# BGGN 213

## Pathway Analysis and the Interpretation of Gene Lists

Barry Grant
UC San Diego

http://thegrantlab.org/bggn213

My high-throughput experiment generated a long list of genes/proteins…

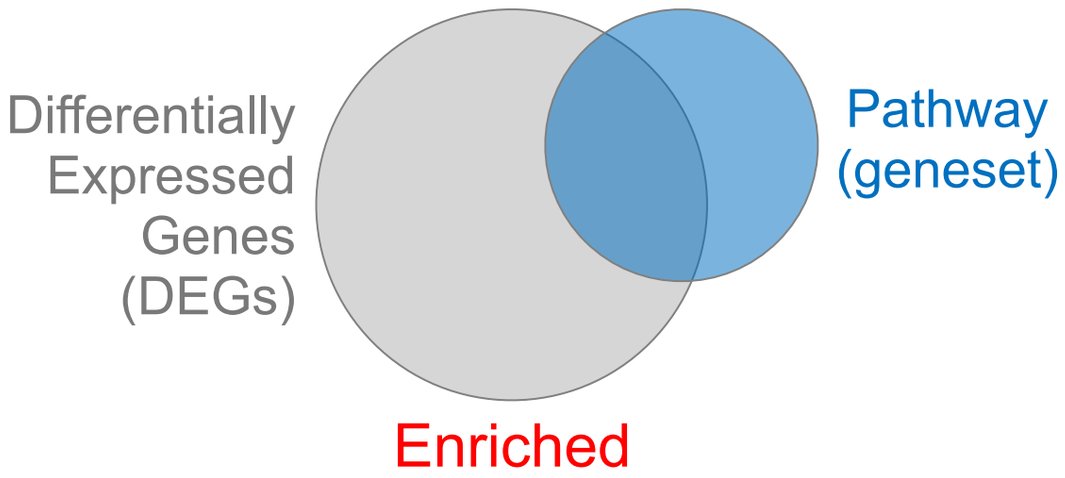
What do I do now?

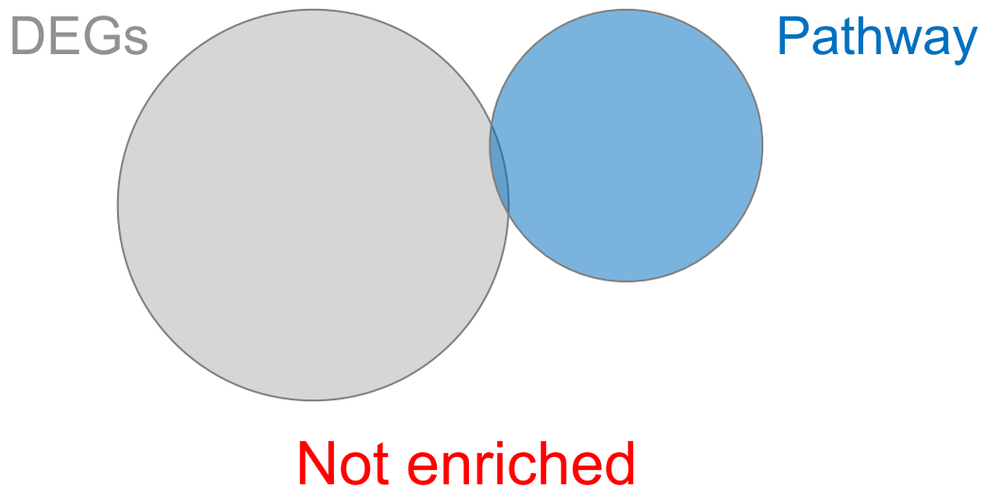
Pathway analysis!

(a.k.a. geneset enrichment)

Use bioinformatics methods to help extract biological meaning from such lists…

# Pathway analysis (a.k.a. geneset enrichment)
# **Principle**




- Variations of the math: overlap, ranking, networks... ➢ *Not critical, different algorithms show similar performances*
- DEGs come from your experiment ➢ *Critical, needs to be as clean as possible*
- Pathway genes (“geneset”) come from annotations ➢ *Important, but typically not a competitive advantage*

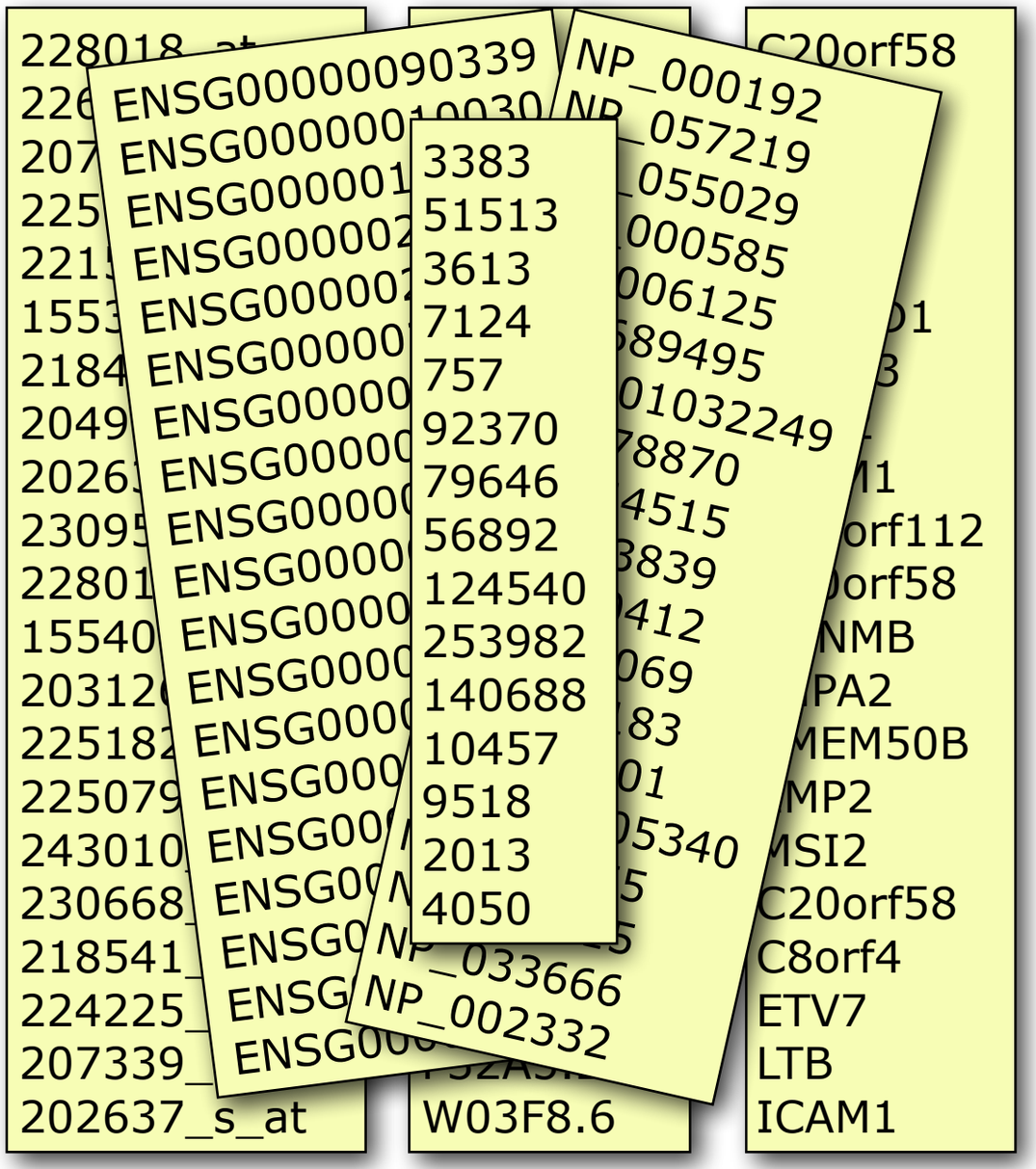
# Pathway analysis (a.k.a. geneset enrichment) **Limitations**

---

- **Post-transcriptional regulation** is neglected
- **Directionality** is hard to capture sensibly
    - e.g. IκBα/NF-κB
- **Tissue-specific** variations of pathways are not annotated
    - e.g. NF-κB regulates metabolism, not inflammation, in adipocytes
- **Size bias**: stats are influenced by the size of the pathway
- **Geneset annotation bias**: can only discover what is already known
- **Non-model organisms**: no high-quality genesets available
- Many pathways/receptors **converge** to few regulators
    - e.g. tens of innate immune receptors activate 4 TFs: NF-kB, AP-1, IRF3/7, NFAT

# Starting point for pathway analysis:
# **Your gene list**

- You have a list of genes/proteins of interest
- You have quantitative data for each gene/protein
  - Fold change
  - p-value
  - Spectral counts
  - Presence/absence



