Data screening and pre-processing multiple testing

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MED3007
Statistical Principles in Genomics: an Introduction with Rstudio
15.01.2024

1. Statistical methods for genomic data screening
    - Two-sample tests

2. Correction for multiple testing
    - Multiple testing correction procedures: FWER vs. FDR

Screening for candidates

Screening is a **testing** problem

A gene is declared **differentially expressed**, if an observed difference between two experimental conditions is greater than what would be expected under the null hypothesis.

Usually **effect** reported as Fold Change $= X/Y$
or $\log_2$ fold change $= \log_2(X/Y) = \log_2(X) - \log_2(Y)$

Two-sample tests

- parametric tests, e.g. t-test
- non-parametric tests, e.g. Wilcoxon rank sum tests
- distribution-free tests, e.g. permutation tests

Student's t-test

- Two samples $x = \{x_1 \ldots, x_{n_x}\}$ and $y = \{y_1, \ldots, y_{n_y}\}$

- Null hypothesis: $H_0 : \mu_x = \mu_y$
  Alternative hypothesis: $H_1 : \mu_x \neq \mu_y$

- The two-sample test statistic is

$$T = \frac{\bar{x} - \bar{y}}{s\sqrt{1/n_x + 1/n_y}} \overset{H_0}{\sim} t_{n_x+n_y-2}$$

where

$$s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x - 1) + (n_y - 1)}$$

is the pooled variance estimate, $\bar{x}$, $\bar{y}$ and s_x^2, s_y^2 are sample means and sample variances, n_x, n_y sample sizes

Student's t-test

- Compute the p-value for the observed value t of test statistic T as follows:

$$p = 1 - P_{H_0}(|T| \leq |t|)$$

- Decision rule: Reject H_0 if $p \leq \alpha$
- State the result: If $p \leq \alpha$, there is a statistically significant difference between group means at the significance level α.

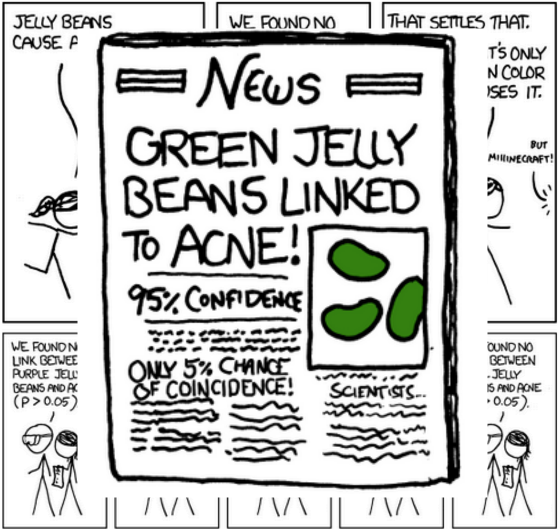
Potential problems when performing 2-sample tests

- Small sample sizes → not for this course!
  - The usual asymptotics might not hold (e.g. assumption of asymptotic normal distributions for t-test)
    → use permutation tests
  - Unreliable estimates of variability
    → stabilise individual variance estimates through shrinkage to global estimate
- Multiplicity problem
  - Thousands of hypotheses are tested simultaneously,
    increasing the chance of false positive findings.

1. Statistical methods for genomic data screening
   - Two-sample tests

2. Correction for multiple testing
   - Multiple testing correction procedures: FWER vs. FDR

Why correct for multiple testing anyway?

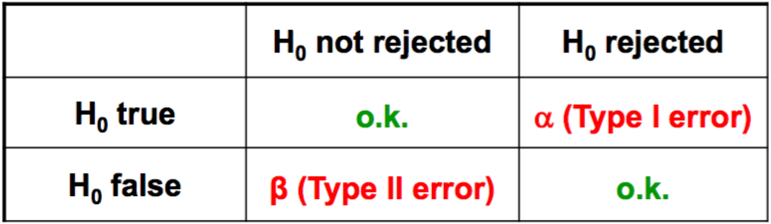

JELLY BEANS CAUSE A
WE FOUND NO
THAT SETTLES THAT.
T'S ONLY N COLOR SES IT.
BUT MINECRAFT!
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...
WE FOUND N LINK BETWEE PURPLE JELL BEANS AND AC (P > 0.05)
OUND NO BETWEEN . JELLY S AND ACNE > 0.05).

copyright: xkcd.com/882/

From single to multiple tests

Test Problem

Null hypothesis H_0 vs. alternative hypothesis H_1



	H_0 not rejected	H_0 rejected
H_0 true	o.k.	α (Type I error)
H_0 false	β (Type II error)	o.k.

Construction of the Test

Control the Type I error at a fixed significance level α (usually 0.05) and choose a test statistic that maximizes the power $1 - \beta$

From single to multiple tests

Suppose we perform **10 tests**, each with significance level $\alpha = 0.05$. Suppose that H_0 is true, so we should never reject. What is the probability that we will get at least one false positive decision?

P(at least one false positive decision) $=$

$= 1$ - P(all true negatives) $= 1 - (1 - 0.05)^{10} = 1 - (0.95)^{10} = 0.401$

Note that: 10 tests $\Rightarrow$ the probability is $1 - (1 - 0.05)^{10}$

If **increasing the number of tests,** probability goes to 1

100 tests $\rightarrow 1 - (1 - 0.05)^{100} = 0.994$
1000 tests $\rightarrow 1 - (1 - 0.05)^{1000} \approx 1$

Take-home message

When performing **many statistical tests,** which means when screening many variables (genes), then we are **certain** to select false positives!

