Pathway analysis

Urmo Võsa

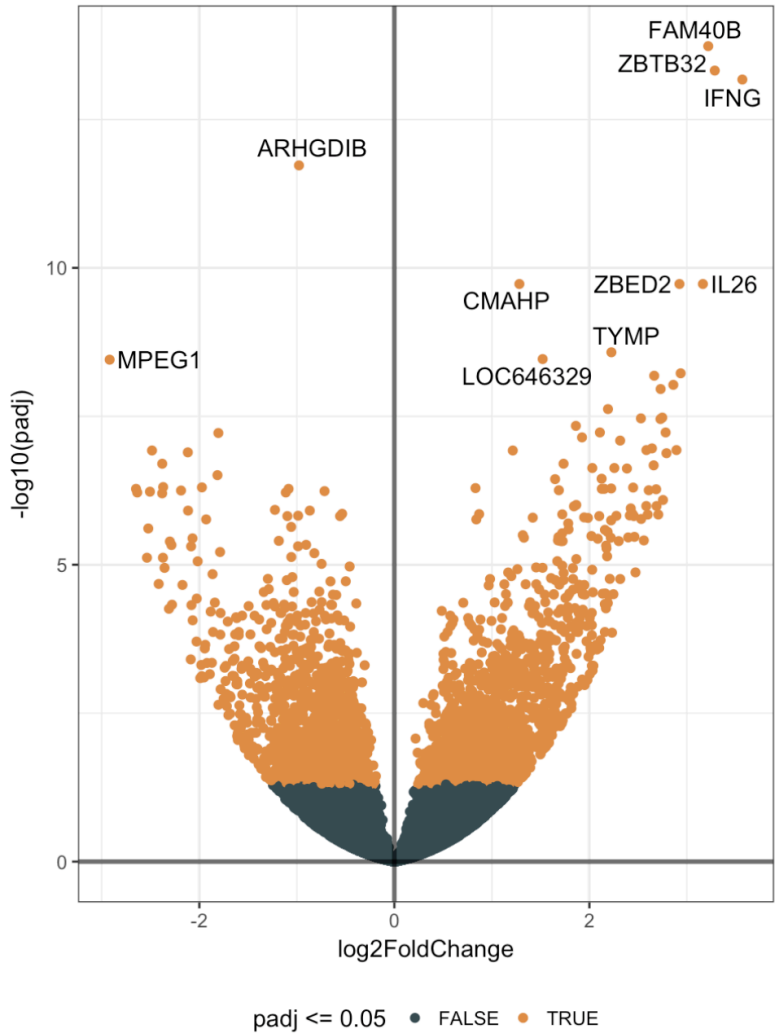
iBMS course 2017

FrankeSwertzLab

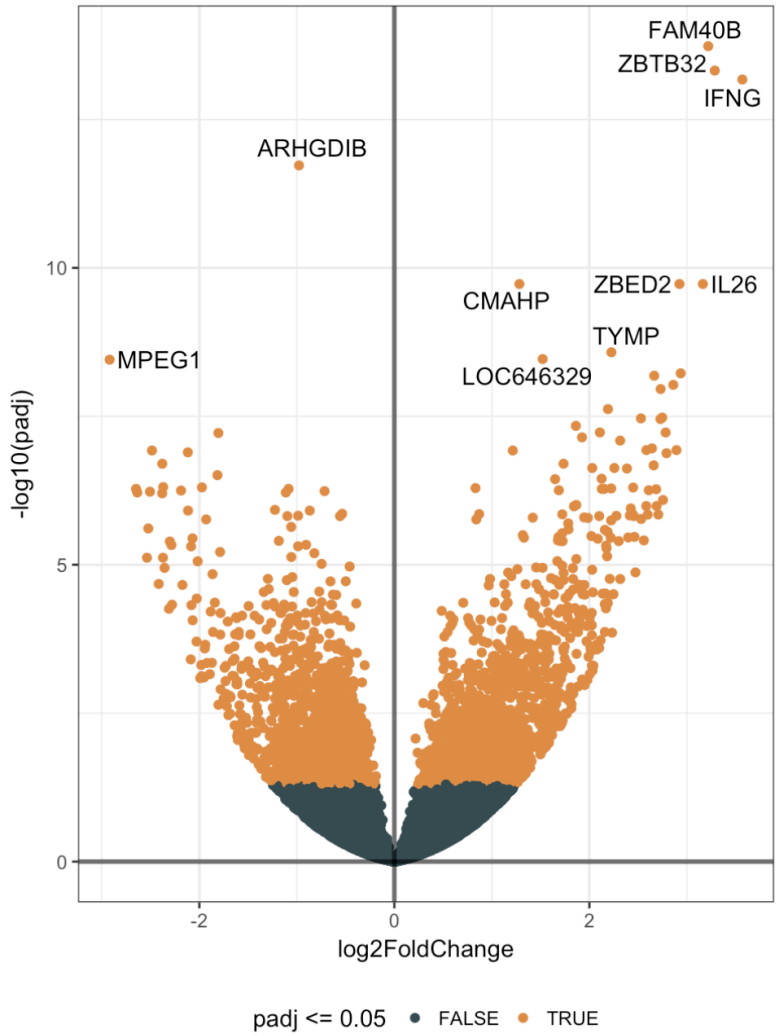
Department of Genetics

UMCG

What next?


FAM40B
ZBTB32
IFNG
ARHGDIB
CMAHP
ZBED2
IL26
TYMP
MPEG1
LOC646329
10
5
0
-2
0
2
-log10(padj)
log2FoldChange
padj <= 0.05
FALSE
TRUE

What next?



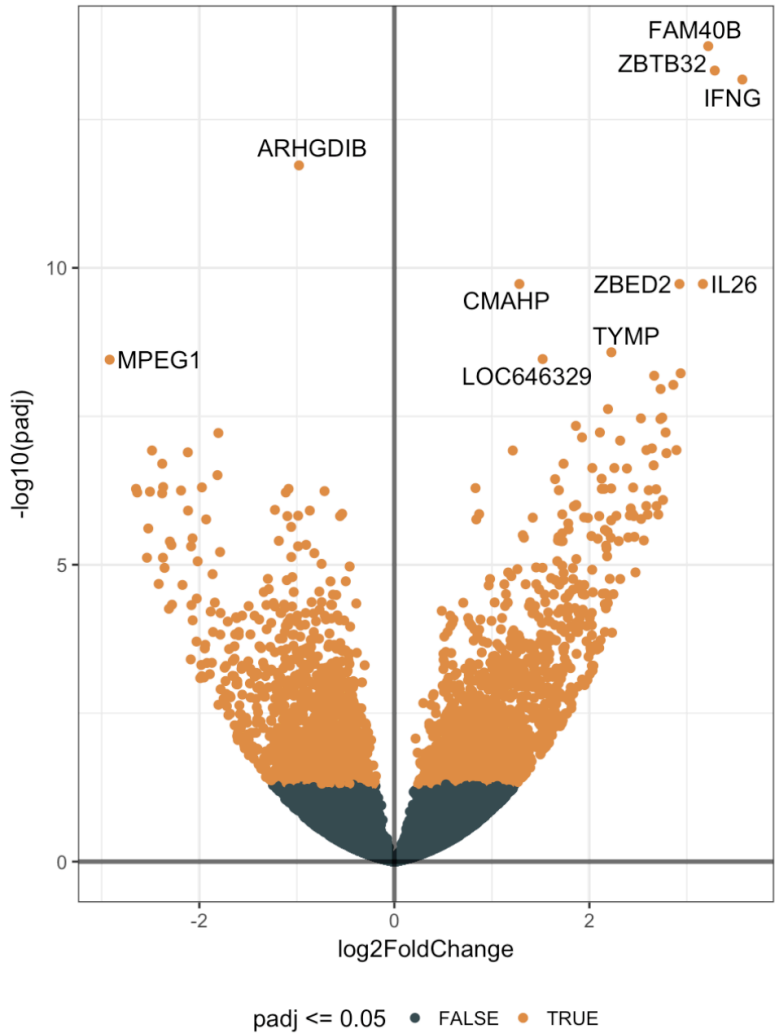
Basic science:

- New knowledge about biological mechanisms of …

Practical aspects:

- Diagnostic/prognostic biomarkers for the disease
- Potential drug targets

What next?



