Pathway and Gene Set Analysis Part 1

Alison Motsinger-Reif, PhD
Branch Chief, Senior Investigator
Biostatistics and Computational Biology Branch
National Institute of Environmental Health Sciences

alison.motsinger-reif@niehs.nih.gove

The early steps of a microarray study

- Scientific Question (biological)
- Study design (biological/statistical)
- Conducting Experiment (biological)
- Preprocessing/Normalizing Data (statistical)
- Finding differentially expressed genes (statistical)

A data example

- Lee et al (2005) compared adipose tissue (abdominal subcutaenous adipocytes) between obese and lean Pima Indians
- Samples were hybridised on HGu95e-Affymetrix arrays (12639 genes/probe sets)
- Available as GDS1498 on the GEO database
- We selected the male samples only
  - 10 obese vs 9 lean

Diabetologia (2005) 48: 1776–1783
DOI 10.1007/s00125-005-1867-3

ARTICLE

Y. H. Lee · S. Nair · E. Rousseau · D. B. Allison ·
G. P. Page · P. A. Tataranni · C. Bogardus ·
P. A. Permana

Microarray profiling of isolated abdominal subcutaneous adipocytes from obese vs non-obese Pima Indians: increased expression of inflammation-related genes

Received: 10 December 2004 / Accepted: 28 April 2005 / Published online: 30 July 2005

Abstract *Aims/hypothesis:* Obesity increases the risk of developing major diseases such as diabetes and cardiovascular disease. Adipose tissue, particularly adipocytes, may play a major role in the development of obesity and its comorbidities. The aim of this study was to characterise, in adipocytes from obese people, the most differentially expressed genes that might be relevant to the development of obesity. *Methods:* We carried out microarray gene profiling of isolated abdominal subcutaneous adipocytes from 20 non-obese (BMI 25±3 kg/m^2) and 19 obese (BMI 55±8 kg/m^2) non-diabetic Pima Indians using Affymetrix HG-U95 GeneChip arrays. After data analyses, we measured the transcript levels of selected genes based on their biological functions and chromosomal positions using quantitative real-time PCR. *Results:* The most differentially expressed genes in adipocytes of obese individuals consisted of 433 upregulated and 244 downregulated genes. Of these, 410 genes could be classified into 20 functional Gene Ontology categories. The analyses indicated that the inflammation/immune response category was over-represented, and that most inflammation-related genes were upregulated in adipocytes of obese subjects. Quantitative real-time PCR confirmed the transcriptional upregulation of representative inflammation-related genes (*CCL2* and *CCL3*) encoding the chemokines monocyte chemoattractant protein-1 and macrophage inflammatory protein 1α. The differential expression levels of eight positional candidate genes, including inflammation-related *THY1* and *C1QTNF5*, were also confirmed. These genes are located on chromosome 11q22–q24, a region with linkage to obesity in the Pima Indians. *Conclusions/interpretation:* This study provides evidence supporting the active role of mature adipocytes in obesity-related inflammation. It also provides potential candidate genes for susceptibility to obesity.

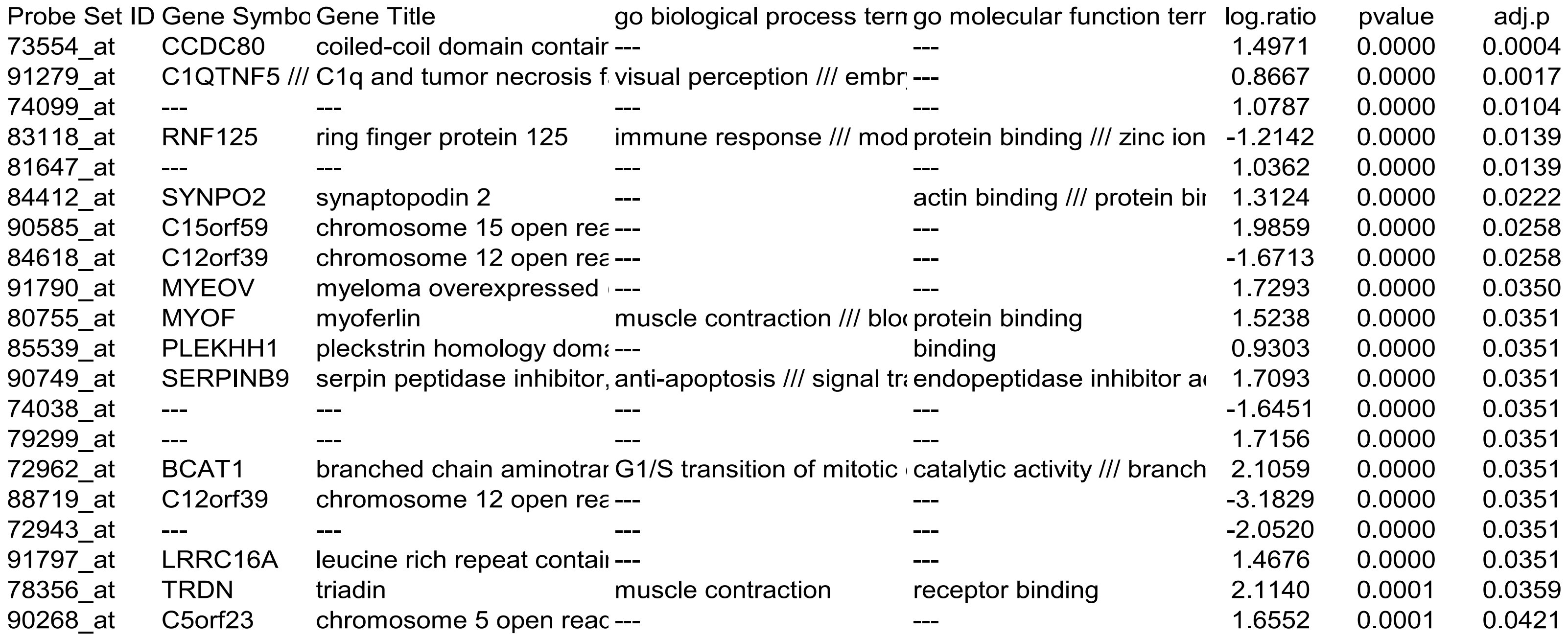
Electronic supplementary material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00125-005-1867-3.

The “Result”

Probe Set ID	log.ratio	pvalue	adj.p
73554_at	1.4971	0.0000	0.0004
91279_at	0.8667	0.0000	0.0017
74099_at	1.0787	0.0000	0.0104
83118_at	-1.2142	0.0000	0.0139
81647_at	1.0362	0.0000	0.0139
84412_at	1.3124	0.0000	0.0222
90585_at	1.9859	0.0000	0.0258
84618_at	-1.6713	0.0000	0.0258
91790_at	1.7293	0.0000	0.0350
80755_at	1.5238	0.0000	0.0351
85539_at	0.9303	0.0000	0.0351
90749_at	1.7093	0.0000	0.0351
74038_at	-1.6451	0.0000	0.0351
79299_at	1.7156	0.0000	0.0351
72962_at	2.1059	0.0000	0.0351
88719_at	-3.1829	0.0000	0.0351
72943_at	-2.0520	0.0000	0.0351
91797_at	1.4676	0.0000	0.0351
78356_at	2.1140	0.0001	0.0359
90268_at	1.6552	0.0001	0.0421

What happened to the Biology???

Slightly more informative results



Probe Set ID	Gene Symbo	Gene Title	go biological process ter	go molecular function ter	log.ratio	pvalue	adj.p
73554_at	CCDC80	coiled-coil domain contai	---	---	1.4971	0.0000	0.0004
91279_at	C1QTNF5 ///	C1q and tumor necrosis f	visual perception /// embr	---	0.8667	0.0000	0.0017
74099_at	---	---	---	---	1.0787	0.0000	0.0104
83118_at	RNF125	ring finger protein 125	immune response /// mod	protein binding /// zinc ion	-1.2142	0.0000	0.0139
81647_at	---	---	---	---	1.0362	0.0000	0.0139
84412_at	SYNPO2	synaptopodin 2	---	actin binding /// protein bi	1.3124	0.0000	0.0222
90585_at	C15orf59	chromosome 15 open re	---	---	1.9859	0.0000	0.0258
84618_at	C12orf39	chromosome 12 open re	---	---	-1.6713	0.0000	0.0258
91790_at	MYEOV	myeloma overexpressed	---	---	1.7293	0.0000	0.0350
80755_at	MYOF	myoferlin	muscle contraction /// blo	protein binding	1.5238	0.0000	0.0351
85539_at	PLEKHH1	pleckstrin homology dom	---	binding	0.9303	0.0000	0.0351
90749_at	SERPINB9	serpin peptidase inhibitor,	anti-apoptosis /// signal tr	endopeptidase inhibitor a	1.7093	0.0000	0.0351
74038_at	---	---	---	---	-1.6451	0.0000	0.0351
79299_at	---	---	---	---	1.7156	0.0000	0.0351
72962_at	BCAT1	branched chain aminotra	G1/S transition of mitotic	catalytic activity /// branch	2.1059	0.0000	0.0351
88719_at	C12orf39	chromosome 12 open re	---	---	-3.1829	0.0000	0.0351
72943_at	---	---	---	---	-2.0520	0.0000	0.0351
91797_at	LRRC16A	leucine rich repeat contai	---	---	1.4676	0.0000	0.0351
78356_at	TRDN	triadin	muscle contraction	receptor binding	2.1140	0.0001	0.0359
90268_at	C5orf23	chromosome 5 open rea	---	---	1.6552	0.0001	0.0421

If we are lucky, some of the top genes mean something to us

But what if they don't?

And how what are the results for other genes with similar biological functions

How to incorporate biological knowledge

- The type of knowledge we deal with is rather simple:

  We know groups/sets of genes that for example
  - Belong to the same pathway
  - Have a similar function
  - Are located on the same chromosome, etc…

- We will assume these groupings to be given, i.e. we will not yet discuss methods used to detect pathways, networks, gene clusters
  - We will later!

What is a pathway?

- No clear definition
  - Wikipedia: “In biochemistry, **metabolic pathways** are series of chemical reactions occurring within a cell. In each pathway, a principal chemical is modified by chemical reactions.”
  - These pathways describe enzymes and metabolites
- But often the word “pathway” is also used to describe gene regulatory networks or protein interaction networks
- In all cases a pathway describes a biological function very specifically

What is a Gene Set?

- Just what it says: a set of genes!
  - All genes involved in a pathway are an example of a Gene Set
  - All genes corresponding to a Gene Ontology term are a Gene Set
  - All genes mentioned in a paper of Smith et al might form a Gene Set
- A Gene Set is a much more general and less specific concept than a pathway
- Still: we will sometimes use two words interchangeably, as the analysis methods are mainly the same

Where Do Gene Sets/Lists Come From?

- Molecular profiling e.g. mRNA, protein
  - Identification → Gene list
  - Quantification → Gene list + values
  - Ranking, Clustering (biostatistics)
- Interactions: Protein interactions, Transcription factor binding sites (ChIP)
- Genetic screen e.g. of knock out library
- Association studies (Genome-wide)
  - Single nucleotide polymorphisms (SNPs)
  - Copy number variants (CNVs)
  - ……..

What is Gene Set/Pathway analysis?

- The aim is to give one number (score, p-value) to a Gene Set/Pathway
  - Are many genes in the pathway differentially expressed (up-regulated/downregulated)
  - Can we give a number (p-value) to the probability of observing these changes just by chance?

