Working with Gene Lists and Over-representation Analysis

MBV-INF4410

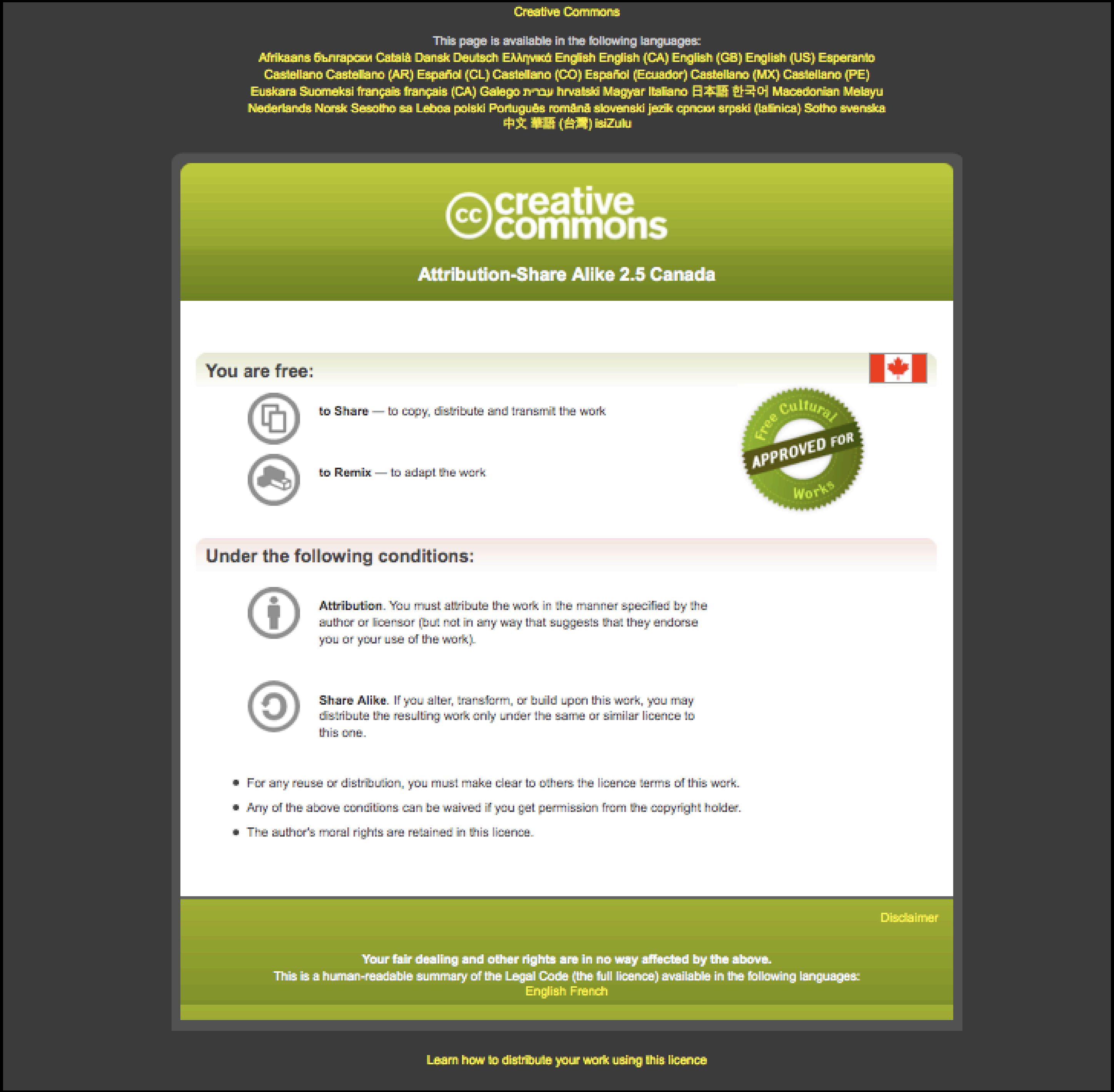
Friday, November 25th, 2011

Ian Donaldson

http://donaldson.uio.no

This talk is a remix of two talks presented in 2009 at the Canadian Bioinformatics Workshops by Gary Bader and Quaid Morris. Many thanks to Gary, Quaid and the CBW for making this material available.

http://baderlab.org
http://morrislab.med.utoronto.ca/
http://www.bioinformatics.ca/workshops/2009/course-content
(see Interpreting Gene Lists from -omics Studies.
Module 1: Introduction to gene lists (Chair: Gary Bader) and
Module 2: Finding over-represented gene functions in gene lists (Chair: Quaid Morris)



Creative Commons

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska 中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

