

# Hypothesis Testing

Wolfgang Huber, EMBL



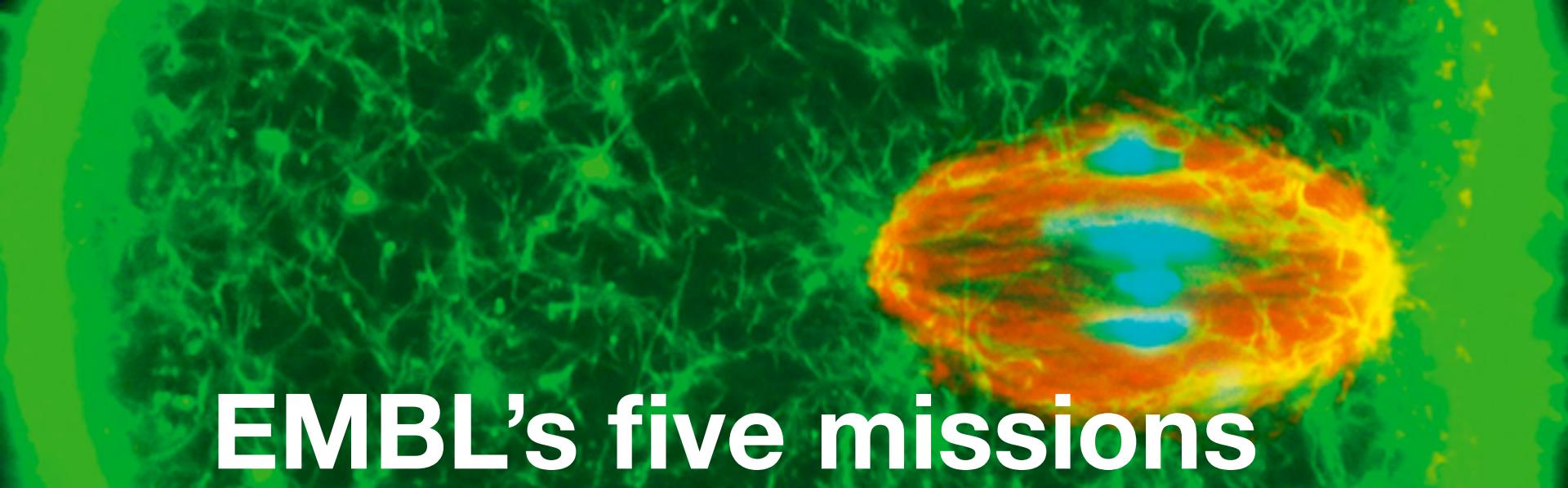
# European Molecular Biology Laboratory (EMBL)



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 ( $\supset$ 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

# What Can You Do at EMBL ?

Biology  
Chemistry  
Medicine  
Physics  
Mathematics  
Informatics  
Engineering



Internships - Phd programme - Postdocs - PIs - Jobs

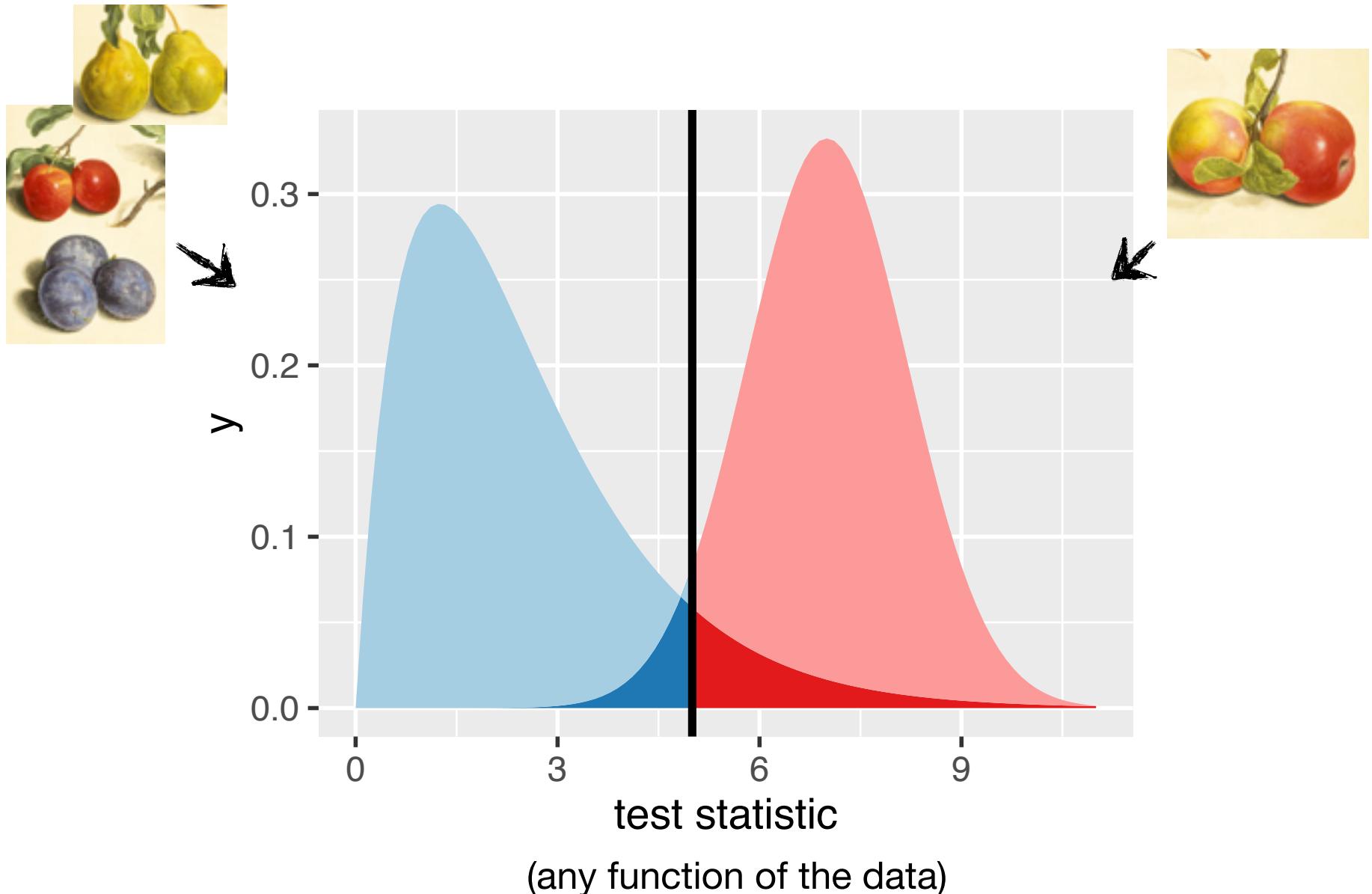
# Aims for this Lecture

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

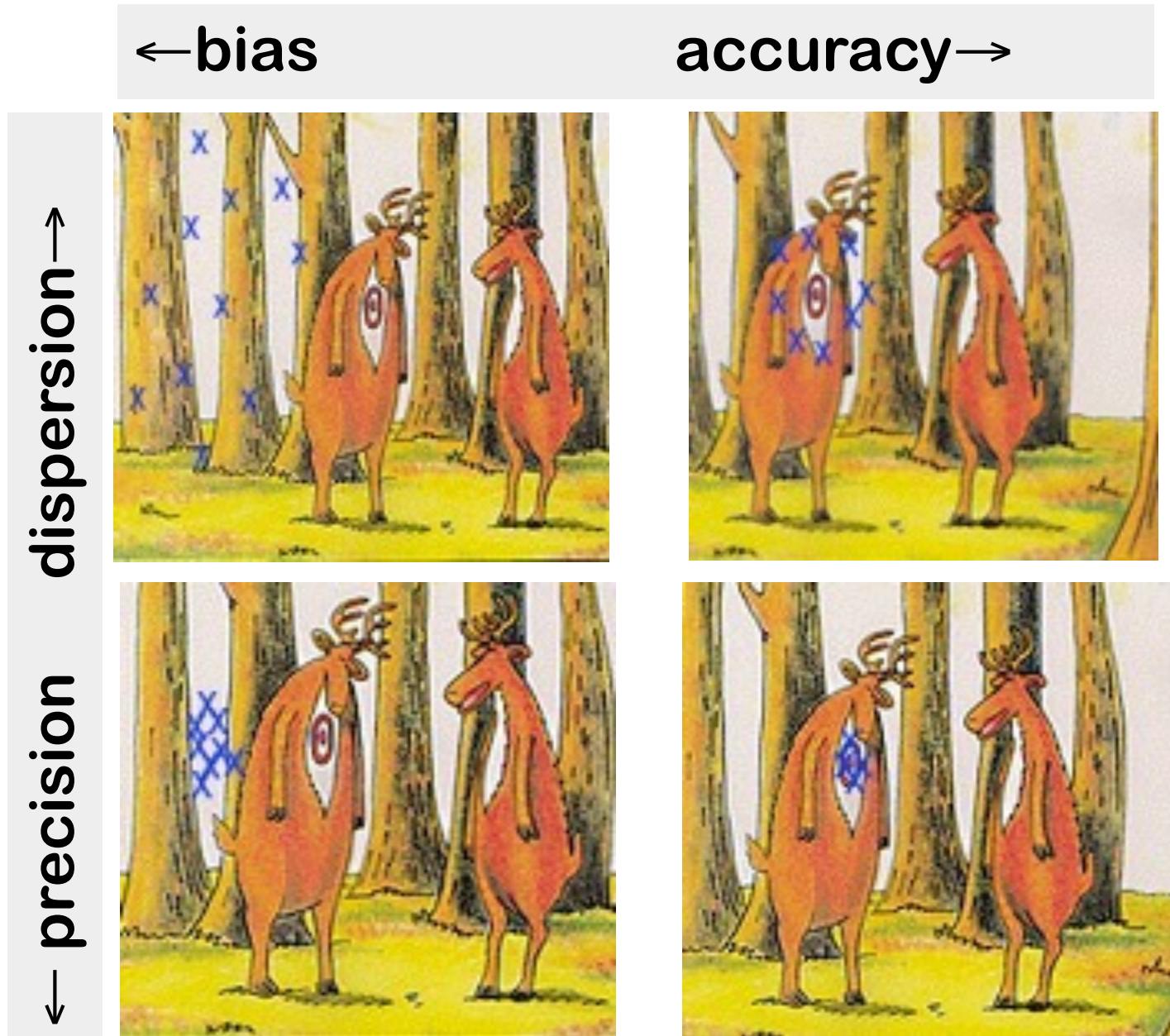
What changes when we go from single to multiple testing?

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

# Testing vs Classification



# 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

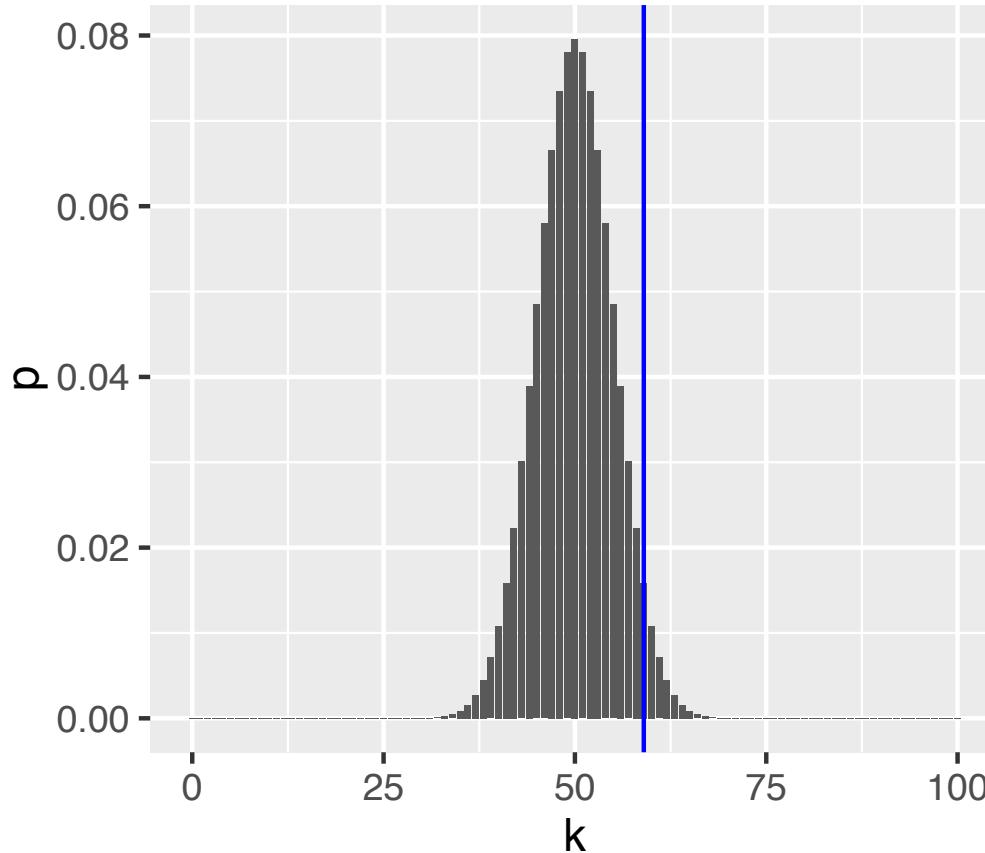


Figure 6.3: The binomial distribution for the parameters  $n = 100$  and  $p = 0.5$ ,

$$P(K = k \mid n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

# Rejection Region

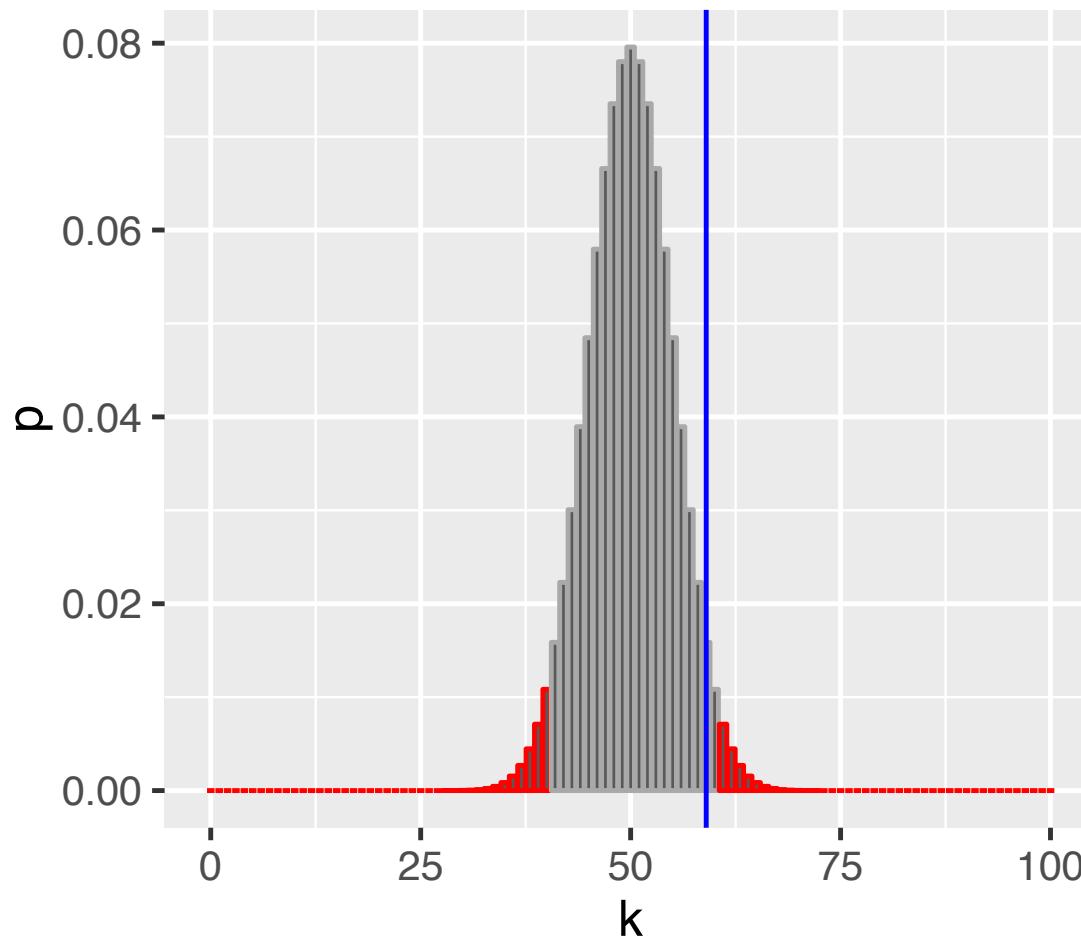


Figure 6.5: As Figure 6.3, with rejection region (red) whose total area is  $\alpha = 0.05$ .

