

Multiple Testing

GWAS

Christian Werner

(Quantitative geneticist and biostatistician) **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)



HerrFalloppio

Oscar González-Recio

(Computational biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)



OscarGenomics



simple testing

- **inference** → is there a **difference** between groups?
 - e.g. AA vs AB vs BB
- **significance** is related to the **size** and **variance** of this difference
- **p-value**: prob of obtaining such an extreme t statistics under H_0 given we repeat the experiment an infinite number of times
 - P-value $< \alpha$ → small likelihood of the data under H_0 → significant difference
 - P-value $> \alpha$ → there is a high chance of observing these data if there is no difference between groups
- $\alpha = 0.05$ → threshold: 5% of rejecting H_0 when it is true (Type I error).
 - **false positive**: significant result when there is no difference (H_0 is true)









multiple testing

- many tests → many false positives
 - e.g. 2000 (independent) tests, $\alpha=0.05$ → How many expected false positives?
100 false positives by chance alone
- multiple testing problem
- GWAS:
 - many SNPs, many statistical tests, many p-values



How to cope with the problem

- Increase the sample size
(e.g. Bio Banks)
- Reduce the number of tests
 - Based on LD
 - Choose relevant regions (functional analysis)
- Decrease the significance threshold
 -  ○ **Arbitrary significance level** (e.g. 5×10^{-8})
 -  ○ **Bonferroni correction**
 -  ○ **False discovery rate**
 -  ○ ***q* values** (*important pitfalls*)
 -  ○ **Permutation analysis**
 -  ○ **Go Bayesian...**

Bonferroni correction

- Bonferroni, mathematician (1892 - 1960)
- **adjust** the significance threshold:
 - **New significance threshold $\leq \alpha/m$**
[m: number of tests (markers)]
- Bonferroni correction tends to be too conservative
 - few false positives
 - many false negatives

False discovery rate (FDR)

- Decrease the significance threshold

0.010
0.025
0.026
0.031
0.042
0.049
0.050
0.065
0.078
0.101
0.125
0.128
...

List of ordered
p-values

- 1) If I apply a threshold α to decide on significance, how much can I trust the results?
- 2) Where should I draw a line (threshold) of significance so that at most e.g. 10% of results are false positives?



False discovery rate (FDR)

- **FDR:** how many of the positive results are false positives?
- Benjamini & Hochberg (1995), Storey (2002), Storey & Tibshirani (2003)
- **Significance level = 0.05** → 5% of **all** tests on average will be false positives (assuming independency)
- **FDR = 0.05** → 5% of **significant** tests will on average be false positives



fewer false positives!

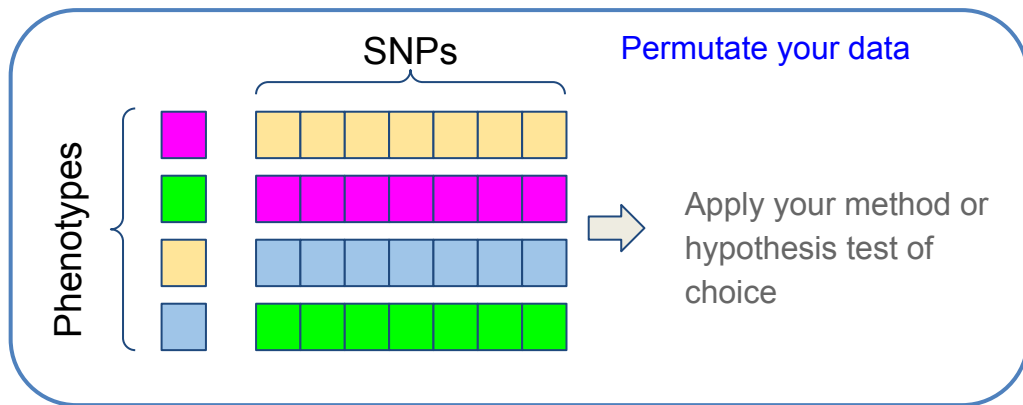
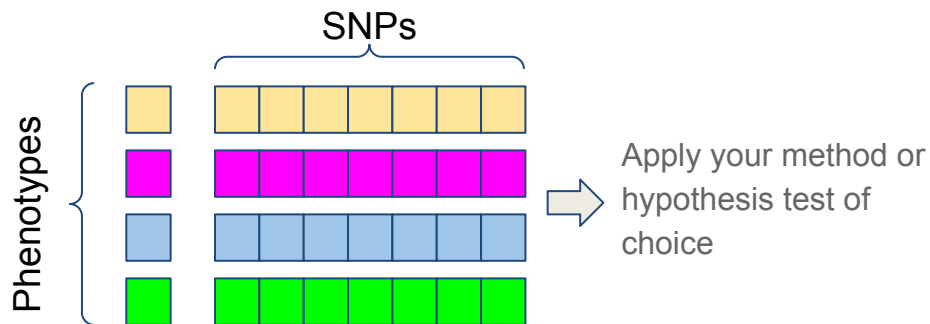


