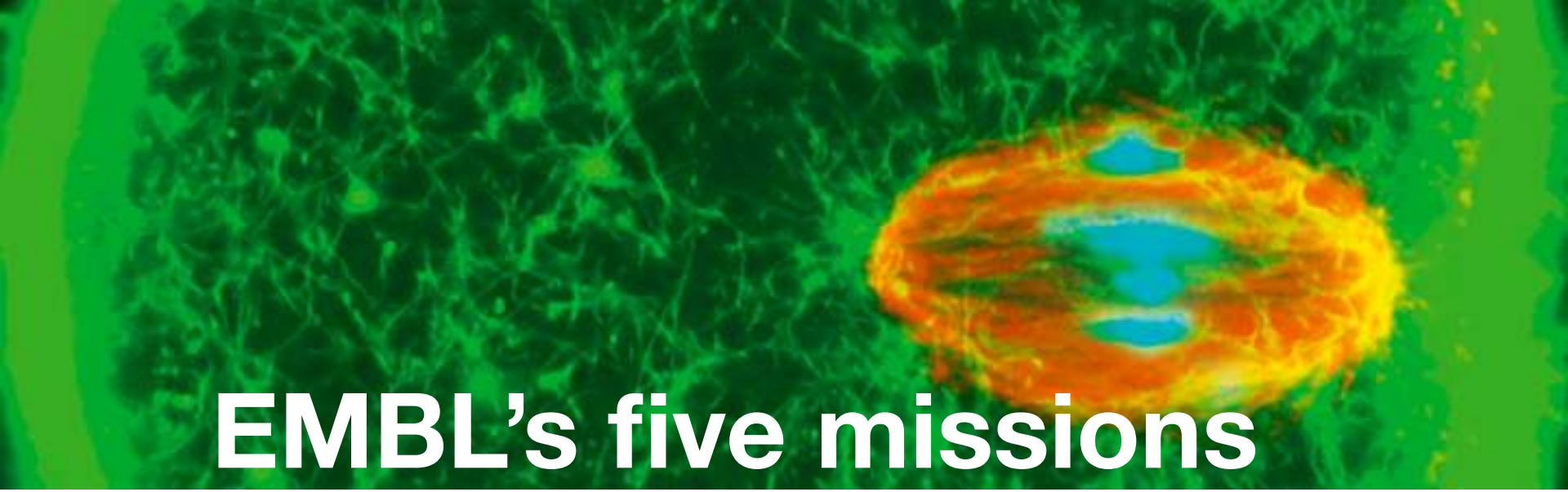
Hypothesis Testing

Wolfgang Huber, EMBL

European Molecular Biology Laboratory (EMBL)

Intergovernmental Research Organisation

- 22 Member States
- Founded in 1974
- Sites in Heidelberg (D), Cambridge (GB), Roma (I), Grenoble (F), Hamburg (D), soon: Barcelona
- ca. 1400 staff (⊃1100 scientists) representing more than 60 nationalities



EMBL's five missions

- Fundamental research
- Develop new technologies and instruments
- Technology transfer
- Services to member states
- Advanced training

EMBL

What Can You Do at EMBL ?

Biology
Chemistry
Medicine
Physics
Mathematics
Informatics
Engineering

Internships - Phd programme - Postdocs - PIs - Jobs

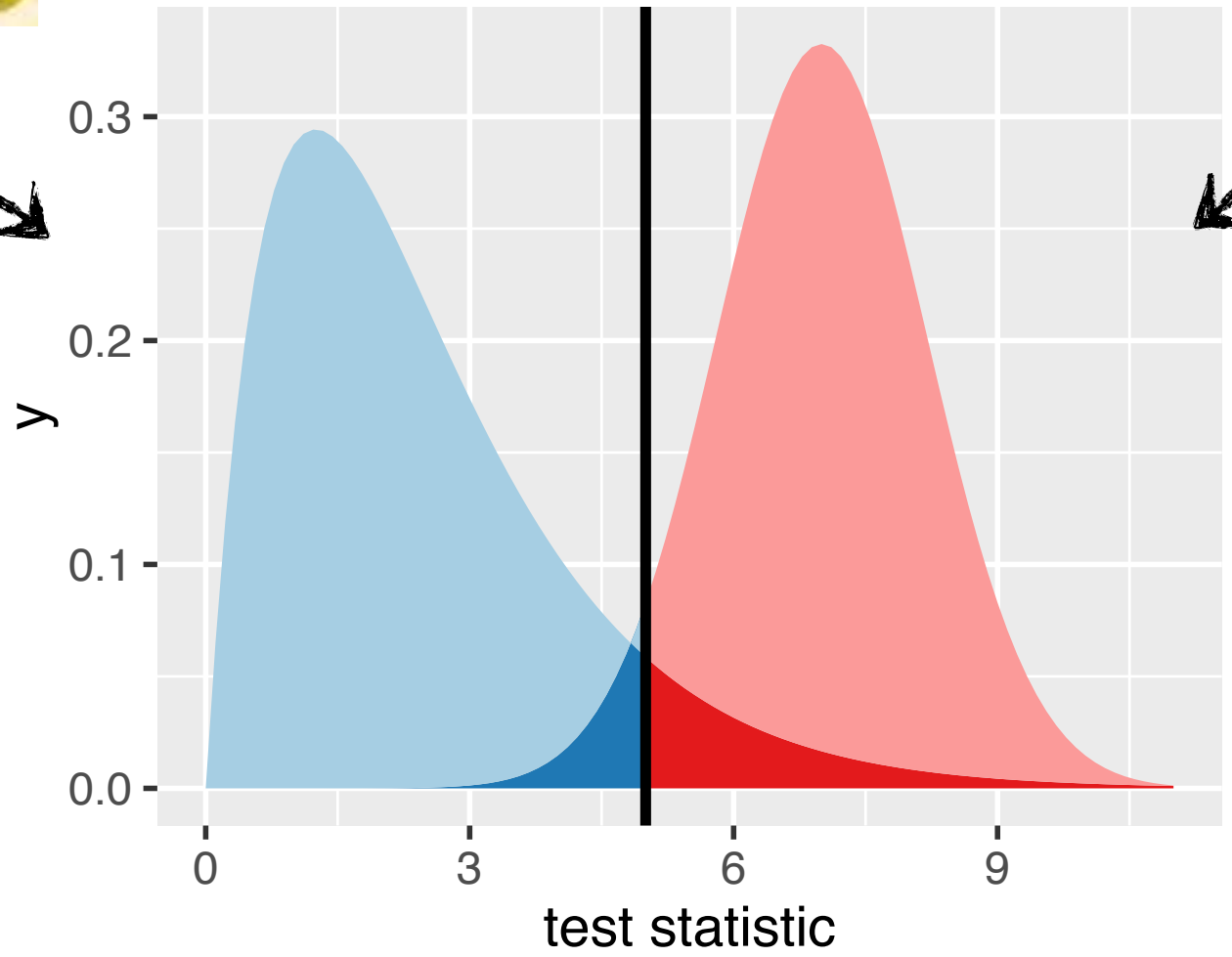
Aims for this Lecture

Understand the basic principles of hypothesis testing, its pitfalls, strengths, use cases and limitations

What changes when we go from single to multiple testing?

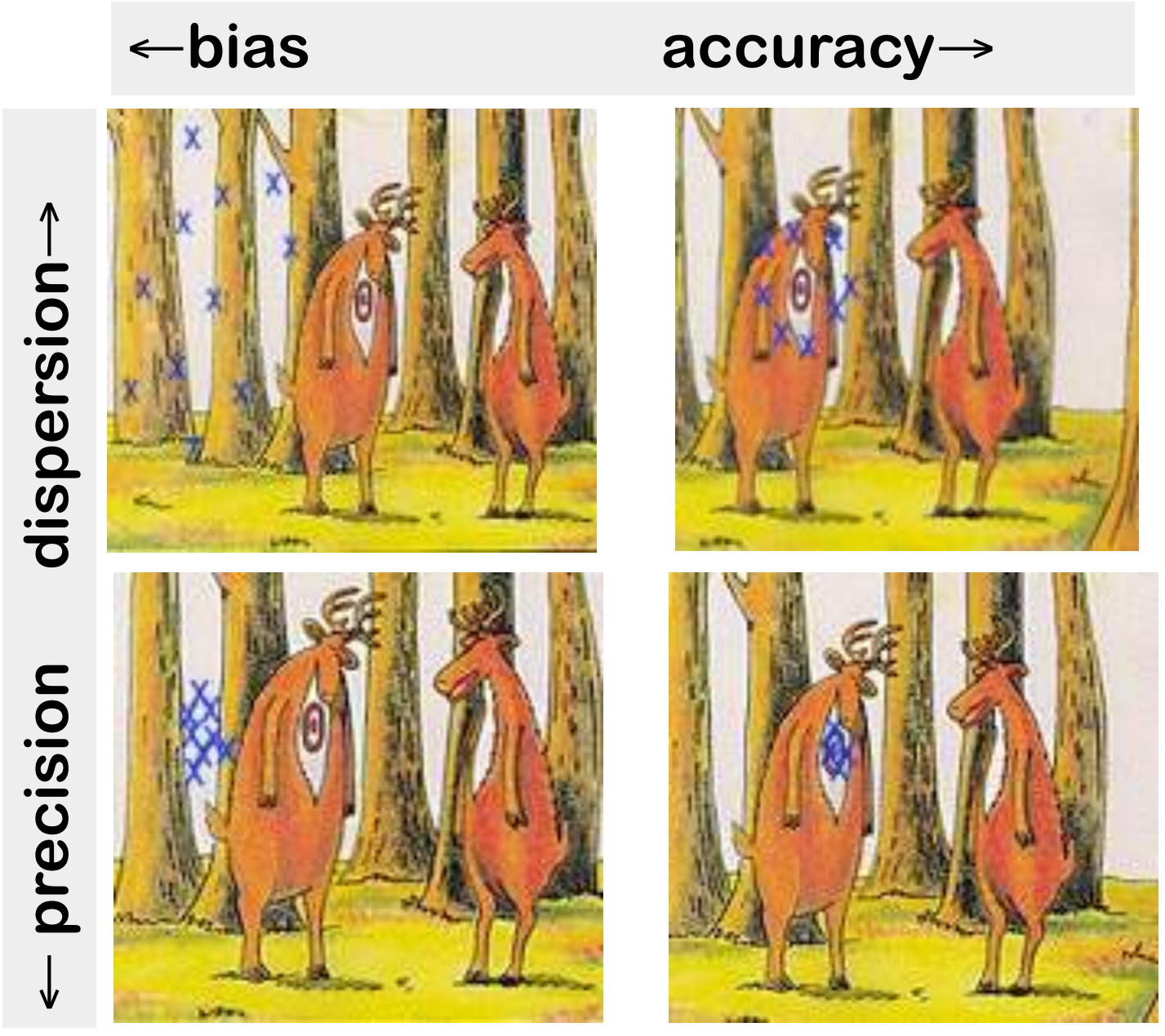
False discovery rates, p-value ‘adjustments’, filtering and weighting

Testing vs Classification



(any function of the data)

Accuracy vs Precision - Bias vs Variance



Karl Popper (1902-1994)

Logical asymmetry between verification and falsifiability.

No number of positive outcomes at the level of experimental testing can confirm a scientific theory, but a single counterexample is logically decisive: it shows the theory is false.

Example

Toss a coin a number of times ⇒

If the coin is fair, then heads should appear half of the time (roughly).

But what is "roughly"? We use combinatorics / probability theory to quantify this.

Suppose we flipped the coin 100 times and got 59 heads. Is this 'significant'?

Binomial Distribution

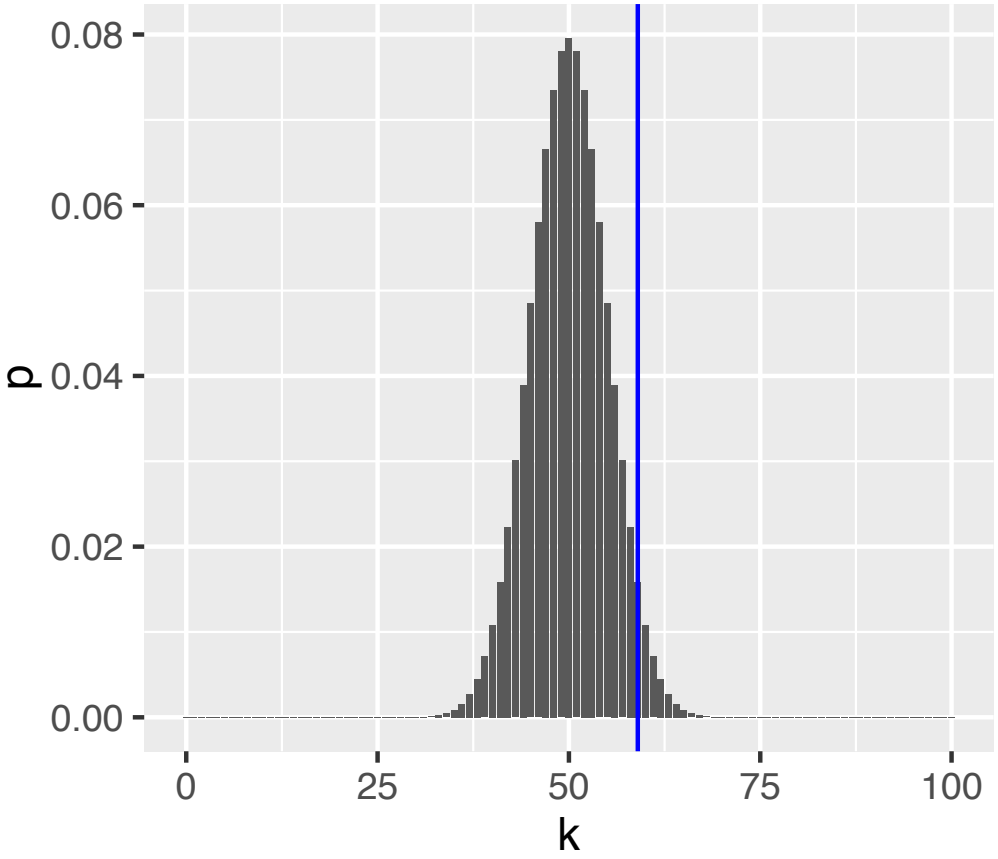


Figure 6.3: The binomial distribution for the parameters $n = 100$ and $p = 0.5$,

$$P(K = k \mid n, p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k \, (1-p)^{n-k}$$

Rejection Region

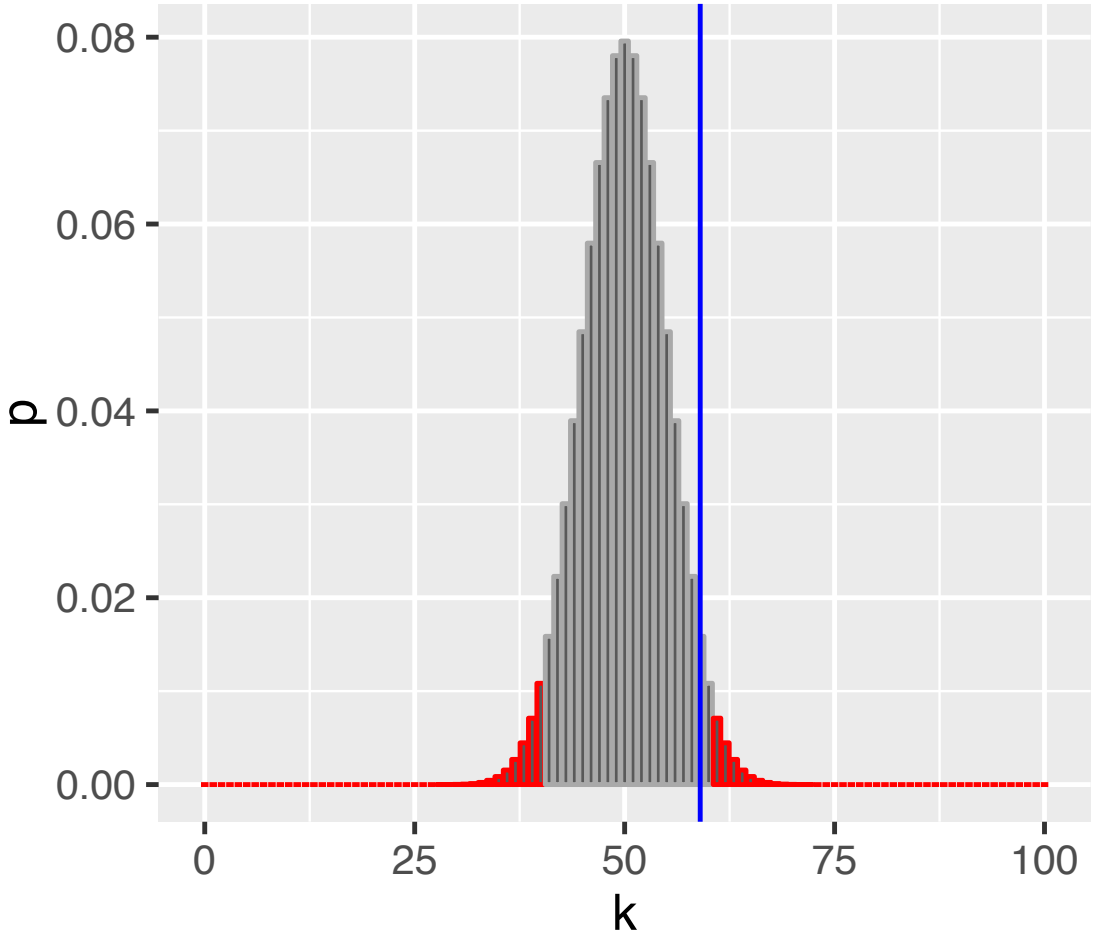


Figure 6.5: As Figure 6.3, with rejection region (red) whose total area is $\alpha = 0.05$.