# 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 ( $\leq$  **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:

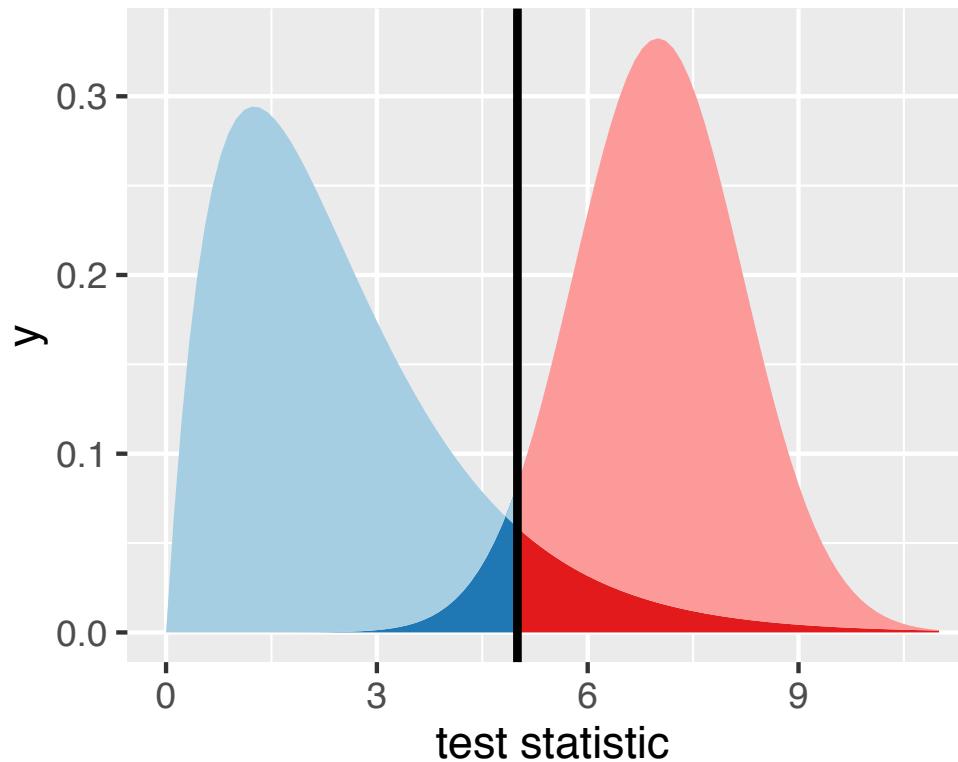
- The number of heads and tails were the same
- The coin is not fair
- The drug is worth its money

# 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

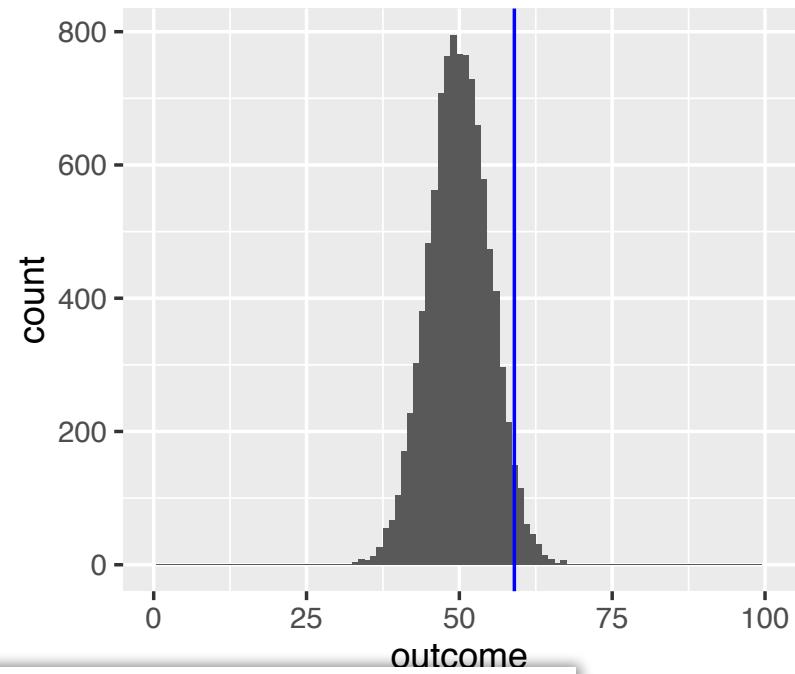
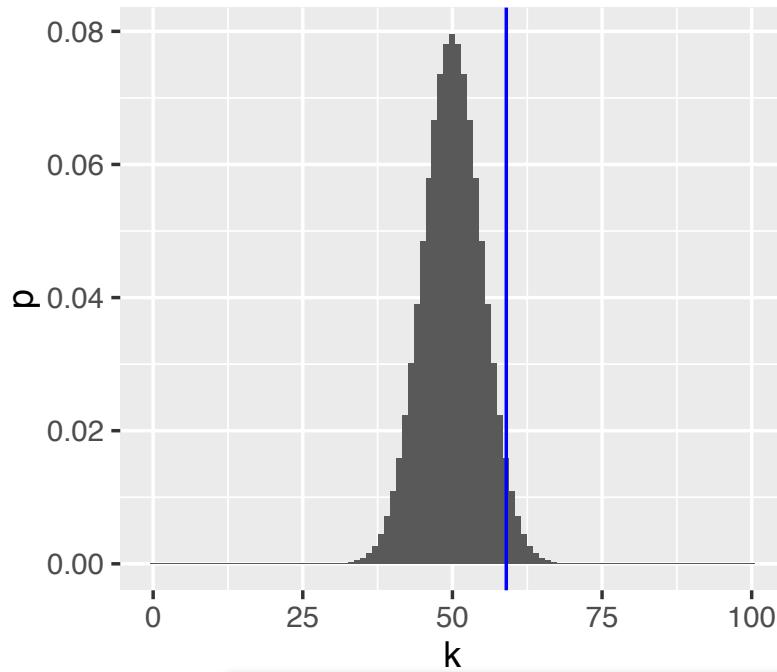


Figure 6.3: The parameter according to

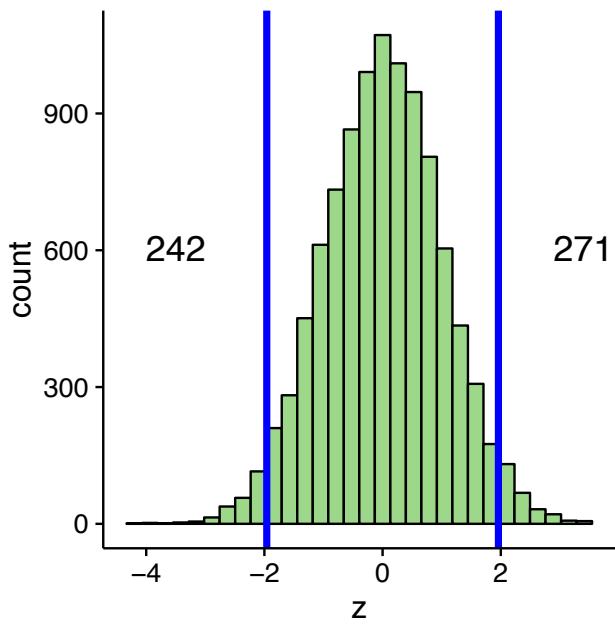
**Q:**

Discuss pros and contras for each

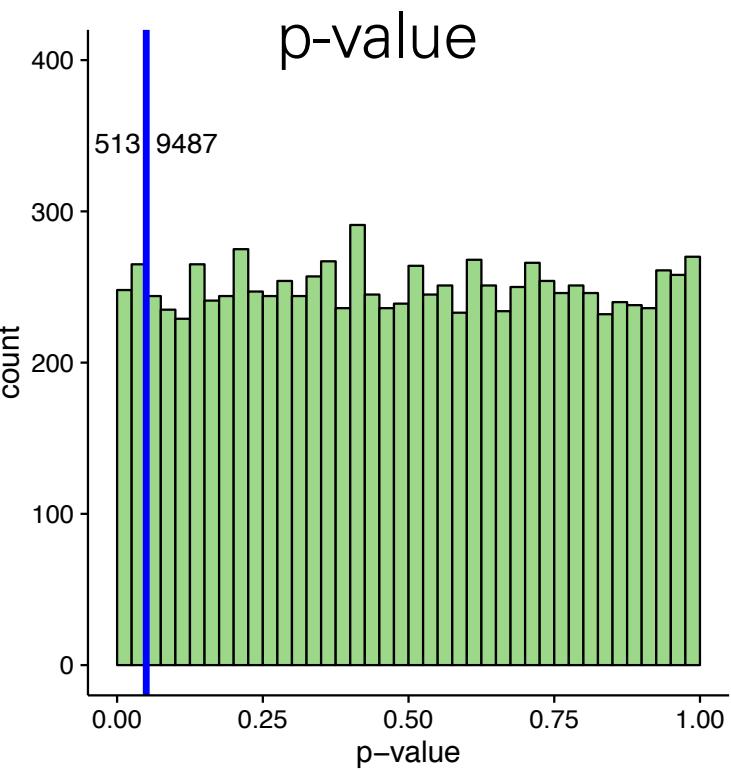
of the simulations).

# p-Values as Random Variables

test statistic



distribution  
function



# 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

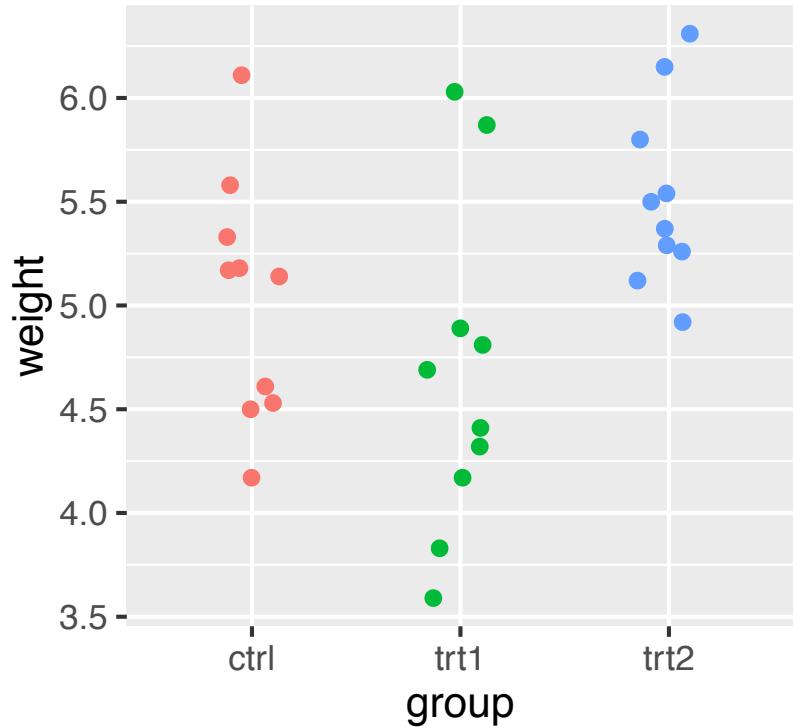


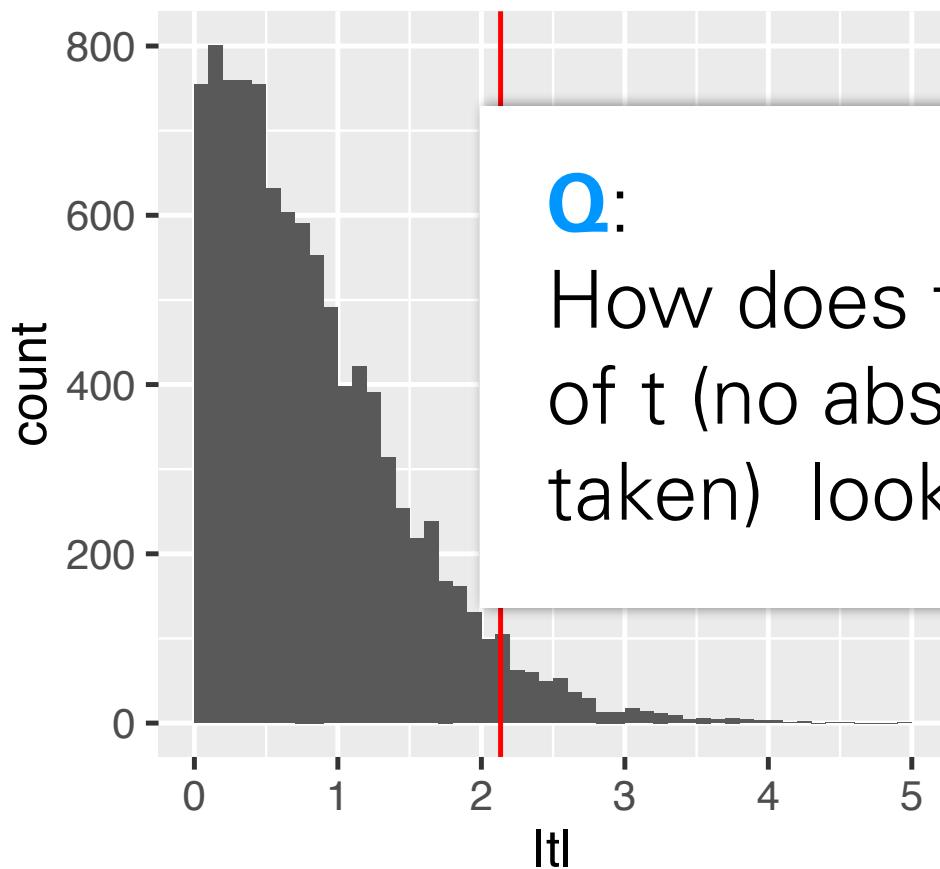
Figure 6.7: The PlantGrowth data.

$$m_g = \frac{1}{n_g} \sum_{i=1}^{n_g} x_{g,i} \quad g = 1, 2$$

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left( \sum_{i=1}^{n_1} (x_{1,i} - m_1)^2 + \sum_{j=1}^{n_2} (x_{2,j} - m_2)^2 \right)$$

$$c = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}.$$

# t- (and |t|-) Distribution



Q:

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

If the data are

then  
is 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

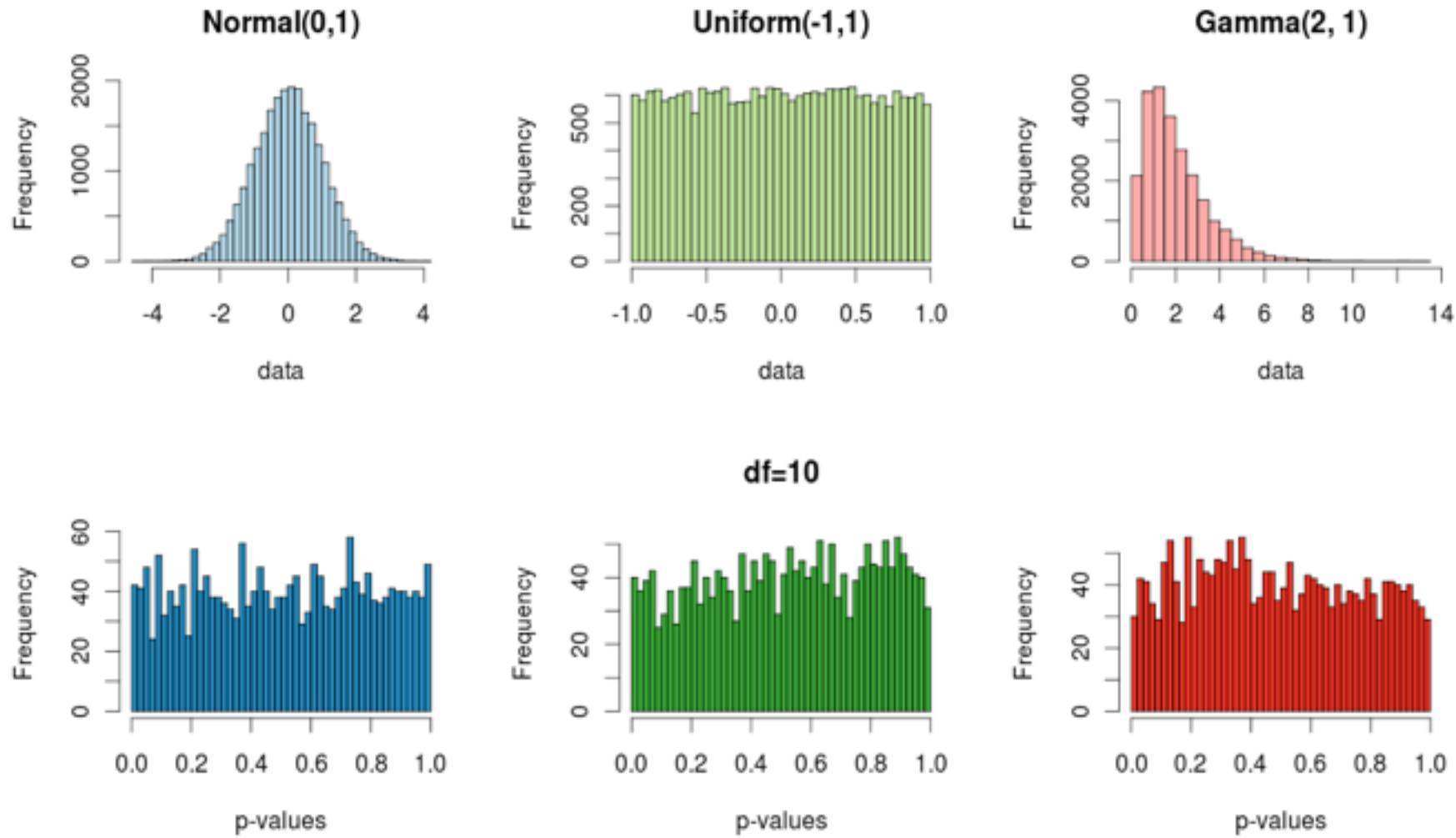
The derivation of the  $t$ -distribution assumes that the observations are independent and that they follow a Normal distribution.

**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 Loses Error Control if Independence Assumption Does not Hold

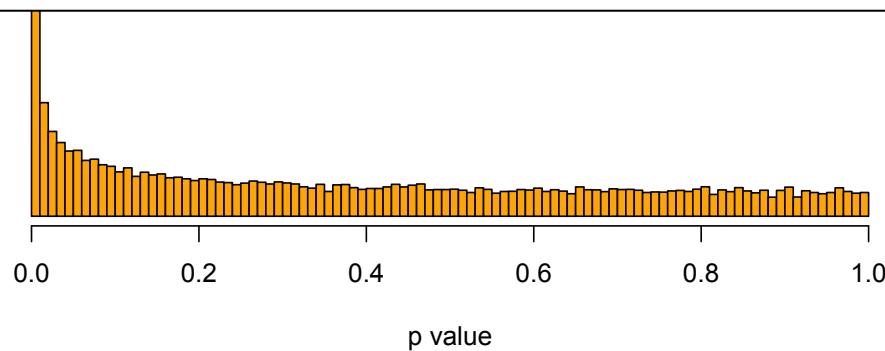
uncorrelated

```
library("mvtnorm")
library("genefilter")
p = 38      ## number of samples
#> [1] 38
#> 
#> #> 
#> #>
```



