

# **EBI web resources I: databases and tools**

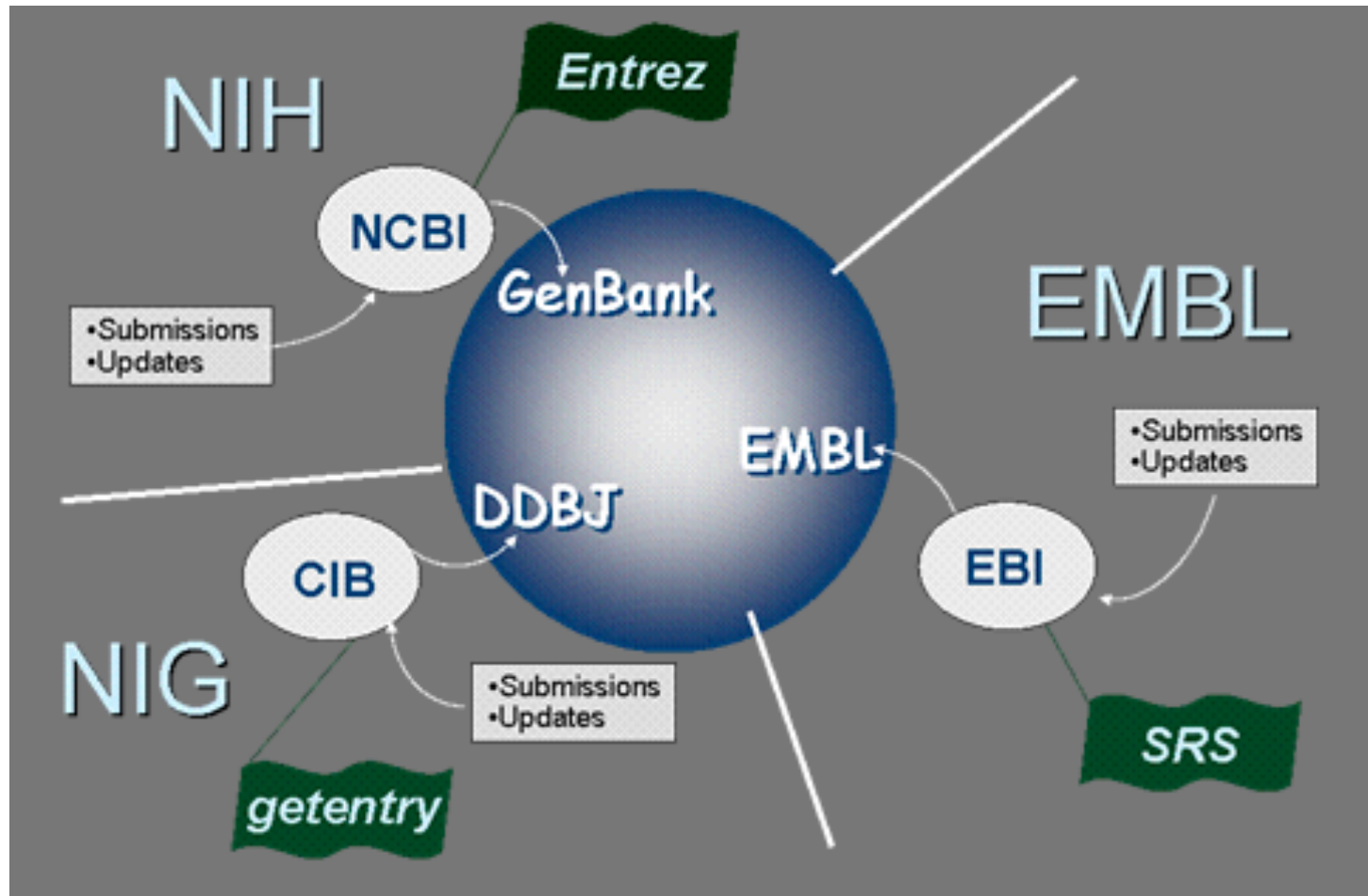
Yanbin Yin  
Spring 2013

# Outline

- Intro to EBI
- Databases and web tools
  - UniProt
  - Gene Ontology
- Hands on Practice

MOST MATERIALS ARE FROM: <http://www.ebi.ac.uk/training/online/course-list>

# Three international nucleotide sequence databases



# The European Bioinformatics Institute (EBI)



Created in 1992 as part of [European Molecular Biology Laboratory](#) (EMBL)

EMBL was created in 1974 and is a [molecular biology](#) research institution supported by 20 European countries and Australia

[Wellcome Trust Genome Campus, Hinxton, Cambridge, UK](#)  
Neighbor of [Wellcome Trust Sanger Institute](#)





Explore the EBI:

<http://www.ebi.ac.uk/>


FIND

Examples: [ROA1](#), [HUMAN](#), [tpi1](#), [Sulston...](#)[Help](#) | [Feedback](#)

## Data Resources and Tools

- [ENA](#)
- [UniProt](#)
- [ArrayExpress](#)
- [Ensembl](#)
- [InterPro](#)
- [PDB](#)
- [Genomes](#)
- [Nucleotide Sequences](#)
- [Protein Sequences](#)
- [Macromolecular Structures](#)
- [Small Molecules](#)
- [Gene Expression](#)
- [Protein Expression](#)
- [Molecular Interactions](#)
- [Reactions & Pathways](#)
- [Protein Families](#)
- [Enzymes](#)
- [Literature](#)
- [Taxonomy](#)
- [Ontologies](#)
- [Patent Resources](#)
- [Sequence Similarity & Analysis](#)
- [Pattern & Motif Searches](#)
- [Structure Tools](#)
- [Text Mining](#)
- [Downloads](#)
- [Web Services](#)

## Latest News

### [DNA storage becomes a reality](#)

Posted: **Jan 23, 2013**

EMBL-EBI researchers have created a way to store data in the form of DNA – a material that lasts for tens of thousands of

## Events & Training

EMBL-EBI [forthcoming courses and conferences](#)

[EMBL-EBI Open Day](#)  
**14 March 2013**

This is the perfect opportunity for young scientists who are

# Research groups in EBI

	Group/team leader	Area of research
<b>Genomes</b>	<a href="#">Ewan Birney</a>	Algorithmic methods for genome analysis <b>InterPro</b>
	<a href="#">Paul Flicek</a>	Vertebrate genomics
	<a href="#">Nick Goldman</a>	Evolutionary tools for sequence analysis
<b>Transcriptomes</b>	<a href="#">Alvis Brazma</a>	Functional genomics <b>miRBase</b>
	<a href="#">Anton Enright</a>	Functional genomics and analysis of small RNA function
	<a href="#">John Marioni</a>	Computational and evolutionary genomics
	<a href="#">Oliver Stegle</a>	Statistical genomics and systems genetics
<b>Proteins</b>	<b>Janet Thornton</b>	Computational biology of proteins: structure, function and evolution
	<a href="#">Rolf Apweiler</a>	Protein sequence analysis and functional annotation <b>UniProt</b>
	<a href="#">Gerard Kleywegt</a>	Structural validation of proteins; protein-ligand interactions
<b>Pathways and systems</b>	<a href="#">Nicolas Le Novère</a>	Computational systems neurobiology
	<a href="#">Nick Luscombe</a>	Genomics and regulatory systems
	<a href="#">Paul Bertone</a>	Pluripotency, reprogramming and differentiation
	<a href="#">Julio Saez-Rodriguez</a>	Systems biomedicine
<b>Literature</b>	<a href="#">Dietrich Rebholz-Schuhmann</a>	Literature analysis and semantic data integration in life science research
<b>Chemistry</b>	<a href="#">Christoph Steinbeck</a>	Cheminformatics and metabolism
	<a href="#">John Overington</a>	Chemogenomics and drug discovery

# Major databases in EBI

GenBank	<a href="#">EMBL-Bank</a> (DNA and RNA sequences)
Genome MapView	<a href="#">Ensembl</a> (genomes)
GEO	<a href="#">ArrayExpress</a> (microarray-based gene-expression data)
GenPept (nr)	<a href="#">UniProt</a> (protein sequences)
CDD	<a href="#">InterPro</a> (protein families, domains and motifs)
MMDB	<a href="#">PDBe</a> (macromolecular structures)

Others, such as

[IntAct](#) (protein–protein interactions)

[Reactome](#) (pathways)

[ChEBI](#) (small molecules)

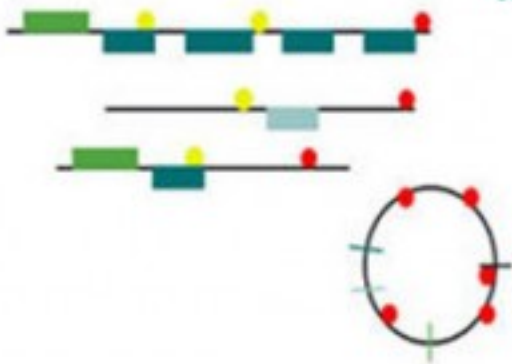
[IntEnz](#) (enzyme classification)

[GO](#) (gene ontology)

