Transcriptome
Lecture 3

Outline

- Multiple Testing Procedures
- Data Visualization, Distance Measures
- Clustering
- Gene Annotation and Enrichment Analysis

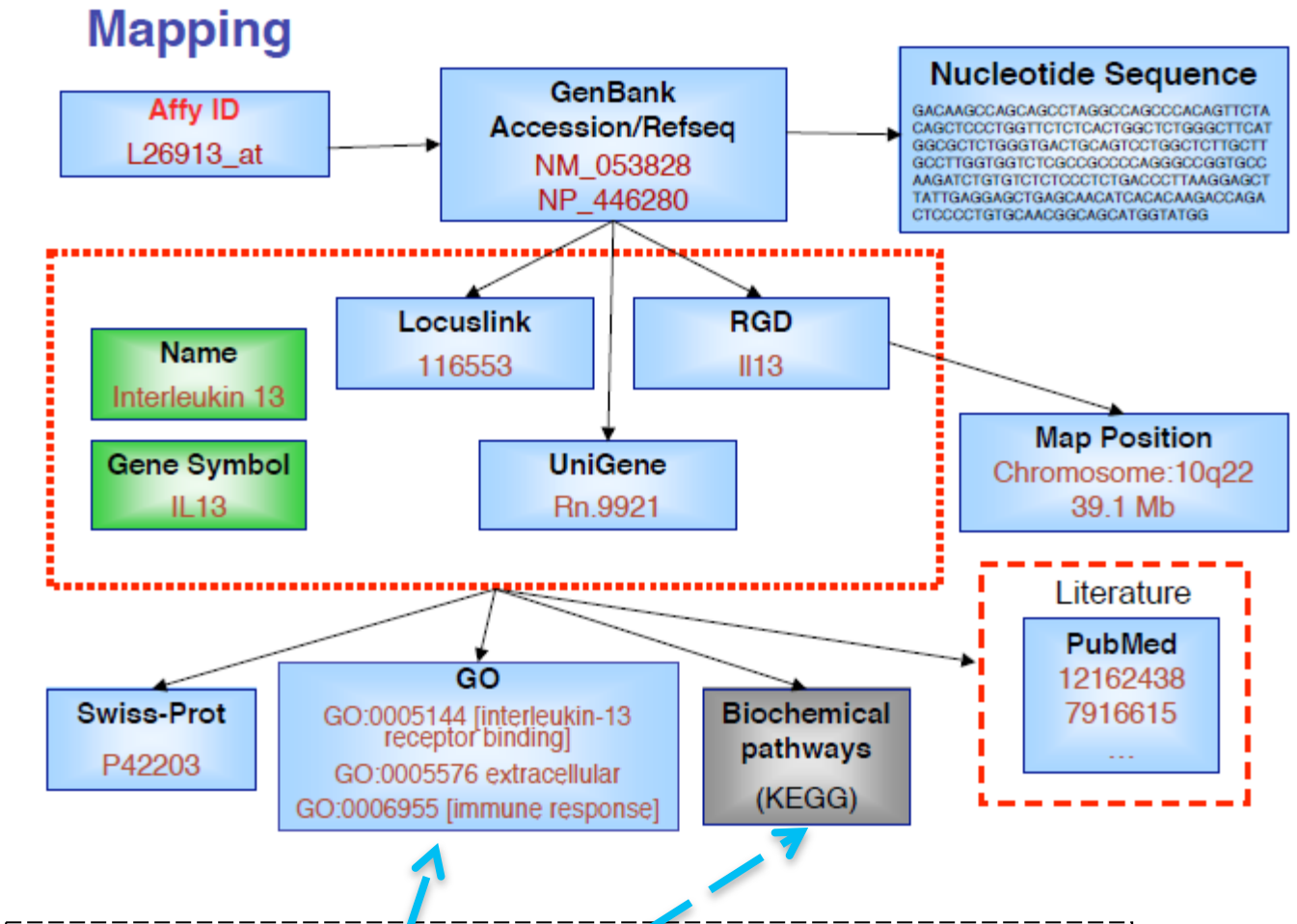
The problem

- After differential expression testing, we obtained a list of significantly differentially expressed probes, controlled for false discovery rate
- We want to understand the biological insight behind this list
  1. we need to map the gene annotation information to these probes or gene IDs
  2. we want to test/infer whether an annotation is significantly enriched in our list

Annotation mapping

- What annotation information can we map probes or gene IDs to?
  - Chromosome, genes, protein family, structure, sequence, variations…
  - Gene Ontology, KEGG Pathway,…
  - Published literatures…

Annotation mapping: example



Gene Ontology (GO) and KEGG pathway are widely used in annotation mapping and testing

