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

Wolfgang Huber

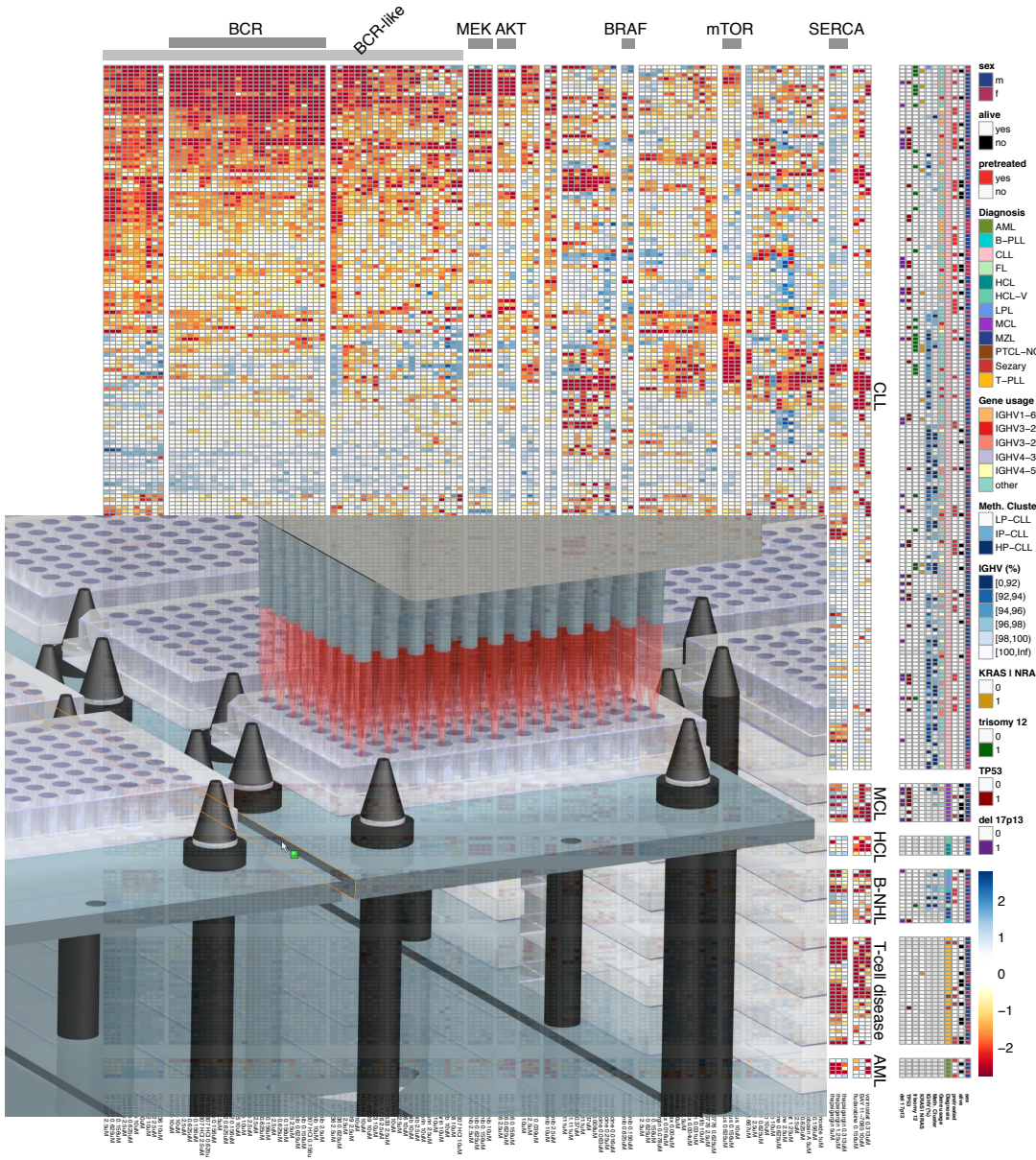
Research Group Leader, Senior Scientist
European Molecular Biology Laboratory (EMBL)

