Multiple testing

Theory

Sarah Kaspar

EMBL

Sources:

- Oehlert (2010) “A first course in design and analysis of experiments” – Chapter 5

- Huber and Holmes: ”Modern statistics for modern bioligists” – Chapter 6

Multiple testing scenarios in biology

EMBL

Scenarios

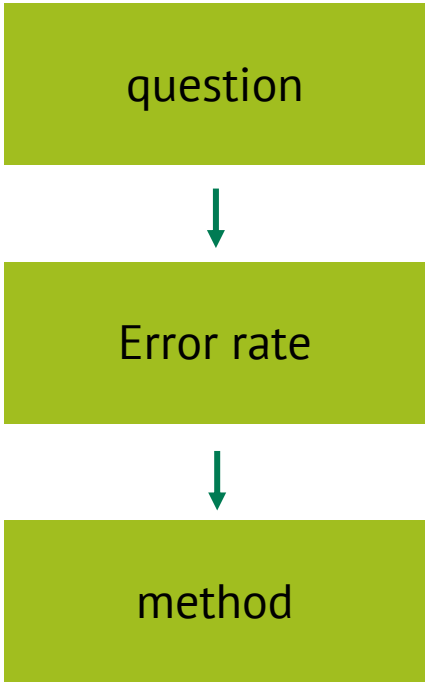
- Expression profiling
- Compound screens
- Drug screens
- Genome-wide association studies
- Proteomics
- …

Questions:

- Which genes are DE due to some condition?
- Which drugs are candidates for targeting a specific pathway/protein?
- Which genetic variants are associated with a particular disease?
- Is any of the compounds in a medication unsafe?

Problem: Many false positives.
Solution: Methods that implement false-positive control.

Workflow summary



The most important part is to know which error rate you want to control for.

The error rate depends on how you phrase your question.

If you know the error rate, the choice of method is (mostly) straight-forward.

Testing a single hypothesis

EMBL

	Not rejected	rejected
H_0 true	true negative	false positive type I error
H_0 false	false negative type II error	true positive

Great page:
https://en.wikipedia.org/wiki/Confusion_matrix

Comparison-wise error rate

- The usual error rate (= significance level) for a t-test / Wilcoxon test / Chi-square test …
- For $\alpha = 0.05$: If H_0 is true, there is a 5% chance of a false rejection.
- Used when all tests are viewed as individual questions.
- Nothing is corrected for.

Example:
- does the high-fat diet have an impact on weight in **female** mice?
- does it have an impact in **male** mice?

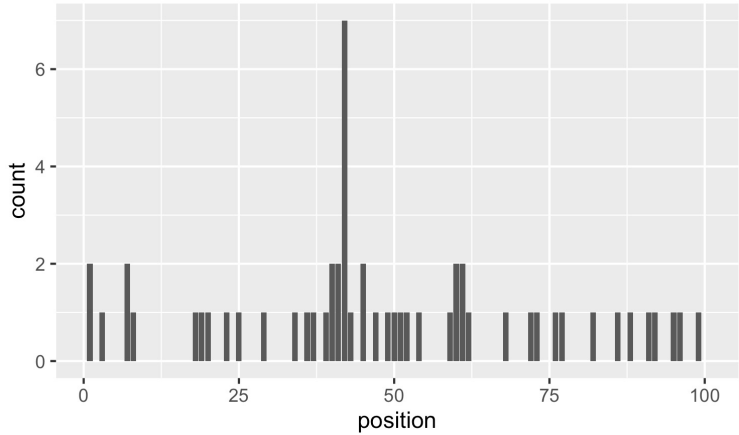
Example: epitopes

100 positions on a protein are tested for a reaction.

Question: Does the protein cause any reaction?

H_{01}: position 1 is no epitope.
H_{02}: position 2 is no epitope.
...
H_0: no position is an epitope.

Question: What is the probability of calling at least one false-positive epitope out of 100?



Data from https://www.huber.embl.de/msmb/

Family-wise error rate

The probability of rejecting one or more H_{0i} (and thus rejecting H_0) when all H_{0i} are true.