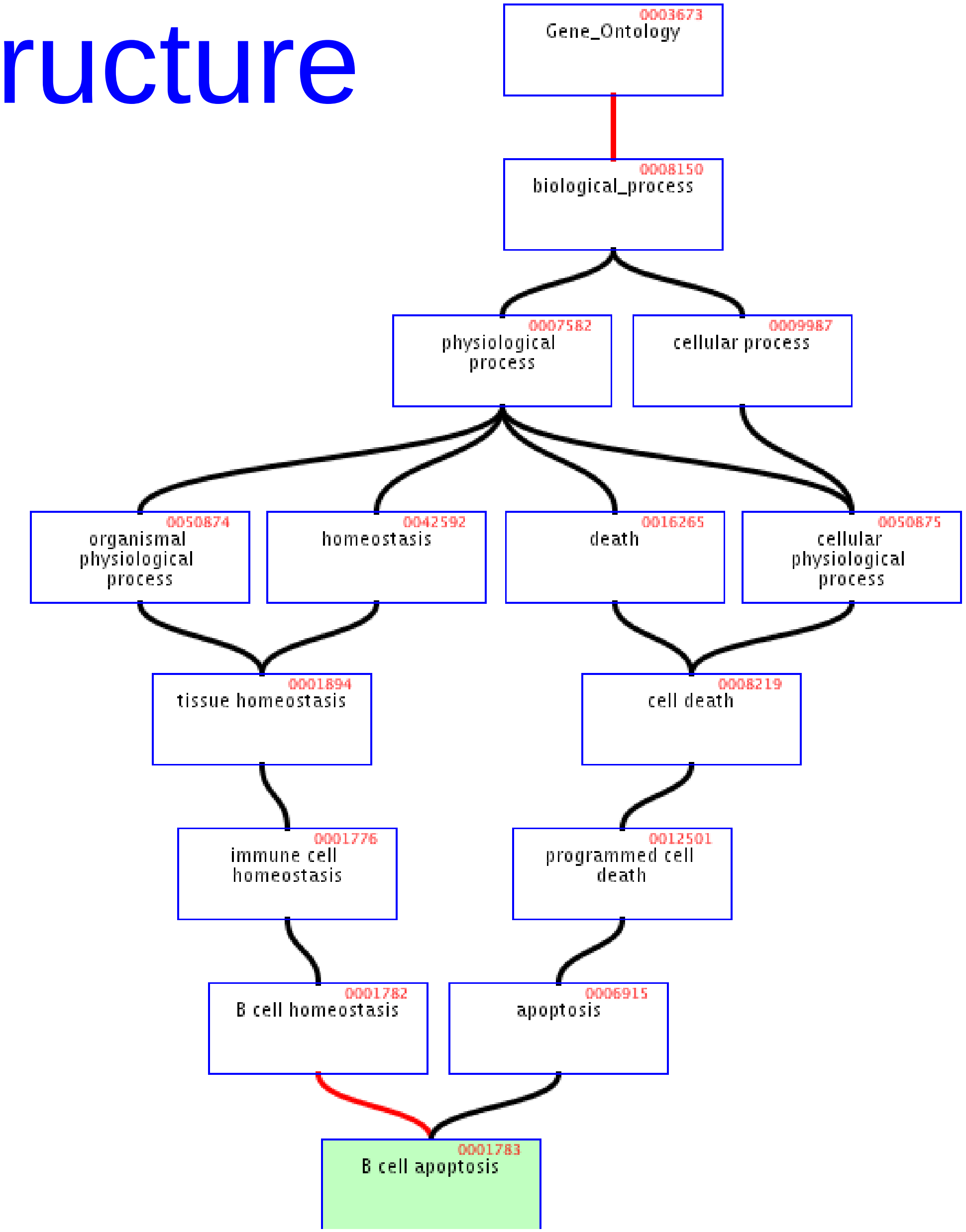
Learn how to distribute your work using this licence

What is the Gene Ontology (GO)?

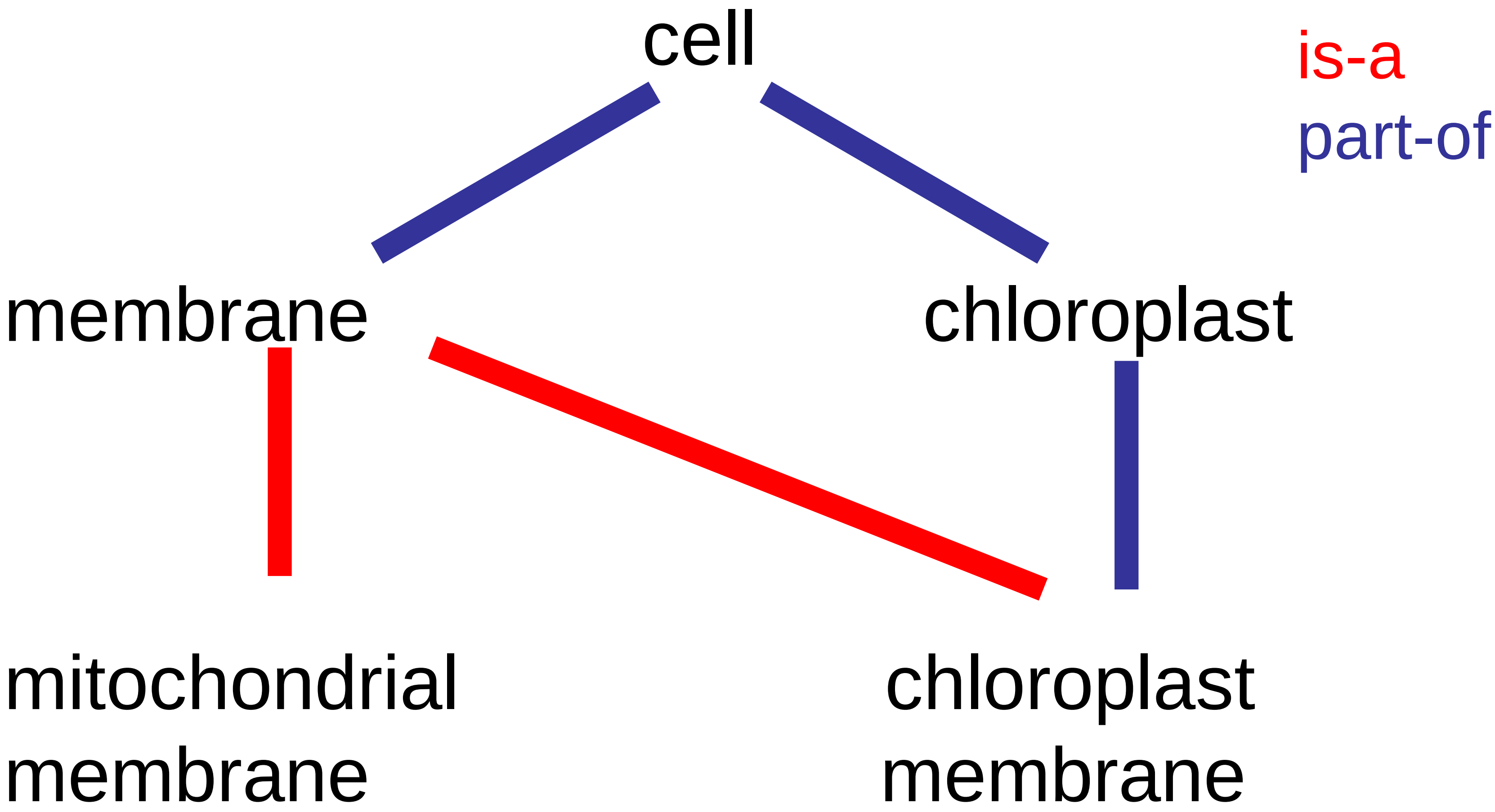
- Set of biological phrases (terms) which are applied to genes:
  - protein kinase
  - apoptosis
  - membrane
- Ontology: A formal system for describing knowledge