Swiss Institute of Bioinformatics  
Sanger Institute



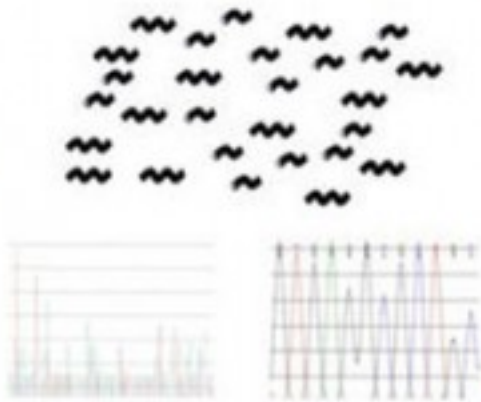
Feature  
annotation



Assembly  
information



Sequencing  
and sampling  
information



1) **EMBL-Bank**

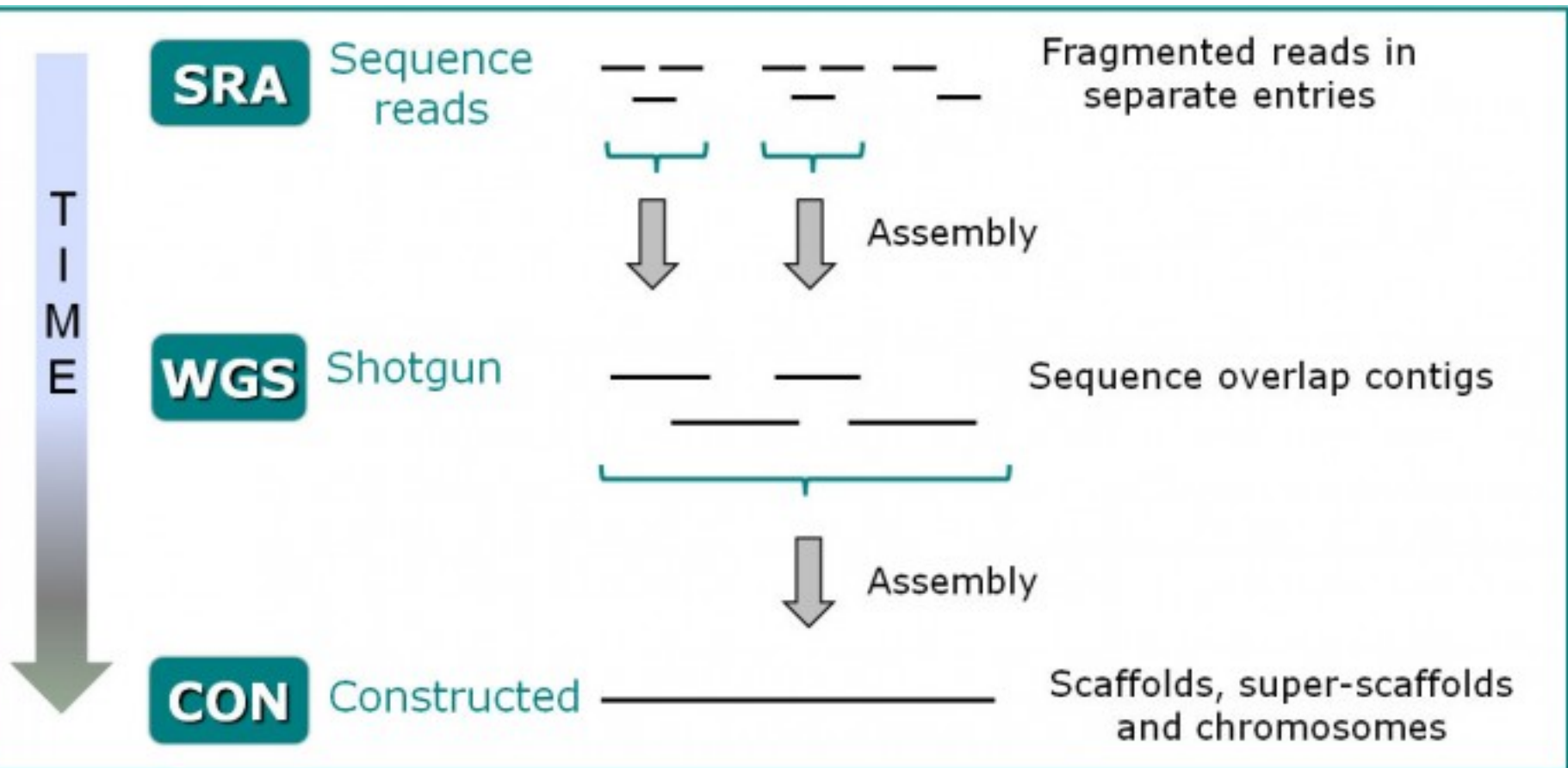
2) **Sequence Read  
Archive**

3) **Trace Archive**





Sequence might first enter ENA as **SRA** (Sequence Read Archive) **fragmented** sequence reads; it might be re-submitted as **assembled WGS** (Whole Genome Shotgun) sequence overlap **contigs**; it might be re-submitted again with **further assembly** as **CON** (Constructed) sequence entries, with the older WGS entries being consigned to the Sequence Version Archive



Data is first split into **classes**, then it is split into intersecting slices by **taxonomy**

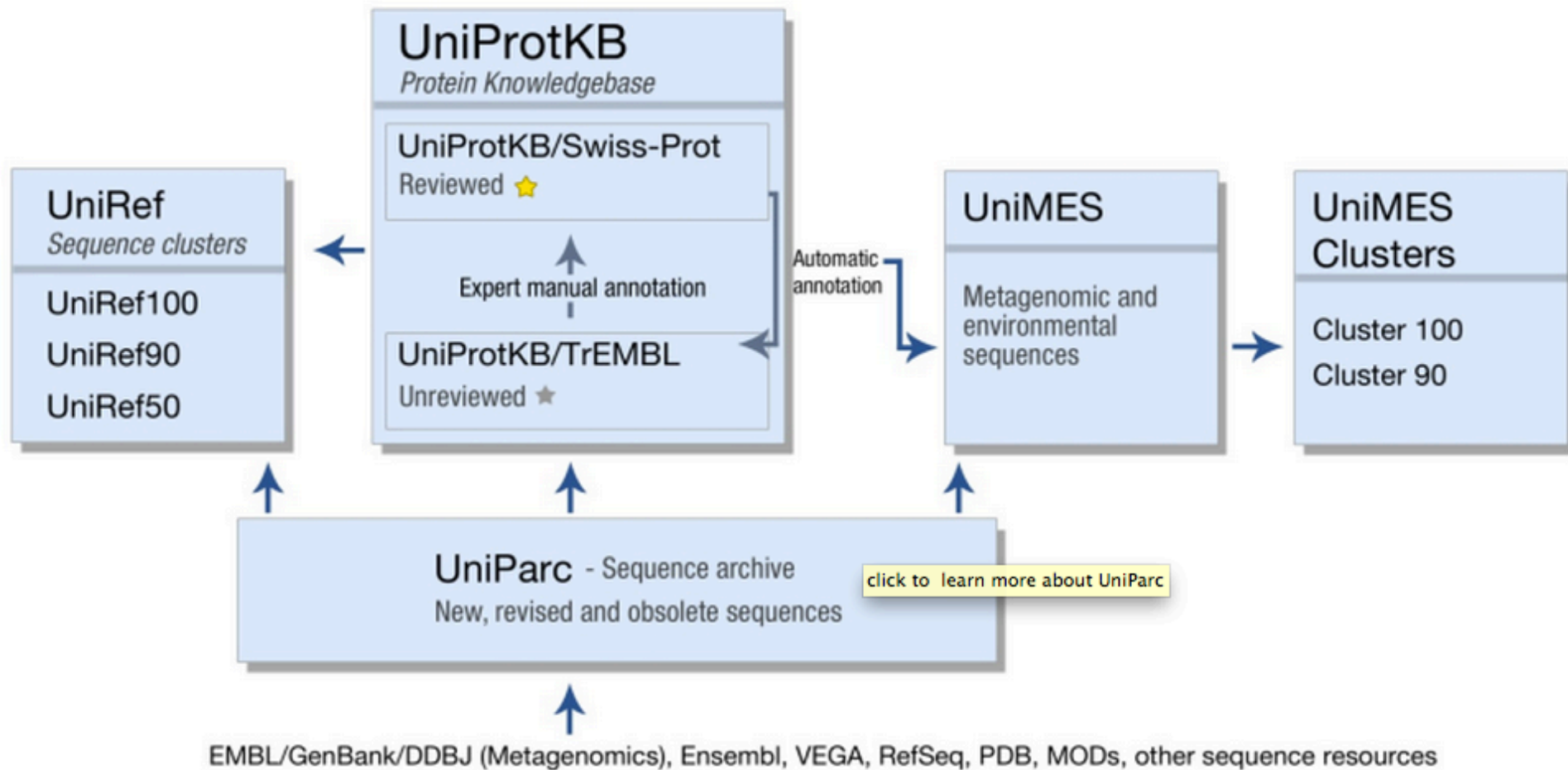
## EMBL-Bank:

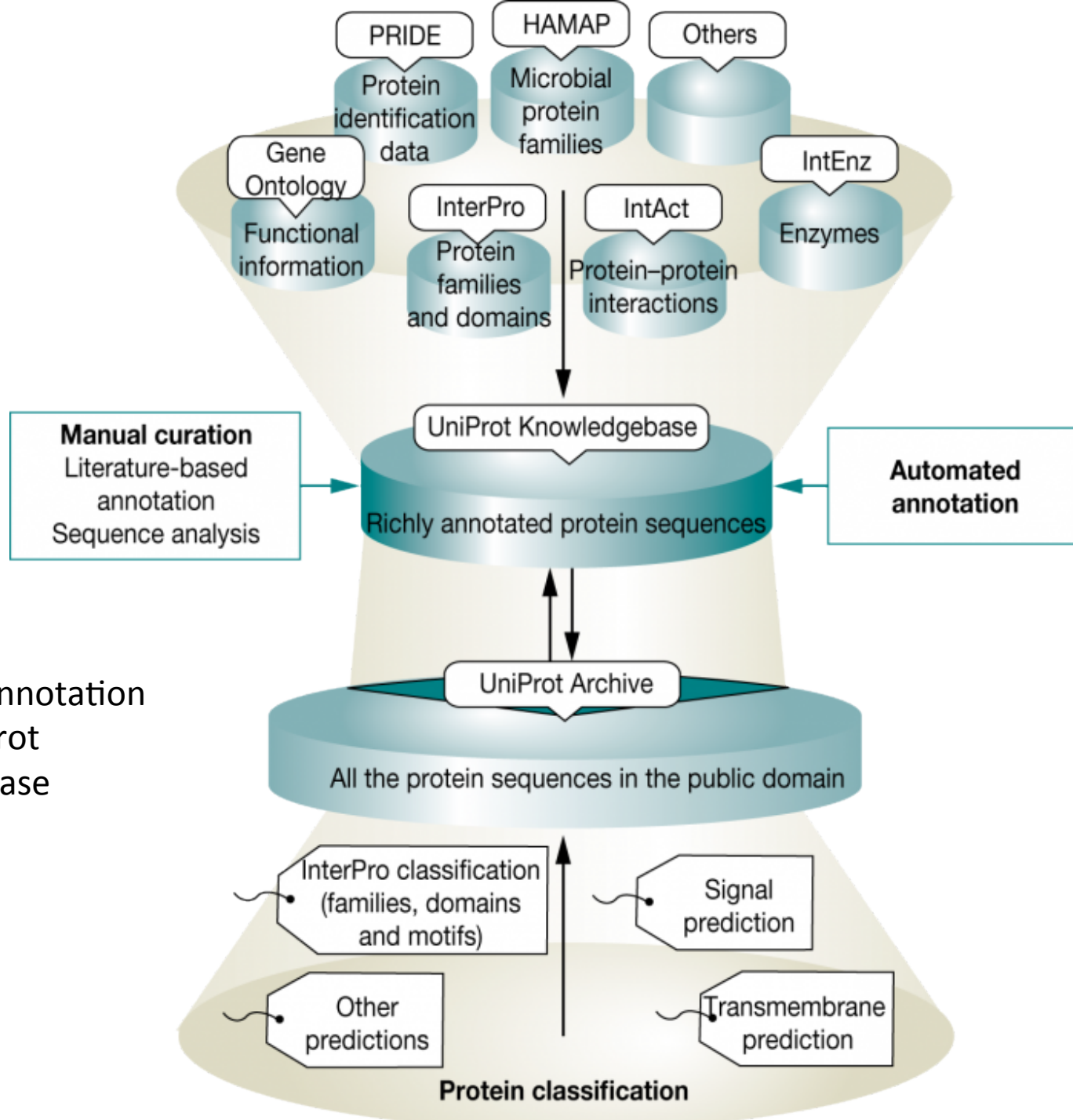
### Data classes

### Taxonomic Divisions

	CON	EST	GSS	HTC	HTG	MSA	PAT	STS	STD	TSA	WGA
A											
HUM											
MUS											
ROD											
MAM											
VRT											
FUN											
INV											
:											

# UniProt





Sources of annotation  
for the UniProt  
Knowledgebase

# Hands on practice 1: UniProt

← → ↻ www.uniprot.org

corecarb – Google Se George Mason Unive Customize Links Free Hotmail RealPlayer Windows Marketplac Windows Media Windows Gmail – Inbox (77) – trav

UniProt

Search Blast Align Retrieve **ID Mapping**

Database identifiers or file

Choose File No file chosen

From

EMBL/GenBank/DDBJ Map

To

UniProtKB AC Swap Clear

## WELCOME

The mission of **UniProt** is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

## What we provide

UniProtKB	<p>Protein knowledgebase, consists of two sections:</p> <ul style="list-style-type: none"><li>★ Swiss-Prot, which is manually annotated and reviewed.</li><li>★ TrEMBL, which is automatically annotated and is <b>not</b> reviewed.</li></ul> <p>Includes <a href="#">complete and reference proteome sets</a>.</p>
-----------	--

## NEWS



### UniProt release 2013\_01 - Jan 9, 2013

Hereditary sensory and autonomic neuropathy type IA: New dietary hope? | UniRef news

- › Statistics for UniProtKB:  
[Swiss-Prot](#) · [TrEMBL](#)
- › [Forthcoming changes](#)
- › [News archives](#)

Follow @uniprot 509 followers

## SITE TOUR

Search

Blast

Align

Retrieve

ID Mapping

Database identifiers

or file

Choose File 1

From

TAIR

Map

To

UniProtKB AC

Swap

Clear

## WELCOME

The mission of **UniProt** is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

## What we provide

UniProtKB	<p>Protein knowledgebase, consists of two sections:</p> <ul style="list-style-type: none"> <li>★ Swiss-Prot, which is manually annotated and reviewed.</li> <li>★ TrEMBL, which is automatically annotated and is <b>not</b> reviewed.</li> </ul> <p>Includes <a href="#">complete and reference proteome sets</a>.</p>
UniRef	Sequence clusters, used to speed up sequence similarity searches.

## NEWS

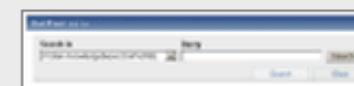
### UniProt release 2

Hereditary sensory an  
dietary hope? | UniRe

- › Statistics for UniProt  
[Swiss-Prot](#) · [TrEMBL](#)
- › [Forthcoming change](#)
- › [News archives](#)

 Follow @uniprot

## SITE TOUR





**Database identifiers** or file

? from file: 1

Choose File No file chosen

**From**

TAIR

**To**

UniProtKB AC

Map

Swap

Clear

⚠ Only IDs from the file are being mapped.

61 out of 61 identifiers mapped to 65 identifiers in the target data set

Download the [mapping table](#) or [target list](#) | UniProtKB (65)

From	To
At2g37040	P35510
At3g53260	P45724
At5g04230	P45725
At3g10340	Q9SS45
At2g30490	P92994
At1g51680	Q42524
At3g21240	Q9S725
At3g21230	Q9LU36
At1g65060	Q9S777
At1g20510	Q84P21
At1g20500	P0C5B6
At1g20490	Q3E6Y4
At1g20480	Q84P25
At1g62940	Q9LQ12
At4g19010	Q84P24
At4g05160	Q9M0X9
At5g63380	Q84P23
At5g38120	Q84P26
At5g48930	Q9FI78
At2g40890	O22203
At1g74540	Q9CA61
At1g74550	Q9CA60
At4g34050	O49499

## Results [Customize](#)

Show only [reviewed \(39\)](#) ★ (UniProtKB/Swiss-Prot) or [unreviewed \(26\)](#) ★ (UniProtKB/TrEMBL) entries

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
<input checked="" type="checkbox"/> <a href="#">Q42524</a>	4CL1_ARATH	★	4-coumarate—CoA ligase 1	<b>4CL1</b> At1g51680 F19C24.11	Arabidopsis thaliana (Mouse-ear cress)	561
<input checked="" type="checkbox"/> <a href="#">Q9S725</a>	4CL2_ARATH	★	4-coumarate—CoA ligase 2	<b>4CL2</b> At3g21240 MXL8.10	Arabidopsis thaliana (Mouse-ear cress)	556
<input checked="" type="checkbox"/> <a href="#">Q9S777</a>	4CL3_ARATH	★	4-coumarate—CoA ligase 3	<b>4CL3</b> At1g65060 F16G16.6	Arabidopsis thaliana (Mouse-ear cress)	561
<input checked="" type="checkbox"/> <a href="#">Q9LU36</a>	4CL4_ARATH	★	4-coumarate—CoA ligase 4	<b>4CL4</b> At3g21230 MXL8_9	Arabidopsis thaliana (Mouse-ear cress)	570
<input type="checkbox"/> <a href="#">Q9LQ12</a>	4CLL1_ARATH	★	4-coumarate—CoA ligase-like 1	<b>4CLL1</b> At1g62940 F16P17.9	Arabidopsis thaliana (Mouse-ear cress)	542
<input type="checkbox"/> <a href="#">Q84P25</a>	4CLL2_ARATH	★	4-coumarate—CoA ligase-like 2	<b>4CLL2</b> At1g20480 F5M15.29	Arabidopsis thaliana (Mouse-ear cress)	565
<input type="checkbox"/> <a href="#">Q3E6Y4</a>	4CLL3_ARATH	★	4-coumarate—CoA ligase-like 3	<b>4CLL3</b> At1g20490 F5M15.30	Arabidopsis thaliana (Mouse-ear cress)	552
<input type="checkbox"/> <a href="#">P0C5B6</a>	4CLL4_ARATH	★	4-coumarate—CoA ligase-like 4	<b>4CLL4</b> At1g20500 F5M15.18	Arabidopsis thaliana (Mouse-ear cress)	550
<input type="checkbox"/> <a href="#">Q84P21</a>	4CLL5_ARATH	★	4-coumarate—CoA ligase-like 5	<b>4CLL5</b> OPCL1 At1g20510 F5M15.17	Arabidopsis thaliana (Mouse-ear cress)	546
<input type="checkbox"/> <a href="#">Q84P24</a>	4CLL6_ARATH	★	4-coumarate—CoA ligase-like 6	<b>4CLL6</b> At4g19010 F13C5.180	Arabidopsis thaliana (Mouse-ear cress)	566
<input type="checkbox"/> <a href="#">Q9M0X9</a>	4CLL7_ARATH	★	4-coumarate—CoA ligase-like 7	<b>4CLL7</b> At4g05160 C17L7.80	Arabidopsis thaliana (Mouse-ear cress)	544
<input type="checkbox"/> <a href="#">Q84P26</a>	4CLL8_ARATH	★	4-coumarate—CoA ligase-like 8	<b>4CLL8</b> At5g38120 MXA21.23 MXA21_10	Arabidopsis thaliana (Mouse-ear cress)	550
<input type="checkbox"/> <a href="#">Q84P23</a>	4CLL9_ARATH	★	4-coumarate—CoA ligase-like 9	<b>4CLL9</b> At5g63380 K9H21.11 K9H21.8	Arabidopsis thaliana (Mouse-ear cress)	562
<input type="checkbox"/> <a href="#">Q42600</a>	C84A1_ARATH	★	Cytochrome P450 84A1	<b>CYP84A1</b> FAH1 At4g36220 F23E13.110	Arabidopsis thaliana (Mouse-ear cress)	520
<input type="checkbox"/> <a href="#">Q22203</a>	C98A3_ARATH	★	Cytochrome P450 98A3	<b>CYP98A3</b> C3'H REF8 At2g40890 T20B5.9	Arabidopsis thaliana (Mouse-ear cress)	508
<input type="checkbox"/> <a href="#">Q9CA61</a>	C98A8_ARATH	★	Cytochrome P450 98A8	<b>CYP98A8</b> At1g74540 F1M20.22	Arabidopsis thaliana (Mouse-ear cress)	497
<input type="checkbox"/> <a href="#">Q9CA60</a>	C98A9_ARATH	★	Cytochrome P450 98A9	<b>CYP98A9</b> At1g74550 F1M20.23	Arabidopsis thaliana (Mouse-ear cress)	487
<input type="checkbox"/> <a href="#">Q9CAI3</a>	CADH1_ARATH	★	Probable cinnamyl alcohol dehydrogenase 1	<b>CAD1</b> CADG At1g72680 F28P22.13	Arabidopsis thaliana (Mouse-ear cress)	355
<input type="checkbox"/> <a href="#">Q9SJ25</a>	CADH2_ARATH	★	Cinnamyl alcohol dehydrogenase 2	<b>CAD2</b> CAD7 CADE LCAD-E At2g21730 F7D8.5	Arabidopsis thaliana (Mouse-ear cress)	376

