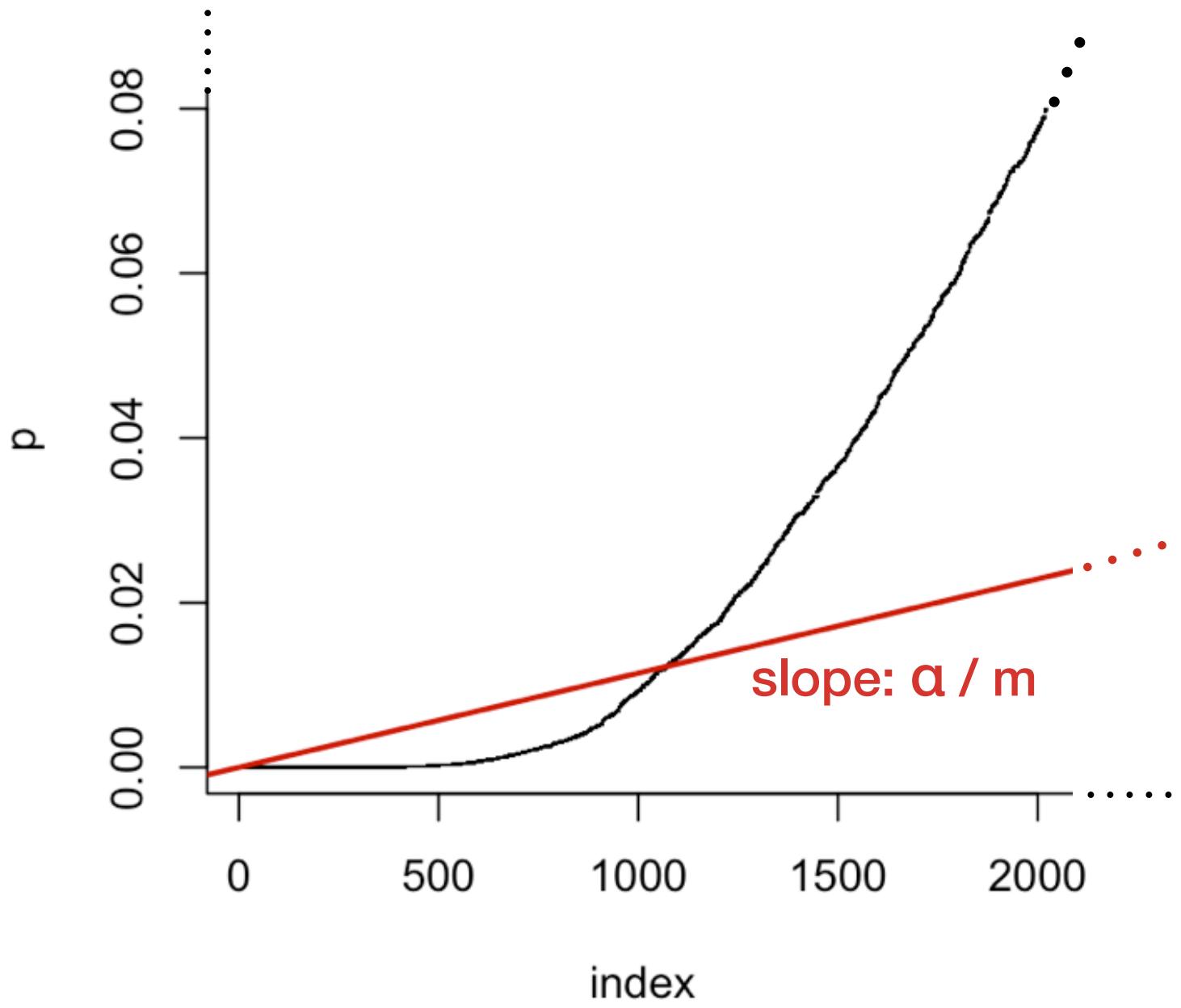


False discovery rate



Method of Benjamini & Hochberg (1995)