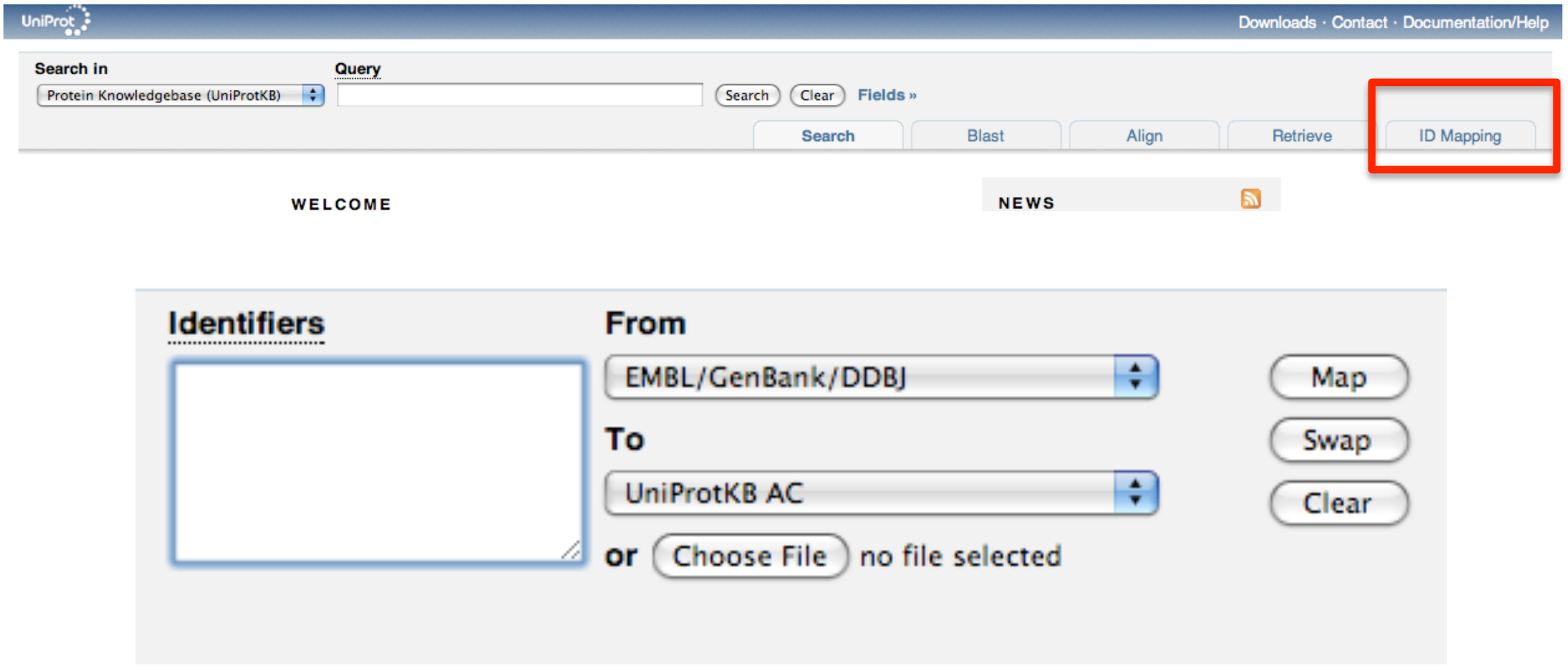
# Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Sometimes you have to translate one set of ids into another
  - A program might only accept certain types of ids
  - You might have a list of genes with one type of id and info for genes with another type of id

# Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Sometimes you have to translate one set of ids into another
  - A program might only accept certain types of ids
  - You might have a list of genes with one type of id and info for genes with another type of id
- **Various web sites translate ids -> *best for small lists***
  - **UniProt < www.uniprot.org>; IDConverter < idconverter.bioinfo.cnio.es >**

# Translating between identifiers:
# UniProt < www.uniprot.org >

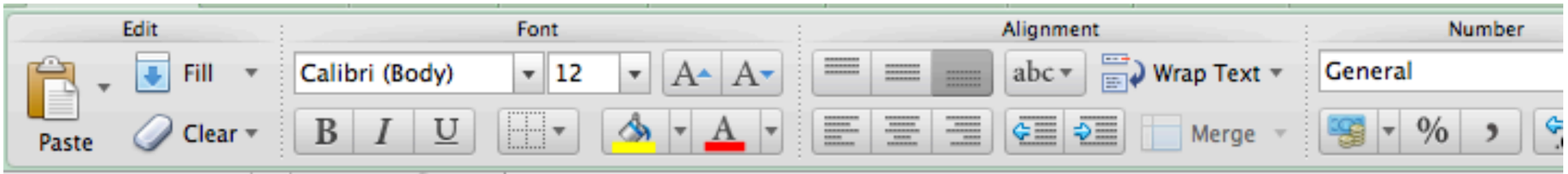


# Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Sometimes you have to translate one set of ids into another
  - A program might only accept certain types of ids
  - You might have a list of genes with one type of id and info for genes with another type of id
- Various web sites translate ids -> *best for small lists*
  - UniProt < www.uniprot.org>; IDConverter < idconverter.bioinfo.cnio.es >
- **VLOOKUP in Excel - *good if you are an excel whizz - I am not!***
  - **Download flat file from Entrez, Uniprot, etc; Open in Excel; Find columns that correspond to the 2 IDs you want to convert between; Sort by ID; Use vlookup to translate your list**

# Translating between identifiers: Excel VLOOKUP

VLOOKUP(lookup_value, table_array, col_index_num)




| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Data Table | | | | | | Annotation Table | | | | |
| 2 | RefSeq | Symbol | Exp1 | Exp2 | Exp3 | | RefSeq | Symbol | Entrez ID | Unigene | RefSeq |
| 3 | NM_153103 | Kif1c | 2.31975457 | 1.24558927 | 2.78816871 | | NM_001001 | Zfp85-rs1 | 22746 | Mm.288396 | NM_001 |
| 4 | NM_146017 | Gabrp | 4.15029735 | 3.08055836 | 1.18919962 | | NM_001001 | Scap | 235623 | Mm.288741 | NM_001 |
| 5 | NM_018883 | Camkk1 | 3.83282512 | 0.0522951 | 0.64684259 | | NM_001001 | Scap | 235623 | Mm.288741 | NM_001 |
| 6 | NM_145936 | Tspyl2 | 0.45449369 | 1.62761318 | 7.59770627 | | NM_001001 | Fbxo41 | 330369 | Mm.38777 | NM_001 |
| 7 | NM_026599 | Cgnl1 | 4.84541871 | 2.84751796 | 1.61595768 | | NM_001001 | Taf9b | 407786 | Mm.19440 | NM_001 |
| 8 | NM_013926 | Cbx8 | 1.22903318 | 0.2863077 | 0.02952665 | | NM_001001 | Taf9b | 407786 | Mm.19440 | NM_001 |
| 9 | NR_015566 | A330023F24 | 1.44695053 | 0.98809479 | 1.59330144 | | NM_001001 | BC051142 | 407788 | Mm.73205 | NM_001 |
| 10 | NM_008623 | Mpz | 0.50749263 | 0.94350028 | 6.10581569 | | NM_001001 | BC051142 | 407788 | Mm.73205 | NM_001 |
| 11 | NM_183127 | Fate1 | 2.45672795 | 4.87960794 | 3.60759511 | | NM_001001 | BC048546 | 232400 | Mm.259234 | NM_001 |
| 12 | NM_008943 | | 4.78701069 | 4.15302647 | 0.85432314 | | NM_001001 | Zfp941 | 407812 | Mm.359154 | NM_001 |
| 13 | NM_025382 | | 0.66397344 | 1.40664187 | 3.09539802 | | NM_001001 | BC031181 | 407819 | Mm.29866 | NM_001 |
| 14 | NM_182841 | | 1.25528938 | 0.20505996 | 2.76879488 | | NM_001001 | Baz2b | 407823 | Mm.486364 | NM_001 |
| 15 | NM_030061 | | 0.17670108 | 2.75415469 | 2.98900691 | | NM_001001 | Tmem204 | 407831 | Mm.34379 | NM_001 |
| 16 | NM_133216 | | 6.572343 | 0.59671282 | 3.84650536 | | NM_001001 | Ccdc111 | 408022 | Mm.217385 | NM_001 |
| 17 | NM_030063 | | 7.05132762 | 0.65043627 | 1.68111836 | | NM_001001 | BC048507 | 408058 | Mm.177840 | NM_001 |

# Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Sometimes you have to translate one set of ids into another
  - A program might only accept certain types of ids
  - You might have a list of genes with one type of id and info for genes with another type of id
- Various web sites translate ids -> *best for small lists*
  - UniProt < www.uniprot.org >; IDConverter < idconverter.bioinfo.cnio.es >
- VLOOKUP in Excel -> *good if you are an excel whizz - I am not!*
  - Download flat file from Entrez, Uniprot, etc; Open in Excel; Find columns that correspond to the 2 ids you want to convert between; Use vlookup to translate your list
- Use the **merge()** or **mapIDs()** functions in **R** - fast, *versatile* & reproducible!
  - Also **clusterProfiler::bitr()** function and many others… [Link to clusterProfiler vignette]

