Spring 2011
BMD6621 – High-Throughput Sequencing Analysis

Data Integration

Shu-Jen Chen, Ph.D.

Department of Biomedical Sciences
Chang Gung University

Jun. 3, 2011 (Friday 8:30 – 12:00)

SJChen/CGU/2011/

To fully utilize the results of contemporary biological research, one would like to analyze data on biological function in addition to sequence information.

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Unfortunately …

- **Compared to sequence information, biological function is much more difficult to analyze.**

- **Biological data is fragmented**
  - **Biologists currently waste a lot of time and effort in searching for all of the available information about each small area of research.**

- **Language used in biological research is not well controlled**
  - **This is hampered further by the wide variations in terminology that may be common usage at any given time, which inhibit effective searching by both computers and people.**

Adopted from http://www.geneontology.org/

A simple example

- If you were searching for new targets for antibiotics, you might want to find
  - all the gene products that are involved in bacterial protein synthesis, and
  - that have significantly different sequences or structures from those in humans.

- If one database describes these molecules as being involved in 'translation' while another uses the phrase 'protein synthesis', it will be difficult for you - and even harder for a computer - to find functionally equivalent terms.

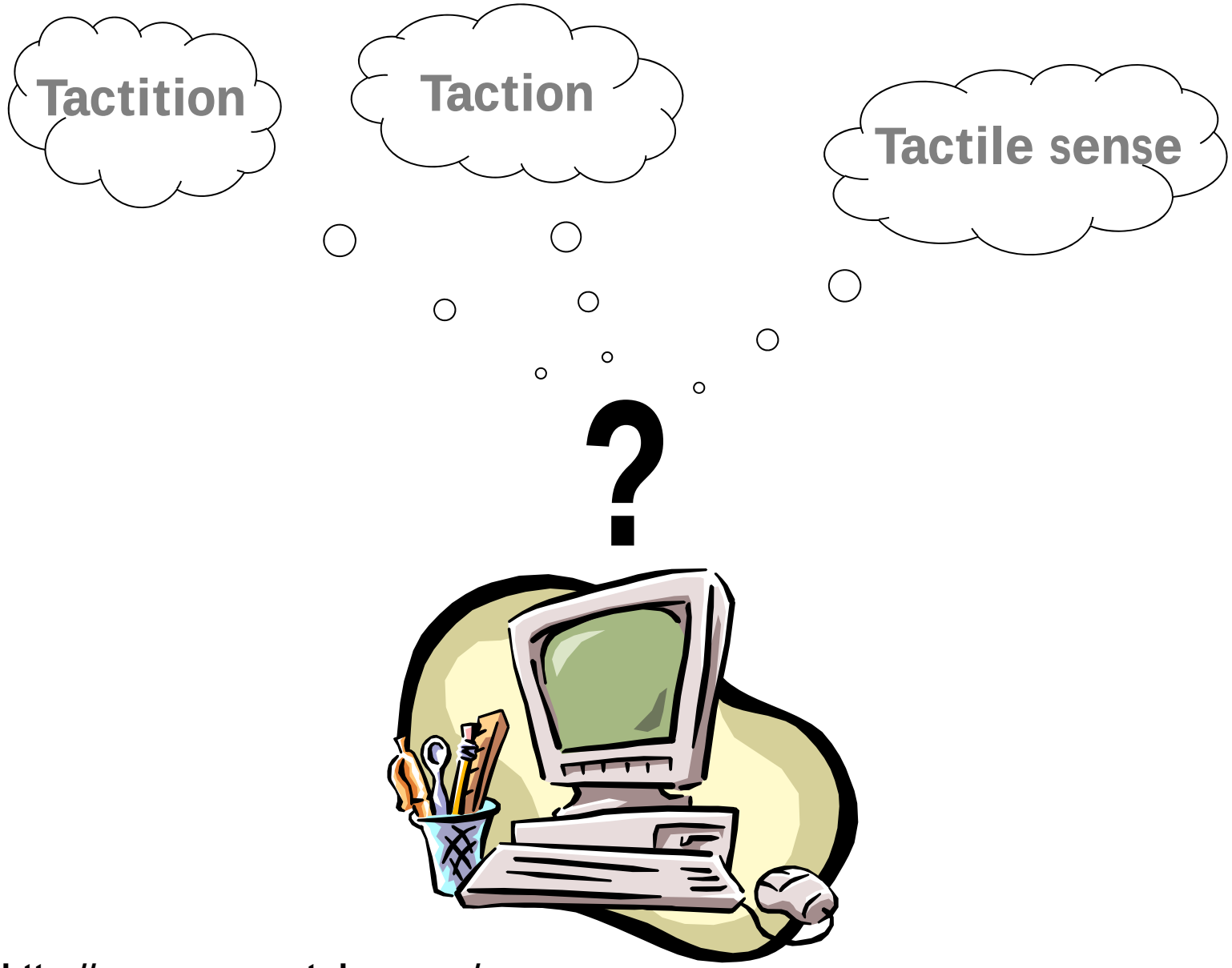


Inconsistent descriptions of biological function makes systemic functional analysis virtually impossible

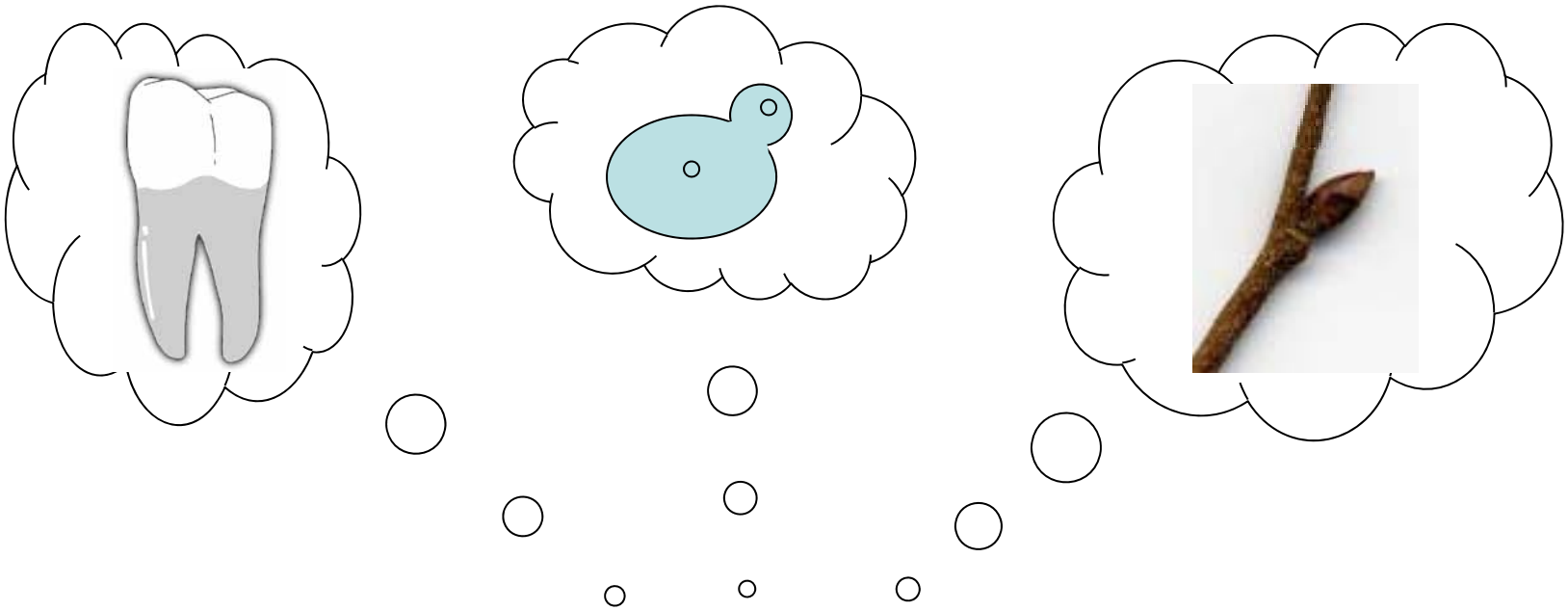
Adopted from http://www.geneontology.org/

SJChen/CGU/2011/

In biology...



Adopted from http://www.geneontology.org/



Bud initiation?



Adopted from http://www.geneontology.org/

The Gene Ontology

http://www.geneontology.org

The Gene Ontology (GO) provides a way to capture and represent biological data and make all this knowledge in a computable form





SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

The Gene Ontology is like a dictionary



Each concept (term) has:

- a name
- a definition
- an ID number

Term: transcription initiation

Definition: Processes involved in the assembly of the RNA polymerase complex at the promoter region of a DNA template resulting in the subsequent synthesis of RNA from that promoter.

ID: GO:0006352

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Tactition

Taction

Tactile sense

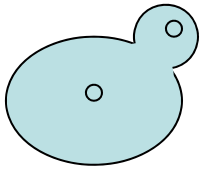
perception of touch ; GO:0050975



Adopted from http://www.geneontology.org/



= tooth bud initiation



= cellular bud initiation



= flower bud initiation

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

What is the Gene Ontology project?

- The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases.

- The project began as a collaboration between three model organism databases, FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Database (MGD), in 1998.

- Since then, the GO Consortium has grown to include many databases, including several of the world's major repositories for plant, animal and microbial genomes.

How does GO work?

What information might we want to capture about a gene product?

- What does the gene product do?
- Where and when does it act?
- Why does it perform these activities?

- GO uses “GO term” to represent these concepts
- Each gene is associated (annotated) with multiple “GO terms” to describe its location and functions
- The information is stored in the GO database

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

The GO project (I)

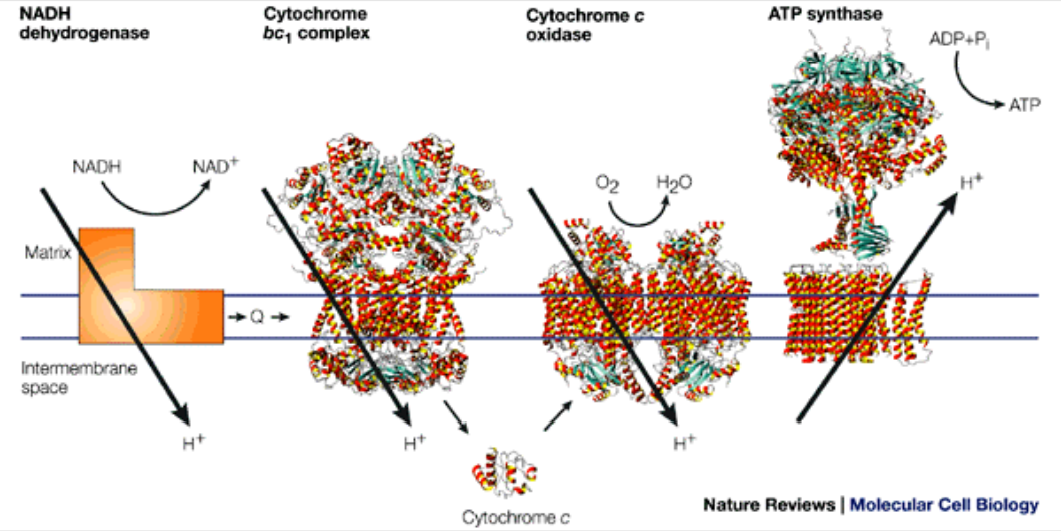
- The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.
- There are three separate aspects to this effort:
  - development and maintenance of the ontologies
  - annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases
  - development of tools that facilitate the creation, maintenance and use of ontologies.
- The use of GO terms by collaborating databases facilitates uniform queries across them.

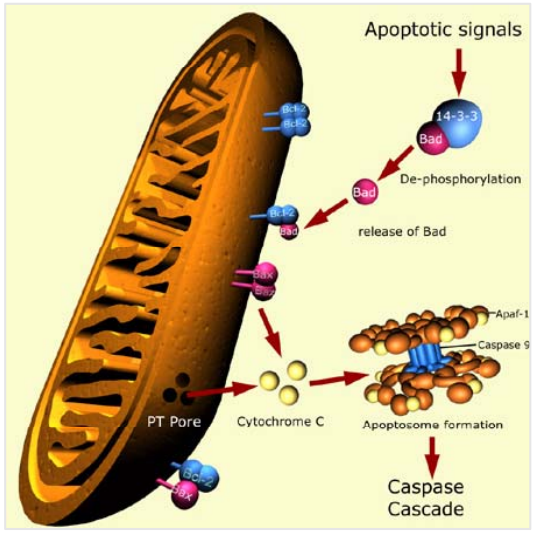
The Gene Ontology

- **The Gene Ontology project provides an ontology of defined terms representing gene product properties.**
- **The ontology covers three domains pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.**
  - **cellular component:**
    the parts of a cell or its extracellular environment
  - **molecular function:**
    the elemental activities of a gene product at the molecular level, such as binding or catalysis
  - **biological process:**
    operations or sets of molecular events with a defined beginning and end

Example: GO terms for cytochrome c

- **The gene product "cytochrome c" can be described by the following GO terms:**
  - **molecular function:**
    oxidoreductase activity
  - **biological process:**
    oxidative phosphorylation and induction of cell death
  - **cellular component:**
    mitochondrial matrix and mitochondrial inner membrane





SJChen/CGU/2011/

The GO project (II)

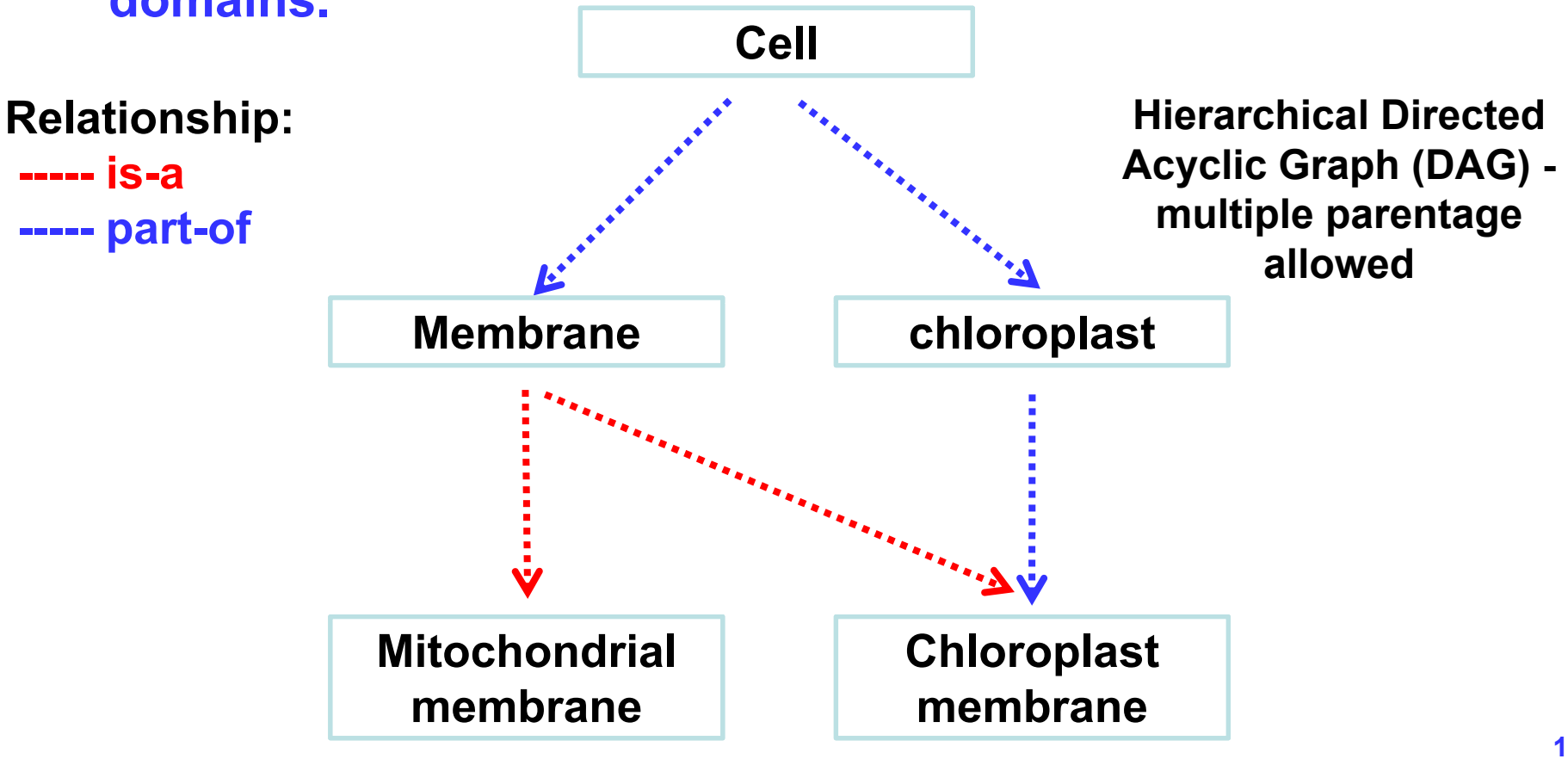
- The controlled vocabularies are structured so that they can be queried at different levels.
- For example, you can use GO to find all the gene products in the mouse genome that are involved in signal transduction, or you can zoom in on all the receptor tyrosine kinases.
- This structure also allows annotators to assign properties to genes or gene products at different levels, depending on the depth of knowledge about that entity.

GO Structure

GO isn’t just a flat list of biological terms.
Terms are related within a hierarchy.

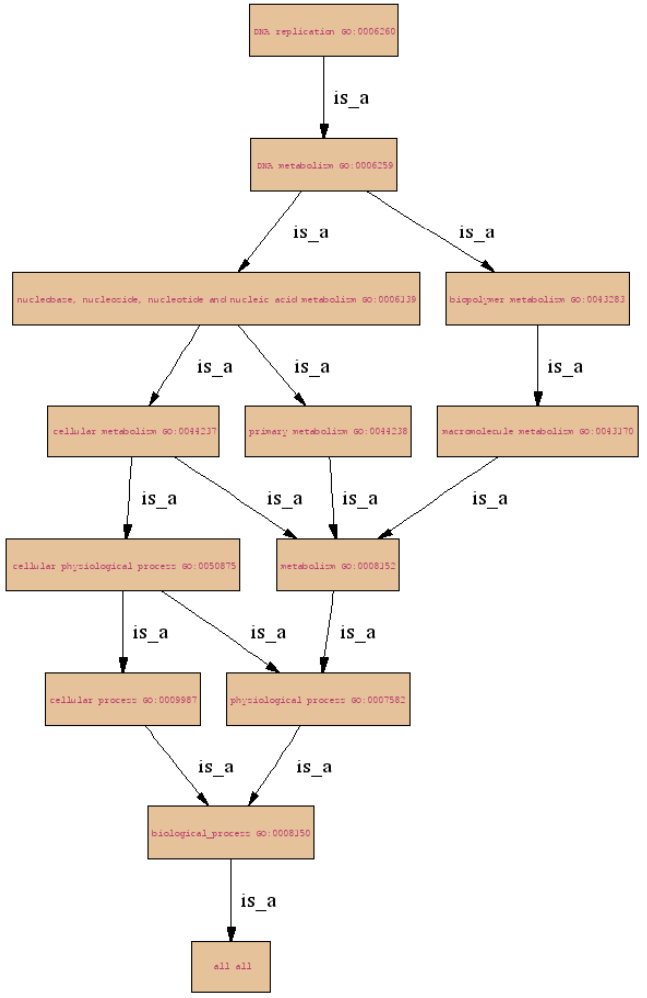
Structure of GO Terms

- The GO ontology is structured as a directed acyclic graph (DAC).
- Each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains.