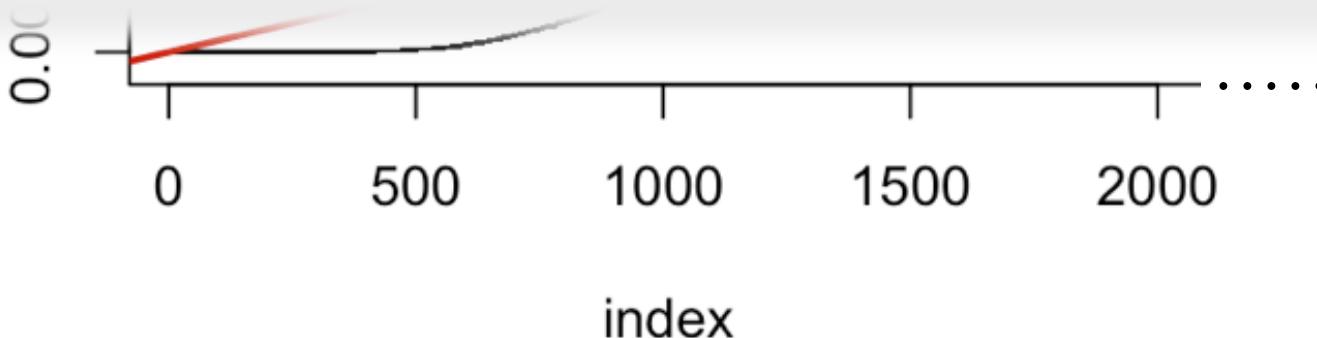
Method of Benjamini & Hochberg



Method of Benjamini & Hochberg

```
BH = {  
    i <- length(p) :1  
    o <- order(p, decreasing = TRUE)  
    ro <- order(o)  
    pmin(1, cummin(n/i * p[o])) [ro]  
}
```

takes a list of p-values as input and returns a matched list of ‘adjusted’ p-values.



A photograph of a vast colony of Emperor penguins gathered on a white, snow-covered ground. In the background, a massive, rugged iceberg dominates the scene, its surface textured with deep blue shadows and bright white highlights from the sun. The penguins are scattered across the frame, some facing towards the camera while others are seen from behind, creating a sense of a large, active community.

Exchangeability?

Covariates - examples

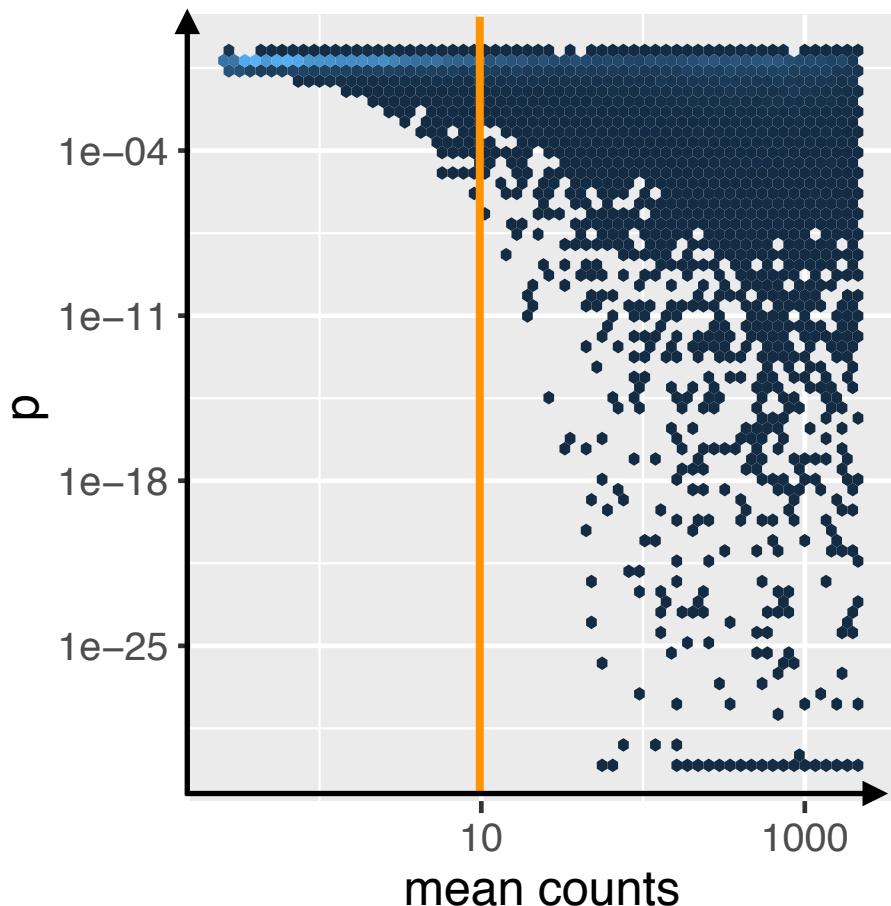
Application	Covariate
Differential RNA-Seq, ChIP-Seq, CLIP-seq, ...	(Normalized) mean of counts for each gene
GWAS	Minor allele frequency
eQTL analysis	SNP – gene distance
<i>t</i> -tests	Overall variance
Two-sided tests	Sign
All applications	Sample size; measures of signal-to-noise ratio

Independent Filtering

Two steps:

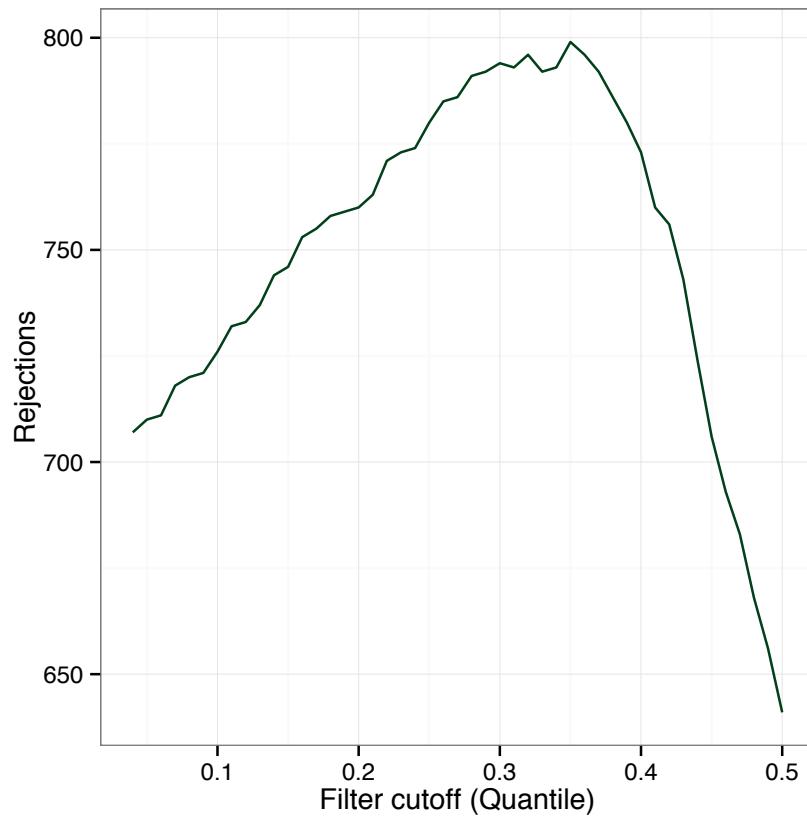
- All hypotheses H_i with $X_i < x$ get filtered.
- Apply BH to remaining hypotheses.

(Bourgon, Gentleman, Huber
PNAS 2010)



Data-driven choice of filtering threshold

- Do Independent Filtering followed by Benjamini-Hochberg procedure with all possible thresholds.
- Report the result with the optimal threshold.
- We have been doing this in *DESeq2* for the last two years.



Weighted Benjamini-Hochberg method

- Let $w_i \geq 0$ and $\frac{1}{m} \sum_{i=1}^m w_i = 1$ (“weight budget”).
- Define $Q_i = P_i/w_i$.
- Apply BH to Q_i instead of P_i .
- Proven Type-I error (FDR) control (Genovese, Roeder, Wasserman *Biometrika* 2006).
- If $w_i > 1$, then H_i is easier to reject.
- $Q_i \leq t \Leftrightarrow P_i \leq w_i t =: t_i$

Weighted Benjamini-Hochberg method

- Let $w_i \geq 0$ and $\frac{1}{m} \sum_{i=1}^m w_i = 1$ (“weight budget”).
- Define $Q_i = P_i/w_i$.
- Apply BH to Q_i instead of P_i .
- Proven Type I error (FDR) control (Genovese, Roeder, Wasserman, 2006)
- If $w_i > 1$, then $Q_i = 1$.
- $Q_i \leq t \Leftrightarrow P_i \leq t w_i$.



