מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

LIFE SCIENCE
CORE FACILITIES

Functional analysis of gene lists using Gene Ontology (GO)

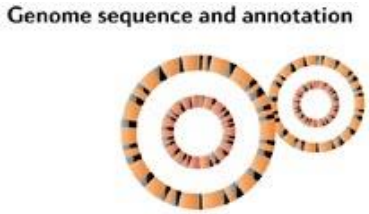
Noa Wigoda

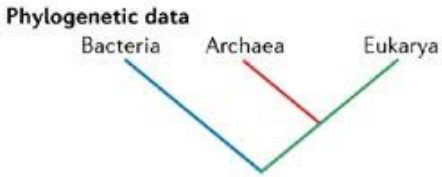
6.12.21

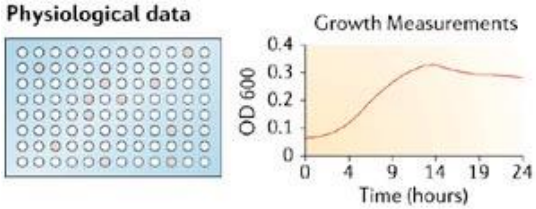
An Introduction to deep-sequencing analysis for biologists

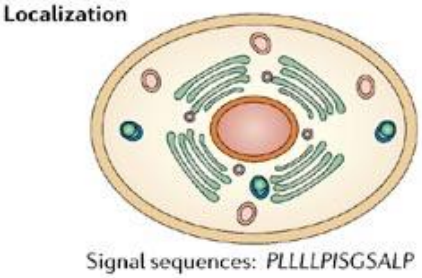
OUTLINE

- Single gene analysis / information
- Analysis of group of genes
- Gene ontology (GO)
- Enrichment analysis
  - Hypergeometric Test and Fisher exact test
  - GO Independence Assumption









20 Questions

Ask question which can be answered with a simple "Yes" or "No."

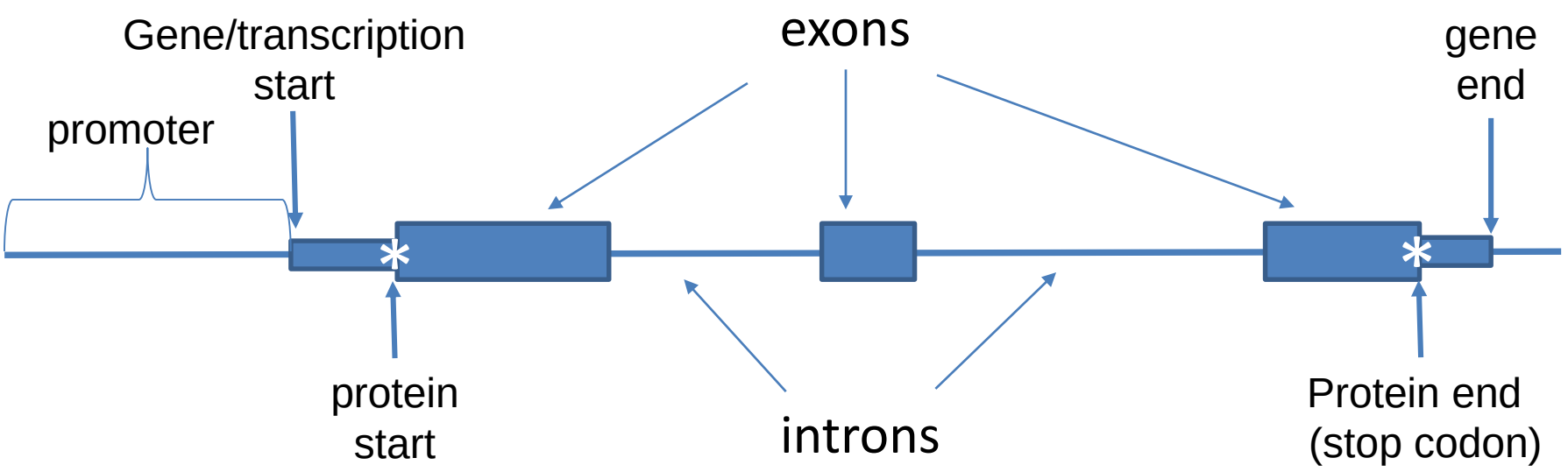
20 Questions

All the answers are “attributes” or characteristics of the item (gene).

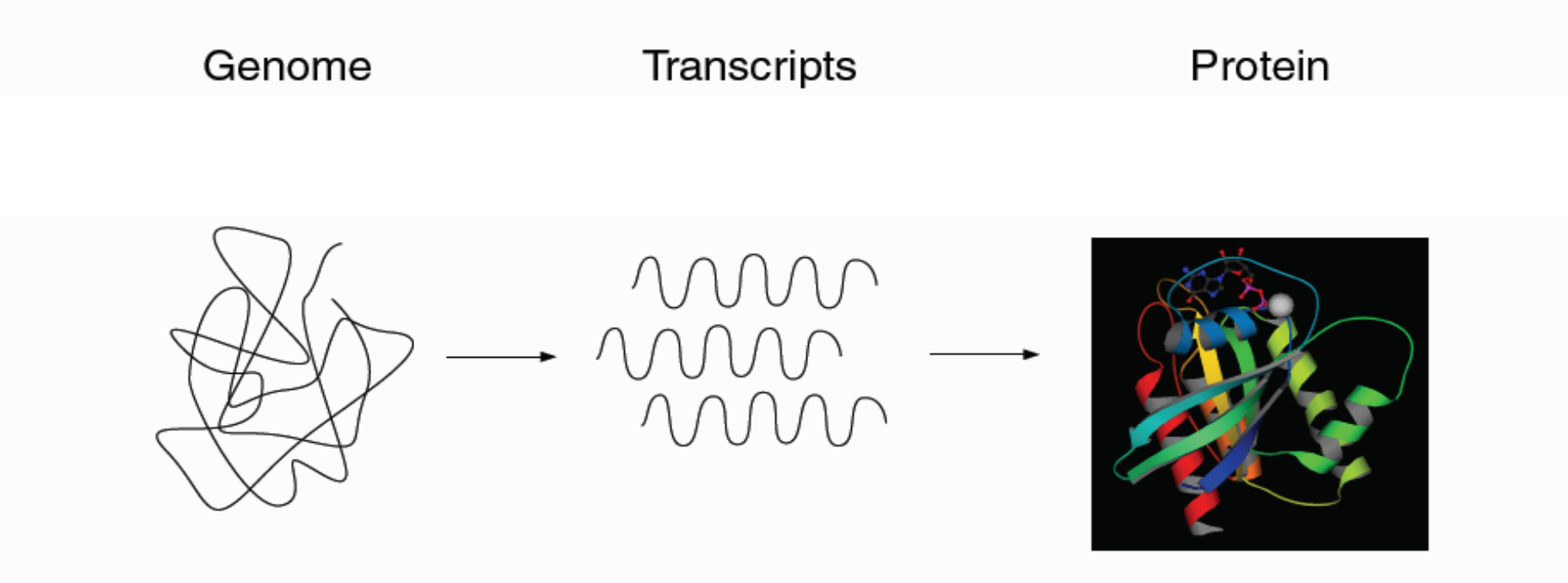
What is a Gene ?

A gene is a region of DNA that encodes instructions for how the cell can make a gene product, which can be:

- a protein
- a noncoding RNA.



Data sources



- There are several kinds of databases, looking at the genome, transcriptome or proteome level.
- The mapping of the different names is not trivial.

Example:

GENE

BRCA2 – BRCA2 DNA repair associated

Homo sapiens (human)

Also known as: BRCC2, BROVCA2, FACD, FAD, FAD1, FANCD, FANCD1, GLM3, PNCA2, XRCC11

Gene ID: 675

Common Identifiers

Gene

Ensembl ENSG00000139618
Entrez Gene 675
Unigene Hs.34012

RNA transcript

GenBank BC026160.1
RefSeq NM_000059
Ensembl ENST00000380152

Protein

Ensembl ENSP00000369497
RefSeq NP_000050.2
UniProt BRCA2_HUMAN or A1YBP1_HUMAN
IPI IPI00412408.1
EMBL AF309413
PDB 1MIU

Species-specific

HUGO HGNC BRCA2
MGI MGI:109337
RGD 2219
ZFIN ZDB-GENE-060510-3
FlyBase CG9097
WormBase WBGene00002299 or ZK1067.1
SGD S000002187 or YDL029W

Annotations

InterPro IPR015252
OMIM 600185
Pfam PF09104
Gene Ontology GO:0000724
SNPs rs28897757

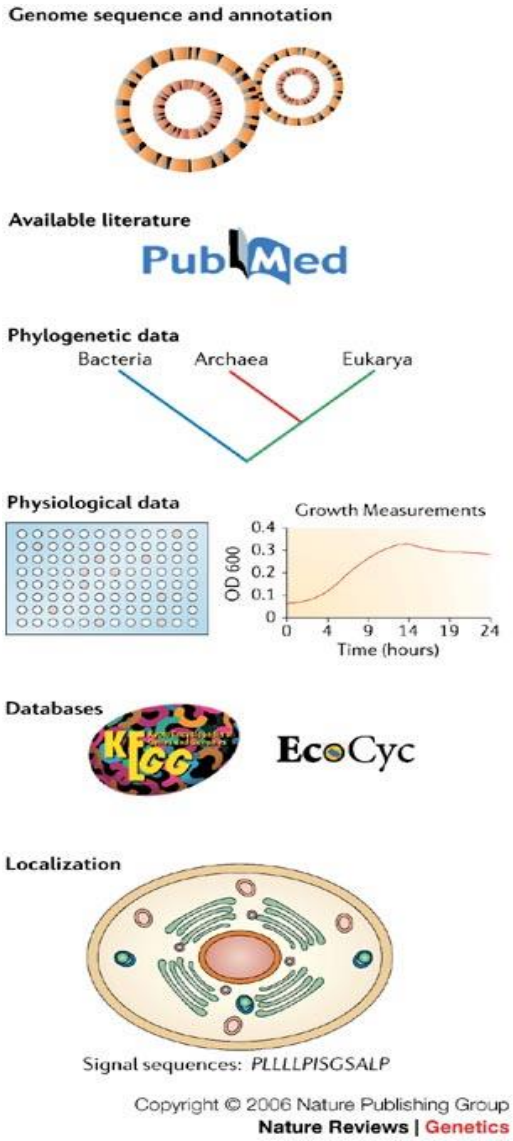
Experimental Platform

Affymetrix 208368_3p_s_at
Agilent A_23_P99452
CodeLink GE60169
Illumina GI_4502450-S

Red = Recommended

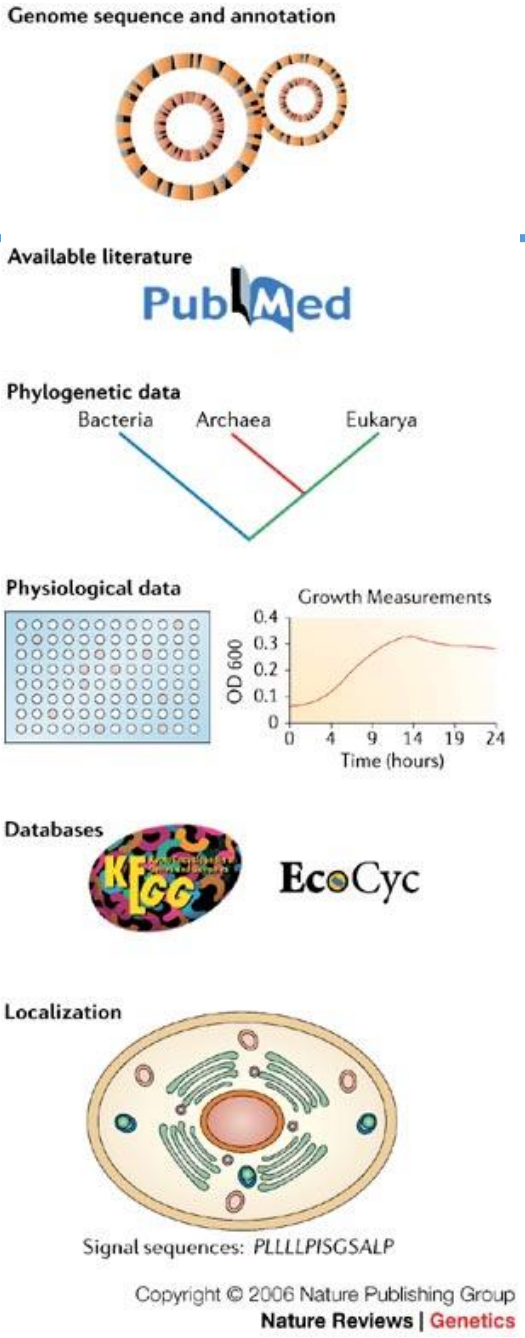
Levels of annotation per gene

Level	Database
Sequence	GenBank SwissProt (curated)
Metabolic pathways	Kegg Transpath MetaCyc
Literature	PubMed
Gene ontology (GO)	Biological process Molecular function Cell compartment
Integrated – Meta databases	GeneCards Entrez Gene OMIM InterPro