GO Structure

- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child



GO Structure



cell

is-a

part-of

membrane

chloroplast

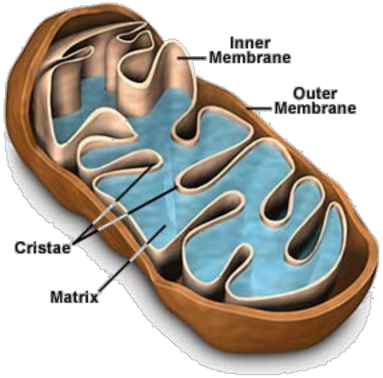
mitochondrial membrane

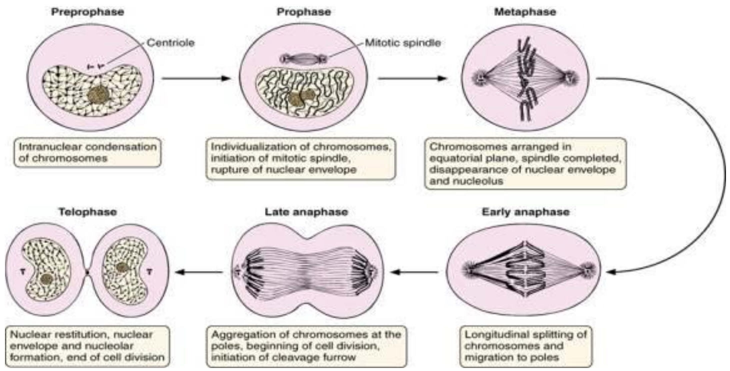
chloroplast membrane

Species independent. Some lower-level terms are specific to a group, but higher level terms are not

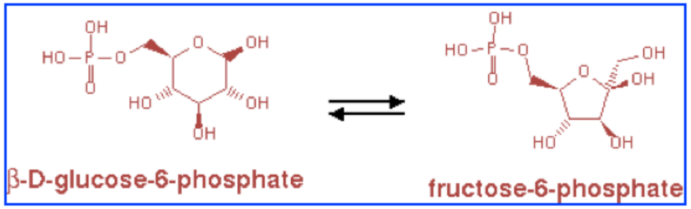
What GO Covers?

- GO terms divided into three aspects:
  - cellular component
  - molecular function
  - biological process




Preprophase
Centriole
Intranuclear condensation of chromosomes
Prophase
Mitotic spindle
Individualization of chromosomes, initiation of mitotic spindle, rupture of nuclear envelope
Metaphase
Chromosomes arranged in equatorial plane, spindle completed, disappearance of nuclear envelope and nucleolus
Early anaphase
Longitudinal splitting of chromosomes and migration to poles
Late anaphase
Aggregation of chromosomes at the poles, beginning of cell division, initiation of cleavage furrow
Telophase
Nuclear restitution, nuclear envelope and nucleolar formation, end of cell division

Cell division



glucose-6-phosphate isomerase activity

Terms

- Where do GO terms come from?
  - GO terms are added by editors at EBI and gene annotation database groups
  - Terms added by request
  - Experts help with major development
  - 27734 terms, 98.9% with definitions.
    - 16731 biological_process
    - 2385 cellular_component
    - 8618 molecular_function
    - As of July 6, 2009

Annotations

- Genes are linked, or associated, with GO terms by trained curators at genome databases
  - Known as ‘gene associations’ or GO annotations
  - Multiple annotations per gene
- Some GO annotations created automatically

Annotation Sources

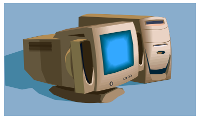
- Manual annotation
  - Created by scientific curators
    - High quality
    - Small number (time-consuming to create)
- Electronic annotation
  - Annotation derived without human validation
    - Computational predictions (accuracy varies)
    - Lower ‘quality’ than manual codes
- Key point: be aware of annotation origin

For your information

Evidence Types

- **ISS**: Inferred from Sequence/Structural Similarity
- **IDA**: Inferred from Direct Assay
- **IPI**: Inferred from Physical Interaction
- **IMP**: Inferred from Mutant Phenotype
- **IGI**: Inferred from Genetic Interaction
- **IEP**: Inferred from Expression Pattern
- **TAS**: Traceable Author Statement
- **NAS**: Non-traceable Author Statement
- **IC**: Inferred by Curator
- **ND**: No Data available