Weighted Benjamini-Hochberg method

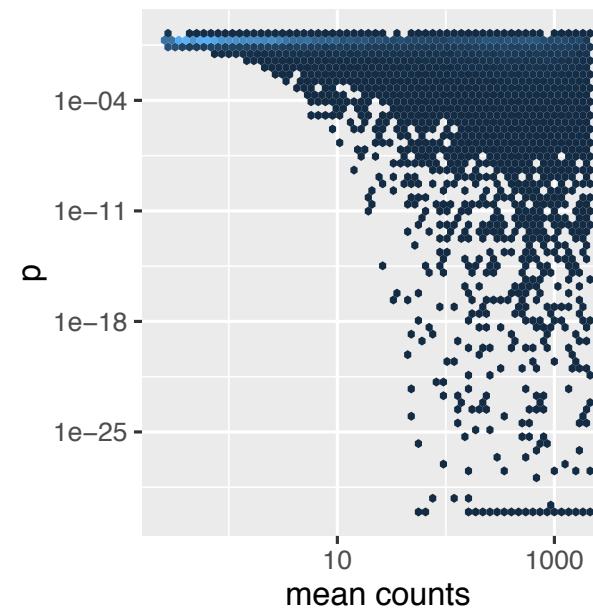
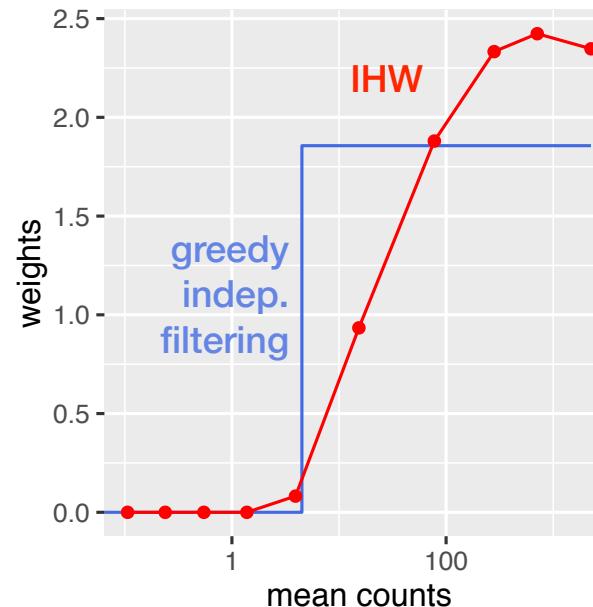
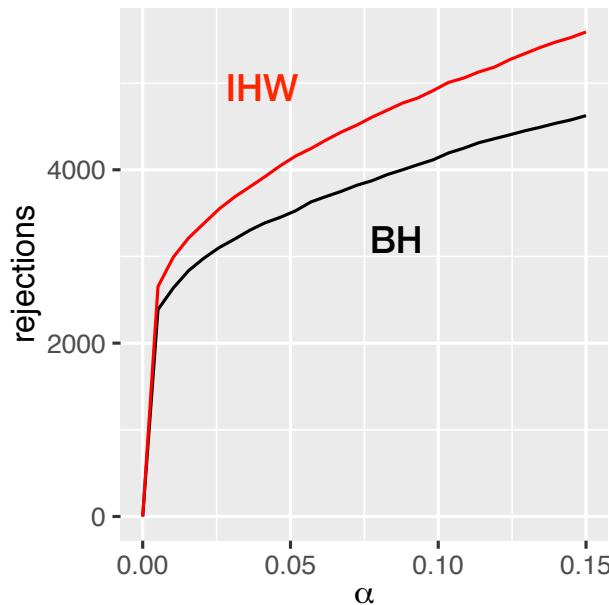
- Let $w_i \geq 0$ and $P_{i,i}$ be the observed p-value for hypothesis H_i .
 - Define $Q_i = \min\{Q_j : P_{j,j} \leq P_{i,i}\}$, where $P_{j,j}$ is the observed p-value for hypothesis H_j .
 - If $w_i Q_i \leq t$ then accept H_i .
- Problem:** how to know the weights?



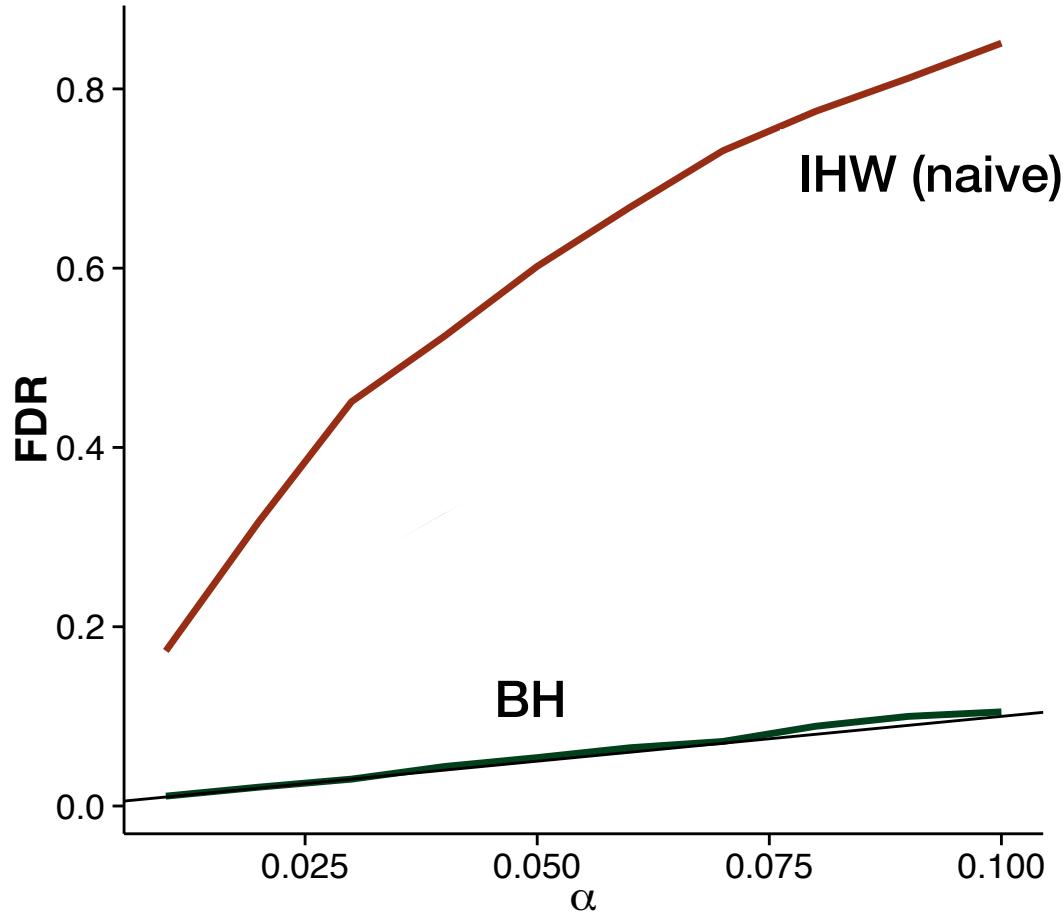
IHW (naive): Independent (data-driven) hypothesis weighting