4 selected: [Q9LU36](#) [Q9S777](#) [Q9S725](#) [Q42524](#) cinnamyl alcohol dehydrogenase 2

CAD3 CAD8 LCAD-F At2g21890 F7D8.21

Arabidopsis thaliana

[Retrieve](#)

[Align](#)

[Blast](#)

[Clear](#)

## UniProt identifiers

Q42524  
Q9S725  
Q9S777  
Q9LU36

## or file

Choose File No file chosen

Retrieve

Clear

4 unique items available for download

 UniProtKB (4)

› Download data [compressed](#) or **uncompressed**

**FASTA**

Sequence data in FASTA format.

[ [Download](#) ( 2 KB\* ) | [Open](#) ]

**GFF**

Sequence features in GFF.

[ [Download](#) ( 7 KB\* ) | [Open](#) ]

**Flat Text**

Complete data in the original flat text format.

[ [Download](#) ( 20 KB\* ) | [Open](#) ]

**XML**

Complete data in XML format.

[ [Download](#) | [Open](#) ]

**RDF/XML**

Complete data in RDF format.

[ [Download](#) | [Open](#) ]

**List**

List of identifiers.

[ [Download](#) | [Open](#) ]

[Search](#)[Blast](#)[Align](#)[Retrieve](#)[ID Mapping](#)

Sequences (in FASTA format) or UniProt identifiers

[Align](#)[Clear](#)


```
>sp|Q42524|4CL1_ARATH 4-coumarate--CoA ligase 1
OS=Arabidopsis thaliana GN=4CL1 PE=1 SV=1
MAPQEQAQVSQVMEKQSNNNNSDVIFRSKLPDIYIPNHLSLHDYIFQNISEFATKPC
LING
PTGHVYTYSDVHVISRQIAANFHKLGVNQNDVVMILLPNCPEFVLSFLAASFRGA
TATAA
```

**Help**

To align several protein sequences

- two or more sequences in FASTA
- two or more UniProt identifiers, e

TPA\_HUMAN  
TPA\_PIG

[text](#) [tree](#) [fas](#)[Alignment](#) · [Tree](#) · [Annotation](#) · [Job information](#) [Customize order](#)**Alignment** Learn how to print this alignment in color

1	MA-----P--QEQA-VSQV-MEKQSNNNNSDVIFRSKLPDIYIPNHLSLHDYIFQNISE	50	<a href="#">Q42524</a>	4CL1_ARATH
1	MT-----T--QDVI-VN---DQNDQKQCSNDVIFRSRLPDIYIPNHLPLHDYIFENISE	48	<a href="#">Q9S725</a>	4CL2_ARATH
1	MITAALHEPQIHKPTDTSVVSDDVLPSPPTPRIFRSKLPDIDIPNHLPLHTYCFEKLSS	60	<a href="#">Q9S777</a>	4CL3_ARATH
1	MV-----LQQQTHFLTKKIDQEDEEEEP SHDFIFRSKLPDIFIPNHLPLTDYVFQRFSG	54	<a href="#">Q9LU36</a>	4CL4_ARATH
	* : . . : . . ***** * * * : *			
51	F----ATKPCLINGPTGHVYTYSDVHVISRQIAANFHKLGVNQNDVVMILLPNCPEFVLS	106	<a href="#">Q42524</a>	4CL1_ARATH
49	F----AAKPCLINGPTGEVYTYADVHVTSRKLAAGLHNLGVKQHDVVMILLPNSPEVVL	104	<a href="#">Q9S725</a>	4CL2_ARATH
61	V----SDKPCLIVGSTGKSYTYGETHLICRRVASGLYKLGIRKGDVIMILLQNSAEFVFS	116	<a href="#">Q9S777</a>	4CL3_ARATH
55	DGDGDSSTTCIIDGATGRILTYADVQTNMRRIAAGIHLRGIRHGDVVMILLPNSPEFALS	114	<a href="#">Q9LU36</a>	4CL4_ARATH
	: . * : * * . * . : : * : : : * : * : * : * : * : * : *			
107	FLAASFRGATATAANPFPTPAEIAKQAKASNTKLIITEARYVDKIKPLQNDGCVVIVCID	166	<a href="#">Q42524</a>	4CL1_ARATH
105	FLAASFIGAITTSANPFPTPAEISQAKASAAKLIIVTSQRYVDKIKNLQNDGVLI-V---	160	<a href="#">Q9S725</a>	4CL2_ARATH
117	FMGASMIGAVSTTANPFYTSQELYQLKSSGAKLIITHSQYVDKLNKLGENTLIT----	172	<a href="#">Q9S777</a>	4CL3_ARATH
115	FLAVAYLGAVSTTANPFYTQPEIAKQAKASAAKMIITKKCLVDKLTNLKNDGVLI-VCLD	173	<a href="#">Q9LU36</a>	4CL4_ARATH
	* : . : * * : : * : * * * : * : * : * : * : * : * : * : * : *			
167	DNESV----PIPEGCLRFTELQSTTEASEVIDSVEISPDVVALPYSSGTTGLPKGVML	222	<a href="#">Q42524</a>	4CL1_ARATH
161	TTDS-----AIPENCLRFSELQSEPRVDSI-PEKISPEDVVALPFSSGTTGLPKGVML	215	<a href="#">Q9S725</a>	4CL2_ARATH
173	-----TDEPTPENCLPFSTLITDDETNP-FQETVDIGDDAAALPFSSGTTGLPKGVVL	225	<a href="#">Q9S777</a>	4CL3_ARATH
174	DDGDNVGVSSSDDGCVSFTELQADETE---LLKPKISPEDTVAMPYSSGTTGLPKGVMI	230	<a href="#">Q9LU36</a>	4CL4_ARATH
	: : * : * * : . . : * . : * . : * : * : * : * : * : * : *			
223	THKGLVTSVAQQVDGENPNLYFHSDDVILCVLPMFHIYALNSIMLCGLRVGAAILMPKF	282	<a href="#">Q42524</a>	4CL1_ARATH
216	THKGLVTSVAQQVDGENPNLYFNRDDVILCVLPMFHIYALNSIMLCGLRVGATILMPKF	275	<a href="#">Q9S725</a>	4CL2_ARATH
226	THKGLVTSVAQQVDGENPNLYFHSDDVILCVLPMFHIYALNSIMLCGLRVGATILMPKF	285	<a href="#">Q9S777</a>	4CL3_ARATH

**Annotation**

- ☐ Mutagenesis
- ☐ Sequence conflict
- ☐ Nucleotide binding
- ☐ Binding site
- ☐ Region
- ☐ Alternative sequence
- ☐ Chain

**Amino acid properties**

- ☐ Similarity
- ☐ Hydrophobic
- ☐ Negative
- ☐ Positive
- ☐ Aliphatic
- ☐ Tiny
- ☐ Aromatic
- ☐ Charged
- ☐ Small
- ☐ Polar
- ☐ Big
- ☐ Serine Threonine

## Sequences

Sequence	Length	Mass (Da)	Tools
<input type="checkbox"/> <b>Isoform 1</b> [UniParc]. Last modified November 1, 1996. Version 1. Checksum: 5A9E20816D0C0D07	FASTA	561	61,053 <div>                         Blast                         <input type="button" value="go"/> </div>

