

UBC Bioinformatics

Topic 8: Functional annotation and ontologies

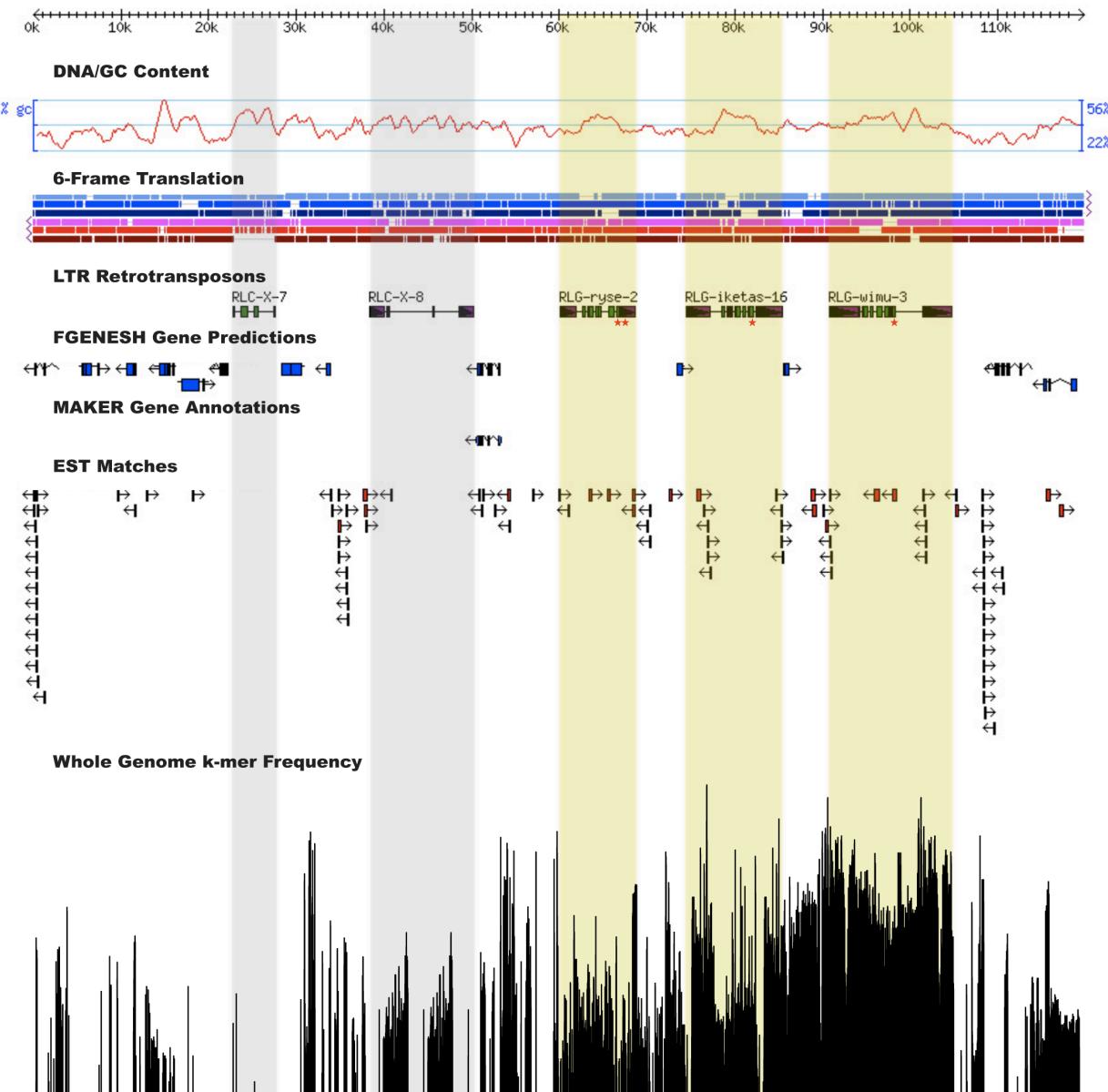
Determine the goal of the study

Do we want to create and share gene models?

Feature descriptions are from combined evidence:

- *ab initio* predictions
- Transcriptome information
- Legacy annotations...

A mapping of all genomic features



Determine the goal of the study

Do we want to create and share gene models?

Feature descriptions are from combined evidence:

- *ab initio* predictions
- Transcriptome information
- Legacy annotations...

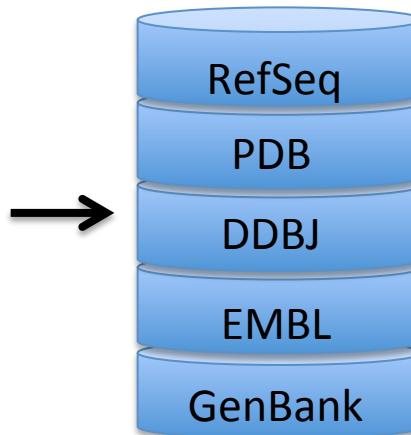
Do we want to annotate a transcript or gene set?

- Assign function to a set of genes
- Place genes in a pathway
- Determine enrichment for a set of functional terms
- Determine change in functional categories

Generate descriptions based on similarity (BLAST)

- First steps: search for homology using BLAST

```
>O14746
MPPAPRCAVRSLLRSHYREVPLATFVRLGPQGWRLVQRCDPAAFRALVAQCLVCVPW
DARPPPAAPSFROVSCILKEVLVARVLQLCERGAKNVLAFGFALLDGARGCPPEAFTTSVR
SYLPNTVTDLRGSGAWGLLRRVGDDVLVHLLARCALFVIVAPSCAYQVCGPPLYQLGA
ATQARPPPHASGPRRRLGCERAWNHSVREAGVPLGLPAPGARRRGGSASRSLLPKRPRR
GAAPEPERTPVQGQSWAHPGTRGPSSDRGFCCVSPARPAEEATSLEGALSCTRISHPSVG
RQHHAGPSTSRPGRPNDTPCPPVVAETKHFLYSSGDKEQQLRPSFLSSLRPSLTGARRL
VETIFLGSRPWMPCTPRLPQRYWQMRPLFLELLGNHAQCPYGVLKTHCPRLRAAVT
PAAGVCAREKPQGSVAPEEEDTDPRRLVQJLLRQHSSPVWVYGFVRACRLRVPGCLWGS
RHNERRFLRNTKFISLGKHAKLSLQELTWKMSVRDCANLRSRGVGVCPAAEHLRREEI
LAKFLHWLMSVYVVELLRSFFYVTTTQKNRLLFYRKSVWSKLQSIGIRQHLKRVQLRE
LSEAEVQRQHREARPALLTSRRLRFIPKPDGLRPIVNMVDYVGARTFRREKERTSRVKA
LFSQLNYERARRPCLLGASVGLDDIHRAWRTFVLRVRAQDPPELYFVKVDVTGAYDTI
PQDRLTEVIASIIKPQNQTYCVRRYAVVOKAAGHGVRKAFKSHVSTLTDLOPYMRQFVAHL
QETSPRLDAVIEQSSSLNEASSGLFDVFLRFMCHHAHVIRGKSYVQCQGIPQGSILSTL
LCSLCYGDMEENKLFACTIRDCCLLRLVDFLLVTBHLTHAKTFLRLTVRGVPYEYGVNL
RKTUVNPVDEALGGTAFVQMPAHGFWCCLLDTRTLLEVSDYSSYARTSIRASLTF
NRGFKA GRNMRRLKGVLRLKCHSLFLDLQVNSLQTVCNTIYKILLQAYRFIAVCLQLP
FHQQVWKNPITFLRVISDTASLCYSILKAKNAGMSLGAKGAAGPLPSEAVQWLCHQAFLL
KLTRHRVTVPLLGLSLRATAQTLQSLRKLPGTTLTALEAAAANPALPSDFKTLID
```



telomerase reverse transcriptase isoform 1 [Homo sapiens]
Sequence ID: [ref|NP_937983.2|](#) Length: 1132 Number of Matches: 1
► See 6 more title(s)

