Gene Ontology

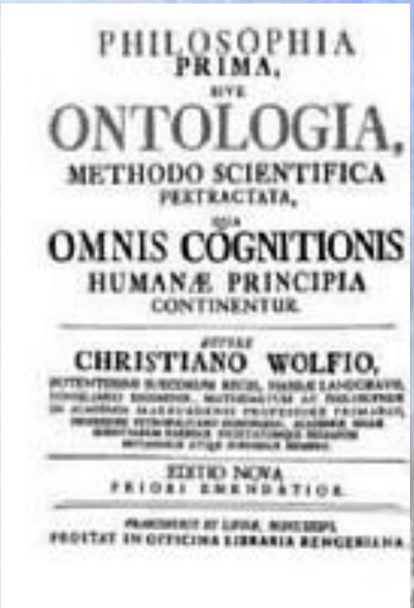
Shifra Ben-Dor

Weizmann Institute of Science

Outline

- What is GO (Gene Ontology)?
- What tools do we use to work with it?
- (Combination of GO with other analyses)

What is Ontology?



PHILOSOPHIA
PRIMA,
SIVE
ONTOLOGIA,
METHODO SCIENTIFICA
PERTRACTATA,
QUA
OMNIS COGNITIONIS
HUMANÆ PRINCIPIA
CONTINENTUR.

AUCTORE
CHRISTIANO WOLFIO,
[illegible]

EDITIO NOVA
PRIORI EMENDATIOR.

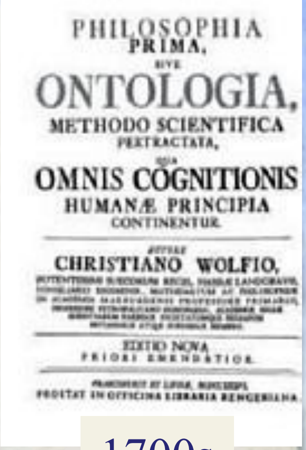
FRANCOFURTI ET LIPSIÆ, MDCCXXXVI.
PROSTAT IN OFFICINA LIBRARIA RENGERIANA.

1700s

Oxford English Dictionary

1. a. Philos. The science or study of being; that branch of metaphysics concerned with the nature or essence of being or existence.

What is Ontology?



PHILOSOPHIA
PRIMA,
SIVE
ONTOLOGIA,
METHODO SCIENTIFICA
PERTRACTATA,
QUA
OMNIS COGNITIONIS
HUMANÆ PRINCIPIA
CONTINENTUR.

AUCTORE
CHRISTIANO WOLFIO,
[illegible]

EDITIO NOVA
PRIORI EMENDATIOR.

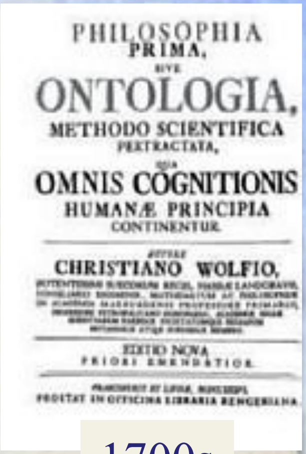
FRANCOFURTI ET LIPSIAE, [illegible]
PROSTAT IN OFFICINA LIBRARIA RENGERIANA.

1700s

WIKIPEDIA
The Free Encyclopedia

Ontology (from the Greek…) is the philosophical study of the nature of being, existence or reality in general, as well as of the basic categories of being and their relations. Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences.

What is Ontology?



PHILOSOPHIA
PRIMA,
SIVE
ONTOLOGIA,
METHODO SCIENTIFICA
PERTRACTATA,
QUA
OMNIS COGNITIONIS
HUMANÆ PRINCIPIA
CONTINENTUR.

AUCTORE
CHRISTIANO WOLFIO,
[illegible]

EDITIO NOVA
PRIORI EMENDATIOR.

FRANCOFURTI ET LIPSIAE, MDCCXXXVI.
PROSTAT IN OFFICINA LIBRARIA RENGERIANA.

1700s

Ontology (from the Greek…) is the philosophical study of the nature of being, existence or reality in general, as well as of the basic categories of being and their relations. Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences.

What is a Gene ?

So what is Gene Ontology?

- Unfortunately, not an ontology of genes, but rather of gene products

- It is an attempt to classify gene products using a structured language (controlled vocabulary) to give a consistent description of characteristics inherent to them.

GENEONTOLOGY
Unifying Biology

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases.

The project provides the controlled vocabulary of terms and gene product annotations from consortium members.