<u>10</u>	<u>20</u>	<u>30</u>	<u>40</u>	<u>50</u>	<u>60</u>
MAPQEQAVSQ	VMEKQSNNNN	SDVIFRSKLP	DIYIPNHL	SL HDYIFQNI	SE FATKPCLING
<u>70</u>	<u>80</u>	<u>90</u>	<u>100</u>	<u>110</u>	<u>120</u>
PTGHVYTYSD	VHVISRQIAA	NFHKLGVNQ	N DVVMLLLP	NC PEFVLSFL	AA SFRGATATAA
<u>130</u>	<u>140</u>	<u>150</u>	<u>160</u>	<u>170</u>	<u>180</u>
NPFFTPAEIA	KQAKASNTKL	IITEARYVDK	IKPLQND	DDGV VIVCIDD	NES VPIPEGCLRF
<u>190</u>	<u>200</u>	<u>210</u>	<u>220</u>	<u>230</u>	<u>240</u>
TELTQSTTEA	SEVIDSVEIS	PDDVVALPYS	SGTTGLPKGV	MLTHKGLVTS	VAQQVDGENP
<u>250</u>	<u>260</u>	<u>270</u>	<u>280</u>	<u>290</u>	<u>300</u>
NLYFHSDDVI	LCVLPMFHIY	ALNSIMLCGL	RVGAAILIMP	KFEINLLEL	IQRCKVTVAP
<u>310</u>	<u>320</u>	<u>330</u>	<u>340</u>	<u>350</u>	<u>360</u>
MVPPIVLAIA	KSSETEKYDL	SSIRVVKSGA	APLGKELEDA	VNAKFPNAKL	GQGYGMTEAG
<u>370</u>	<u>380</u>	<u>390</u>	<u>400</u>	<u>410</u>	<u>420</u>
PVLAMSLGFA	KEPFPVKSGA	CGTVVRNAEM	KIVDPDTGDS	LSRNQPGEIC	IRGHQIMKGY
<u>430</u>	<u>440</u>	<u>450</u>	<u>460</u>	<u>470</u>	<u>480</u>
LNNPAATAET	IDKDGWLHTG	DIGLIDDDDE	LFIVDRLKEL	IKYKGFQVAP	AELEALLIGH
<u>490</u>	<u>500</u>	<u>510</u>	<u>520</u>	<u>530</u>	<u>540</u>

## Search in

Protein Knowledgebase (UniProtKB)

## Query



### Field

Organism [OS]

### Term

human



## WELCOME

The mission of **UniProt** is to provide the scientific community with comprehensive, high-quality and freely accessible protein sequence and functional information.

## What we provide

UniProtKB	Protein knowledgebase, containing: <ul style="list-style-type: none"> <li>★ Swiss-Prot, which is manually reviewed.</li> <li>★ TrEMBL, which is automatically reviewed.</li> </ul> Includes <a href="#">complete and reference proteomes</a> .
UniRef	Sequence clusters, used to speed up searches.
UniParc	Sequence archive, used to keep their identifiers.
Supporting data	<a href="#">Literature citations</a> , <a href="#">taxonomy</a> , <a href="#">locations</a> , <a href="#">cross-referenced data</a>

## Getting started

- [Text search](#)

Human [9606]  
 Human adenovirus A [129875]  
 Human adenovirus B [108098]  
 Human adenovirus D [130310]  
 Human adenovirus B2 [565303]  
 Human rhinovirus A [147711]  
 Human calicivirus  
 HU/NLV/Wortley/90/UK [122922]  
 Human adenovirus 50 [107462]  
 Human adenovirus E [130308]  
 Human enterovirus A [138948]  
 Human enterovirus C [138950]  
 Human enteric calicivirus [82658]  
 Human calicivirus HU/NLV/Rbh/93/UK [122921]  
 Human calicivirus  
 HU/NLV/Winchester/94/UK [122913]  
 Human calicivirus  
 HU/NLV/Thistlehall/90/UK [122925]  
 Southampton virus (strain  
 GI/Human/United  
 Kingdom/Southampton/1991) [37129]  
 Human calicivirus  
 HU/NLV/Whiterose/96/UK [122914]  
 Human calicivirus  
 HU/NLV/Birmingham/93/UK [122916]  
 Human adenovirus F [130309]  
 Human calicivirus  
 HU/NLV/Musgrove/89/UK [122918]  
 Human enterovirus B [138949]

## NEWS

### UniProt release 2013\_01 - Jan

Hereditary sensory and autonomic neuropathy: dietary hope? | UniRef news

- › [Statistics for UniProtKB: Swiss-Prot · TrEMBL](#)
- › [Forthcoming changes](#)
- › [News archives](#)

Follow @uniprot

509 followers

## SITE TOUR



Learn how to make best use of the tool



Search

Blast

Align

Retrieve

ID Mapping \*

Search in

Query

Protein Knowledgebase (UniProtKB) ▾

organism:"Human [9606]"

Search

Advanced Search »

Clear

1 - 25 of 133,808 results for organism:"Homo sapiens (Human) [9606]" in UniProtKB sorted by score descending ☒

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50%

Results [Customize](#)

- › Show only [reviewed \(20,232\)](#) ★ (UniProtKB/Swiss-Prot) or [unreviewed \(113,576\)](#) ★ (UniProtKB/TrEMBL) entries
- › Expand search to "Homo sapiens (Human) [9606]" to include lower taxonomic ranks
- › Show only entries from a [complete proteome set \(70,584\)](#)
- › Show only entries from a [reference proteome set \(70,584\)](#)

	Entry	Entry name	Status	Protein names	Gene names	
<input type="checkbox"/>	<a href="#">P31946</a>	1433B_HUMAN	★	14-3-3 protein beta/alpha	YWHAB	Ho
<input type="checkbox"/>	<a href="#">P62258</a>	1433E_HUMAN	★	14-3-3 protein epsilon	YWHAE	Ho
<input type="checkbox"/>	<a href="#">Q04917</a>	1433F_HUMAN	★	14-3-3 protein eta	YWHAH YWHA1	Ho
<input type="checkbox"/>	<a href="#">P61981</a>	1433G_HUMAN	★	14-3-3 protein gamma	YWHAG	Ho
<input type="checkbox"/>	<a href="#">P31947</a>	1433S_HUMAN	★	14-3-3 protein sigma	SFN HME1	Ho
<input type="checkbox"/>	<a href="#">P27348</a>	1433T_HUMAN	★	14-3-3 protein theta	YWHAQ	Ho
<input type="checkbox"/>	<a href="#">P63104</a>	1433Z_HUMAN	★	14-3-3 protein zeta/delta	YWHAZ	Ho
<input type="checkbox"/>	<a href="#">P30443</a>	1A01_HUMAN	★	HLA class I histocompatibility antigen, A-1 a...	HLA-A HLAA	Ho
<input type="checkbox"/>	<a href="#">P01892</a>	1A02_HUMAN	★	HLA class I histocompatibility antigen, A-2 a...	HLA-A HLAA	Ho
<input type="checkbox"/>	<a href="#">P04439</a>	1A03_HUMAN	★	HLA class I histocompatibility antigen, A-3 a...	HLA-A HLAA	Ho
<input type="checkbox"/>	<a href="#">P13746</a>	1A11_HUMAN	★	HLA class I histocompatibility antigen, A-11 ...	HLA-A HLAA	Ho
<input type="checkbox"/>	<a href="#">Q96QU6</a>	1A1L1_HUMAN	★	1-aminocyclopropane-1-carboxylate synthase-li...	ACCS PHACS	Ho
<input type="checkbox"/>	<a href="#">Q4AC99</a>	1A1L2_HUMAN	★	1-aminocyclopropane-1-carboxylate synthase-li...	ACCSL	Ho
<input type="checkbox"/>	<a href="#">P30447</a>	1A23_HUMAN	★	HLA class I histocompatibility antigen, A-23 ...	HLA-A HLAA	Ho
<input type="checkbox"/>	<a href="#">P05534</a>	1A24_HUMAN	★	HLA class I histocompatibility antigen, A-24 ...	HLA-A HLAA	Ho



Search in

Query

Protein Knowledgebase (UniProtKB)

organism:9606 AND keyword:1185

Search

Advanced Search »

Clear

1 - 25 of 70,584 results for organism:"Homo sapiens (Human) [9606]" AND keyword:"Reference proteome [1185]" in UniProtKB sorted by score descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50%

Results [Customize](#)

- Show only reviewed (20,226) ★ (UniProtKB/Swiss-Prot) or unreviewed (50,358) ★ (UniProtKB/TrEMBL) entries
- Expand search to "Homo sapiens (Human) [9606]" to include lower taxonomic ranks