Range 1: 1 to 1132 GenPept Graphics				
Score	Expect Method	Identities	Positives	Gaps
2294 bits(5945) 0.0	Compositional matrix adjust. 1132/1132(100%) 1132/1132(100%) 0/1132(0%)			
Query 1	MPPAPRCAVRSLLRSHYREVPLATFVRLGPQGWRLVQRCDPAAFRALVAQCLVCVPW	60		
Sbjct 1	MPPAPRCAVRSLLRSHYREVPLATFVRLGPQGWRLVQRCDPAAFRALVAQCLVCVPW	60		
Query 61	DARPPPAAPSFQVSCLEVLVARVLQLCERGAKNVLAFGFALLDGARGCPPEAFTTSVR	120		
Sbjct 61	DARPPPAAPSFQVSCLEVLVARVLQLCERGAKNVLAFGFALLDGARGCPPEAFTTSVR	120		
Query 121	SYLPNTVTDLRGSGAWGLLRRVGDDVLVHLLARCALFVIVAPSCAYQVCGPPLYQLGA	180		
Sbjct 121	SYLPNTVTDLRGSGAWGLLRRVGDDVLVHLLARCALFVIVAPSCAYQVCGPPLYQLGA	180		
Query 181	ATQARPPPHASGPRRRLCERAWNHSVREAGVPLGLPAPGARRGGGSASRSLPLPKRPR	240		
Sbjct 181	ATQARPPPHASGPRRRLCERAWNHSVREAGVPLGLPAPGARRGGGSASRSLPLPKRPR	240		
Query 241	GAAPPERTPVQGQSWAHPGTRGPSDRGFCVVSPARPAEEATSLEGALSCTRISHPSVG	300		
Sbjct 241	GAAPPERTPVQGQSWAHPGTRGPSDRGFCVVSPARPAEEATSLEGALSCTRISHPSVG	300		
Query 301	RQHHAGPSTSRPGRPFWDTCPCPVYAEATHFLYSSGDKEQQLRPSFLLSLRPSLTGARRL	360		
Sbjct 301	RQHHAGPSTSRPGRPFWDTCPCPVYAEATHFLYSSGDKEQQLRPSFLLSLRPSLTGARRL	360		

Generate descriptions based on similarity (BLAST)

- First steps: search for homology using BLAST

Caveats:

- What is the evidence for the match?
- What is the source of the match?
- How was the annotation determined?
- How to get phylogenomic information?

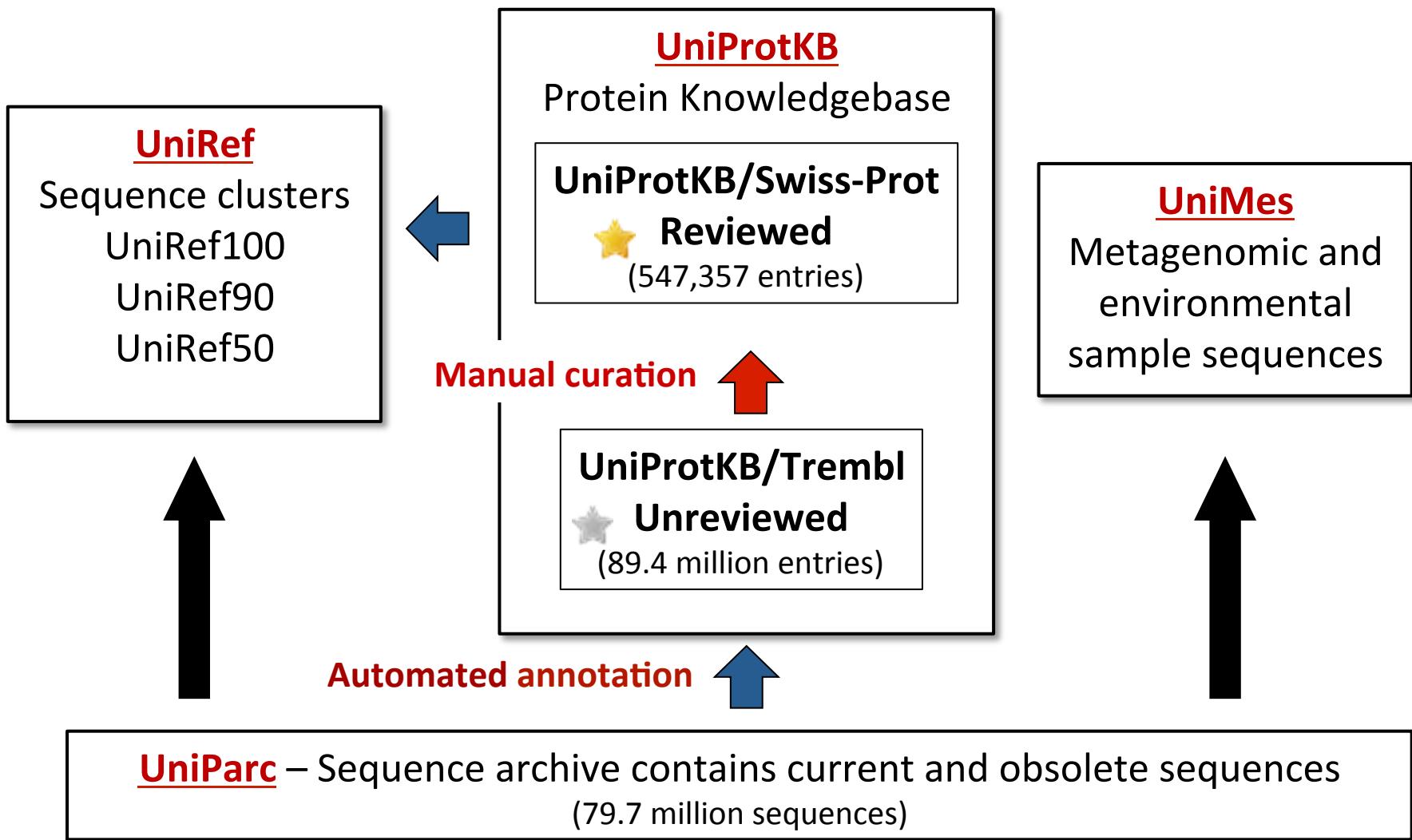
Overview of annotation procedures

- Evidence based – combined evidence from multiple, curated datasets

Common approaches:

- 1) Sequence-based – Search UniProt with a set of proteins
- 2) Model-based – Search Pfam with a set of proteins or models

Four components to UniProt for different needs



Names & Taxonomy

Protein names ⁱ	Recommended name: MADS-box transcription factor ANR1
Alternative name(s):	<ul style="list-style-type: none"> • Protein AGAMOUS-LIKE 44 • Protein ARABIDOPSIS NITRATE REGULATED 1
Gene names ⁱ	Name: ANR1 Synonyms: AGL44 Ordered Locus Names: At2g14210 ORF Name:
Organism ⁱ	Arabidopsis thaliana (Mouse-ear cress)
Taxonomic identifier ⁱ	3702 [NCBI]
Taxonomic lineage ⁱ	Eukaryota > Plantae > Magnoliophyta > Brassicales > Brassicaceae > Arabidopsis
Proteomes ⁱ	UP000006548: Chromosome 2

Protein and gene names, taxonomic information

Function

Probable transcription factor. Required for root plasticity in response to nitrate, NO₃⁻. Promotes lateral root growth in a NRT1.1-dependent manner.  3 Publications

GO - Molecular functionⁱ
 Source: UniProtKB-KW
 Source: TAIR
 Source: UniProtKB

GO - Biological processⁱ
 Source: UniProtKB
 Response to nitrate
 Transcription, DNA-templated
Complete GO annotation

Keywords - Molecular function: Developmental protein pattern

Keywords - Biological process: Transcription, Transcription regulation

Keywords - Ligand: DNA-binding

Enzyme and pathway databases

Reactomeⁱ REACT_206508. ERK/MAPK targets.

General annotation: subcellular location, tissue specificity

Organism-specific databases

TAIRⁱ AT2G14210.

Q9SI38 - ANR1_ARATH

Protein: **MADS-box transcription factor ANR1**
Gene: **ANR1**
Organism: *Arabidopsis thaliana (Mouse-ear cress)*
Status: Reviewed -  Experimental evidence at protein levelⁱ

Display: None

BLAST Align Format Add to basket History Show feature tables

Cross-references

Sequence databases

Select the link destinations:

- EMBLⁱ
- GenBankⁱ
- DDBJⁱ

Z97057 mRNA. Translation: CAB09793.1 .
AC007210 Genomic DNA. Translation: AAD25638.1 .
CP002685 Genomic DNA. Translation: AEC06290.1 .

