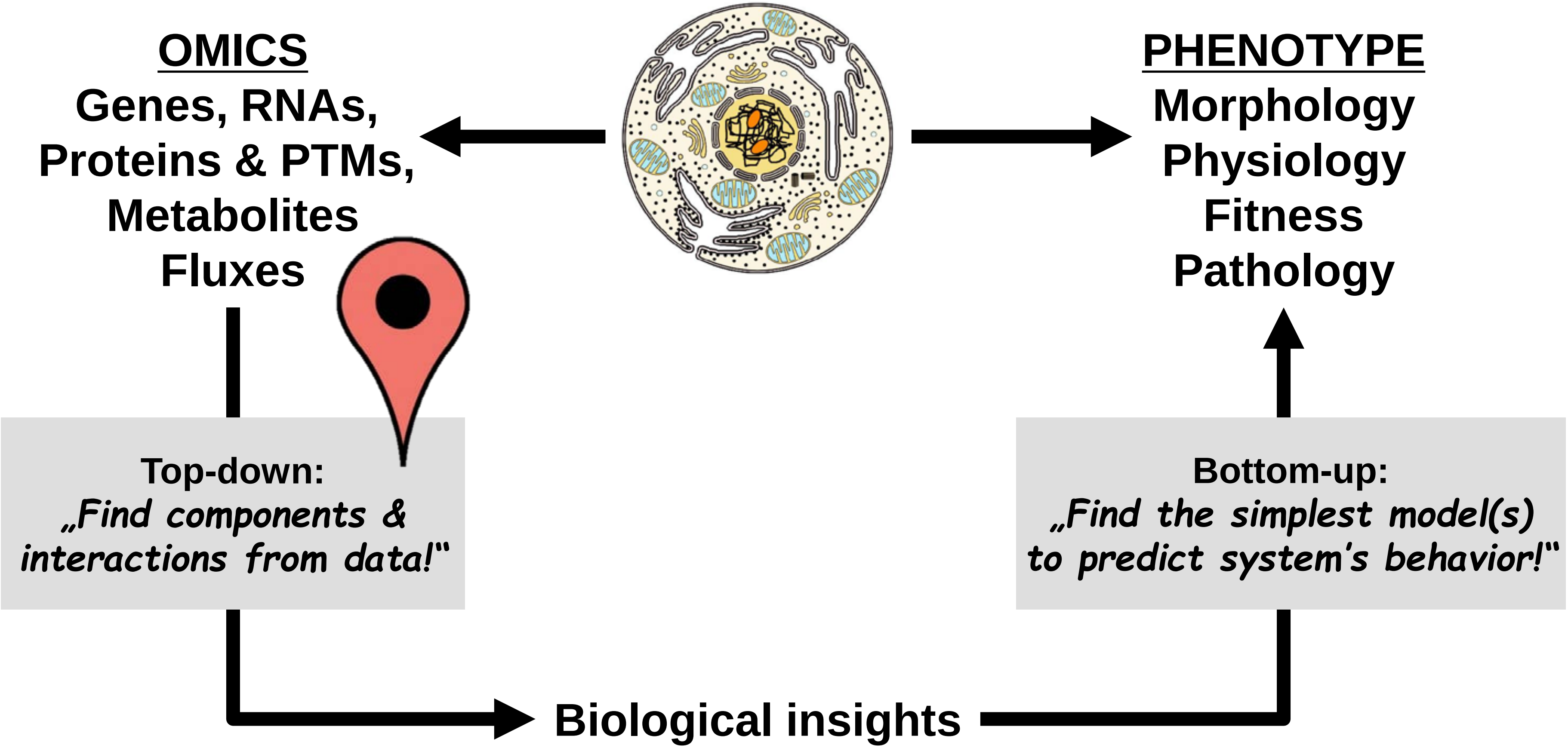
Learning Goals Lecture 8

- **Intro on data mining of omics data**
  - Know what biological insights are to be obtained by data mining of omics data
  - Describe the typical problems associated with omics data?
  - Describe the role of statistics in biological context
- **Differential, univariate analysis between two groups**
  - Understand the basic form of univariate analysis
  - Understand pathway Enrichment analysis
  - Correctly assess statistical significance

The big picture



OMICS

Genes, RNAs,
Proteins & PTMs,
Metabolites
Fluxes

PHENOTYPE

Morphology
Physiology
Fitness
Pathology

Top-down:
„Find components & interactions from data!“

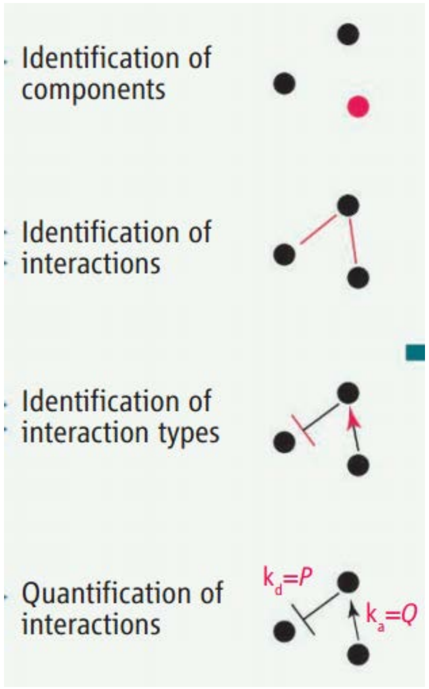
Bottom-up:
„Find the simplest model(s) to predict system's behavior!“

Biological insights

What are *biological insights*?

OMICS → Data mining →

1) Molecular level


Identification of components
Identification of interactions
Identification of interaction types
Quantification of interactions
k_d=P
k_a=Q

2) Phenotype

e.g. growth, disease, …

Exemplary BIO questions

COMPONENTS ID

- What is the composition of a cell?
- What is the function of a gene / gene product?
- What are the variants of a protein?

NETWORK RECONSTRUCTION (LINK PREDICTION)

- What cellular regulators are active?
- Infer a transcriptional network

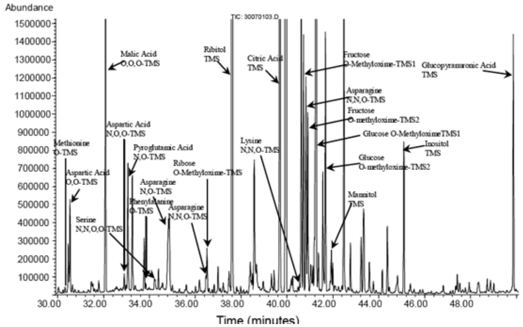
CLASSIFICATION, FEATURE SELECTION

- Group cells/organs/patients based on molecular features (markers)
- Describe molecular response of cells to treatments
- Predict behavior of cells to treatments

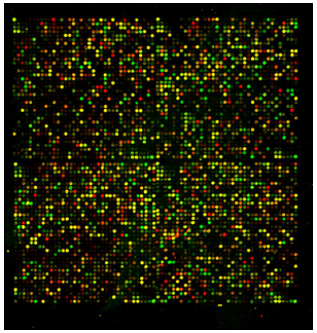
…

OMICS data are large

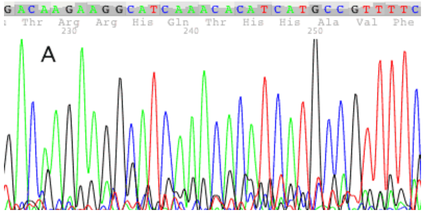
It doesn’t have to be omics data all the time. However, with the current state of the art in technology, it’s fairly <u>common</u> to obtain data on



- **100s** of metabolites/lipids
  - Mass spectrometry
- **1’000s** of proteins
  - Mass spectrometry or antibodies
- **1’000s – 10’000s** genes and RNAs
  - Microarrays and NextGen Sequencing



… **for <u>EACH</u> tested condition.**



Big data, low sample number

Most frequent problems in analyzing OMICS data:

1) **# of detected features >>> # of samples**

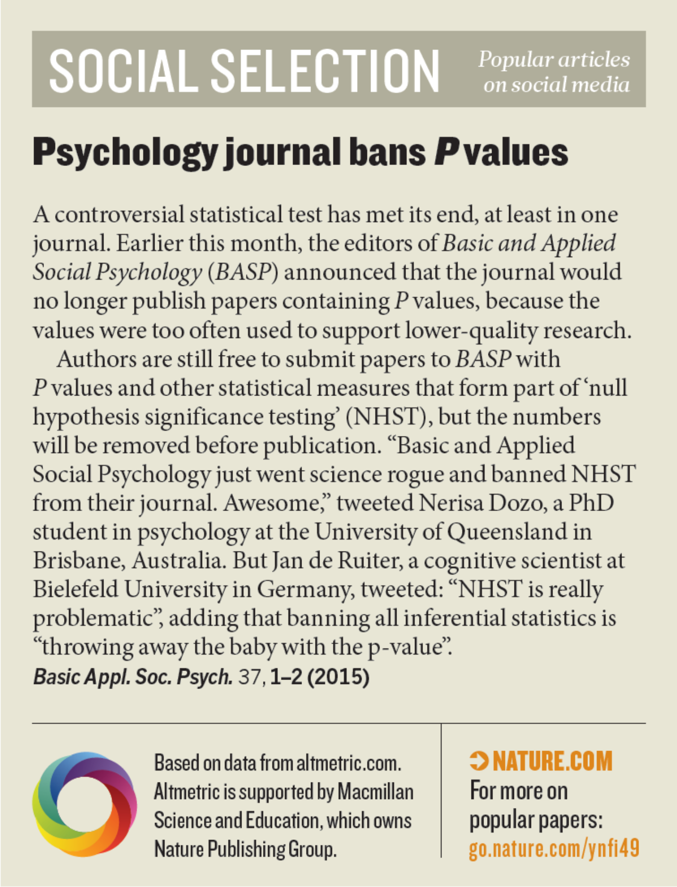
2) Data are noisy

3) Technical reproducibility of experiments

There is a serious risk of **overfitting** the data, i.e. to identify patterns that capture NOISE of the data and are hardly reproducible. Such models are excessively complicated and lead to poor predictions on new data sets.

Good statistics is a must but not sufficient

Opportunity and quality of statistical tests are under debate…



SOCIAL SELECTION *Popular articles on social media*

Psychology journal bans *P* values

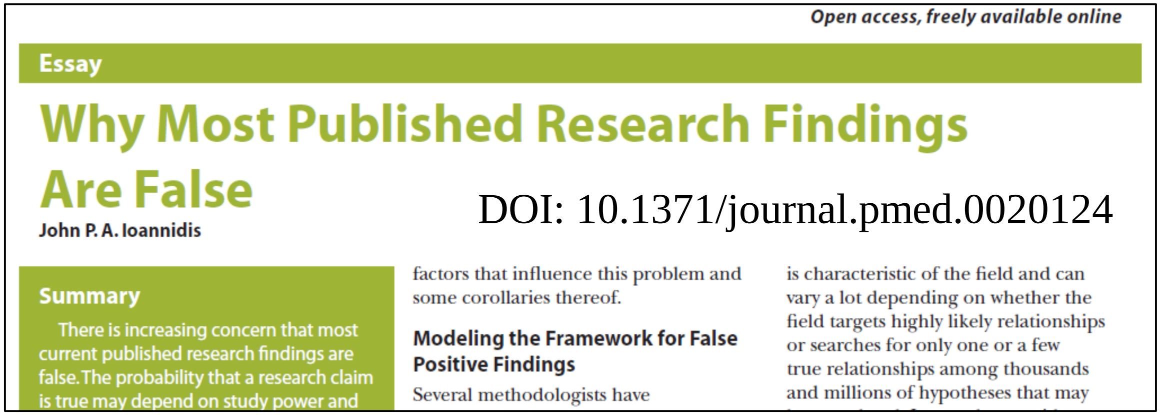
A controversial statistical test has met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values, because the values were too often used to support lower-quality research.

