# Píndoles estadístiques UEB-VHIR

## *Comparacions multiples I Multiple testing*
## *Perque, Quan, Com?*


***Alex Sanchez-Pla***
***Unitat d'Estadística i Bioinformàtica (VHIR)***
***Dept. de Genetica, Micro I Estadistca (UB)***


Dilluns 16 de Desembre de 12:30 a 13:30
Sala d'Actes de Traumatologia i Rehabilitació

Les píndoles estadístiques son sessions divulgatives, organitzades per la Unitat d'Estadística i Bioinformàtica (UEB) del VHIR, on es presenten problemes i solucions estadístiques dirigides als professionals interessats del Campus Vall d'Hebron

Vall d'Hebron Institut de Recerca VHIR

ueb.vhir.org

Vall d'Hebron Institut de Recerca

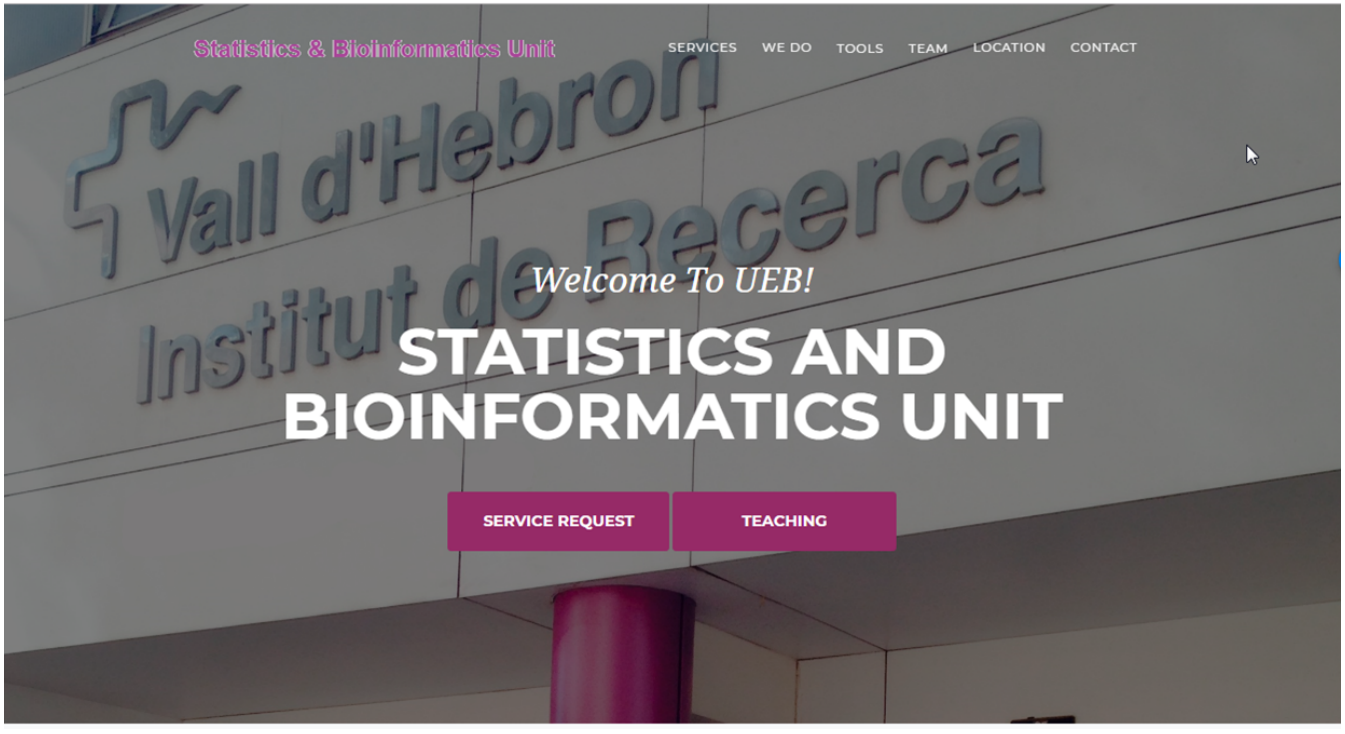
UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

UNIVERSITAT DE BARCELONA

GRBIO

# Statistics and Bioinformatics

Unit (UEB)