batch effects!

```
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.

| <u>P-VALUE</u> | <u>INTERPRETATION</u>                    |
|----------------|--|
| 0.001          |  |
| 0.01           |  |
| 0.02           | HIGHLY SIGNIFICANT                       |
| 0.03           |  |
| 0.04           |  |
| 0.049          | SIGNIFICANT                              |
| 0.050          | OH CRAP. REDO<br>CALCULATIONS.           |
| 0.051          |  |
| 0.06           | ON THE EDGE<br>OF SIGNIFICANCE           |
| 0.07           |  |
| 0.08           | HIGHLY SUGGESTIVE,<br>SIGNIFICANT AT THE |
| 0.09           | $p < 0.10$ LEVEL                         |
| 0.099          | HEY, LOOK AT                             |
| $\geq 0.1$     | THIS INTERESTING<br>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

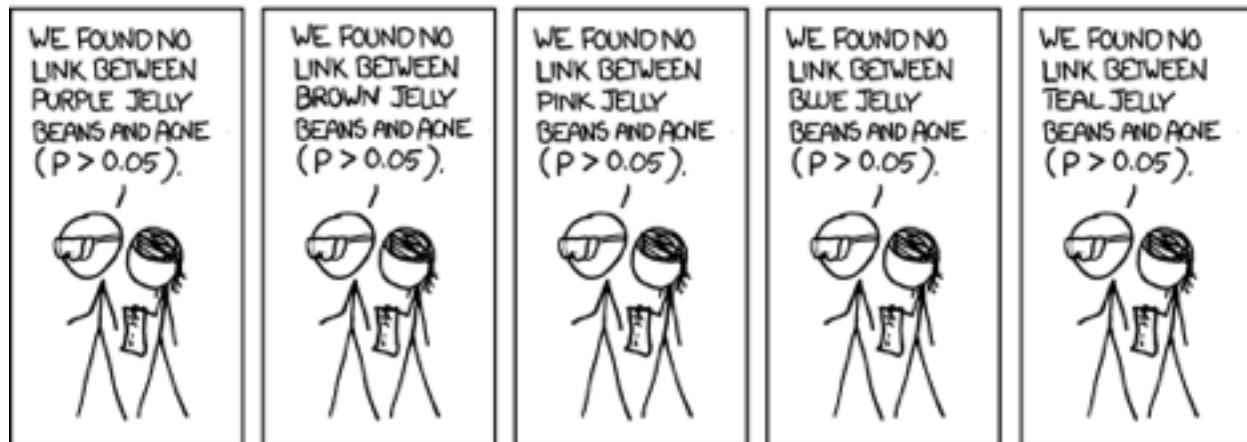
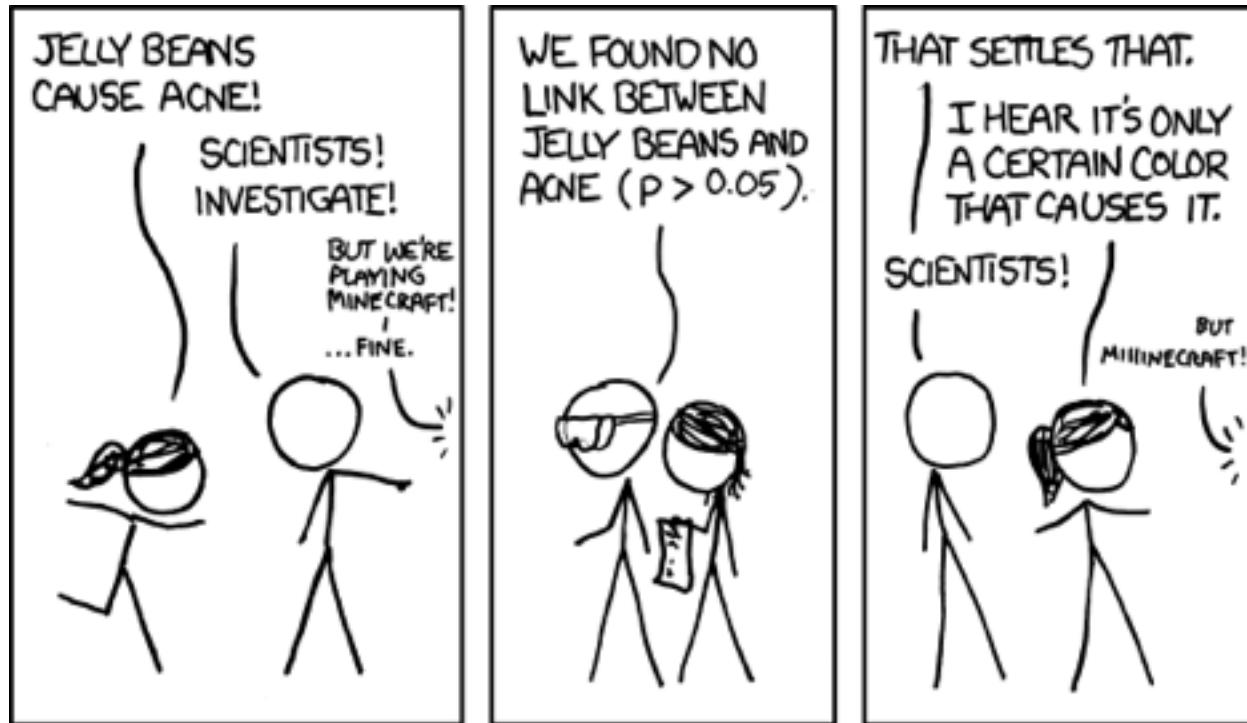
'outlier' removal

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](https://doi.org/10.1080/00031305.2016.1154108)

## What can we do about this?

# Multiple Testing



# Multiple Testing

WE FOUND NO  
LINK BETWEEN  
SALMON JELLY  
BEANS AND ACNE  
( $P > 0.05$ ).



WE FOUND NO  
LINK BETWEEN  
RED JELLY  
BEANS AND ACNE  
( $P > 0.05$ ).



WE FOUND NO  
LINK BETWEEN  
TURQUOISE JELLY  
BEANS AND ACNE  
( $P > 0.05$ ).



WE FOUND NO  
LINK BETWEEN  
MAGENTA JELLY  
BEANS AND ACNE  
( $P > 0.05$ ).



WE FOUND NO  
LINK BETWEEN  
YELLOW JELLY  
BEANS AND ACNE  
( $P > 0.05$ ).



WE FOUND NO  
LINK BETWEEN  
GREY JELLY  
BEANS AND ACNE  
( $P > 0.05$ ).



WE FOUND NO  
LINK BETWEEN  
TAN JELLY  
BEANS AND ACNE  
( $P > 0.05$ ).



WE FOUND NO  
LINK BETWEEN  
CYAN JELLY  
BEANS AND ACNE  
( $P > 0.05$ ).



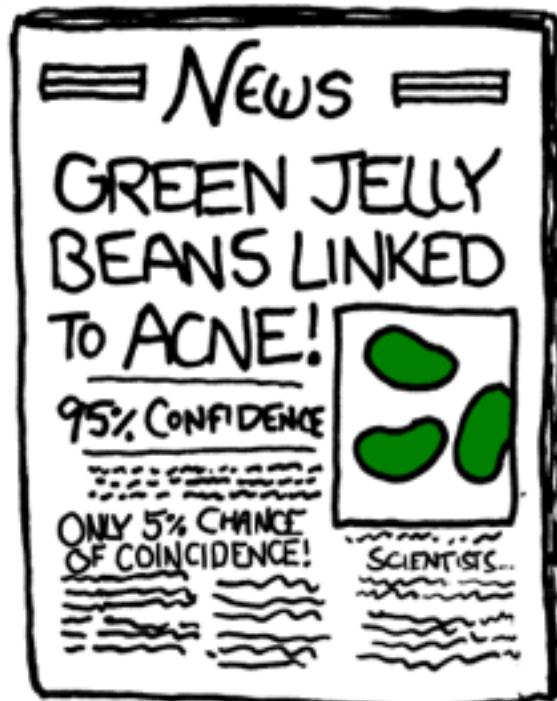
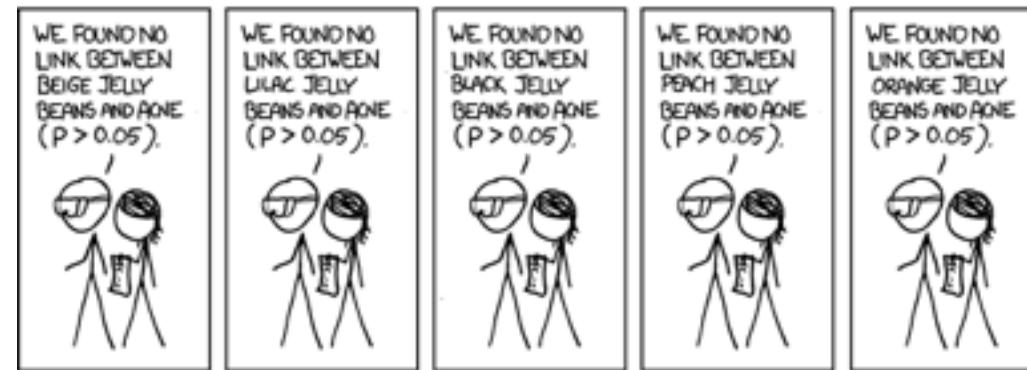
WE FOUND A  
LINK BETWEEN  
GREEN JELLY  
BEANS AND ACNE  
( $P < 0.05$ ).



WE FOUND NO  
LINK BETWEEN  
MAUVE JELLY  
BEANS AND ACNE  
( $P > 0.05$ ).



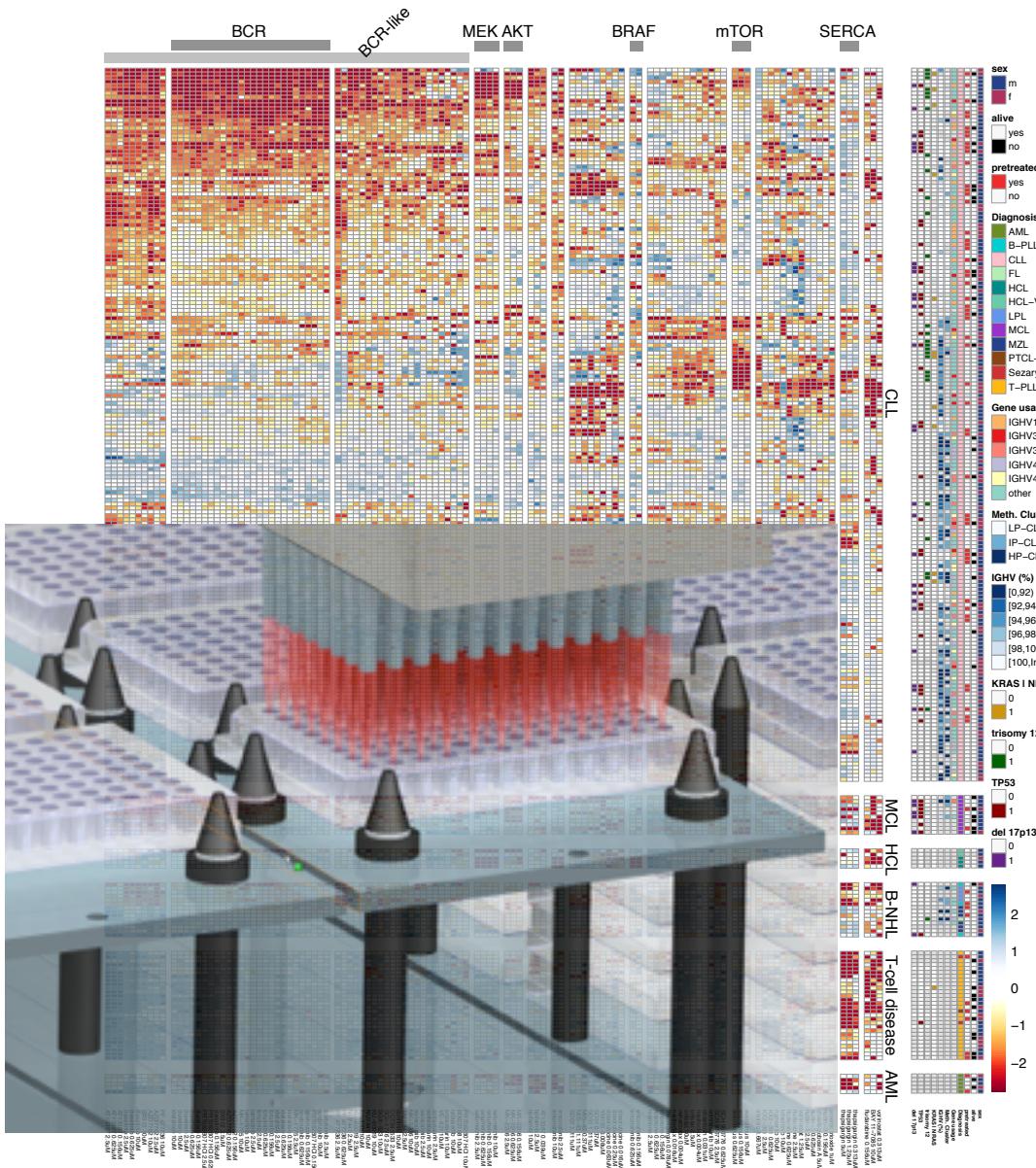
# Multiple Testing



# 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  $\alpha = 0.05$  and  $n$  indep. 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.

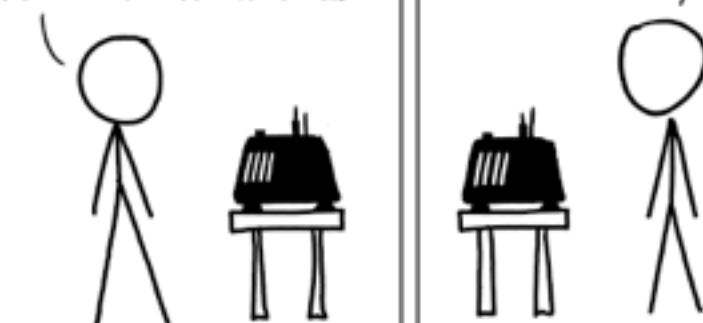
LET'S TRY.  
DETECTOR! HAS THE  
SUN GONE NOVA?

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 Positive Rate and False Discovery Rate

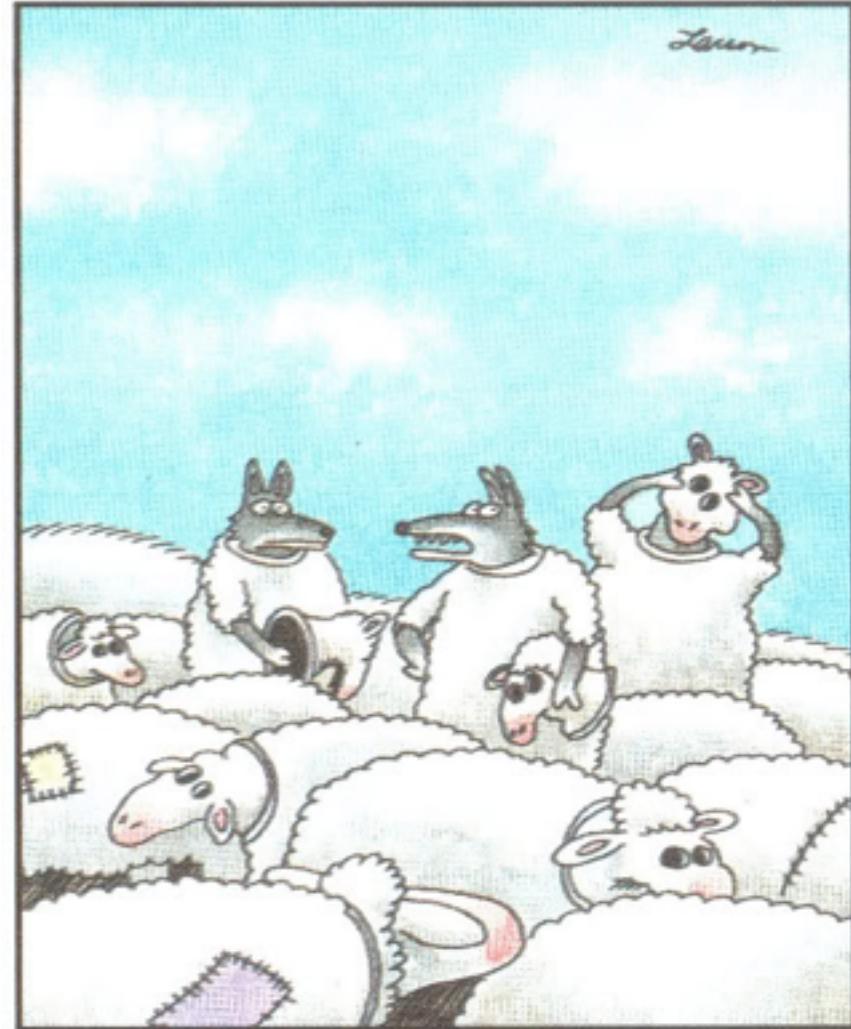
FPR: fraction of FP among all genes (etc.) tested

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

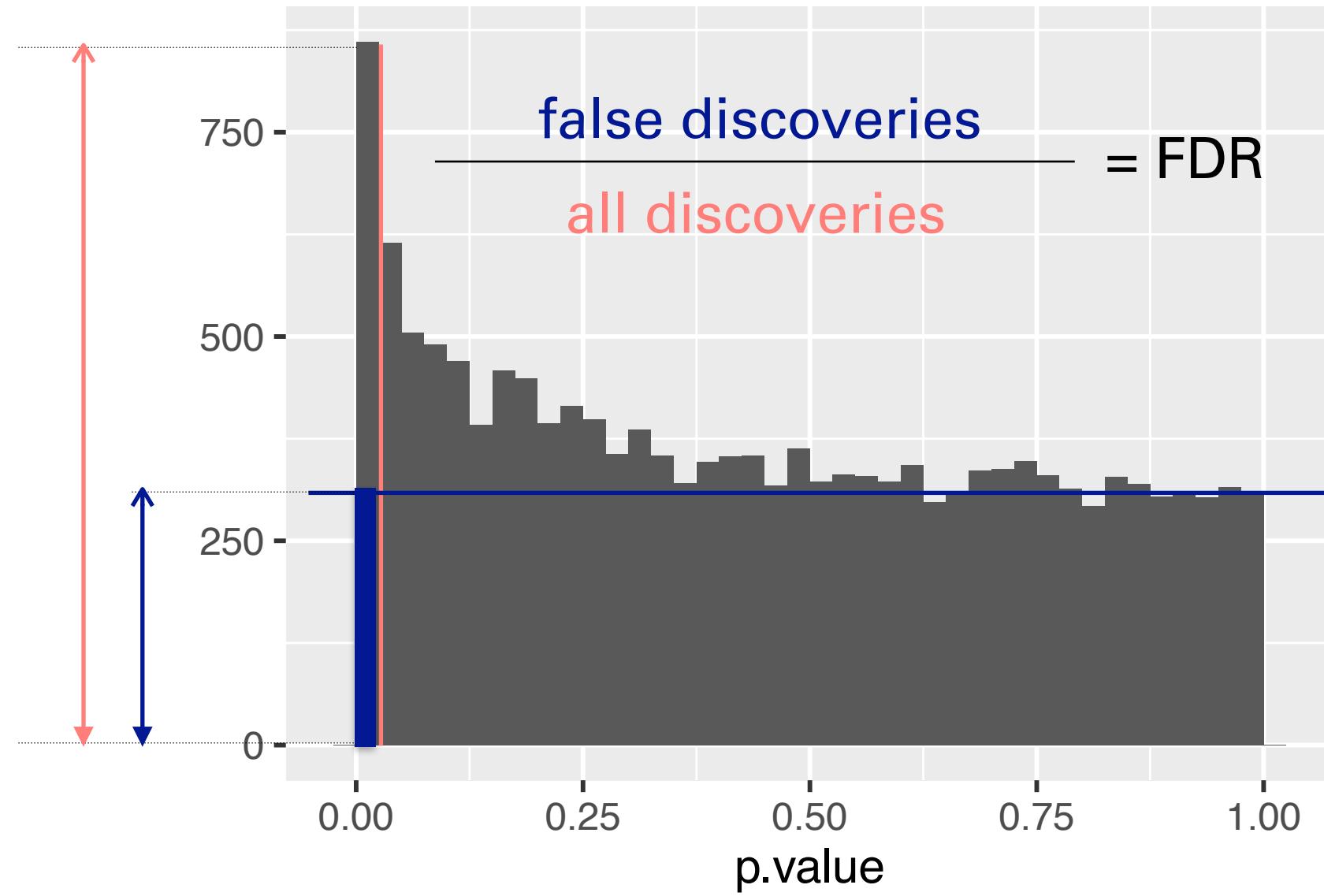


For  $m$  tests, multiply each  $p$ -value with  $m$ .  
Then see if anyone still remains below  $\alpha$ .

# False Discovery Rate



# False Discovery Rate

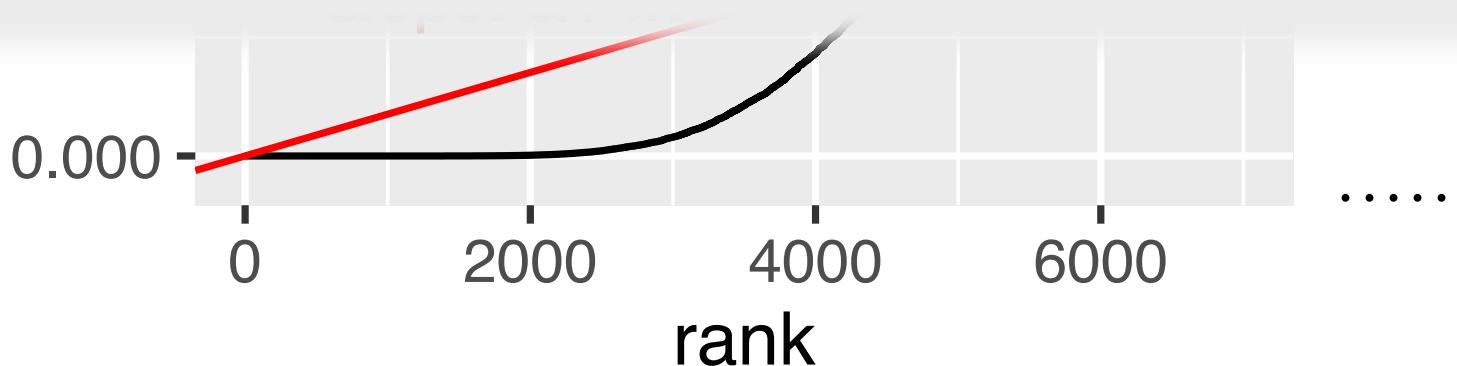


Method of Benjamini & Hochberg (1995)

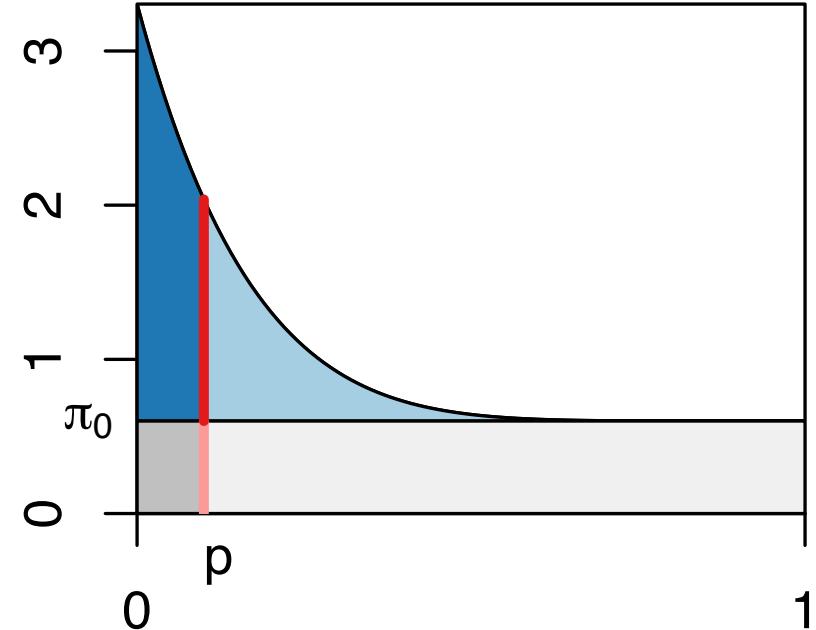
# Method of Benjamini & Hochberg

```
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



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

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

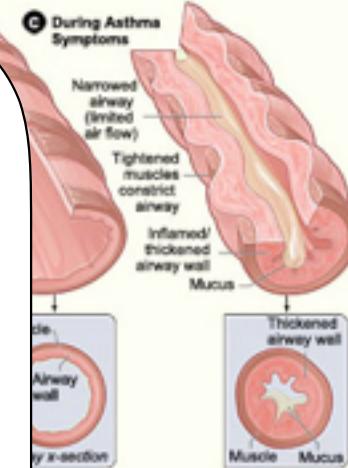
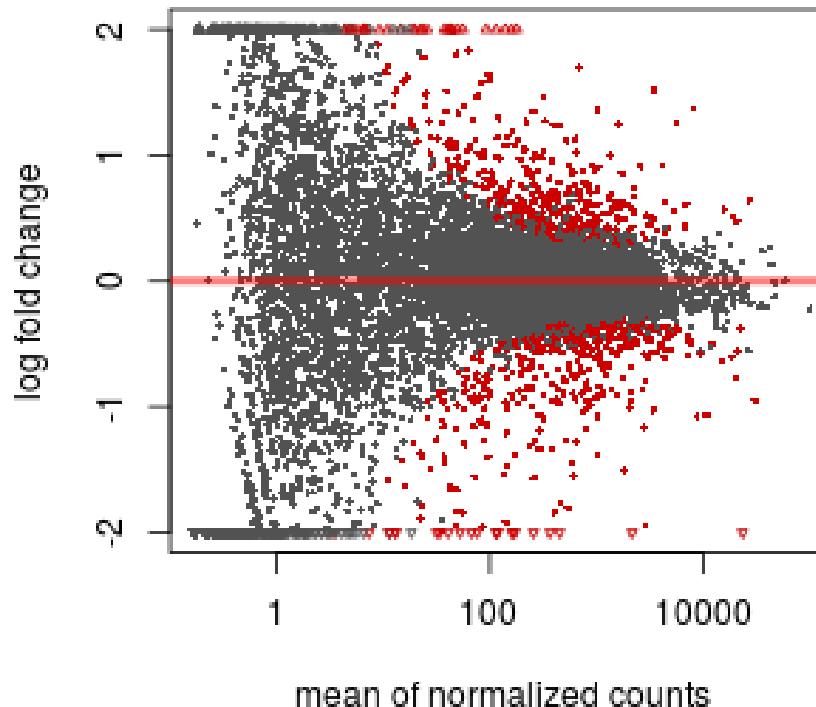
**FDR**: a set property. A single number that applies to a whole set of discoveries.