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

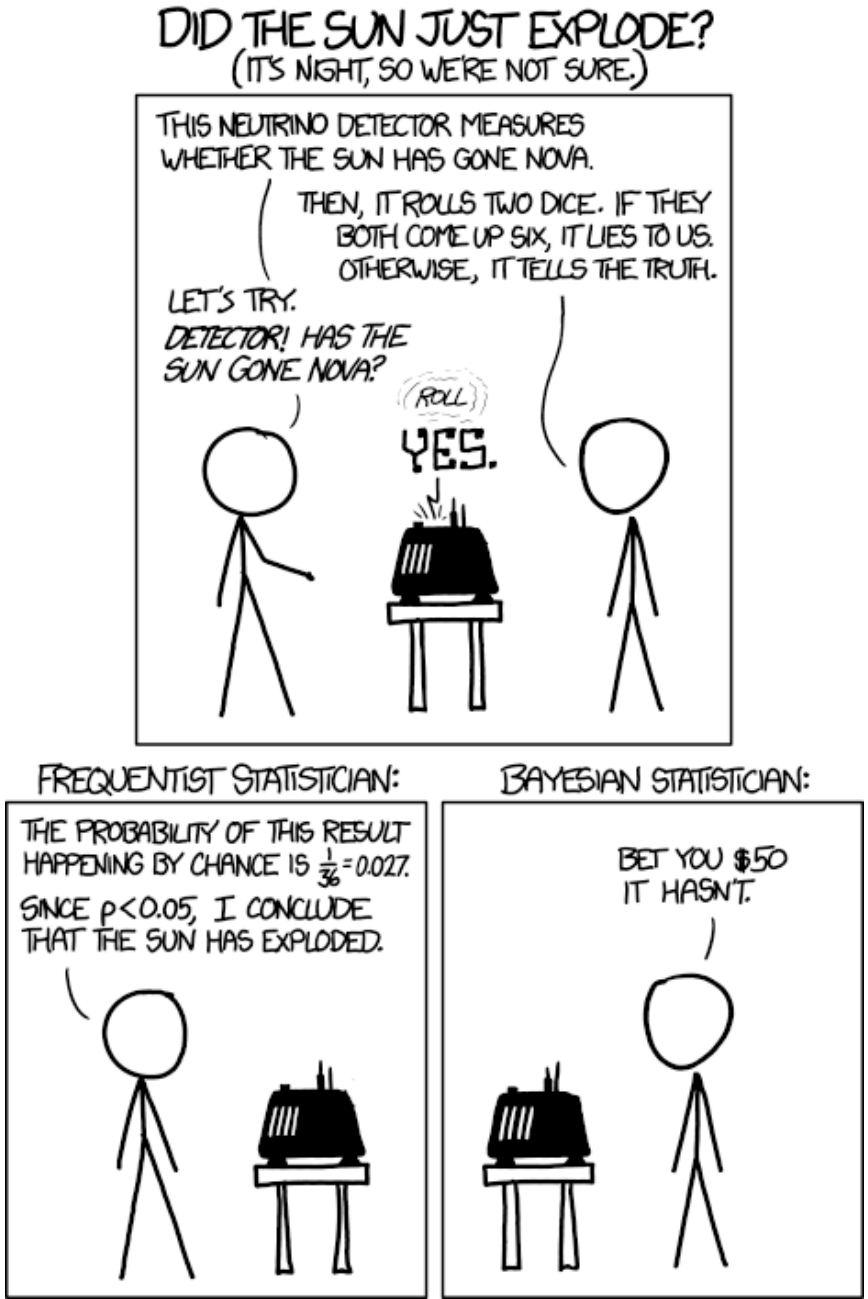


The Multiple Testing Burden

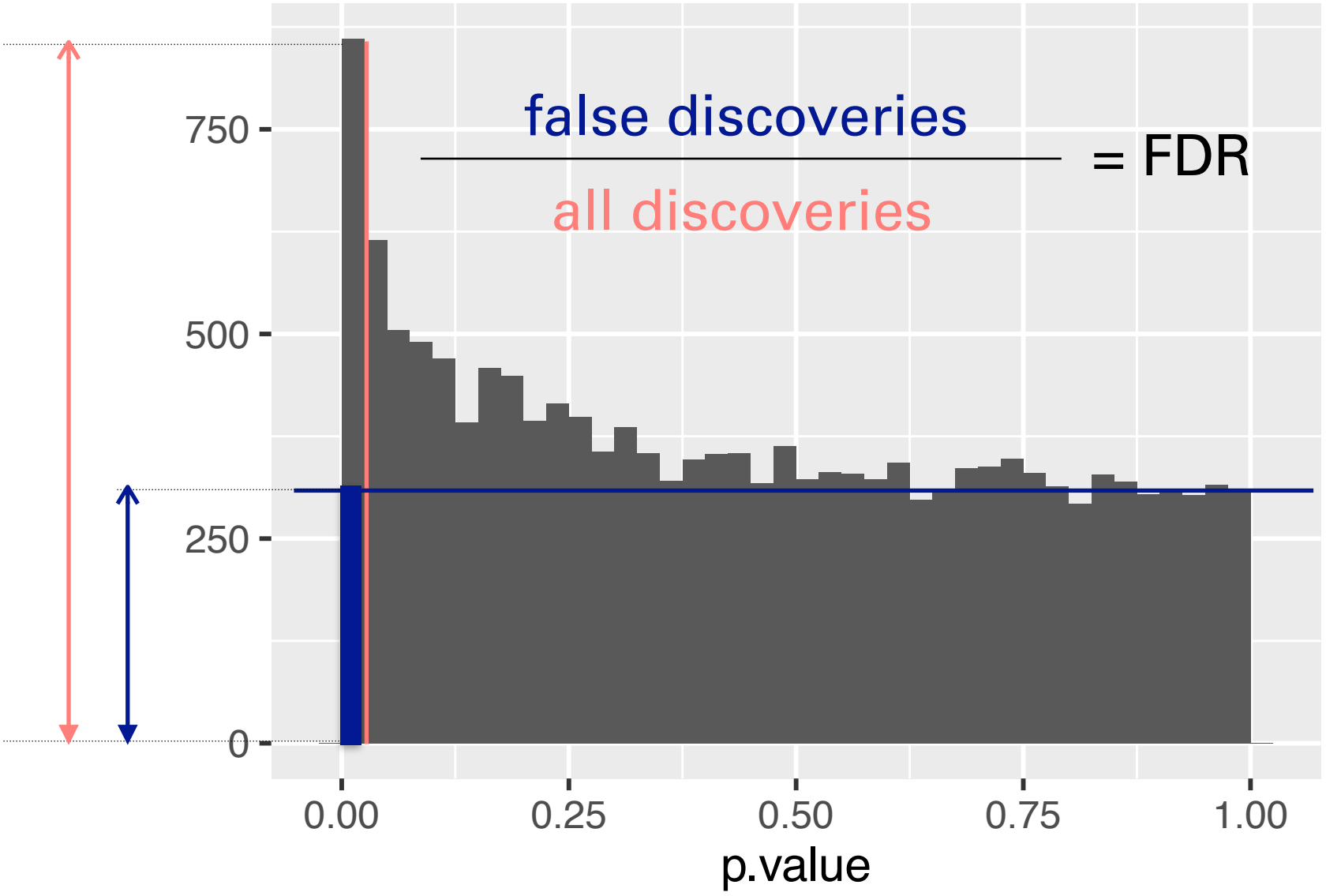
When performing several tests, type I error goes up: for $\alpha = 0.05$ and n independent tests, probability of no false positive result is

$$\underbrace{0.95 \cdot 0.95 \cdot \ldots \cdot 0.95}_{\text{n–times}} \lll 0.95$$