Example type I error Inflation:

If we test every position with $\alpha = 0.05$:

$$\begin{aligned} P(\text{false rejection of } H_0) &= 1 - P(\text{no rejection of any } H_{0i}) \\ &= 1 - 0.95^{100} \\ &= 0.994 \end{aligned}$$

Bonferroni correction

- Used to control the **FWER** at a desired value α_{FWER}
- Adjusted p-values: $\boldsymbol{p_{adj} = p_{unadj} \times m}$, where m is the number of tests
- Adjusted p-values give the chance of seeing any value as extreme as this *within m tests* under H_0.
- We call the test significant if $p_{adj} \leq \alpha_{FWER}$.
- Equivalent: Reject every test at α_{FWER}/m

Side note

The epitope data are discrete.

For **discrete test statistics**, p-values are **conservative** (i.e. too large).

$$P(p < 0.05) \leq 0.05$$

This also affects multiple testing:

$$P(\text{false rejection of } H_0) \leq 0.994$$

Ways to address this problem:

- Mid p-values
- Randomized p-values
- Modified adjustment procedures

Further reading:
Chapter 1.1.4 in Agresti, A. (2006). An Introduction to Categorical Data Analysis: Second Edition. https://doi.org/10.1002/0470114754

Example: Screening for differentially expressed genes

Question: Which genes are differentially expressed under some condition?

H_{01}: gene 1 is not DE.
H_{02}: gene 2 is not DE.
...
$H_{0,10000}$: gene 10000 is not DE.

Problem: If controlling for family-wise error rate, the probability of finding anything at all is small (low power).

Family-wise error rate for 10^4 tests:

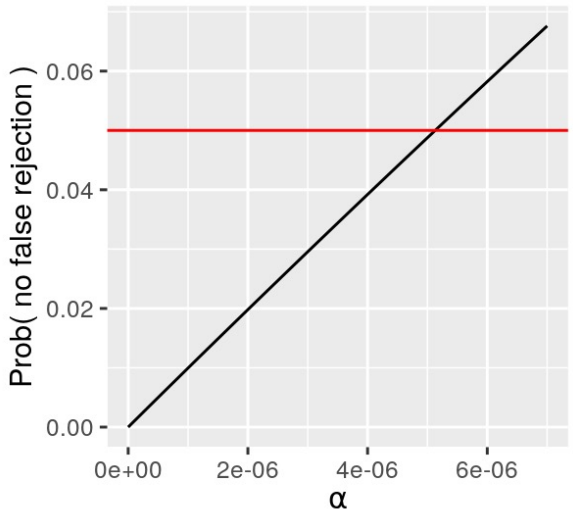


Figure: https://www.huber.embl.de/msmb/Chap-Testing.html#multiple-testing

False discovery rate

FDR: expected value of the false discovery fraction $\frac{FP}{FP+TP}$.

What percentage of hits are false positives?

Scenario: We expect some of the H_{0i} to be true and some to be false.

Trade-off between type I and type II error:

- if controlling for family-wise error rate, the probability of finding anything at all is small (low power).
- we allow some percentage of false discoveries to increase power.

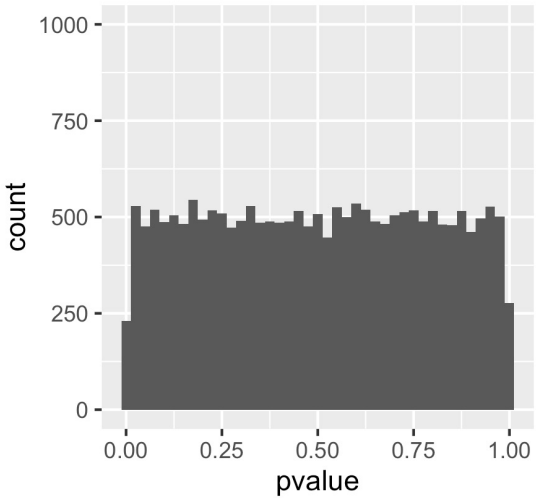
Benjamini-Hochberg algorithm:

