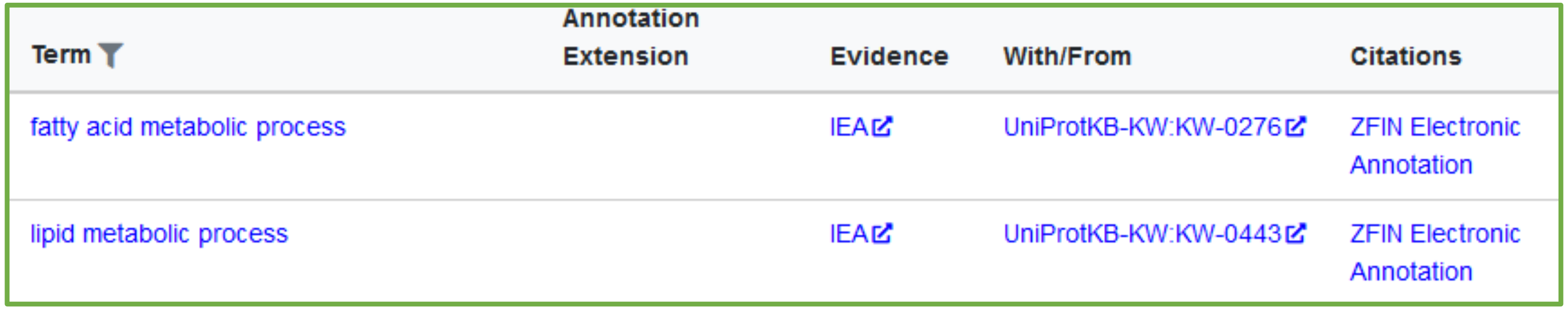
Gene Ontology tools and resources
Antonia Lock

Objectives

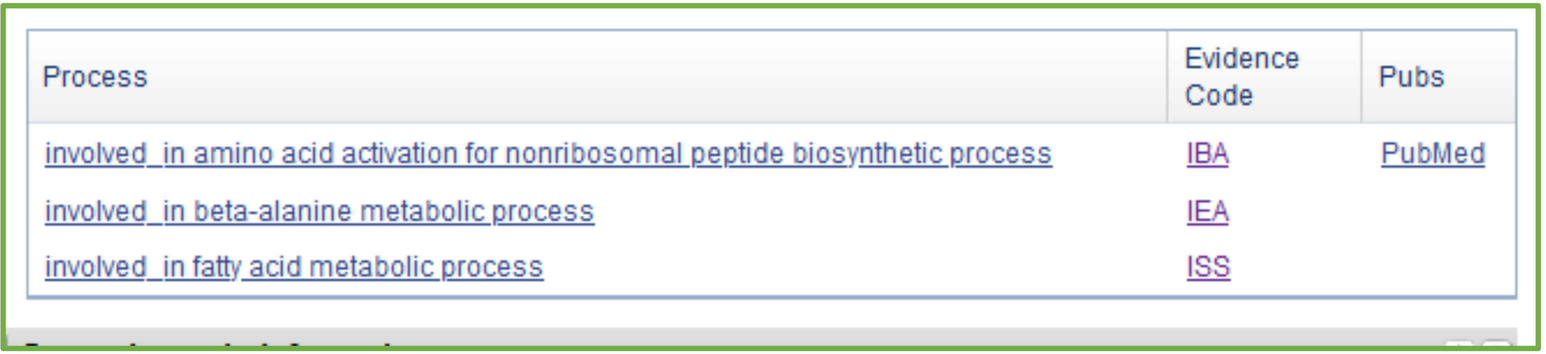
- To explain
  - What the Gene Ontology (GO) is and its limitations
  - Where GO annotations come from
  - How it is accessed
  - How it is used

AASDH aminoadipate-semialdehyde dehydrogenase



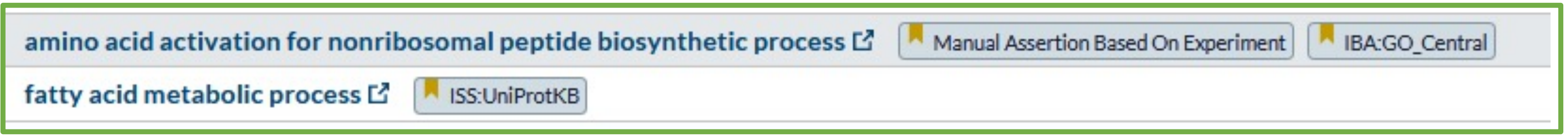
Term	Annotation Extension	Evidence	With/From	Citations
fatty acid metabolic process		IEA	UniProtKB-KW:KW-0276	ZFIN Electronic Annotation
lipid metabolic process		IEA	UniProtKB-KW:KW-0443	ZFIN Electronic Annotation

ZFIN



Process	Evidence Code	Pubs
involved_in amino acid activation for nonribosomal peptide biosynthetic process	IBA	PubMed
involved_in beta-alanine metabolic process	IEA	
involved_in fatty acid metabolic process	ISS	

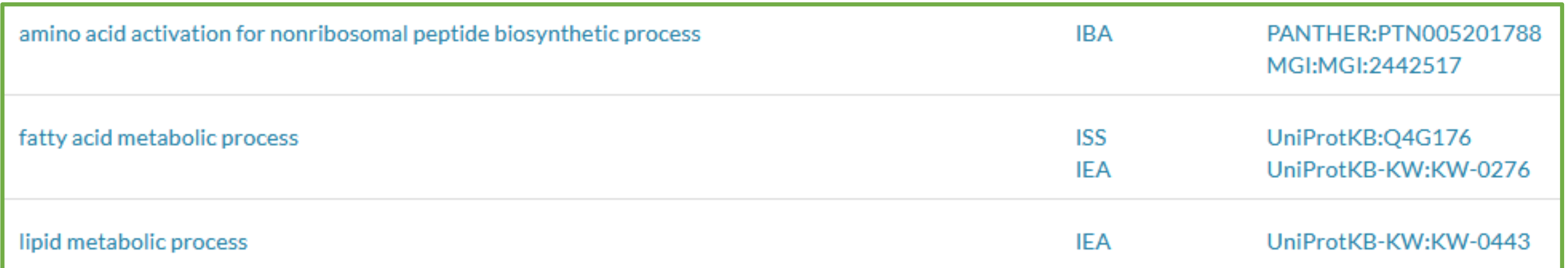
NCBI gene



amino acid activation for nonribosomal peptide biosynthetic process Manual Assertion Based On Experiment IBA:GO_Central

fatty acid metabolic process ISS:UniProtKB

UniProt



amino acid activation for nonribosomal peptide biosynthetic process	IBA	PANTHER:PTN005201788 MGI:MGI:2442517
fatty acid metabolic process	ISS IEA	UniProtKB:Q4G176 UniProtKB-KW:KW-0276
lipid metabolic process	IEA	UniProtKB-KW:KW-0443

Alliance

The Gene Ontology (GO)

The world’s largest resource on gene functions (http://geneontology.org/)

Aims to capture biological knowledge and to describe attributes of gene products:

- In a species agnostic manner - allows transfer of knowledge between organisms
- In language interpretable by humans and machines (ontologies)
  - Enables computational analyses for interpretation of high-throughput experimental and clinical data
  - Enables access to data at scale

Freely available

Two parts; the **ontology** and the **annotations**

- An ontology is a dictionary of defined terms, but with hierarchal relationships to each other
- An annotation is an association of a gene product with an ontology term

What GO is not

- GO only describes ‘normal’ biological processes and functions
- It does NOT cover pathological processes (e.g. disease) or functions only observed in vitro
- It is not static
  - Both the ontology and annotations are reviewed and updated regularly
  - **Always ensure you use up-to-date releases in analysis**

The Ontology

GO provides systematic language to describe gene products in three domains:

Molecular Function (MF)

- Single step tasks
  - fumarase, protein kinase, cargo receptor

Biological Process (BP)

- A recognized series of events
  - citric acid cycle, cell division, kidney development

Cellular Component (CC)

- Where gene products localize
  - mitochondrion, lysosomal membrane, synapse

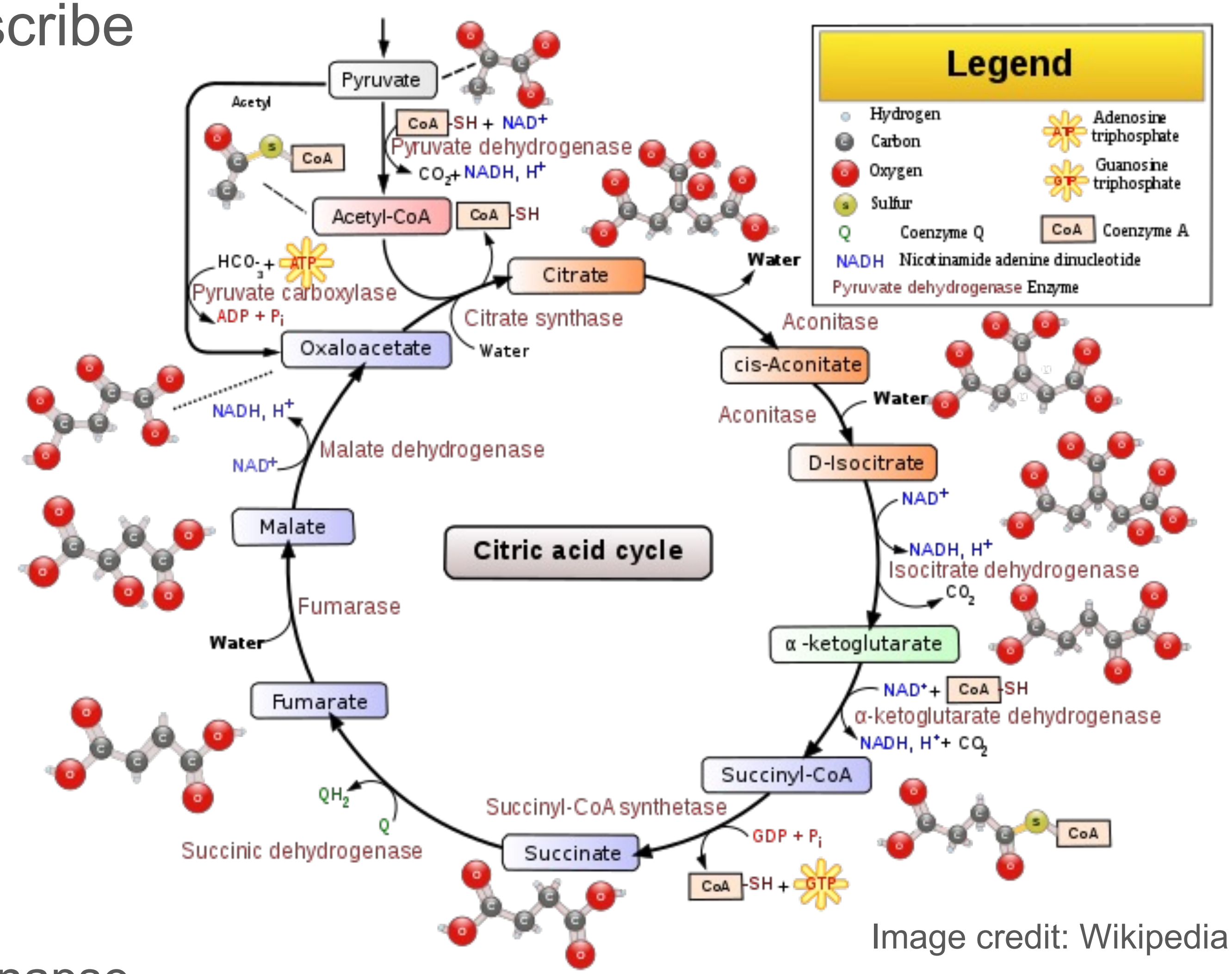


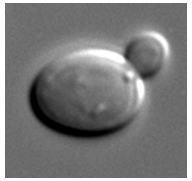
Image credit: Wikipedia

Why is GO useful?

- Disambiguate language
  - The **same name** is used for **different concepts**

Tooth **bud initiation**

≠



Cell **bud initiation**

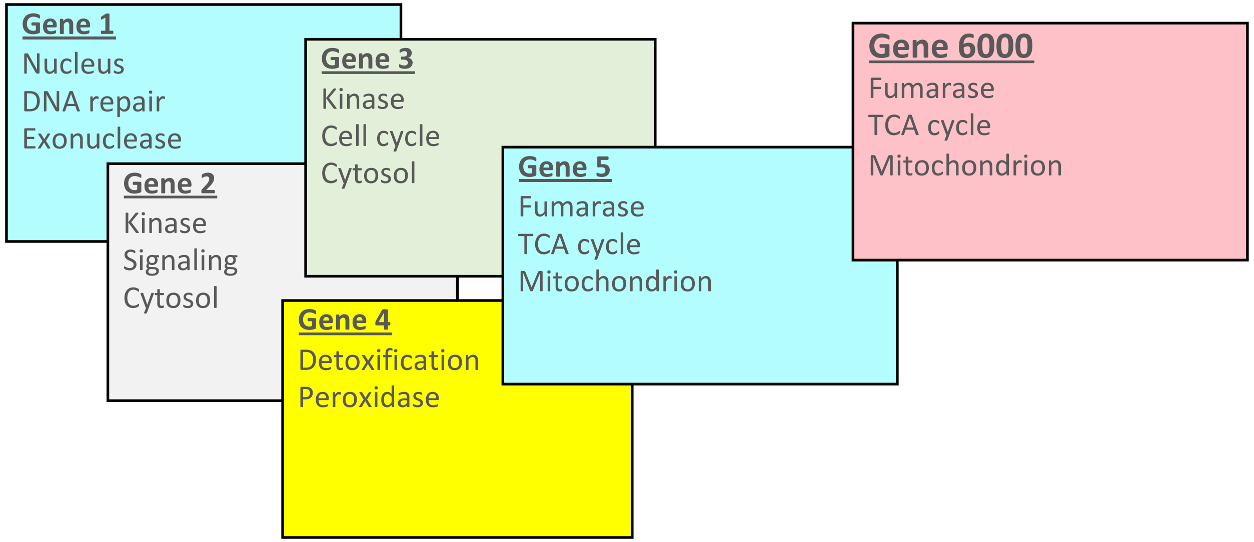
≠

Flower **bud initiation**

  - **Different names** describe the **same concept**
    - Citric acid cycle, tricarboxylic acid cycle, TCA cycle, Krebs cycle
  - This makes comparisons across species or databases difficult

GO facilitates gene specific data aggregation

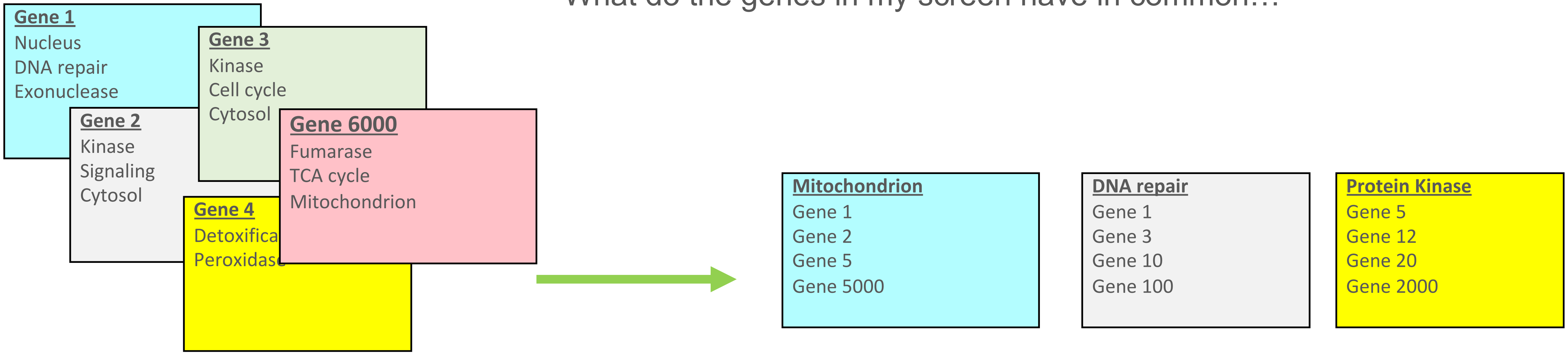
- Allows systematic capture of gene-specific knowledge
  - Curate detail, annotating genes to ‘terms’
- Involves time-consuming curation/annotation activities
  - Literature searching, database searching, sequence analysis…
  - 165.000+ publications annotated for GO



Data aggregation on this scale is not feasible for individual research groups

GO allows grouping genes by feature

- Essentially creating lists of genes with similar features
  - Useful to plan experiments
    - What genes encode protein kinases…
  - Analysing data
    - Identifying trends (enrichments)
      - What do the genes in my screen have in common…



Gene 1
Nucleus
DNA repair
Exonuclease

Gene 2
Kinase
Signaling
Cytosol

Gene 3
Kinase
Cell cycle
Cytosol

Gene 4
Detoxifica
Peroxidase

Gene 6000
Fumarase
TCA cycle
Mitochondrion

Mitochondrion
Gene 1
Gene 2
Gene 5
Gene 5000

DNA repair
Gene 1
Gene 3
Gene 10
Gene 100

Protein Kinase
Gene 5
Gene 12
Gene 20
Gene 2000

FAIR - four founding principles for good data management

Guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets.

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure.