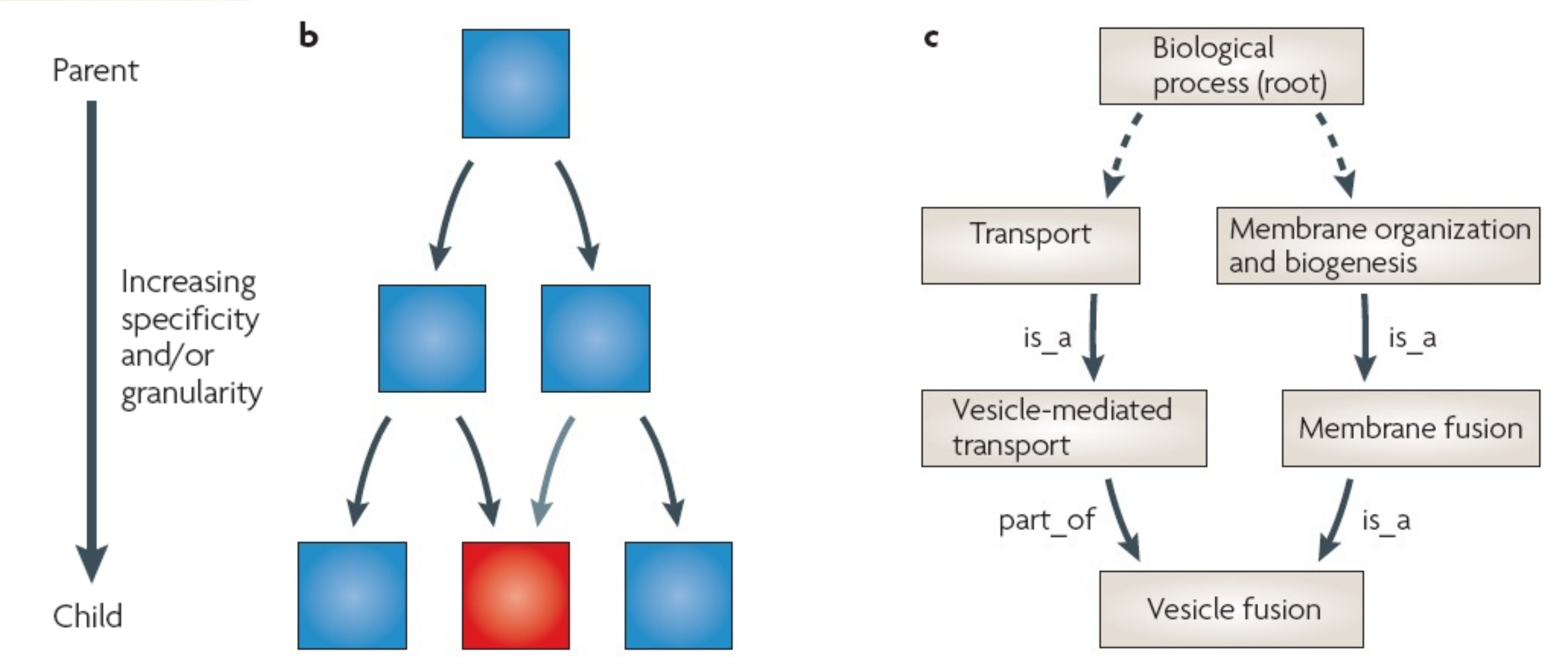
GENEONTOLOGY
Unifying Biology

- Gene ontology is an annotation system which tries to describe attributes of gene products (what does it do? where? how?)

- It represents a unified consistent system, i.e. terms occur only once, and there is a dictionary of allowed words, which is consistent across species

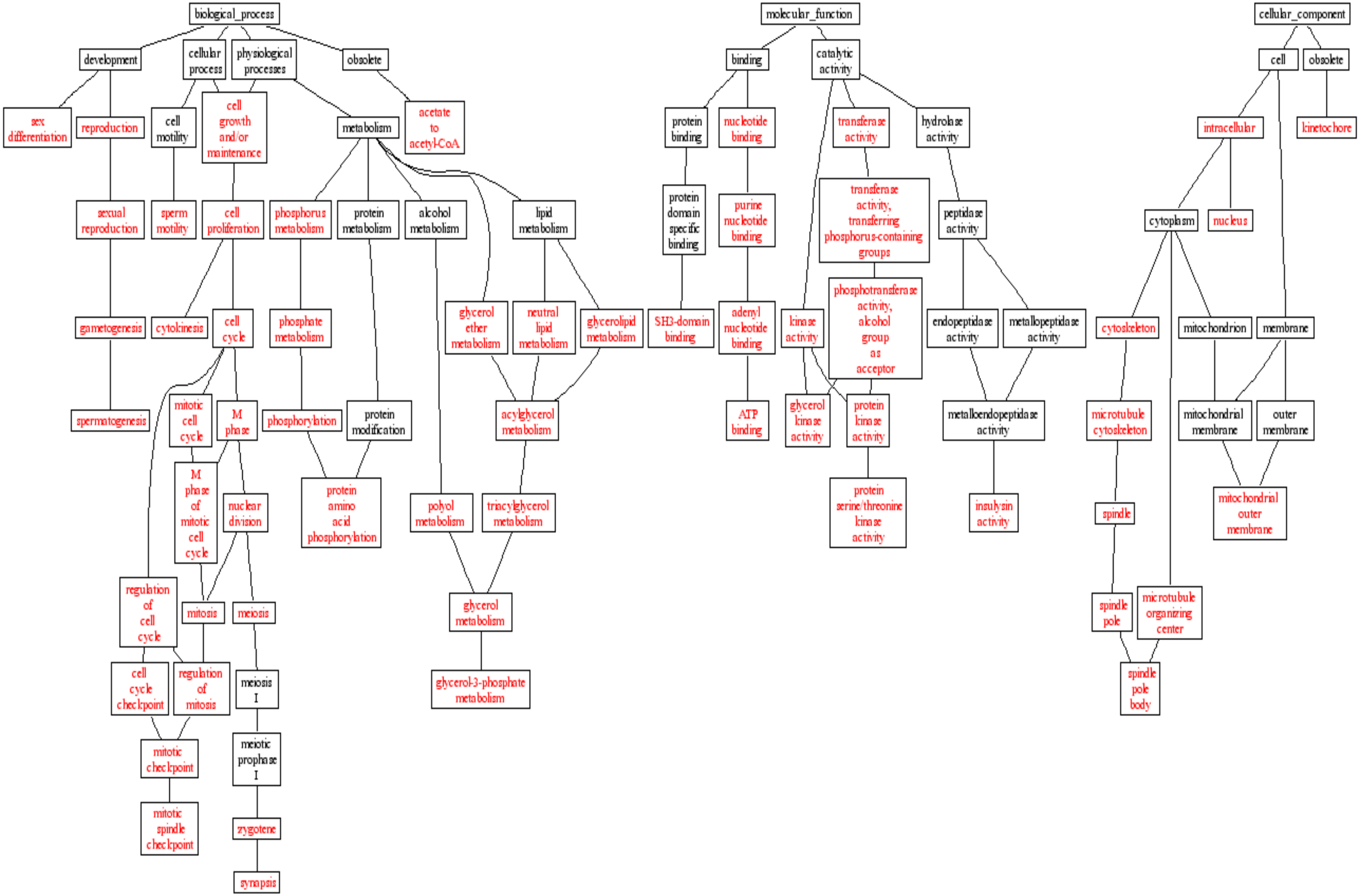
- Furthermore, terms are related to each other: the hierarchy goes from very general terms to very detailed ones

Gene ontology is represented as a directed acyclic graph (DAG)


Parent
Increasing specificity and/or granularity
Child
b
c
Biological process (root)
Transport
Membrane organization and biogenesis
is_a
is_a
Vesicle-mediated transport
Membrane fusion
part_of
is_a
Vesicle fusion

Taken from: Nature Reviews Genetics 9:509-515 (2008)

- A child can have more than one parent (parents are closer to the root and are more general, children are further from the root and more specific)
- There are no cycles - there is a root
- It is a directed graph
- You can skip levels in the graph



biological_process
development
cellular process
physiological processes
obsolete
sex differentiation
reproduction
cell motility
cell growth and/or maintenance
metabolism
acetate to acetyl-CoA
sexual reproduction
sperm motility
cell proliferation
phosphorus metabolism
protein metabolism
alcohol metabolism
lipid metabolism
gametogenesis
cytokinesis
cell cycle
phosphate metabolism
glycerol ether metabolism
neutral lipid metabolism
glycerolipid metabolism
spermatogenesis
mitotic cell cycle
M phase
phosphorylation
protein modification
acylglycerol metabolism
M phase of mitotic cell cycle
nuclear division
protein amino acid phosphorylation
polyol metabolism
triacylglycerol metabolism
regulation of cell cycle
mitosis
meiosis
glycerol metabolism
cell cycle checkpoint
regulation of mitosis
meiosis I
glycerol-3-phosphate metabolism
mitotic checkpoint
meiotic prophase I
mitotic spindle checkpoint
zygotene
synapsis

molecular_function
binding
catalytic activity
protein binding
nucleotide binding
transferase activity
hydrolase activity
protein domain specific binding
purine nucleotide binding
transferase activity, transferring phosphorus-containing groups
peptidase activity
SH3-domain binding
adenyl nucleotide binding
kinase activity
phosphotransferase activity, alcohol group as acceptor
endopeptidase activity
metallopeptidase activity
ATP binding
glycerol kinase activity
protein kinase activity
metalloendopeptidase activity
protein serine/threonine kinase activity
insulysin activity