Basic science:

- New knowledge about biological mechanisms of …

Practical aspects:

- Diagnostic/prognostic biomarkers for the disease
- Potential drug targets

- Find common themes from the results
- Pinpoint potentially relevant genes
- Put results into biological context

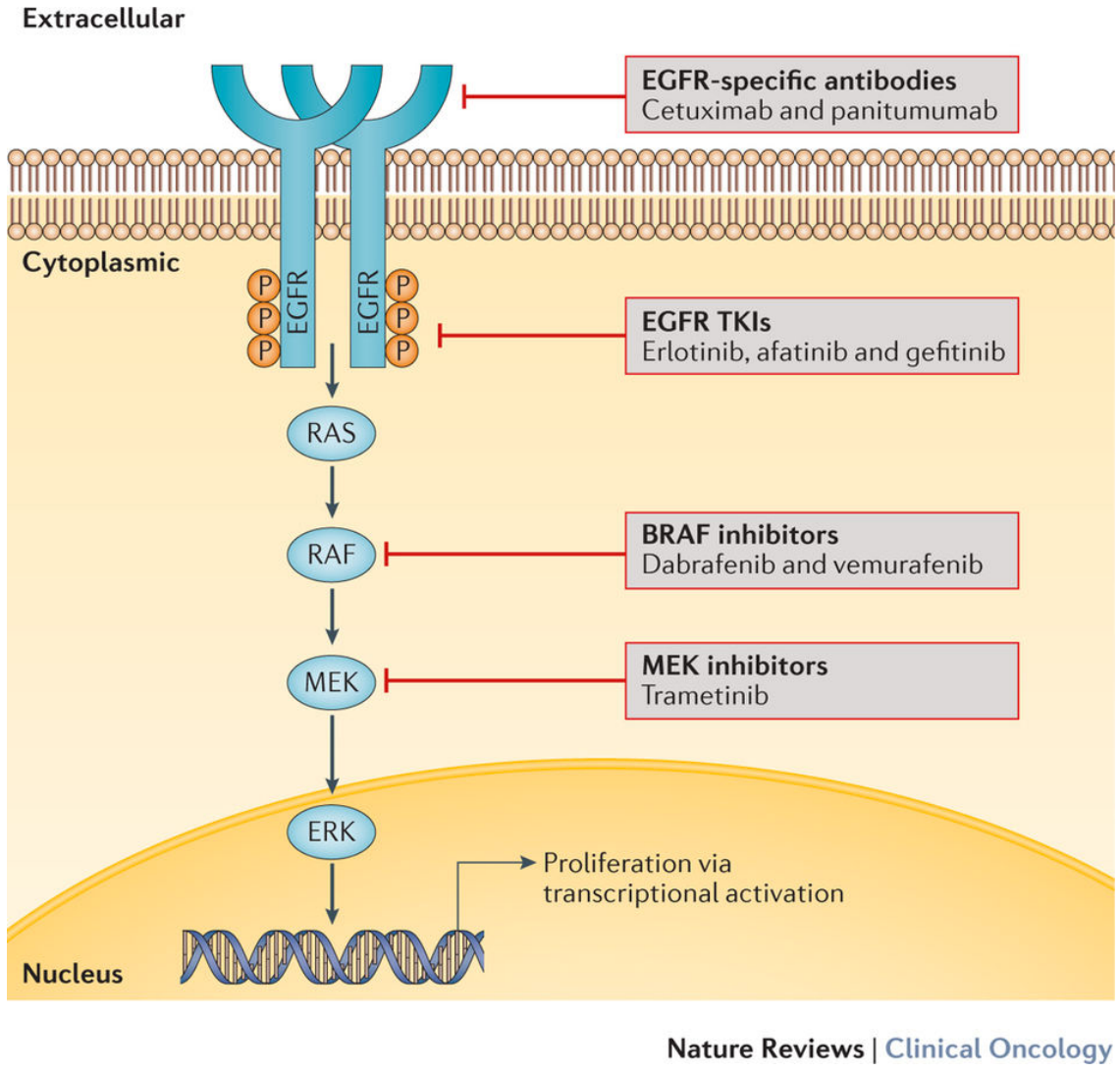
Pathway analysis

What is molecular pathway?

*A **molecular pathway** is a series of interactions that occurs between the molecules and proteins within a cell and between cells.*

http://www.geneticalliance.org/

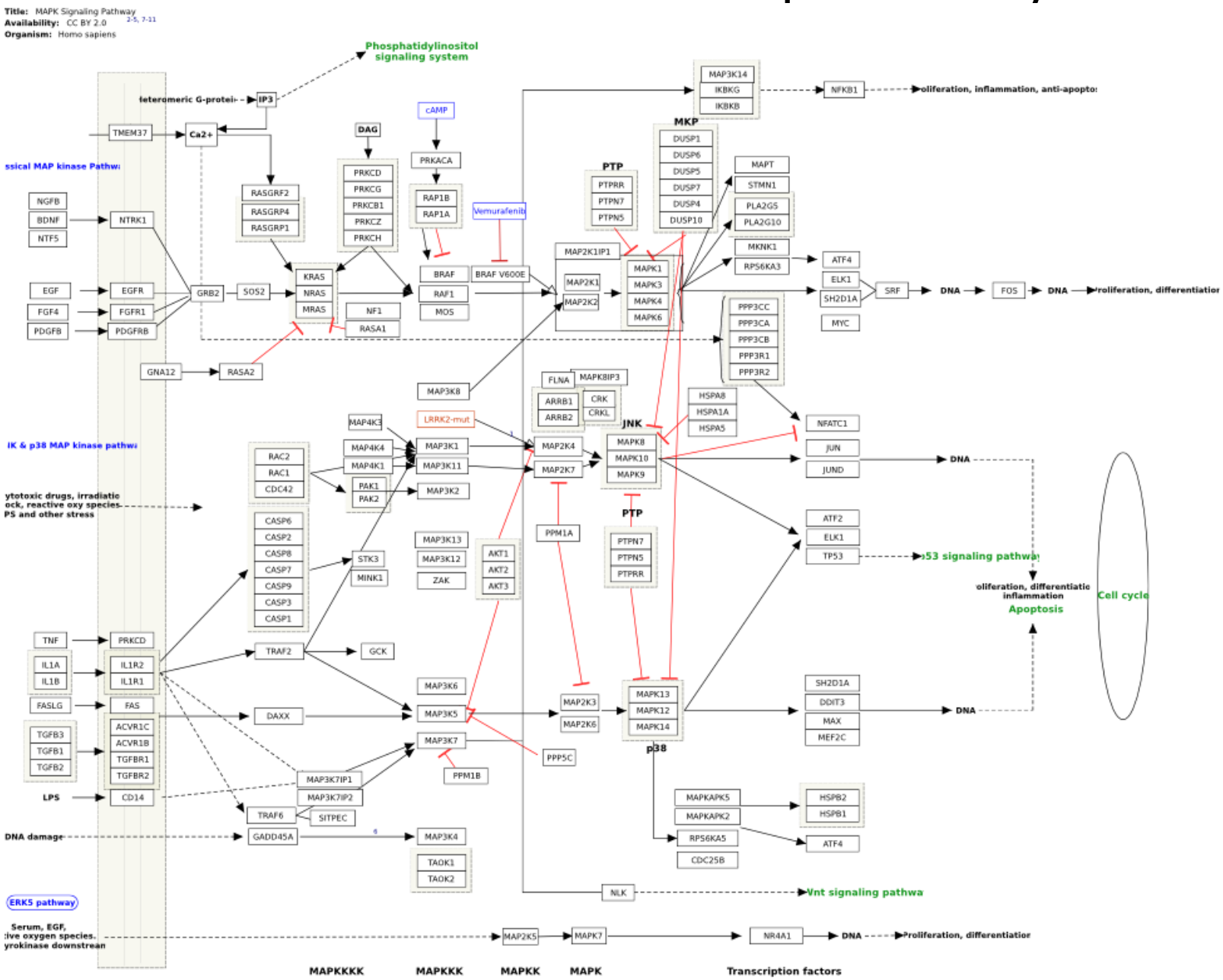
What is molecular pathway?