Authors are still free to submit papers to *BASP* with *P* values and other statistical measures that form part of 'null hypothesis significance testing' (NHST), but the numbers will be removed before publication. "Basic and Applied Social Psychology just went science rogue and banned NHST from their journal. Awesome," tweeted Nerisa Dozo, a PhD student in psychology at the University of Queensland in Brisbane, Australia. But Jan de Ruiter, a cognitive scientist at Bielefeld University in Germany, tweeted: "NHST is really problematic", adding that banning all inferential statistics is "throwing away the baby with the p-value".

Basic Appl. Soc. Psych. 37, **1–2 (2015)**

Based on data from altmetric.com. Altmetric is supported by Macmillan Science and Education, which owns Nature Publishing Group.

NATURE.COM
For more on popular papers:
go.nature.com/ynfi49



Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

DOI: 10.1371/journal.pmed.0020124

Summary

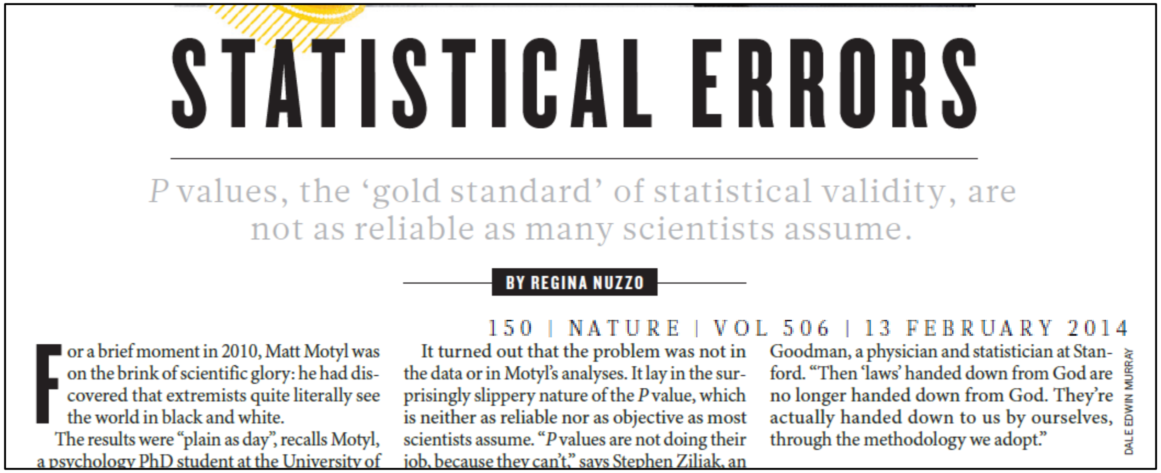
There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may



STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

150 | NATURE | VOL 506 | 13 FEBRUARY 2014

For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white.

The results were "plain as day", recalls Motyl, a psychology PhD student at the University of

It turned out that the problem was not in the data or in Motyl's analyses. It lay in the surprisingly slippery nature of the *P* value, which is neither as reliable nor as objective as most scientists assume. "*P* values are not doing their job, because they can't," says Stephen Ziliak, an

Goodman, a physician and statistician at Stanford. "Then 'laws' handed down from God are no longer handed down from God. They're actually handed down to us by ourselves, through the methodology we adopt."

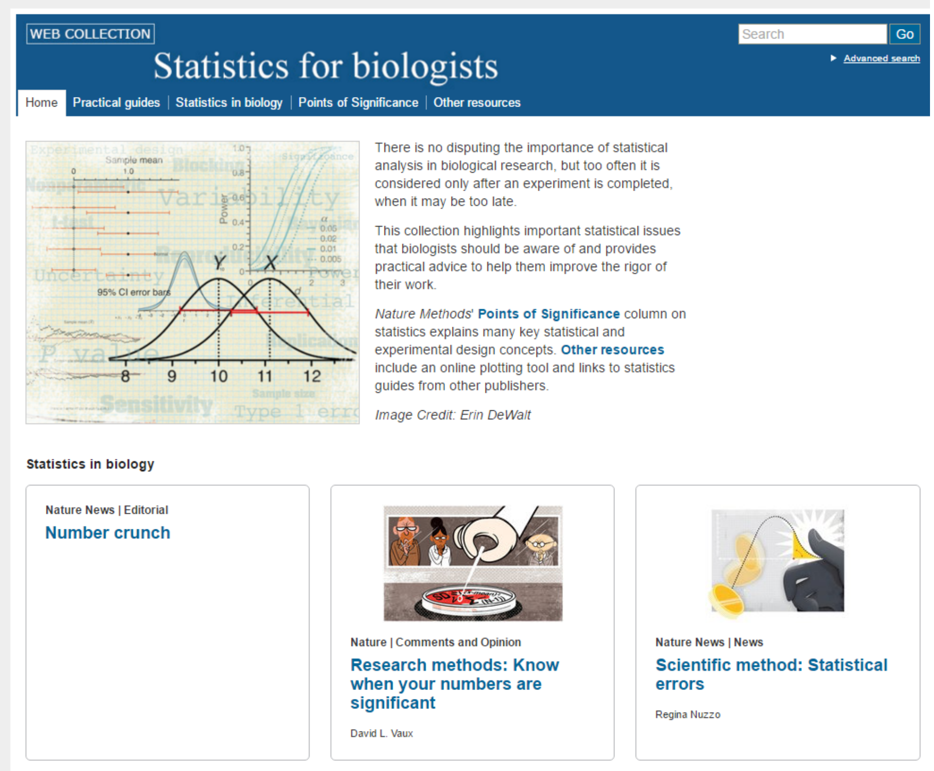
DALE EDWIN MURRAY

Great resource on stats: http://www.nature.com/collections/qghhqm/

Recommended resource

http://www.nature.com/collections/qghhqm/

- Practical guides
- Primers for biologists



WEB COLLECTION

Statistics for biologists

Search
Go
Advanced search
Home | Practical guides | Statistics in biology | Points of Significance | Other resources

There is no disputing the importance of statistical analysis in biological research, but too often it is considered only after an experiment is completed, when it may be too late.

This collection highlights important statistical issues that biologists should be aware of and provides practical advice to help them improve the rigor of their work.

Nature Methods' **Points of Significance** column on statistics explains many key statistical and experimental design concepts. **Other resources** include an online plotting tool and links to statistics guides from other publishers.

Image Credit: Erin DeWalt

Statistics in biology

Nature News | Editorial

Number crunch

Nature | Comments and Opinion

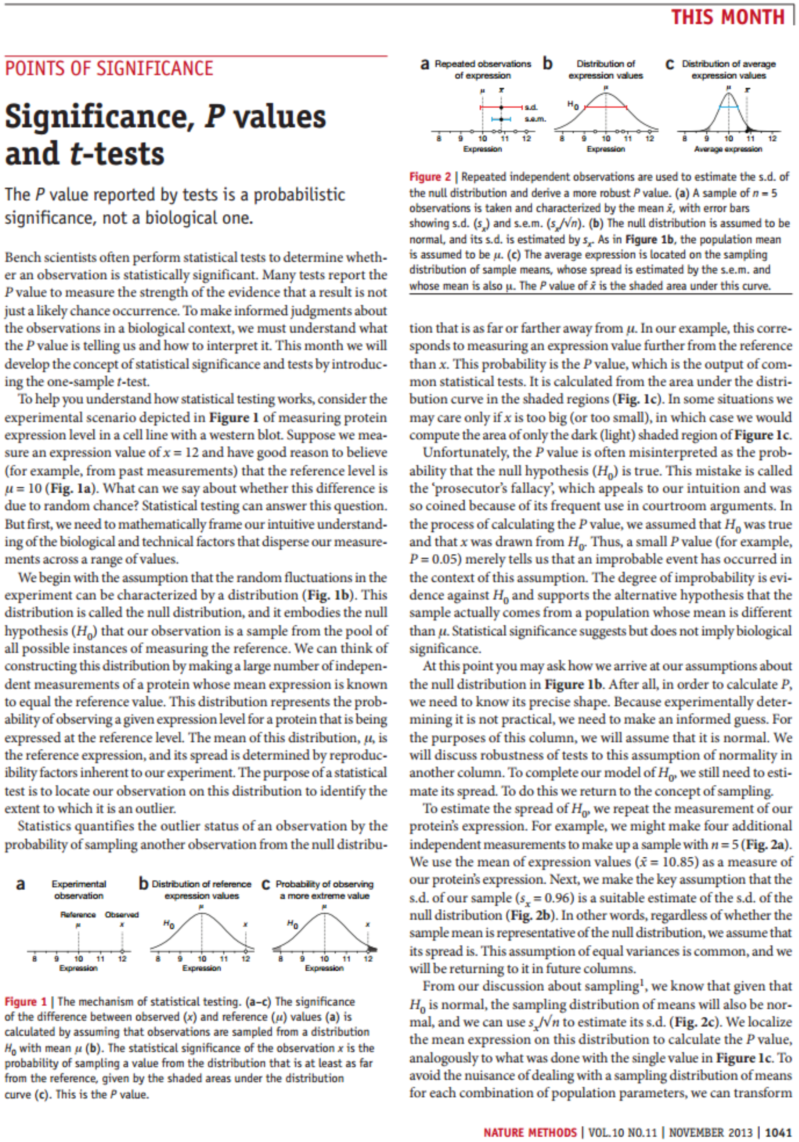
Research methods: Know when your numbers are significant

David L. Vaux

Nature News | News

Scientific method: Statistical errors

Regina Nuzzo



THIS MONTH

POINTS OF SIGNIFICANCE

Significance, *P* values and *t*-tests

The *P* value reported by tests is a probabilistic significance, not a biological one.

Bench scientists often perform statistical tests to determine whether an observation is statistically significant. Many tests report the *P* value to measure the strength of the evidence that a result is not just a likely chance occurrence. To make informed judgments about the observations in a biological context, we must understand what the *P* value is telling us and how to interpret it. This month we will develop the concept of statistical significance and tests by introducing the one-sample *t*-test.

To help you understand how statistical testing works, consider the experimental scenario depicted in **Figure 1** of measuring protein expression level in a cell line with a western blot. Suppose we measure an expression value of $x = 12$ and have good reason to believe (for example, from past measurements) that the reference level is $\mu = 10$ (**Fig. 1a**). What can we say about whether this difference is due to random chance? Statistical testing can answer this question. But first, we need to mathematically frame our intuitive understanding of the biological and technical factors that disperse our measurements across a range of values.