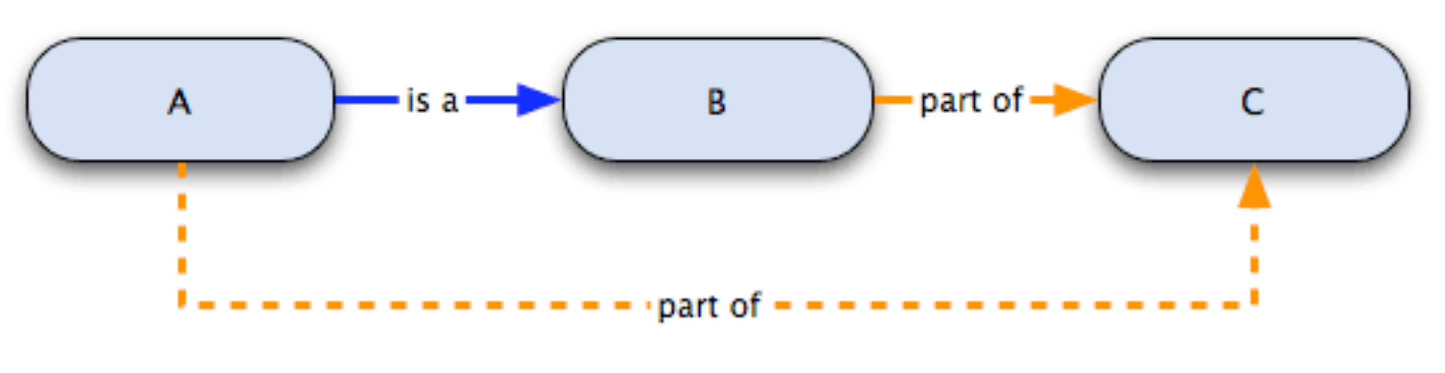
cellular_component
cell
obsolete
intracellular
kinetochore
cytoplasm
nucleus
cytoskeleton
mitochondrion
membrane
microtubule cytoskeleton
mitochondrial membrane
outer membrane
spindle
mitochondrial outer membrane
spindle pole
microtubule organizing center
spindle pole body

Ontology Relations

- Just as the ontology terms are defined, so are the relationships between them (the arrows). The terms are linked by three relationships:
  - is_a
  - part_of
  - regulates, positively regulates, negatively regulates

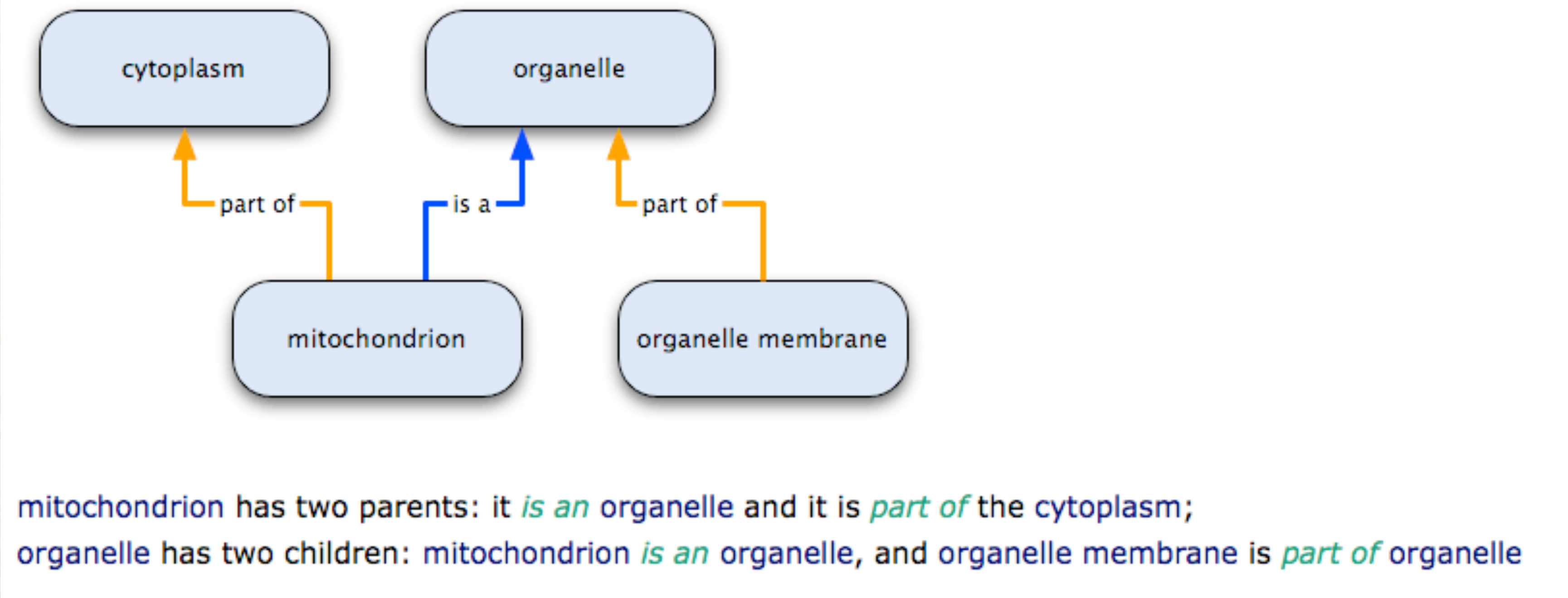
Ontology Relations

- is_a is a simple class-subclass relationship, for example, nuclear chromosome is_a chromosome.
- part_of is slightly more complex; C part_of D means that whenever C is present, it is always a part of D. An example would be nucleus part_of cell; nuclei are always part of a cell, but not all cells have nuclei.