GO structure

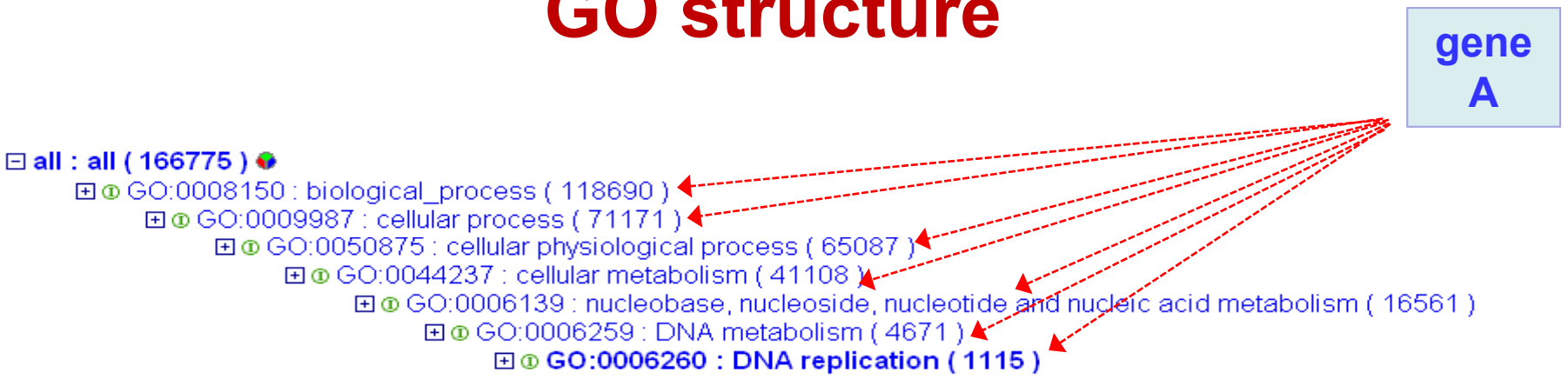
- all : all (166775)
  - GO:0008150 : biological_process (118690)
    - GO:0009987 : cellular process (71171)
      - GO:0050875 : cellular physiological process (65087)
        - GO:0044237 : cellular metabolism (41108)
          - GO:0006139 : nucleobase, nucleoside, nucleo
            - GO:0006259 : DNA metabolism (4671)
              - **GO:0006260 : DNA replication (**
    - GO:0007582 : physiological process (73658)
      - GO:0050875 : cellular physiological process (65087)
        - GO:0044237 : cellular metabolism (41108)
          - GO:0006139 : nucleobase, nucleoside, nucleo
            - GO:0006259 : DNA metabolism (4671)
              - **GO:0006260 : DNA replication (**
      - GO:0008152 : metabolism (44953)
        - GO:0044237 : cellular metabolism (41108)
          - GO:0006139 : nucleobase, nucleoside, nucleo
            - GO:0006259 : DNA metabolism (4671)
              - **GO:0006260 : DNA replication (**
        - GO:0043170 : macromolecule metabolism (23499)
          - GO:0043283 : biopolymer metabolism (1352
            - GO:0006259 : DNA metabolism (4671)
              - **GO:0006260 : DNA replication (**
        - GO:0044238 : primary metabolism (36601)
          - GO:0006139 : nucleobase, nucleoside, nucleo
            - GO:0006259 : DNA metabolism (4671)
              - **GO:0006260 : DNA replication (1115)**



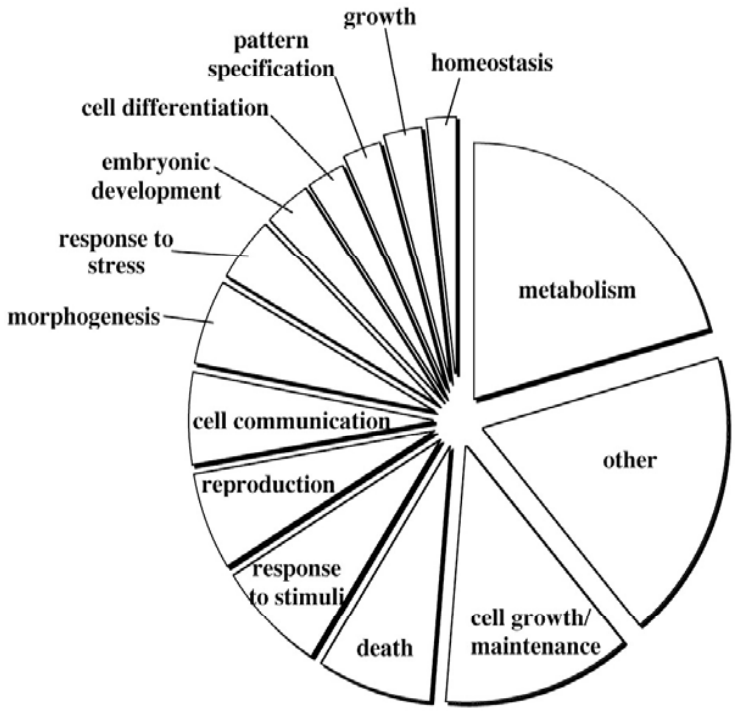
SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

GO structure



- **This means genes can be grouped according to user-defined levels**
- **Allows broad overview of gene set or genome**



SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

GO namespace

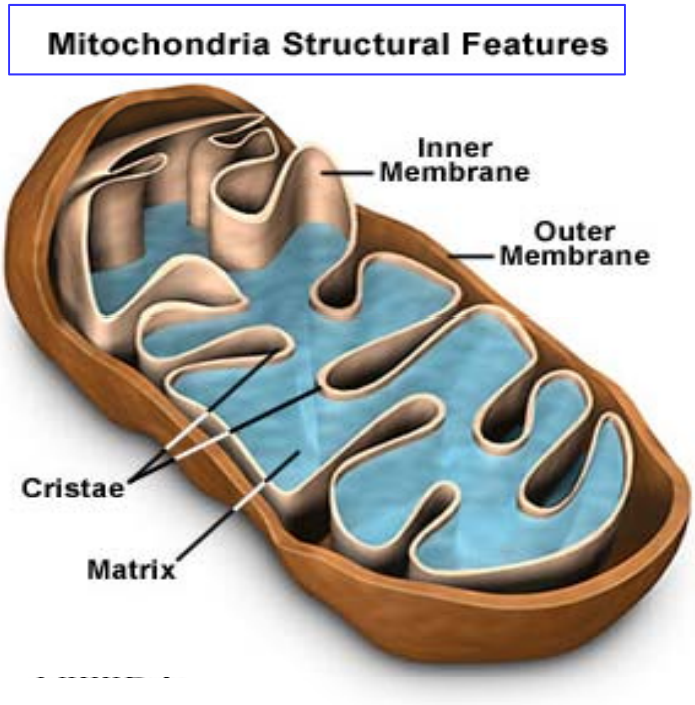
- **GO terms are divided into three types:**
  - **Cellular component : where and when does it act?**
  - **Molecular function : what does the gene product do?**
  - **Biological process : why does it perform these activities?**

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Cellular Component

- where a gene product acts

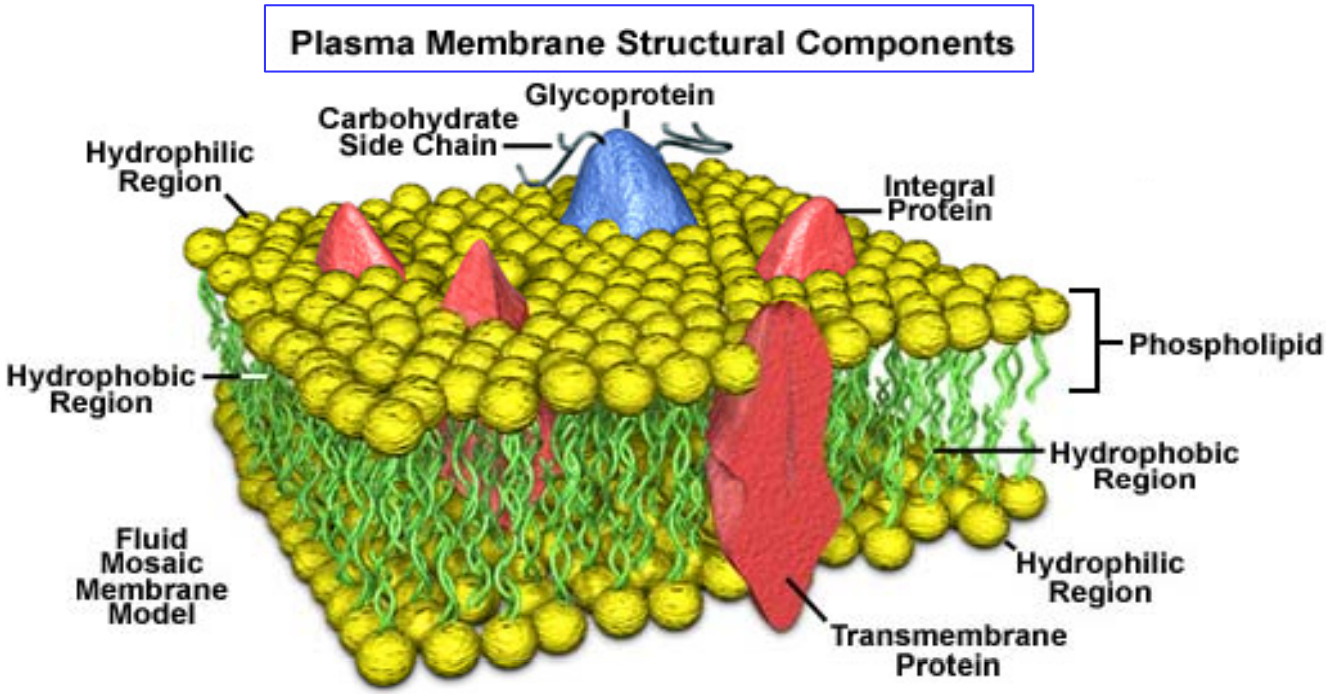


Adopted from http://www.geneontology.org/

SJChen/CGU/2011/

Cellular Component

- where a gene product acts

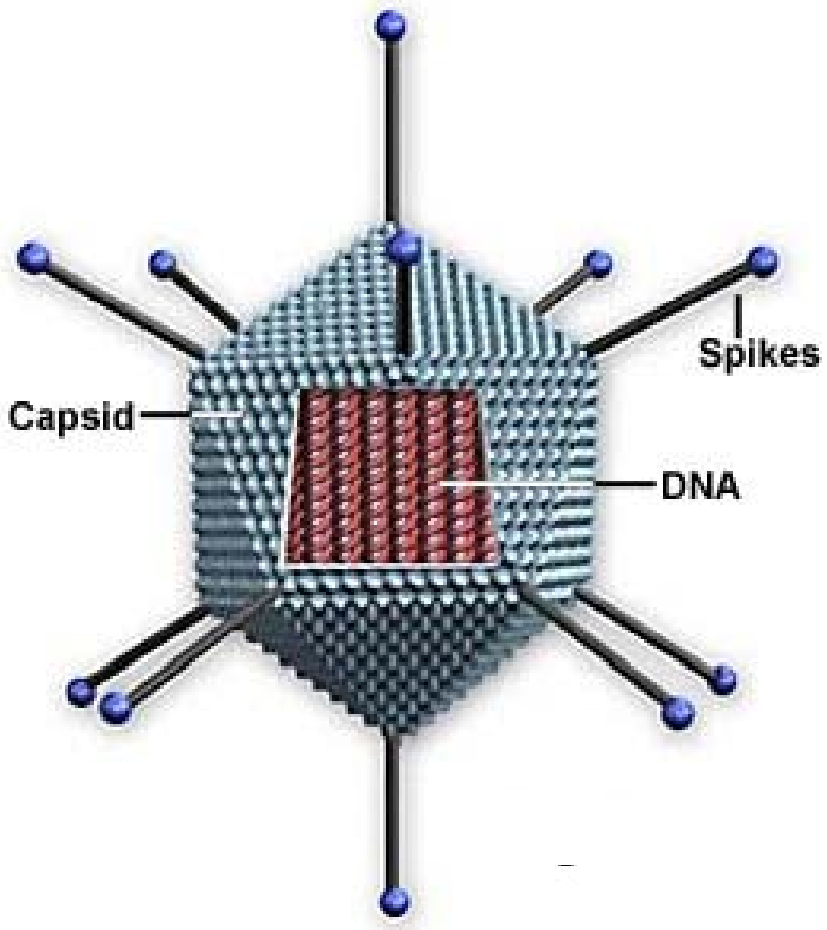


SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Cellular Component

- where a gene product acts

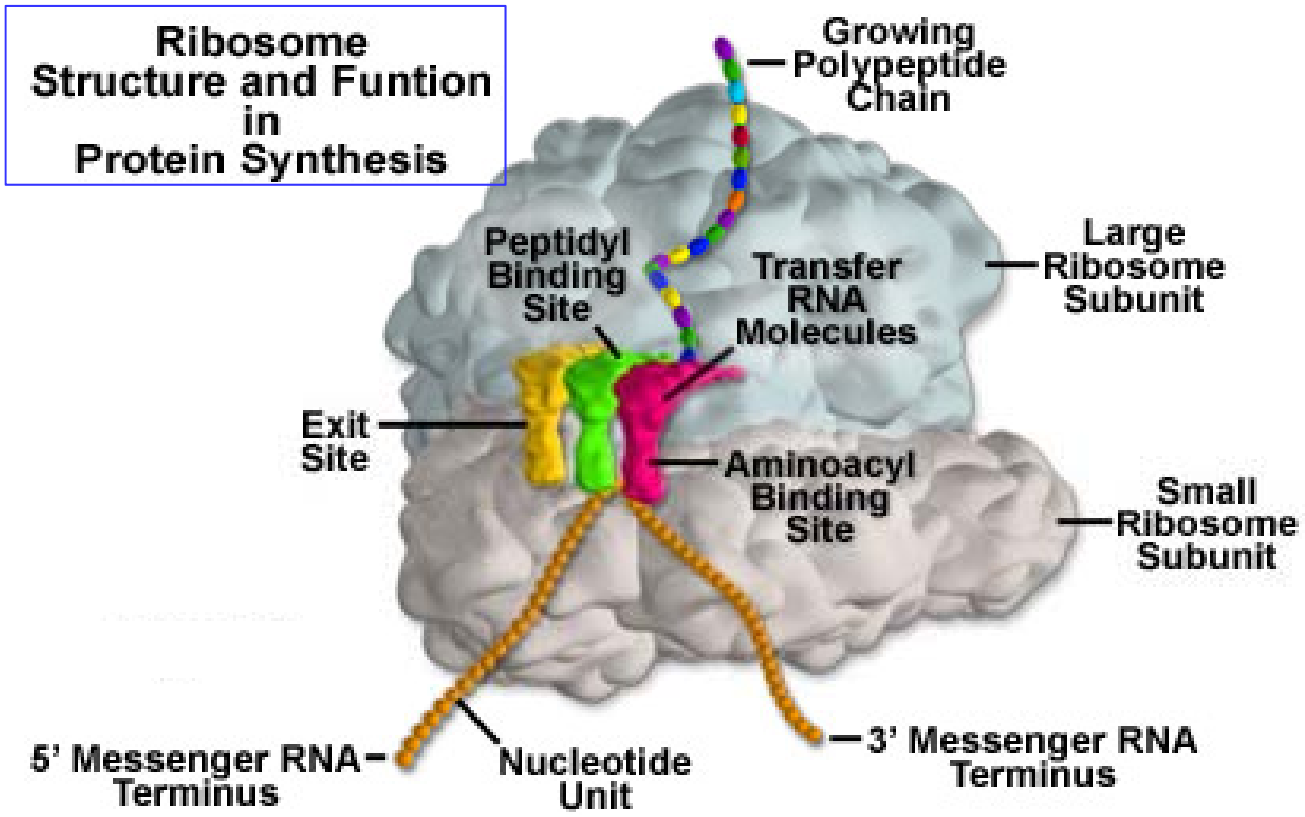


Adopted from http://www.geneontology.org/

SJChen/CGU/2011/

Cellular Component

- where a gene product acts



- Enzyme complexes in the component ontology refer to places, not activities.

Adopted from http://www.geneontology.org/

SJChen/CGU/2011/

Molecular Function & Biological Process

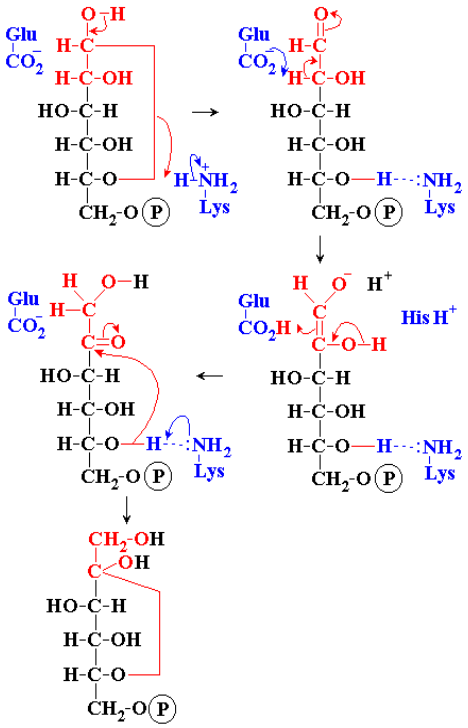
- A gene product may have several functions.
- A function term refers to a reaction or activity, not a gene product
  → How ?
- Sets of functions make up a biological process
  → Why ?