We begin with the assumption that the random fluctuations in the experiment can be characterized by a distribution (**Fig. 1b**). This distribution is called the null distribution, and it embodies the null hypothesis (H_0) that our observation is a sample from the pool of all possible instances of measuring the reference. We can think of constructing this distribution by making a large number of independent measurements of a protein whose mean expression is known to equal the reference value. This distribution represents the probability of observing a given expression level for a protein that is being expressed at the reference level. The mean of this distribution, μ, is the reference expression, and its spread is determined by reproducibility factors inherent to our experiment. The purpose of a statistical test is to locate our observation on this distribution to identify the extent to which it is an outlier.

Statistics quantifies the outlier status of an observation by the probability of sampling another observation from the null distribution that is as far or farther away from μ. In our example, this corresponds to measuring an expression value further from the reference than x. This probability is the *P* value, which is the output of common statistical tests. It is calculated from the area under the distribution curve in the shaded regions (**Fig. 1c**). In some situations we may care only if x is too big (or too small), in which case we would compute the area of only the dark (light) shaded region of **Figure 1c**.

a Experimental observation
Reference μ
Observed x
b Distribution of reference expression values
c Probability of observing a more extreme value
H0
Expression
8 9 10 11 12

Figure 1 | The mechanism of statistical testing. (**a–c**) The significance of the difference between observed (x) and reference (μ) values (**a**) is calculated by assuming that observations are sampled from a distribution H_0 with mean μ (**b**). The statistical significance of the observation x is the probability of sampling a value from the distribution that is at least as far from the reference, given by the shaded areas under the distribution curve (**c**). This is the *P* value.

a Repeated observations of expression
b Distribution of expression values
c Distribution of average expression values
s.d.
s.e.m.
H0
Expression
Average expression
8 9 10 11 12

Figure 2 | Repeated independent observations are used to estimate the s.d. of the null distribution and derive a more robust *P* value. (**a**) A sample of $n = 5$ observations is taken and characterized by the mean $\bar{x}$, with error bars showing s.d. (s_x) and s.e.m. ($s_x/\sqrt{n}$). (**b**) The null distribution is assumed to be normal, and its s.d. is estimated by s_x. As in **Figure 1b**, the population mean is assumed to be μ. (**c**) The average expression is located on the sampling distribution of sample means, whose spread is estimated by the s.e.m. and whose mean is also µ. The *P* value of $\bar{x}$ is the shaded area under this curve.

Unfortunately, the *P* value is often misinterpreted as the probability that the null hypothesis (H_0) is true. This mistake is called the 'prosecutor's fallacy', which appeals to our intuition and was so coined because of its frequent use in courtroom arguments. In the process of calculating the *P* value, we assumed that H_0 was true and that x was drawn from H_0. Thus, a small *P* value (for example, $P = 0.05$) merely tells us that an improbable event has occurred in the context of this assumption. The degree of improbability is evidence against H_0 and supports the alternative hypothesis that the sample actually comes from a population whose mean is different than μ. Statistical significance suggests but does not imply biological significance.

At this point you may ask how we arrive at our assumptions about the null distribution in **Figure 1b**. After all, in order to calculate P, we need to know its precise shape. Because experimentally determining it is not practical, we need to make an informed guess. For the purposes of this column, we will assume that it is normal. We will discuss robustness of tests to this assumption of normality in another column. To complete our model of H_0, we still need to estimate its spread. To do this we return to the concept of sampling.

To estimate the spread of H_0, we repeat the measurement of our protein's expression. For example, we might make four additional independent measurements to make up a sample with $n = 5$ (**Fig. 2a**). We use the mean of expression values ($\bar{x} = 10.85$) as a measure of our protein's expression. Next, we make the key assumption that the s.d. of our sample ($s_x = 0.96$) is a suitable estimate of the s.d. of the null distribution (**Fig. 2b**). In other words, regardless of whether the sample mean is representative of the null distribution, we assume that its spread is. This assumption of equal variances is common, and we will be returning to it in future columns.

From our discussion about sampling[1], we know that given that H_0 is normal, the sampling distribution of means will also be normal, and we can use $s_x/\sqrt{n}$ to estimate its s.d. (**Fig. 2c**). We localize the mean expression on this distribution to calculate the *P* value, analogously to what was done with the single value in **Figure 1c**. To avoid the nuisance of dealing with a sampling distribution of means for each combination of population parameters, we can transform

NATURE METHODS | VOL.10 NO.11 | NOVEMBER 2013 | 1041

The biologist’s view on statistics

- **Research won’t proceed if solely based on absolutely proven facts.** It’s ok to allow for uncertain hypotheses as long as we are honest about that (to us and others).
  - Depending on the study, it might be more important to minimize false positives before maximizing true positives. This is particularly true if the follow up experiments are very tedious/complex.

- **Biologists use data mining to generate hypotheses.** Some of these are for sure wrong.
  - Hypotheses are a start for research, not an endpoint.
  - Hypotheses are tested/proven with ad-hoc follow-up experiments.

- Good stats are an indicator of good lab work (reproducibility).

Best practice in (bio) data mining

- **Start from a biological question!**
  - Explorative analyses are difficult and, typically, lead nowhere.

Nice reading: 10.1126/science.aaa6146

- Don’t expect an answer at all costs!
  Because of lack of data, noise, lack of response, wrong question

- Always think about **positive and negative controls**

- It’s a good habit to (try to) keep data mining **simple**!

- Use different techniques and parameters

Self check

- What do we expect to obtain from omics data?
- What are the typical problems associated with omics data?
- Why do we need computers?
- What is the role of statistics?

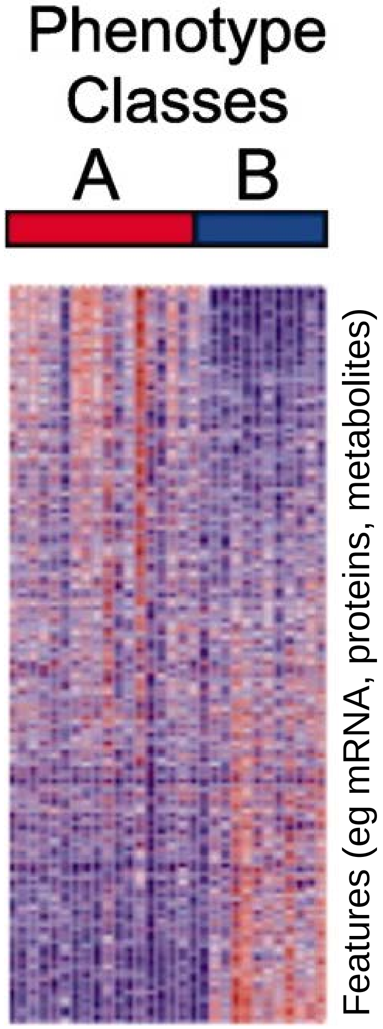
Two groups problem

Samples are divided a priori in two groups/classes.
Typical examples:

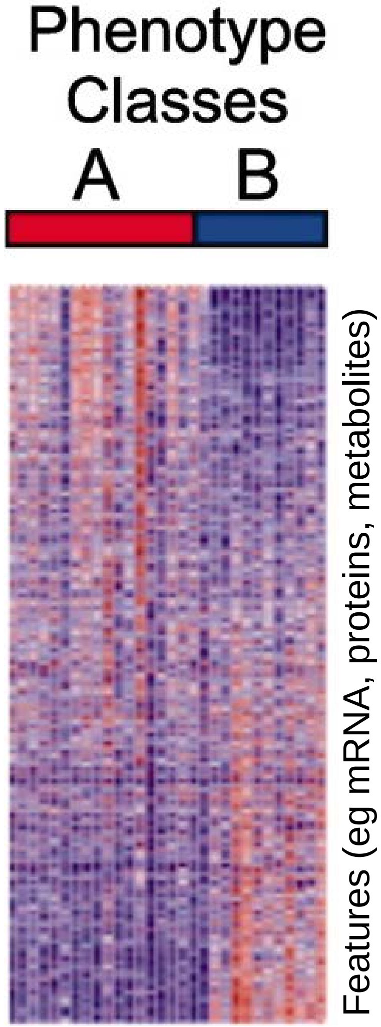
- mutant vs. wild-type
- disease vs. health
- diet A vs. diet B
- treated vs. untreated (by a drug)
- resistant vs. sensitive (to a drug)
- old vs. young
- …

BIO questions:

- **Identify significantly changed features**
- **Identify significantly changed cellular processes**



How would you identify significantly changed features between A and B?



Univariate analysis

The simplest approach is to use univariate analysis as we know it from low-dimensional data.

That is: consider 1 feature at the time and compare the levels in the two groups. The test is repeated independently for all features.

Typical approach:

- Calculate magnitude of change > fold-change between group means or medians
- Verify statistical significance > p-value for null hypothesis

Calculate magnitude of change

Fold-change:

$$FC = \frac{mean(GroupA)}{mean(GroupB)}$$

[0…1…+∞]

For better (symmetric) visualization:

$$log2FC = log2\left(\frac{mean(GroupA)}{mean(GroupB)}\right)$$

[-∞…0…+∞]

Verify statistical significance

Statistical hypothesis testing is often used to the likelihood to obtain the measured data given the hypothesis that the two groups originate from the same distribution = they are NOT different (= null hypothesis).

Common variants:

- **T-test** (usually using the two-tailed, unequal variance form)
  assumes normal distribution
- **Mann-Whitney/Wilcoxon/Ranksum test**
  uses ranking (no normality/distribution assumed)
- **Permutation test**
  For studies with many samples, it counts how frequently the result is better to that obtained with randomized sample labels.

Such tests produce a p-value. It indicates that under the assumption that the null hypothesis is true (= no difference), there are [p-value] chances to obtain the same results. The p-value is NOT the probability that the null hypothesis is true.

