Transcriptome
Lecture 4

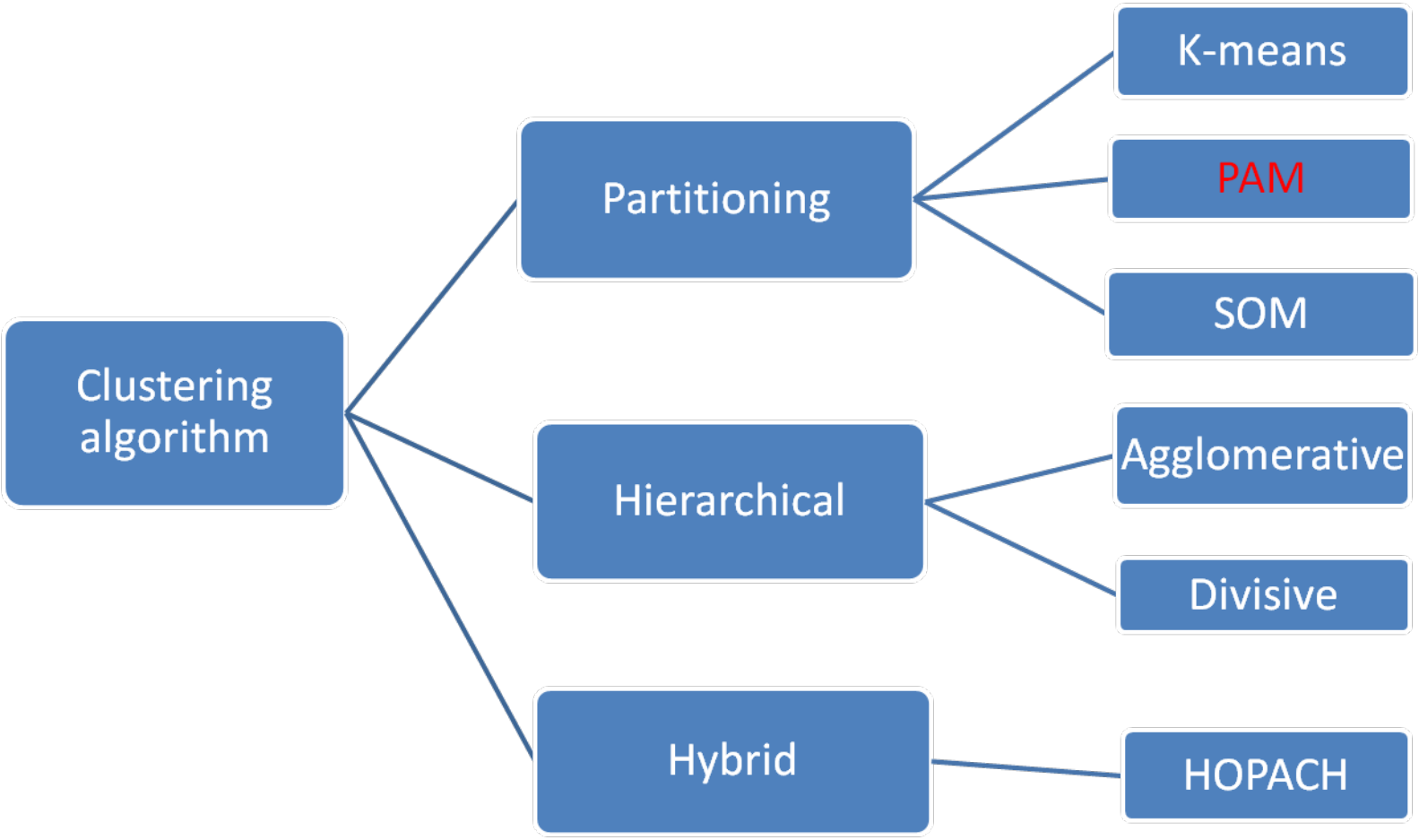
Outline

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

Clustering: Basic principles

- Issues to be consider before performing a cluster analysis
  - ❑Which genes/arrays to be used?
  - ❑Which distance (similarity) measures?
    - Correlation coefficient based distance or Minkowski metric
  - ❑Which method is used to join clusters/ observations?
    - Single-link, Complete-link, Average-link, Centroid-link
  - ❑Which clustering algorithm is applied?

