

# Bioinformatics Databases

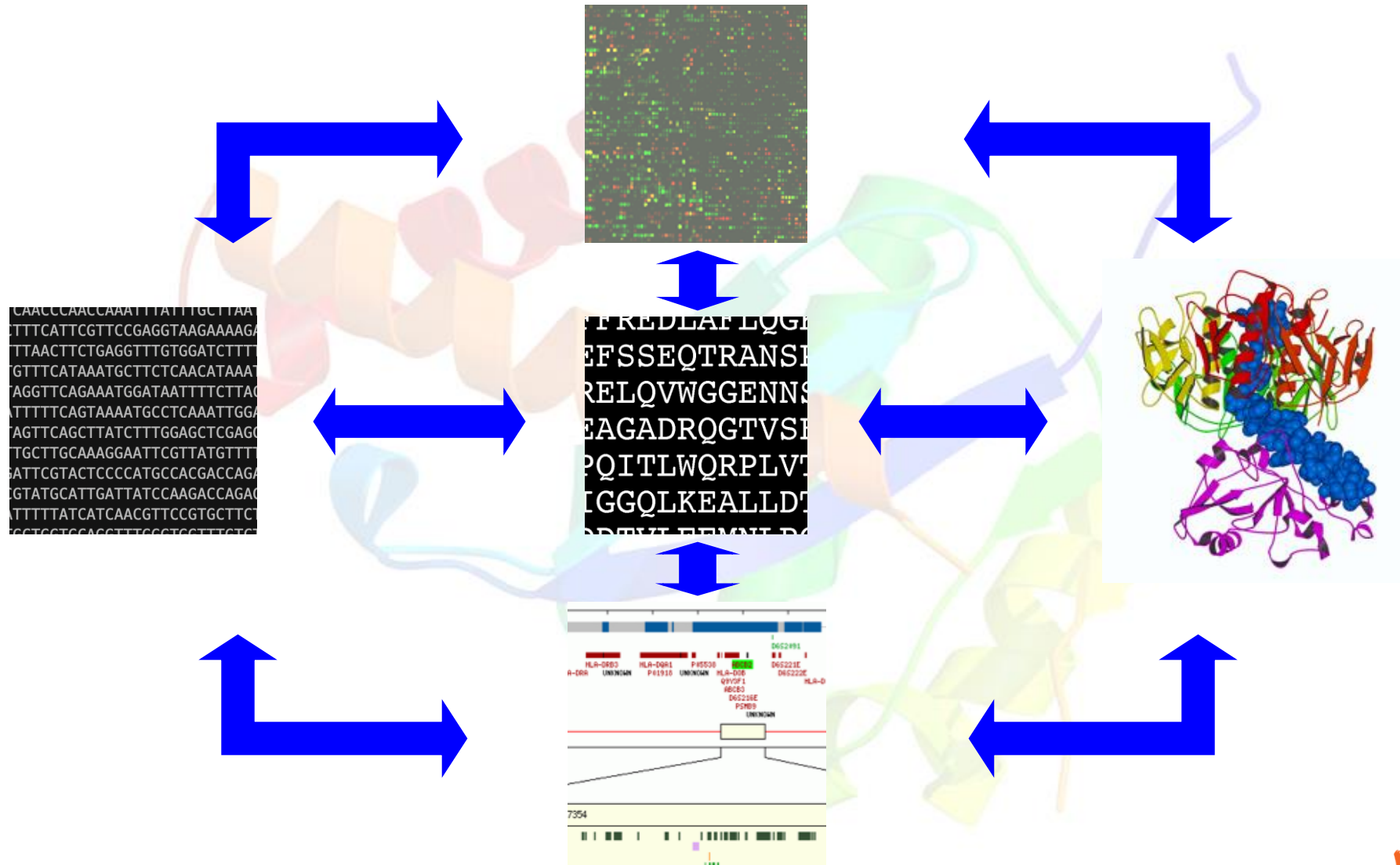
Michael Landi & Andreas Gisel

IITA- Bioinformatics  
Kenya & Nigeria



African Star Apple Assembly workshop 2 – 6 June 2025

# Bioinformatics Databases



# Bioinformatics Databases



UniProtKB

Advanced

BLAST Align Retrieve/ID mapping Help Contact

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.