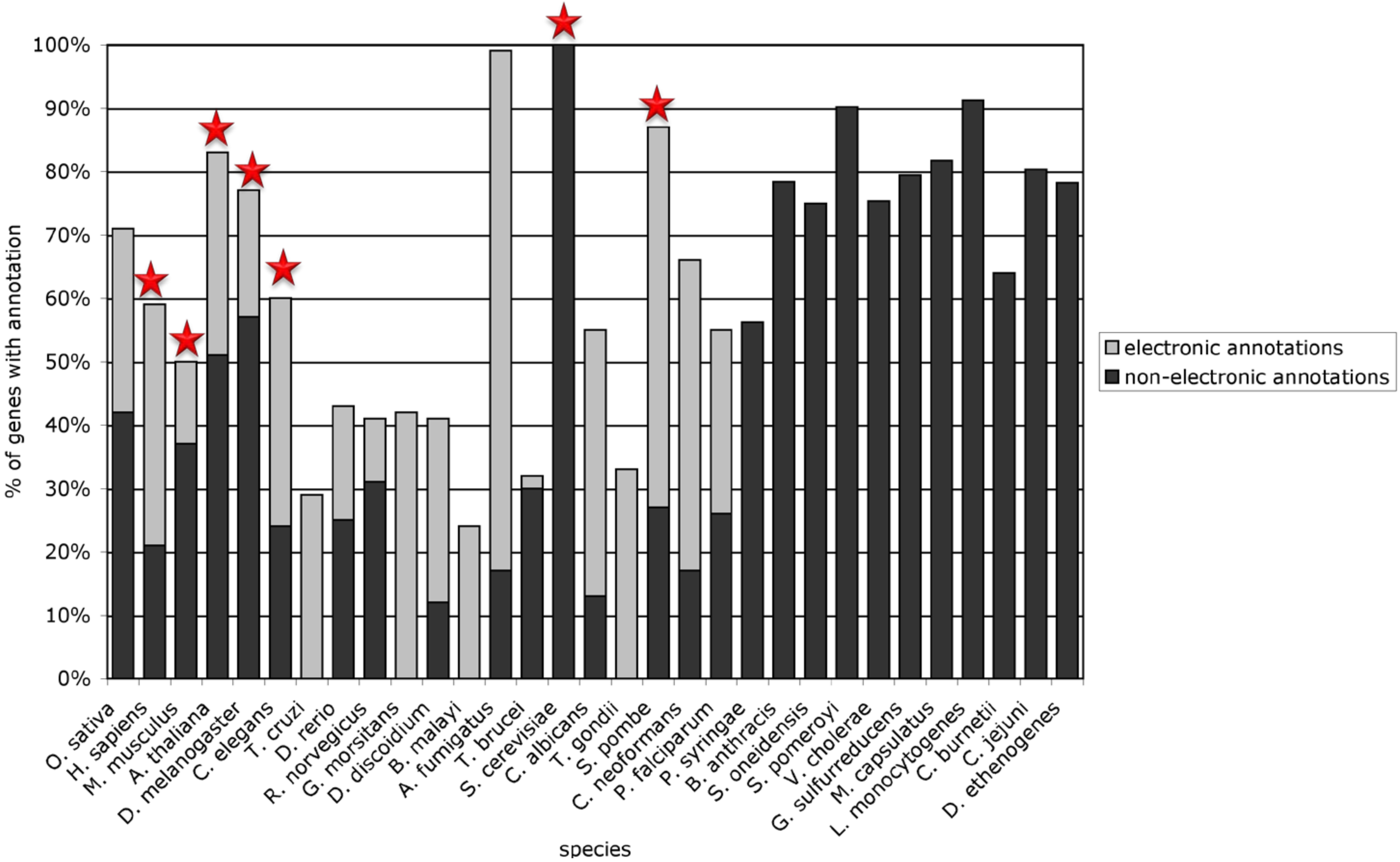
- **IEA**: Inferred from electronic annotation



Species Coverage

- All major eukaryotic model organism species
- Human via GOA group at UniProt
- Several bacterial and parasite species through TIGR and GeneDB at Sanger
- New species annotations in development

Variable Coverage



Lomax J. Get ready to GO! A biologist's guide to the Gene Ontology. Brief Bioinform. 2005 Sep;6(3):298-304.

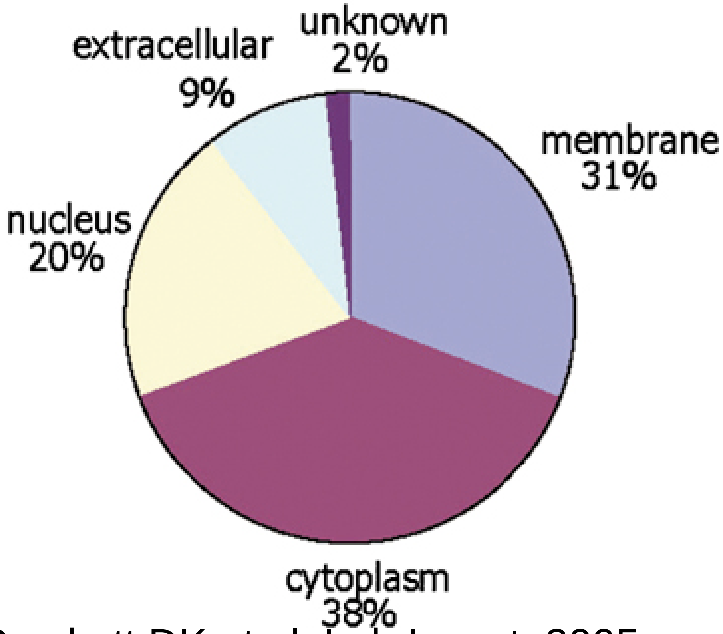
For your information

Contributing Databases

- Berkeley *Drosophila* Genome Project (BDGP)
- dictyBase (*Dictyostelium discoideum)*
- FlyBase (*Drosophila melanogaster)*
- GeneDB (*Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Leishmania major* and *Trypanosoma brucei)*
- UniProt Knowledgebase (Swiss-Prot/TrEMBL/PIR-PSD) and InterPro databases
- Gramene (grains, including rice, *Oryza*)
- Mouse Genome Database (MGD) and Gene Expression Database (GXD) (*Mus musculus)*
- Rat Genome Database (RGD) (*Rattus norvegicus)*
- Reactome
- *Saccharomyces* Genome Database (SGD) (*Saccharomyces cerevisiae)*
- The *Arabidopsis* Information Resource (TAIR) (*Arabidopsis thaliana)*
- The Institute for Genomic Research (TIGR): databases on several bacterial species
- WormBase (*Caenorhabditis elegans)*
- Zebrafish Information Network (ZFIN): (*Danio rerio)*

GO Slim Sets

- GO has too many terms for some uses
  - Summaries (e.g. Pie charts)
- GO Slim is an official reduced set of GO terms
  - Generic, plant, yeast