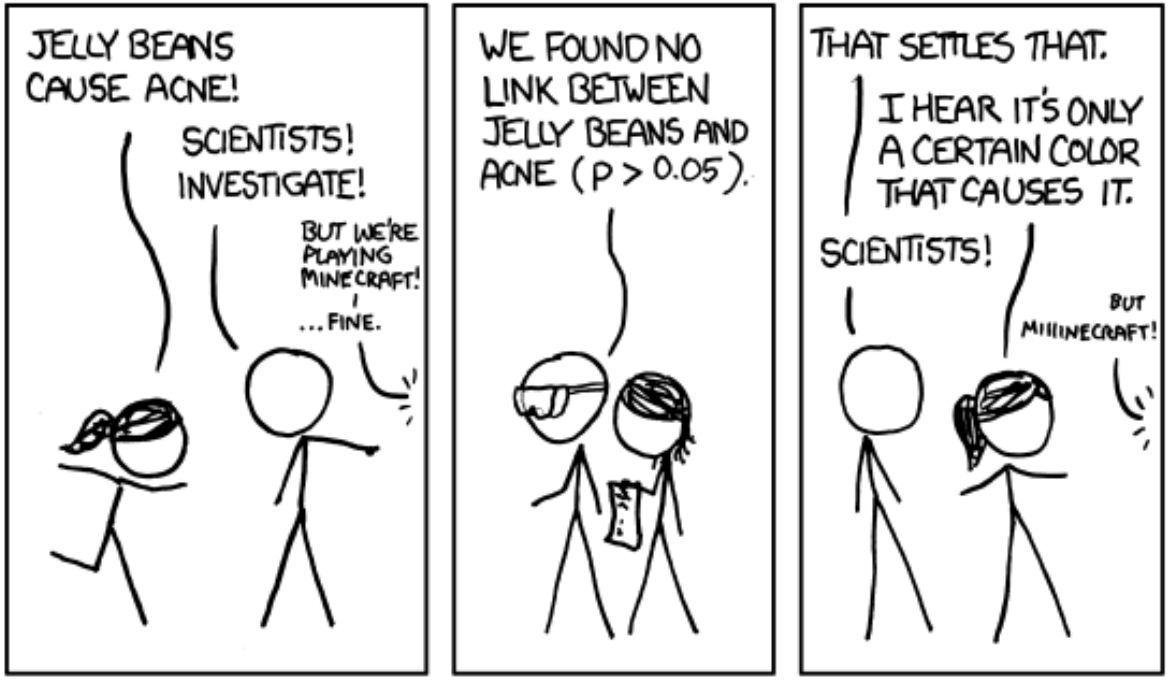

http://ueb.vhir.org

# Outline of the talk

- Why this pill. Motivation & Cases
- From Type I errors to Multiple Error Rates
- Strategies for Multiple Testing Adjustments
- Multiple testing adjustment in practice
- Variations on a theme
- Should we correct or not? When?
- Recomendations (guidelines)
- Summary

# When multiplicity is ignored ...

(Bad management) of multiplicity can yield potentially spurious results



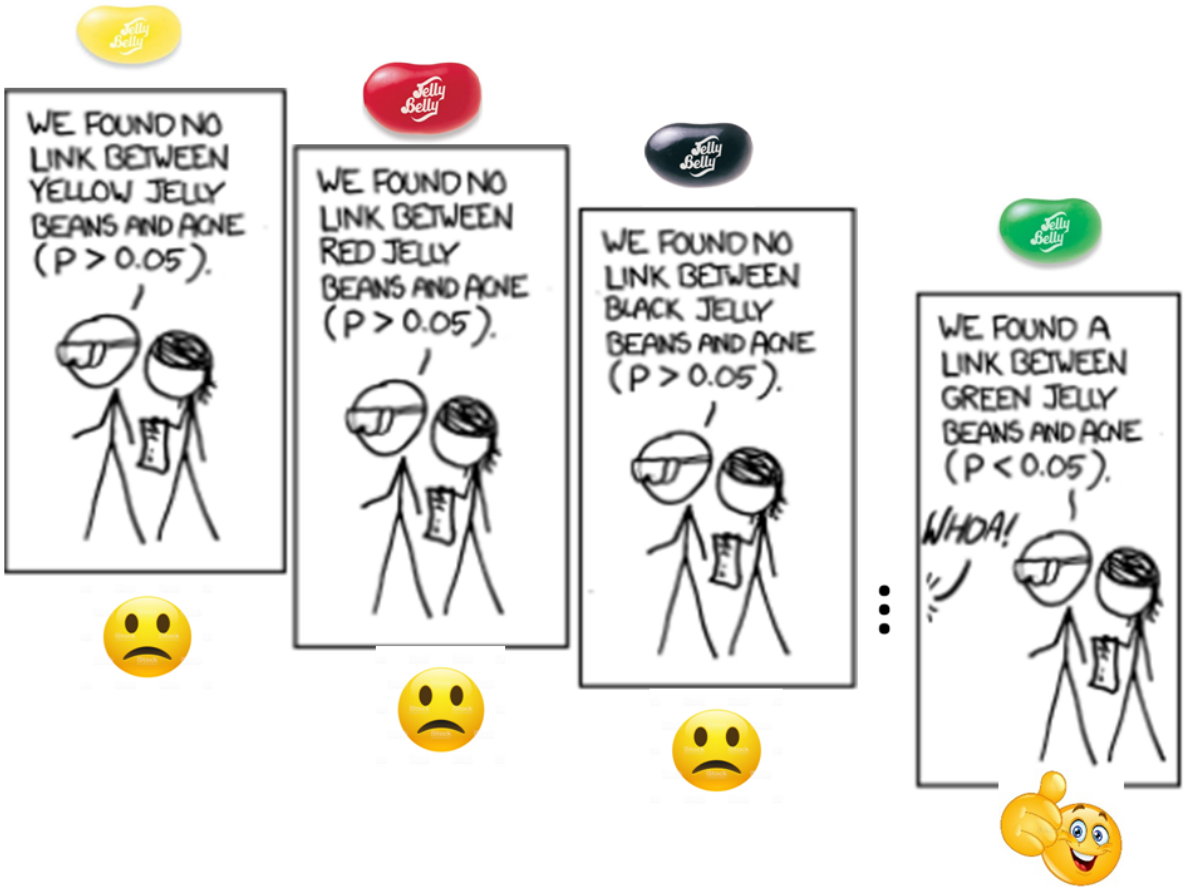

http://xkcd.com/882/

# When multiplicity is ignored ...

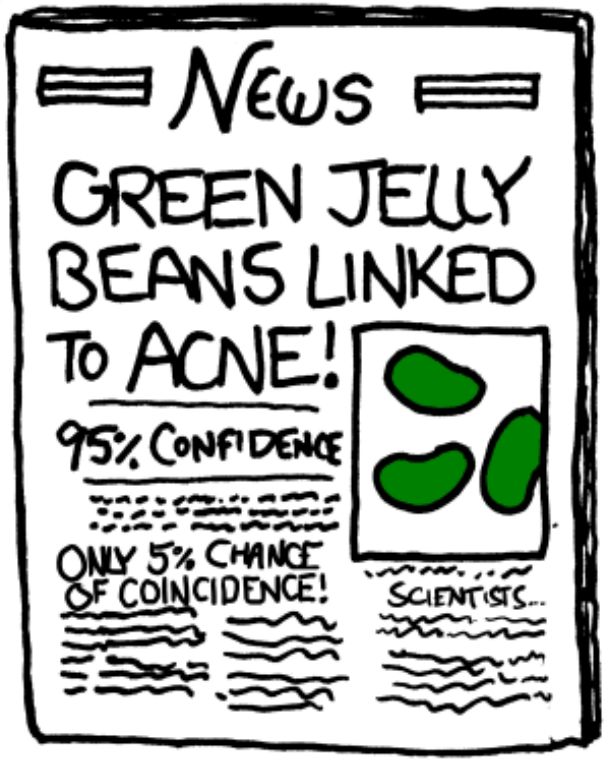
# When multiplicity is ignored ...