Type of Clustering algorithm



Outline

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

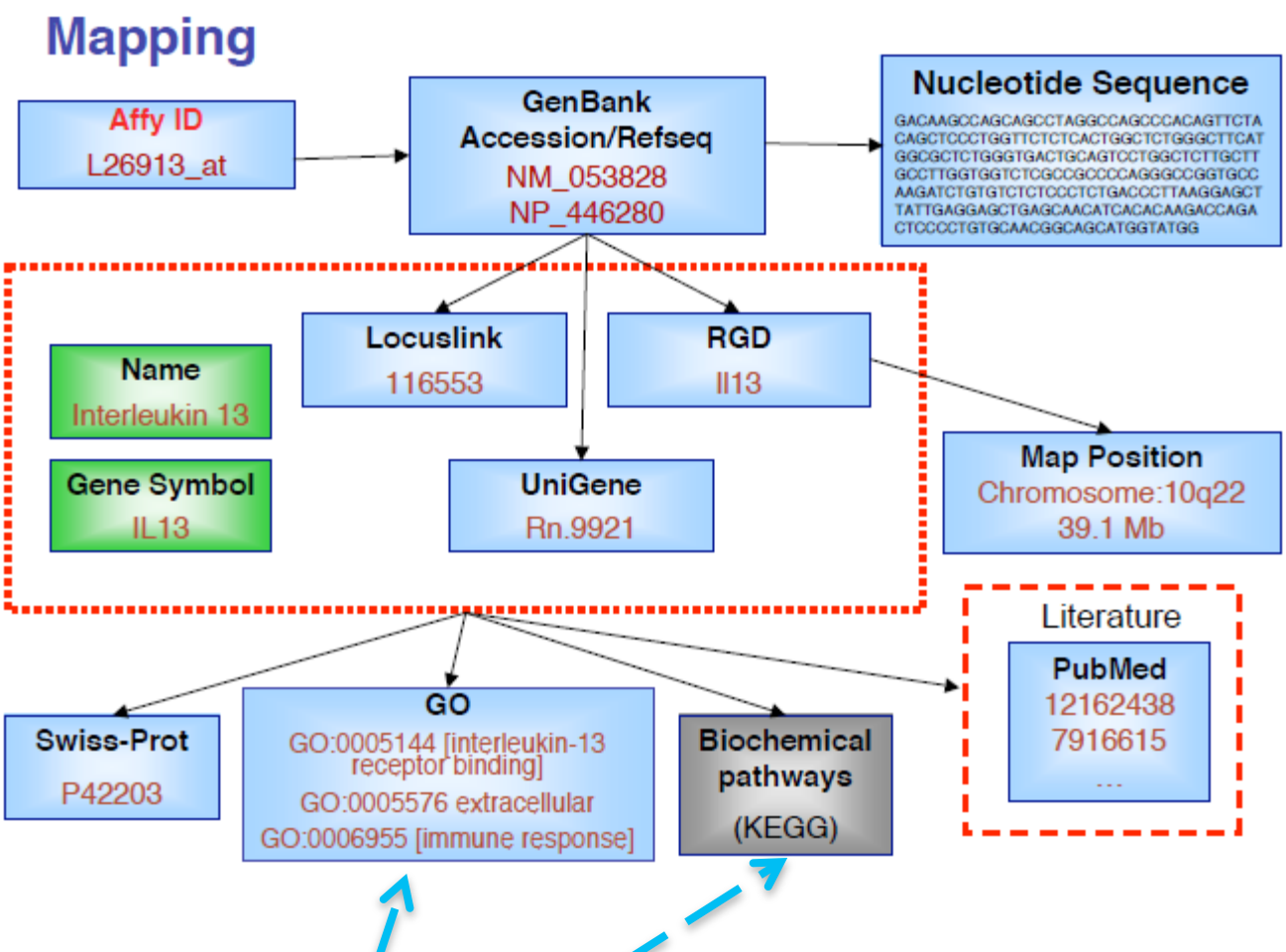
The problem

- 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…

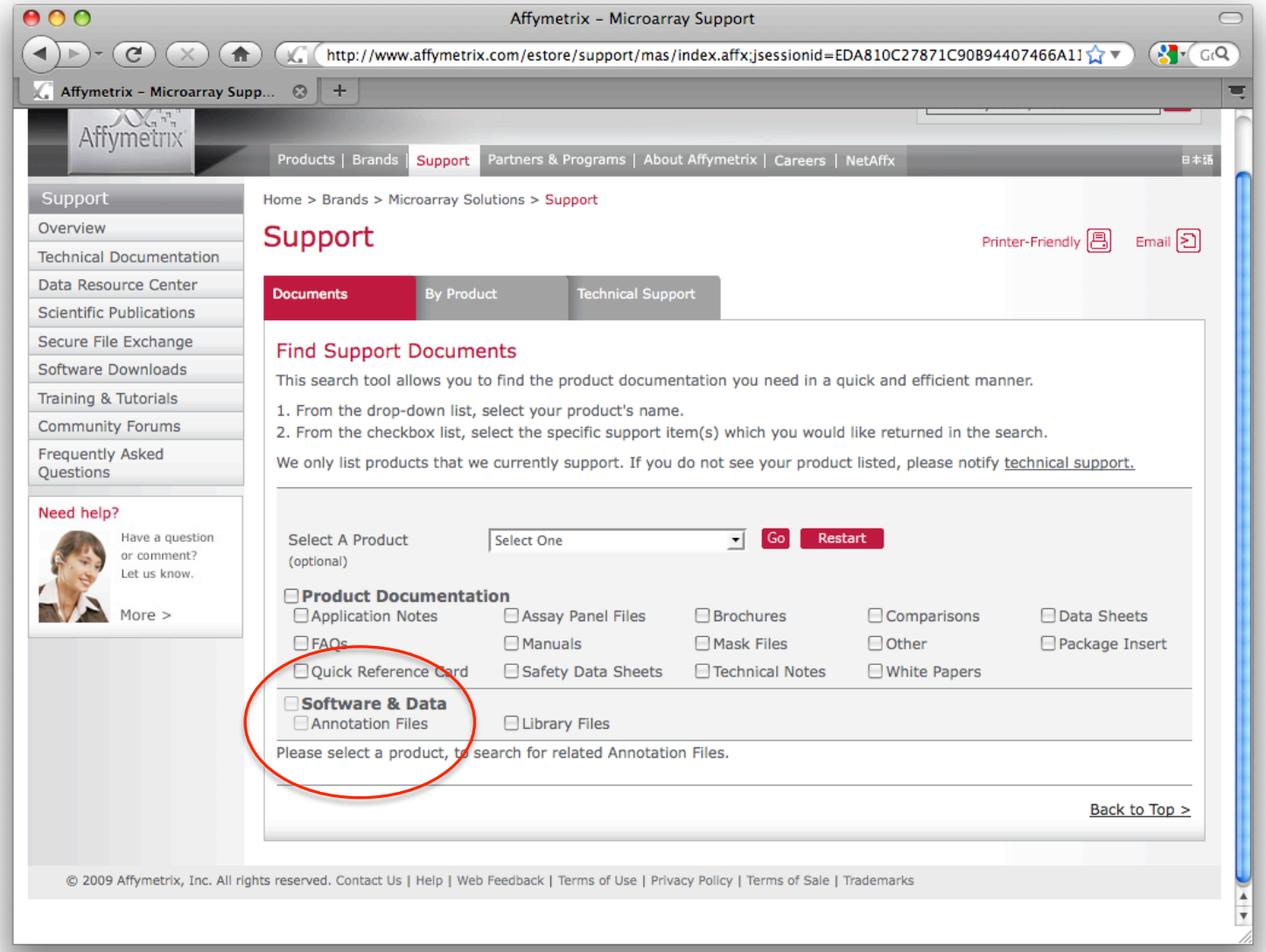
Annotation mapping: example



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

Annotation mapping

http://www.affymetrix.com/



Annotation mapping

Probe ID	Unigene	SwissProt	RefSeq	Entrez	Gene Symbo	Gene Title
Zm.1.1.A1_at	Zm.80960	B6T8E4 // / Q41804	NP_001105349	542280	eps5	embryo specific protein5

Annotation mapping in R

- What (bioconductor) packages are available for us to the mapping?

Annotation mapping

- The Bioconductor project provides comprehensive annotation data packages, that contain many different ID mappings to interesting data
  - http://www.bioconductor.org/packages/2.6/data/annotation/
  - E.g., "hgu95av2" provides the mapping between between Affy IDs and IDs like gene IDs, GO, KEGG pathway…
- These packages are updated and expanded regularly as new data become available.

Annotation package

Installation:

```
> source("http://bioconductor.org/biocLite.R")
>biocLite("hgu95av2.db")
```

```
>library("hgu95av2.db")
> hgu95av2()
```

This package has the following mappings:

hgu95av2ACCNUM has 12625 mapped keys (of 12625 keys)

…

hgu95av2GENENAME has 11725 mapped keys (of 12625 keys)

…

Annotation mapping: Hash

- “Mapping” is basically the role of a hash table in most programming languages. In R, we can use “environment” object.

- The annotation data packages provide R environment objects containing key (e.g., affy probe set ID)and value (e.g., GO ID) pairs for the mappings between two sets of probe identifiers.

Annotation mapping: Hash

- > library(hgu95av2)

- > get("41046_s_at", env = hgu95av2GENENAME)
  [1] "zinc finger protein 261“

- > get("41046_s_at", env = hgu95av2GO)
  "GO:0003677" "GO:0007275" "GO:0016021"

Annotation mapping: Hash

- Alternatively, instead of relying on the general R functions for environments, new user friendly functions have been written for accessing and working with specific identifiers.
  - E.g. getGO, getGOdesc, getSYMBOL, …

Annotation mapping: Hash

- > library(hgu95av2)

- > getSYMBOL("41046_s_at",data="hgu95av2")

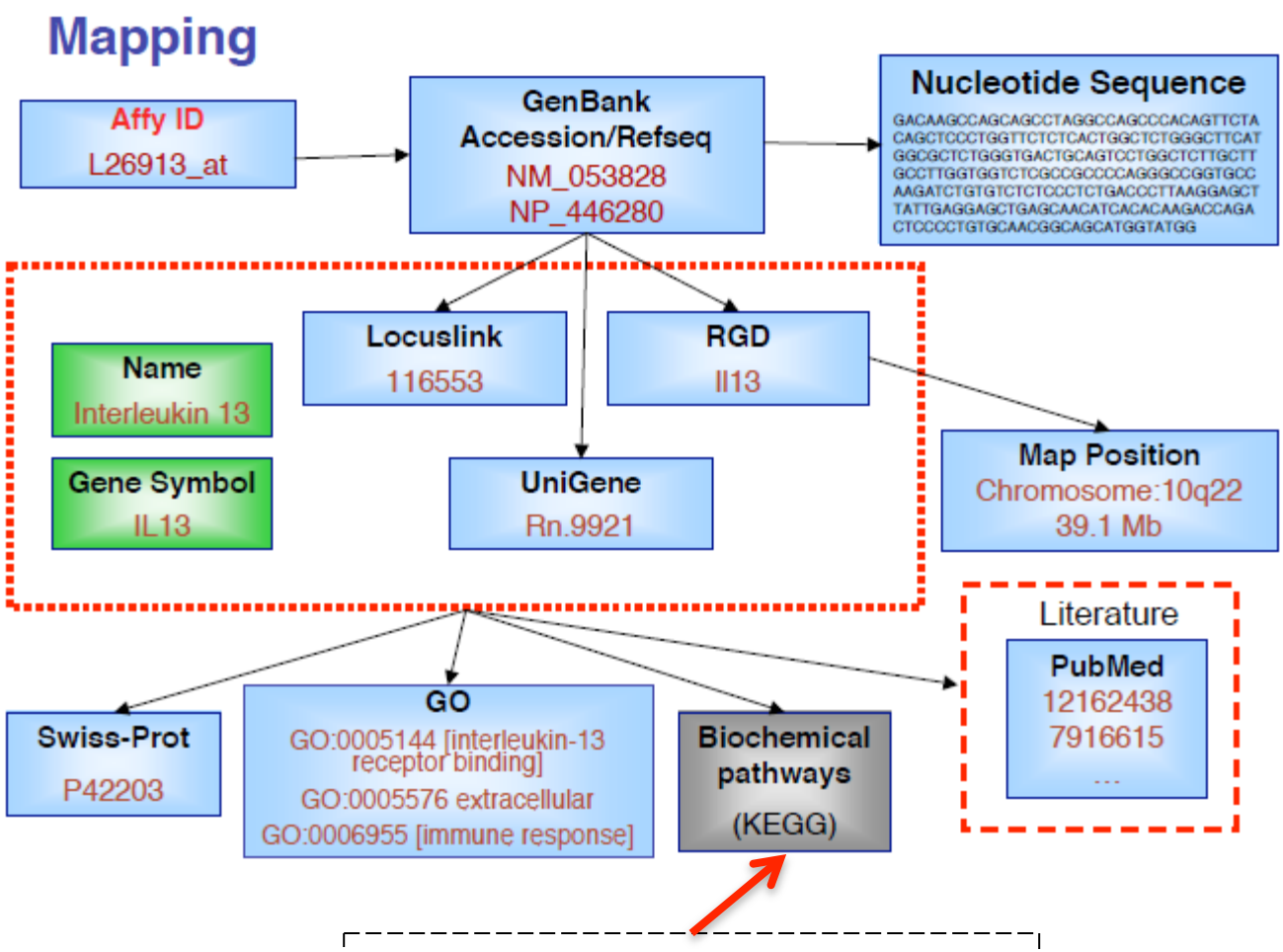
  41046_s_at "ZNF261"