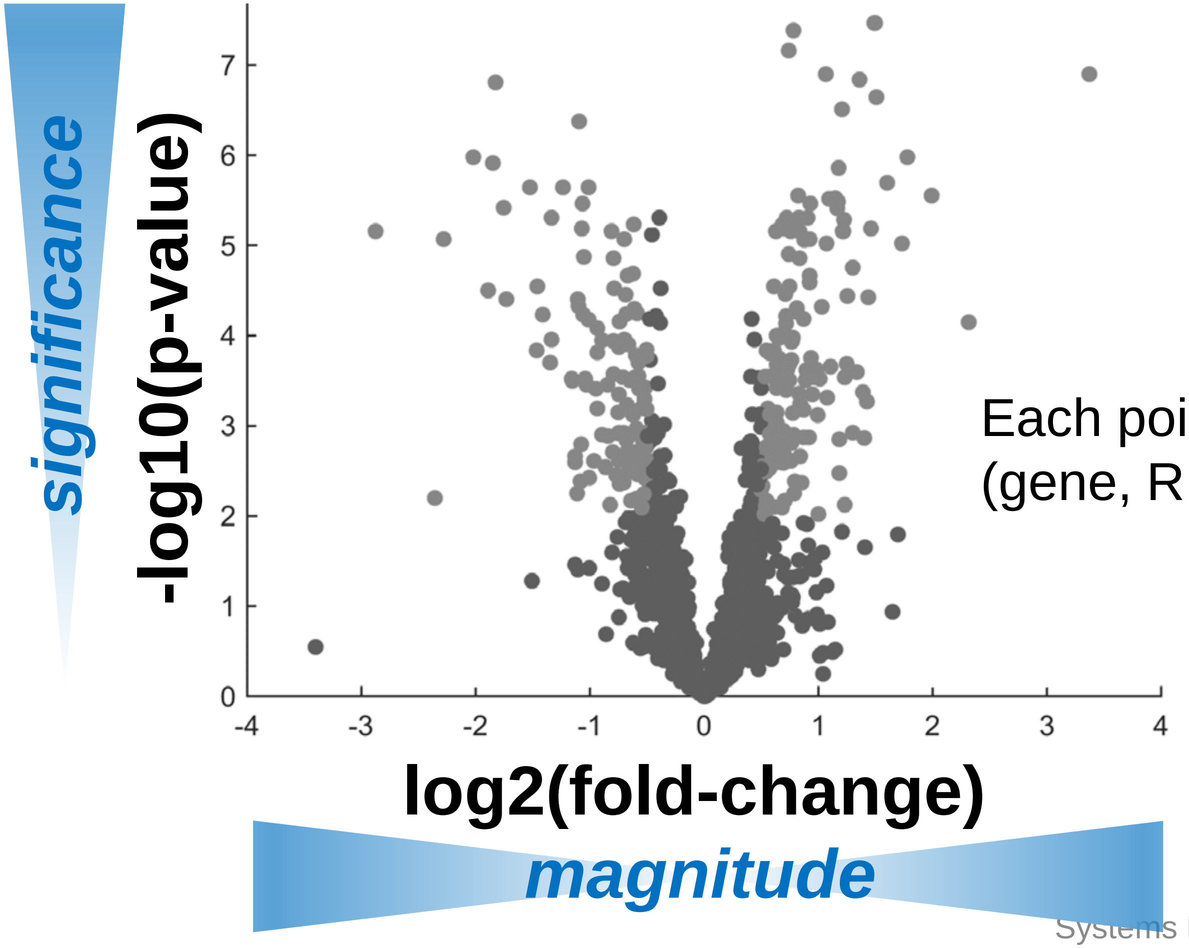
Traditionally in biology:

p-value < 0.05 : acceptable!

p-value < 0.01 : nice to have!

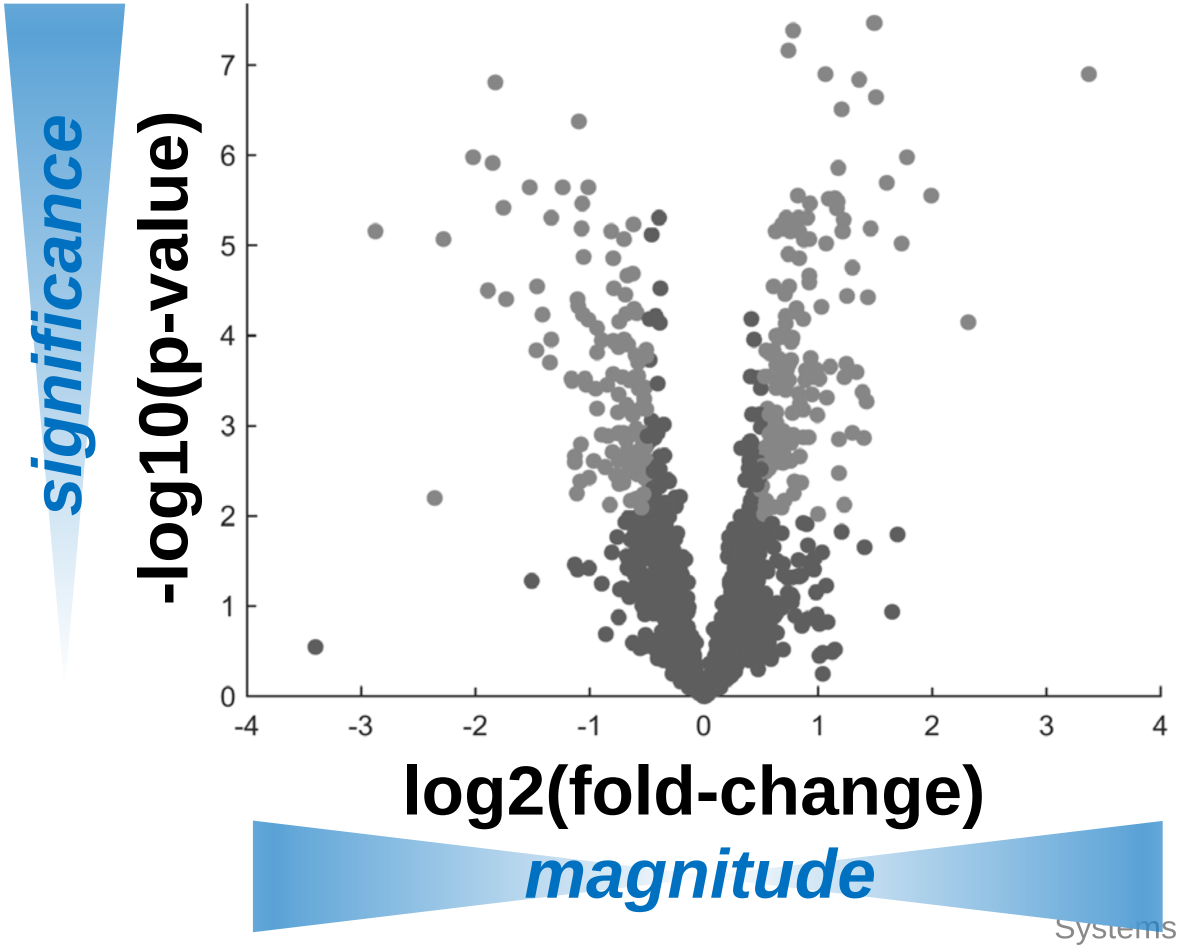
Volcano plot

It’s a common visualization method to evaluate the FC and statistical significance of a two-group comparison



Each point is one feature (gene, RNA, metabolite, …)

How to choose thresholds?



Fold Change threshold

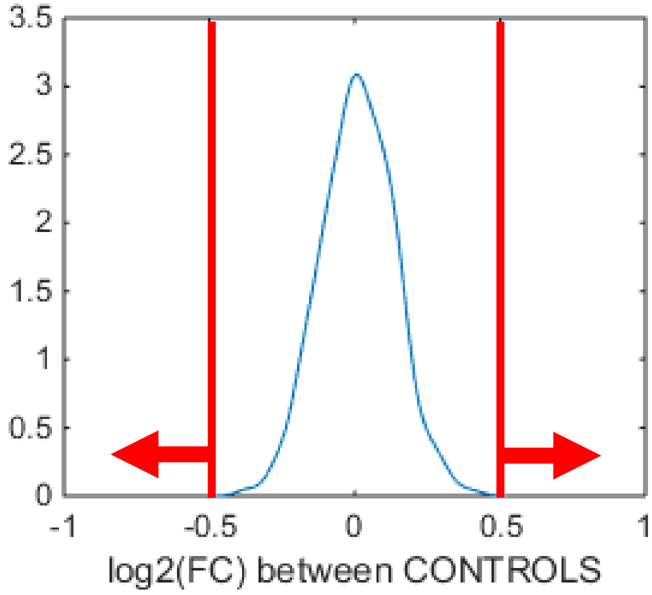
Pick an arbitrary value: given e.g. overall reproducibility, object, sample size (example: |log2FC| > 1 in transcriptomics)

Empirical procedure: choose value such that things that are supposed to be identical are not called significantly different.

Example of negative controls:

- all wild-types
- all healthy samples
- all non-treated samples

Thresholds can be determined by estimating the distribution of fold-changes between random (all) subsets of negative controls.

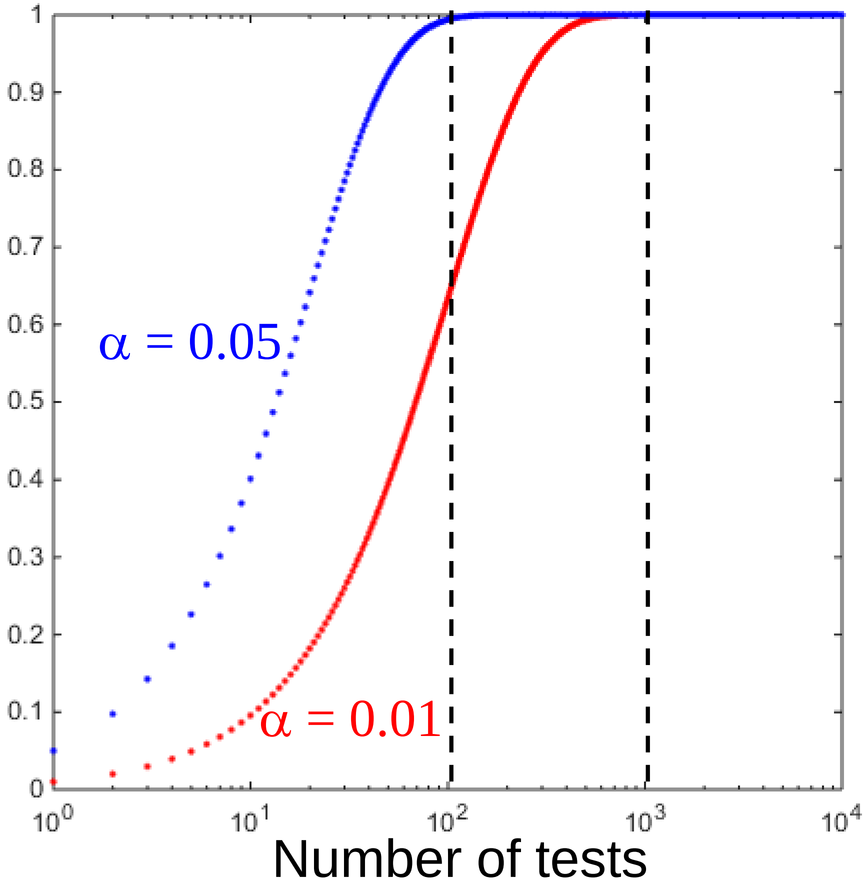


p-value threshold

Multiple testing problem: Choosing a significance cutoff α = 0.05 means that in a single test there is a 5% chance that the result is a false positive. This is ok if we do only 1 test, but…