The Multiple Testing Opportunity

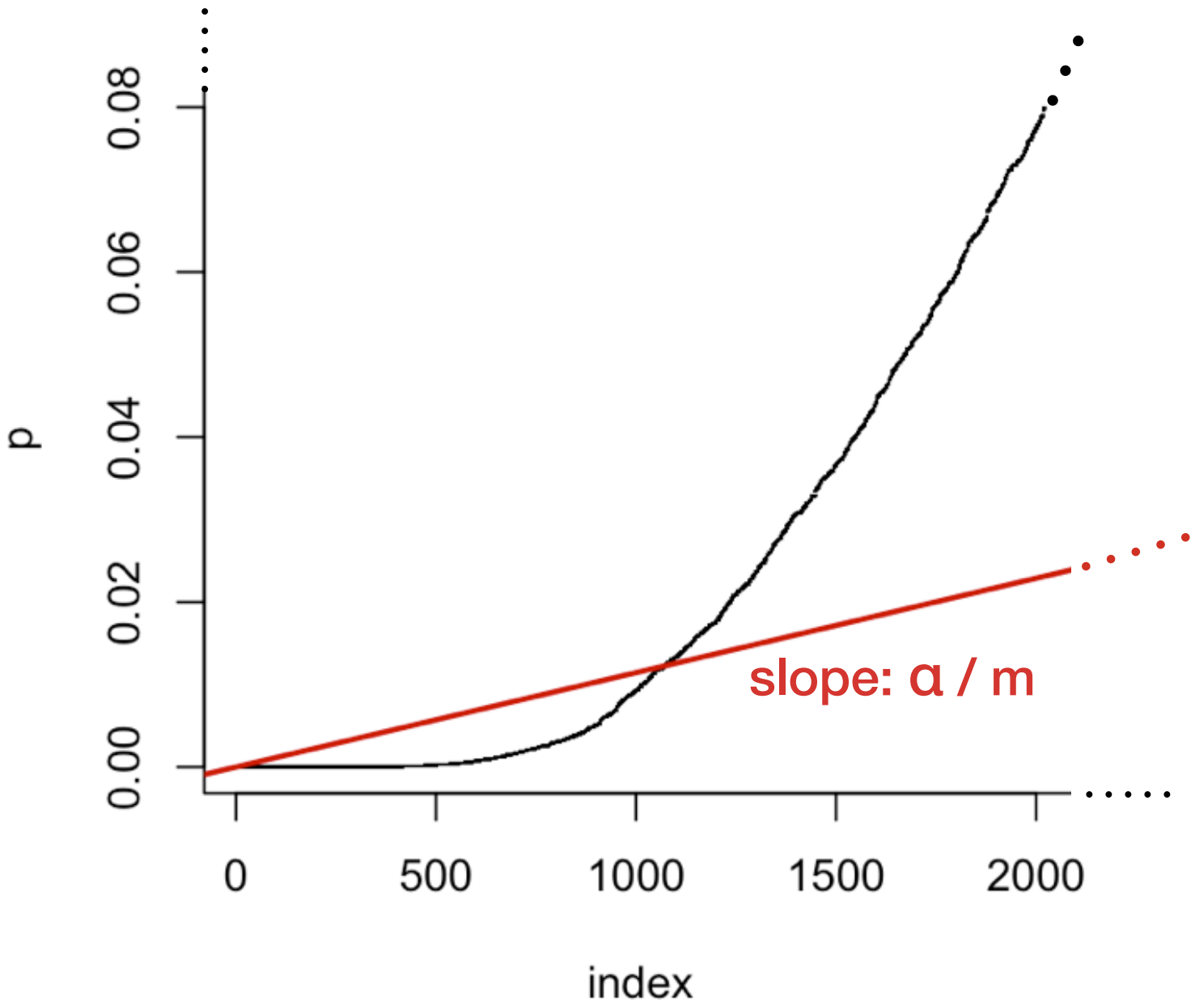


False discovery rate



Method of Benjamini & Hochberg (1995)

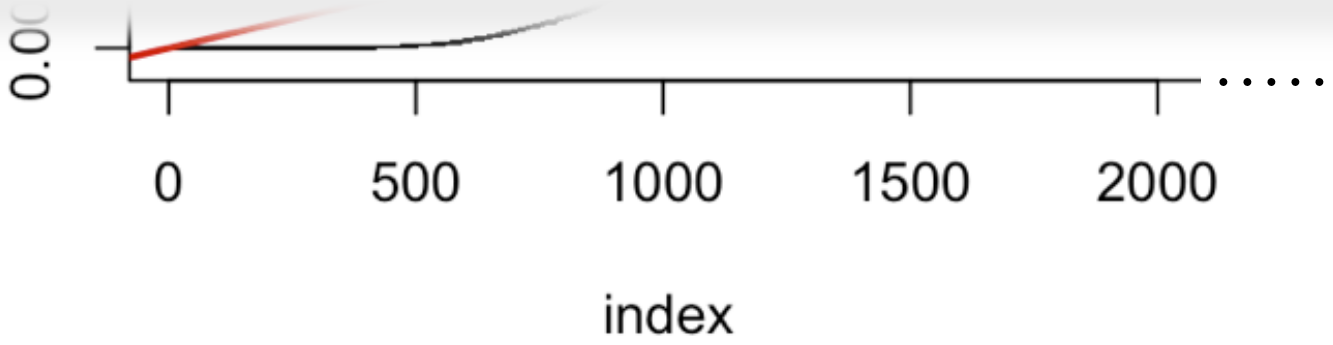
Method of Benjamini & Hochberg



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.