Goals

- Pathway and gene set data resources
    - Gene attributes
    - Database resources
        - GO, KeGG, Wikipathways, MsigDB
    - Gene identifiers and issues with mapping

- Differences between pathway analysis tools
    - Self contained vs. competitive tests
    - Cut-off methods vs. global methods
    - Issues with multiple testing

Goals

- Pathway and gene set data resources
  - Gene attributes
  - Database resources
    - GO, KeGG, Wikipathways, MsigDB
  - Gene identifiers and issues with mapping

- Differences between pathway analysis tools
  - Self contained vs. competitive tests
  - Cut-off methods vs. global methods
  - Issues with multiple testing

Gene Attributes

- Functional annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3’ UTR, microRNA binding sites
- Protein properties
  - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Gene Attributes

- Functional annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3' UTR, microRNA binding sites
- Protein properties
  - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

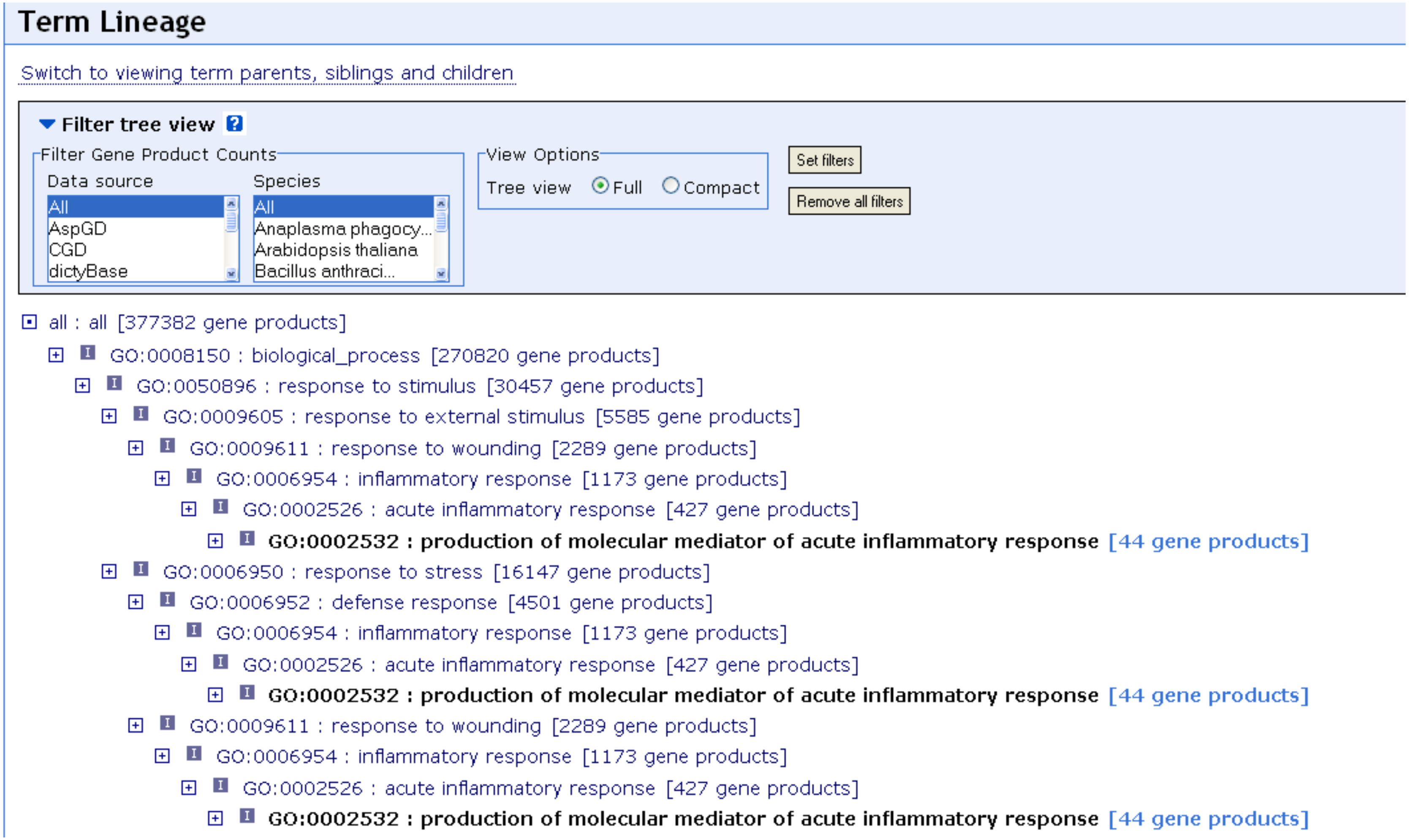
Database Resources

- Use functional annotation to aggregate genes into pathways/gene sets

- A number of databases are available
  - Different analysis tools link to different databases
  - Too many databases to go into detail on every one
  - Commonly used resources:
    - GO
    - KeGG
    - MsigDB
    - WikiPathways

Pathway and Gene Set data resources

- The Gene Ontology (GO) database
  - http://www.geneontology.org/
  - GO offers a relational/hierarchical database
  - Parent nodes: more general terms
  - Child nodes: more specific terms
  - At the end of the hierarchy there are genes/proteins
  - At the top there are 3 parent nodes: biological process, molecular function and cellular component
- Example: we search the database for the term “inflammation”



Term Lineage

Switch to viewing term parents, siblings and children

```
⊡ all : all  [377382 gene products]
   ⊞ I GO:0008150 : biological_process  [270820 gene products]
      ⊞ I GO:0050896 : response to stimulus  [30457 gene products]
         ⊞ I GO:0009605 : response to external stimulus  [5585 gene products]
            ⊞ I GO:0009611 : response to wounding  [2289 gene products]
               ⊞ I GO:0006954 : inflammatory response  [1173 gene products]
                  ⊞ I GO:0002526 : acute inflammatory response  [427 gene products]
                     ⊞ I GO:0002532 : production of molecular mediator of acute inflammatory response  [44 gene products]
         ⊞ I GO:0006950 : response to stress  [16147 gene products]
            ⊞ I GO:0006952 : defense response  [4501 gene products]
               ⊞ I GO:0006954 : inflammatory response  [1173 gene products]
                  ⊞ I GO:0002526 : acute inflammatory response  [427 gene products]
                     ⊞ I GO:0002532 : production of molecular mediator of acute inflammatory response  [44 gene products]
            ⊞ I GO:0009611 : response to wounding  [2289 gene products]
               ⊞ I GO:0006954 : inflammatory response  [1173 gene products]
                  ⊞ I GO:0002526 : acute inflammatory response  [427 gene products]
                     ⊞ I GO:0002532 : production of molecular mediator of acute inflammatory response  [44 gene products]
```

The genes on our array that code for one of the 44 gene products would form the corresponding “inflammation” gene set

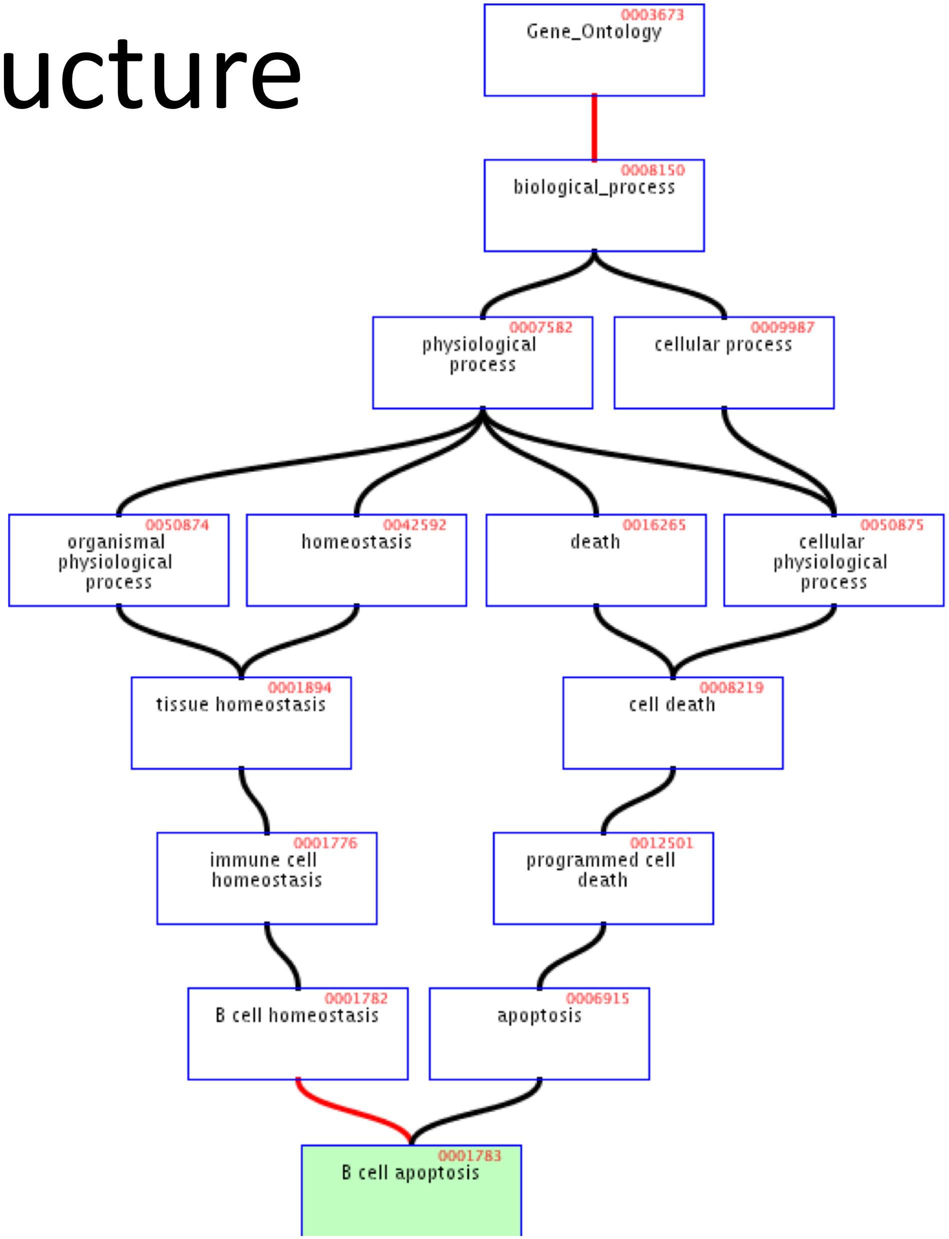
What is the Gene Ontology (GO)?

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

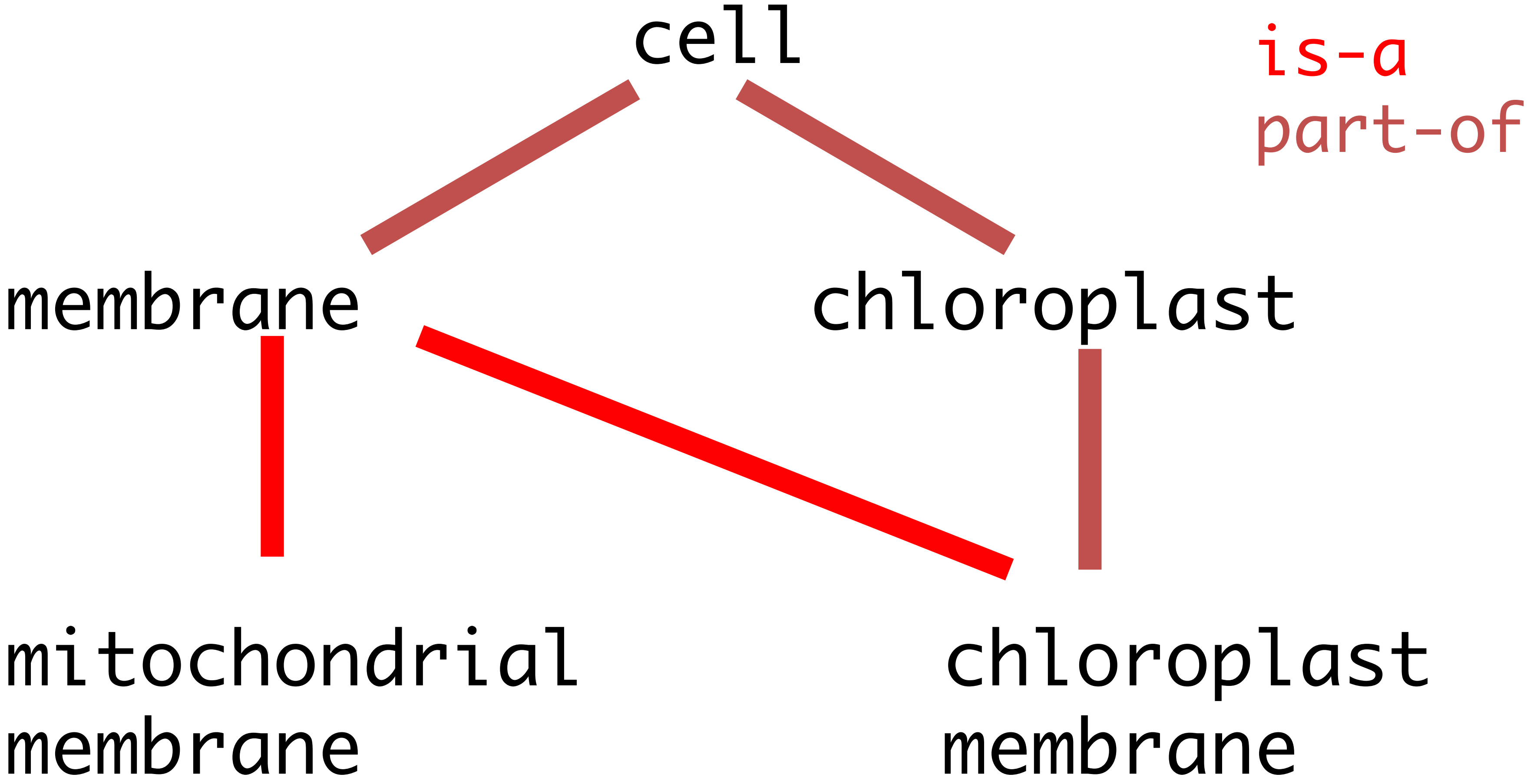
- Gaudet P., Dessimoz C. (2017) Gene Ontology: Pitfalls, Biases, and Remedies. In: Dessimoz C., Škunca N. (eds) The Gene Ontology Handbook. Methods in Molecular Biology, vol 1446. Humana Press, New York, NY