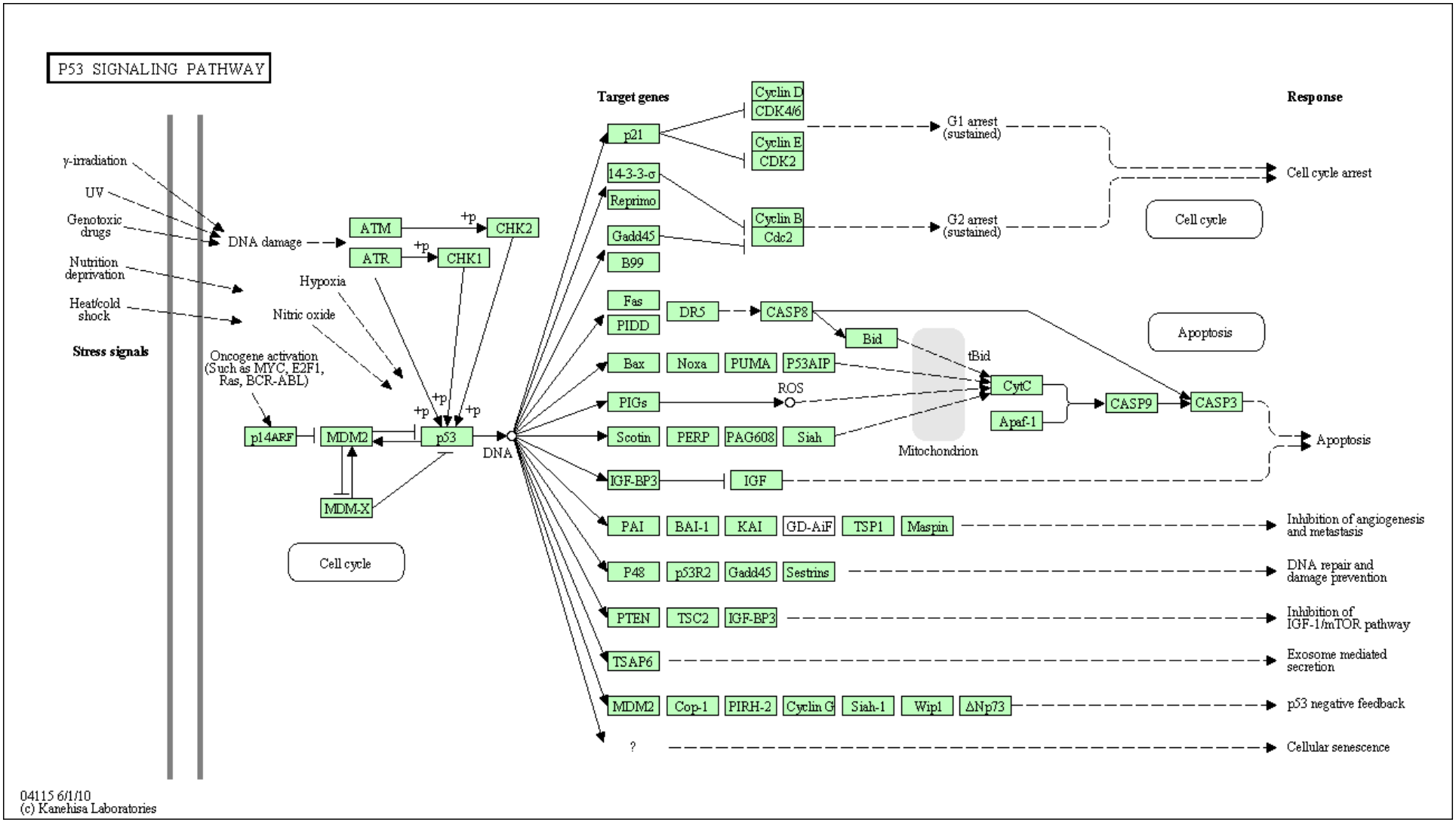
Metabolic Pathways

- PMN: Plant Metabolic Network (http://www.plantcyc.org/)
- MetaCyc (http://metacyc.org/)
- KEGG: Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/kegg2.html)
- Reactome (http://www.reactome.org/)
- PANTHER PATHWAYS (http://www.pantherdb.org/pathway/)
- Pathways Commons (http://www.pathwaycommons.org/pc/home.do)