Questions

- Does the fact that we don't reject the null hypothesis mean that the coin is fair?
- Would we have a better chance of detecting that the coin is not fair if we did more coin tosses? How many?
- If we repeated the whole procedure and again tossed the coin 100 times, might we **then** reject the null hypothesis?
- Our rejection region is asymmetric - its left part ends with 40, while its right part starts with 61. Why is that? Which other ways of defining the rejection region might be useful?

The Five Steps of Hypothesis Testing

Choose an experimental design and a data summary function for the effect that you are interested in: the **test statistic**

Set up a **null hypothesis**: a simple, computationally tractable model of reality that lets you compute the null distribution of the test statistic, i.e. the possible outcomes and each of their probabilities.

Decide on the **rejection region**, i.e., a subset of possible outcomes whose total probability is small (<= **significance level**).

Do the experiment, collect data, compute the test statistic.

Make a **decision**: reject null hypothesis if the test statistic is in the rejection region.

Examples of Null Hypotheses:

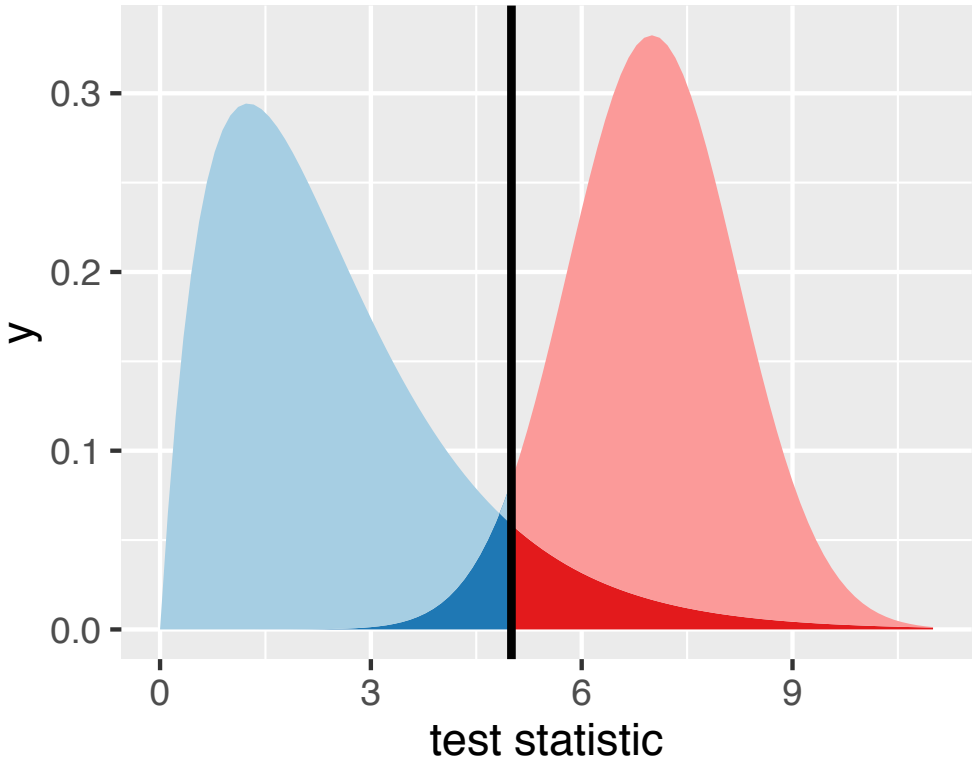
- The coin is fair
- The new drug is no better or worse than a placebo
- The effect of that RNAi-treatment on my cells is no different than that of a negative control treatment

These are not Null Hypotheses:

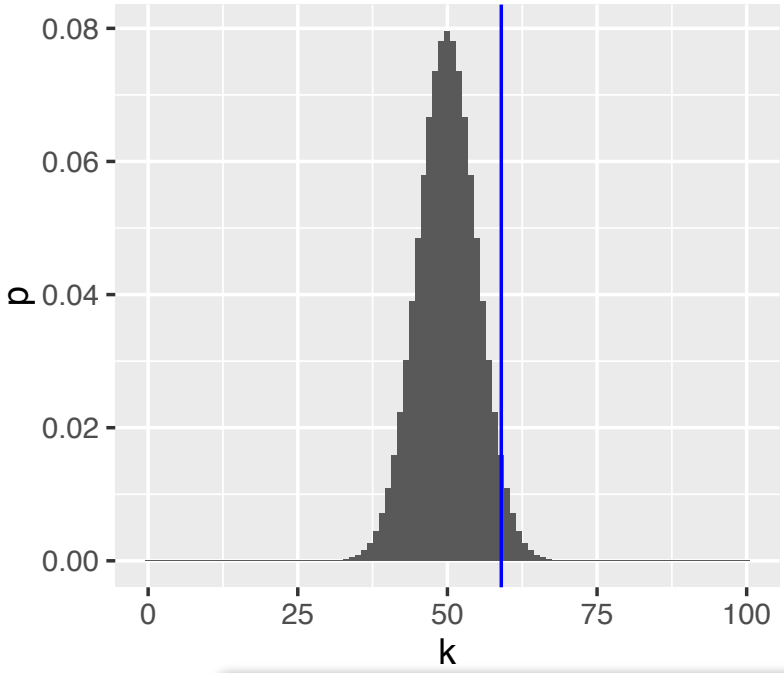
- The number of heads and tails were the same
- The coin is not fair
- The drug is worth its money

Types of Error in Testing

Test vs reality	**Null hypothesis is true**	**. . . is false**
Reject null hypothesis	Type I error (false positive)	True positive
Do not reject	True negative	Type II error (false negative)



Parametric Theory vs Simulation



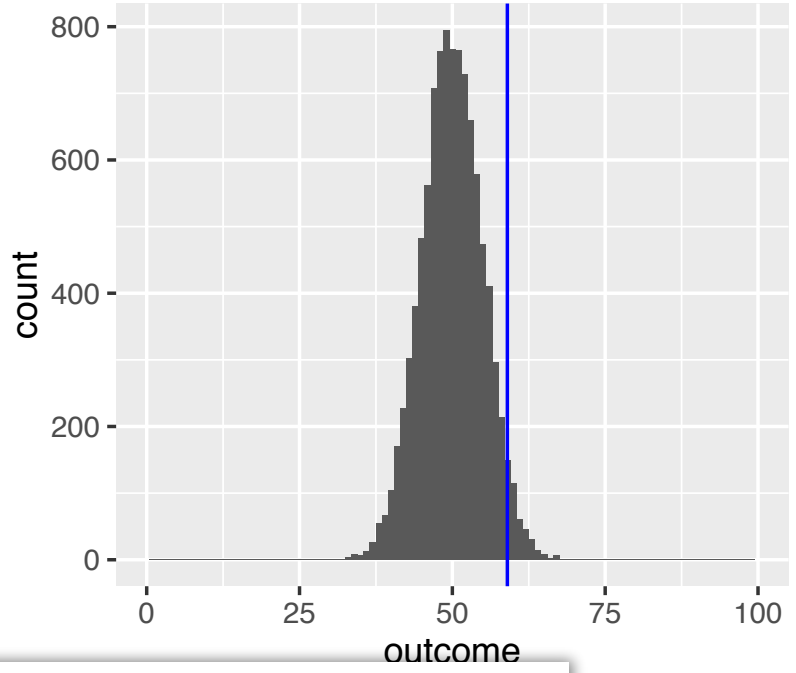
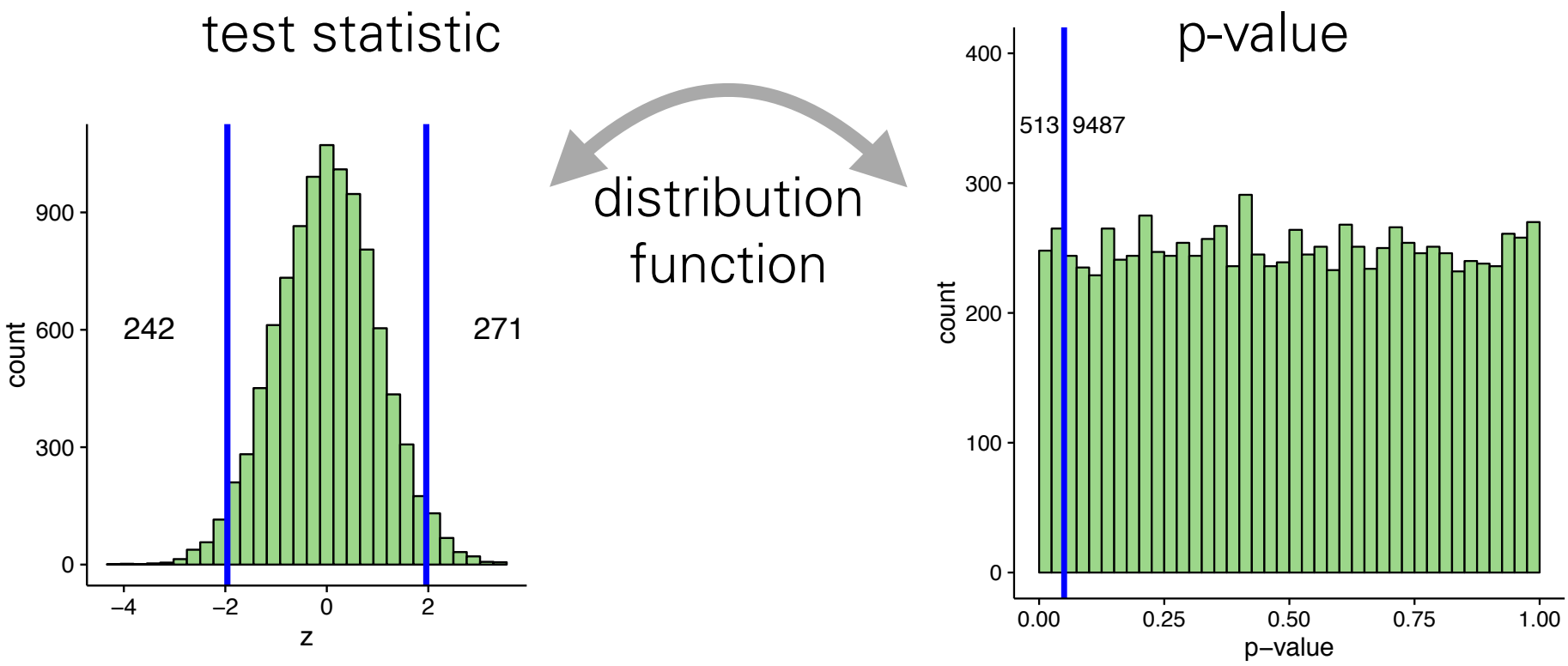


Figure 6.3: T
the paramete
according to

of the
simulations
).

Q:
Discuss pros and contras for each

p-Values as Random Variables



The Test Statistic

Suppose we observed 50 tails in a row, and then 50 heads in a row. Is this a perfectly fair coin?

We could use a different test statistic: number of times we see two tails in a row

Is this statistic generally and always preferable?

Power

There can be several test statistics, with different power, for different types of alternative

Continuous Data: the t-Statistic

$$t = c\,\frac{m_1 - m_2}{s}$$

- Can also be adapted to one group only
- Relation to z-score

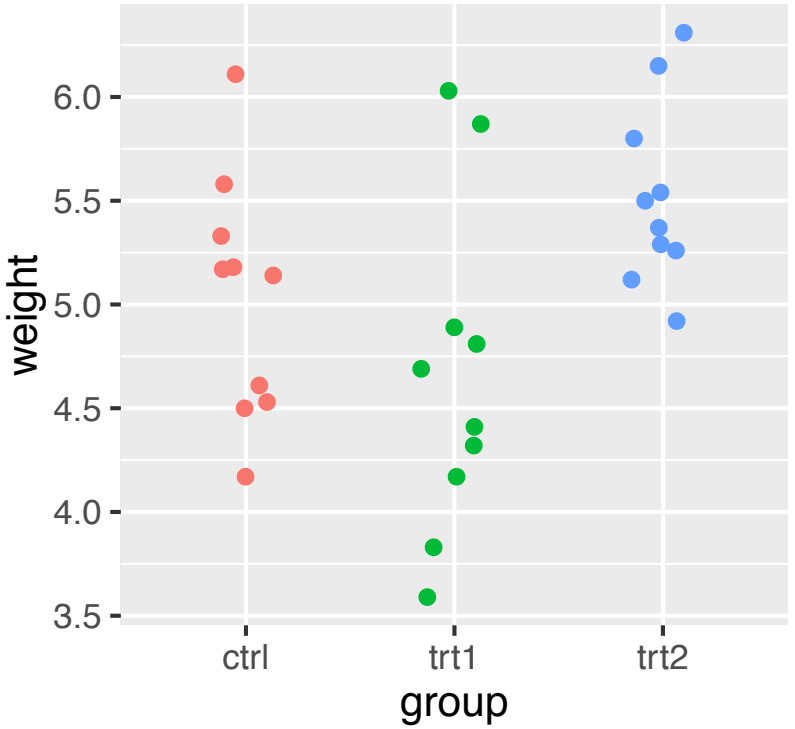
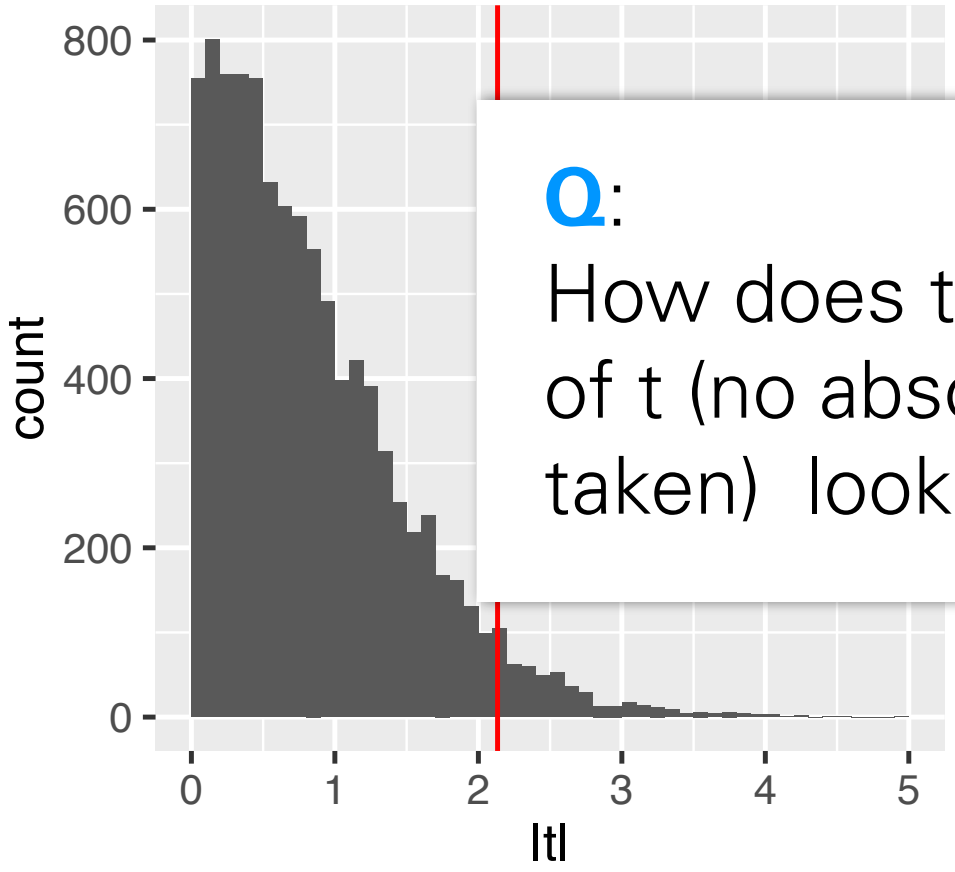


Figure 6.7: The `PlantGrowth` data.

$$\begin{aligned}
m_g &= \frac{1}{n_g}\sum_{i=1}^{n_g} x_{g,i} \qquad g = 1,2 \\
s^2 &= \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(x_{1,i} - m_1)^2 + \sum_{j=1}^{n_2}(x_{2,j} - m_2)^2\right) \\
c &= \sqrt{\frac{n_1 n_2}{n_1 + n_2}}.
\end{aligned}$$

t- (and |t|-) Distribution



Q:
How does the distribution of t (no absolute value taken) look?

If the data are

s a

'*t*-distribution' with parameter n_1+n_2 (a.k.a. degrees of freedom)

Figure 6.8: The null distribution of the (absolute) *t*-statistic determined by simulations – namely, by random permutations of the group labels.