X-refs (151 databases)

RefSeqⁱ NP_179033.1. NM_126990.3.
UniGeneⁱ At.126.

3D structure databases

ProteinModelPortalⁱ Q9SI38.
SMRⁱ Q9SI38. Positions 2-68.
ModBaseⁱ Search...
MobiDBⁱ Search...

uniprot.org/uniprot/Q9SI38

Changes in root architecture."

Science 279:407-409(1998) [PubMed] [Europe PMC] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [MRNA], FUNCTION, INDUCTION BY NITRATE, DISRUPTION PHENOTYPE, TISSUE SPECIFICITY.
Strain: cv. C24 and cv. Landsberg erecta.

2. "Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*."
Lin X., Kaul S., Rounsley S.D., Shea T.P., Pape G., Haas B.J., Feltus D.A., Lai J., Liang H., Pertea M., et al. Nature 402:761-768(1999) [PubMed] [Europe PMC] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
Strain: cv. Columbia.

3. The Arabidopsis Information Resource (TAIR).
Submitted (APR-2011) to the EMBL/GenBank/DDBJ databases
Cited for: GENOME REANNOTATION.
Strain: cv. Columbia.

4. "Empirical analysis of transcriptional activity in the *Arabidopsis* genome."
Yamada K., Lim J., Dale J.M., Chen H., Shinn P., Palm C.J., Southwick A.M., Wu H.C., Kim C.J., Nguyen M., Pham P.K., Cheuk R.F., Karlin-Newmann G., Liu S.X., Lam B., Sakano H., Wu T., Yu G., Ecker J.R. Science 302:842-846(2003) [PubMed] [Europe PMC] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
Strain: cv. Columbia.

5. "Large-scale analysis of RIKEN *Arabidopsis* full-length (RAFL) cDNAs."
Totoki Y., Seki M., Ishida J., Nakajima M., Enju A., Kamiya A., Narusaka M., Shin-i T., Nakagawa M., Sakamoto N., Oishi K., Kohara K., Kobayashi M., Toyoda A., Sakaki Y., Sakurai T., Iida K., Akiyama K., Shinozaki K.
Submitted (JUL-2006) to the EMBL/GenBank/DDBJ databases
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
Strain: cv. Columbia.

References

Origin of the sequences in UniProtKB

- ✓ International Nucleotide Sequence Database Collection (INSDC)
- ✓ Ensembl or EnsemblGenomes
- ✓ RefSeq
- ✓ Direct submissions (protein sequences)
- ✓ Literature
- ✓ Protein Data Bank

Controlled vocabularies

- Keywords provide a summary of the entry content
- Uniprot is annotated using the Gene Ontology (GO)

Ontologies

Keywords

Cellular component	Endoplasmic reticulum Membrane
Coding sequence diversity	Alternative splicing
Domain	Signal-anchor Transmembrane Transmembrane helix
Molecular function	Acyltransferase Transferase
PTM	Glycoprotein
Technical term	Complete proteome Reference proteome

Gene Ontology (GO)

Biological process	peptidyl-pyroglutamic acid biosynthetic process, using glutaminyl-peptide cyclotransferase Inferred from direct assay Ref.6 . Source: TAIR
Cellular component	endoplasmic reticulum membrane Inferred from electronic annotation. Source: UniProtKB-SubCell
	integral to membrane Inferred from electronic annotation. Source: UniProtKB-KW
Molecular function	glutaminyl-peptide cyclotransferase activity Inferred from direct assay Ref.6 . Source: TAIR
Complete GO annotation...	

UniProtKB, complete proteome sequence sets

- Genome completely sequenced
- Proteins mapped to the genome

5,933 complete proteomes

Fully manually reviewed (e.g. *S. cerevisiae*)

Partially manually reviewed (e.g. *A. thaliana*)

Unreviewed (e.g. *Chlorella variabilis*)

UniProtKB, reference proteome sequence sets

A reference proteome is the complete proteome of a representative, well-studied model organism or an organism of interest for biomedical research.

2,351 reference proteomes

Summary

UniProtKB/Swiss-Prot, the manually curated knowledgebase:

- **Protein sequence database covering all kingdoms of life**
- **Manually annotated**
- **Non-redundant:** all products of one gene in one species in a single entry
- **Highly cross-referenced** (links to ~150 databases).

Plant protein annotation:

- **Complete proteome for *Arabidopsis thaliana***
- **Synchronization with TAIR**

How to get annotations for distant homologies, or get phylogenetic information for protein families?

- 1) Search **Pfam**
- 2) Map Pfam identifiers to the **Gene Ontology (GO)**

Why do we need to use GO?



Reasons for the Gene Ontology

- Inconsistency in language

- Different names for the **same** concept

Eggplant



Brinjal

Aubergine

Melongene

Same for biological concepts

→ Comparison is difficult – in particular across species or across databases

Just one reason why the Gene Ontology (GO) is needed...



Reasons for the Gene Ontology

- Inconsistency in language
- Increasing amounts of biological data available
- Increasing amounts of biological data to come

Increasing amounts of biological data available

PubMed DNA repair RSS Save search Advanced

Article types Clinical Trial Review Customize ...

Text availability Abstract Free full text Full text

PubMed Commons Reader comments Trending articles

Publication dates 5 years 10 years Custom range...

Species Humans Other Animals

[Clear all](#) [Show additional filters](#)

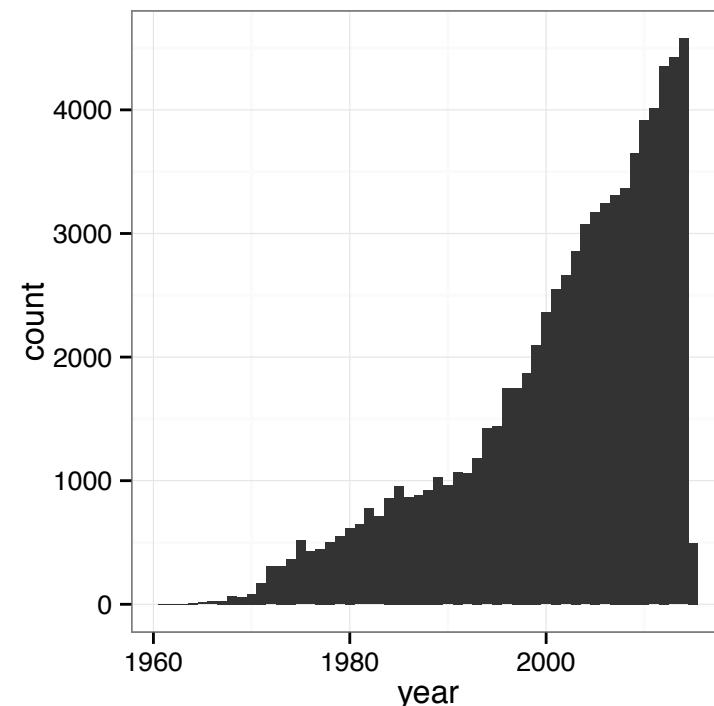
Display Settings: Summary, 20 per page, Sorted by Recently Added Send to:

Results: 1 to 20 of 78775 << First < Prev Page 1 of 3939 Next > Last >>

- [Environmental, Dietary, Maternal, and Fetal Predictors of Bulky DNA Adducts in Cord Blood: A European Mother-Child Study \(NewGeneris\).](#)
Pedersen M, Mendez MA, Schoket B, Godschalk RW, Espinosa A, Landström A, Villanueva CM, Merlo DF, Fthenou E, Gracia-Lavedan E, van Schooten FJ, Hoek G, B Nielsen JK, Sunyer J, Wright J, Kovács K, de Hoogh K, Gutzkov Anna L, Ketzel M, Haugen M, Botsvall M, Nieuwenhuijsen MJ, Fleming S, Agramunt S, Kyrtopoulos SA, Lukács V, Kleinjans JC Environ Health Perspect. 2015 Jan 27. [Epub ahead of print] PMID: 25626179 [PubMed - as supplied by publisher]
- [MEN1 Mutations in Hürthle Cell \(Oncocytic\) Thyroid Carcinoma.](#)
Kasaian K, Chindris AM, Wiseman SM, Mungall KL, Zeng T, Tse K, Sch Kachergus JM, Casler JD, Mungall AJ, Moore RA, Marra MA, Copland RC, Jones SJ. J Clin Endocrinol Metab. 2015 Jan 27:jc20143622. [Epub ahead of print] PMID: 25625803 [PubMed - as supplied by publisher]
- [Spatio-temporal regulation of RAG2 following genotoxic stress.](#)
Rodgers W, Byrum JN, Sapkota H, Rahman NS, Cail RC, Zhao S, Scha DNA Repair (Amst). 2015 Jan 8;27C:19-27. doi: 10.1016/j.dnarep.2014.12.008. [Epub ahead of print] PMID: 25625798 [PubMed - as supplied by publisher]