### UniProtKB

Swiss-Prot (548,454)  
 Manually annotated and reviewed.

TrEMBL (47,452,313)  
 Automatically annotated and not reviewed.

### UniRef

Sequence clusters

### UniParc

Sequence archive

### Proteomes

### Supporting data

Literature citations

Cross-ref. databases

Taxonomy

Diseases  
XXX

Subcellular locations

Keywords

### News

A never-ending race between evolution and genomic integrity | Removal of IPI species proteome data sets from FTP site | UniProtKB XSD ch...  
[UniProt release 2015\\_05](#)

---

Of CAT tails and protein translation by-products | Reducing redundancy in proteomes | Retirement of UniProt Metagenomic and Environmenta...  
[UniProt release 2015\\_04](#)

---

Regulation of translation initiation through folding | New...  
[News archive](#)

### Getting started

**Text search**  
Our basic text search allows you to search all the resources available

**BLAST**  
Find regions of similarity between your sequences

### UniProt data

**Download latest release**  
Get the UniProt data

**Statistics**  
View Swiss-Prot and TrEMBL statistics

### Protein spotlight

#### A Question Of Perspective

April 2015

Paradigms are meant to be broken. In the 1980s, biology students were taught "the one gene = one protein" dogma which has since stepped down from its pedestal, as we now know that one gene,

# Bioinformatics Databases



UniProtKB resistance protein arabidopsis

BLAST Align Retrieve/ID mapping Help Contact

Results Quote terms: "resistance protein"

Filter by

Reviewed (716) Swiss-Prot

Unreviewed (3,701) TrEMBL

Popular organisms

A. thaliana (1,873)

Rice (1,716)

S. cerevisiae (1)

NICPL (1)

MELIC (1)

Other organisms

Subcellular location

Multi-pass membrane protein (144)

Peripheral membrane protein (25)

Protein storage vacuole

Entry	Entry name	Protein names	Gene names	Organism	Length
Q9S7N2	TAA1_ARATH	L-tryptophan--pyruvate aminotransfe...	TAA1, CKRC1, SAV3, TIR2, WEI8, At1g70560, F24J13.13, F5A18.26	Arabidopsis thaliana (Mouse-ear cress)	391
Q3S4A7	AHK5_ARATH	Histidine kinase 5	AHK5, CKI2, At5g10720, MAJ23.80	Arabidopsis thaliana (Mouse-ear cress)	922
Q39176	ERLI1_ARATH	Lipid transfer protein EARLI 1	EARLI1, At4g12480, T1P17.70	Arabidopsis thaliana (Mouse-ear cress)	168
Q9LND1	ERF94_ARATH	Ethylene-responsive transcription f...	ERF094, ORA59, At1g06160, F9P1		
P0DKH5	WR52C_ARATH	Disease resistance protein RRS1	RRS1, RCH2, RRS1-S, RSH4, SLH: At5g45260/At5g45270, K9E15.2/K		
Q9XGM3	RPS4C_ARATH	Disease resistance protein RPS4	RPS4, At5g45250, K9E15.1		
E1B328	WR52N_ARATH	Disease resistance protein RRS1	RRS1, RCH2, RRS1-R, RSH4, SLH:		
C4B7M5	WR52W_ARATH	Disease resistance protein RRS1	RRS1, RCH2, RRS1-R, WRKY52		
O64973	RPS5_ARATH	Disease resistance protein RPS5	RPS5, At1g12220, T28K15.5		
Q42371	ERECT_ARATH	LRR receptor-like serine/threonine-...	ERECTA, ER, QRP1, QRS1, TE1, A		
Q42484	RPS2_ARATH	Disease resistance protein RPS2	RPS2, At4g26090, F20B18.200		

C4B7M5 - WR52W\_ARATH

Protein | Disease resistance protein RRS1

Gene | RRS1

Organism | Arabidopsis thaliana (Mouse-ear cress)

Sequence features | View only features (sites, domains, PTMs ...)