Probability of making at least 1 error

$1 - (1-\alpha)^m$

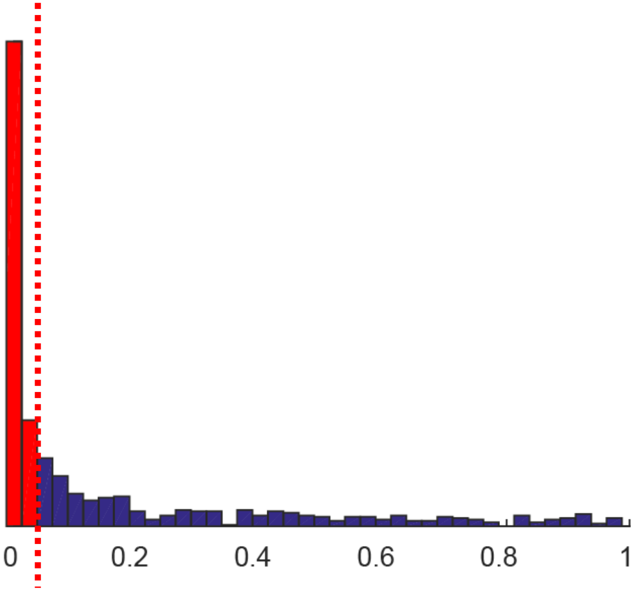


p-value threshold

Example: 821 features, significance cutoff < 0.05

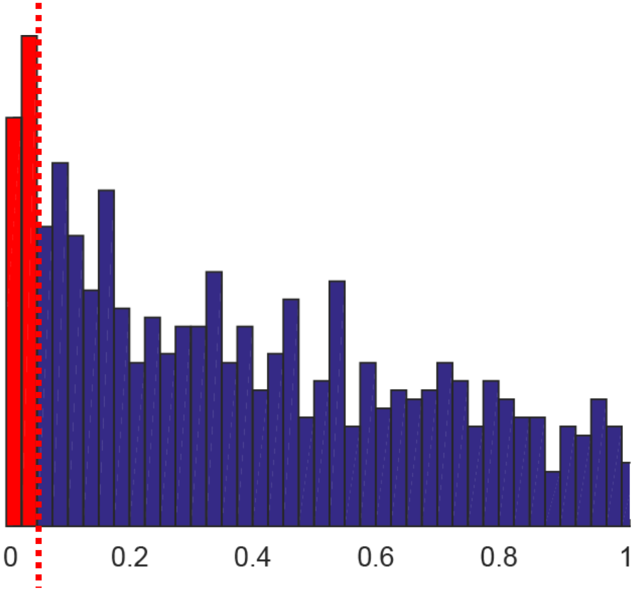
(t-test)
p-value
distribution

Case 1: Group A vs. B



56% significant

Case 2: Group C vs. C



14% significant

Error types (more in future…)

False negatives

“TRUTH”

“DECISION”

	H_0 true	H_0 false	Total
Do not reject H_0	Correct U $1-\alpha$	**Type II Error** T β	m-R
Reject H_0	**Type I Error** V α	Correct S $1-\beta$	R
	m_0	$m-m_0$	m

False positives

False Discovery Rate (FDR) = V/R

False Positive Rate (FPR) = V/m_0

Correcting for multiple testing

Control Family-Wise Error Rate

Guard against ANY false positives FWER = P(V ≥ 1).

Bonferroni: adjust p-value by number of tests ***m***: reject H_i if $p_i \leq \frac{\alpha}{m}$ very strict!!

Less stringent, sequential procedures exist (e.g. Holm), but don't change the issue that in biology we don't really want to control for FWER. In discovery, we can live with some false positives. Hence, all FWER tend to be overly restrictive.

Better approaches:

Control False Discovery Rate (**Benjamini and Hochberg** 1995)

Controls the proportion of false positives among the set of rejected hypotheses

p-values > **FDR-adjusted p-values**, reject H_i if $adj.p_i \leq \alpha$

OK if differences are rare ($pi_0 \approx 1$)

Control positive False Discovery Rate (**Storey and Tibshirani** 2003)

Controls the rate that discoveries are false (pFDR). Uses all p-values to derive ***q*-values** (minimum FDR that can be attained when calling that feature significant)

OK if differences are frequent ($pi_0 < 1$)

(example: if gene X has q = 0.02 it means that 2% of genes that show p-values at least as small as gene X are false positives).

[More info: Noble, *How does multiple testing correction work?* Nature Biotech 2009]

The effect of MT correction

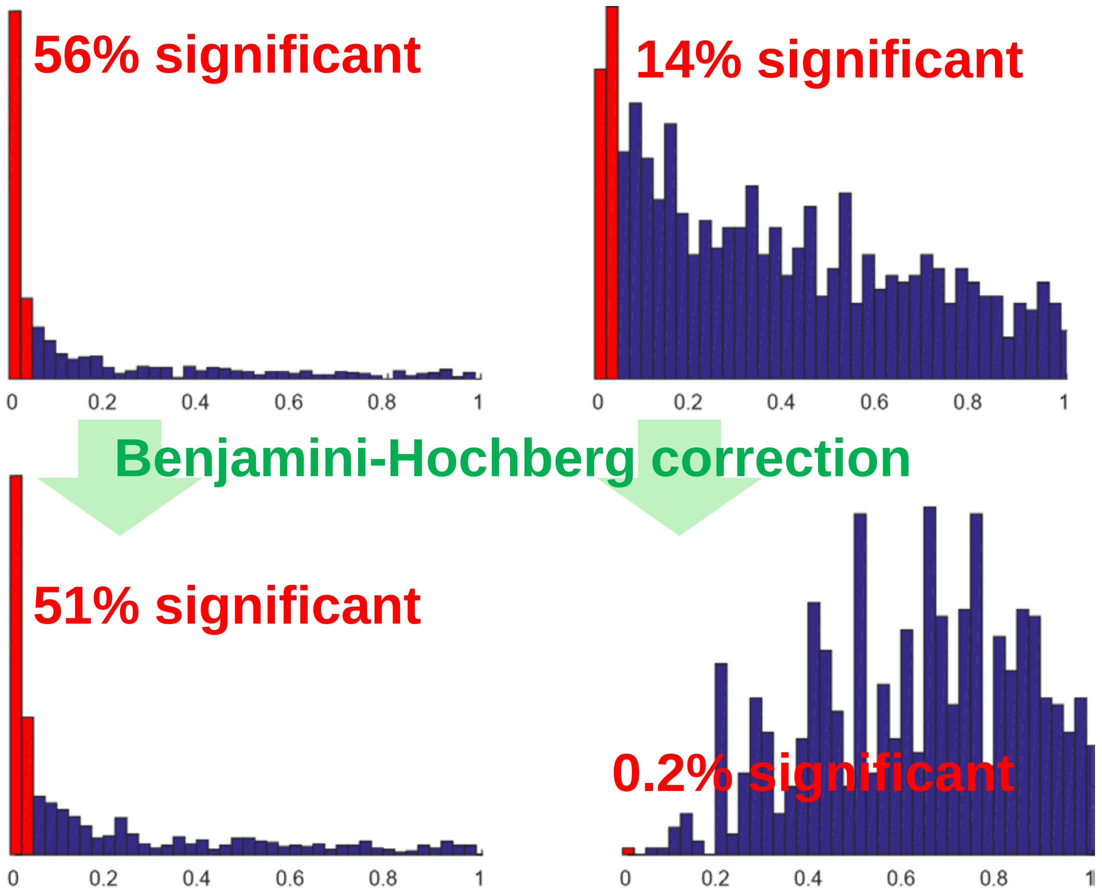
Example: 821 features, significance < 0.05