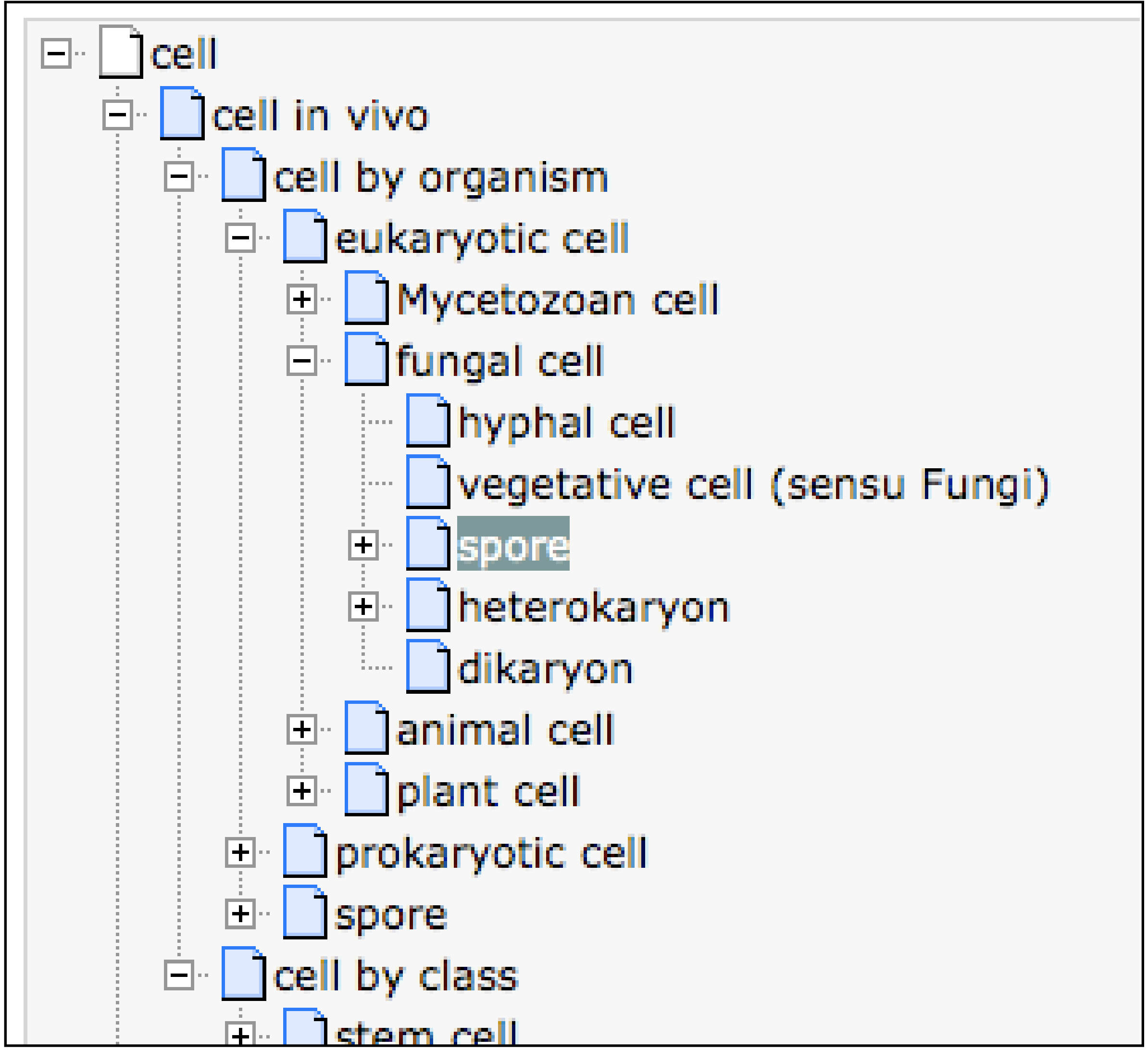


Crockett DK et al. Lab Invest. 2005 Nov;85(11):1405-15

GO Software Tools

- GO resources are freely available to anyone without restriction
  - Includes the ontologies, gene associations and tools developed by GO
- Other groups have used GO to create tools for many purposes
  - http://www.geneontology.org/GO.tools

Other Ontologies



- ⊟ cell
  - ⊟ cell in vivo
    - ⊟ cell by organism
      - ⊟ eukaryotic cell
        - ⊞ Mycetozoan cell
        - ⊟ fungal cell
          - hyphal cell
          - vegetative cell (sensu Fungi)
          - ⊞ spore
          - ⊞ heterokaryon
          - dikaryon
        - ⊞ animal cell
        - ⊞ plant cell
      - ⊞ prokaryotic cell
      - ⊞ spore
    - ⊟ cell by class
      - ⊞ stem cell

http://www.ebi.ac.uk/ontology-lookup

Overview of over representation analysis

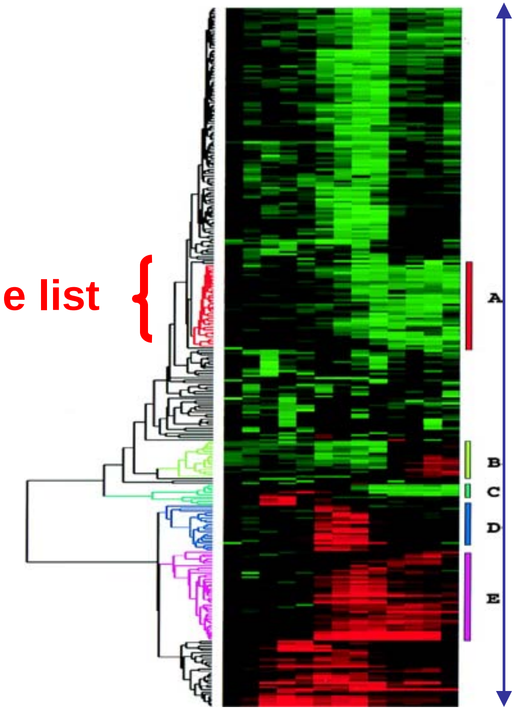
- Theory:
  - Review: What is a P-value? The good ole' T-test.
  - Fisher's Exact Test, the bread and butter of ORA
  - Correcting for multiple testing
  - Enrichment analysis with gene rankings

Over-representation analysis (ORA) in a nutshell

- Given:
  1. Gene list: e.g. RRP6, MRD1, RRP7, RRP43, RRP42 (yeast), or Gene Scores: RRP6 (4.0), MRD1 (3.0) etc
  2. Gene annotations: e.g. Gene ontology, transcription factor binding sites in promoter
- ORA Question: *Are any of the gene annotations* ***surprisingly*** *enriched in the gene list?*
- Details:
  - How to assess “surprisingly” (statistics)
  - How to correct for repeating the tests