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

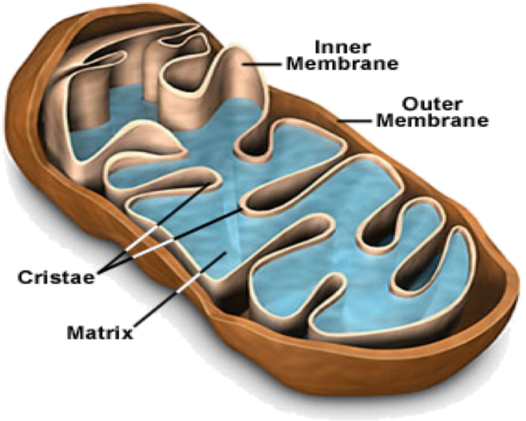
mitochondrial membrane

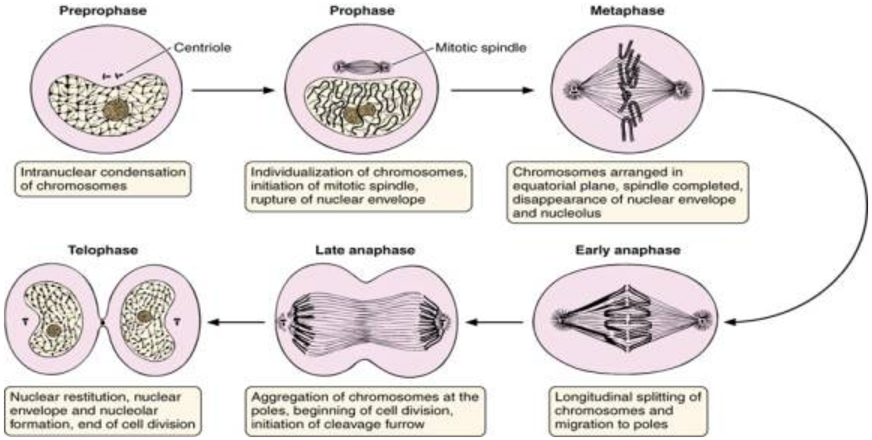
chloroplast membrane

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

What GO Covers?

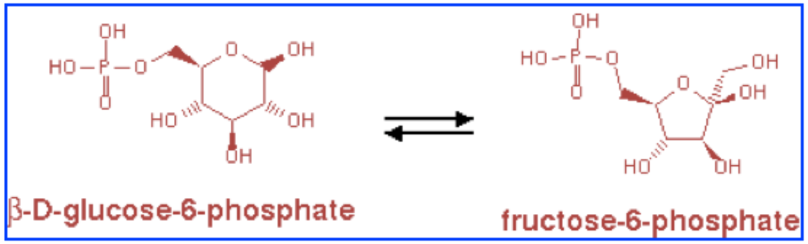
- GO terms divided into three aspects:
  - cellular component
  - molecular function
  - biological process




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

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

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)

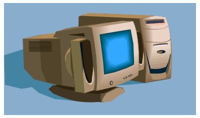
- Electronic annotation
  - Annotation derived without human validation
    - Computational predictions (accuracy varies)
    - Lower ‘quality’ than manual codes

- Key point: be aware of annotation origin

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

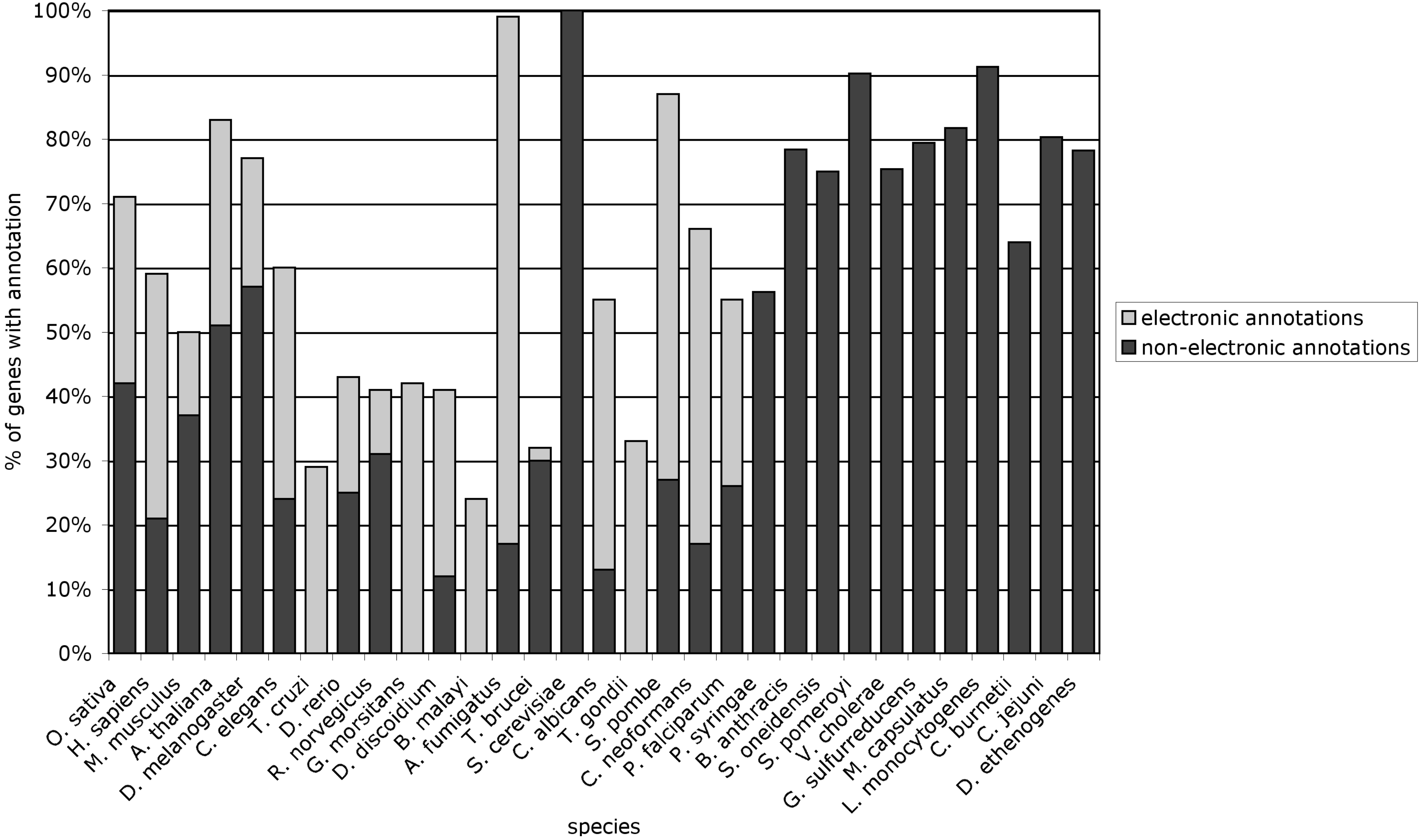
- **IEA:** Inferred from electronic annotation



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



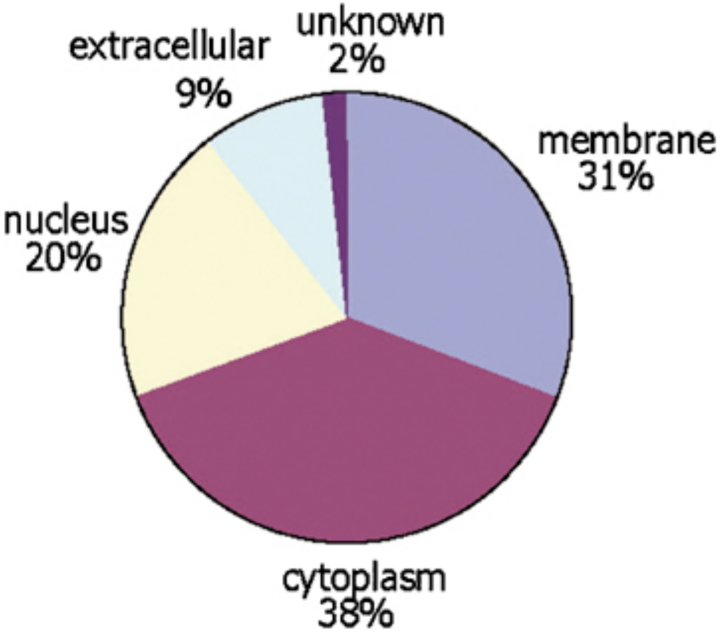
Lomax J. Get ready to GO! A biologist's guide to the Gene Ontology. Brief Bioinform. 2005 Sep;6(3):298-304.

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

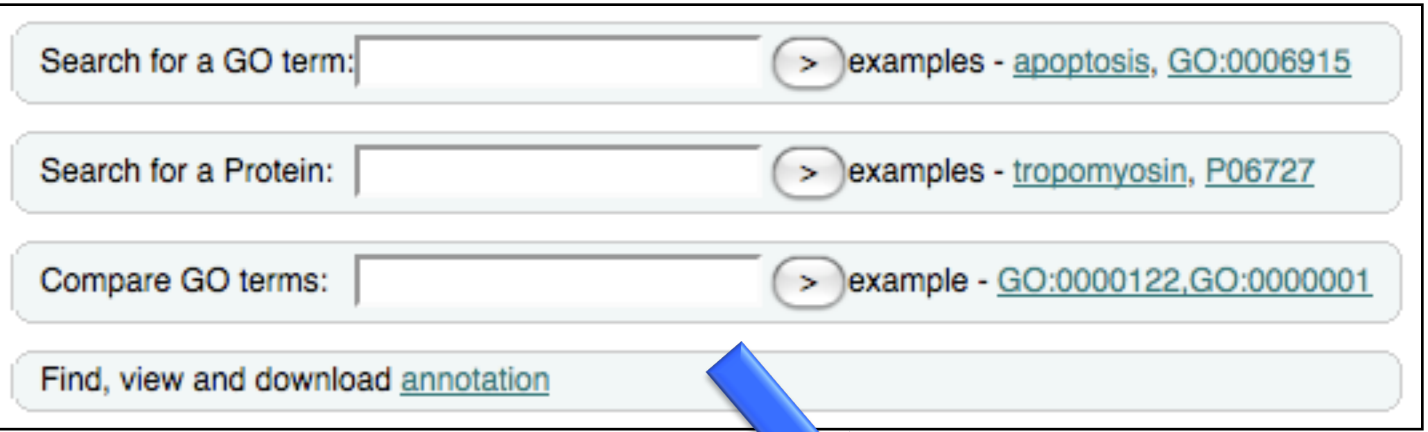
- GO has too many terms for some uses
  - Summaries (e.g. Pie charts)
- GO Slim is an official reduced set of GO terms
  - Generic, plant, yeast