Comments and Pitfalls

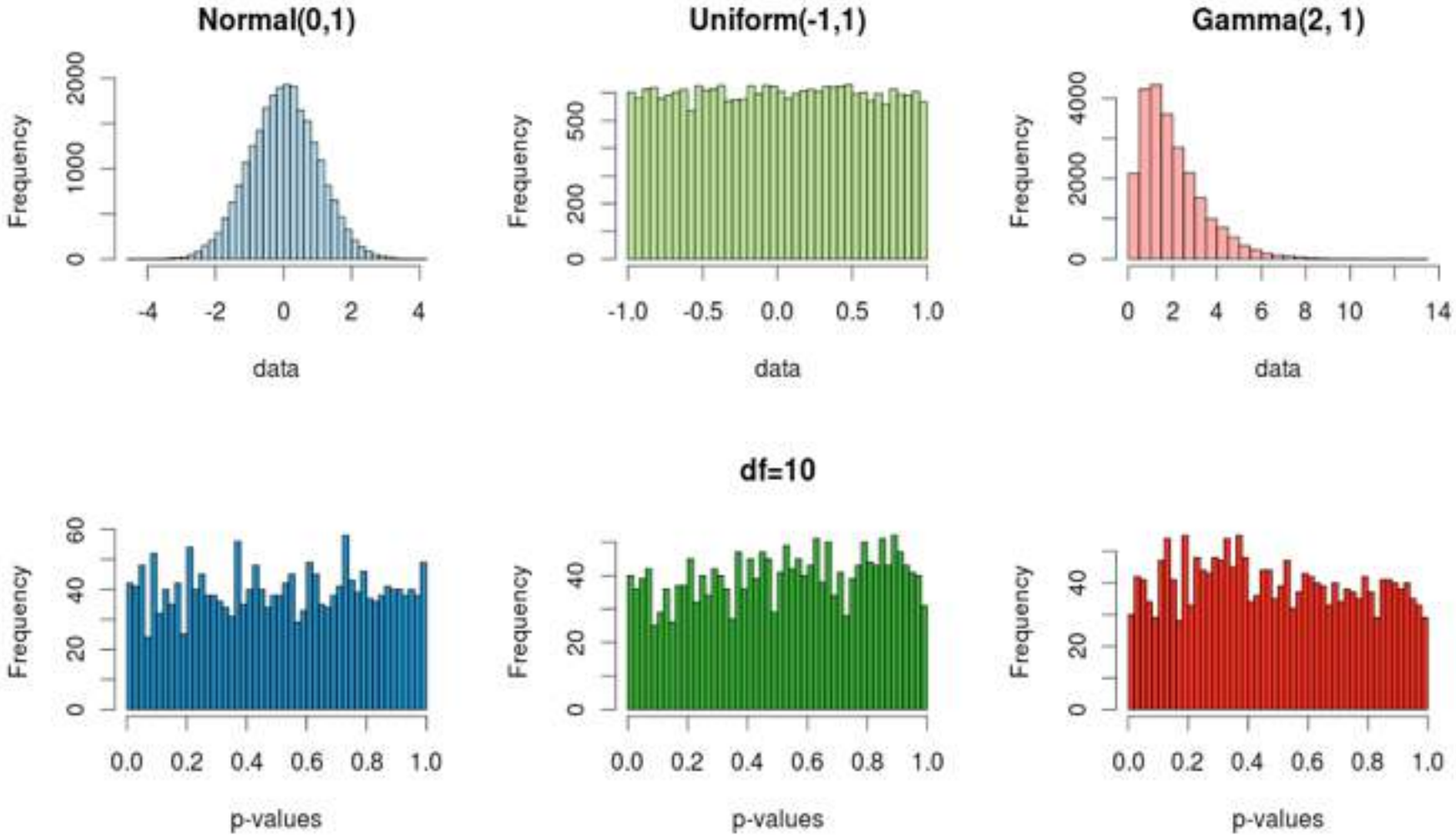
The derivation of the *t*-distribution assumes that the observations are independent and that they follow a Normal distribution.

Deviation from Normality - heavier tails: test still maintains type-I error control, but may no longer have optimal power.

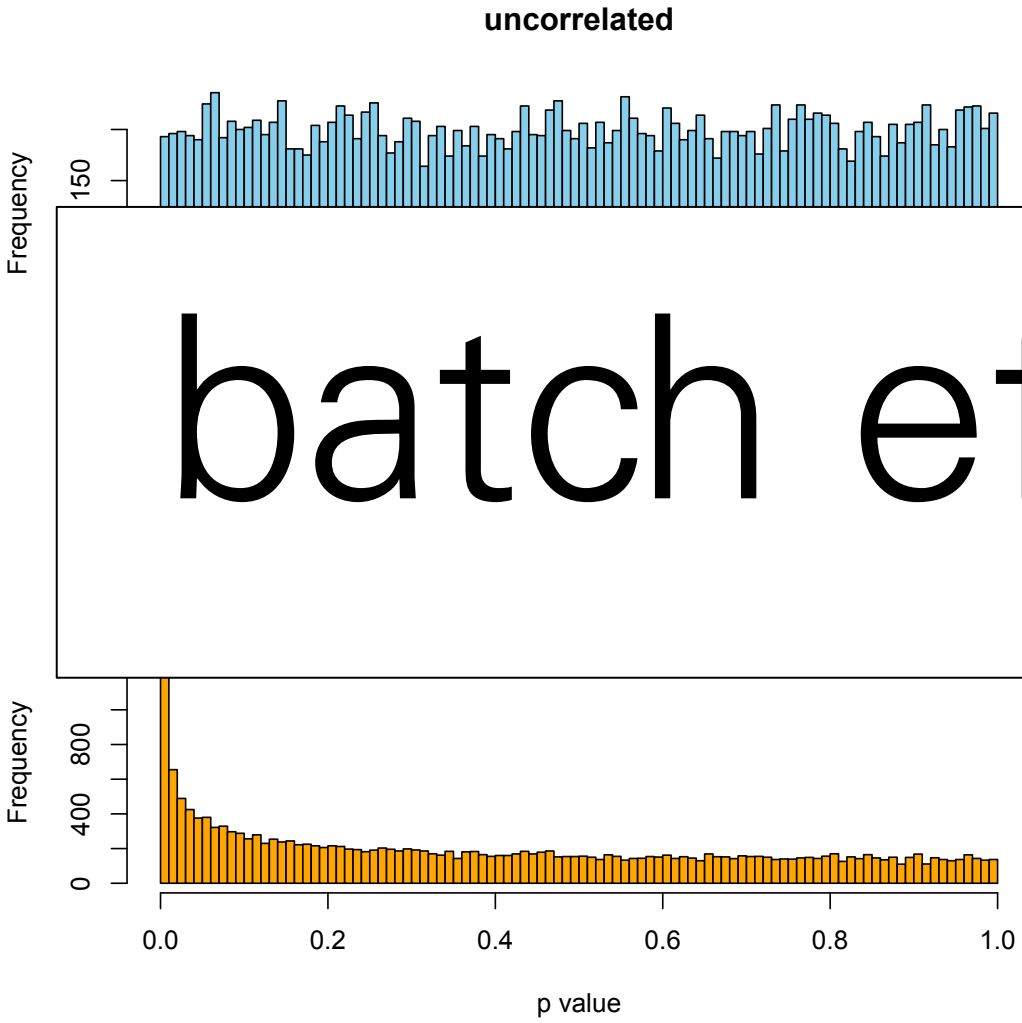
Options: use permutations; transform (e.g. ranks - Wilcoxon test)

If the data are dependent, then p-values will likely be totally wrong (e.g., for positive correlation, too optimistic).

Different Data Distributions – Independent Case



t-Test Looses Error Control
if Independence Assumption Does not Hold



batch effects!

```
library("mvtnorm")
library("genefilter")

p = 30      ## number of samples
```

```
= s))
```

```
for(i in seq(along=tt))
    hist(tt[[i]]$p.value, breaks=100, col=c("skyblue", "orange")[i],
         main=names(tt)[i], xlab="p value")
```

Avoid Fallacy

The p-value is the probability that the data could happen, under the condition that the null hypothesis is true.

It is not the probability that the null hypothesis is true.

Absence of evidence ≠ evidence of absence

Recap: Single Hypothesis Testing

p-values are random variables: uniformly distributed if the null hypothesis is true - and should be close to zero if the alternative holds.

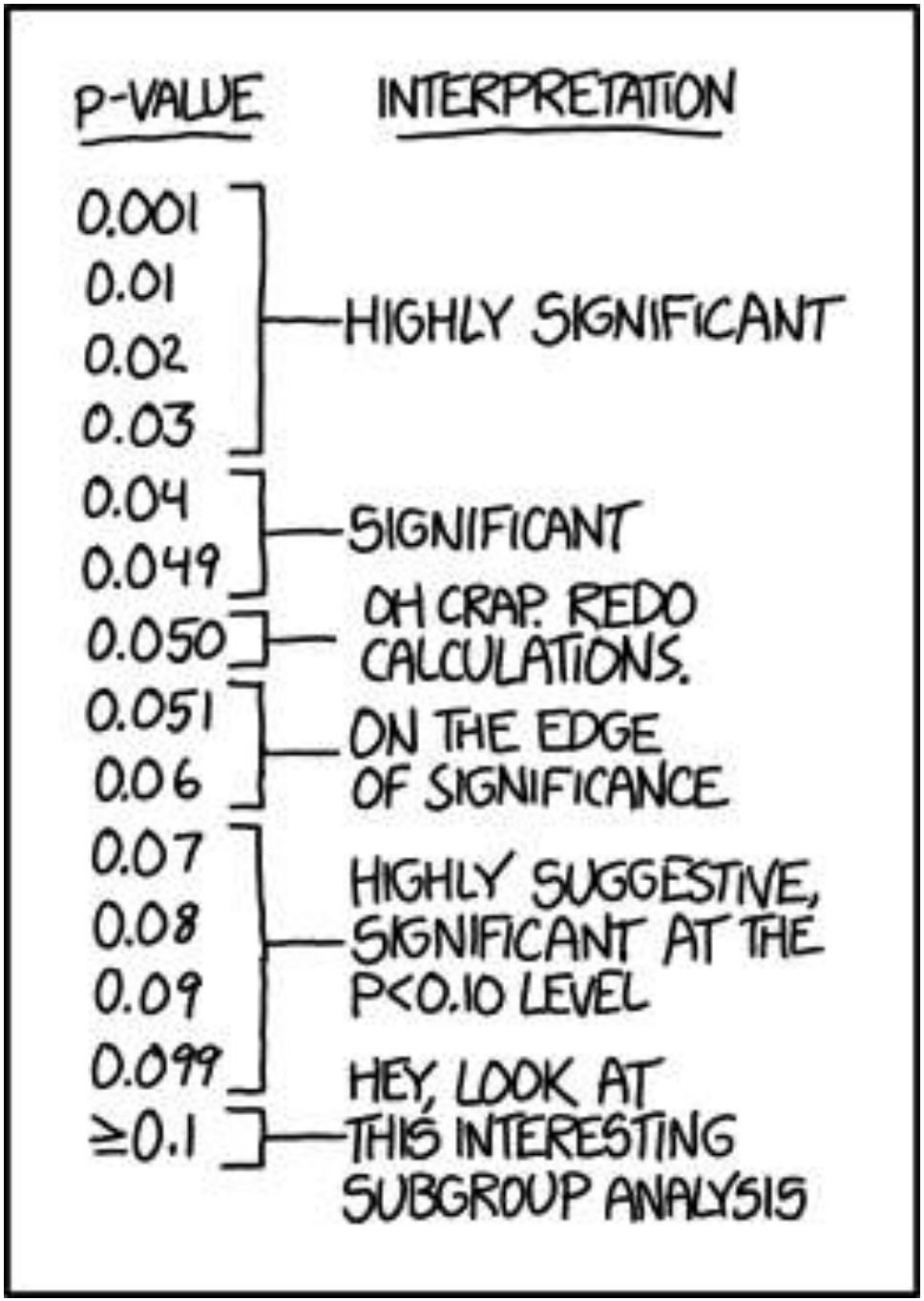
Note: We only observe one draw.

We prove something by disproving ('rejecting') the opposite (the null hypothesis)

Not rejecting does not prove the null hypothesis

Repeating the experiment (under the null): Around 5% of the times the p-value will be less than 0.05 by chance

All this reasoning is probabilistic. Testing & p-values are for rational decision making in uncertain contexts.


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

What is p-Value Hacking ?

On the same data, try different tests until one is significant

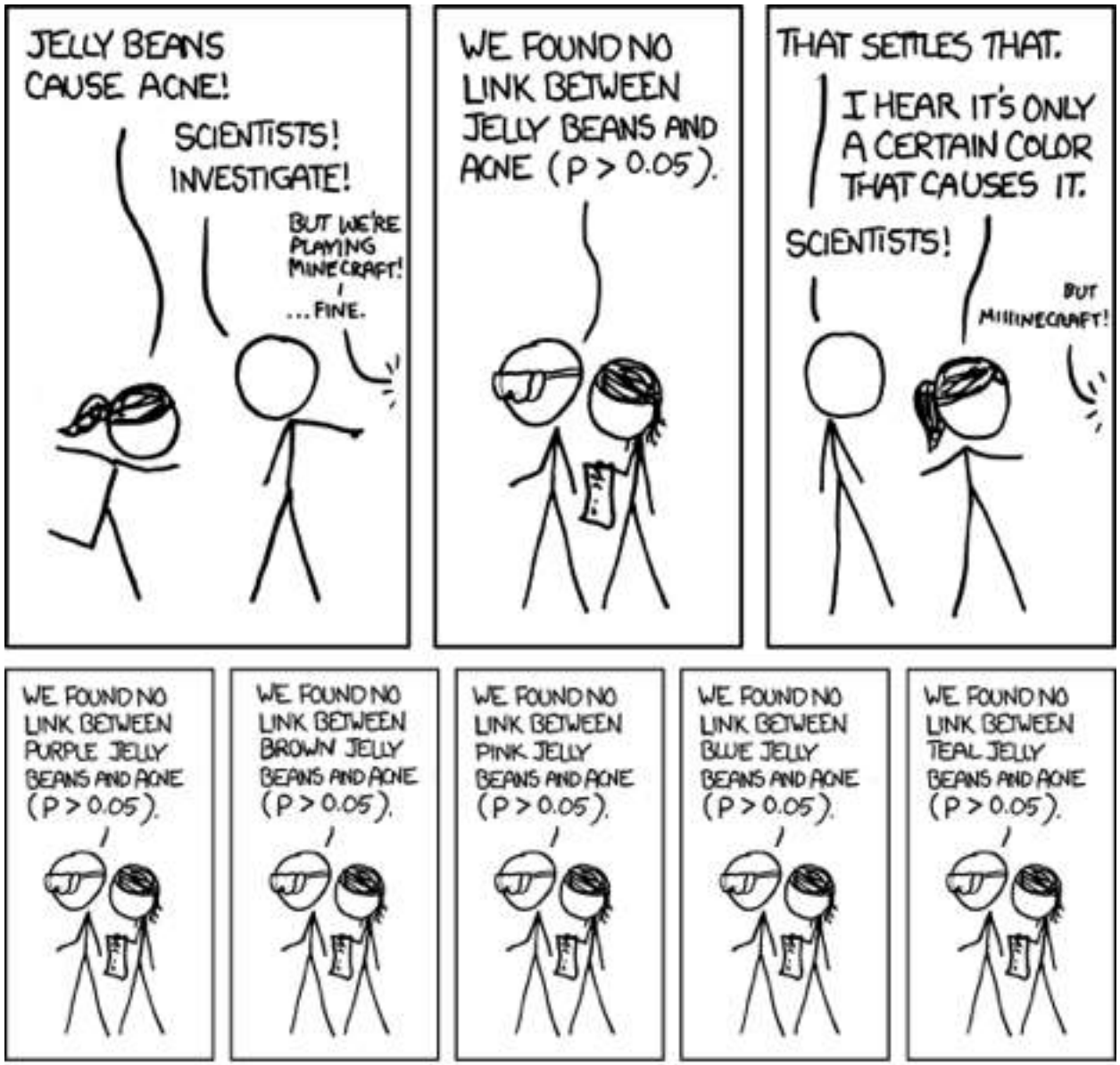
On the same data, try different hypotheses until one is significant (HARKing - hypothesizing after results are known)

Moreover…:
retrospective data picking
'outlier' removal
the 5% threshold and publication bias

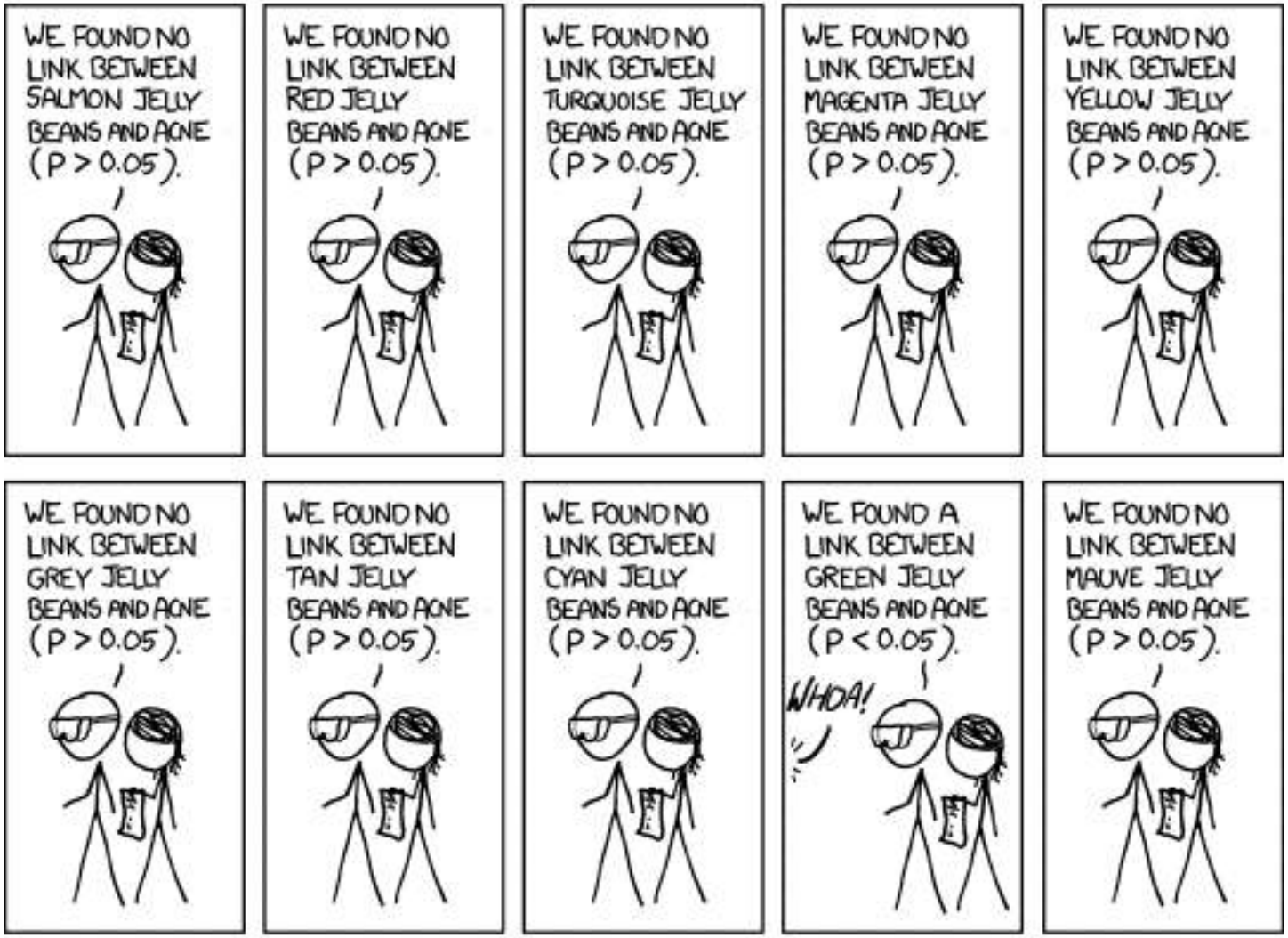
The ASA's Statement on p-Values: Context, Process, and Purpose
Ronald L. Wasserstein & Nicole A. Lazara DOI: 10.1080/00031305.2016.1154108

What can we do about this?