A dotted line means an inferred relationship, e.g. one that has not been expressly stated

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

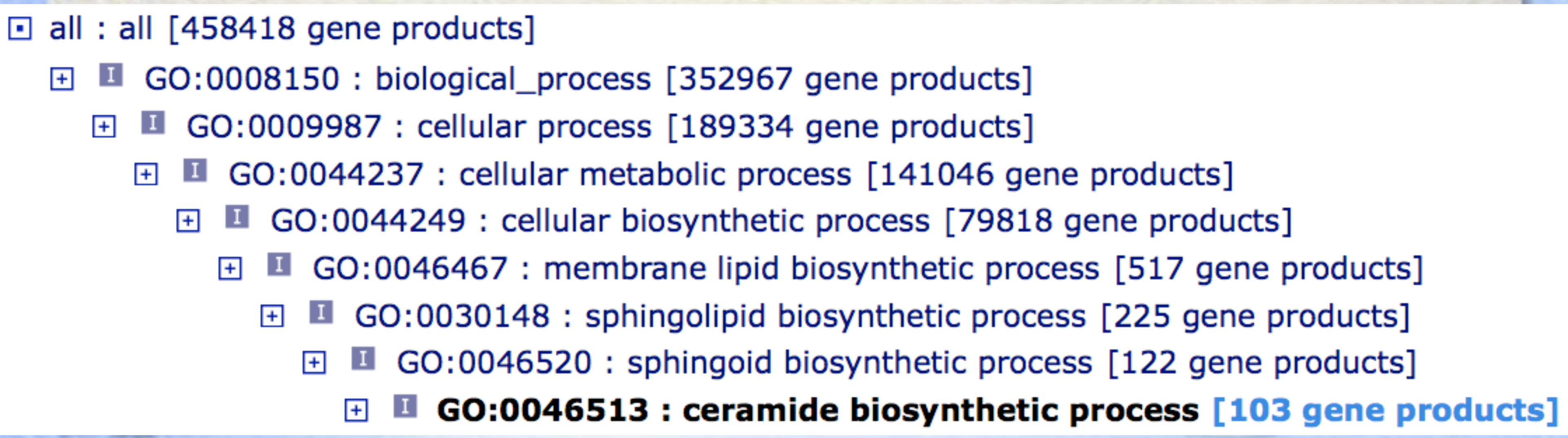


mitochondrion has two parents: it *is an* organelle and it is *part of* the cytoplasm;
organelle has two children: mitochondrion *is an* organelle, and organelle membrane is *part of* organelle

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

Ontology Structure

Every GO term must obey “the true path rule”: if the child term describes the gene product, then all its parent terms must also apply to that gene product.



⊡ all : all [458418 gene products]

⊞ I GO:0008150 : biological_process [352967 gene products]

⊞ I GO:0009987 : cellular process [189334 gene products]

⊞ I GO:0044237 : cellular metabolic process [141046 gene products]

⊞ I GO:0044249 : cellular biosynthetic process [79818 gene products]

⊞ I GO:0046467 : membrane lipid biosynthetic process [517 gene products]

⊞ I GO:0030148 : sphingolipid biosynthetic process [225 gene products]

⊞ I GO:0046520 : sphingoid biosynthetic process [122 gene products]

⊞ I **GO:0046513 : ceramide biosynthetic process** [103 gene products]

GO has 3 major divisions (roots)

- Biological Process
- Molecular Function
- Cellular Component

Biological Process

A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are cellular physiological process or signal transduction. Examples of more specific terms are pyrimidine metabolic process or alpha-glucoside transport. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

Biological Process

A biological process is not equivalent to a pathway; at present, GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

Molecular Function

Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are catalytic activity, transporter activity, or binding; examples of narrower functional terms are adenylate cyclase activity or Toll receptor binding.

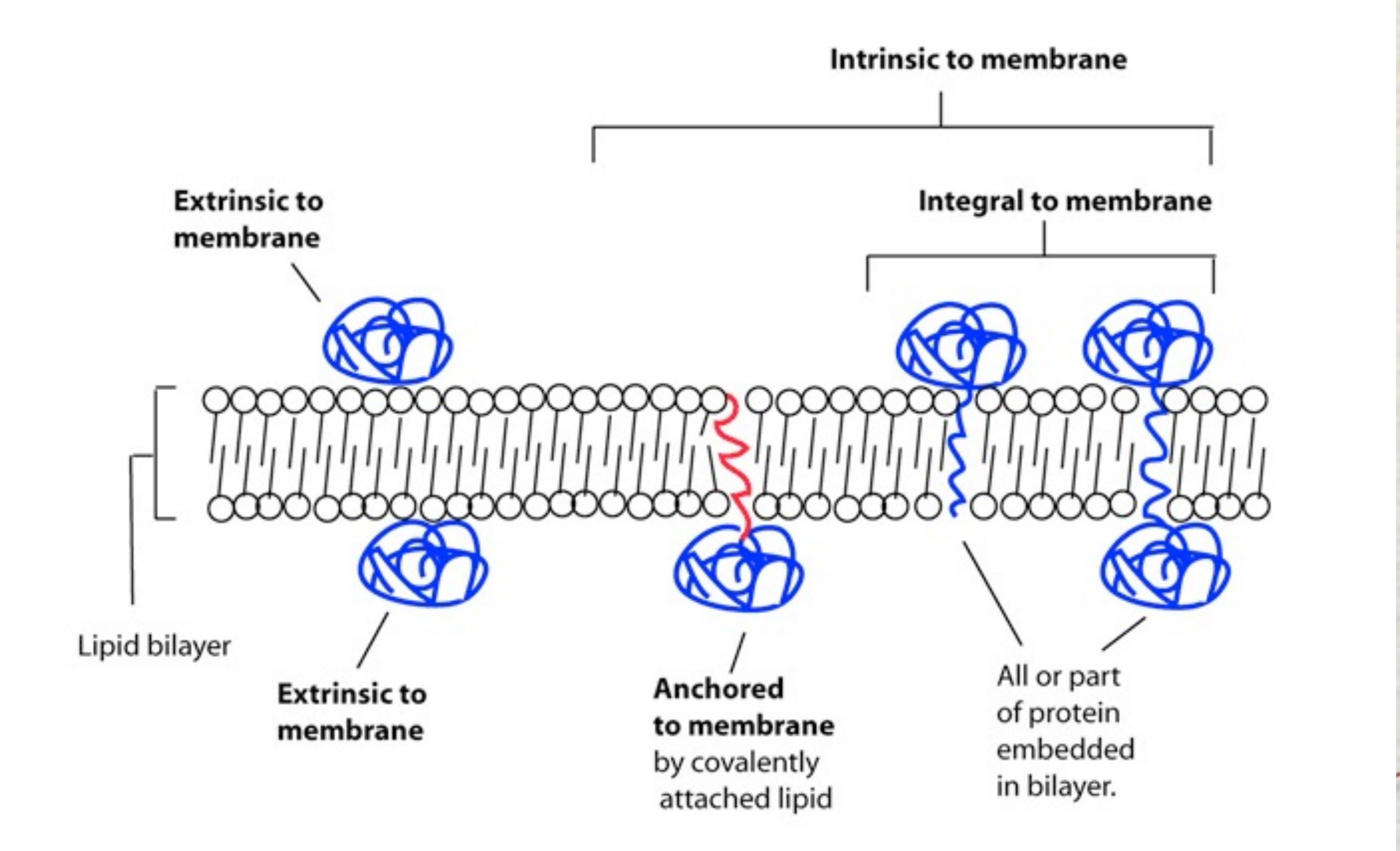
Molecular Function

It is easy to confuse a gene product name with its molecular function, and for that reason many GO molecular functions are appended with the word "activity".