Exchangeability?

Covariates - examples

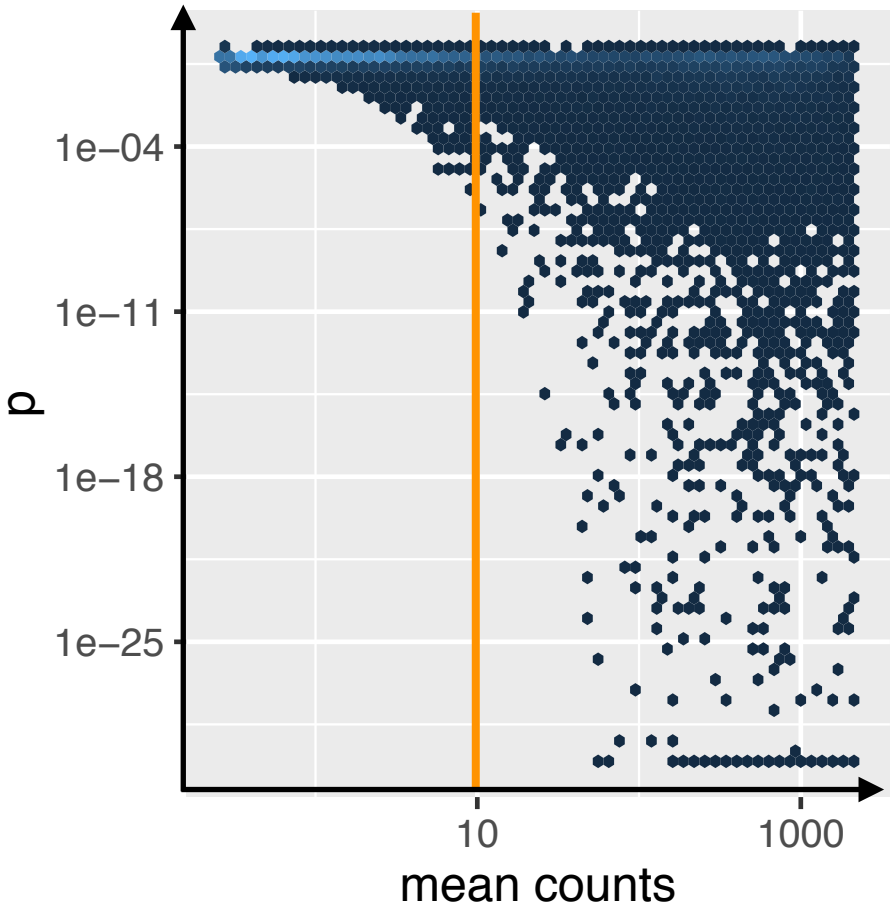
Application	Covariate
Differential RNA-Seq, ChIP-Seq, CLIP-seq, ...	(Normalized) mean of counts for each gene
GWAS	Minor allele frequency
eQTL analysis	SNP – gene distance
t-tests	Overall variance
Two-sided tests	Sign
All applications	Sample size; measures of signal-to-noise ratio

Independent Filtering

Two steps:

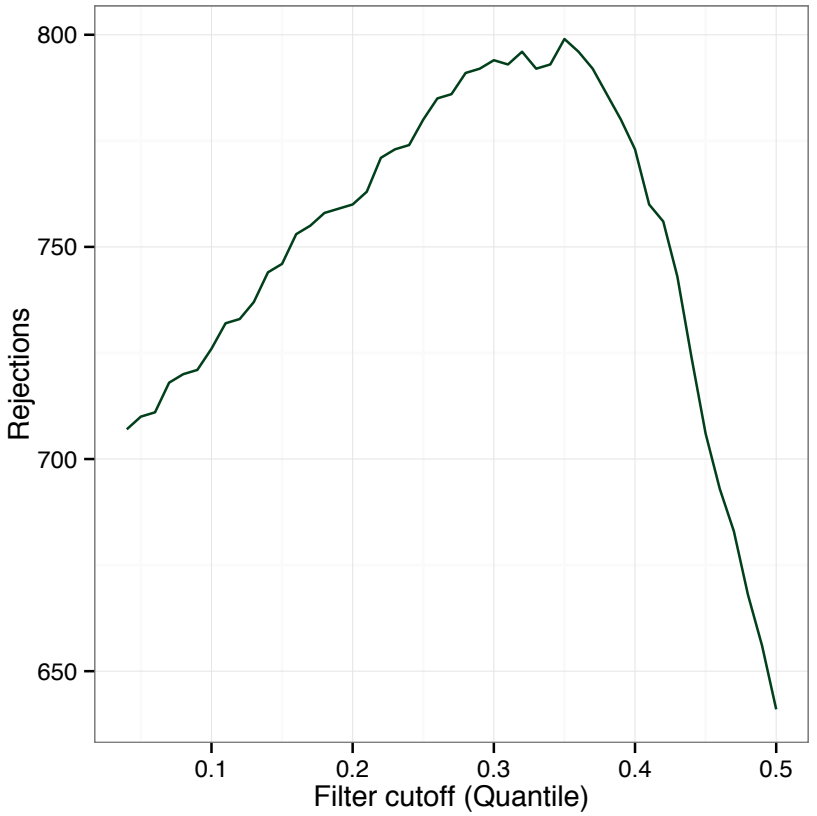
- All hypotheses H_i with $X_i < x$ get filtered.
- Apply BH to remaining hypotheses.