# What functional set databases do you want?

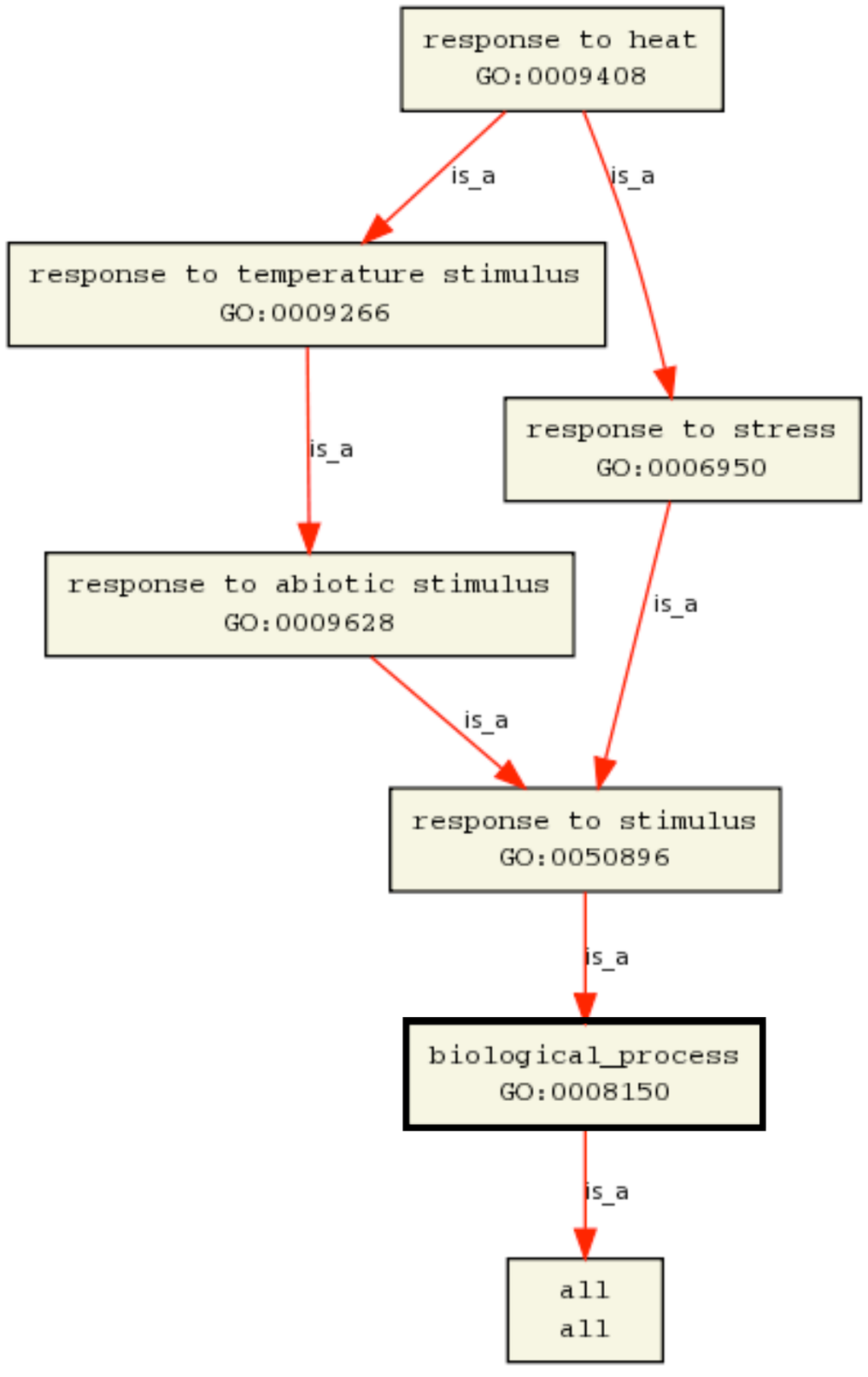
- Commonly used
  - **Gene Ontology (GO)**
  - **KEGG Pathways** (mostly metabolic)
  - **GeneGO MetaBase**
  - **Ingenuity Pathway Analysis (IPA)**
  - **MSigDB** (gene sets based on chromosomal position, cis-regulatory motifs, GO terms, etc)
- Many others...
  - Enzyme Classification, Pfam families
  - Open Biomedical Ontologies (OBO, www.obofoundry.org)

# GO database < www.geneontology.org >

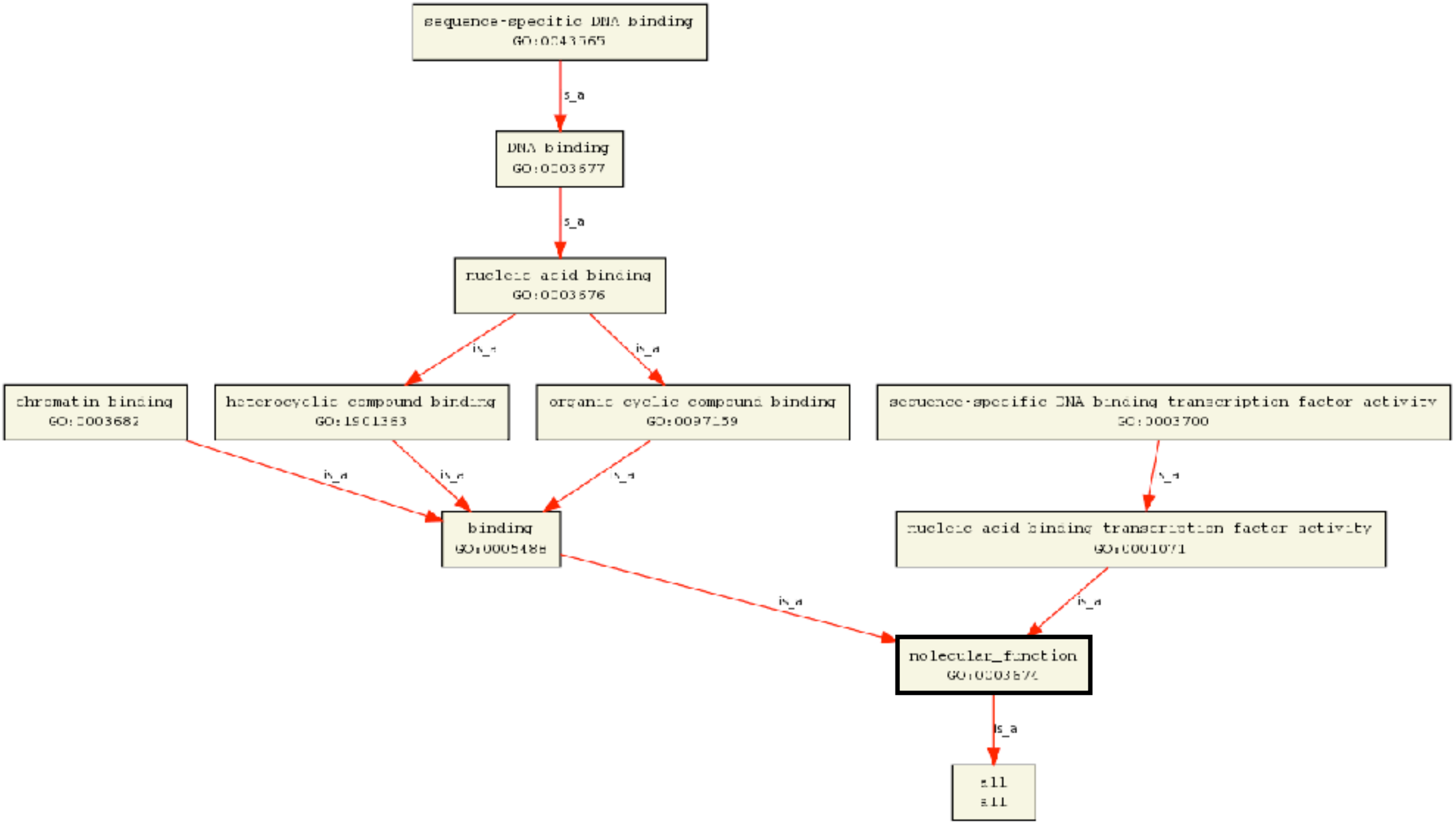
- **What function does HSF1 perform?**
  - *response to heat; sequence-specific DNA binding; transcription; etc*

- **Ontology** => a structured and controlled vocabulary that allows us to annotate gene products consistently, interpret the relationships among annotations, and can easily be *handled by a computer*

- GO database consists of 3 ontologies that describe gene products in terms of their associated **biological processes**, **cellular components** and **molecular functions**