- Stratify the tests into G bins, by covariate X
- Choose α
- For each possible weight vector $\mathbf{w} = (w_1, \dots, w_G)$ apply weighted BH procedure. Choose \mathbf{w} that maximizes the number of rejections at level α .
- Report the result with the optimal weight vector \mathbf{w}^* .

RNA-Seq example (DESeq2)



But naive IHW does not always control the FDR (e.g. $\pi_0 = 1$)



Modified IHW



Nikos Ignatiadis

Data splitting: randomly split hypotheses into k folds. Learn weights for the hypotheses in a fold from the other $k-1$ folds

Regularisation:

- for ordered covariate: $\sum_g |w_g - w_{g-1}| \leq \lambda$
- for categorical covariate: $\sum_g |w_g - 1| \leq \lambda$

Convex relaxation: for weight optimisation (only), replace ECDFs of the p-values with Grenander estimators (least concave majorant of the ECDF)

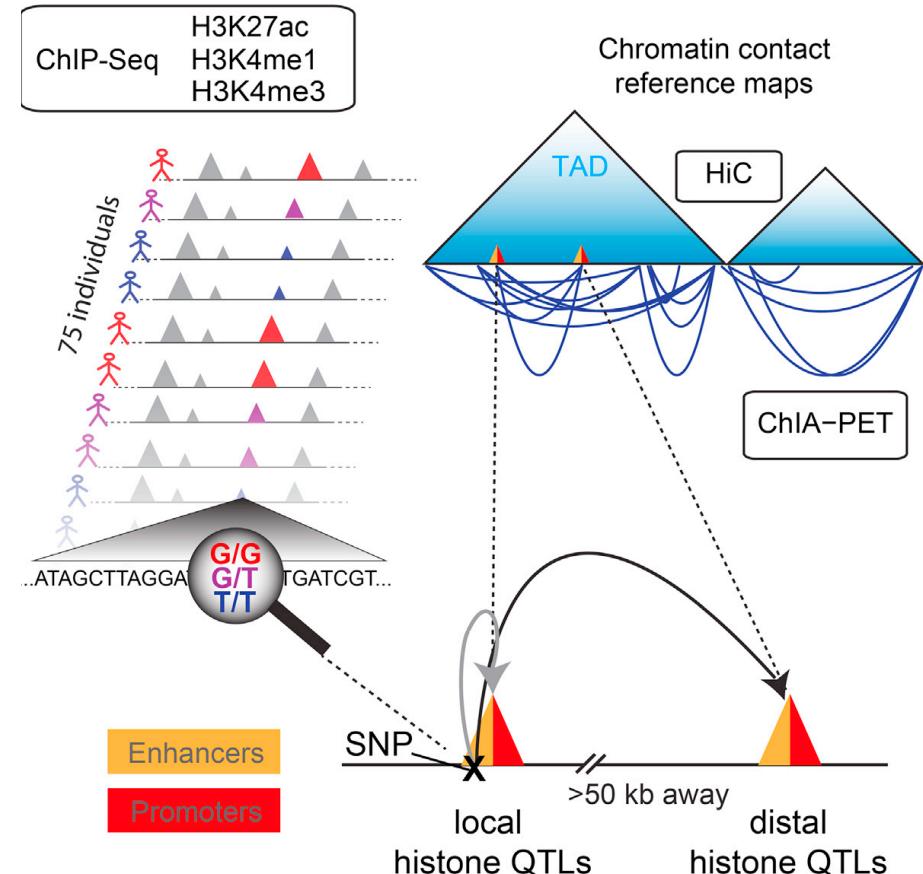
Data set 2: hQTL

ChIP-seq for histone marks in lymphoblastoid cell lines from 75 sequenced individuals.