Expansion of sequence information

Search on ‘DNA repair’ ...
get almost 79,000 results



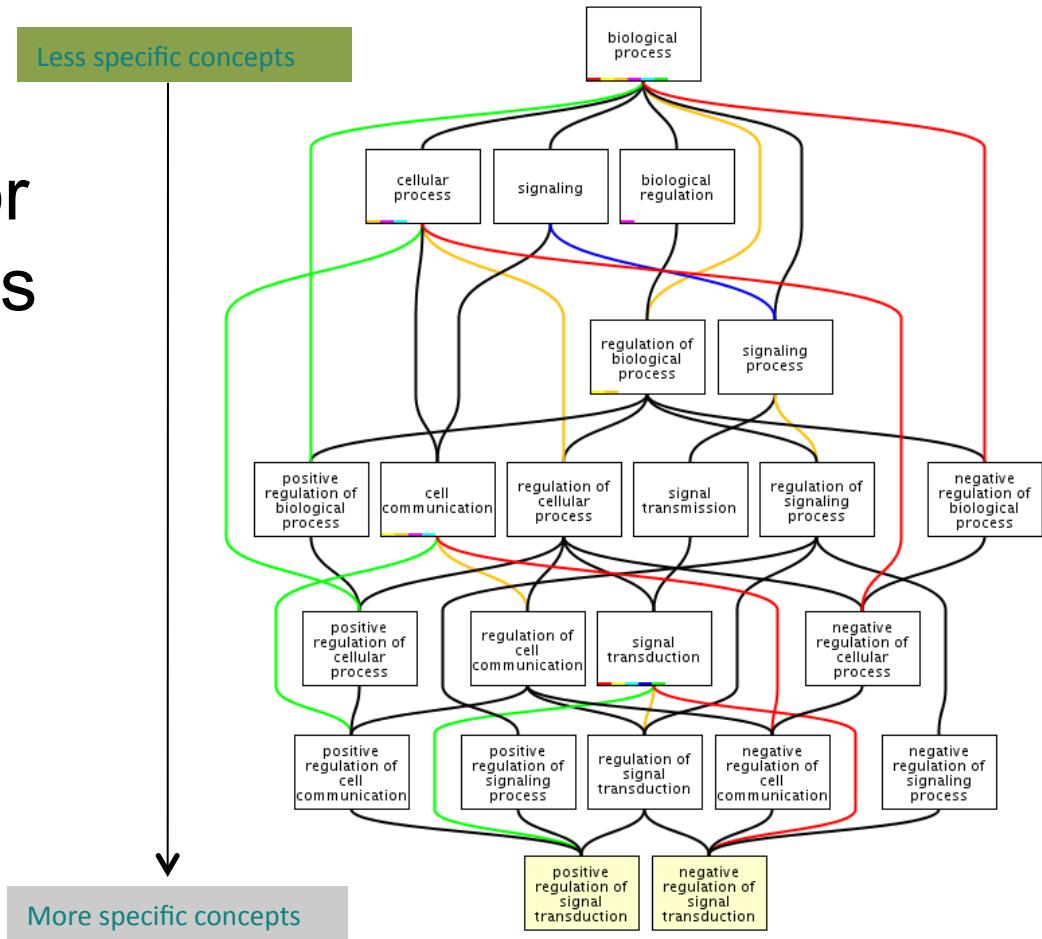


Reasons for the Gene Ontology

- Inconsistency in English language
- Increasing amounts of biological data available
- Increasing amounts of biological data to come
- Large datasets need to be interpreted quickly

The Gene Ontology

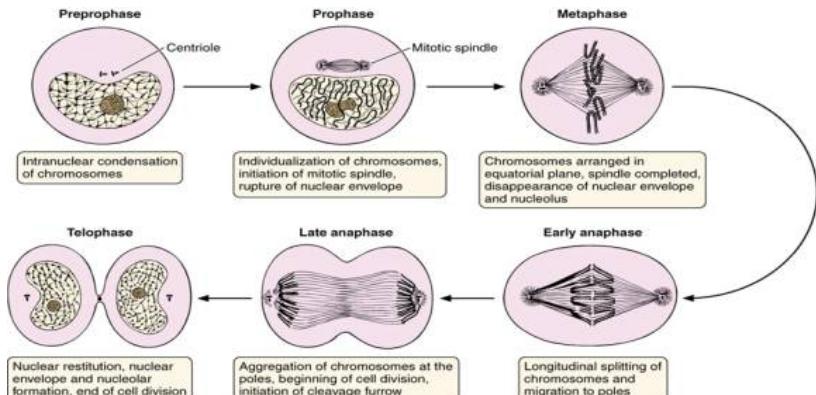
- A way to capture biological knowledge for individual gene products in a written and computable form
- A set of concepts and their relationships to each other arranged as a hierarchy



The Concepts in GO

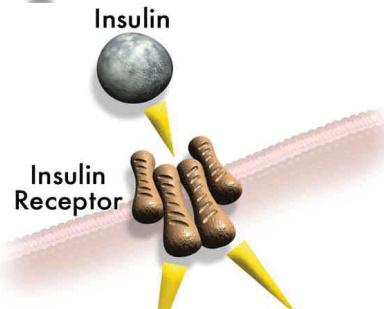
1. Molecular Function

An elemental activity or task or job



3. Cellular Component

Where a gene product is located

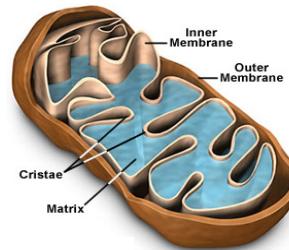


- protein kinase activity
- insulin receptor activity

2. Biological Process

A commonly recognized series of events

- cell division



- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

Anatomy of a GO term

		Unique identifier
		Term name
		Definition
i ID		GO:0005634
i Name		nucleus
i Definition		A membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. In most cells, the nucleus contains all of the cell's chromosomes except the organellar chromosomes, and is the site of RNA synthesis and processing. In some species, or in specialized cell types, RNA metabolism or DNA replication may be absent.
i Comment		
i Synonyms		Synonyms
Type	Synonym	Definition
exact	cell nucleus	
Cross-references associated with this term:		
Database	ID	Cross-references
INTERPRO	IPR000003	
INTERPRO	IPR000116	
INTERPRO	IPR000135	