Entry	Entry name	Status	Protein names	Gene names
<input type="checkbox"/> <a href="#">P31946</a>	1433B_HUMAN	★	14-3-3 protein beta/alpha	YWHAB
<input type="checkbox"/> <a href="#">P62258</a>	1433E_HUMAN	★	14-3-3 protein epsilon	YWHAE
<input type="checkbox"/> <a href="#">Q04917</a>	1433F_HUMAN	★	14-3-3 protein eta	YWHAH YWHA1
<input type="checkbox"/> <a href="#">P61981</a>	1433G_HUMAN	★	14-3-3 protein gamma	YWHAG
<input type="checkbox"/> <a href="#">P31947</a>	1433S_HUMAN	★	14-3-3 protein sigma	SFN HME1
<input type="checkbox"/> <a href="#">P27348</a>	1433T_HUMAN	★	14-3-3 protein theta	YWHAQ
<input type="checkbox"/> <a href="#">P63104</a>	1433Z_HUMAN	★	14-3-3 protein zeta/delta	YWHAZ
<input type="checkbox"/> <a href="#">P30443</a>	1A01_HUMAN	★	HLA class I histocompatibility antigen, A-1 a...	HLA-A HLAA
<input type="checkbox"/> <a href="#">P01892</a>	1A02_HUMAN	★	HLA class I histocompatibility antigen, A-2 a...	HLA-A HLAA
<input type="checkbox"/> <a href="#">P04439</a>	1A03_HUMAN	★	HLA class I histocompatibility antigen, A-3 a...	HLA-A HLAA
<input type="checkbox"/> <a href="#">P13746</a>	1A11_HUMAN	★	HLA class I histocompatibility antigen, A-11 ...	HLA-A HLAA
<input type="checkbox"/> <a href="#">Q96QU6</a>	1A1L1_HUMAN	★	1-aminocyclopropane-1-carboxylate synthase-li...	ACCS PHACS
<input type="checkbox"/> <a href="#">Q4AC99</a>	1A1L2_HUMAN	★	1-aminocyclopropane-1-carboxylate synthase-li...	ACCSL
<input type="checkbox"/> <a href="#">P30447</a>	1A23_HUMAN	★	HLA class I histocompatibility antigen, A-23 ...	HLA-A HLAA
<input type="checkbox"/> <a href="#">P05534</a>	1A24_HUMAN	★	HLA class I histocompatibility antigen, A-24 ...	HLA-A HLAA

Search in

Query

Protein Knowledgebase (UniProtKB)

organism:9606 AND keyword:1185

Search

Advanced Search »

Clear

70,584 results for **organism:"Homo sapiens (Human) [9606]"** AND **keyword:"Reference proteome [1185]"** in UniProtKB sorted by **score** descending

› Download data [compressed](#) or [uncompressed](#)

› Limit to [1,000](#) results


### Tab-Delimited

Summary information from the result view.

[ [Download](#) | [Open](#) | [Open first 10](#) ]

### Excel

Summary information from the result view for MS Excel™.

 The result set will be truncated in Excel versions older than 2007 and in OpenOffice, as it contains [too many rows](#).

[ [Download](#) | [Open](#) | [Open first 10](#) ]

### FASTA

Canonical sequence data in FASTA format.

[ [Download](#) (40 MB\*) | [Open](#) | [Open first 10](#) ]

Canonical and isoform sequence data in FASTA format.

[ [Download](#) (40 MB\*) | [Open](#) | [Open first 10](#) ]

### GFF

Sequence annotation in GFF format.

[ [Download](#) (200 MB\*) | [Open](#) | [Open first 10](#) ]

### Flat Text

Complete data in the original flat text format.

[ [Download](#) (200 MB\*) | [Open](#) | [Open first 10](#) ]

### XML

Complete data in XML format.

[ [Download](#) (400 MB\*) | [Open](#) | [Open first 10](#) ]

### RDF/XML

Complete data in RDF format.

[ [Download](#) (700 MB\*) | [Open](#) | [Open first 10](#) ]

### List

Search in

Query

Protein Knowledgebase (UniProtKB)

Search

Advanced Search »

Clear

## WELCOME

The mission of **UniProt** is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

## What we provide

UniProtKB	<p>Protein knowledgebase, consists of two sections:</p> <ul style="list-style-type: none"> <li>★ Swiss-Prot, which is manually annotated and reviewed.</li> <li>★ TrEMBL, which is automatically annotated and is <b>not</b> reviewed.</li> </ul> <p>Includes <b>complete and reference proteome sets</b>.</p>
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, <b>taxonomy</b> , <b>keywords</b> , <b>subcellular locations</b> , <b>cross-referenced databases</b> and more.

## Getting started

- [Text search](#)
- [Sequence similarity searches \(BLAST\)](#)
- [Sequence alignments](#)

## NEWS

### UniProt release 2013\_01 - Jan 9,

Hereditary sensory and autonomic neuropathy: dietary hope? | UniRef news

- › [Statistics for UniProtKB: Swiss-Prot · TrEMBL](#)
- › [Forthcoming changes](#)
- › [News archives](#)



Follow @uniprot

509 followers

## SITE TOUR



Learn how to make best use of the tools a

Search

Blast

Align

Retrieve

ID Mapping

Search in




Taxonomy

Query

Search

Advanced Search »

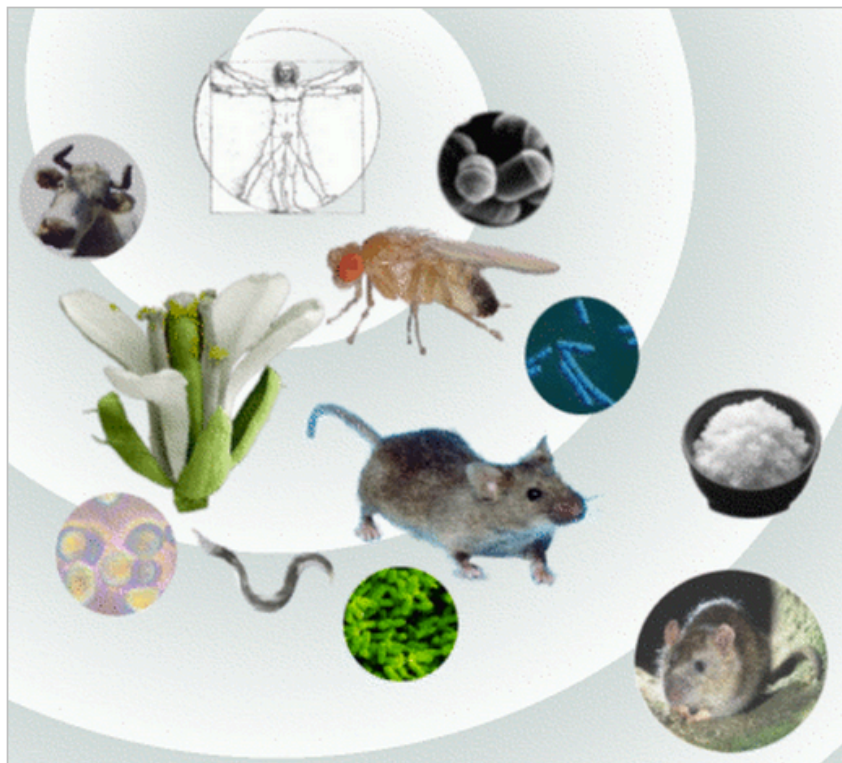
Clear

 Browse by [hierarchy](#) | 
  [List all taxa](#) ( 952,849 ) | 
  [Complete and reference proteomes](#)

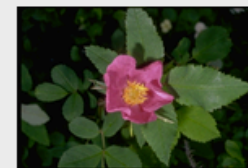
## TAXONOMY

Organisms are classified in a hierarchical tree structure. Our taxonomy database contains every node (taxon) of the tree. UniProtKB taxonomy data is manually curated: next to manually verified [organism names](#), we provide a selection of external links, organism [strains](#) and [viral host](#) information. [More »](#)

[Archaea](#) · [Bacteria](#) · [Eukaryota](#) · [Viruses](#)



## FEATURING



[Rosa acicularis](#)

## FAQ

- › [What are complete proteome sets?](#)
- › [What are reference proteome sets?](#)
- › [What is the human complete proteome?](#)
- › [How can I download a list of all viruses infecting humans?](#)

[More »](#)

## HEADLINES

- › [Sex by deception | Update to Reference proteomes in UniProtKB](#)
- › [What's in a \(species\) name? | Clustal Omega](#)
- › [Complete proteome sets for Homo sapiens and Mus musculus](#)
- › [Viral reference strains | Changes in cross-references to WormBase](#)

[More »](#)



Search

Blast

Align

Retrieve

ID Mapping

Search in

Taxonomy

Query

\* AND complete:yes

Search

Advanced Search »

C











## COMPLETE PROTEOMES AND REFERENCE PROTEOMES

A [complete proteome](#) consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

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

These organisms can be searched via the taxonomy pages, which provide links to download complete and reference proteome sets when available, as well as links to the HAMAP web site.

Browse or list organisms with:

Complete proteomes	Reference proteomes
 <a href="#">Browse by hierarchy</a>	 <a href="#">Browse by hierarchy</a>
 <a href="#">List all Bacteria</a>	 <a href="#">List all Bacteria</a>
 <a href="#">List all Archaea</a>	 <a href="#">List all Archaea</a>
 <a href="#">List all Eukaryota</a>	 <a href="#">List all Eukaryota</a>
 <a href="#">List all Viruses</a>	 <a href="#">List all Viruses</a>

Search organisms with complete proteomes:

Search organisms with reference proteomes:

## FAQ

› [What are complete proteome sets?](#)

› [What are reference proteome sets?](#)

› [How to retrieve sets of protein sequences?](#)

› [What is HAMAP?](#)

HAMAP is a system, based on manual protein annotation, that identifies and semi-automatically annotates proteins... [More](#)

Search

Blast

Align

Retrieve

ID Mapping

Search in

Query

Protein Knowledgebase (UniProtKB)

Search

Advanced Search »

Clear

## Downloads

UniProt is updated every four weeks (see FAQ on [how to be notified automatically of updates](#)). You can download small data sets and subsets directly from ~~this web site~~ by following the download link on any search result page. For downloading complete data sets we recommend using <ftp.uniprot.org>. If you are located in Europe, the Middle East or Africa, you may want to download data from our mirror site in the [United Kingdom](#) or in [Switzerland](#) instead.