Cellular Component

A cellular component is just that, a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

Cellular Component



Available GO Information

Current ontology statistics, as of June 15, 2022:

43,613	terms
28,199	biological_process
4,184	cellular_component
11,230	molecular_function
3718	obsolete terms (not counted above)

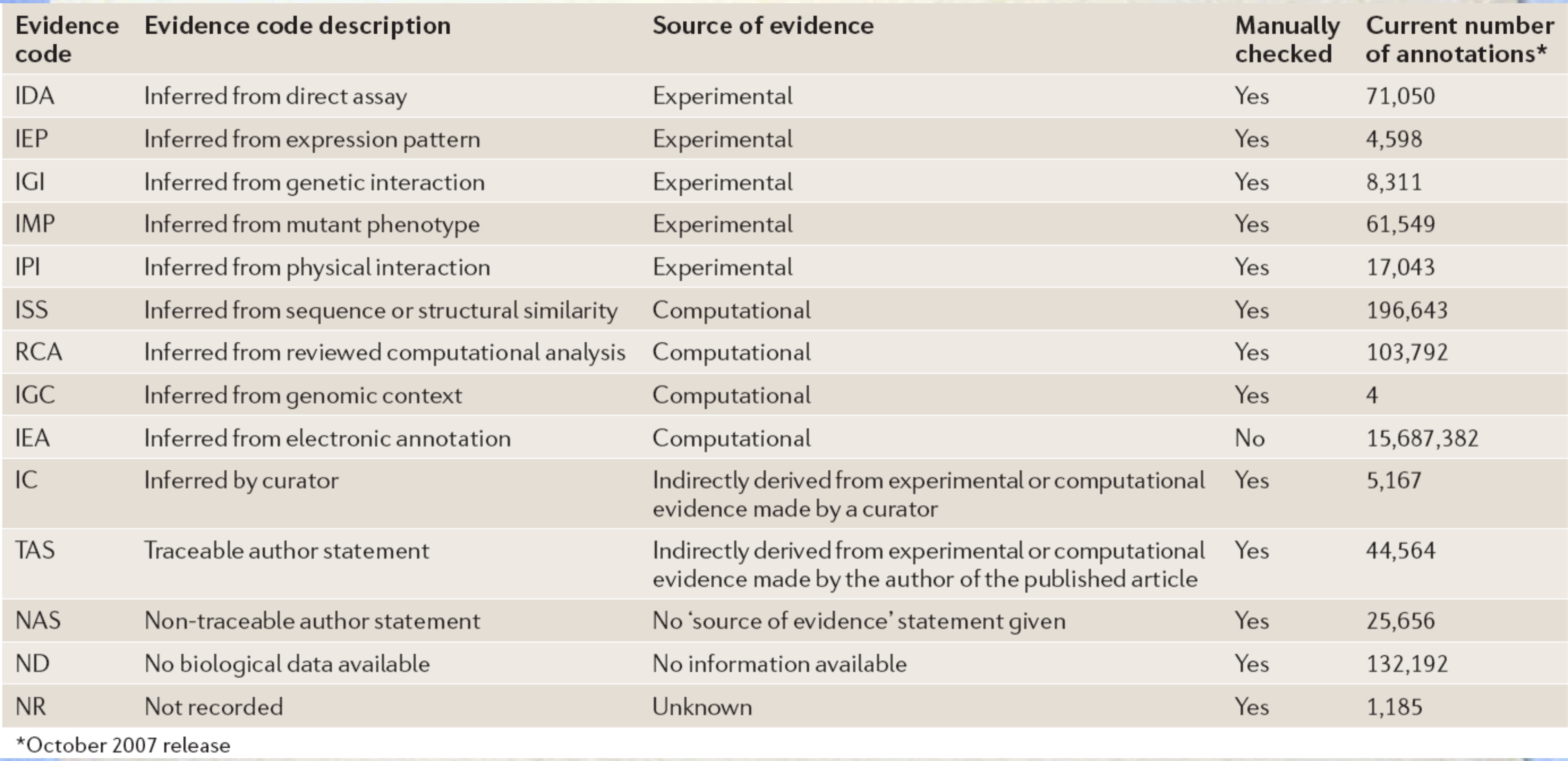
What is not GO?

- Gene products: e.g. cytochrome c is not in the ontologies, but attributes of cytochrome c, such as oxidoreductase activity, are
- Processes, functions or components that are unique to mutants or diseases: e.g. oncogenesis
- Attributes of sequence such as intron/exon parameters
- Protein domains or structural features
- Protein-protein interactions
- Environment, evolution and expression
- It is not **complete,** it is done “by hand” by curators

Annotation

- What connects the GO terms to specific gene products
- Annotation is carried out by curators in a range of bioinformatics database resource groups. These groups then contribute their data to the central GO repository for storage and redistribution.
- There are two general principles: first, annotations should be attributed to a source; second, each annotation should indicate the evidence on which it is based.

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

Evidence codes — not all annotations are created equal

Taken from: Nature Reviews Genetics 9:509-515 (2008)

GO Pitfalls

- Not complete
- Computational annotations
- NOT qualifier
- Splice variants
- Identifier flagged as ‘obsolete’