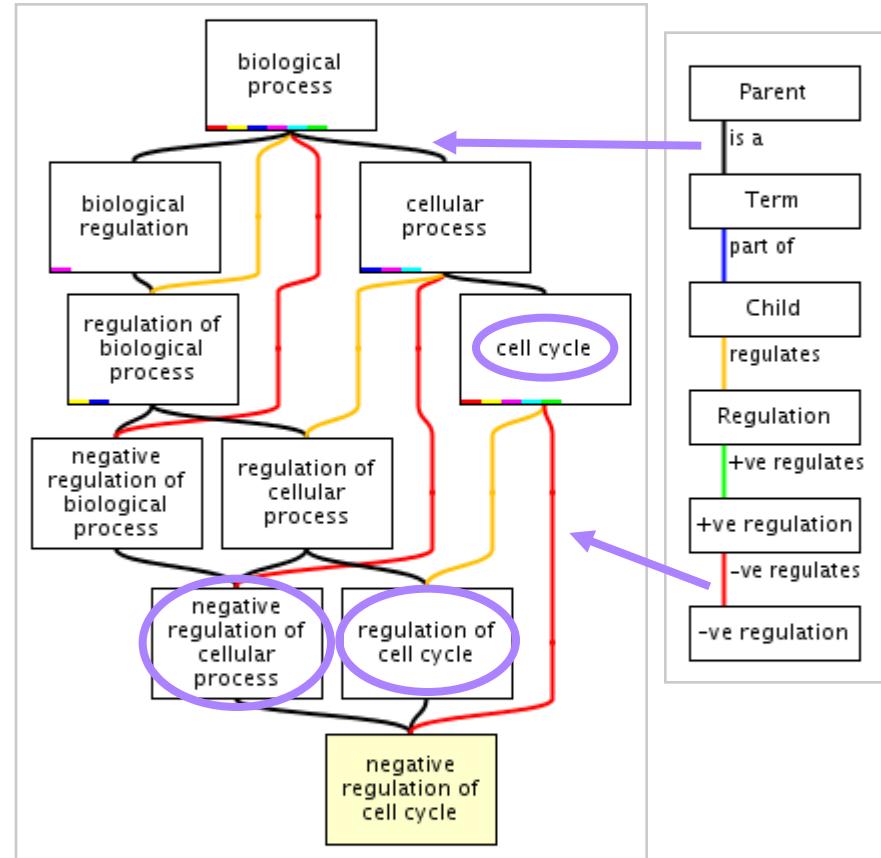
Ontology structure

- Directed acyclic graph

Terms can have more than one parent

- Terms are linked by relationships

is_a
part_of
regulates (and +/- regulates)
has_part
occurs_in



www.ebi.ac.uk/QuickGO

These relationships allow for complex analysis of large datasets

A GO annotation is ...

...a statement that a gene product;

1. has a particular molecular function
or is involved in a particular biological process
or is located within a certain cellular component
2. as determined by a particular method
3. as described in a particular reference

<i>Accession</i>	<i>Name</i>	<i>GO ID</i>	<i>GO term name</i>	<i>Reference</i>	<i>Evidence code</i>
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA



GTCGGCAATCCCTAAGTAGCCAAATTATTATTGTTAGATACTCAC
AGCGTAAACAGTGCGAGCCTTGAGGTGTTTGAATCAGTGAAATT
ATGCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT
ATTAATAAAACAAACAGTGCACACAGCCGGGGCATCTTCATAGA

FlyBase



GENE ONTOLOGY™ CONSORTIUM
<http://www.geneontology.org>



InterPro



Reactome



AgBase



J. Craig VenterTM
INSTITUTE



Aims of the GO project

- Compile the ontologies
 - currently over 40,000 terms
 - constantly increasing and improving
- Annotate gene products using ontology terms
 - around 30 groups provide annotations
- Provide a public resource of data and tools
 - regular releases of annotations
 - tools for browsing/querying annotations and editing the ontology

GO Annotation

UniProt-Gene Ontology Annotation (UniProt-GOA) database at the EBI

- Largest open-source contributor of annotations to GO
- Provides annotation for more than 60,000 species
- Their main priority is to annotate the human proteome



UniProt-GOA incorporates annotations made using two methods



Electronic Annotation

- Quick way of producing large numbers of annotations
- Annotations use less-specific GO terms
- Only source of annotation for many non-model organism species

Manual Annotation

- Time-consuming process producing lower numbers of annotations
- Annotations tend to use very specific GO terms

Electronic annotation methods

1. Mapping of external concepts to GO terms

e.g. InterPro2GO, UniProt Keyword2GO, Enzyme Commission2GO

UniProt > UniProtKB

P00519 (ABL1_HUMAN) ★ Reviewed, UniProtKB/Swiss-Prot
Last modified October 5, 2010. Version 162. [History...](#)

Clusters with 100%, 90%, 50% identity | Documents (7) | Third-party data

Names · Attributes · General annotation · Ontologies · Interactions · Alt products · Sequence annotation · Sequenc
Customize order

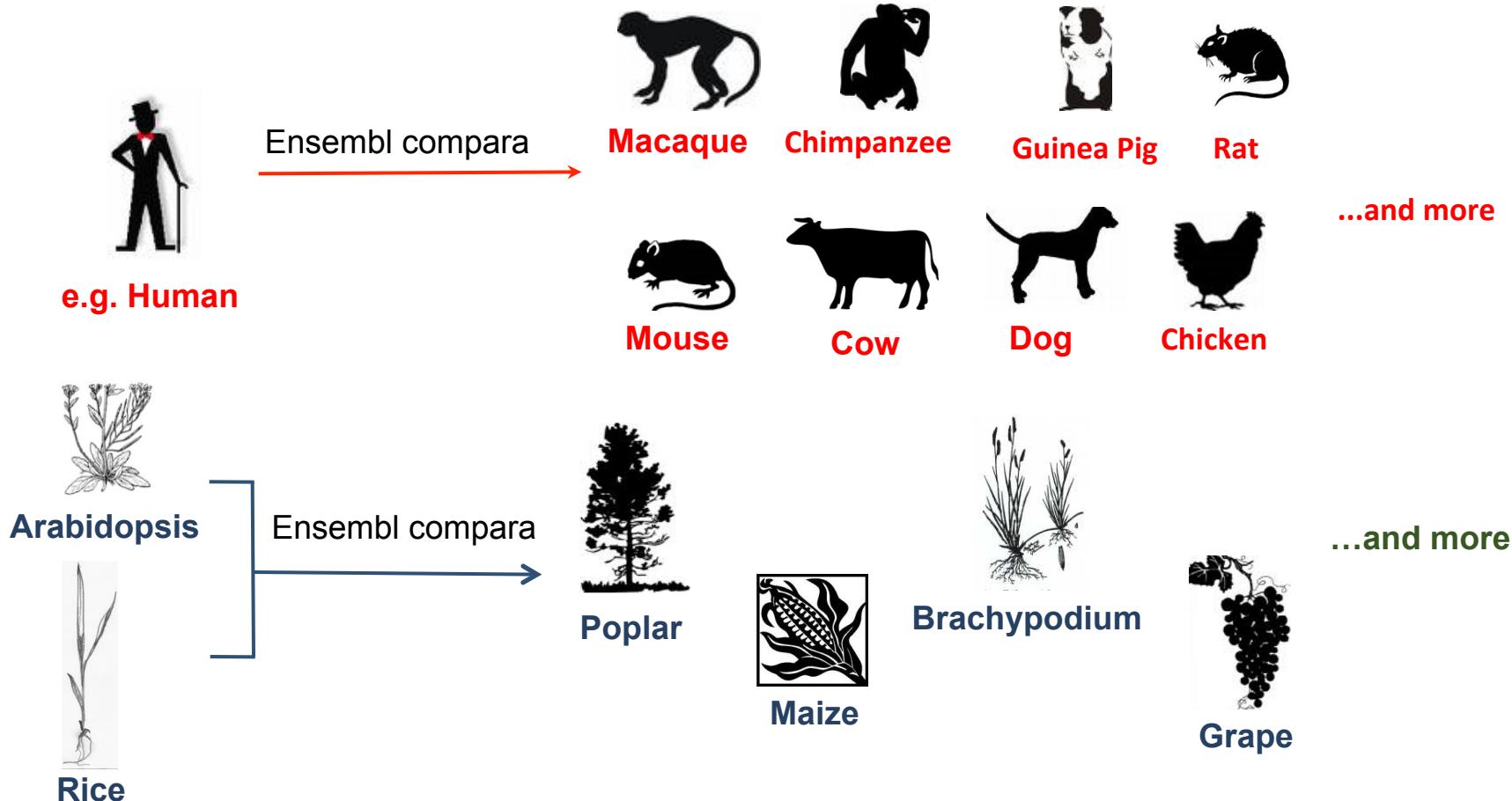
Names and origin

Protein names	<p>Recommended name: Tyrosine-protein kinase ABL1 EC=2.7.10.2</p> <p>Alternative name(s): Abelson murine leukemia viral oncogene homolog 1 Proto-oncogene c-Abl p150</p>
Gene names	Name: ABL1 Synonyms: ABL, JTK7
Organism	Homo sapiens (Human) [Complete proteome]
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota > Metazoa > Chordata > Craniata > Vertebrata > Euteleostomi > Mammalia > Eutheria

GO:0004715 ; non-membrane spanning protein tyrosine kinase activity

Electronic annotation methods

2. Automatic transfer of manual annotations to orthologs



Annotations are high-quality and have an explanation of the method (GO_REF)

Manual annotation by GOA

High-quality, specific annotations made using:

- Full text peer-reviewed papers
- A range of evidence codes to categorise the types of evidence found in a paper,

Manual annotation by GOA

High-quality, specific annotations made using:

- Full text peer-reviewed papers
- A range of evidence codes to categorise the types of evidence found in a paper, e.g.,
 - Inferred from Experiment (EXP)
 - Inferred from Direct Assay (IDA)
 - Inferred from Physical Interaction (IPI)
 - Inferred from Mutant Phenotype (IMP)
 - Inferred from Genetic Interaction (IGI)
 - Inferred from Expression Pattern (IEP)

How to access and use GO annotation data

Where can you find annotations?