Permutation tests

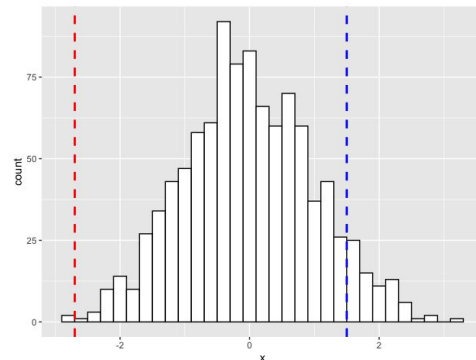
- Determine the significance of a result by randomly reshuffling the data and recalculating the test statistic.
 - This allows to test the null hypothesis that the observed difference between two groups is due to chance, rather than a real difference between the groups.
- Permutation tests are often used when the assumptions of traditional parametric tests are not met, or when the sample size is small.
- They are also useful when the data is not normally distributed, or when the groups being compared are not independent.



Permutation tests



x large number of times
(e.g. x1000)



Non-significant, within 95%HPD

Significant, out of 95%HPD



Bayesian inference

- In Bayesian inference, the probability of a hypothesis is updated using Bayes' theorem as follows

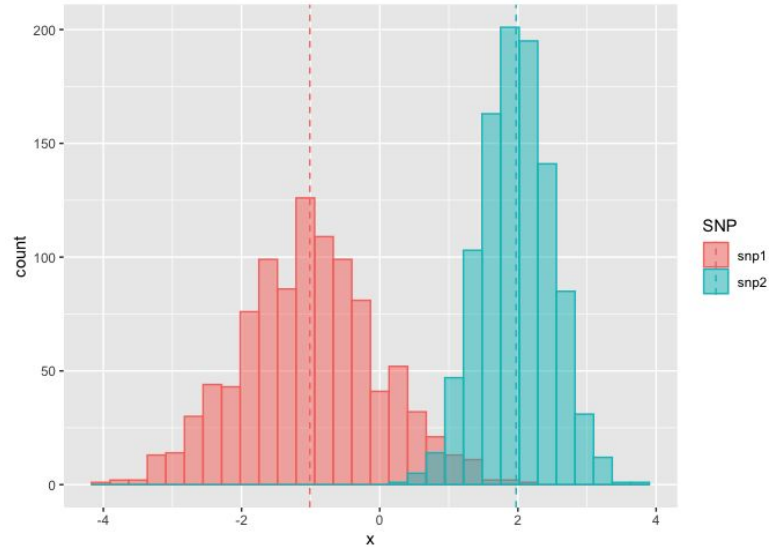


$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- where $P(\boldsymbol{\theta} | y)$ is the updated probability of the effect given the new evidence, $P(y|\boldsymbol{\theta})$ is the likelihood of the evidence given the effect, $P(\boldsymbol{\theta})$ is the prior probability of the effect, and $P(y)$ is the probability of the evidence.
- Make inferences of the posterior distribution using MCMC algorithms (Gibbs sampling, acceptance rejection, Metropolis-Hasting)