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Molecular Function

- activities or “jobs” of a gene product



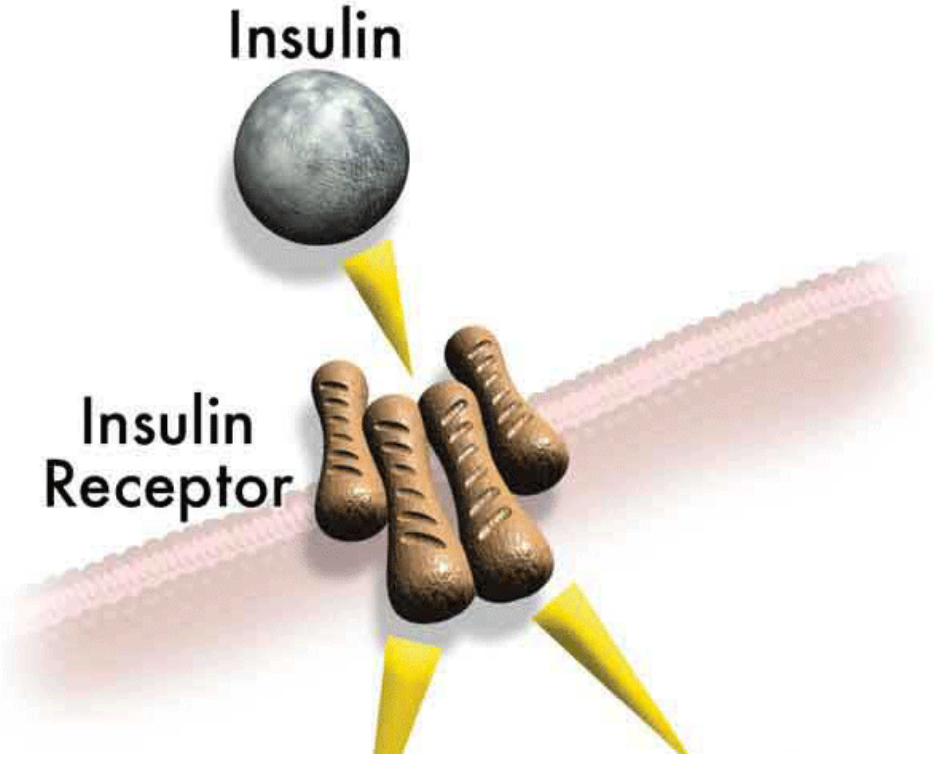
glucose-6-phosphate isomerase activity

Adopted from http://www.geneontology.org/

SJChen/CGU/2011/

Molecular Function

- activities or “jobs” of a gene product

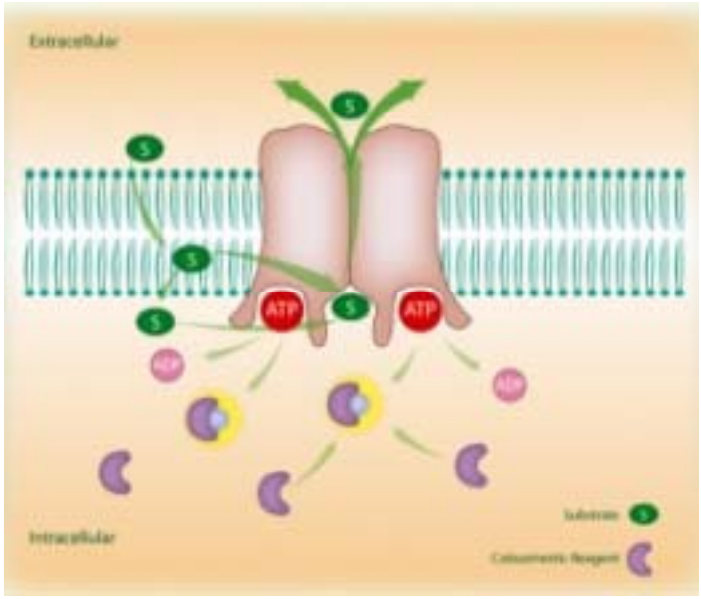


insulin binding
insulin receptor activity

Adopted from http://www.geneontology.org/

Molecular Function

- activities or “jobs” of a gene product



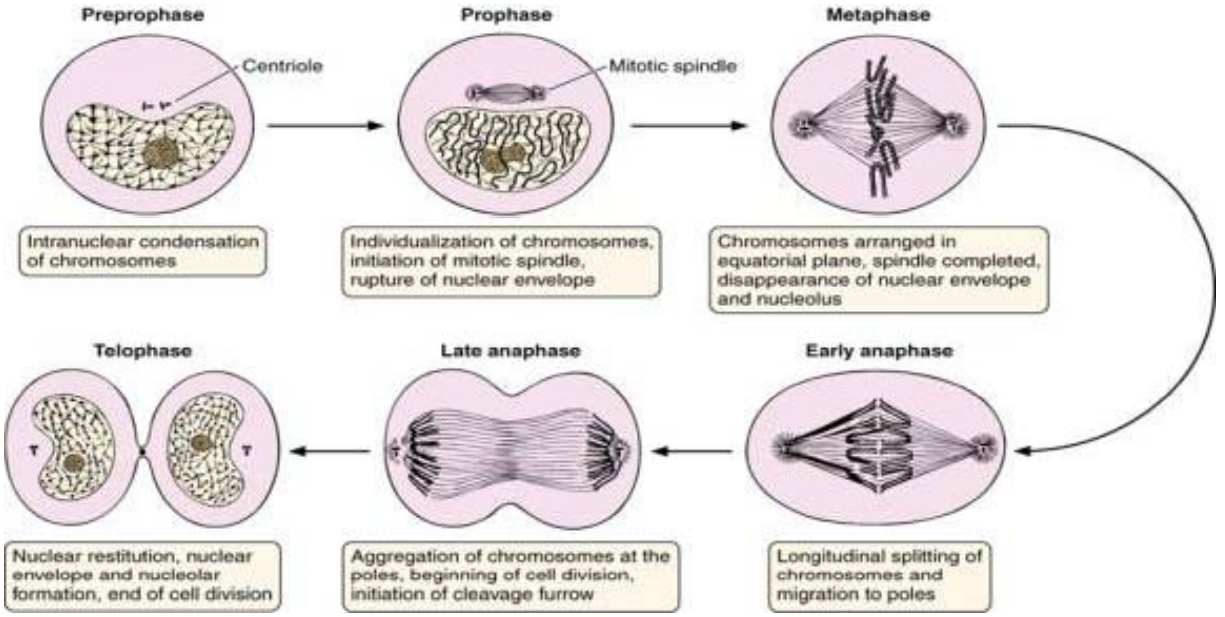
drug transporter activity

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Biological Process

- a commonly recognized series of events



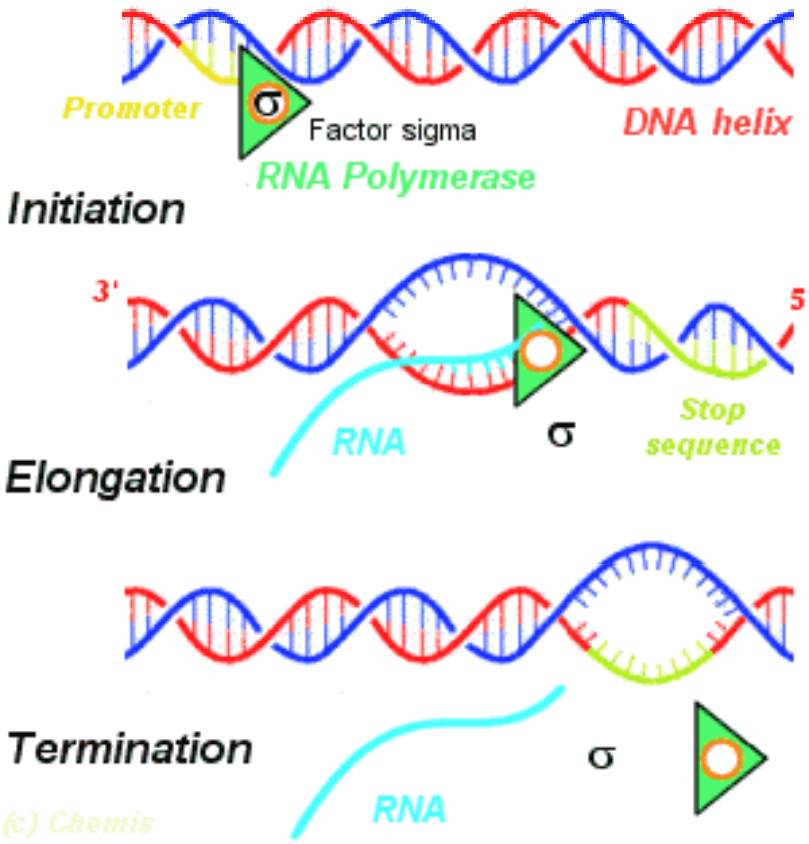
cell division

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Biological Process

- a commonly recognized series of events



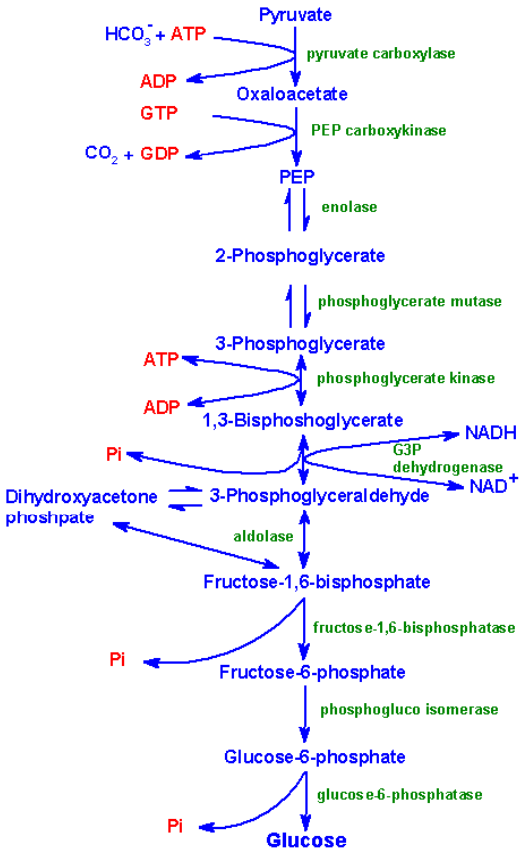
transcription

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Biological Process

- a commonly recognized series of events



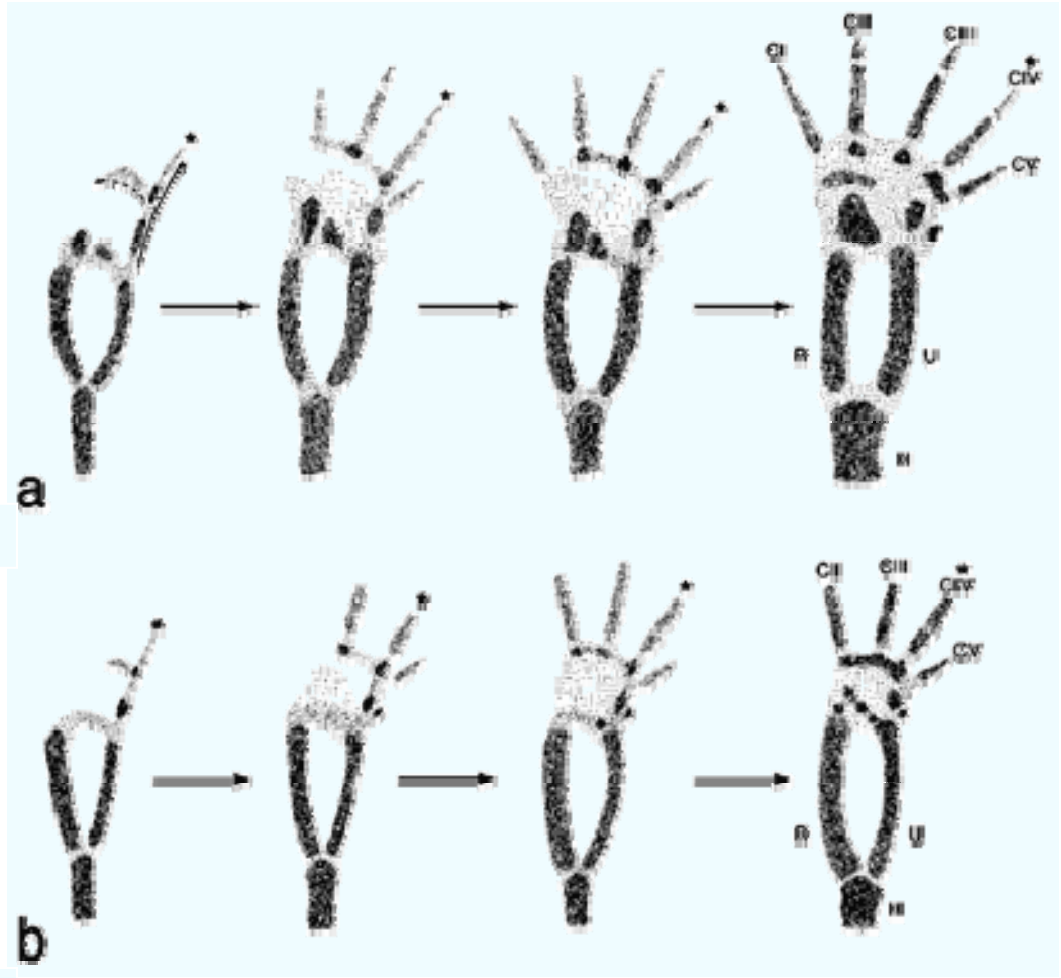
regulation of gluconeogenesis

Adopted from http://www.geneontology.org/

SJChen/CGU/2011/

Biological Process

- a commonly recognized series of events



limb development

Adopted from http://www.geneontology.org/

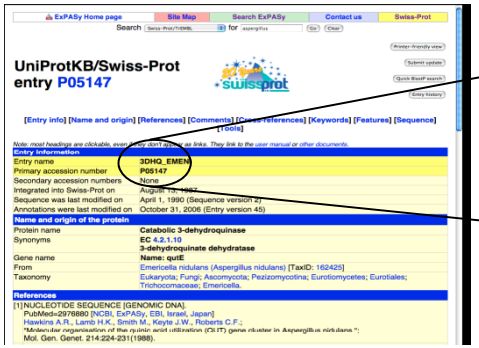
SJChen/CGU/2011/

Categorization of gene products using GO is called annotation.

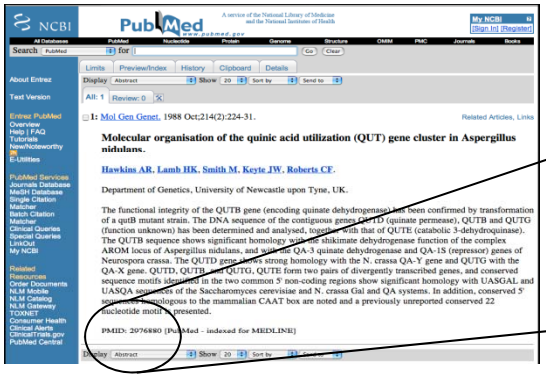
So how does that happen?

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/



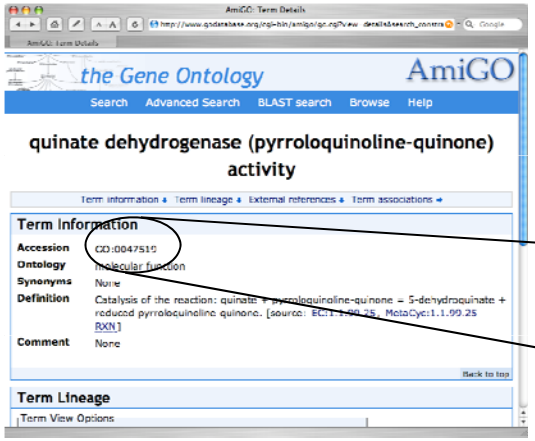
P05147



What evidence do they show?

IDA

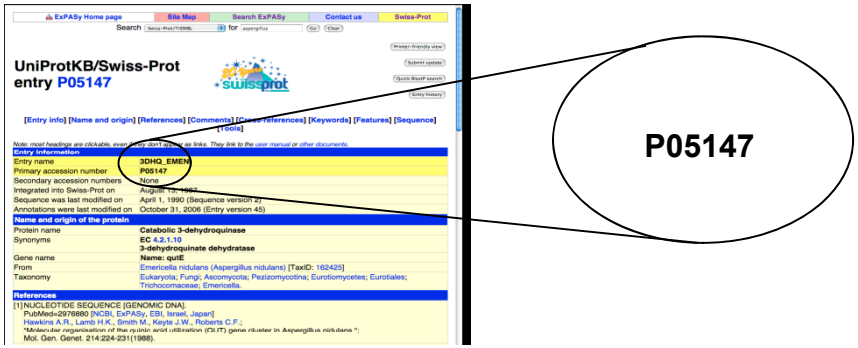
Code	Definition
IEA	Inferred from **E**lectronic **A**nnotation
IDA	Inferred from **D**irect **A**ssay
IEP	Inferred from **E**xpression **P**attern
IGI	Inferred from **G**enetic **I**nteraction
IMP	Inferred from **M**utant **P**henotype
IPI	Inferred from **P**hysical **I**nteraction
ISS	Inferred from **S**equence **S**imilarity
TAS	**T**raceable **A**uthor **S**tatement
NAS	**N**on-traceable **A**uthor **S**tatement
RCA	**R**eviewed **C**omputational **A**nalysis
IC	Inferred from **C**urator
ND	**N**o **D**ata

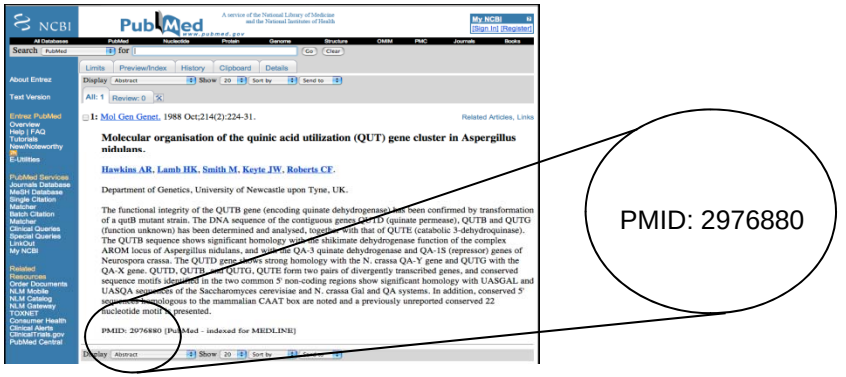


GO:0047519

Adopted from http://www.geneontology.org/

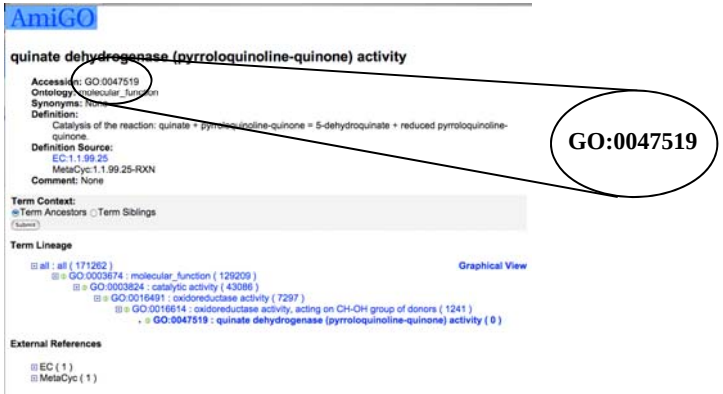
SJChen/CGU/2011/





Record these:

P05147	**GO:0047519**	**IDA**	**PMID:2976880**



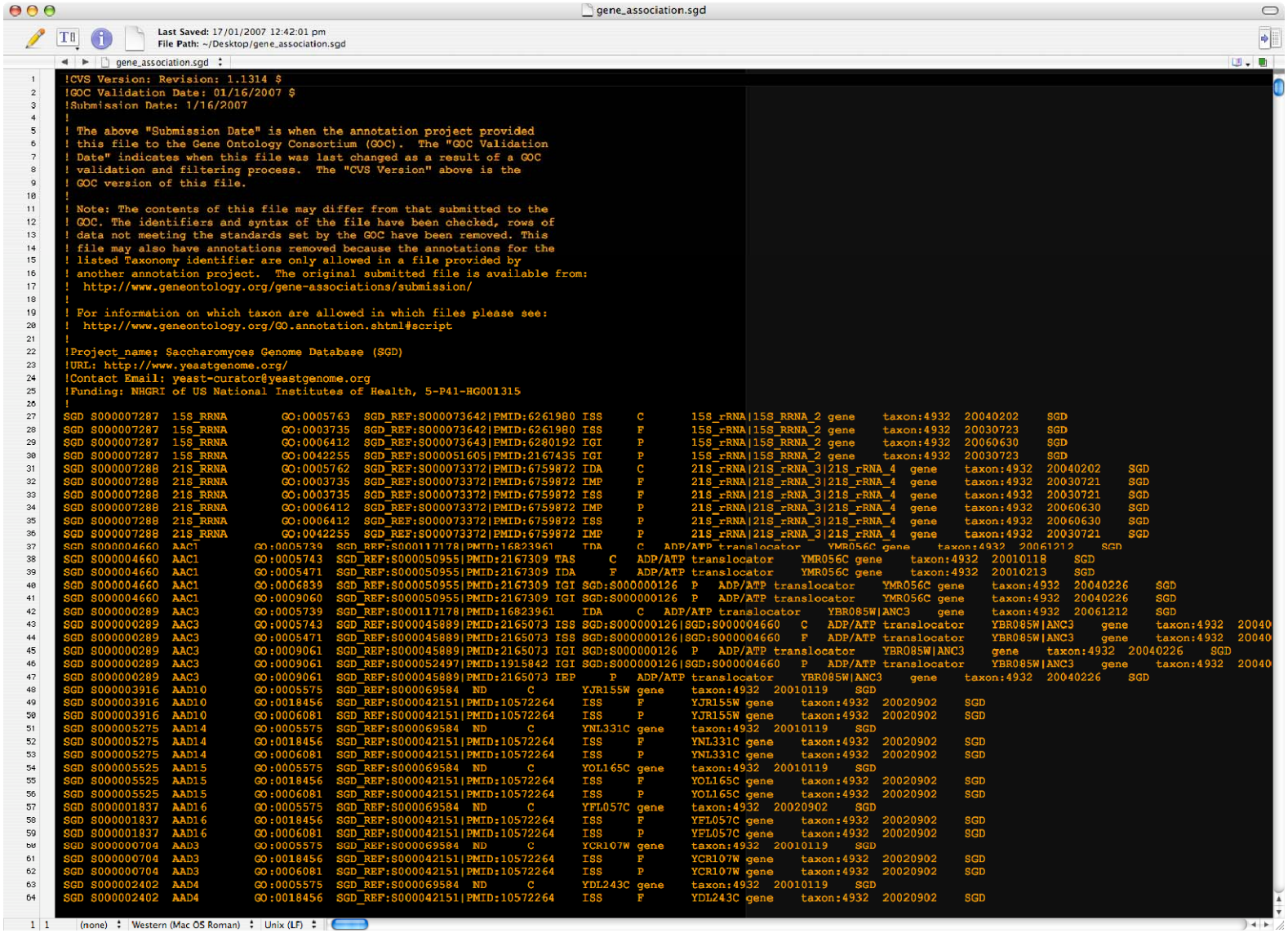
IDA

Code	Definition
IEA	Inferred from Electronic Annotation
IDA	Inferred from Direct Assay
IEP	Inferred from Expression Pattern
IGI	Inferred from Genetic Interaction
IMP	Inferred from Mutant Phenotype
IPI	Inferred from Physical Interaction
ISS	Inferred from Sequence Similarity
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
RCA	Reviewed Computational Analysis
IC	Inferred from Curator
ND	No Data

