

SIB
Swiss Institute of
Bioinformatics



Enrichment analysis

Linda Dib

21st of march 2018

Bioinformatics
Core FacilitySwiss Institute of
Bioinformatics

[Home](#)
[People](#)
[Research](#)
[Publications](#)
[Services](#)
[Teaching](#)
[Resources](#)
[Partners](#)
[Contact](#)

Welcome to **BCF-SIB**

[About](#)[History](#)[Location](#)

About **BCF-SIB**

The Bioinformatics Core Facility (BCF) is a research and service group within the [Swiss Institute of Bioinformatics \(SIB\)](#). Our core competence and activities reside in the interface between biomedical sciences, statistics and computation, particularly in the application of high-throughput omics technologies, such as gene-expression microarray, to problems of clinical importance, such as development of cancer biomarkers. The BCF offers consulting, teaching and training, data analysis support and research collaborations for both academic and industrial partners.

History

The BCF was initially founded in 2002 as a data analysis support group within the [NCCR Molecular Oncology](#), serving mostly biomedical research groups in Lausanne, Switzerland, mainly at the Institute of Experimental Cancer Research



Bioinformatics
Core Facility



Swiss Institute of
Bioinformatics

[Home](#)
[People](#)
[Research](#)
[Publications](#)
[Services](#)
[Teaching](#)
[Resources](#)
[Partners](#)
[Contact](#)

Teaching and Training

[Upcoming](#) [Past](#)



The BCF provides researchers with educational support and practical training in the use of software and analysis methods. This includes the organization of seminars, workshops, statistical software training courses, and teaching in the regular curriculum at the University of Geneva, the University of Lausanne and the EPFL.

The range of topics we have covered includes:

- Introduction to statistics in biomedical sciences
- R statistical software and BioConductor
- Transcriptomics analysis (microarray analysis, RNAseq and qPCR)

These courses are available at both introductory or advanced level. Most courses are taught over a full week; some specialized workshops can be organized over one day, including:

- General statistics in biomedical sciences (for people who want to understand statistics but won't use them directly)
- Multivariate Analysis
- Integration of data from several sources
- Graphical representation of life science data
- Data analysis and reproducible research

We can also offer these courses "in-house", or develop custom courses tailored to your needs and level, according to your requirements. Please contact stat@isb-sib.ch if you have any question.

Upcoming

Our courses upcoming courses are announced on the [SIB education web page](#). You can also [sign up](#) to remain informed about the education activities at the SIB.

The organization of our courses depends strongly on the interest of potential participants. If you have any question or suggestion, please contact stat@isb-sib.ch



Bioinformatics
Core Facility



Swiss Institute of
Bioinformatics

[Home](#)
[People](#)
[Research](#)
[Publications](#)
[Services](#)
[Teaching](#)
[Resources](#)
[Partners](#)
[Contact](#)

Services

SIB Biostat Teaching Consulting Analysis Collaboration Embedding



SIB Biostatistics Support

The BCF provides a consulting service on biostatistics matters, on a mandate from (and partially funded by) the SIB and the Swiss Confederation. This service is aimed at all people active in life sciences in Switzerland. It includes training and teaching, consulting, data analysis, and research collaboration, with a focus on high-throughput technologies in genomics or proteomics.

The service can be provided on a collaborative basis or for a fee, depending on the circumstances: among other factors, the origin and goals of the request (academy or industry), the amount of work involved and our current workload will be taken into account in determining the service provided. For academic groups that require long-term support, we strongly advise to start a discussion at the grant-submission step, and to include a request for a part-time bioinformatician in the grant. By pooling such part-time positions, the BCF is able to offer a longer-term dedicated support.

Consulting usually starts with a short meeting discussing the questions asked. Often, this is enough to help the researcher solve the problem. In other cases, the meeting allows us to define the different possibilities for a forthcoming collaboration.

For more information, please contact us at stat@isb-sib.ch or by calling Frédéric Schütz at +41 21 692 40 94 or Charlotte Soneson at +41 21 692 40 91.

Teaching and Training

We provide short courses and workshops, as well as longer but low-intensity semester courses. More information about recent and upcoming courses is available on the [SIB education web page](#). The [Teaching](#) page holds information about courses up to 2011. You can also [sign up](#) to remain informed about the education activities at the SIB.

Schedule

9:00	- 10:30	<i>Recall differential expression</i> <i>Recall statistical tests</i> <i>Exercise</i>
10:30	- 10:45	coffee break
10:45	- 12:30	<i>Threshold-based versus Threshold-free enrichment methods</i> <i>GSEA advantages and drawbacks</i> <i>Classification of available gene enrichment methods</i> <i>Exercise</i>
12:30	- 13:30	lunch (on your own)
13:30	- 15:30	<i>Generalizing enrichment</i> <i>Exercise</i>
15:30	- 15:45	coffee break
15:45	- 17:00	<i>Ontologies and enrichment</i> <i>Exercise</i>
17:00		end of day

Questions

Anytime, by raising your hands

Course web-page

Course page: <https://edu.sib.swiss/course/view.php?id=333>

Login: ea18

Password: SIB-ea18

Credits

Who?

This course worth 0.25 credits

Pre-requisites

R beginner level,
Elementary statistics

Suppose that two classes of students had grade scores in Reading Comprehension at the end of the third grade. Each class followed a different teaching method. Considering that the grades are normally distributed and of the same variance. How would you assess the efficiency of the two teaching methods in R?

29 responses

t test (2)

T-test (2)

A simple t-test would be enough.

`t.test(grades_class1,grades_class2,var.equal=TRUE)`

Les données sont stockées dans deux vecteurs différents (x et y).

`q1<-t.test(x, y, alternative=c("two.sided", "less", "greater"),var.equal=TRUE)`

with a student's t-test

Two Sample t Test with equal variances

xx

I would assess the efficiency of the two teaching methods by performing a Student's t test in R and set the p-value to 0.05 (if the obtained p-value is smaller, the two teaching methods differ in their efficiency).

you could do a t-test, or use a linear model.

using a linear model (`lm()` function) or anova that compares means between groups

Suppose that two classes of students had grade scores in Reading Comprehension at the end of the third grade. Each class followed a different teaching method. Considering that the grades are normally distributed and of the same variance. How would you assess the efficiency of the two teaching methods in R?

29 responses

t test (2)

T-test (2)

A simple t-test would be enough.

`t.test(grades_class1,grades_class2,var.equal=TRUE)`

Les données sont stockées dans deux vecteurs différents (x et y).

`q1<-t.test(x, y, alternative=c("two.sided", "less", "greater"),var.equal=TRUE))`

with a student's t-test

Two Sample t Test with equal variances

xx

I would assess the efficiency of the two teaching methods by performing a Student's t test in R and set the p-value to 0.05 (if the obtained p-value is smaller, the two teaching methods differ in their efficiency).

you could do a t-test, or use a linear model.

using a linear model (`lm()` function) or anova that compares means between groups

Now suppose that the two classes of students had several grade scores (a) one in Reading Comprehension (b) one in writing skills (c) one in math. How would you assess the efficiency of the two teaching methods in R? Hint: we have to compare the two groups of students several times - what would you do once the p-values are extracted? (The grades are assumed to be normally distributed and of the same variance)

28 responses

ANOVA (2)

Multiple t test or paired t-test, but I would have to google ... :-/

The extracted p-values must be corrected for multiple comparisons in order to avoid Type-I errors.

p.adjust(vector_w_pvalues)

p<c(pvalue1,pvalue2)
q2<-p.adjust(p, method = bonferroni, n = 2)

Multiple comparison testing (ANOVA)

perform three student's t-tests and correct the p-values (for example with bonferroni correction)

xx

I would perform multiple Student t-tests (one per comparison) and perform a multiple-testing correction such as Bonferroni's correction (dividing the p-value by the number of tests carried, here $0.05/3 = 0.0166$ and take that as the p-value threshold for significance for each test)

You would need to correct for multiple comparisons. The easiest way is Bonferroni's correction, where you divide the threshold of significance by the number of tests. There's also the Benjamini-Hochberg correction,

Now suppose that the two classes of students had several grade scores (a) one in Reading Comprehension (b) one in writing skills (c) one in math. How would you assess the efficiency of the two teaching methods in R? Hint: we have to compare the two groups of students several times - what would you do once the p-values are extracted? (The grades are assumed to be normally distributed and of the same variance)

28 responses

ANOVA (2)

Multiple t test or paired t-test, but I would have to google ... :-/

The extracted p-values must be corrected for multiple comparisons in order to avoid Type-I errors.

```
p.adjust(vector_w_pvalues)
```

```
p<-c(pvalue1,pvalue2)  
q2<-p.adjust(p, method = bonferroni, n = 2)
```

Multiple comparison testing (ANOVA)

perform three student's t-tests and correct the p-values (for example with bonferroni correction)

xx

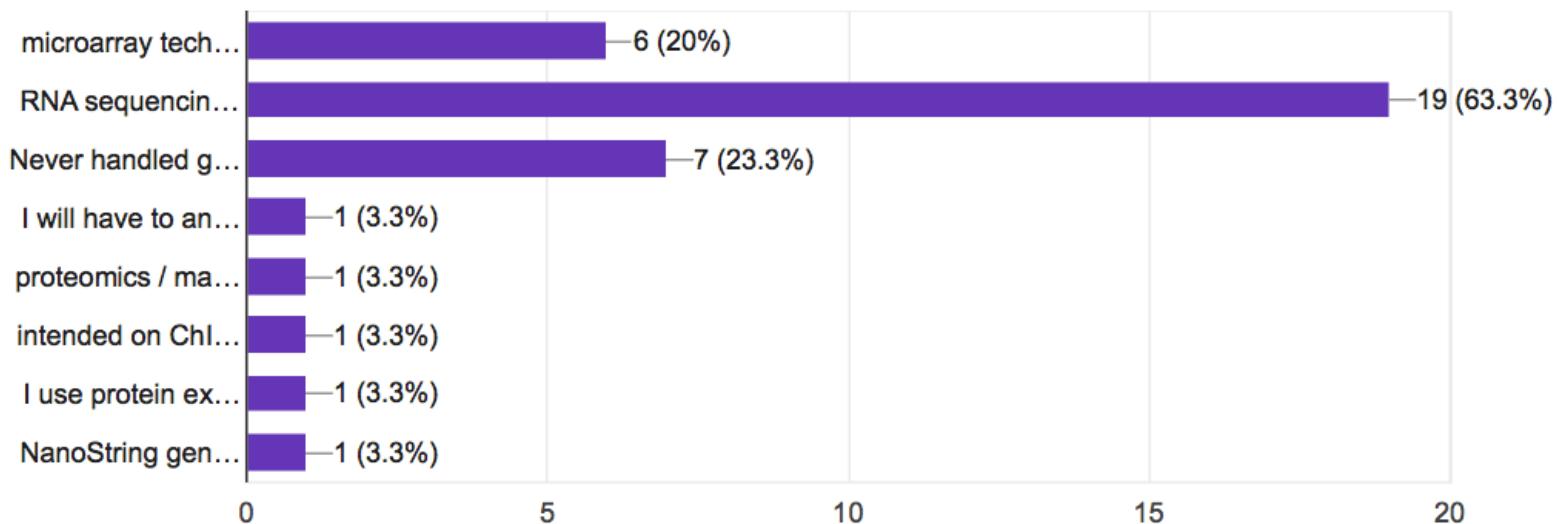
I would perform multiple Student t-tests (one per comparison) and perform a multiple-testing correction such as Bonferroni's correction (dividing the p-value by the number of tests carried, here $0.05/3 = 0.0166$ and take that as the p-value threshold for significance for each test)

You would need to correct for multiple comparisons. The easiest way is Bonferroni's correction, where you divide the threshold of significance by the number of tests. There's also the Benjamini-Hochberg correction,

Did you analyse gene expression issued from



30 responses



RNA-seq pipeline

1. Check the quality of the reads

- FastQC
- cutAdapt to trimm

2. Map to your favorite genome

- TopHat, star, Hisat2

3. Sort , create, index bam files

- SAMTOOLS

4. Control mapping and quality

- RNAseq QC, Qualimap, noiseQC

5. Generate count matrix

- summarizeOverlaps , featureCounts, tximport , htseq-count

6. Check for batch effect, normalization and correction

7. Differential expression of counts - **based on statistics**

- using Limma*, edgeR, DESeq2,...

8. Enrichment analysis given a phenotype - **based on statistics**

Overview

Count matrix

Differential expression

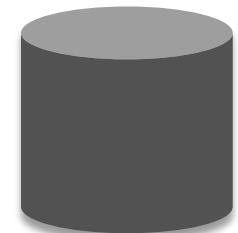
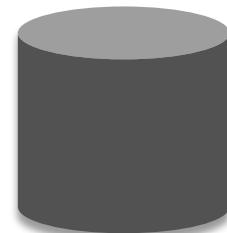
Enrichment

SEA

GSEA

MEA

Knowledge



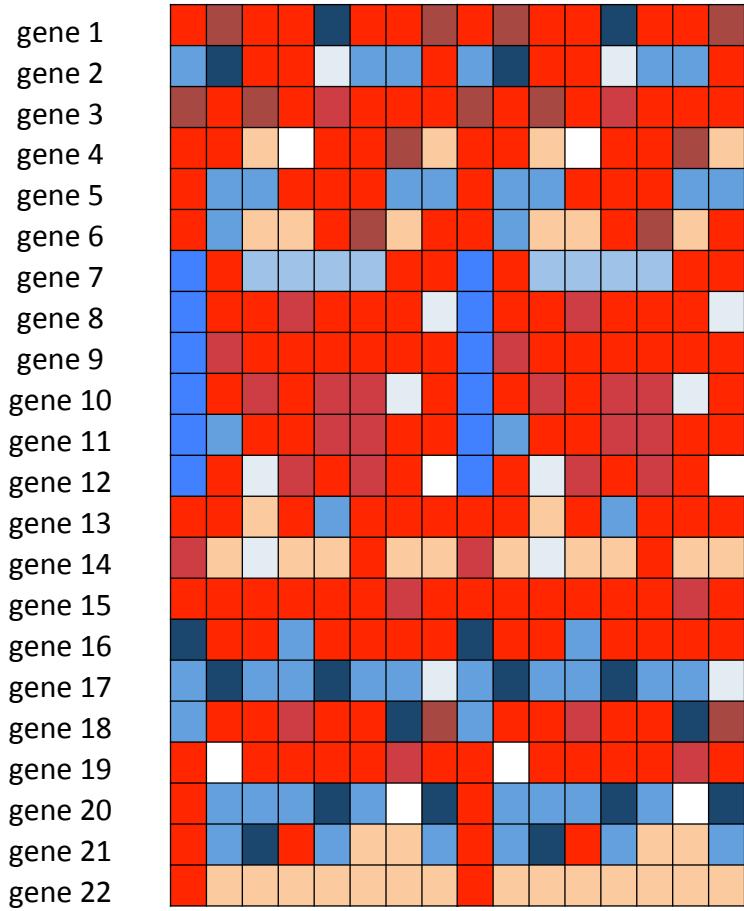
Overview

Count matrix

Differential
expression

High-throughput **expression** data

mRNA

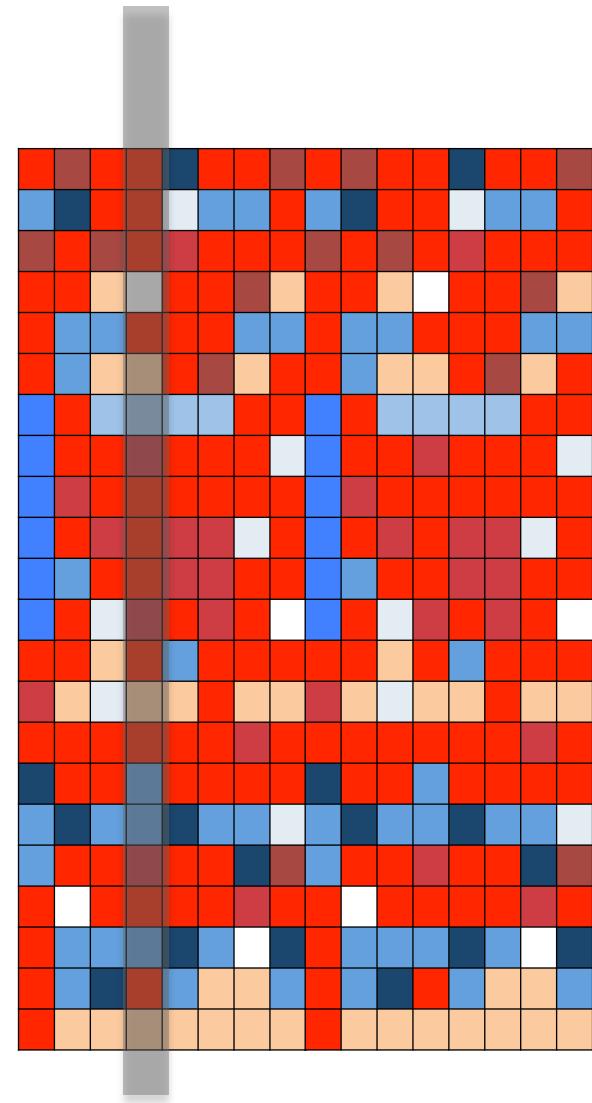


High-throughput expression data

Count matrix

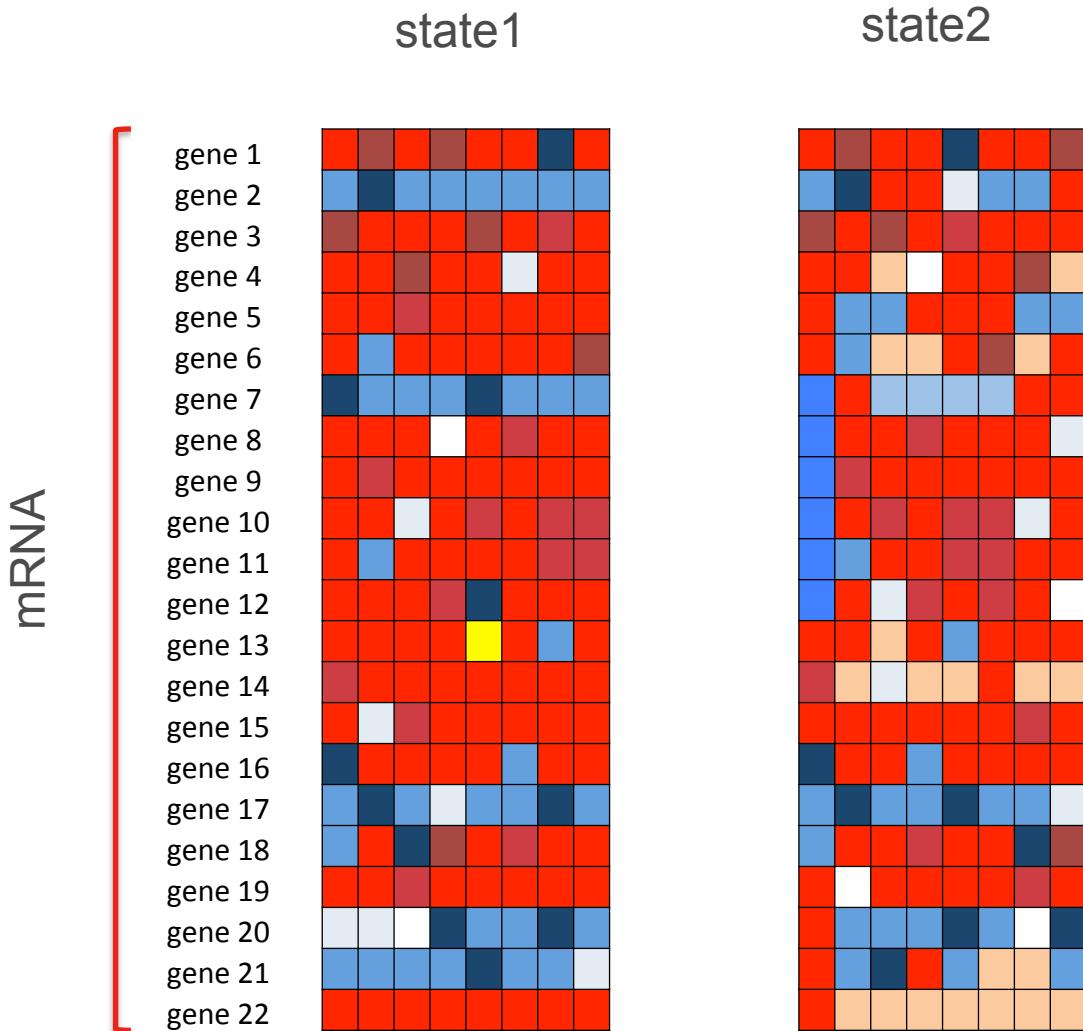
mRNA

Patient, mouse, cell, ...