- Allows FDR control

	Not rejected	rejected
H_0 true	TN	FP
H_0 false	FN	TP

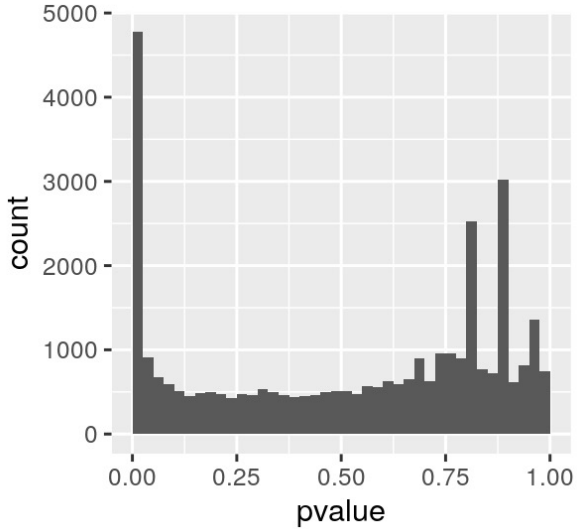
p-value histogram

No genes is DE:



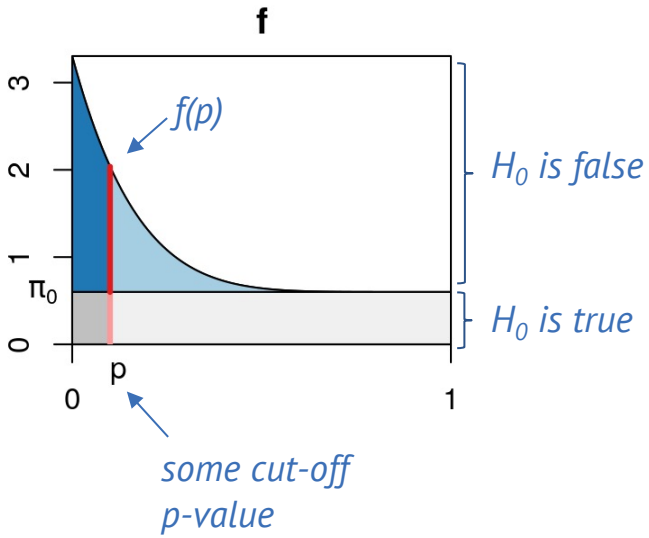
p-values are uniformly distributed.

Some genes are DE:



Peak at low p-values.

p-value histogram decomposition



Density of p-values

$$f(p) = \underbrace{\pi_0}_{\text{uniform component}} + \underbrace{(1-\pi_0) f_{alt}(p)}_{\text{alternative component } (H_A \text{ is true})}$$

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

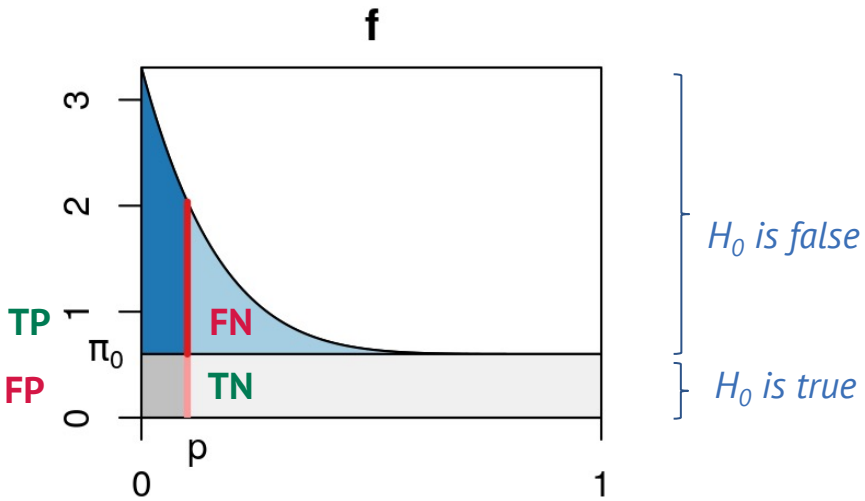
Applies to tests rejected just at this threshold

FDR: $\text{FDR}(p) = \frac{\pi_0 p}{\int_0^p f(t)dt}$

An average property of all tests rejected below the threshold

Figure: https://www.huber.embl.de/msmb/Chap-Testing.html#the-local-fdr

p-value histogram decomposition



Benjamini Hochberg algorithm:

- Estimates the uniform component (null is true)
- Finds a critical p-value *c*, so that rejecting everything below *c* will lead to the desired FDR
- Equivalently: Gives adjusted p-values, so that rejecting everything below p_{adj} will lead to an FDR of p_{adj}.