Findable:

- Metadata and data should be easy to find for both humans and computers.
- Machine-readable metadata are essential for automatic discovery of datasets and services.
- Associate data with identifiers, metadata, & deposit/index in a searchable repository

Accessible

- Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation
- Metadata is retrievable and accessible using standardized protocols

Interoperable

- Data needs to be integratable with other data & interoperate with applications or workflows for analysis, storage, and processing.
- Metadata and data described using accessible and widely acceptable/domain-specific language

Reusable

- The ultimate goal of FAIR is to optimise the reuse of data.
- Well-described metadata and data so they can be replicated and/or combined in different settings.
- Clear data provenance & usage license

The FAIR Guiding Principles for scientific data management and stewardship. Wilkinson et al (2016) PMID:26978244

Ontologies

- **Help us build standards for terms used in a particular domain**

- Consists of several parts:
    - Classes / entities / terms
    - Metadata
    - Relationships
    - File formats
    - Axioms

Ontology parts: Classes

Basic unit within an ontology, representing things within a domain

- Eg. endocytosis, hydrogen peroxide, fragile X syndrome

Associated with **stable** and **semantics-free** identifiers to promote stability even though ontology representation may change with new scientific knowledge

- e.g. GO:0006897, CHEBI:16240 and MONDO:0010383 or OMIM:300624 or DOID:14261)

Classes may be **merged**

- Alternate IDs maintained as secondary identifiers

Classes may be **obsoleted**

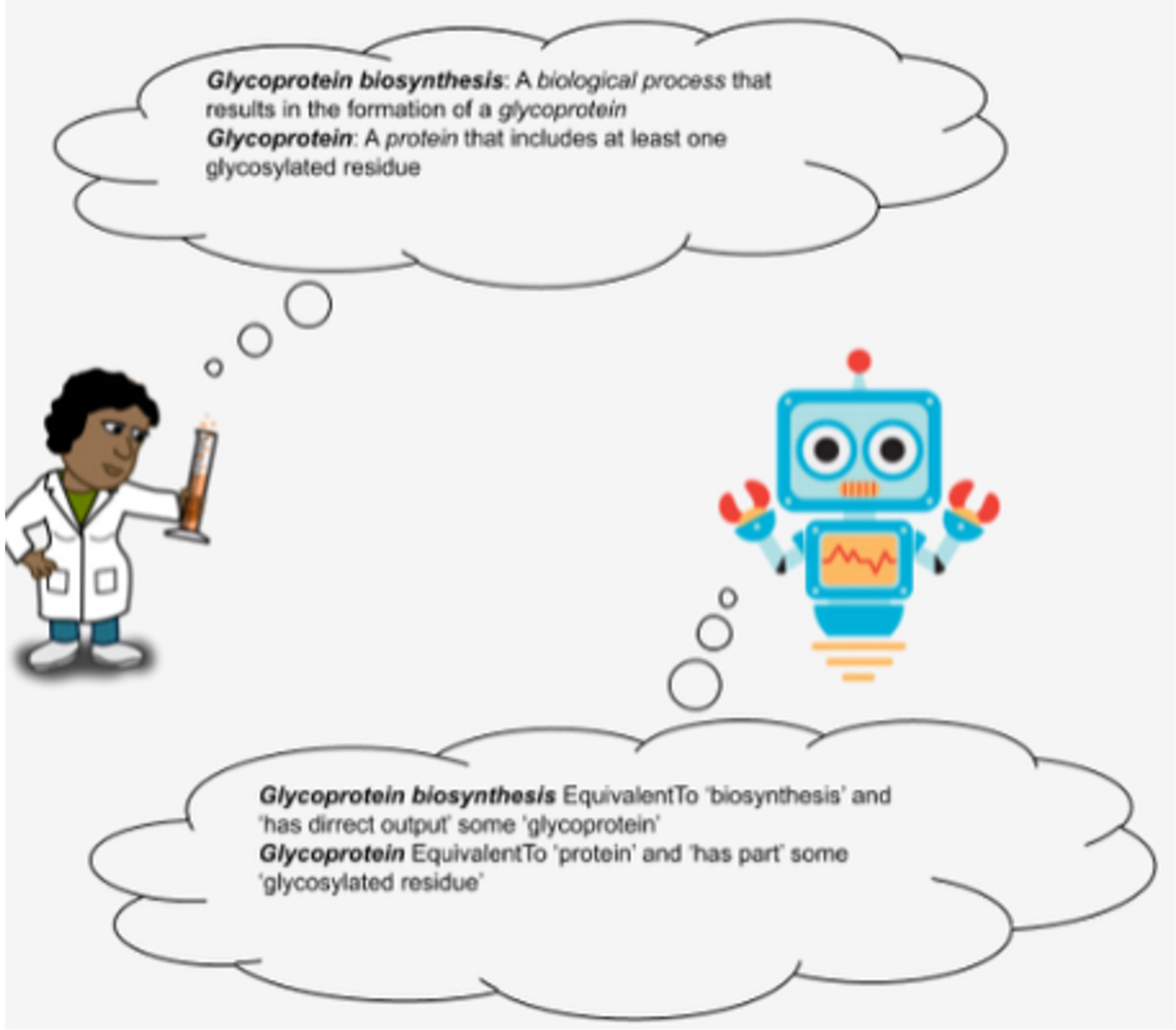
- Shouldn’t be used for further annotation, but kept for historical records
- Obsolete classes may contain metadata pointing to alternative classes

Ontology parts: Metadata

Information associated with classes

Term definition: Both a **textual definition** (aimed primarily at humans) and **logical definition** (aimed primarily at machines)

- Curators must always annotate to the term definition, not the term name
- **Good** definitions are important:
    - Promote inter-curator annotation agreement
    - Facilitates integration: Allows non-experts and experts in adjacent disciplines to understand unfamiliar terms, enabling use of terms from external ontologies
    - Allow ontology developers to use assisting tools such as “reasoners” (more later in “tools”)



Glycoprotein biosynthesis: A *biological process* that results in the formation of a *glycoprotein*
Glycoprotein: A *protein* that includes at least one glycosylated residue

Glycoprotein biosynthesis EquivalentTo 'biosynthesis' and 'has direct output' some 'glycoprotein'
Glycoprotein EquivalentTo 'protein' and 'has part' some 'glycosylated residue'

From C. Mungall’s blog post: OntoTip: Write simple, concise, clear, operational textual definitions https://douroucouli.wordpress.com/2019/07/08/ontotip-write-simple-concise-clear-operational-textual-definitions/

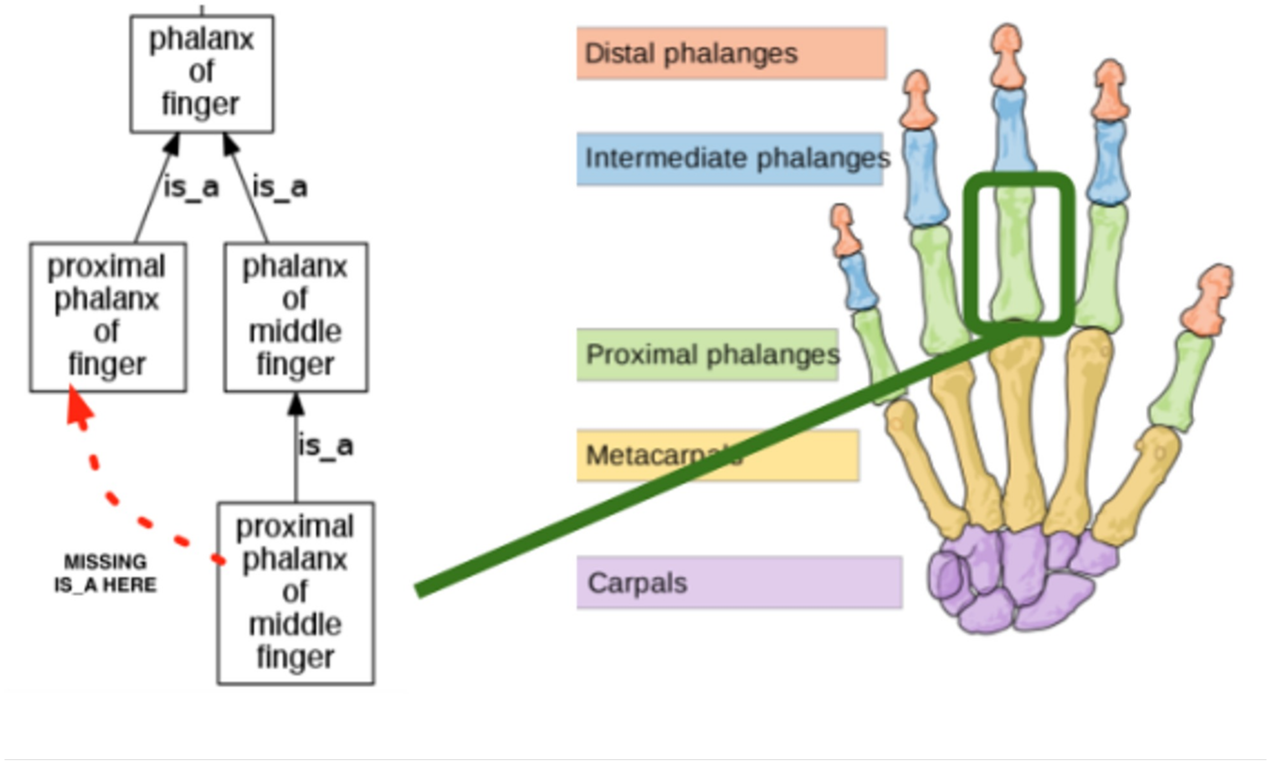
Why is multiple inheritance preferred over single inheritance?

Classification of *proximal phalanx of middle finger* (PPoMF)

- *is-a* links are shown as black arrows
- missing is-a relationship indicated with a dashed red line

If a phenotype involving the *proximal phalanx of middle finger* (PPoMF) is annotated, and a user queries for *proximal phalanx of [any] finger* (PPoF), **the user will not get the expected results**.

The missing relationship is from a single-inheritance ontology - the missing relationship is by design.

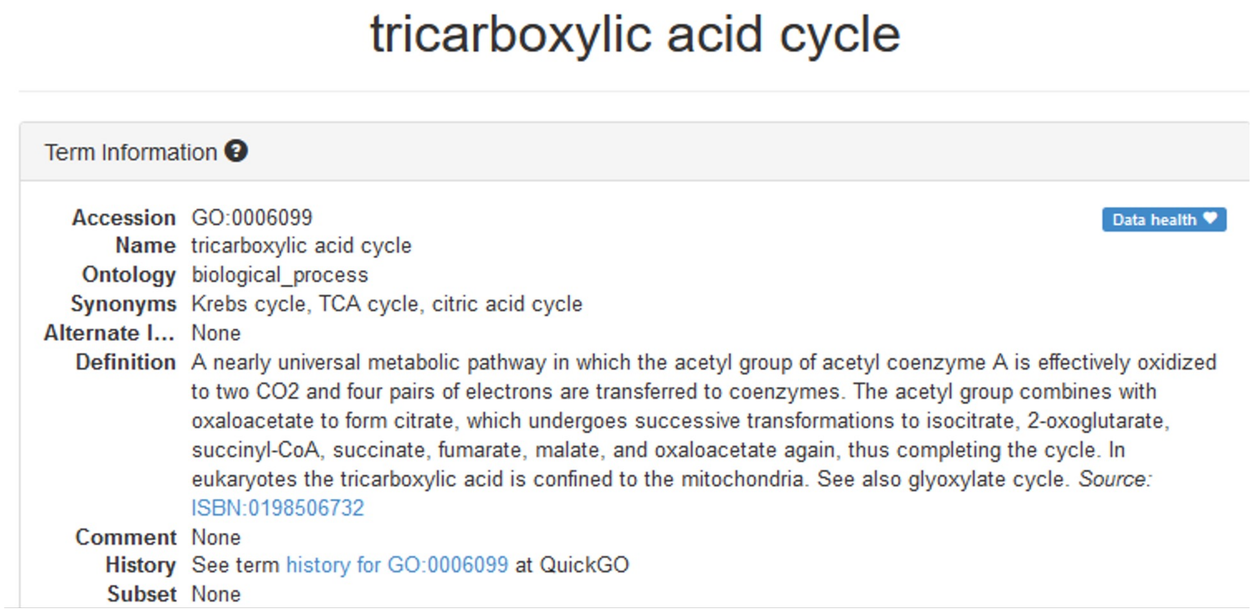


Example from C. Mungall's blog post "**Single-inheritance principle considered dangerous**: https://douroucouli.wordpress.com/2019/05/10/ontotip-single-inheritance-principle-considered-dangerous/

Ontology parts: Metadata

Information associated with classes

- **Secondary IDs:** resulting from term merges), obsoletion flags, synonyms, cross-references to alternative databases
  - Synonyms come in different types, e.g. exact, related, narrow
  - Different ontologies might cross-reference other ontologies for different reasons!
- e.g. GO cross-references EC and Rhea and links to publications supporting the existence of the class, other ontologies may cross-reference chemical compounds participating in a reaction)
- May contain **comments** and examples of intended usage



tricarboxylic acid cycle

Term Information

Accession GO:0006099

Name tricarboxylic acid cycle

Ontology biological_process

Synonyms Krebs cycle, TCA cycle, citric acid cycle

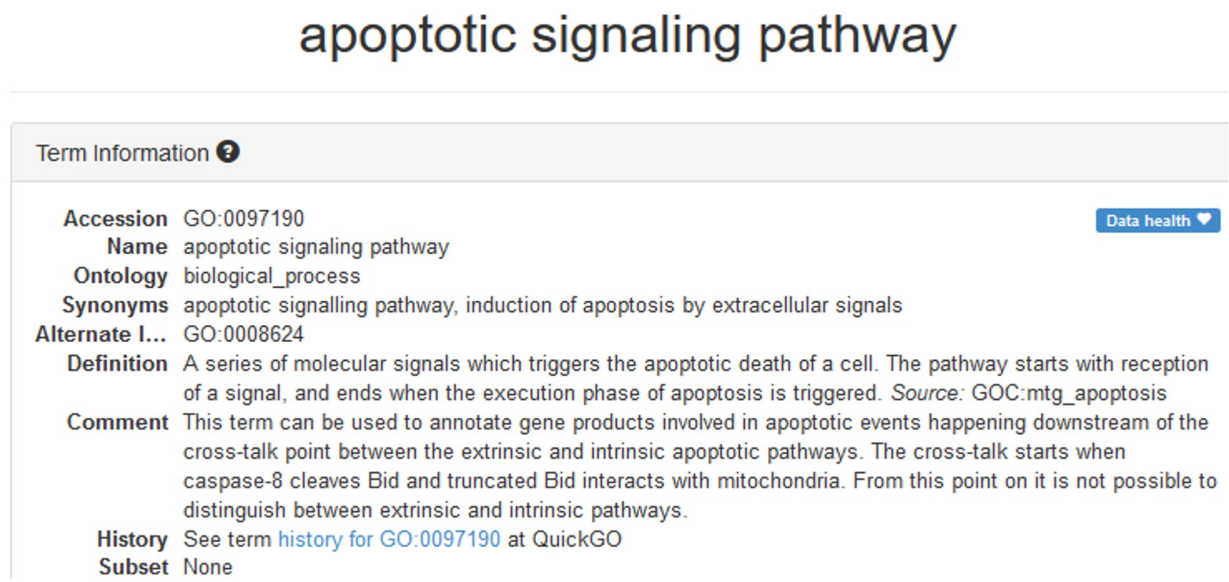
Alternate I... None

Definition A nearly universal metabolic pathway in which the acetyl group of acetyl coenzyme A is effectively oxidized to two CO2 and four pairs of electrons are transferred to coenzymes. The acetyl group combines with oxaloacetate to form citrate, which undergoes successive transformations to isocitrate, 2-oxoglutarate, succinyl-CoA, succinate, fumarate, malate, and oxaloacetate again, thus completing the cycle. In eukaryotes the tricarboxylic acid is confined to the mitochondria. See also glyoxylate cycle. *Source:* ISBN:0198506732

Comment None

History See term history for GO:0006099 at QuickGO

Subset None



apoptotic signaling pathway

Term Information

Accession GO:0097190

Name apoptotic signaling pathway

Ontology biological_process

Synonyms apoptotic signalling pathway, induction of apoptosis by extracellular signals

Alternate I... GO:0008624

Definition A series of molecular signals which triggers the apoptotic death of a cell. The pathway starts with reception of a signal, and ends when the execution phase of apoptosis is triggered. *Source:* GOC:mtg_apoptosis

Comment This term can be used to annotate gene products involved in apoptotic events happening downstream of the cross-talk point between the extrinsic and intrinsic apoptotic pathways. The cross-talk starts when caspase-8 cleaves Bid and truncated Bid interacts with mitochondria. From this point on it is not possible to distinguish between extrinsic and intrinsic pathways.

History See term history for GO:0097190 at QuickGO

Subset None

Ontology parts: Relationships

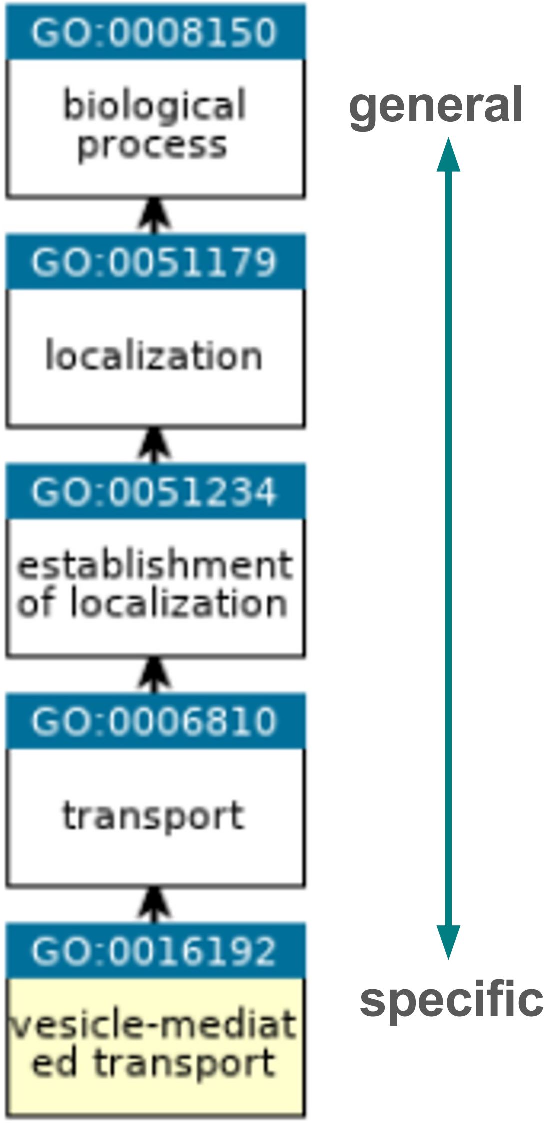
Classes are arranged in a **hierarchy** from general to specific

Relations between classes provide structure to the ontology

- Broad ***general/ancestor/parent*** terms give rise to ***specific/descendant/child*** terms

Important since relationships within the ontology enable transitive **inheritance of annotations**

- I.e. when annotating to a more specific class, the annotated feature (such as a protein, or a drug) is automatically annotated to all the more general classes
- A search to a parent term can retrieve all children
- Every possible path from a term back to the root node must be accurate, or the ontology must be revised



Ontology parts: Relationships & ontology structure

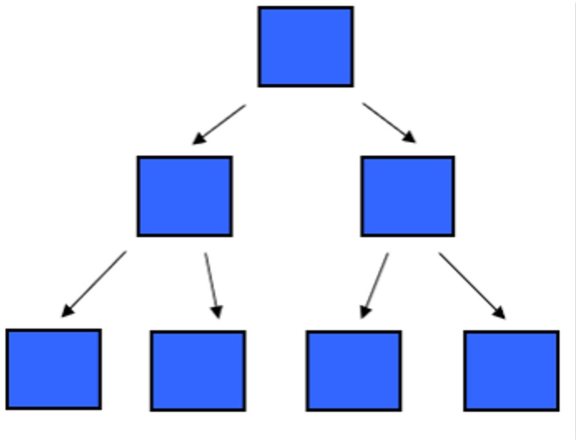
Relationships link together terms in the ontology