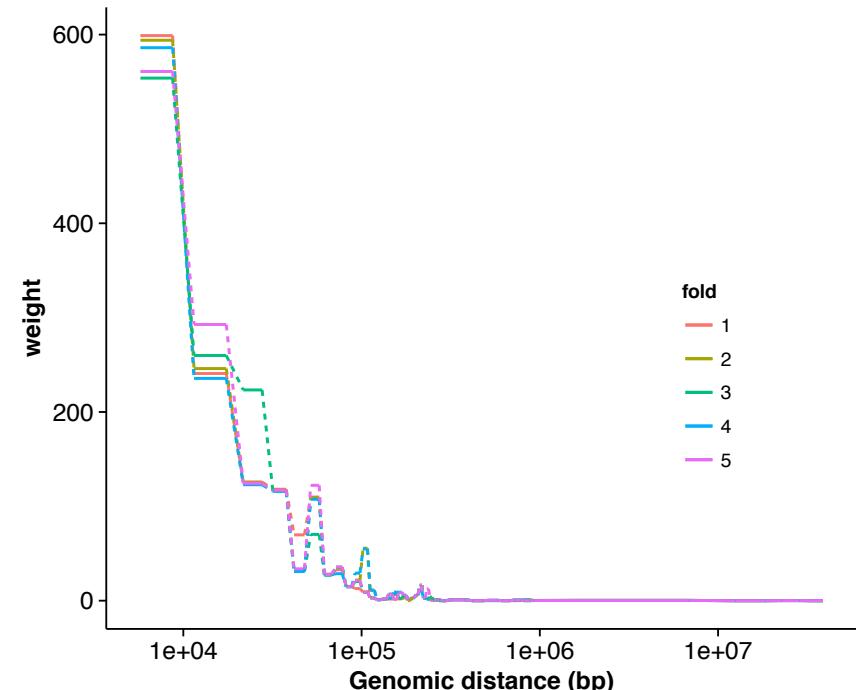
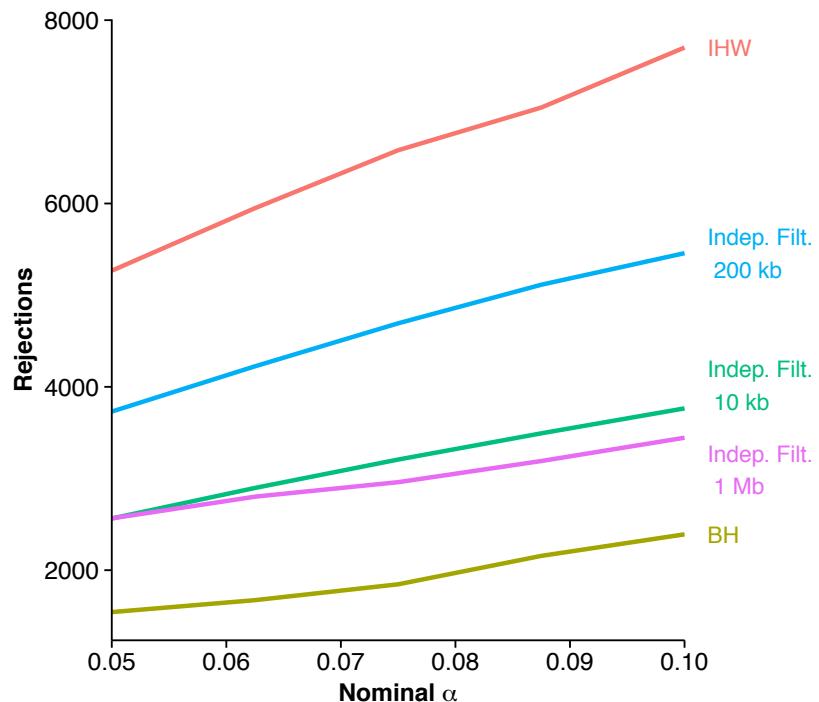
Local QTLs: find best-correlated SNP within 2kb of peak boundaries/promoters.

14,142 local hQTLs linked to ~10% of H3K27ac peaks (FDR 10%, permutations)