UniProtKB

P00519 (ABL1_HUMAN) ★ Reviewed, UniProtKB/Swiss-Prot
Last modified October 5, 2010. Version 162. History...

Clusters with 100%, 90%, 50% identity | Documents (7) | Third-party data

Names · Attributes · General annotation · Ontologies · Interactions · Alt products · Se

Customize order

Names and origin

Protein names

Recommended name:
Tyrosine-protein kinase ABL1
EC=2.7.10.2
Alternative name(s):
Abelson murine leukemia viral oncogene homolog 1
Proto-oncogene c-Abl
p150

Gene names

Name: **ABL1**
Synonyms: ABL, JTK7

Transcript-based displays

- Transcript summary
- Supporting evidence (24)
- Sequence
 - Exons (11)
 - cDNA
 - Protein
- External References
 - General identifiers (105)
 - Oligo probes (38)
 - **Gene ontology (48)**
- Genetic Variation
- Population comparison
- Comparison image
- Protein Information
 - Protein summary
 - Domains & features (43)
 - Variations (48)

Transcript: ABL1-001 (ENST00000318560)
c-abl oncogene 1, receptor tyrosine kinase [Source:HGNC Symbol;Acc:76]

Location Chromosome 9: 133,710,453-133,763,062 forward strand.

Gene This transcript is a product of gene ENSG00000097007 - There are 6 transcripts in this gene

Gene ontology help

The following GO terms have been mapped to this entry via UniProt and/or RefSeq:

GO Accession	GO Term	Evidence Annotation Source
GO:0000115	regulation of transcription involved in S-phase of mitotic cell cycle	TAS
GO:0000166	nucleotide binding	IEA
GO:0000287	magnesium ion binding	IDA
GO:0003672	DNA binding	NAS
GO:0004672	protein kinase activity	IEA

Ensembl

Entrez gene

NCBI Resources How To

Entrez Gene Genes and mapped phenotypes

Search: Gene Limits Advanced search Help

Display Settings Full Report

ABL1 c-abl oncogene 1, non-receptor tyrosine kinase [Homo sapiens]

Gene ID: 25, updated on 5-Oct-2010

Summary

Official Symbol ABL1 provided by HGNC

Official Full Name c-abl oncogene 1, non-receptor tyrosine kinase provided by HGNC

Primary source HGNC_76

Locus tag RP11-83J21.1

See related Ensembl ENSG00000097007; HPRD:01809; MIM:189980

Gene type protein coding

RefSeq status REVIEWED

Gene Ontology provided by GOA

Function	Evidence	Pubs
	Evidence Code	
ATP binding	IDA	PubMed
DNA binding	NAS	PubMed
SH3 domain binding	IPI	PubMed
magnesium ion binding	IDA	PubMed
manganese ion binding	IDA	PubMed
non-membrane spanning protein tyrosine kinase activity	IEA	
nucleotide binding	IEA	
proline-rich region binding	IPI	PubMed
protein C-terminus binding	IPI	PubMed
protein binding	IPI	PubMed

Gene Association Files

17 column files containing all information for each annotation

the Gene Ontology

Open menus

Home

FAQ

Downloads

- Ontologies
- Annotations
- Database
- Mappings to GO
- Teaching Resources

Annotation Details and Downloads

- Filtered files
- Unfiltered files
- gp2protein files

Current Annotations

GO Consortium website

Species, Database	Gene Products Annotated	Annotations	Submission date MM/DD/YYYY	Download filtered files
<i>Anaplasma phagocytophilum</i> HZ JCVI	1289	3471 (3471 non-IEA)	8/27/2010	annotations [39.7 kb] README
<i>Agrobacterium tumefaciens</i> str. C58 PAMGO	83			
<i>Arabidopsis thaliana</i> TAIR	31339			
<i>Aspergillus nidulans</i> AspGD	11368			
<i>Bacillus anthracis</i> Ames JCVI	5280			

EBI Databases Tools Research Training Industry About Us Help Site Index

Enter Text Here Find Help Feedback

UniProtKB-GOA Home New to UniProtKB-GOA? Who uses GO? Downloads Searching UniProtKB-GOA

EBI > Databases > Biological Ontologies > UniProtKB-GOA > UniProtKB-GOA Downloads

Search QuickGO

Query examples - [apoptosis](#), [GO:0006915](#), [tropomyosin](#), [P06727](#).

UniProt-GOA - Downloads

Set of Proteins	Gene Association File	Identifier Cross-Reference File
UniProtKB	Gene association	UniProtKB ID mapping Readme
Human	Gene association	UniProtKB ID mapping Readme
Mouse	Gene association	UniProtKB ID mapping Readme
Rat	Gene association	UniProtKB ID mapping Readme
Arabidopsis	Gene association	UniProtKB ID mapping Readme
Zebrafish	Gene association	UniProtKB ID mapping Readme
Chicken	Gene association	UniProtKB ID mapping Readme
Cow	Gene association	UniProtKB ID mapping Readme
Proteomes	Gene association Browse	n/a
PDB	Gene association	n/a

UniProt-GOA website

Numerous species-specific files →



Search GO

terms genes or proteins exact match

ABL1

GO browsers

Tyrosine-protein kinase ABL1

protein from *Homo sapiens* (human)

Term associations Peptide Sequence

Term Associations

Download all association information in:

Perform an action with this page's selected terms...

Accession, Term	Ontology	Qualifier	Evidence	Reference	Assigned by
<input type="checkbox"/> GO:0030036 : actin cytoskeleton organization	2336 gene products view in tree	biological process	ISS With UniProtKB:P00520	GO REF:0000024	UniProtKB
<input type="checkbox"/> GO:0007155 : cell adhesion	4735 gene products view in tree	biological process	IEA With SP_KWIKW-0130	GO REF:0000004	UniProtKB
<input type="checkbox"/> GO:0008630 : DNA damage response, signal transduction resulting in induction of apoptosis	145 gene products view in tree	biological process	TAS	PMID:10391249	Proteome Inc.
<input type="checkbox"/> GO:0006298 : mismatch repair					

EMBL-EBI  EB-eye Search All Databases Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index 

EBI > Databases > QuickGO

ABL1 *Homo sapiens* P00519



Quick GO Click for example search   

Accession P00519
Gene ABL1
Taxonomy Homo sapiens
Description Tyrosine-protein kinase ABL1

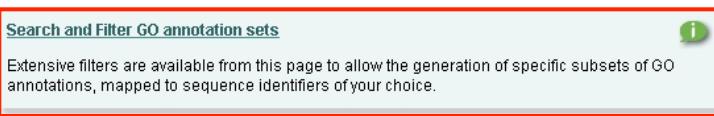
Annotation Results: 1 to 219 of 219   Page size: 25 