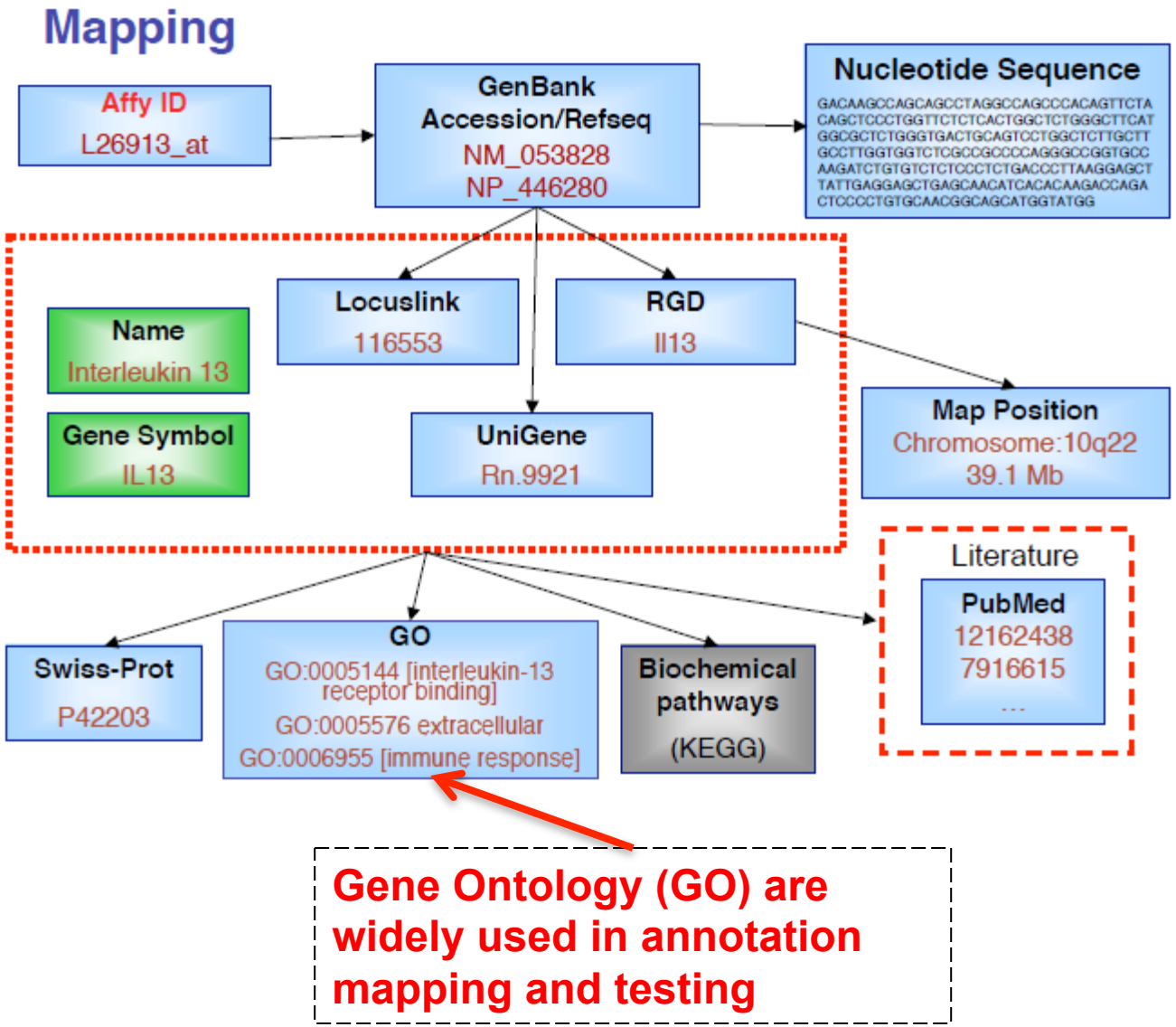
KEGG Pathway

- KEGG Pathways:
  - Manually curated pathway maps representing the knowledge on the molecular interaction and reaction networks, for a large selection of organisms
  - The KEGG pathways include a collection of pathways important in:
    - Metabolism
    - Genetic Information Processing
    - Environmental Information Processing
    - Cellular Processes
    - Human Disease
    - …

KEGG Pathway: An example



Annotation mapping: example



Gene Ontology (GO)

- Gene Ontology (GO) is a collection of controlled vocabularies describing the biology of a gene product in any organism
- http://www.geneontology.org/
- Very useful for interpreting biological insight of microarray data – and it is computable!

So what is ontology?

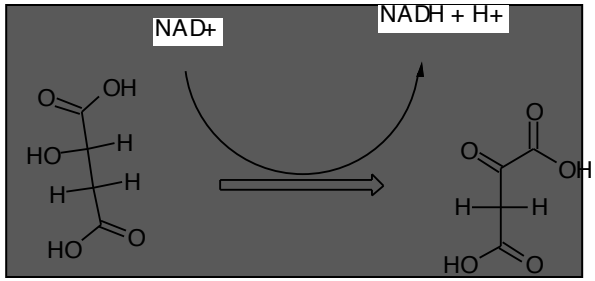
From a practical view, ontology is the representation of something we know about. "Ontologies" consist of a representation of things, that are detectable or directly observable, and the relationships between those things.

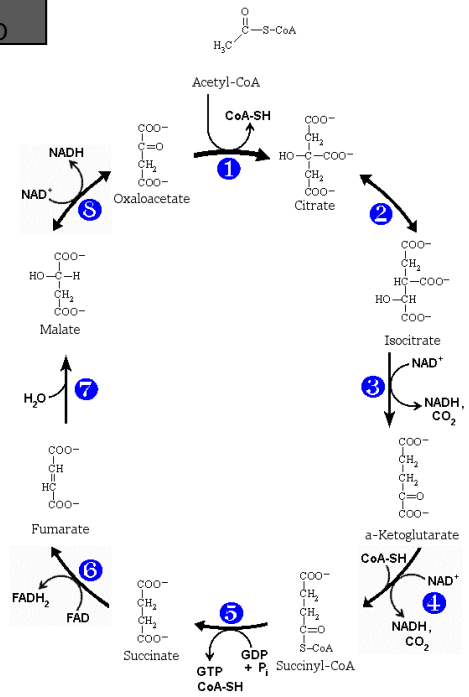
is part of

The GO has three Ontologies

Molecular Function

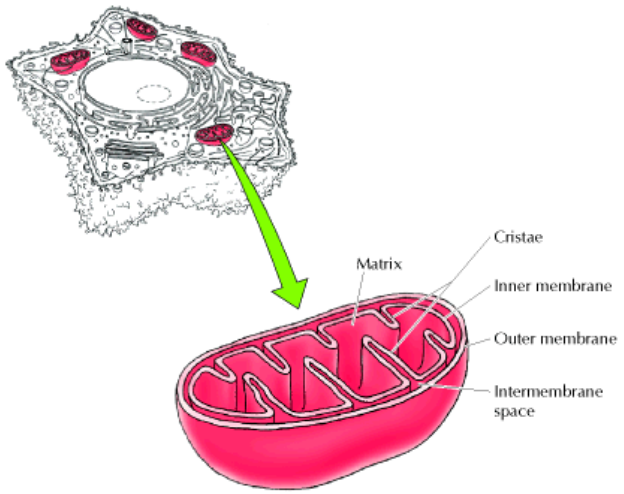
GO term: Malate dehydrogenase.
GO id: GO:0030060
(S)-malate + NAD(+) = oxaloacetate + NADH.