- > gg<- getGO("41046_s_at",data="hgu95av2")
- > getGOdesc(gg[[1]], "MF")

  $"GO:0003677“

  "DNA binding activity"

Annotation mapping: example



Metabolic pathway are widely used in annotation mapping and testing

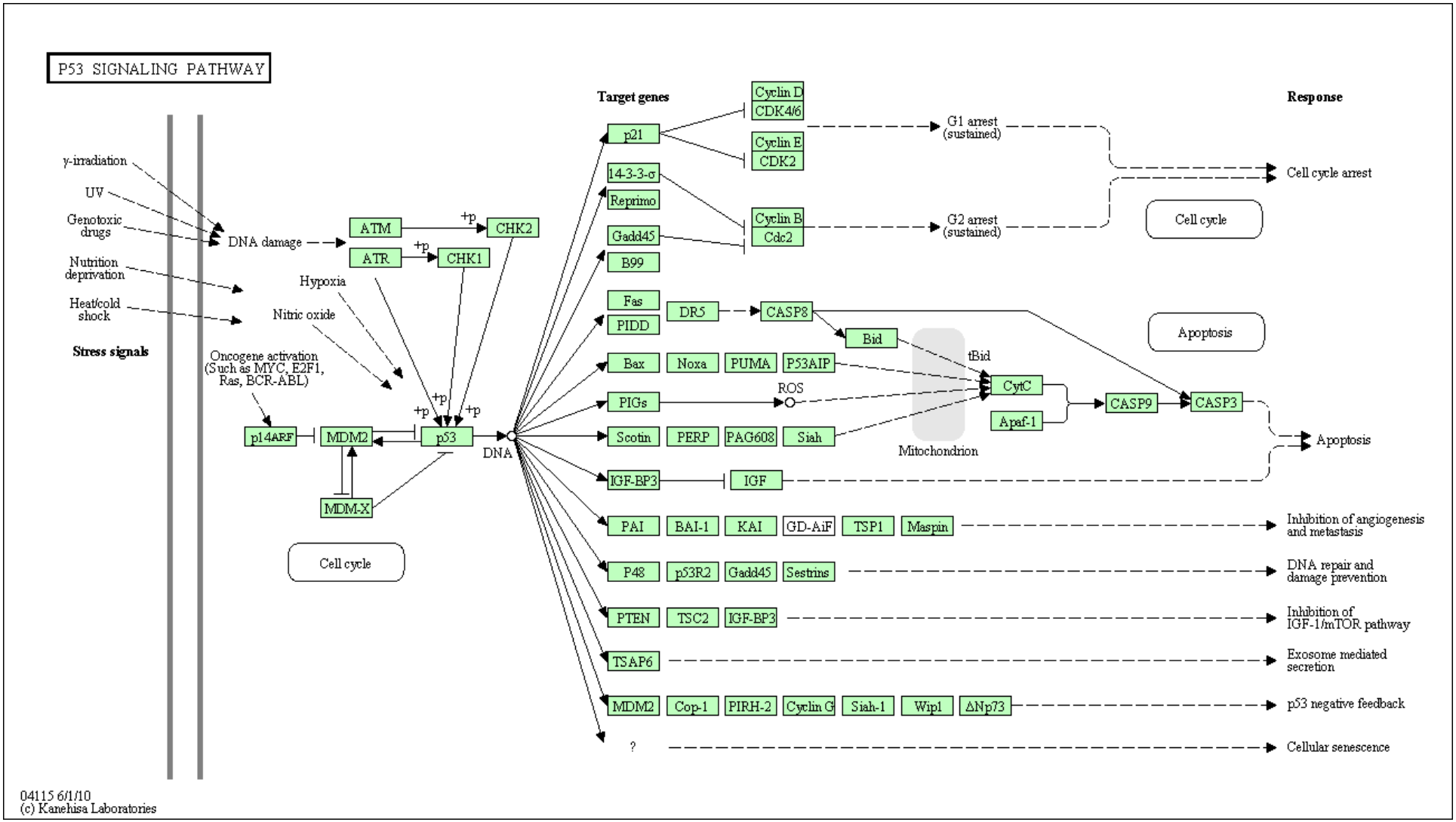
Metabolic Pathways

- 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)

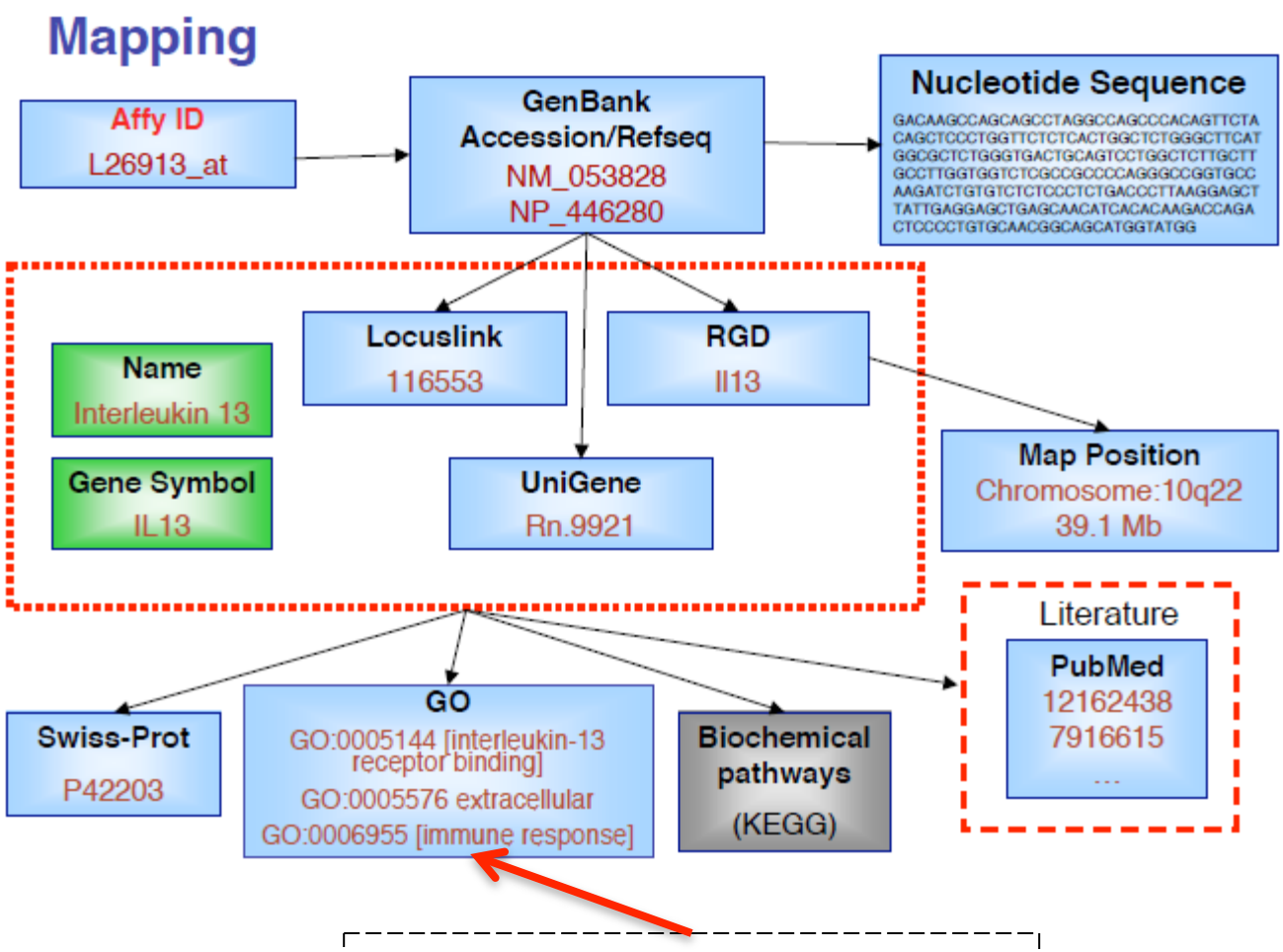
KEGG Pathway

- KEGG Pathways:
  - Manually curated pathway maps representing our 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) are widely used in annotation mapping and testing

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 does that mean?

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.