Here are some direct links to frequently downloaded files:

UniProtKB	UniProtKB/Swiss-Prot	<a href="#">xml</a> <a href="#">fasta</a> <a href="#">text</a>
	UniProtKB/TrEMBL	<a href="#">xml</a> <a href="#">fasta</a> <a href="#">text</a>
	<a href="#">Isoform sequences</a>	<a href="#">fasta</a>
	<a href="#">Taxonomic divisions</a>	<a href="#">text</a>
	<a href="#">Proteomes</a> (Only a selected set of species are available on the FTP site. All species can be downloaded from the web site.)	<a href="#">fasta</a>
	<a href="#">ID mapping data</a>	<a href="#">tab</a>
	<a href="#">Documents</a>	
	<a href="#">XML Schema</a>	
UniRef	UniRef100	<a href="#">xml</a> <a href="#">fasta</a>

# Hands on practice 2:

## UniProt ftp site



command → There is a space here! → This is called argument

<b>lftp</b> addr	command to connect to a remote ftp server
<b>cd</b> dir	change to the directory
<b>cd</b> ..	change to the upper folder (..)
<b>ls</b>	list files and folders in the current directory at once
<b>ls</b> dir	list files and folders in dir at once
<b>ls</b>   less	list page by page (good if the list is too long)
<b>get</b> file	get a file
<b>mirror</b> dir	get a folder
<b>zmore</b> file	view the file content
<b>by</b> or <b>bye</b>	exit lftp

Suppose you have lftp to <ftp.uniprot.org>:

right click to paste to the command line

You are in the **root folder (/)** after connected

You want to change to `/pub/databases/uniprot/current_release/knowledgebase/idmapping`

You can do

`cd /pub/databases/uniprot/current_release/knowledgebase/idmapping`

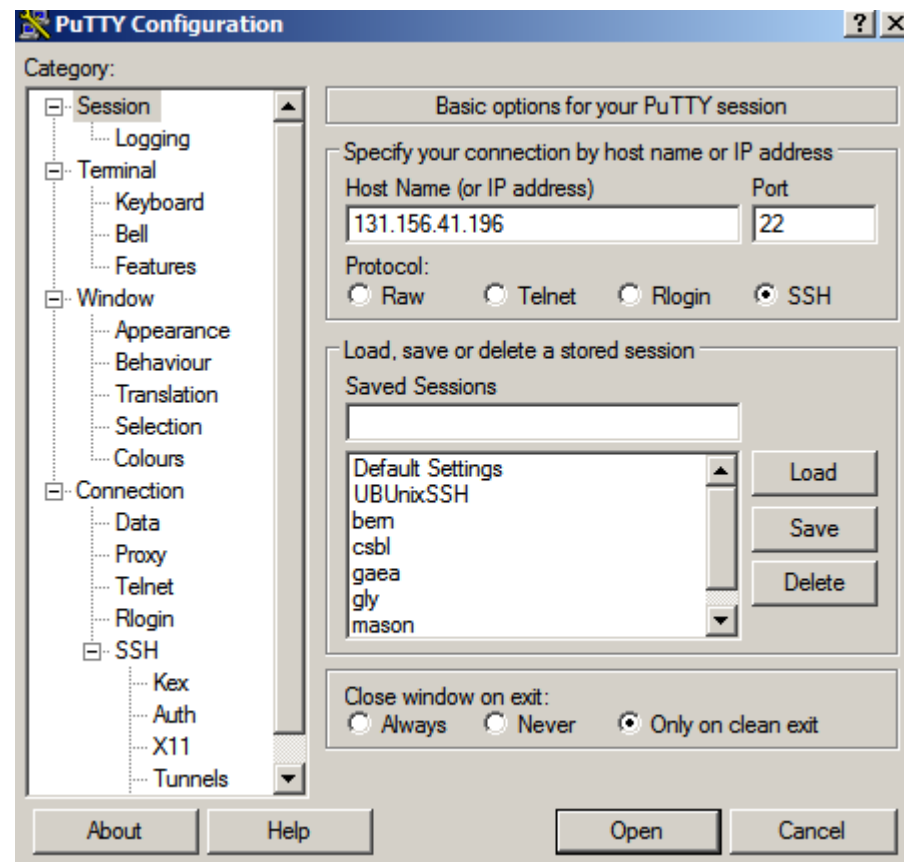
You **do not** need to type in the **full word**, instead type in **the first few letters** of a word and then press **tab** key to auto-complete the word (if exist)

Try to find human UniProt dataset and download it to glu

# login info

IP 131.156.41.196  
Account student ID (e.g. z1003529)  
Pswd student ID (e.g. z1003529)

z117479  
z1245176  
z1608050  
z1576493  
z1598039  
z1559435  
z1660438  
z1003529  
z1678230



PuTTY

```

yyin@gly: ~
lrwxrwxrwx    1 147      59          16 Jan 09 15:28 LICENSE -> ../../../../LICENSE
-rw-rw-r--    1 147      59      13186969 Jan 09 15:21 MOUSE.fasta.gz
-rw-rw-r--    1 147      59      7445498 Jan 09 15:21 PIG.fasta.gz
-rw-rw-r--    1 147      59     12401242 Jan 09 15:21 RAT.fasta.gz
-rw-rw-r--    1 147      59        3153 Jan 09 15:21 README
-rw-rw-r--    1 147      59     2002902 Jan 09 15:21 YEAST.fasta.gz
-rw-rw-r--    1 147      59        615 Jan 09 15:21 relnotes.txt
lftp ftp.uniprot.org:/pub/databases/uniprot/current_release/knowledgebase/teomes> ls
-rw-rw-r--    1 147      59      8625090 Jan 09 15:21 ARATH.fasta.gz
-rw-rw-r--    1 147      59      8258494 Jan 09 15:21 BOVIN.fasta.gz
-rw-rw-r--    1 147      59     6916006 Jan 09 15:21 CAEEL.fasta.gz
-rw-rw-r--    1 147      59     9465302 Jan 09 15:21 CANFA.fasta.gz
-rw-rw-r--    1 147      59     7180146 Jan 09 15:21 CHICK.fasta.gz
-rw-rw-r--    1 147      59     12920295 Jan 09 15:21 DANRE.fasta.gz
-rw-rw-r--    1 147      59     6330711 Jan 09 15:21 DROME.fasta.gz
-rw-rw-r--    1 147      59     17827662 Jan 09 15:21 HUMAN.fasta.gz
lrwxrwxrwx    1 147      59          16 Jan 09 15:28 LICENSE -> ../../../../LICENSE
-rw-rw-r--    1 147      59      13186969 Jan 09 15:21 MOUSE.fasta.gz
-rw-rw-r--    1 147      59      7445498 Jan 09 15:21 PIG.fasta.gz
-rw-rw-r--    1 147      59     12401242 Jan 09 15:21 RAT.fasta.gz
-rw-rw-r--    1 147      59        3153 Jan 09 15:21 README
-rw-rw-r--    1 147      59     2002902 Jan 09 15:21 YEAST.fasta.gz
-rw-rw-r--    1 147      59        615 Jan 09 15:21 relnotes.txt
lftp ftp.uniprot.org:/pub/databases/uniprot/current_release/knowledgebase/teomes>

```

# Gene Ontology

<http://www.geneontology.org/GO.doc.shtml>

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

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

Three structured **controlled vocabularies** (ontologies) that describe gene products in terms of their associated **biological processes**, **cellular components** and **molecular functions** in a **species-independent** manner.

There are three separate aspects to this effort:

- 1, the development and maintenance of the **ontologies** themselves;
- 2, the **annotation** of gene products, which entails **making associations between the ontologies and the genes and gene products** in the collaborating databases; and
- 3, development of **tools** that facilitate the creation, maintenance and use of ontologies.

# The scope of GO

Gene Ontology covers three domains:

**cellular component**, the parts of a cell or its extracellular environment;

**molecular function**, the elemental activities of a gene product at the molecular level, such as binding or catalysis;

**biological process**, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms

GO is **not a database of gene sequences**, nor a catalog of gene products. Rather, GO **describes how gene products behave** in a cellular context.

GO is not a dictated standard, mandating nomenclature across databases. Groups participate because of self-interest, and cooperate to arrive at a **consensus**.

GO is not a way to unify biological databases (i.e. GO is not a 'federated solution'). Sharing vocabulary is a step towards unification, but is not, in itself, sufficient.

id: GO:0000016  
name: lactase activity namespace: molecular\_function  
def: "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose." [EC:3.2.1.108]  
synonym: "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]  
synonym: "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]  
xref: EC:3.2.1.108  
xref: MetaCyc:LACTASE-RXN  
xref: Reactome:20536  
is\_a: GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds

## What can I do with GO?

### What can I do with GO?

One of the most popular uses of **GO** is to find significant shared GO terms (or parents of those GO terms) that are annotated to **genes** in a particular query set (e.g. a set of genes that are overexpressed in a microarray experiment). This process helps you to find out what those genes may have in common and is known as a **GO enrichment analysis**.