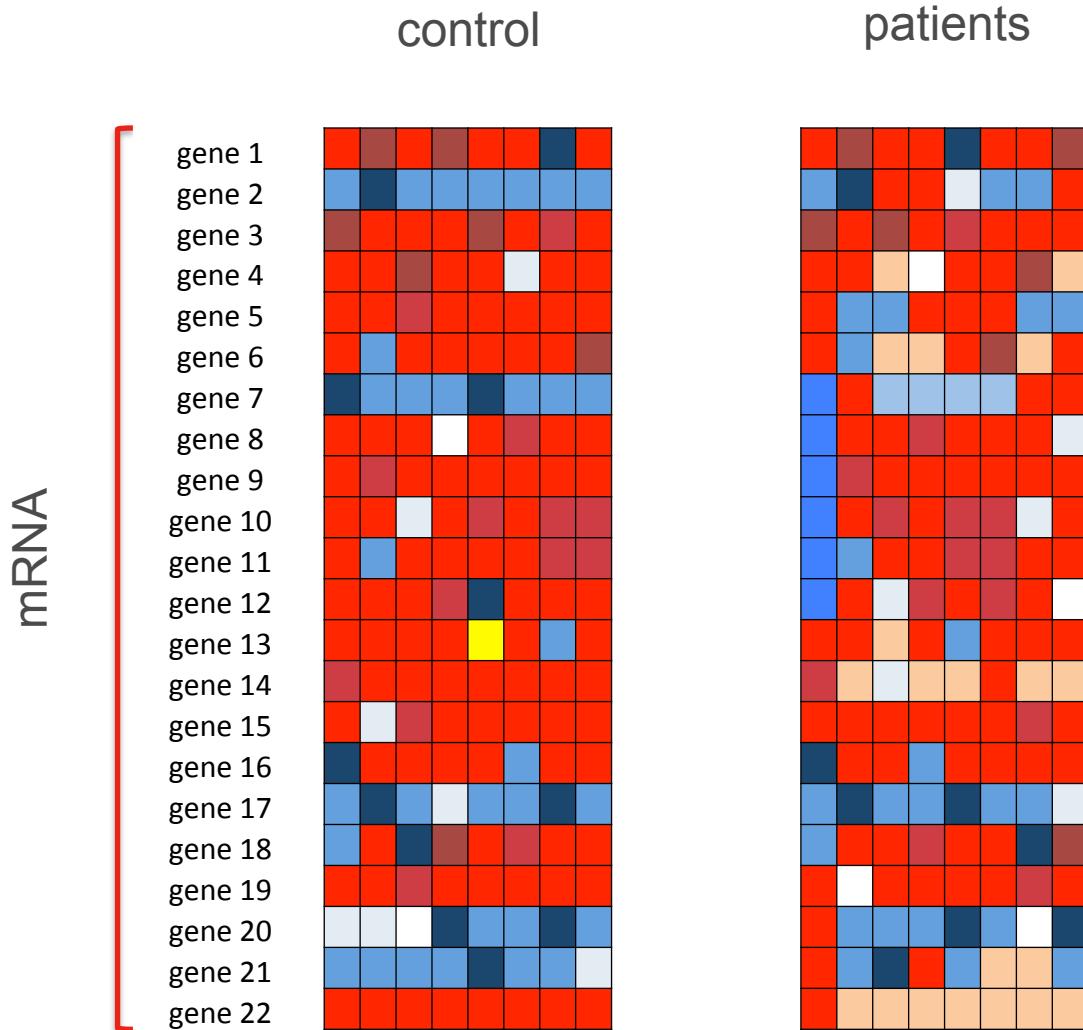
Differential expression

Comparing
two
biological
states



Differential expression

Comparing two biological states



Comparing two groups

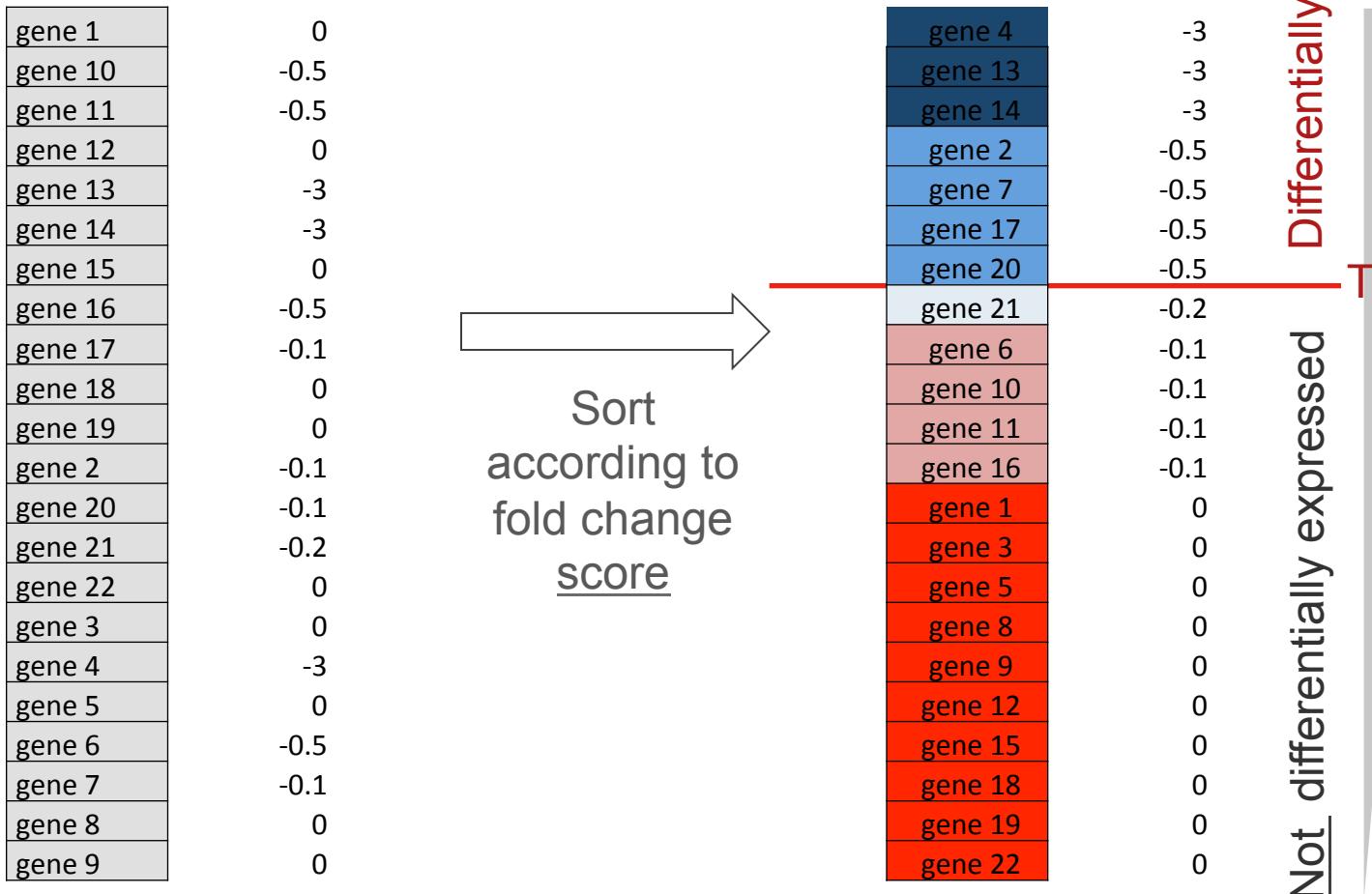
For each gene i , is there a difference in expression between the condition1 (healthy controls) and condition2 (patients)?

Fold change approach

$$\log(\pi_{i1}/\pi_{i2}) = \log(\pi_{i1}) - \log(\pi_{i2})$$

gene 1	0	gene 4	-3
gene 10	-0.5	gene 13	-3
gene 11	-0.5	gene 14	-3
gene 12	0	gene 2	-0.5
gene 13	-3	gene 7	-0.5
gene 14	-3	gene 17	-0.5
gene 15	0	gene 20	-0.5
gene 16	-0.5	gene 21	-0.2
gene 17	-0.1	gene 6	-0.1
gene 18	0	gene 10	-0.1
gene 19	0	gene 11	-0.1
gene 2	-0.1	gene 16	-0.1
gene 20	-0.1	gene 1	0
gene 21	-0.2	gene 3	0
gene 22	0	gene 5	0
gene 3	0	gene 8	0
gene 4	-3	gene 9	0
gene 5	0	gene 12	0
gene 6	-0.5	gene 15	0
gene 7	-0.1	gene 18	0
gene 8	0	gene 19	0
gene 9	0	gene 22	0

Sort
according to
fold change
score



Fisher exact test

Hypergeometric

Chi-square

Binomial

T-Test

...

T-test is a statistical test that compares
the mean of two states

T-test

For each gene i , is there a **significant difference** in mean expression between the condition1 (healthy controls) and condition2 (patients)?

Hypothesis testing

\mathcal{H}_0 : Healthy controls and patients have similar **gene i** expression

$$\mathcal{H}_0 : \pi_{i1} = \pi_{i2}$$

T-test

For each gene i , is there a **significant difference** in mean expression between the condition1 (healthy controls) and condition2 (patients)?

Hypothesis testing

\mathcal{H}_0 : Healthy controls and patients have similar **gene i** expression

$$\mathcal{H}_0 : \pi_{i1} = \pi_{i2}$$

\mathcal{H}_1 : Healthy controls and patients don't have a similar gene i expression

$$\mathcal{H}_1 : \pi_{i1} \neq \pi_{i2}$$

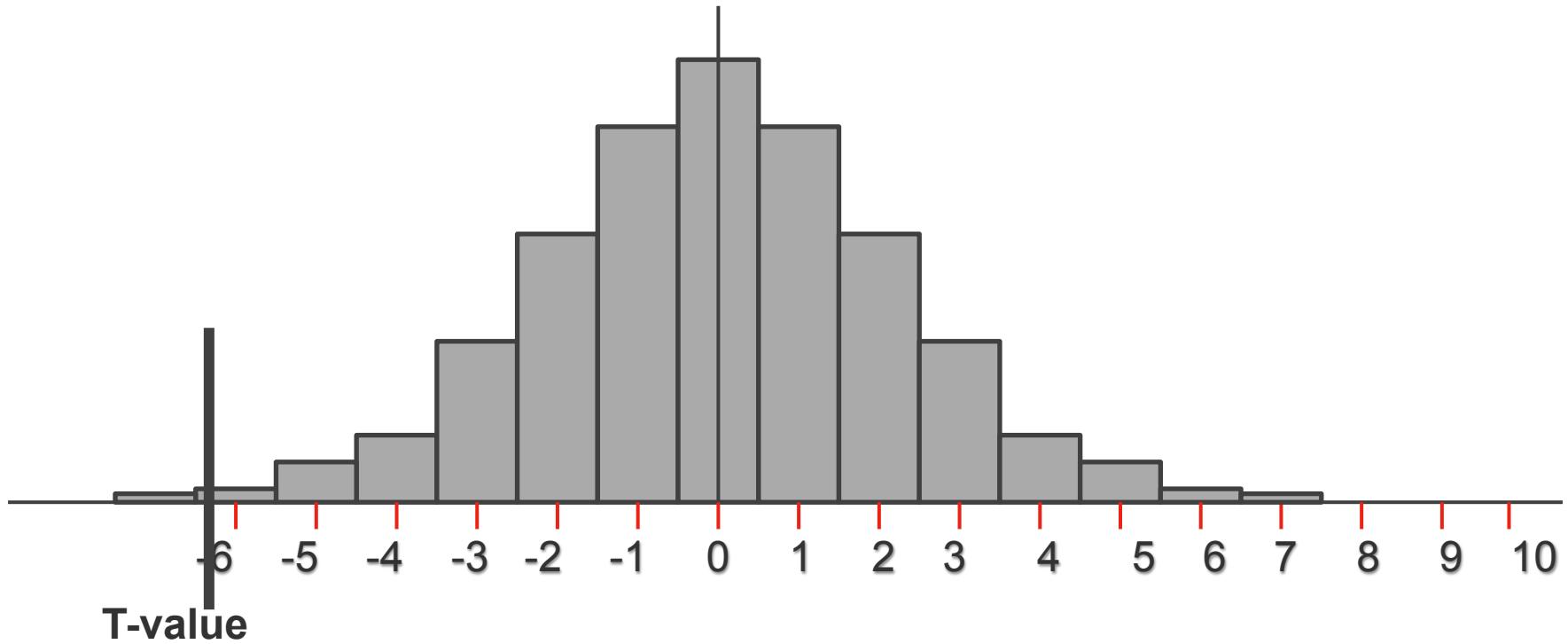
In R

```
>?t.test  
>t.test(g1,g2)
```

Welch Two Sample t-test

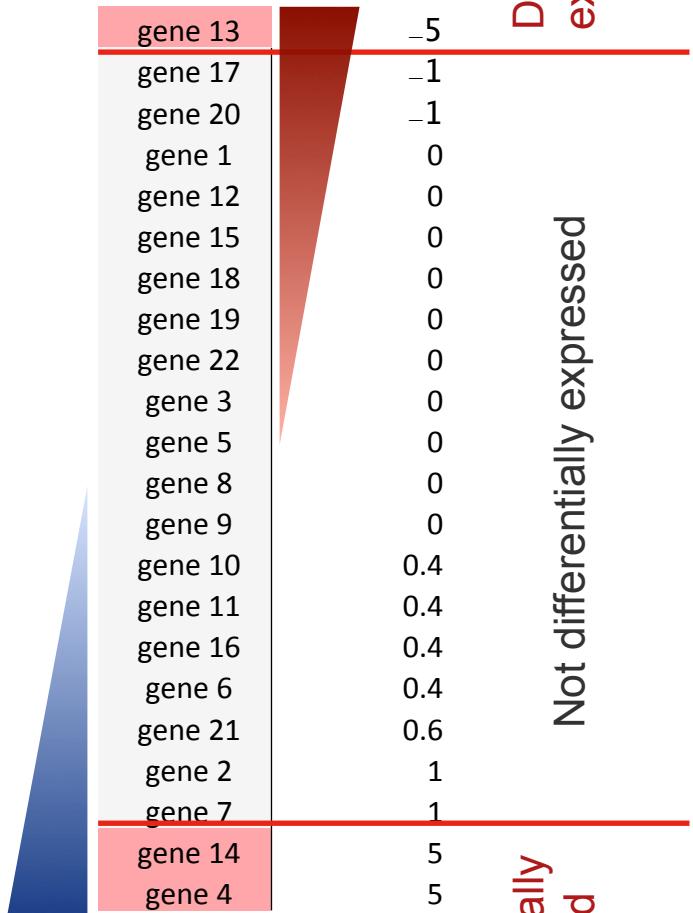
```
data: g1 and g2  
t = -6.7969, df = 7.1146, p-value = 0.0002361  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-117.84184 -57.15816  
sample estimates:  
mean of x mean of y  
12.9      100.4
```

T-distribution with group size =8



gene 1	0
gene 2	0.4
gene 3	0.4
gene 4	0
gene 5	-5
gene 6	5
gene 7	0
gene 8	0.4
gene 9	-1
gene 10	0
gene 11	0
gene 12	1
gene 13	-1
gene 14	0.6
gene 15	0
gene 16	0
gene 17	5
gene 18	0
gene 19	0.4
gene 20	1
gene 21	0
gene 22	0

Sort
according to
T score



Differentially
expressed

Not differentially expressed

Differentially
expressed

P-value

The p-value is the probability of getting a result that is as or more extreme than the observed result, assuming that the null hypothesis is true.

The p-value reflects the magnitude of the difference between the study groups

AND

the sample size

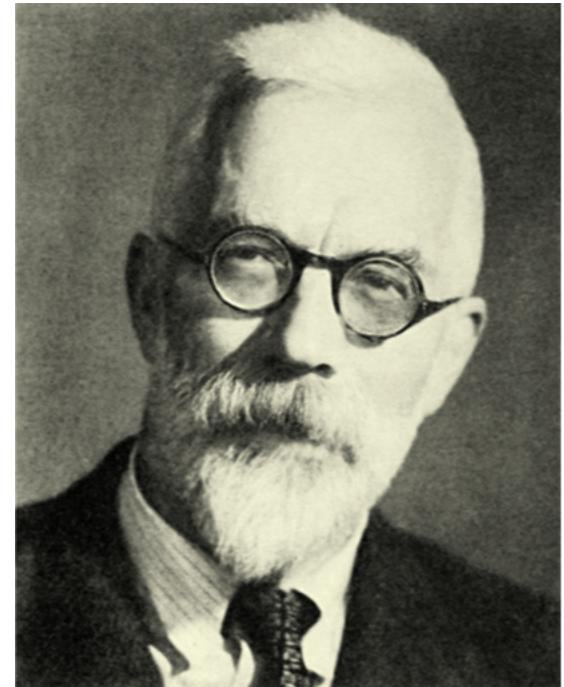
AND

the variability within each group

P-value and decision

By convention, if $p < 0.05$, then the association between the exposure and disease is considered to be “**statistically significant.**” (e.g. we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1))

Why 0.05?



Fisher

P-value and decision

What does $p < 0.05$ mean?

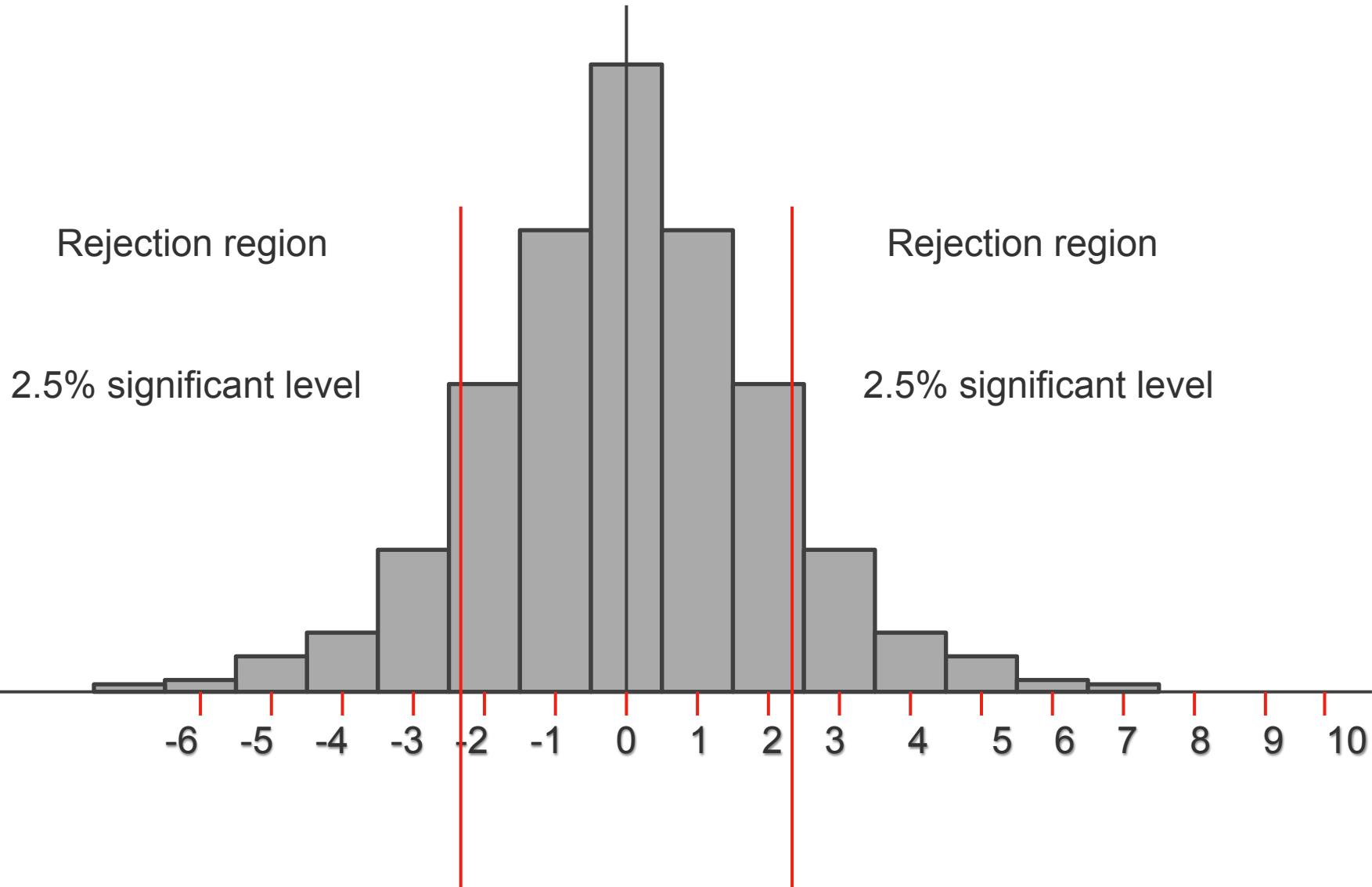
Indirectly, it means that we suspect that the magnitude of effect observed (e.g. odds ratio) is not due to chance alone

(in the absence of biased data collection or analysis)

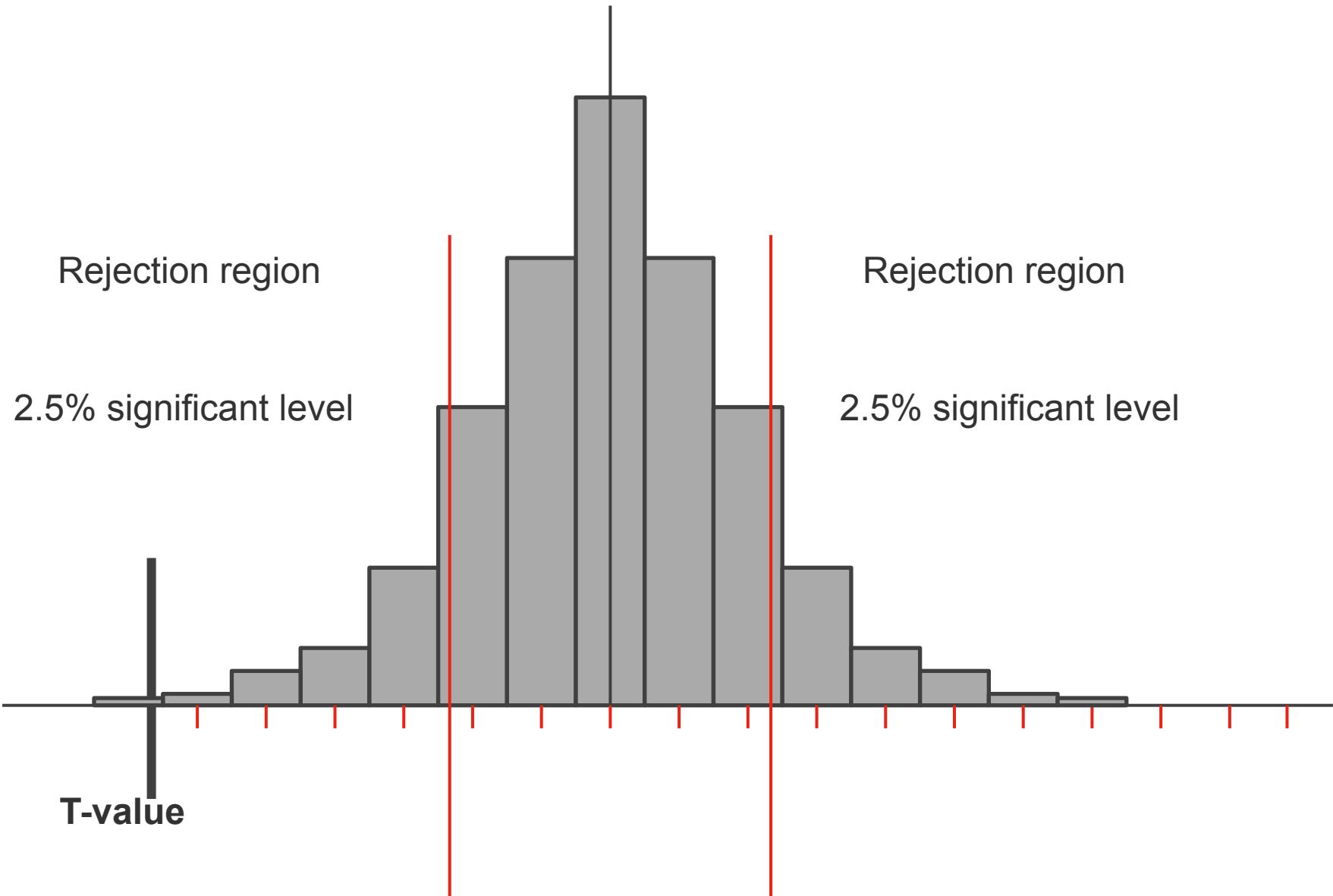
Directly, $p=0.05$ means that one test result out of twenty results would be expected to occur due to chance (random error) alone