(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, Wasserm
- If $w_i > 1$
- $Q_i \leq t \Leftarrow$

Weighted Benjamini-Hoc method

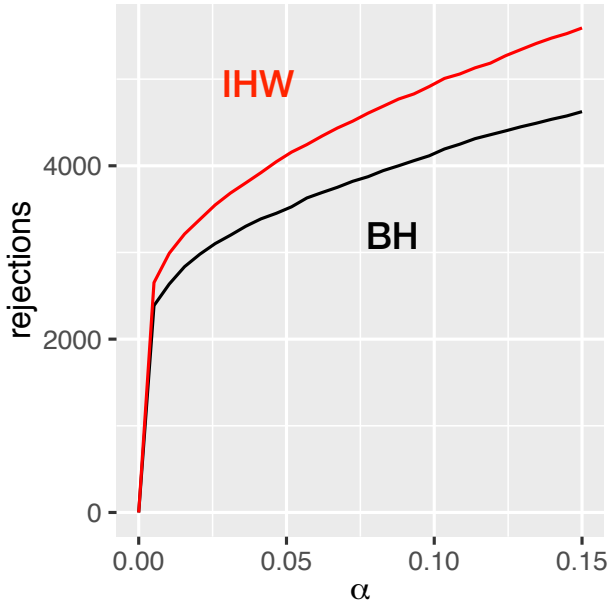
- Let $w_i \geq 0$ an budget").
- Define
- A P_i.
- P r, Wa
- If w_i
- $Q_i \leq t \Leftarrow$

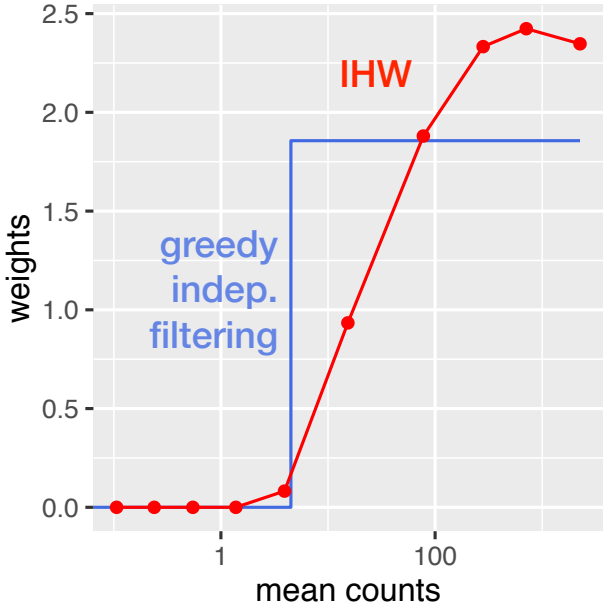
Problem: how to know the weights?

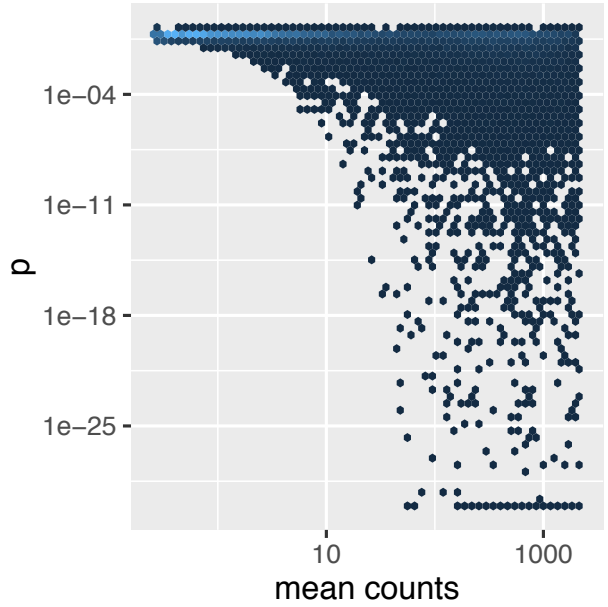
IHW (naive): Independent (data-driven) hypothesis weighting