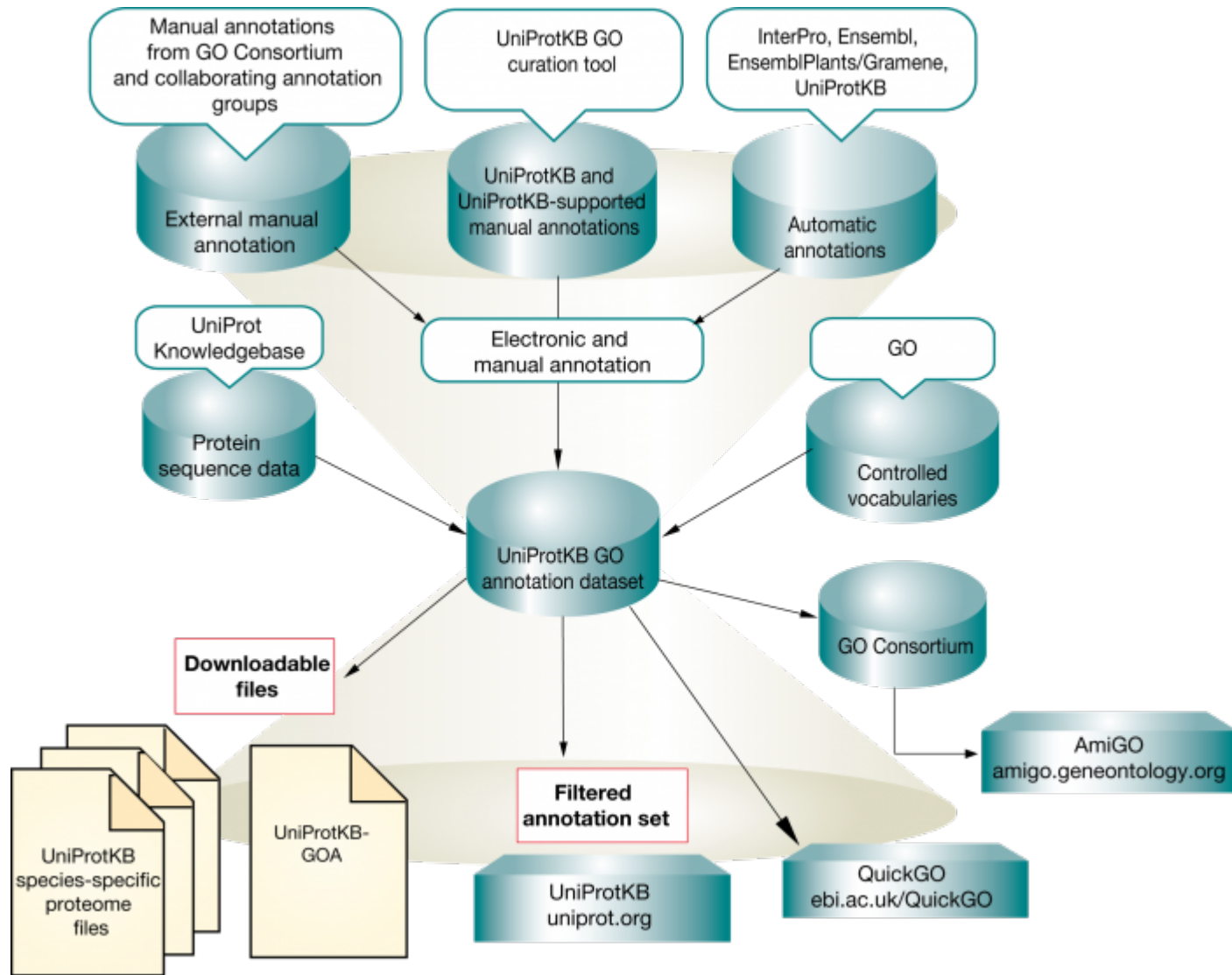
GO is also used for purposes as diverse as:

- integrating proteomic information from different organisms;
- assigning functions to protein domains;
- finding functional similarities in genes that are overexpressed or underexpressed in diseases and as we age;
- analysing groups of genes that are co-expressed during development;
- developing automated ways of deriving information about gene function from the literature;
- verifying models of genetic, metabolic and product interaction networks.

The [GO tools web page](#) lists the tools that you can use to analyse the data from GO.

<http://www.ebi.ac.uk/training/online/course/go-quick-tour/what-can-i-do-go>

# UniProt-GO annotation (GOA)





# UniProt-GOA format

The *reference* used to make the annotation (e.g. a journal article)

An *evidence code* denoting the type of evidence upon which the annotation is based

The date and the creator of the annotation

Gene product: Actin, alpha cardiac muscle 1, [UniProtKB:P68032](#)

GO term: [heart contraction](#) ; [GO:0060047](#) (biological process)

Evidence code: Inferred from Mutant Phenotype (IMP) Reference: [PMID 17611253](#)

Assigned by: UniProtKB, June 6, 2008

# The idea of GO annotation for new sequences

If you have a new genome/transcriptome sequenced, how do you perform a GO annotation for it?

1. Find a closet model organism which has been annotated by GO
2. BLAST your data against this closest organisms
3. Transfer the GO annotation of the best match to your query sequences

For instance, if we want to annotate fern transcriptome with GO function descriptions ....

1. Find Arabidopsis UniProt protein dataset
2. Find the Arabidopsis GOA association file
3. BLASTx fern reads (or assembled UniGenes) against the UniProt set
4. Analyze BLAST result to link fern reads GO terms

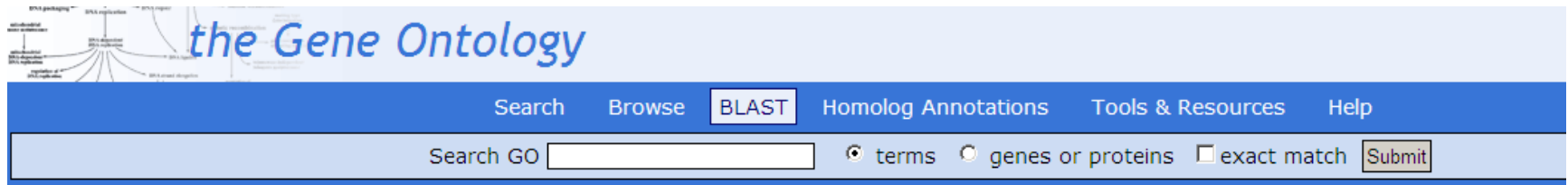
# Hands on practice 3: UniProt-GOA ftp site

lftp ftp.ebi.ac.uk:/pub/databases/GO/goa

```
yyin@gly: ~  
lftp ftp.ebi.ac.uk:/pub/databases/GO/goa> ls  
drwxrwxr-x  2 ftp      ftp          77 Jul 28  2011 ARABIDOPSIS  
drwxrwxr-x  2 ftp      ftp          93 Jul 28  2011 CHICKEN  
drwxrwxr-x  2 ftp      ftp          69 Jul 28  2011 COW  
drwxrwxr-x  2 ftp      ftp          71 Jul 28  2011 DICTY  
drwxrwxr-x  2 ftp      ftp          69 Jul 28  2011 DOG  
drwxrwxr-x  2 ftp      ftp          69 Jul 28  2011 FLY  
drwxrwxr-x  2 ftp      ftp          71 Jul 28  2011 HUMAN  
drwxrwxr-x  2 ftp      ftp          71 Jul 28  2011 MOUSE  
drwxrwxr-x  2 ftp      ftp          69 Sep 01  2006 PDB  
drwxrwxr-x  2 ftp      ftp          69 Jul 28  2011 PIG  
drwxrwxr-x  2 ftp      ftp          69 Jul 28  2011 RAT  
drwxrwxr-x  2 ftp      ftp        245 Jun 03  2010 UNIPROT  
drwxrwxr-x  2 ftp      ftp          70 Jul 28  2011 WORM  
drwxrwxr-x  2 ftp      ftp          71 Jul 28  2011 YEAST  
drwxrwxr-x  2 ftp      ftp          75 Jul 28  2011 ZEBRAFISH  
drwxrwxr-x  2 ftp      ftp          73 Feb 24  2009 bhf-ucl  
drwxrwxr-x  2 ftp      ftp        311 Jul 11  2012 external2go  
drwxrwxr-x  2 ftp      ftp        202 Mar 28  2011 goslim  
drwxrwxr-x  2 ftp      ftp        362 May 31  2012 gp2protein  
drwxrwxr-x 19 ftp      ftp         400 Jul 07  2011 old  
drwxrwxr-x 20 ftp      ftp      148726 Jan 09 09:29 proteomes  
drwxrwxr-x  2 ftp      ftp          32 Apr 16  2010 tools  
lftp ftp.ebi.ac.uk:/pub/databases/GO/goa>
```

**Gene Association Files;** Tab-delimited files of associations between gene products, GO terms and associated annotation information. The UniProt-  
GOA project uses GAF2.0, which is the standard annotation format used  
within the GO Consortium

# http://amigo.geneontology.org



*the Gene Ontology*

Search Browse **BLAST** Homolog Annotations Tools & Resources Help

Search GO  ☒ terms ☐ genes or proteins ☐ exact match

## BLAST Search

The sequence search is performed using either BLASTP or BLASTX (from the [WU-BLAST](#) package), depending on the type of the input sequence.

### BLAST Query

#### Enter your query

Enter a UniProtKB accession **or** upload a text file of queries **or** paste in FASTA sequence(s)

UniProtKB accession:

Text file (maximum file size 500K):  No file chosen

FASTA sequence(s):

Sequences should be separated with an empty line.

```
>AT5G22740.1|AT5G22740.1|cs1A
MDGVSPKFEVLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLL
MSVMLLCERVYMGIVIVLVKLFWKPKDKRYKFEPHDDDEELGSSNFPVVL
VQIPMFNEREVYKLSIGAACGLSWPSDRLVIQVLDSDTDPTVKQMVEVEC
QRWASKGINIRYQIRENRVGYKAGALKEGLKRSYVKHCEYVVIFDADFQP
EPDFLRRSIPFLMHNPNIALVQARWRFVNSDECLTRMQEMSLDYHFTVE
QEVGSSTHAFFGFNGTAGIWRIAAINEAGGWKDRITVEDMDLAVRASLRG
MKREYLGNTQVKSRTDSTEDAEEDQCHDMSCQDANTFDKXMMFTVDNKKV
```

**Next lecture:** EBI web  
resources II (ENSEMBL  
and InterPro)