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

A photograph of a vast colony of penguins, likely Emperor penguins, gathered on a snow-covered slope. In the background, a massive, rugged iceberg dominates the scene, its white surface textured with deep blue shadows. The penguins are scattered across the foreground and middle ground, their dark heads and bodies contrasting with the bright snow.

# *Exchangeability?*

# Data set 1: RNA-Seq



expression analysis:

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

N061011

trt

design <- ~ cellline + dexamethasone

# Not all Hypothesis Tests are Created Equal

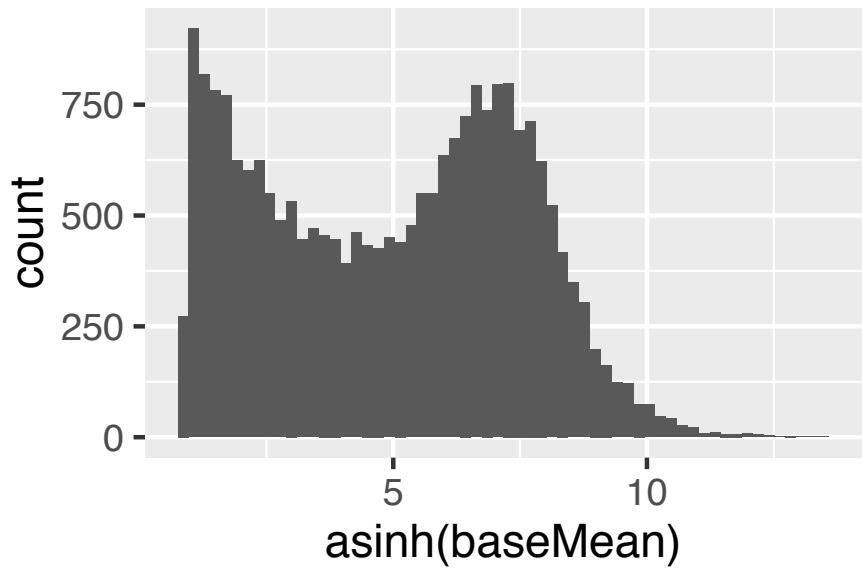
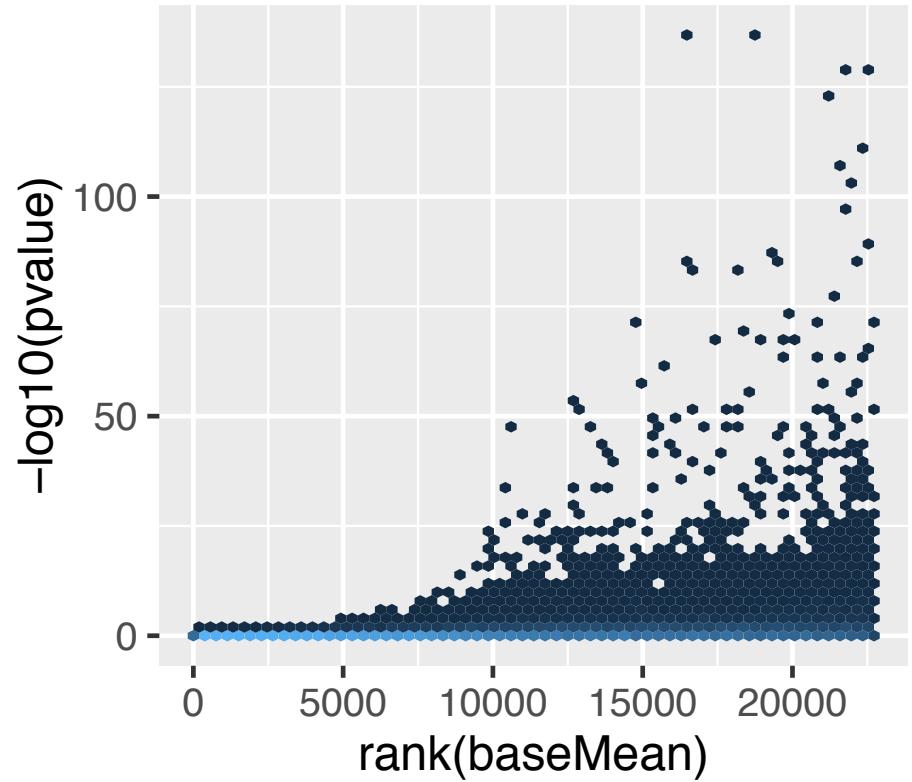


Figure 6.15: Histogram of  $\text{baseMean}$ . We see that it covers a large dynamic range, from close to 0 to around  $3.3 \times 10^5$ .



# Covariates - examples

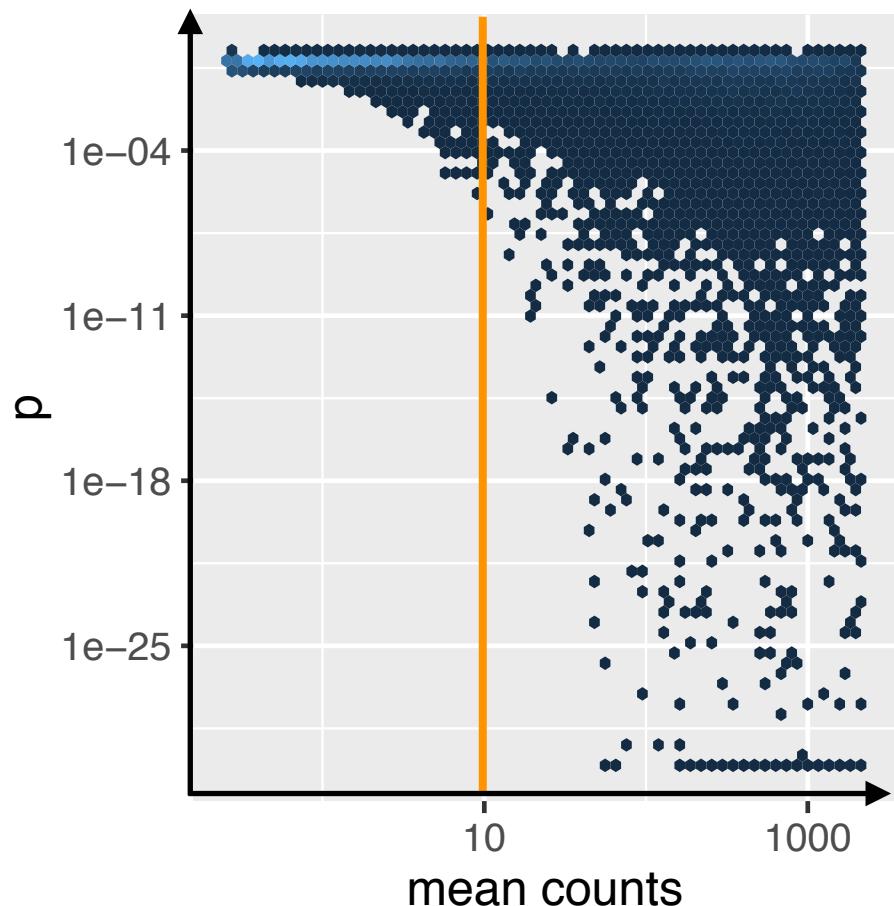
| Application                                      | Covariate   |
|--|---|
| Differential RNA-Seq,<br>ChIP-Seq, CLIP-seq, ... | (Normalized) mean of<br>counts for each gene      |
| GWAS   | Minor allele frequency                            |
| eQTL analysis                                    | SNP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

# Independent Filtering

Two steps:

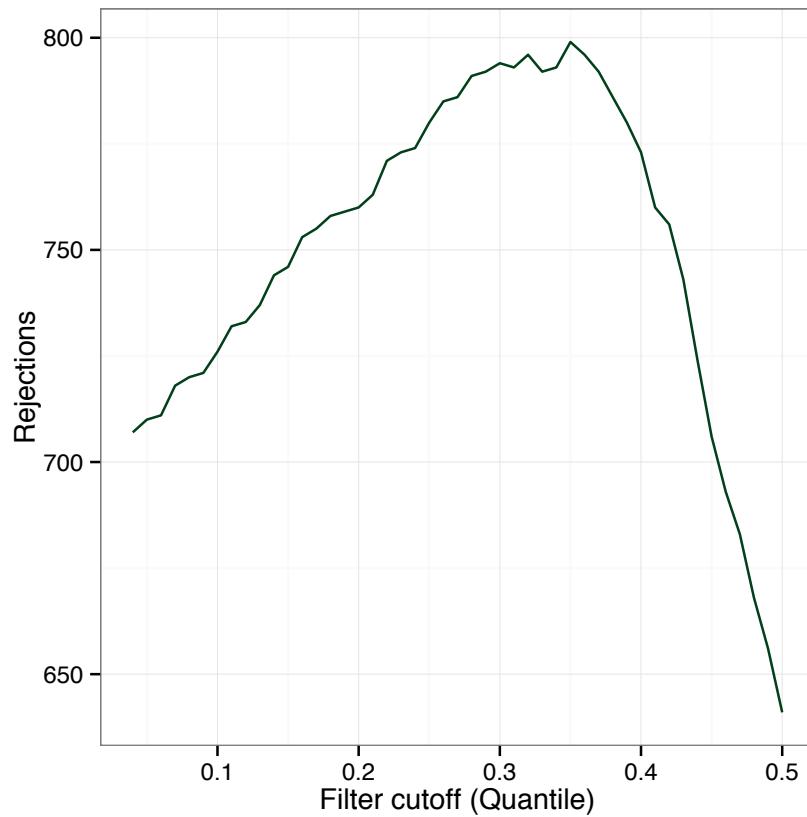
- All hypotheses  $H_i$  with  $X_i < x$  get filtered.
- Apply BH to remaining hypotheses.

(Bourgon, Gentleman, Huber  
PNAS 2010)



# Data-driven choice of filtering threshold

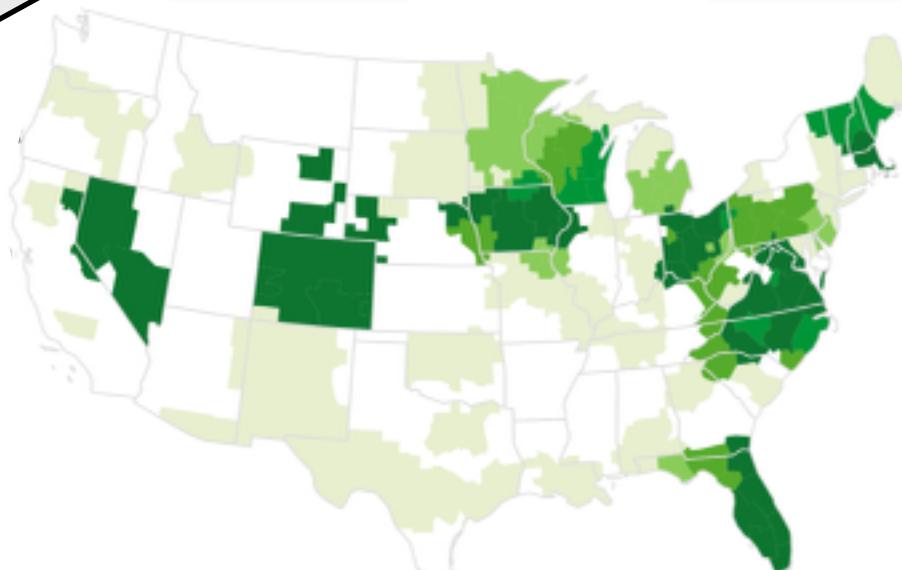
- Do Independent Filtering followed by Benjamini-Hochberg procedure with all possible thresholds.
- Report the result with the optimal threshold.
- We have been doing this in *DESeq2* for the last two years.



# Weighted Benjamini-Hochberg method

- Let  $w_i \geq 0$  and  $\alpha = \alpha(w)$  be the “alpha budget”).
- Define  $\hat{\alpha}_i = \min\{\alpha, w_i\}$ .
- A threshold  $\hat{P}_i$ .
- FDR control (Genovese, Roeder,

Problem: how to know the weights?

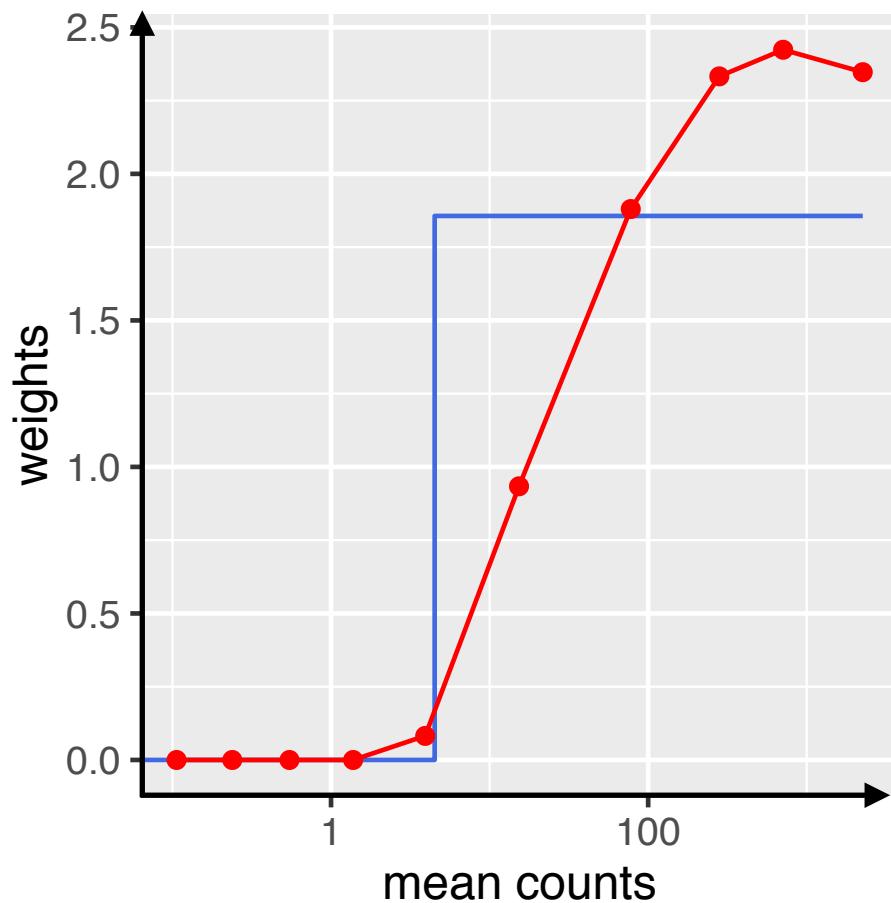


‘alpha  
investing’