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

Accessing GO: QuickGO

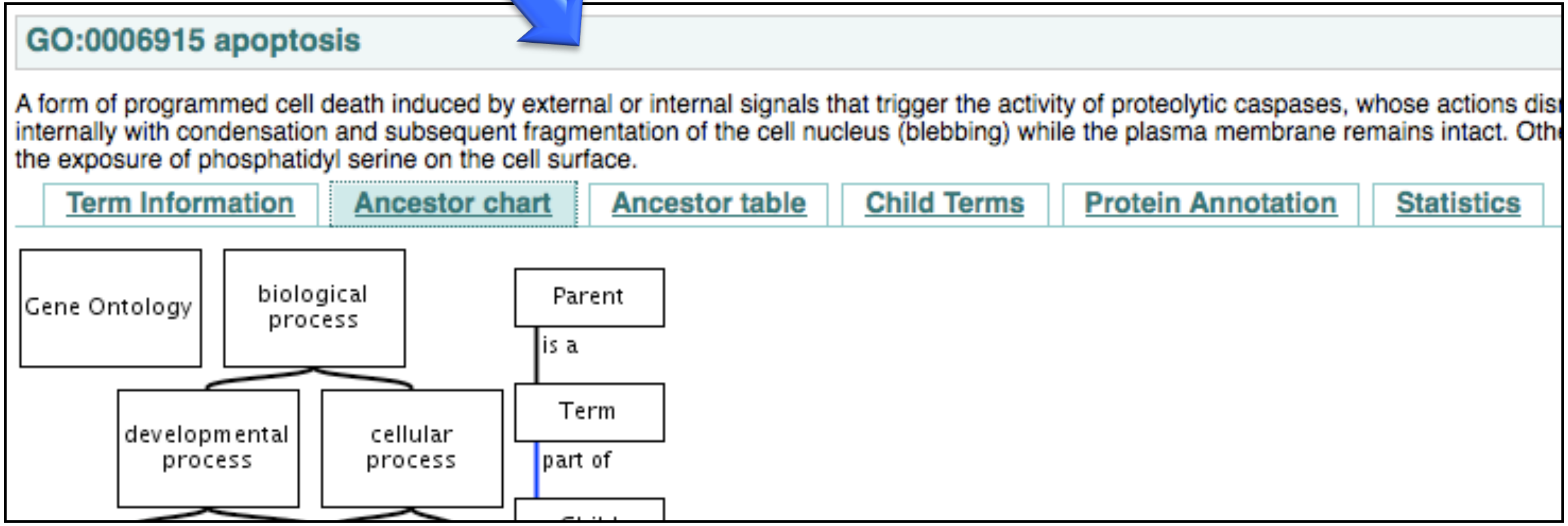


Search for a GO term: > examples - apoptosis, GO:0006915

Search for a Protein: > examples - tropomyosin, P06727

Compare GO terms: > example - GO:0000122,GO:0000001

Find, view and download annotation



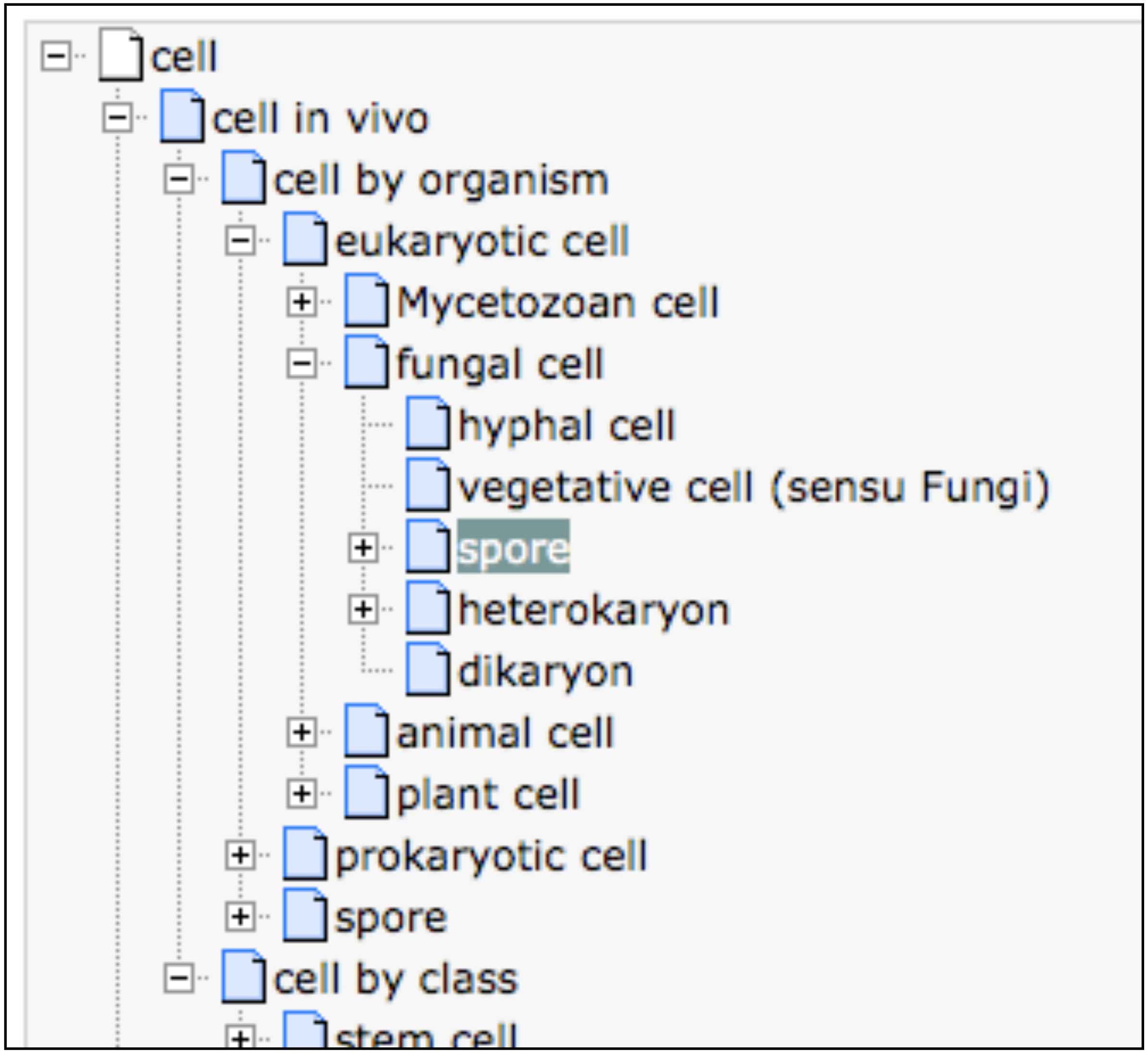
GO:0006915 apoptosis

A form of programmed cell death induced by external or internal signals that trigger the activity of proteolytic caspases, whose actions dis internally with condensation and subsequent fragmentation of the cell nucleus (blebbing) while the plasma membrane remains intact. Oth the exposure of phosphatidyl serine on the cell surface.

Term Information | Ancestor chart | Ancestor table | Child Terms | Protein Annotation | Statistics

http://www.ebi.ac.uk/ego/

Other Ontologies

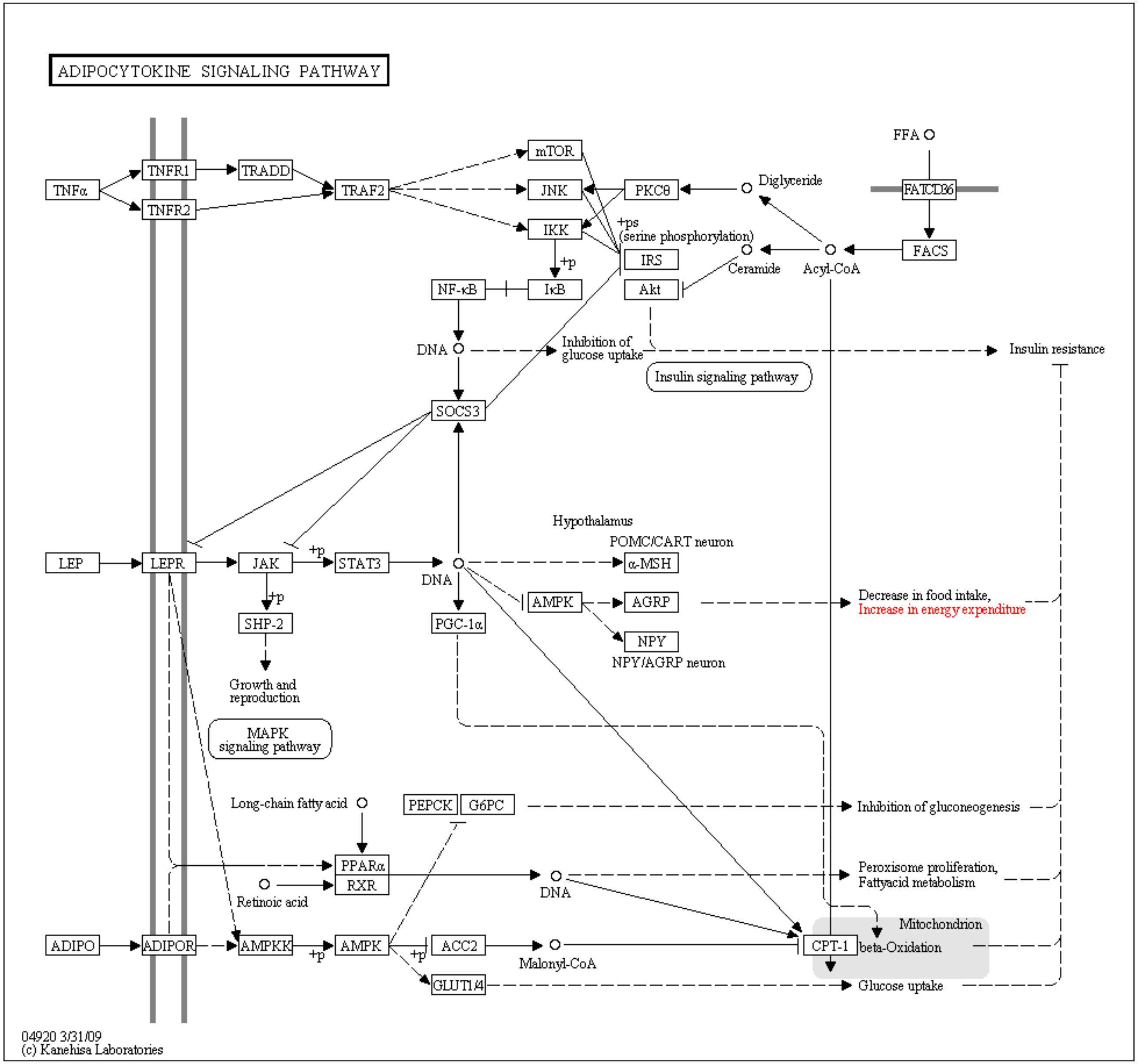


- cell
  - cell in vivo
    - cell by organism
      - eukaryotic cell
        - Mycetozoan cell
        - fungal cell
          - hyphal cell
          - vegetative cell (sensu Fungi)
          - spore
          - heterokaryon
          - dikaryon
        - animal cell
        - plant cell
      - prokaryotic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

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

KEGG pathway database

- KEGG = Kyoto Encyclopedia of Genes and Genomes
  - http://www.genome.jp/kegg/pathway.html
  - The pathway database gives far more detailed information than GO
    - Relationships between genes and gene products
  - But: this detailed information is only available for selected organisms and processes
  - Example: Adipocytokine signaling pathway