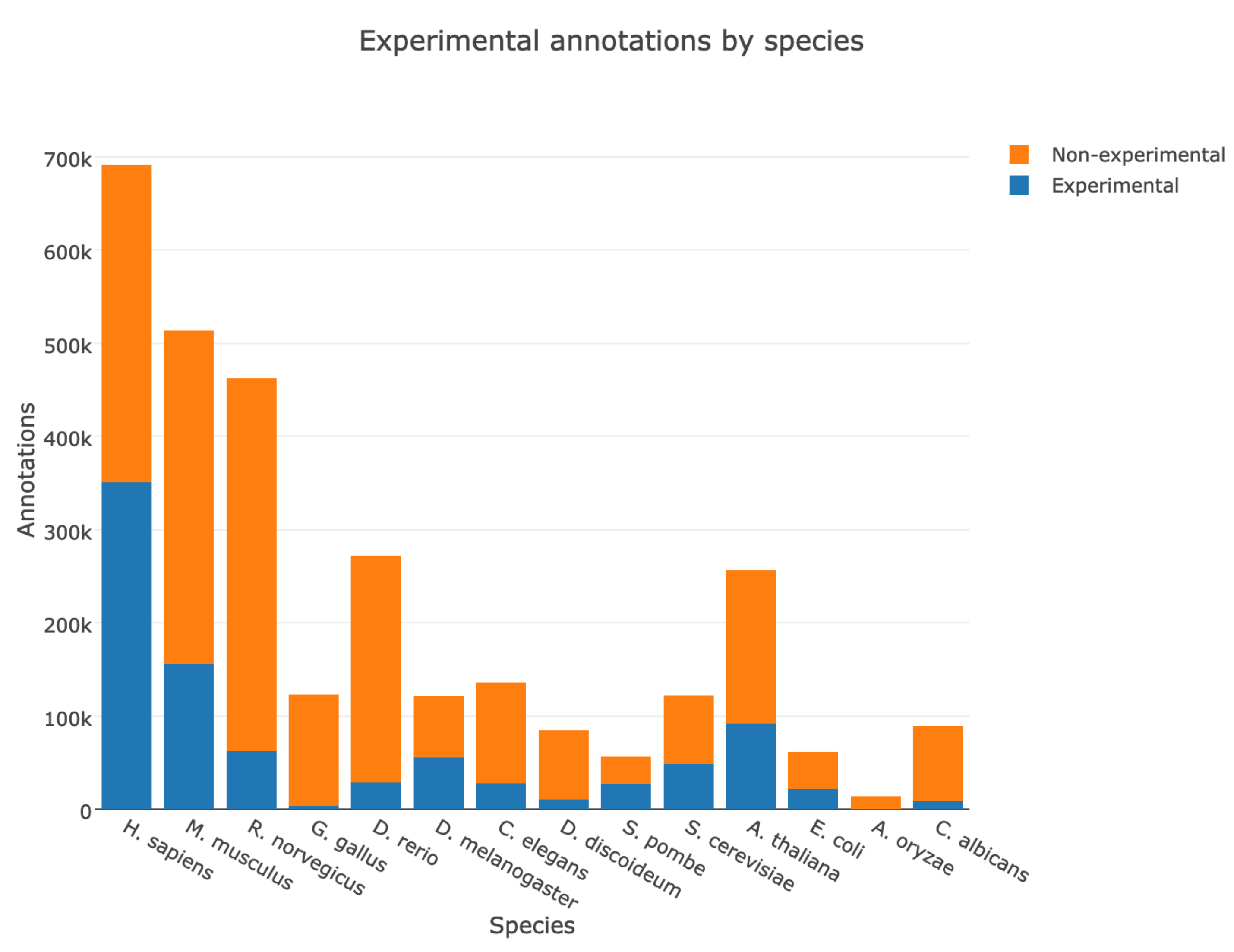
GO Pitfalls

- Not complete
- Computational annotations
- NOT qualifier
- Splice variants
- Identifier flagged as ‘obsolete’

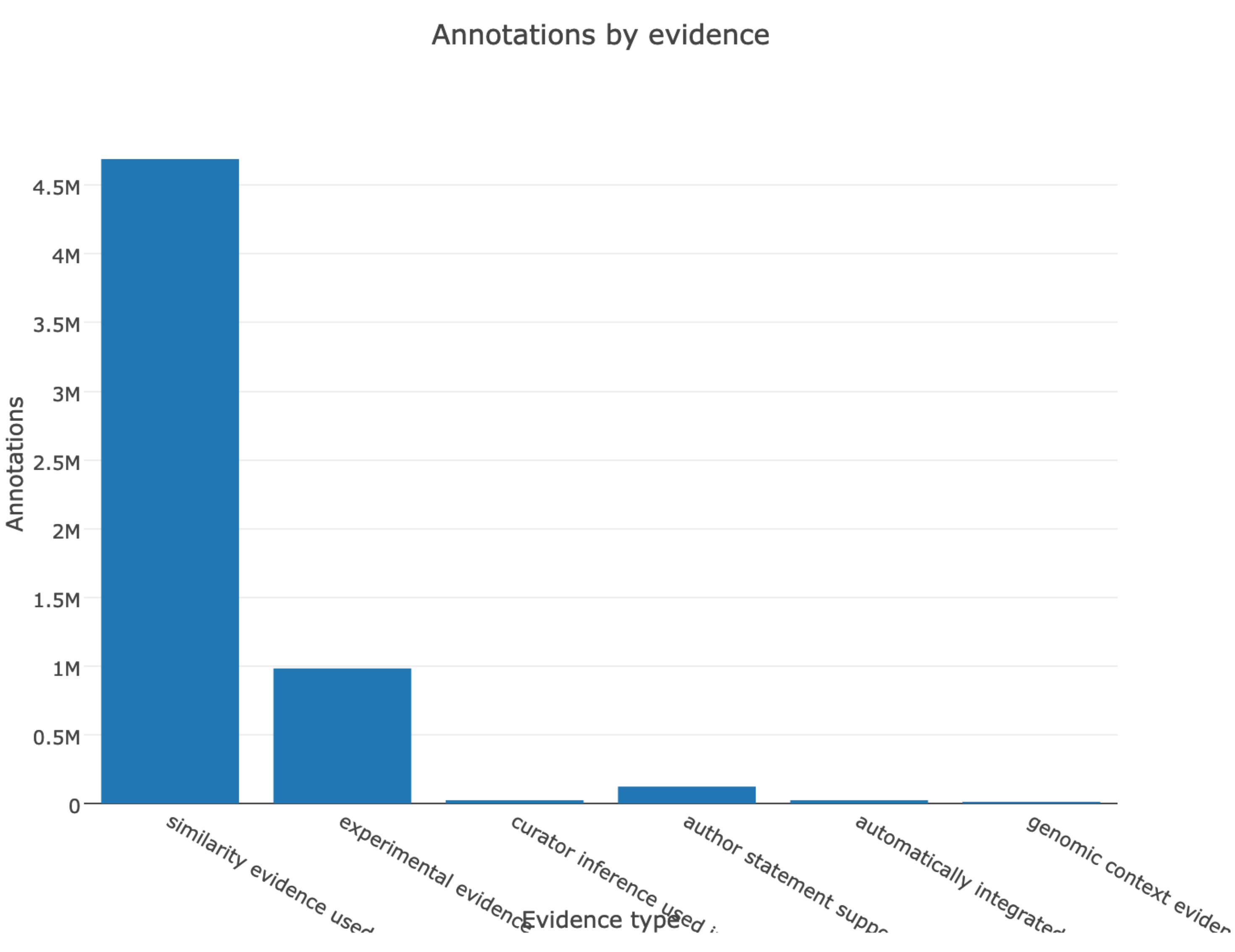
GO Pitfalls

- Not complete
- Computational annotations
- NOT qualifier
- Splice variants
- Identifier flagged as ‘obsolete’