# Independent filtering is a special case of weighted BH

$S$  = set of hypotheses  
retained by filtering step

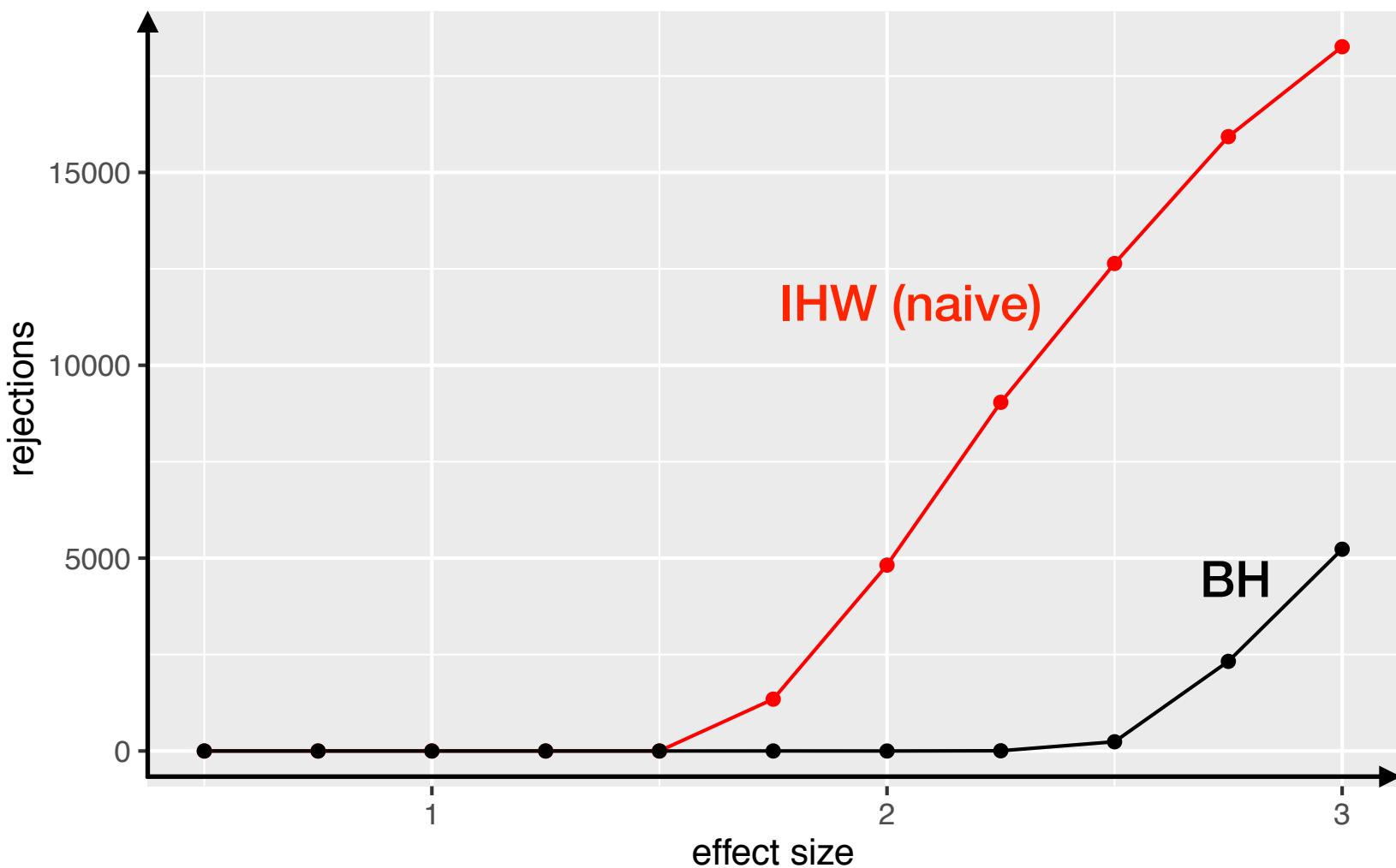
$$w_i = \begin{cases} m/|S| & \forall i \in S \\ 0 & \forall i \notin S \end{cases}$$



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

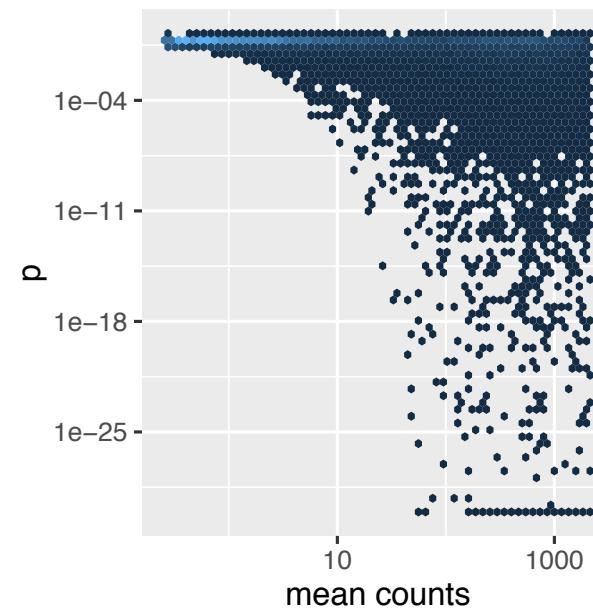
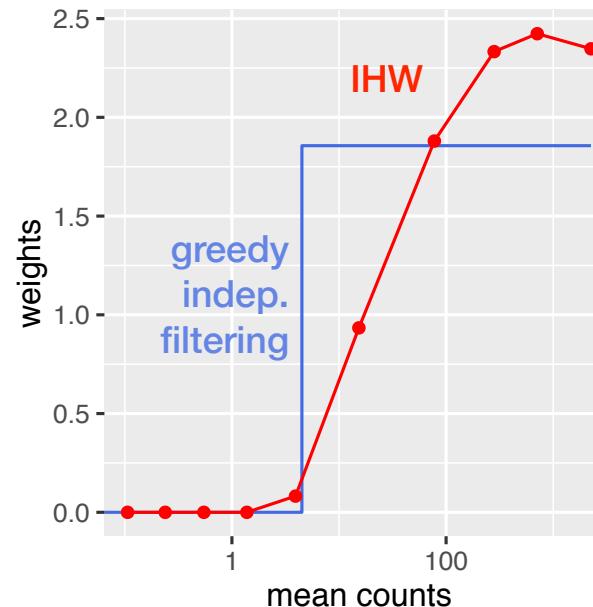
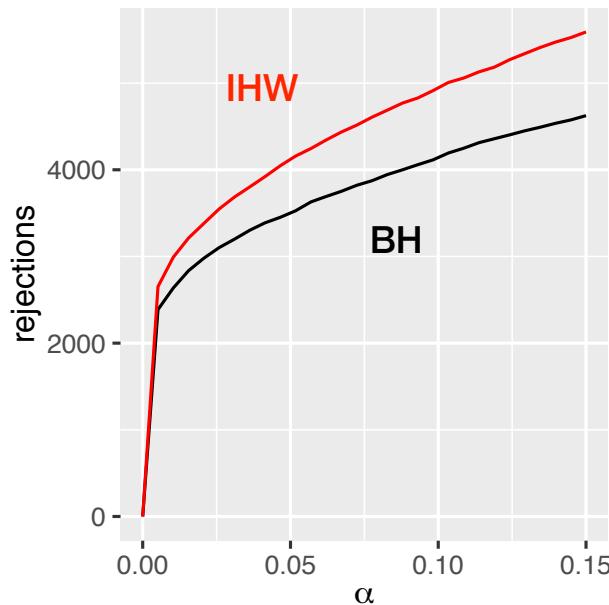
- Stratify the tests into  $G$  bins, by covariate  $X$
- Choose  $\alpha$
- For each possible weight vector  $\mathbf{w} = (w_1, \dots, w_G)$  apply weighted BH procedure. Choose  $\mathbf{w}$  that maximizes the number of rejections at level  $\alpha$ .
- Report the result with the optimal weight vector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

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

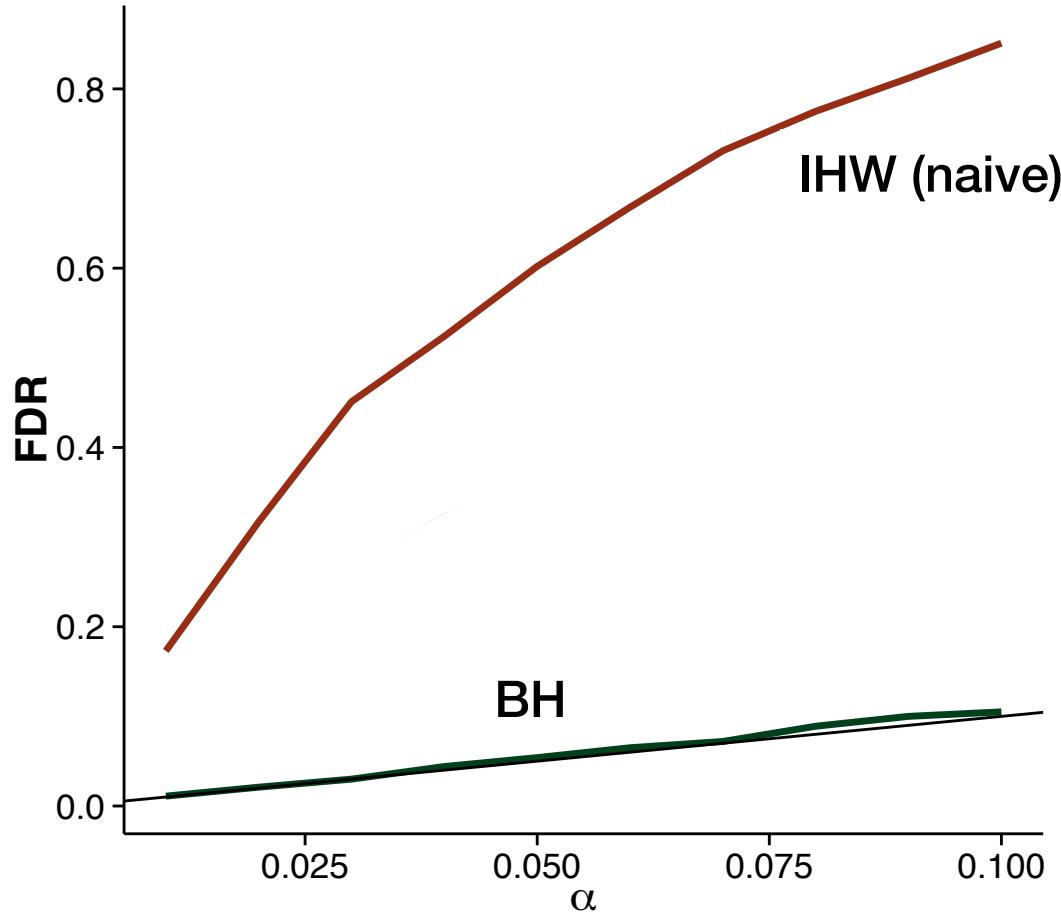


$$m = 500,000 \quad m_1 = 20,000 \quad \alpha = 0.1 \quad n = 2 \times 4$$

# RNA-Seq example (DESeq2)



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



# Modified HW



Nikos Ignatiadis

‘Pre-validation’: randomly split hypotheses into  $k$  folds. Learn weights for the hypotheses in a fold from the other  $k-1$  folds

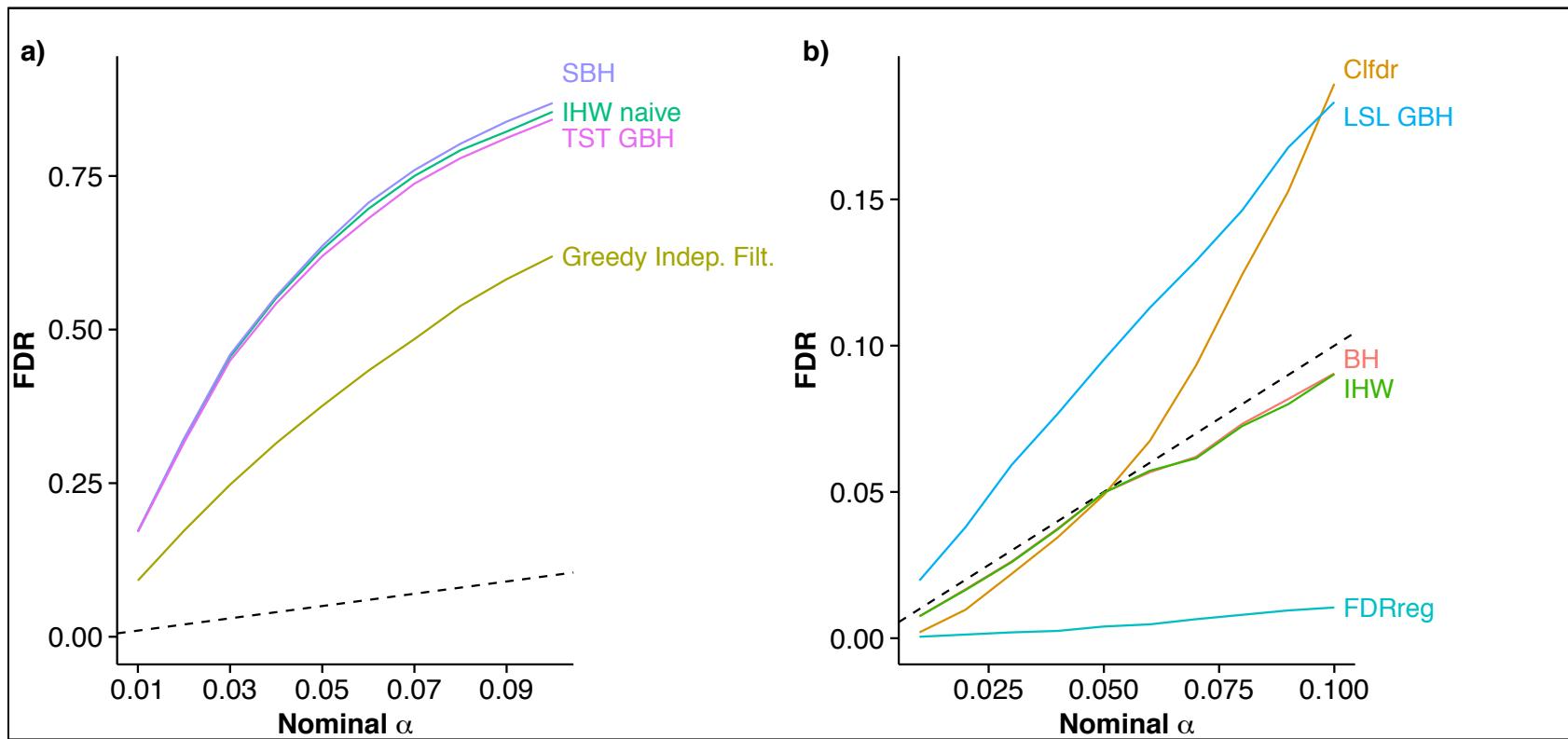
## Regularisation:

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

# IHW controls FDR

Nulls only



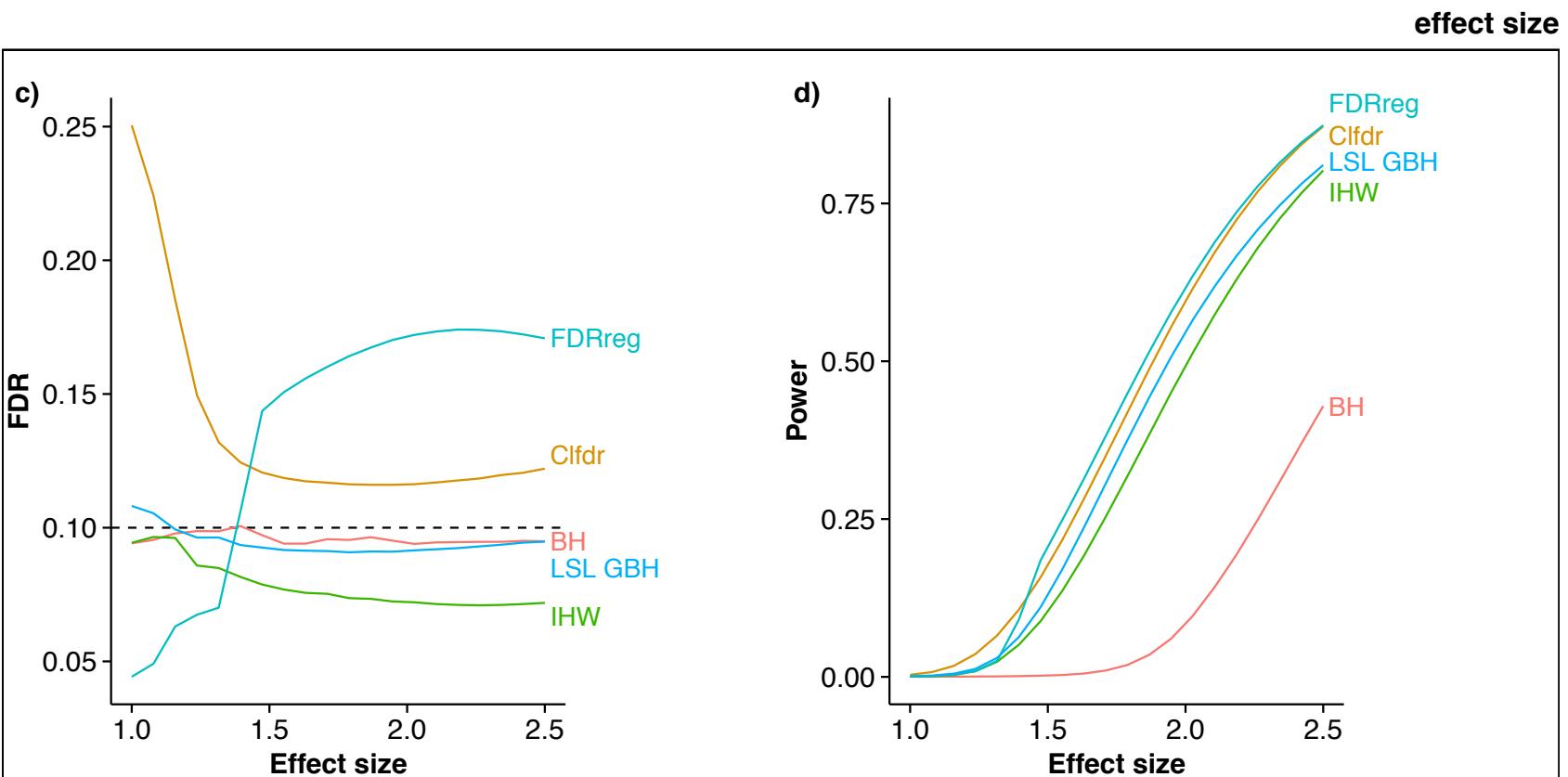
**SBH:** stratified BH (e.g. Yoo, Bull, ... Sun, Genet. Epidem 2010)

**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



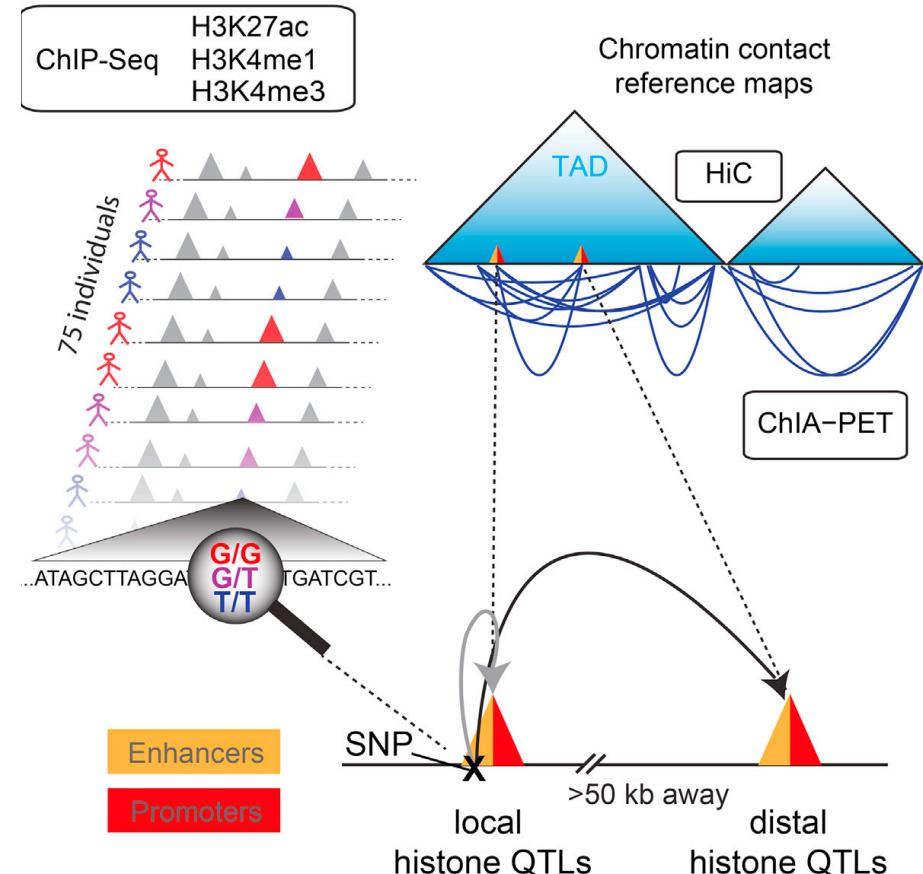
# Data set 2: hQTL

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

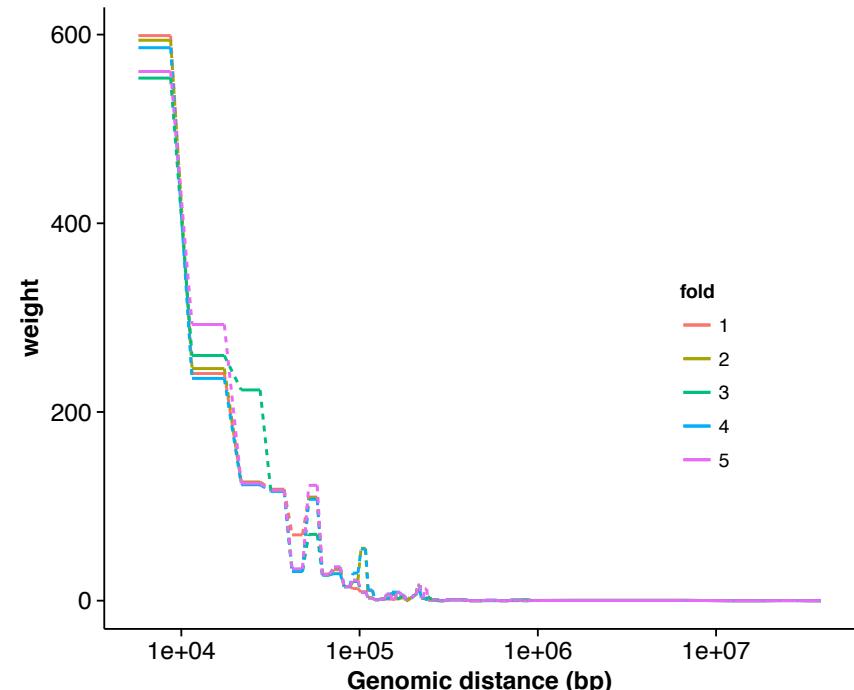
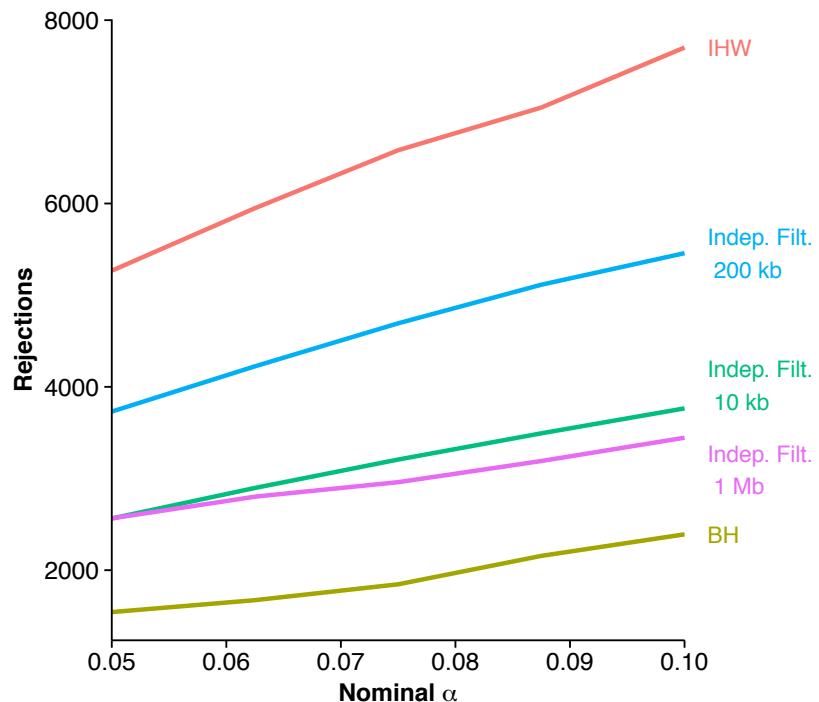
**Local QTLs:** find best-correlated SNP within 2kb of peak boundaries/promoters.

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

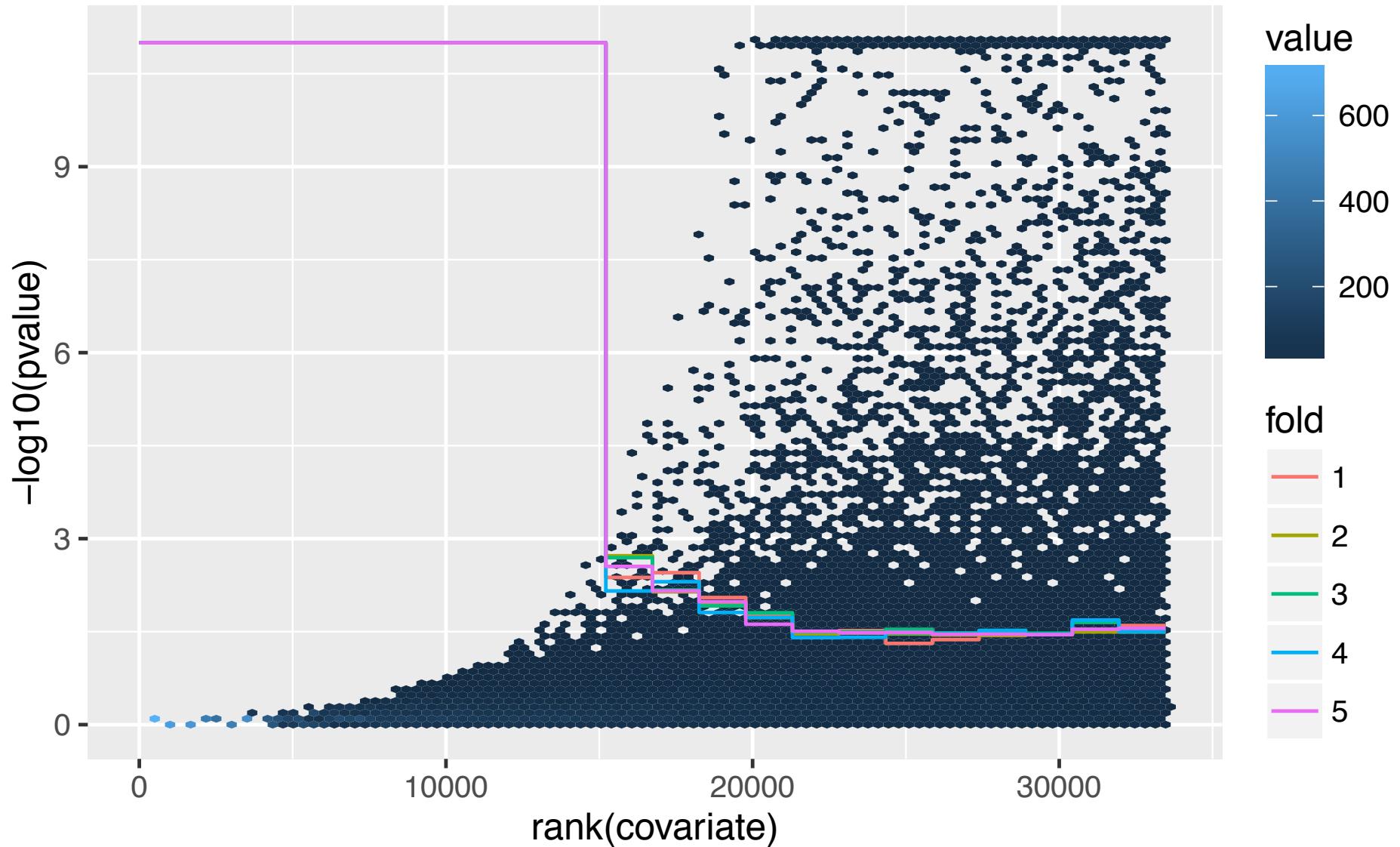
**Distal:** distance cutoffs from 50 to 300 kb; also HiC



# histone-QTL example: H3K27ac



# 2D decision boundaries



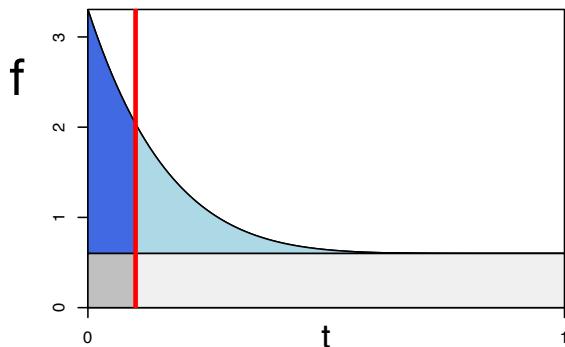
# Why does IHW work?