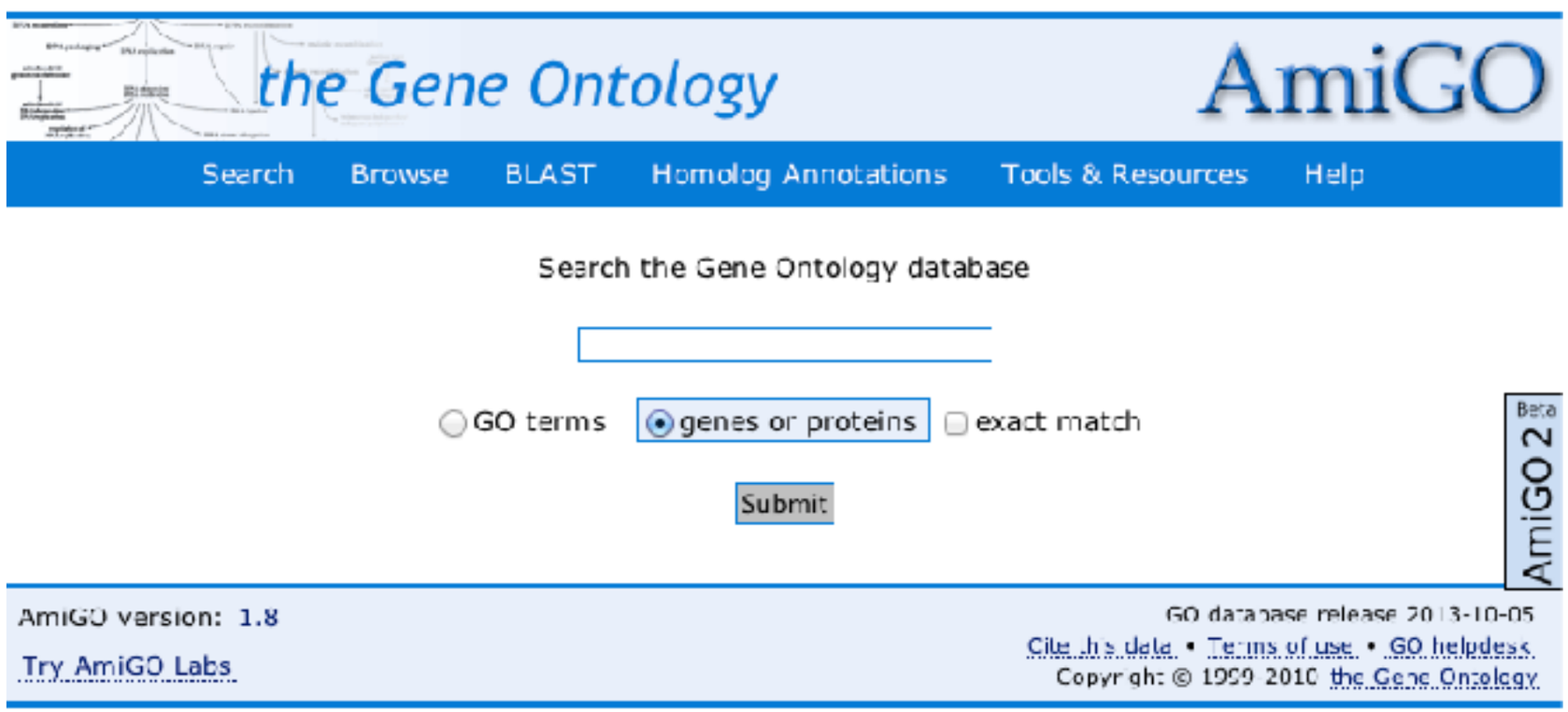



- Terms are nodes
- Relationships are edges
- Parent terms are more general
- Terms can have multiple parents


sequence-specific DNA binding
GO:0043565
is_a
DNA binding
GO:0003677
is_a
nucleic acid binding
GO:0003676
is_a
is_a
chromatin binding
GO:0003682
heterocyclic compound binding
GO:1901363
organic cyclic compound binding
GO:0097159
sequence-specific DNA binding transcription factor activity
GO:0003700
is_a
is_a
is_a
is_a
binding
GO:0005488
nucleic acid binding transcription factor activity
GO:0001071
is_a
is_a
molecular_function
GO:0003674
is_a
all
all

# GO Annotations

- GO is not a database of genes/proteins or sequences
- Gene products get annotated with GO terms by organism specific databases, such as Flybase, Wormbase, MGI, ZFIN, UniProt, etc
- Annotations are available through AmiGO < amigo.geneontology.org >



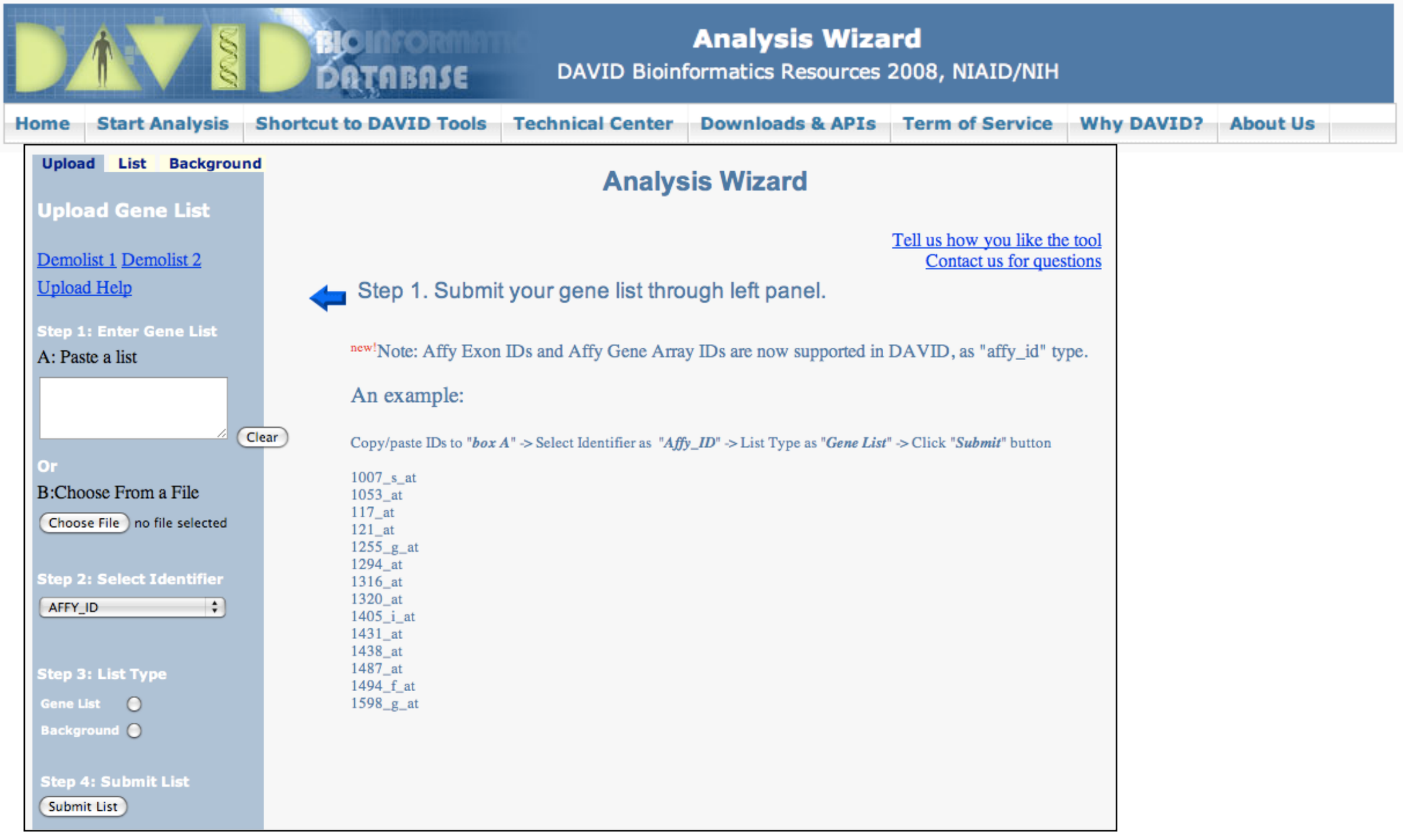
# GO evidence codes

| Evidence code | Evidence code description | Source of evidence | Manually checked | Current number of annotations* |
|---|---|---|---|---|
| IDA | Inferred from direct assay | Experimental | Yes | 71,050 |
| IEP | Inferred from expression pattern | Experimental | Yes | 4,598 |
| IGI | Inferred from genetic interaction | Experimental | Yes | 8,311 |
| IMP | Inferred from mutant phenotype | Experimental | Yes | 61,549 |
| IPI | Inferred from physical interaction | Experimental | Yes | 17,043 |
| ISS | Inferred from sequence or structural similarity | Computational | Yes | 196,643 |
| RCA | Inferred from reviewed computational analysis | Computational | Yes | 103,792 |
| IGC | Inferred from genomic context | Computational | Yes | 4 |
| IEA | Inferred from electronic annotation | Computational | No | 15,687,382 |
| IC | Inferred by curator | Indirectly derived from experimental or computational evidence made by a curator | Yes | 5,167 |
| TAS | Traceable author statement | Indirectly derived from experimental or computational evidence made by the author of the published article | Yes | 44,564 |
| NAS | Non-traceable author statement | No 'source of evidence' statement given | Yes | 25,656 |
| ND | No biological data available | No information available | Yes | 132,192 |
| NR | Not recorded | Unknown | Yes | 1,185 |