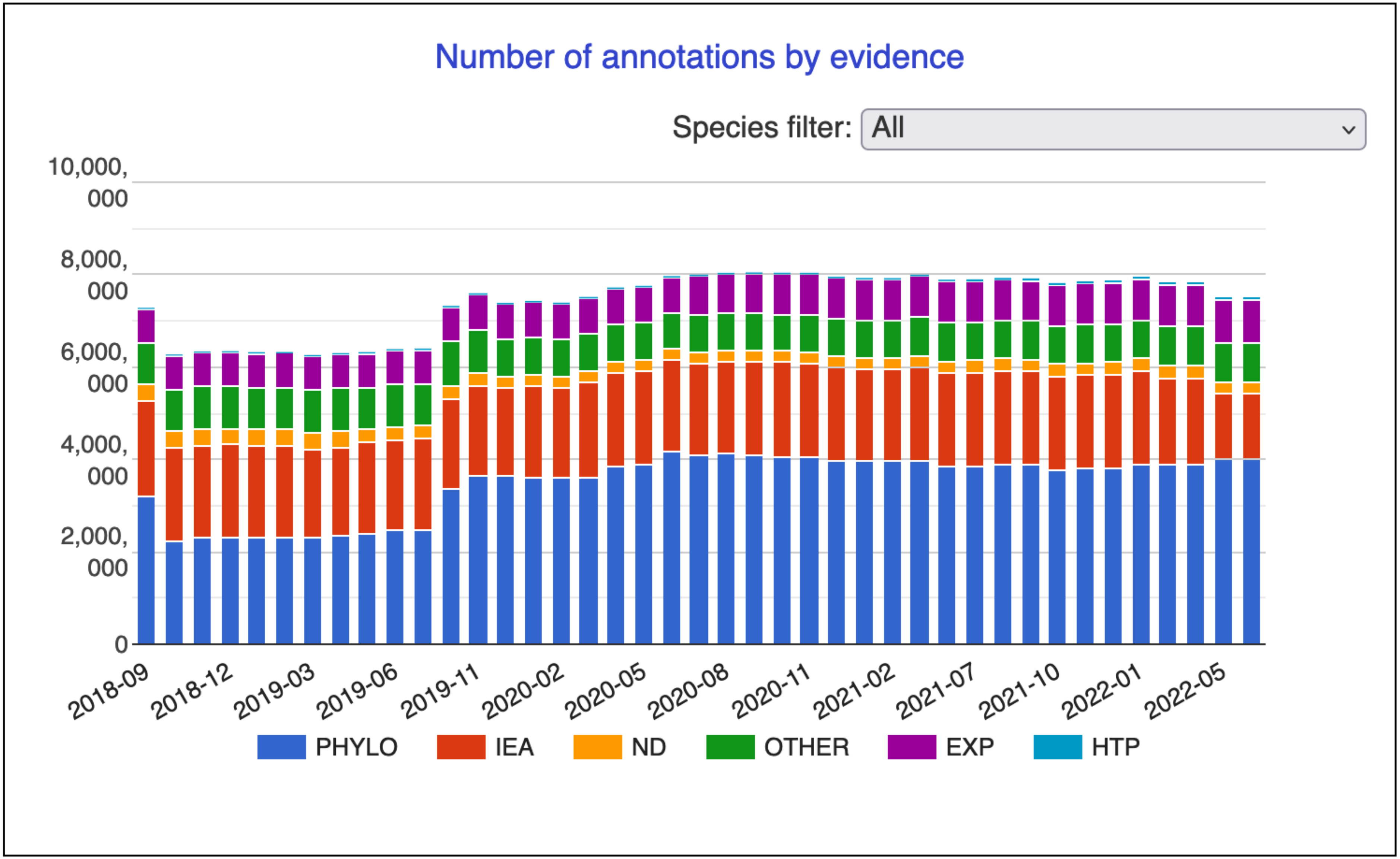
Type of annotation per species



Type of annotation per evidence



Type of annotation per evidence



Number of annotations by evidence

Species filter: All

GO Pitfalls

- Not complete
- Computational annotations
- NOT qualifier
- Splice Variants
- Identifier flagged as ‘obsolete’

NOT annotations in the gene ontology (GO) database

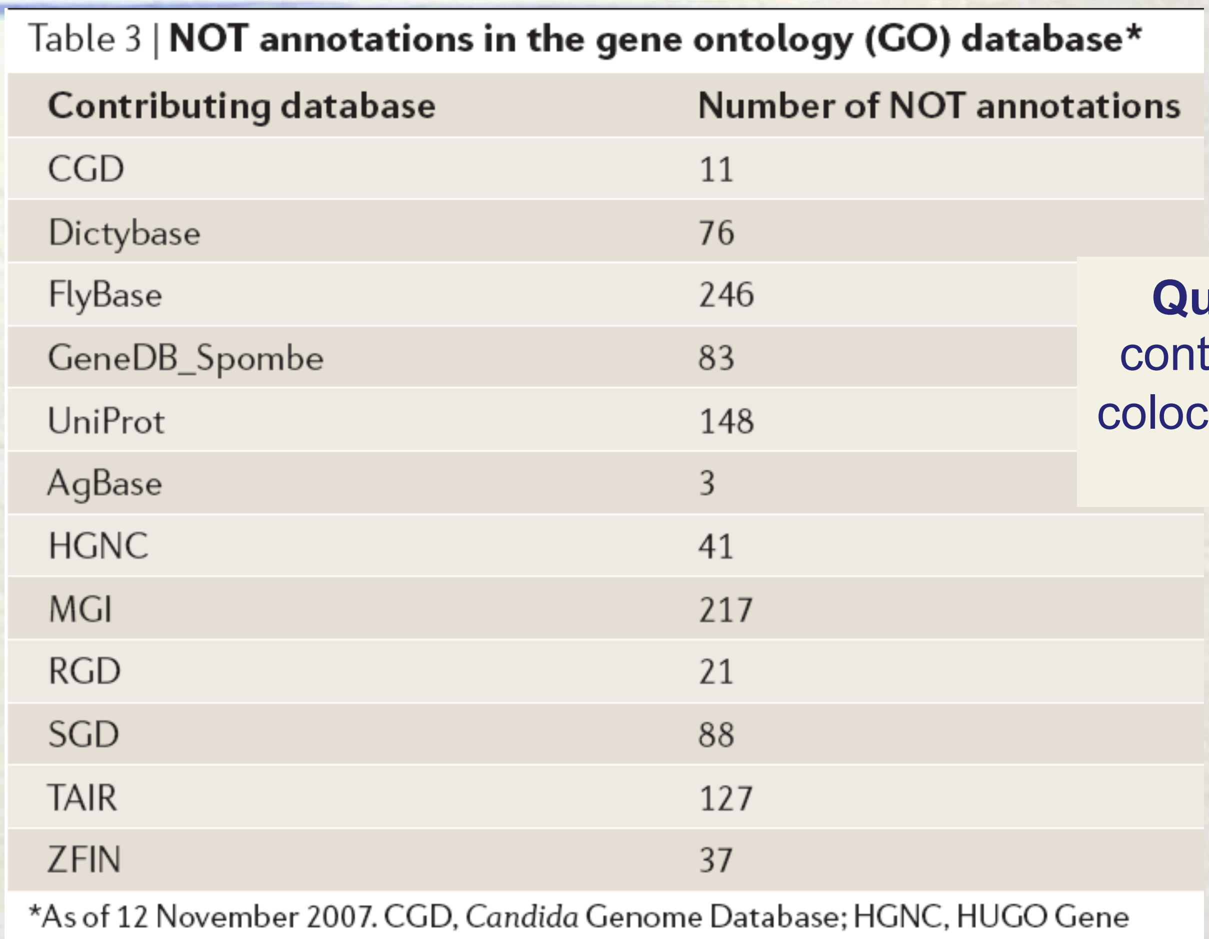


Table 3 | **NOT annotations in the gene ontology (GO) database***

Contributing database	Number of NOT annotations
CGD	11
Dictybase	76
FlyBase	246
GeneDB_Spombe	83
UniProt	148
AgBase	3
HGNC	41
MGI	217
RGD	21
SGD	88
TAIR	127
ZFIN	37