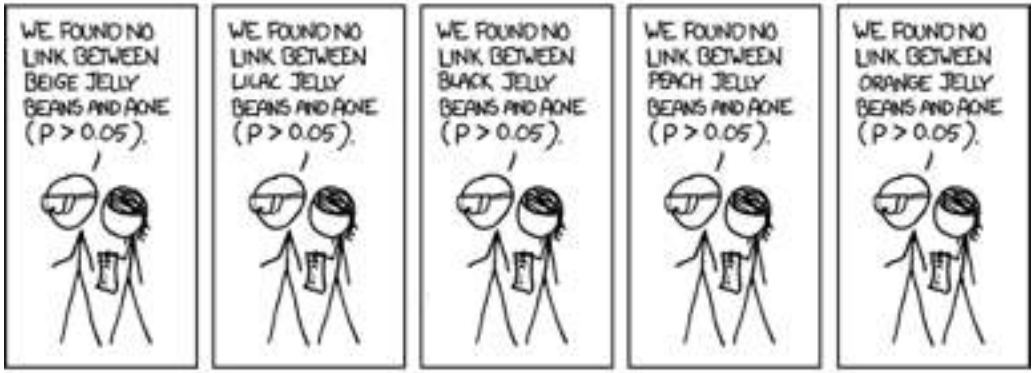
Multiple Testing

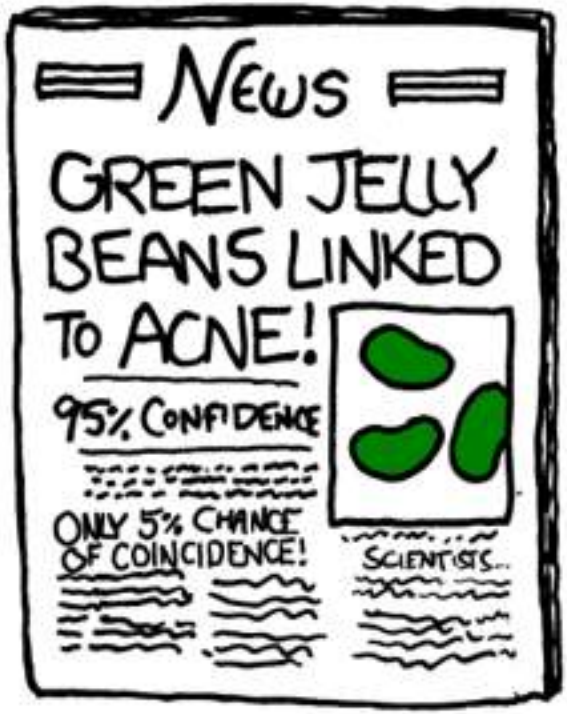


Multiple Testing



Multiple Testing

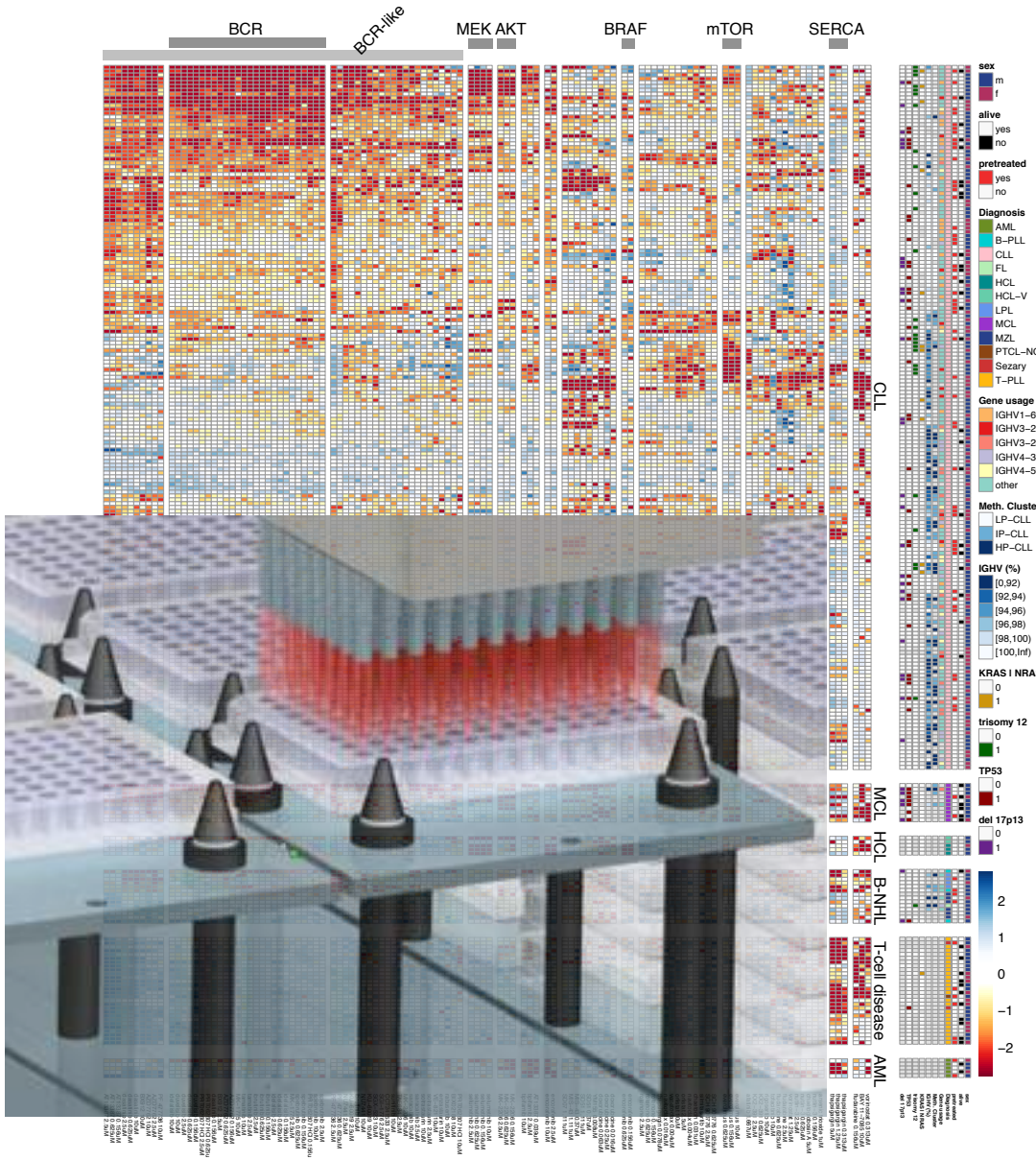




Multiple Testing

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

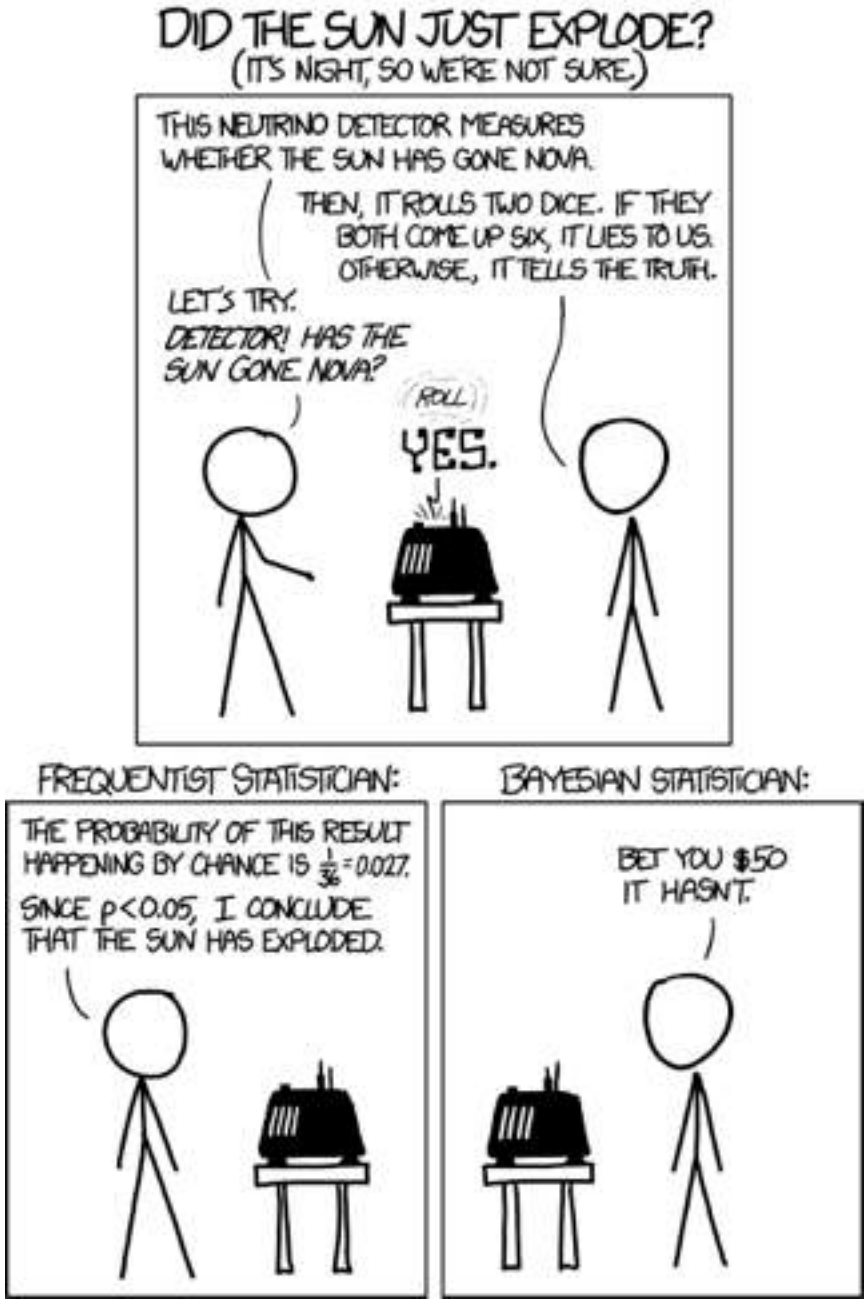


The Multiple Testing Burden

When performing several tests, type I error goes up: for α = 0.05 and *n* indep. 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



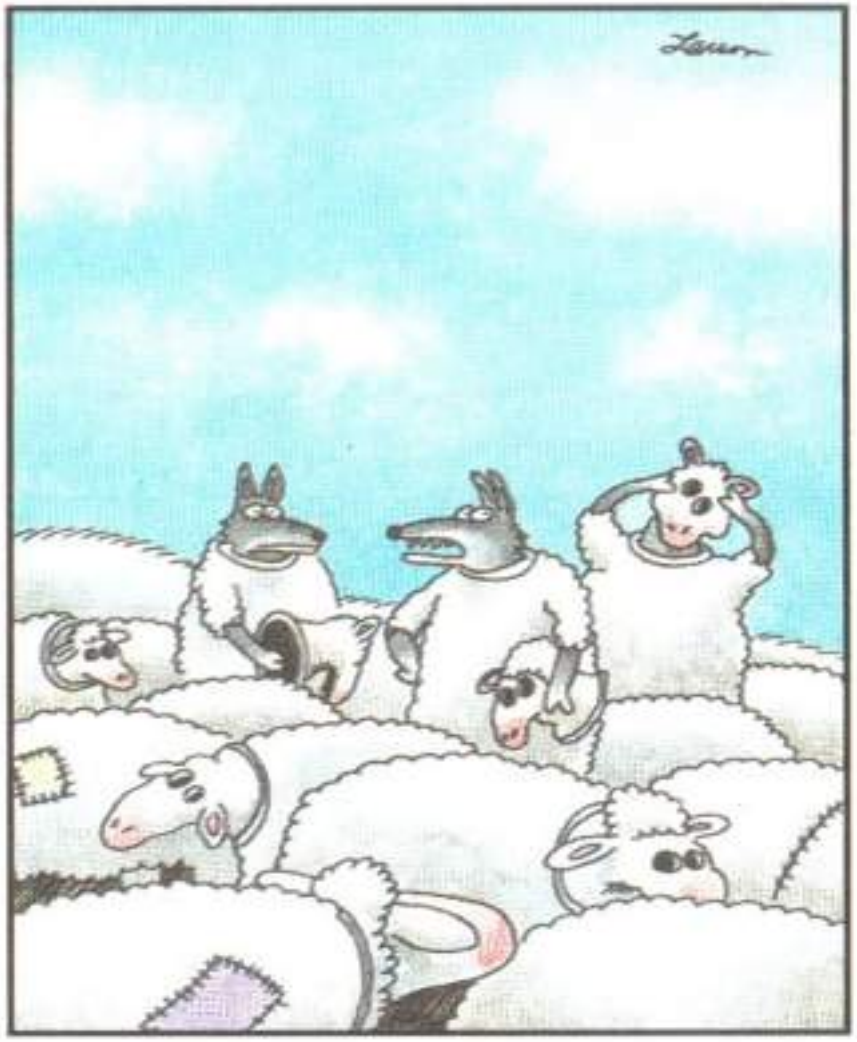
False Positive Rate and False Discovery Rate

FPR: fraction of FP among all genes (etc.) tested

FDR: fraction of FP among hits called

Example:
20,000 genes, 100 hits, 10 of them wrong.

FPR: 0.05%
FDR: 10%



"Wait a minute! Isn't anyone here a real sheep?"

Experiment-Wide Type I Error Rates

Test vs Reality	**Null Hypothesis is true**	**. . . is false**	**Total**
Rejected	V	S	R
Not rejected	U	T	$m - R$
Total	m_0	$m - m_0$	m

- m: total number of hypotheses
- m_0: number of null hypotheses
- V: number of false positives (a measure of type I error)

Family-wise error rate (FWER): The probability of one or more false positives, P(V > 0). For large m_0, this is difficult to keep small.

False discovery rate (FDR): The expected fraction of false positives among all discoveries, E[V / max {R, 1}].