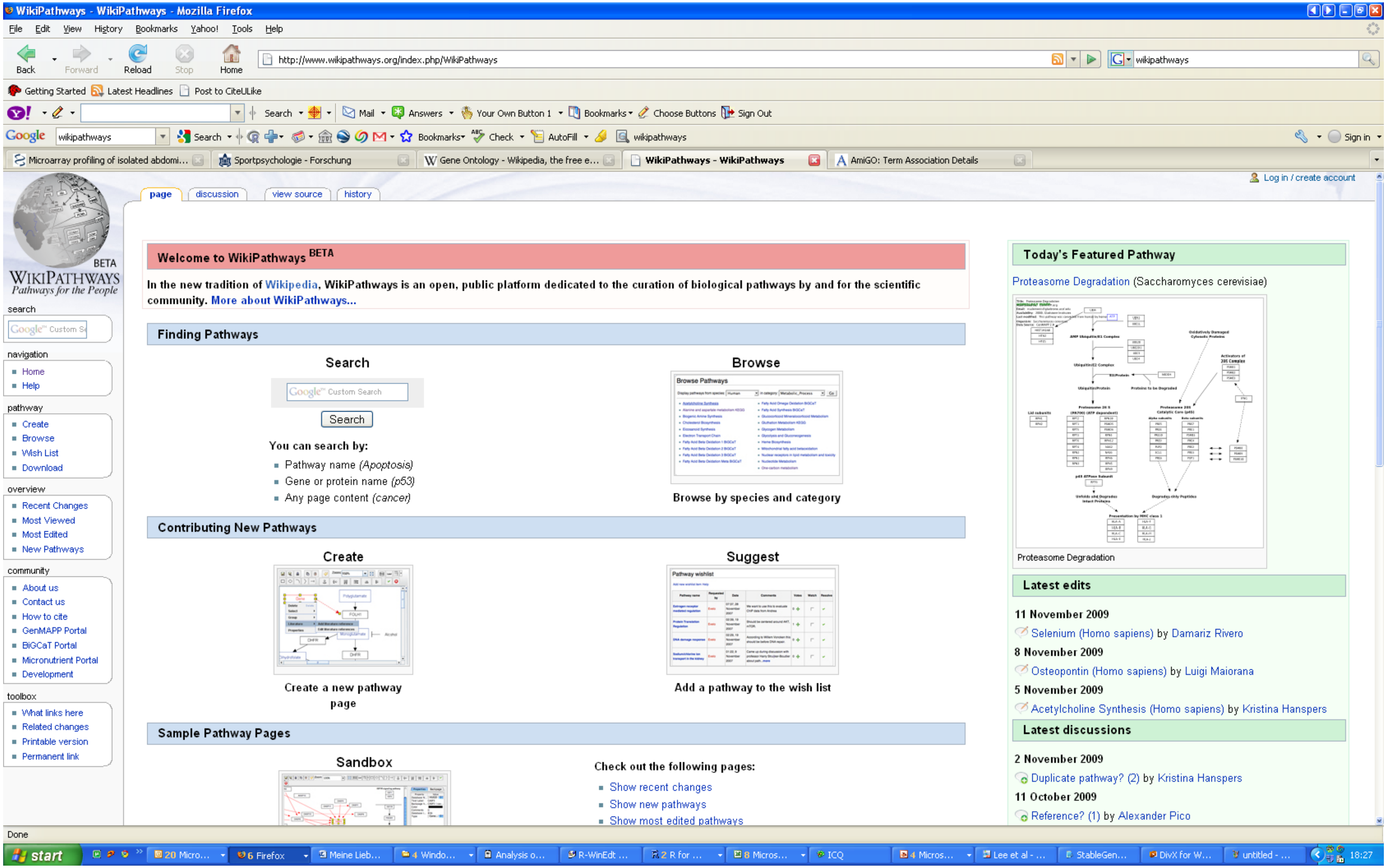

ADIPOCYTOKINE SIGNALING PATHWAY
TNFα
TNFR1
TNFR2
TRADD
TRAF2
mTOR
JNK
IKK
+p
IκB
NF-κB
PKCθ
Diglyceride
FFA
FAT/CD36
FACS
Acyl-CoA
Ceramide
+ps
(serine phosphorylation)
IRS
Akt
DNA
Inhibition of glucose uptake
Insulin resistance
Insulin signaling pathway
SOCS3
Hypothalamus
POMC/CART neuron
α-MSH
LEP
LEPR
JAK
+p
STAT3
DNA
AMPK
AGRP
NPY
NPY/AGRP neuron
Decrease in food intake,
Increase in energy expenditure
+p
SHP-2
PGC-1α
Growth and reproduction
MAPK signaling pathway
Long-chain fatty acid
PEPCK
G6PC
Inhibition of gluconeogenesis
PPARα
RXR
Retinoic acid
DNA
Peroxisome proliferation, Fattyacid metabolism
Mitochondrion
ADIPO
ADIPOR
AMPKK
+p
AMPK
+p
ACC2
Malonyl-CoA
CPT-1
beta-Oxidation
GLUT1/4
Glucose uptake
04920 3/31/09
(c) Kanehisa Laboratories

KEGG pathway database

- Clicking on the nodes in the pathway leads to more information on genes/proteins
  - Other pathways the node is involved with
  - Entries in Gene/Protein databases
  - References
  - Sequence information
- Ultimately this allows to find corresponding genes on the microarray and define a Gene Set for the pathway

Wikipathways

- http://www.wikipathways.org
- A wikipedia for pathways
  - One can see and download pathways
  - But also edit and contribute pathways
- The project is linked to the GenMAPP and Pathvisio analysis/visualisation tools


WikiPathways - WikiPathways - Mozilla Firefox
File Edit View History Bookmarks Yahoo! Tools Help
Back Forward Reload Stop Home
http://www.wikipathways.org/index.php/WikiPathways
wikipathways
Getting Started Latest Headlines Post to CiteULike
Search Mail Answers Your Own Button 1 Bookmarks Choose Buttons Sign Out
Google wikipathways Search Bookmarks Check AutoFill wikipathways Sign in
Microarray profiling of isolated abdomi...
Sportpsychologie - Forschung
Gene Ontology - Wikipedia, the free e...
WikiPathways - WikiPathways
AmiGO: Term Association Details
Log in / create account
page discussion view source history
BETA
WikiPathways
Pathways for the People
Done
start
20 Micro... 6 Firefox Meine Lieb... 4 Windo... Analysis o... R-WinEdt ... 2 R for ... 8 Micros... ICQ 4 Micros... Lee et al - ... StableGen... DivX for W... untitled - ...
18:27

search

navigation
- Home
- Help

pathway
- Create
- Browse
- Wish List
- Download

overview
- Recent Changes
- Most Viewed
- Most Edited
- New Pathways

community
- About us
- Contact us
- How to cite
- GenMAPP Portal
- BiGCaT Portal
- Micronutrient Portal
- Development

toolbox
- What links here
- Related changes
- Printable version
- Permanent link

Welcome to WikiPathways BETA

In the new tradition of Wikipedia, WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community. More about WikiPathways...

Finding Pathways

Search

Search

You can search by:

- Pathway name *(Apoptosis)*
- Gene or protein name *(p53)*
- Any page content *(cancer)*

Browse

Browse Pathways

Browse by species and category

Contributing New Pathways

Create

Create a new pathway page

Suggest

Pathway wishlist

Add a pathway to the wish list

Sample Pathway Pages

Sandbox

Check out the following pages:

- Show recent changes
- Show new pathways
- Show most edited pathways

Today's Featured Pathway

Proteasome Degradation (Saccharomyces cerevisiae)

Proteasome Degradation

Latest edits

11 November 2009

Selenium (Homo sapiens) by Damariz Rivero

8 November 2009

Osteopontin (Homo sapiens) by Luigi Maiorana

5 November 2009

Acetylcholine Synthesis (Homo sapiens) by Kristina Hanspers

Latest discussions

2 November 2009

Duplicate pathway? (2) by Kristina Hanspers

11 October 2009

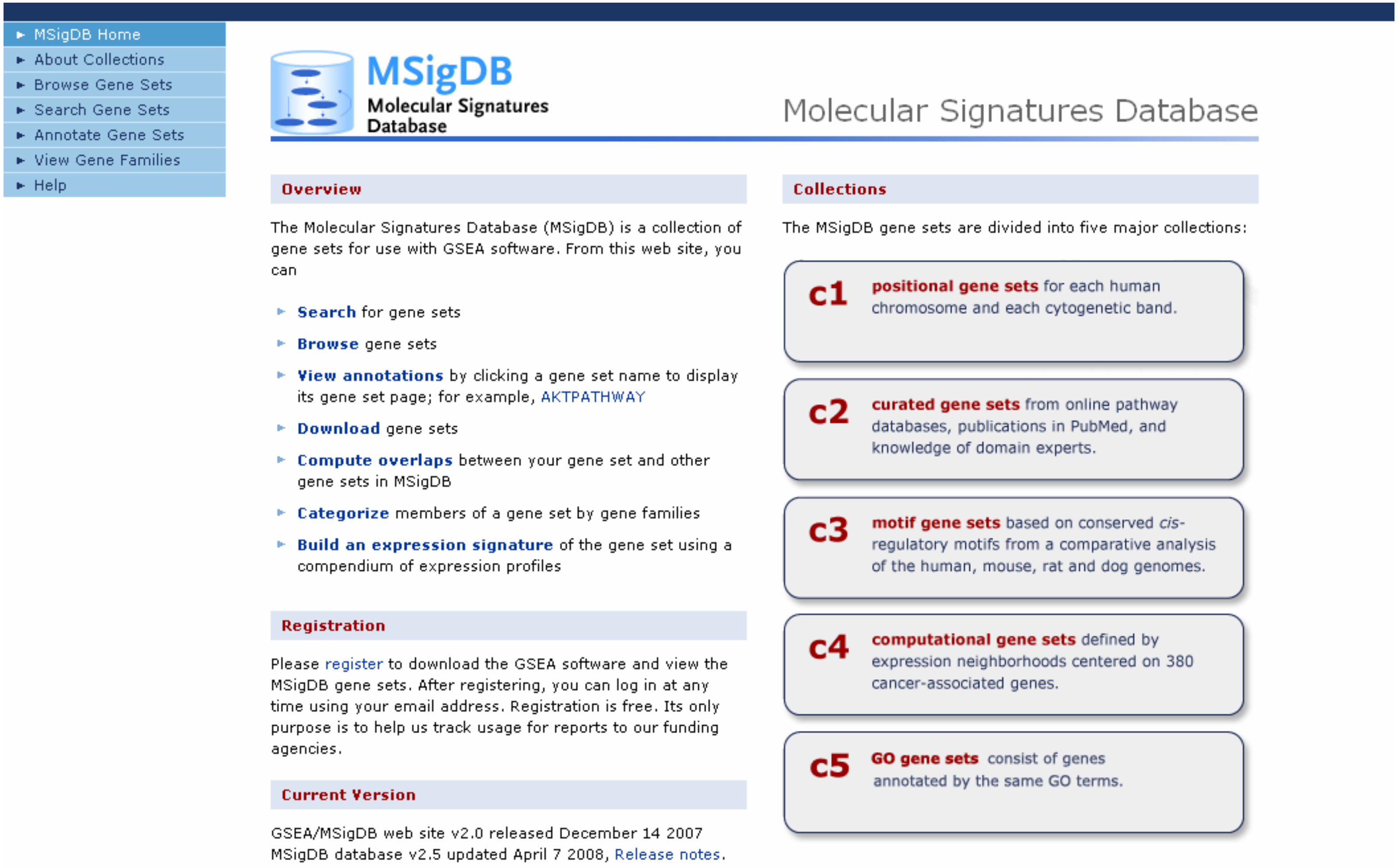
Reference? (1) by Alexander Pico

MSigDB

- MSigDB = Molecular Signature Database

http://www.broadinstitute.org/gsea/msigdb

- Related to the the analysis program GSEA
- MSigDB offers gene sets based on various groupings
  - Pathways
  - GO terms
  - Chromosomal position,…



Molecular Signatures Database

Overview

The Molecular Signatures Database (MSigDB) is a collection of gene sets for use with GSEA software. From this web site, you can

- **Search** for gene sets
- **Browse** gene sets
- **View annotations** by clicking a gene set name to display its gene set page; for example, AKTPATHWAY
- **Download** gene sets
- **Compute overlaps** between your gene set and other gene sets in MSigDB
- **Categorize** members of a gene set by gene families
- **Build an expression signature** of the gene set using a compendium of expression profiles

Registration

Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

GSEA/MSigDB web site v2.0 released December 14 2007
MSigDB database v2.5 updated April 7 2008, Release notes.