Biological Process

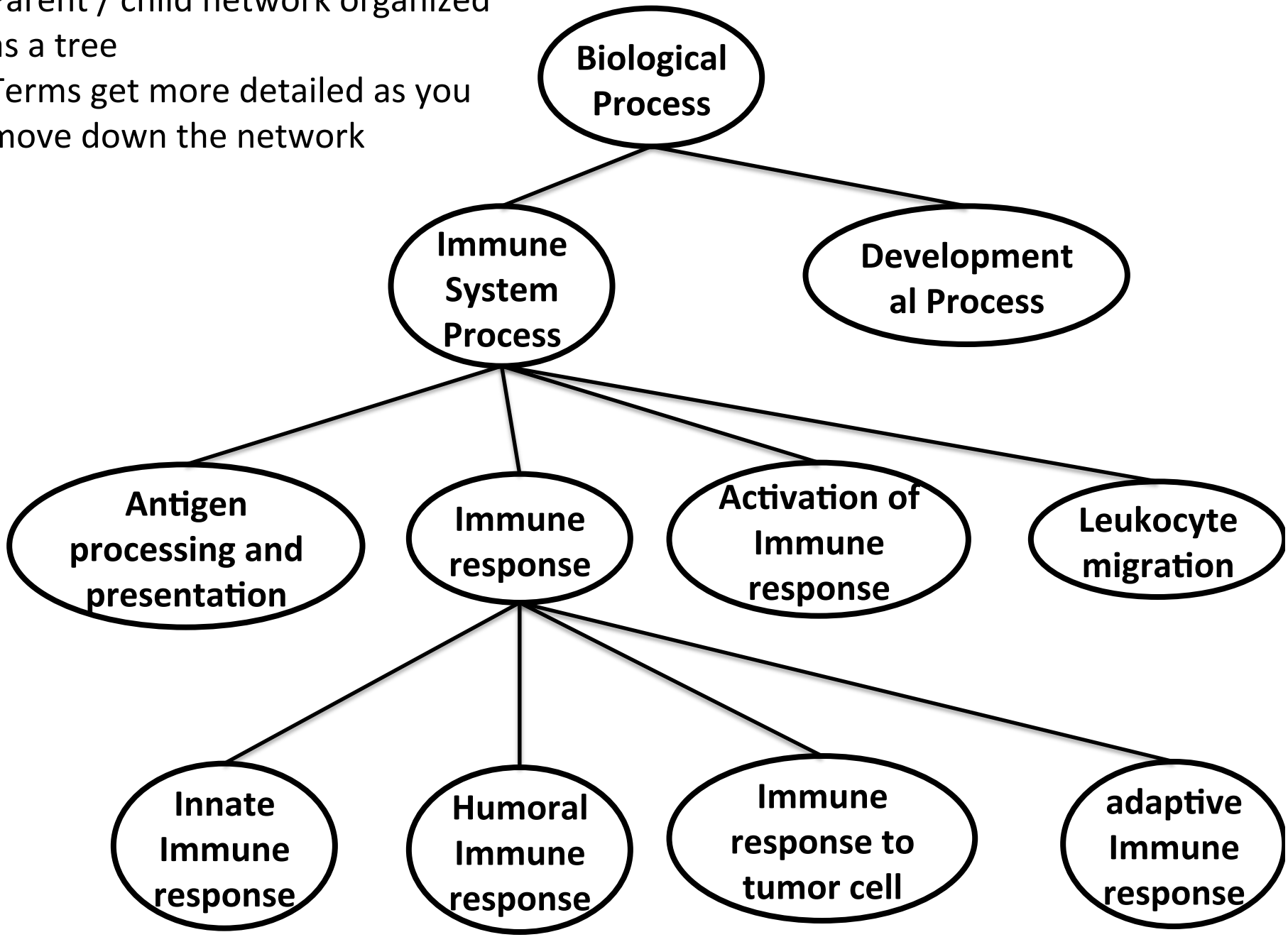
GO term: tricarboxylic acid cycle
Synonym: Krebs cycle
Synonym: citric acid cycle
GO id: GO:0006099



Cellular Component

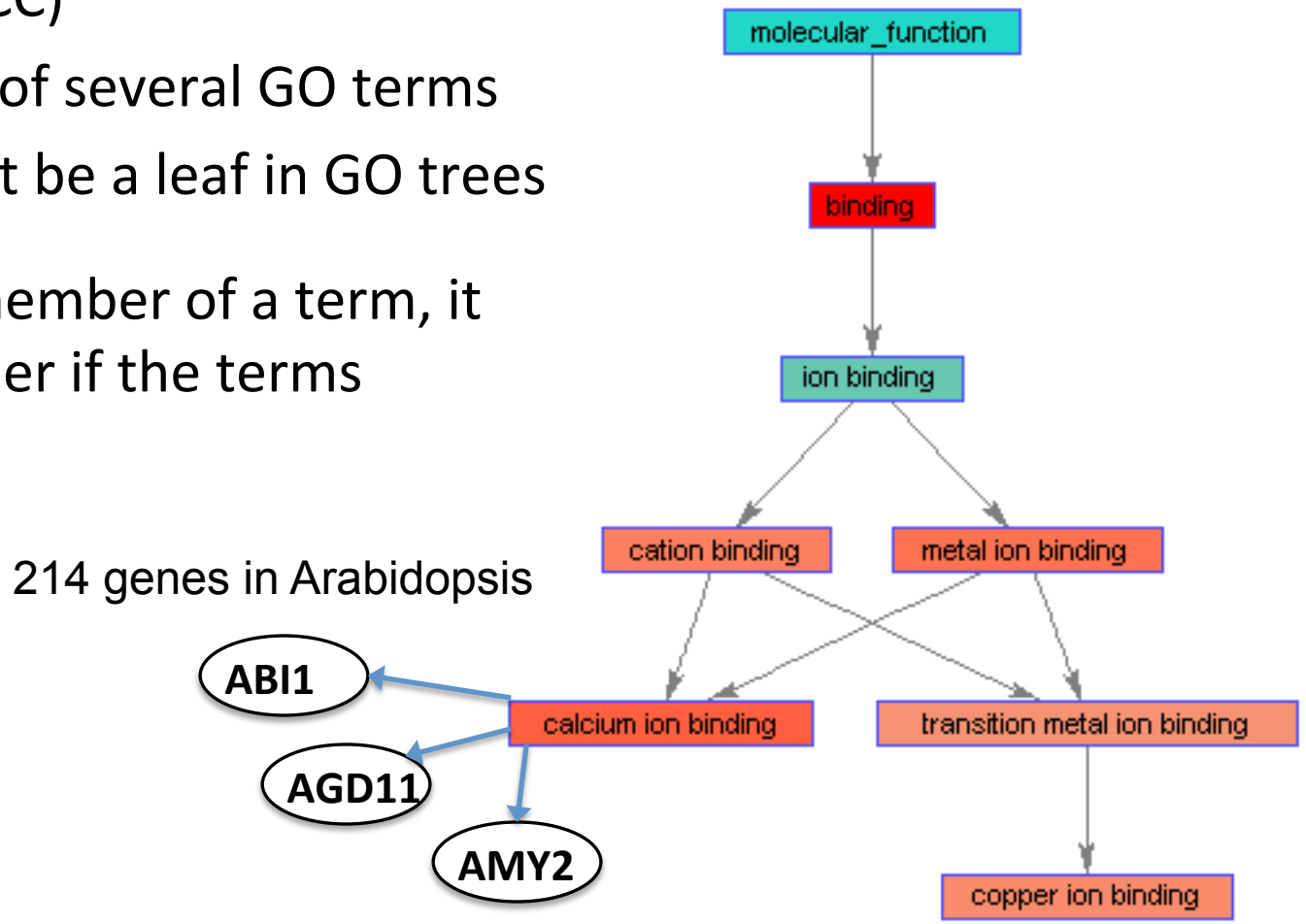
GO term: mitochondrion
GO id: GO:0005739