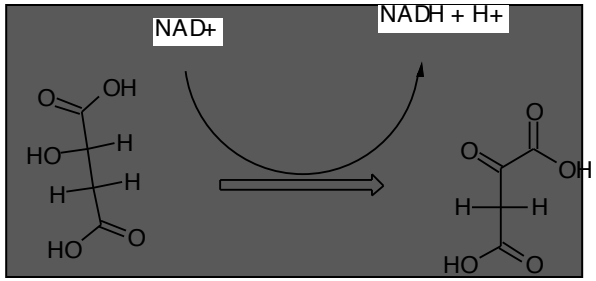
is part of

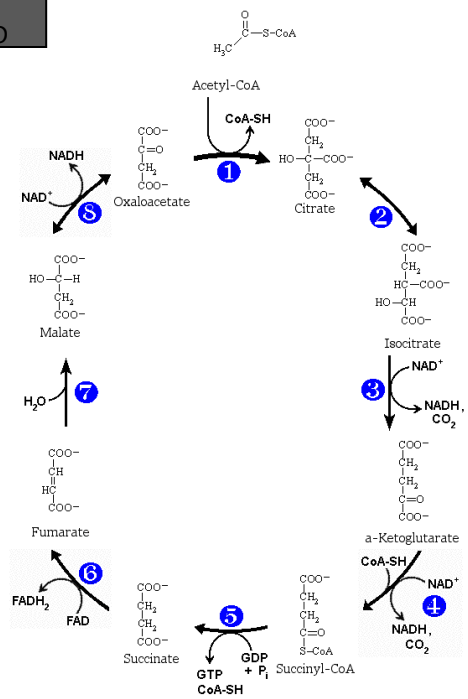
Gene Ontology (GO)

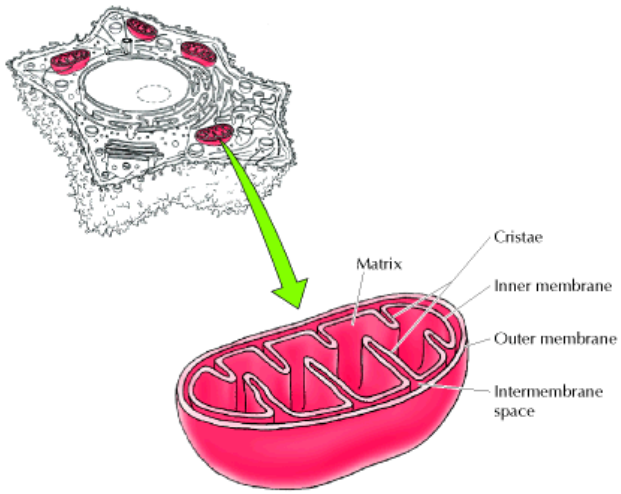
- Organized in 3 independent sets of ontologies in a tree structure
  - Molecular function (MF),
  - Biological process (BP),
  - Cellular compartment (CC)

The GO is Actually Three Ontologies

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





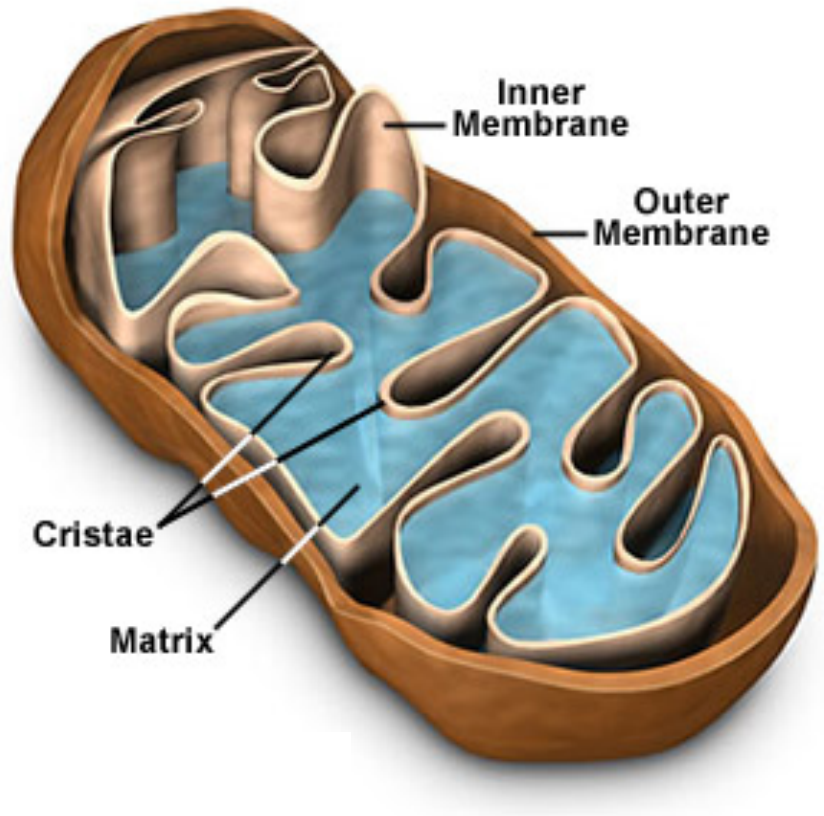


Biological Process
GO term: tricarboxylic acid cycle
Synonym: Krebs cycle
Synonym: citric acid cycle
GO id: GO:0006099

Cellular Component
GO term: mitochondrion
GO id: GO:0005739
Definition: A semiautonomous, self replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration.