*As of 12 November 2007. CGD, *Candida* Genome Database; HGNC, HUGO Gene

Qualifiers:
contributes_to
colocalizes_with
NOT

Annotation qualifiers — to be or not to be is crucial for GO

GO Pitfalls

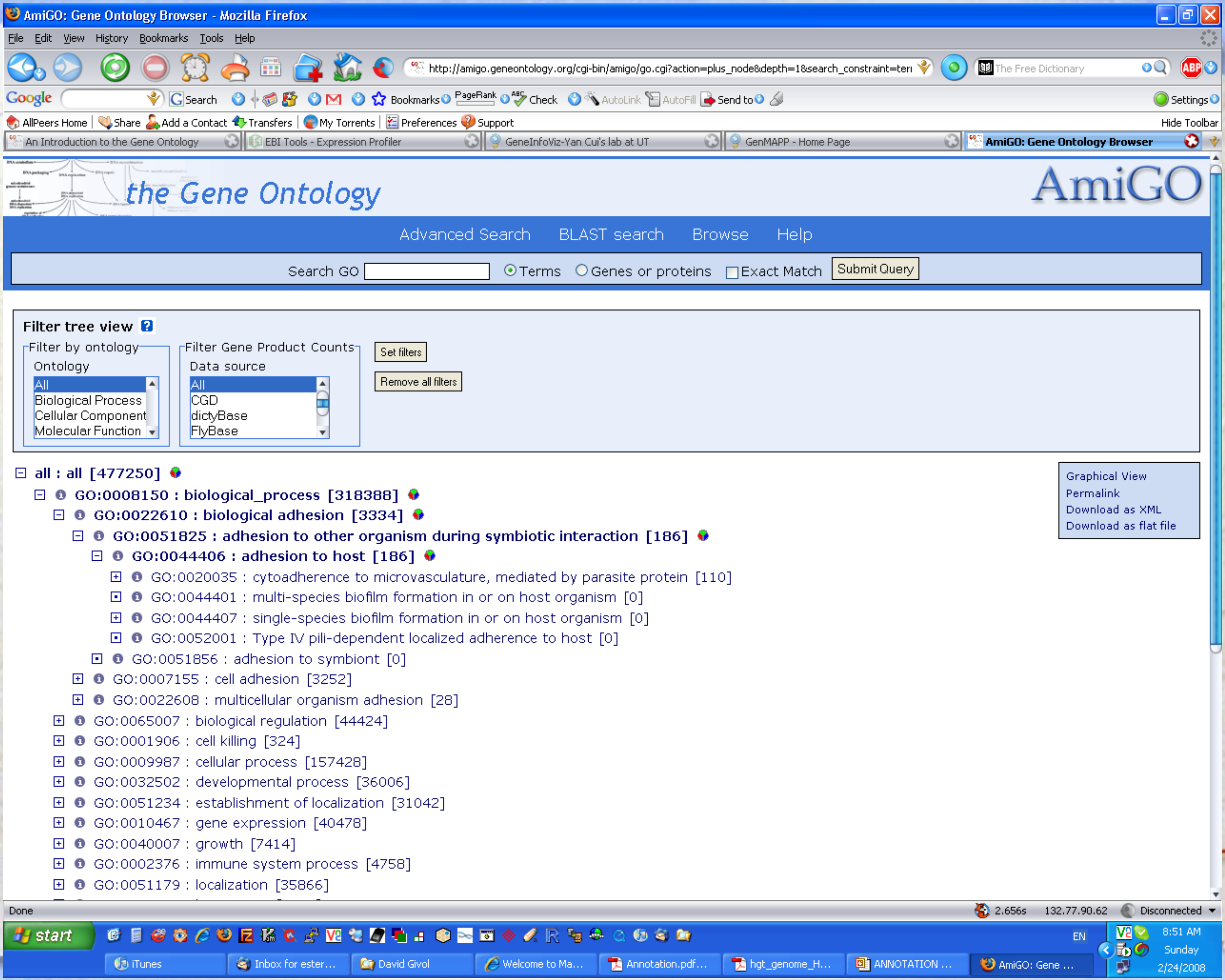
- Not complete
- Computational annotations
- NOT qualifier
- Splice Variants
- Identifier flagged as ‘obsolete’

Splice Variants

- GO annotation is related to gene products, not proteins, so the defining unit is the gene
- If you have different splice variants that have opposite effects, you will have opposing annotation for the same gene, for example BCLX – the long form is anti-apoptotic, the short form is pro-apoptotic, but they are from the same gene...

GO Pitfalls

- Not complete
- Computational annotations
- NOT qualifier
- Splice Variants
- Identifier flagged as ‘obsolete’



all : all [477250]
- GO:0008150 : biological_process [318388]
  - GO:0022610 : biological adhesion [3334]
    - GO:0051825 : adhesion to other organism during symbiotic interaction [186]
      - GO:0044406 : adhesion to host [186]
        - GO:0020035 : cytoadherence to microvasculature, mediated by parasite protein [110]
        - GO:0044401 : multi-species biofilm formation in or on host organism [0]
        - GO:0044407 : single-species biofilm formation in or on host organism [0]
        - GO:0052001 : Type IV pili-dependent localized adherence to host [0]
      - GO:0051856 : adhesion to symbiont [0]
    - GO:0007155 : cell adhesion [3252]
    - GO:0022608 : multicellular organism adhesion [28]
  - GO:0065007 : biological regulation [44424]
  - GO:0001906 : cell killing [324]
  - GO:0009987 : cellular process [157428]
  - GO:0032502 : developmental process [36006]
  - GO:0051234 : establishment of localization [31042]
  - GO:0010467 : gene expression [40478]
  - GO:0040007 : growth [7414]
  - GO:0002376 : immune system process [4758]
  - GO:0051179 : localization [35866]

THANKS TO:

- Dr. Esti Feldmesser, for slides, ideas, and encouragement
- GO consortium website
- Nature Genetics Review article (reference given on earlier slides and on the webpage)