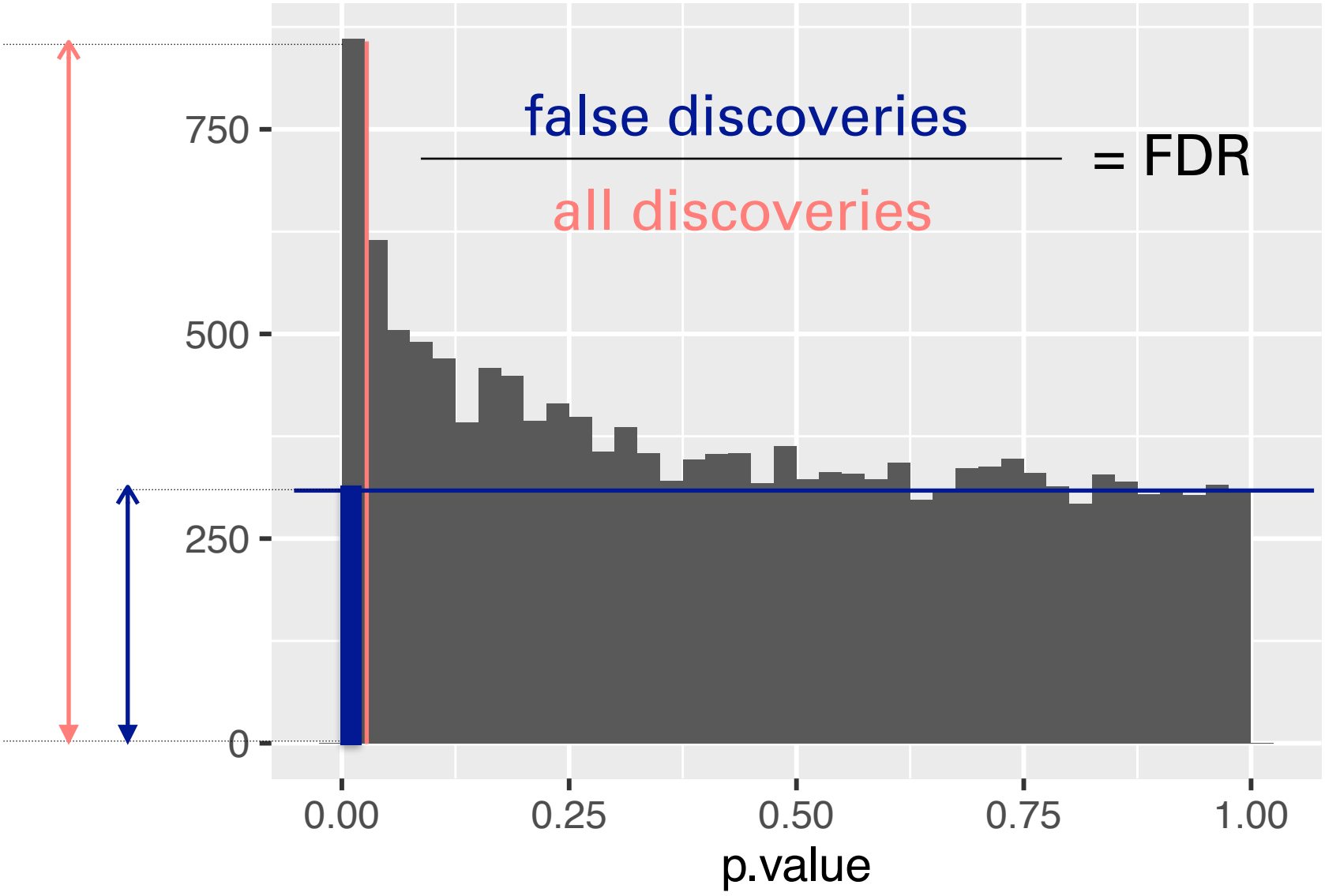
Bonferroni Correction

For m tests, multiply each p-value with m.

Then see if anyone still remains below α.

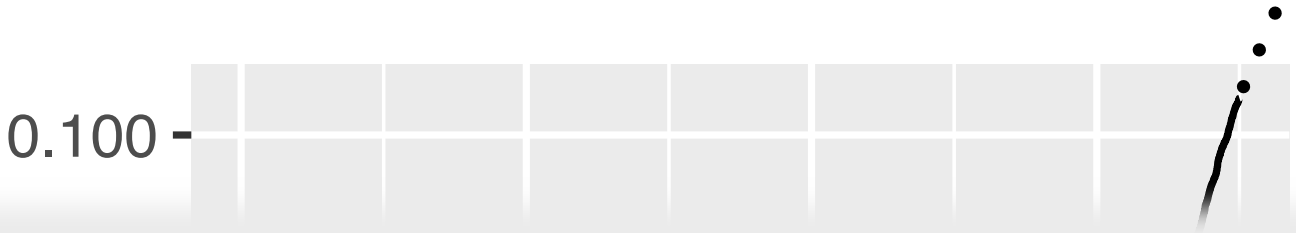
False Discovery Rate

False Discovery Rate



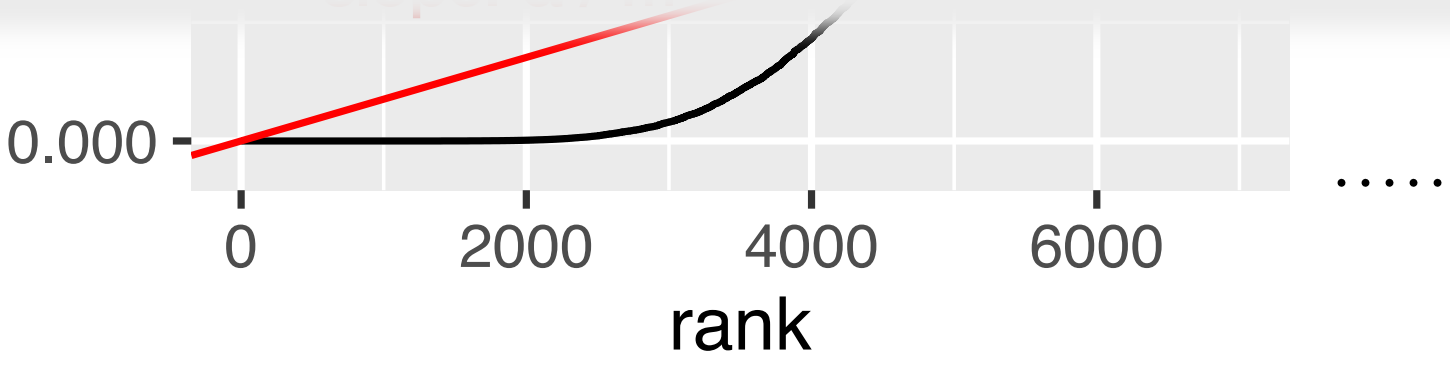
Method of Benjamini & Hochberg (1995)

Method of Benjamini & Hochberg

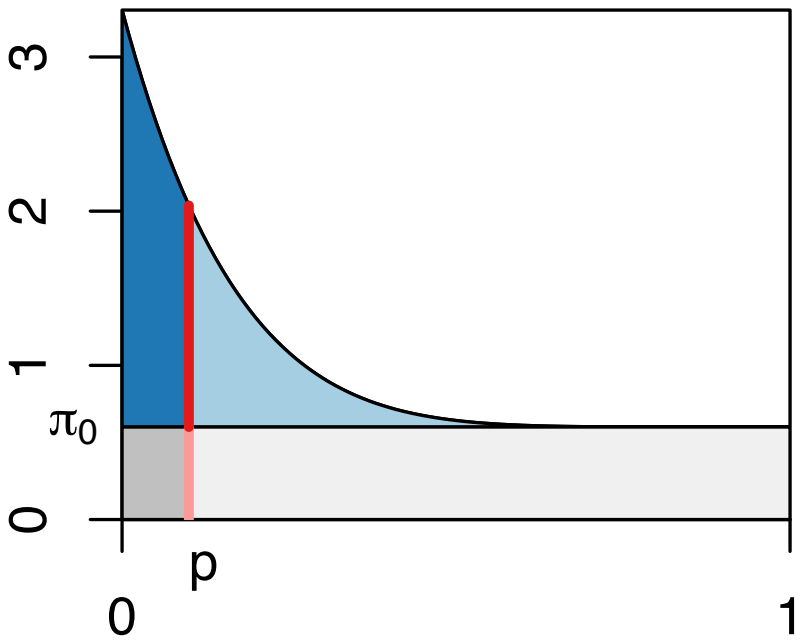



```
BH = {
        i <- length(p):1
        o <- order(p, decreasing = TRUE)
        ro <- order(o)
        pmin(1, cummin(n/i * p[o]))[ro]
    }
```

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



The Two-Groups Model and the Local False Discovery Rate



$$f(p) = \pi_0 + (1 - \pi_0) f_{\text{alt}}(p)$$

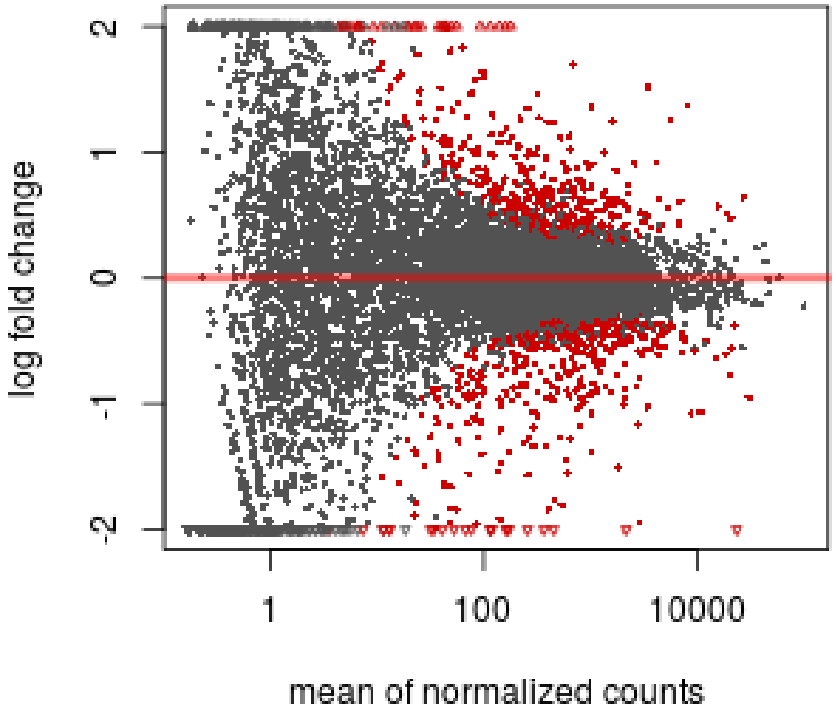
$$\text{fdr}(p) = \frac{\pi_0}{f(p)}$$

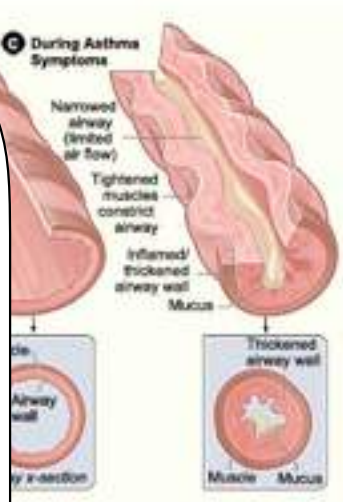
FDR: a set property. A single number that applies to a whole set of discoveries.

fdr: a local property. It applies to individual hypothesis.


Exchangeability?

Data set 1: RNA-Seq





ression analysis:

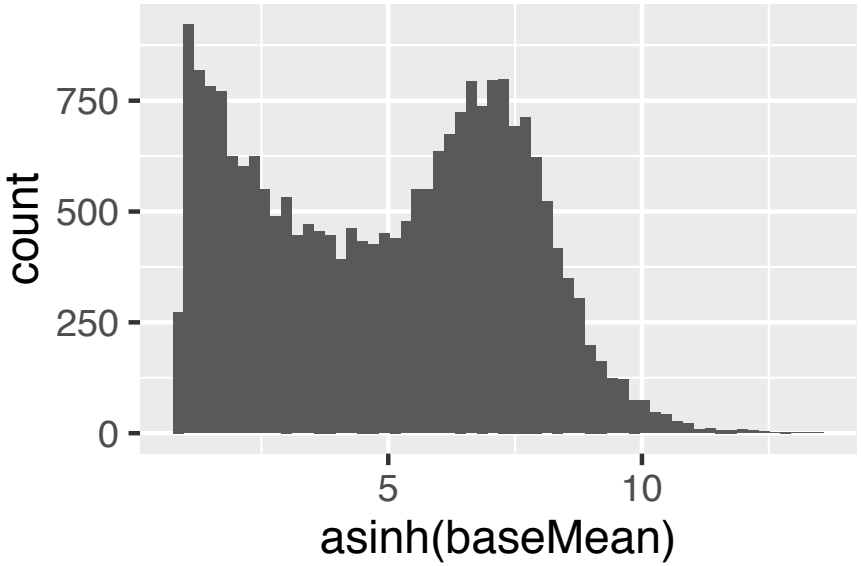
$= \mu_{ij}, \text{dispersion} = \alpha_j)$

```
N061011    trt
```

```
design <-  ~ cellline + dexamethasone
```

Himes et al. “RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene that Modulates Cytokine Function in Airway Smooth Muscle Cells.” PLoS One. 2014 GEO: GSE52778.

Not all Hypothesis Tests are Created Equal



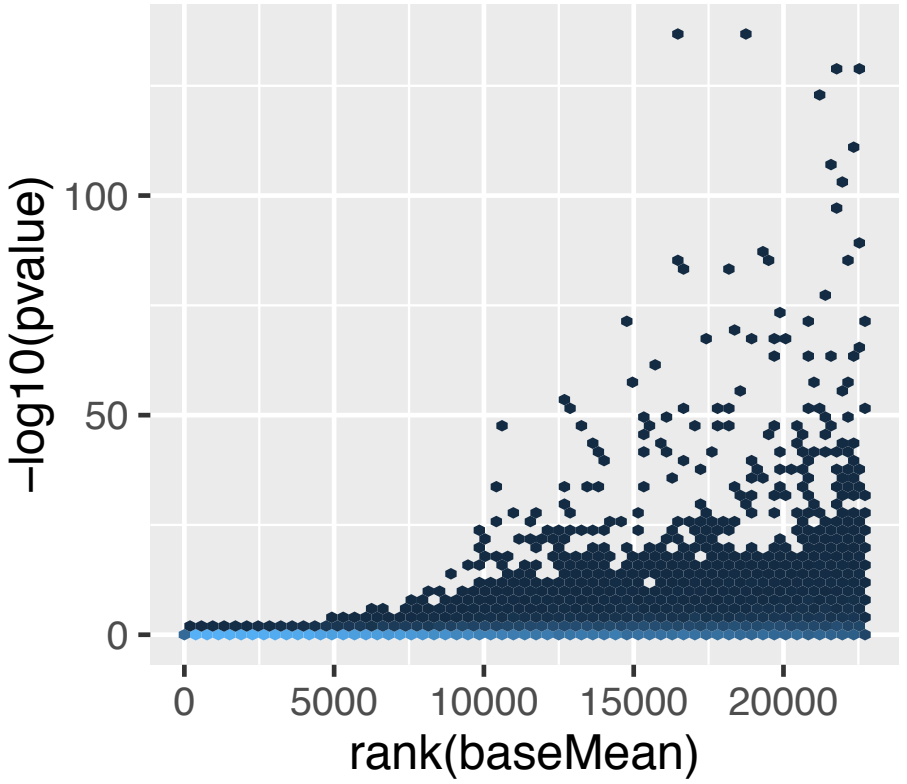


Figure 6.15: Histogram of `baseMean`. We see that it covers a large dynamic range, from close to 0 to around 3.3×10^5.

Covariates - examples

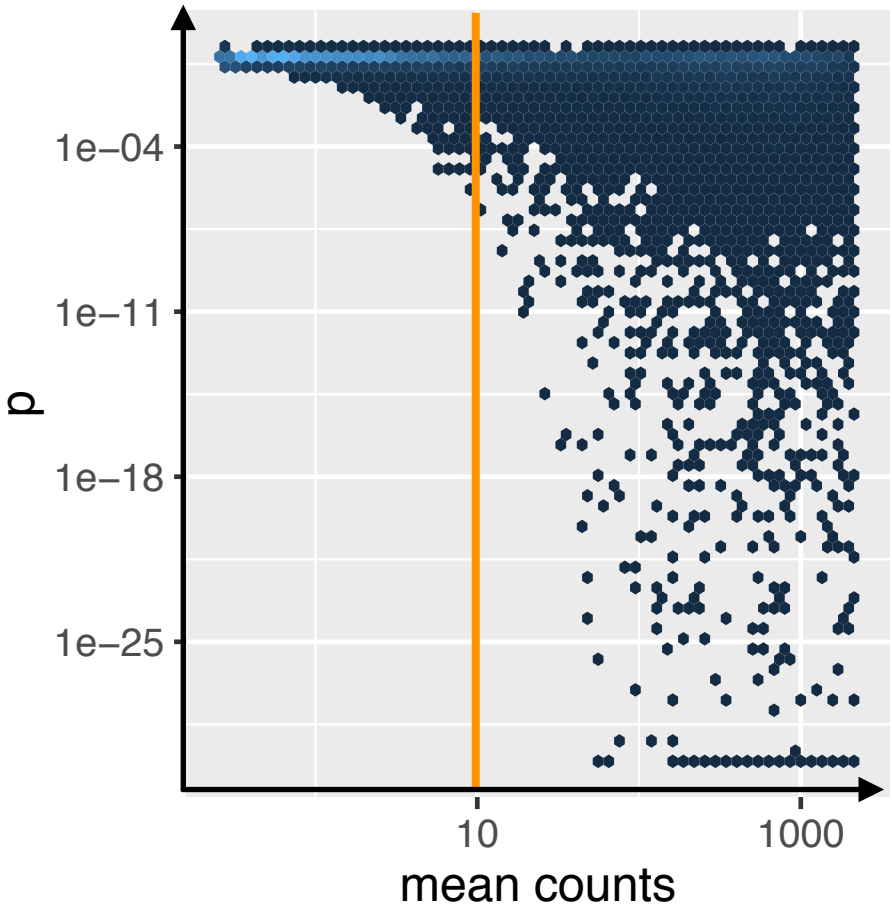
Application	Covariate
Differential RNA-Seq, ChIP-Seq, CLIP-seq, ...	(Normalized) mean of counts for each gene
GWAS	Minor allele frequency
eQTL analysis	SNP – gene distance
t-tests	Overall variance
Two-sided tests	Sign
All applications	Sample size; measures of signal-to-noise ratio

Independent Filtering

Two steps:

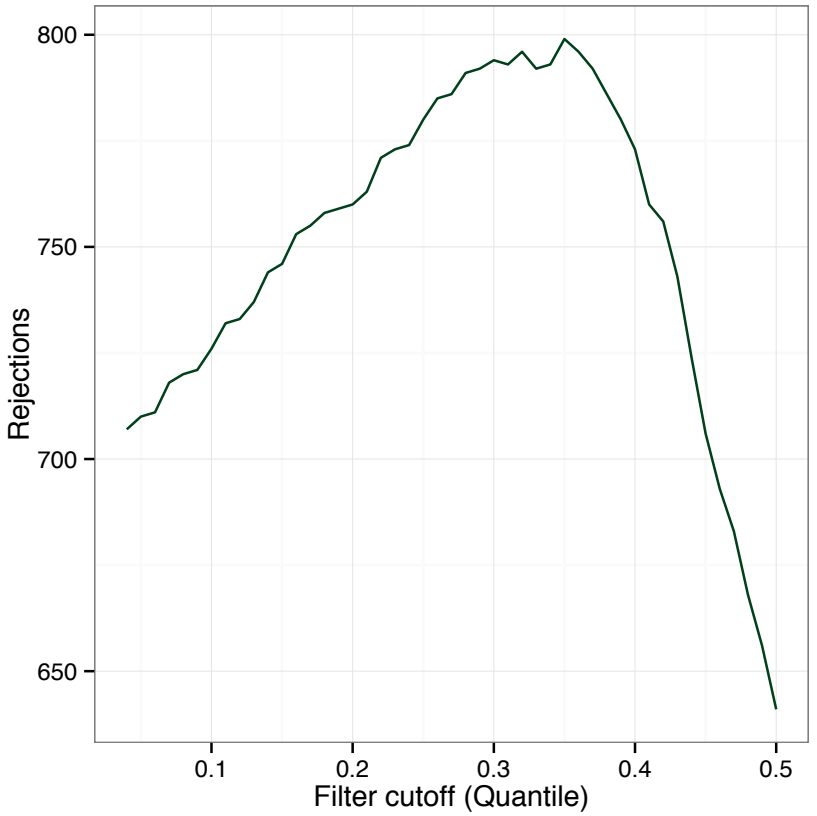
- All hypotheses H_i with $X_i < x$ get filtered.
- Apply BH to remaining hypotheses.