- Gene Ontology: Tree Structure
- Parent / child network organized as a tree
- Terms get more detailed as you move down the network



Gene Ontology: Rule

- In GO, a gene can be
  - present in any of the ontologies (MF / BP / CC)
  - a member of several GO terms
  - a gene must be a leaf in GO trees
- If a gene is a member of a term, it is also a member if the terms parents



Gene Ontology: files

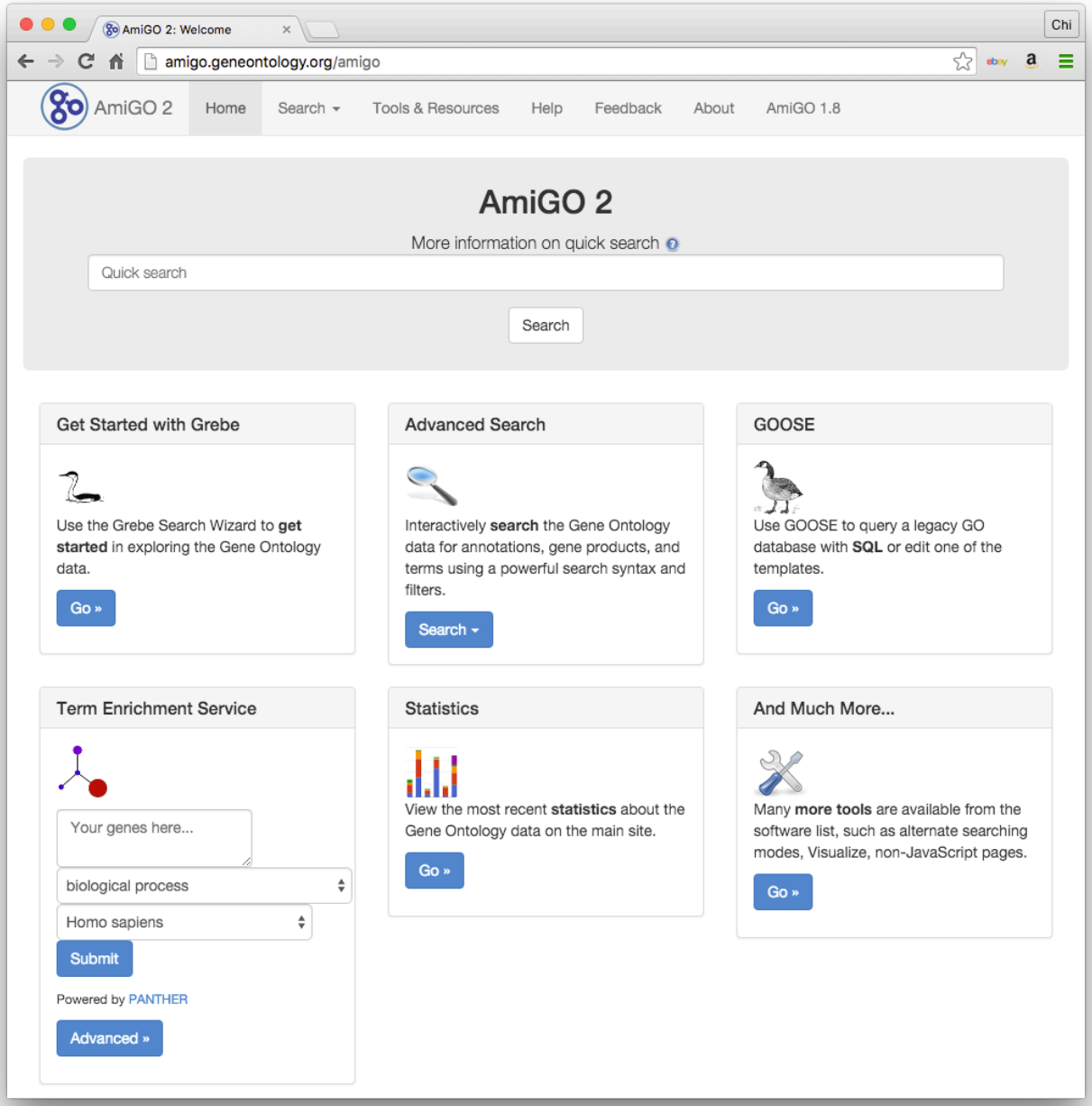
- Ontology file: GO terms and relationships in a variety of formats. The ontology file is unique for all species.
- Annotation files: associations between gene products and GO terms submitted by members and associates of the GO consortium. Different species have different annotation files.
  - ✓ gene_association.tair
  - ✓ gene_association.goa_human