# When multiplicity is ignored ...

# When multiplicity is ignored ...



# When multiplicity is ignored ...



# But, what is multiplicity?

- The **multiple comparisons**, **multiplicity** or **multiple testing** problem occurs when one considers a set of statistical inferences simultaneously ...
- The more inferences are made, the more likely erroneous inferences are to occur.
- Multiplicity appears in many distinct situations

# Example: ***Multiple outcomes***

Frantic paresis association with distinct outcomes

| **Frantic paresis association with distinct outcomes** | | |
|---|---|---|
| *Categorical outcome* | *Odds-Ratio* | *P-value* |
| Tracheobronchitis | 2.80 | 0.0121 |
| Neumonia | 1.60 | 0.335 |
| Tracheostomy | 5.10 | 1.8E-07 |
| ICU-Mortality | 0.48 | 0.222 |
| *Numeric Outcome* | *Mean difference* | *P-value* |
| Mechanic Ventilation Days | 19 | 8E-06 |
| ICU days | 22 | 3.9E-06 |
| Hospital days | 27 | 0.00035 |

# Example: ***Several groups***

Raw p-values of post-Hoc (after ANOVA) pairwise comparisons

| | B | C | D |
|---|---|---|---|
| **A** | 0.045 | 0.098 | 0.062 |
| **B** | | 0.683 | 0.891 |
| **C** | | | 0.638 |

# Example: ***Omics data***

A simple microarray analysis yields tables with thousands of test results

| gen | SYMBOL | GENENAME | logFC | AveExpr | t | P.Value | adj.P.Val |
|---|---|---|---|---|---|---|---|
| 1 | Dcst1 | DC-STAMP domain containing 1 | 3.67559 | 4.69707 | 14.65697 | 0.000000011 | 0.000024208 |
| 2 | Dio2 | deiodinase, iodothyronine, type II | -3.56532 | 9.09379 | -14.60544 | 0.000000011 | 0.000024208 |
| 3 | Cdhr5 | cadherin-related family member 5 | 2.28073 | 6.70668 | 14.50082 | 0.000000012 | 0.000024208 |
| 4 | D630039A03Rik | RIKEN cDNA D630039A03 gene | 2.20739 | 2.46498 | 11.92132 | 0.000000097 | 0.000120170 |
| 5 | Gys2 | glycogen synthase 2 | -3.36608 | 7.13253 | -11.88752 | 0.000000100 | 0.000120170 |
| 6 | Dcst2 | DC-STAMP domain containing 2 | -1.81633 | 4.45705 | -11.15264 | 0.000000195 | 0.000188620 |
| 7 | Them7 | thioesterase superfamily member 7 | 2.11593 | 5.41287 | 11.02483 | 0.000000220 | 0.000188620 |
| 8 | Gk | glycerol kinase | -2.05576 | 9.74829 | -10.42198 | 0.000000394 | 0.000295272 |
| 9 | Cpn2 | carboxypeptidase N, polypeptide 2 | -1.77256 | 6.05494 | -9.95801 | 0.000000629 | 0.000418672 |
| 10 | St8sia6 | ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyl | 1.55094 | 6.50413 | 9.52587 | 0.000000987 | 0.000591609 |
| 11 | Eci3 | enoyl-Coenzyme A delta isomerase 3 | -2.16774 | 5.34872 | -9.12197 | 0.000001527 | 0.000832279 |
| 12 | Dhrs9 | dehydrogenase/reductase (SDR family) member | -3.00131 | 7.77390 | -9.02358 | 0.000001703 | 0.000850492 |
| 13 | Got1 | glutamic-oxaloacetic transaminase 1, soluble | 1.19586 | 9.55212 | 8.30667 | 0.000003868 | 0.001783300 |
| 14 | Cntnap1 | contactin associated protein-like 1 | -1.32228 | 6.13394 | -7.90574 | 0.000006264 | 0.002566810 |
| 15 | Slc4a4 | solute carrier family 4 (anion exchanger), memb | -1.19605 | 10.94534 | -7.88522 | 0.000006423 | 0.002566810 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 5690 | Cul5 | cullin 5 | -0.00052 | 7.44326 | -0.00337 | 0.99737 | 0.99859 |
| 5691 | Gm15753 | predicted gene 15753 | -0.00100 | 3.68334 | -0.00330 | 0.99742 | 0.99859 |
| 5692 | Bdh1 | 3-hydroxybutyrate dehydrogenase, type 1 | 0.00051 | 2.19023 | 0.00264 | 0.99794 | 0.99879 |
| 5693 | Pnn | pinin | 0.00061 | 7.59266 | 0.00261 | 0.99796 | 0.99879 |
| 5694 | Slco2a1 | solute carrier organic anion transporter family, | -0.00060 | 6.46110 | -0.00240 | 0.99812 | 0.99879 |
| 5695 | Nudt18 | nudix (nucleoside diphosphate linked moiety X)- | 0.00073 | 6.02751 | 0.00198 | 0.99845 | 0.99879 |
| 5696 | Mzt2 | mitotic spindle organizing protein 2 | 0.00055 | 4.56552 | 0.00198 | 0.99846 | 0.99879 |
| 5697 | Fig4 | FIG4 phosphoinositide 5-phosphatase | -0.00029 | 6.54977 | -0.00128 | 0.99900 | 0.99917 |
| 5698 | Rfc4 | replication factor C (activator 1) 4 | -0.00020 | 6.61395 | -0.00106 | 0.99918 | 0.99918 |

# Hypothesis Testing Refresher

- Most situations desbribed above can be described or related with a **test of hypothesis**.
- Tests use to be summarized with **p-values**.
- **p-value** : Probability, assuming no effect $(H_0)$, to obtain a difference greater or equal than the one observed on a given sample.
- Standard criterion: "reject $H_0$ if $p \geq \alpha$".