Ontologies can take on a simple tree structure (**hierarchy**)

Most bio-ontologies are structured as more complex **directed acyclic graphs** (DAGs)

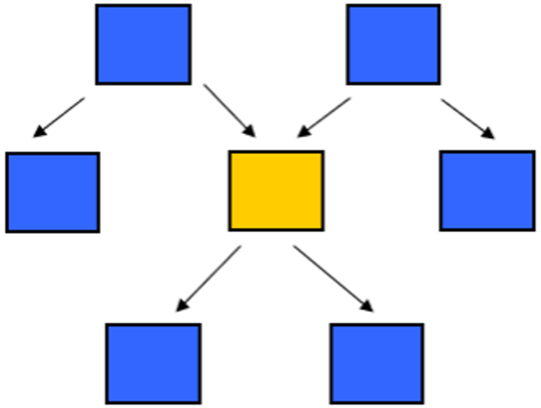
Hierarchy

- Parent: one to many children
  - Child: only one parent
  - Relationship: ***is a***



DAG

  - Many to many each direction
  - Many relationship types
- No circles

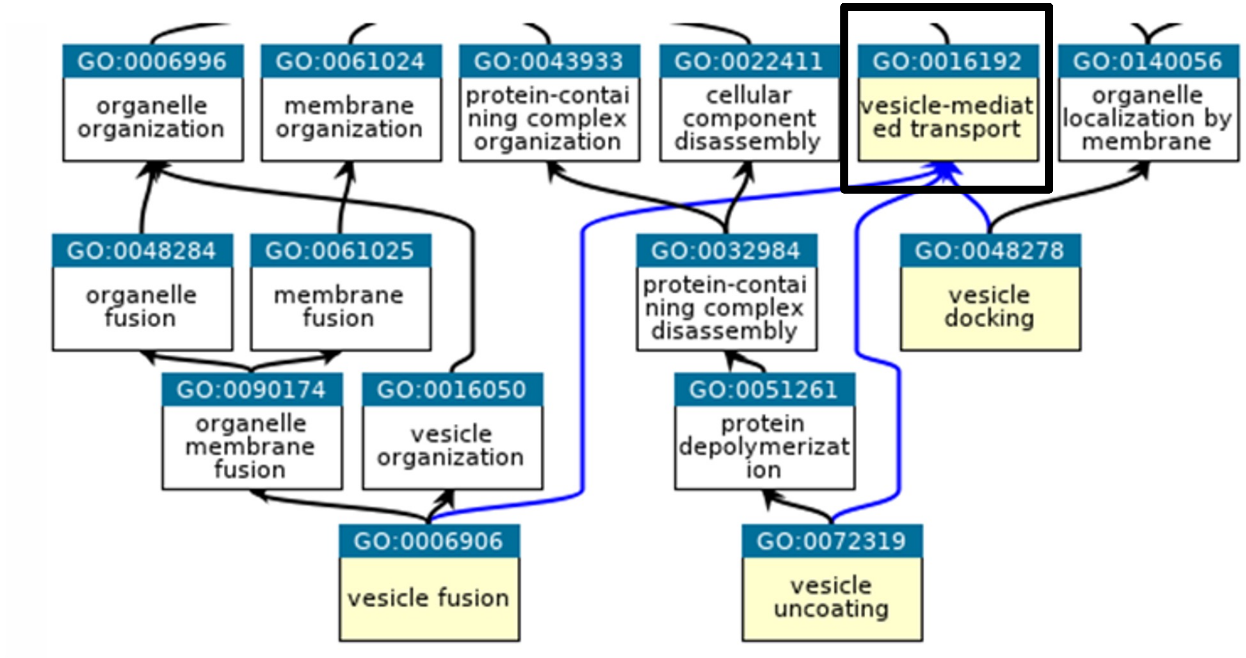


Ontology parts: Relationships

Relationships common to many ontologies

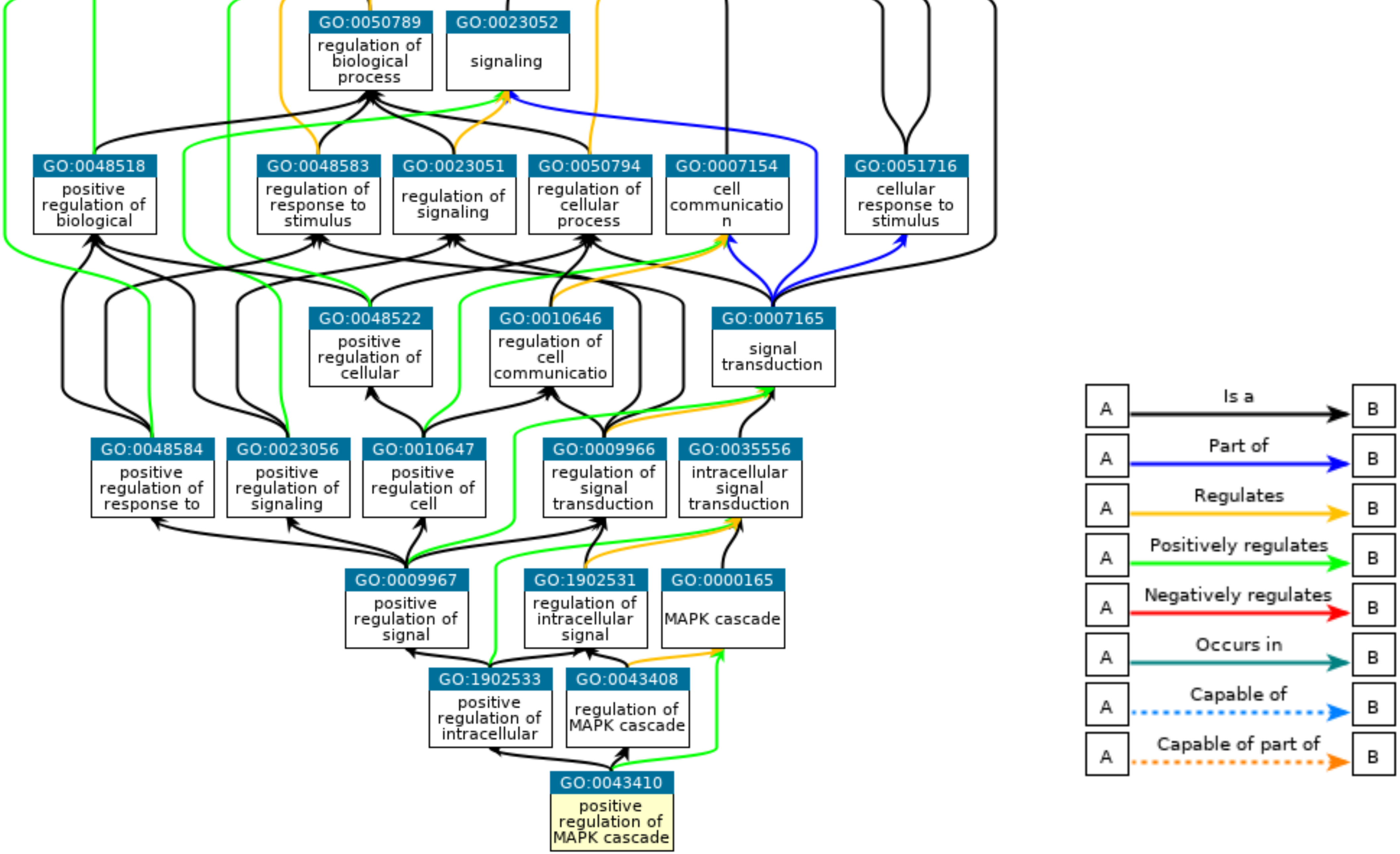
- **Is_a:** is a subtype of
  - A volvo is_a car
  - The mitochondrion is an organelle
- **Part_of:**
  - A wheel is part_of a car
  - The mitochondrial lumen is part_of the mitochondrion

- **Vesicle mediated transport** consists of several ordered steps
- All steps (e.g. docking, fusion, uncoating) are *part_of* vesicle mediated transport



- Each step may be connected to their own separate parts of the ontology
  - Vesicle fusion *is_a* organelle membrane organization
  - Vesicle uncoating *is_a* protein depolymerization

- Acyclic – no circles allowed
- **Ancestors are inherited**
- >40.000 terms across the three ontologies


GO:0050789
regulation of biological process
GO:0023052
signaling
GO:0048518
positive regulation of biological
GO:0048583
regulation of response to stimulus
GO:0023051
regulation of signaling
GO:0050794
regulation of cellular process
GO:0007154
cell communicatio n
GO:0051716
cellular response to stimulus
GO:0048522
positive regulation of cellular
GO:0010646
regulation of cell communicatio
GO:0007165
signal transduction
GO:0048584
positive regulation of response to
GO:0023056
positive regulation of signaling
GO:0010647
positive regulation of cell
GO:0009966
regulation of signal transduction
GO:0035556
intracellular signal transduction
GO:0009967
positive regulation of signal
GO:1902531
regulation of intracellular signal
GO:0000165
MAPK cascade
GO:1902533
positive regulation of intracellular
GO:0043408
regulation of MAPK cascade
GO:0043410
positive regulation of MAPK cascade
A
Is a
B
A
Part of
B
A
Regulates
B
A
Positively regulates
B
A
Negatively regulates
B
A
Occurs in
B
A
Capable of
B
A
Capable of part of
B

Annotation inheritance

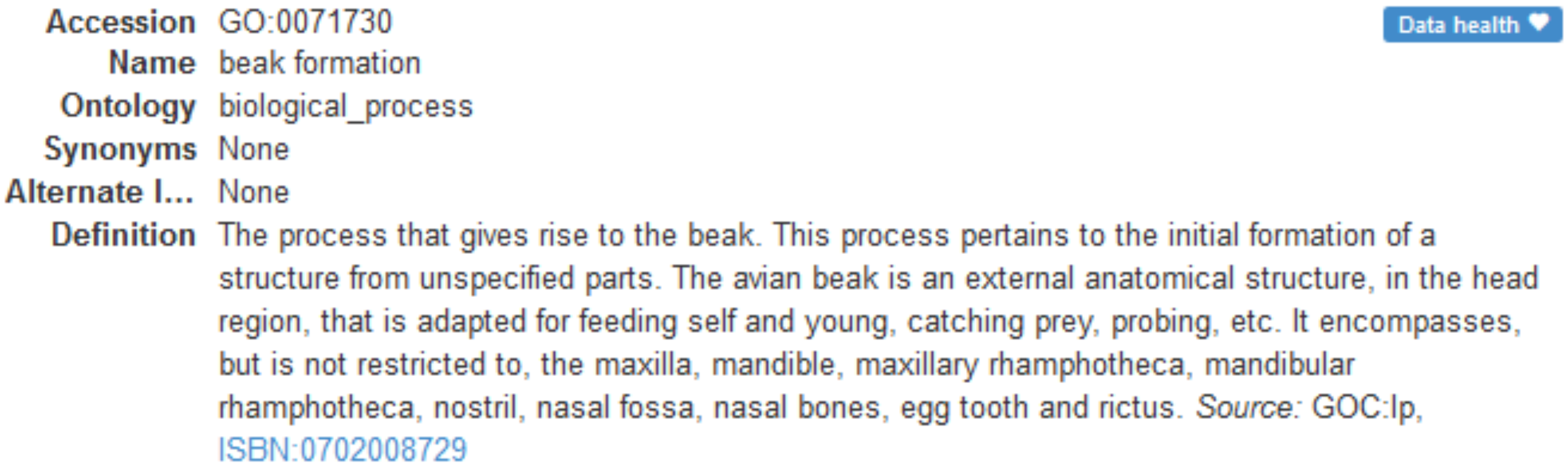
- An important feature of GO is that broader parents give rise to more specific children.
- When a gene is directly annotated to a term it is automatically indirectly annotated to all its parent terms …therefore **a search on a parent retrieves children**
  - E.g. all genes annotated to “mitochondrial lumen” are also annotated to “mitochondrion”
- Every possible path from any term back to the root node must be biologically accurate, or the ontology must be revised

Functional biocuration

- **A functional annotation is the association of a feature** (e.g. a gene) **with an ontology term**
  - A statement that a gene product performs a specific function, localizes to a particular structure, or participates in a specific pathway
  - Currently 1 billion+ GO annotations
- Two main aspects
  - Literature curation (reading full text publications and associating novel biological information with the appropriate genes or features)
  - Sequence analysis (infer information for unpublished genes)
- Manual curation
  - Curators from annotation groups (e.g. ZFIN, UniProt) create annotations based on published experiments
  - Low throughput
  - Annotations are then transferred to orthologous genes (manually or automatically)
- Electronic pipelines (bulk annotation)
  - Many are based on interpro signatures (InterPro2GO) or 1:1 orthologs from Ensembl gene trees.
  - Typically results in high level, but high quality, annotations (PMID:22693439)
  - High throughput

Caution

- Not all annotations are transferred to all species, even if genes have a 1-to-1 orthologous relationship.
- Would be incorrect to transfer this annotation from chick *BMP4* to human *BMP4*



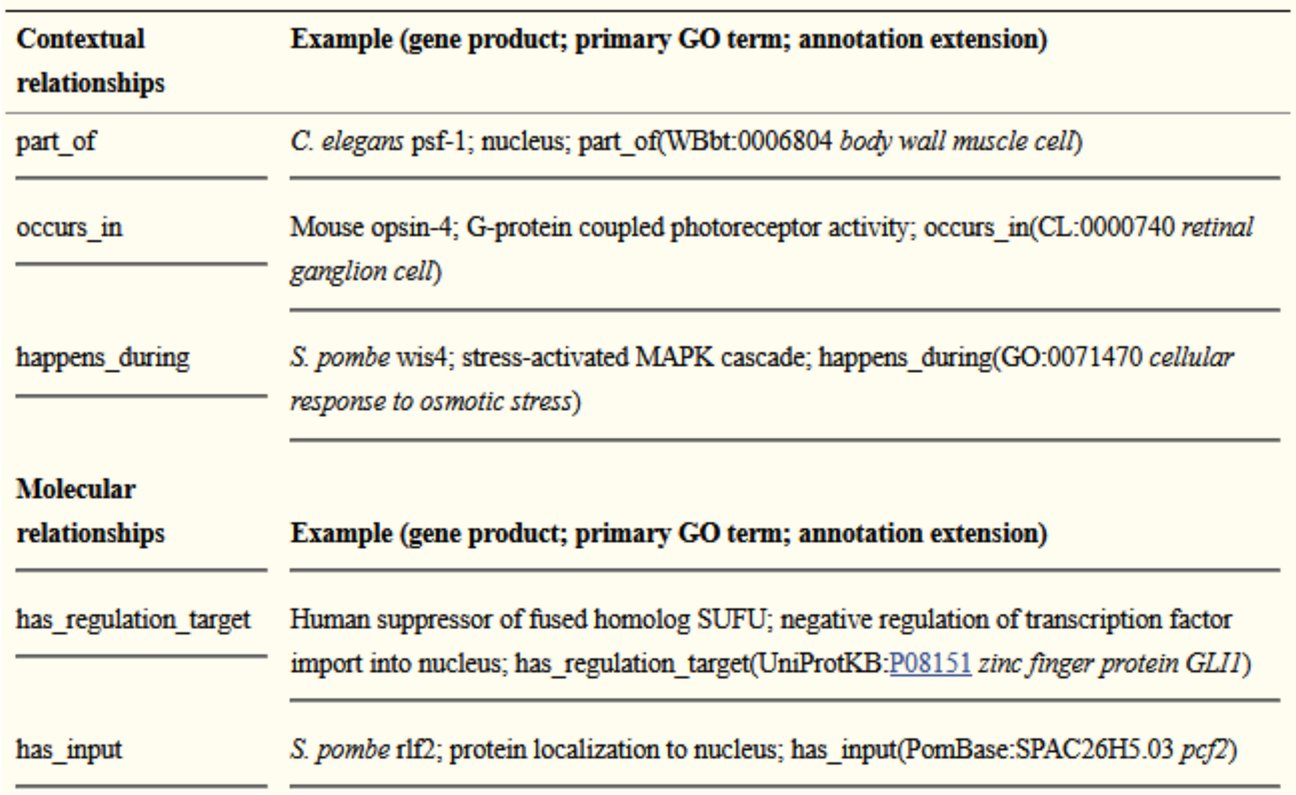
Accession	GO:0071730
Name	beak formation
Ontology	biological_process
Synonyms	None
Alternate I...	None
Definition	The process that gives rise to the beak. This process pertains to the initial formation of a structure from unspecified parts. The avian beak is an external anatomical structure, in the head region, that is adapted for feeding self and young, catching prey, probing, etc. It encompasses, but is not restricted to, the maxilla, mandible, maxillary rhamphotheca, mandibular rhamphotheca, nostril, nasal fossa, nasal bones, egg tooth and rictus. *Source:* GOC:lp, ISBN:0702008729

Evidence codes – describes how an annotation is supported

- Many types, only a subset listed below. http://geneontology.org/docs/guide-go-evidence-codes/
- **Experimental** – evidence from wet-lab experiments - e.g:
  - IDA: Inferred from Direct Assay – enzyme assays, immunofluorescence
  - IMP: Inferred from Mutant Phenotype – comparison of differences in phenotypes from two different alleles
- **Computational**
  - ISO / ISS – Inferred from Sequence Orthology / Similarity
- **Phylogenetic**
  - IBA – Inferred from Biological aspect of Ancestor - phylogenetic analyses to transfer annotations amongst related sequences based on common ancestry
- **“No evidence”**
  - IC – Inferred by Curator
- **Author statements**
- **Automatic**
  - IEA – Inferred from Electronic annotation – automatic transfer of annotations based on the presence of sequence features or on orthology. The annotations are not manually reviewed.