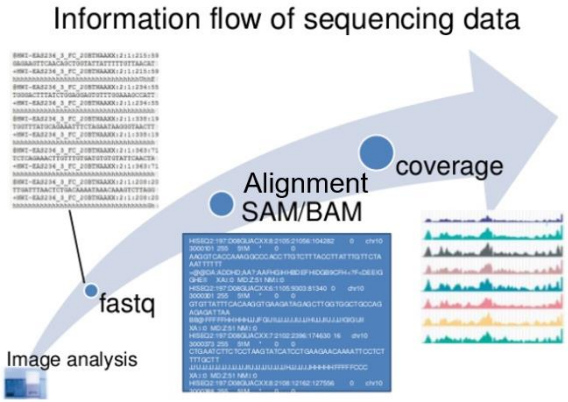


OUTLINE

- Single gene analysis / information
- Analysis of group of genes
- Gene ontology (GO)
- Enrichment analysis
  - Hypergeometric Test and Fisher exact test
  - GO Independence Assumption



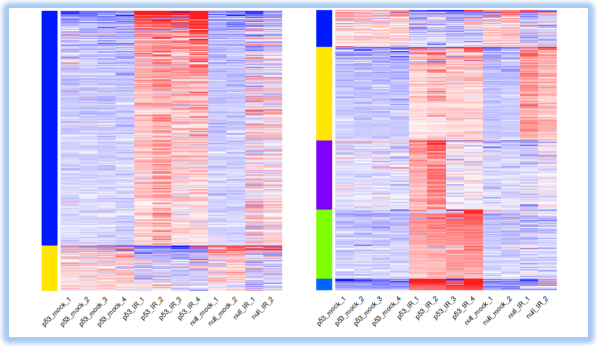
What have we done until now?



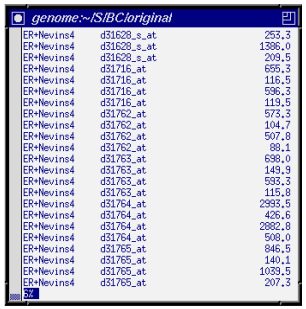
A complex high-throughput experiment:
Deep Sequencing
Proteomics
Microarrays
…

What did we get?

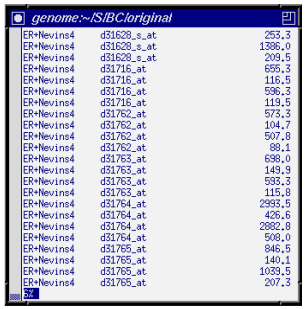
Lists of genes



Clusters of differential genes



Up regulated



Down regulated

Goal

To gain a higher view

not only deal with individual genes

perform Functional Enrichment Analysis

WHY?

Input:
a list of differentially expressed genes.

What do we need?

Output:
functions that are enriched in our set of differentially expressed genes.

Linking between genes and biological functions

The problem / challenge

- Vast amounts of biological data
- Different names/terms for the same concepts

  For example: the same function can be called

  translation or protein synthesis.

- Cross-species comparison is difficult

Part of the solution

Gene Ontology

GENEONTOLOGY
Unifying Biology

http://geneontology.org/

Ontology is a formal system for describing knowledge.

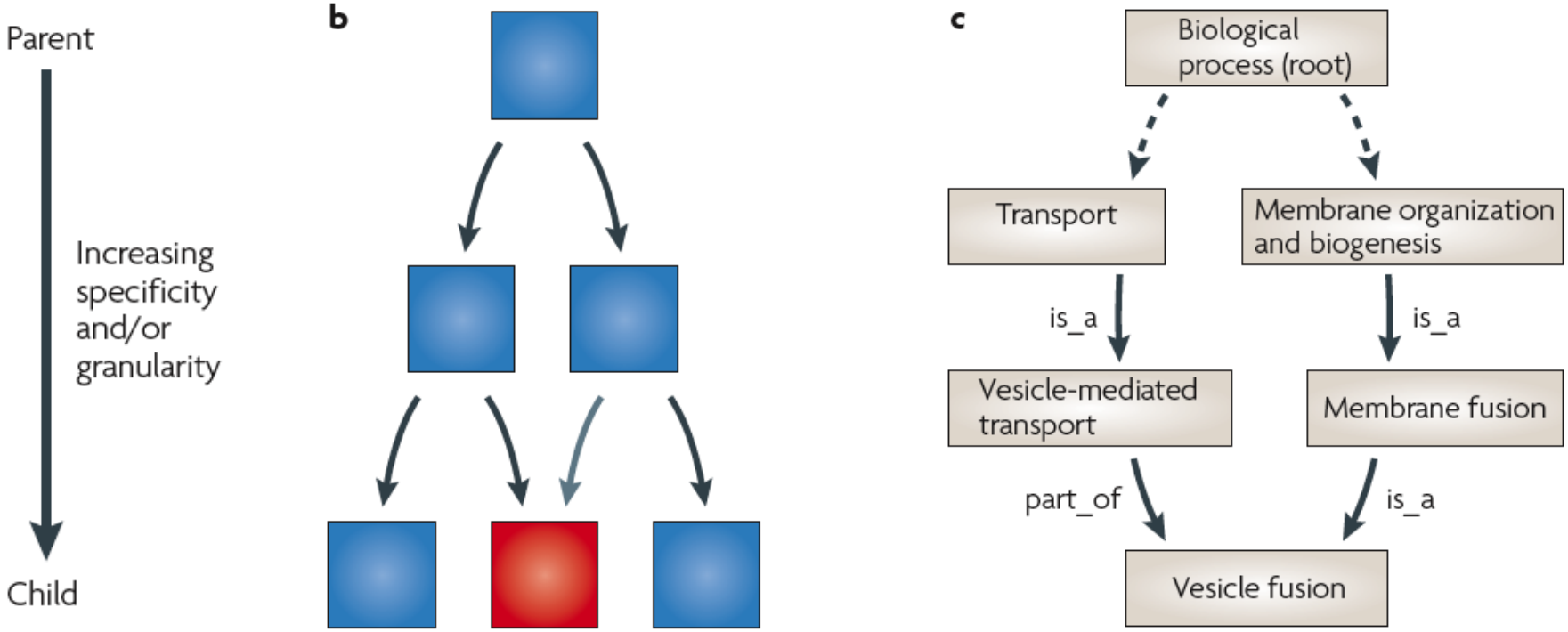
- Gene Ontology (GO) describes the functions of gene products, using a dictionary of allowed words.
- GO is a series of relations between controlled vocabulary terms.
- The knowledge is available to both humans and computers.

GENEONTOLOGY
Unifying Biology

There are three structured, controlled vocabularies (ontologies) that use terms to describe gene products in a species-independent manner:

- Biological processes
- Cellular components
- Molecular functions

Gene ontology is represented as a directed acyclic graph (DAG)

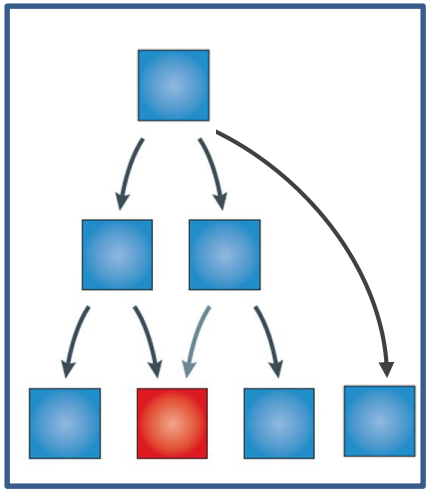


Taken from: Nature Reviews Genetics 9:509-515 (2008)

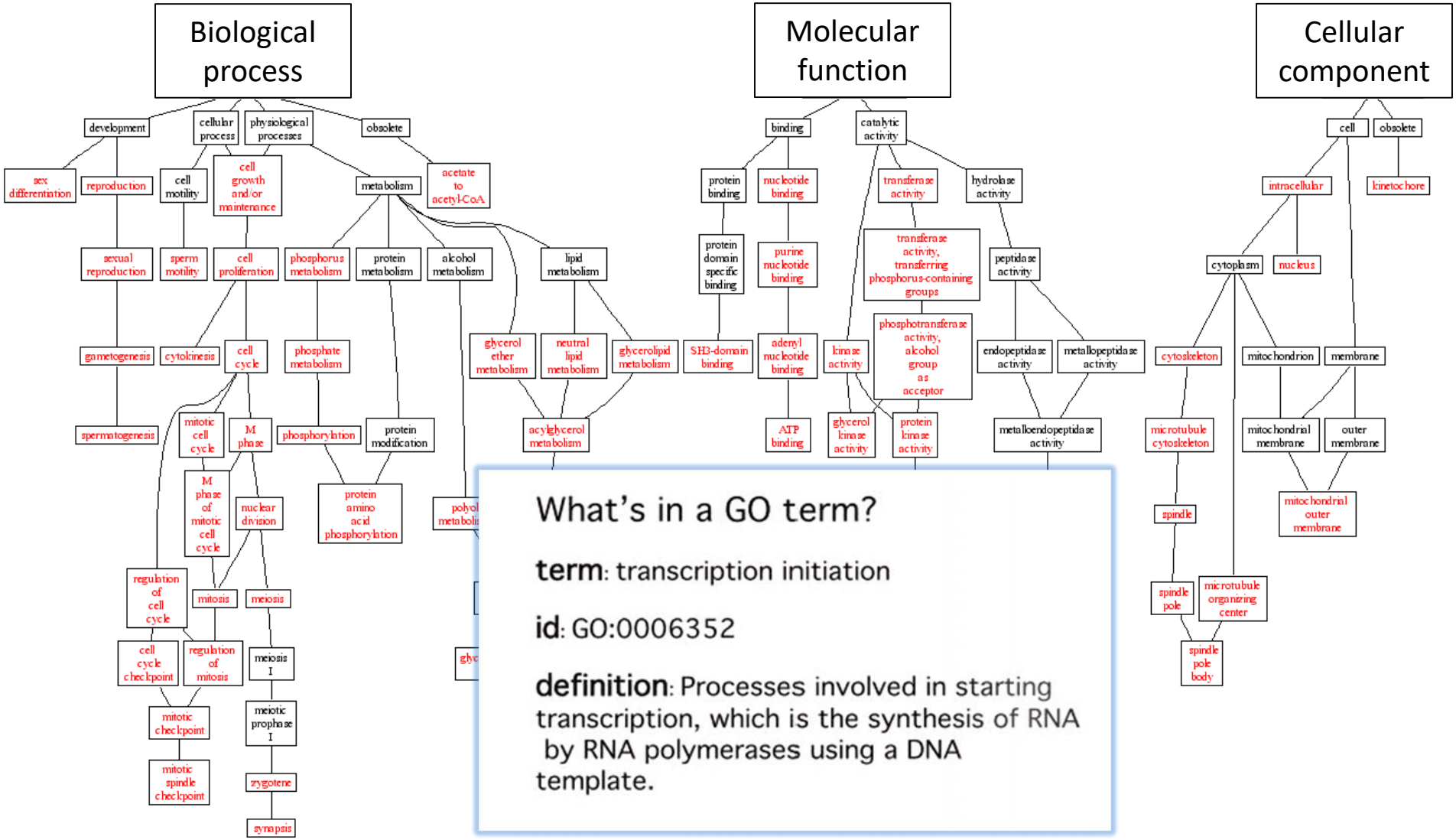
Gene Ontology is a series of relations between controlled vocabulary terms.

Directed Acyclic Graph (DAG)

- A child can have more than one parent
  - parents are closer to the root and are more general
  - children are further from the root and more specific
- There are no cycles - there is a root
- It is a directed graph
- You can skip levels in the graph



Example



What's in a GO term?