# Decision table and error types

- When decisions are made, based on data, one can take right or wrong decisions
- Wrong decisions: **type I** or **type II errors**.

| | | *Reported decision* | |
|---|---|---|---|
| | | **H0 is Accepted (No effect claimed)** | **H0 is Rejected (Effect claimed)** |
| ***State of the nature ("Truth")*** | **H0 is true (No Effect)** | TN , prob: $\beta$ 😊 | FP: Prob: $\alpha$ Type I error ☹ |
| | **H0 is false (Effect)** | FN: prob: 1-$\beta$ Type II error ☹ | TP , prob: 1-$\alpha$ 😊 |

# Controlling (type I) Errors

- A test is said to *control type I error* if the probability of wrongly rejecting $H_0$ is smaller than the significance level of the test.

$$P[\text{Reject } H_0 | H_0 \text{ true}] = P[FP] \leq \alpha$$

- This does not guarantee anything on the power of the test.
    - A test can control type I error while having small power

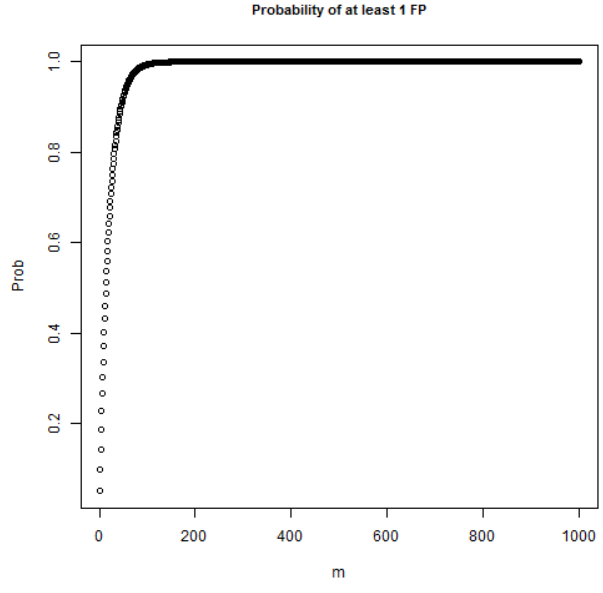
# From 1 to $> 1$ hypotheses

- As more hypothesis are tested simultaneously, the probability of wrongly rejecting **at least one true null** increases:

- P(Making 1 error) = $\alpha$
- P(**Not** Making 1 error) = $1 - \alpha$
- P(**Not** Making 1 error in 2 tests) = $(1 - \alpha)^2$
- 3, 4, ..., m tests
- P(**Not** Making 1 error in $m$ tests) = $(1 - \alpha)^m$
- P(Making **at least** 1 error in $m$ tests) = $1 - (1 - \alpha)^m$

# From 1 to $>$ 1 hypotheses

| m | Prob |
|---|---|
| 1 | 0.05000 |
| 2 | 0.09750 |
| 4 | 0.18549 |
| 8 | 0.33658 |
| 12 | 0.45964 |
| 100 | 0.99408 |
| 1000 | 1.00000 |



# Implications for our examples

- If we test multiple hypothesis simultaneously the overall type I error probability is not controlled anymore.
- Testing 12 tests simultaneosly yields almost a 50% chance of a statistically significant result **even if none of the hypothesis tested is false**
- How do we incorporate the impact of multiple testing on our inference?

# A simulated example (1)

- We simulated an omics study with 6000 genes whose expression has been measured on 8 cases and 8 controls, and where **no gene shows real difference between them**.
- What happens if we call *differentially expressed* any gene with $p < 0.05$
- The number of genes falsely rejected will be on average of $6000 \times \alpha$.

# We start with no differences ...

| Gene | p-value | t-Statistic | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | Y7 | Y8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.7079 | 0.3824 | 0.29 | 1.49 | 0.43 | -1.4 | -0.3 | -1.2 | 1.27 | 0.33 | -0.5 | 0.74 | -1.3 | 1.4 | 0.8 | -0.1 | -1.3 | -0.3 |
| 2 | 0.4381 | 0.7981 | -1.5 | -0.7 | 0.52 | -1.3 | -1.2 | -0.4 | -0.3 | 0.36 | -0.2 | 0.81 | -1.8 | -0.6 | -0.2 | -0.1 | -0.7 | 0.87 |
| 3 | 0.8077 | 0.2480 | -0.7 | 0.11 | 0.72 | ° | 1.4 | 0.24 | 0.98 | -0.3 | -0.5 | -0.2 | -0.5 | 1.53 | 1.93 | 0.74 | 0.86 | -0.1 |
| 4 | 0.8186 | 0.2337 | 0.23 | -0.5 | -1.3 | 0.05 | 0.3 | 0.26 | -1 | 0.1 | -0.9 | 0.22 | 0.22 | 0.68 | 0.74 | 0.48 | -2 | -0.6 |
| 5 | 0.4029 | 0.8626 | 0.59 | -0.2 | 0.5 | -0.5 | -1.7 | -1 | -1.4 | -0.2 | 0.22 | 0.11 | 1.96 | 0.19 | -0.9 | -0.2 | -0.1 | -1.7 |
| 6 | 0.6181 | 0.5099 | 0.25 | 0.95 | -0.1 | -0.9 | -1.2 | 0.78 | -0.5 | 0.12 | -0.2 | -0.5 | 2.08 | 1.09 | 0.56 | 0.32 | -1 | -1.1 |
| 7 | 0.7516 | 0.3229 | 0.31 | -1.5 | -0.5 | -0.4 | -2.5 | -0.6 | 0.25 | -0.8 | -1 | -0.2 | 0.58 | -0.5 | -0.8 | -0.9 | 1 | -2.5 |
| 8 | 0.2748 | 1.1365 | 0.72 | -2 | -0.4 | 0.28 | -0.8 | -0.4 | 0.94 | -0.8 | 0.82 | -0.2 | -2.3 | -0.1 | 1.19 | 0.32 | 1.33 | 1.34 |
| 9 | 0.7172 | 0.3696 | 0.52 | -0.1 | 1.2 | 1.74 | 0.66 | -0.4 | -0.8 | -0.2 | -0.9 | 0.65 | 0.05 | -0.2 | 0.07 | 1.49 | 0.42 | 0 |
| 10 | 0.5794 | 0.5675 | 1.38 | 0.38 | 0.56 | -1 | -0.9 | 0.71 | -0.1 | -0.3 | 2.65 | 2.38 | 1.15 | -0.7 | 0.19 | -1.7 | -0.3 | -0.3 |
| 11 | 0.1472 | 1.5346 | 1.09 | -1.4 | -1.3 | -0.3 | 0.99 | 1.35 | 0.59 | -1.9 | 1.9 | 0.98 | 0.8 | 1.34 | -0.6 | 0.13 | 0.07 | 1.25 |
| 12 | 0.3745 | 0.9173 | -1.7 | 2.84 | 1.85 | 1.43 | 1.32 | 0.13 | 1.41 | -0.9 | 1.6 | 0.05 | -0.9 | 0.32 | 0.66 | -0.6 | 0.54 | 0.24 |
| 13 | 0.2181 | 1.2896 | -0.9 | -1.8 | -1.9 | 0.4 | 0.08 | 0 | 0.68 | -0.6 | -2 | 0.01 | -0.5 | 0.35 | -0.8 | 1.67 | 1.7 | 1.59 |
| 14 | 0.9019 | 0.1256 | -0.5 | 1.24 | 1.21 | 0.88 | 1.94 | -0.7 | 1.01 | 2.19 | 0.52 | -1.3 | 1.57 | 0.84 | 3.3 | 0.73 | 0.83 | 1.34 |
| 15 | 0.1497 | 1.5242 | 1.12 | 1.29 | -1 | 0.06 | 0.58 | 0.4 | 1.21 | 0.65 | 1.85 | -1.4 | -0.4 | -0.9 | 0.76 | -1.5 | -0.3 | 0.29 |
| 16 | 0.7913 | 0.2697 | 1.72 | 0.49 | 0.41 | 0 | 0.26 | -0.1 | 0.67 | 0 | -0.1 | -0.4 | 0.18 | 0.76 | 0.25 | 1.68 | 0.12 | 1.59 |
| 17 | 0.3928 | 0.8818 | -0.5 | 0.79 | -0.2 | 0.16 | 1.65 | 2.06 | 0.27 | -0.8 | 0.24 | -0.1 | 0 | 1.93 | -1.5 | 1.19 | -1.4 | -0.9 |