Beyond basic GO annotation: Annotation extensions

- Provides additional detail to GO annotations
  - E.g. the target gene or the location of a molecular function
- Provides curated, directional links between gene products
  - supports network construction



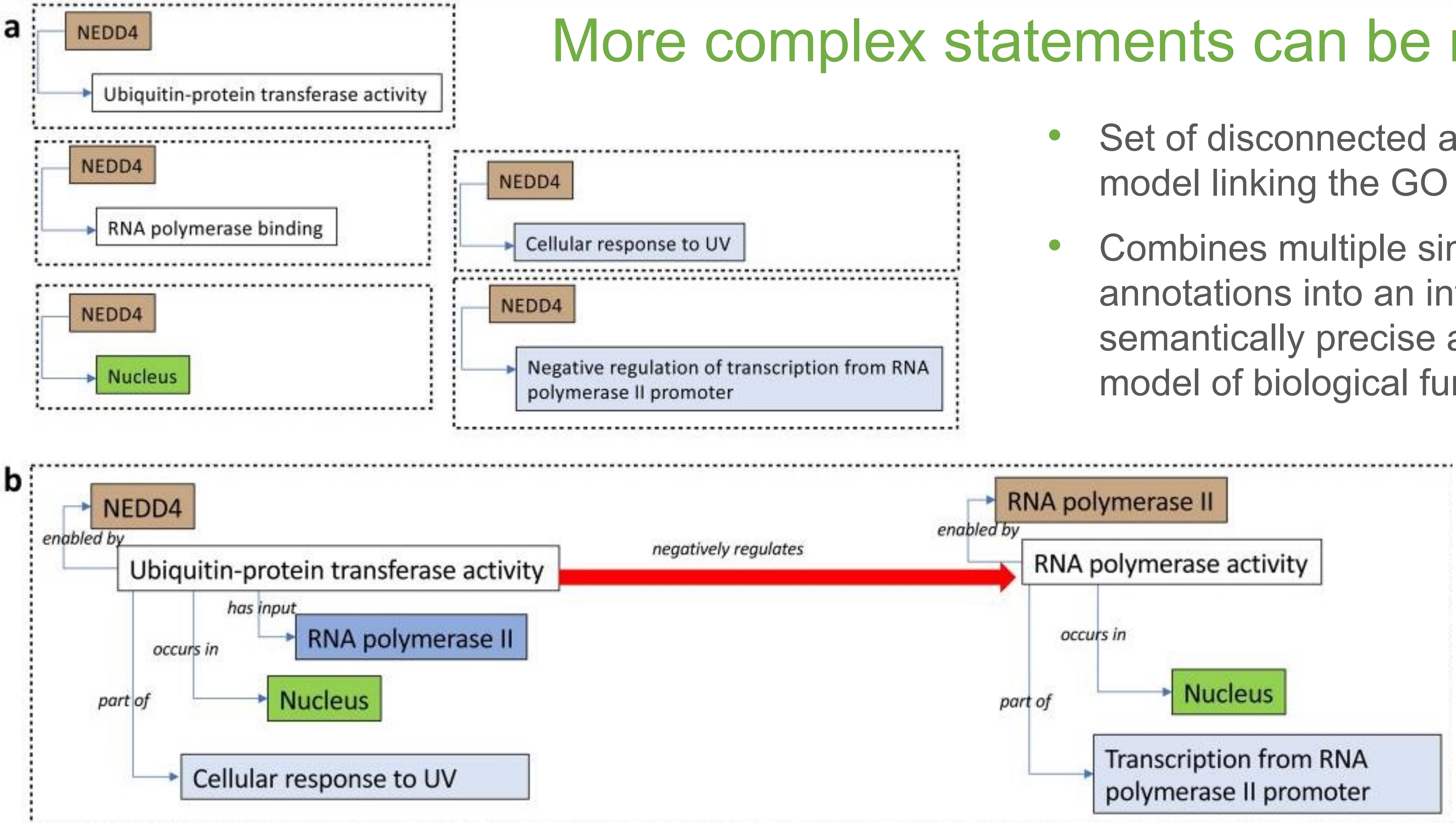
Contextual relationships	**Example (gene product; primary GO term; annotation extension)**
part_of	*C. elegans* psf-1; nucleus; part_of(WBbt:0006804 *body wall muscle cell*)
occurs_in	Mouse opsin-4; G-protein coupled photoreceptor activity; occurs_in(CL:0000740 *retinal ganglion cell*)
happens_during	*S. pombe* wis4; stress-activated MAPK cascade; happens_during(GO:0071470 *cellular response to osmotic stress*)
Molecular relationships	**Example (gene product; primary GO term; annotation extension)**
has_regulation_target	Human suppressor of fused homolog SUFU; negative regulation of transcription factor import into nucleus; has_regulation_target(UniProtKB:P08151 *zinc finger protein GLI1*)
has_input	*S. pombe* rlf2; protein localization to nucleus; has_input(PomBase:SPAC26H5.03 *pcf2*)

Huntley et al.
PMID:24885854

Beyond basic GO annotation:
Causal Activity Modeling (GO-CAM)

- Provides a framework to link GO annotations in integrated models of biological systems
  - A GO annotation is the association of a gene product with a term in the 3 ontologies
    - In GO, what gene products do (Molecular Functions) are distinct from the larger biological programs they are part of (Biological Processes) and where they take place (Cellular Components)
  - For traditional GO annotation there has been little representation of how annotations to one gene fit together
  - Provides a system for formalizing and extending GO annotations
  - Enables pathway visualisation
  - Documentation: http://geneontology.org/go-cam

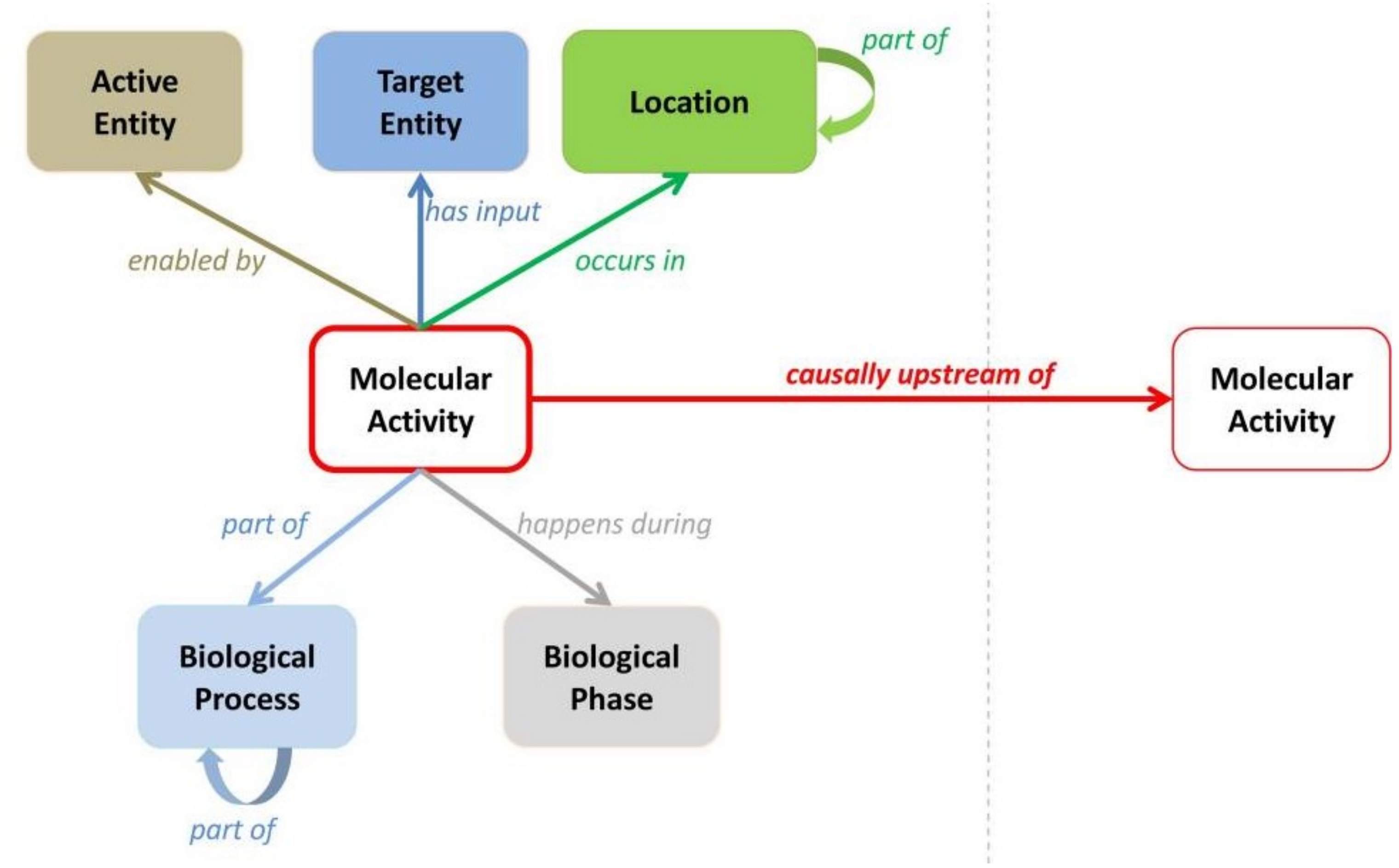
More complex statements can be represented



- Set of disconnected annotations vs. a model linking the GO annotations
- Combines multiple simple GO annotations into an integrated, semantically precise and computable model of biological function

Thomas et. al PMID:31548717

More complex statements can be represented



Thomas et. al PMID:31548717

Centres on a molecular activity

….carried out by a gene product or complex (brown)

…that may act on another entity (dark blue)

…the activity occurs in a location (green)

…and part of a specific biological program (blue)

……which may be part of a larger process or phase, or anatomical structure or cell type (curved arrow)

…and may have causal effects on other molecular activities (red arrow)

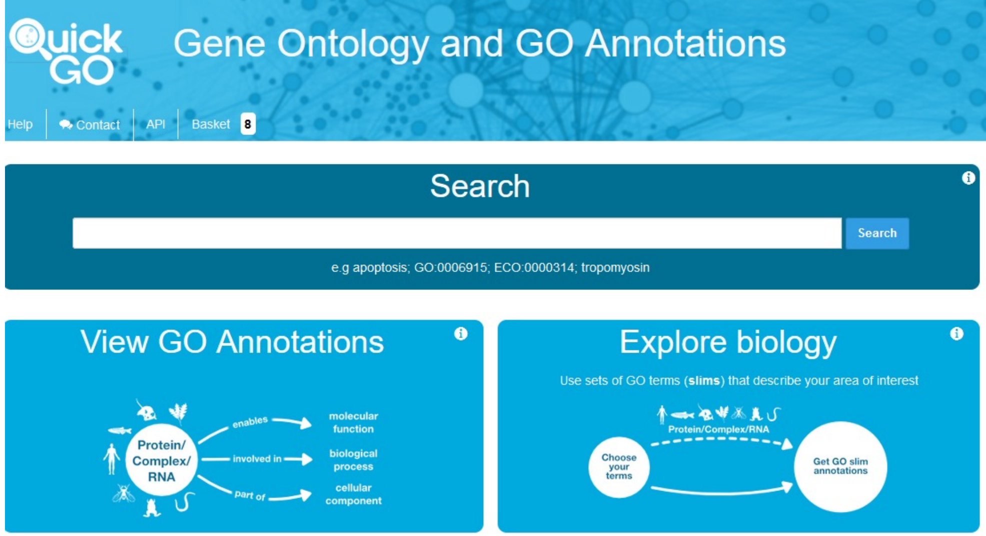
……effectively creating networks of activities

Accessing GO annotations

- GO downloads http://current.geneontology.org/products/pages/downloads.html
- Browsers
  - ZFIN https://zfin.org/
  - QuickGO https://www.ebi.ac.uk/QuickGO/
  - AmiGO 2 http://amigo.geneontology.org/amigo
- File formats
  - A GAF (Gene Association File) is a GO annotation file containing annotations made to the GO
  - The GPAD - Gene Product Association File Format - is an alternative means of exchanging annotations from the GAF. The GPAD format is designed to be more normalized than GAF.

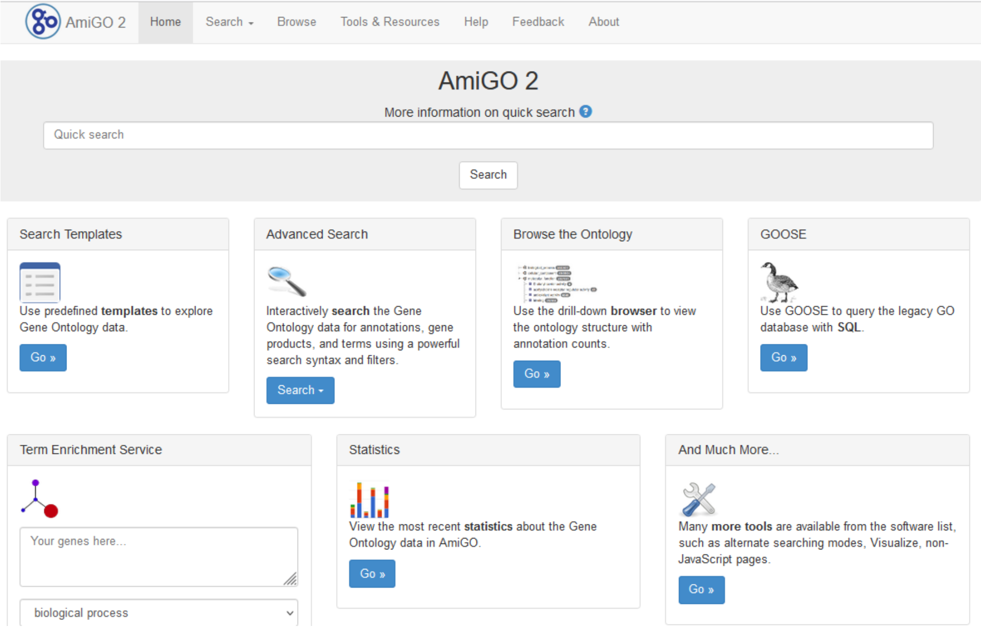
Ontology and annotation browsers

- Two main generic browsers: QuickGO and AmiGO
- QuickGO (EMBL-EBI)
  - https://www.ebi.ac.uk/QuickGO/
  - Nice graphical view for term browsing
  - **Webinar**: https://www.ebi.ac.uk/training/events/quickgo-gene-ontology-annotation/
  - **Tutorial**: https://www.ebi.ac.uk/training/online/courses/goa-and-quickgo-quick-tour/#vf-tabs__section--gettingstarted
- AmiGO (GOC)
  - http://amigo.geneontology.org/amigo
  - Easy access to annotations with faceted search



e.g apoptosis; GO:0006915; ECO:0000314; tropomyosin

Use sets of GO terms (**slims**) that describe your area of interest



Use predefined **templates** to explore Gene Ontology data.

Interactively **search** the Gene Ontology data for annotations, gene products, and terms using a powerful search syntax and filters.

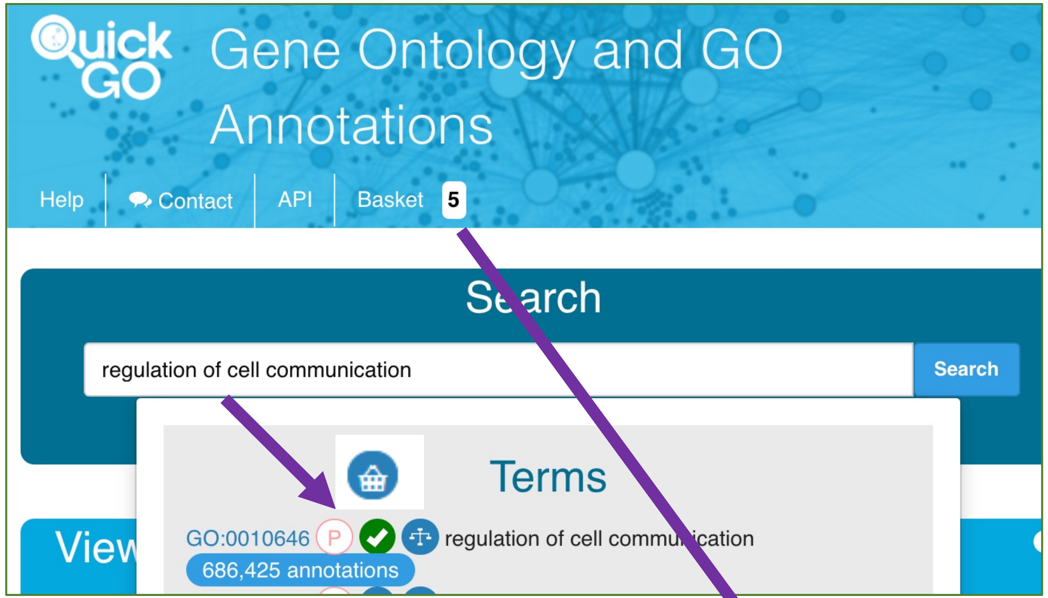
Use the drill-down **browser** to view the ontology structure with annotation counts.

Use GOOSE to query the legacy GO database with **SQL**.

View the most recent **statistics** about the Gene Ontology data in AmiGO.

Many **more tools** are available from the software list, such as alternate searching modes, Visualize, non-JavaScript pages.