Examples of sources of gene lists

Clustering



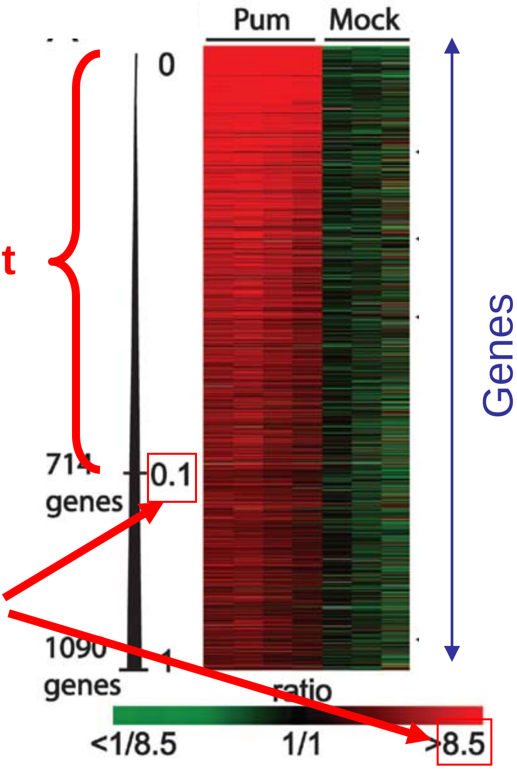
Gene list

Genes

Time

Source Eisen et al. (1998) PNAS 95

Thresholding a gene “score”



Gene list

Examples of gene scores

Genes

Source: Gerber et al. (2006) PNAS103

Overview

- Theory:
  - Review: What is a P-value?
  - Fisher’s Exact Test, the bread and butter of ORA
  - Correcting for multiple testing
  - Enrichment analysis with gene rankings

What is a P-value?

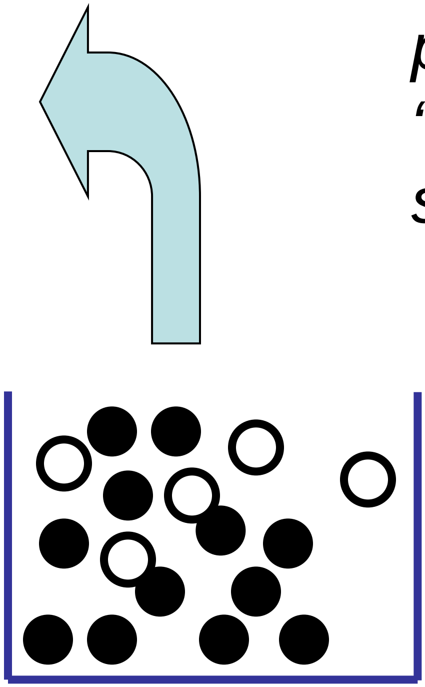
- The P-value is (a bound) on the probability that the "null hypothesis" is true,
- Calculated by calculating statistics using the data and testing the probability of observing those statistics, or ones more extreme, given a sample of the same size distributed according to the null hypothesis,
- Intuitively: *P-value is the probability of a false positive result* (aka "Type I error")

Fisher's exact test: the bread and butter of ORA

a.k.a., the hypergeometric test

Gene list

○ RRP6
● MRD1
○ RRP7
○ RRP43
○ RRP42



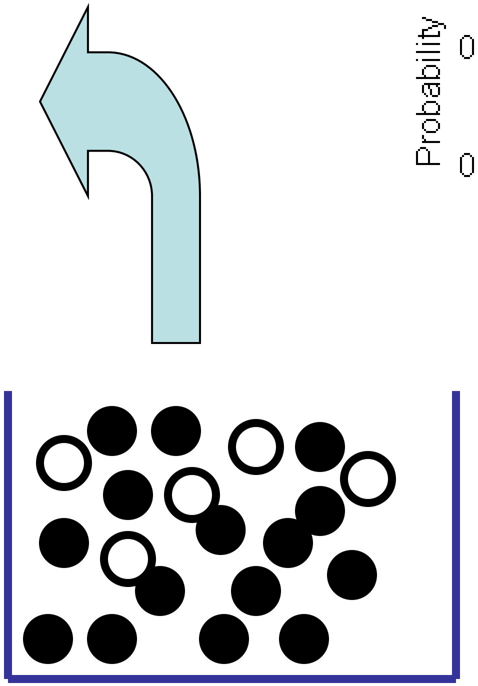
Formal question: *What is the probability of finding 4 or more "white" genes in a random sample of 5 genes?*

Background population:
500 white genes,
4500 black genes

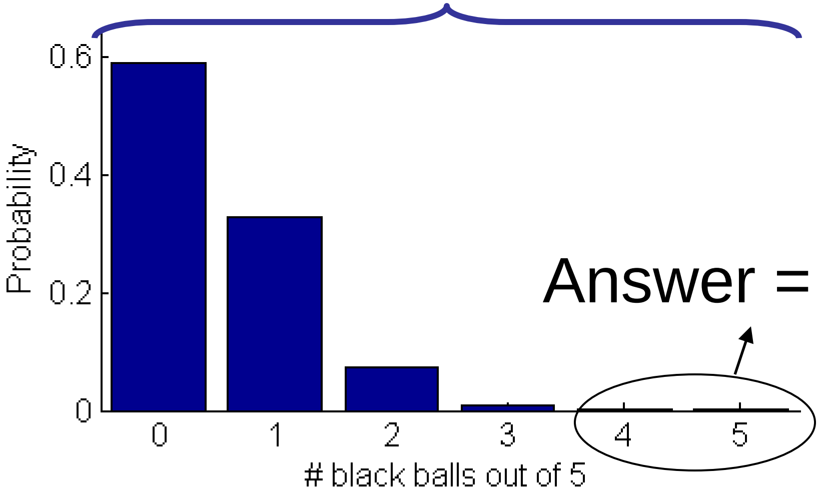
Fisher's exact test cont.

Gene list

○ RRP6
● MRD1
○ RRP7
○ RRP43
○ RRP42



Null distribution



P-value

Answer = 4.6×10^{-4}

Background population:
500 white genes,
4500 black genes

Fisher’s exact test:
a.k.a., the hypergeometric test

In R, use the dhyper function to calculate the probability of seeing x white balls drawn from an urn of white balls (m) and black balls (n) when you randomly draw k balls.

```
dhyper(x, m, n, k)
```

You can calculate the probability of seeing x (or more – up to k) white balls using:

```
sum(dhyper(x:k , m, n, k))
```

Important details

- To test for *under-enrichment* of “white”, test for *over-enrichment* of “black”.
- Need to choose “background population” appropriately, e.g., if only portion of the total gene complement is queried (or available for annotation), only use that population as background.
- To test for enrichment of more than one independent types of annotation, apply Fisher’s exact test separately for each type. ***More on this later***