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

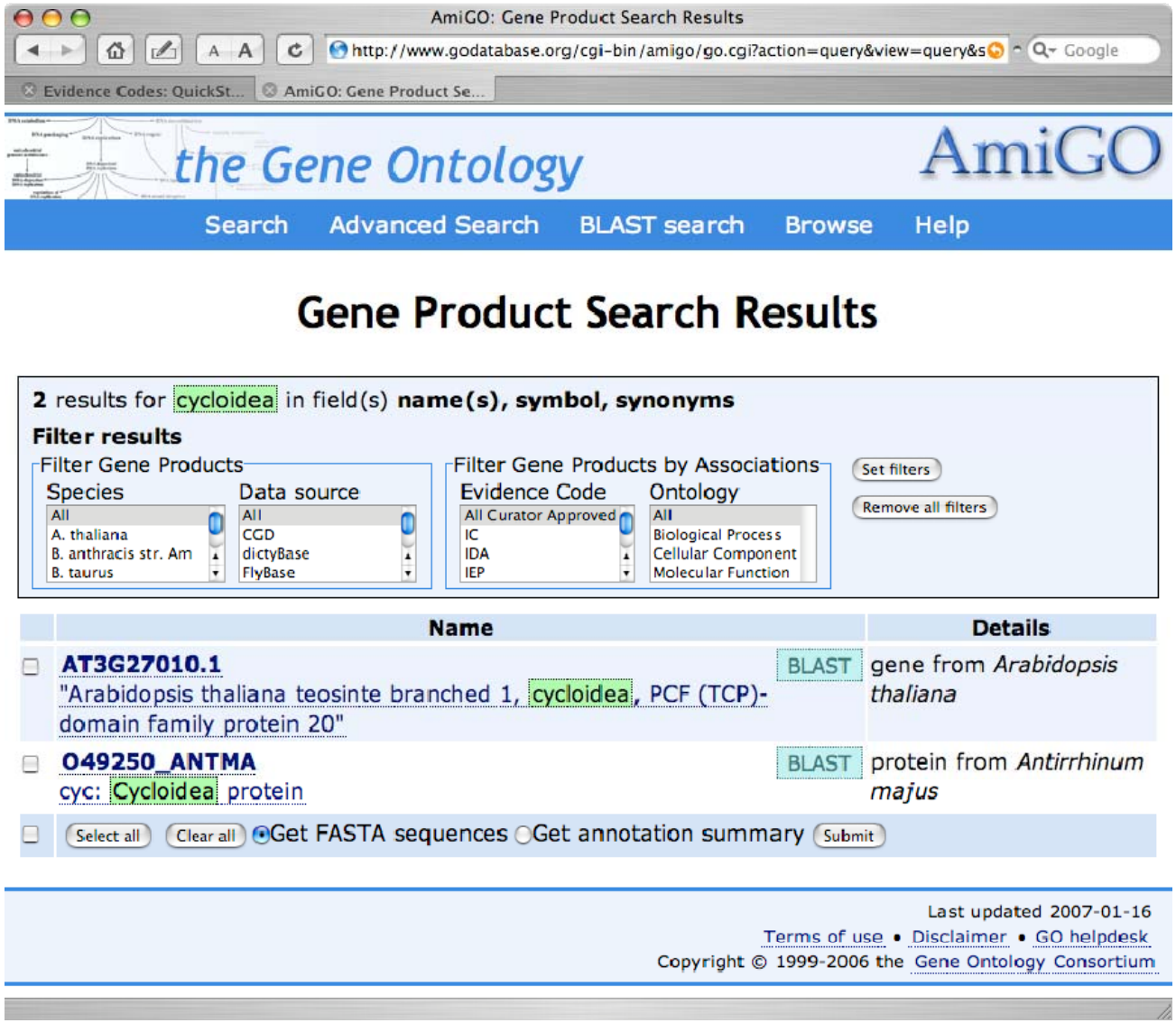
Submit to the GO Consortium



SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

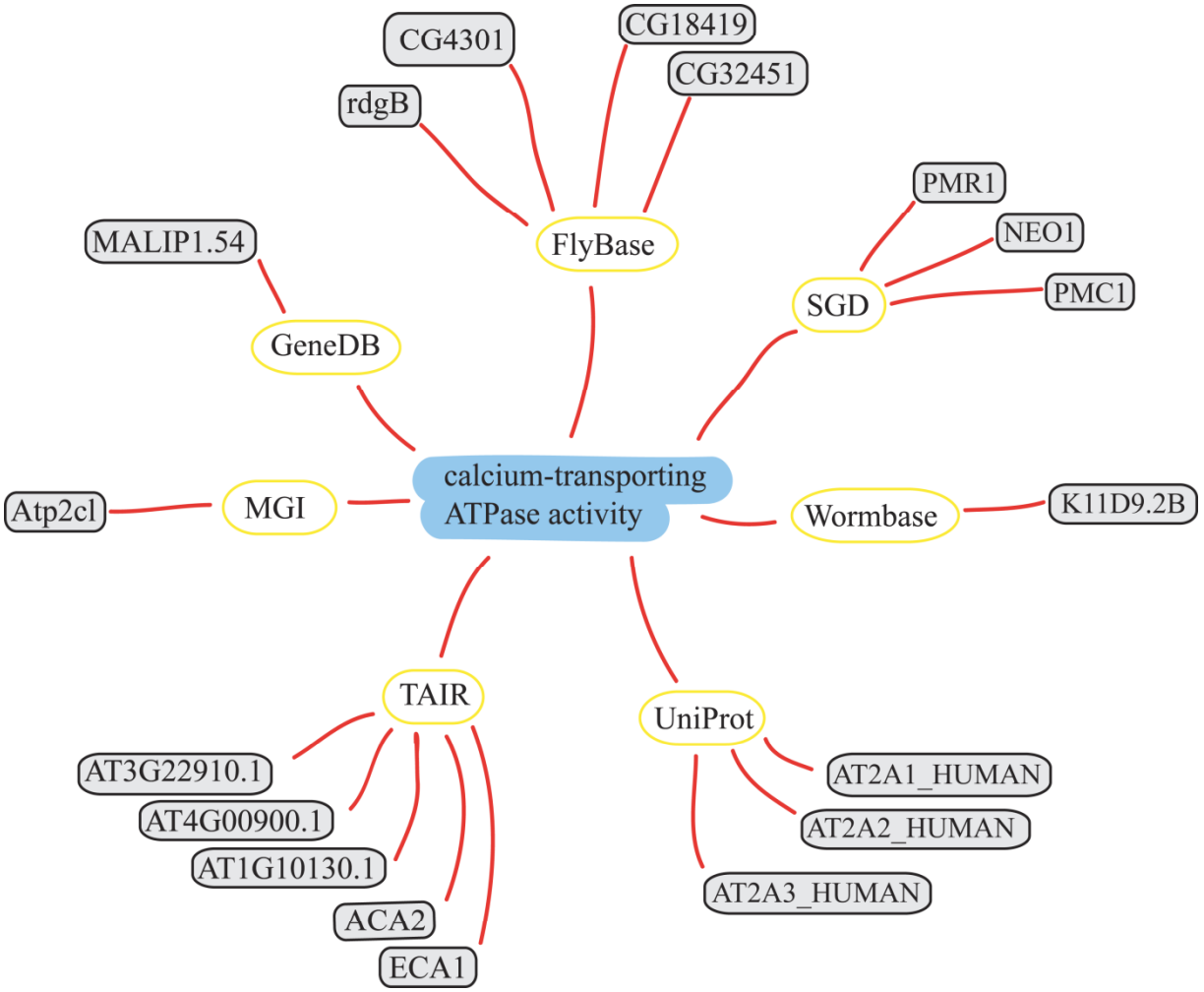
Annotation appears in GO database



Adopted from http://www.geneontology.org/

SJChen/CGU/2011/

Many species groups annotate



We see the research of one function across all species

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Scope of GO Terms

- **The GO vocabulary is designed to be species-neutral, and includes terms applicable to prokaryotes and eukaryotes, single and multicellular organisms.**

Species, Database	Gene Products Annotated	Annotatic
*Agrobacterium tumefaciens*str. C58 PAMGO	82	248 (248 non-
Arabidopsis thaliana TAIR	31766	133010 (102197 n
Aspergillus nidulans AspGD	21035	73646 (56612 no
Comprehensive Microbial Resource [multispecies] JCVI	61752	155509 (155509 n
Bos taurus GO Annotations @ EBI	25495	118905 (8535 non
Caenorhabditis elegans WormBase	17477	107900 (58271 no
Candida albicans CGD	6334	30347 (17624 no
Gallus gallus GO Annotations @ EBI	16235	75366 (4139 non-IEA)	4/8/2011
Homo sapiens GO Annotations @ EBI	18255	231096 (135905 non-IEA)	4/8/2011
Leishmania major Sanger GeneDB	10	28 (28 non-IEA)	5/31/2010
Magnaporthe grisea PAMGO	11274	27631 (27631 non-IEA)	3/18/2011
Mus musculus MGI	33917	285214 (195161 non-IEA)	4/8/2011
Oomycetes PAMGO	30	126 (126 non-IEA)	5/31/2010
Oryza sativa Gramene	41142	49638 (49638 non-IEA)	3/18/2011
Protein Data Bank [multispecies] GO Annotations @ EBI	93982	879623 (31390 non-IEA)	4/8/2011

SJChen/CGU/2011/

Example 1

Using GO to identify all genes involved in a specific biological process.

There is a lot of biological research output

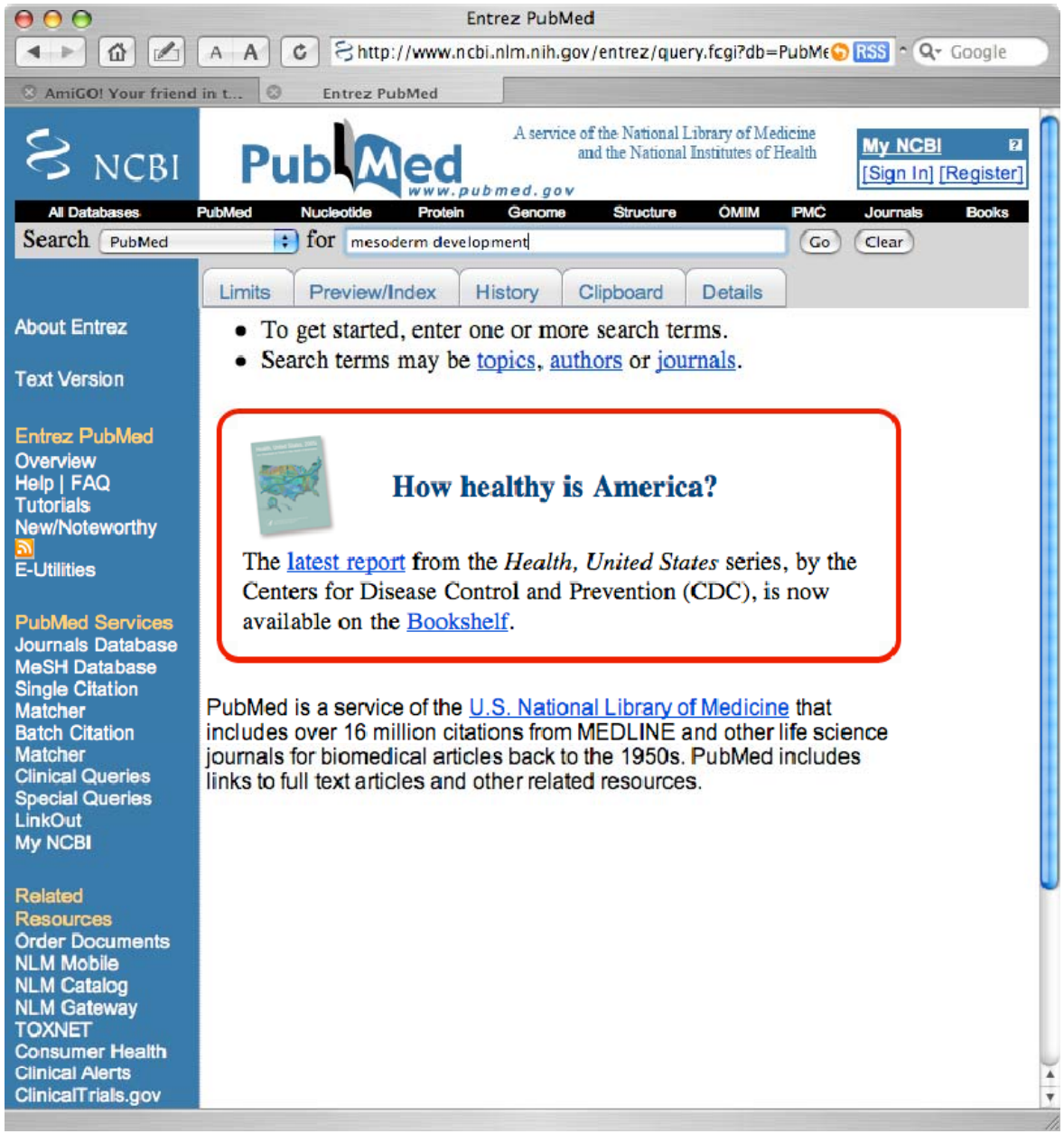


SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

You're interested in which genes control **mesoderm development**...

You conduct a term search in PubMed

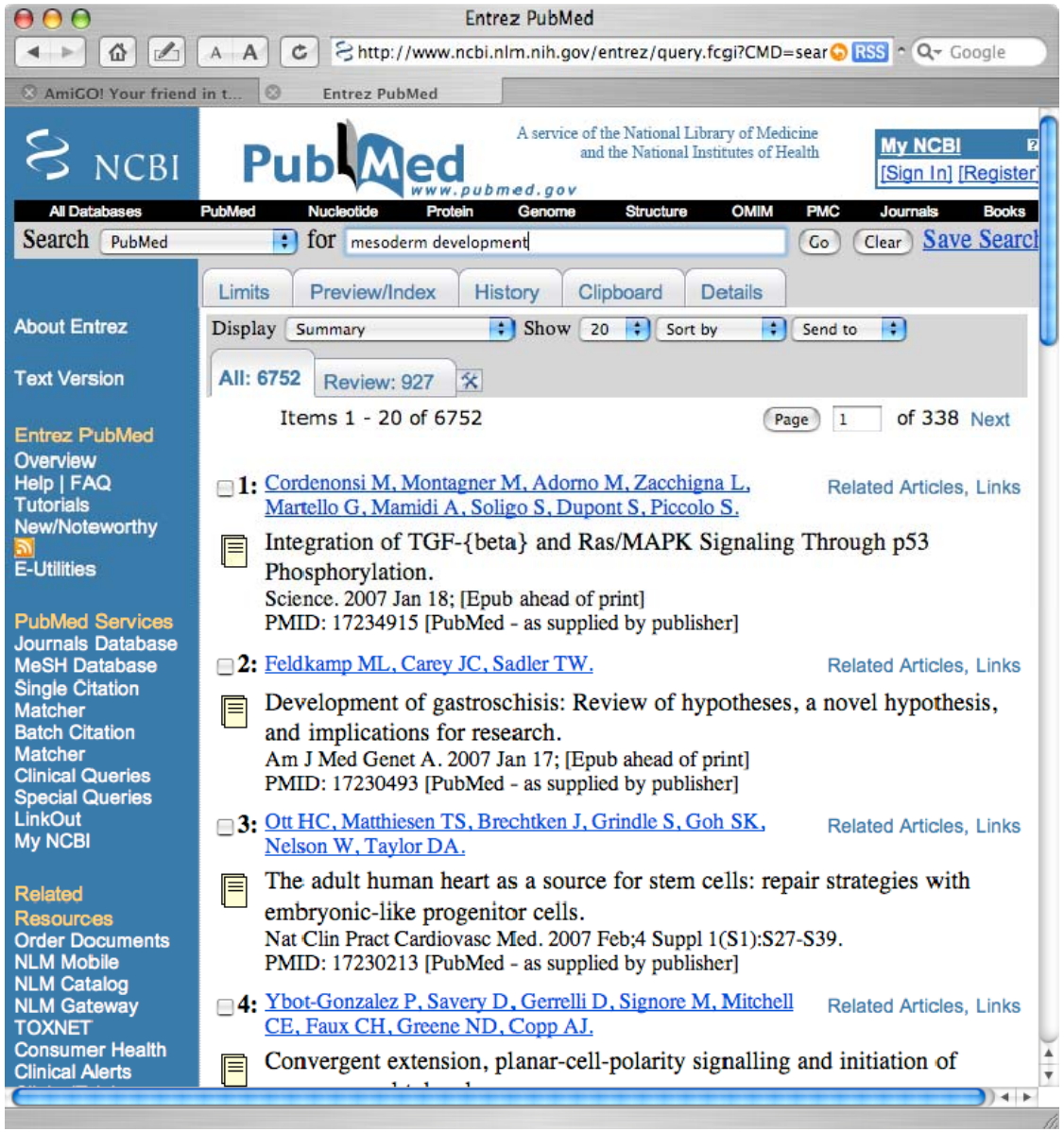


SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

You get 6752 results!