Collections

The MSigDB gene sets are divided into five major collections:

c1 **positional gene sets** for each human chromosome and each cytogenetic band.

c2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

c3 **motif gene sets** based on conserved *cis*-regulatory motifs from a comparative analysis of the human, mouse, rat and dog genomes.

c4 **computational gene sets** defined by expression neighborhoods centered on 380 cancer-associated genes.

c5 **GO gene sets** consist of genes annotated by the same GO terms.

Some Warnings

- In many cases the definition of a pathway/gene set in a database might differ from that of a scientist

- The nodes in pathways are often proteins or metabolites; the activity of the corresponding gene set is not necessarily a good measurement of the activity of the pathway

- There are many more resources out there (BioCarta, BioPax)

- Commercial packages often use their own pathway/gene set definitions (Ingenuity, Metacore, Genomatix,…)

- Genes in a gene set are usually not given by a Probe Set ID, but refer to some gene data base (Entrez IDs, Unigene IDs)
  - Conversion can lead to errors!

Some Warnings

- In many cases the definition of a pathway/gene set in a database might differ from that of a scientist
- The nodes in pathways are often proteins or metabolites; the activity of the corresponding gene set is not necessarily a good measurement of the activity of the pathway
- There are many more resources out there (BioCarta, BioPax)
- Commercial packages often use their own pathway/gene set definitions (Ingenuity, Metacore, Genomatix,…)
- Genes in a gene set are usually not given by a Probe Set ID, but refer to some gene data base (Entrez IDs, Unigene IDs)
    - Conversion can lead to errors!

Gene Attributes

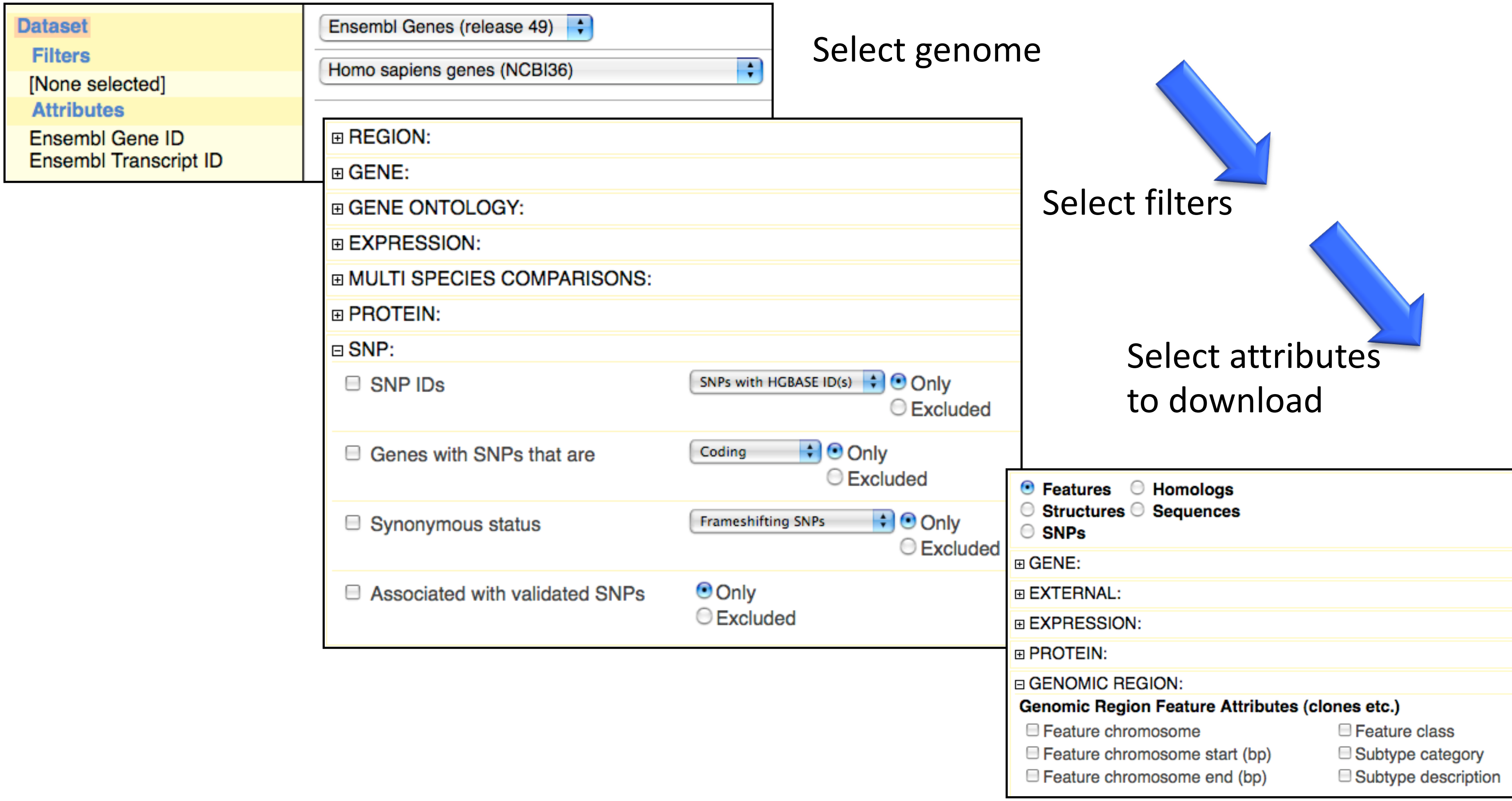
- Functional annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3’ UTR, microRNA binding sites
- Protein properties
  - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Sources of Gene Attributes

- Ensembl BioMart (eukaryotes)
  - http://www.ensembl.org
- Entrez Gene (general)
  - http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
- Model organism databases
  - E.g. SGD: http://www.yeastgenome.org/
- Many others…..

Ensembl BioMart

- Convenient access to gene list annotation



Dataset
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID

Ensembl Genes (release 49)
Homo sapiens genes (NCBI36)

Select genome

⊞ REGION:
⊞ GENE:
⊞ GENE ONTOLOGY:
⊞ EXPRESSION:
⊞ MULTI SPECIES COMPARISONS:
⊞ PROTEIN:
⊟ SNP:

□ SNP IDs — SNPs with HGBASE ID(s) — ● Only ○ Excluded
□ Genes with SNPs that are — Coding — ● Only ○ Excluded
□ Synonymous status — Frameshifting SNPs — ● Only ○ Excluded
□ Associated with validated SNPs — ● Only ○ Excluded

Select filters

Select attributes to download

● Features ○ Homologs
○ Structures ○ Sequences
○ SNPs

⊞ GENE:
⊞ EXTERNAL:
⊞ EXPRESSION:
⊞ PROTEIN:
⊟ GENOMIC REGION:

Genomic Region Feature Attributes (clones etc.)

□ Feature chromosome
□ Feature chromosome start (bp)
□ Feature chromosome end (bp)
□ Feature class
□ Subtype category
□ Subtype description

Gene and Protein Identifiers

- Identifiers (IDs) are ideally unique, stable names or numbers that help track database records
  - E.g. Social Insurance Number, Entrez Gene ID 41232

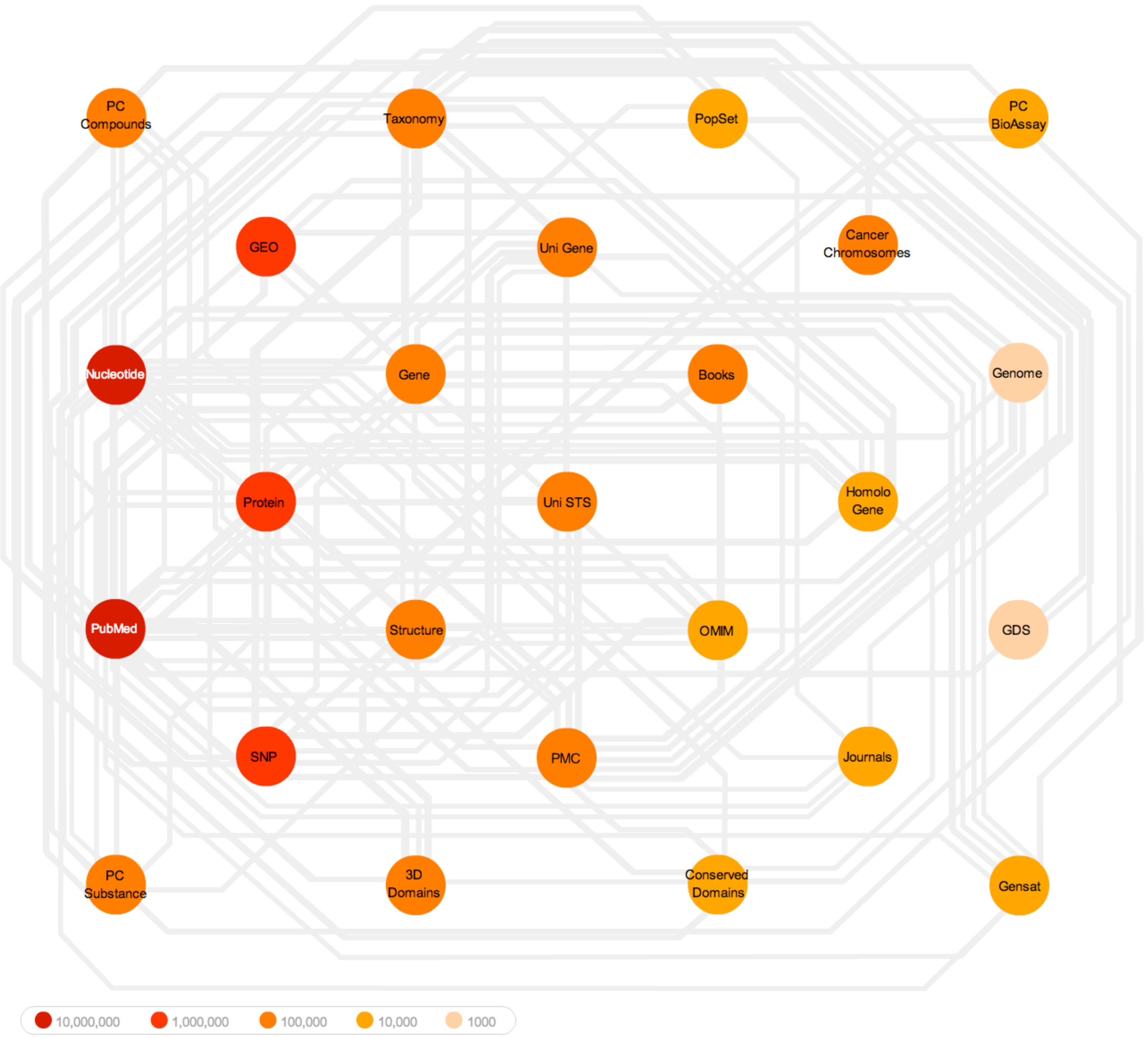
- Gene and protein information stored in many databases
  - → Genes have many IDs

- Records for: Gene, DNA, RNA, Protein
  - Important to recognize the correct record type
  - E.g. Entrez Gene records don’t store sequence. They link to DNA regions, RNA transcripts and proteins.