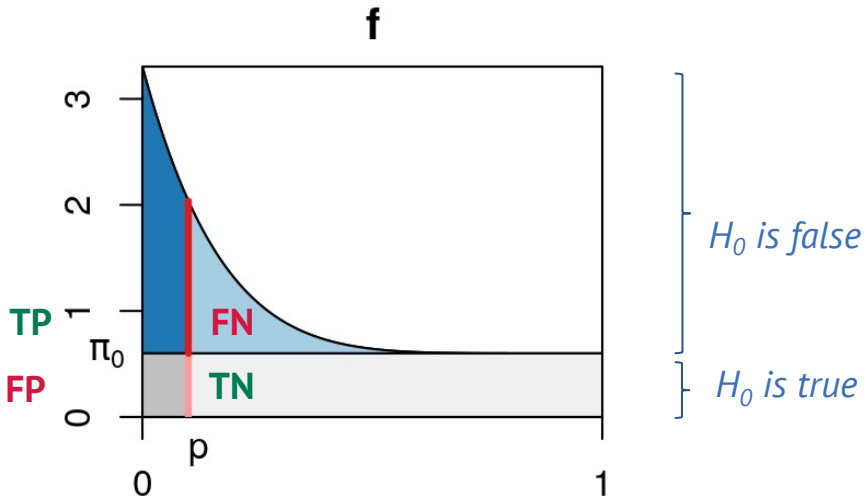
Figure: https://www.huber.embl.de/msmb/Chap-Testing.html#the-local-fdr

Multiple testing opportunity

Multiple testing doesn’t have to be a burden:

- Helps to estimate the uniform component (null is true)
- Helps to prevent over-optimism (type I error)
- The FDR has a much more useful interpretation than the p-value
    - Closer to the question “What is the probability that this hit is wrong?”

p-value histogram decomposition

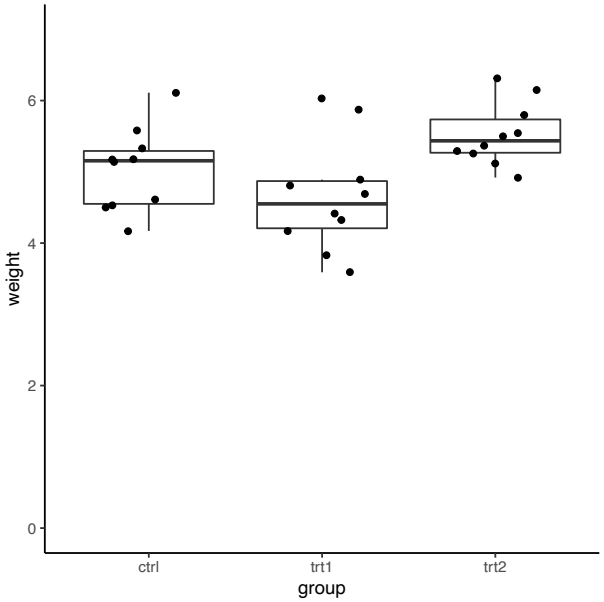


Question: What would change in the histogram if the test has a low power?

Figure: https://www.huber.embl.de/msmb/Chap-Testing.html#the-local-fdr

Pairwise comparisons

Plant growth data (from R)

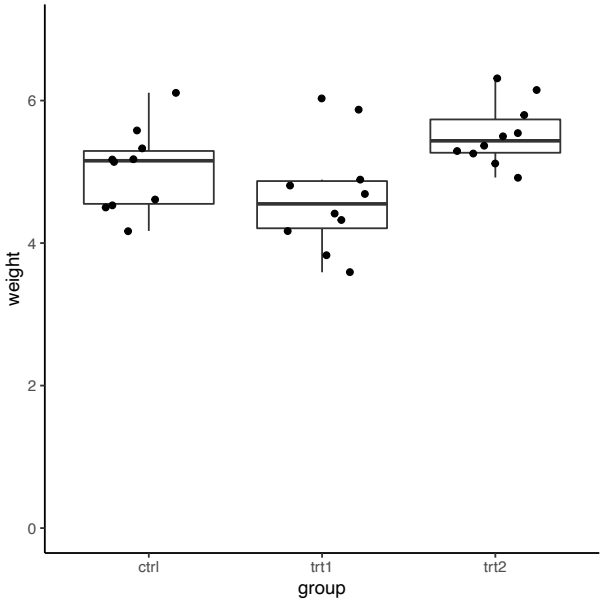


Questions:

- How can we test the individual differences?
- How many comparisons are possible in this data set?
- If we choose $\alpha = 0.05$, what is the probability of seeing at least one significant difference, if in fact all differences are 0?

$P(\text{at least one false rejection}) = 1 - (0.95)^3 = 0.14$

Pairwise comparisons



ANOVA F-test:

H_0: All differences are zero.
H_A: Not all differences are zero.

- ANOVA is used for comparing more than one mean.
- In case of two groups, ANOVA is equivalent to t-test.
- ANOVA controls for the family-wise error rate.

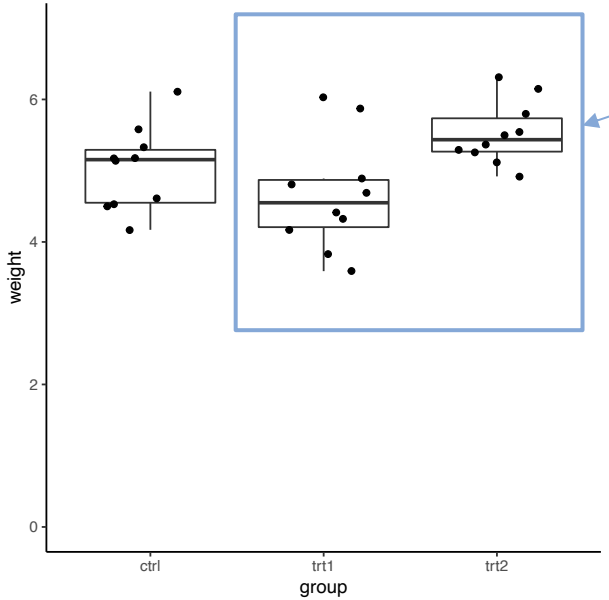
p=0.016

Tukey HSD:

Gives adjusted p-values for the individual comparisons.
Controls for the FWER.
H_0: all differences are zero.

Data snooping

Only performing tests that were suggested by the data.



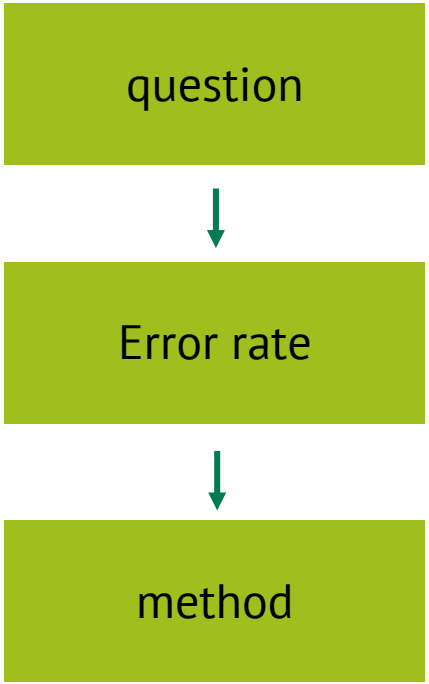
"Hmmm, these look different..."

p=0.009

Question: Can you think of scenarios where you

- don't have to adjust the p-values
- don't have to do ALL comparisons?

Workflow summary



The most important part is to know which error rate you want to control for.

If you know the error rate, the choice of method is mostly straight-forward.

Be honest to yourself when it comes to data snooping.

Summary error rates

EMBL

Error rate	Scenario	Examples
Comparison-wise (no adjustment)	Each test is an individual question.	Test for female and male mice separately whether the diet has an effect.
Family-wise	Control the probability of at least one false positive.	A protein is considered safe if no epitope causes a reaction. Pairwise comparisons: Is any of the differences non-zero?
False-discovery rate	Allow a few false positives to increase power.	Drug screens Screens for differentially expressed genes