*October 2007 release

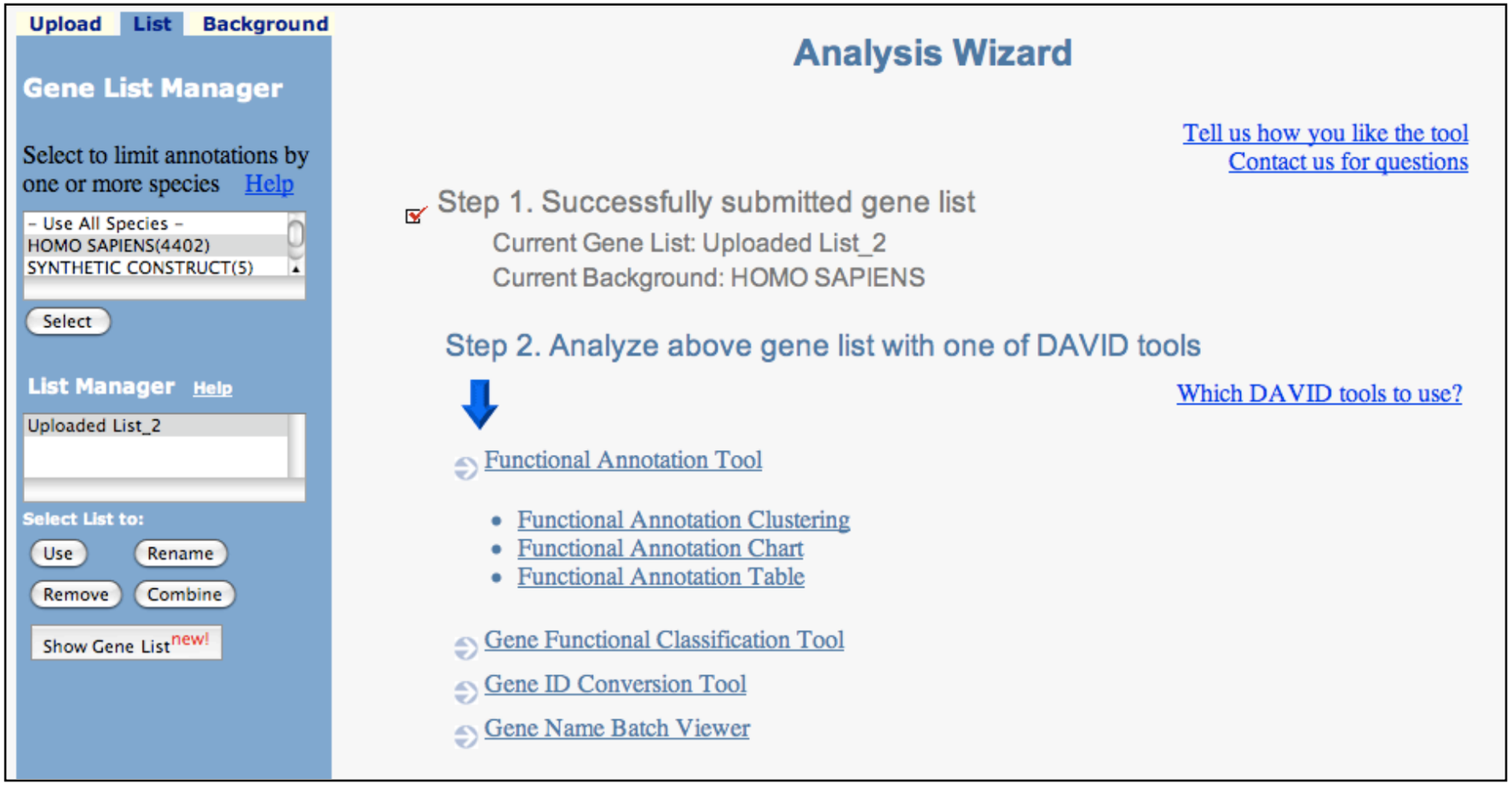
**Use and misuse of the gene ontology annotations**
Seung Yon Rhee, Valerie Wood, Kara Dolinski & Sorin Draghici
*Nature Reviews Genetics* **9**, 509-515 (2008)

# **DAVID** at NIAID < david.abcc.ncifcrf.gov >



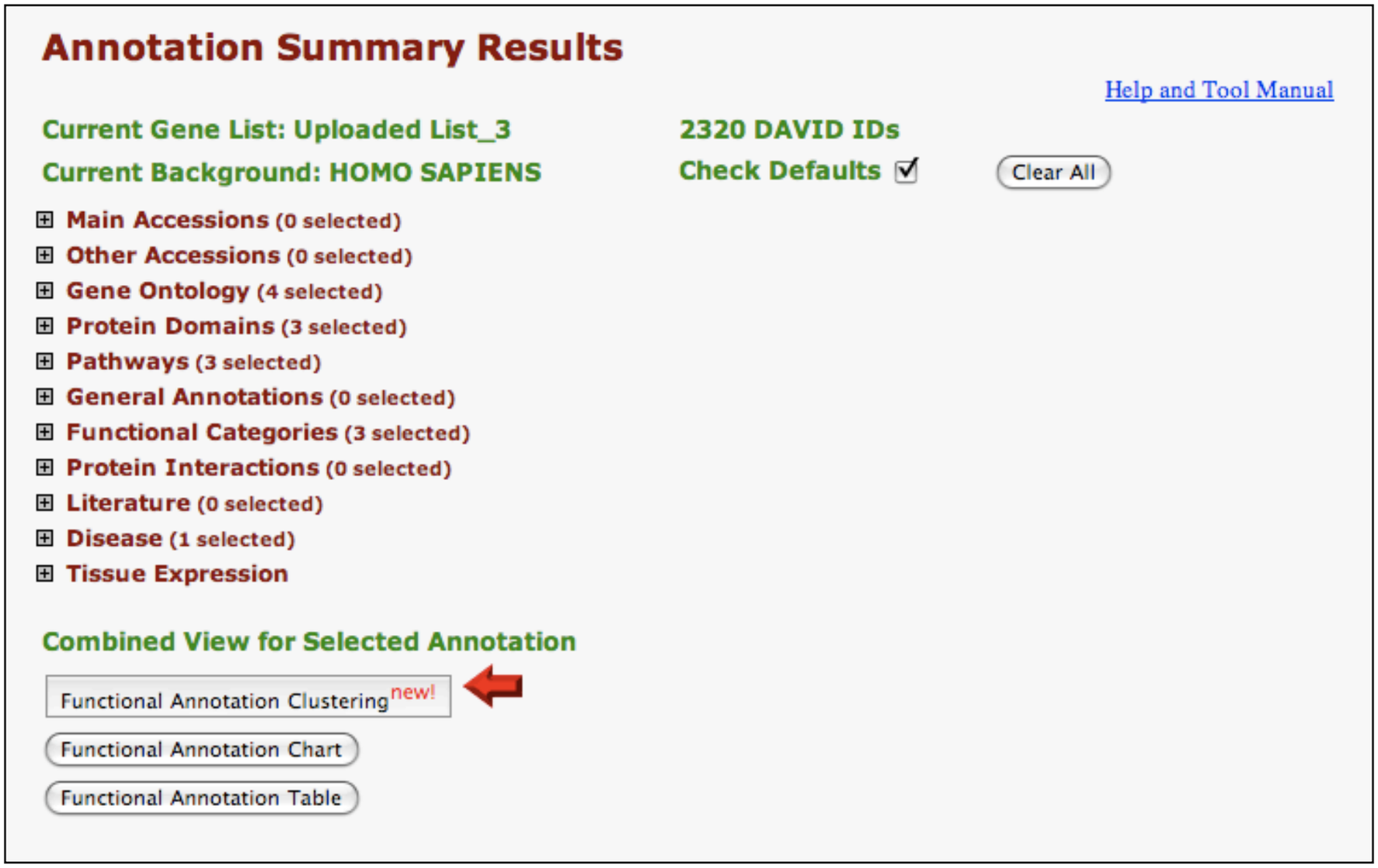
# DAVID

- Notice that you can pick a *Background* (Universe)