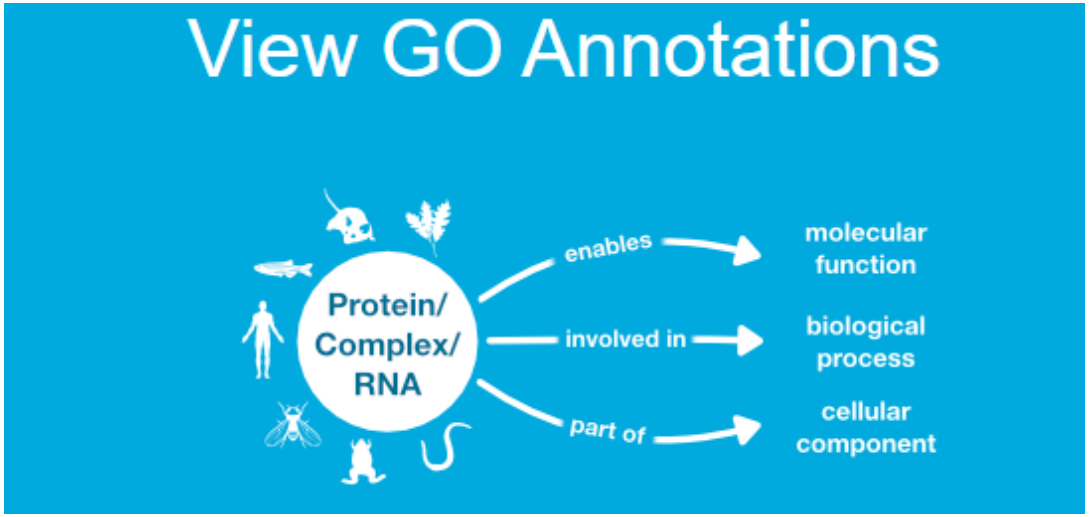
term: transcription initiation

id: GO:0006352

definition: Processes involved in starting transcription, which is the synthesis of RNA by RNA polymerases using a DNA template.

GO annotation

- A GO annotation is a statement about the function of a particular gene.
- GO annotations are associations made between gene products or protein complexes and the GO terms that describe them.
  - attributed to a source
  - indicate the evidence upon which it is based.



https://www.ebi.ac.uk/QuickGO/

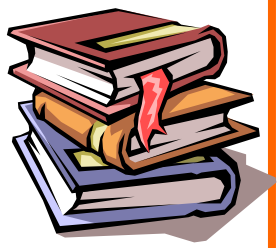
Evidence codes

not all annotations are created equal

	EXP	Inferred from Experiment
HTP		Inferred from High Throughput Experiment

BLAST

HDA	IDA	Inferred from Direct Assay
	IPI	Inferred from Physical Interaction
HMP	IMP	Inferred from Mutant Phenotype
HGI	IGI	Inferred from Genetic Interaction
HEP	IEP	Inferred from Expression Pattern

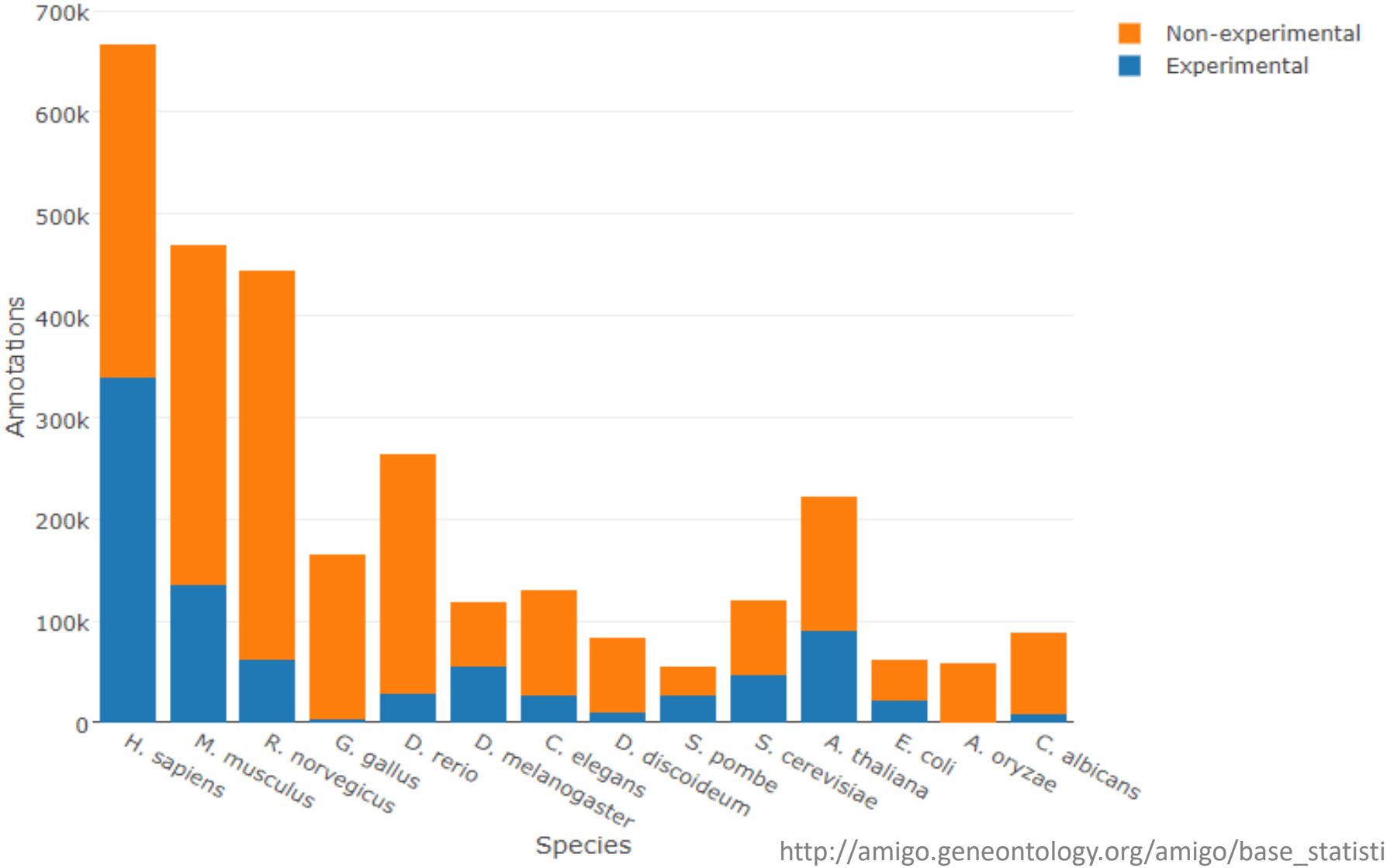


ISS	Inferred from Sequence/Structural Similarity
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
IC	Inferred by Curator
ND	No Data available

IEA	Inferred from electronic annotation

http://geneontology.org/docs/guide-go-evidence-codes/

Type of annotation per species



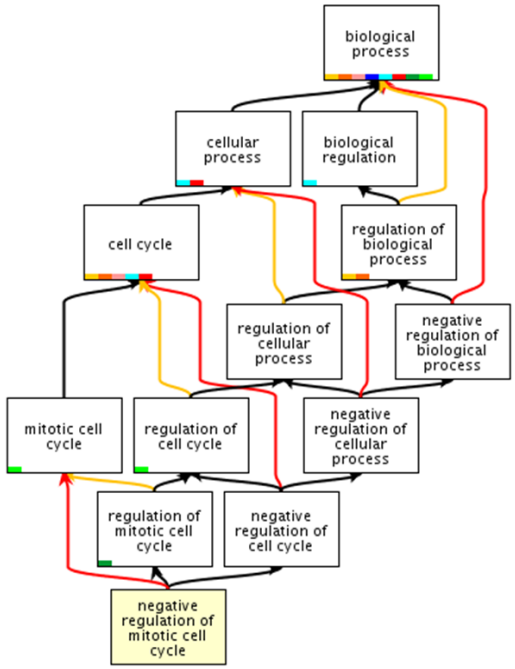
http://amigo.geneontology.org/amigo/base_statistics

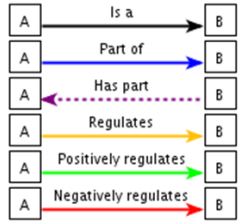
Ontology Relations

Defines the relationships (the arrows) between the ontology terms.

The commonly used relationships in GO:

- is_a
- part_of
- regulates:
  - positively regulates
  - negatively regulates

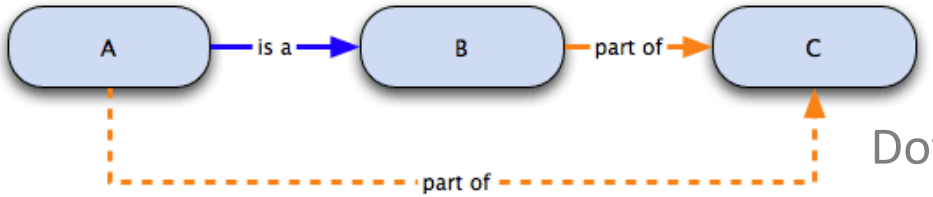




Ontology Relations

- is_a is a simple class-subclass relationship

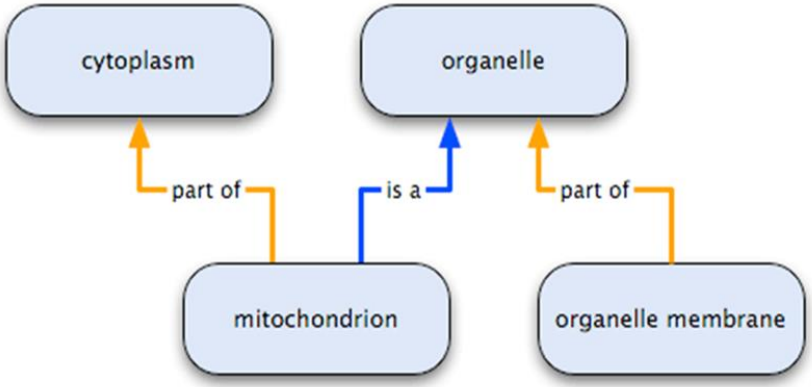
  Example: nuclear chromosome is_a chromosome.



Dotted line: an inferred relationship, e.g. one that has not been expressly stated

- part_of represent part-whole relationships;
  C part_of D means that whenever C is present, it is always a part of D.

  Example: nucleus part_of cell; nuclei are always part of a cell, but not all cells have nuclei.

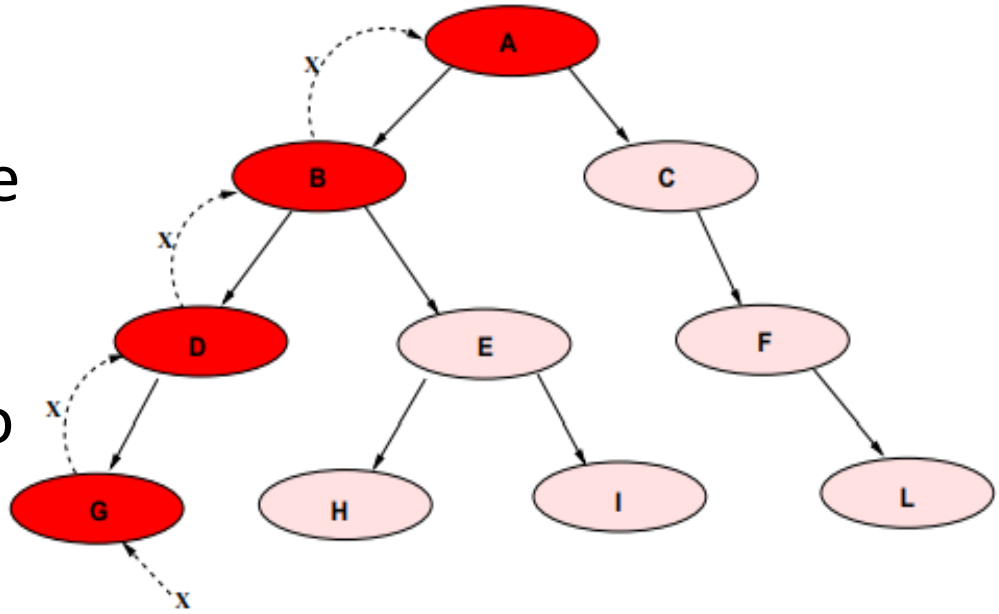


http://geneontology.org/docs/ontology-relations/

Ontology Structure

Every GO term obeys “the true path rule”:

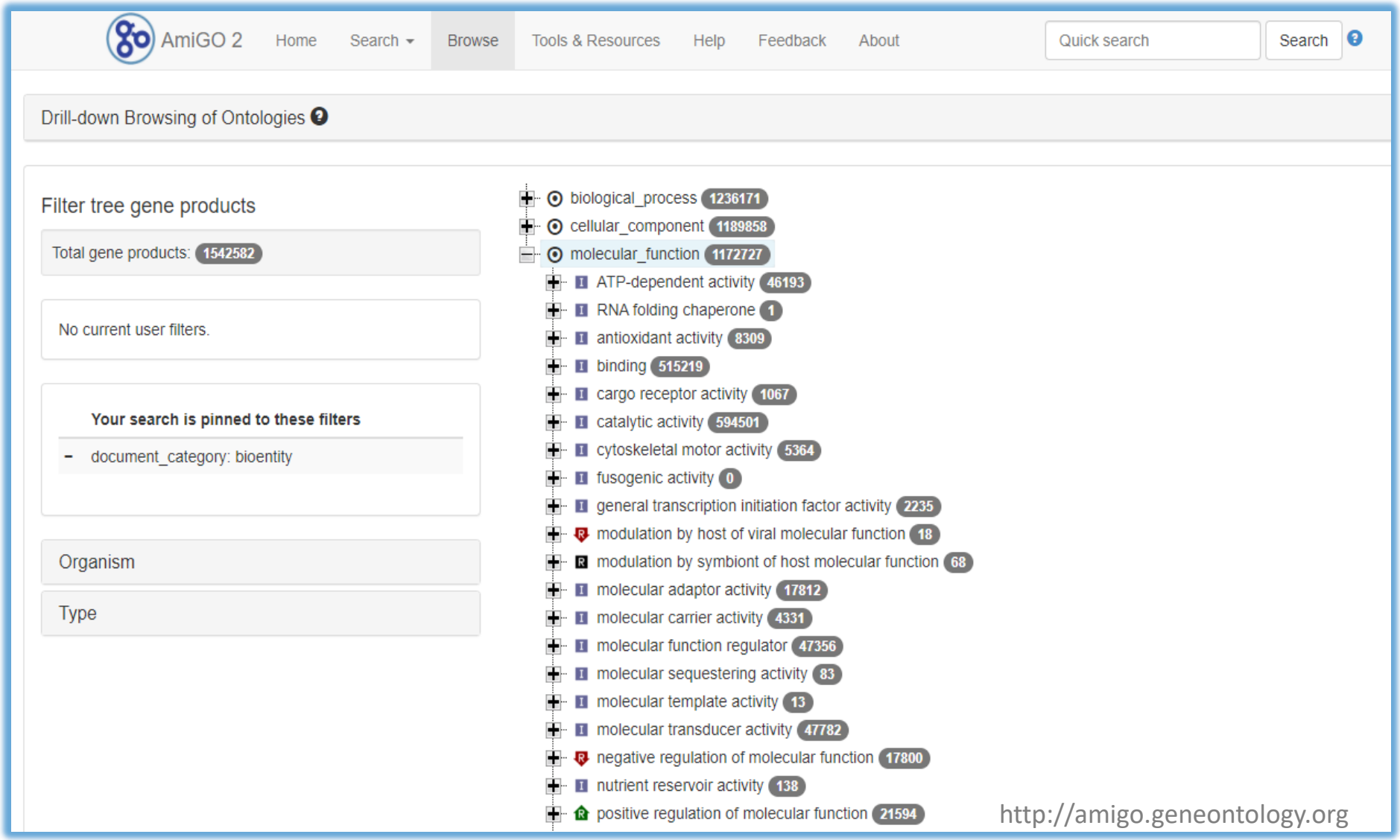
- If a child term describes the gene product, then all its ancestors (parent) terms must also apply to that gene product.
- If a gene is not annotated to a term, it cannot be annotated to its offsprings.



https://homes.di.unimi.it/~valentini/papers/vale.TPR.hier.revised.pdf

AmiGO

a web application to query, browse and visualize ontologies



http://amigo.geneontology.org

Available GO Information

Current ontology statistics: as of Nov., 2021:

43791 terms, 100.0% defined

- 28,438 Biological process terms
- 11,170 Molecular function terms
- 4,183 Cellular component terms
- 3444 obsolete terms (not included in terms above)

http://geneontology.org/stats.html

20 Questions

Which attribute is not a GO term?

Is it part of a complex?

Is it a protein coding gene?

Is it a regulator – transcription factor?

Is it in the nucleus?

Is it an enzyme?

Is it related to a disease?

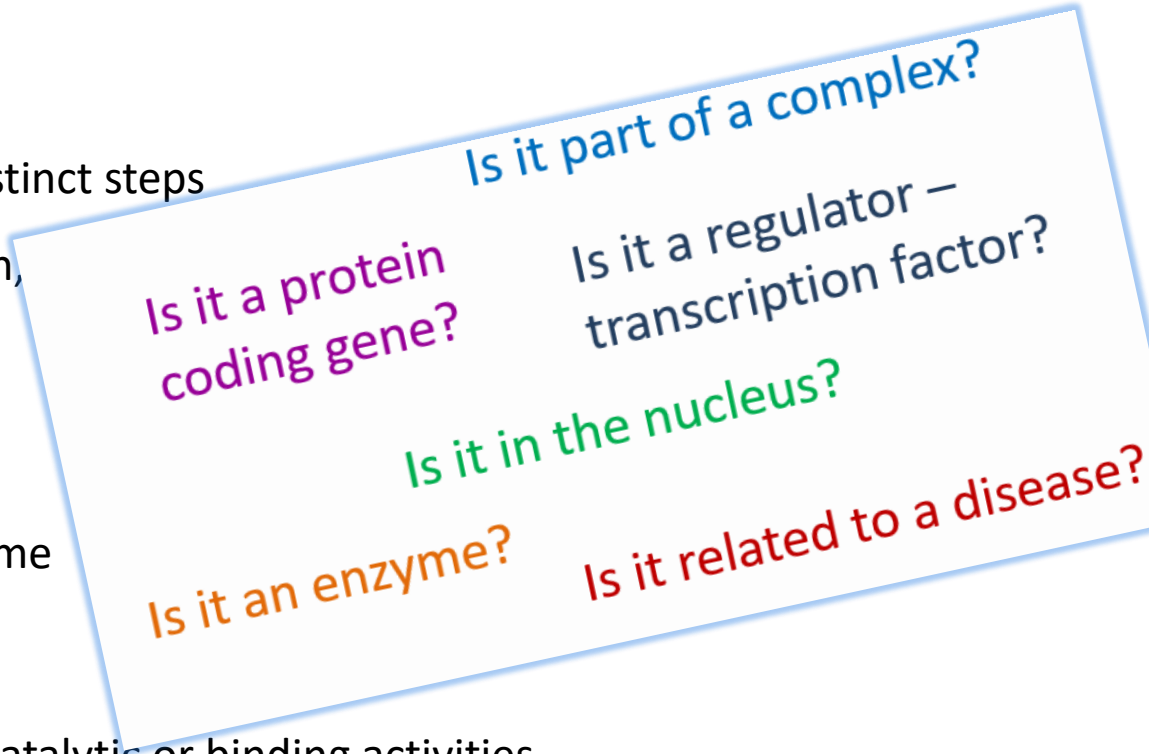
All the answers are “attributes” or characteristics of the item (gene).

Reminder

GENEONTOLOGY
Unifying Biology

There are three structured, controlled vocabularies (ontologies) that use terms to describe gene products in a species-independent manner:

- Biological processes
  - must have more than one distinct steps
  - Examples: signal transduction,
- Cellular components
  - an anatomical structure
  - Examples: nucleus, proteasome
- Molecular functions
  - describes activities, such as catalytic or binding activities



What is not GO?

- Gene products: e.g. cytochrome c is not in the ontologies, but attributes of cytochrome c, such as oxidoreductase activity, are
- Processes, functions or components that are unique to mutants or diseases: e.g. oncogenesis
- Attributes of sequence such as intron/exon parameters
- Protein domains or structural features
- Protein-protein interactions
- Environment, evolution and expression

- A pathway

GO Pitfalls

- Not complete, it is done “by hand” by curators
- Computational annotations
- Identifier flagged as ‘obsolete’, some tools do not update their databases
- NOT qualifier

Required in Gene Association File (GAF) 2.2 introduced in March 2021

("Qualifier") is now required

Annotation qualifiers

Some annotations are modified by qualifiers:

- *NOT* – gene product has been shown experimentally **not** to have the activity.

SKOR	NOT|enables	GO:0005242 F inward rectifier potassium channel activity

- *contributes_to* - Molecular function of individual subunits working as complex, in which no individual subunit has the activity. (Not a biological process).

- *colocalizes_with* - indicate a transient or peripheral association.

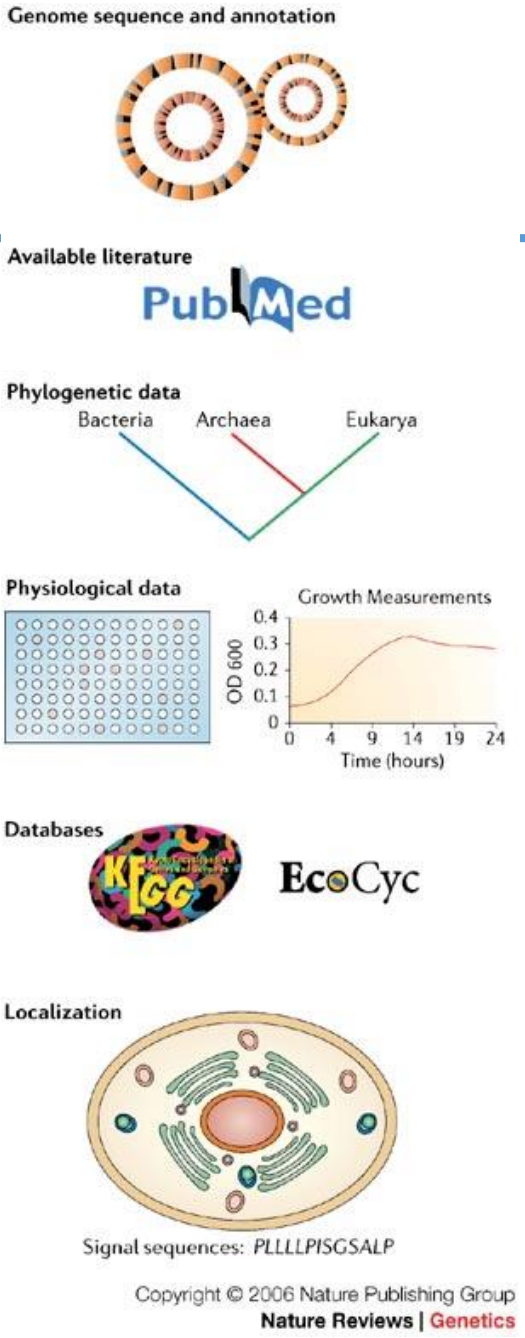
http://geneontology.org/docs/go-annotations/

Summary – GO

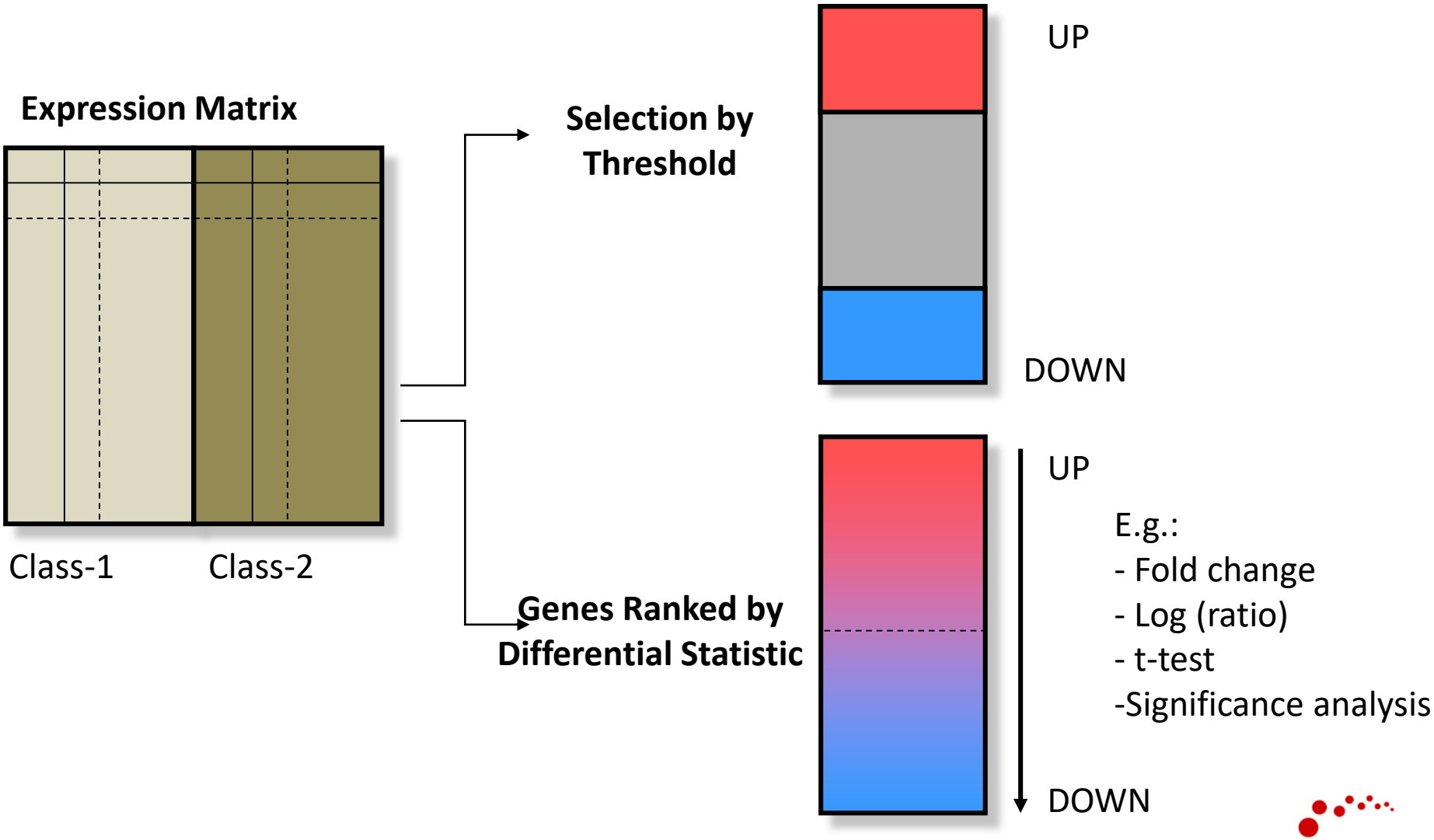
- Gene Ontology (GO) is human-readable and machine-readable
  - Ontology – a dictionary of related terms, for
    - Biological processes
    - Cellular compartments
    - Molecular functions
  - Annotation – Statement (based on evidence) associating a specific gene product to particular GO term.
- A gene can have more than one annotation term.
- The research community actively participates and contributes to the GO project.
- Annotation comes from manual and electronic sources.