GO tools

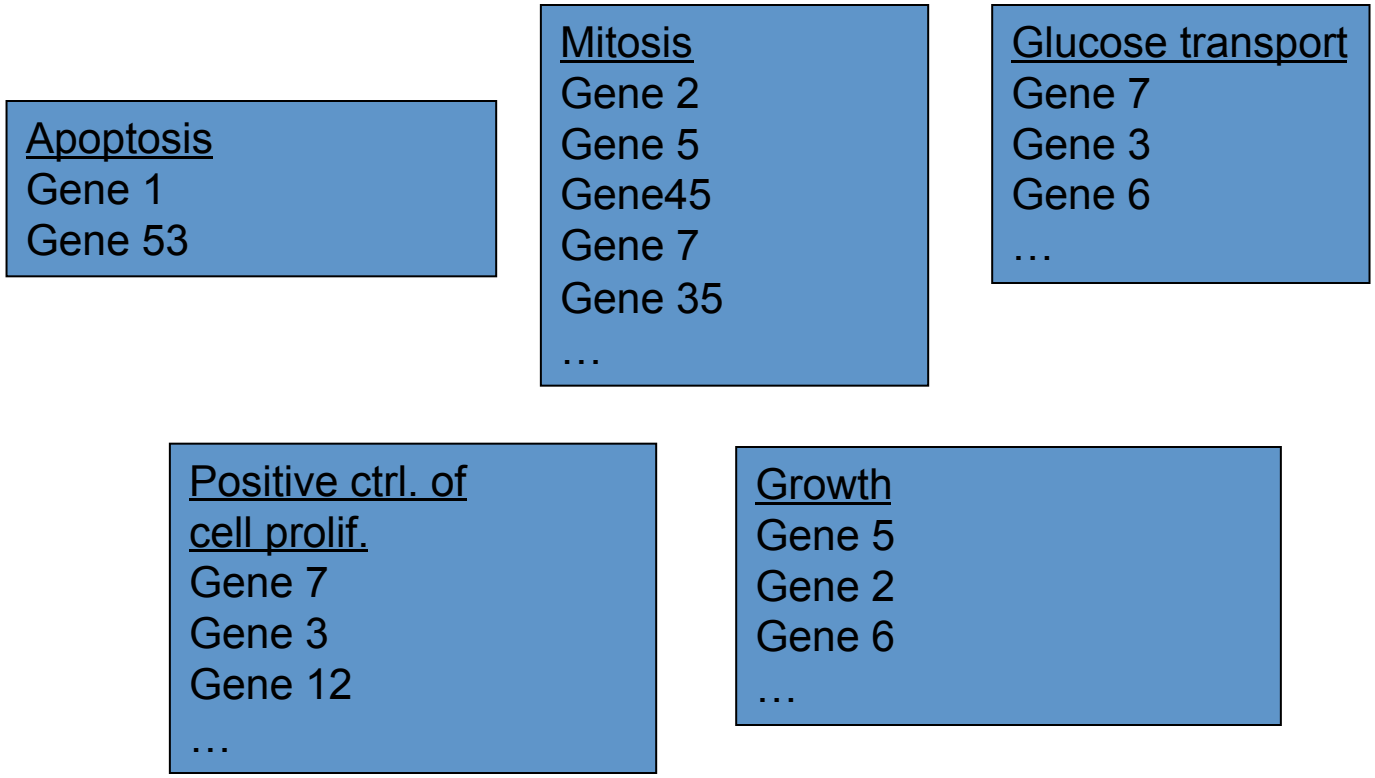
- GO resources are freely available to anyone to use without restriction
  - Includes the ontologies, gene associations and tools developed by GO
- Other groups have used GO to create tools for many purposes:

    http://amigo.geneontology.org/amigo

Gene Ontology: tools

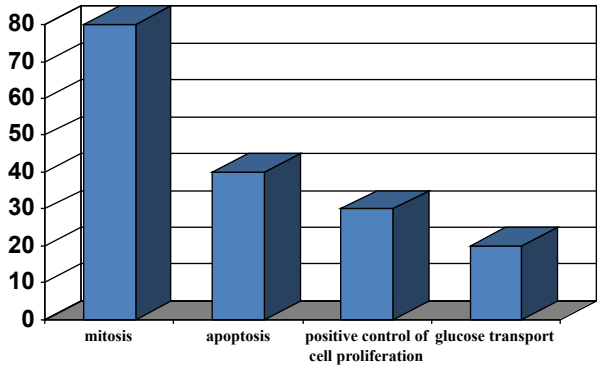


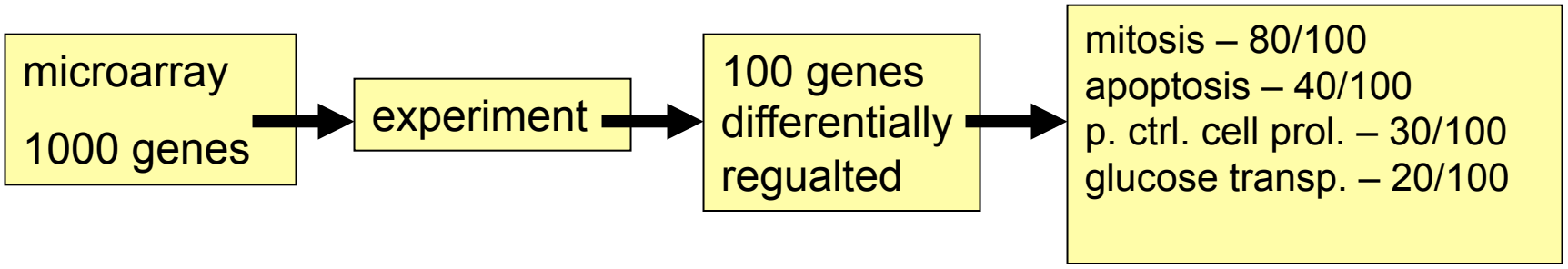
Grouping by Biological process



Using GO in practice

- statistical measure
  - how likely your differentially regulated genes fall into that category by chance