QuickGO ancestor view

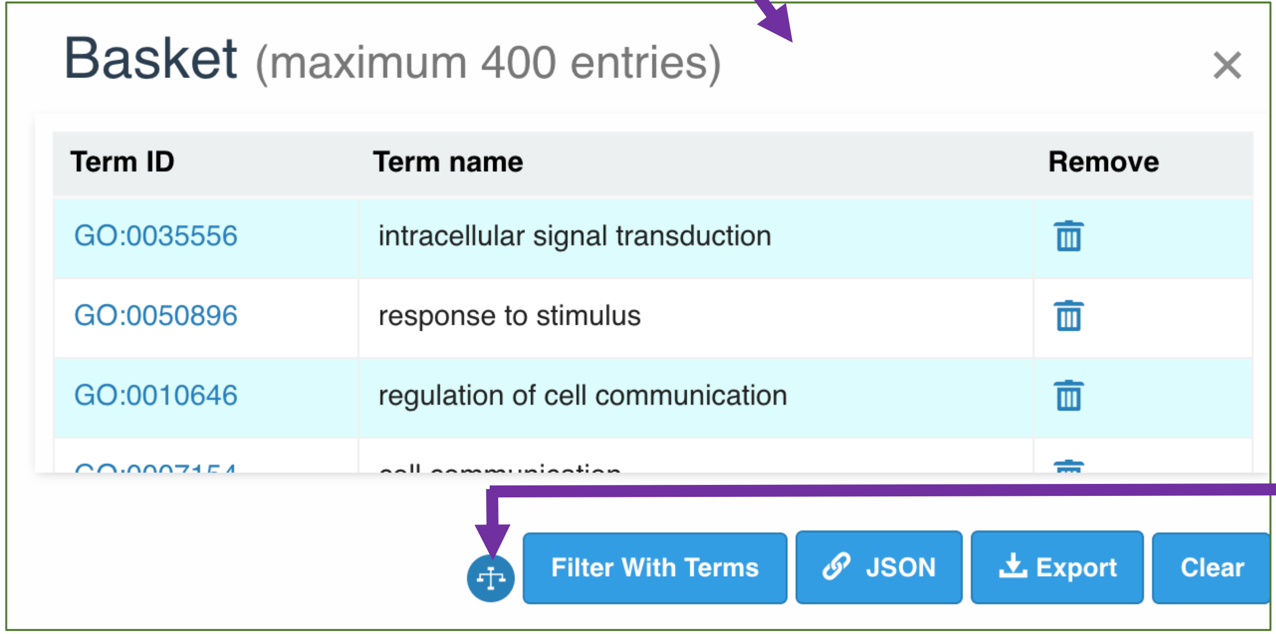


regulation of cell communication

Search

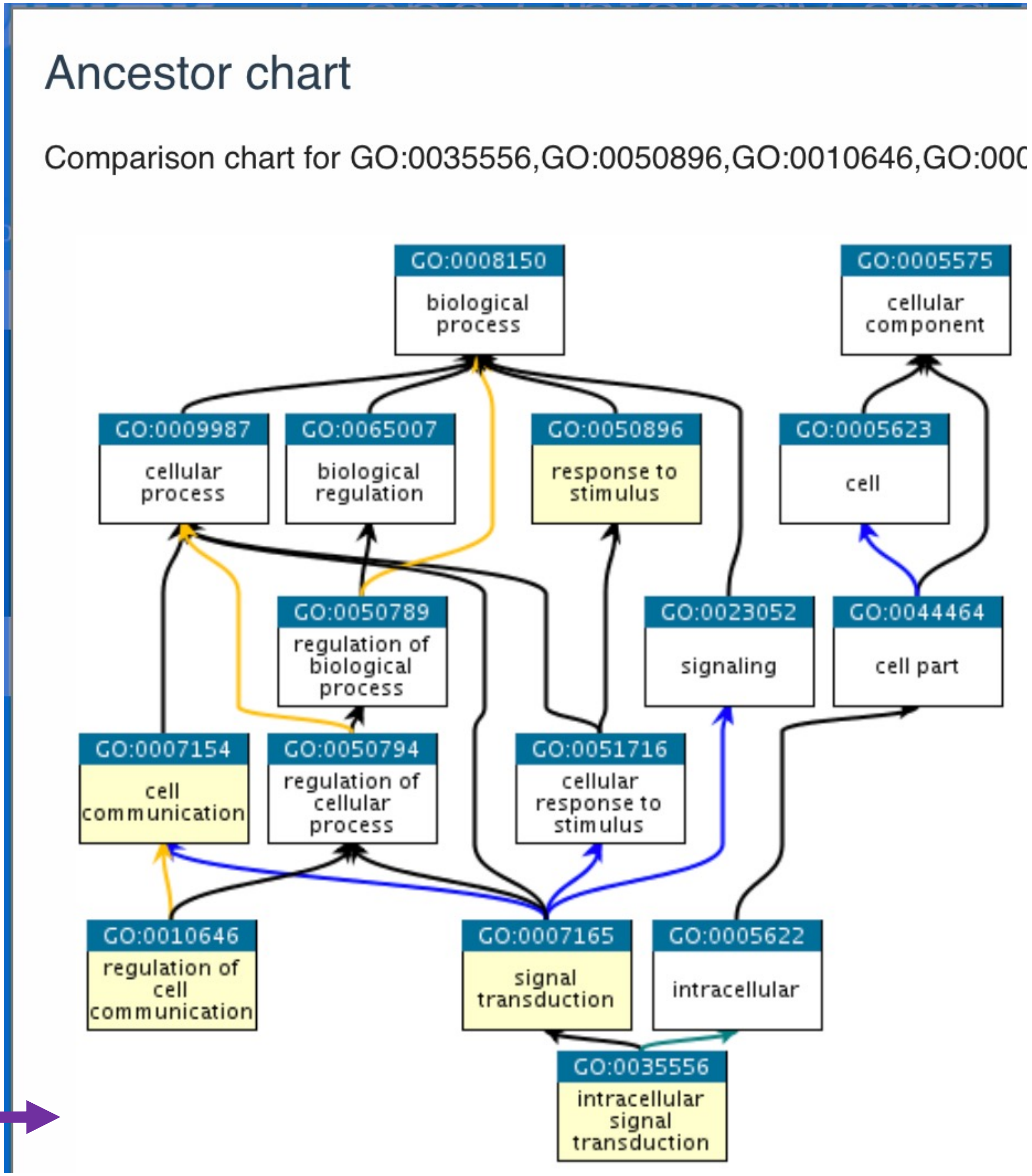
GO:0010646 regulation of cell communication

686,425 annotations



Basket (maximum 400 entries)

Term ID	Term name	Remove
GO:0035556	intracellular signal transduction	🗑
GO:0050896	response to stimulus	🗑
GO:0010646	regulation of cell communication	🗑
GO:0007154	cell communication	🗑

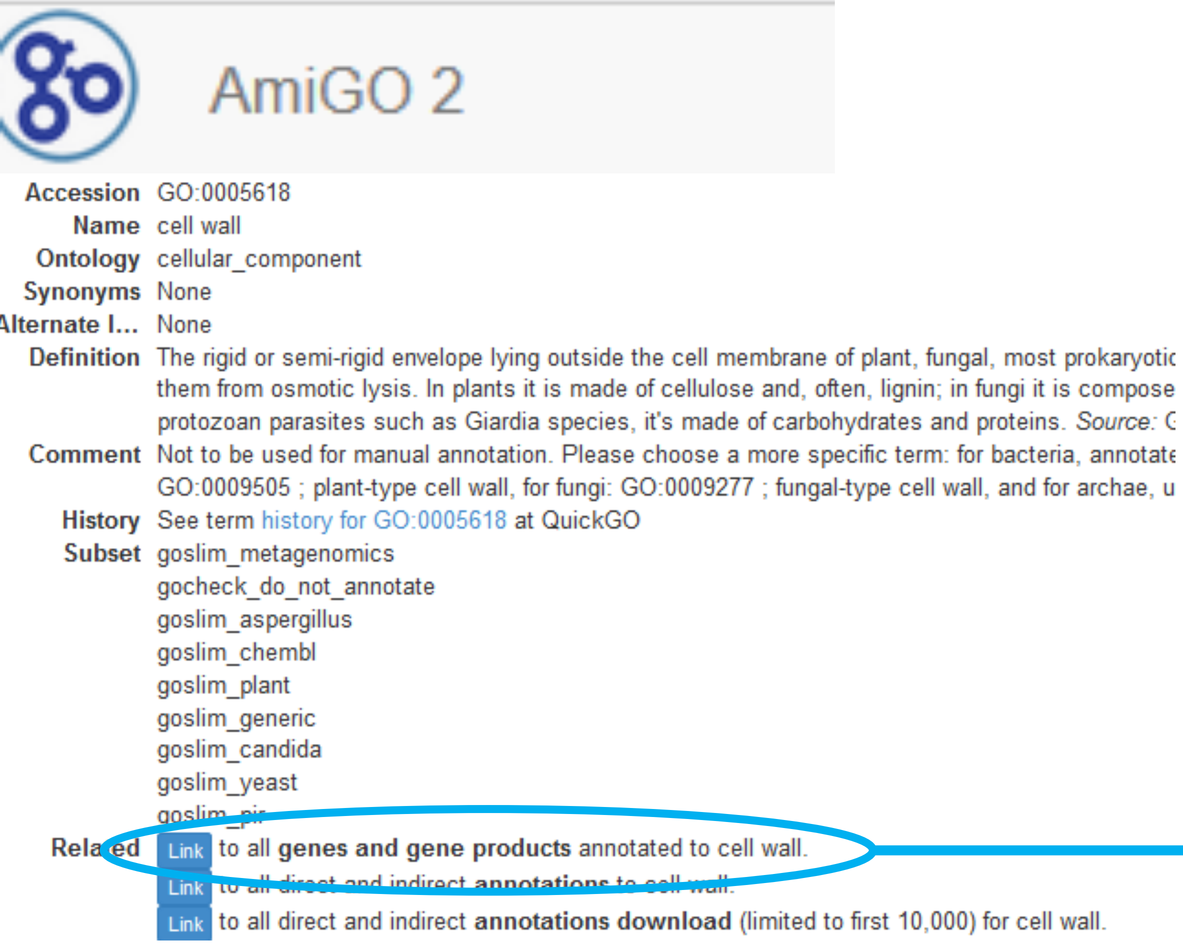


Ancestor chart

Comparison chart for GO:0035556,GO:0050896,GO:0010646,GO:000

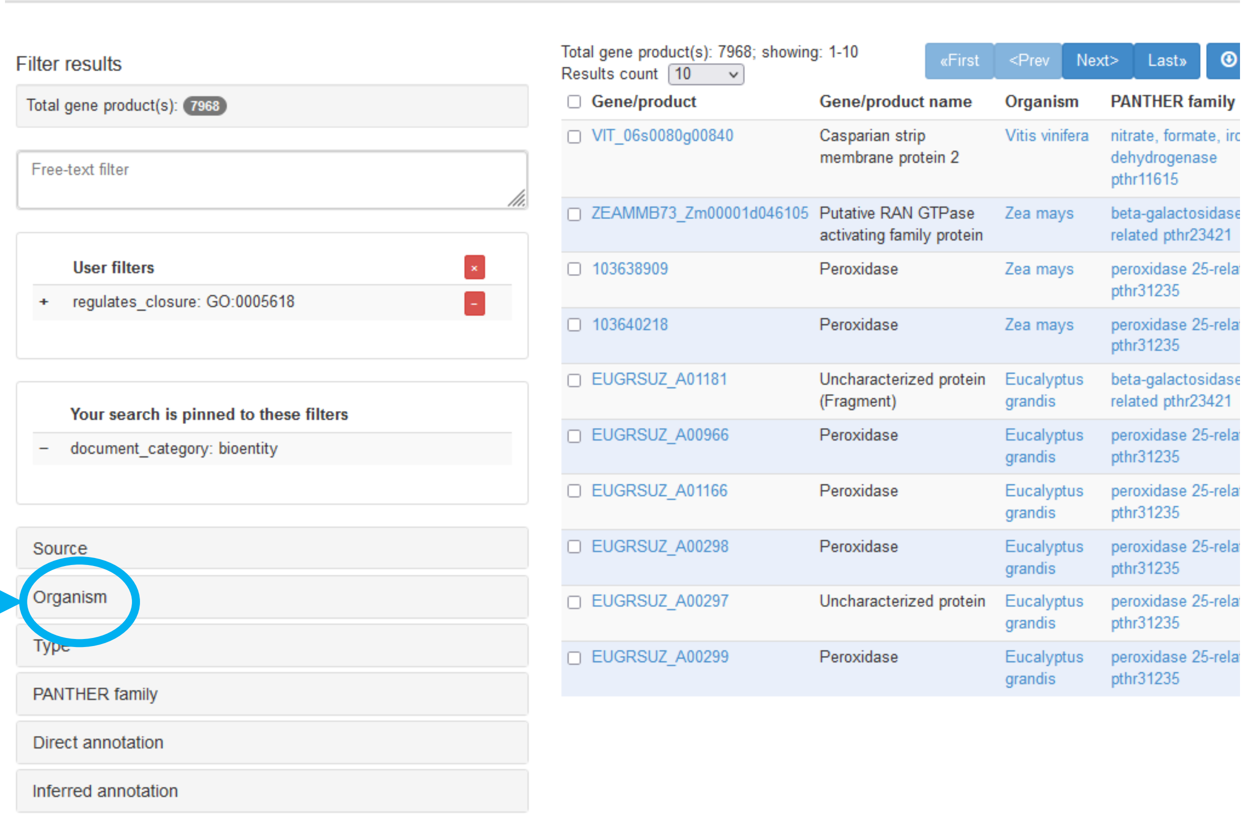
GO:0008150 biological process
GO:0005575 cellular component
GO:0009987 cellular process
GO:0065007 biological regulation
GO:0050896 response to stimulus
GO:0005623 cell
GO:0050789 regulation of biological process
GO:0023052 signaling
GO:0044464 cell part
GO:0007154 cell communication
GO:0050794 regulation of cellular process
GO:0051716 cellular response to stimulus
GO:0010646 regulation of cell communication
GO:0007165 signal transduction
GO:0005622 intracellular
GO:0035556 intracellular signal transduction

AmiGO – species filter



AmiGO 2

Accession	GO:0005618
Name	cell wall
Ontology	cellular_component
Synonyms	None
Alternate I...	None
Definition	The rigid or semi-rigid envelope lying outside the cell membrane of plant, fungal, most prokaryotic them from osmotic lysis. In plants it is made of cellulose and, often, lignin; in fungi it is compose protozoan parasites such as Giardia species, it's made of carbohydrates and proteins. *Source:* G
Comment	Not to be used for manual annotation. Please choose a more specific term: for bacteria, annotate GO:0009505 ; plant-type cell wall, for fungi: GO:0009277 ; fungal-type cell wall, and for archae, u
History	See term history for GO:0005618 at QuickGO
Subset	goslim_metagenomics gocheck_do_not_annotate goslim_aspergillus goslim_chembl goslim_plant goslim_generic goslim_candida goslim_yeast goslim_pir
Related	Link to all **genes and gene products** annotated to cell wall. Link to all direct and indirect **annotations** to cell wall. Link to all direct and indirect **annotations download** (limited to first 10,000) for cell wall.



Filter results

Total gene product(s): 7968

Free-text filter

User filters

+ regulates_closure: GO:0005618

Your search is pinned to these filters

– document_category: bioentity

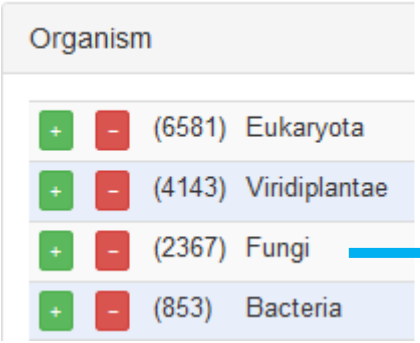
Total gene product(s): 7968; showing: 1-10

Results count 10

□ Gene/product	Gene/product name	Organism	PANTHER family
□ VIT_06s0080g00840	Casparian strip membrane protein 2	Vitis vinifera	nitrate, formate, irc dehydrogenase pthr11615
□ ZEAMMB73_Zm00001d046105	Putative RAN GTPase activating family protein	Zea mays	beta-galactosidase related pthr23421
□ 103638909	Peroxidase	Zea mays	peroxidase 25-rela pthr31235
□ 103640218	Peroxidase	Zea mays	peroxidase 25-rela pthr31235
□ EUGRSUZ_A01181	Uncharacterized protein (Fragment)	Eucalyptus grandis	beta-galactosidase related pthr23421
□ EUGRSUZ_A00966	Peroxidase	Eucalyptus grandis	peroxidase 25-rela pthr31235
□ EUGRSUZ_A01166	Peroxidase	Eucalyptus grandis	peroxidase 25-rela pthr31235
□ EUGRSUZ_A00298	Peroxidase	Eucalyptus grandis	peroxidase 25-rela pthr31235
□ EUGRSUZ_A00297	Uncharacterized protein	Eucalyptus grandis	peroxidase 25-rela pthr31235
□ EUGRSUZ_A00299	Peroxidase	Eucalyptus grandis	peroxidase 25-rela pthr31235

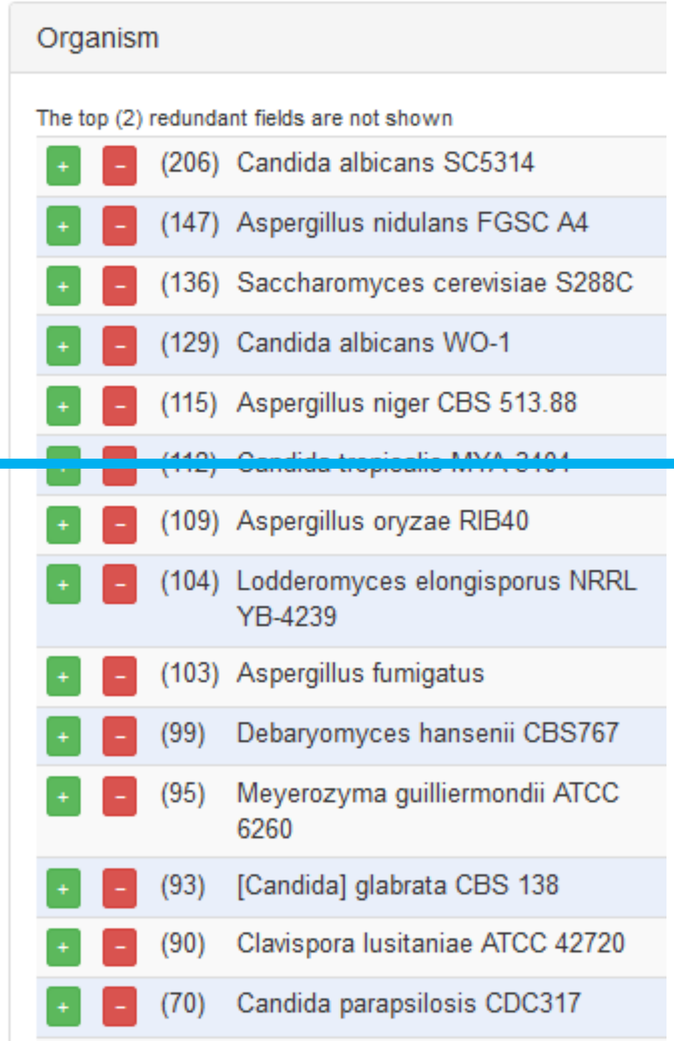
AmiGO – species filter

AmiGO 2



Organism

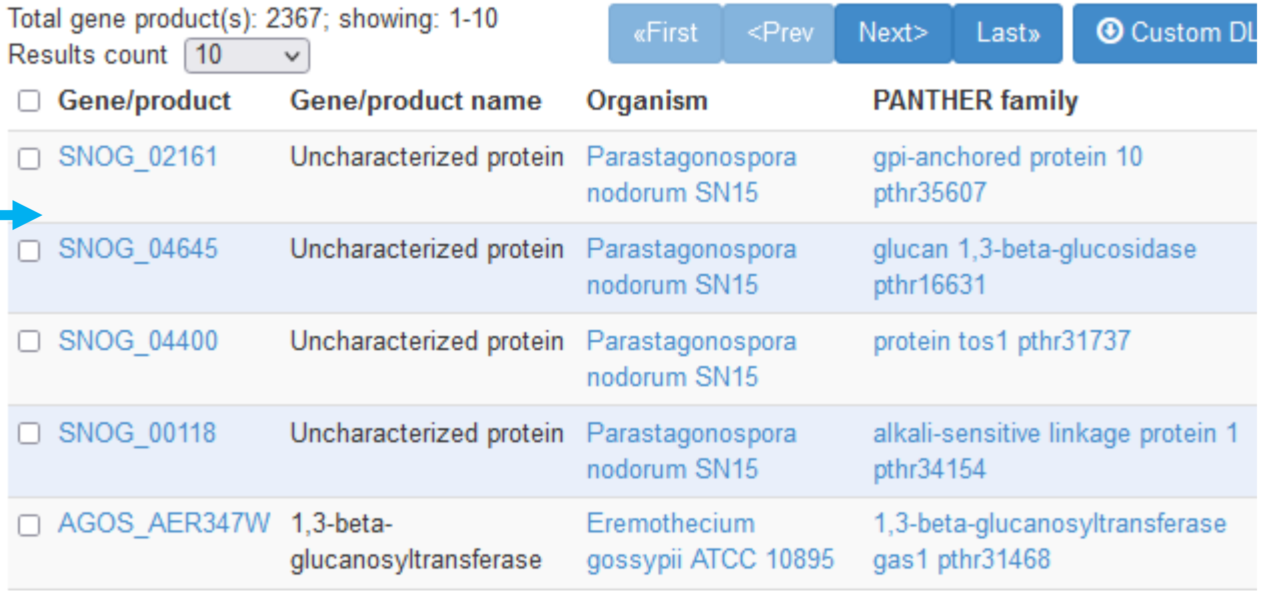
+	−	(6581)	Eukaryota
+	−	(4143)	Viridiplantae
+	−	(2367)	Fungi
+	−	(853)	Bacteria



Organism

The top (2) redundant fields are not shown

+	−	(206)	Candida albicans SC5314
+	−	(147)	Aspergillus nidulans FGSC A4
+	−	(136)	Saccharomyces cerevisiae S288C
+	−	(129)	Candida albicans WO-1
+	−	(115)	Aspergillus niger CBS 513.88
+	−	(112)	Candida tropicalis MYA-3404
+	−	(109)	Aspergillus oryzae RIB40
+	−	(104)	Lodderomyces elongisporus NRRL YB-4239
+	−	(103)	Aspergillus fumigatus
+	−	(99)	Debaryomyces hansenii CBS767
+	−	(95)	Meyerozyma guilliermondii ATCC 6260
+	−	(93)	[Candida] glabrata CBS 138
+	−	(90)	Clavispora lusitaniae ATCC 42720
+	−	(70)	Candida parapsilosis CDC317



Total gene product(s): 2367; showing: 1-10

Results count 10

□	Gene/product	Gene/product name	Organism	PANTHER family
□	SNOG_02161	Uncharacterized protein	Parastagonospora nodorum SN15	gpi-anchored protein 10 pthr35607
□	SNOG_04645	Uncharacterized protein	Parastagonospora nodorum SN15	glucan 1,3-beta-glucosidase pthr16631
□	SNOG_04400	Uncharacterized protein	Parastagonospora nodorum SN15	protein tos1 pthr31737
□	SNOG_00118	Uncharacterized protein	Parastagonospora nodorum SN15	alkali-sensitive linkage protein 1 pthr34154
□	AGOS_AER347W	1,3-beta-glucanosyltransferase	Eremothecium gossypii ATCC 10895	1,3-beta-glucanosyltransferase gas1 pthr31468

Difference between AmiGO and QuickGO (beware)

- Gene-centric or protein-centric worldview
  - In UniProt GOA (and consequently in QuickGO) annotations are made to proteins, and there may be multiple proteins per gene, sometimes representing different isoforms.
  - GO Central omits the majority of the sequences and IEA [electronic] annotations from UniProtKB from the weekly database builds due to the large size of the data set.