(Bourgon, Gentleman, Huber *PNAS* 2010)



Data-driven choice of filtering threshold

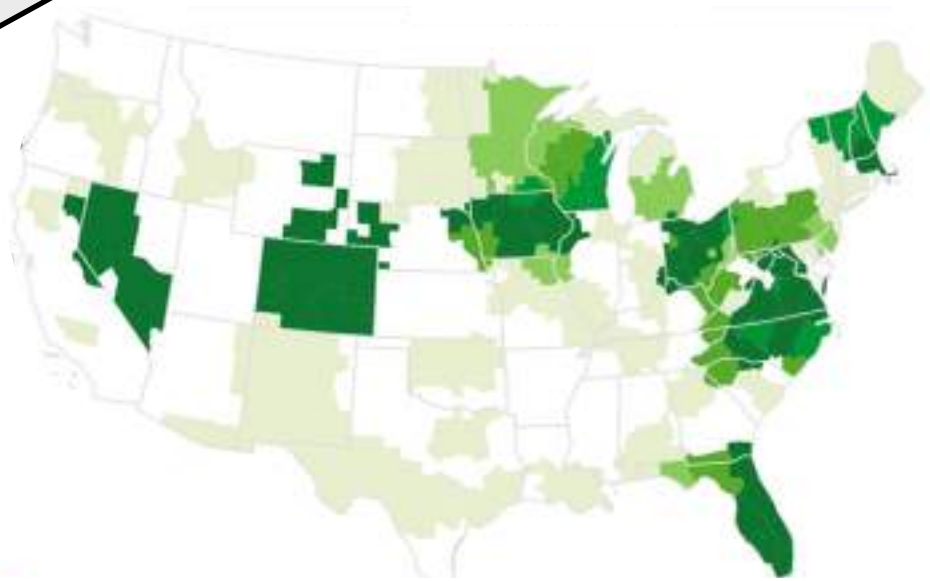
- Do Independent Filtering followed by Benjamini-Hochberg procedure with all possible thresholds.
- Report the result with the optimal threshold.
- We have been doing this in *DESeq2* for the last two years.



Weighted Benjamini-Hoc[illegible] method

- Let $w_i \geq 0$ an[illegible] budget").
- Define [illegible]
- A[illegible] P_i.
- R[illegible] (FDR) control (Genovese, Roeder,

Problem: how to know the weights?



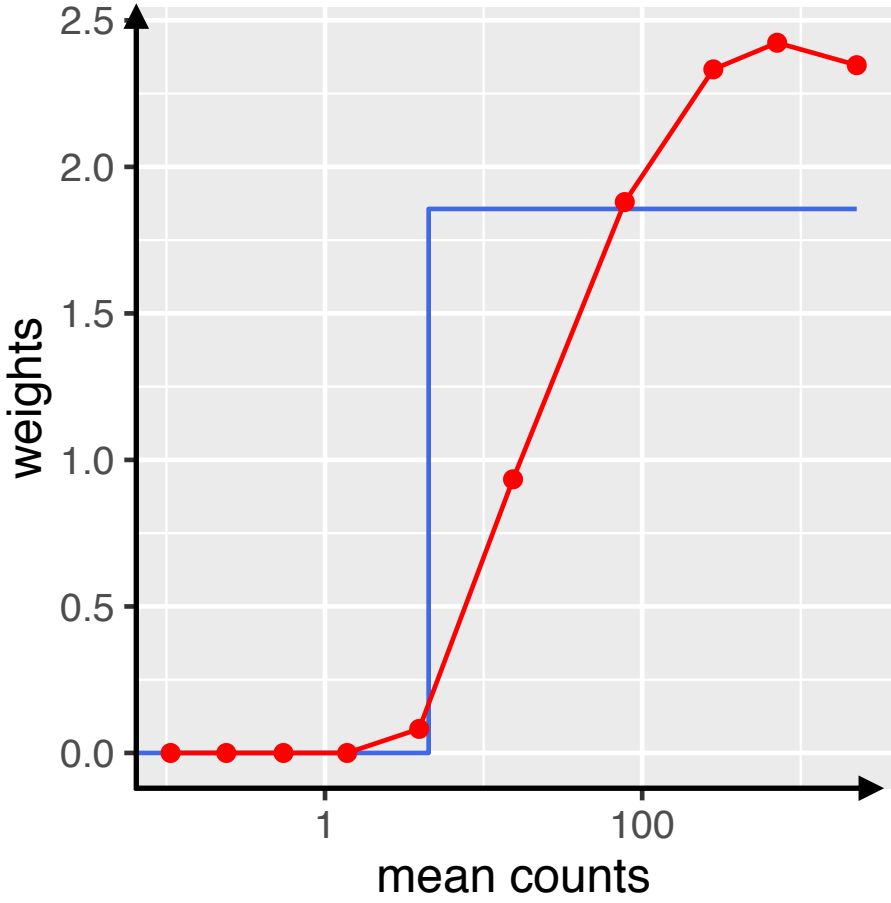
http://www.washingtonpost.com/wp-srv/special/politics/track-presidential-campaign-ads-2012/

'alpha investing'

Independent filtering is a special case of weighted BH

$S =$ set of hypotheses retained by filtering step

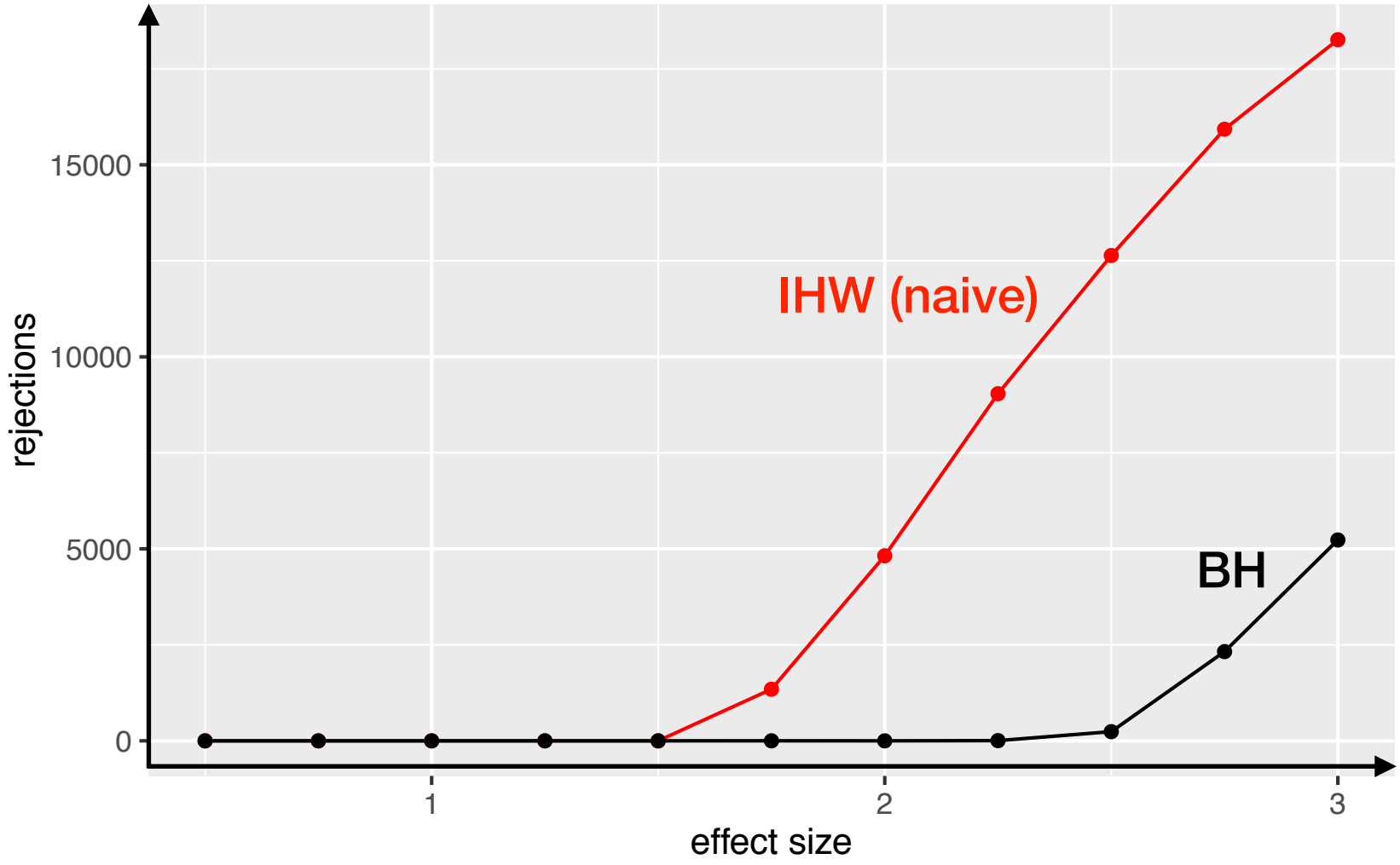
$$w_i = \begin{cases} m/|S| & \forall i \in S \\ 0 & \forall i \notin S \end{cases}$$



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

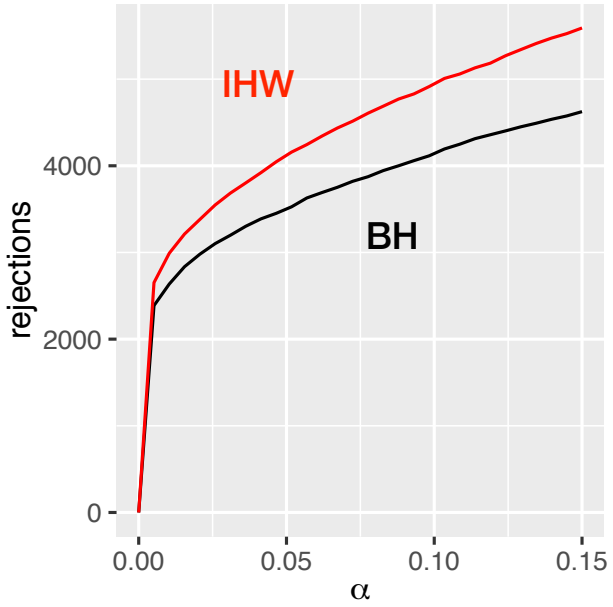
- Stratify the tests into G bins, by covariate X
- Choose α
- For each possible weight vector $\mathbf{w} = (w_1, \ldots, w_G)$ apply weighted BH procedure. Choose $\mathbf{w}$ that maximizes the number of rejections at level α.
- Report the result with the optimal weight vector $\mathbf{w}^*$.

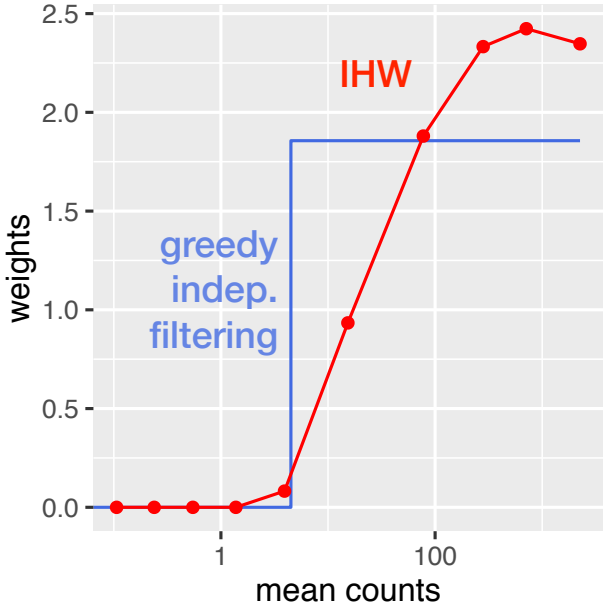
IHW (naive) is powerful (t-test simulation)

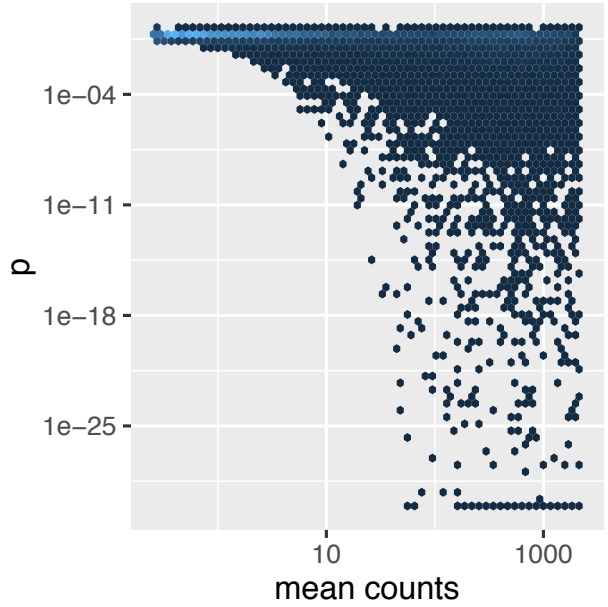


$$m = 500,000 \quad m_1 = 20,000 \quad \alpha = 0.1 \quad n = 2 \times 4$$

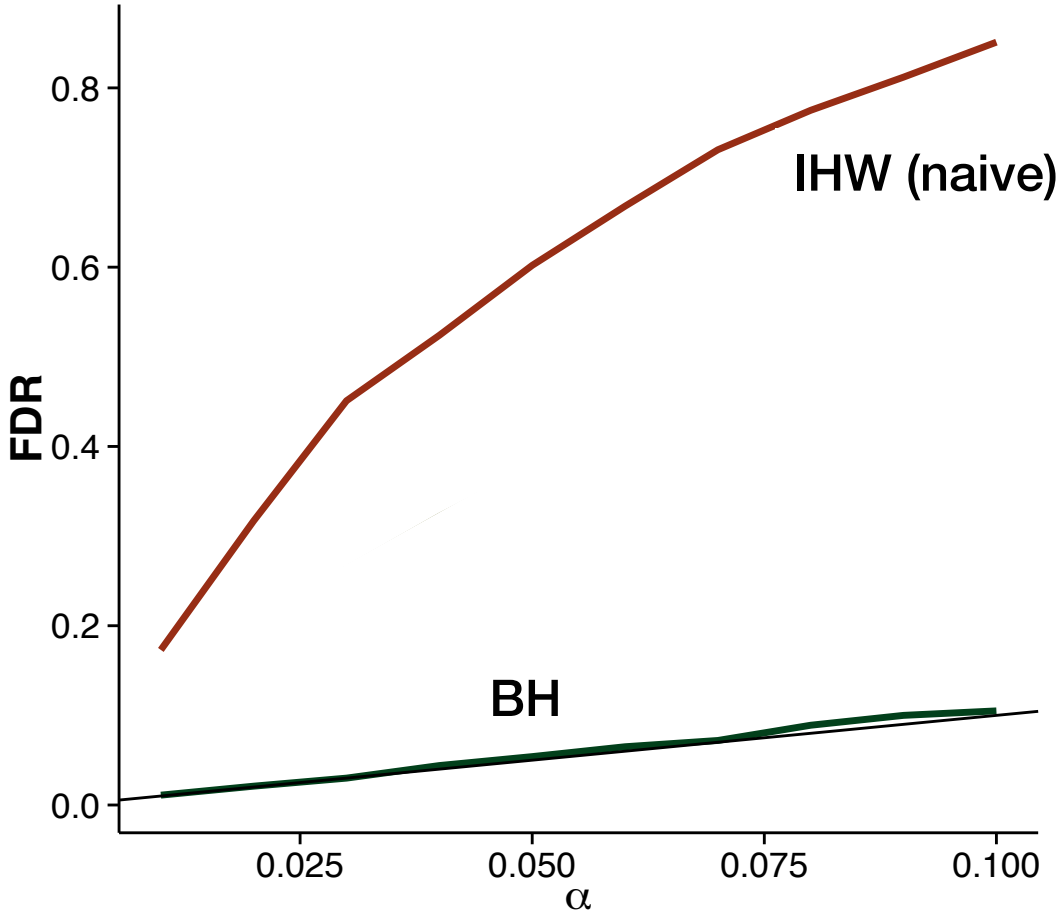
RNA-Seq example (DESeq2)