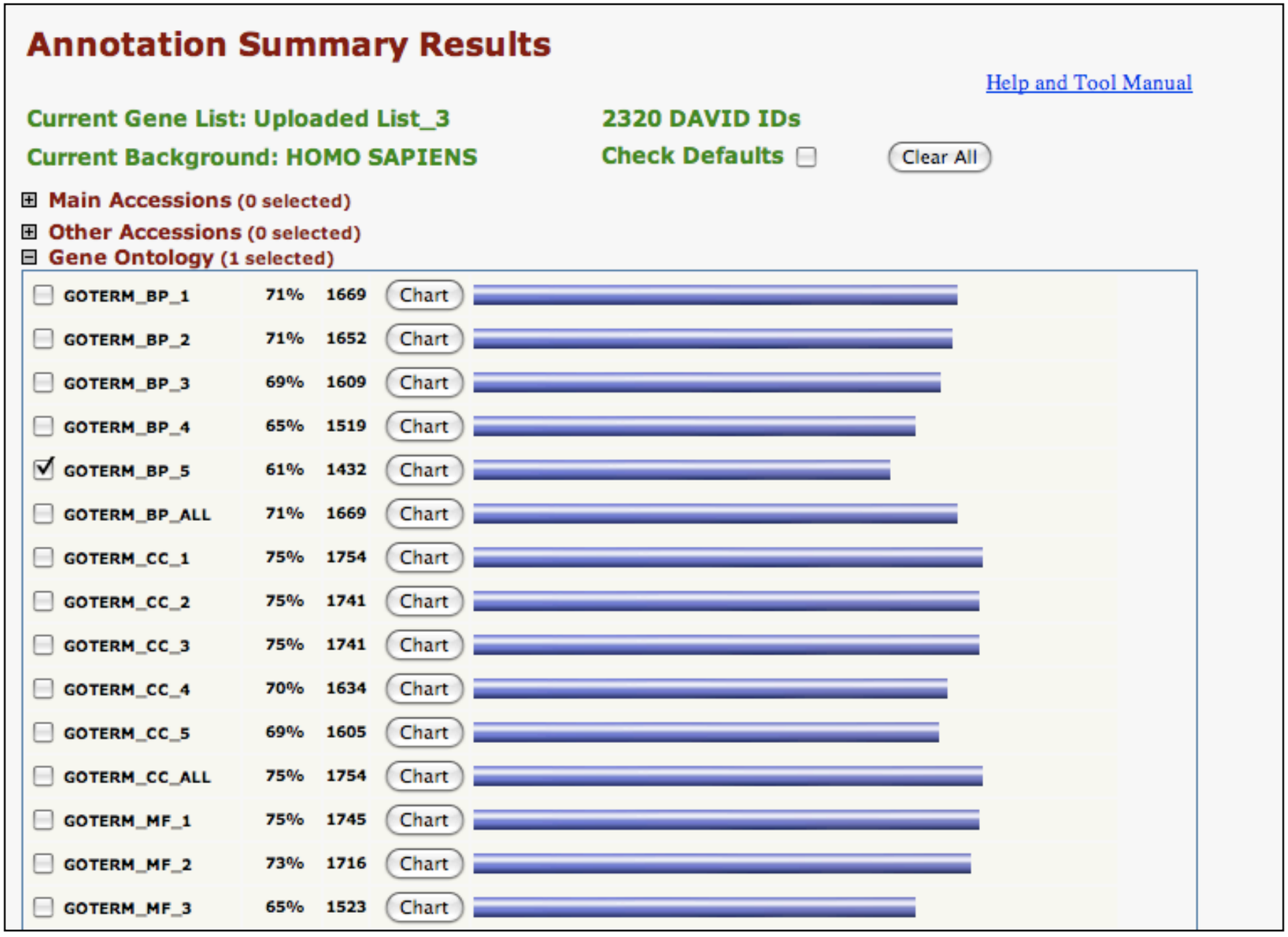
# DAVID

- *Functional Annotation Tool*



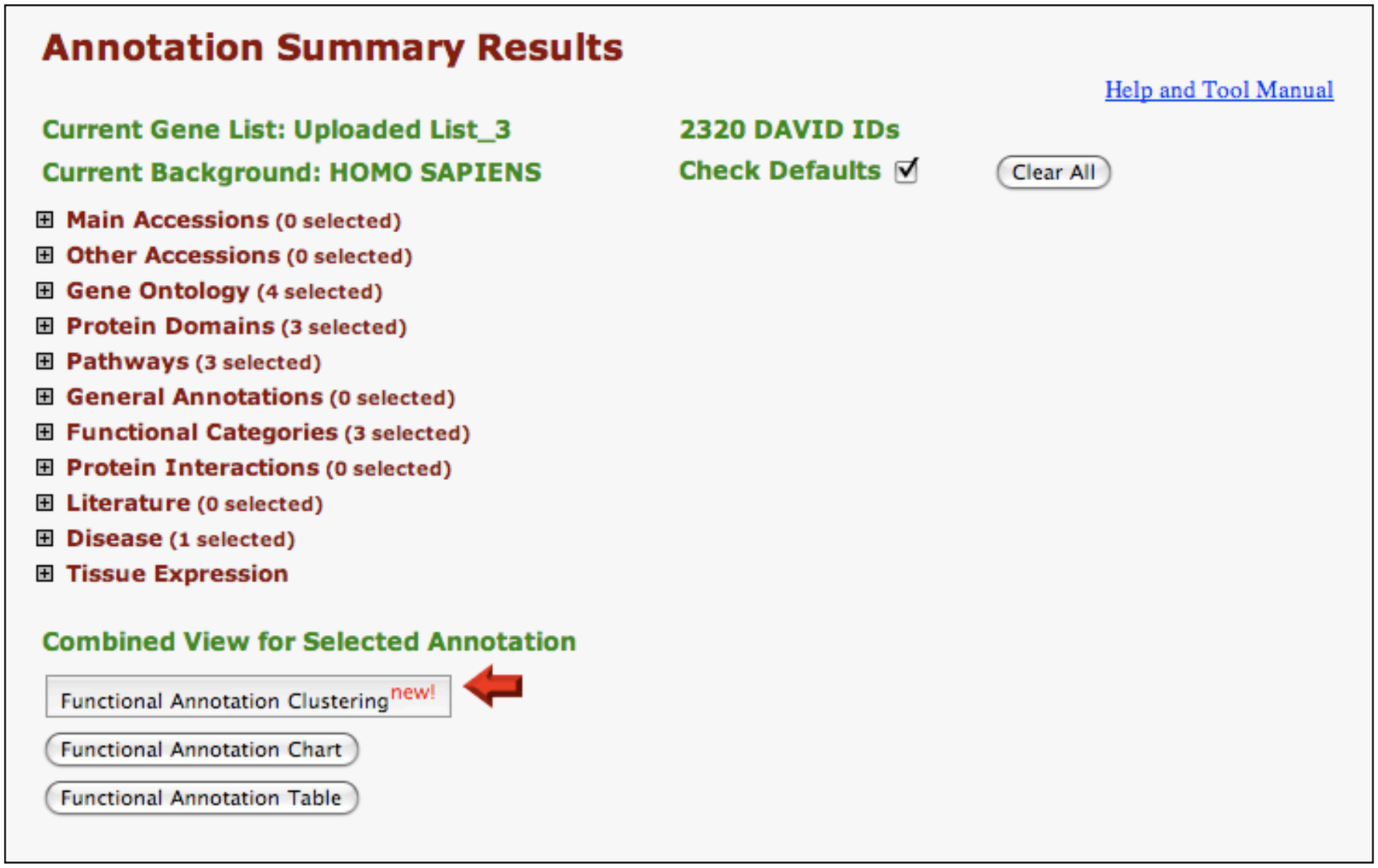
# DAVID

- Specify functional sets



# DAVID

- Let's look at the *Functional Annotation Chart*



# DAVID

- *Functional Annotation Chart*

## Functional Annotation Chart

Help and Manual

**Current Gene List: Uploaded List_1**
**Current Background: Homo sapiens**
**2316 DAVID IDs**

⊞ **Options**

Rerun Using Options | Create Sublist

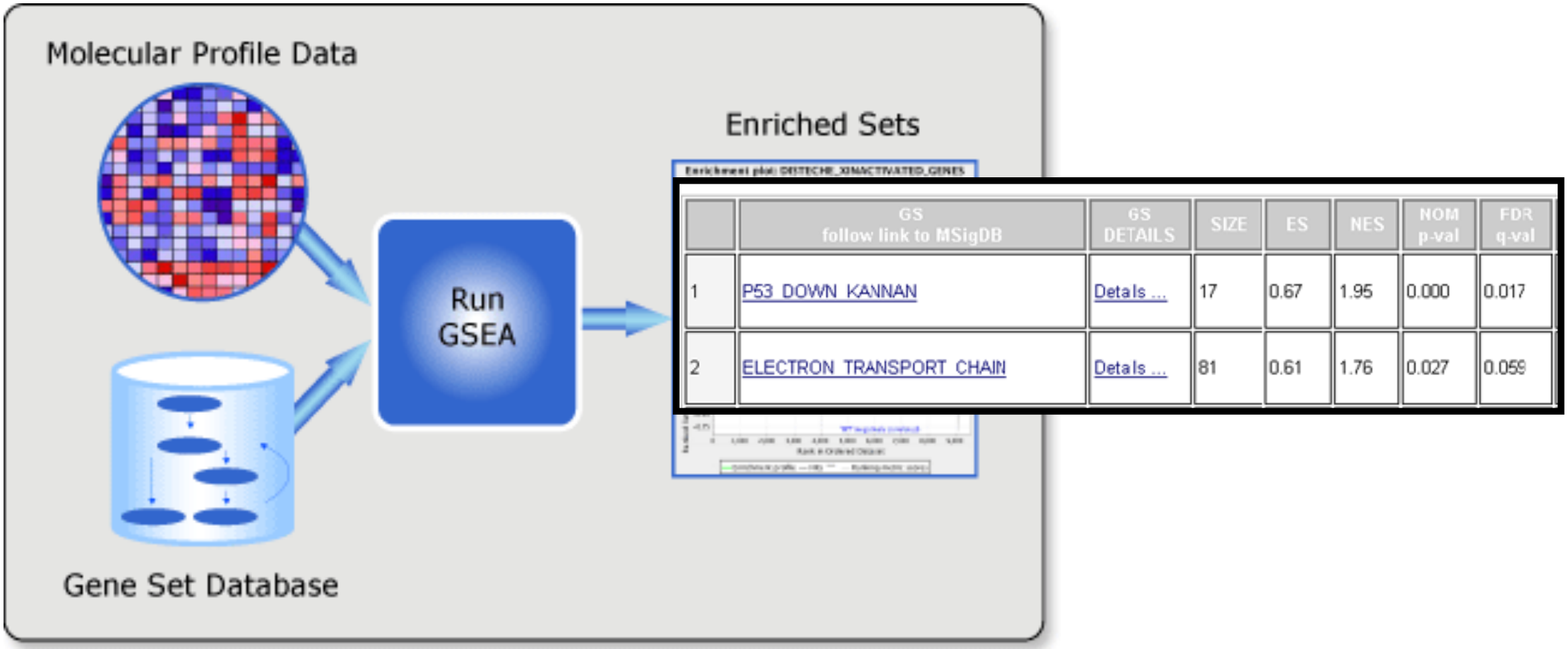
**Download File**

| Sublist | Category | Term | RT | Genes | Count | % | P-Value | Benjamini |
|---|---|---|---|---|---|---|---|---|
| ☐ | GOTERM_BP_5 | regulation of progression through cell cycle | RT | | 98 | 4.2 | 3.3E-7 | 8.6E-4 |
| ☐ | GOTERM_BP_5 | apoptosis | RT | | 131 | 5.7 | 1.6E-6 | 2.1E-3 |
| ☐ | GOTERM_BP_5 | cell death | RT | | 136 | 5.9 | 3.8E-6 | 3.3E-3 |
| ☐ | GOTERM_BP_5 | regulation of transcription from RNA polymerase II promoter | RT | | 83 | 3.6 | 3.7E-5 | 2.4E-2 |
| ☐ | GOTERM_BP_5 | protein kinase cascade | RT | | 71 | 3.1 | 4.7E-5 | 2.4E-2 |
| ☐ | GOTERM_BP_5 | regulation of kinase activity | RT | | 48 | 2.1 | 5.4E-5 | 2.3E-2 |
| ☐ | GOTERM_BP_5 | negative regulation of cell proliferation | RT | | 48 | 2.1 | 1.0E-4 | 3.7E-2 |