P-value and decision

T-distribution with group size =8



P-value and decision

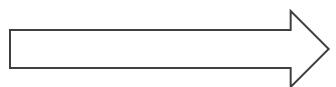


p-value =0.000001 & p-value =0.049

0.01 and 0.1
are also possible threshold

gene 1	1
gene 2	0.01
gene 3	1
gene 4	0.0001
gene 5	1
gene 6	0.6
gene 7	0.01
gene 8	1
gene 9	1
gene 10	0.6
gene 11	0.6
gene 12	1
gene 13	0.0001
gene 14	0.0001
gene 15	1
gene 16	0.6
gene 17	0.01
gene 18	1
gene 19	1
gene 20	0.01
gene 21	0.4
gene 22	1

Sort
according to
p-value



	T-score	p-value
gene 4	5	0.0001
gene 13	5	0.0001
gene 14	-5	0.0001
gene 2	5	0.01
gene 7	1	0.01
gene 17	1	0.01
gene 20	-1	0.01
gene 21	-1	0.4
gene 6	0.6	0.6
gene 10	0.4	0.6
gene 11	0.4	0.6
gene 16	0.4	0.6
gene 1	0.4	1
gene 3	0	1
gene 5	0	1
gene 8	0	1
gene 9	0	1
gene 12	0	1
gene 15	0	1
gene 18	0	1
gene 19	0	1
gene 22	0	1

Differentially
expressed

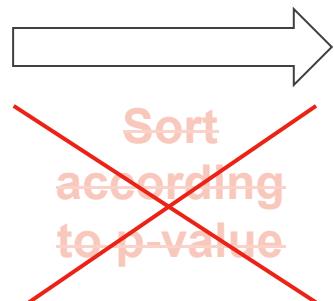
P-value and decision

Truth \ Decision	H_0 not rejected (negative)	H_0 Rejected (positive)
Truth		
H_0 is true (no signal in the data)	 specificity True negative TN	 Type I error False Positive α
H_0 is false (there is something to find)	 Type II error False Negative β	 Power $1 - \beta$;  sensitivity True Positive TP

P-value and decision

gene 1	1
gene 2	0.01
gene 3	1
gene 4	0.0001
gene 5	1
gene 6	0.6
gene 7	0.01
gene 8	1
gene 9	1
gene 10	0.6
gene 11	0.6
gene 12	1
gene 13	0.0001
gene 14	0.0001
gene 15	1
gene 16	0.6
gene 17	0.01
gene 18	1
gene 19	1
gene 20	0.01
gene 21	0.4
gene 22	1

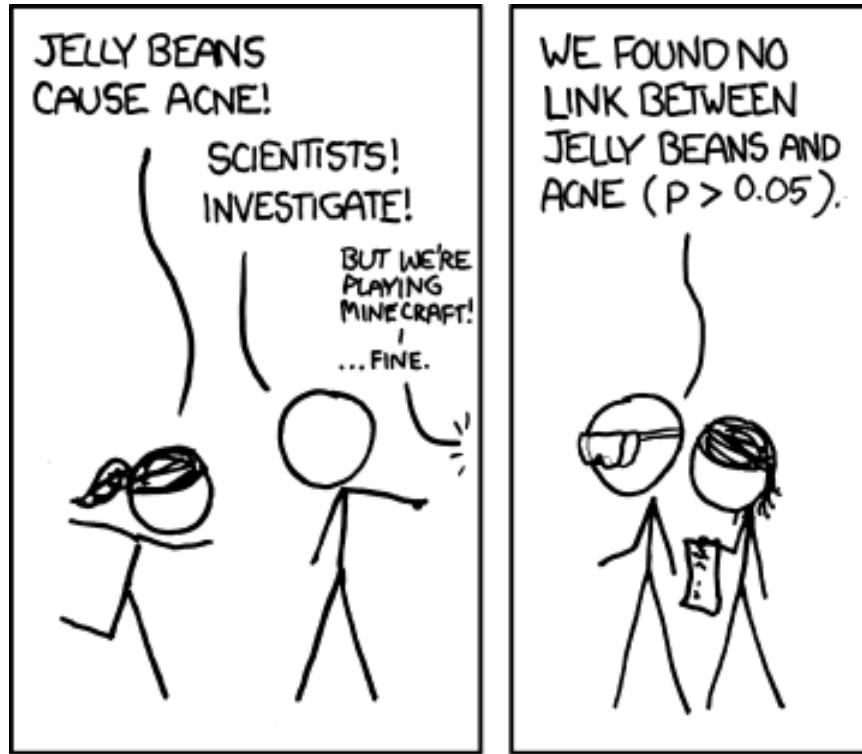
Sort
according to
adj. p-value



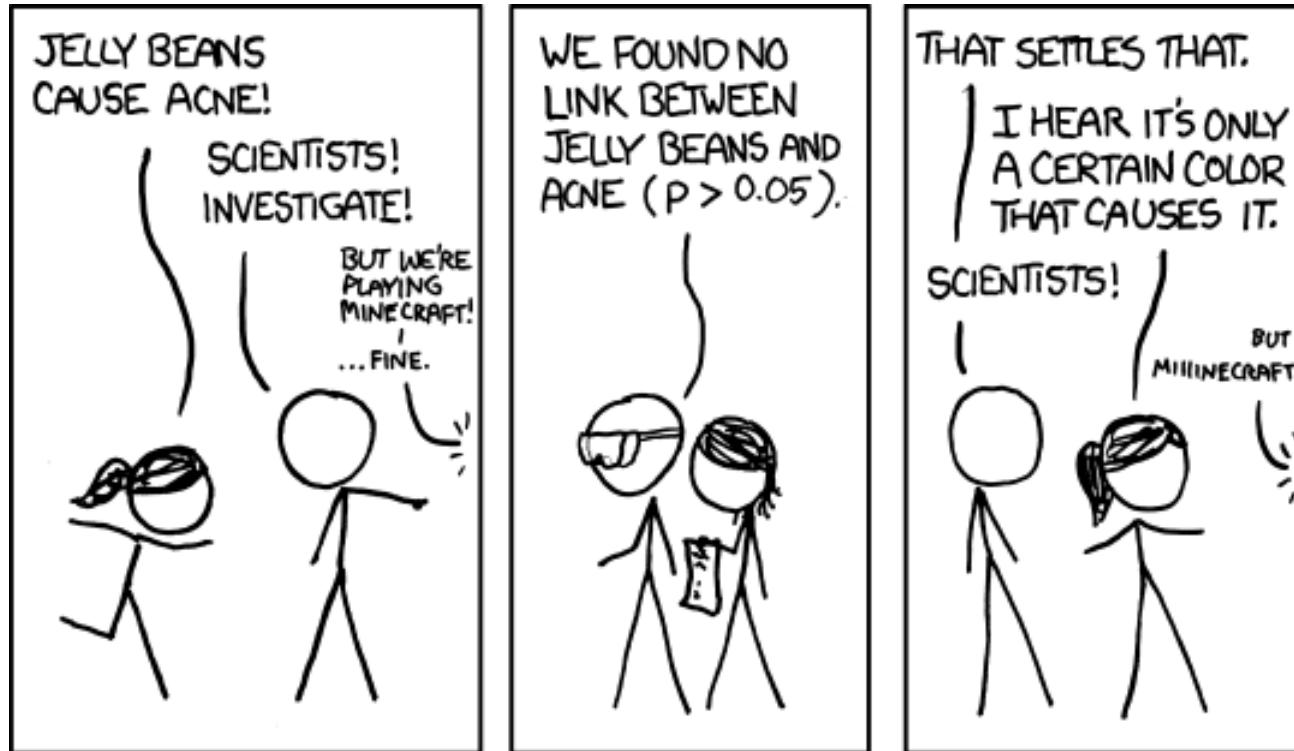
	T-score	p-value	Adj. p-value
gene 4	5	0.0001	0.0022
gene 13	5	0.0001	0.0022
gene 14	-5	0.0001	0.0022
gene 2	5	0.01	0.19
gene 7	1	0.01	0.19
gene 17	1	0.01	0.19
gene 20	-1	0.01	0.19
gene 21	-1	0.4	1
gene 6	0.6	0.6	1
gene 10	0.4	0.6	1
gene 11	0.4	0.6	1
gene 16	0.4	0.6	1
gene 1	0.4	1	1
gene 3	0	1	1
gene 5	0	1	1
gene 8	0	1	1
gene 9	0	1	1
gene 12	0	1	1
gene 15	0	1	1
gene 18	0	1	1
gene 19	0	1	1
gene 22	0	1	1

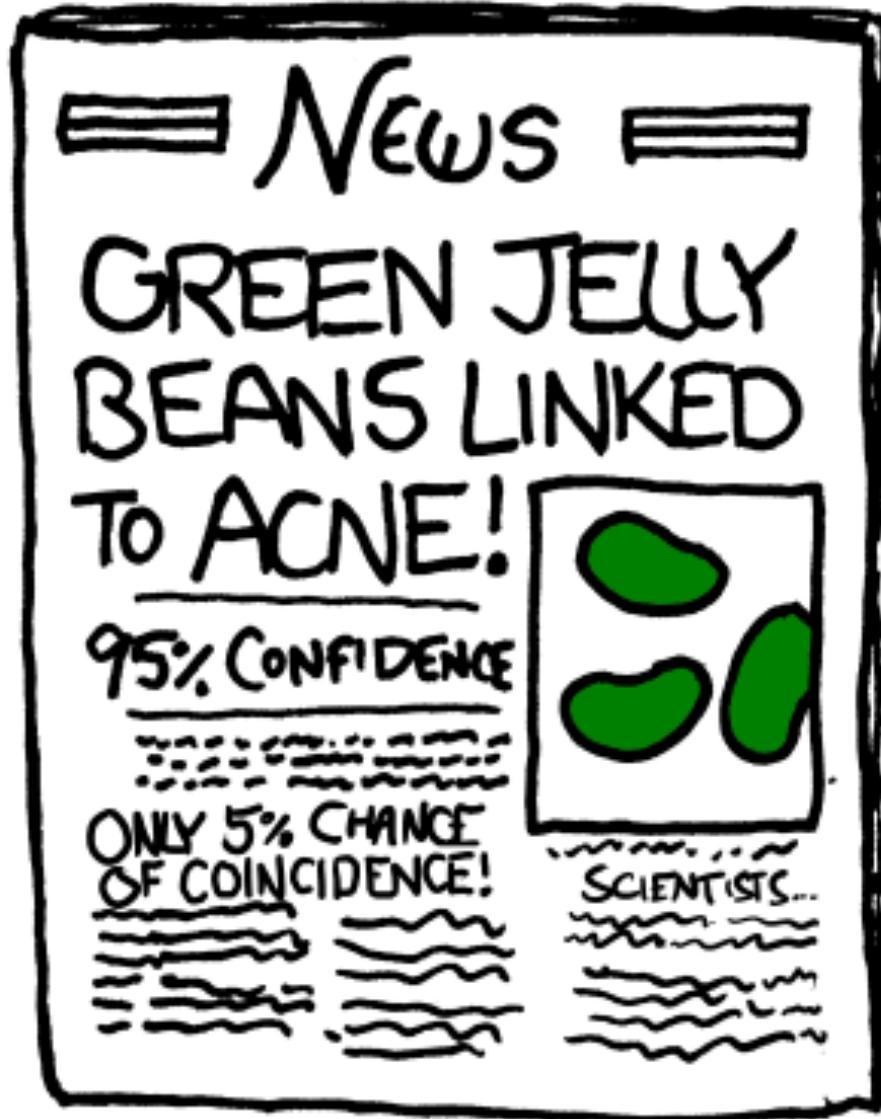
Differentially
expressed

Adjusting p-value, why?



Adjusting p-value, why?



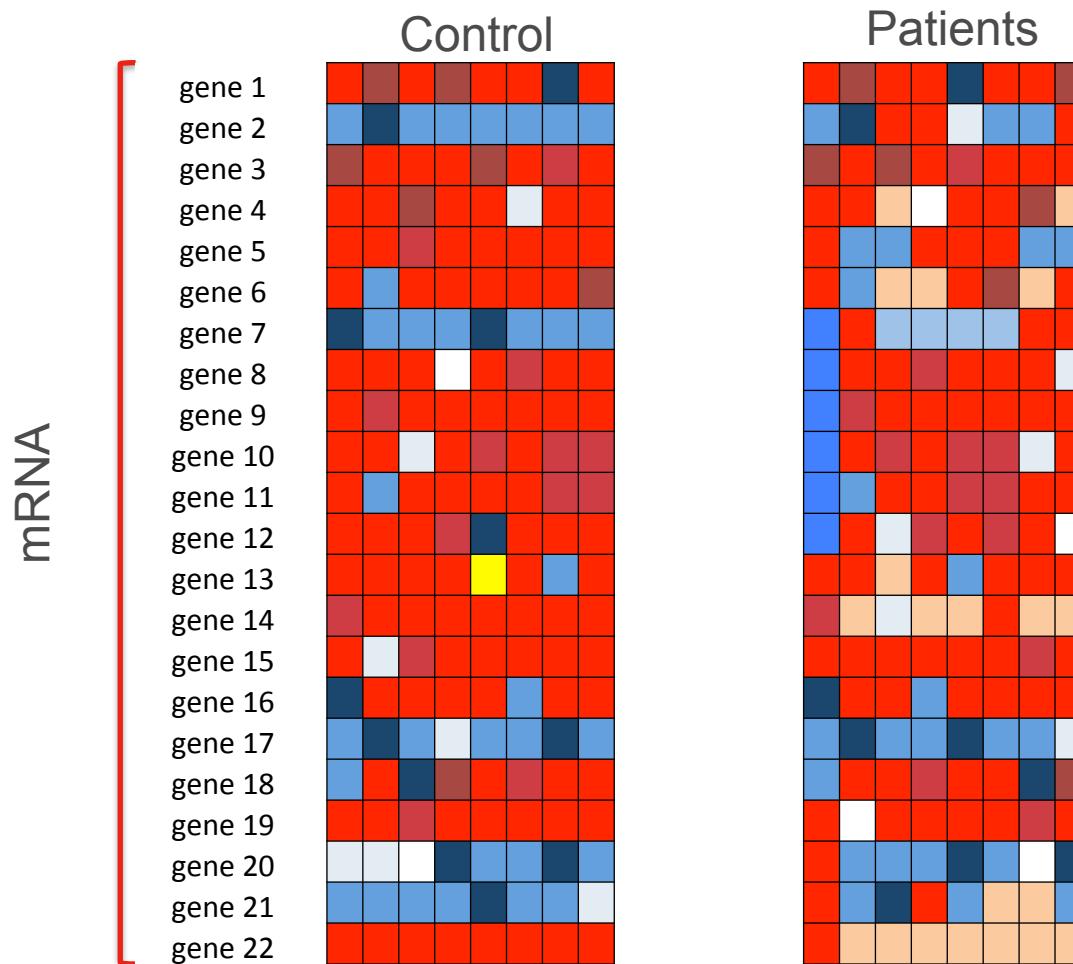


So, uh, we did the green study again and got no link. It was probably a—
'RESEARCH CONFLICTED ON GREEN JELLY BEAN/ACNE LINK;
MORE STUDY RECOMMENDED!'

Multiple testing correction

Experiment

- Imagine if we perform a test on each of the **10'000 genes**
- None of the genes is differentially expressed



Significance level
 $\alpha = 5\%$

Consequences:
we expect to find
around **500 p-values**
below 0.05!

adjust the p-values to take
the number of tests into
account

Multiple testing correction

FWER

Control the probability of obtaining
any false positives

k is the rank

Bonferroni

α =significance level (ex: 0.05)

Change α for each test

$$\begin{aligned}\alpha' &= \alpha/k, \\ p_k &= 1 - (1 - \alpha')^k, \\ p_{\text{bonferroni}} &= \min(p_k, 1).\end{aligned}$$

The probability of
getting **at least**
one significant
p-value

⇒The probability of obtaining any
false positive is controled.

⇒Very stringent, we may miss
many true positives

Multiple testing correction

FWER

Control the probability of obtaining any false positives

k is the rank

Bonferroni

α = significance level (ex: 0.05)

Change α for each test

$$\begin{aligned}\alpha' &= \alpha/k, \\ p_k &= 1 - (1 - \alpha')^k, \\ p_{\text{bonferroni}} &= \min(p_k, 1).\end{aligned}$$

The probability of getting **at least one significant p-value**

p-value	k	Probability (p_k)
	1	0.05
	5	0.23
	10	0.4
	20	0.64
	50	0.92
	100	0.99
	500	1

⇒ The probability of obtaining any false positive is controled.

⇒ Very stringent, we may miss many true positives

Multiple testing correction

FWER

Control the probability of obtaining any false positives

Bonferroni

α =significance level (ex: 0.05)

Change α for each test

$$\alpha' = \alpha/k,$$

$$p_k = 1 - (1 - \alpha')^k,$$

$$p_{\text{bonferroni}} = \min(p_k, 1).$$

FDR

Controls the expected number of false discoveries

k is the rank

Benjamini-Hochberg

Order the p-values from the smallest to the largest

$$\text{q-value}_{(1)} = \text{p-value}_{(1)} \cdot n/(n-1)$$

$$\text{q-value}_{(2)} = \text{p-value}_{(2)} \cdot n/(n-2)$$

$$\text{q-value}_{(k)} = \text{p-value}_{(k)} \cdot n/(n-k)$$

Where

n is number of genes

Correct less and less as the p-values get larger

⇒The probability of obtaining any false positive is controled.

⇒Very stringent, we may miss many true positives

⇒Less stringent than Bonferroni.

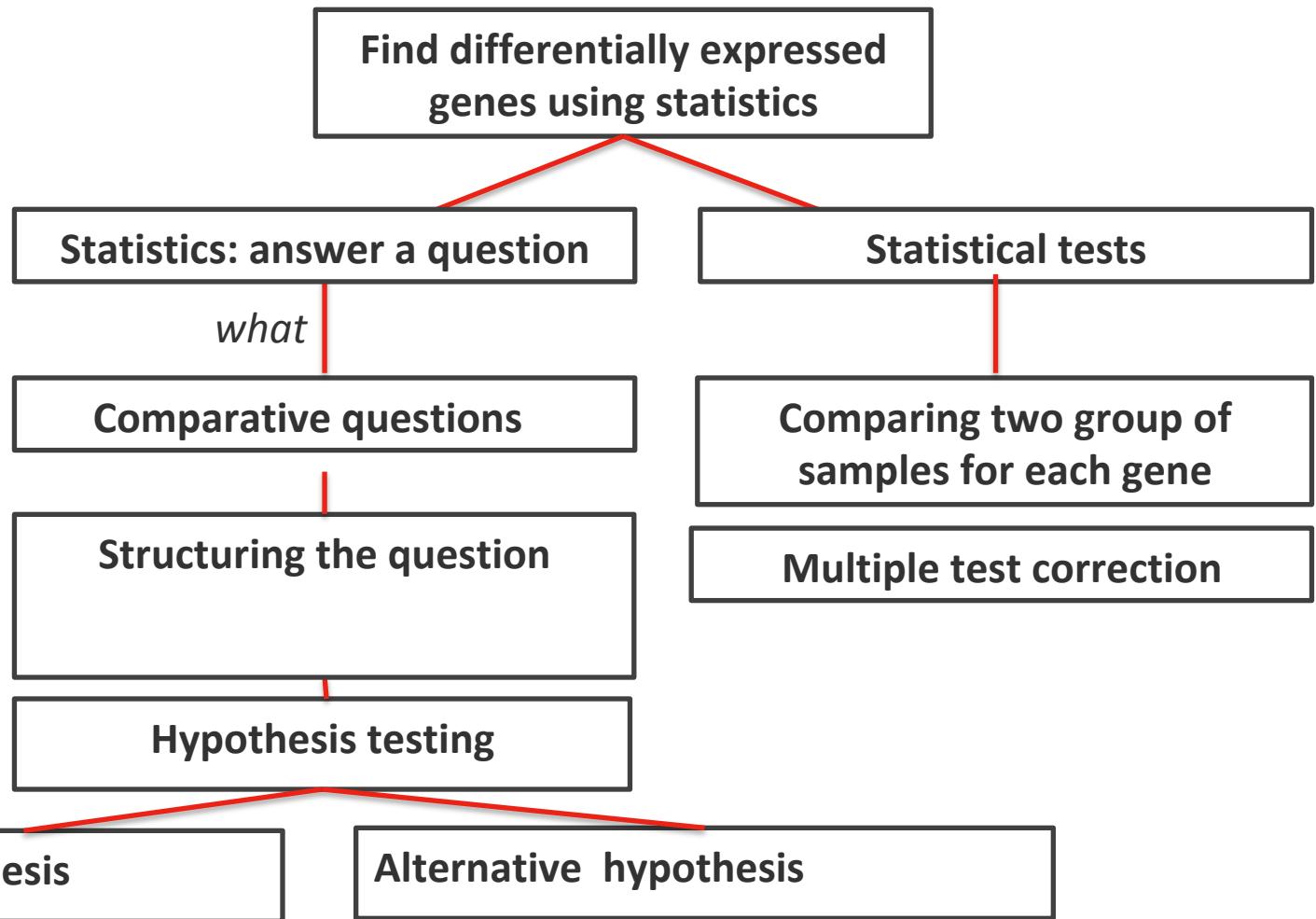
In R:

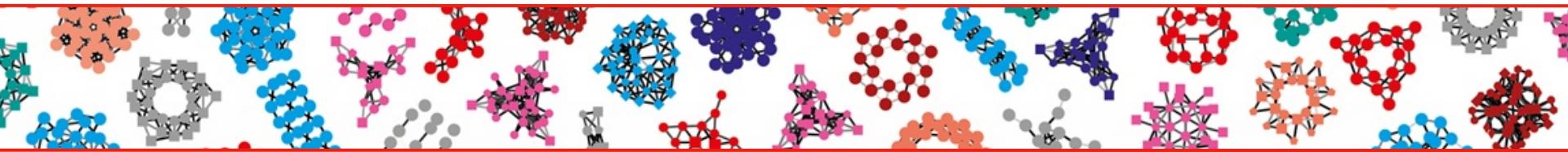
```
>?p.adjust  
>p.adjust.methods
```

Example

```
>p_bonf <- p.adjust(sort(rawp),method="bonf")  
>p_bh   <- p.adjust(sort(rawp),method="BH")  
>p_holm <- p.adjust(sort(rawp),method="holm")  
>p_holm <- p.adjust(sort(rawp),method=p.adjust.methods)
```

Wrap up





EXERCICE 1: Differential expression

Analysing microarray expression of rat Affymetrix probes

Download *rat_KD.txt* from course web-page.