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



Modified IHW


Nikos Ignatiadis

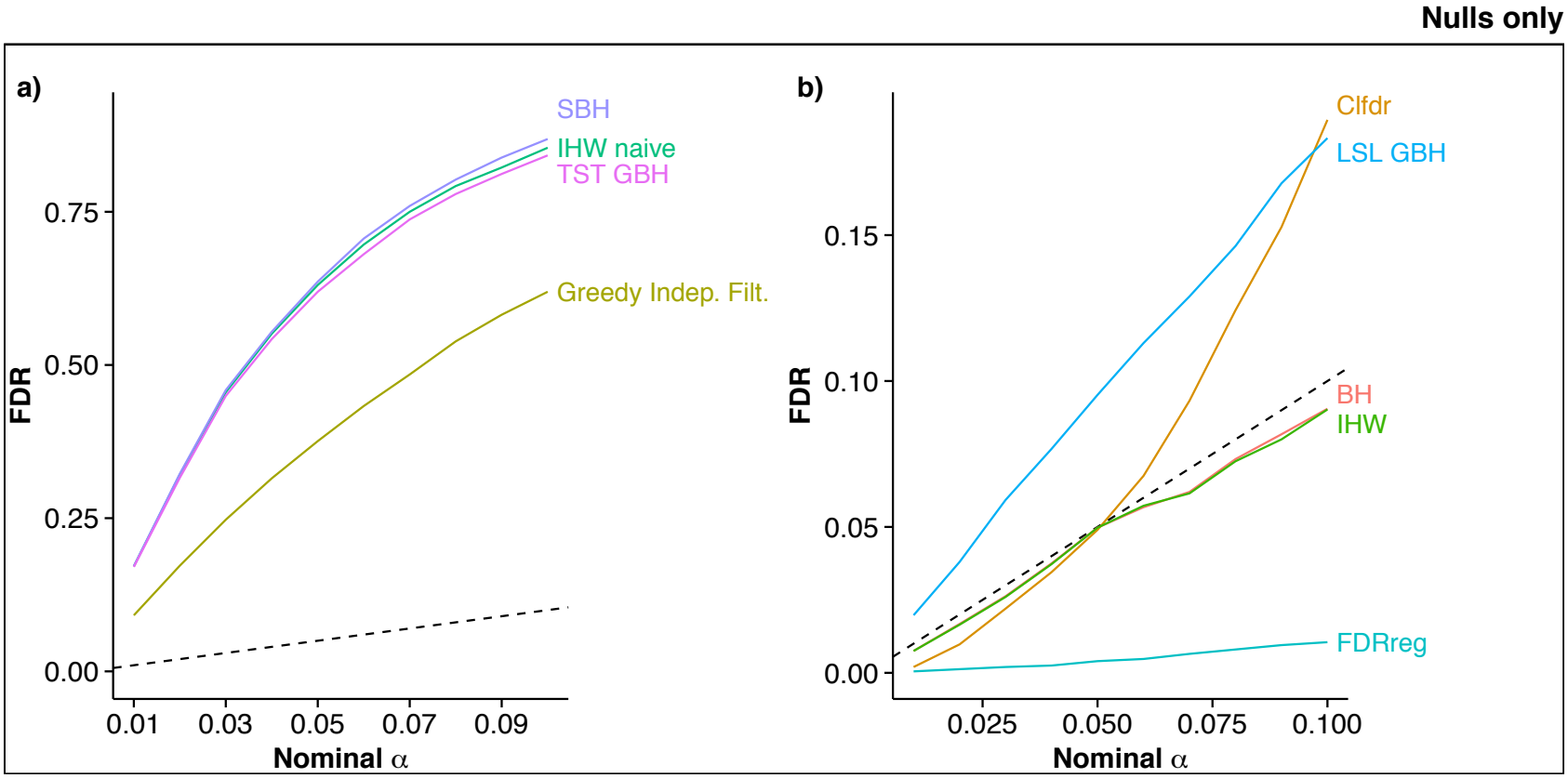
'Pre-validation': randomly split hypotheses into *k* folds. Learn weights for the hypotheses in a fold from the other *k–1* folds

Regularisation:

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

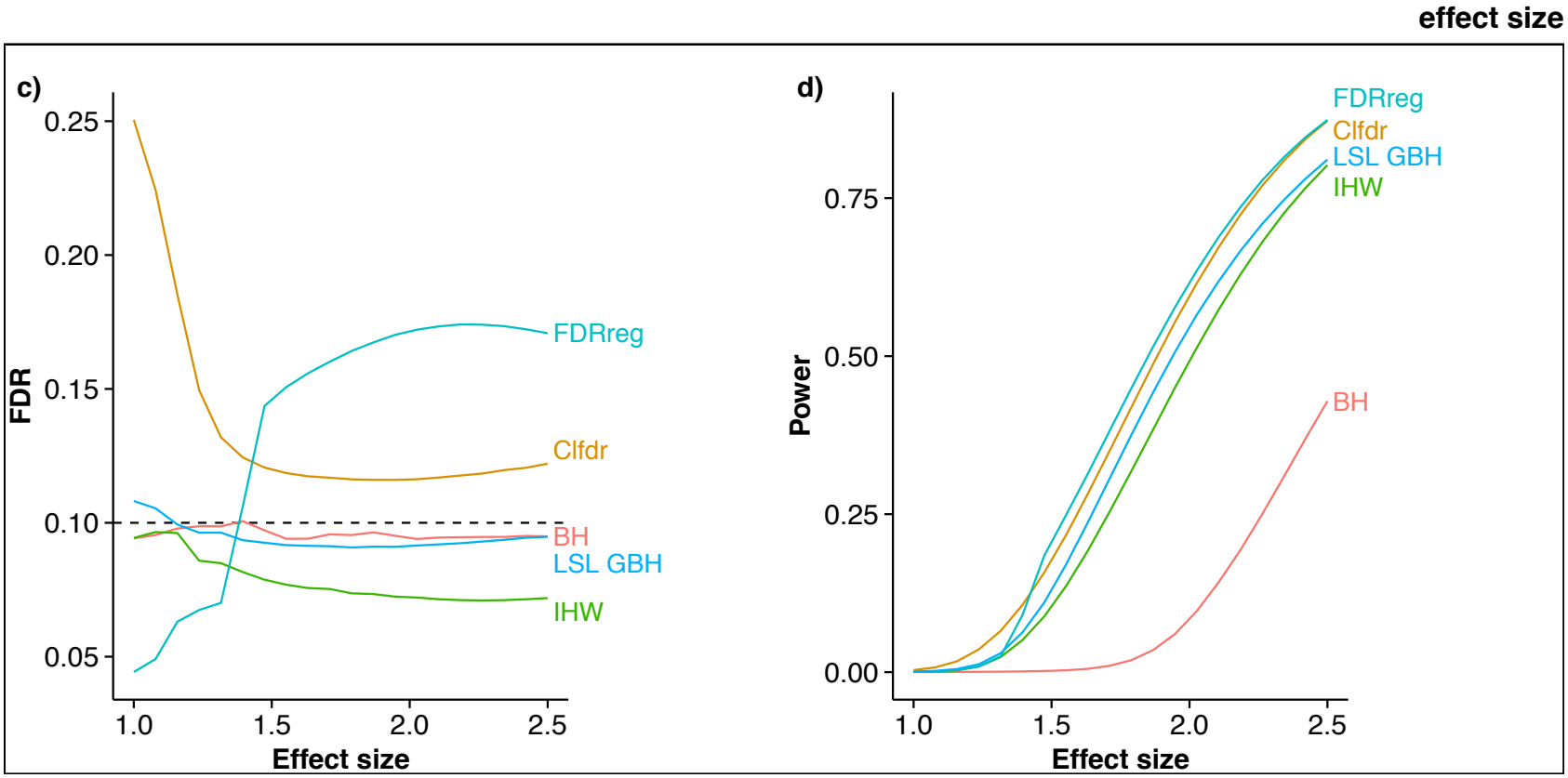
Convex relaxation: for weight optimisation (only), replace ECDFs of the p-values with Grenander estimators (least concave majorant of the ECDF)

IHW controls FDR



SBH: stratified BH (e.g. Yoo, Bull, …Sun, Genet. Epidem 2010)
GBH: grouped BH (Hu, Zhao, Zhou, JASA 2010)
Clfdr: conditional local fdr (Cai, Sun, JASA 2009)
FDRreg (J. Scott JASA 2015)

IHW controls FDR and is powerful



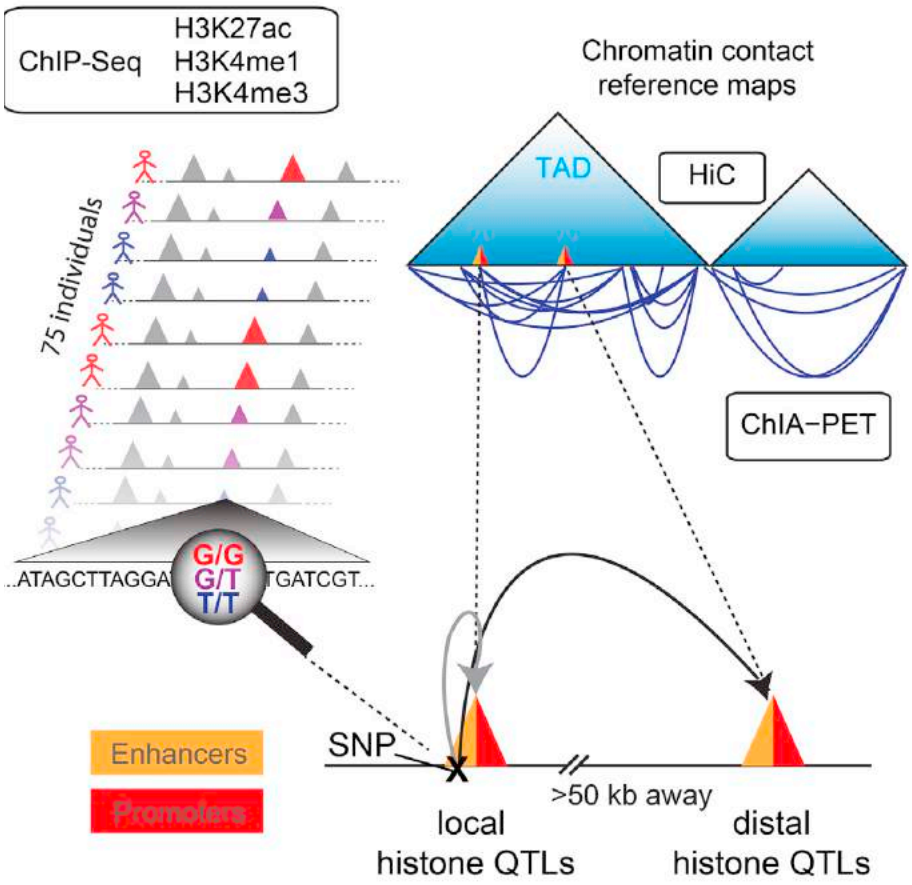
Data set 2: hQTL

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.

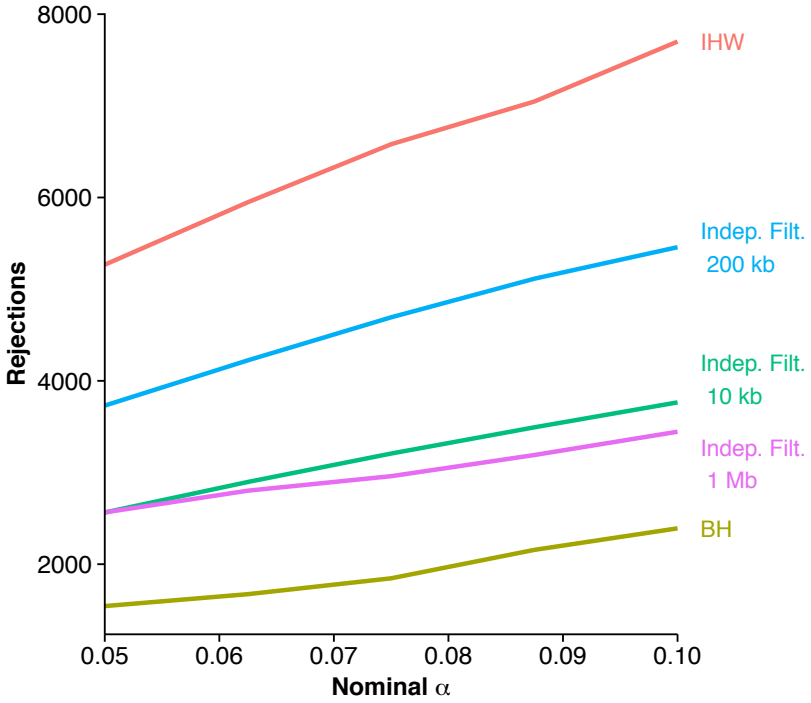
14,142 local hQTLs linked to ~10% of H3K27ac peaks (FDR 10%, permutations)

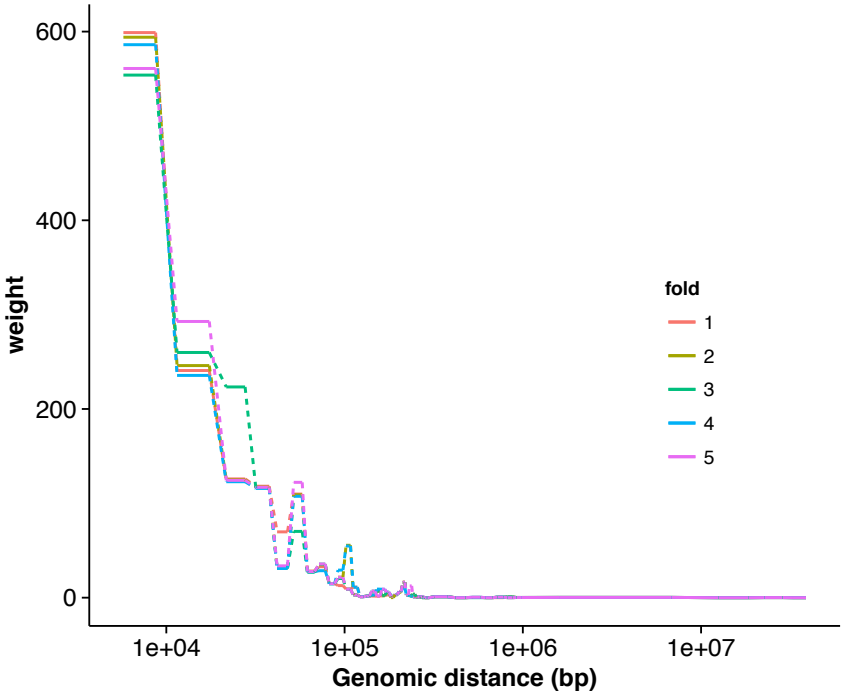
Distal: distance cutoffs from 50 to 300 kb; also HiC



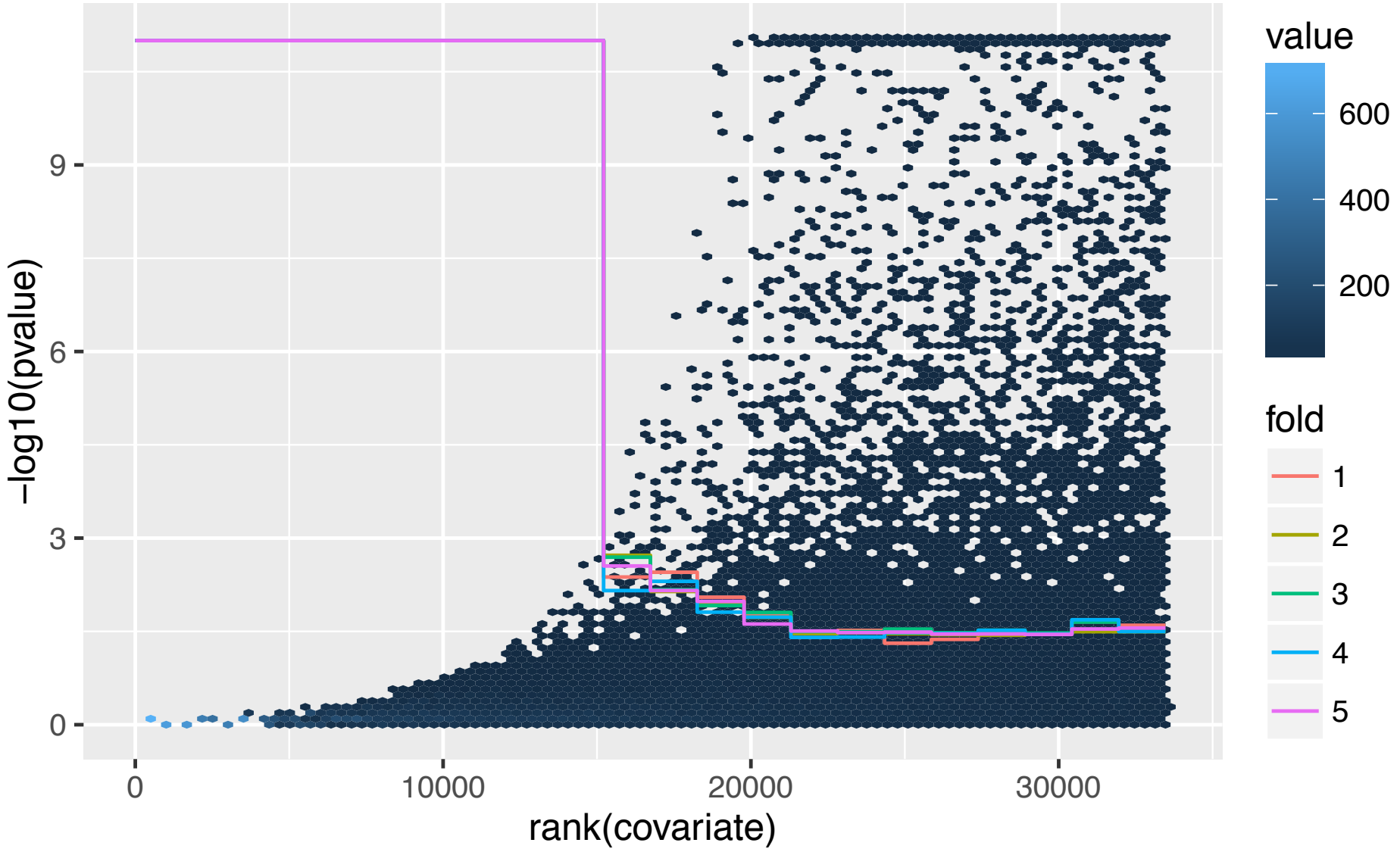
Grubert, Zaugg, Kasowski, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell (2015).