| ☐ | GOTERM_BP_5 | regulation of cell size | RT | | 41 | 1.8 | 1.2E-4 | 3.9E-2 |
| ☐ | GOTERM_BP_5 | monocarboxylic acid metabolic process | RT | | 48 | 2.1 | 1.3E-4 | 3.6E-2 |
| ☐ | GOTERM_BP_5 | positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process | RT | | 61 | 2.6 | 1.5E-4 | 3.8E-2 |
| ☐ | GOTERM_BP_5 | positive regulation of cellular metabolic process | RT | | 72 | 3.1 | 1.7E-4 | 3.8E-2 |

**Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**
Da Wei Huang, Brad T Sherman & Richard A Lempicki
*Nature Protocols* **4**, *44 - 57 (2009)*

# GSEA < www.broadinstitute.org/gsea >

- Download GSEA desktop application




| | GS follow link to MSigDB | GS DETAILS | SIZE | ES | NES | NOM p-val | FDR q-val |
|---|---|---|---|---|---|---|---|
| 1 | P53_DOWN_KANNAN | Details ... | 17 | 0.67 | 1.95 | 0.000 | 0.017 |
| 2 | ELECTRON_TRANSPORT_CHAIN | Details ... | 81 | 0.61 | 1.76 | 0.027 | 0.059 |

- Excellent tutorial, user's guide and example datasets to work through

**Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**
Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, ...
*PNAS **102**, 15545-15550 (2005)*

# Overlapping functional sets

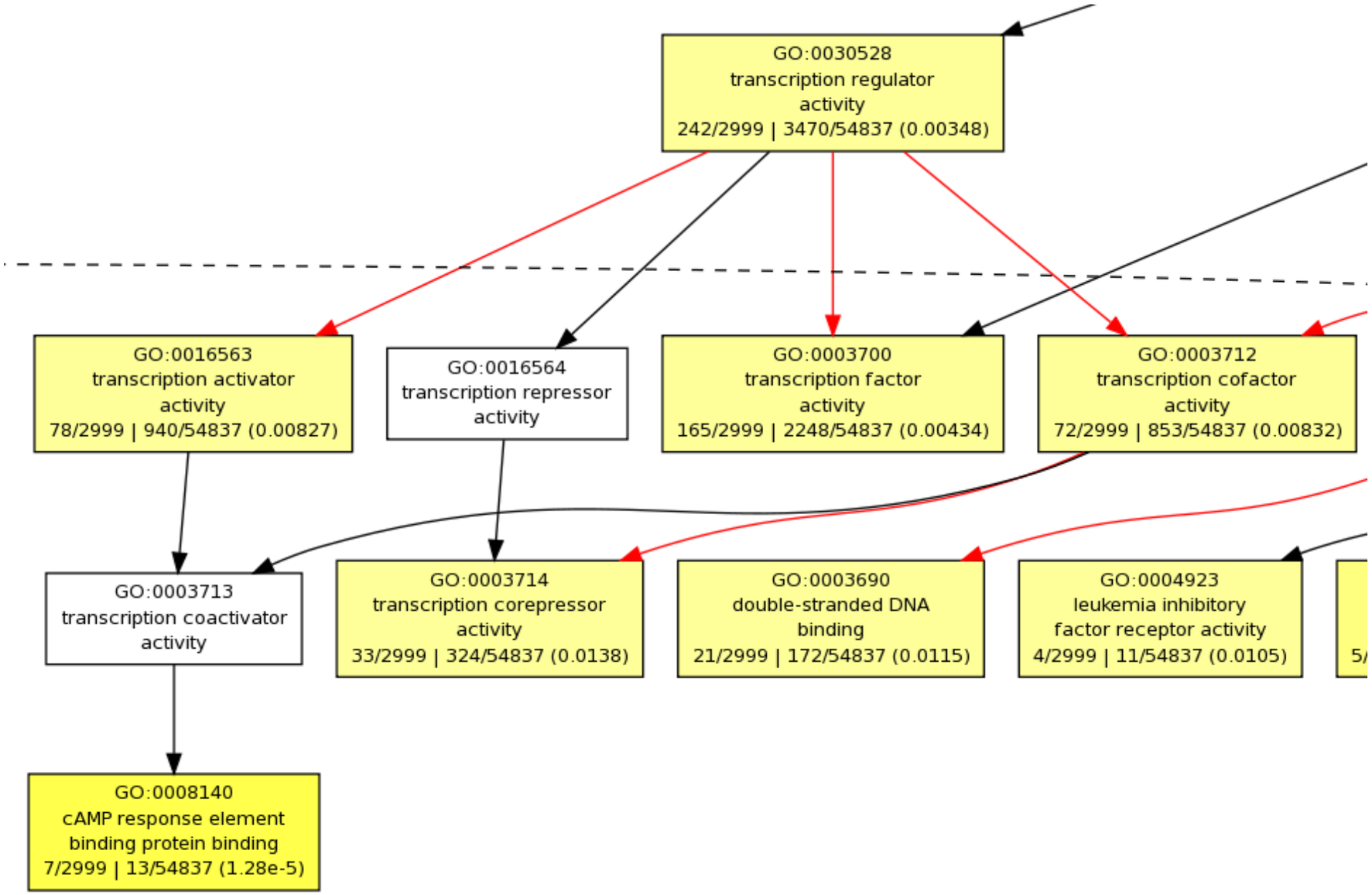
- Many functional sets overlap, in particular those from databases that are hierarchical in nature (e.g. GO)