1. Is probe *1398751_at* differentially expressed considering a significance value of 0.01?
2. How many probes are differentially expressed considering a significance value of 0.01?

In R: solution

Get data

```
>rat <- read.table("rat_KD.txt", sep = "\t", header = T, stringsAsFactors=FALSE)
>dimnames(rat)[[1]] <- rat[,1]
```

Question1

```
>rowNb<-which(rat[,1] == "1398751_at")
>v1<- t.test(rat[rowNb,2:7], rat[rowNb,8:12])
```

Question2

```
>ttestRat <- function(df, grp1, grp2) {
  x = df[grp1]
  y = df[grp2]
  x = as.numeric(x)
  y = as.numeric(y)
  results = t.test(x, y)
  results$p.value }

>rawp <- apply(rat, 1, ttestRat, grp1 = c(2:7), grp2 = c(8:12))
>p_holm <- p.adjust(sort(rawp),method="BH")
>hist(p_holm)
```

Overview

Count matrix

Differential expression

Enrichment

SEA

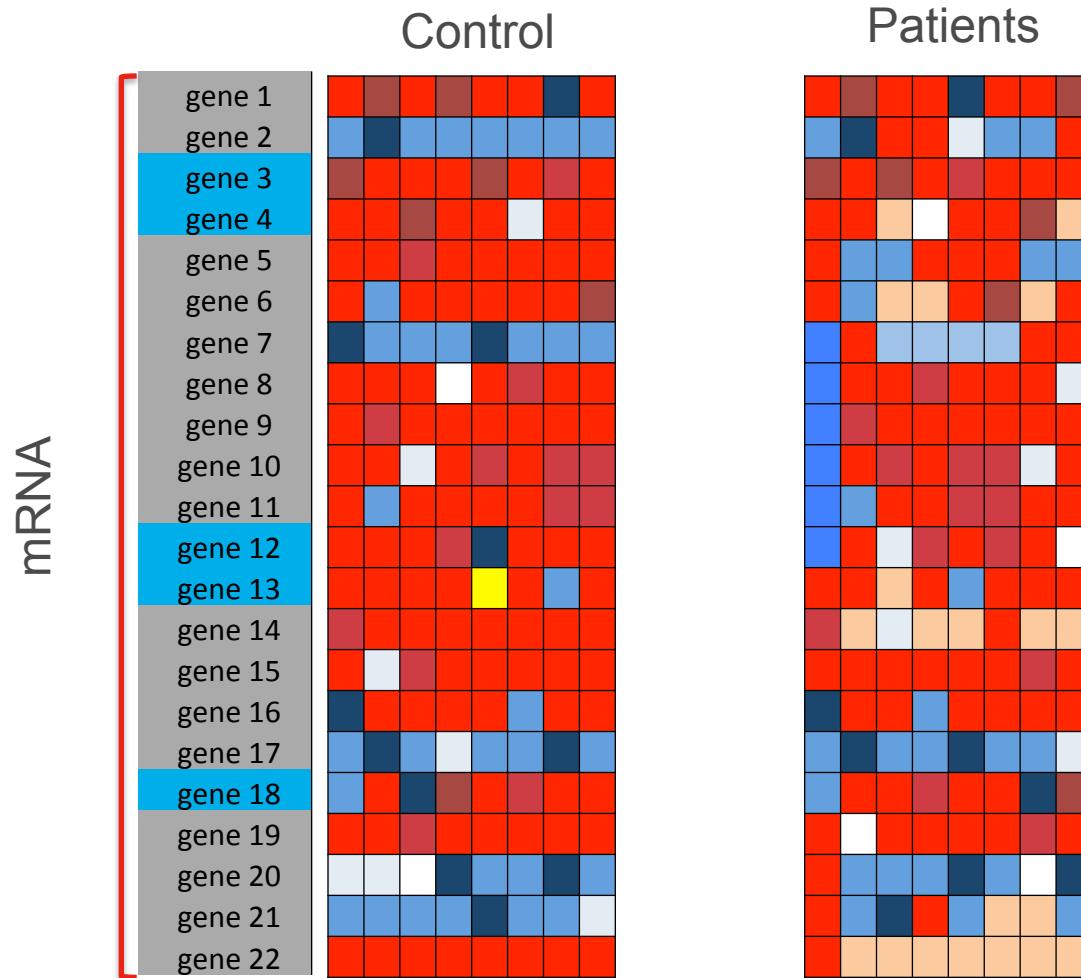
GSEA

MEA

Knowledge



Are genes belonging to blue set differentially expressed?



Are genes belonging to blue set differentially expressed?

gene 1	0
gene 2	0.4
gene 3	0.4
gene 4	0
gene 5	-5
gene 6	5
gene 7	0
gene 8	0.4
gene 9	-1
gene 10	0
gene 11	0
gene 12	1
gene 13	-1
gene 14	0.6
gene 15	0
gene 16	0
gene 17	5
gene 18	0
gene 19	0.4
gene 20	1
gene 21	0
gene 22	0

0
0.4
0.4
0
-5
5
0
0.4
-1
0
0
1
-1
0.6
0
0
0
5
0
0.4
1
0
0
0

Sort
according to
T score

gene 5	-5
gene 9	-1
gene 13	-1
gene 1	0
gene 4	0
gene 7	0
gene 10	0
gene 11	0
gene 15	0
gene 16	0
gene 18	0
gene 21	0
gene 22	0
gene 2	0.4
gene 3	0.4
gene 8	0.4
gene 19	0.4
gene 14	0.6
gene 12	1
gene 20	1
gene 6	5
gene 17	5

-5
-1
-1
0
0
0
0
0
0
0
0
0
0
0
0
0
0.4
0.4
0.4
0.4
0.6
1
1
5
5



Fisher exact test

<i>2x2 count table</i>	Differentially expressed	Not Differentially expressed	total
blue	2	3	5
Not blue	5	12	17
total	7	15	22

Fisher exact test

\mathcal{H}_0 : The proportion of blue genes differentially expressed set is the same as the proportion of blue genes in non-differentially expressed

$$\mathcal{H}_0: \frac{\pi_{b1}}{\pi_D} = \frac{\pi_{b2}}{\pi_{ND}}$$

\mathcal{H}_1 : The proportion of blue genes differentially expressed set is not the same as the proportion of blue genes in non-differentially expressed

$$\mathcal{H}_1: \frac{\pi_{b1}}{\pi_D} \neq \frac{\pi_{b2}}{\pi_{ND}}$$

In R

```
>dat2 <- matrix(c(2,3,5,12), ncol=2)
>dat2
>fisher.test(dat2)
```

Fisher's Exact Test for Count Data

```
data: dat2
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1012333 18.7696686
sample estimates:
odds ratio
 1.56456
```

gene 1	0
gene 2	0.4
gene 3	0.4
gene 4	0
gene 5	-5
gene 6	5
gene 7	0
gene 8	0.4
gene 9	-1
gene 10	0
gene 11	0
gene 12	1
gene 13	-1
gene 14	0.6
gene 15	0
gene 16	0
gene 17	5
gene 18	0
gene 19	0.4
gene 20	1
gene 21	0
gene 22	0

Which gene class
(blue, pink, purple, green)
is differentially expressed?

Enrichment analysis of
several phenotypes/classes:

multiple testing!

Fisher exact test
is a Threshold-based test

Are genes belonging to blue set differentially expressed?

gene 1	0
gene 2	0.4
gene 3	0.4
gene 4	0
gene 5	-5
gene 6	5
gene 7	0
gene 8	0.4
gene 9	-1
gene 10	0
gene 11	0
gene 12	1
gene 13	-1
gene 14	0.6
gene 15	0
gene 16	0
gene 17	5
gene 18	0
gene 19	0.4
gene 20	1
gene 21	0
gene 22	0

0
0.4
0.4
0
-5
5
0
0.4
-1
0
0
1
-1
0.6
0
0
0
5
0
0.4
1
0
0
0

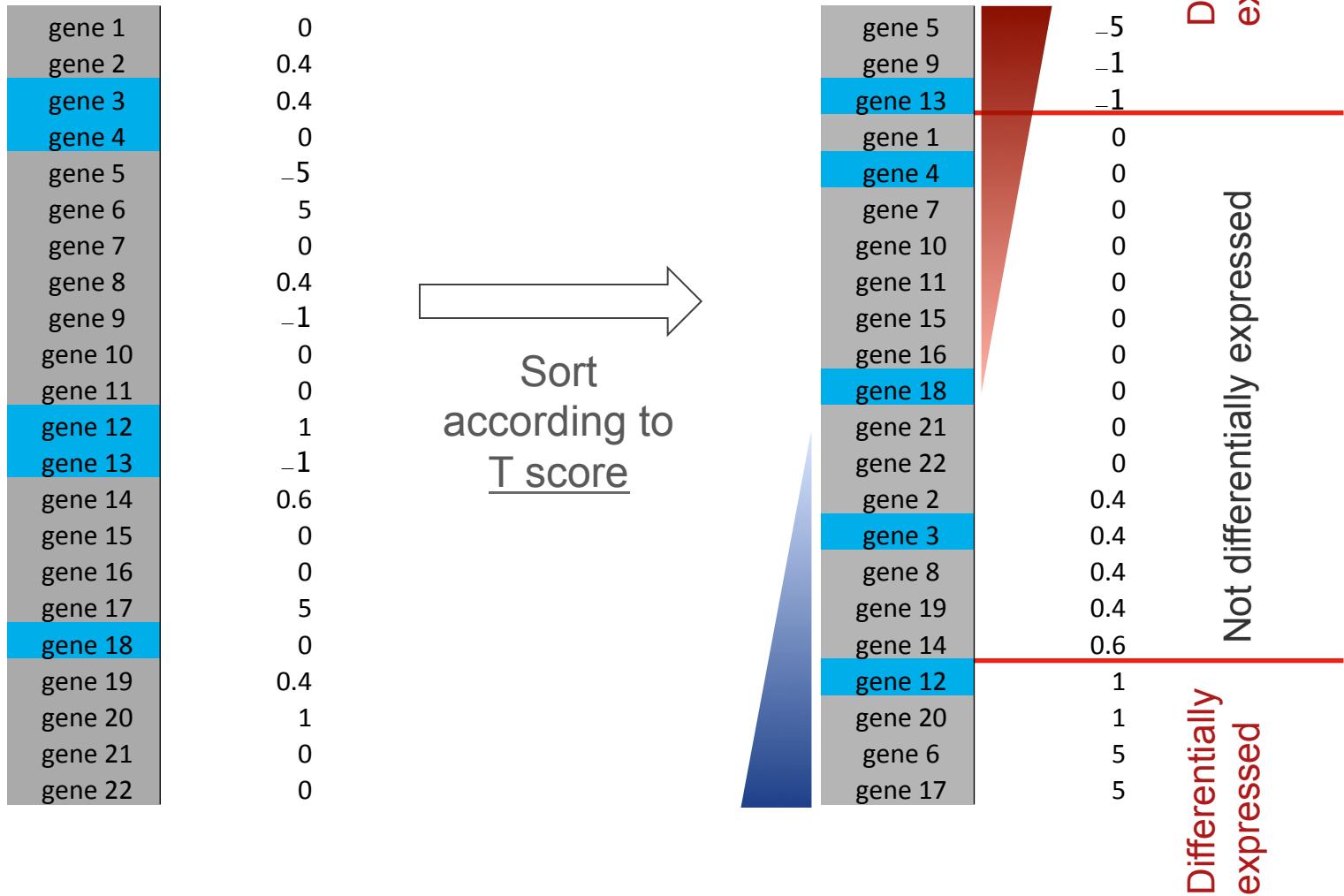
Sort
according to
T score

gene 5	-5
gene 9	-1
gene 13	-1
gene 1	0
gene 4	0
gene 7	0
gene 10	0
gene 11	0
gene 15	0
gene 16	0
gene 18	0
gene 21	0
gene 22	0
gene 2	0.4
gene 3	0.4
gene 8	0.4
gene 19	0.4
gene 14	0.6
gene 12	1
gene 20	1
gene 6	5
gene 17	5

-5
-1
-1
0
0
0
0
0
0
0
0
0
0
0
0
0
0.4
0.4
0.4
0.4
0.6
1
1
5
5



Are genes belonging to blue set differentially expressed?



Threshold-**free**?

Kolmogorov-Smirnov-like

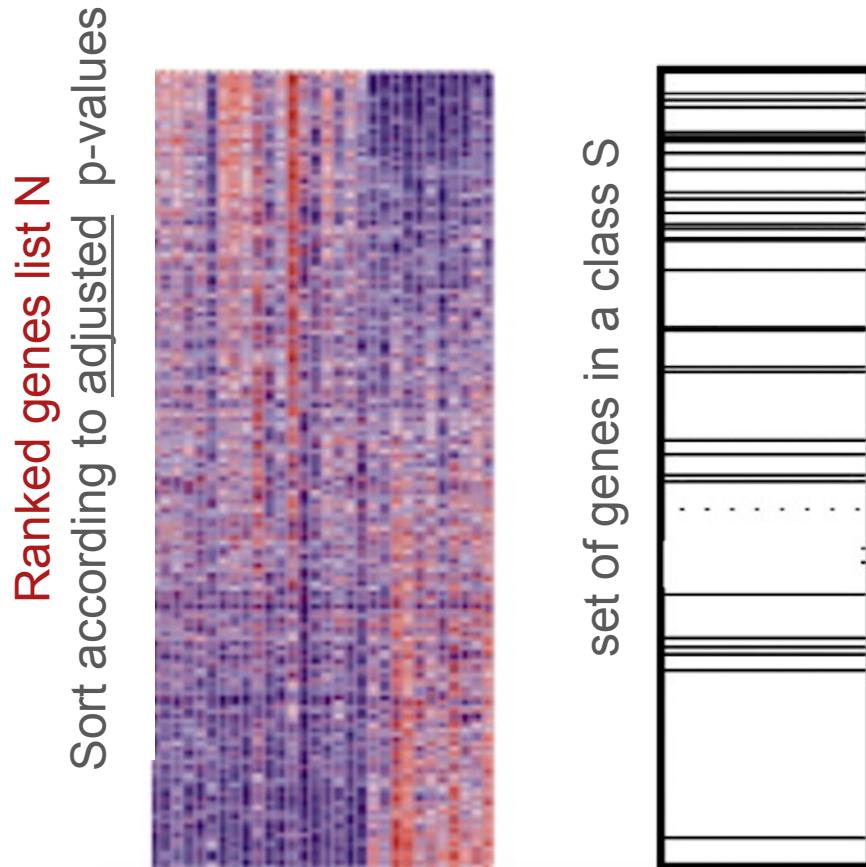
Permutation

Z-score

...

Gene Set Enrichment Analysis

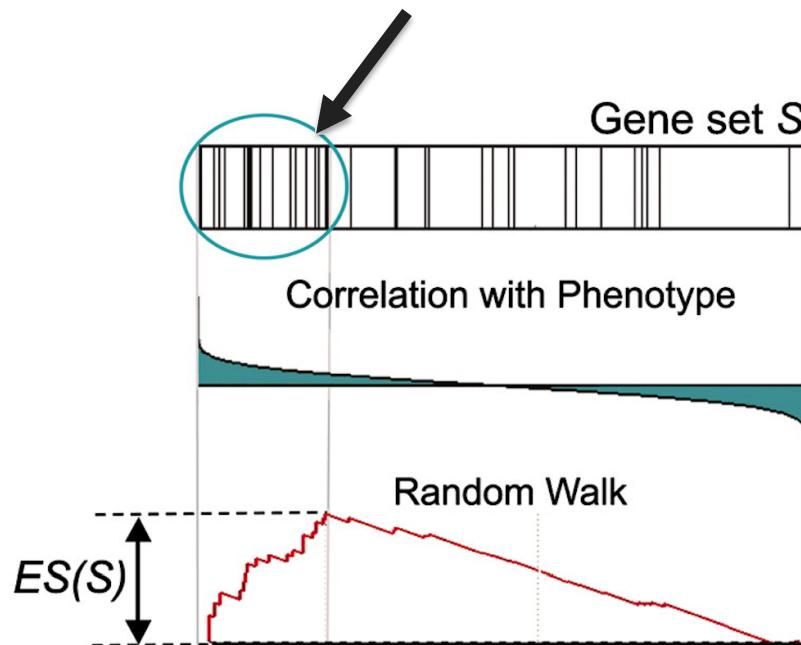
Given



Three steps:
Evaluate, Estimate, Adjust

Evaluate

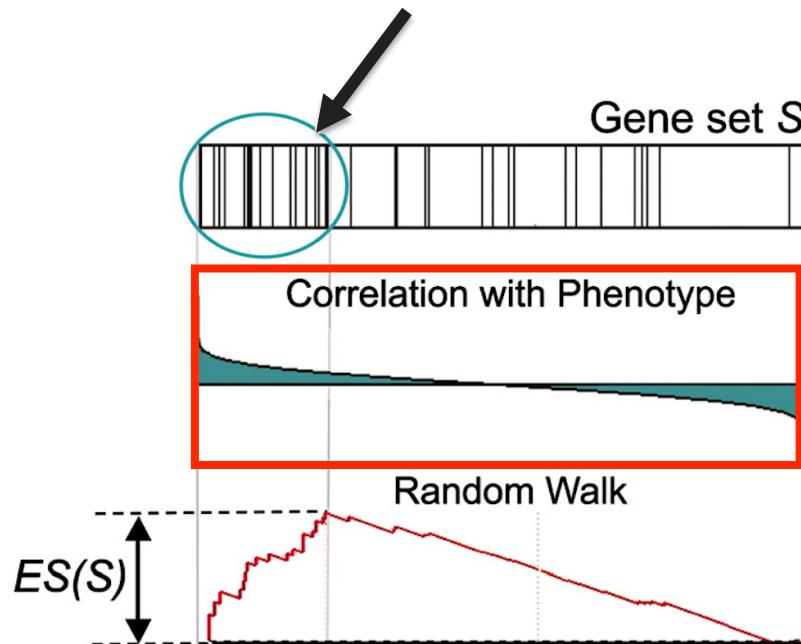
the enrichment score using a
Kolmogorov-Smirnov-like score