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

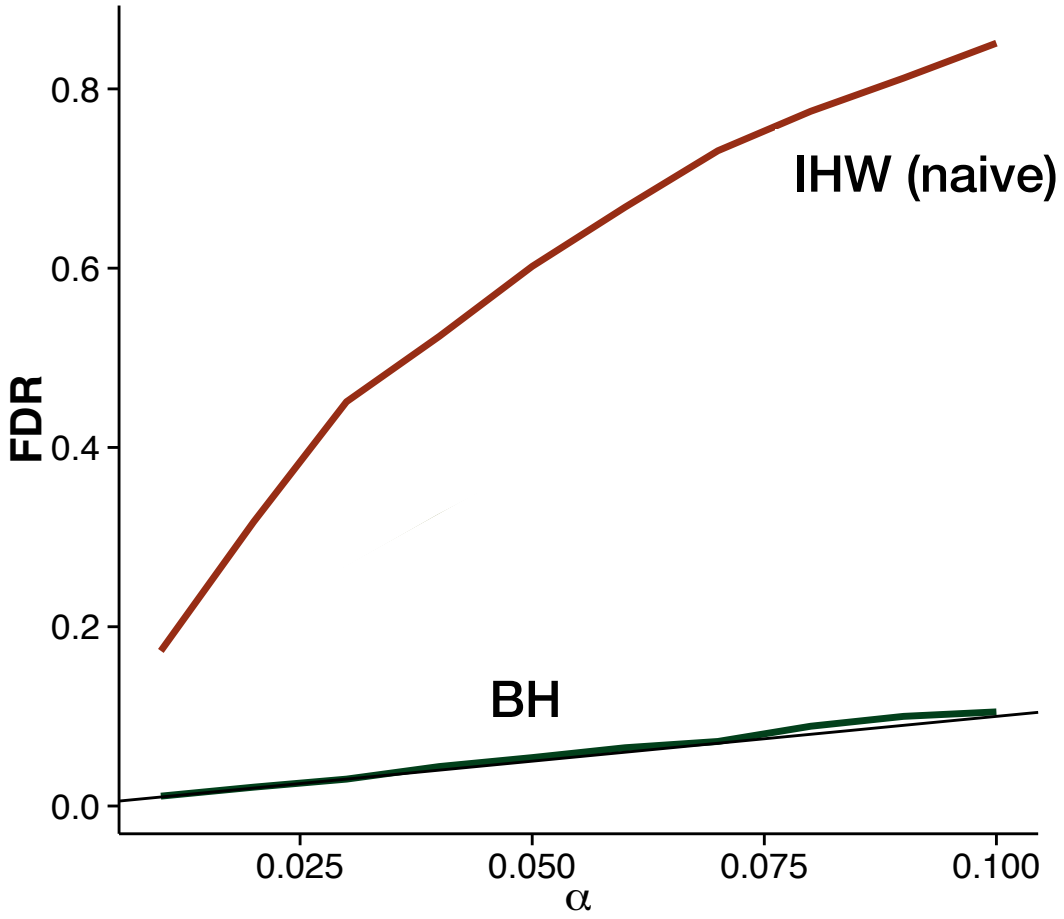
RNA-Seq example (DESeq2)







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

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

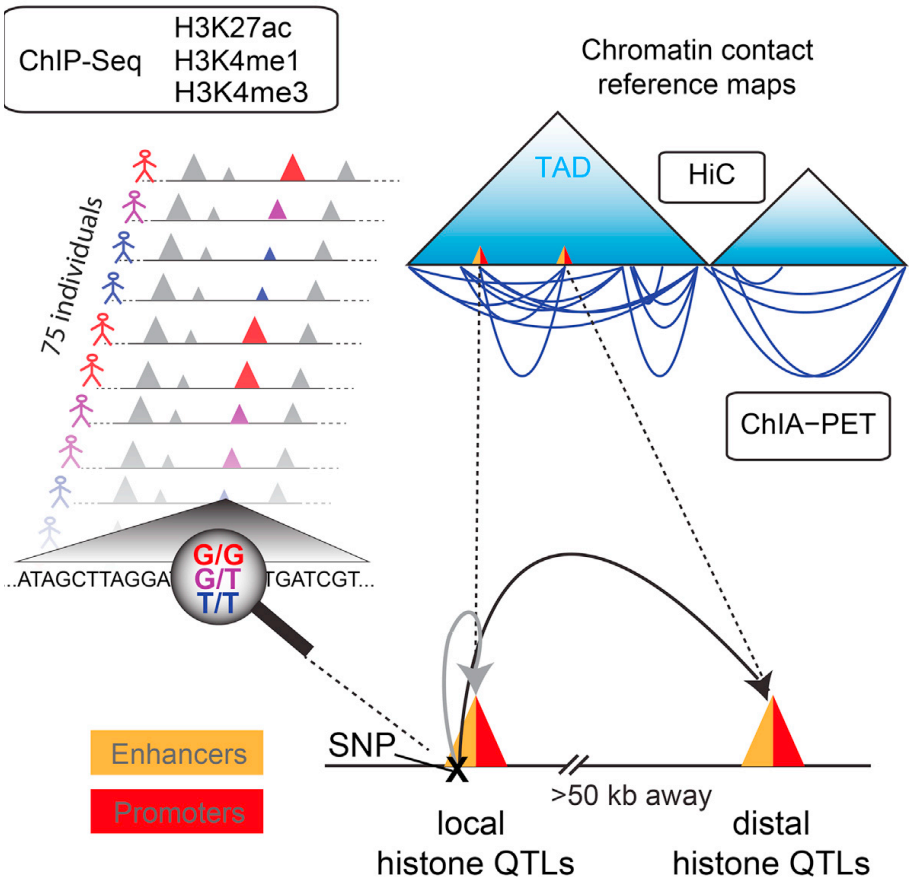
Data set 2: hQTL

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.