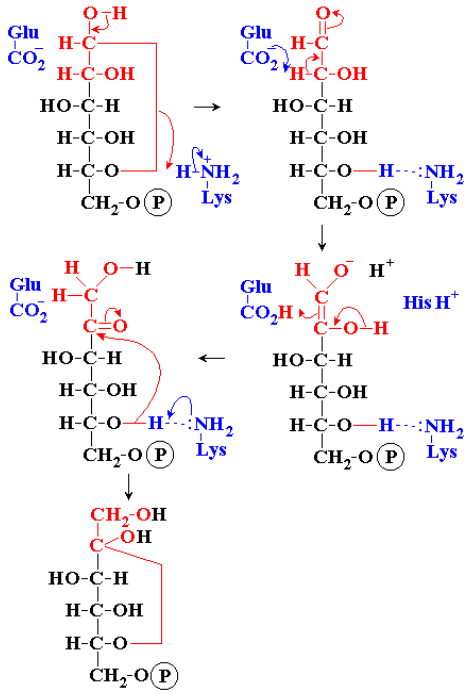
Cellular Component

- where a gene product acts



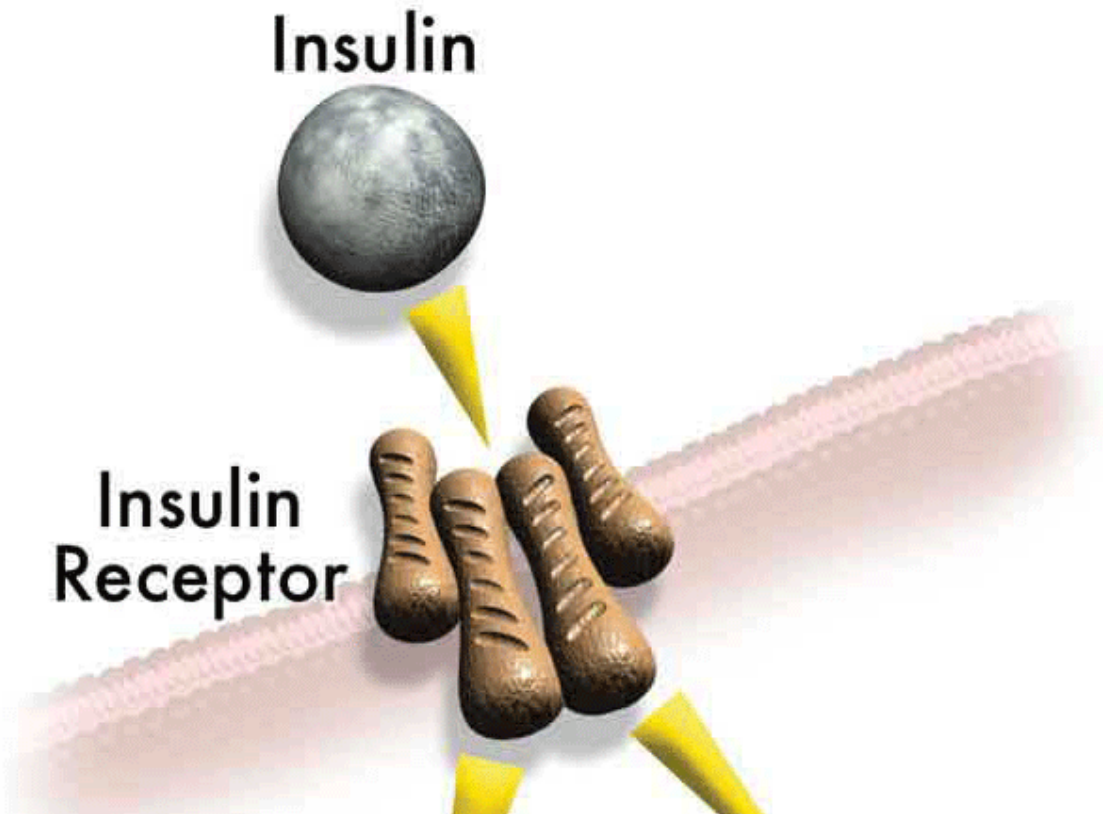
Molecular Function

- activities or "jobs" of a gene product



glucose-6-phosphate isomerase activity

Molecular Function



insulin binding

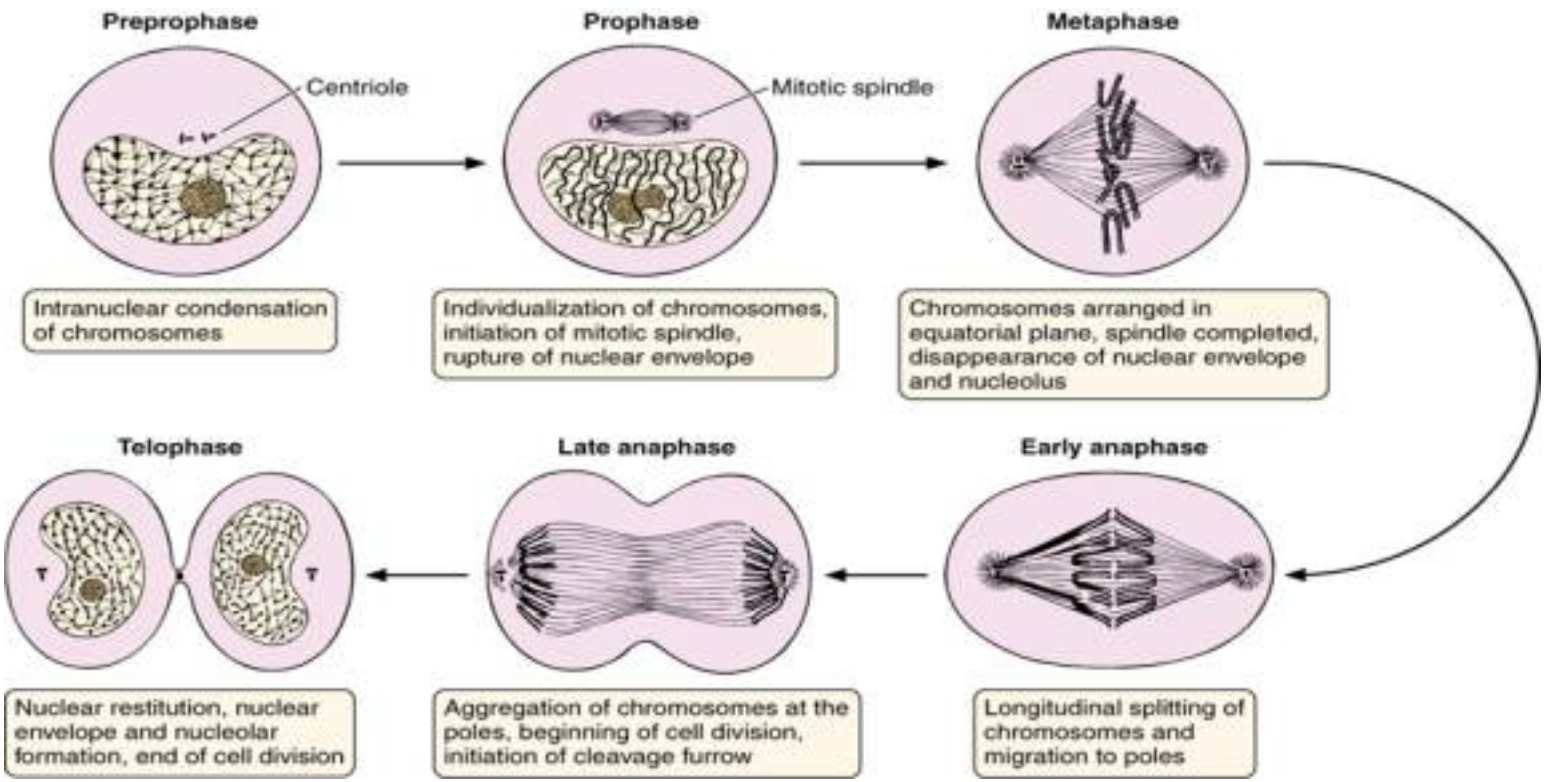
insulin receptor activity

Molecular Function

- A gene product may have several functions
- Sets of functions make up a biological process.

Biological Process

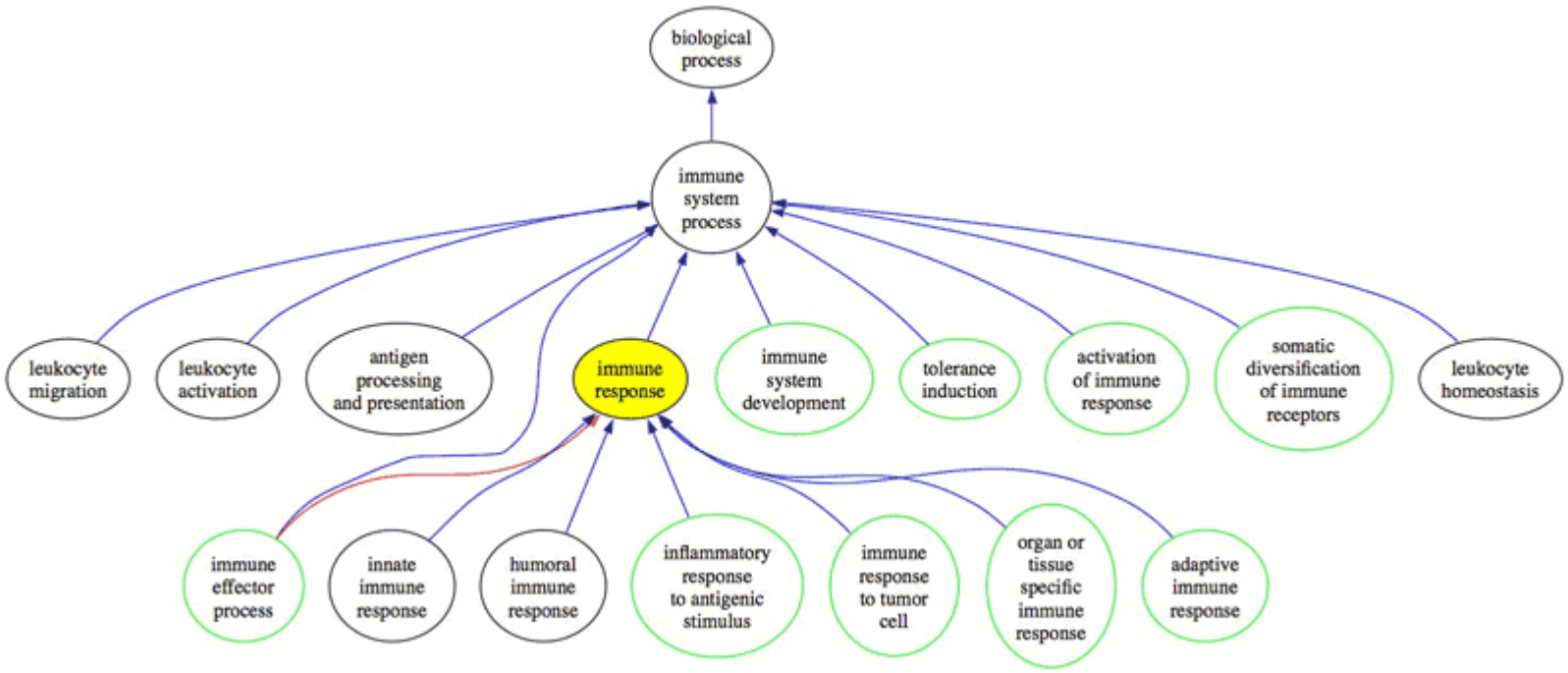
a commonly recognized series of events



cell division

Gene Ontology: Tree Structure

- Controlled networked terms
  - Parent / child network organized as a tree
  - Terms get more detailed as you move down the network



GO terms

Gene Ontology Browser

Term Detail

GO term:	**cell differentiation**
GO id:	**GO:0030154**
Definition:	**The process whereby relatively unspecialized cells, e.g. embryonic or regenerative cells, acquire specialized structural and/or functional features that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history.**