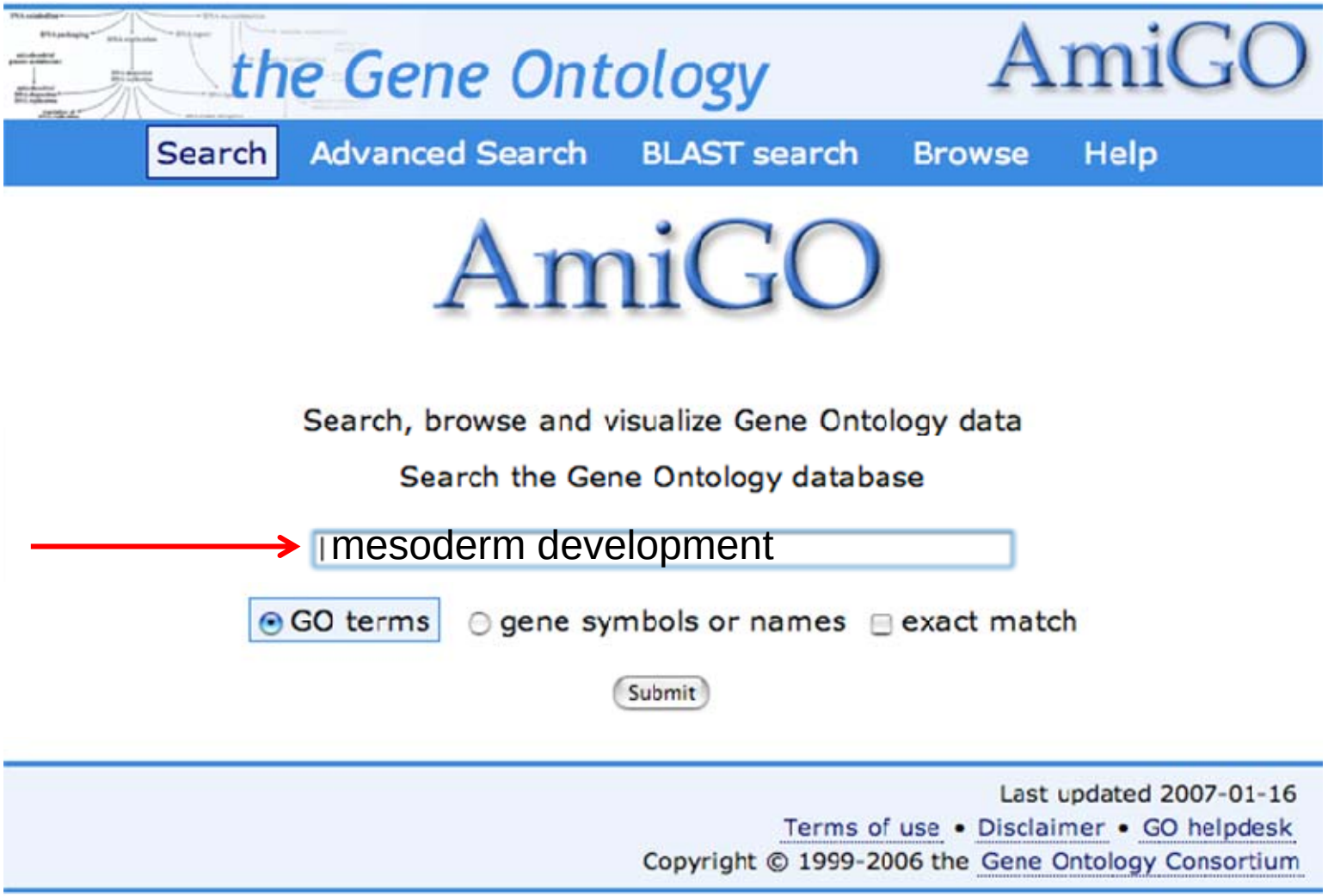
How will you ever find what you want?



SJChen/CGU/2011/

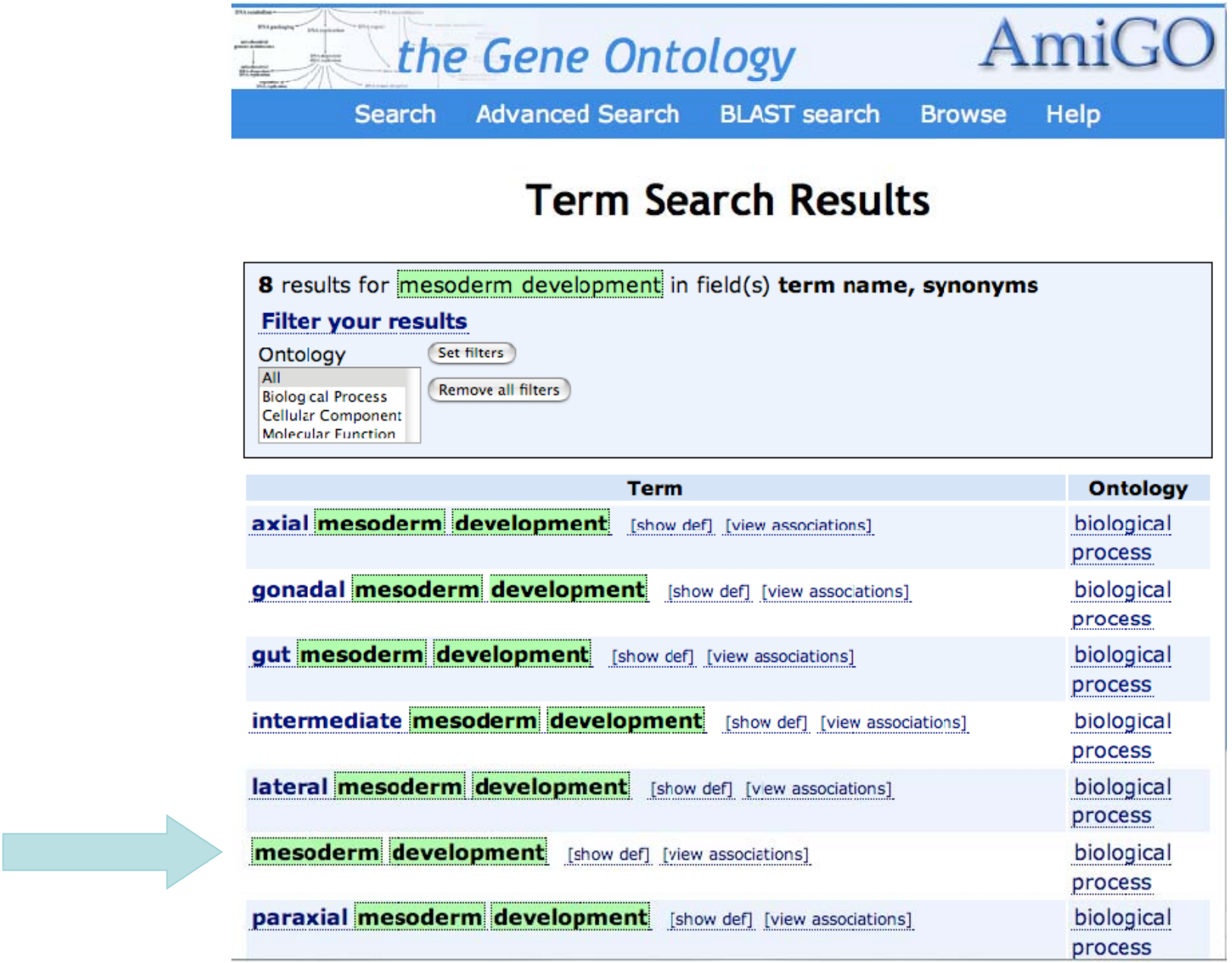
Adopted from http://www.geneontology.org/

GO browser



SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

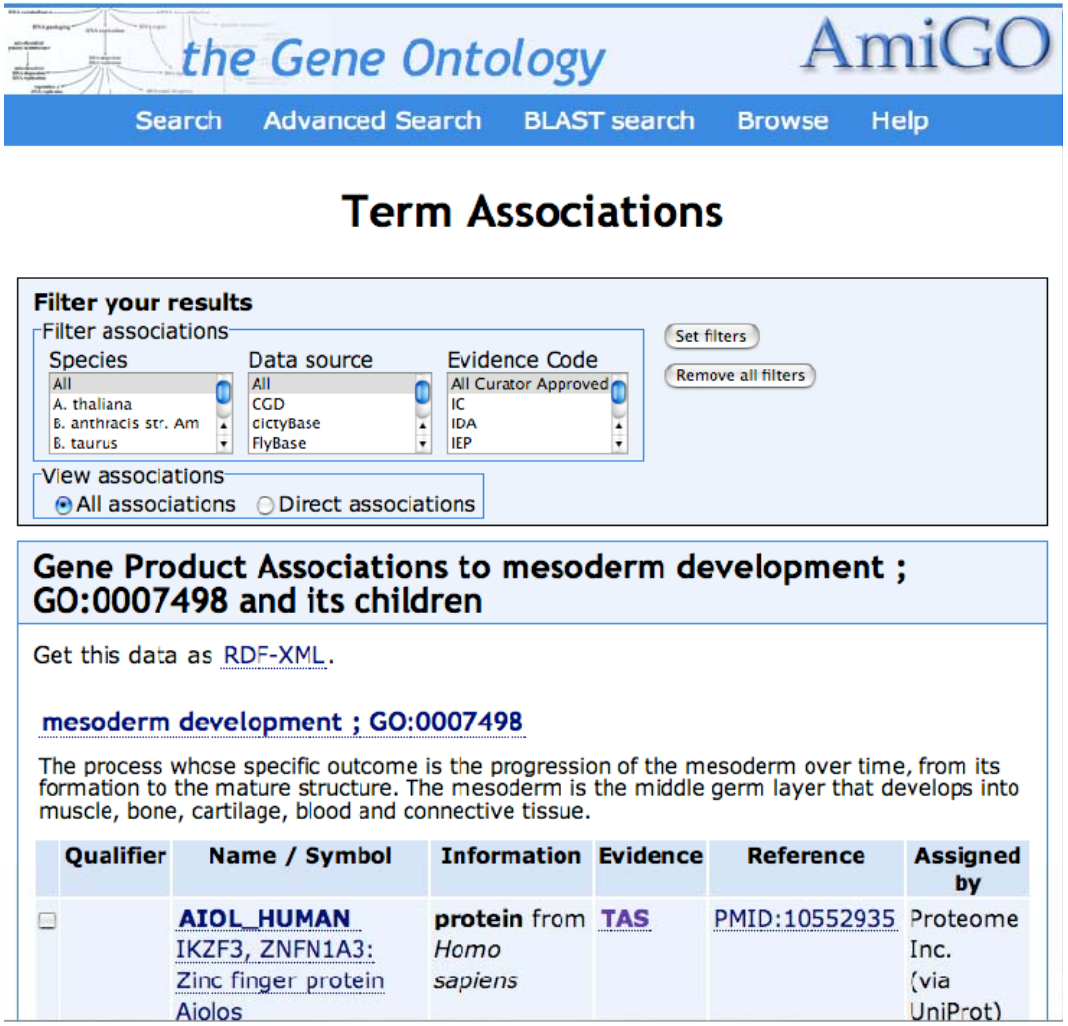


Adopted from http://www.geneontology.org/

SJChen/CGU/2011/

Definition of mesoderm development

Gene products involved in mesoderm development



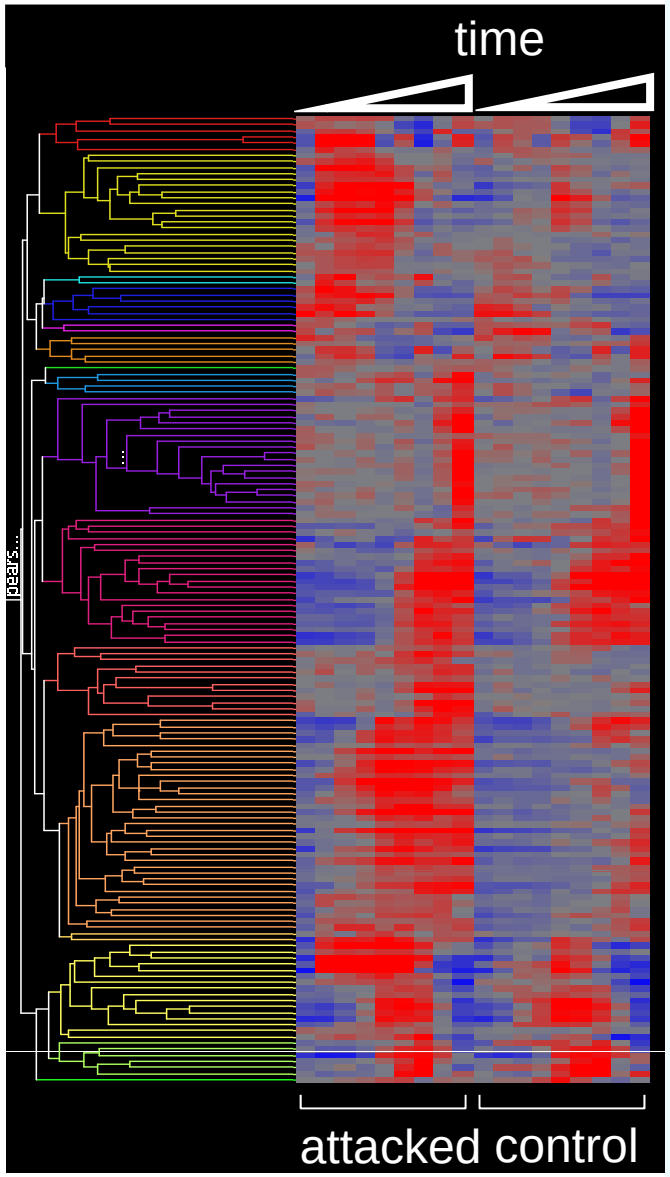
SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Example 2

Using GO to classify genes differentially expressed from microarray study

Bregje Wertheim at the Centre for Evolutionary Genomics, Department of Biology, UCL and Eugene Schuster Group, EBI.

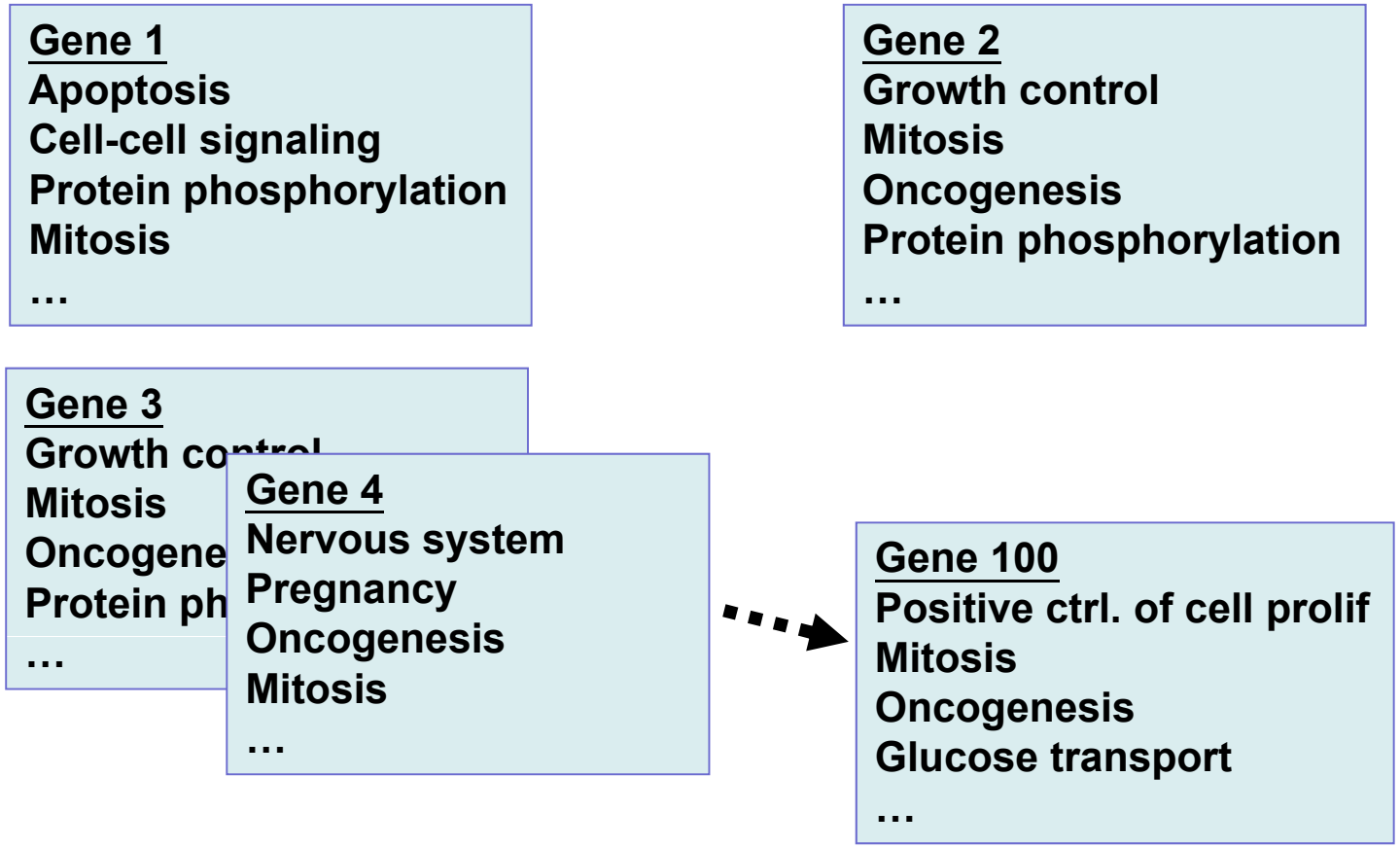


Microarray data shows changed expression of thousands of genes.

How will you spot the patterns?