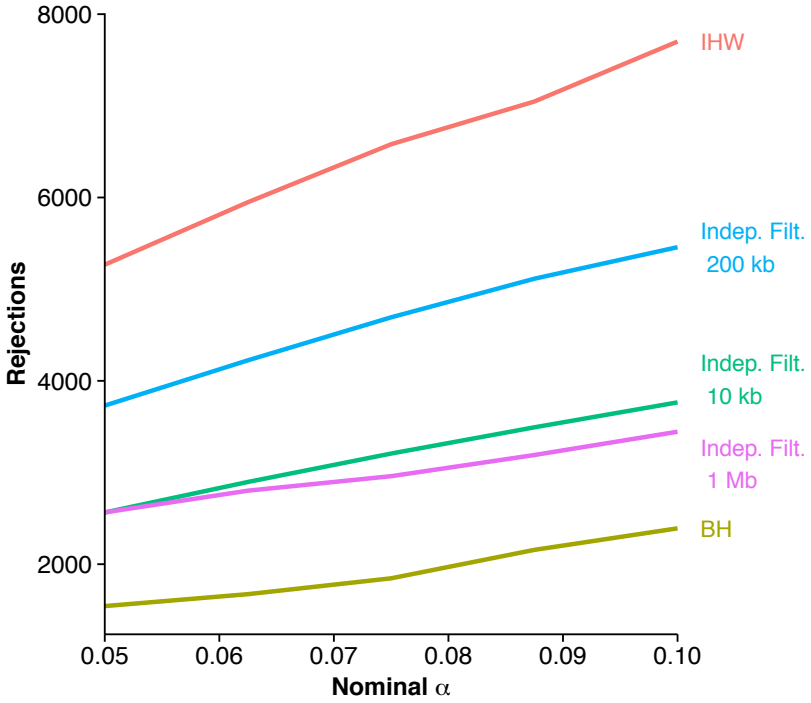
14,142 local hQTLs linked to ~10% of H3K27ac peaks (FDR 10%, permutations)

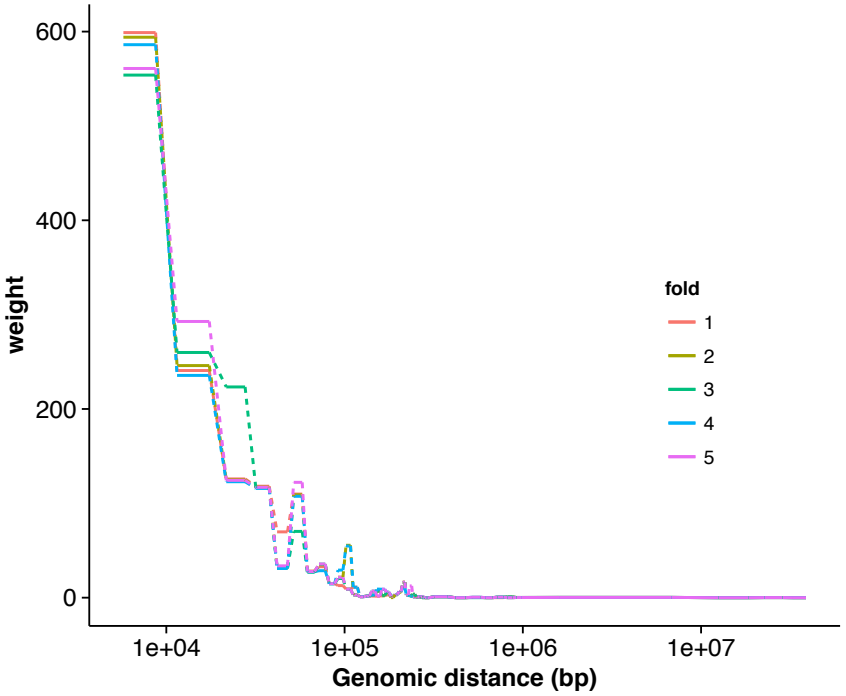
Distal: distance cutoffs from 50 to 300 kb; also HiC



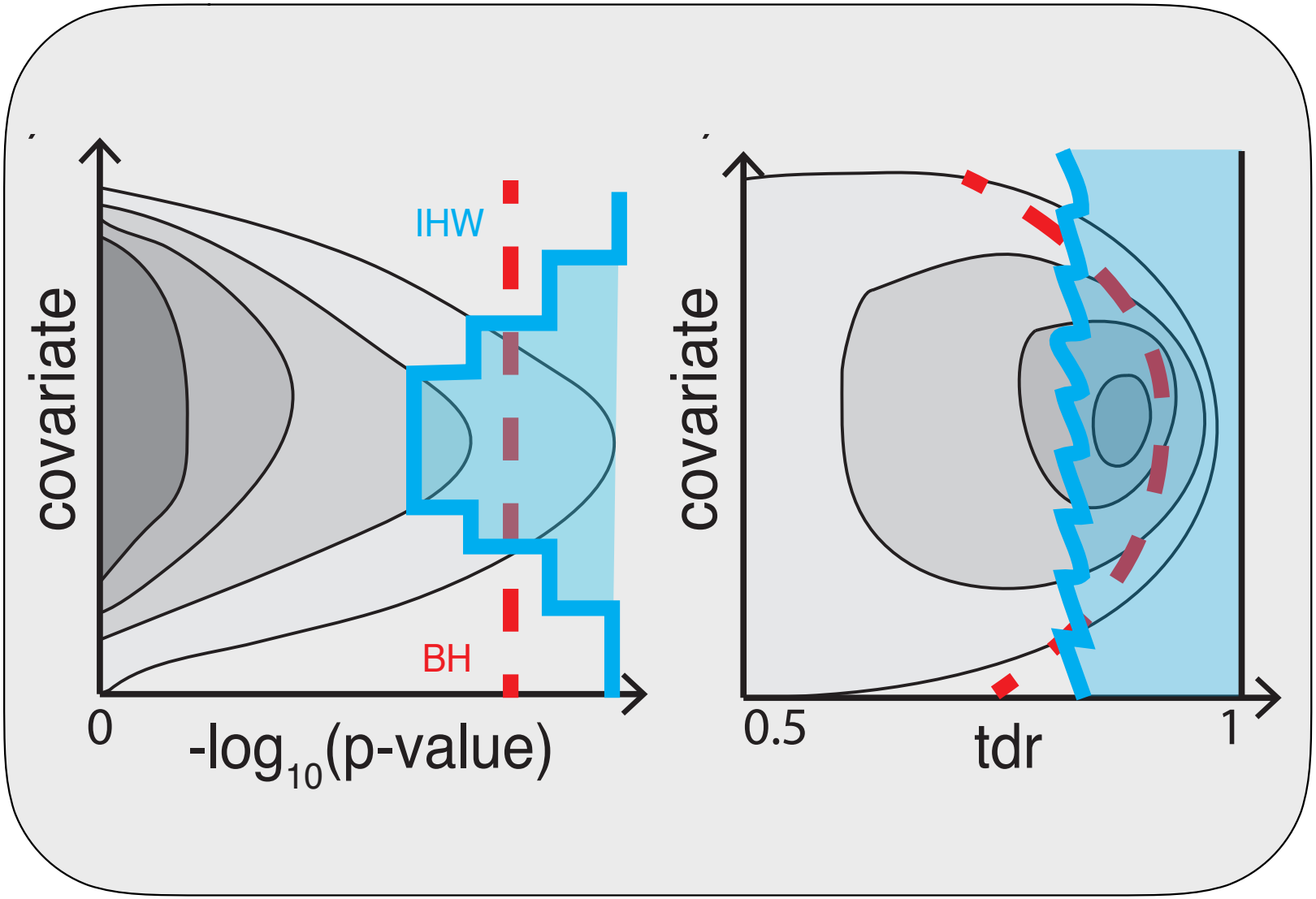
Grubert, Zaugg, Kasowski, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell (2015).