Database ID	Symbol Qualifier GO Identifier GO Term Name	Aspect Evidence Reference			With	Taxon Date	Assigned By	Product ID
Process								
UniProtKB P00519 ABL1	GO:0000115 regulation of transcription involved in S phase of mitotic cell cycle	P	TAS	PMID:8242749		9606 20030904 PINC		
UniProtKB P00519 ABL1	GO:0007165 signal transduction	P	IEA	InterPro2GO	IPR020700	9606 20101009 InterPro		
UniProtKB P00519 ABL1	GO:0051353 positive regulation of oxidoreductase activity	P	IDA	PMID:12893824		9606 20071108 BHF-UCL		
UniProtKB P00519 ABL1	GO:0006468 protein amino acid phosphorylation	P	IEA	InterPro2GO	IPR001245	9606 20101009 InterPro		
UniProtKB P00519 ABL1	GO:0018108 peptidyl-tyrosine phosphorylation	P	IDA	PMID:7590236		9606 20100505 UniProtKB		
UniProtKB P00519 ABL1	GO:0006464 protein modification process	P	NAS	PMID:8242749		9606 20040707 UniProtKB		
UniProtKB P00519 ABL1	GO:0030036 actin cytoskeleton organization	P	ISS	GO_REF:0000024	Ab1 (ABL1_MOUSE)	9606 20061011 UniProtKB		
UniProtKB P00519 ABL1	GO:0006355 regulation of transcription, DNA-dependent	P	TAS	PMID:8242749		9606 20030904 PINC		
UniProtKB P00519 ABL1	GO:0006468 protein amino acid phosphorylation	P	IEA	InterPro2GO	IPR008266	9606 20101009 InterPro		
UniProtKB P00519 ABL1	GO:0018108 peptidyl-tyrosine phosphorylation	P	IEA	InterPro2GO	IPR020700	9606 20101009 InterPro		
UniProtKB P00519 ABL1	GO:0018108 peptidyl-tyrosine phosphorylation	P	IDA	PMID:9144171		9606 20061004 UniProtKB		
UniProtKB P00519 ABL1	GO:0006468 protein amino acid phosphorylation	P	IEA	InterPro2GO	IPR000719	9606 20101009 InterPro		
UniProtKB P00519 ABL1	GO:0006298 mismatch repair	P	TAS	PMID:10391249		9606 20030904 PINC		
UniProtKB P00519 ABL1	GO:0008630 DNA damage response, signal transduction resulting in induction of apoptosis	P	TAS	PMID:10391249		9606 20030904 PINC		

The EBI's QuickGO browser

Search GO terms or proteins → 

Find sets of GO annotations → 

EMBL-EBI  **EB-eye Search** All Databases Reset  Advanced Search

Databases Tools EBI Groups Training Industry About Us Help Site Index  

QuickGO Help Reference FAQs

EBI > Databases > QuickGO

QuickGO

QuickGO is a fast web-based browser for [Gene Ontology](#) terms and annotations, which is provided by the [UniProtKB-GOA group](#) at the EBI.

QuickGO  Click for example search Web Services Dataset Your Terms: 16 

Search and Filter GO annotation sets 
Extensive filters are available from this page to allow the generation of specific subsets of GO annotations, mapped to sequence identifiers of your choice.

Investigate GO slims 
GO slims are lists of GO terms that have been selected from the full set of terms available from the Gene Ontology project.

GO slims can be used to generate a focused view of part of the GO, or with annotation data they can be used to see how a set of proteins/genes can be broadly categorized (using annotation data and the relationships that exist between terms in the ontologies).

Further information on GO slims can be found at the [GO Consortium web site](#).

Example queries
QuickGO can be queried for both **GO terms** and **proteins**:

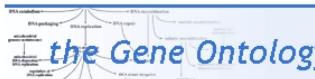
- Search for terms by keyword or ID: [apoptosis](#), [GO:0006915](#)
- Search for proteins by name or accession: [tropomyosin](#), [P06727](#)
- List all terms in an ontology: [Biological Process](#), [Molecular Function](#), [Cellular Component](#)

QuickGO Tips

- QuickGO can help make customized GO slims, which can be used to 'map up' your gene/protein identifiers to broadly categorise their functions or subcellular locations. Click [here](#) for more information.
- QuickGO's search facility can be used to find information on GO terms as well as information on annotation to distinct proteins. Just type a partial GO term name/synonym or a gene symbol, protein name or sequence identifier into the Search box.

Tutorial
[Try an interactive demo](#)

Numerous Third Party Tools

 the Gene Ontology

Downloads Tools Documentation Projects About Contact

Search
gene or protein name

Tools for Searching and Browsing GO

The following tools make use of the GO ontologies or the gene associations provided by Consortium members. Being listed on this page does not represent an endorsement by the GO Consortium, nor has the Consortium tested the tool or found that it uses the Consortium information accurately. This page is provided to promote an exchange of information between users and software developers.

 the Gene Ontology

Search
gene or protein name

Downloads Tools Documentation Projects About Contact



Unless stated otherwise, tools are free for academic use.

AmiGO

[The Gene Ontology Consortium](#)

No publication

AmiGO provides an interface to search and browse the GO ontologies and view the terms with which they are associated; and gene product annotations. AmiGO also provides annotations annotated to a GO term and submitted to the GO Consortium.



Unless stated otherwise, tools are free for academic use.

Avadis

[Strand Genomics](#)

No publication

Avadis is a data analysis and visualization tool for gene expression data. Avadis has a built-in Gene Ontology browser to view ontology hierarchies. There are common ontology paths for multiple genes. Genes can be clustered based on ontology terms to identify functional signatures in gene expression clusters.



Annotating novel sequences

- Can use BLAST queries to find similar sequences with GO annotation which can be transferred to the new sequence
- Several tools are currently available;

AmiGO BLAST – searches the GO Consortium database

BLAST2GO – searches the NCBI database

HMMER2GO – searches ORFs against a set of HMMs

Trinotate – finds ORFs, does a lot of searching....

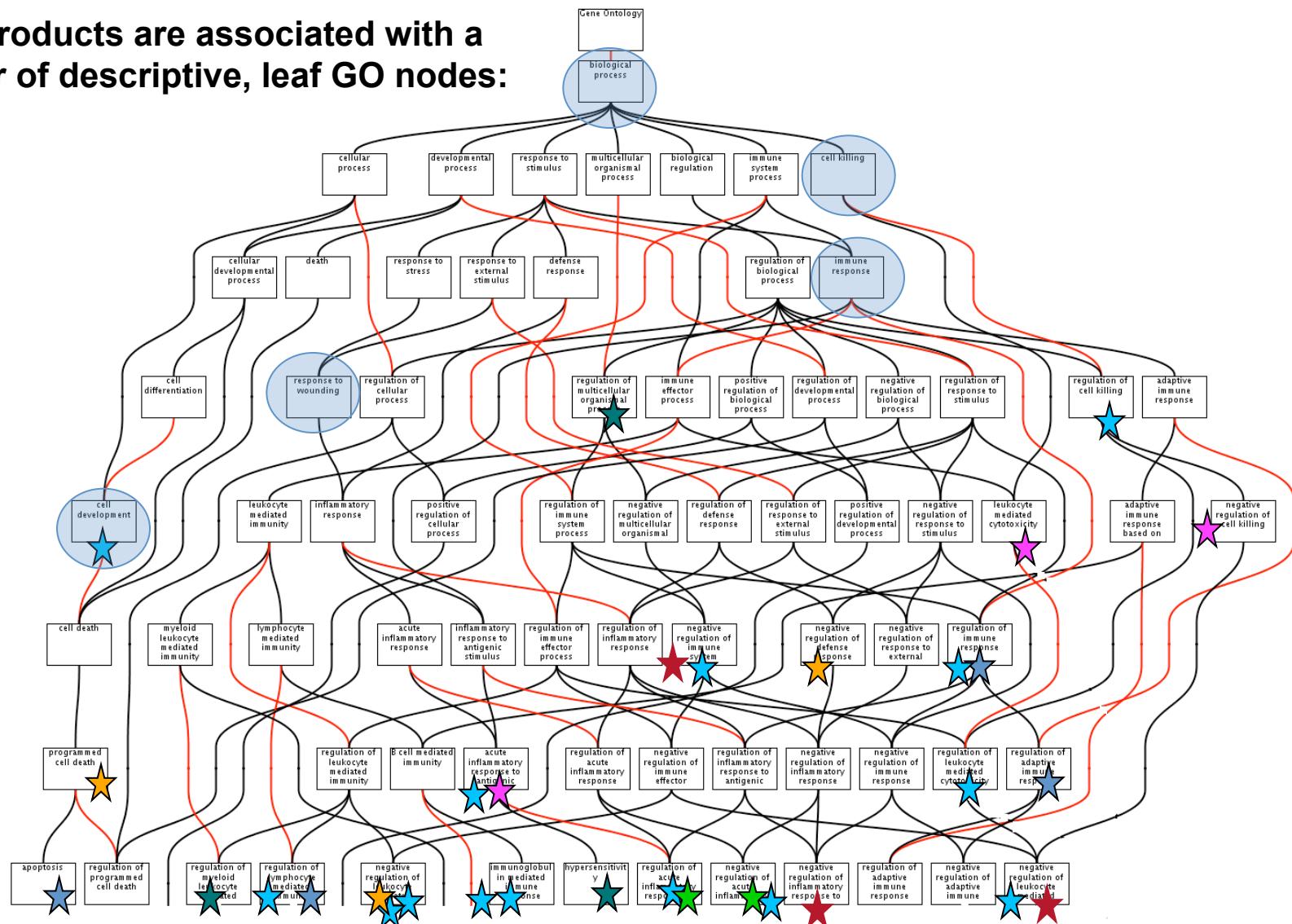


Using the GO to provide a functional overview for a large dataset

- Many GO analysis tools use GO slims to give a broad overview of the dataset
- GO slims are cut-down versions of the GO and contain a subset of the terms in the whole GO
- GO slims usually contain less-specialised GO terms