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

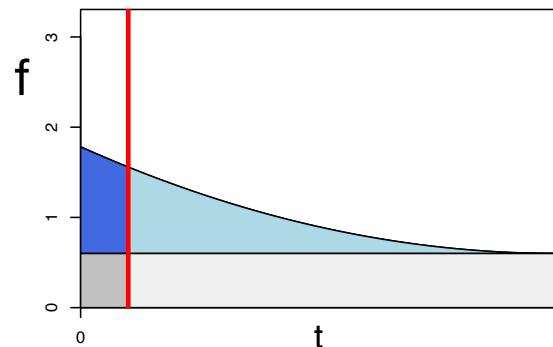
$$f(t) = \pi_0 + (1 - \pi_0)f_{\text{alt}}(t)$$

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

$$\text{tdr}(t) = 1 - \text{fdr}(t)$$

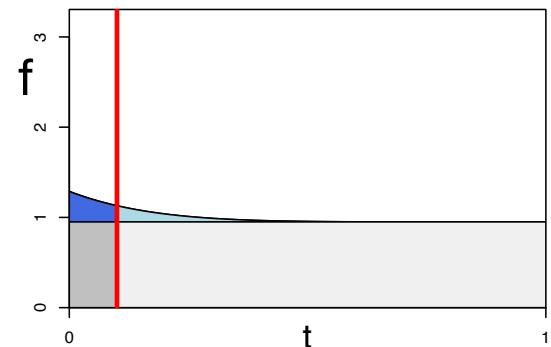


$$\pi_0 = 0.6$$



$$\pi_0 = 0.6$$

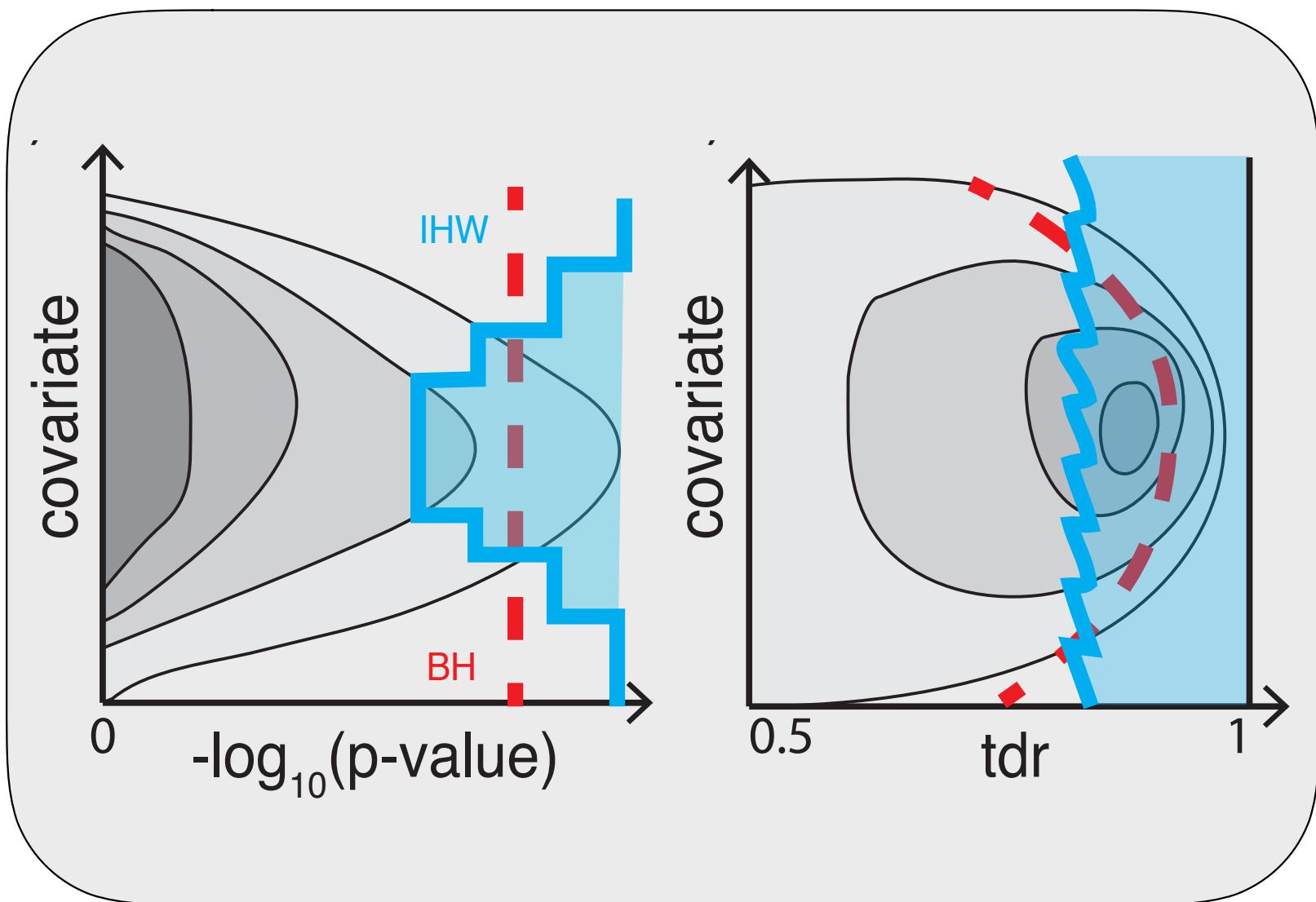
different  $f_{\text{alt}}$



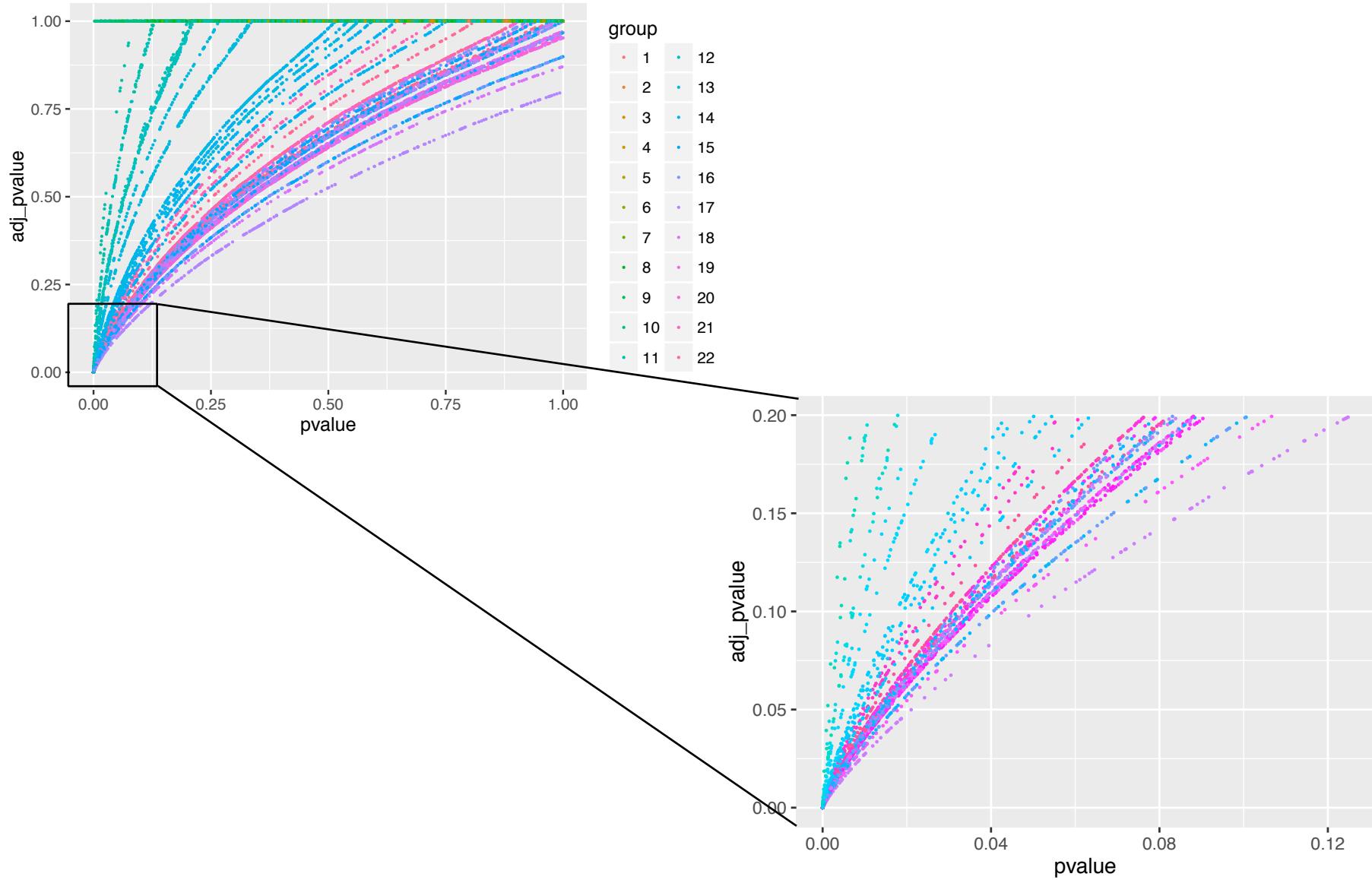
$$\pi_0 = 0.95$$

same  $f_{\text{alt}}$

# 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  $\alpha$  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... X + https://www.bioconductor.org/packages/devel/IHW/ Search

**Bioconductor**  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

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 man
- [Workflows](#) for learning and
- [Course and conference mat](#)
- [Videos](#).
- Community [resources](#) and [tutorials](#).

[R / CRAN](#) ↗

Support  
Please read questions and answers in the following:

- [Support](#): Bioconduc
- [BioC-dev](#): develop

**Joint work with**  
**Nikos Ignatiadis**  
**Bernd Klaus**  
**Judith Zaugg**

**Thanks also to**  
**Robert Gentleman**  
**Richard Bourgon**  
**Misha Savitski**  
**Oliver Stegle**  
**Vlad Kim**

# Summary

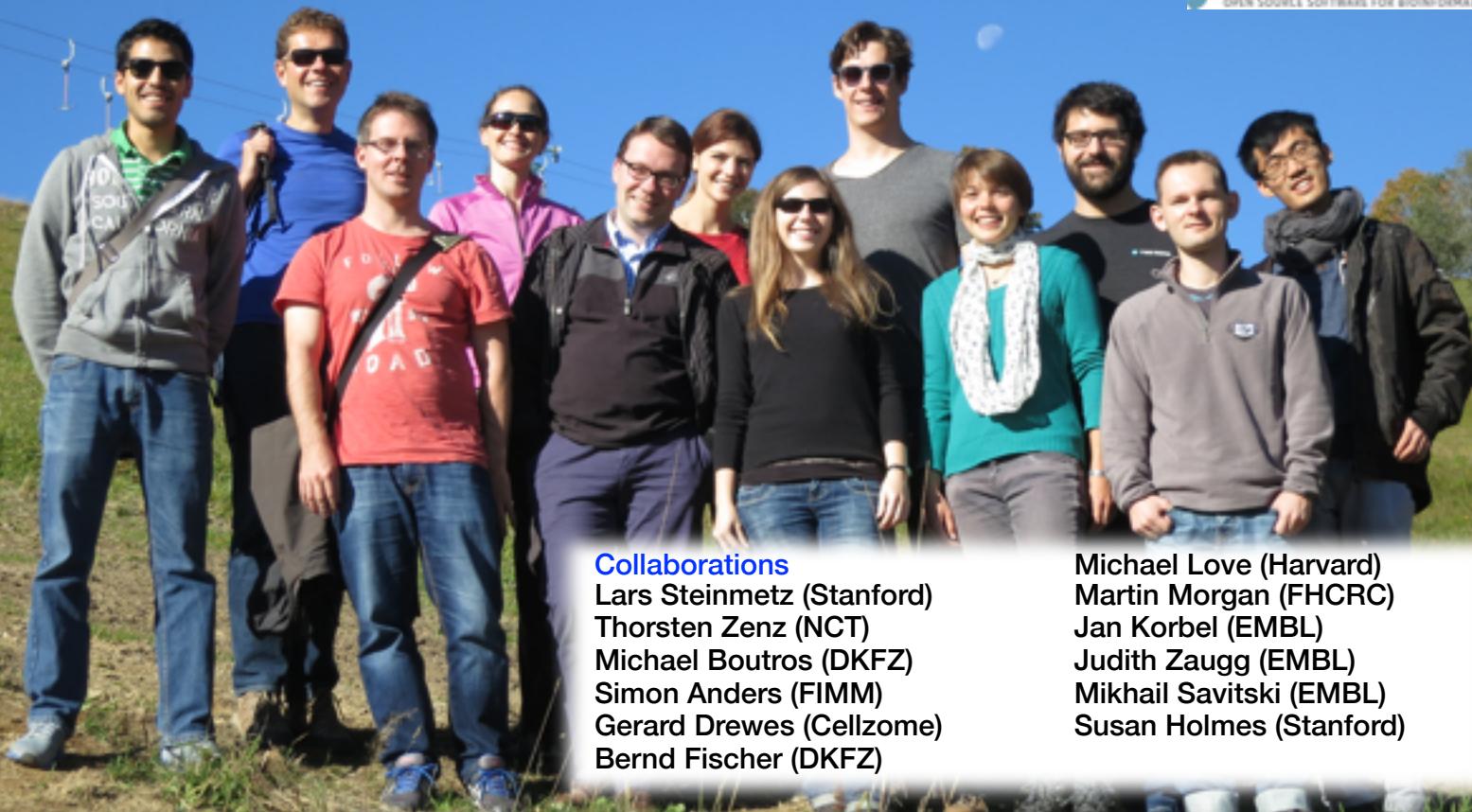
- Multiple testing is not a problem but an opportunity
- Heterogeneity across tests
- Informative covariates are often apparent to domain scientists
  - independent of test statistic under the null
  - informative on  $\pi_1, F_{\text{alt}}$
- Data-driven weighting
- Scales well to millions of hypotheses
- 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



#### Collaborations

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

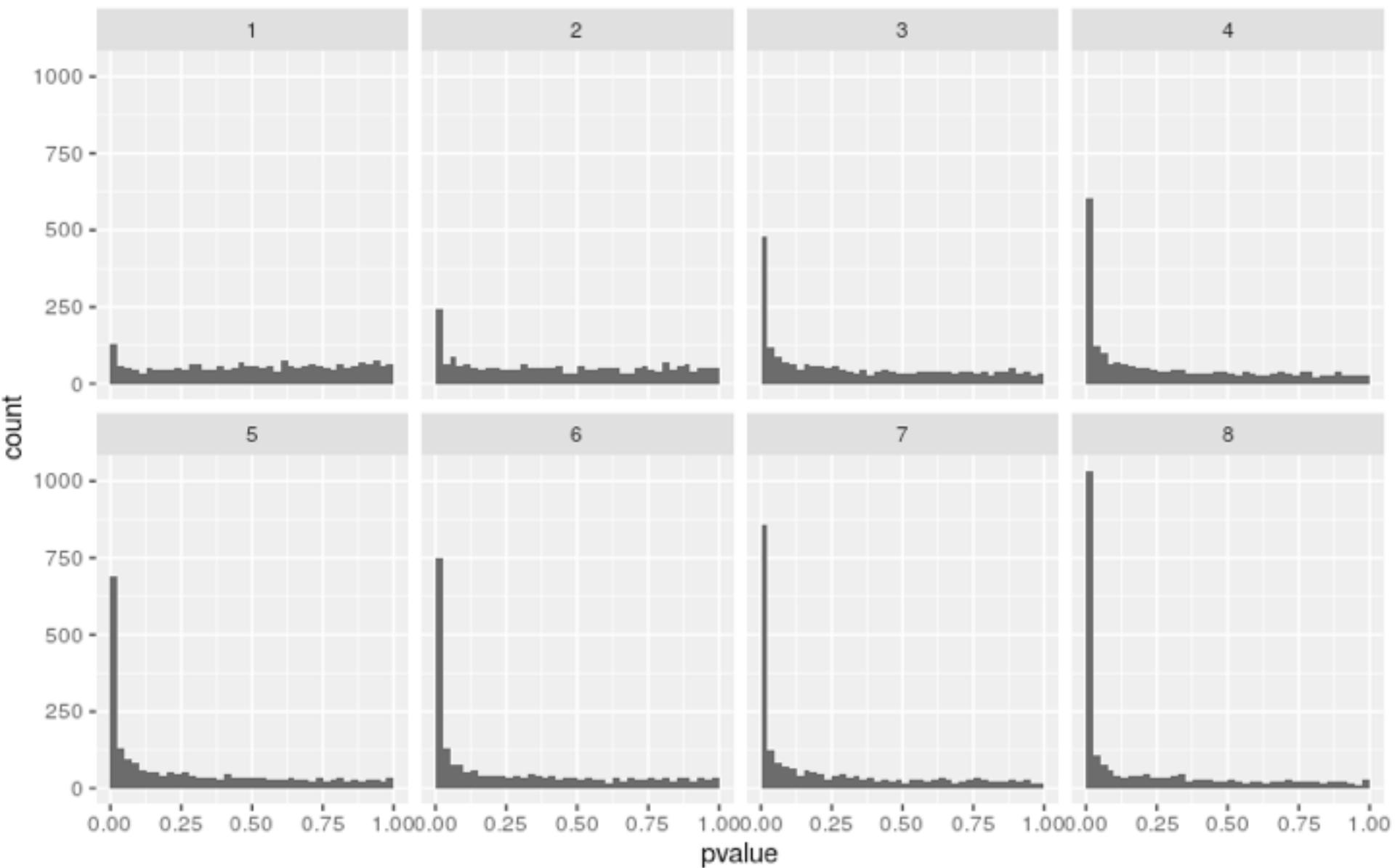
Michael Love (Harvard)  
Martin Morgan (FHCRC)  
Jan Korbel (EMBL)  
Judith Zaugg (EMBL)  
Mikhail Savitski (EMBL)  
Susan Holmes (Stanford)

#### Funding

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



# 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.



STUDY AT HOME

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., Providence, RI 02904