ES reflects the degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L

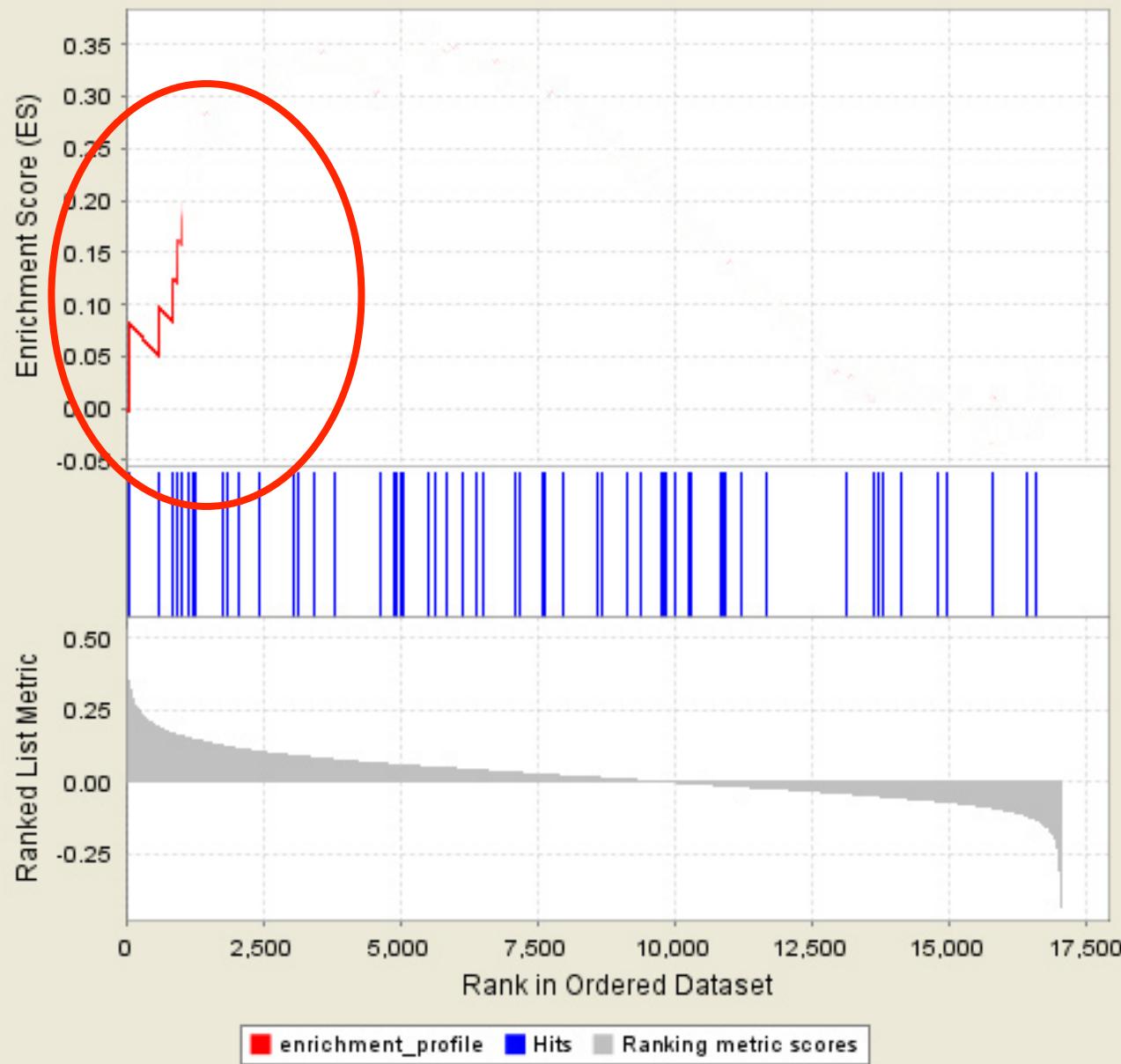
Evaluate

the enrichment score using a
Kolmogorov-Smirnov-like score

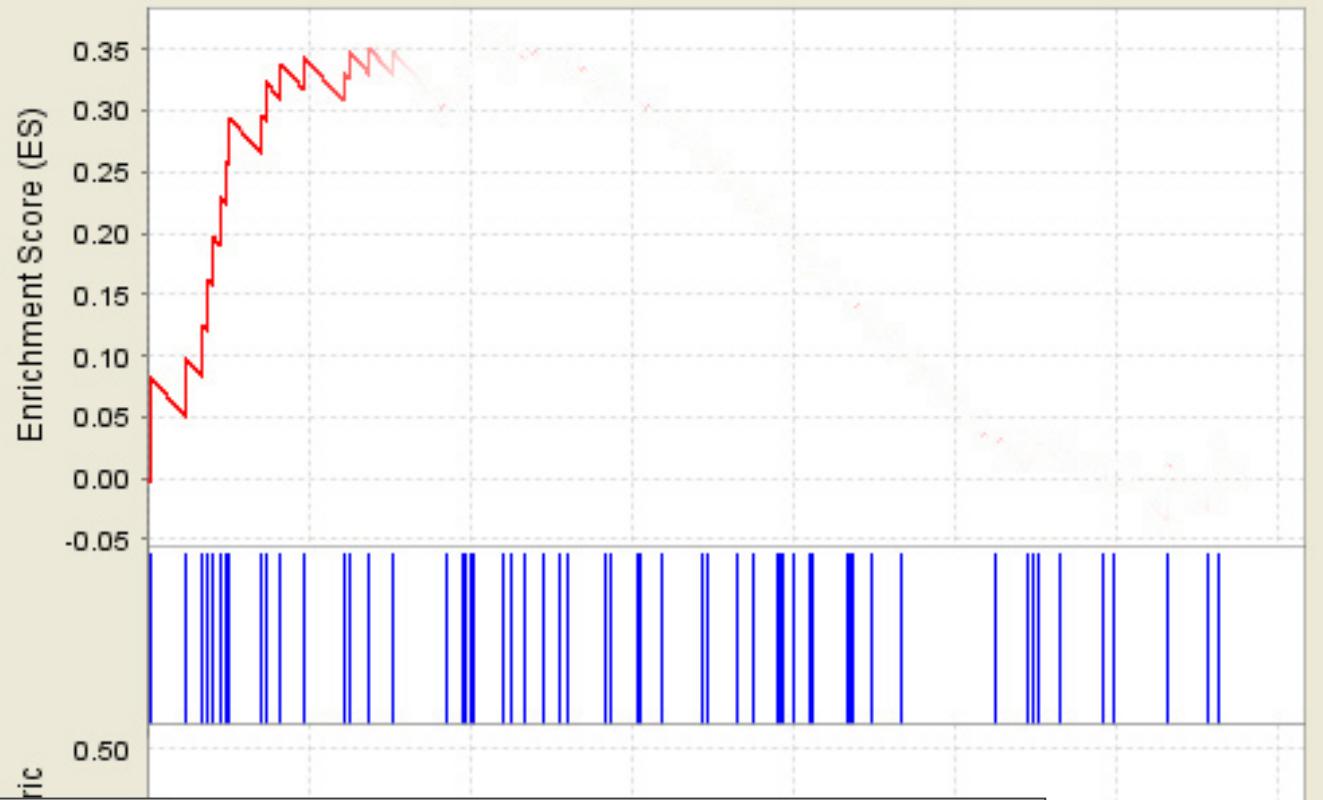


ES reflects the degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L

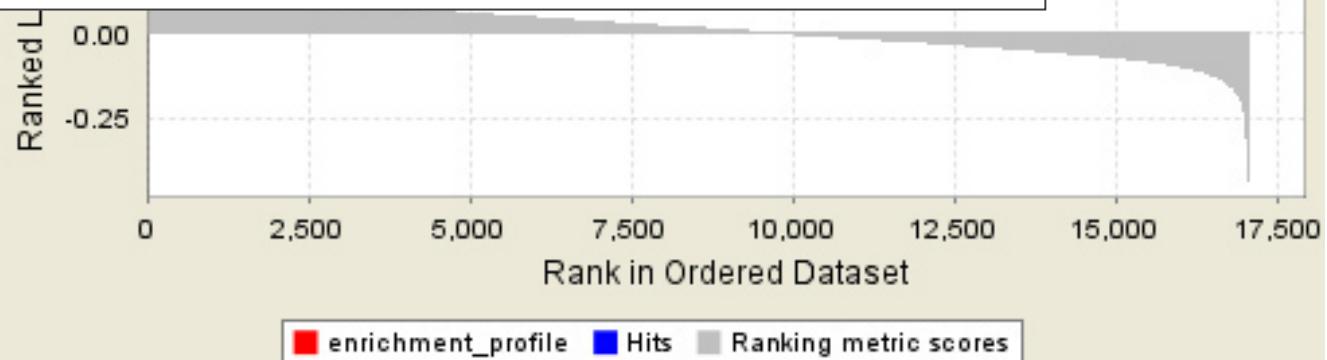
GSEA_Results



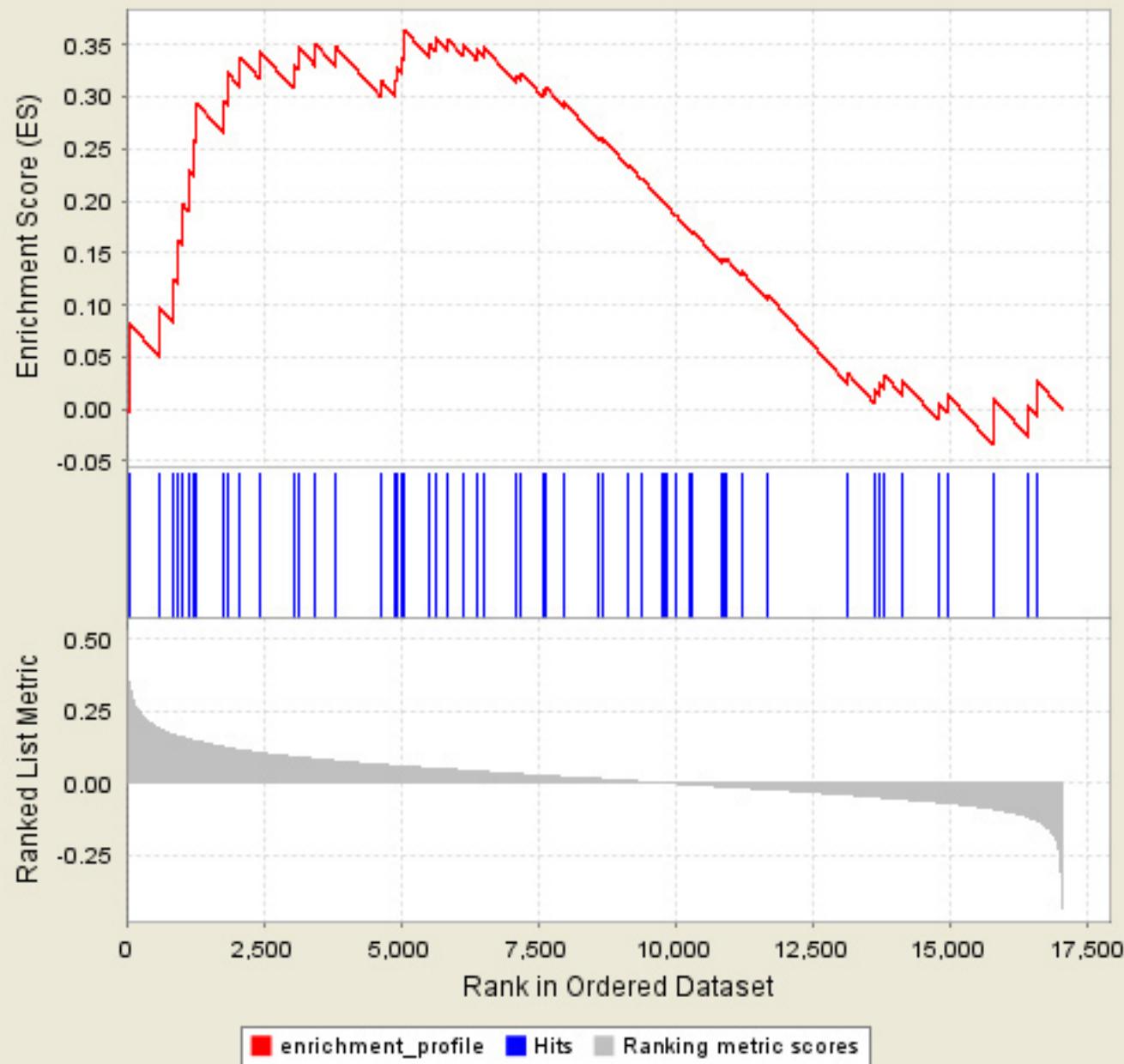
GSEA_Results



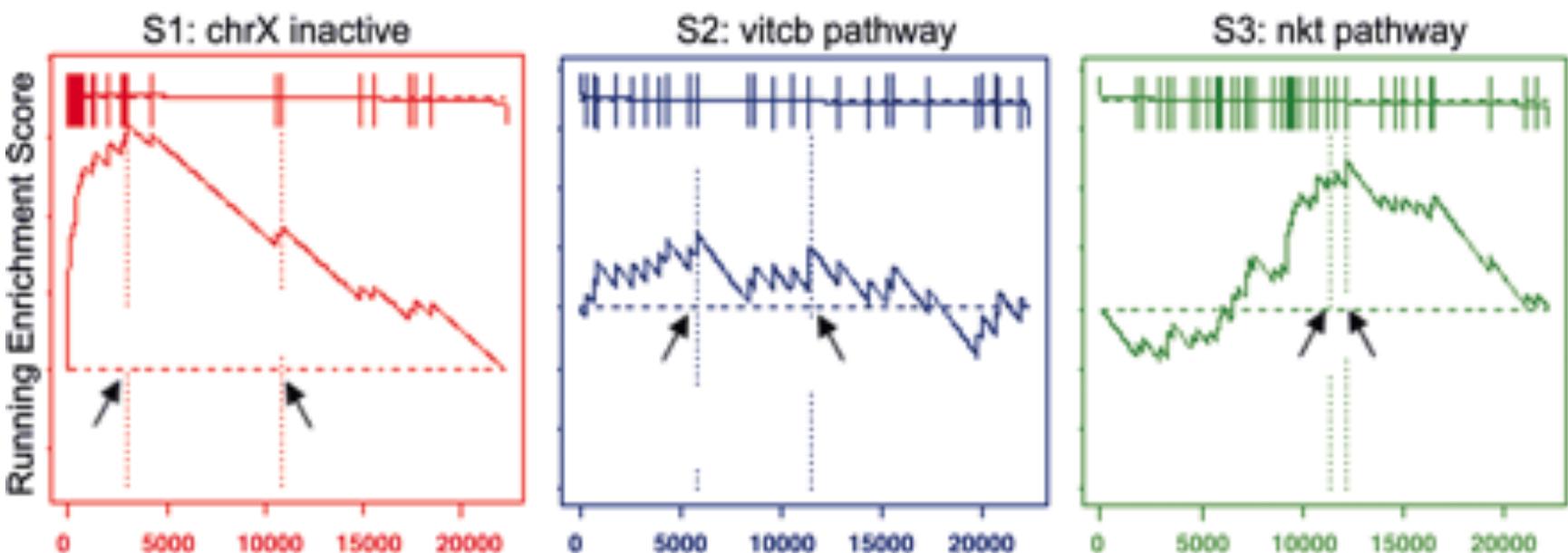
The more S genes is found, the higher the ES



GSEA_Results

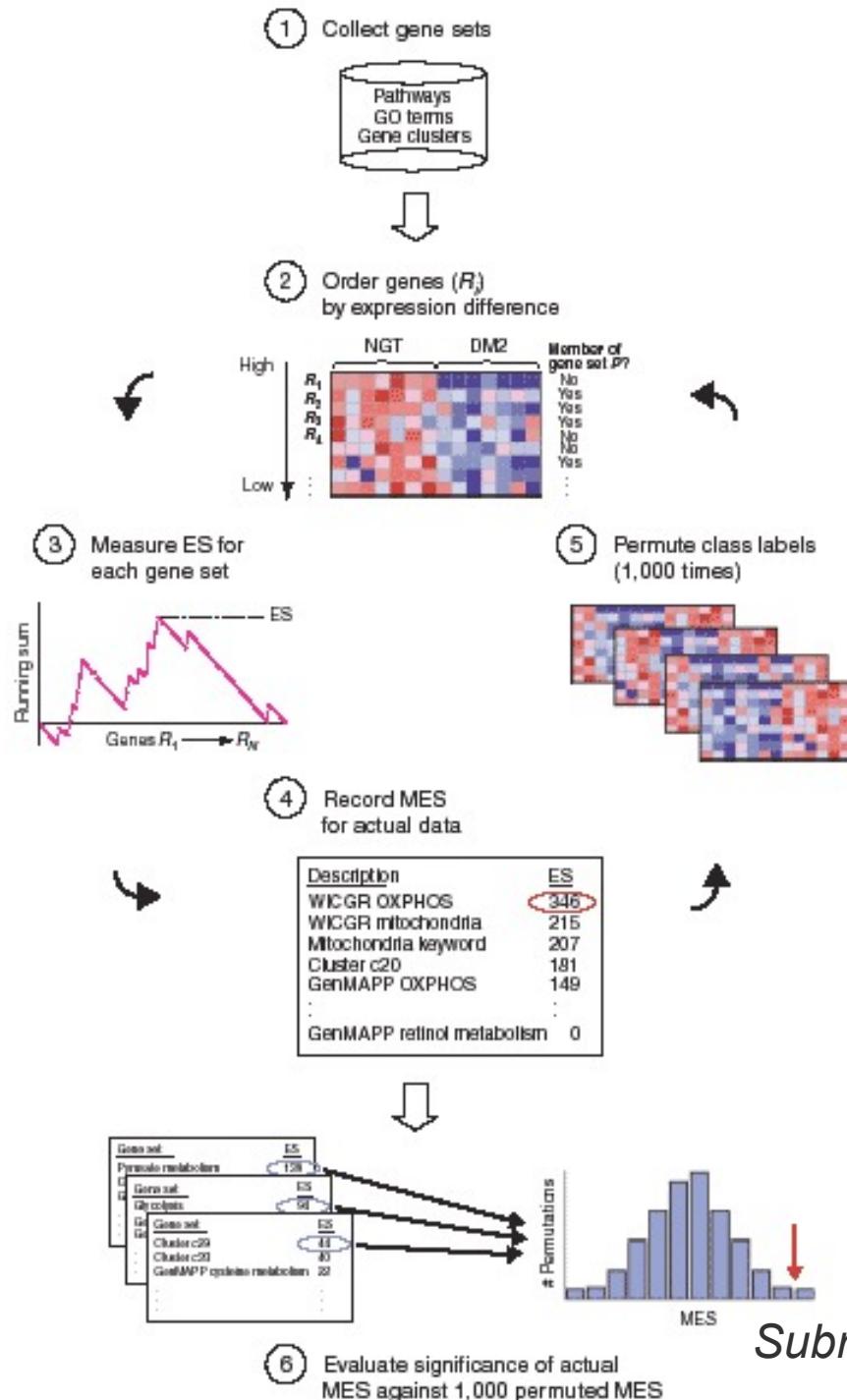


Different KS outcomes



Estimate
the statistical significance

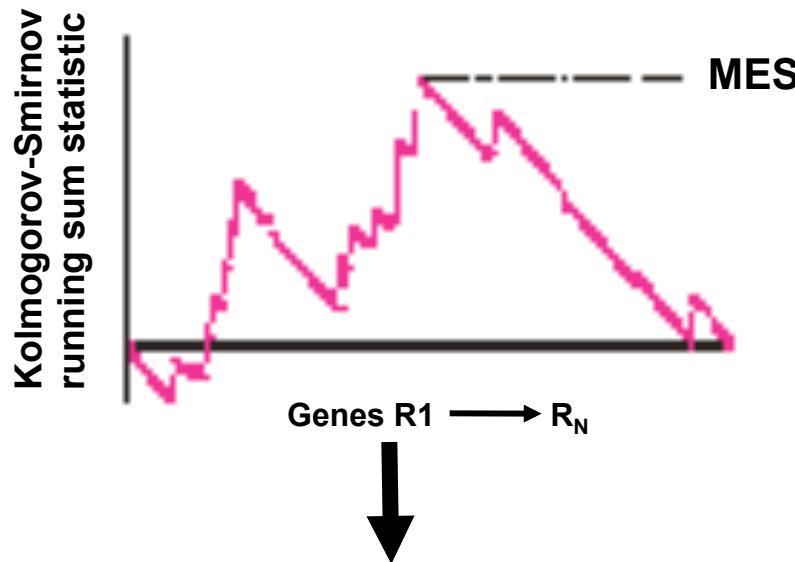
The statistical significance
(P-value) for each gene set is
calculated based on permutation of
genes labels



Subramanian et al. - 2005

Kolmogorov-Smirnov test statistic

- ③ Measure ES for each gene set S



- ④ Record Maximum Enrichment Score (MES)

Description	ES
WICGR OXPHOS	346
WICGR mitochondria	215
Mitochondria keyword	207
Cluster c20	181
GenMAPP OXPHOS	148
-	.
-	.
GenMAPP retinol metabolism	0

if R_i is not a member of S

$$X_i = -\sqrt{\frac{G}{N-G}}$$

if R_i is a member of S

$$X_i = \sqrt{\frac{N-G}{G}}$$

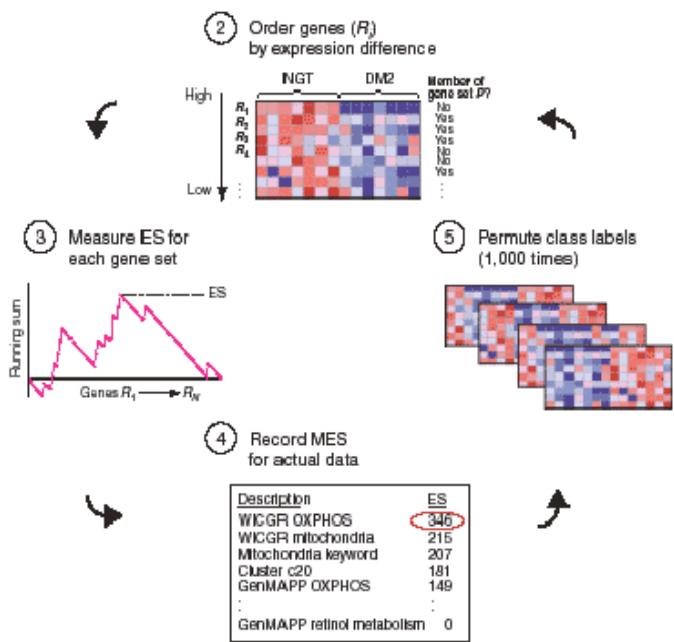
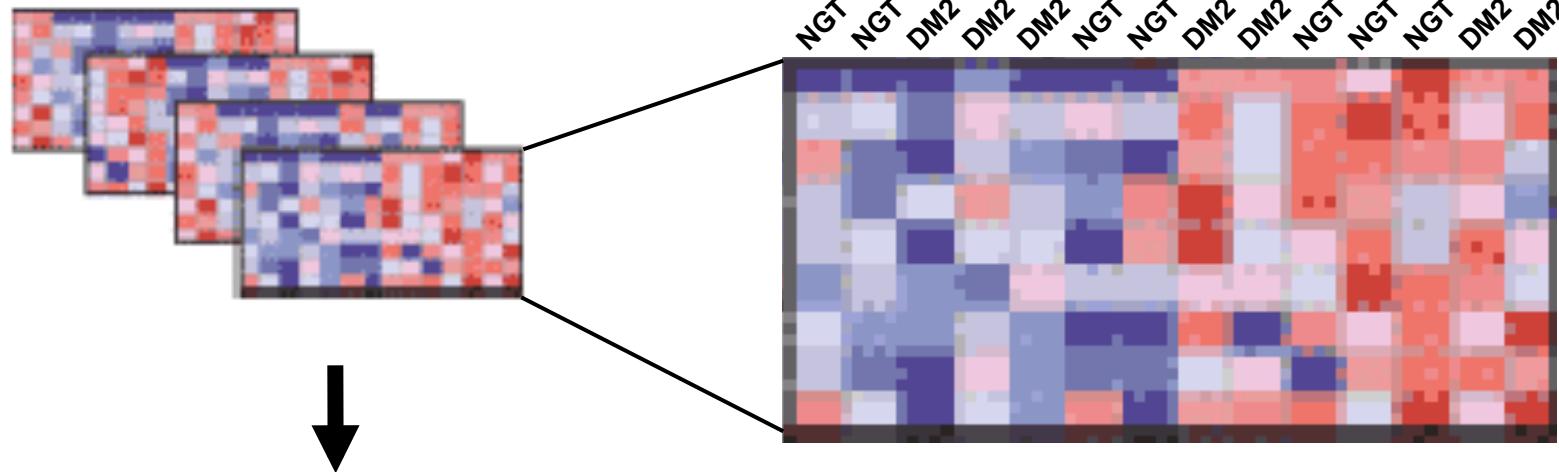