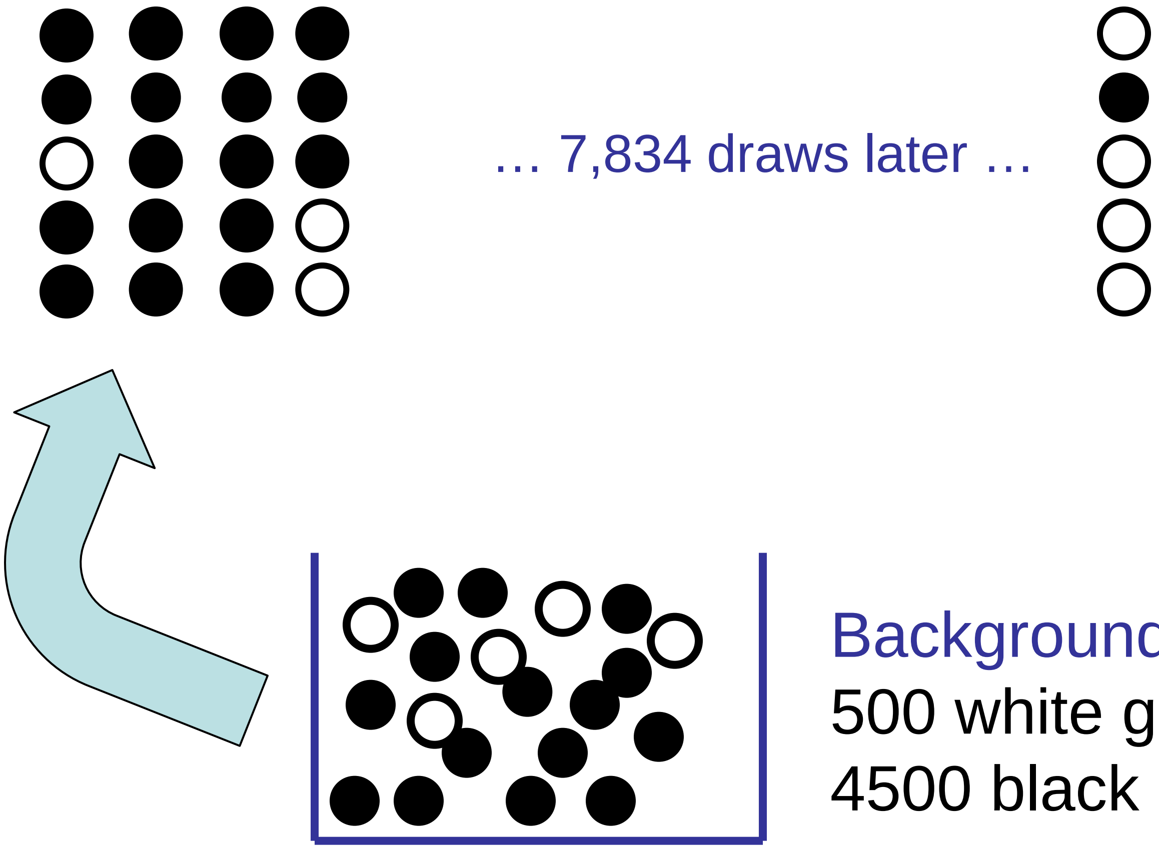
What have we learned?

- Fisher’s exact test is used for ORA of gene lists for a single type of annotation,
- P-value for Fisher’s exact test
  - is “the probability that a random draw of the same size as the gene list from the background population would produce the observed number of annotations in the gene list or more.”,
  - and depends on size of both gene list and background population as well and # of “white” genes in gene list and background.

Correcting for multiple testing: overview

- – Why do we need to correct? Winning the P-value lottery.
- – Controlling the Family-wise Error Rate (FWER) with the Bonferroni-correction
- – Controlling the false-discovery rate (FDR): Benjamini-Hochberg, Storey-Tibshirani, Q-values and all that