histone-QTL example: H3K27ac

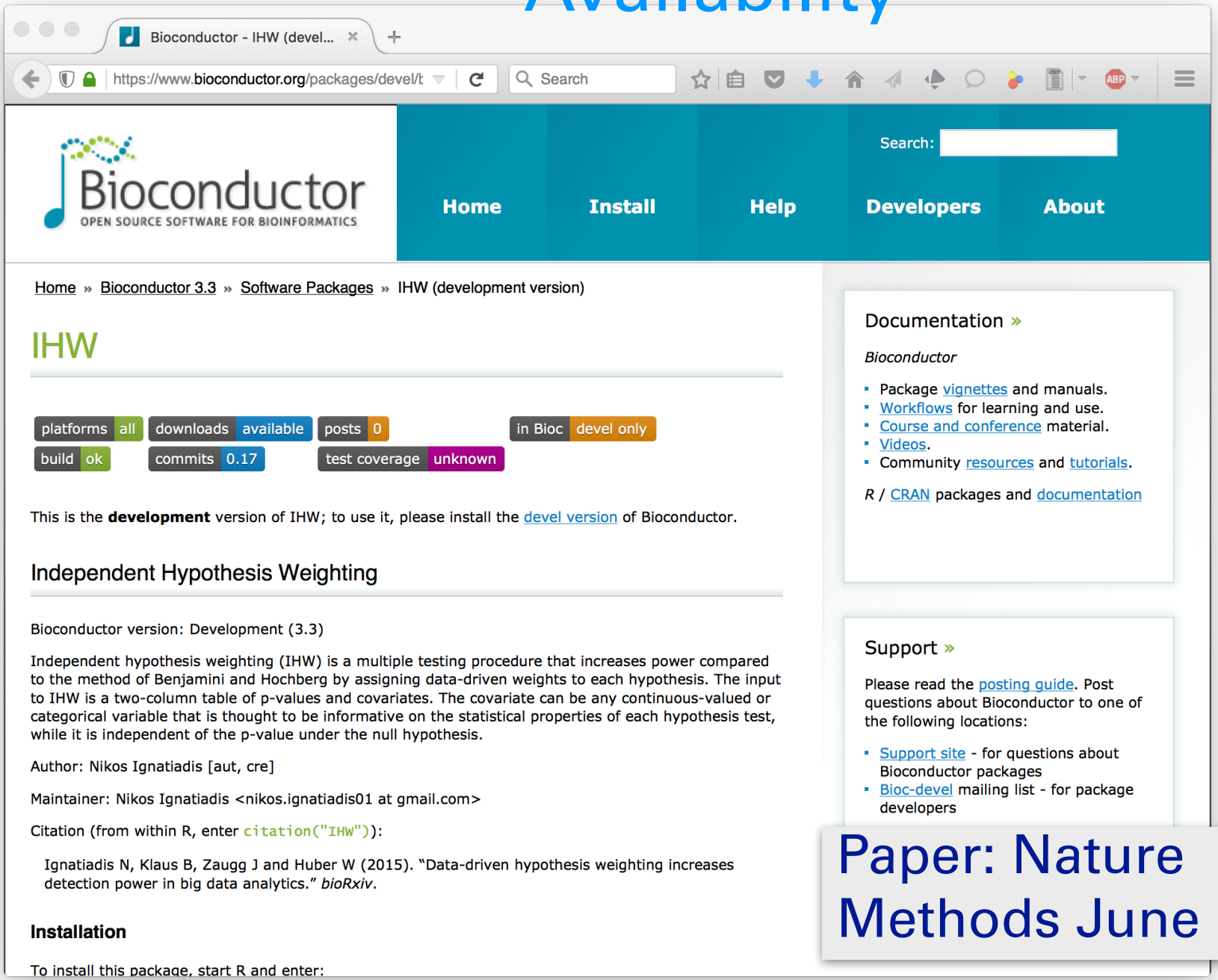




2D decision boundaries



Availability



Paper: Nature Methods June 2016

Summary

Joint work with
Nikos Ignatiadis

- 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’

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

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

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)

EMBL