Group A vs. B

Group C vs. C

(t-test)
p-value
distribution

adjusted p-value
distribution



56% significant

14% significant

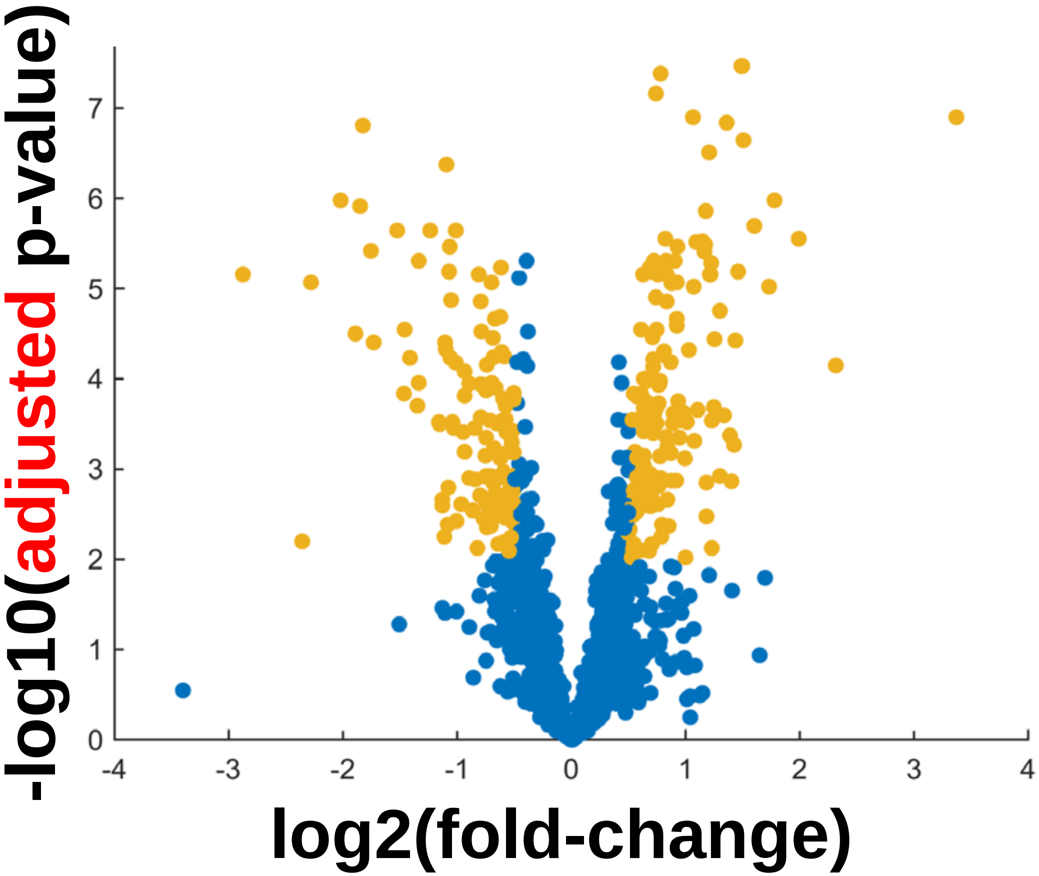
Benjamini-Hochberg correction

51% significant

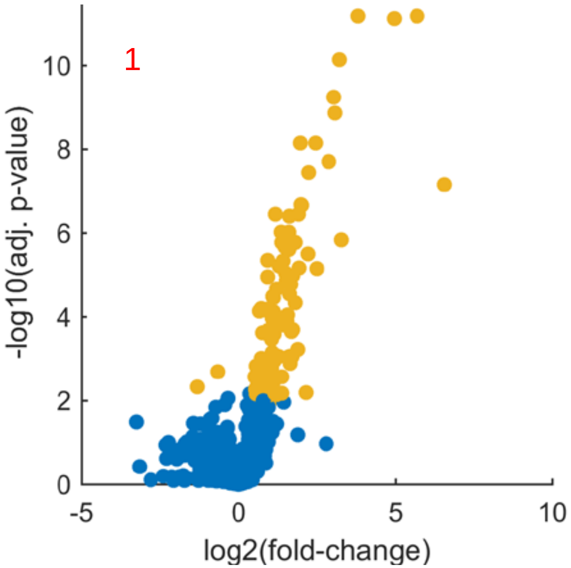
0.2% significant

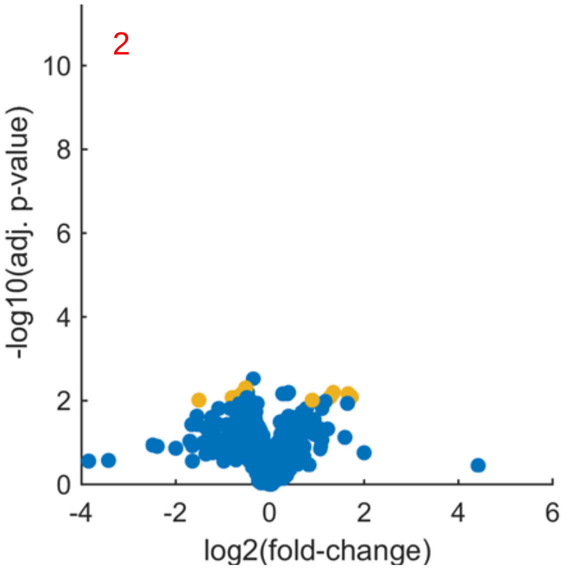
Volcano plot

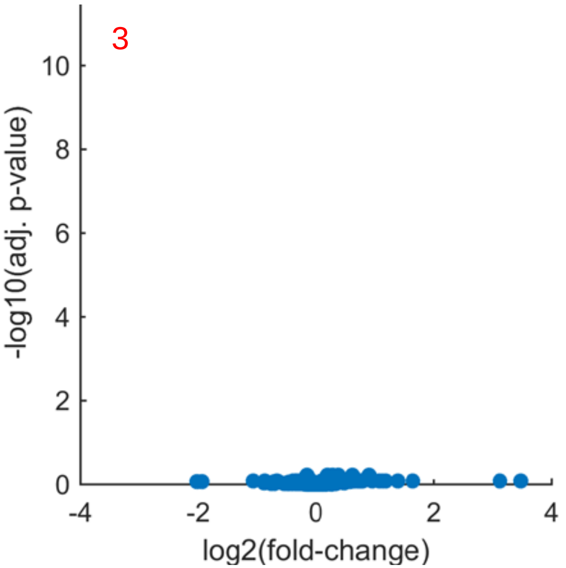
It’s a common visualization method to evaluate the FC and statistical significance of a two-group comparison



What do we conclude from these volcano plots?







What do we learn from that?

Zero significant changes 3	Check data quality Low p-values, high FC > increase number of replicates
Very **few** significant changes or very few “strong” changes 2	Simplest case > What is the identity of the markers?
A lot of significant changes No clear ranking 1	Quite common, but hard to generate hypotheses > Too many hits > Combination of primary and/or secondary effects > **Enrichment analysis!**

Enrichment analysis

*Provided that **many** significantly changed features were found (between two groups),*

how would you identify significantly changed cellular processes?

Cellular ”Processes”

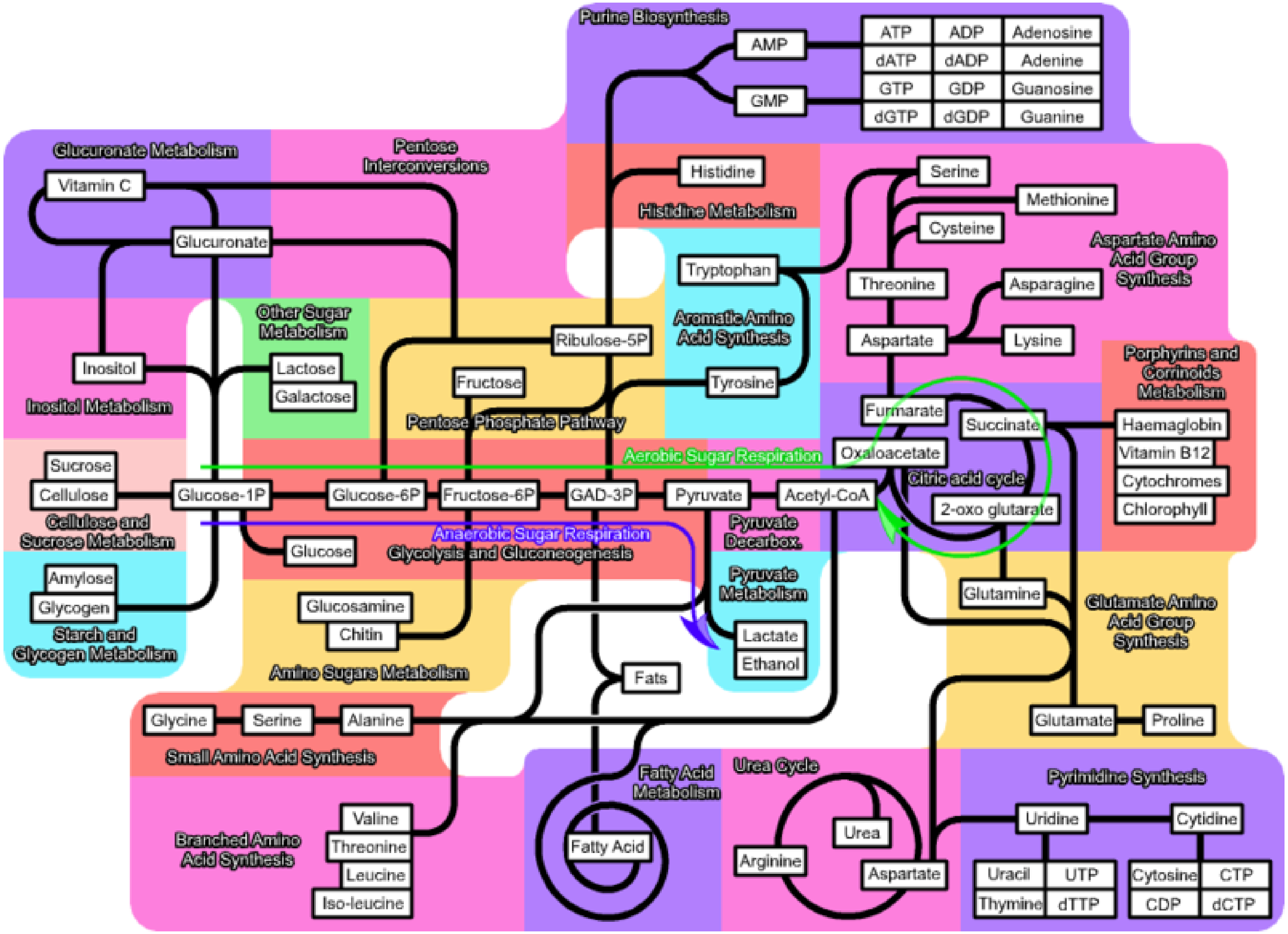
In a (systems-) biological context, we frequently identify processes with the molecular features that we know to be involved.

Examples:

- Enzymes or metabolite belonging to a pathway
- Proteins belonging to a complex (eg Ribosome)
- Genes controlled by a transcription factor
- GO terms
- Molecular Signatures Database (GSEA)

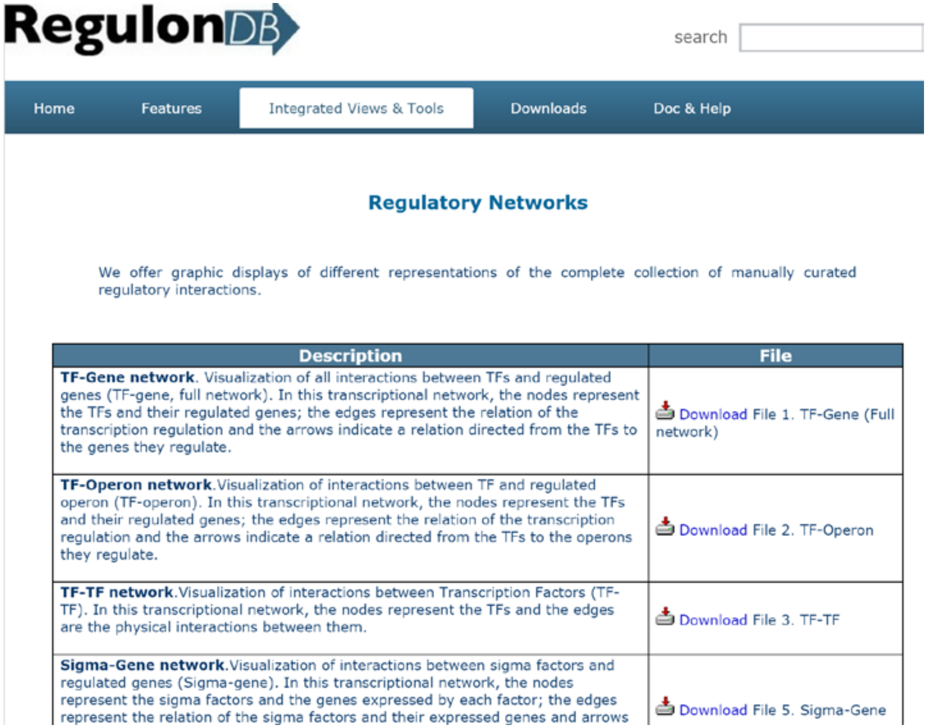
Important: A feature can belong to several groups/processes

Biochemical pathways



Transcriptional Network

Example: RegulonDB *(E. coli)*

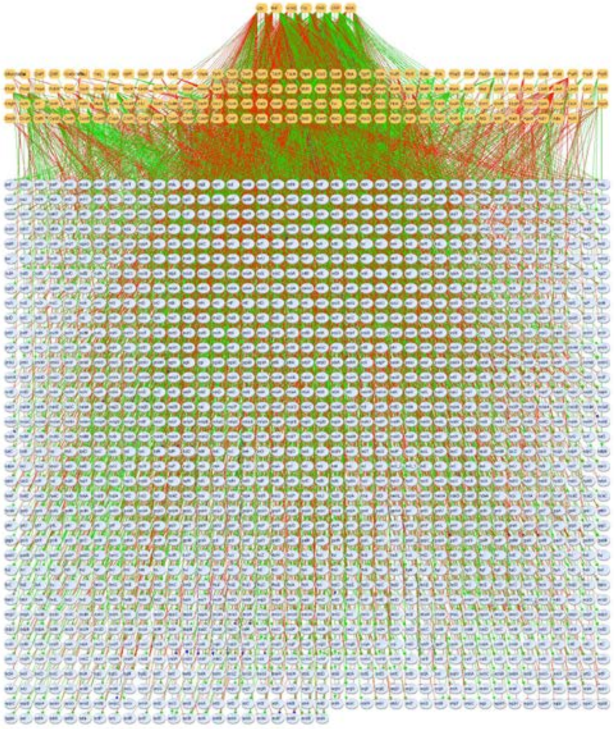


Regulatory Networks

We offer graphic displays of different representations of the complete collection of manually curated regulatory interactions.