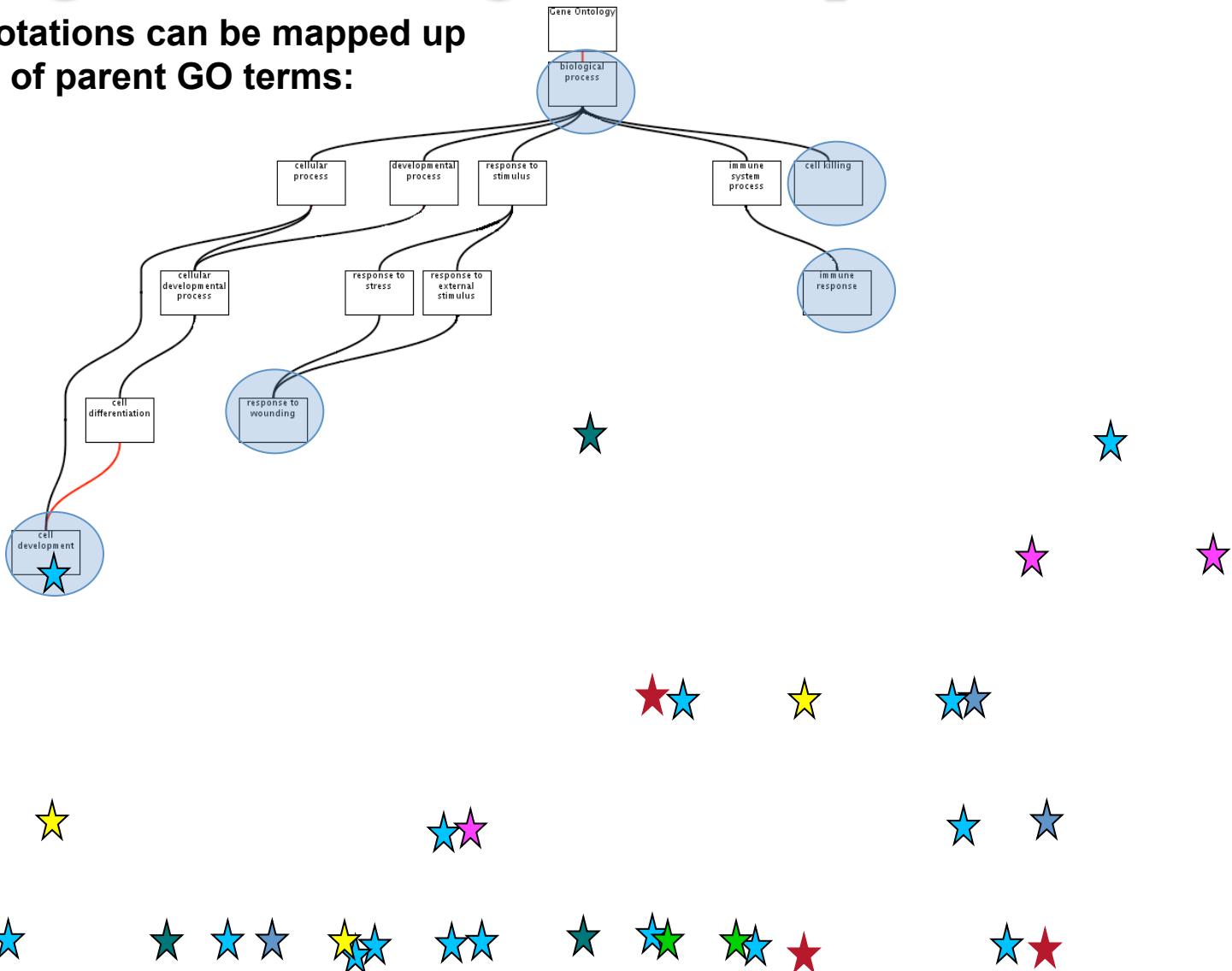
Slimming the GO using the ‘true path rule’

Many gene products are associated with a large number of descriptive, leaf GO nodes:



Slimming the GO using the ‘true path rule’

...however annotations can be mapped up to a smaller set of parent GO terms:



GO slims

Custom slims are available for download;

geneontology.org/page/go-slim-and-subset-guide

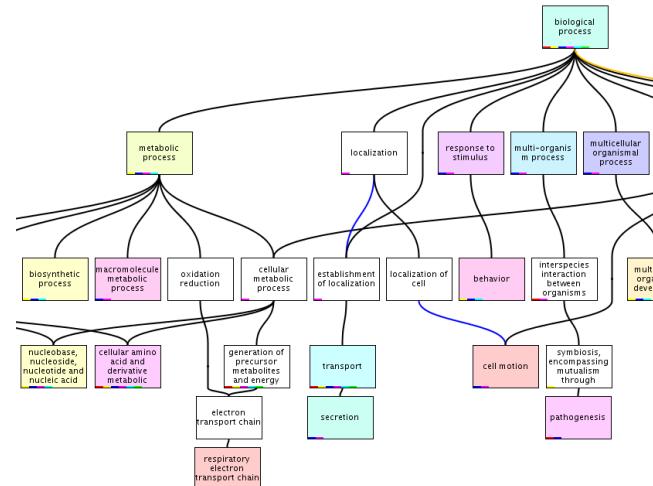
or you can make your own using;

- QuickGO

ebi.ac.uk/QuickGO

- AmiGO's GO slimmer

amigo.geneontology.org/cgi-bin/amigo/slimmer



The EBI's QuickGO browser

The screenshot shows the EBI's QuickGO browser interface. At the top, there is a navigation bar with links for Databases, Tools, EBI Groups, Training, Industry, About Us, Help, Site Index, and Give us feedback. A search bar is also present.

On the left, three main sections are highlighted with arrows pointing to specific features:

- Search GO terms or proteins** points to the main search area, which includes a "Quick GO" logo, a search input field with placeholder "Click for example search", a "Search!" button, and links for Web Services, Dataset, and Your Terms: 16.
- Find sets of GO annotations** points to a section titled "Search and Filter GO annotation sets" containing text about extensive filters for generating specific subsets of GO annotations.
- Map-up annotations with GO slims** points to a section titled "Investigate GO slims" containing text about GO slims being lists of GO terms selected from the full set, and how they can be used to categorize proteins/genes.

On the right side, there is a "QuickGO Tips" box with two bullet points:

- QuickGO can help make customized GO slims, which can be used to 'map up' your gene/protein identifiers to broadly categorise their functions or subcellular locations. Click [here](#) for more information.
- QuickGO's search facility can be used to find information on GO terms as well as information on annotation to distinct proteins. Just type a partial GO term name/synonym or a gene symbol, protein name or sequence identifier into the Search box.

At the bottom right, there is a "Tutorial" link and a "Try an interactive demo" link.

Why to use the GO

- Access gene product functional information
- Analyse high-throughput genomic or proteomic datasets
- Validation of experimental techniques
- Get a broad overview of a proteome
- Obtain functional information for novel gene products

GO term enrichment

How to determine the significance of a result?

GO term enrichment

How to determine the significance of a result?

Find terms significantly enriched in a subset of the data using the GO.

Many methods:

- R -> GOstats
- Ontologizer
- David (david.abcc.ncifcrf.gov) - for humans
- QuickGO

Practical session

[github.com/UBCBio525/Bio525D/Day 7/exercises](https://github.com/UBCBio525/Bio525D/Day%207/exercises)