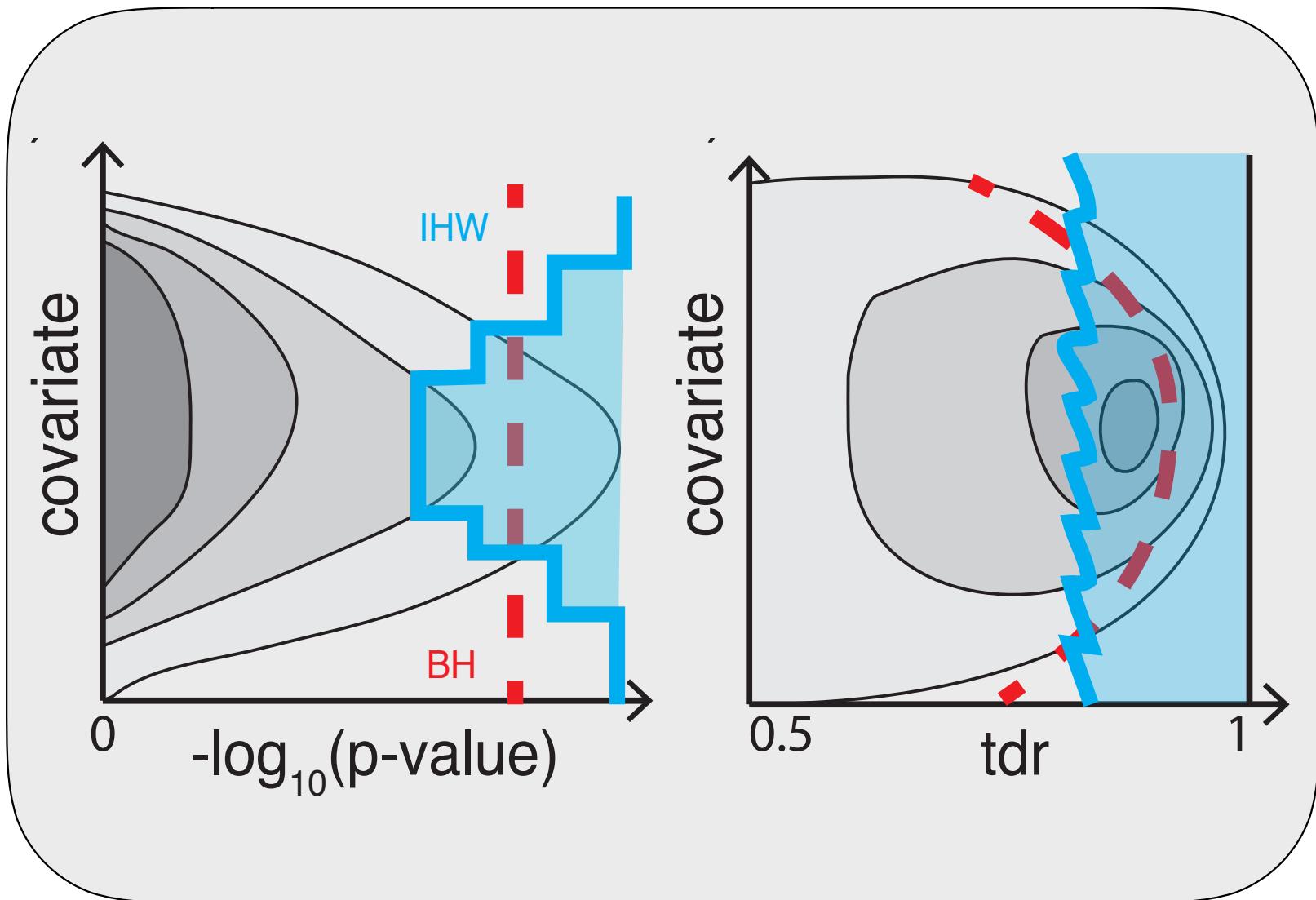
Distal: distance cutoffs from 50 to 300 kb; also HiC



histone-QTL example: H3K27ac



2D decision boundaries



Availability

Bioconductor - IHW (devel...)

https://www.bioconductor.org/packages/devel/t/

Search

Home Install Help Developers About

Search:

Home » Bioconductor 3.3 » Software Packages » IHW (development version)

IHW

platforms all downloads available posts 0 in Bioc devel only
build ok commits 0.17 test coverage unknown

This is the **development** version of IHW; to use it, please install the [devel version](#) of Bioconductor.

Independent Hypothesis Weighting

Bioconductor version: Development (3.3)

Independent hypothesis weighting (IHW) is a multiple testing procedure that increases power compared to the method of Benjamini and Hochberg by assigning data-driven weights to each hypothesis. The input to IHW is a two-column table of p-values and covariates. The covariate can be any continuous-valued or categorical variable that is thought to be informative on the statistical properties of each hypothesis test, while it is independent of the p-value under the null hypothesis.

Author: Nikos Ignatiadis [aut, cre]

Maintainer: Nikos Ignatiadis <nikos.ignatiadis01 at gmail.com>

Citation (from within R, enter `citation("IHW")`):

Ignatiadis N, Klaus B, Zaugg J and Huber W (2015). "Data-driven hypothesis weighting increases detection power in big data analytics." *bioRxiv*.