- Hierarchy enables:
  - Annotation flexibility (e.g. allow different degrees of annotation completeness based on what is known)
  - Computational methods to “understand” function relationships (e.g. ATPase function is a subset of enzyme function)

- Unfortunately, this also makes functional profiling trickier

# **GOEast** < omicslab.genetics.ac.cn/GOEAST >

- Graphical view of enriched GO terms and their relationships



# GO SLIMs

- Cut-down versions of the GO ontologies containing a subset of the terms in the whole GO
- GO FAT (DAVID):
    - filters out very broad GO terms based on a measured specificity of each term

# DAVID Functional Annotation Clustering

- Based on shared genes between functional sets

## Functional Annotation Clustering

Help and Manual

**Current Gene List: Uploaded List_3**
**2320 DAVID IDs**

⊞ **Options** **Classification Stringency** Medium

Rerun using options | Create Sublist

**Download File**

| | Annotation Cluster 1 | Enrichment Score: 3.72 | | | Count | P_Value | Benjamini |
|---|---|---|---|---|---|---|---|
| ☐ | GOTERM_BP_5 | regulation of transcription from RNA polymerase II promoter | RT | | 83 | 3.7E-5 | 2.4E-2 |
| ☐ | GOTERM_BP_5 | positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process | RT | | 61 | 1.5E-4 | 3.8E-2 |
| ☐ | GOTERM_BP_5 | positive regulation of cellular metabolic process | RT | | 72 | 1.7E-4 | 3.8E-2 |
| ☐ | GOTERM_BP_5 | positive regulation of transcription | RT | | 58 | 3.8E-4 | 5.0E-2 |
| ☐ | GOTERM_BP_5 | positive regulation of transcription, DNA-dependent | RT | | 48 | 7.4E-4 | 7.6E-2 |
| | **Annotation Cluster 2** | **Enrichment Score: 3.54** | | | **Count** | **P_Value** | **Benjamini** |
| ☐ | GOTERM_BP_5 | regulation of cell size | RT | | 41 | 1.2E-4 | 3.9E-2 |
| ☐ | GOTERM_BP_5 | regulation of cell growth | RT | | 33 | 3.7E-4 | 5.1E-2 |
| ☐ | GOTERM_BP_5 | cell morphogenesis | RT | | 81 | 5.2E-4 | 5.7E-2 |
| | **Annotation Cluster 3** | **Enrichment Score: 3.37** | | | **Count** | **P_Value** | **Benjamini** |
| ☐ | GOTERM_BP_5 | apoptosis | RT | | 131 | 1.6E-6 | 2.1E-3 |
| ☐ | GOTERM_BP_5 | cell death | RT | | 136 | 3.8E-6 | 3.3E-3 |
| ☐ | GOTERM_BP_5 | regulation of programmed cell death | RT | | 88 | 3.2E-4 | 5.8E-2 |
| ☐ | GOTERM_BP_5 | positive regulation of apoptosis | RT | | 48 | 3.3E-4 | 5.6E-2 |
| ☐ | GOTERM_BP_5 | regulation of apoptosis | RT | | 87 | 3.5E-4 | 5.2E-2 |
| ☐ | GOTERM_BP_5 | positive regulation of programmed cell death | RT | | 48 | 4.0E-4 | 5.0E-2 |

# Want more?

- **GeneGO** < portal.genego.com >
  - MD/PhD curated annotations, great for certain domains (eg, Cystic Fibrosis)
  - Nice network analysis tools
  - Email us for access
- **Oncomine** < www.oncomine.org >
  - Extensive cancer related expression datasets
  - Nice concept analysis tools
  - Research edition is free for academics, Premium edition $$$
- **Lots of other Bioconductor packages in this area!**

Do it Yourself!

# Hands-on time!

https://tinyurl.com/bggn213-pathways

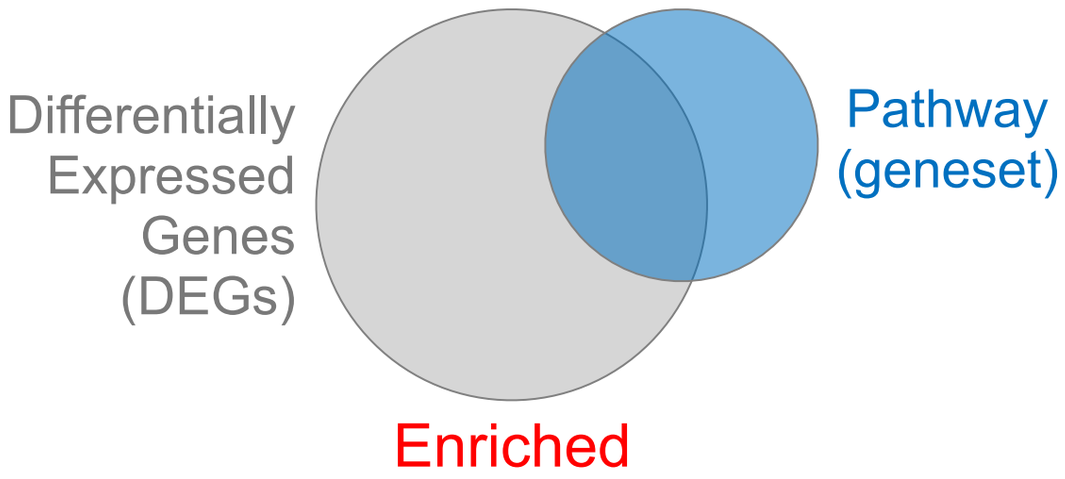
# Advice:
# Figure out “**What do I want to do with my list?**”

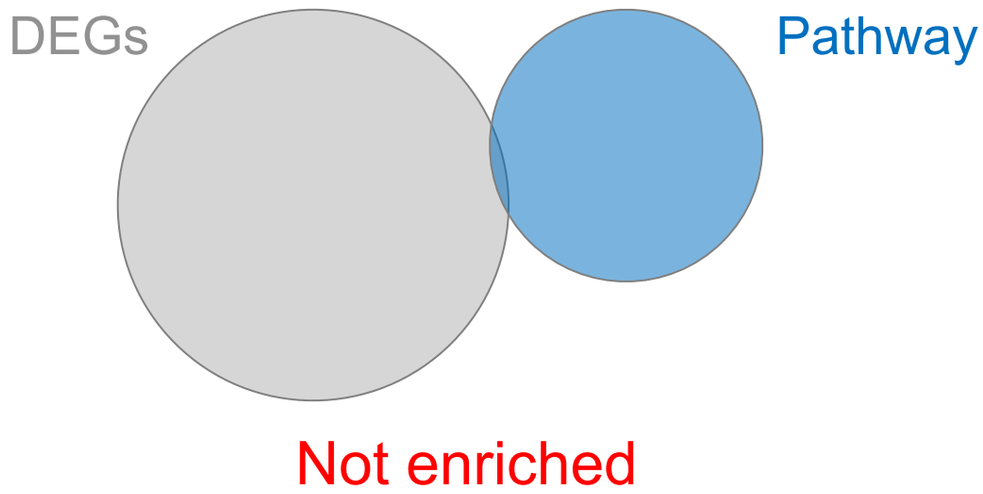
- Organize/summarize data for presentation or manuscript
  - DAVID: GO_FAT -> Functional Annotation Clustering -> Pick threshold

- Infer biological processes from the list
  - DAVID: Functional Annotation Chart -> explore functional databases and see which make sense
  - GSEA: Select MSigDB sets of interest -> e.g., immunologic signatures
  - Use domain specific database it at all possible!

- Find “missing” genes/proteins not detected by experiment
  - ConceptGen: Gene-gene enrichment

# Pathway analysis (a.k.a. geneset enrichment)
# **Principle**




- Variations of the math: overlap, ranking, networks... ➢ *Not critical, different algorithms show similar performances*
- DEGs come from your experiment ➢ *Critical, needs to be as clean as possible*
- Pathway genes (“geneset”) come from annotations ➢ *Important, but typically not a competitive advantage*

# Pathway analysis (a.k.a. geneset enrichment) **Limitations**

---

- **Post-transcriptional regulation** is neglected
- **Directionality** is hard to capture sensibly
    - e.g. IκBα/NF-κB
- **Tissue-specific** variations of pathways are not annotated
    - e.g. NF-κB regulates metabolism, not inflammation, in adipocytes
- **Size bias**: stats are influenced by the size of the pathway
- **Geneset annotation bias**: can only discover what is already known
- **Non-model organisms**: no high-quality genesets available
- Many pathways/receptors **converge** to few regulators
    - e.g. tens of innate immune receptors activate 4 TFs: NF-kB, AP-1, IRF3/7, NFAT