Adopted from http://www.geneontology.org/

Traditional Analysis



- **After searching all information about these 100 genes, it is still difficult to know which biological processes are most significantly altered**

Adopted from http://www.geneontology.org/

Using GO Annotations

- But by using GO annotations, this work has already been done



Adopted from http://www.geneontology.org/

SJChen/CGU/2011/

Grouping Genes by Biological Process

Apoptosis
Gene 1
Gene 53

Mitosis
Gene 2
Gene 5
Gene45
Gene 7
Gene 35
…

Glucose transport
Gene 7
Gene 3
Gene 6
…

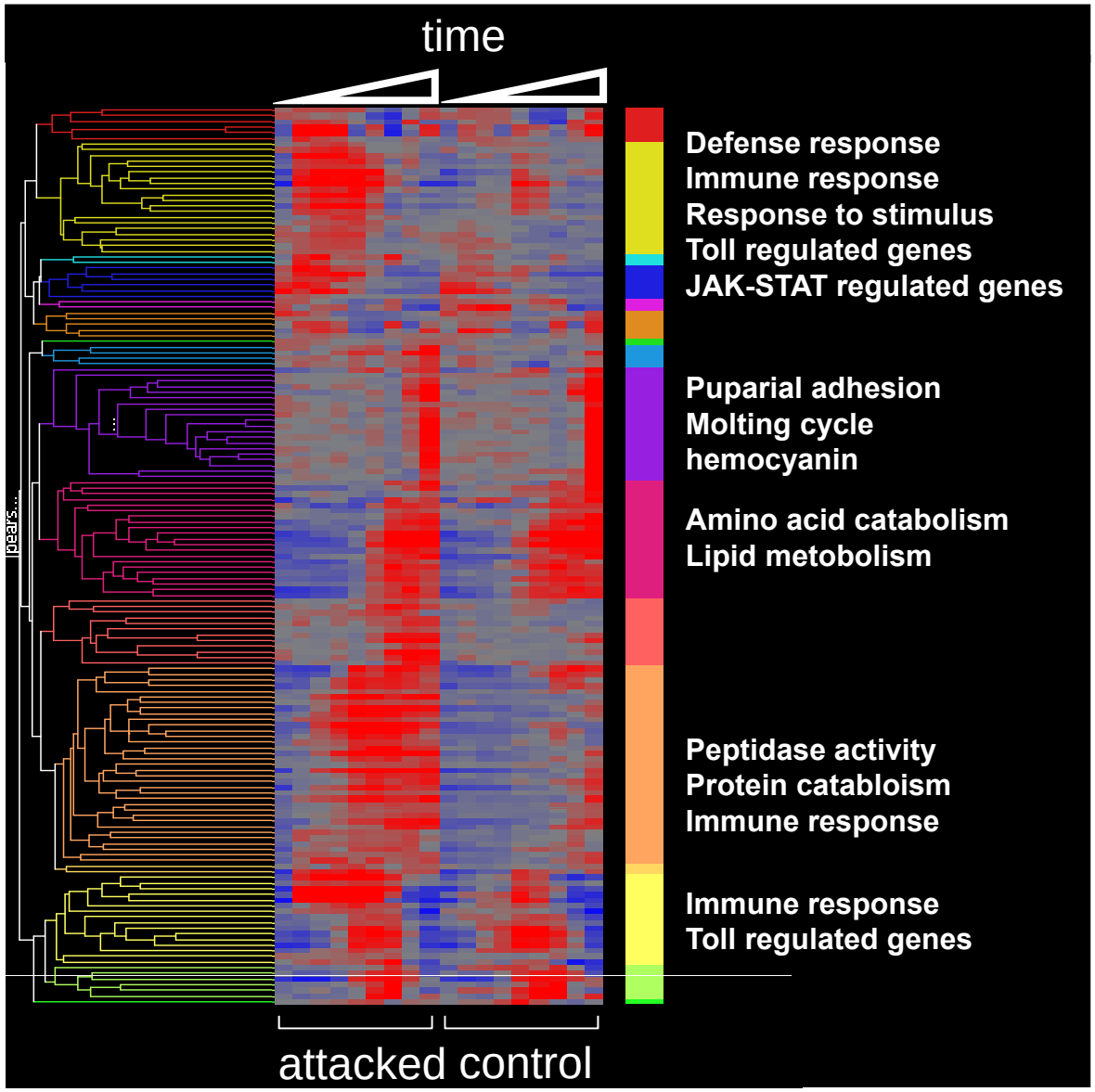
Positive ctrl. of cell prolif.
Gene 7
Gene 3
Gene 12
…

Growth
Gene 5
Gene 2
Gene 6
…

SJChen/CGU/2011/

Adopted from http://www.geneontology.org/

Bregje Wertheim at the Centre for Evolutionary Genomics, Department of Biology, UCL and Eugene Schuster Group, EBI.



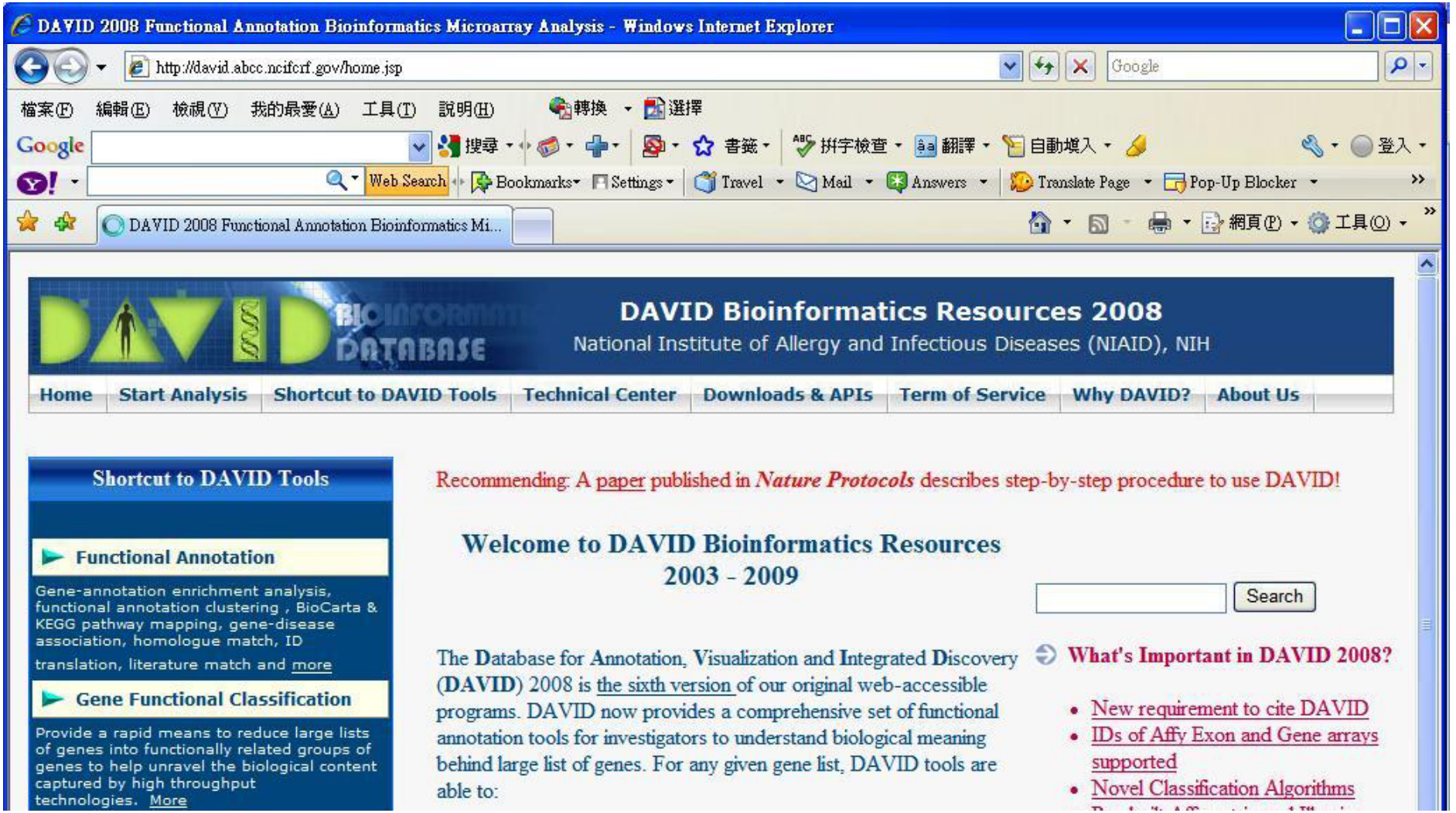
Adopted from http://www.geneontology.org/

SJChen/CGU/2011/

How to spot biological functions embedded in a gene list?

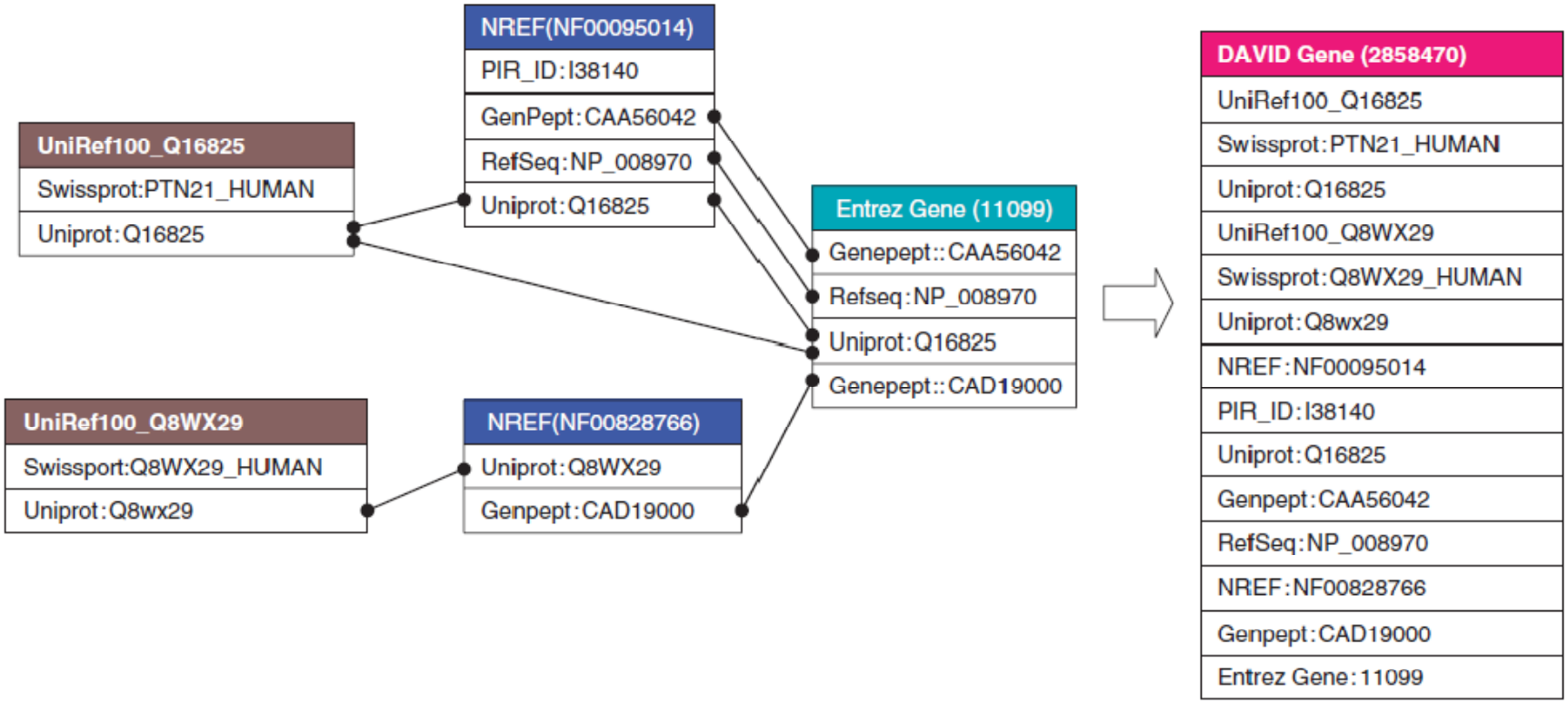
DAVID Bioinformatics Resources

- **DAVID web server : http://david.abcc.ncifcrf.gov/home.jsp**



SJChen/CGU/2011/

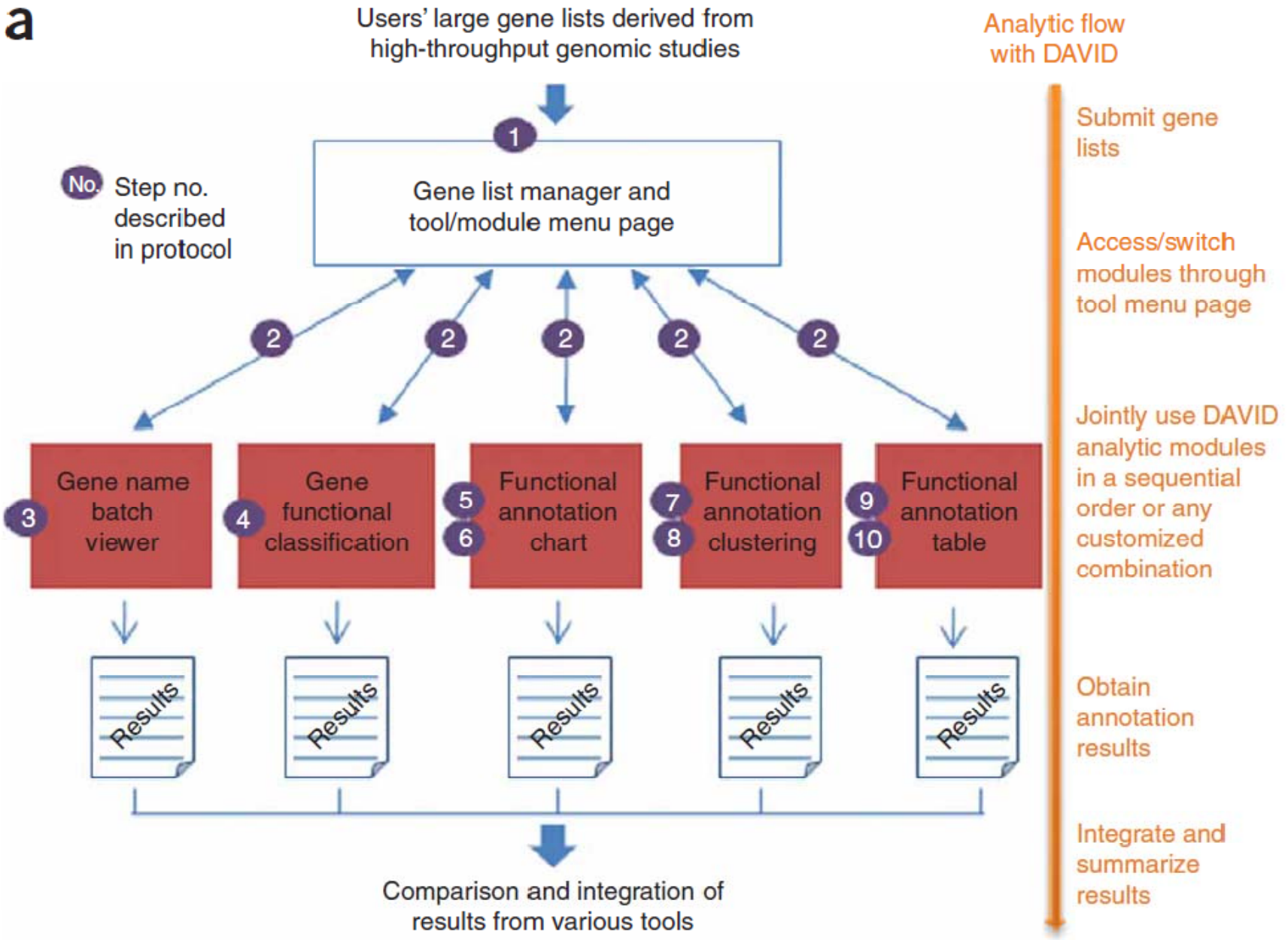
Construction of a DAVID Gene



SJChen/CGU/2011/

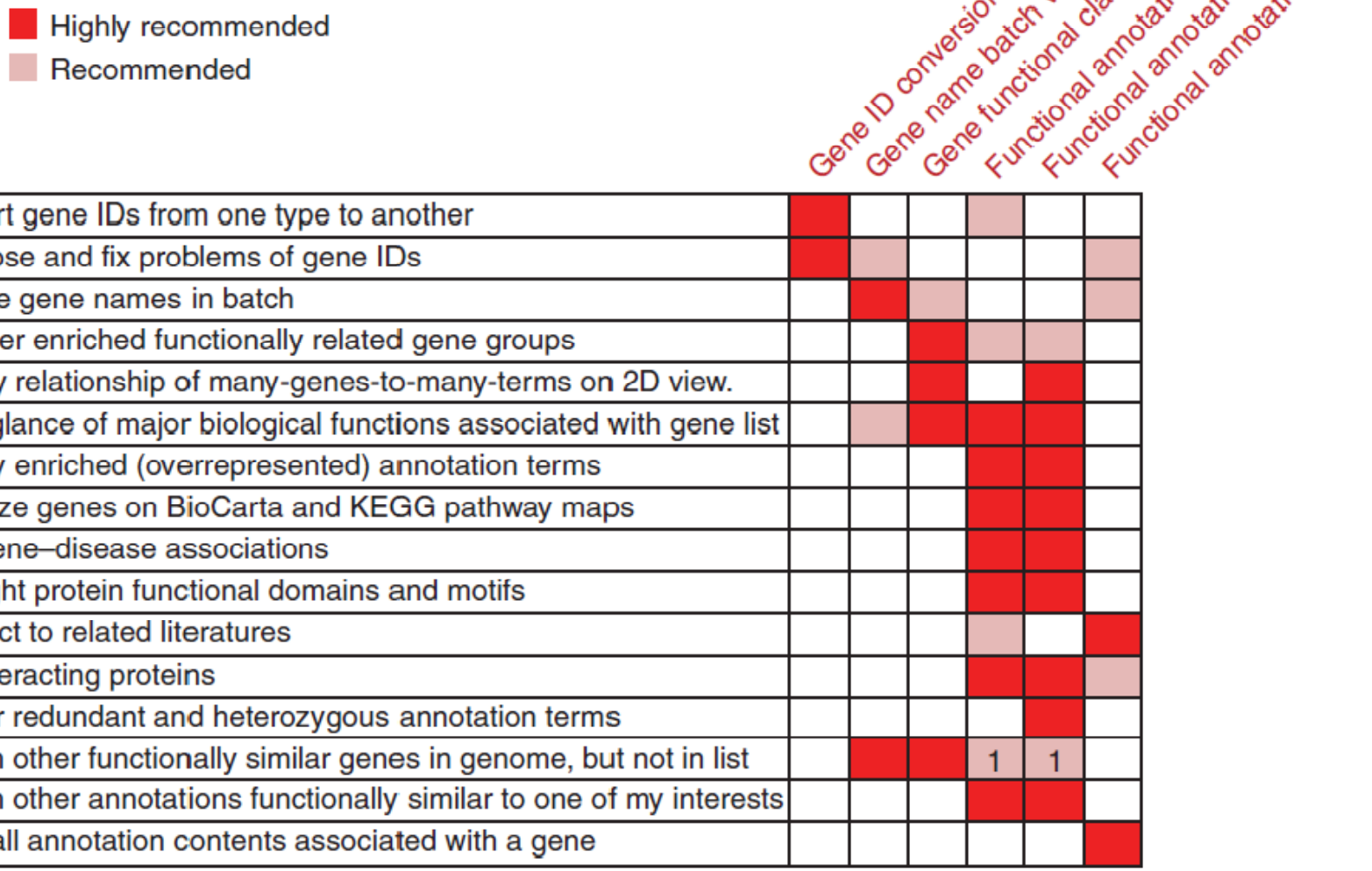
Nucleic Acid Res (2007) 35:W169

Analytic tools/modules in DAVID



SJChen/CGU/2011/

DAVID analytic modules



Gene List – Quality Control

- Reasonable number of genes ranging from hundreds to thousands (e.g., 100–2,000 genes), not extremely low or high.
- Most of the genes significantly pass the statistical threshold for selection (e.g., selecting genes by comparing gene expression between control and experimental cells with t-test statistics: fold changes ≥ 2 and P-values ≤0.05).
- A 'good' gene list should consistently contain more enriched biology than that of a random list in the same size range during analysis in DAVID.

Background List - Definition

- To decide the degree of enrichment, a certain background must be set up to be compared with the user's gene list.
- For example, 10% of user's genes are kinases versus 3% of genes in human genome (this is population background) are kinases.
- Thus, the conclusion is obvious in the particular example that the user's study is highly related to kinase.
- However, 10% itself alone cannot provide such a conclusion without comparing it with the background information.

Background List – How to use

- A general guideline is to set up the reference background as the pool of genes that have a chance to be selected for the studied annotation category under the scope of users' particular study
- Default background is the entire genome-wide genes of the species matching the user's input IDs.
- Pre-built backgrounds, such as genes in Affymetrix chips and so on, are available for the user's choice
- In principle, a larger gene background tends to give smaller P-values.
- As most of the high throughput studies are, or at least are close to, genome-wide scope, the default background is good for regular cases in general

Classification Stringency

- To control the behavior of DAVID Fuzzy clustering

- A general guideline is to choose higher stringency settings for tight, clean and smaller numbers of clusters; otherwise, lower for loose, broader and larger numbers of clusters
- Default setting is medium
- Five predefined levels from lowest to highest for user's choices
- Users may want to play with different stringency for more satisfactory results

Enrichment Score - Definition

- To rank overall enrichment of gene groups.
- It is the geometric mean of all the enrichment P-values (EASE scores) for each annotation term associated with the gene members in the group.
- To emphasize that the geometric mean is a relative score instead of an absolute P-value, minus log transformation is applied on the average P-values.
- A higher score for a group indicates that the gene members in the group are involved in more important (enriched) terms in a given study; therefore, more attention should go to them

Fold Enrichment – How to use ?