Installation

To install this package, start R and enter:

Documentation »

Bioconductor

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages
- [Bioc-devel](#) mailing list - for package developers

Paper: Nature Methods June 2016

Summary

- Multiple testing is not a problem but an opportunity
- Heterogeneity across tests
- Informative covariates are often apparent to domain scientists
 - independent of test statistic under the null
 - informative on π_1, F_{alt}

Data-driven weighting

- Scales well to millions of hypotheses
- Controls ‘overoptimism’



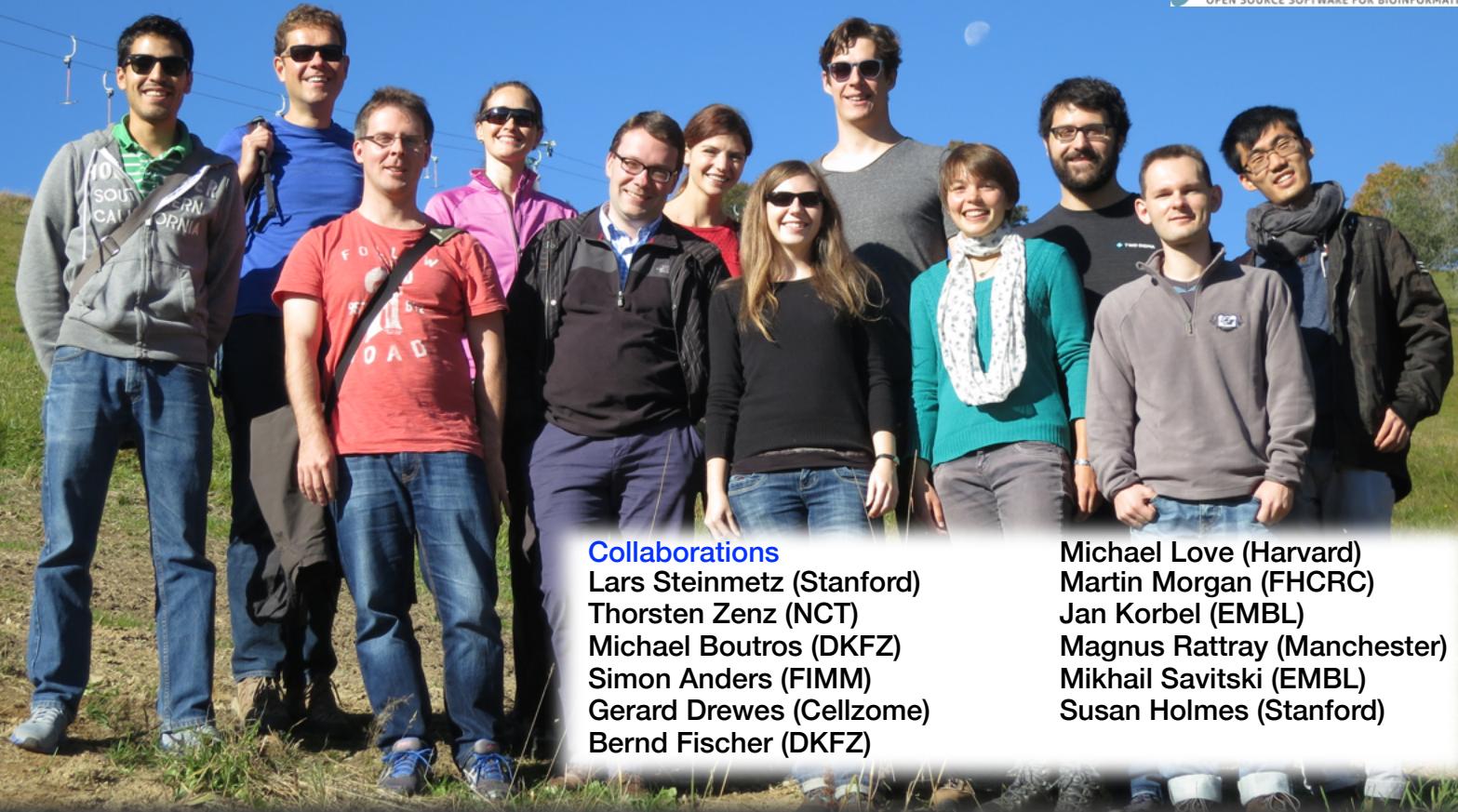
Joint work with
Nikos Ignatiadis

Simone Bell
Dorothee Childs
Sascha Dietrich
Nikos Ignatiadis

Vlad Kim
Bernd Klaus
Junyan Lu
Andrzej Oles

Malgorzata Oles
Thomas Schwarzl
Arne Smits
Mike Smith

Britta Velten



Collaborations

Lars Steinmetz (Stanford)
Thorsten Zenz (NCT)
Michael Boutros (DKFZ)
Simon Anders (FIMM)
Gerard Drewes (Cellzome)
Bernd Fischer (DKFZ)

Michael Love (Harvard)
Martin Morgan (FHCRC)
Jan Korbel (EMBL)
Magnus Rattray (Manchester)
Mikhail Savitski (EMBL)
Susan Holmes (Stanford)

Funding

EC: CancerPathways, Systems Microscopy, RADIANT, SOUND
HFSP NSF - BIGDATA
EMBL Cellzome (GSK)