Kolmogorov-Smirnov
running sum statistic : ES

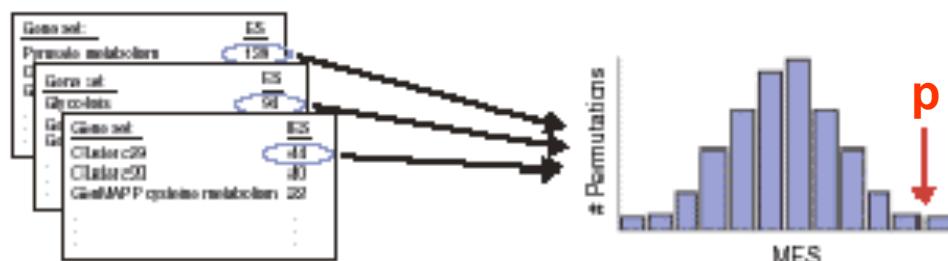
$$\text{MES} = \max_{1 \leq j \leq N} \sum_{i=1}^j X_i$$

N = Number of genes
G = number of members in a gene set S
MES = Maximum Enrichment Score

⑤ Permute class labels (1000 times)



⑥ Evaluate significance of actual MES against 1000 permuted MES



$$p = \frac{\text{nb permuted MES} > \text{MES}}{\text{nb permutations}}$$

Adjust for multiple hypothesis testing

To take into account multiple hypotheses
testing of multiple gene sets

Gene Set Enrichment Analysis

Advantages

- It only requires gene set membership information to compute enrichment scores
- It considers the entire ranked list of genes
- Threshold-free model

Gene Set Enrichment Analysis

Drawbacks

Significance is measured using a permutation-based procedure: Incorporates the permutation of pathway labels

=>thereby not preserving the “biological” correlation structure of the markers



A null distribution considering samples permutation would be computationally expensive

In R

```
>source("https://bioconductor.org/biocLite.R")
>biocLite("clusterProfiler")
>require(clusterProfiler)

# Get data
>table<-read.csv("GSEA_data_input.csv")
>df_case1 <- data.frame(table$gene.ID, table$scores, table$case1)
>colnames(df_case1) <- c("ID", "score", "S")
>head(df_case1)
```

	ID	score	S
1	17	0.65033	PATHWAY
2	42	0.65033	PATHWAY
3	29	0.43832	PATHWAY
4	30	0.43832	PATHWAY
5	159	0.43366	NO
6	178	0.43366	NO

In R

```
>source(https://bioconductor.org/biocLite.R)
>biocLite("clusterProfiler")
>require(clusterProfiler)

# Get data
>table<-read.csv("GSEA_data_input.csv")
>df_case1 <- data.frame(table$gene.ID, table$scores, table$case1)
>colnames(df_case1) <- c("ID", "score", "PATHWAY")
>head(df_case1)

# set score (those you get from a t-test or any other statistical test)
>SCORE=df_case1$score
>names(SCORE)=df_case1$ID
>SCORE=sort(SCORE,decreasing=TRUE)  17      42      29      30      159      178
>head(SCORE)
```

```
# get phenotype (term)
>term2gene_case1=data.frame(term=df_case1$PATHWAY,
                               name=df_case1$ID)
>head(term2gene_case1)
```

	term	name
1	PATHWAY	17
2	PATHWAY	42
3	PATHWAY	29
4	PATHWAY	30
5	NO	159
6	NO	178
7	PATHWAY	2
8	NO	179
9	NO	158
10	NO	157
11	PATHWAY	3
12	NO	4

In R

```
>source(https://bioconductor.org/biocLite.R)
>biocLite("clusterProfiler")
>require(clusterProfiler)

# Get data
>table<-read.csv("GSEA_data_input.csv")
>df_case1 <- data.frame(table$gene.ID, table$scores, table$case1)
>colnames(df_case1) <- c("ID","score","PATHWAY")
>head(df_case1)

# set score (those you get from a t-test or any other statistical test)
>SCORE=df_case1$score
>names(SCORE)=df_case1$ID
>SCORE=sort(SCORE,decreasing=TRUE)
>head(SCORE)

# get phenotype (term)
>term2gene_case1=data.frame(term=df_case1$PATHWAY,
                               name=df_case1$ID)
>head(term2gene_case1)

# run GSEA
>gsea.out_case1=GSEA(SCORE,
                      TERM2GENE=term2gene_case1,
                      nPerm=10000,
                      pvalueCutoff=1,
                      pAdjustMethod = "BH")
>gseaplot(gsea.out_case1,"PATHWAY")
```

Enrichment Analysis classification

Gene Set Enrichment Analysis

Singular Enrichment Analysis

Modular Enrichment Analysis

Overview

Count matrix

Differential expression

Enrichment

SEA

GSEA

MEA

Knowledge



Gene Set Enrichment Analysis

- All genes are included in analysis
- Pairwise comparisons (e.g., disease vs. control)



No need to select list

Example

GSEA of broad institute

GSA

SAFE

GeneTrail

FatiScan

Singular Enrichment Analysis

- P-value calculated on each term from pre-selected list
- Enrichment terms are listed

Example

ClueGO

GOStat

DAVID: Provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes

FatiGO

Marmite

Babelomics Suite: Suite of web tools for the functional profiling of genome scale experiments

Modular Enrichment Analysis

- Predetermined list of genes
- term-term or gene-gene relationships included in enrichment P-value calculation

→ Closest to nature of biological data structure

We could consider the gene-gene relationship

Example

DAVID
GOtoolBox



EXERCICE 2

1. Using exercise 1 evaluated p-values, what is the outcome of GSEA on pathway 1 (*Pathway.csv*)?

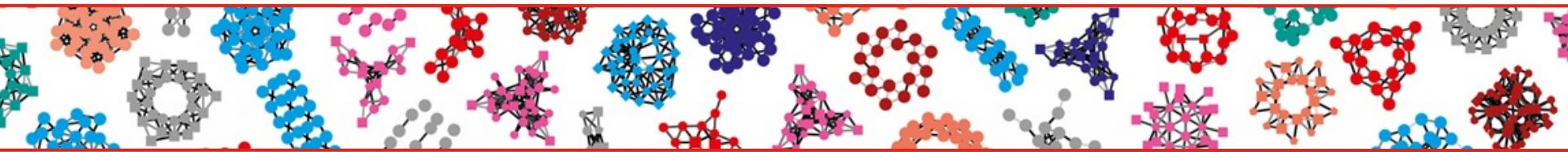
In column pathways of *Pathway.csv* designates the probes of interest (“yes” means in pathway, “no” means not in pathway)

2. Transform affymetrix probes into genes names

Given a list of Affymetrix probes, in R

```
>library("AnnotationDbi")
>library("rat2302.db")
>PROBES  <- rat$row.names
>OUT      <- select(rat2302.db,keys= PROBES, columns=c("SYMBOL", "ENTREZID", "GENENAME"))
```

3. What is the enrichment outcome on probes coding for ribosomal proteins?



Note:

If the following command line raises an error
->?GSEA

Then install
bit
AnnotationDbi
DO.db
stringi

biocLite("tibble")
biocLite("clusterProfiler")

In R solution

```
>require(clusterProfiler)
>pathway<-read.csv(file="Pathway.csv", stringsAsFactor=FALSE, header=TRUE)

# set score (those you get from a t-test or any other statistical test)
>rawp <- apply(rat, 1, ttestRat, grp1 = c(2:7), grp2 = c(8:12))
>names(rawp)           <-rat$row.names
>sortedrawp            <-sort(rawp)
>p_holm                <-p.adjust(sortedrawp,method="BH")
>names(p_holm)          <-names(sortedrawp)
>SCORE                  <-p_holm
>SCORE                  <-sort(SCORE,decreasing=TRUE)
>head(SCORE)

# get phenotype (term)
>term2gene              <-data.frame(term=pathway$pathways, name=pathway$row.names)
>head(term2gene)

# run GSEA
>gsea.out                <-GSEA(SCORE, TERM2GENE=term2gene, nPerm=10000, pvalueCutoff=1,
pAdjustMethod = "BH")
>gseaplot(gsea.out,"yes")
>summary(gsea.out)
```

In R solution

```
>library("AnnotationDbi")
>library("rat2302.db")
>library("DescTools")

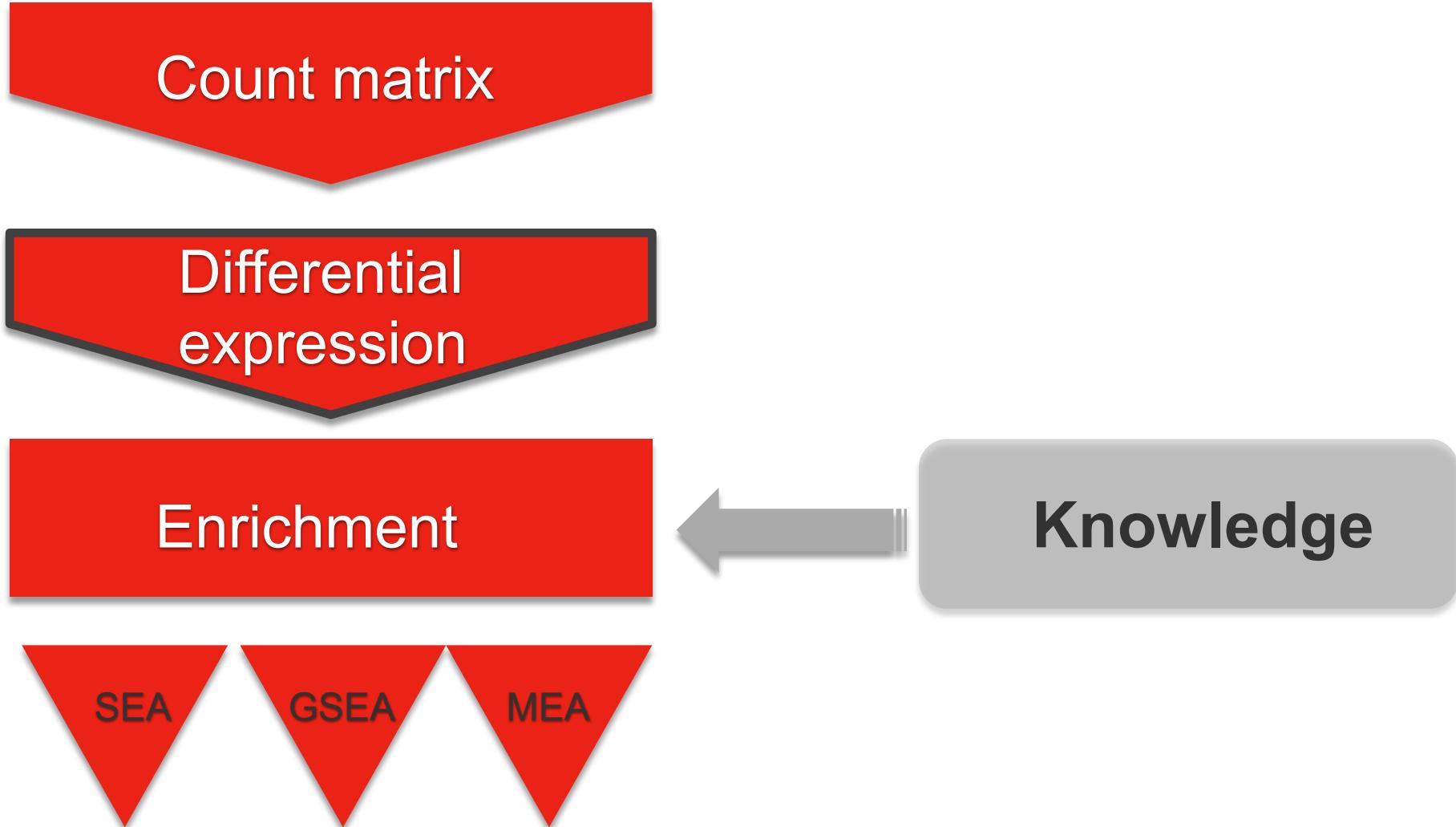
>PROBES<- rat$row.names
>OUT    <- select(rat2302.db,keys= PROBES, columns=c("SYMBOL", "ENTREZID", "GENENAME"))
>ribosomal<-OUT[ which(OUT$GENENAME %like% "ribosomal protein"),]

# set score (those you get from a t-test or any other statistical test)
>rawp <- apply(rat, 1, ttestRat, grp1 = c(2:7), grp2 = c(8:12))
>names(rawp)<-rat$row.names
>sortedrawp<-sort(rawp)
>p_holm <- p.adjust(sortedrawp,method="BH")
>names(p_holm)<-names(sortedrawp)
>SCORE<-p_holm
>SCORE=sort(SCORE,decreasing=TRUE)
>head(SCORE)

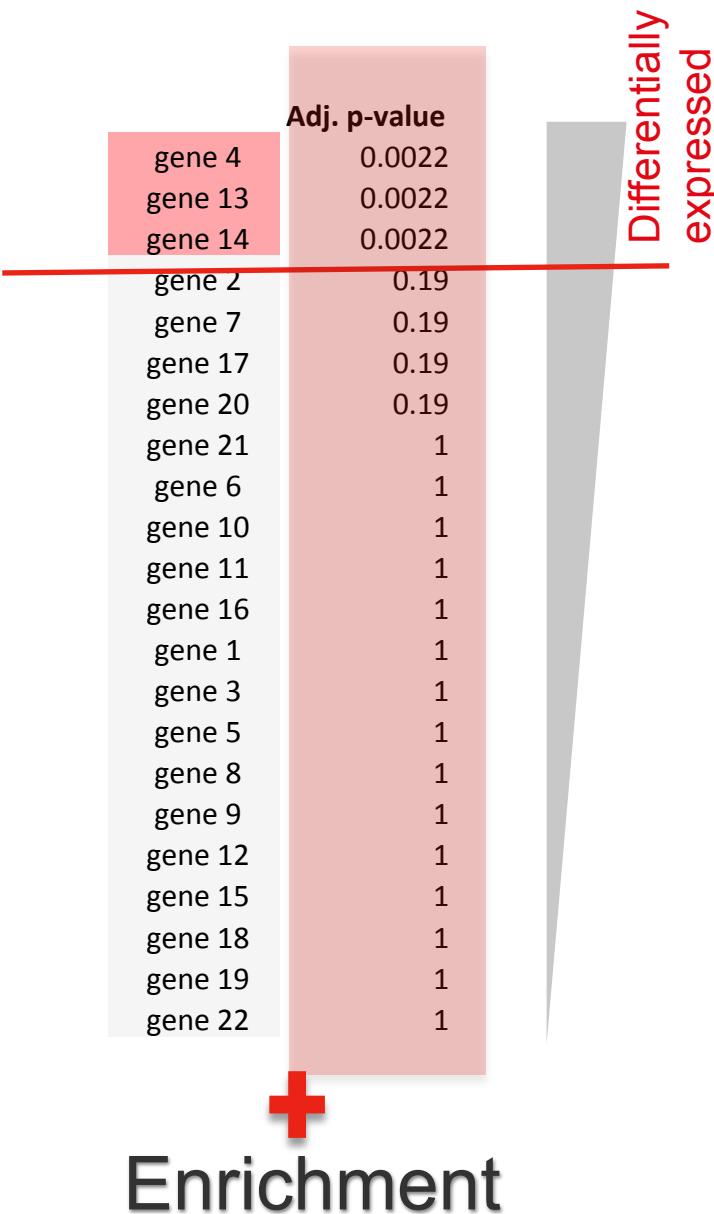
# get phenotype (term)
>term2gene<-data.frame(term="no",name=rat$row.names,stringsAsFactors=FALSE)
>term2gene[which(term2gene$name %in% ribosomal$PROBEID),1]<-"yes"
>head(term2gene)

# run GSEA
>gsea.out<-GSEA(SCORE, TERM2GENE=term2gene, nPerm=10000, pvalueCutoff=1, pAdjustMethod = "BH")
>gseaplot(gsea.out, "yes")
>summary(gsea.out)
```

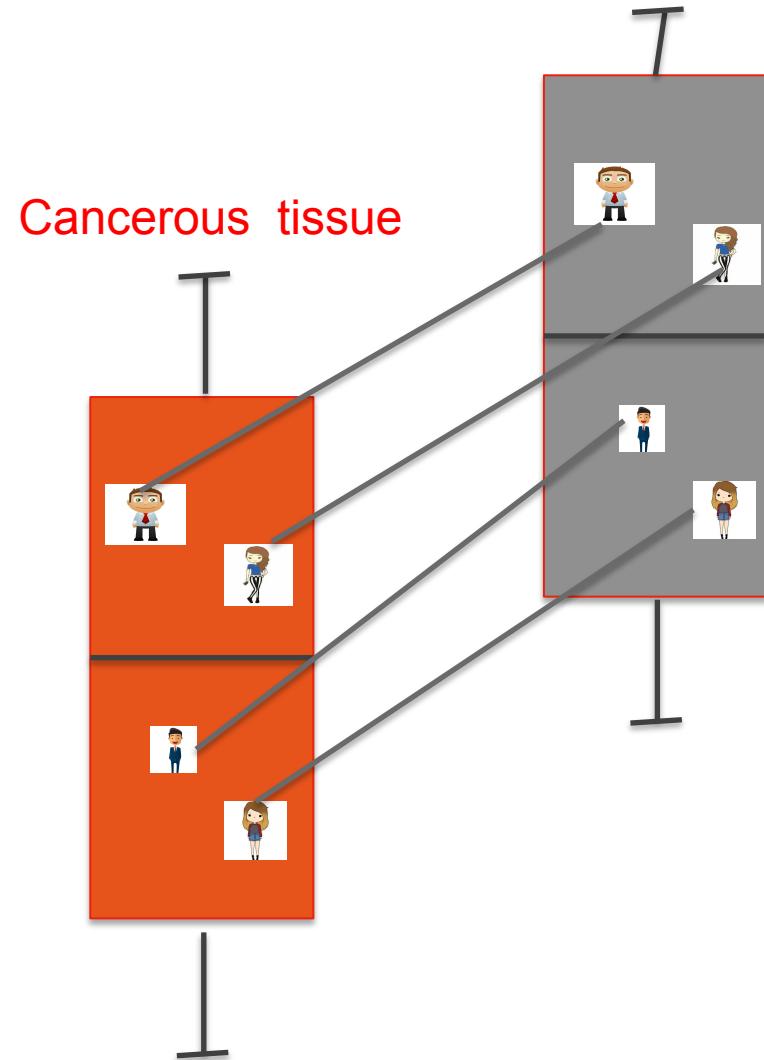
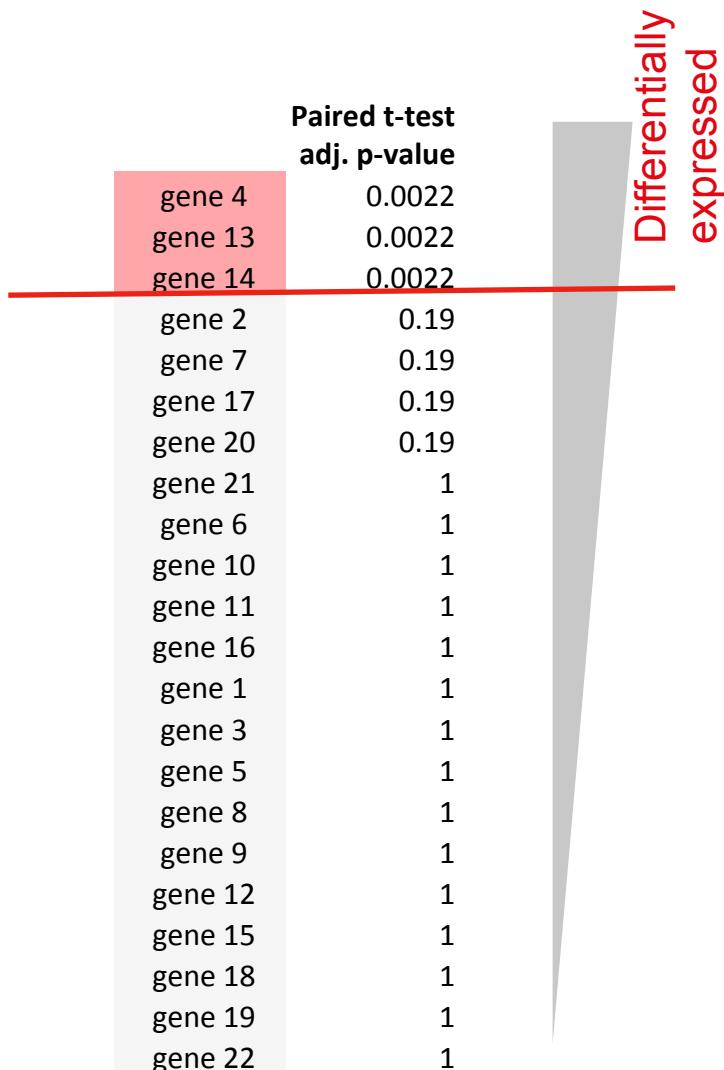
Overview



Application to any type of data!