- **Enrichment score of 1.3 is equivalent to non-log scale 0.05. Fold enrichment 1.5 and above are suggested to be considered as interesting.**
- **Caution should be taken when big fold enrichments are obtained from a small number of genes (e.g., ≤3). This situation often happens to the terms with a few genes (more specific terms) or of smaller size (e.g.,<100) of user's input gene list. In this case, the reliability is not as much as those fold enrichment scores obtained from larger numbers of genes**

P-vlaue (EASE score)

- To examine the significance of gene–term enrichment with a modified Fisher's exact test (EASE score).

- The smaller the P-values, the more significant they are
- Default cutoff is 0.1
- Users could set different levels of cutoff through option panel on the top of result page.
- Owing to the complexity of biological data mining of this type, P-values are suggested to be treated as score systems, i.e., suggesting roles rather than decision-making roles.
- Users themselves should play critical roles in judging 'are the results making sense or not for expected biology

Benjamini

- To globally correct enrichment P-values to control family-wide false discovery rate under certain rate (e.g., ≤0.05).
- It is one of the multiple testing correction techniques (Bonferroni, Benjamini and FDR) provided by DAVID
- More terms examined, more conservative the corrections are. As a result, all the P-values get larger
- It is great if the interesting terms have significant P-values after the corrections.
- But as the multiple testing correction techniques are known as conservative approaches, it could hurt the sensitivity of discovery if overemphasizing them.

% - Defintion

- Number of genes involved in given term is divided by the total number of user's input genes, i.e., percentage of user's input gene hitting a given term.
- For example, 10% of user's genes hit 'kinase activity
- It gives overall idea of gene distributions among the terms
- The higher percentage does not necessarily have a good EASE score because it also depends on the percentage of background genes

Data Interpretation

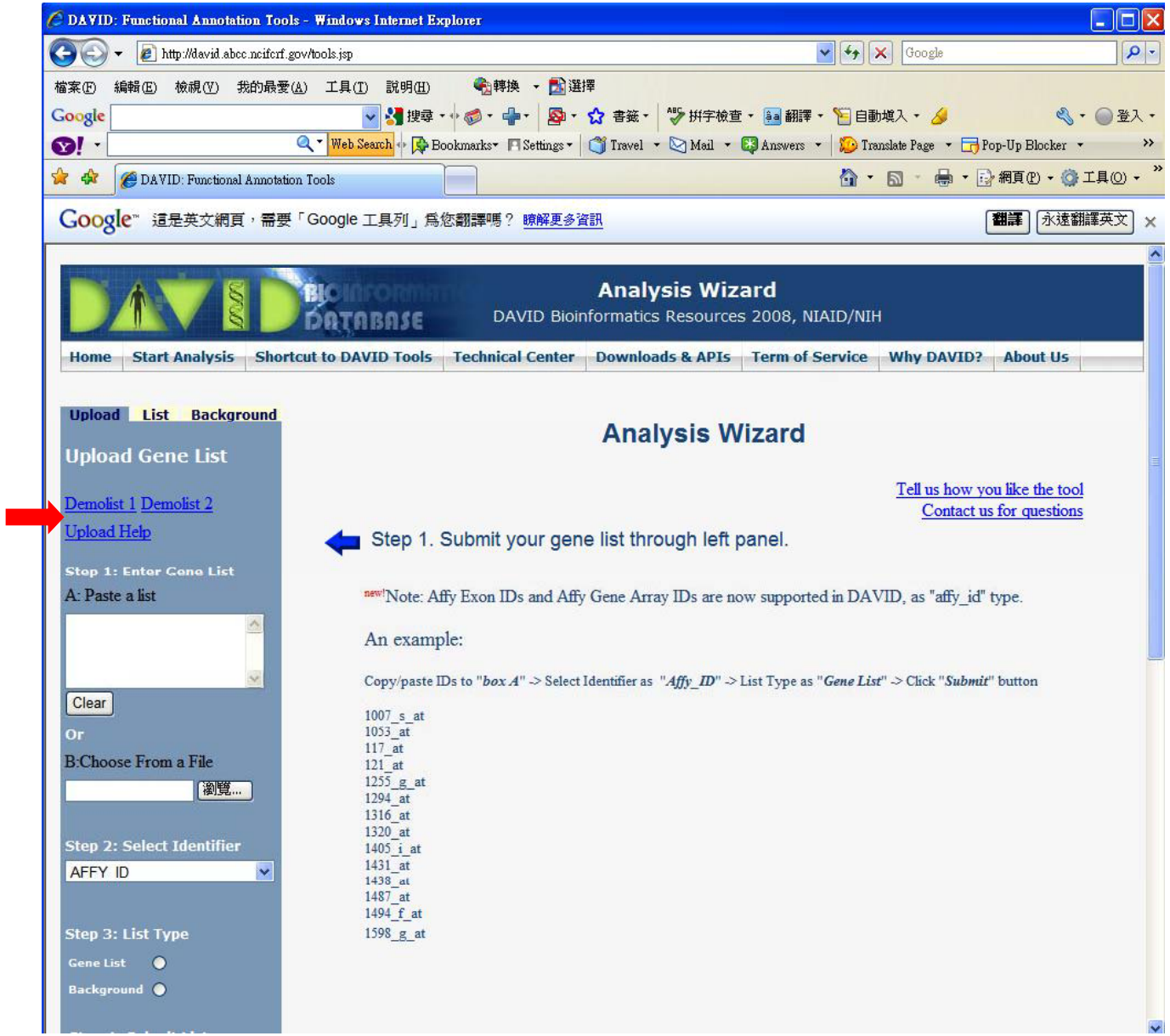
- **Fold enrichment and EASE score should always be examined side by side.**
- **Terms with larger fold enrichments and smaller EASE score may be interesting.**

Start analysis wizard

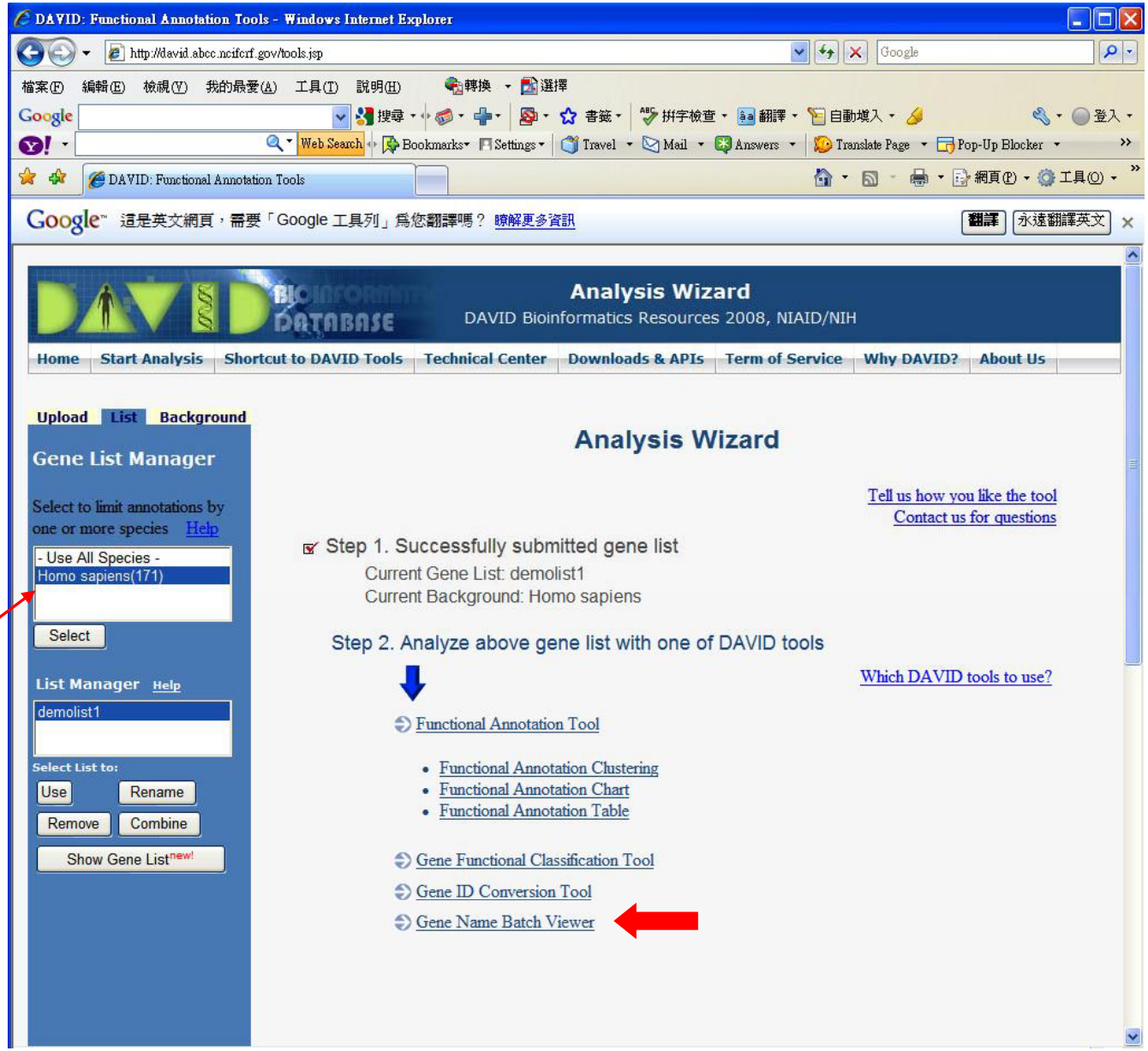
Click "Start Analysis" from anywhere within the website