Ontology and annotation uses

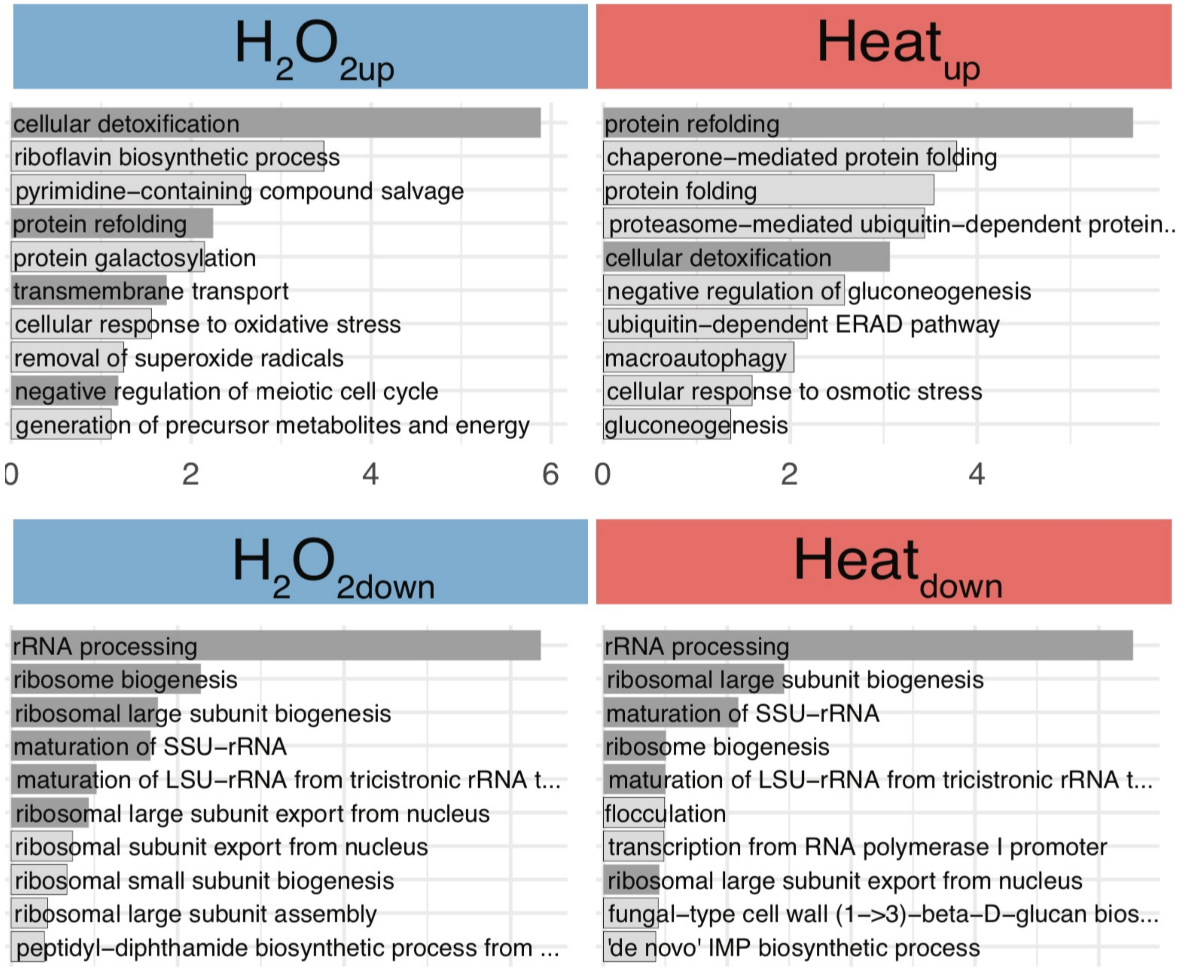
- Aids searching
- Data integration
- Data analysis
- Data visualization

Enrichment analysis

Find significantly overrepresented terms among a list of genes

- What do genes in a gene list have in common?

- Compare the annotations to your gene list with the annotations to the entire gene set to see if any terms are over or under represented compared to the background
  - E.g. 100 genes identified in an experiment, if 4/5 annotated to 'galactose metabolism' are identified, this is highly significant
- An enrichment provides enriched terms and a statistical measure of how likely your genes (result set) fall into that category by chance



Thodberg et. al 2019 PMID:30566651

Components

- **GO annotations**
- **Gene list of interest -** phenotype screen, differentially expressed genes across conditions or WT vs mutant, interacting genes, etc
- **Background** (/reference) **gene set -**Set of all genes that *could* have been detected in the experiment. Do NOT include genes in the background set that couldn't have been identified in the screen
  - e.g. if you identified proteins with a specific protein modification then do not include ncRNA genes in your background
  - e.g. if a microarray only contained a subset of genes, then these are the background
- Some tools:

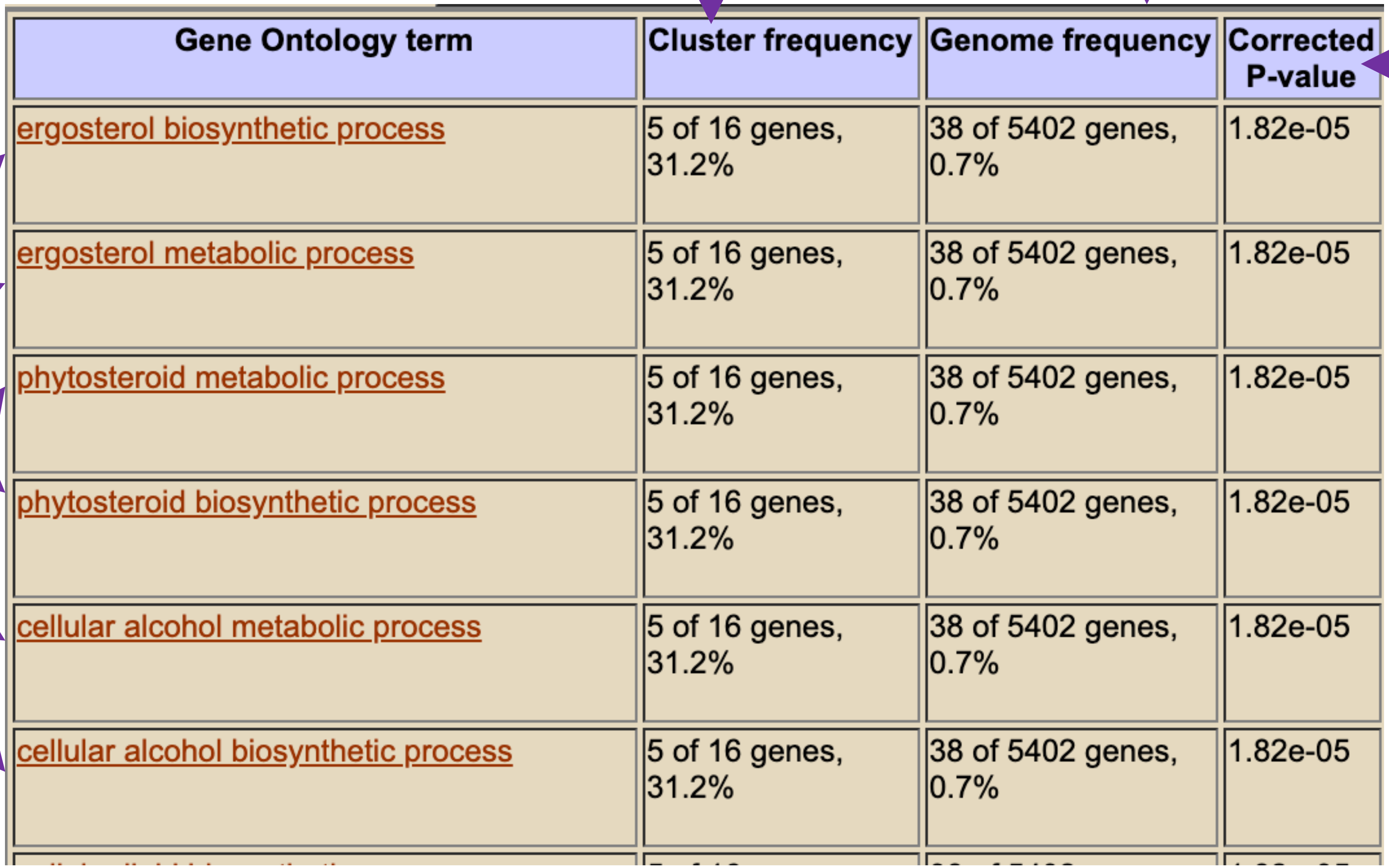
Enrichment Tools

- Many different tools available
  - GO tool (Panther): http://pantherdb.org/webservices/go/overrep.jsp
  - Princeton: http://go.princeton.edu/cgi-bin/GOTermFinder
  - G:profiler: http://biit.cs.ut.ee/gprofiler/gost
- It is worth comparing results from multiple tools (look/feel, agreement, etc)

Results

Number of genes annotated to the GO term in the list of interest

Number of genes annotated to the term in the background set



Gene Ontology term	Cluster frequency	Genome frequency	Corrected P-value
ergosterol biosynthetic process	5 of 16 genes, 31.2%	38 of 5402 genes, 0.7%	1.82e-05
ergosterol metabolic process	5 of 16 genes, 31.2%	38 of 5402 genes, 0.7%	1.82e-05
phytosteroid metabolic process	5 of 16 genes, 31.2%	38 of 5402 genes, 0.7%	1.82e-05
phytosteroid biosynthetic process	5 of 16 genes, 31.2%	38 of 5402 genes, 0.7%	1.82e-05
cellular alcohol metabolic process	5 of 16 genes, 31.2%	38 of 5402 genes, 0.7%	1.82e-05
cellular alcohol biosynthetic process	5 of 16 genes, 31.2%	38 of 5402 genes, 0.7%	1.82e-05

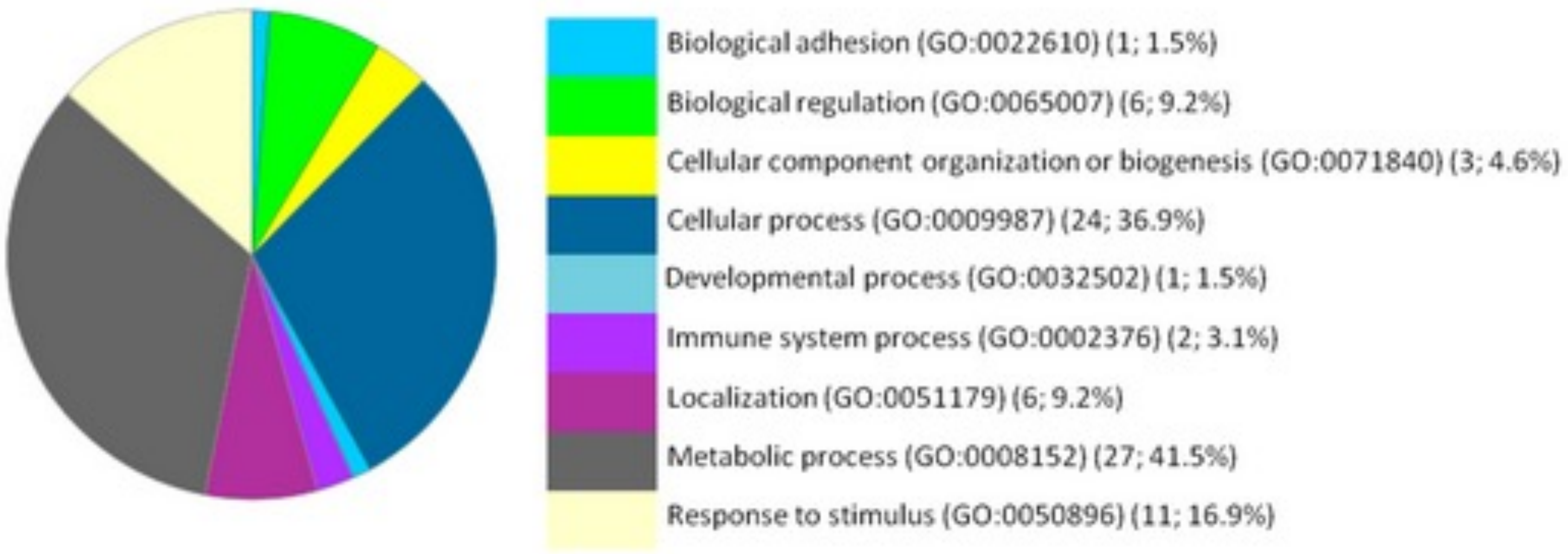
Repetition

The probability of seeing at least X genes in the list of interest annotated to a specific GO term, considering how commonly it is annotated in the background set

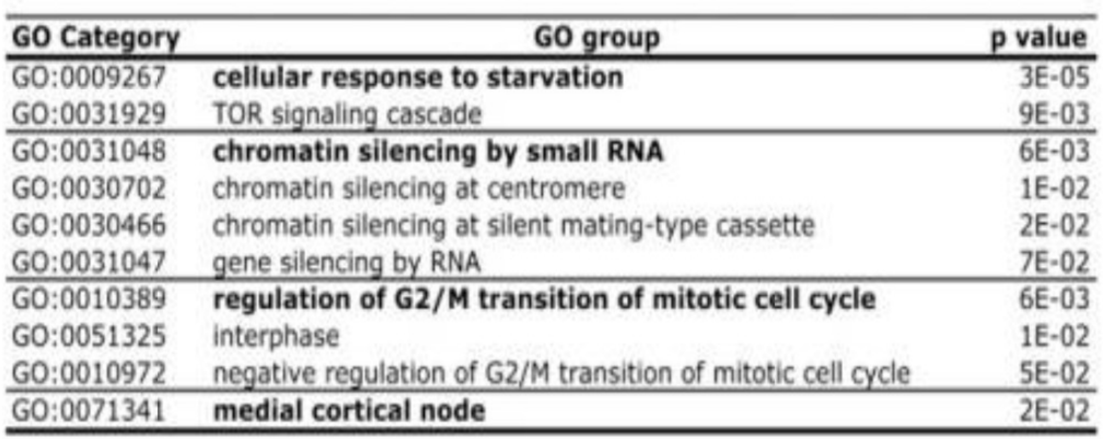
The closer the value is to zero the less likely it is to be observed by chance

Results

- List of terms enriched in data-set
- List needs to be evaluated
  - Physiological relevance?
    - High level terms are often uninformative (for physiological role)
  - Pinpoint ancestor/child terms



- Filter very similar terms out to narrow down the list
- Grouping related terms and concepts

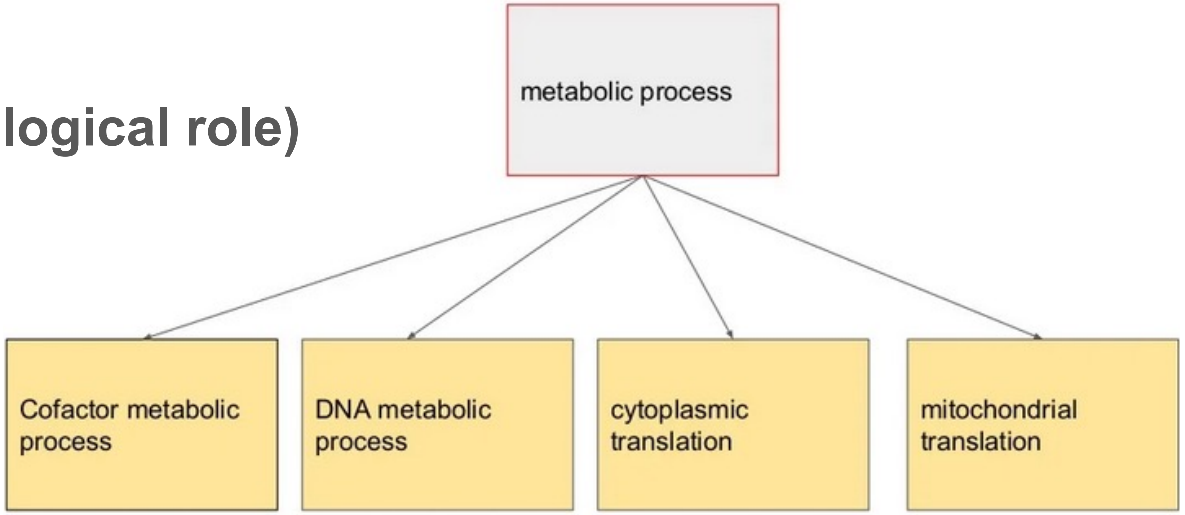


GO Category	GO group	p value
GO:0009267	**cellular response to starvation**	3E-05
GO:0031929	TOR signaling cascade	9E-03
GO:0031048	**chromatin silencing by small RNA**	6E-03
GO:0030702	chromatin silencing at centromere	1E-02
GO:0030466	chromatin silencing at silent mating-type cassette	2E-02
GO:0031047	gene silencing by RNA	7E-02
GO:0010389	**regulation of G2/M transition of mitotic cell cycle**	6E-03
GO:0051325	interphase	1E-02
GO:0010972	negative regulation of G2/M transition of mitotic cell cycle	5E-02
GO:0071341	**medial cortical node**	2E-02