Thes are "fake" microarray data
They have been generated from a unique normal distribution

Mean: -0.0090342

Standard deviation: 1.004293235

# As more genes are checked ...

| Gene | p-value | t-Statistic | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | Y7 | Y8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 211 | 0.0380 | 2.2915 | 0 | -0.2 | 1.23 | 0.76 | 0 | -0.3 | 0.1 | -1.1 | -1.7 | -0.6 | -1.3 | 0 | -1.3 | -0.8 | 0.53 | -0.8 |
| 212 | 0.4444 | 0.7870 | 0.44 | -0.7 | -0.3 | 0.1 | 0.52 | -1.5 | -0.7 | 1.02 | -2.8 | -2 | -0.8 | -0.9 | -1.2 | 1.29 | -0.6 | 1.97 |
| 213 | 0.1631 | 1.4720 | 1.21 | -0.9 | -0.5 | 1.48 | -0.1 | -0.4 | 1.49 | 0.78 | 0.01 | -2.3 | 0.1 | -1.3 | 0.18 | -0.2 | 0.06 | 0.82 |
| 214 | 0.0259 | 2.4921 | -1.2 | -1.5 | 1.27 | -0.2 | -0.9 | -1 | -2 | -1.3 | 0 | -0.9 | 1.02 | 1.11 | 0.07 | 0.91 | -0.5 | 0.39 |
| 215 | 0.9993 | 0.0008 | 0.89 | -0.3 | -0.4 | -0.1 | 0.33 | 0.5 | -1.7 | 0.74 | -0.2 | 2.04 | 0.1 | -0.4 | -0.4 | -0.3 | -1.2 | 0.47 |
| 216 | 0.6719 | 0.4326 | 1.78 | 1.18 | -2.2 | 1.5 | 0.83 | -1.3 | 0.53 | -0.7 | -0.3 | 1.26 | 0.59 | -0.8 | 0.64 | -1.4 | -1.2 | 0.79 |
| 217 | 0.9994 | 0.0008 | -0.6 | -0.7 | 0.07 | -2.1 | -0.8 | -1.2 | 2.38 | -1.7 | -0.9 | -0.9 | -0.4 | 0.49 | -1.4 | -0.7 | -0.3 | -0.7 |
| 218 | 0.2880 | 1.1044 | -0.6 | 0.48 | 1.48 | 0.32 | -2.1 | 0.31 | -0.4 | -0.3 | 0.34 | -0.2 | -0.5 | -1.3 | -0.5 | -0.6 | -1.6 | -0.2 |
| 219 | 0.2511 | 1.1972 | -0.2 | 0.16 | 0.77 | -1.2 | -1.7 | -0.4 | 0.73 | 0.92 | -0.2 | -1.1 | 0.27 | -0.1 | -0.9 | -1 | -0.6 | -0.9 |
| 220 | 0.9673 | 0.0418 | -0.6 | -1.9 | -0.6 | 1.13 | -2.1 | 0 | 1.41 | 1.13 | -0.2 | 0.03 | -0.1 | -0.8 | -0.2 | -1.3 | -0.1 | 0.96 |
| 221 | 0.8993 | 0.1288 | 0.2 | 0.36 | -0.2 | 2.08 | 0.28 | -0.8 | -0.6 | 0.3 | 0.9 | 1.35 | -0.3 | 0.64 | 0.5 | -0.7 | -0.1 | -0.2 |
| 222 | 0.4814 | 0.7233 | -0.3 | 2.14 | -1.3 | 0.89 | 0.3 | 0 | -0.6 | -0.4 | -1.4 | 0.12 | -0.2 | -1.8 | 1.43 | -0.9 | 0.15 | 0.38 |
| 223 | 0.7884 | 0.2736 | -0.8 | -0.5 | -0.8 | 0.42 | -1.6 | 0.02 | -0.3 | -0.8 | -0.6 | -1.2 | -1.1 | -0.4 | -0.6 | 0.53 | 0.44 | -0.7 |
| 224 | 0.6595 | 0.4502 | -1.1 | -1 | 0.97 | -0.3 | -0.6 | -0.1 | -0.8 | 0.77 | -1.8 | -0.3 | 1.14 | 0.86 | -0.4 | -0.6 | -0.1 | 0.62 |
| 225 | 0.9715 | 0.0363 | 1.67 | 1.59 | -0.6 | 1.95 | -0.9 | -1.6 | 0.74 | 0.52 | 0.91 | 1.04 | 0.68 | 1.84 | -0.6 | 0.73 | -0.1 | -1.4 |
| 226 | 0.2204 | 1.2828 | -0.7 | 0.42 | -0.5 | -0.6 | -1 | -0.7 | -0.2 | -0.4 | 1.4 | -0.3 | 1.24 | 0.11 | 1.18 | -1.5 | -0.8 | -0.7 |
| 227 | 0.0258 | 2.4928 | 0.19 | -0.1 | -0.8 | 0.02 | -0.2 | 0.62 | -1.4 | -0.3 | 0.42 | 0.83 | 0 | 1.97 | 0.58 | 0.49 | -0.4 | 0.83 |
| 228 | 0.8879 | 0.1436 | 0.16 | -0.9 | 0.94 | 2.05 | -1.8 | 0.59 | -0.2 | 0.07 | 0.01 | 0 | 1.76 | -1.6 | -0.5 | -0.3 | 0.62 | 0.28 |
| 229 | 0.1398 | 1.5654 | -1.9 | -0.7 | 1.27 | 0.38 | 1.75 | -0.1 | 1.11 | 1.72 | -1 | -1.1 | -1.7 | 0.21 | -0.8 | 0.68 | 0.42 | -0.1 |
| 230 | 0.2118 | 1.3084 | -1.2 | 1.44 | -3 | 0.75 | -0.4 | 0.99 | 0.75 | -1.2 | 0.59 | -0.9 | 1.75 | 1.59 | -1 | 1.83 | 1.18 | 0.15 |
| 231 | 0.4661 | 0.7493 | -0.2 | -0.6 | 0.68 | 0.12 | 0.02 | -1.1 | 0.16 | -0.6 | 0.88 | 0.2 | 0.01 | 0.75 | -1.2 | -0.6 | 0.69 | -0.3 |
| 232 | 0.6779 | 0.4242 | -1.5 | 0.24 | 1.38 | 0.53 | -0.9 | 0.07 | 1.01 | -0.9 | 0.42 | -0.3 | -0.9 | 0.17 | 0.08 | 0.51 | -0.2 | -1.4 |
| 233 | 0.5496 | 0.6131 | 1.7 | -1.4 | 0.08 | -1 | 0.03 | 0.72 | 0.06 | 0.2 | -0.6 | -0.3 | -0.4 | -0.5 | -0.4 | 0.28 | 0.87 | -0.7 |
| 234 | 0.8899 | 0.1410 | 0.69 | -0.1 | -1.7 | 0.24 | -0.5 | 0.77 | 1.61 | -2.1 | 0.18 | 0.46 | -0.4 | -1.4 | -0.7 | -0.2 | 1.67 | -0.1 |
| 235 | 0.0238 | 2.5346 | -1.3 | 0.38 | -0.2 | -0.9 | -1.6 | -0.4 | -2.4 | -0.5 | -0.2 | -0.4 | 0.05 | 0.58 | 0.21 | -1.1 | 1.5 | 1.95 |
| 236 | 0.9046 | 0.1220 | 0.42 | -0.5 | -0.6 | -1.5 | -1.1 | 0.25 | -0.6 | -0.6 | 1.03 | -1.5 | 0.05 | -0.2 | 0.62 | -1.8 | -1.3 | -0.6 |