Paired data



Paired T-test: Equivalent to testing whether the difference between the pairs is different from zero

Enrichment

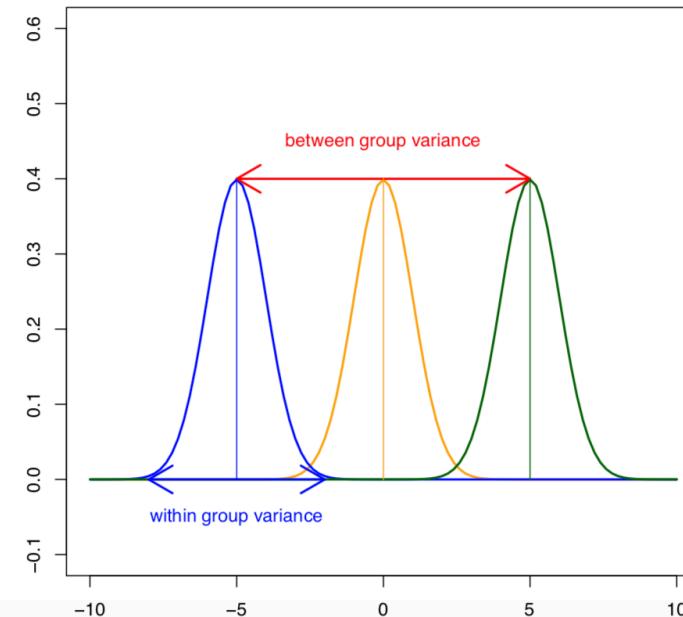
ANOVA one factor

	F score or p-value of ANOVA	Differentially expressed
gene 4	0.0022	
gene 13	0.0022	
gene 14	0.0022	
gene 2	0.19	
gene 7	0.19	
gene 17	0.19	
gene 20	0.19	
gene 21	1	
gene 6	1	
gene 10	1	
gene 11	1	
gene 16	1	
gene 1	1	
gene 3	1	
gene 5	1	
gene 8	1	
gene 9	1	
gene 12	1	
gene 15	1	
gene 18	1	
gene 19	1	
gene 22	1	

ANOVA

=

analysis of variance mean



ANOVA determines whether there are any statistically significant differences between the means of three or more independent (unrelated) groups

`aov(expression~ patient type)`

where patient type either healthy, sick without nodules, sick with nodules

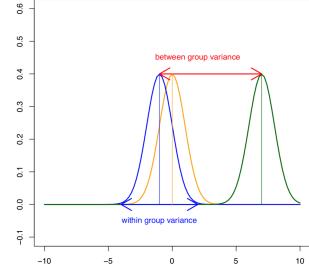
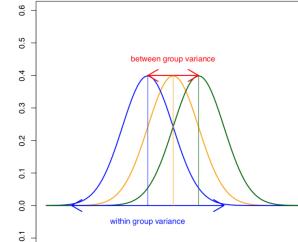
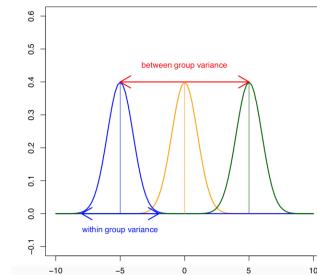


Enrichment

ANOVA one factor

	F score or p-value of ANOVA
gene 4	0.0022
gene 13	0.0022
gene 14	0.0022
gene 2	0.19
gene 7	0.19
gene 17	0.19
gene 20	0.19
gene 21	1
gene 6	1
gene 10	1
gene 11	1
gene 16	1
gene 1	1
gene 3	1
gene 5	1
gene 8	1
gene 9	1
gene 12	1
gene 15	1
gene 18	1
gene 19	1
gene 22	1

Differentially expressed



within group variance = SS_{error}

between group variance = SS_{group}

if $SS_{\text{group}} > SS_{\text{error}}$

⇒ at least two means are different!



Enrichment

ANOVA two factors

	F score or p-value of ANOVA
gene 4	0.0022
gene 13	0.0022
gene 14	0.0022
gene 2	0.19
gene 7	0.19
gene 17	0.19
gene 20	0.19
gene 21	1
gene 6	1
gene 10	1
gene 11	1
gene 16	1
gene 1	1
gene 3	1
gene 5	1
gene 8	1
gene 9	1
gene 12	1
gene 15	1
gene 18	1
gene 19	1
gene 22	1

Differentially expressed

Patient type

Gender

	Male	Female
Healthy		
Patient without nodular aspect		
Patient with nodular aspect		

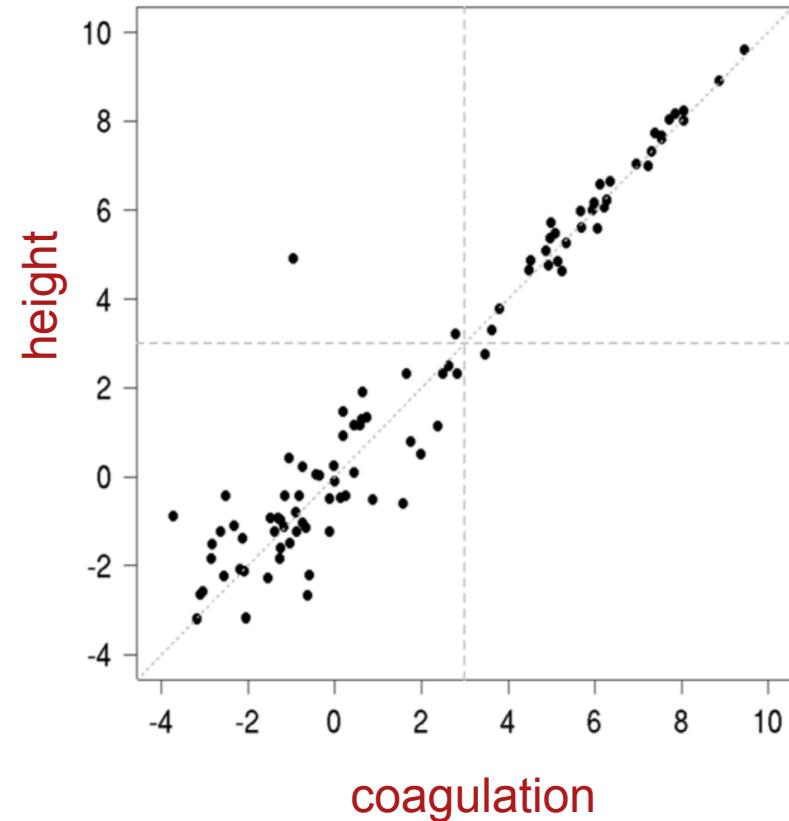
$aov(expression \sim patient\ type * gender)$



Enrichment

Linear model

	Adj. p-value of LM	Differentially expressed
gene 4	0.0022	
gene 13	0.0022	
gene 14	0.0022	
gene 2	0.19	
gene 7	0.19	
gene 17	0.19	
gene 20	0.19	
gene 21	1	
gene 6	1	
gene 10	1	
gene 11	1	
gene 16	1	
gene 1	1	
gene 3	1	
gene 5	1	
gene 8	1	
gene 9	1	
gene 12	1	
gene 15	1	
gene 18	1	
gene 19	1	
gene 22	1	



$\text{lm(expression} \sim \text{height + coagulation)}$

$\text{lm(expression} \sim \text{height * coagulation)}$



Enrichment

Linear model

	Adj. p-value of LM	Differentially expressed
gene 4	0.0022	
gene 13	0.0022	
gene 14	0.0022	
gene 2	0.19	
gene 7	0.19	
gene 17	0.19	
gene 20	0.19	
gene 21	1	
gene 6	1	
gene 10	1	
gene 11	1	
gene 16	1	
gene 1	1	
gene 3	1	
gene 5	1	
gene 8	1	
gene 9	1	
gene 12	1	
gene 15	1	
gene 18	1	
gene 19	1	
gene 22	1	

Rotation-based GSEA

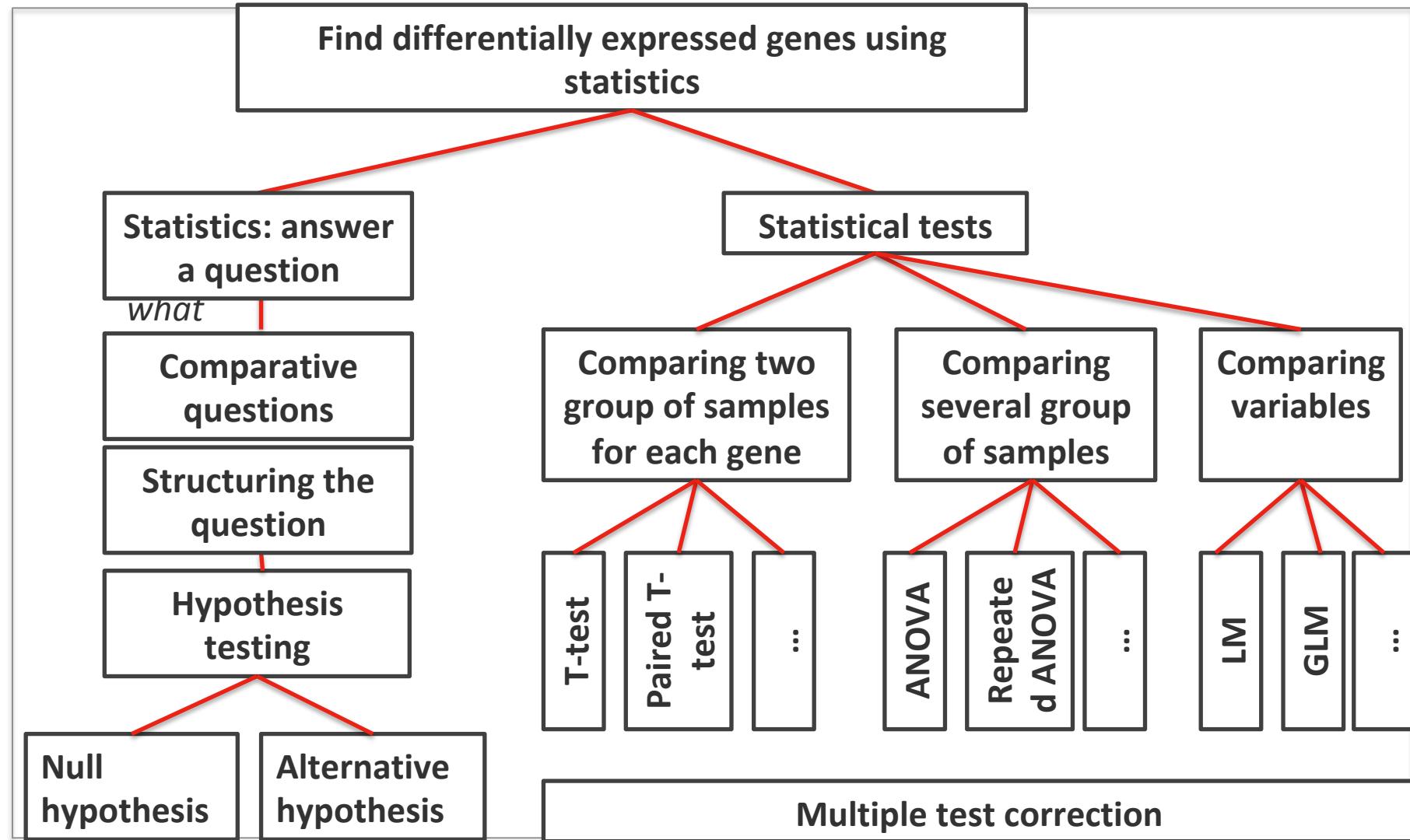
Implemented in the “ROMER” functionality within the limma package from Bioconductor

[Langsrud, 2005; Wu et al, 2010]



Enrichment

Wrap up



Application to any type of data!

Homo sapiens
Gorilla gorilla
Macaca mulatta
Callithrix jacchus
Bos taurus
Felis caritius
Canis lupus
Myotis lifugus

EKAHVAVSALWHKPLVTTTGVNNLPPHVEEFGGEALGRPPLLVNNLPVYPW
EKAHVAVSALWHKPLVTTTGVNNLPPHVEEFGGEALGRPPLLVNNLPVYPW
EKA..AVSALWHK..V.....HVEEFGGEALGR..LLV....VYPW
EKT..AVLALWNN..S.....DVEDCGGEALGR..LLV....VYPW
EKT..QVTNMWGK..V.....NVKELGGEALSR..LLV....VYPW
EKT..QVTNLWGK..V.....NVKELGGEALSR..LLV....VYPW
EKT..QVTNLWGK..P.....NVKELGGEALSR..LLV....VYPW
EKT..QVTNLWGK..V.....NVKELGGEALSR..LLV.\..VYPW

Conserved sites

Application to any type of data!

Homo sapiens
Gorilla gorilla
Macaca mulatta
Callithrix jacchus
Bos taurus
Felis caritius
Canis lupus
Myotis lifugus

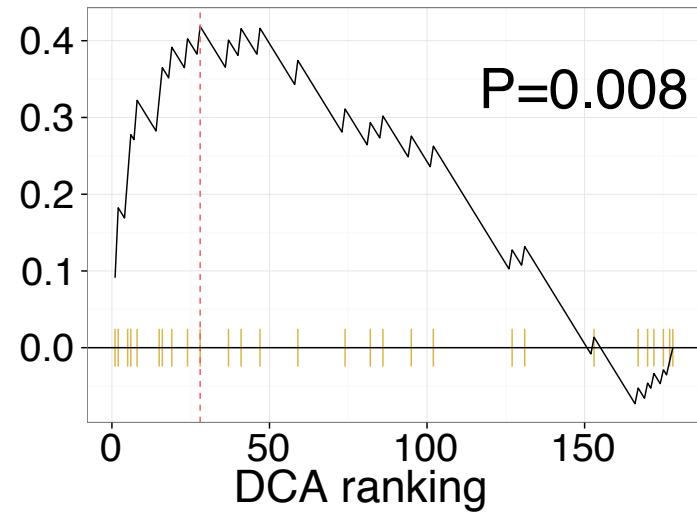
Homo sapiens	EKAHVAVSALWHKPLVTTTGVNNLPPHVEEFGGEALGRPPPLLVNNLPVYPW
Gorilla gorilla	EKAHVAVSALWHKPLVTTTGVNNLPPHVEEFGGEALGRPPPLLVNNLPVYPW
Macaca mulatta	EKA .. AVSALWHK .. V HVEEFGGEALGR .. LLV VYPW
Callithrix jacchus	EKT .. AVLALWNN .. S DVEDCGGEALGR .. LLV VYPW
Bos taurus	EKT .. QVTNMWGK .. V NVKELGGEALSR .. LLV VYPW
Felis caritius	EKT .. QVTNLWGK .. V NVKELGGEALSR .. LLV VYPW
Canis lupus	EKT .. QVTNLWGK .. P NVKELGGEALSR .. LLV VYPW
Myotis lifugus	EKT .. QVTNLWGK .. V NVKELGGEALSR .. LLV .. \ . . VYPW

coevolution

Application to any type of data!

Coevolution		
	score	polymorphic
position 4	0.2	no
position 13	0.2	no
position 14	0.15	no
position 2	0.15	no
position 7	0.14	yes
position 17	0.14	no
position 20	0.14	no
position 21	0.14	no
position 6	0.06	no
position 10	0.06	yes
position 11	0.06	yes
position 16	0.06	no
position 1	0.001	yes
position 3	0.001	no
position 5	0.001	yes
position 8	0.001	yes
position 9	0.001	yes
position 12	0.001	no
position 15	0.001	yes
position 18	0.001	yes
position 19	0.001	no
position 22	0.001	yes

Polymorphism in human and coevolving constraints





EXERCICE 3

How does GSEA deal with genes sets enrichment when they in the following configurations



GSEA_data_input.csv



EXERCICE 3

A. Are the pathway's genes, pinpointed in case2 of *GSEA_data_input.csv* dataset, highly differentially expressed?

Answer using

- a. Fisher test and a threshold of 0.17 on scores
- b. GSEA

B. Repeat question 1 using pathway association of cases 3, 4 and 5 of *GSEA_data_input.csv* dataset.

In R solution

```
>source(https://bioconductor.org/biocLite.R)
>biocLite("clusterProfiler")
>require(clusterProfiler)

# Get data
>table<-read.csv("GSEA_data_input.csv")

>case2.mat=matrix(c(
  length(which(table$scores<0.17 & table$case2=="PATHWAY")) ,
  length(which(table$scores<0.17 & table$case2=="NO")) ,
  length(which(table$scores>0.17 & table$case2=="PATHWAY")) ,
  length(which(table$scores>0.17 & table$case2=="NO")))
, nrow=2)

>fisher.test(case2.mat)
```

In R solution

```
>df_case2 <- data.frame(table$gene.ID, table$scores, table$case2)
>colnames(df_case2) <- c("ID", "score", "S")
>head(df_case2)

>SCORE=df_case2$score
>names(SCORE)=df_case2$ID
>SCORE=sort(SCORE, decreasing=TRUE)
>head(SCORE)

>term2gene_case2=data.frame(term=df_case2$S, name=df_case2$ID)
>head(term2gene_case2)

>gsea.out_case2<-GSEA(SCORE,
                        TERM2GENE=term2gene_case2,
                        nPerm=10000,
                        pvalueCutoff=1,
                        pAdjustMethod = "BH")

>gseaplot(gsea.out_case1,"PATHWAY")

...
```

Overview

Count matrix

Differential expression

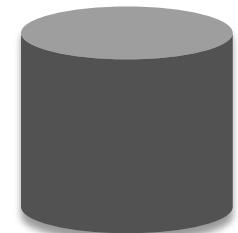
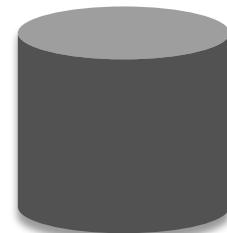
Enrichment

SEA

GSEA

MEA

Knowledge

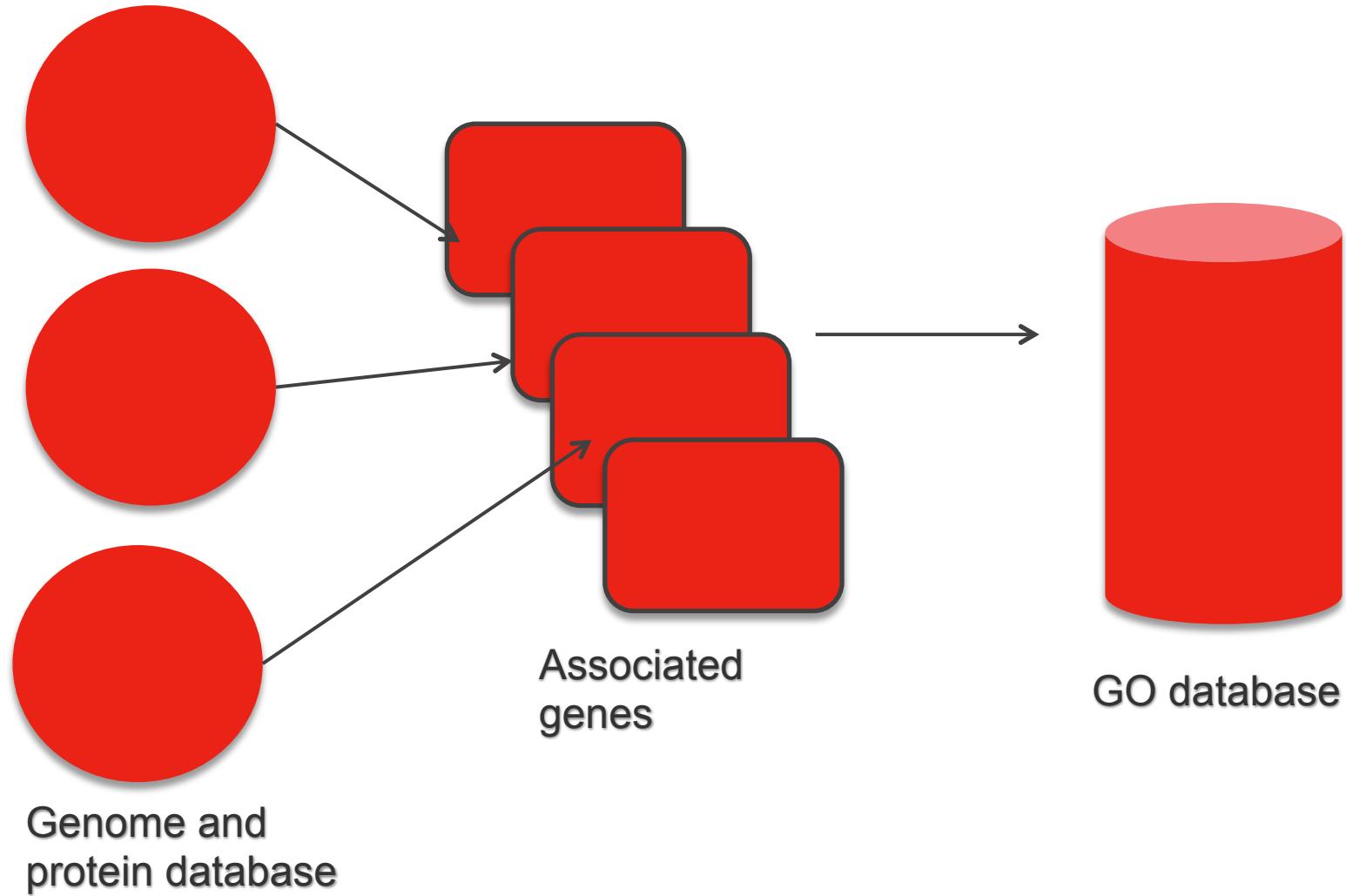


Enrichment analysis & ontologies

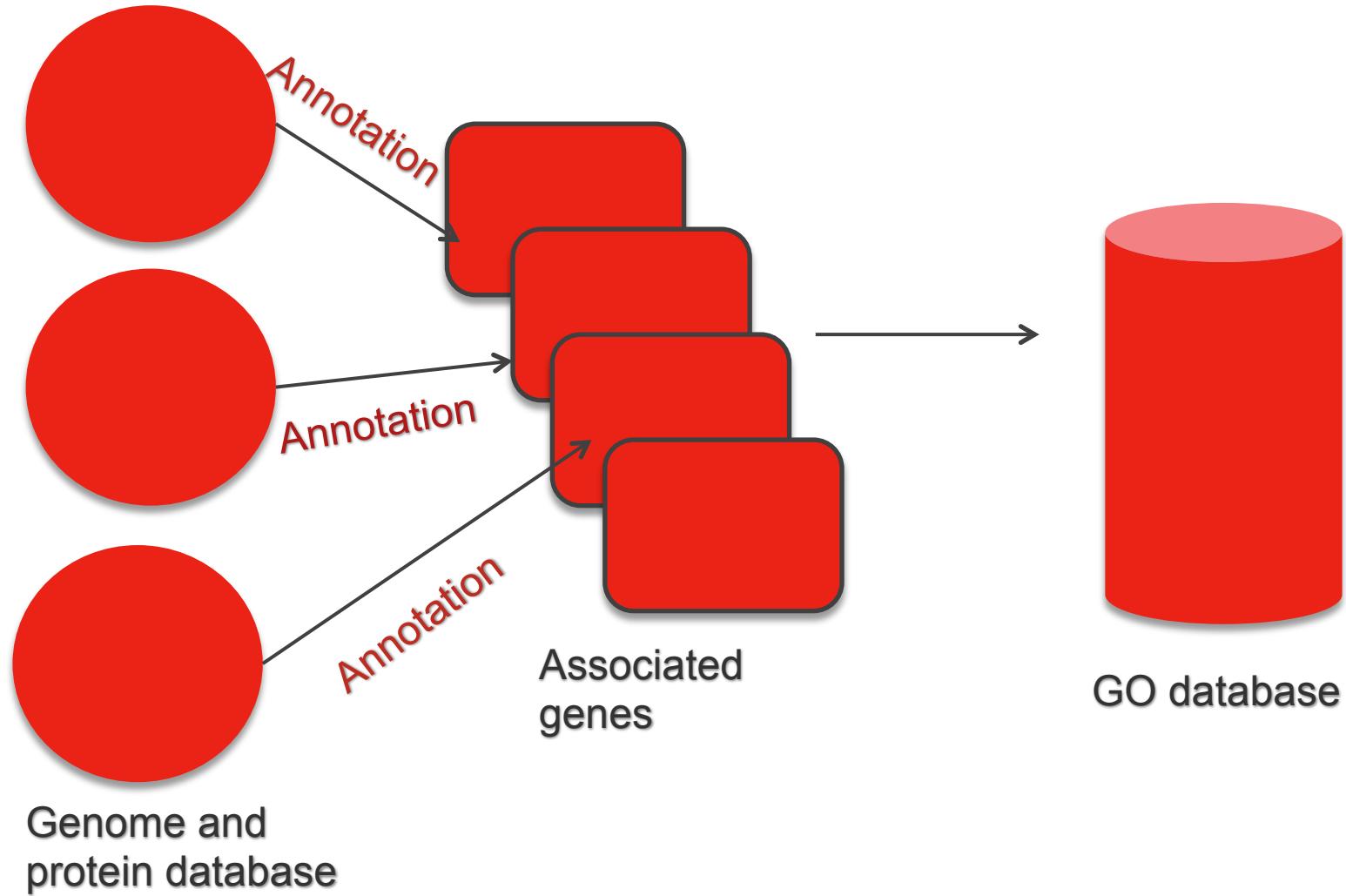
An ontology is a **specification of the concepts & relationships** that can exist in a domain of discourse.