Simplified MAPK pathway

Schmitt et al., 2016

What is molecular pathway?



http://www.wikipathways.org/

Gene set

Any set of genes which shares some common theme:

- **on the shared pathway**
- share biological process/theme
- share cellular localization
- share common disease annotation
- share common regulating transcription factor/microRNA
- share common genomic location (chromosome band)

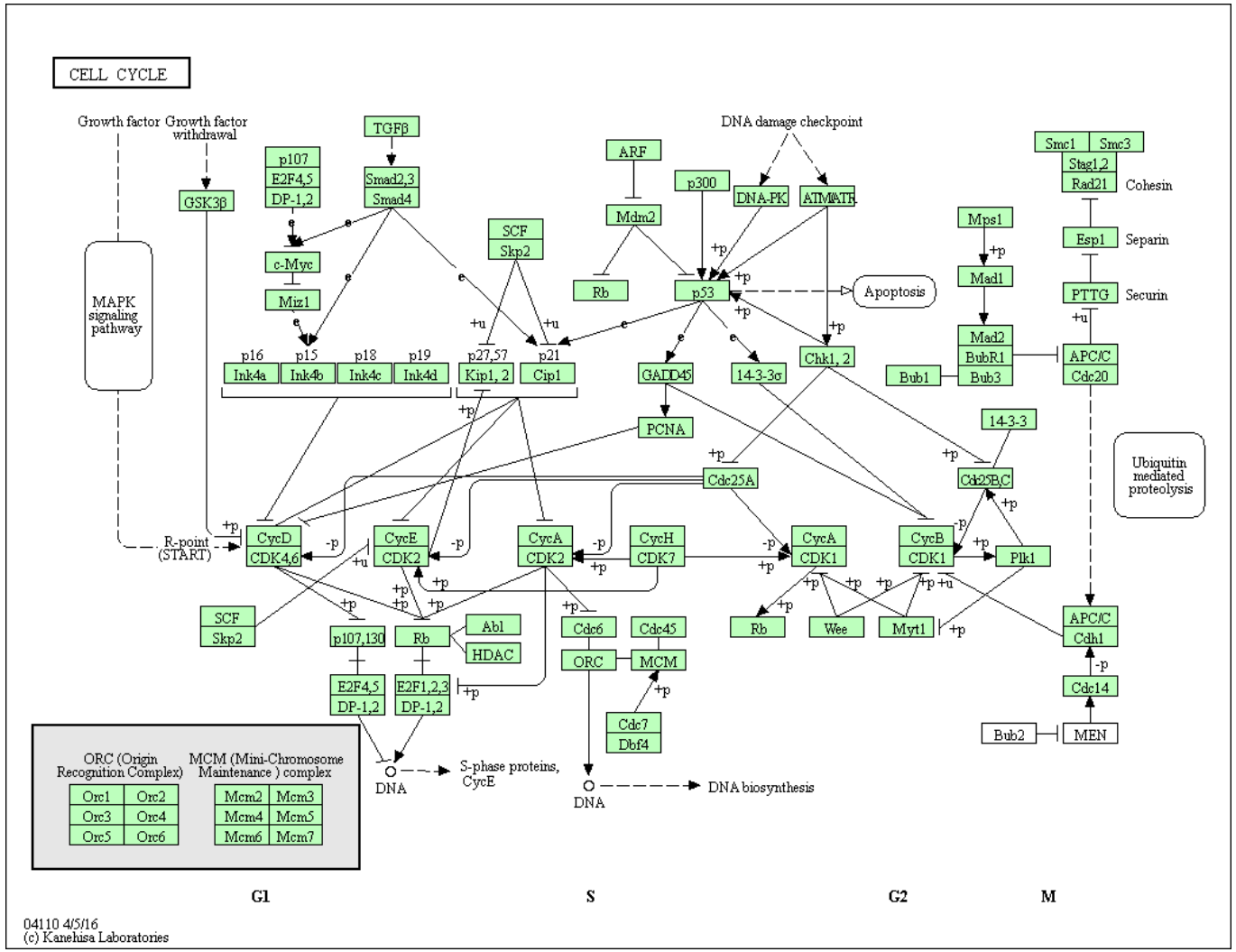
- “genes associated with NIH grants”
- …

Gene set is not pathway!

~~Pathway analysis~~
Gene enrichment analysis

Some most widely used databases

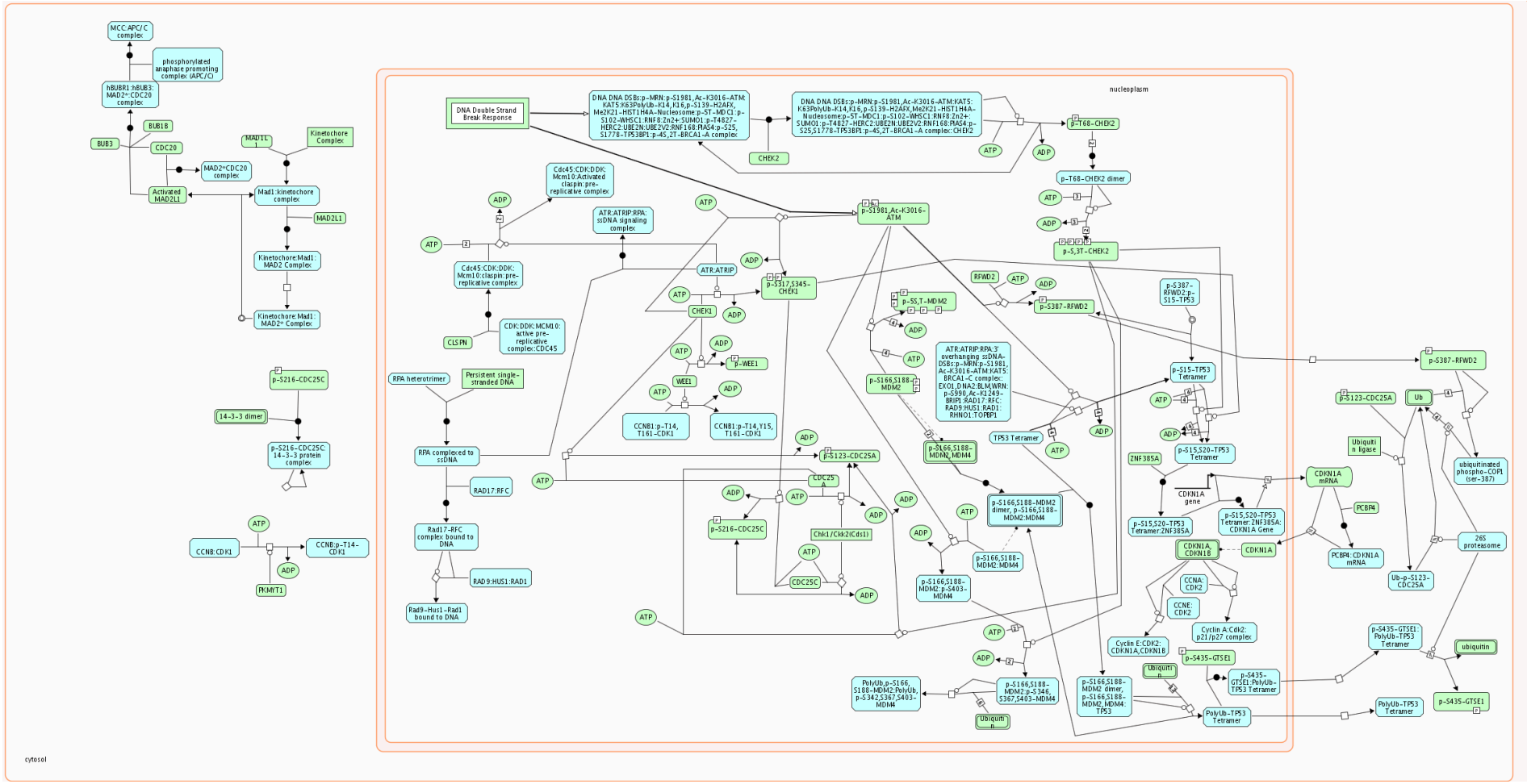
- Kyoto Encyclopedia of Genes and Genomes (**KEGG**)