histone-QTL example: H3K27ac





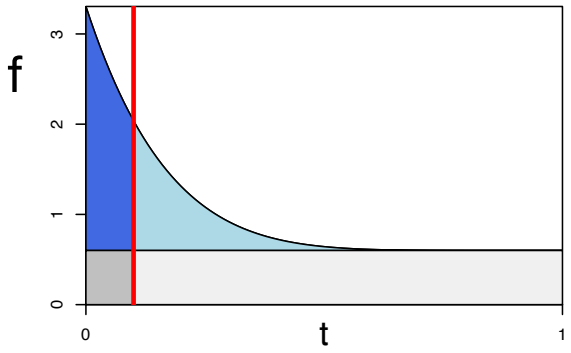
2D decision boundaries



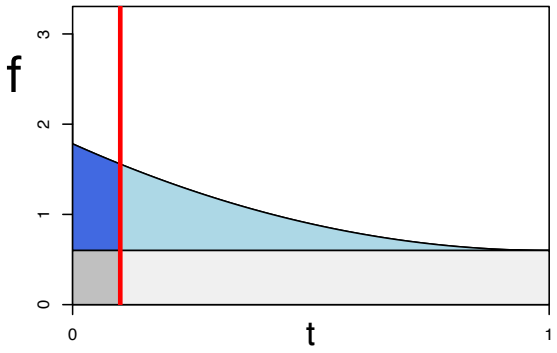
Why does IHW work?

Rank (and reject) hypotheses by local true (false) discovery rate, not by p-value

$$
\begin{aligned}
f(t) &= \pi_0 + (1-\pi_0) f_{\text{alt}}(t) \\
\text{fdr}(t) &= \frac{\pi_0}{f(t)} \\
\text{tdr}(t) &= 1 - \text{fdr}(t)
\end{aligned}
$$

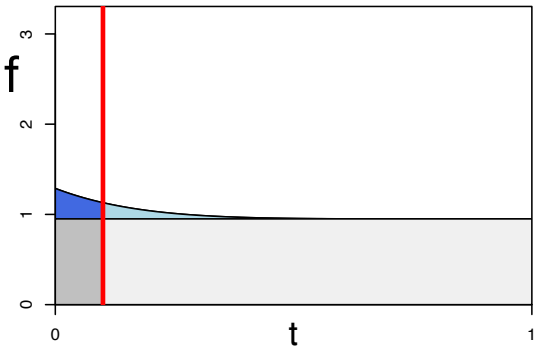


$\pi_0 = 0.6$



$\pi_0 = 0.6$

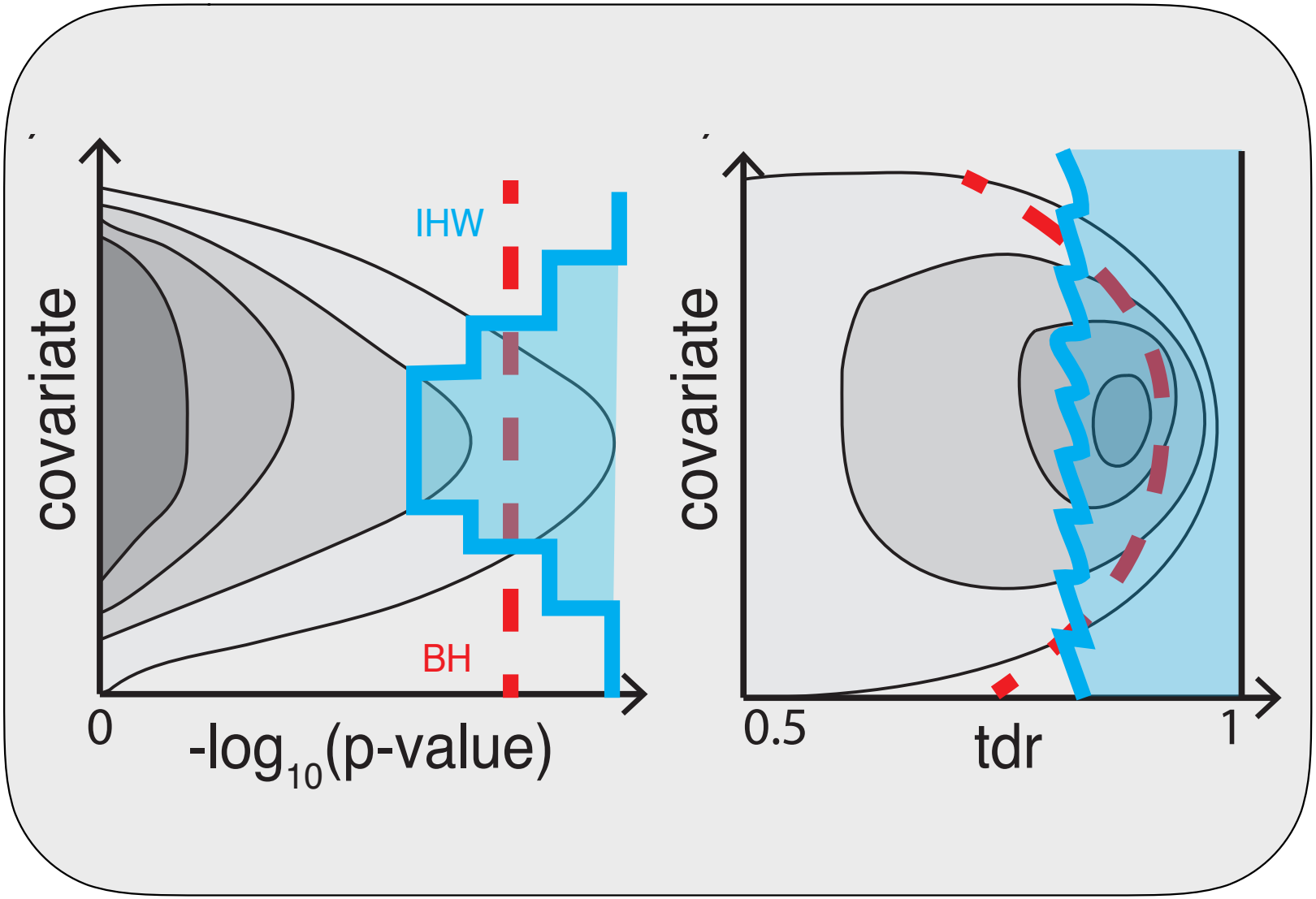
different f_{alt}



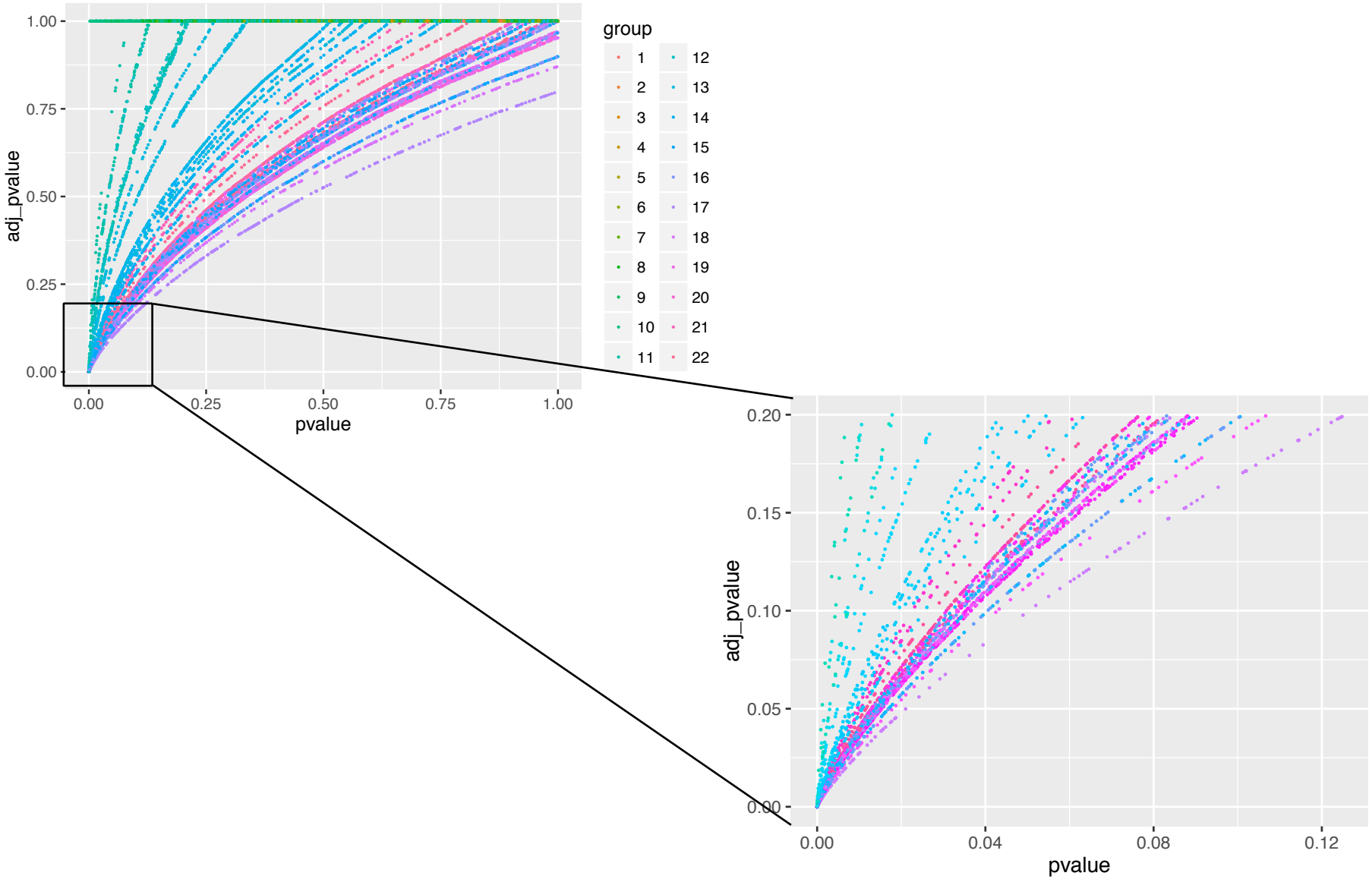
$\pi_0 = 0.95$

same f_{alt}

2D decision boundaries



Ranking is not monotonous in raw p-values



Formal results

IHW is asymptotically consistent: it controls the FDR at the nominal level α as the number of hypotheses becomes large.

Proof: generalisation of Storey, Taylor, Siegmund, JRSSB (2004)

Variant “IHW-Bonferroni” has finite sample FWER control

Availability

Paper in Nature Methods

Bioconductor - IHW (devel...

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

Search

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | Ab

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 man
- Workflows for learning and
- Course and conference mat
- Videos.
- Community resources and tutorials.

R / CRAN

Support

Please rea
questions
the followi

- Support
- Bioc-dev

Joint work with

Nikos Ignatiadis

Bernd Klaus

Judith Zaugg

Thanks also to

Robert Gentleman

Richard Bourgon

Misha Savitski

Oliver Stegle

Vlad Kim

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
- Controlling ‘overoptimism’

Simone Bell
Dorothee Childs
Sascha Dietrich
Julian Gehring
Nikos Ignatiadis
Vlad Kim
Bernd Klaus
Junyan Lu
Andrzej Oles
Malgorzata Oles
Aleks Pekowska
Alejandro Reyes
Thomas Schwarzl
Mike Smith
Britta Velten

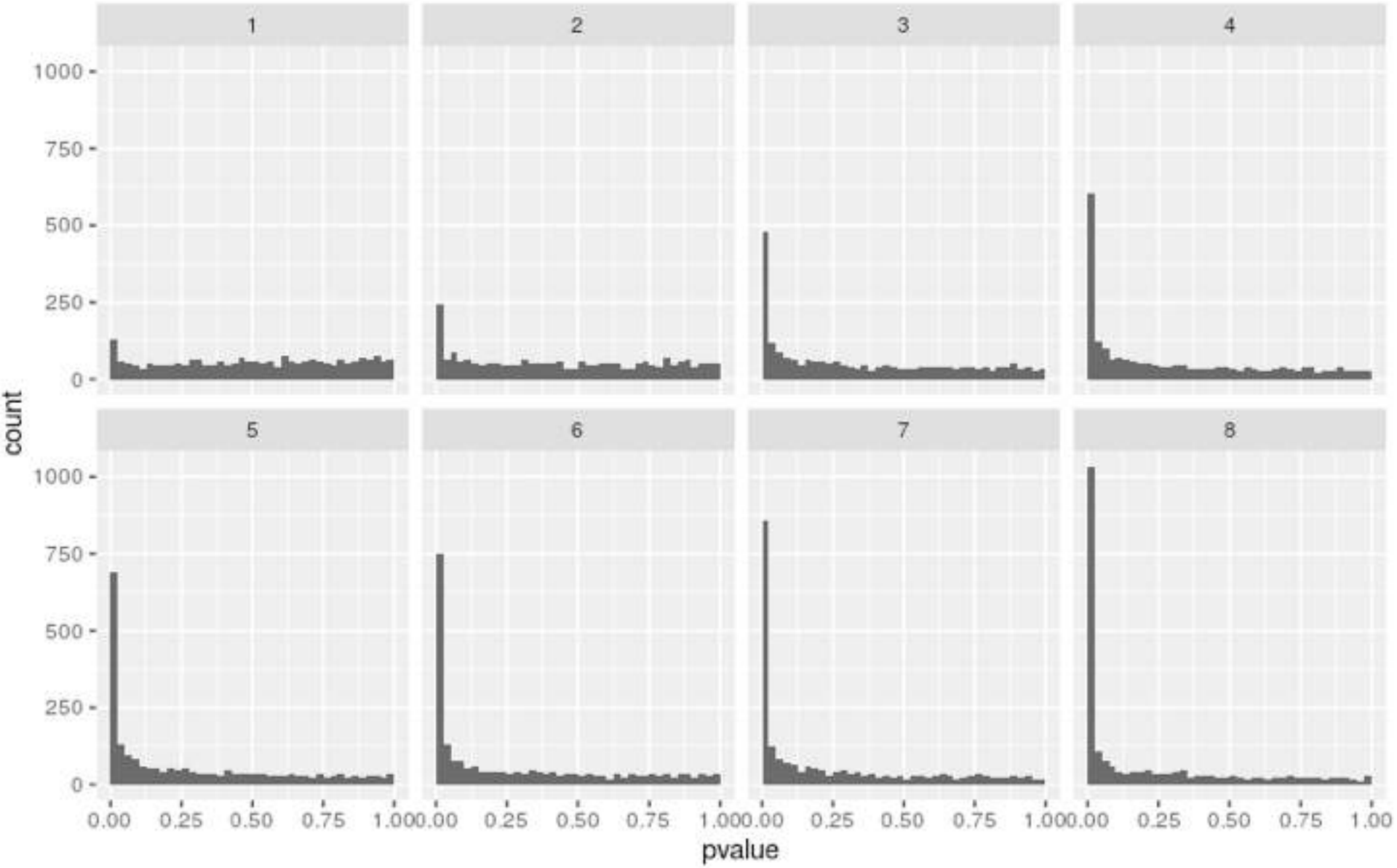
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

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)
Judith Zaugg (EMBL)
Mikhail Savitski (EMBL)
Susan Holmes (Stanford)

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

EMBL

RNA-Seq p-value histogram stratified by average read count



STUCK IN A DULL, LOW PAYING JOB?
WANT TO MAKE ***BIG MONEY?***

BE A QUANTUM MECHANIC!

. . . EVEN IF YOU NEVER FINISHED HIGH SCHOOL!

STUDY AT HOME!

THE COLUMBIA INSTITUTE OF QUANTUM MECHANICS, INC.

Not affiliated with the Columbia Broadcasting System, Columbia University, the District of Columbia, or Columbia, Gem of the Ocean.

CUT OUT AND SEND

☐ Yes! I want to get in on the ground floor of this exciting new field. I understand no salesman will call.

NAME _____

ADDRESS _____

CITY, STATE, ZIP _____

COLUMBIA INSTITUTE OF QUANTUM MECHANICS
Suite 293, 1100 Back St., Improvidence, RI 02904