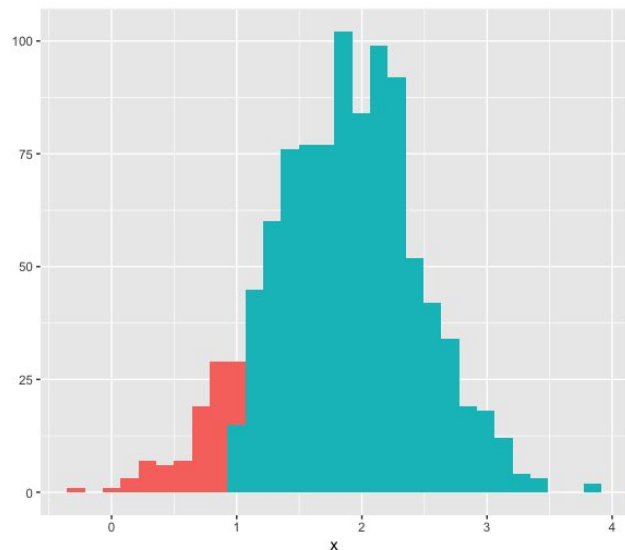
Bayesian inference

- What is the mean of the posterior distribution and its standard deviation?
- Does it contain zero?



Bayesian inference

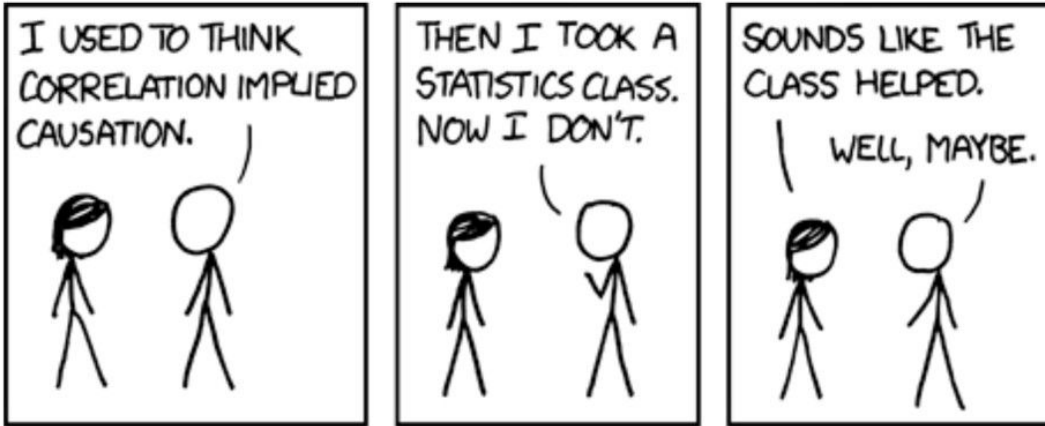
- What is the probability of the effect being larger than a relevant magnitude (e.g. +1).
- Is it a sufficient probability (e.g. 80%)



Possible to combine Bayesian inference and permutation test

REMEMBER

- Correlation does not imply causation



<https://xkcd.com/552/>

Make your rationale choice

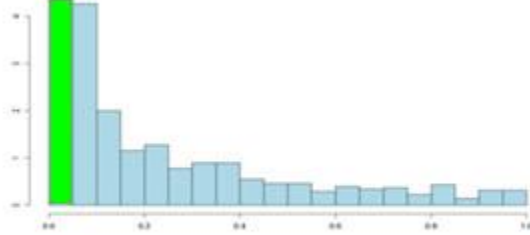
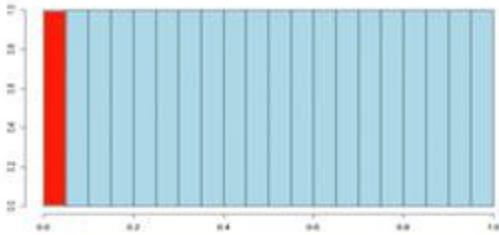
NEXT LECTURE