Gene_Ontology
- biological_process
  - cellular process
    - cell communication +
    - cell differentiation [GO:0030154] *(493 genes, 649 annotations)*
      - adipocyte differentiation +
      - antipodal cell differentiation +
      - cardiac cell differentiation +

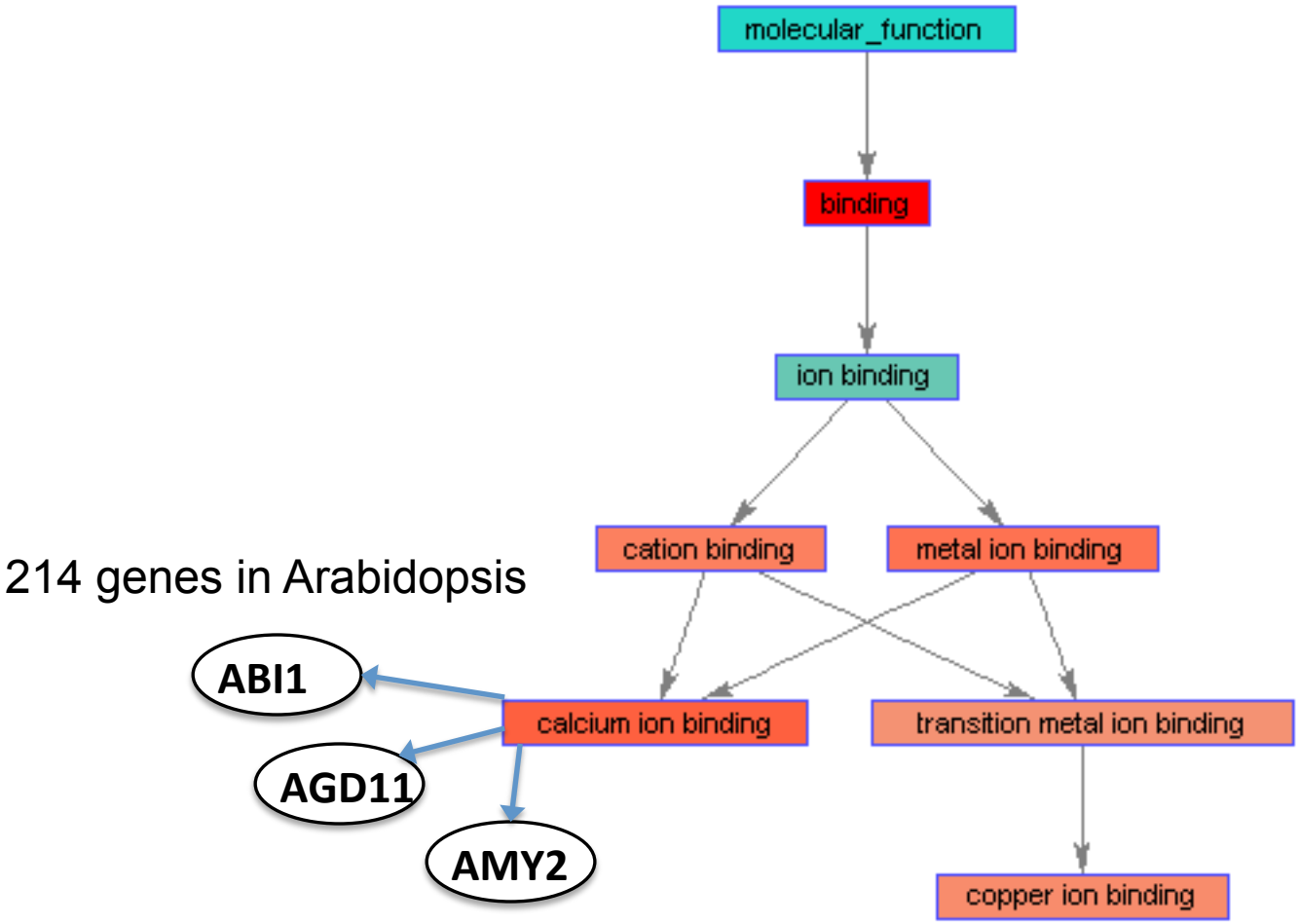
Gene Ontology: Rule

- 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

- The rule is that if a gene is a member of a term, it is also a member of the term's parents (or ancestors).

Gene Ontology: Rule

- The rule is that if a gene is a member of a term, it is also a member if the terms parents



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

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

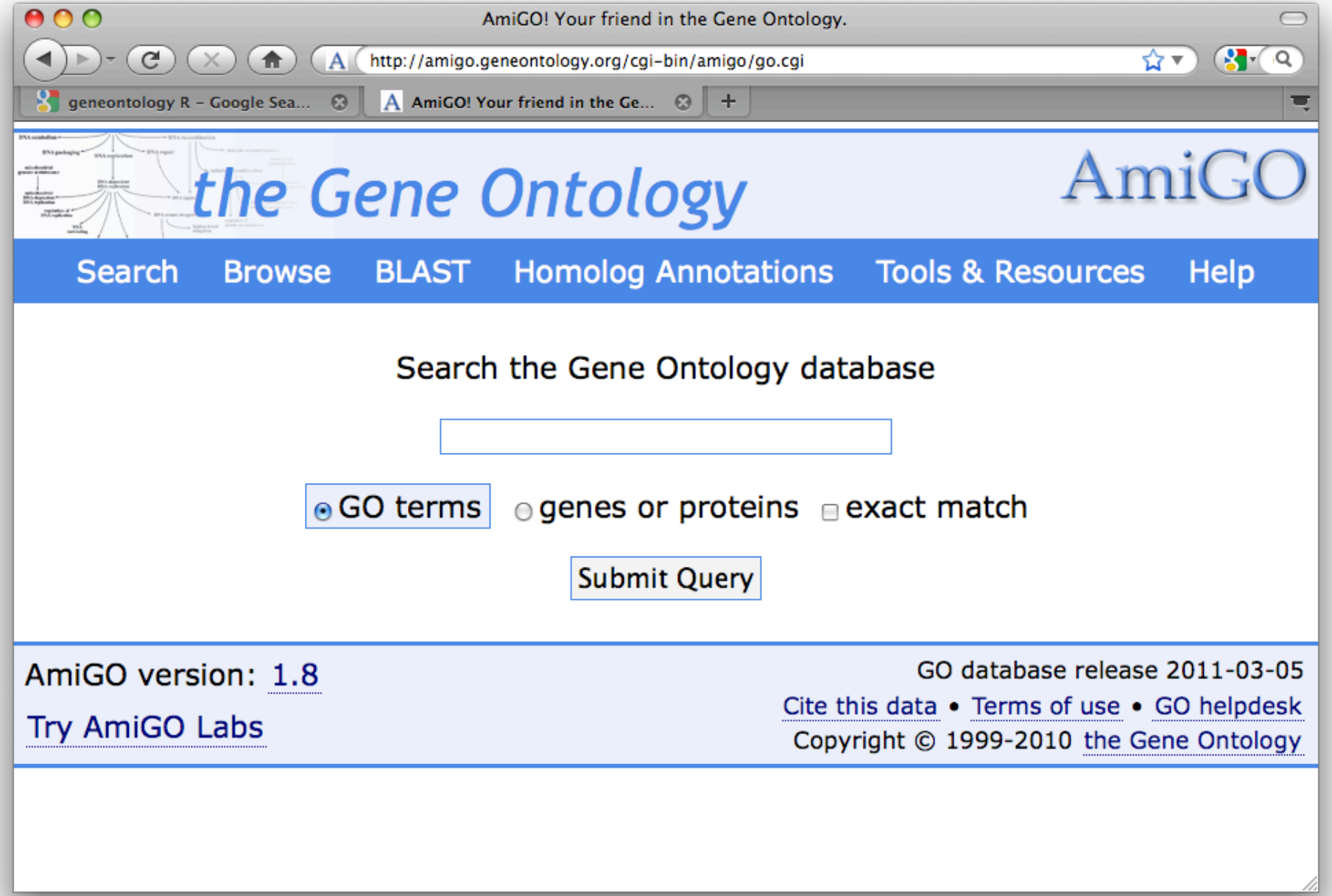
GO tools

- 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://www.geneontology.org/GO.tools

http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools

Gene Ontology: tools



Gene Ontology: tools

AmiGO: Term Details for GO:0005509

http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0005509

Apple Yahoo! Google Maps YouTube Wikipedia News (1191) Popular

calcium ion binding

Term information ↓ Term neighborhood ↓ External references ↓ 5907 gene product associations →

Term Information

Accession	GO:0005509
Ontology	**Molecular Function**
Synonyms	**related**: calcium ion storage activity
Definition	Interacting selectively and non-covalently with calcium ions (Ca2+). *Source:* GOC:ai
Comment	None
Subset	Prokaryotic GO subset
Community	Add usage comments for this term at GONUTS.