NCBI Database Links

NCBI:
U.S. National Center for Biotechnology Information

Part of National Library of Medicine (NLM)



http://www.ncbi.nlm.nih.gov/Database/datamodel/data_nodes.swf

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

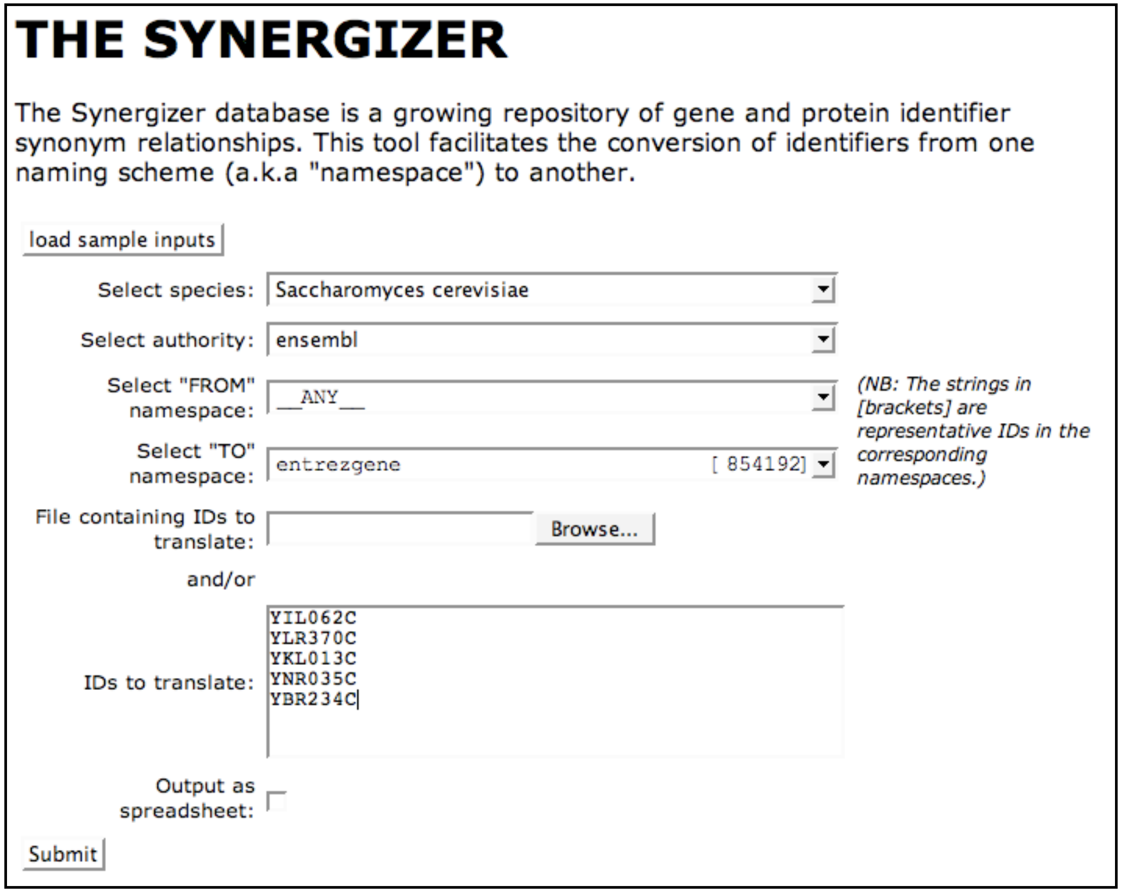
Red = Recommended

Identifier Mapping

- So many IDs!
  - Mapping (conversion) is a headache

- Four main uses
  - Searching for a favorite gene name
  - Link to related resources
  - Identifier translation
    - E.g. Genes to proteins, Entrez Gene to Affy
  - Unification during dataset merging
    - Equivalent records

ID Mapping Services



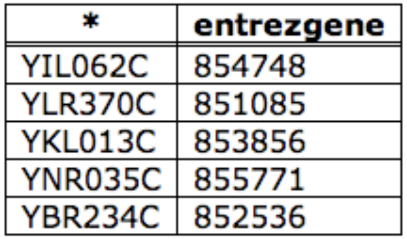
THE SYNERGIZER

The Synergizer database is a growing repository of gene and protein identifier synonym relationships. This tool facilitates the conversion of identifiers from one naming scheme (a.k.a "namespace") to another.

Select species:	Saccharomyces cerevisiae
Select authority:	ensembl
Select "FROM" namespace:	__ANY__
Select "TO" namespace:	entrezgene [854192]
File containing IDs to translate:	
and/or	
IDs to translate:	YIL062C YLR370C YKL013C YNR035C YBR234C
Output as spreadsheet:	□

(NB: The strings in [brackets] are representative IDs in the corresponding namespaces.)

Browse...
Submit



*	entrezgene
YIL062C	854748
YLR370C	851085
YKL013C	853856
YNR035C	855771
YBR234C	852536

- Synergizer
  - http://llama.med.harvard.edu/synergizer/translate/
- Ensembl BioMart
  - http://www.ensembl.org
- UniProt
  - http://www.uniprot.org/

ID Mapping Challenges

- Avoid errors: map IDs correctly
- Gene name ambiguity – not a good ID
  - e.g. FLJ92943, LFS1, TRP53, p53
  - Better to use the standard gene symbol: TP53
- Excel error-introduction
  - OCT4 is changed to October-4
- Problems reaching 100% coverage
  - E.g. due to version issues
  - Use multiple sources to increase coverage

Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics BMC Bioinformatics. 2004 Jun 23;5:80

Goals

- Pathway and gene set data resources
    - Gene attributes
    - Database resources
        - GO, KeGG, Wikipathways, MsigDB
    - Gene identifiers and issues with mapping

- Differences between pathway analysis tools
    - Self contained vs. competitive tests
    - Cut-off methods vs. global methods
    - Issues with multiple testing

Goals

- Pathway and gene set data resources
    - Gene attributes
    - Database resources
        - GO, KeGG, Wikipathways, MsigDB
    - Gene identifiers and issues with mapping

- Differences between pathway analysis tools
    - Self contained vs. competitive tests
    - Cut-off methods vs. global methods
    - Issues with multiple testing

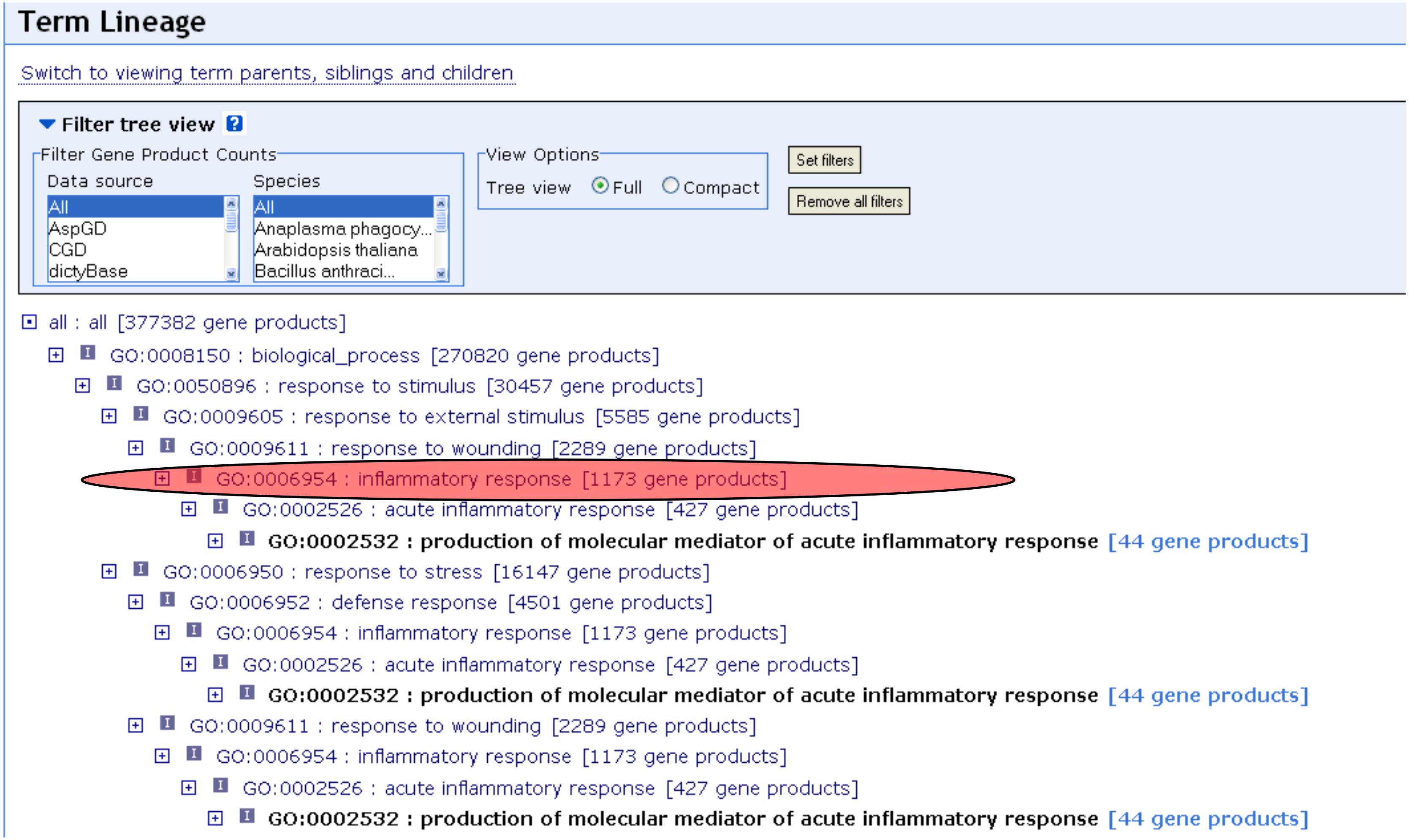
Aims of Analysis

- Reminder: The aim is to give one number (score, p-value) to a Gene Set/Pathway
  - Are many genes in the pathway differentially expressed (up-regulated/downregulated)?
  - Can we give a number (p-value) to the probability of observing these changes just by chance?
  - Similar to single gene analysis statistical hypothesis testing plays an important role

General differences between analysis tools

- Self contained vs competitive test
  - The distinction between “self-contained” and “competitive” methods goes back to Goeman and Buehlman (2007)
  - A self-contained method only uses the values for the genes of a gene set
    - The null hypothesis here is: H = {“No genes in the Gene Set are differentially expressed”}
  - A competitive method compares the genes within the gene set with the other genes on the arrays
    - Here we test against H: {“The genes in the Gene Set are not more differentially expressed than other genes”}

Example: Analysis for the GO-Term “inflammatory response” (GO:0006954)



Term Lineage

Switch to viewing term parents, siblings and children

▼ Filter tree view ?

Filter Gene Product Counts

Data source	Species
All	All
AspGD	Anaplasma phagocy...
CGD	Arabidopsis thaliana
dictyBase	Bacillus anthraci...

View Options

Tree view ● Full ○ Compact

Set filters

Remove all filters

```
⊡ all : all [377382 gene products]
  ⊞ I GO:0008150 : biological_process [270820 gene products]
    ⊞ I GO:0050896 : response to stimulus [30457 gene products]
      ⊞ I GO:0009605 : response to external stimulus [5585 gene products]
        ⊞ I GO:0009611 : response to wounding [2289 gene products]
          ⊞ I GO:0006954 : inflammatory response [1173 gene products]
            ⊞ I GO:0002526 : acute inflammatory response [427 gene products]
              ⊞ I GO:0002532 : production of molecular mediator of acute inflammatory response [44 gene products]
      ⊞ I GO:0006950 : response to stress [16147 gene products]
        ⊞ I GO:0006952 : defense response [4501 gene products]
          ⊞ I GO:0006954 : inflammatory response [1173 gene products]
            ⊞ I GO:0002526 : acute inflammatory response [427 gene products]
              ⊞ I GO:0002532 : production of molecular mediator of acute inflammatory response [44 gene products]
        ⊞ I GO:0009611 : response to wounding [2289 gene products]
          ⊞ I GO:0006954 : inflammatory response [1173 gene products]
            ⊞ I GO:0002526 : acute inflammatory response [427 gene products]
              ⊞ I GO:0002532 : production of molecular mediator of acute inflammatory response [44 gene products]
```

Back to the Real Data Example

- Using Bioconductor software we can find 96 probesets on the array corresponding to this term

- 8 out of these have a p-value < 5%

- How many significant genes would we expect by chance?

- Depends on how we define "by chance"

The “self-contained” version

- By chance (i.e. if it is NOT differentially expressed) a gene should be significant with a probability of 5%
- We would expect 96 x 5% = 4.8 significant genes
- Using the binomial distribution we can calculate the probability of observing 8 or more significant genes as p = 0.108, i.e. not quite significant

The “competitive” version

- Overall 1272 out of 12639 genes are significant in this data set (10.1%)
- If we randomly pick 96 genes we would expect 96 x 10.1% = 9.7 genes to be significant “by chance”
- A p-value can be calculated based on the 2x2 table
- Tests for association: Chi-Square-Test or Fisher’s exact test

	In GS	**Not in GS**
sig	8	1264
non-sig	88	11 279

P-value from Fisher’s exact test (one-sided): 0.733, i.e very far from being significant

Competitive Tests

- Competitive results depend highly on how many genes are on the array and previous filtering
  - On a small targeted array where all genes are changed, a competitive method might detect no differential Gene Sets at all

- Competitive tests can also be used with small sample sizes, even for n=1
  - BUT: The result gives no indication of whether it holds for a wider population of subjects, the p-value concerns a population of genes!