Power of GWAS experiments



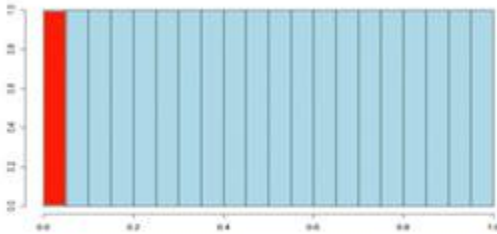
q-values

- q-values: proxies for FDR based on the **distribution of p-values**

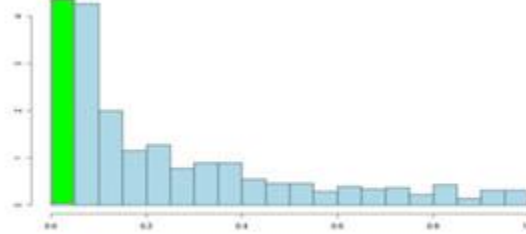


q-values

- q-values: proxies for FDR based on the **distribution of p-values**



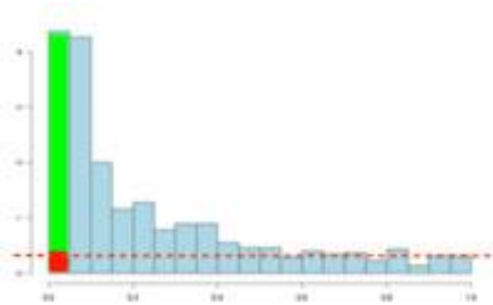
no significant
differences



significant
differences

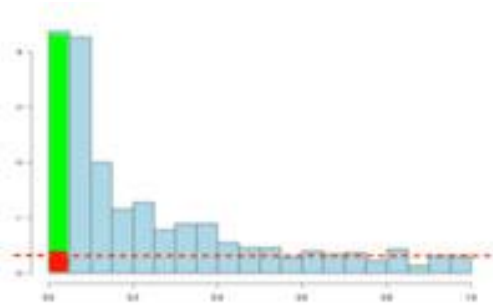
q-values

- the q-value approach tries to find the proportion of significant results which are likely to be false positives
- intuitively, it finds the height (density) at which the distribution of p-values flattens out

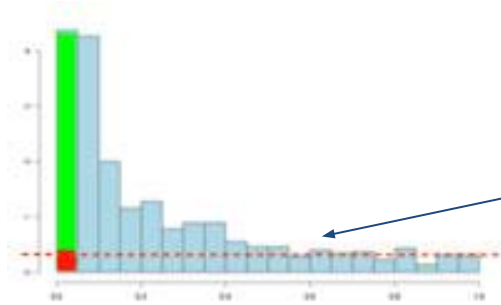


q-values

- the q-value approach tries to find the proportion of significant results which are likely to be false positives
- intuitively, it finds the height (density) at which the distribution of p-values flattens out



q-values



here the distribution is similar to the case where there is no actual difference

- this proportion of false positives is then incorporated in the calculation of adjusted p-values (**q-values**)



interpretation of q-values

- *Significance level* = 0.01 \rightarrow probability of the p-value under H_0
- q-value = 0.02 \rightarrow probability of the SNP being a false positive
- *Significance level* = 0.01 \rightarrow 1% chance of false positives (e.g. 7900 SNPs \rightarrow 79 false positives expected)
- q-value = 0.02 \rightarrow 2% of positive results may be false positives (e.g. 800 SNPs with q-value ≤ 0.02 \rightarrow 16 false positives expected)

interpretation of the
single SNP

interpretation of the
distribution of SNPs



q-values

- What's **wrong** with **q-values**?
 - They assume p-value is the probability of rejecting the null hypothesis when it is true
 - They do not consider that p-values are drawn from a probability distribution, and assume an infinite repetition of the experiment (obtaining different p-values for each experiment).