# All together, sorted by p-values

| Gene | p-value | t-Statistic | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | Y7 | Y8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2132 | 9.2801E-05 | 5.40482888 | -0.9 | 0.01 | -0.9 | -0.3 | -0.6 | -1.5 | -1.6 | -1.5 | 0.1 | 0.9 | 0.32 | 0.32 | 0.99 | 0.13 | 0.97 | 0.61 |
| 2381 | 0.000422777 | 4.587026656 | -0.1 | 0.4 | 1.8 | 0.6 | 0.6 | 1.8 | 0.3 | 0.9 | 0.13 | 1.13 | 1.17 | 0.01 | 0.05 | 0.78 | 0.43 | 0.6 |
| 707 | 0.000446161 | 4.558778883 | -( | | | | | | | | | | | | .4 | -0.7 | -0.8 | -0.3 |
| 2945 | 0.000662585 | 4.352658186 | -1 | | | | | | | | | | | | 01 | 0.15 | 1.73 | -0.7 |
| 1306 | 0.000682671 | 4.337186373 | -( | | | | | | | | | | | | .1 | 0.99 | 0.78 | 0.19 |
| 664 | 0.001271354 | 4.017625809 | 2.26 | 0.73 | 1.72 | 0.77 | 0.59 | -0.2 | 1.1 | 0.77 | -0.8 | 0 | -1.7 | -0.1 | 0.65 | -1.2 | -0.2 | -0.8 |
| 3906 | 0.001743125 | 3.857057903 | -0.6 | -0.3 | -0.3 | -1.3 | -0.4 | -0.1 | -0.8 | -0.3 | -0.8 | 1.51 | 0.93 | 0.76 | 1.12 | 0.86 | 0.41 | 0.49 |
| 2882 | 0.002499024 | 3.67479096 | 1.46 | 0.93 | 2.09 | 0.34 | 0.92 | 0.77 | 2.2 | 0.57 | -0.8 | -1.6 | -1.2 | 1.61 | -0.9 | -0.5 | 0.3 | -0.7 |
| 4366 | 0.002643532 | 3.64642387 | 0.74 | -1.1 | -1.5 | 0 | -0.4 | -2.1 | -2.1 | -1.4 | 0.22 | 1.55 | 1.38 | -0.3 | 0.41 | -0.1 | 0.81 | 1.04 |
| 2428 | 0.002736875 | 3.628922003 | -1.7 | -0.2 | -0.9 | -2.2 | -0.8 | -0.6 | 0.67 | -0.8 | -0.7 | 1.35 | 1.92 | 0.58 | 1.59 | 0.32 | 0.71 | 0.19 |
| 2925 | 0.003053207 | 3.573796247 | 2.07 | 0.92 | 1.09 | -0.3 | 0 | 3.38 | 0.76 | -0.4 | -1.2 | -1.9 | -1.4 | -1.1 | -0.9 | -0.2 | -1 | -0.2 |
| 3884 | 0.003195558 | 3.550845365 | 0.5 | 1.41 | 1.1 | 0 | 0.52 | 1.74 | 0.21 | 0.77 | 0.38 | -0.9 | -1.3 | -0.4 | -0.3 | -0.9 | 0.87 | -0.7 |
| 5367 | 0.003260925 | 3.54064957 | -1.3 | -0.9 | -0.5 | 0.98 | -2.2 | -1.2 | -0.8 | -1.3 | -1.1 | 1.48 | 0.24 | 0.48 | 1.63 | 0.57 | 0.75 | 1.55 |
| 545 | 0.00328505 | 3.536938555 | 1.79 | -0.1 | 0.37 | 0.81 | 0.27 | 0.32 | 1.85 | 1.07 | 0.16 | -0.2 | -1.1 | -1.3 | 0.25 | -0.8 | -0.1 | 0 |
| 1209 | 0.003300778 | 3.534534023 | -0.1 | 1.45 | -0.1 | 1.64 | 0.92 | 1.03 | 1.21 | 0.75 | 0 | 0.19 | -1.6 | 0.09 | -0.5 | -2.3 | 0.14 | -0.5 |
| 1072 | 0.003392546 | 3.520730005 | 0.38 | -0.5 | -1.8 | -2.5 | -0.4 | 0.41 | -0.8 | -1.3 | 0.9 | 1.75 | 2.27 | 1.96 | 1.44 | -0.1 | 0.75 | -0.9 |
| 2186 | 0.003412249 | 3.517815306 | -0.2 | 0.67 | 0.33 | 0.71 | -0.5 | 0.17 | 0.05 | 0.68 | -0.7 | -0.8 | -0.9 | -0.4 | -0.6 | -2.1 | -0.7 | 0.15 |
| 4168 | 0.00355642 | 3.496989878 | -1.8 | -1.4 | -2.5 | 0.11 | -1.5 | -0.4 | 0.59 | -0.1 | 0.84 | 0.7 | 0.1 | 1.39 | 2.1 | 0.19 | 0.85 | -0.1 |
| 3045 | 0.003654273 | 3.483333503 | -0.2 | 1.18 | -0.4 | -1.6 | -0.3 | -0.5 | -1.4 | -1.5 | 0.14 | 0.64 | 0.91 | 0 | 0.96 | 1.74 | 1.98 | 0.52 |