www.genome.jp/kegg/

Some most widely used databases

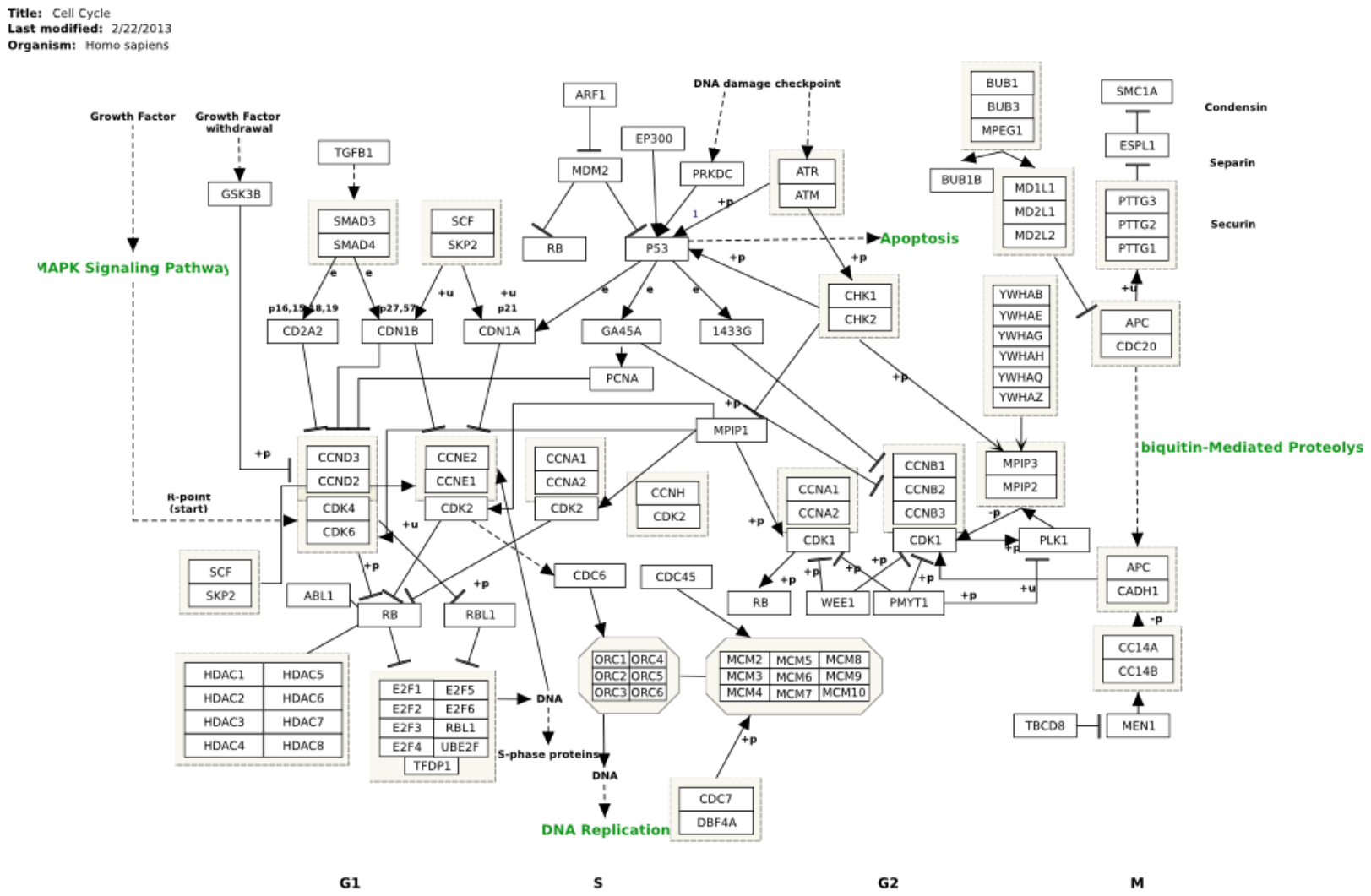
- REACTOME



www.reactome.org

Some most widely used databases

- WikiPathways



www.wikipathways.org

Some most widely used databases

- Gene Ontology (GO):
  - Biological processes
  - Molecular function
  - Cellular composition

http://www.geneontology.org/
http://cbl-gorilla.cs.technion.ac.il/

Some most widely used databases

- Gene Ontology (GO):
  - Biological processes
  - Molecular function
  - Cellular composition

Not exactly the pathway but gene set!

http://www.geneontology.org/
http://cbl-gorilla.cs.technion.ac.il/

Some most widely used databases

- Gene Ontology (GO):
  - Biological processes
  - Molecular function
  - Cellular composition

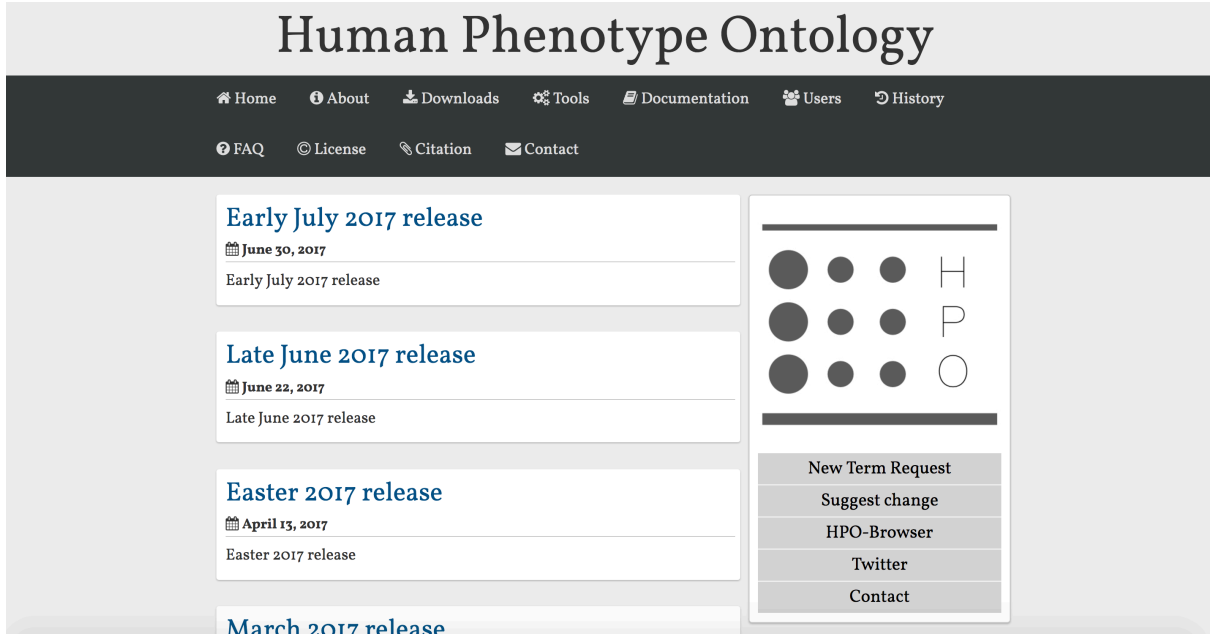
Not exactly the pathway but gene set!