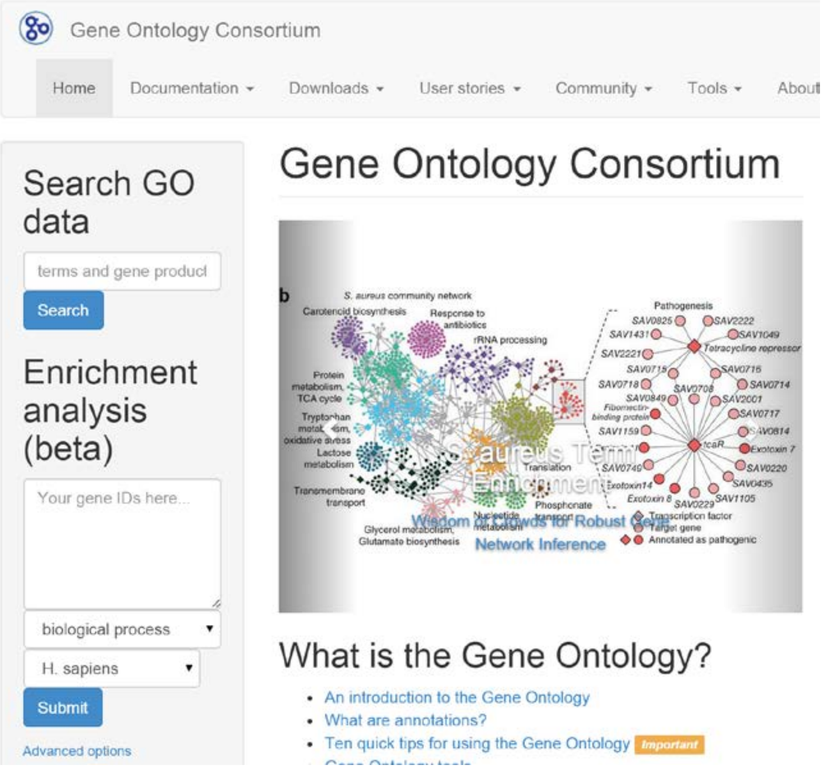
Description	File
TF-Gene network. Visualization of all interactions between TFs and regulated genes (TF-gene, full network). In this transcriptional network, the nodes represent the TFs and their regulated genes; the edges represent the relation of the transcription regulation and the arrows indicate a relation directed from the TFs to the genes they regulate.	Download File 1. TF-Gene (Full network)
TF-Operon network.Visualization of interactions between TF and regulated operon (TF-operon). In this transcriptional network, the nodes represent the TFs and their regulated genes; the edges represent the relation of the transcription regulation and the arrows indicate a relation directed from the TFs to the operons they regulate.	Download File 2. TF-Operon
TF-TF network.Visualization of interactions between Transcription Factors (TF-TF). In this transcriptional network, the nodes represent the TFs and the edges are the physical interactions between them.	Download File 3. TF-TF
Sigma-Gene network.Visualization of interactions between sigma factors and regulated genes (Sigma-gene). In this transcriptional network, the nodes represent the sigma factors and the genes expressed by each factor; the edges represent the relation of the sigma factors and their expressed genes and arrows	Download File 5. Sigma-Gene

E. coli TF-Gene Network



Gene Ontology & Collections

Reference of controlled vocabularies and annotations representing gene product functions (10’000s entries, hierarchical).



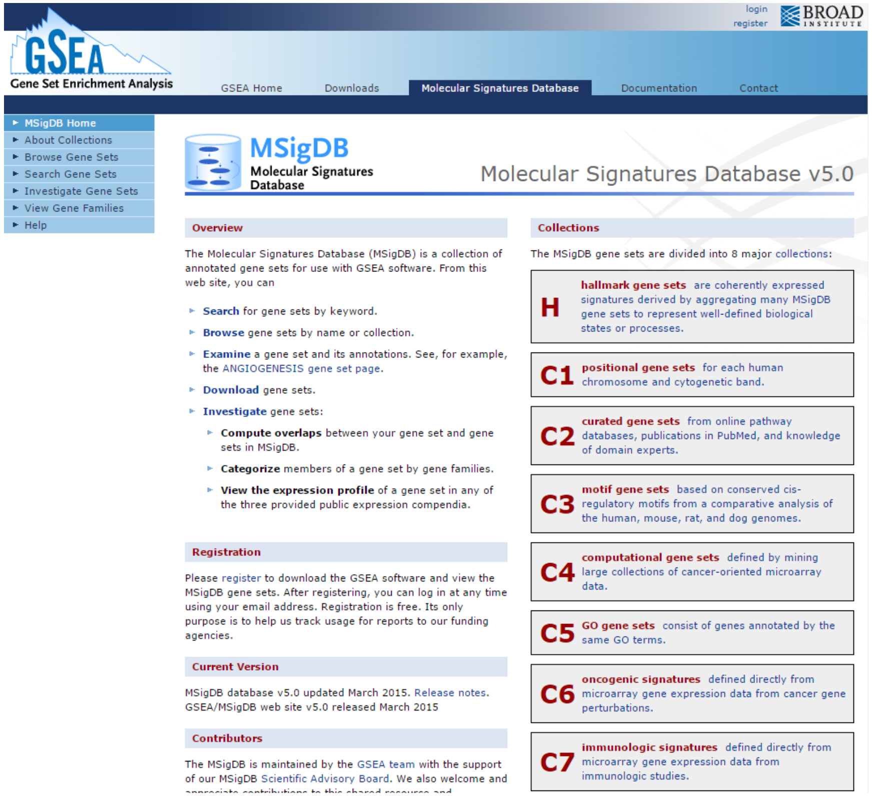
Search GO data

Enrichment analysis (beta)

Gene Ontology Consortium

What is the Gene Ontology?

- An introduction to the Gene Ontology
- What are annotations?
- Ten quick tips for using the Gene Ontology Important
- Gene Ontology tools



Molecular Signatures Database v5.0

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- **Search** for gene sets by keyword.
- **Browse** gene sets by name or collection.
- **Examine** a gene set and its annotations. See, for example, the ANGIOGENESIS gene set page.
- **Download** gene sets.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in any of the three provided public expression compendia.

Registration

Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

MSigDB database v5.0 updated March 2015. Release notes.
GSEA/MSigDB web site v5.0 released March 2015

Contributors

The MSigDB is maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. We also welcome and appreciate contributions to this shared resource and

Collections

The MSigDB gene sets are divided into 8 major collections:

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **GO gene sets** consist of genes annotated by the same GO terms.

C6 **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.

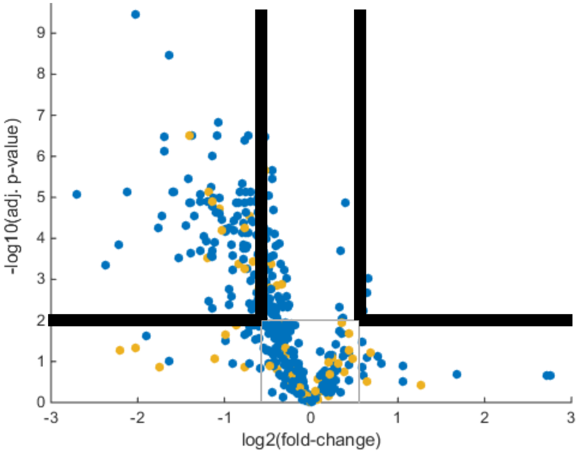
C7 **immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

Enrichment analysis (EA)

We want to test whether in the set of significantly changed features (i.e. genes, proteins, metabolites), the members of a specific group/process/pathway are over-represented.

Volcano Plot > significance threshold > count > **Contingency table:**

From definitions (GO term)



	In group (G+)🟡	**Not in group (G-) 🔵**
Significant (S+)	A	B
Not significant (S-)	C	D

[Fisher’s Exact Test]

	In group (G+)	Not in group (G-)
Significant (S+)	A	B
Not significant (S-)	C	D

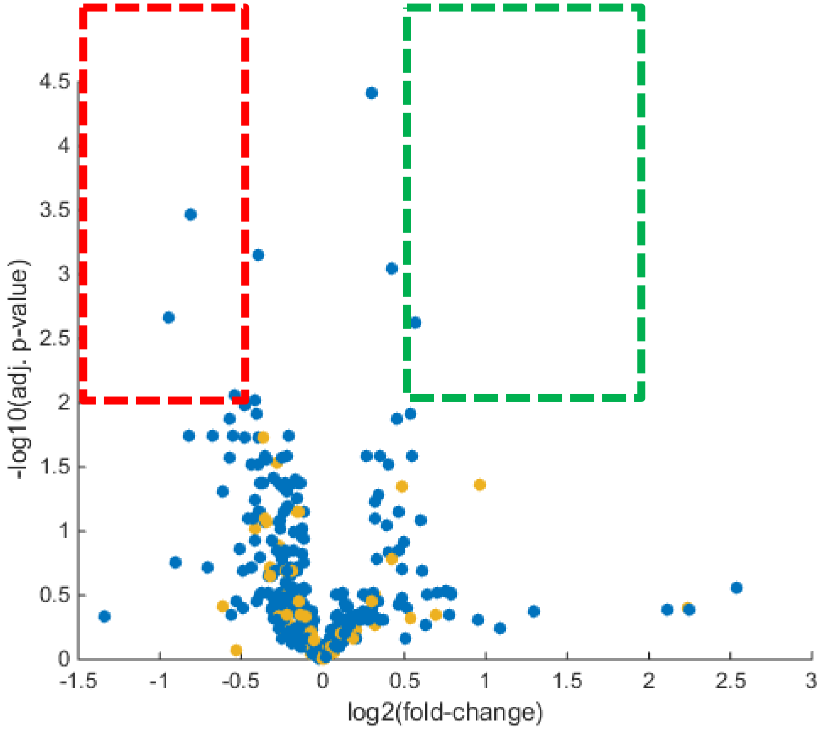
From the contingency table, we can directly calculate the likelihood that the same results could be obtained assuming the null hypothesis according to a Fisher’s exact test (or an hypergeometric distribution)

In MATLAB: *(all formula give the same result)*

```
      >> [~,p] = fishertest([A B;C D],’Tail’,’right’)
or    >> p = sum(hygepdf(A:A+B, A+B+C+D, A+C, A+B))
or    >> p = 1-hygecdf(A-1, A+B+C+D, A+C, A+B)
```

Problem: How to choose cutoffs?