The problem

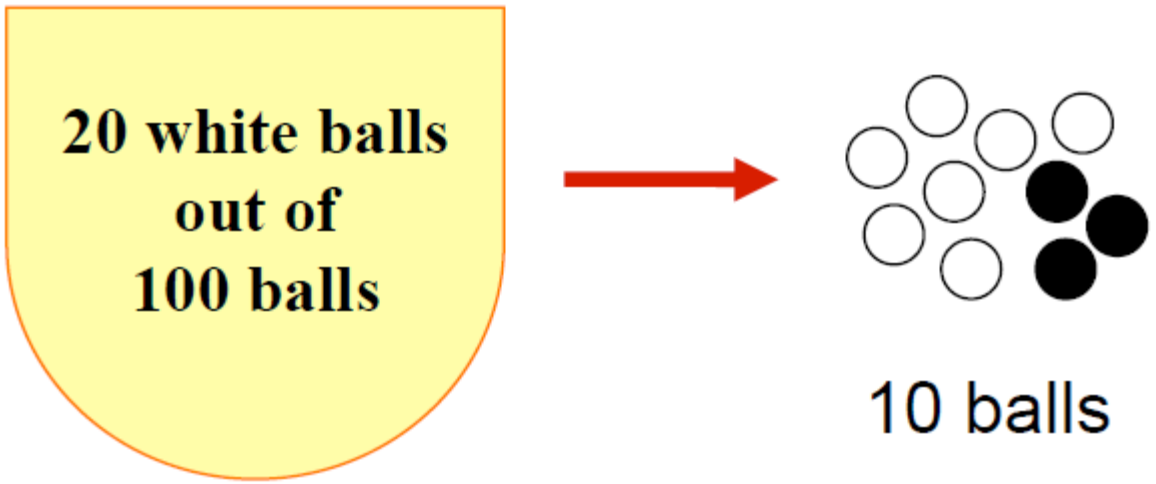
- After differential expression testing, we obtained a list of significantly differentially expressed probesets, controlled for false discovery
- We want to understand the biological insight behind this list
  1. we need to map the gene annotation information to these probesets
  2. we need to test/infer whether an annotation is significantly enriched in our list

Annotation Testing (enrichment analysis)

- We want to ask:
  - Are there any GO terms overrepresented in the obtained gene list, compared with what would happen by chance?
    - Hypergeometric test
    - Fisher's exact test fisher.test()
    - Binomial test binom.test()
    - Chi-squared test chisq.test()
    - Kolmogorov-Smirnov test

Hypergeometric distribution

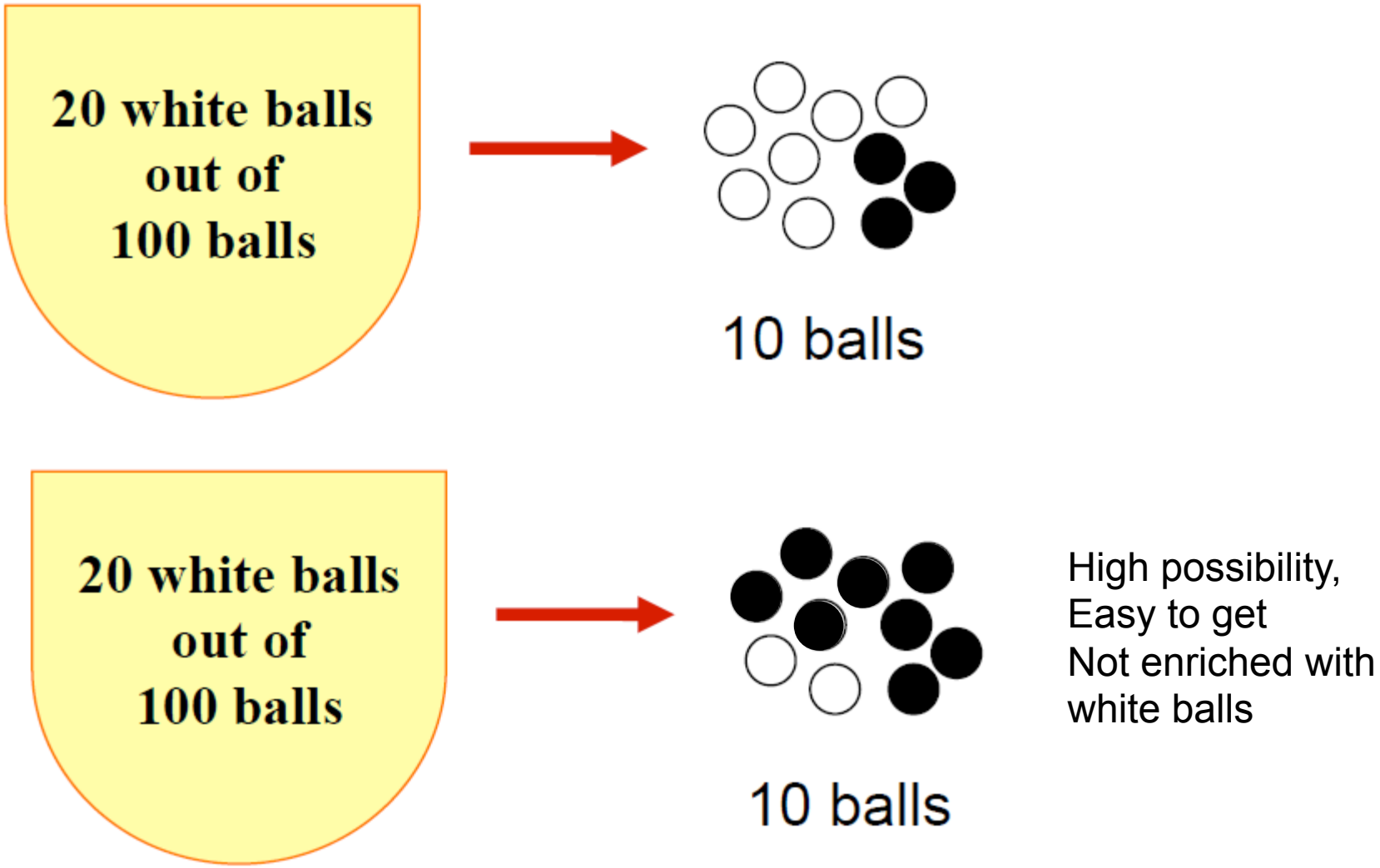
- The hypergeometric distribution arises from sampling from a fixed population.