How to correct for this? Intuition

Adjusting for M tests (AKA Bonferroni correction)

Adjust the significance level α_i **of each test** so that globally the significance level is the wanted ($\alpha =$ global significance level):

$$\alpha_i = \frac{\alpha}{M}, \qquad i = 1, \ldots, M$$

Increasing M (number of tests) **decreases** significance level α_i of each single test

10 tests → $1 - (1 - \frac{.05}{10})^{10} = 0.049$
100 tests → $1 - (1 - \frac{.05}{100})^{100} = 0.049$
1000 tests → $1 - (1 - \frac{.05}{1000})^{1000} = 0.049$

Intuitive take-home message

Multiple Testing Procedures protect against false positive conclusions

Multiple Testing Procedures: Counting Errors

Assume we are testing M null hypotheses: $H_{0i}, i = 1, \ldots, M$

Possible scheme of the situation:

	nr. **NOT rejected** H_{0i}	nr. **rejected** H_{0i}	tot
nr. **TRUE** H_{0i}	U	V	h_0
nr. **FALSE** H_{0i}	T	S	h_1
	G - R	R	G

with:

- $h_0 =$ number of true null hypotheses
- R $=$ number of rejected null hypotheses
- V $=$ number of type I errors (false positives)
- T $=$ number of type II errors (false negatives)

Controlling for Type I error **rates**

Family-wise error rate (FWER)

Probability of at least one false positive (type I error)

$$\text{FWER} := P(V \geq 1)$$

False discovery rate (FDR)

Expected proportion of false positives (type I error) among the total number of rejected null hypotheses

$$\text{FDR} := E(Q), \quad Q := \begin{cases} V/R, & \text{if} \quad R > 0 \\ 0, & \text{if} \quad R = 0 \end{cases}$$

Comparison FWER vs FDR

FWER

- extremely **conservative**, only few genes are called significant
- used when we **need to be certain** that all findings are truly positive (example: when making decisions about the admittance of medical treatments)
- **can miss out** on potentially important genes (false negatives)

FDR

- **used if FWER is too stringent,** that is, when more interested in having more true positives (the false positives can be sorted out in subsequent expensive experiments)
- **Cool fact:** by controlling the FDR one can choose how many of the subsequent experiments one is willing to perform in vain

Adjusting p-values for multiple testing

- For each variable (ex: gene) $i = 1, \ldots, M$ we test the null hypothesis H_{0i} and obtain the (unadjusted) p-value p_i
- We then apply a correction method (next slide) and obtain the **adjusted p-value** p_i^*
- We **reject** H_{0i} at significance level α if $p_i^* < \alpha$

How? Two possibilities

Single Step Procedures

Take M unadjusted p-values and adjust them independently

Step-Wise Procedures

Adjust p-values sequentially (ex: from the smallest to the largest)
More powerful

Common adjustment methods

For controlling FWER$< \alpha$: Bonferroni correction (remember the intuition!)

- **single-step procedure**
- $p_i^* = \min(M \times p_i, 1)$

For controlling FDR$< \alpha$: Benjamini & Hochberg correction

- **step-wise procedure,** independence assumption
- how to adjust?
  1. first order observed p_i's such that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M)}$
  2. $p_i^* = \min_{k=i,\ldots,M} \left(min(\frac{M}{k} \times p_{(k)}, 1) \right)$

Example: Adjusting p-values

Suppose you have tested 5 genes and got these p-values:
0.001, 0.021, 0.34, 0.88, 0.011

$rank(k)$	p_i	FWER (Bonferroni) p_i^*	FDR (Benj.-Hochb.) p_i^*
1	0.001		
2	0.011		
3	0.021		
4	0.34		
5	0.88		

* significant at 0.05 level

Bonferroni: $p_i^* = \min(M \times p_i, 1)$
Benjamini-Hochberg: $p_i^* = \min_{k=i,\ldots,M} \left(min(\frac{M}{k} \times p_{(k)}, 1) \right)$

Example: Adjusting p-values

Suppose you have tested 5 genes and got these p-values:
0.001, 0.021, 0.34, 0.88, 0.011

rank(k)	p_i	FWER (Bonferroni) p_i^*	FDR (Benj.-Hochb.) p_i^*
1	0.001	0.005*	
2	0.011	0.055	
3	0.021	0.105	
4	0.34	1	
5	0.88	1	

* significant at 0.05 level

Bonferroni: $p_i^* = \min(M \times p_i, 1)$
Benjamini-Hochberg: $p_i^* = \min_{k=i,\dots,M}\left(min(\frac{M}{k} \times p_{(k)}, 1)\right)$

Example: Adjusting p-values

Suppose you have tested 5 genes and got these p-values:
0.001, 0.021, 0.34, 0.88, 0.011

$rank(k)$	p_i	FWER (Bonferroni) p_i^*	FDR (Benj.-Hochb.) p_i^*
1	0.001	0.005^*	0.005^*
2	0.011	0.055	0.0275^*
3	0.021	0.105	0.035^*
4	0.34	1	0.425
5	0.88	1	0.88

* significant at 0.05 level

Bonferroni: $p_i^* = \min(M \times p_i, 1)$
Benjamini-Hochberg: $p_i^* = \min_{k=i,\ldots,M} \left(min(\frac{M}{k} \times p_{(k)}, 1)\right)$

Take-home messages

Screening genes (for ex. differentially expressed ones) is a statistical testing problem: we simultaneously test thousands of null hypotheses

- Unspecific gene filtering can reduce the number of tests
- **Multiple testing procedures** control for the different kinds of type I error rates such as FWER and FDR

> *"For outcome-related gene finding, the most common and serious flaw was an inadequate, unclear, or unstated method for controlling the number of false-positive differentially expressed genes." (Dupuy and Simon, 2007)*[1]

[1] Dupuy A., & Simon R. (2007). Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting, *J Natl Cancer Inst*, 99, 147–157.