“significant enrichment of four distinct functional groups encompassing the regulation of the cell cycle, cytokinesis, response to nutrients, and chromatin modifications at the centromere” - Bauer et. al 2012 PMID:22768388

High level terms are often uninformative (for physiological role)

- Very high level terms lump unrelated concepts
- Biological “modules” generally do not overlap

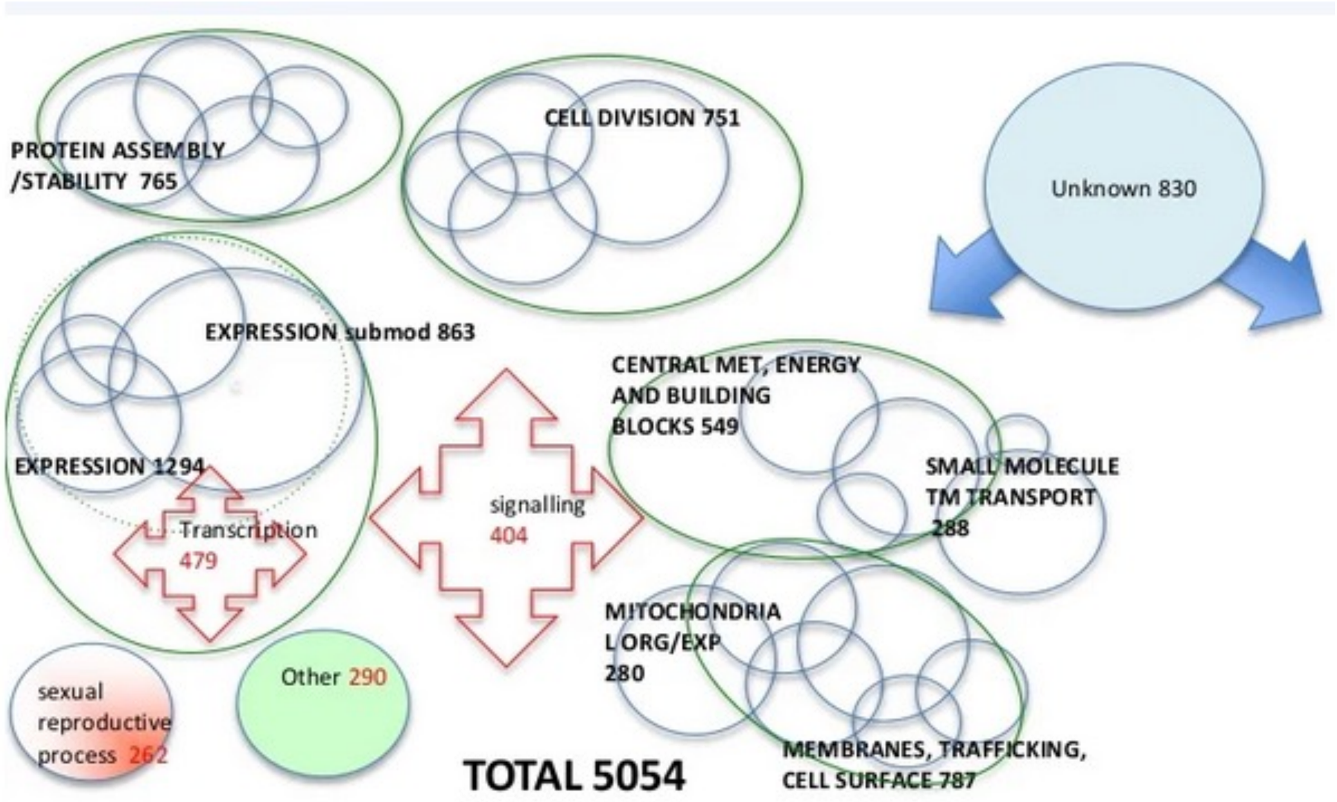

metabolic process

Cofactor metabolic process

DNA metabolic process

cytoplasmic translation

mitochondrial translation


PROTEIN ASSEMBLY /STABILITY 765
CELL DIVISION 751
Unknown 830
EXPRESSION submod 863
EXPRESSION 1294
Transcription 479
signalling 404
CENTRAL MET, ENERGY AND BUILDING BLOCKS 549
SMALL MOLECULE TM TRANSPORT 288
MITOCHONDRIA ORG/EXP 280
sexual reproductive process 262
Other 290
TOTAL 5054
MEMBRANES, TRAFFICKING, CELL SURFACE 787

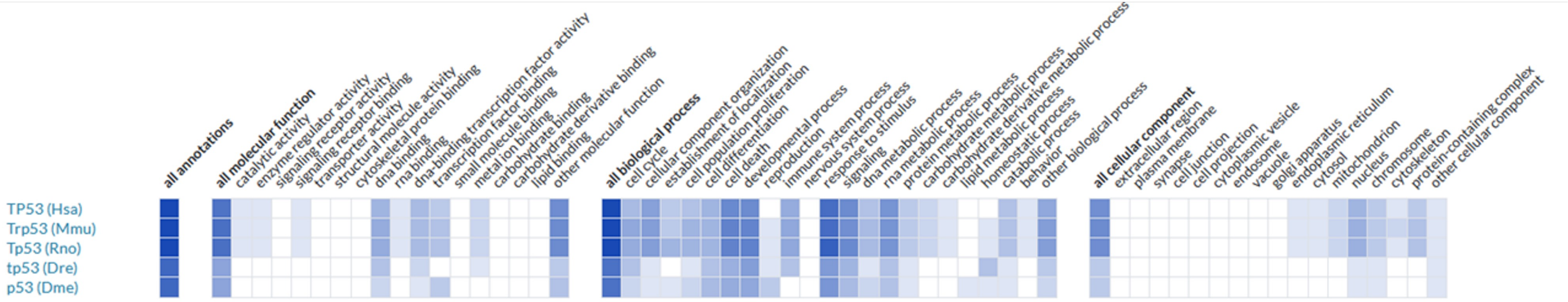
GO subsets (GO slims)

- Mapping granular annotations of a set of genes to high-level (more general) parent terms
  - A coarse-grained overview of detailed annotation
  - Possible because high-level terms are inherited
- Useful for visualizing broad annotation categories for a gene list / entire genome, or to summarize a dataset

Common uses of subsets/slims

1. Whole genome slims

- Summarises an organism’s biology
- A resource for planning curation (or research) – enables identification of unannotated or uncharacterized genes
- Needs to be biologically relevant and redundancy minimized
- Coverage needs to be as good as possible

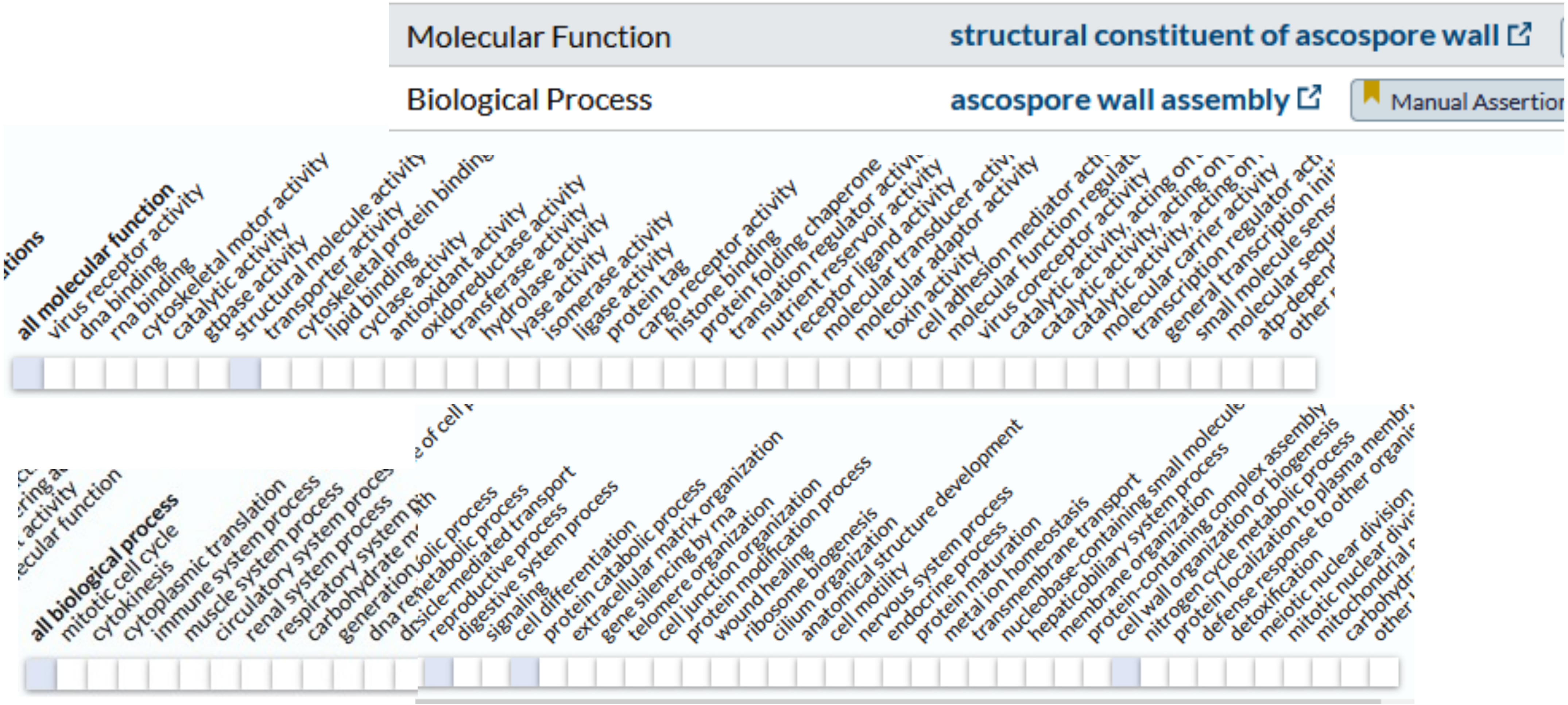


Cell color indicative of annotation volume

https://www.alliancegenome.org/gene/HGNC:11998#function---go-annotations

2. Single gene overview

- E.g. ribbon display
- Indicates branches of GO applicable to gene
- High level grouping terms – redundancy matters less



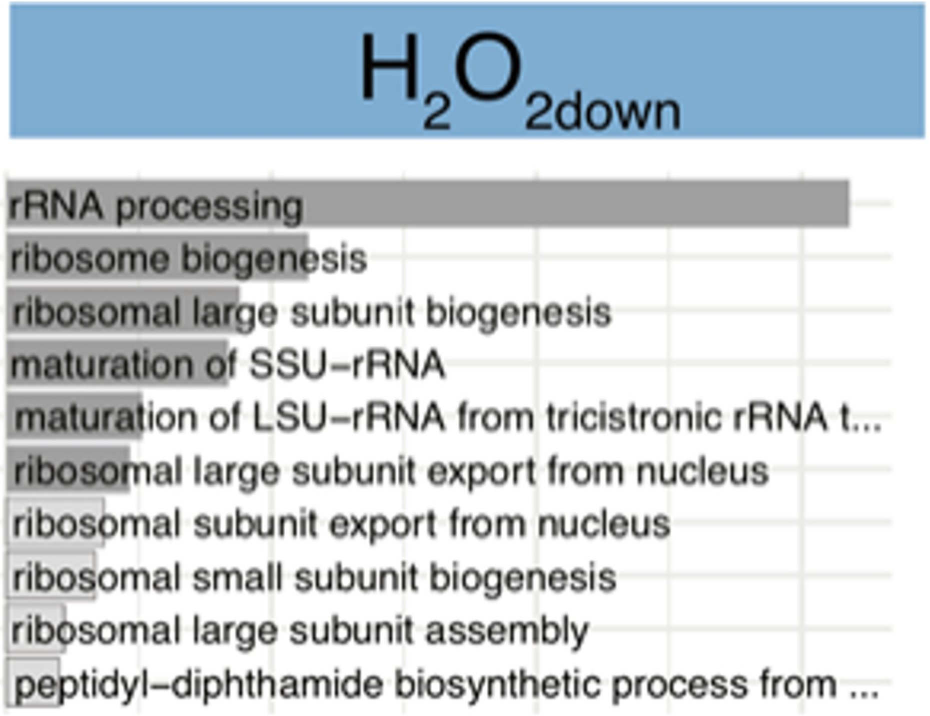
Molecular Function	structural constituent of ascospore wall
Biological Process	ascospore wall assembly

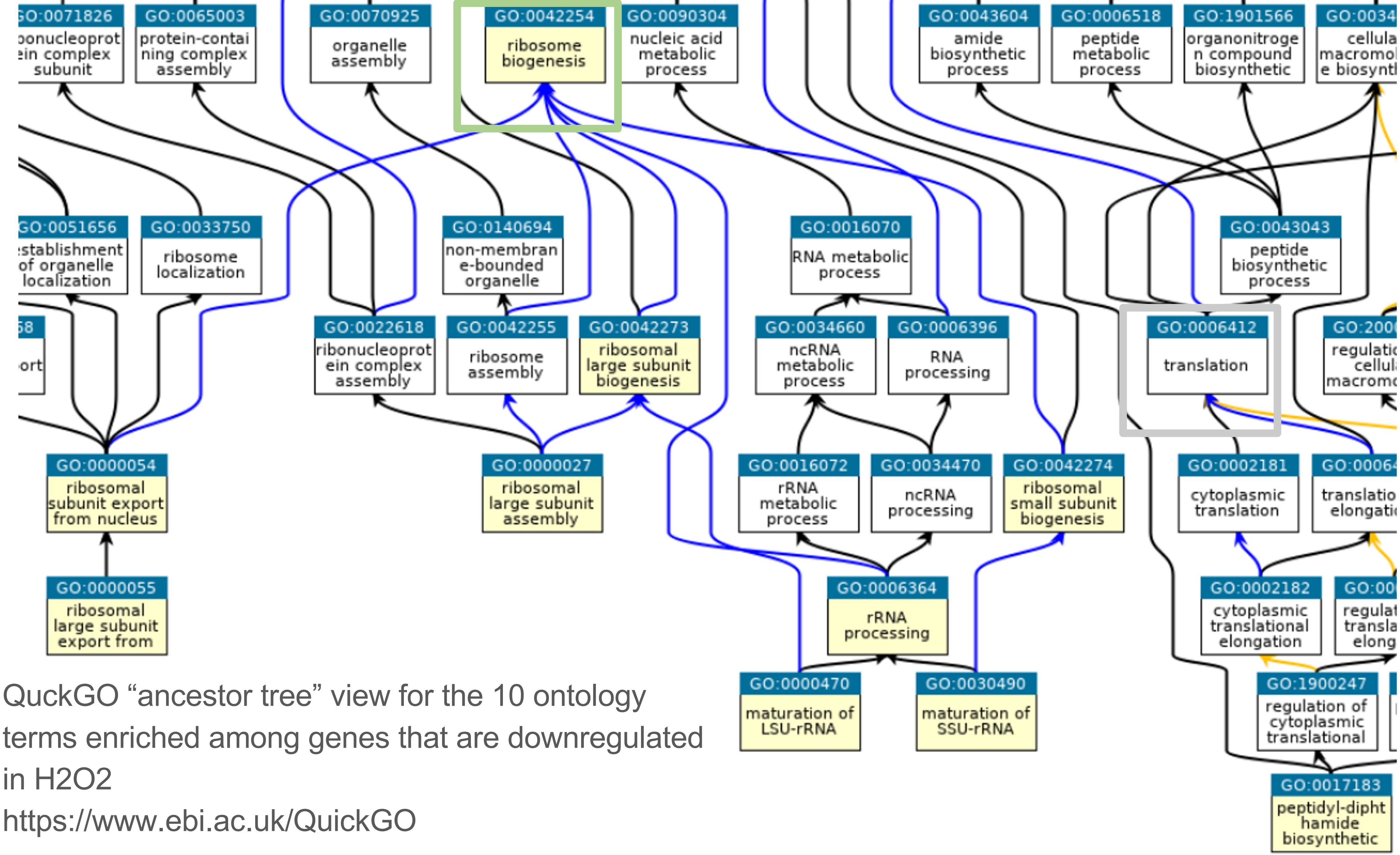
3. Slimming results from an analysis

- To aid interpretation of results (e.g. results from an enrichment)
- Data display/result summary – smallest possible table that will convey the message