Gene set:
e.g. all genes part of BP GO:0000075
Cell cycle checkpoint

http://www.geneontology.org/
http://cbl-gorilla.cs.technion.ac.il/

Some most widely used databases

- Human Phenotype Ontology (HPO):
  - 11,000 terms
  - 115,000 hereditary diseases
  - 4,000 common diseases

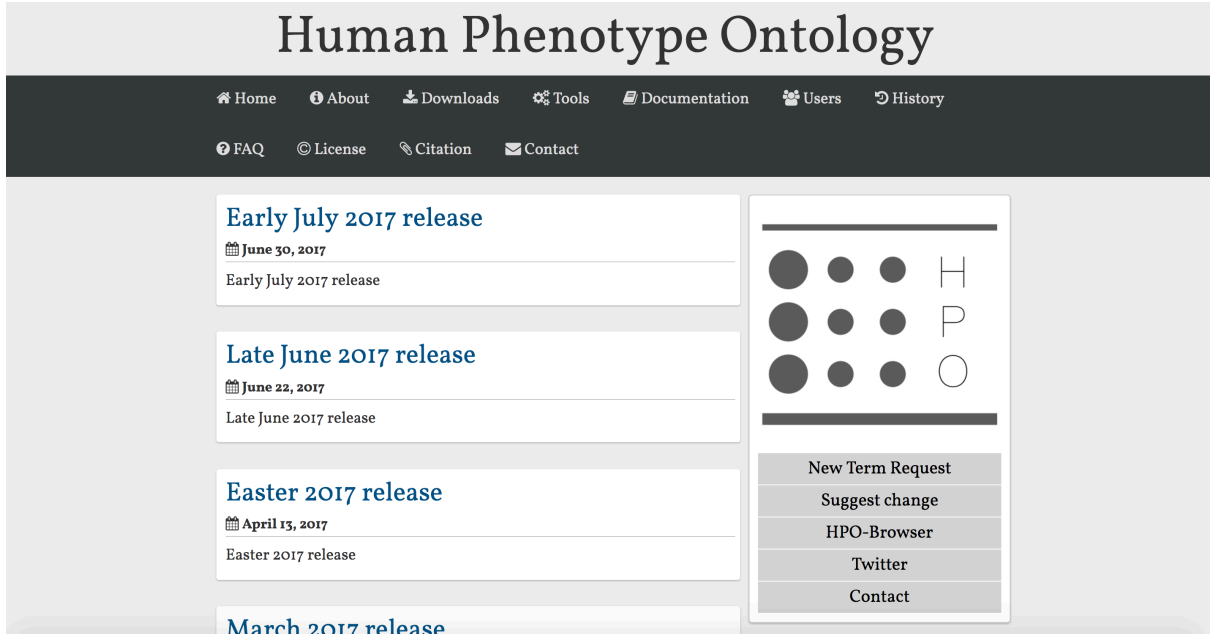


Not exactly the pathway but gene set!

Köhler et al., 2017
http://human-phenotype-ontology.github.io/

Some most widely used databases

- Human Phenotype Ontology (HPO):
  - 11,000 terms
  - 115,000 hereditary diseases
  - 4,000 common diseases



Not exactly the pathway but gene set!

Gene set:
e.g. all genes part of HP:0030359
Squamous cell lung carcinoma

Köhler et al., 2017
http://human-phenotype-ontology.github.io/

Gene enrichment analysis

Question: are genes of interest significantly enriched by some gene set (pathway/process/…)?

Gene enrichment analysis

Question: are genes of interest significantly enriched by some gene set (pathway/process/…)?

In our case: are the genes differentially expressed between Coeliac disease patients and control individuals over-represented by some specific biological themes?

Classification of pathway analysis methods

Overrepresentation analysis (ORA)

- Hypergeometric test
- Fisher's exact test
- Chi-squared test
- Bayesian statistics
- …

Functional class scoring (FCS)

- GSEA
- CAMERA
- …

Topology-based methods (PT)

- SPIA
- …

Semantic similarity analyses

- GoSemSim
- …

Miscallaneous

- GeneNetwork
- …

Khatri et al., 2012 (Plos Comp. Biol.)

Classification of pathway analysis methods

Overrepresentation analysis (ORA)

- Hypergeometric test
- Fisher's exact test
- Chi-squared test
- Bayesian statistics
- …

Functional class scoring (FCS)

- GSEA
- CAMERA
- …

Topology-based methods (PT)

- SPIA
- …

Semantic similarity analyses

- GoSemSim
- …

Miscallaneous

- GeneNetwork
- …

Khatri et al., 2012 (Plos Comp. Biol.)

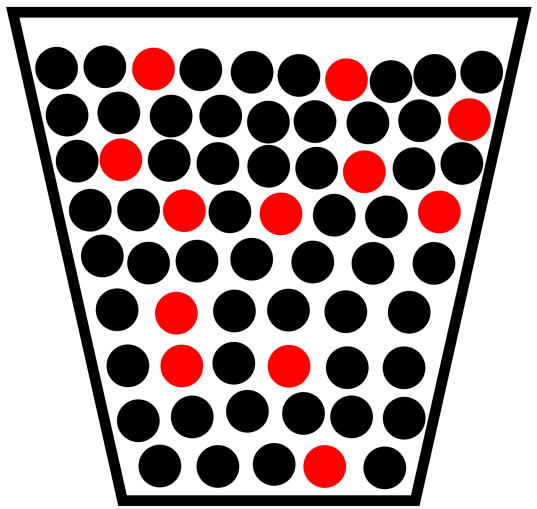
Over-representation analysis (ORA)

- Uses the defined set **genes of interest** (e.g. diff. expressed genes)

Over-representation analysis (ORA)

- Uses the defined set **genes of interest** (e.g. diff. expressed genes)

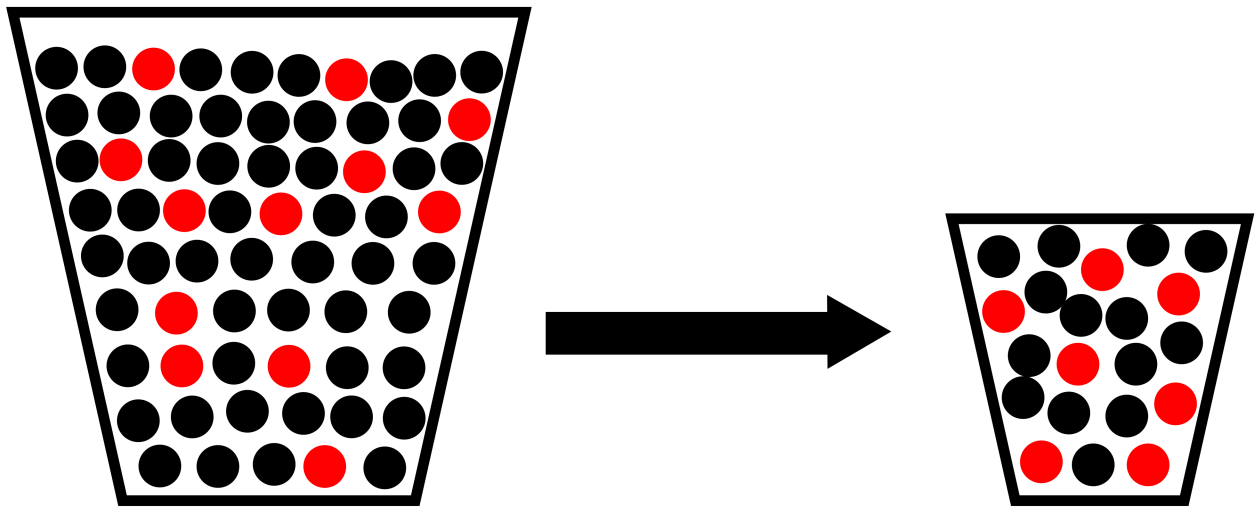
Urn analogy:



Over-representation analysis (ORA)

- Uses the defined set **genes of interest** (e.g. diff. expressed genes)

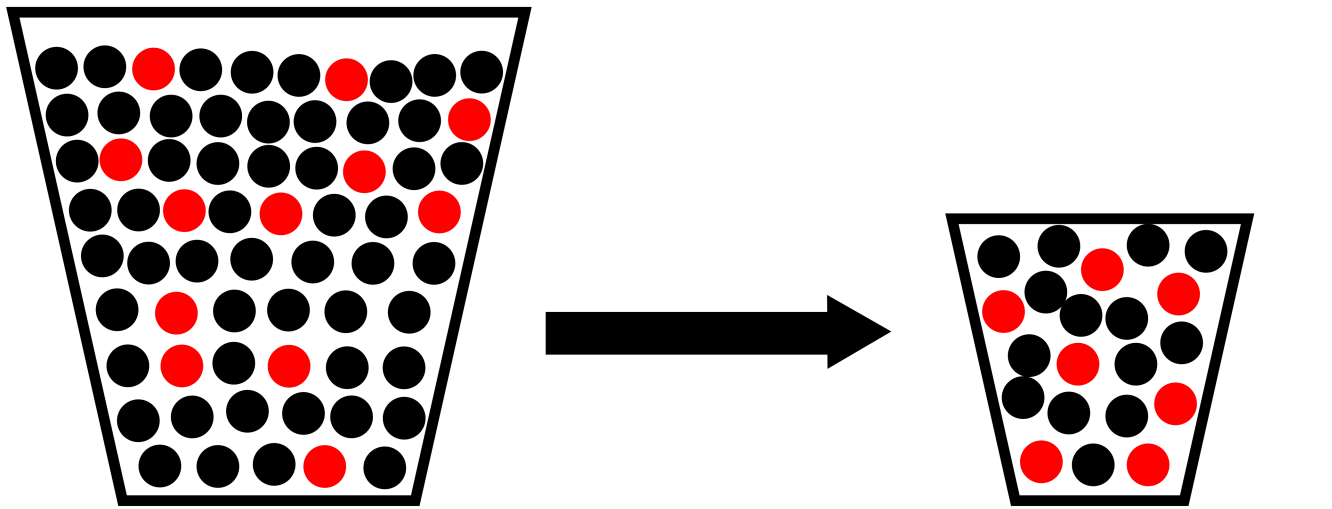
Urn analogy:



Over-representation analysis (ORA)

- Uses the defined set **genes of interest** (e.g. diff. expressed genes)

Urn analogy:



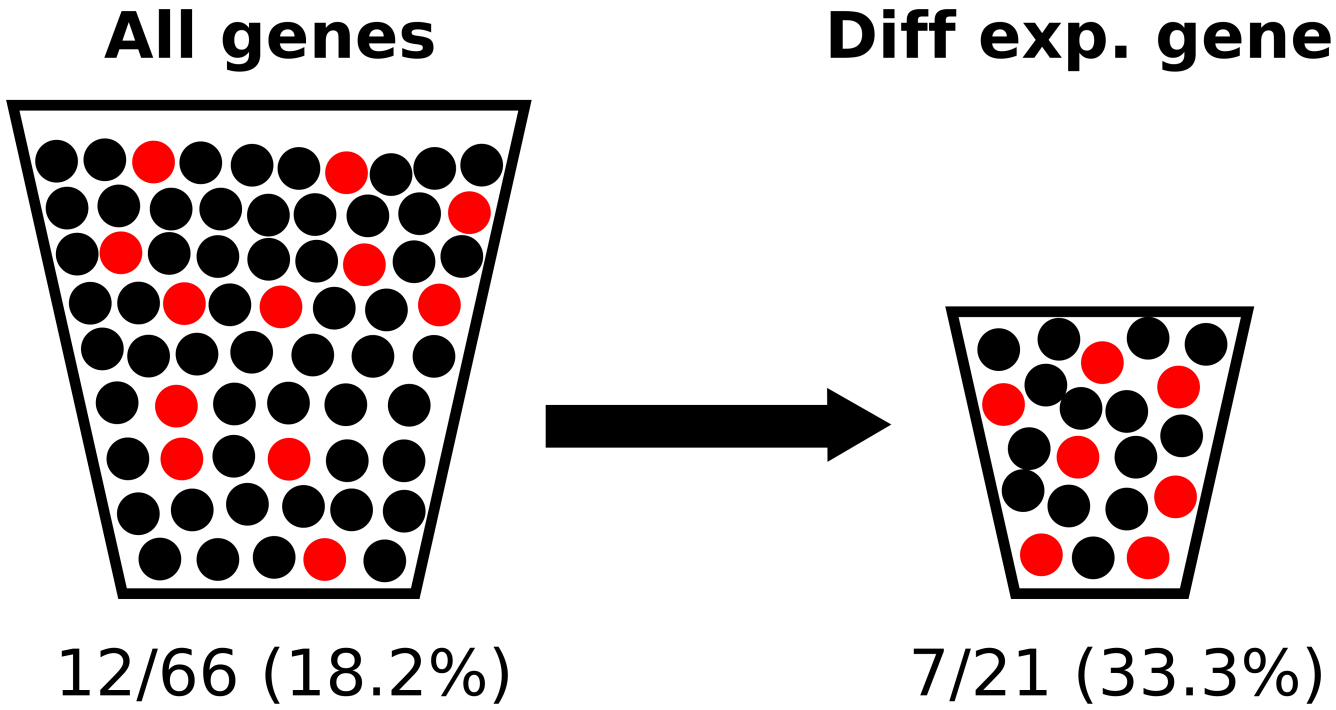
12/66 (18.2%)

7/21 (33.3%)

Over-representation analysis (ORA)

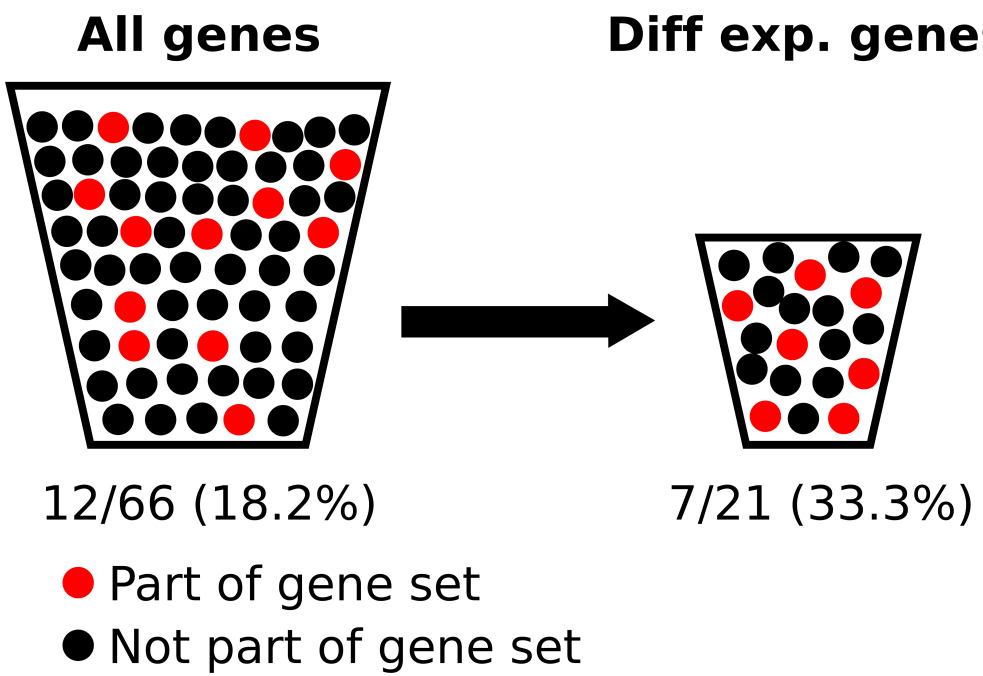
- Uses the defined set **genes of interest** (e.g. diff. expressed genes)