There are 291 genes with a p-value of less than 0.05,
We know however that all samples come from the same distribution
So we have detected 291 false postives.
We expected 6000 * 0.05 = 300 FP on average

# So what can be done?

- Intuitive idea: Doing many tests increases the chances of calling false positives,
- This may be compensated
  - Using *more restrictive error rates*, for instance 0.01 or 0.005 instead of 0.05.
  - Adjusting ("correcting") the p-values to compensate for the number of tests.

# Distinct Error Rates

- Individual error rate (IER)
  - Error rate of a single test.
  - For a test with 5% significance level the IER is 0.05
- Global Error Rate
  - Error rate for one or several groups of tests.
  - For a group of tests each with 5% significance the global error rate is > 5%

# Decision table for many tests

- With many tests we count discoveries

| | | *Reported decision* | | |
|---|---|---|---|---|
| | | **Accepted ("Non Discoveries")** | **Rejected ("Discoveries")** | **Total** |
| ***State of the nature ("Truth")*** | **True Null Hypotheses** | TN 😊 | FD ☹ | $m_0$ |
| | **False Null Hypotheses** | FN ☹ | TD 😊 | $m_1$ |
| | **Total** | ***N*** | ***D*** | $m$ |

**FWER = P (FD > 1)**

**FDR = Avg (FD / D)** (D>0)

# Two main error rate extensions

- Family Wise Error Rate (FWER)
  - FWER is probability of at least 1 False Discovery
  - FWER = P (FD > 0)
- False Discovery Rate (FDR)
  - FDR is expected value of proportion of False Discoveries among all Discoveries .
  - FDR = E (FD/D; D>0)

# FWER / FDR control procedures

- FWER
  - Bonferroni
  - Holm (1979)
  - Hochberg (1986)
- FDR
  - Benjamini & Hochberg (1995)
  - Benjamini & Yekutieli (2001)

# Controlling the FWER (Bonferroni)

- Bonferroni procedure: Adjust significance level for number of tests performed (m)
  - Use significance level $\alpha/m$,
- Equivalently, adjust p-values multiplying all p-values by m.
- Other, more efficient procedures available: See a statistician

# Example. Presenting data.

- García-Arenzana et al. (2014) tested associations of 25 dietary variables with mammographic density, an important risk factor for breast cancer, in Spanish women.
- They found the following results (only first 10 are shown)

| Dietary variable | *P* value |
|---|---|
| Total calories | <0.001 |
| Olive oil | 0.008 |
| Whole milk | 0.039 |
| White meat | 0.041 |
| Proteins | 0.042 |
| Nuts | 0.06 |
| Cereals and pasta | 0.074 |
| White fish | 0.205 |
| Butter | 0.212 |
| Vegetables | 0.216 |

See complete example

# Example. Bonferroni (FWER)

| | | number of tests: | 10 |
|---|---|---|---|
| critical value: | 0.05 | Adjusted critical value: | 0.005 |
| | | | |
| **Dietary variable** | **↓ P-values ↓** | **Bonferroni-corrected significance** | Bonferroni-adjusted P-value |
| Total calories | 0.001 | **significant** | 0.01 |
| Olive oil | 0.008 | **not significant** | 0.08 |
| Whole milk | 0.039 | **not significant** | 0.39 |
| White meat | 0.041 | **not significant** | 0.41 |
| Proteins | 0.042 | **not significant** | 0.42 |
| Nuts | 0.06 | **not significant** | 0.6 |
| Cereals and pasta | 0.074 | **not significant** | 0.74 |
| White fish | 0.205 | **not significant** | 1 |
| Butter | 0.212 | **not significant** | 1 |
| Vegetables | 0.216 | **not significant** | 1 |

# Controlling the FDR (B & H)

- Benjamini-Hochberg procedure: Provides control of FDR for a fixed FDR value
    - 5% FDR: On average, 5% of your significant findings will be false
- Important: FDR is not an individual error rate.
    - A number higher than 0.05, such as 0.10 or 0.25 can be used

# Benjamini & Hochberg

- Procedure is relatively simple
  - Order the p-values
  - To provide control at a $Q$ FDR value compare i-th smallest p-value to $i \times Q/m$
- Instead of setting the FDR at a fixed value and establishing significance/non significance an, **adjusted p-value** may be computed.