OUTLINE

- Single gene analysis / information
- Analysis of group of genes
- Gene ontology (GO)
- Enrichment analysis
  - Hypergeometric Test and Fisher exact test
  - GO Independence Assumption



Two-class Design



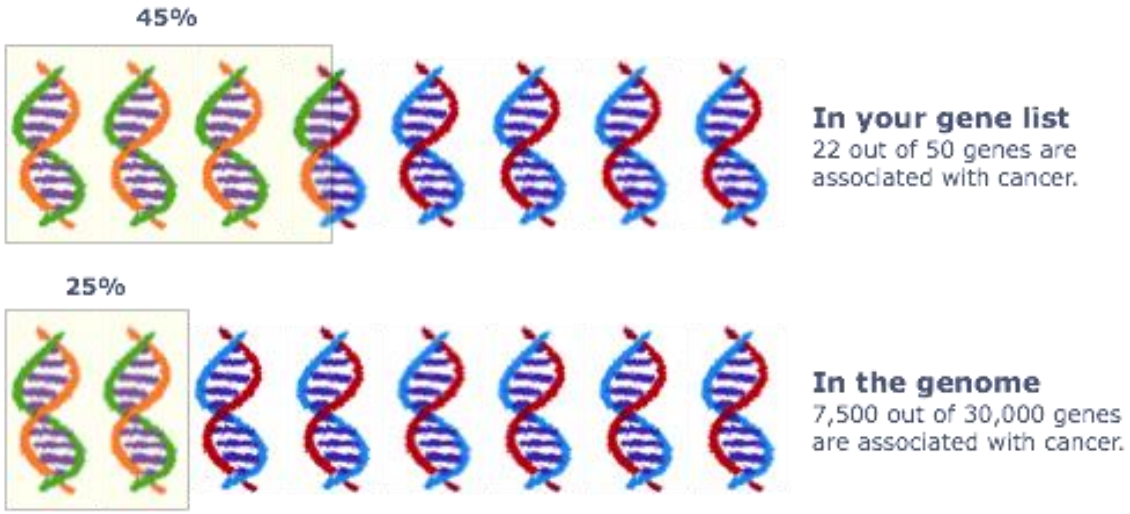
(modified) bioinformatics.ca

What is functional enrichment?

- It is a measure of how much a group of gene products is found in our data set
- It requires some type of background measure, as a basis for comparison
- What we look at is how many we have (observed) as opposed to how many we would expect to see at random, given our background.

Background

The choice of an appropriate background is critical to get meaningful results

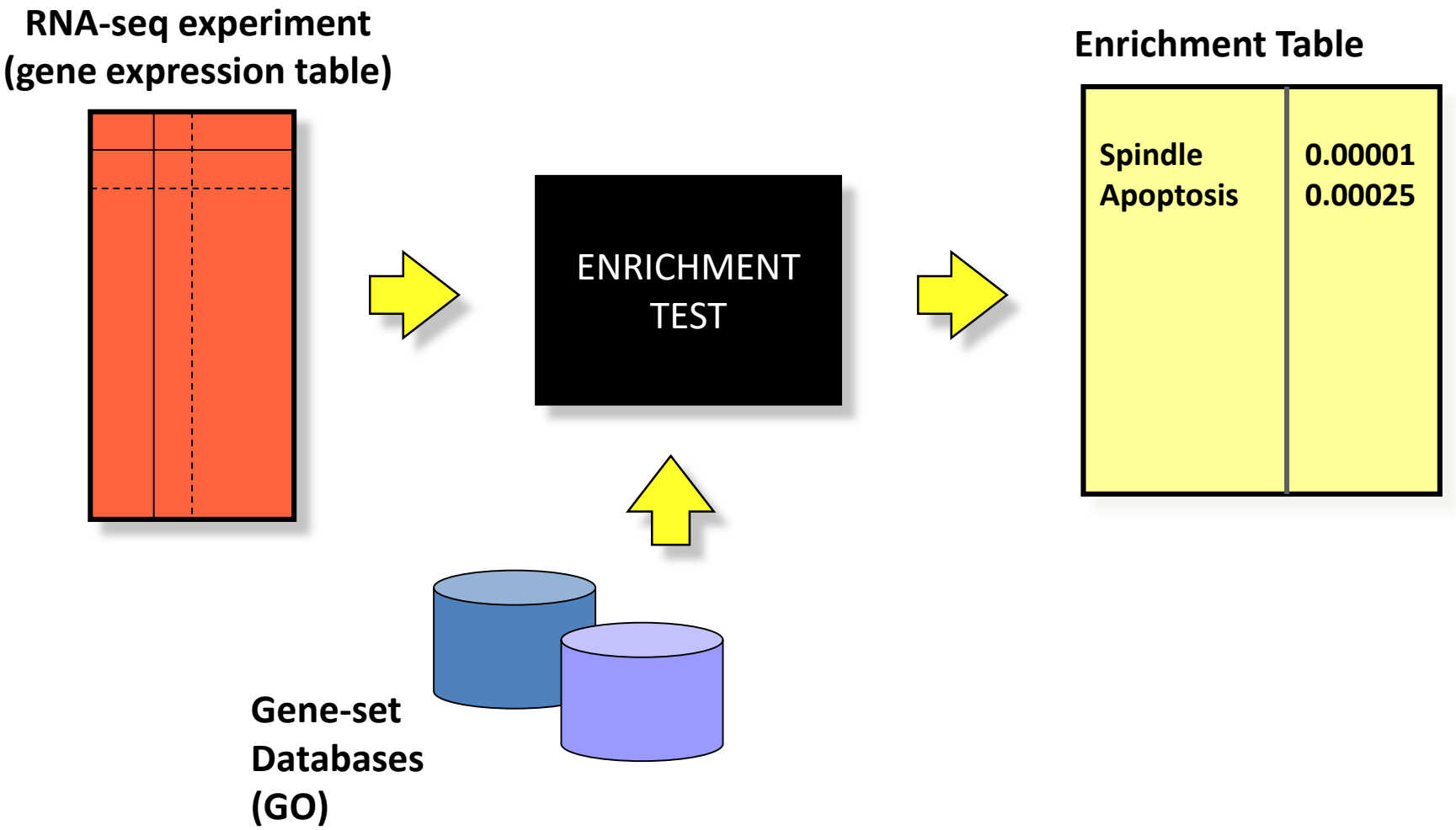


Fold of enrichment = 45% / 25% = 1.8

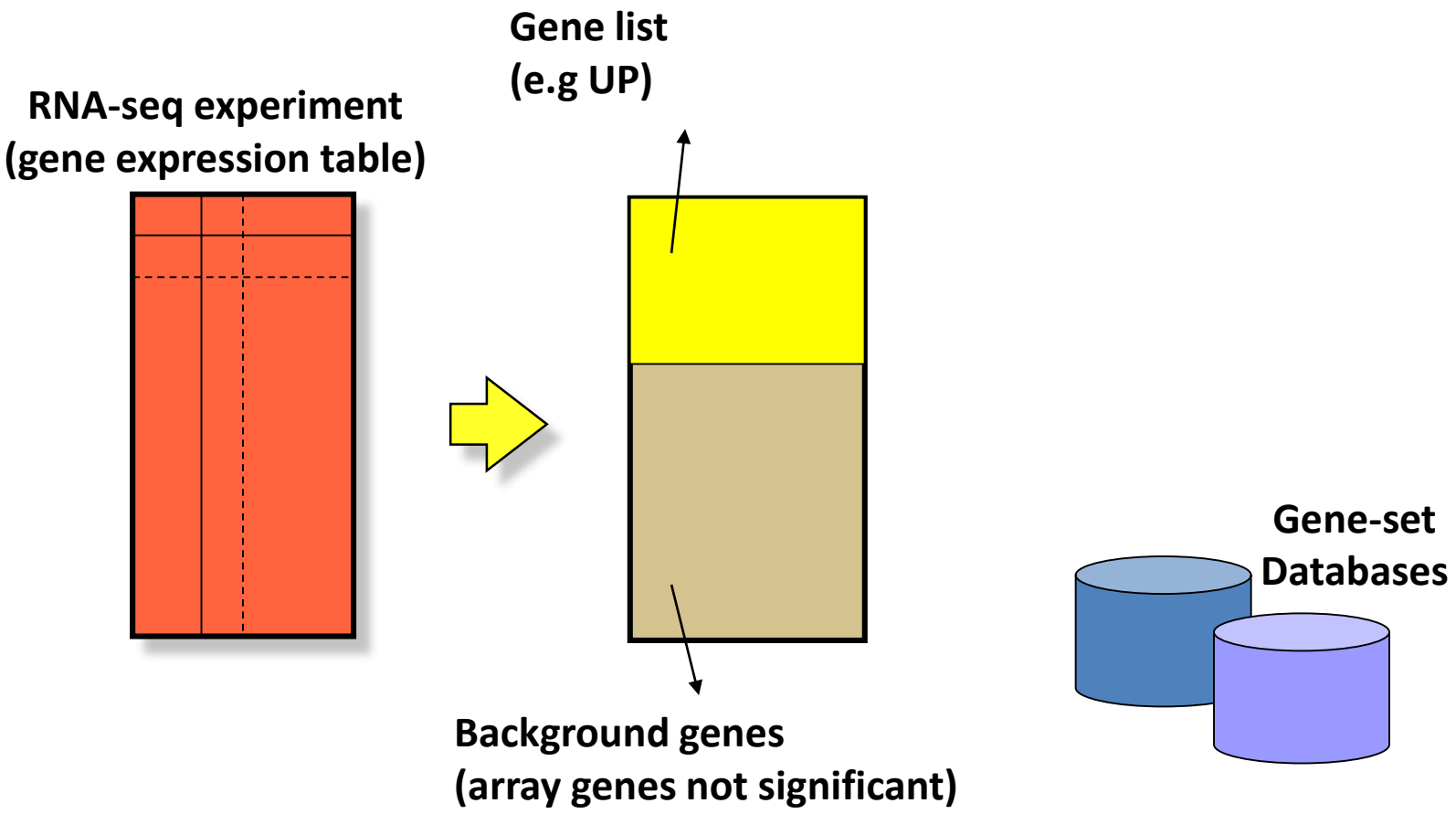
You should use all the genes detected by the method used in your experiment,

not all the genes in the genom, if possible.