Urn analogy:



Most widely used method

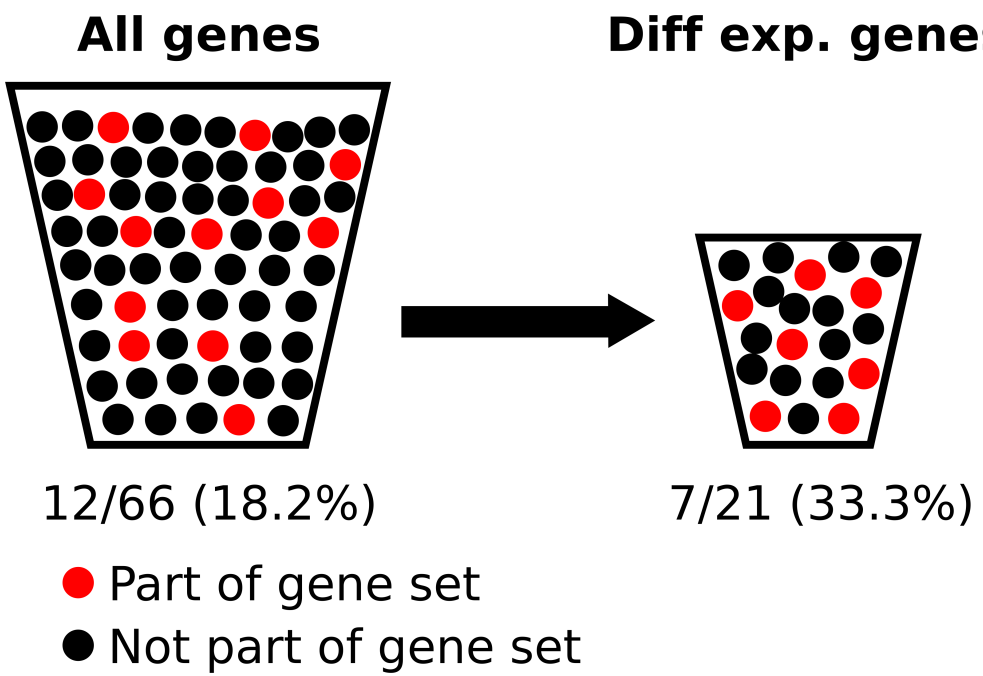
Hypergeometric test (one-sided Fisher's Exact Test)



	Part of gene set	Not part of gene set
Diff. expressed	**7**	21-7=**14**
Not. diff expressed	12-7=**5**	66-21-(12-7)=**40**

Most widely used method

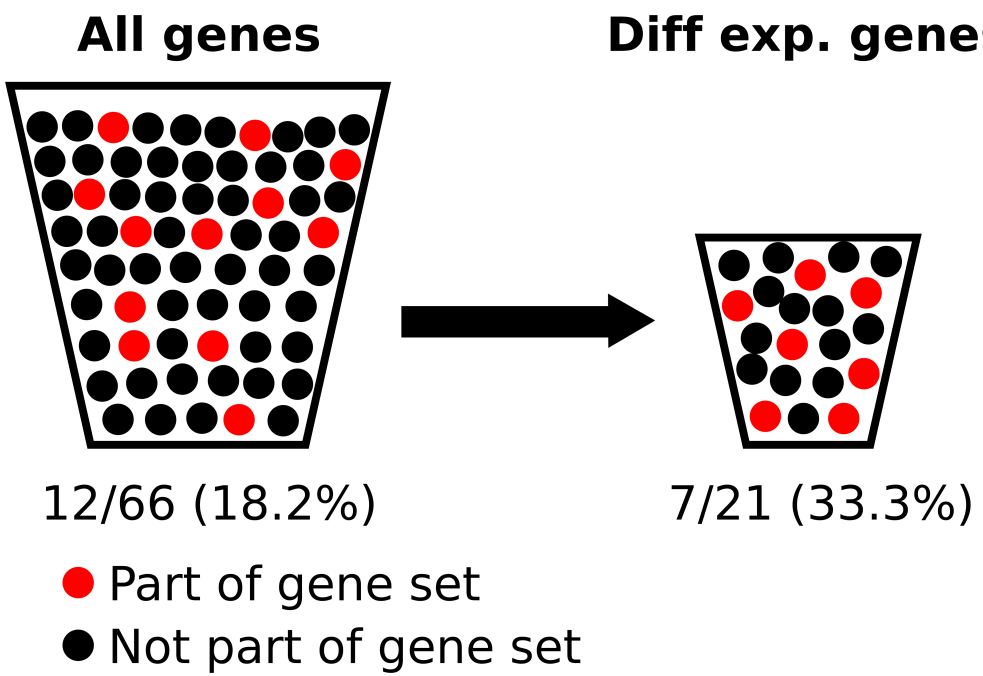
Hypergeometric test (one-sided Fisher's Exact Test)



	Part of gene set	Not part of gene set	∑
Diff. expressed	**7**	21-7=**14**	21
Not. diff expressed	12-7=**5**	66-21-(12-7)=**40**	45
∑	12	54	66

Most widely used method

Hypergeometric test (one-sided Fisher's Exact Test)



	Part of gene set	Not part of gene set	∑
Diff. expressed	**7**	21-7=**14**	21
Not. diff expressed	12-7=**5**	66-21-(12-7)=**40**	45
∑	12	54	66

One-sided FET: **P=0.036**

Odds ratio: **OR=3.9**

Multiple testing!

Usually many gene sets tests are done in parallel.

Multiple testing correction is needed to limit false positive results!

Multiple testing!

Usually many gene sets tests are done in parallel.

Multiple testing correction is needed to limit false positive results!

- Bonferroni
- FDR

Gene Universe

"*Gene Universe*" AKA "*Background*"

The set of genes which could be identified by your analysis.

- Often used: all annotated genes for investigated organism.
- For diff. exp. analysis: all genes which were tested by your method.
- It is not trivial to choose suitable background for more complex designs.

Incorrect selection of background may bias your analysis and lead to wrong conclusions!

Limitations of ORA

- Defining significance of the diff. expressed genes is often arbitrary.
- All genes are treated with equal weight in the analysis.
- Assumes the independence of the genes.

GSEA

- Uses the full information from genes in the analysis
- Small, non-significant but consistent expression changes are picked up by this analysis.
- Ranks genes based on their effect size (e.g. strength of expression change)
- Analyses whether genes in the top/bottom of the ranked list contain more pathway members as you would expect by chance.