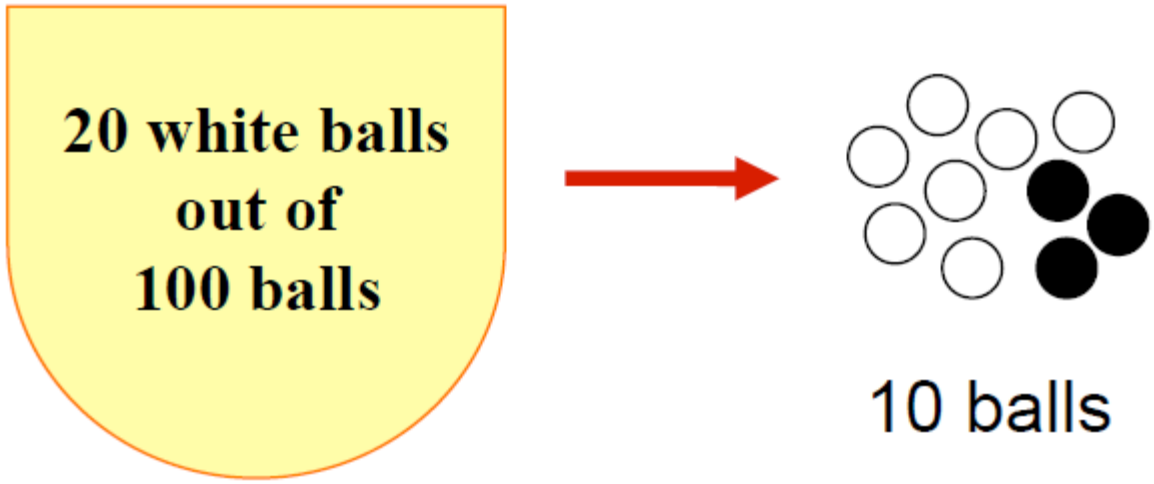
$$P(k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

N=100
m=20
n=10
k=7

Hypergeometric test



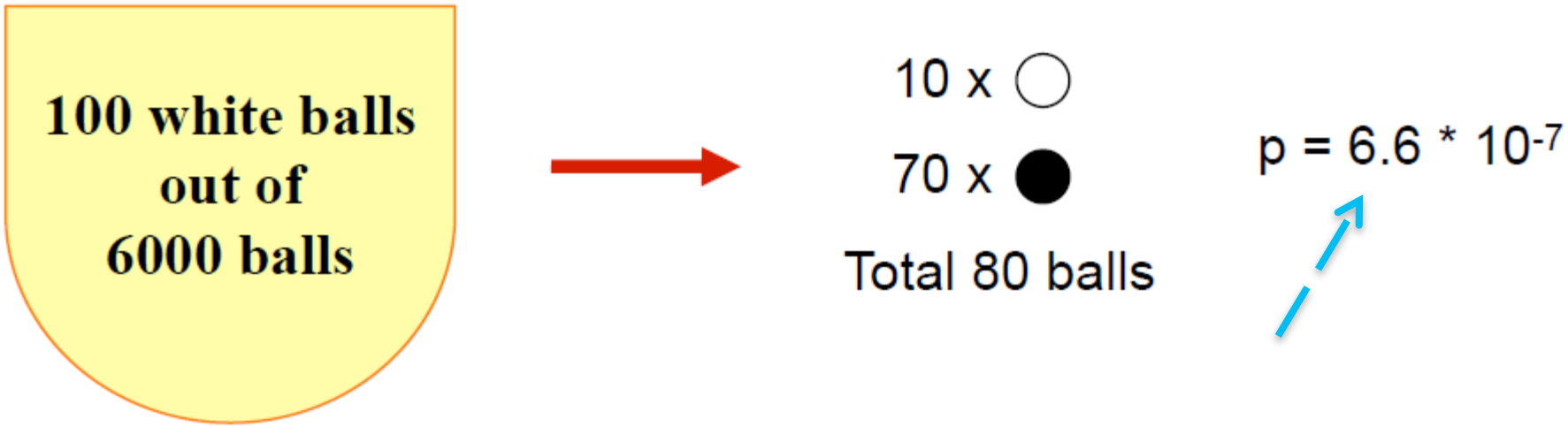
Hypergeometric test



- **TEST:** We want to calculate the probability for drawing 7 or more white balls out of 10 balls given the distribution of balls in the urn.
- The smaller the possibility is, the more significantly enriched.

Annotation Testing (Hypergeometric test)

- Example: we obtained a list of 80 significant genes from a gene expression experiment of yeast.
- Yeast has 6000 genes, and 100 of them can be mapped to a GO term called "Cell cycle". For the 80 significant differentially expressed genes from micrroarray/RNA-seq, 10 are mapped to this GO term.
  - Is this observation a significant event? Or, is the GO term "Cell cycle" significantly over-represented in our list of 80



GO enrichment analysis in R

- Gostat
  - http://www.bioconductor.org/packages/2.3/bioc/html/GOstats.html
- PGSEA
  - http://www.bioconductor.org/packages/2.4/bioc/html/PGSEA.html

Online tools

- DAVID http://david.abcc.ncifcrf.gov/
- GoMiner: http://discover.nci.nih.gov/gominer
- GOstat: http://gostat.wehi.edu.au
- GSEA: http://www.broadinstitute.org/gsea/index.jsp
- BIOBASE (Whitehead has license)
- BiNGO (uses Cytoscape)

DAVID: a function annotation tool

- http://david.abcc.ncifcrf.gov/