The Gene Ontology (GO) project is an effort to provide **consistent descriptions of gene products**

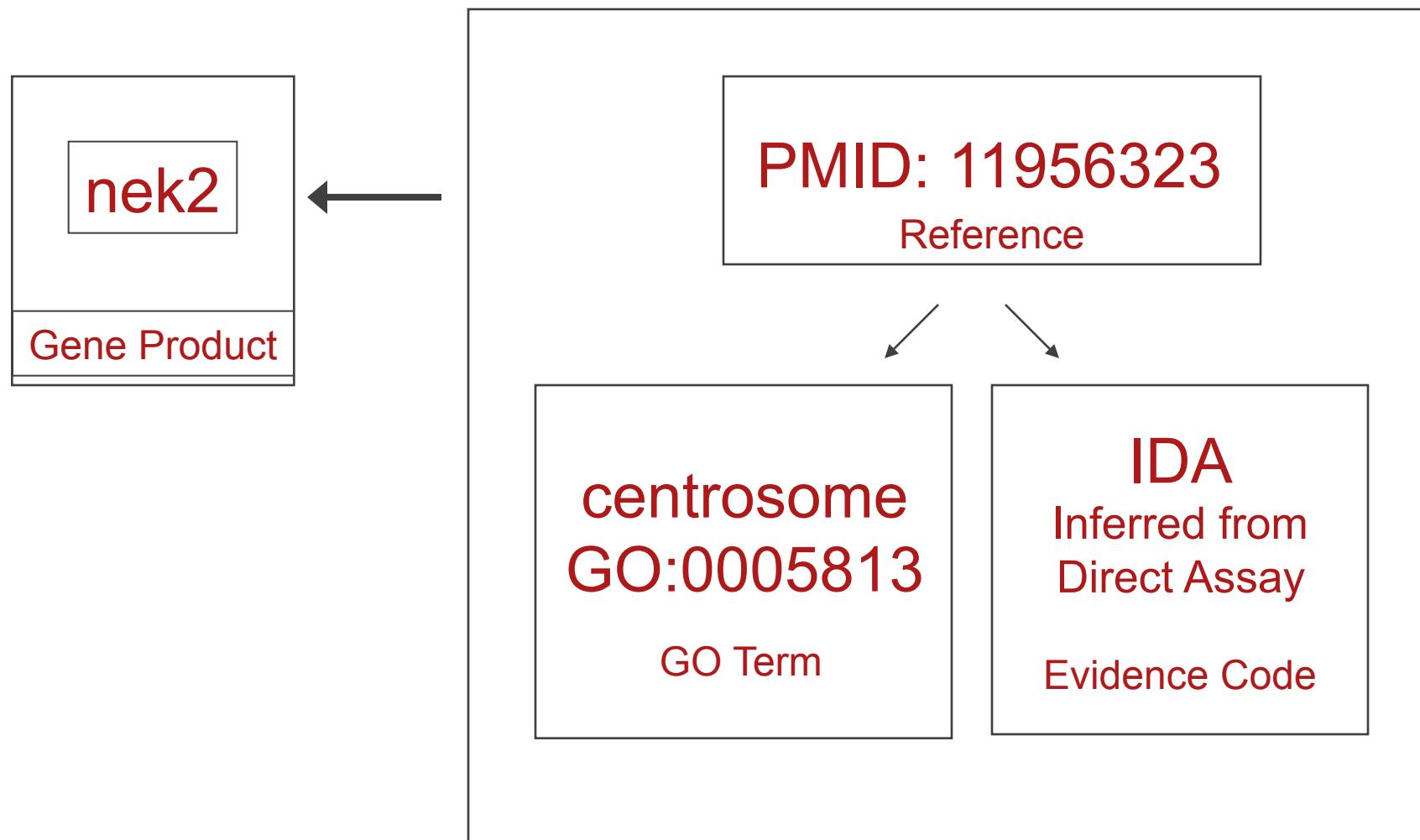
The Gene Ontology (GO)



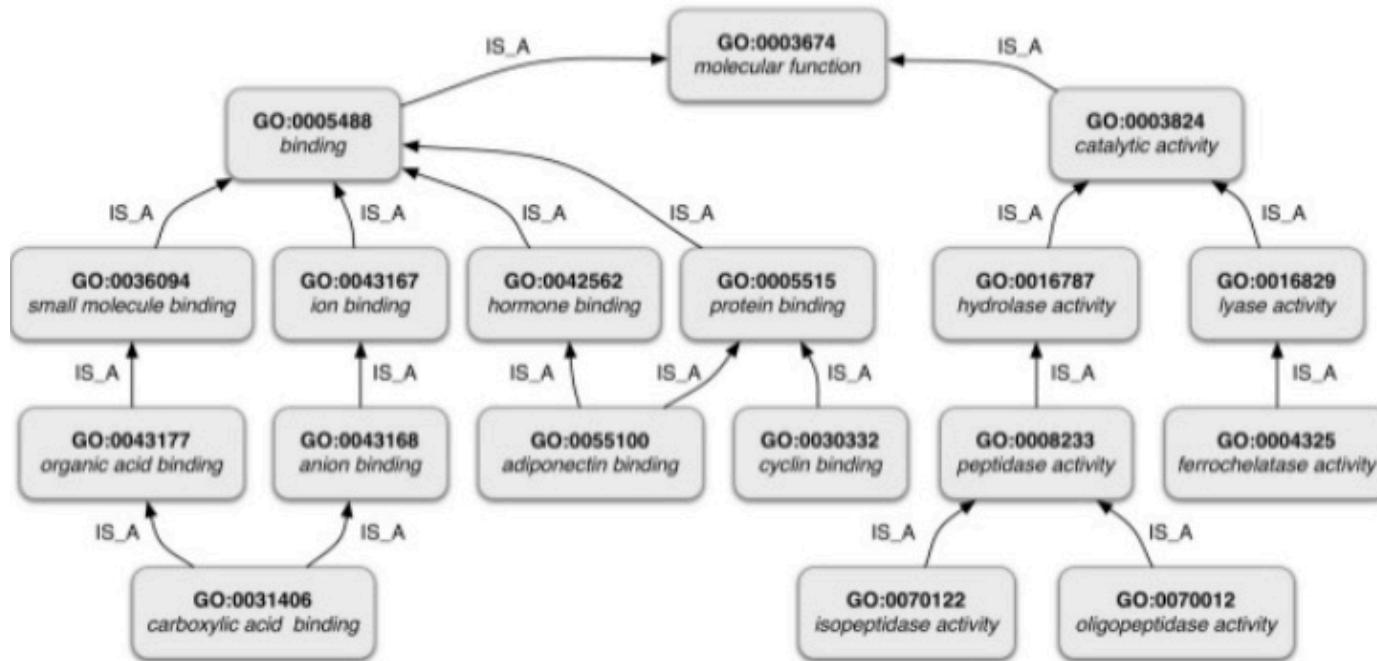
The Gene Ontology (GO)



Example Annotation



Links between GO terms



GO domain

1. cellular component (CC)
2. biological process (BP)
3. molecular function (MF)

In enrichment several GO terms are checked



multiple testing

Comparison to other GO enrichment tools (as of late 2008)

Table I: A comparison of web-based GO enrichment tools.

Tool	P-value and statistical method	Flexible threshold	Graphical visualization	Multiple organisms	Running time
GOrilla	Exact mHG p-value computation (no need for simulations)	+	+	+	7 Sec
Fatiscan [13]	Fischer Exact (FDR corrected for number of thresholds)	+ (predetermined steps of 30)	-	+	30 Min
GO-stat [14]	Wilcoxon Rank-Sum/ Kolmogorov Smirnov	+	-	+	2 Min
GOEAST [9]	Hypergeometric	-	+	+	20 Min
SGD [11]	Hypergeometric	-	+	- (only yeast)	2 Min
DAVID [7]	Modified Fischer Exact	-	-	+	2 Min
GOTM [10]	Hypergeometric	-	+	+	2 Min
GoMiner [3]	Fisher Exact	-	-	+	7 Min
			(only in the downloadable version)		

In R

```
>source(https://bioconductor.org/biocLite.R)
>biocLite("clusterProfiler")
>require(clusterProfiler)

# Use GSEA to evaluate the Gene set enrichment and find an ontology that is
differentially expressed in our dataset
>? gseGO
>gsecc<- gseGO(geneList      = geneList,
                  OrgDb        = org.Hs.eg.db,
                  ont          = "ALL",
                  nPerm        = 10000,
                  pvalueCutoff = 1,
                  verbose      = FALSE)

>gseaplot(gsecc, geneSetID="GO:0000779")
```

→ Ranked adj. p-value scores

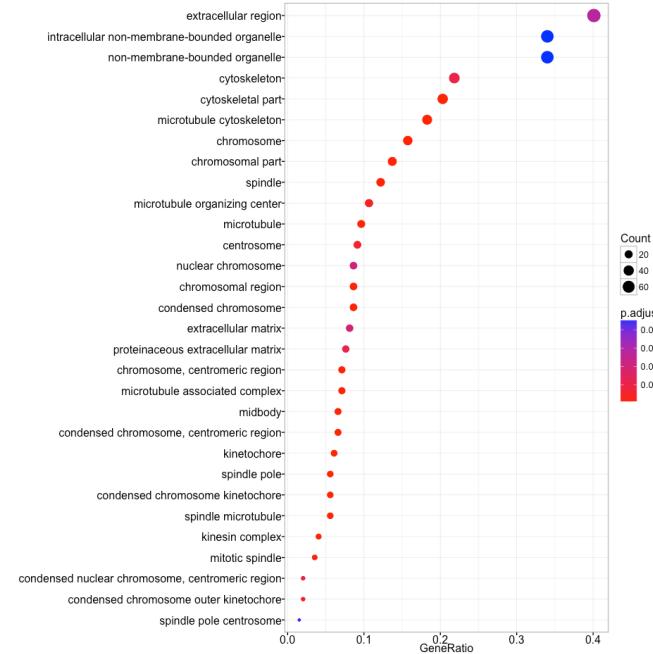
In R

```
>source(https://bioconductor.org/biocLite.R)
>biocLite("clusterProfiler")
>require(clusterProfiler)

# Use GSEA to evaluate the Gene set enrichment and find an ontology that is
differentially expressed in our dataset
>? gseGO
>gsecc<- gseGO(geneList      = geneList,
                  OrgDb        = org.Hs.eg.db,
                  ont          = "ALL",
                  nPerm        = 10000,
                  pvalueCutoff = 1,
                  verbose      = FALSE)

>gseaplot(gsecc, geneSetID="GO:0000779")

# Visualize
>?dotplot
>dotplot(ego, showCategory=30)
```



In R

```
>source(https://bioconductor.org/biocLite.R)
>biocLite("clusterProfiler")
>require(clusterProfiler)

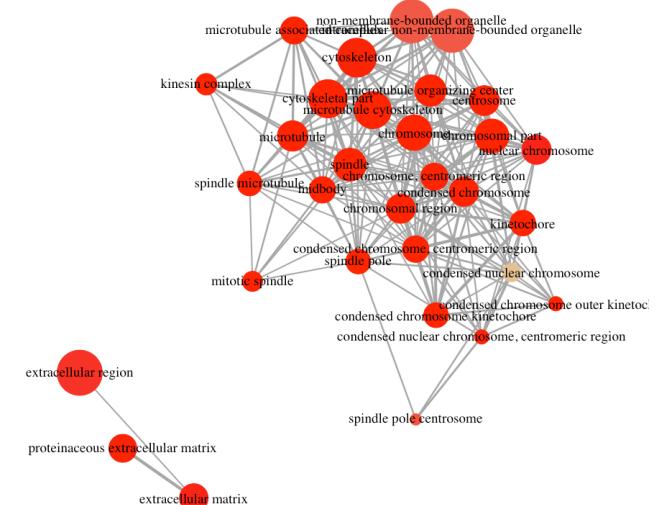
# Use GSEA to evaluate the Gene set enrichment and find an ontology that is
differentially expressed in our dataset
>? gseGO
>gsecc<- gseGO(geneList      = geneList,
                  OrgDb        = org.Hs.eg.db,
                  ont          = "ALL",
                  nPerm        = 10000,
                  pvalueCutoff = 1,
                  verbose      = FALSE)

>gseaplot(gsecc, geneSetID="GO:0000779")

# Visualize
>?dotplot
>dotplot(ego, showCategory=30)

>?enrichMap
```

→ Ranked adj. p-value scores



In R

```
>source(https://bioconductor.org/biocLite.R)
>biocLite("clusterProfiler")
>require(clusterProfiler)

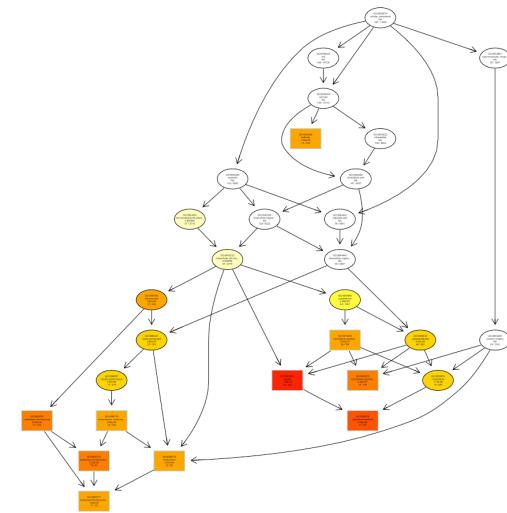
# Use GSEA to evaluate the Gene set enrichment and find an ontology that is
differentially expressed in our dataset
>? gseGO
>gsecc<- gseGO(geneList      = geneList,
                  OrgDb        = org.Hs.eg.db,
                  ont          = "ALL",
                  nPerm        = 10000,
                  pvalueCutoff = 1,
                  verbose      = FALSE)

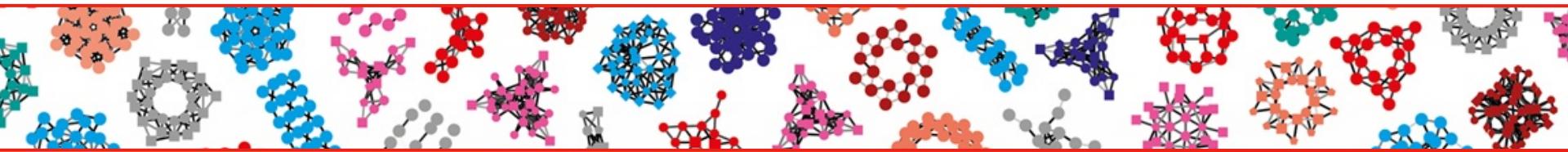
>gseaplot(gsecc, geneSetID="GO:0000779")

# Visualize
>?dotplot
>dotplot(ego, showCategory=30)

>?enrichMap

>plotGOgraph(ego)
```





EXERCICE 4:

Enrichment and ontologies

HS_pvalues.csv is a file containing the adj. p-values issued from an ANOVA statistical test for several ENTREZ gene names.

1. Use *HS_pvalues.csv* dataset and look for GO ontologies that are enriched with a significant value
2. What are the gene names that enriched the best GO ontology?

In R

```
>require(clusterProfiler)
>library(org.Hs.eg.db)
>keytypes(org.Hs.eg.db)

>table      <-read.csv("HS_pvalues.csv")
>SCORE      <-table$score
>names(SCORE)<-table$gene.ENTREZ.ID
>SCORE      <-sort(SCORE ,decreasing=T)

>ego <- gseGO(geneList      = SCORE,
                OrgDb        = org.Hs.eg.db,
                ont          = "ALL",
                nPerm        = 1000,
                pvalueCutoff = 1,
                verbose      = FALSE)

>head(ego)

>gseaplot(ego, geneSetID="GO:0048518")
>dotplot(ego, showCategory=30)
>enrichMap(ego)
>plotGOgraph(ego)

#can be used on the outcome of enrichGO function
#barplot(ego, showCategory=30)
```



EXERCICE extra: Enrichment and ontologies

1. Use *Rat_KS.txt* dataset and look for GO ontologies that are enriched with a significant value
2. What are the gene names that enriched the best GO ontology?
3. What is the enrichment outcome using KEGG database?
4. Can you distinguish up and down-regulated genes enrichments ?

In R: solution

```
>require(clusterProfiler)
>library(org.Rn.eg.db)
>keytypes(org.Rn.eg.db)
# USE RAT DATABASES AND ANNOTATION
>rat<-read.csv("rat_KD.txt")
>PROBES<- rat$row.names
>OUT  <- select(rat2302.db,keys= PROBES, columns=c("SYMBOL", "ENTREZID",
"ENSEMBL"))
>duplicated(OUT$PROBEID)
>OUT<-OUT [-which(duplicated(OUT$PROBEID)),]
>dim(OUT)

>rawp <- apply(rat, 1, ttestRat, grp1 = c(2:7), grp2 = c(8:12))
>names(rawp)           <-OUT$ENSEMBL
>sortedrawp            <-sort(rawp)
>p_holm                <-p.adjust(sortedrawp,method="BH")
>names(p_holm)          <-names(sortedrawp)
>SCORE                  <-p_holm
>SCORE                  <-sort(SCORE, decreasing=TRUE)
>head(SCORE)

>egoGSECC <- gseGO(geneList= SCORE,
  OrgDb        = org.Rn.eg.db,
  keyType      = 'ENSEMBL',
  ont          = "CC",
  nPerm        = 1000,
  minGSSize    = 10,
  maxGSSize    = 500,
  pvalueCutoff = 1,
  verbose      = FALSE)
>head(egoGSECC)
```

In R: solution

```
>gseaplot(egoGSECC , geneSetID="GO:0014069")
>dotplot(egoGSECC , showCategory=30)
>enrichMap(egoGSECC )
>plotGOgraph(egoGSECC )

# KEGG ONLY WORKS WITH ENTREZ ID
>rawp <- apply(rat, 1, ttestRat, grp1 = c(2:7), grp2 = c(8:12))
>names(rawp)           <-OUT$ENTREZ
>sortedrawp            <-sort(rawp)
>p_holm                <-p.adjust(sortedrawp,method="BH")
>names(p_holm)          <-names(sortedrawp)
>SCORE                  <-p_holm
>SCORE                  <-sort(SCORE, decreasing=TRUE)
>head(SCORE)

>kk2 <- gseKEGG(geneList      = SCORE,
                  organism       = 'rat',
                  nPerm         = 1000,
                  minGSSize     = 10,
                  pvalueCutoff  = 1,
                  verbose       = FALSE)
>head(kk2)
```

In R: solution

```
library(gtools)
fcRat <- function(df, grp1, grp2) {
  x = df[grp1]
  y = df[grp2]
  x = as.numeric(x)
  y = as.numeric(y)
  x = mean(x)
  y = mean(y)
  foldchange(x, y)
  rawp      <- apply(rat, 1, ttestRat, grp1 = c(2:7), grp2 = c(8:12))
  Fc        <- apply(rat, 1, fcRat   , grp1 = c(2:7), grp2 = c(8:12))

  resExp           <- data.frame(pValues=rawp, log2FC=Fc, name=OUT$ENSEMBL)
  topDEGenesDetails <- resExp[which(resExp[,1] < 0.01 & abs(resExp[,2])>2), ]
  topDEGenesFC     <- resExp[which(resExp[,1] < 0.01 & abs(resExp[,2])>2),3]

mydf <- data.frame(ENSEMBL=topDEGenesFC, FC=topDEGenesDetails[,2])
mydf$group
mydf$group[mydf$FC < 0] <- "upregulated"
mydf$group[mydf$FC > 0] <- "downregulated"

formula_res <- compareCluster(ENSEMBL~group,
                                data      = mydf,
                                fun       = "enrichGO",
                                keyType   = 'ENSEMBL',
                                OrgDb     = org.Rn.eg.db,
                                ont       = "ALL",
                                pAdjustMethod = "BH",
                                pvalueCutoff = 0.01,
                                qvalueCutoff = 0.05,
                                readable   = TRUE)
dotplot(formula_res, font.size=10)
```

Learning objectives

At the end of the course, the participants are expected to be able to:

1. identify statistical methods that could be used to pinpoint differentially expressed genes
2. determine whether a set of genes shows statistically significant differences between two classes
3. apply GSEA using R
4. distinguish available enrichment analysis methods
5. apply enrichment analysis implementations using R
6. do an Ab initio exploration of transcript data
7. determine whether the genes of a GO term have a statistically significant difference in expression.

Feedbacks
through course web-page

Thank you for your attention



Swiss Institute of
Bioinformatics