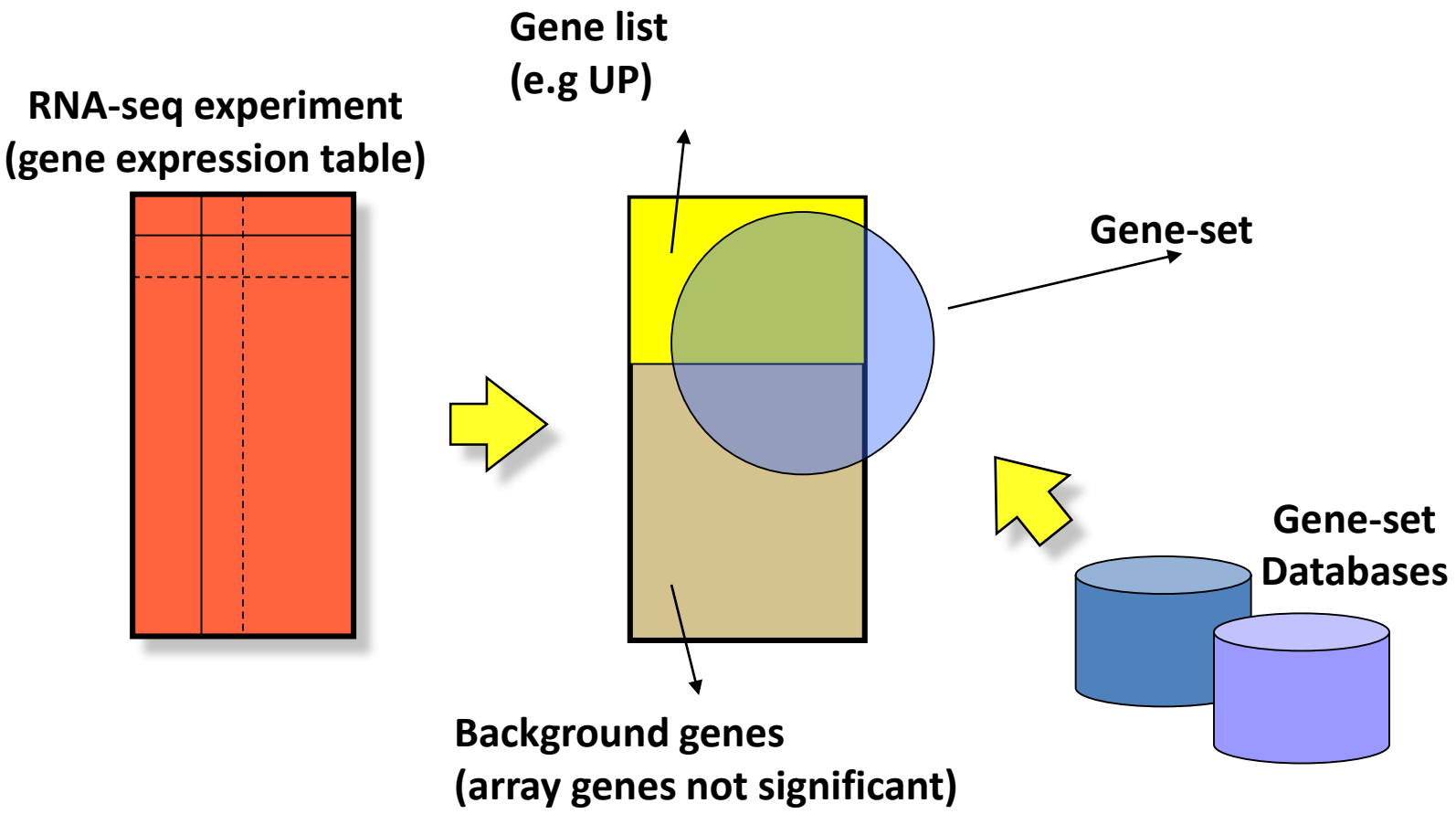
Enrichment test



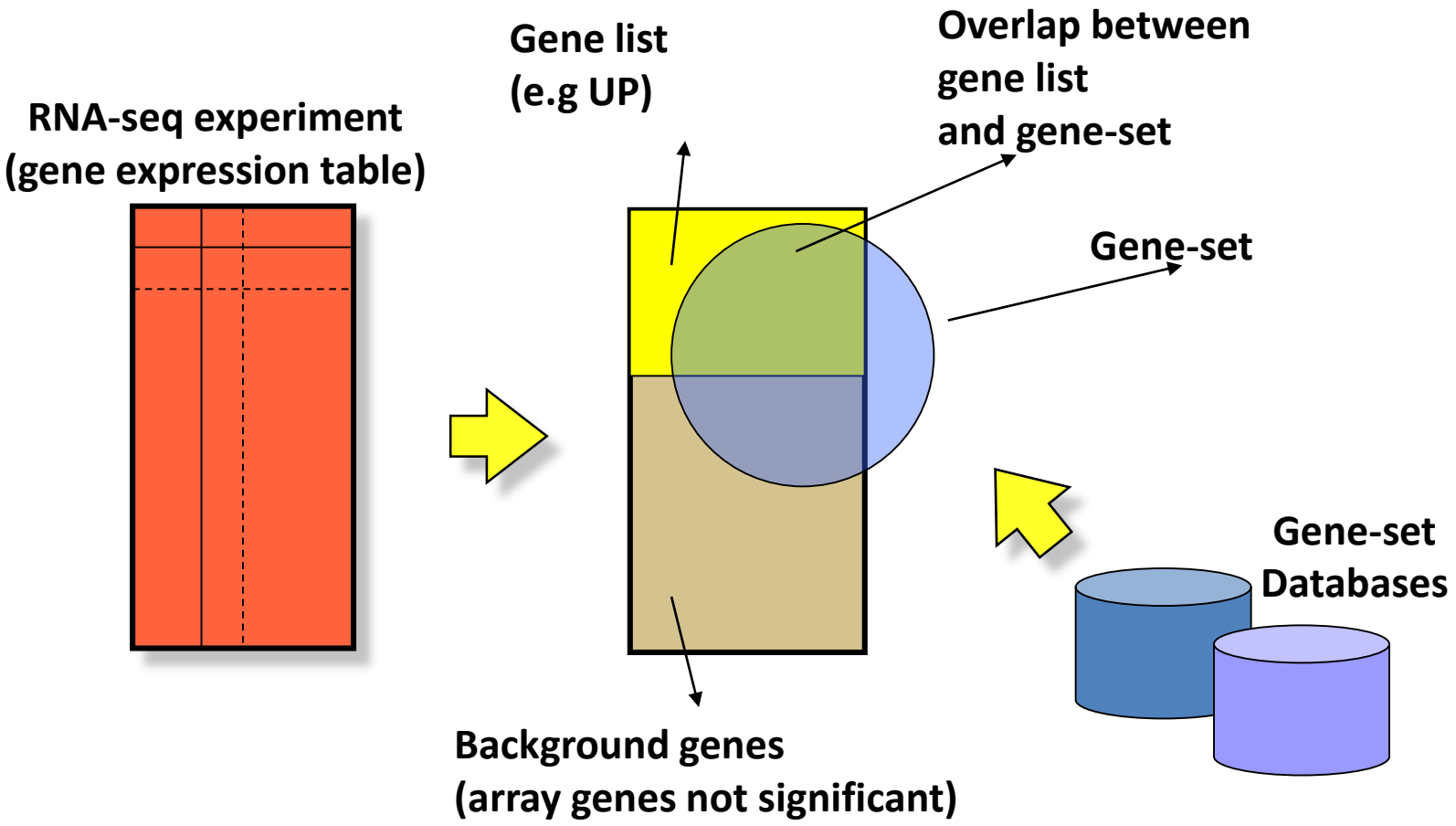
Enrichment test



Enrichment test



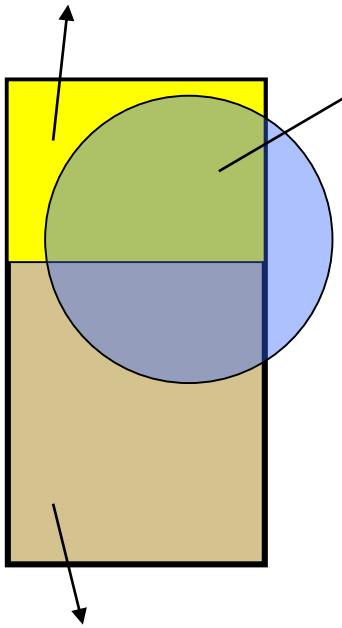
Enrichment test



Enrichment test

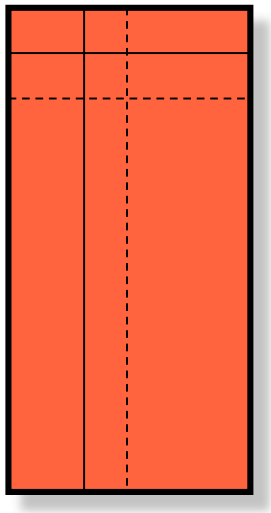
Significant genes (e.g UP)

Overlap between gene list and gene-set

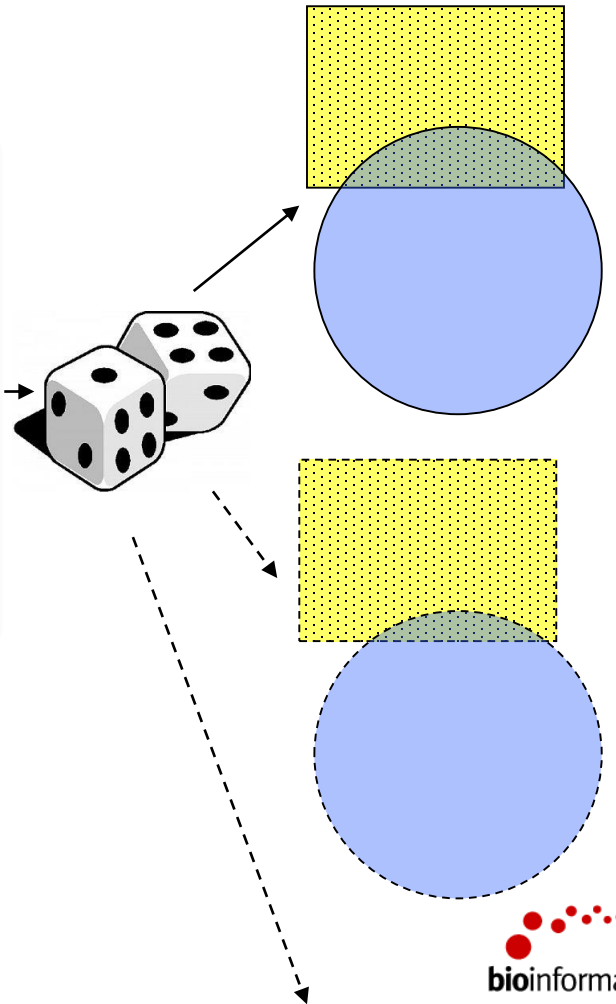


Background genes (array genes not significant)

Is this overlap larger than expected by ***random sampling*** *the array genes?*



Random samples of array genes



Enrichment analysis

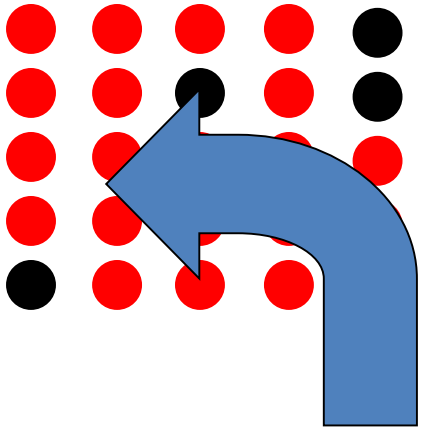
- Given:
  1. Gene list: e.g. RRP6, MRD1, RRP7, RRP43, RRP42
  2. Gene sets or annotations: e.g. Gene ontology, transcription factor binding sites in promoter

- Question: *Are any of the gene annotations surprisingly enriched in the gene list?*

- Details:
  - Where do the gene lists come from?
  - How to assess “surprisingly” (statistics)
  - How to correct for repeating the tests

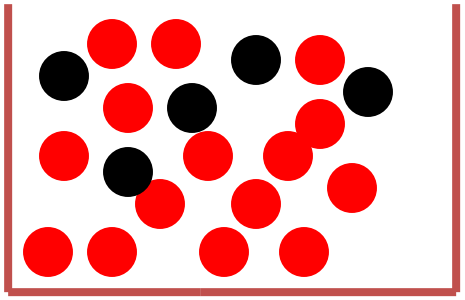
Randomization test

Random draws



… 7,834 draws later …

Expect a random draw with observed enrichment once every 1 / P-value draws



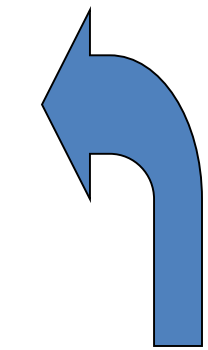
Background population:
500 black genes
4500 red genes

Fisher's exact test

a.k.a., the hypergeometric test

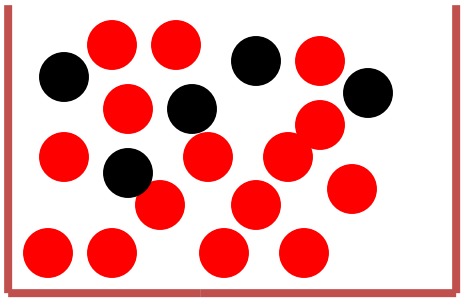
Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



Null hypothesis: List is a random sample from population

Alternative hypothesis: More black genes than expected



Background population:

500 black genes

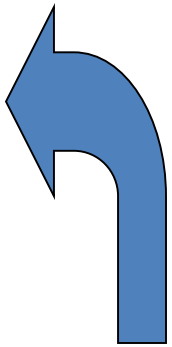
4500 red genes

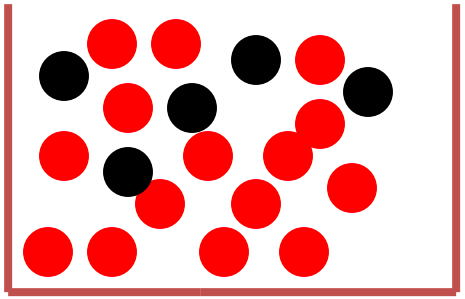
Fisher's exact test

a.k.a., the hypergeometric test

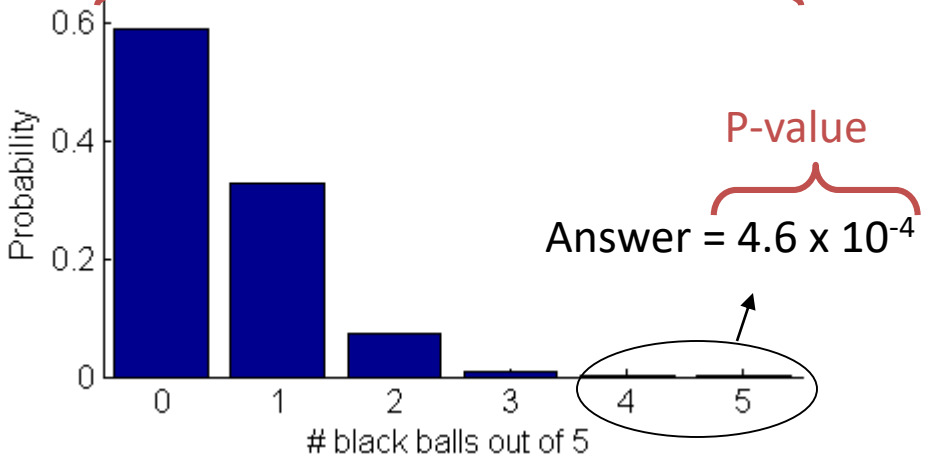
Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42





Null distribution



P-value

Answer = 4.6×10^{-4}

Background population:

500 black genes

4500 red genes

Group testing: Hypergeometric Test

Given:

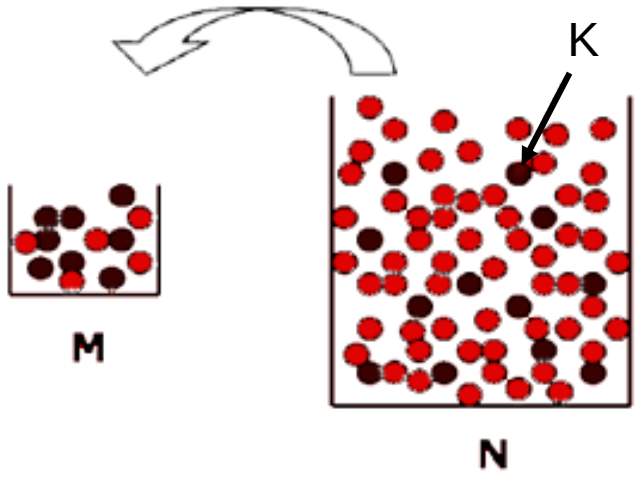
N genes identified in the experiment

M genes in a gene list (example:- up regulated genes)

K "interesting" genes (related to a specific GO annotation)

what is the probability of having x from K interesting genes in the gene list?

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$$



- A p-value for the gene list corresponds to $P(X \geq x \mid N;M;K)$

Fisher's Exact Test

- The hypergeometric test is equivalent to Fisher's exact test
- Fisher-test and similar tests based on gene counts are very often used in Gene Ontology analysis

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$$

	In gene list	**Not in gene list**	
In GO term	x	K-x	K
Not in GO term	M-x	(N-M) - (K-x)	N-K
	M	N-M	N

K "interesting" genes

N genes identified in the experiment

M genes in a gene list

Fisher's Exact Test- Example

Input the parameters to calculate the p-value for under- or over-enrichment based on the cumulative distribution function (CDF) of the hypergeometric distribution.

number of successes **k**		Genes related to GO term in our gene list
sample size **s**	300	Genes in a gene list (DE)
number of successes in the population **M**	100	Genes related to GO term
population size **N**	20000	Genes in the experiment

Submit Reset

https://systems.crump.ucla.edu/hypergeometric/index.php

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$$

Number of genes related to GO term in our gene list	p value	Conclusion
3	0.19	could be random
6	0.004	not likely random

Problems working with large data sets

- The more comparisons we make, the more there is a chance that we will get random hits
- We need to correct for multiple tests, using statistical methods such as Bonferroni, FDR (Benjamini-Hochberg)
- Statistical significance doesn’t necessarily mean biological significance

Reducing multiple test correction stringency

- The correction to the P-value threshold α depends on the # of tests that you do, so, no matter what, the more tests you do, the more sensitive the test needs to be

- Can control the stringency by reducing the number of tests: e.g. use GO slim; restrict testing to the appropriate GO annotations; or filter gene sets by size.

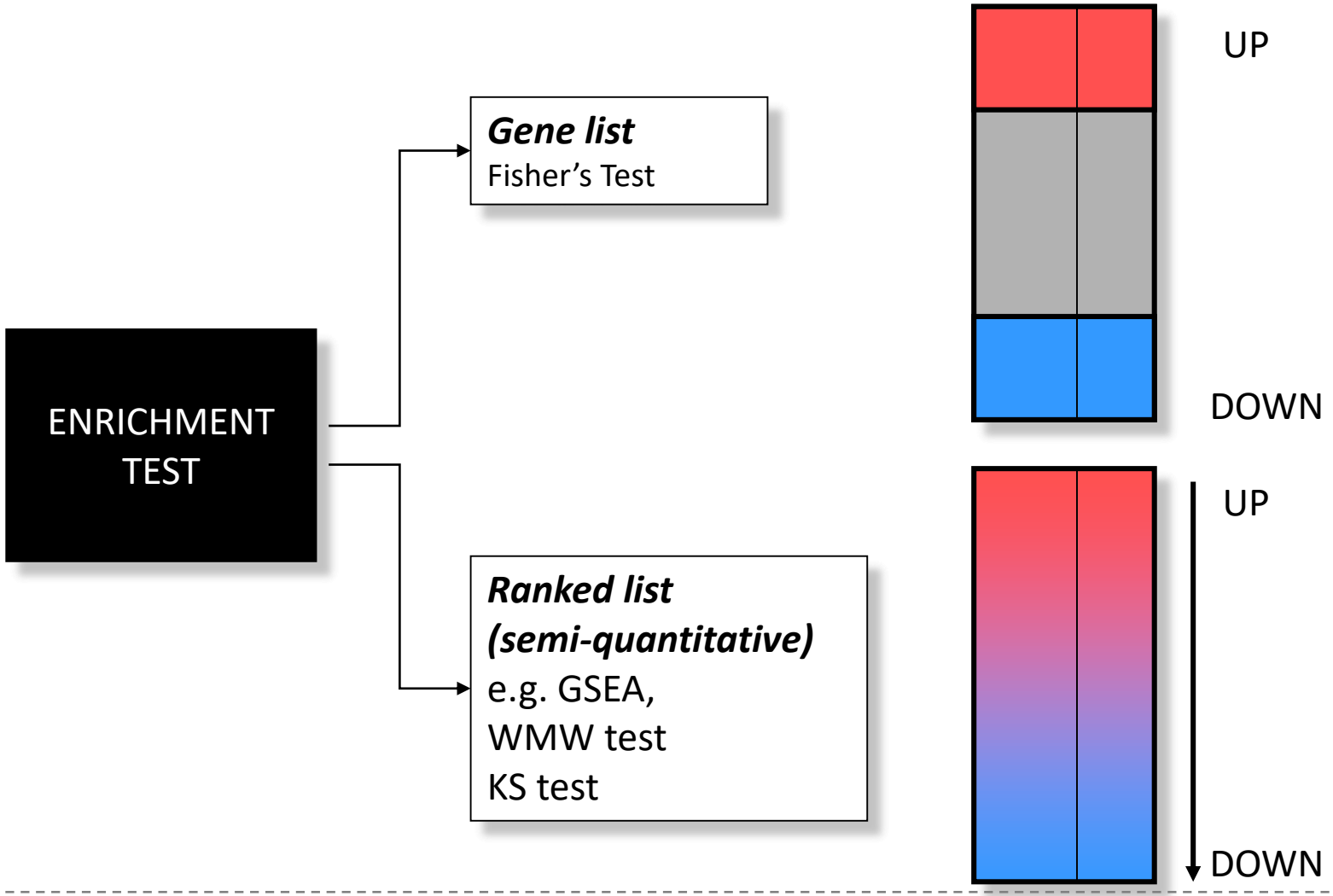
Beyond Fisher's Exact Test

Possible problems with Fisher's Exact Test:

- No "natural" value for the threshold
- Different results at different threshold settings
- Possible loss of statistical power due to thresholding
  - No resolution between significant signals with different strengths
  - Weak signals neglected