Back to top

Term Neighborhood for calcium ion binding (GO:0005509)

Filter lineage gene product counts

Data source: No filter, ASAP, AspGD, CGD

Species: No filter, A. fumigatus, A. thaliana, B. anthracis str. Ames

Inferred Tree View | Ancestors and Children | Graph View | Other Views | Downloads | Mappings

- GO:0003674 molecular_function [381106 gene products]
  - GO:0005488 binding [166439 gene products]
    - GO:0043167 ion binding [54418 gene products]
      - GO:0043169 cation binding [54332 gene products]
        - GO:0046872 metal ion binding [52505 gene products]
          - **GO:0005509 calcium ion binding [5907 gene products]**

Back to top

Grouping by Biological process

<u>Apoptosis</u>
Gene 1
Gene 53

<u>Mitosis</u>
Gene 2
Gene 5
Gene45
Gene 7
Gene 35
…

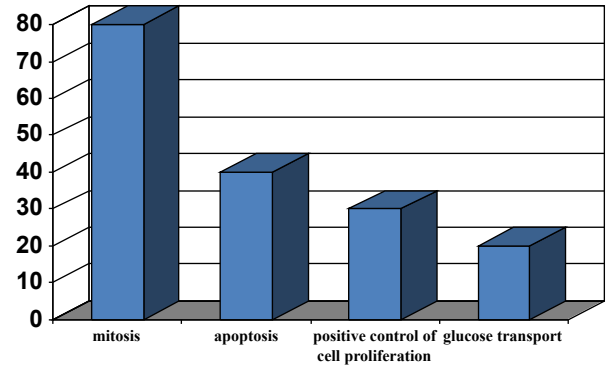
<u>Glucose transport</u>
Gene 7
Gene 3
Gene 6
…

<u>Positive ctrl. of cell prolif.</u>
Gene 7
Gene 3
Gene 12
…

<u>Growth</u>
Gene 5
Gene 2
Gene 6
…

Using GO in practice

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



microarray
1000 genes

→ experiment

→ 100 genes differentially regualted

→ mitosis – 80/100
apoptosis – 40/100
p. ctrl. cell prol. – 30/100
glucose transp. – 20/100

The problem

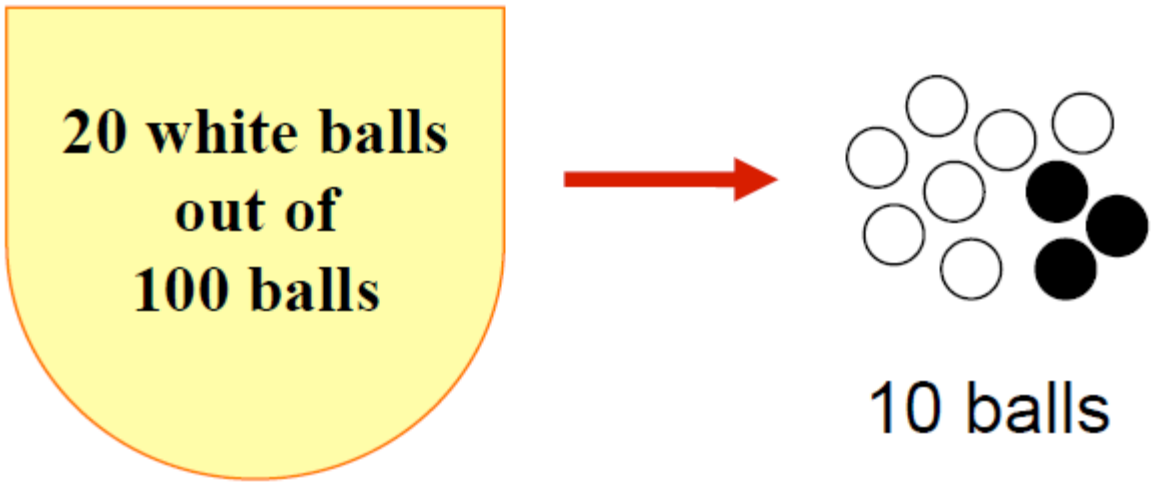
- After differential expression testing, we obtained a list of significantly differentially expressed probesets, controlled for false discovery
- We want to understand the biological insight behind this list
  1. we need to map the gene annotation information to these probesets
  2. we need to test/infer 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 testing or Fisher's exact test
    - Kolmogorov-Smirnov test or Wilcoxon signed rank test

Hypergeometric distribution

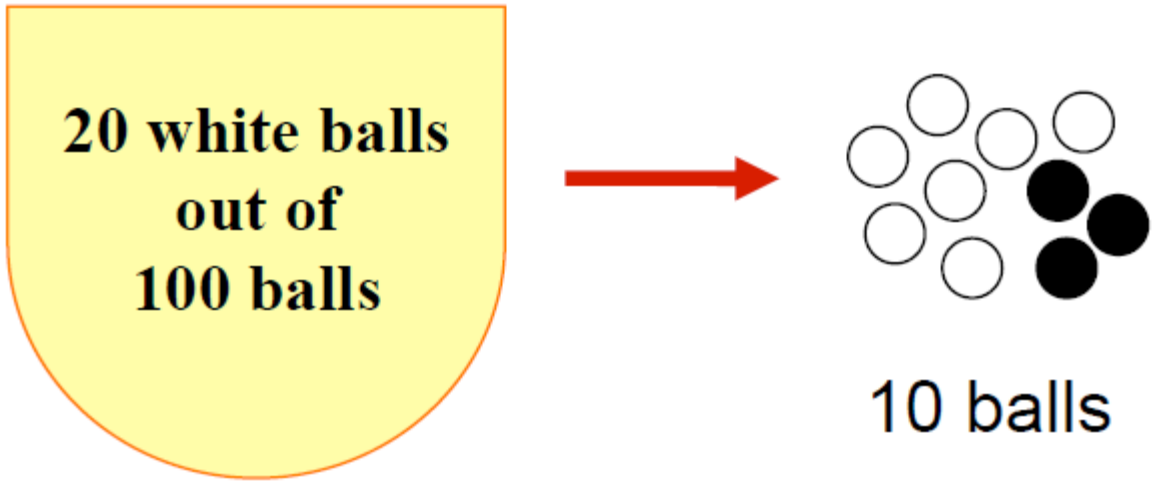
- The hypergeometric distribution arises from sampling from a fixed population.