The contingency table depends on pre-defined cutoffs. These



Arguments against using the “normal” cutoffs: [Subramanian PNAS 2005]

After correcting for multiple hypotheses testing, no individual gene may meet the threshold for statistical significance, because the ***relevant biological differences are modest relative to the noise.***

Single-gene analysis may miss important effects on pathways. Cellular processes often affect sets of genes acting in concert. ***An increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene.***

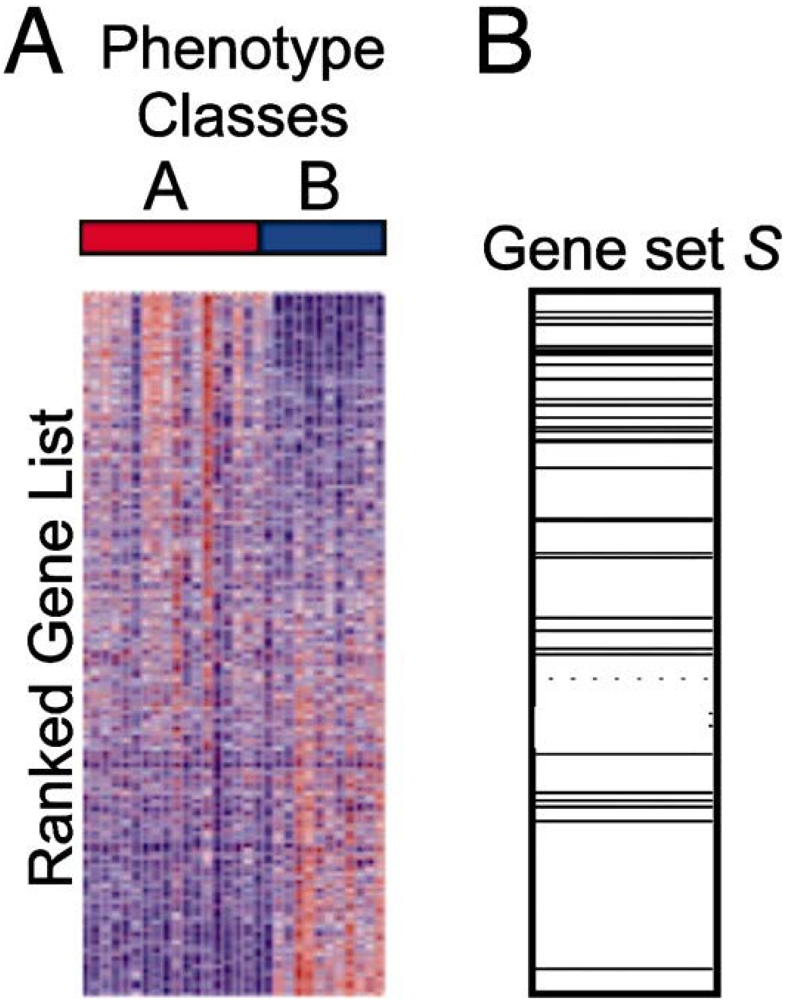
Gene Set Enrichment Analysis

Subramanian et al, PNAS 2005

(A) Genes are ranked in a list L based on the correlation between their expression and the class distinction by using any suitable metric, e.g. log2(fold-change) or t-test p-value.

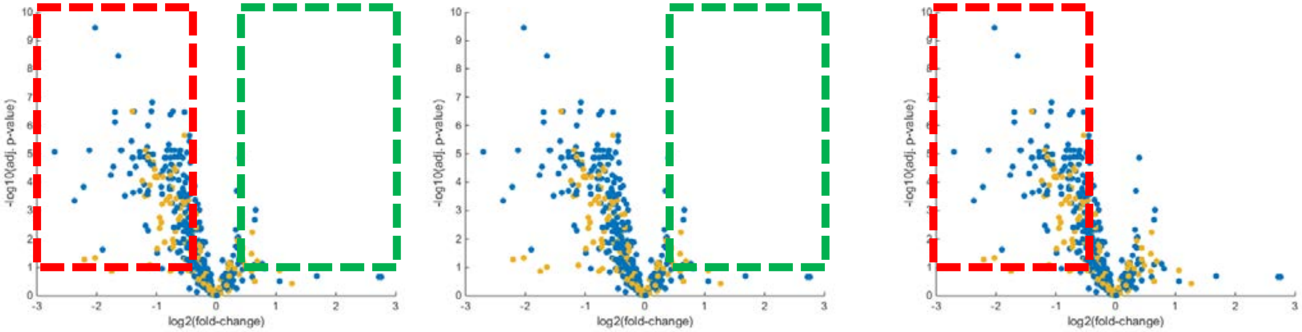
(B) Given an a priori defined set of genes S, the goal of GSEA is to determine whether the members of S are randomly distributed or primarily found at the top or bottom.

We expect that sets related to the phenotypic distinction will tend to show the latter distribution.



The GSEA procedure

1. **Decide on what signs do we want to include:**



POS or NEG

POS only

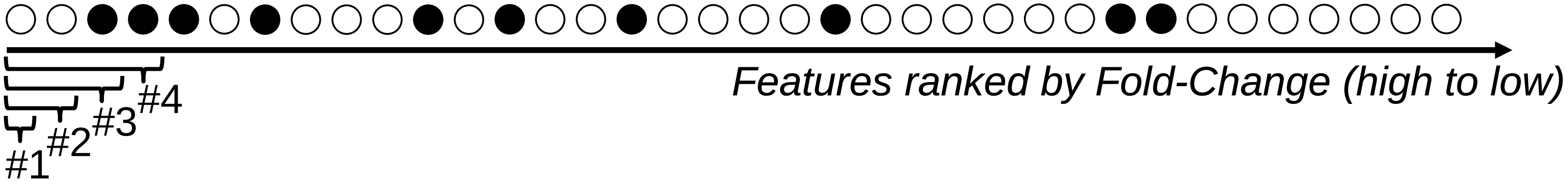
NEG only

2. **Procedure for 1 pathway/group/process:**
   1. Choose very permissive thresholds (e.g. log2FC > 0.2 and p-value < 0.1)
   2. Rank all hits on either p-value (preferred) or log2FC
   3. Build contingency tables with either 2, 3, 4, …., ALL top hits
   4. For each table, calculate p-value using Fisher’s exact test
   5. Keep lowest p-value = “best” enrichment

3. **Repeat for all pathways**
4. **FDR-correction (Benjamini-Hochberg or Storey)**

Example

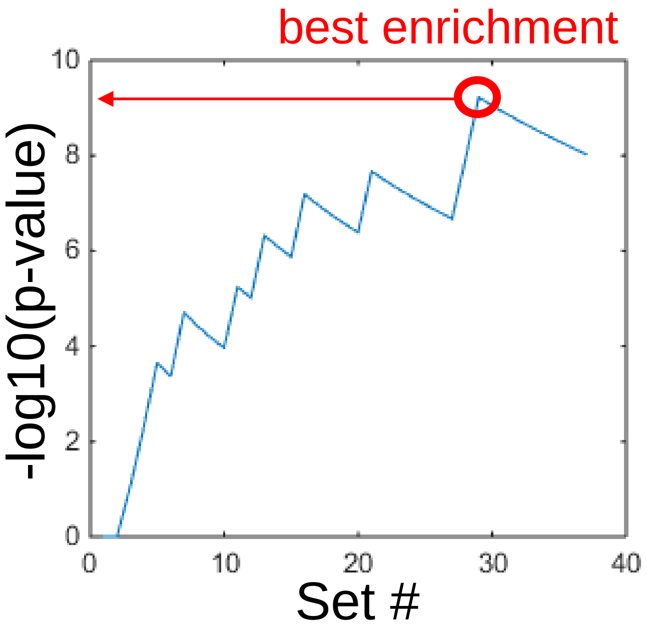
Out of 650 features detected, we know that 20 belong to a pathway/GO term.
Out of 650, 36 features passed our criteria abs(log2FC) > 0.2 and adj. p-value < 0.1.



Features ranked by Fold-Change (high to low)

set	A+B+C+D	A+C	A ●	B ○	C	D	Enrichment p-value (Fisher)
#1	650	20	0	1	20	629	1.00
#2	650	20	0	2	20	628	1.00
#3	650	20	1	2	19	628	0.087
#4	650	20	2	2	18	628	0.049
#5	650	20	3	2	17	628	0.00022
#6	650	20	3	3	17	627	0.00043
#7	650	20	4	3	16	627	0.000019
#8	650	20	4	4	16	626	0.00038
…							

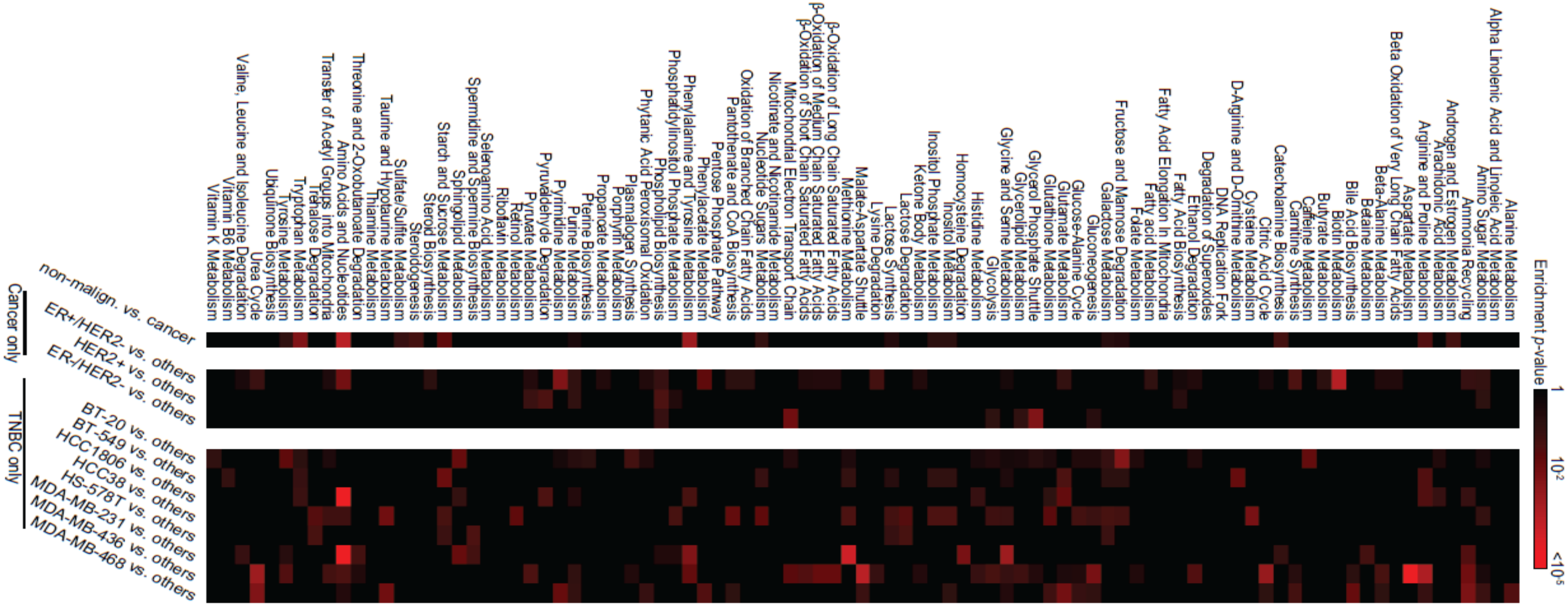
```
[~,p] = fishertest([A B;C D],'Tail','right')
```



EA across pathways

A full analysis is typically repeated on tens/hundreds of pathways/GO terms.

1. For each pathway, we keep the lowest p-value
2. FDR correction of p-values (e.g. Benjamini-Hochberg or Storey)
3. Enrichments are significant if adj. p-value or q-value < 0.01 (0.05)



Self check

- What if no enrichment is found?

- What happens if the enrichment analysis is done without any log2FC or p-value cutoff?

Summary

- **Define the question in biological and data mining terms**.
- Data mining of omics data is about hypothesis generation.
- Seeking for absolute certainty is likely to prevent discovery of testable hypotheses.
- In the simplest approach, significant features are found by univariate analysis of all features (independently)
  > t-test > FDR correction
- Significantly changed “processes” are identified by enrichment analysis
  > rank by p > test subsets > Fisher’s exact test > FDR correction

Exercises

- This week:
  - two groups problem, p-values, multiple hypothesis testing

What are the significantly changed features?

- Next week:
  - gene set enrichment analysis

What are the significantly changed processes?