Solution: enrichment tests based on ranked lists

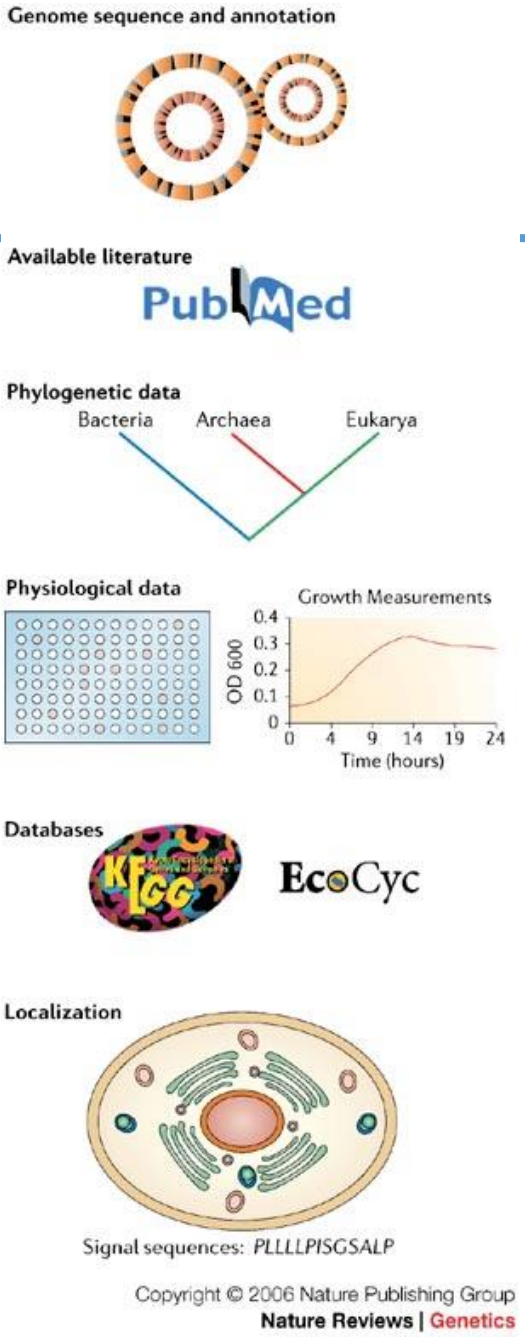
Beyond Fisher's Exact Test



WMW - the Wilcoxon-Mann-Whitney test KS - the Kolmogorov–Smirnov test

OUTLINE

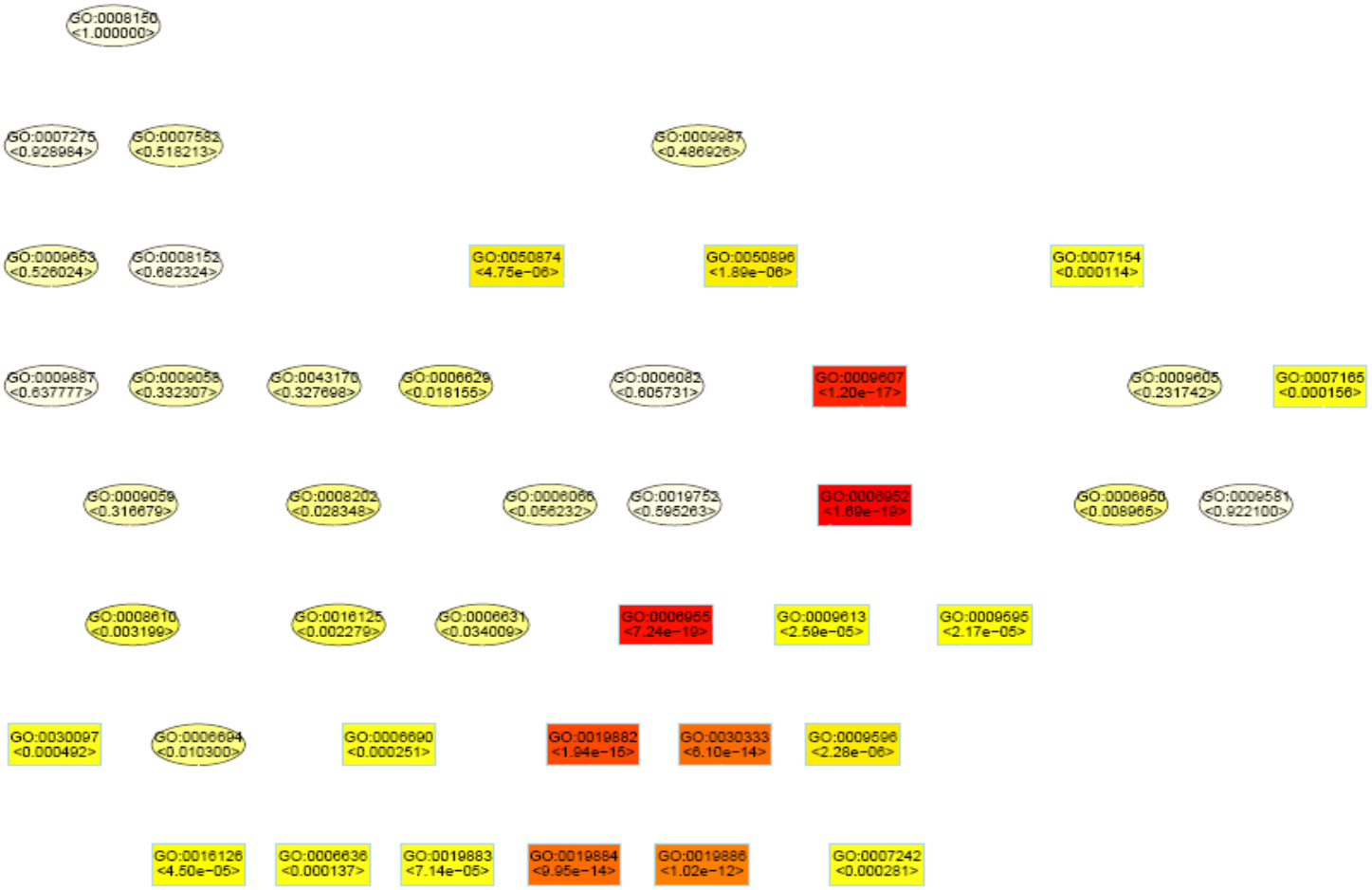
- Single gene analysis / information
- Analysis of group of genes
- Gene ontology (GO)
- Enrichment analysis
  - Hypergeometric Test and Fisher exact test
  - GO Independence Assumption



Term-for-term

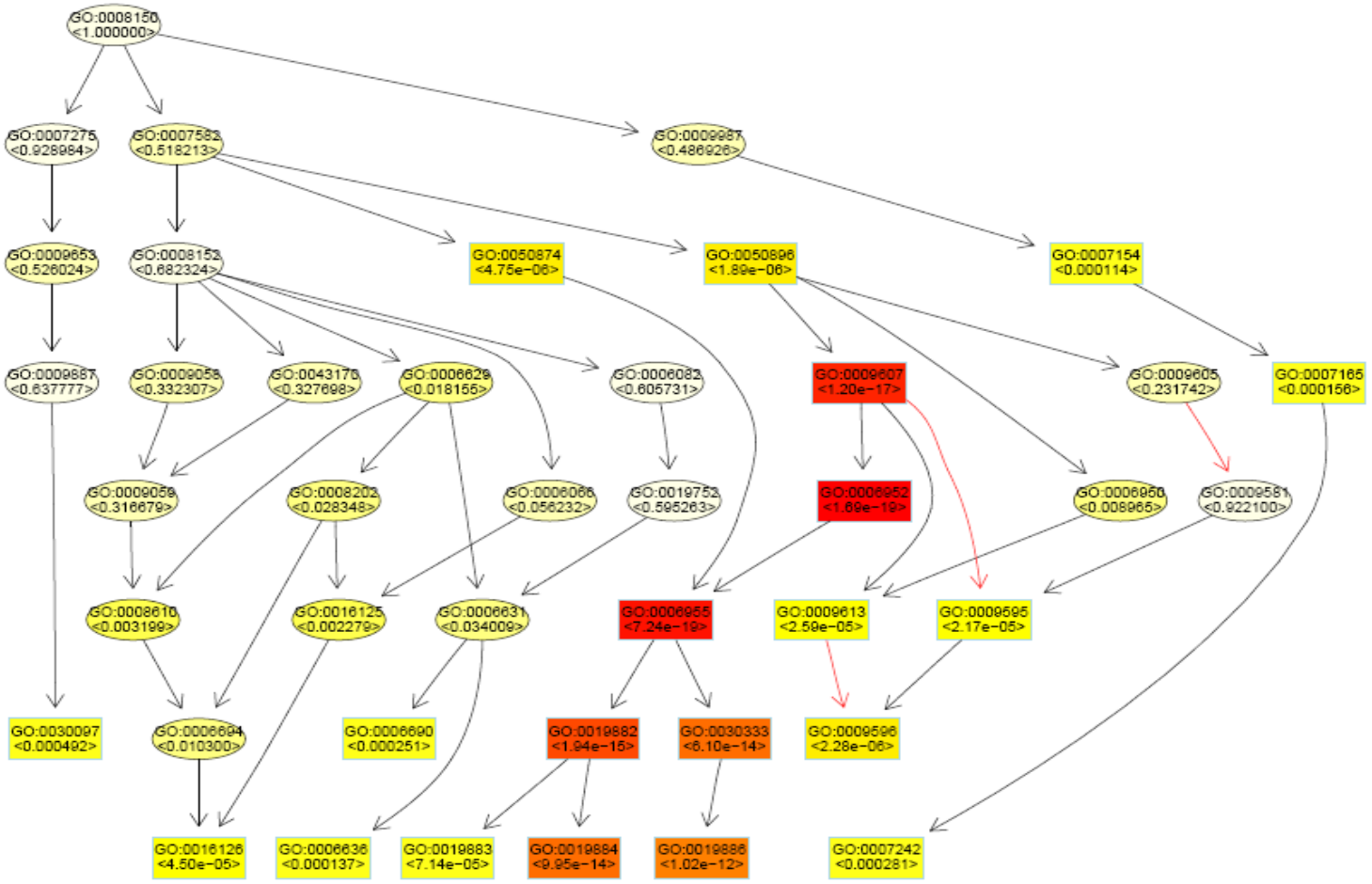
- The most common type of analysis
- Each term is considered independently of its neighbors in the GO tree
- Compares observed to expected and calculates significance

GO Independence Assumption



Note: The coloring of the nodes represent the *relative* significance of the GO terms: dark red is the most significant, light yellow is the least significant from the graph

GO Independence Assumption



Note: The coloring of the nodes represent the *relative* significance of the GO terms: dark red is the most significant, light yellow is the least significant from the graph

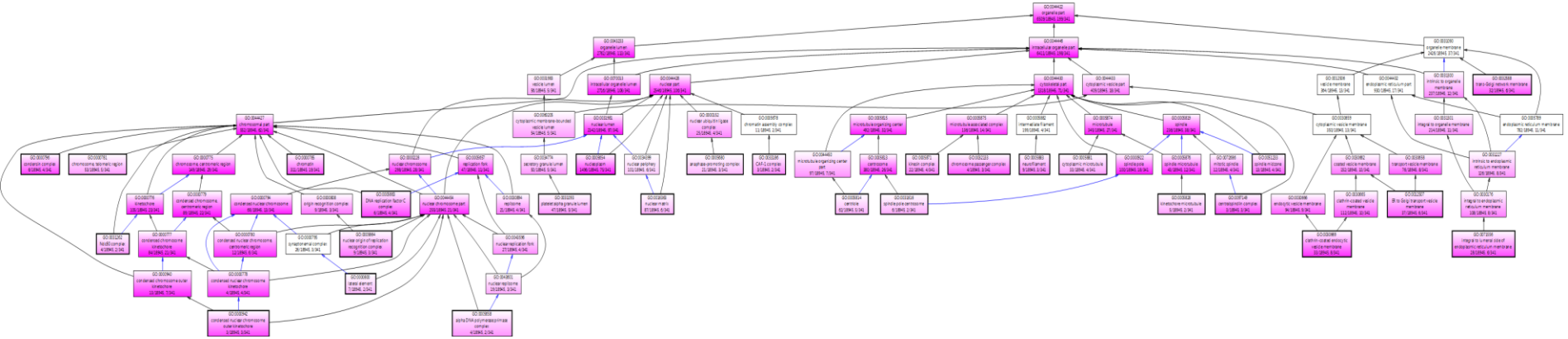
Algorithms review

- classic **algorithm**
  - Calculate significance of each GO term independently.
  - Adjust pvalues for multiple testing (Bonferroni, FDR, etc.).
  - Kolmogorov-Smirnov test can easily be used in this case
- elim **algorithm**
  - Nodes are processed bottom-up in the GO graph.
  - It iteratively removes the genes annotated to significant GO terms from more general GO terms.
  - Intuitive and simple to interpret.
- weight **algorithm**
  - The genes obtain weights that denote the gene relevance in the significant nodes.
  - To decide if a GO term u better represents the interesting genes, the enrichment score of node u is compared with the scores of its children.
  - Children with a better score than u better represent the interesting genes; their significance is increased
  - Children with a lower score than u have their significance reduced.

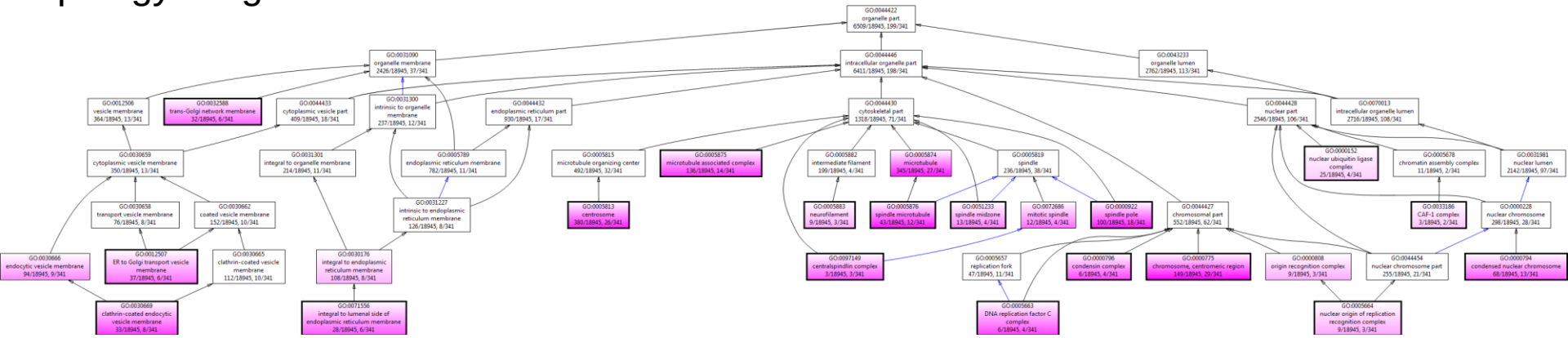
Alexa A, Rahnenführer J, Lengauer T. Bioinformatics. 2006 Jul 1;22(13):1600-7

Same input data – different results….

Term for term



Topology weighted



Many available tools for GO analysis

Summary of model organism web-based enrichment analysis tools

Tool	Interactive	Unique result URL	API	Ortholog conversion	User background upload	Model organisms
AmiGO	+		+			104
DAVID	+		+		+	65 000
g:Profiler	+	+	+	+	+	467
KOBAS						
LRpath						
Lynx						
modEnrichr	+	+	+	+		6
modPhEA	+			+	+	6
STRING	+	+	+			50
ToppFun	+	+		+		1
WebGestalt			+		+	12
WormBase	+					1

Nat Methods. 2016 Aug 30; 13(9): 705–706.
PMID: 27575621
doi: 10.1038/nmeth.3963

Impact of outdated gene annotations on pathway enrichment analysis

Lina Wadi,[1] Mona Meyer,[1] Joel Weiser,[1] Lincoln D Stein,[1,2] and Jüri Reimand[1,3]

Nucleic Acids Research, July 2019, https://doi.org/10.1093/nar/gkz347

Summary – part 1

- Gene Ontology (GO) is human-readable and machine-readable
  - Ontology – a dictionary of related terms, for
    - Biological processes
    - Cellular compartments
    - Molecular functions
  - Annotation – Statements (based on evidence) relating specific gene products to particular GO terms.

- Enrichment analysis assists in interpreting our high throughput experimental results.
  - Use the relevant background when possible.
  - Perform multiple test correction.

Thanks to:

Dr. Esti Feldmeser & Dr. Shifra Ben Dor

for interchanging and improving slides

for
your attention

Questions?