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

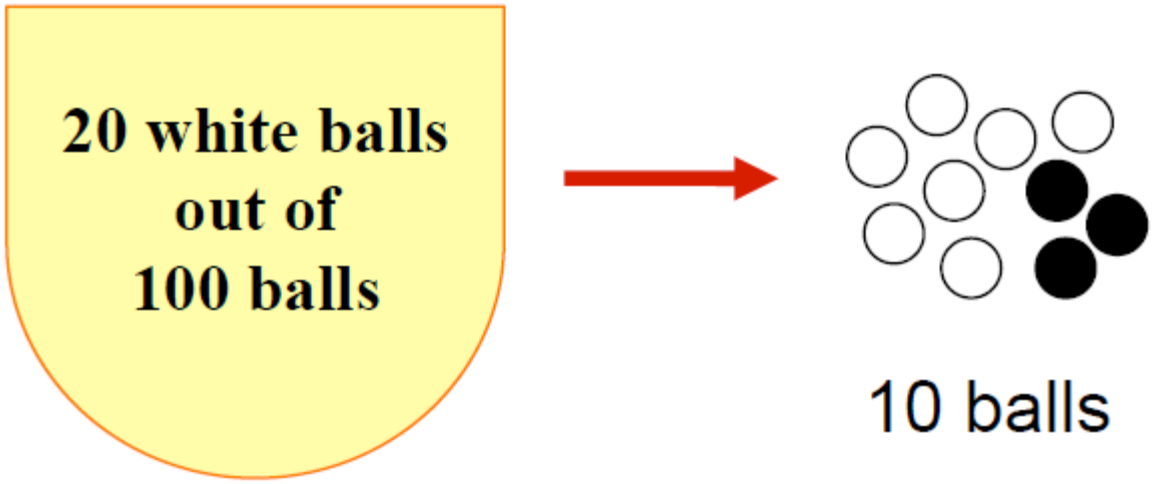
N=100
m=20
n=10
k=7

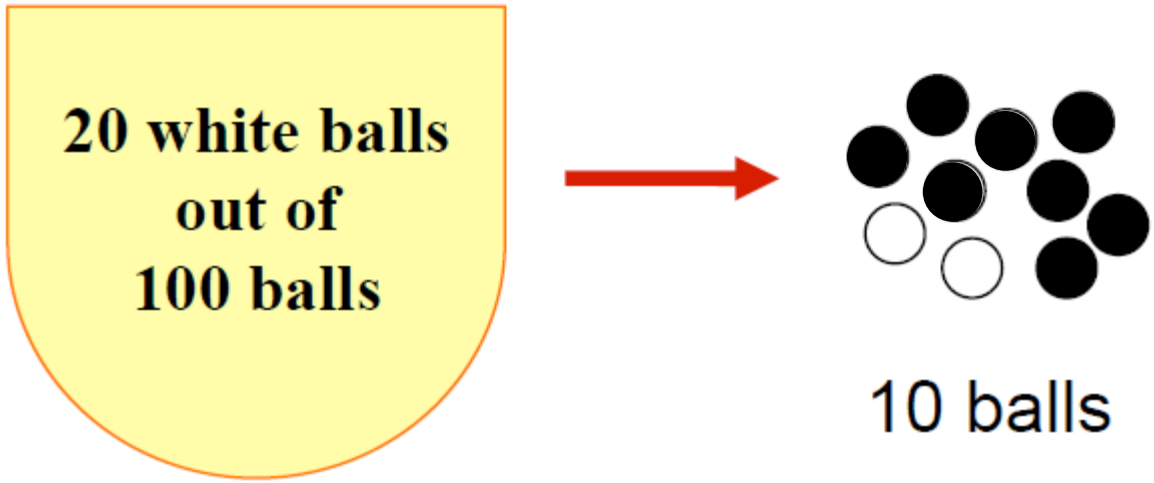
Hypergeometric test



- **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.

Hypergeometric test



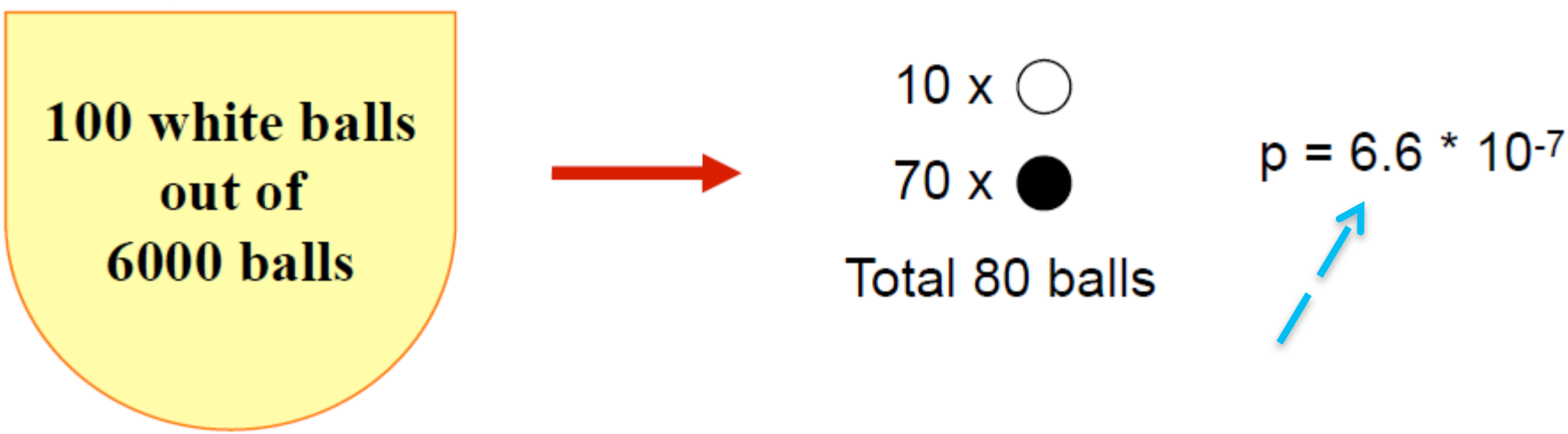


High possibility,
Easy to get
Not enriched with
white balls

Background

Annotation Testing (Hypergeometric test)

- Example: we obtained a list of 80 significant genes from a microarray 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 genes from micrroarray, 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 genes derived from microarray?



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 testing or Fisher's exact test
    - Kolmogorov-Smirnov test or Wilcoxon signed rank test

Annotation Testing (K-S test)

- Yeast has 6000 genes, and 100 of them can be mapped to a GO term called "Cell cycle". For the 80 significant genes from micrroarray, 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 genes derived from microarray?
  - K-S test will test the null hypothesis that x and y were drawn from the same *continuous* distribution
    - > ?ks.test
    - > ks.test (x, y)

10 genes in 100 genes mapped to "Cell Cycle"

70 genes in the rest 5900 genes not mapped to "Cell Cycle"

Annotation Testing (enrichment analysis)

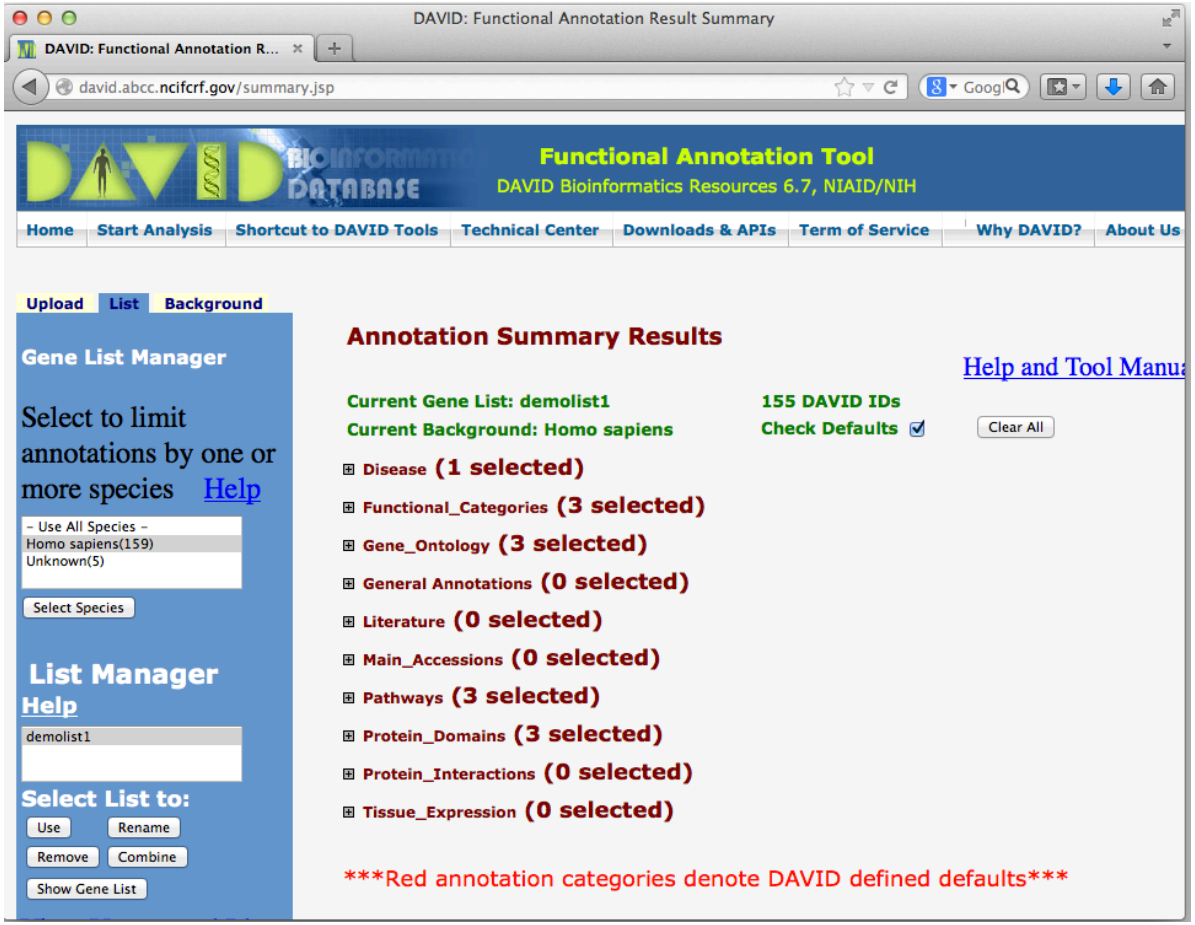
- Bioconductor tools using Hypergeometric testing or Fisher's exact test for enrichment analysis:
  - Gostat
    - » http://www.bioconductor.org/packages/2.3/bioc/html/GOstats.html
- Bioconductor tools using variant of K-S test for enrichment analysis:
  - PGSEA
    - » http://www.bioconductor.org/packages/2.4/bioc/html/PGSEA.html

Summary

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

DAVID: a function annotation tool

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



Midterm

- Will be posted before this Saturday.
- Midterm Exam is due by 3/23, Sunday, 11:59PM. Late submission is not accepted.
- Open book
- You can ask me, but cannot discuss with any other people.
- Including some topics, such as limma package, multiple test, enrichment test etc.