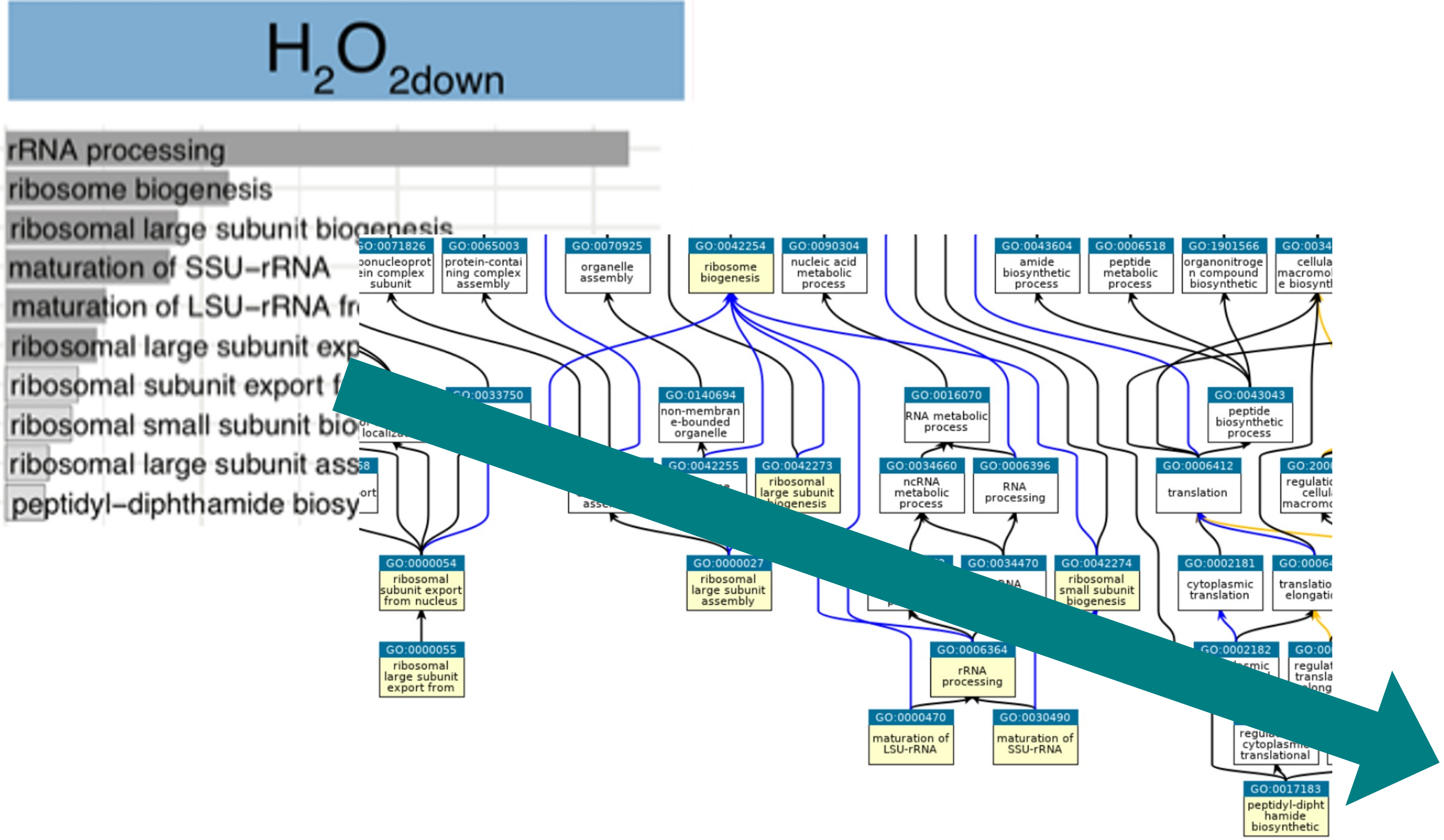
Thodberg et al. (2019) PMID:30566651

- Compared differentially expressed genes across different conditions (genes downregulated in hydrogen peroxide shown)




GO:0071826 ibonucleoprot ein complex subunit
GO:0065003 protein-containing complex assembly
GO:0070925 organelle assembly
GO:0042254 ribosome biogenesis
GO:0090304 nucleic acid metabolic process
GO:0043604 amide biosynthetic process
GO:0006518 peptide metabolic process
GO:1901566 organonitrogen compound biosynthetic
GO:0034 cellula macromol e biosynt
GO:0051656 establishment of organelle localization
GO:0033750 ribosome localization
GO:0140694 non-membrane-bounded organelle
GO:0016070 RNA metabolic process
GO:0043043 peptide biosynthetic process
58 ort
GO:0022618 ribonucleoprotein complex assembly
GO:0042255 ribosome assembly
GO:0042273 ribosomal large subunit biogenesis
GO:0034660 ncRNA metabolic process
GO:0006396 RNA processing
GO:0006412 translation
GO:200 regulati cellul macromo
GO:0000054 ribosomal subunit export from nucleus
GO:0000027 ribosomal large subunit assembly
GO:0016072 rRNA metabolic process
GO:0034470 ncRNA processing
GO:0042274 ribosomal small subunit biogenesis
GO:0002181 cytoplasmic translation
GO:00064 translatio elongati
GO:0000055 ribosomal large subunit export from
GO:0006364 rRNA processing
GO:0002182 cytoplasmic translational elongation
GO:00 regulat transla elong
GO:0000470 maturation of LSU-rRNA
GO:0030490 maturation of SSU-rRNA
GO:1900247 regulation of cytoplasmic translational
GO:0017183 peptidyl-diphthamide biosynthetic

- QuckGO “ancestor tree” view for the 10 ontology terms enriched among genes that are downregulated in H2O2
- https://www.ebi.ac.uk/QuickGO



H2O2down
rRNA processing
ribosome biogenesis
ribosomal large subunit biogenesis
maturation of SSU−rRNA
maturation of LSU−rRNA fr
ribosomal large subunit exp
ribosomal subunit export f
ribosomal small subunit bio
ribosomal large subunit ass
peptidyl−diphthamide biosy
GO:0071826 ribonucleoprotein complex subunit
GO:0065003 protein-containing complex assembly
GO:0070925 organelle assembly
GO:0042254 ribosome biogenesis
GO:0090304 nucleic acid metabolic process
GO:0043604 amide biosynthetic process
GO:0006518 peptide metabolic process
GO:1901566 organonitrogen compound biosynthetic
GO:0034 cellular macromolecule biosynth
GO:0033750
GO:0140694 non-membrane-bounded organelle
GO:0016070 RNA metabolic process
GO:0043043 peptide biosynthetic process
GO:0042255
GO:0042273 ribosomal large subunit biogenesis
GO:0034660 ncRNA metabolic process
GO:0006396 RNA processing
GO:0006412 translation
GO:200 regulation cellular macromo
GO:0000054 ribosomal subunit export from nucleus
GO:0000027 ribosomal large subunit assembly
GO:0034470
GO:0042274 ribosomal small subunit biogenesis
GO:0002181 cytoplasmic translation
GO:0006 translation elongation
GO:0000055 ribosomal large subunit export from
GO:0006364 rRNA processing
GO:0002182
GO:0000470 maturation of LSU-rRNA
GO:0030490 maturation of SSU-rRNA
regulation cytoplasmic translational
GO:0017183 peptidyl-diphthamide biosynthetic

Take the knowledge of the structure of the ontology to “slim down” the results for condensed data-display

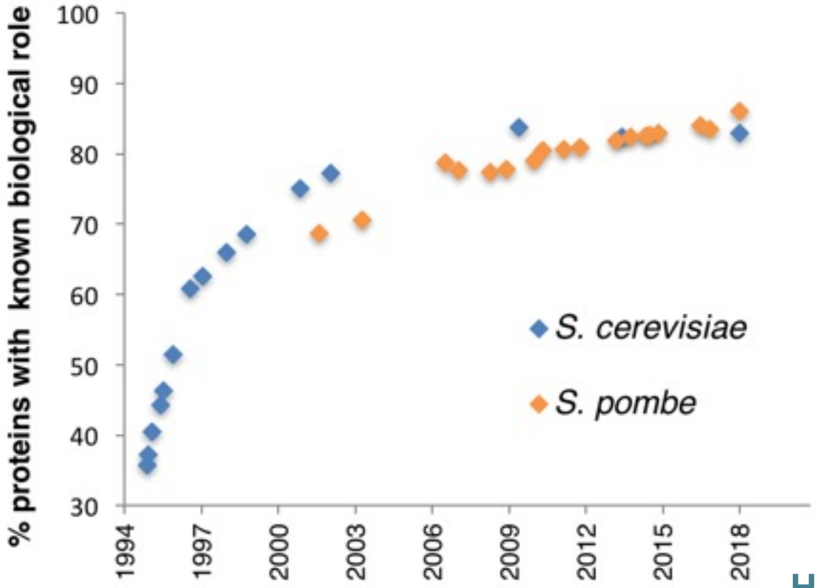
H_2O_{2down}

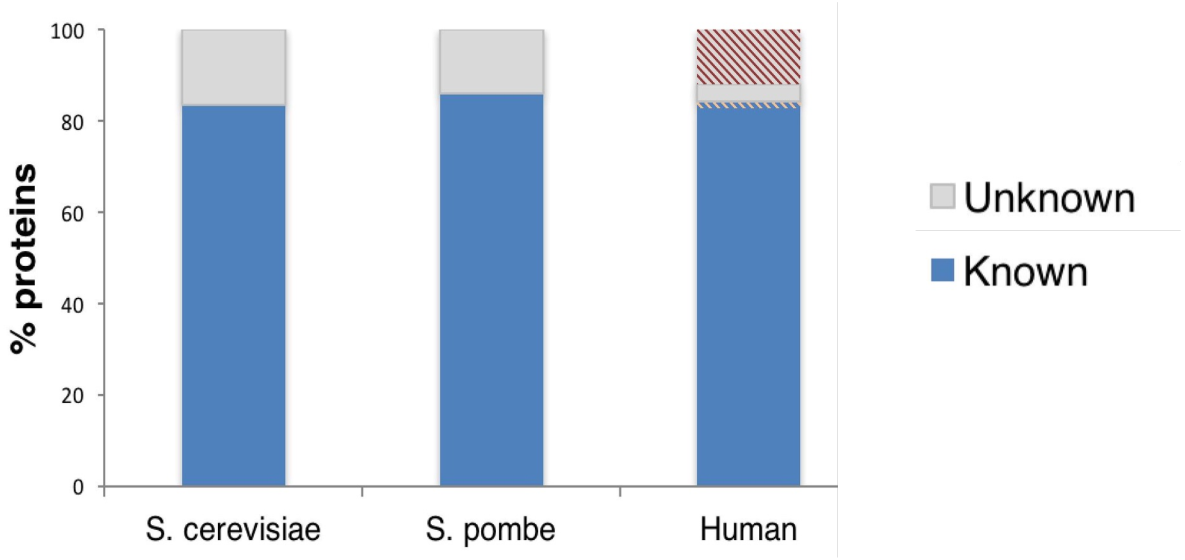
- Ribosome biogenesis
- Translation

- Ready made subsets are available at the GO site
  - Includes (among others) species-specific genome subsets as well as a generic GO subset
  - http://geneontology.org/docs/download-ontology/#subsets
- No “one size fits all” - you can create your own GO slim with terms of interest
  - http://go.princeton.edu/cgi-bin/GOTermMapper
  - Map2Slim script https://github.com/owlcollab/owltools/wiki/Map2Slim

Slimming tips

- Only slim one aspect (molecular function, biological process or cellular component) at a time
- Minimise genes that do not slim – **unannotated, unknown, unslimmed?**





Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? PMID:30938578

New gene characterization is slow!

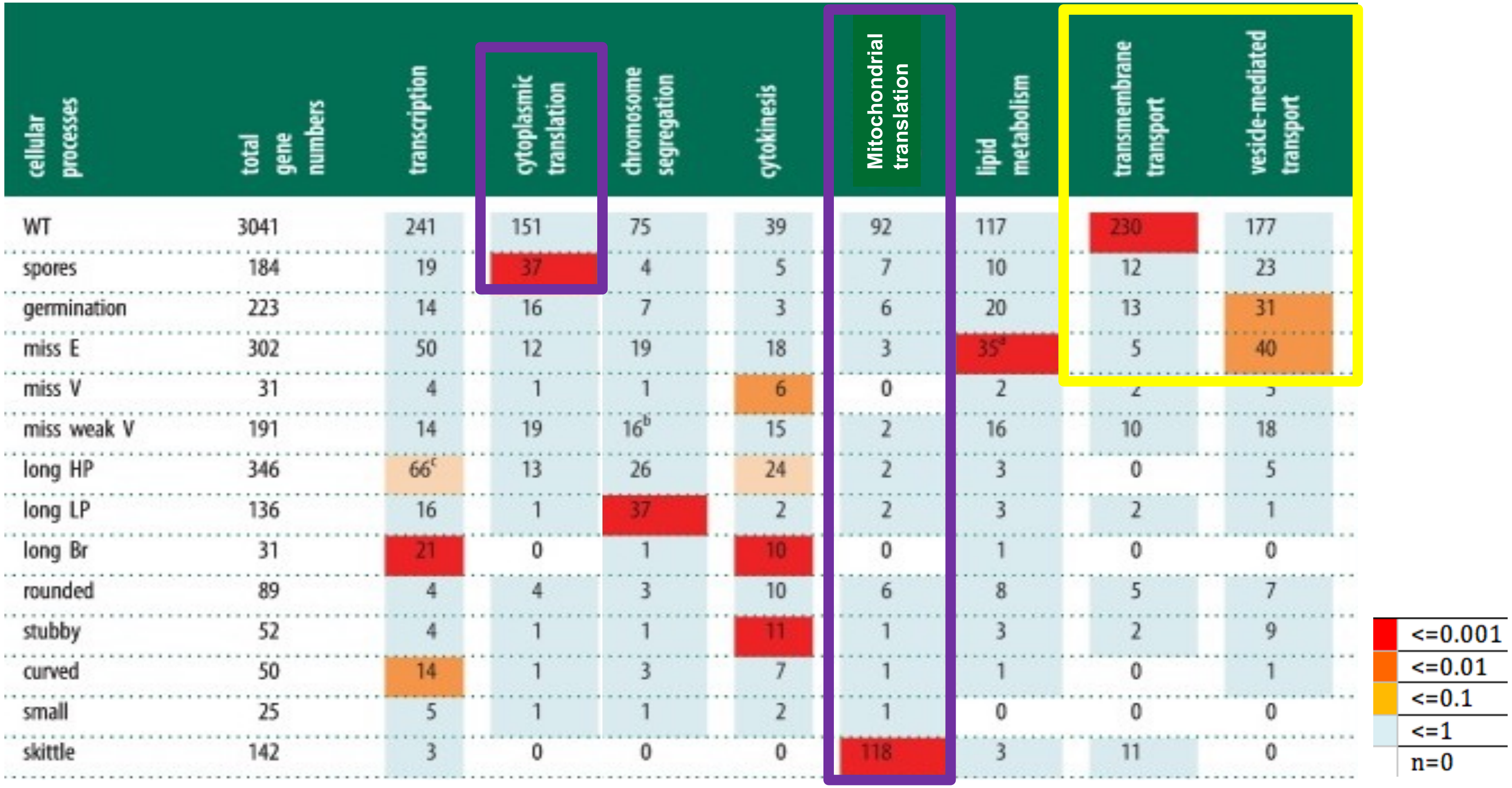
20% of genes are still unknown Biological Process (- do not slim to any ‘biologically informative’ genome slim terms)

70% of human research is on proteins known pre-genome sequencing (PMID:21307913)

Slimming tips

- Only slim one aspect (molecular function, biological process or cellular component) at a time
- Minimise genes that do not slim – **unannotated, unknown, unslimmed?**
- Terms must be picked manually
  - There are no readily identifiable ‘levels’ in the ontology
  - Avoid going too high (terms that slim most genes - biologically uninformative)
  - Lump vs. split – Some terms have a common parent in the ontology, but are largely unconnected in terms of gene sets
    - Nuclear gene expression vs. mitochondrial gene expression
    - Vesicle-mediated transport vs. nucleocytoplasmic transport

Lump vs. split, real example



cellular processes	total gene numbers	transcription	cytoplasmic translation	chromosome segregation	cytokinesis	Mitochondrial translation	lipid metabolism	transmembrane transport	vesicle-mediated transport
WT	3041	241	151	75	39	92	117	230	177
spores	184	19	37	4	5	7	10	12	23
germination	223	14	16	7	3	6	20	13	31
miss E	302	50	12	19	18	3	35[d]	5	40
miss V	31	4	1	1	6	0	2	2	3
miss weak V	191	14	19	16[b]	15	2	16	10	18
long HP	346	66[c]	13	26	24	2	3	0	5
long LP	136	16	1	37	2	2	3	2	1
long Br	31	21	0	1	10	0	1	0	0
rounded	89	4	4	3	10	6	8	5	7
stubby	52	4	1	1	11	1	3	2	9
curved	50	14	1	3	7	1	1	0	1
small	25	5	1	1	2	1	0	0	0
skittle	142	3	0	0	0	118	3	11	0

🟥	<=0.001
🟧	<=0.01
🟨	<=0.1
▨	<=1
	n=0

Adapted from Hayles et al 2013 PMID:23697806

- Genome-wide deletion mutant phenotype screen
- GO analysis to identify genes annotated to cellular processes enriched within phenotype categories.
- The enrichment results were mapped to ‘GO slim’ (high level) terms

Some phenotypes (horizontal) are significantly enriched for specific GO processes (vertical)

Problems & pitfalls

- **Versioning**: GO (the ontology) and GO annotations are changing daily, **check the date stamp** on the ontology and the dataset you are using
- Not all genomes are comprehensively annotated
- ***Unknowns*** are ignored by enrichment tools, and these might be interesting genes in your list
- Be aware of evidence codes – e.g. include RCA (computational predictions) – more false positives
- Don't exclude electronic annotation codes (important for unannotated genes)
    - Analysis shows that they are as accurate as manual annotation, just not as specific (PMID:22693439)
- **Enrichment:** use the correct background set (the set of genes in your "experiment") - If you used the whole protein set you need to filter for ncRNAs…
- **Slimming:** GO annotations are NOT mutually exclusive, therefore it doesn't make sense to put in a pie chart and show as distribution as percentages (it is an annotation number total, not a gene product count)
- **Slimming:** Consider genes that are annotated but not to terms in your slim set (you may need to define a better slim set)

Further reading
PMID:18475267 Use and misuse of the Gene Ontology
PMID: 24244145 Ten quick tips for using the gene ontology.

GO help and contact

- GO helpdesk (questions/feedback): http://help.geneontology.org
- FAQ: http://geneontology.org/docs/faq/
- Documentation: http://geneontology.org/search.html
- Technical questions and accessing GO codes: https://github.com/geneontology
- GO REST API (support common queries to GO): http://api.geneontology.org/api
- GO SPARQL (support common queries to GO-CAMs): http://sparql.geneontology.org

What about genomes with no GO annotation?

- Currently, GO recommends groups submit their transcriptomes to NCBI. These submissions will reach UniProt, where InterPro2GO automatically creates GO annotations. These annotations, made with the IEA evidence codes (Inferred from Electronic Annotation), will be made available in a future GO release.
- GO does not recommend groups create their own IEAs with internal tools due to reproducibility and accuracy concerns.

http://geneontology.org/docs/faq/#how-do-i-annotate-a-novel-genome-with-go-annotations

But what if people want to summarize it's functional capabilities using GO as part of a publication? I would recommend:

1. For a very basic first pass run interProscan and then use InterPro to GO mappings.
2. 2. If there is a well annotated model, and they can generate a good orthology mapping they can do direct transfer - similar to what we did for japonicusDB https://academic.oup.com/genetics/advance-article/doi/10.1093/genetics/iyab223/6481558?login=false
3. 3. For high quality annotations use PAINT and tree grafter (best way to get specific annotation if there is no close relative with good annotation coverage) https://academic.oup.com/bioinformatics/article/35/3/518/5056037
4. None are trivial, but presumably a project of that sort would have some provision for bioinfomatics support.
5. Personally I would not trust any tools which used blast or similar but others may have recommendations of tools that work well.