# Example. B-H (FDR)

| false discovery rate (Q) | 0.1 | | number of P-values (m) | 10 |
|---|---|---|---|---|
| | | | | |
| ↓ labels (optional) ↓ | ↓ P-values ↓ | Benjamini-Hochberg significance | Benjamini-Hochberg P-value | |
| Total calories | 0.001 | **significant** | 0.0100 | |
| Olive oil | 0.008 | **significant** | 0.0400 | |
| Whole milk | 0.039 | **significant** | 0.0840 | |
| White meat | 0.041 | **significant** | 0.0840 | |
| Proteins | 0.042 | **significant** | 0.0840 | |
| Nuts | 0.06 | **not significant** | 0.1000 | |
| Cereals and pasta | 0.074 | **not significant** | 0.1057 | |
| White fish | 0.205 | **not significant** | 0.2160 | |
| Butter | 0.212 | **not significant** | 0.2160 | |
| Vegetables | 0.216 | **not significant** | 0.2160 | |

# Example: Adjustments with R

Statistics in Action with R

Hypothesis testing ▾ Regression models ▾ PK modelling ▾ Mixed effects models ▾ Mixture models ▾

## Statistical tests: multiple comparisons

*Marc Lavielle*
*January 14th, 2019*

# What error rate to control for

- FWER Controls for no (0) false positives
  - Rejects many fewer hypothese (less false positives),
  - but you are likely to miss many.
  - Adequate if goal is to identify few cases that differ between two groups.

# What error rate to control for

- FDR Controls the (expected) proportion of false positives
  - if you can tolerate more false positives
  - you will get many fewer false negatives
  - Adequate if goal is to pursue the study e.g. to determine functional relationships among genes

# Should we adjust for MT?

- This a controversial issue.
- Many authors are in favour
  - Moyé:“ *Type I error accumulates with each executed hypothesis test and must be controlled by the investigators*”
  - Blakesley et al. : “ *Failure to control type I errors when examining multiple outcomes may yield false inferences, which may slow or sidetrack research progress*”

# Or shouldn't we?

- Rothman: "No adjustments are needed for multiple comparisons"

- *Reducing the type I error for null associations increases the type II error for those associations that are not null*

- *A policy of not making adjustments for multiple comparisons is preferable because it will lead to fewer errors of interpretation when the data under evaluation are not random numbers but actual observations on nature.*

- *Scientists should not be so reluctant to explore leads that may turn out to be wrong that they penalize themselves by missing possibly important findings.*

# Summary: Basic principles

1. The multiple comparisons problem (MCP) **should not be ignored**.

2. **Limiting the number of outcomes and subgroups** is one of the best ways to address the MCP.

3. The MCP should be addressed by **first structuring the data**.
   Furthermore, protocols for addressing the MCP **should be made before data analysis is undertaken**.

# Developing a Strategy for MT

1. Delineate separate outcome domains in the study protocols.
2. Define confirmatory and exploratory analysis components prior to data analysis.
3. As a general rule consider adjusting for multiple testing in confirmatory analysis.
4. As a general rule exploratory analysis does not require adjusting for multiple testing.
5. Specify which subgroups will be part of the confirmatory analysis and which ones will be part of the exploratory analysis.

# Developing a Strategy (II)

6. Apply multiplicity adjustments in experimental designs with multiple treatment groups.
7. Design the study to have sufficient statistical power for examining intervention effects for all prespecified confirmatory analyses.
8. Qualify confirmatory and exploratory analysis findings in the study reports.

# References


1. Blakesley RE, Mazumdar S, Dew MA, et al. Comparisons of Methods for Multiple Hypothesis Testing in Neuropsychological Research. Neuropsychology. 2009;23(2):255-264. doi:10.1037/a0012850
2. Moyé LA. P-value interpretation and alpha allocation in clinical trials. Ann Epidemiol. 1998;8(6):351-357. doi:10.1016/s1047-2797(98)00003-9
3. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990;1(1):43-46. doi:10.1097/00001648-199001000-00010
4. Streiner DL. Best (but oft-forgotten) practices: The multiple problems of multiplicity-whether and how to correct for many statistical tests. Am J Clin Nutr. 2015;102(4):721-728. doi:10.3945/ajcn.115.113548
5. View of Planned Hypothesis Tests Are Not Necessarily Exempt From Multiplicity Adjustment | Journal of Research Practice. http://jrp.icaap.org/index.php/jrp/article/view/514/417. Accessed November 25, 2019.
6. Asante I, Pei H, Zhou E, et al. Exploratory metabolomic study to identify blood-based biomarkers as a potential screen for colorectal cancer. Mol Omi. 2019;15(1):21-29. doi:10.1039/c8mo00158h
7. Matsui S. Confirmatory and Exploratory Analyses in Omics Studies with Particular Focus on Multiple Testing and P-value. Japanese J Biometrics. 2018;38(2):127-139. doi:10.5691/jjb.38.127
8. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between Exploratory and Confirmatory Preclinical Research Will Improve Translation. PLoS Biol. 2014;12(5). doi:10.1371/journal.pbio.1001863
9. Sethuraman A, Gonzalez NM, Grenier CE, et al. Continued misuse of multiple testing correction methods in population genetics-A wake-up call? Mol Ecol Resour. 2019;19(1):23-26. doi:10.1111/1755-0998.12969
10. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29(4):1165-1188. doi:10.1214/aos/1013699998
11. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x
12. Hauser S, Wakeland K, Leberg P. Inconsistent use of multiple comparison corrections in studies of population genetic structure: Are some type I errors more tolerable than others? Mol Ecol Resour. 2019;19(1):144-148. doi:10.1111/1755-0998.12947
13. 10 Things to Know About Multiple Comparisons | Egap. https://egap.org/methods-guides/10-things-you-need-know-about-multiple-comparisons. Accessed November 9, 2019.
14. Konishi T. Microarray test results should not be compensated for multiplicity of gene contents. BMC Syst Biol. 2011;5(Suppl 2):S6. doi:10.1186/1752-0509-5-S2-S6

# QUESTIONS?