Submit gene list or use built-in demo gene lists

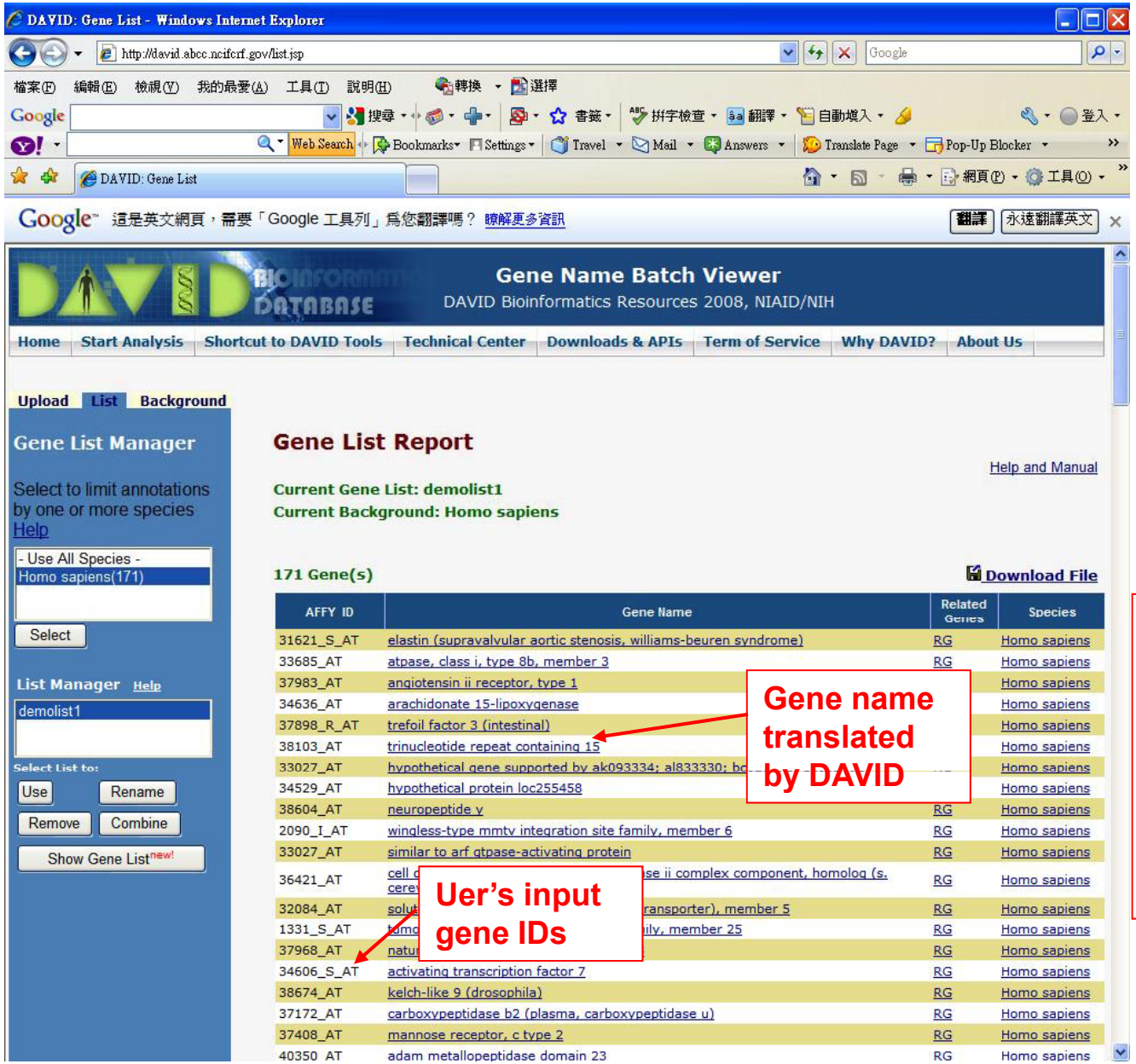


Select one of the DAVID Tools



Gene List Manager Panel

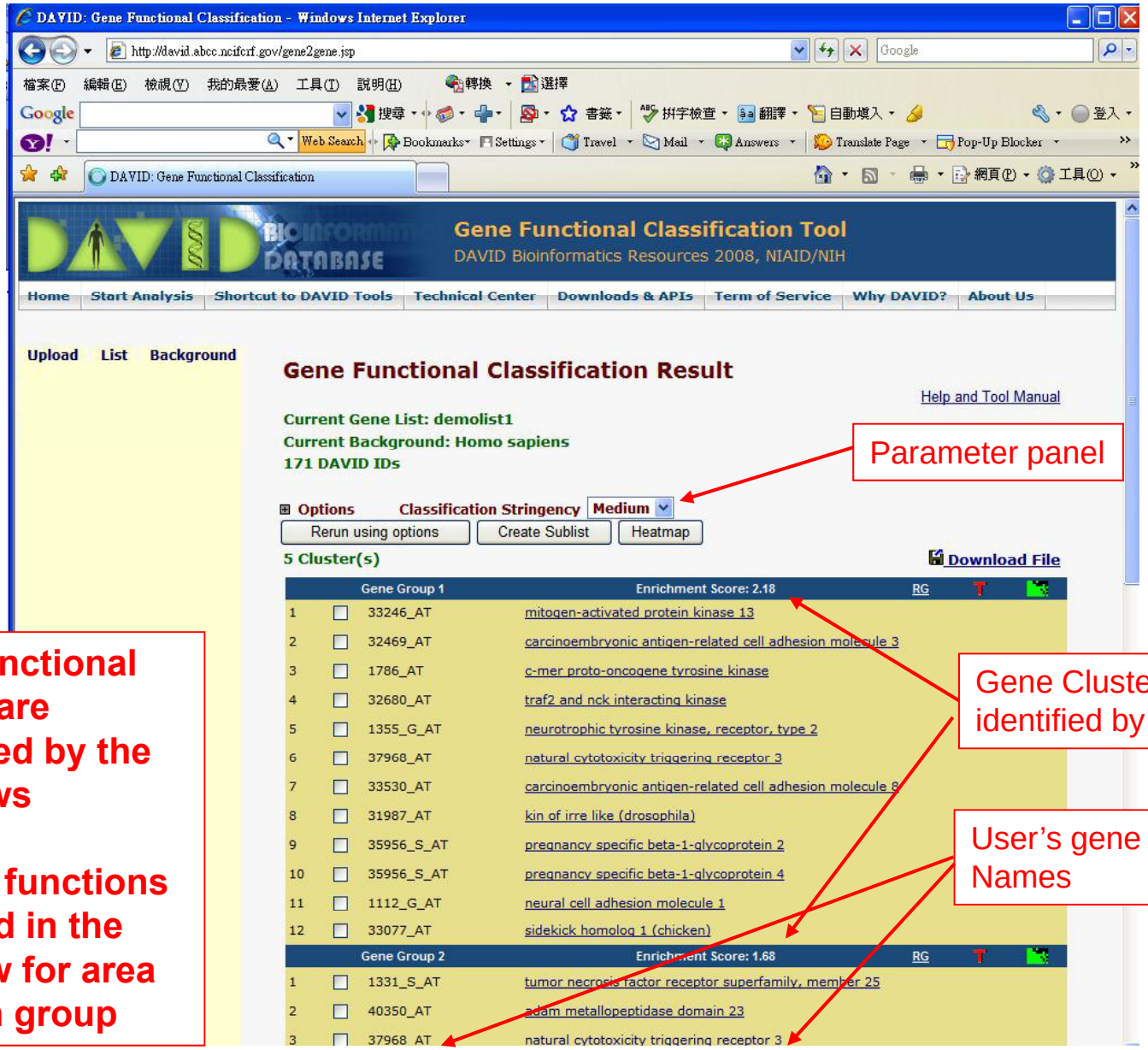
Gene Name Batch Viewer



Click on gene name will lead to more detail info

"RG" means "Related Genes" search fucntion

Gene Functional Classification



Parameter panel

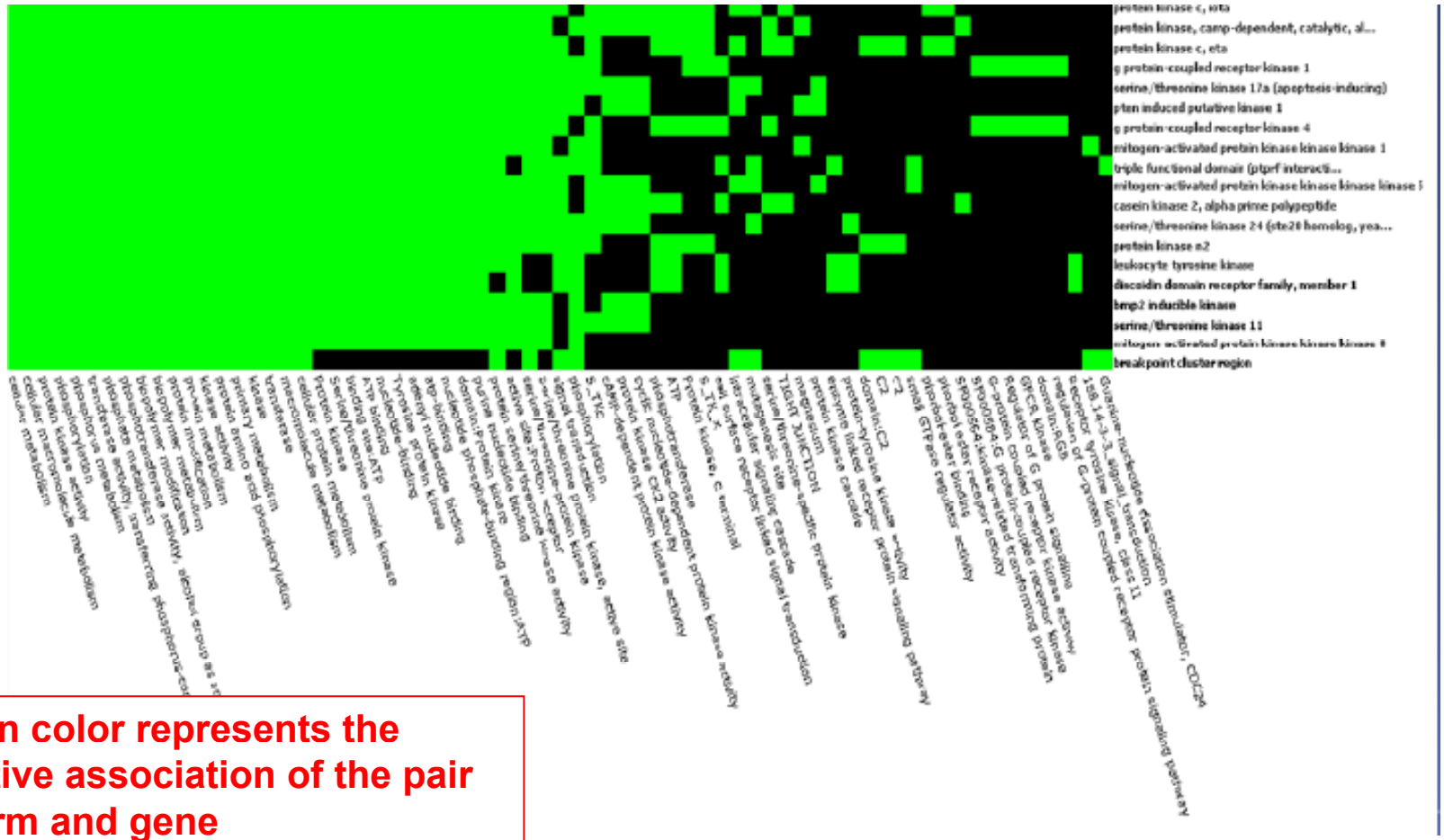
Gene Clusters identified by DAVID

User's gene IDs & Names

Gene functional groups are separated by the blue rows

A set of functions provided in the blue row for area for each group

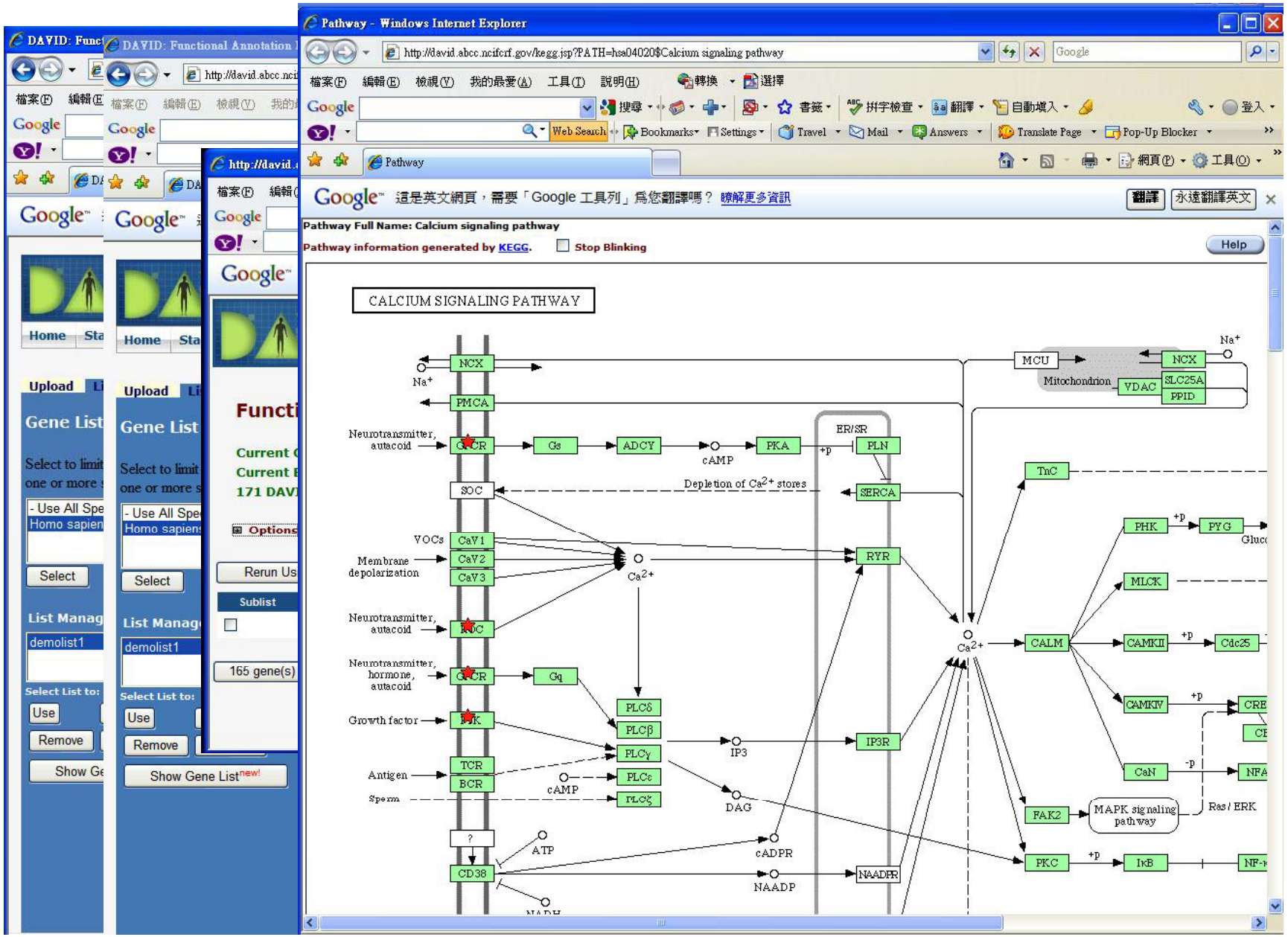
2D View of Gene Function Classification



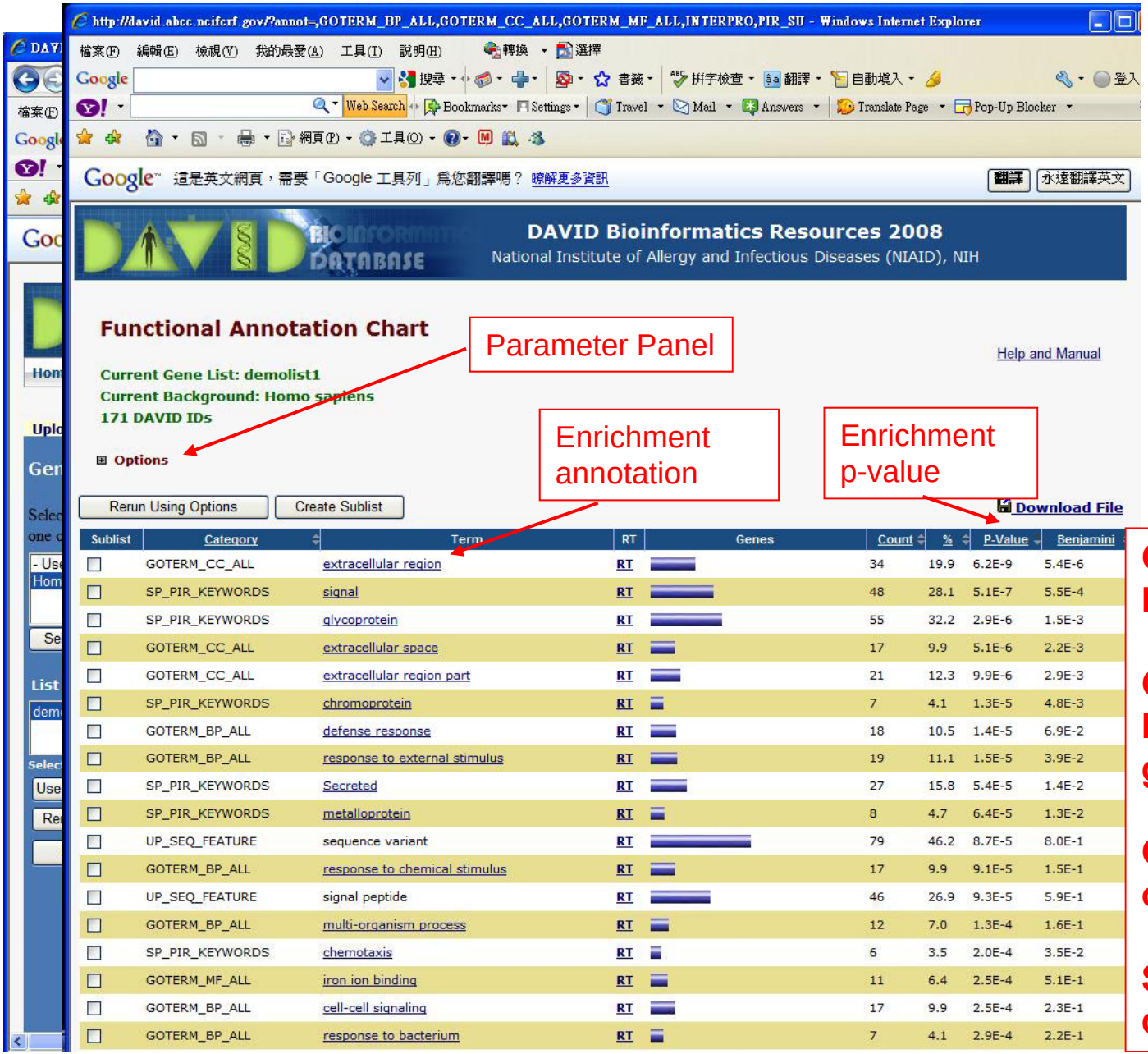
Green color represents the positive association of the pair of term and gene

Blank color represents the negative or no association of the pair of term and gene

Select annotation category and run Functional Annotation Chart



Select annotation category and run Functional Annotation Chart



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	GOTERM_CC_ALL	extracellular region	RT		34	19.9	6.2E-9	5.4E-6
☐	SP_PIR_KEYWORDS	signal	RT		48	28.1	5.1E-7	5.5E-4
☐	SP_PIR_KEYWORDS	glycoprotein	RT		55	32.2	2.9E-6	1.5E-3
☐	GOTERM_CC_ALL	extracellular space	RT		17	9.9	5.1E-6	2.2E-3
☐	GOTERM_CC_ALL	extracellular region part	RT		21	12.3	9.9E-6	2.9E-3
☐	SP_PIR_KEYWORDS	chromoprotein	RT		7	4.1	1.3E-5	4.8E-3
☐	GOTERM_BP_ALL	defense response	RT		18	10.5	1.4E-5	6.9E-2
☐	GOTERM_BP_ALL	response to external stimulus	RT		19	11.1	1.5E-5	3.9E-2
☐	SP_PIR_KEYWORDS	Secreted	RT		27	15.8	5.4E-5	1.4E-2
☐	SP_PIR_KEYWORDS	metalloprotein	RT		8	4.7	6.4E-5	1.3E-2
☐	UP_SEQ_FEATURE	sequence variant	RT		79	46.2	8.7E-5	8.0E-1
☐	GOTERM_BP_ALL	response to chemical stimulus	RT		17	9.9	9.1E-5	1.5E-1
☐	UP_SEQ_FEATURE	signal peptide	RT		46	26.9	9.3E-5	5.9E-1
☐	GOTERM_BP_ALL	multi-organism process	RT		12	7.0	1.3E-4	1.6E-1
☐	SP_PIR_KEYWORDS	chemotaxis	RT		6	3.5	2.0E-4	3.5E-2
☐	GOTERM_MF_ALL	iron ion binding	RT		11	6.4	2.5E-4	5.1E-1
☐	GOTERM_BP_ALL	cell-cell signaling	RT		17	9.9	2.5E-4	2.3E-1
☐	GOTERM_BP_ALL	response to bacterium	RT		7	4.1	2.9E-4	2.2E-1

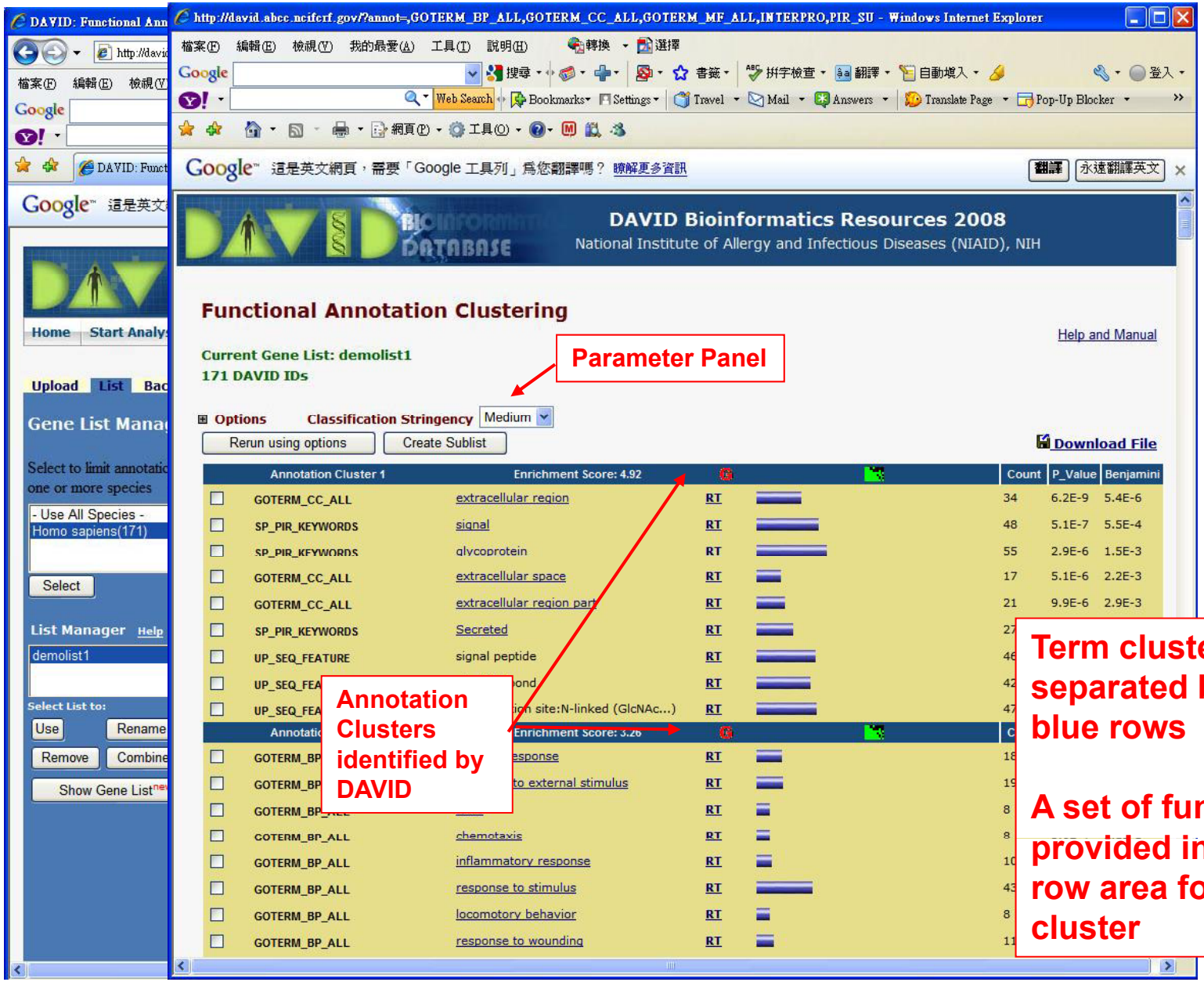
Click on term name lead to details

Click on blue bar to list all associated genes

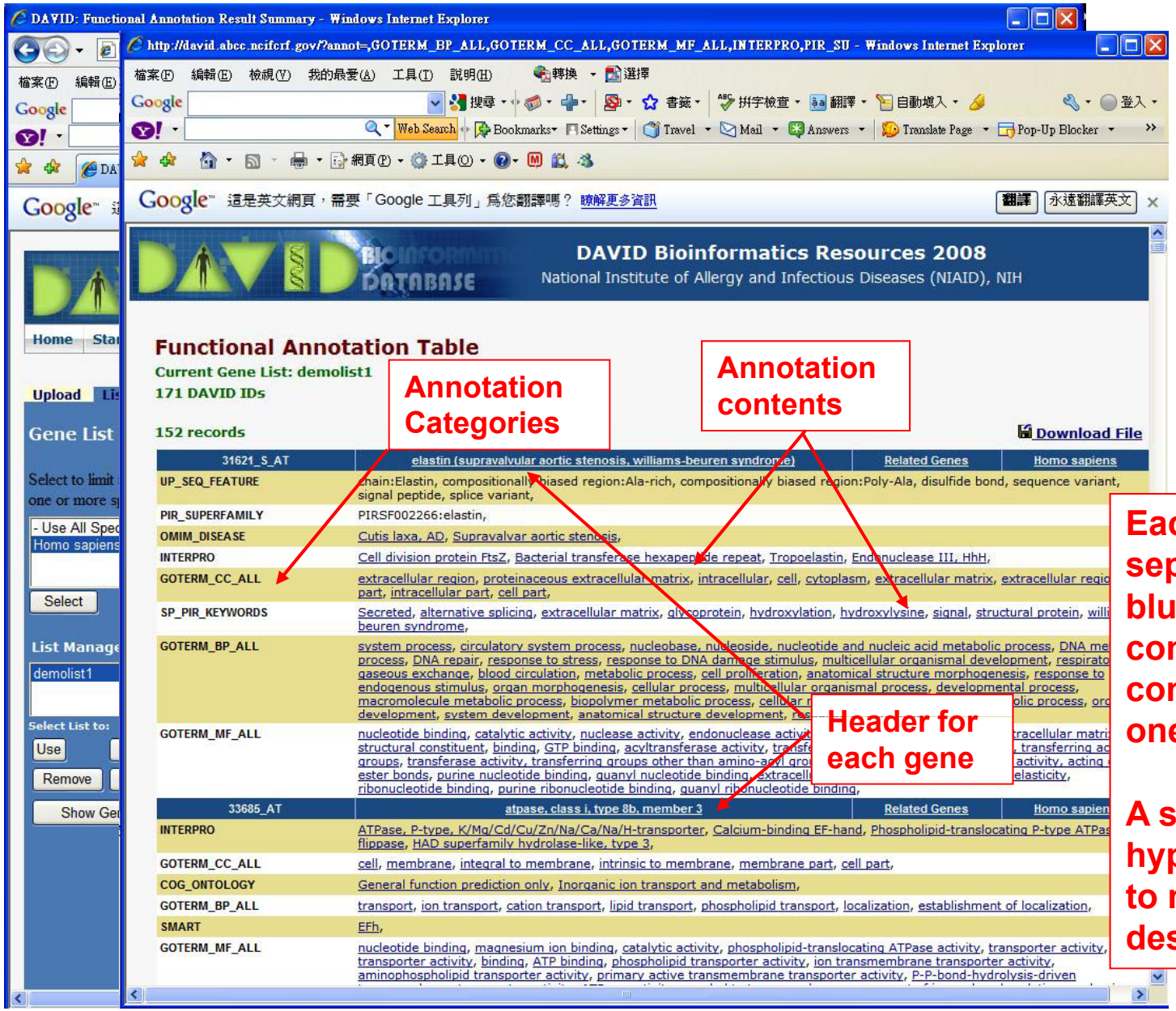
Click on RT to list other related terms

Sort results by different columns

Select annotation category and run Functional Clustering



Functional Table



Each block separated by blue rows contains the contents for one gene

A set of hyperlinks lead to more detailed descrptions

DAVID Bioinformatics Resources

- DAVID web server : http://david.abcc.ncifcrf.gov/home.jsp



SJChen/CGU/2011/