How to win the P-value lottery, part 1



Random draws

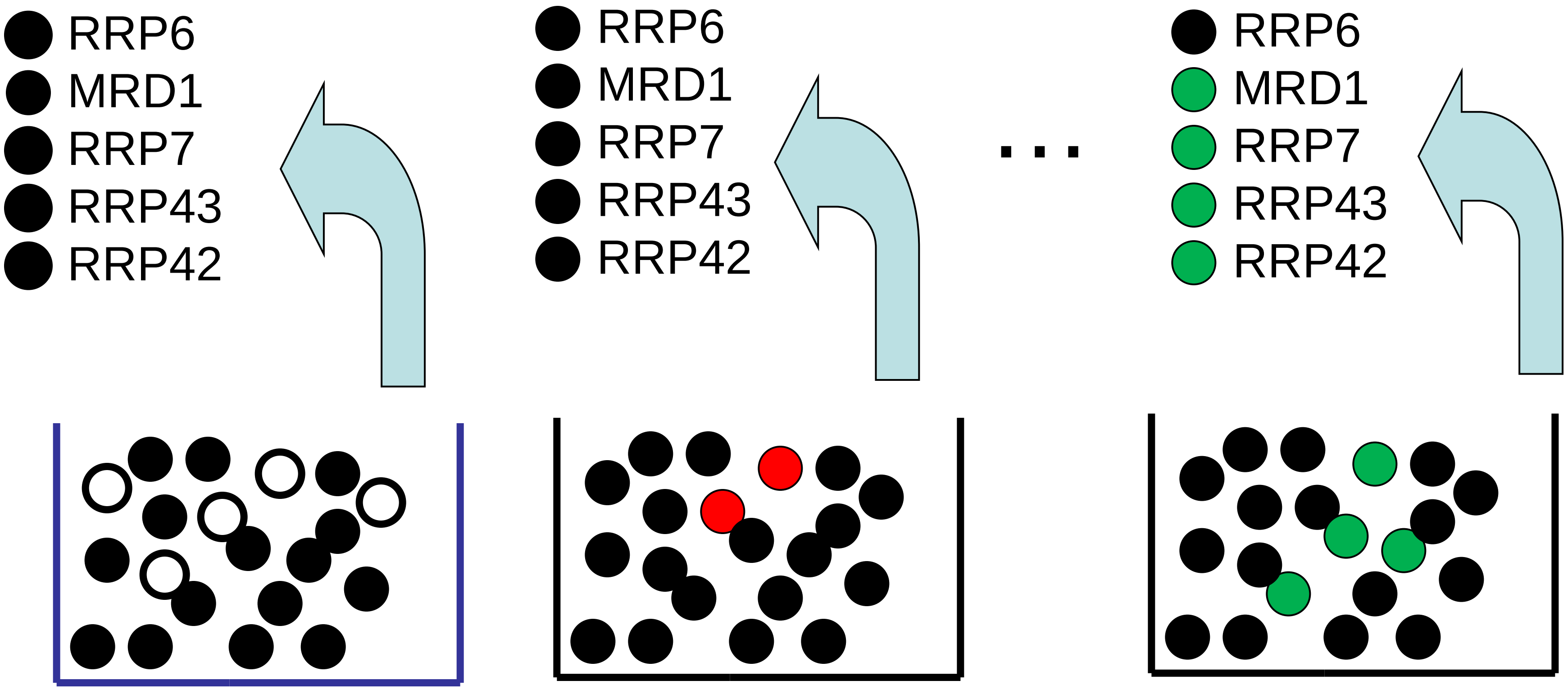
… 7,834 draws later …

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

Background population:
500 white genes,
4500 black genes

How to win the P-value lottery, part 2

Keep the gene list the same, evaluate different annotations



● RRP6
● MRD1
● RRP7
● RRP43
● RRP42

● RRP6
● MRD1
● RRP7
● RRP43
● RRP42

. . .

● RRP6
🟢 MRD1
🟢 RRP7
🟢 RRP43
🟢 RRP42

Bingo!...you win.

ORA tests need correction

From the Gene Ontology website:

Current ontology statistics: **25206** terms

- **14825** biological process
- **2101** cellular component
- **8280** molecular function

Two types of multiple test corrections

- Controlling the Family-Wise Error Rate (FWER) controls the probability that any test is a false positive
- Controlling the False Discovery Rate (FDR) controls the proportion of positive tests (i.e. rejections of the null hypothesis) that are false positives

Controlling Family-Wise Error Rate using the Bonferroni correction

If *M* = # of annotations tested:

Corrected P-value = *M* x original P-value

Corrected P-value is greater than or equal to the probability that any single one of the observed enrichments could be due to random draws. The jargon for this correction is **“controlling for the *Family-Wise Error Rate (FWER)*”**

Bonferroni correction caveats

- Bonferroni correction is very stringent and can “wash away” real enrichments.
- Often users are willing to accept a less stringent condition, the “false discovery rate” (FDR), which leads to a gentler correction when there are real enrichments.

False discovery rate (FDR)

- FDR is *the expected **proportion** of the observed enrichments that are due to random chance.*

- Compare to Bonferroni correction which is *the probability that **any one** of the observed enrichments is due to random chance.*

Benjamini-Hochberg example

Rank	Category	P-value	Adjusted P-value	FDR / Q-value
1	*Transcriptional regulation*	0.001		
2	*Transcription factor*	0.01		
3	*Initiation of transcription*	0.02		
…	…	…		
50	*Nuclear localization*	0.04		
51	*RNAi activity*	0.055		
52	*Cytoplasmic localization*	0.06		
53	*Translation*	0.07		

Sort P-values of all tests in decreasing order

Benjamini-Hochberg example

Rank	Category	P-value	Adjusted P-value	FDR / Q-value
1	*Transcriptional regulation*	0.001	0.001 x 53/1 = 0.053	
2	*Transcription factor*	0.01	0.01 x 53/2 = 0.27	
3	*Initiation of transcription*	0.02	0.02 x 53/3 = 0.35	
…	…	…	…	
50	*Nuclear localization*	0.04	0.04 x 53/50 = 0.042	
51	*RNAi activity*	0.055	0.05 x 53/51 = 0.052	
52	*Cytoplasmic localization*	0.06	0.06 x 53/52 = 0.061	
53	*Translation*	0.07	0.07 x 53/53 = 0.07	

Adjusted P-value = P-value X [# of tests] / Rank

Benjamini-Hochberg example

Rank	Category	P-value	Adjusted P-value	FDR / Q-value
1	*Transcriptional regulation*	0.001	0.001 x 53/1 = 0.053	0.042
2	*Transcription factor*	0.01	0.01 x 53/2 = 0.27	0.042
3	*Initiation of transcription*	0.02	0.02 x 53/3 = 0.35	0.042
…	…	…	…	…
50	*Nuclear localization*	0.04	0.04 x 53/50 = 0.042	0.042
51	*RNAi activity*	0.055	0.05 x 53/51 = 0.052	0.052
52	*Cytoplasmic localization*	0.06	0.06 x 53/52 = 0.061	0.061
53	*Translation*	0.07	0.07 x 53/53 = 0.07	0.07

Q-value = minimum adjusted P-value at given rank or below

Benjamini-Hochberg example

Rank	**Category**	**P-value**	**Adjusted P-value**	**FDR / Q-value**	**FDR < 0.05?**
1	*Transcriptional regulation*	0.001	0.001 x 53/1 = 0.053	0.042	Y
2	*Transcription factor*	0.01	0.01 x 53/2 = 0.27	0.042	Y
3	*Initiation of transcription*	0.02	0.02 x 53/3 = 0.35	0.042	Y
…	…	…	…	…	…
50	*Nuclear localization*	**0.04**	0.04 x 53/50 = 0.042	0.042	Y
51	*RNAi activity*	0.055	0.05 x 53/51 = 0.052	0.052	N
52	*Cytoplasmic localization*	0.06	0.06 x 53/52 = 0.061	0.061	N
53	*Translation*	0.07	0.07 x 53/53 = 0.07	0.07	N

P-value threshold for FDR < 0.05

P-value threshold is highest ranking P-value for which corresponding Q-value is below desired significance threshold

Reducing multiple test correction stringency

- The correction to the P-value threshold 〈 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 select only larger GO categories.

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 or restrict testing to the appropriate GO annotations.

What have we learned

- When testing multiple annotations, need to correct the P-values (or, equivalently, α) to avoid winning the P-value lottery.
- There are two types of corrections:
  - **Bonferroni** controls the probability any one test is due to random chance (aka FWER) and is very stringent
  - **B-H** controls the FDR, i.e., expected proportion of “hits” that are due to random chance
- Can control stringency by carefully choosing which annotation categories to test.