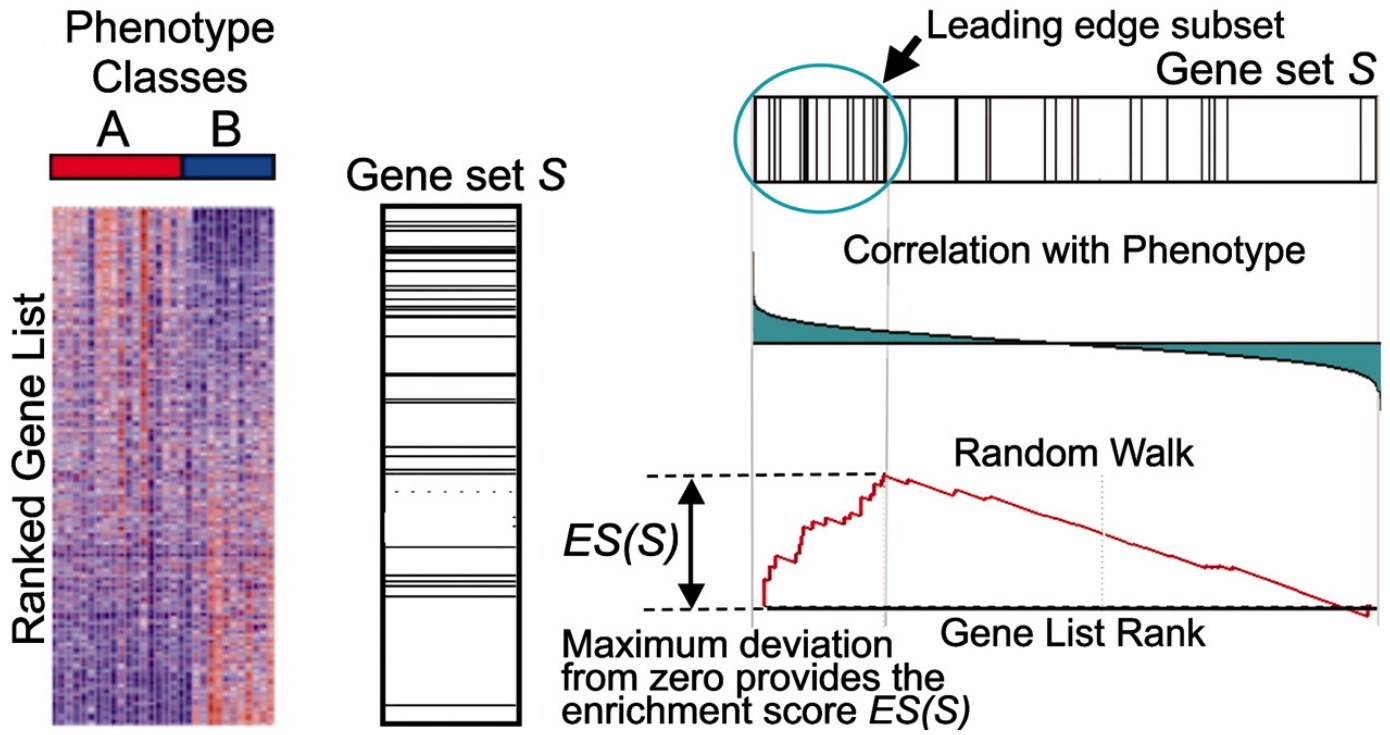
NB! The term is often used interchangeably to designate ORA (*confusing!*)

Subramanian et al., 2005

GSEA algorithm

For each Gene Set:

- Walk down from the ranked list of diff. exp. genes
  - Each time the gene is part of Gene Set: add value to the running sum. The increment is weighted by the correlation with phenotype (or fold change of diff expression).
  - Each time the gene is not part of Gene Set: subtract the weighted value from the running sum.
  - Enrichment score (**ES**) is the max. deviation of running sum from 0.



Subramanian et al., 2005

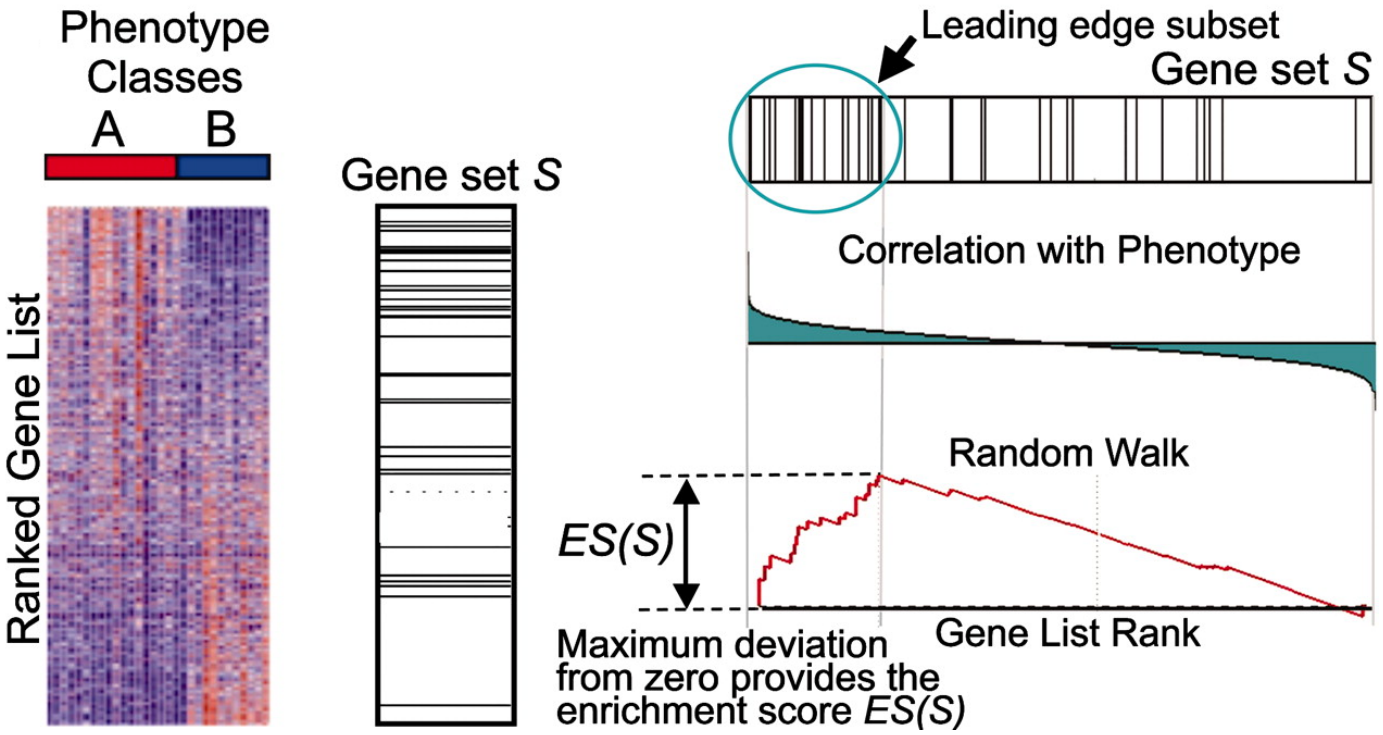
GSEA algorithm

For each Gene Set:

- Calculate P-value for ES
  - Permute phenotypes or gene labels to get null distribution of ES
- Correct for multiple testing

- Normalized enrichment score (**NES**) is more proper comparison metric between gene sets. It is defined as follows:

$$NES = \frac{actual\ ES}{mean\ ES\ over\ permutations}$$



Subramanian et al., 2005

Some recommendations

- Gene sets which are too small are usually not very trustworthy in ORA and GSEA.
- Likewise, gene sets which are very big are difficult to interpret.

Web- or command-line tool?

Web-based tools

+ Easy to use
+ Often gives comprehensive report

- Not replicable
- Often out of date
- Often “black box”
- Often not flexible
- Prone to mistakes

Command-line tools

+ Replicable
+ Very flexible
+ Parallelizable
+ Integratable to the pipelines
+ Smaller possibility to make mistakes

- Steep learning curve