- Competitive tests typically give less significant results than self-contained (as seen with the example)

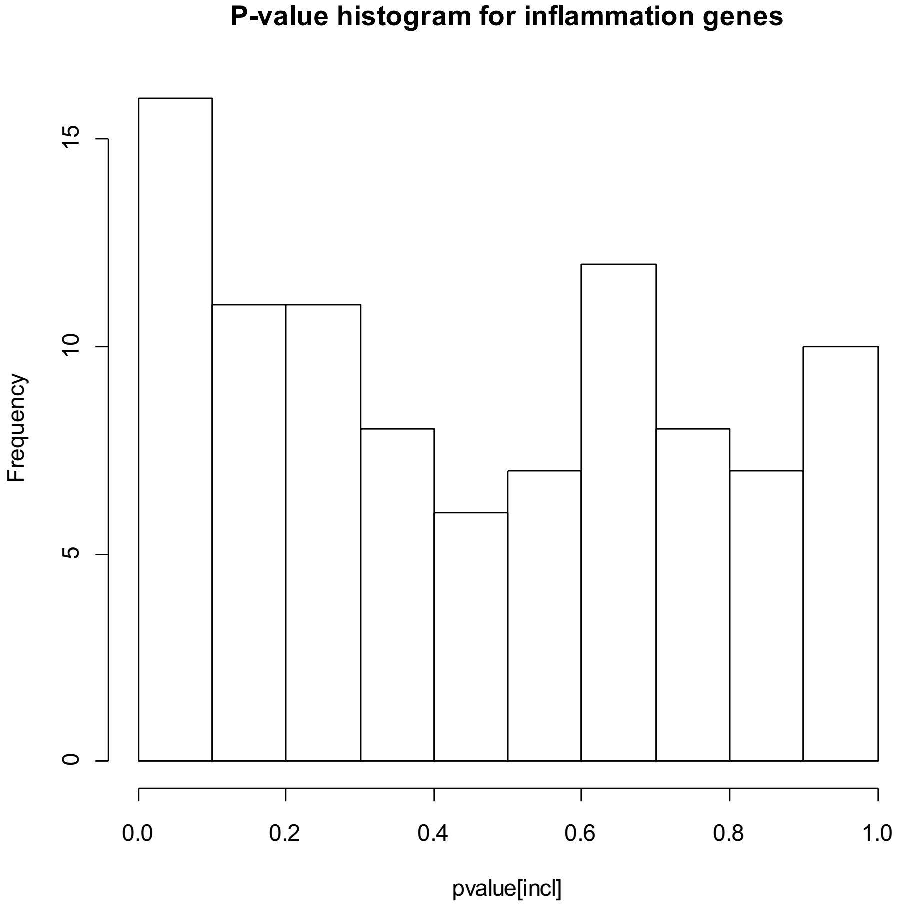
- Fisher’s exact test (competitive) is probably the most widely used method!

Cut-off methods vs whole gene list methods

- A problem with both tests discussed so far is, that they rely on an arbitrary cut-off

- If we call a gene significant for 10% alpha threshold the results will change
  - In our example the binomial test yields p= 0.022, i.e. for this cut-off the result is significant!

- We also lose information by reducing a p-value to a binary (“significant”, “non-significant”) variable
  - It should make a difference, whether the non-significant genes in the set are nearly significant or completely unsignificant



- We can study the distribution of the p-values in the gene set

- If no genes are differentially expressed this should be a uniform distribution

- A peak on the left indicates, that some genes are differentially expressed

- We can test this for example by using the Kolmogorov-Smirnov-Test

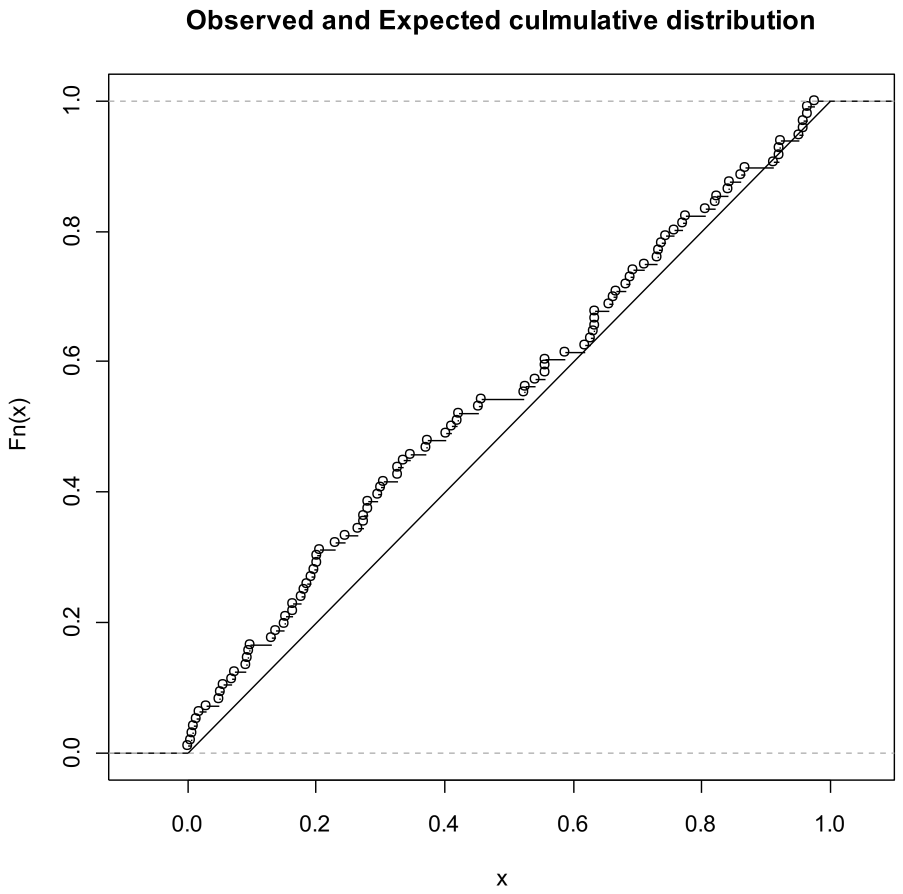
- Here p = 0.082, i.e. not quite significant

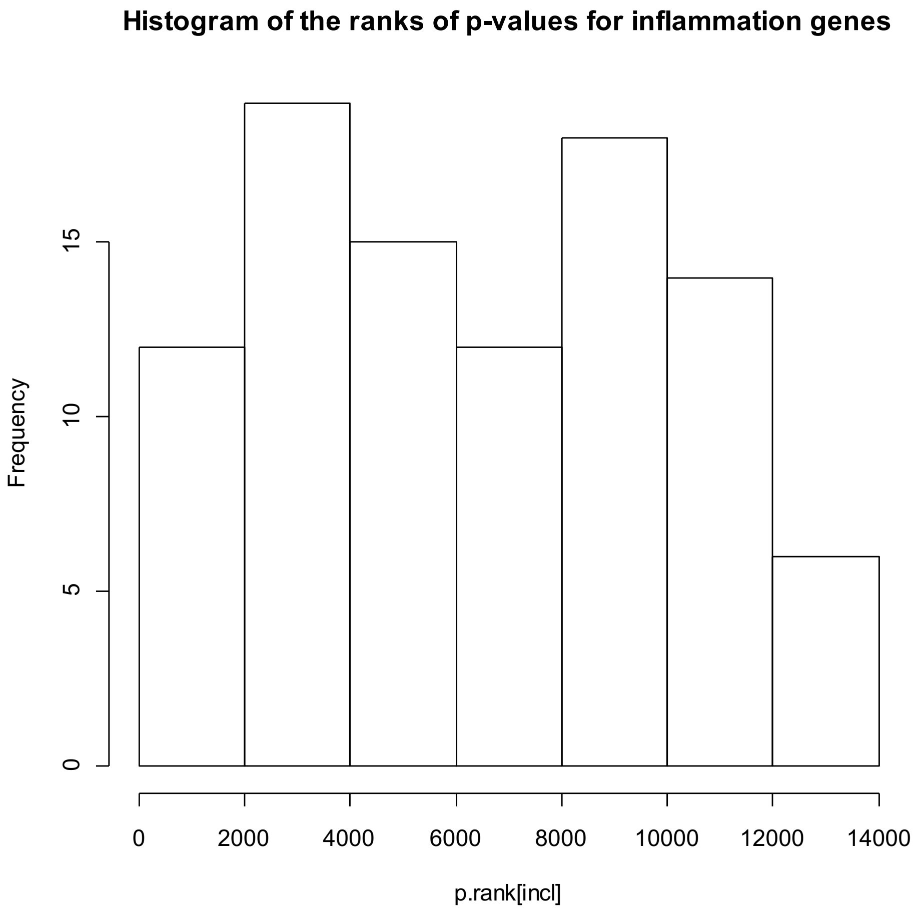
- This would be a “self-contained” test, as only the genes in the gene set are being used

Kolmogorov-Smirnov Test

- The KS-test compares an observed with an expected cumulative distribution

- The KS-statistic is given by the maximum deviation between the two





- Alternatively we could look at the distribution of the RANKS of the p-values in our gene set
- This would be a competitive method, i.e we compare our gene set with the other genes
- Again one can use the Kolmogorov-Smirnov test to test for uniformity
- Here: p= 0.851, i.e. very far from significance

Other general issues

- Direction of change
  - In our example we didn’t differentiate between up or down-regulated genes
  - That can be achieved by repeating the analysis for p-values from one-sided test
    - Eg. we could find GO-Terms that are significantly up-regulated
  - With most software both approaches are possible

- Multiple Testing
  - As we are testing many Gene Sets, we expect some significant findings “by chance” (false positives)
  - Controlling the false discovery rate is tricky: The gene sets do overlap, so they will not be independent!
    - Even more tricky in GO analysis where certain GO terms are subset of others
  - The Bonferroni-Method is most conservative, but always works!

Multiple Testing for Pathways

- Resampling strategies (dependence between genes)
  - The methods we used so far in our example assume that genes are independent of each other…if this is violated the p-values are incorrect
  - Resampling of group/phenotype labels can correct for this
  - We give an example for our data set

Example Resampling Approach

1. Calculate the test statistic, e.g. the percentage of significant genes in the Gene Set

2. Randomly re-shuffle the group labels (lean, obese) between the samples

3. Repeat the analysis for the re-shuffled data set and calculate a re-shuffled version of the test statistic

4. Repeat 2 and 3 many times (thousands…)

5. We obtain a distribution of re-shuffled % of significant genes: the percentage of re-shuffled values that are larger than the one observed in 1 is our p-value

Resampling Approach

- The reshuffling takes gene to gene correlations into account
- Many programs also offer to resample the genes: This does NOT take correlations into account
- Roughly speaking:
  - Resampling phenotypes: corresponds to self-contained test
  - Resampling genes: corresponds to competitive test

Resampling Approaches

- Genes being present more than once
  - Common approaches
    - Combine duplicates (average, median, maximum,…)
    - Ignore (i.e treat duplicates like different genes)

- Using summary statistics vs using all data
  - Our examples used p-values as data summaries
  - Other approaches use fold-changes, signal to noise ratios, etc…
  - Some methods are based on the original data for the genes in the gene set rather than on a summary statistic

Resampling Approaches

- The resampling approaches are highly computationally intensive

- New methods are being developed to speed this up
  - Empirical approximations of permutations
  - Empirical pathway analysis, without permutation.
    - Zhou YH, Barry WT, Wright FA.Biostatistics. 2013 Jul;14(3):573-85. doi: 10.1093/biostatistics/kxt004. Epub 2013 Feb 20.

Summary

- Databases
- Choice makes a difference
- Not all use the same IDs – watch out ☺
- Major differences between methods
- Issues with multiple testing

- Next lecture, will go into more detail on a few methods

Questions?