Status | Reviewed - Annotation score: - Experimental evidence at transcript level

Display

Function

Names & Taxonomy

Subcellular location

Pathology & Biotech

PTM / Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Cross-references

Publications

Entry information

Miscellaneous

Function

Transcription factor. Interacts specifically with the W box (5'-(T)TGAC[CT]-3'), a frequently occurring elicitor-responsive cis-acting element. Acts also as a disease resistance protein involved in resistance to fungal and bacterial pathogens, including R.solanacearum, P.syringae pv. tomato and C.higginsianum. In presence of RPS4, elicits an EDS1-dependent hypersensitive response (PubMed:24146667). By similarity Curated 3 Publications

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding	179 - 186	8	ATP PROSITE-ProRule annotation			
DNA binding	1204 - 1272	69	WRKY PROSITE-ProRule annotation			Add BLAST

GO - Molecular function

ADP binding Source: InterPro

sequence-specific DNA binding Source: InterPro

ATP binding Source: UniProtKB-KW

sequence-specific DNA binding transcription factor activity Source: InterPro

GO - Biological process

defense response Source: UniProtKB-KW

transcription, DNA-templated Source: UniProtKB-KW

signal transduction Source: InterPro

Complete GO annotation...

# Bioinformatics Databases



## Keywords<sup>i</sup>

Molecular function	<a href="#">DNA-binding</a>
Biological process	<a href="#">Plant defense</a> , <a href="#">Transcription</a> , <a href="#">Transcription regulation</a>
Ligand	<a href="#">ATP-binding</a> , <a href="#">Nucleotide-binding</a>

## Names & Taxonomy<sup>i</sup>

Protein names <sup>i</sup>	<b>Recommended name:</b> <b>Disease resistance protein RRS1</b> <a href="#">1 Publication</a> ▼ <b>Alternative name(s):</b> <ul style="list-style-type: none"><li>Disease resistance protein RCH2 <a href="#">1 Publication</a> ▼</li><li>Probable WRKY transcription factor 52</li><li>Resistance to Colletotrichum higginsianum 2 protein <a href="#">1 Publication</a> ▼</li><li>Resistance to Ralstonia solanacearum 1 protein <a href="#">1 Publication</a> ▼</li></ul>
Gene names <sup>i</sup>	<b>Name:</b> <b>RRS1</b> <a href="#">1 Publication</a> ▼ <b>Synonyms:</b> RCH2 <a href="#">1 Publication</a> ▼, RRS1-R, WRKY52
Organism <sup>i</sup>	<a href="#">Arabidopsis thaliana (Mouse-ear cress)</a> <a href="#">Imported</a> ▼
Taxonomic identifier <sup>i</sup>	<a href="#">3702 [NCBI]</a>
Taxonomic lineage <sup>i</sup>	<a href="#">Eukaryota</a> > <a href="#">Viridiplantae</a> > <a href="#">Streptophyta</a> > <a href="#">Embryophyta</a> > <a href="#">Tracheophyta</a> > <a href="#">Spermatophytes</a> > <a href="#">Gunneridae</a> > <a href="#">Pentapetalae</a> > <a href="#">rosids</a> > <a href="#">malvids</a> > <a href="#">Brassicales</a> > <a href="#">Brassicaceae</a> > <a href="#">Camelineae</a>

## Subcellular location<sup>i</sup>

- [Nucleus](#) [PROSITE-ProRule annotation](#) ▼

## Pathology & Biotech<sup>i</sup>

### Disruption phenotype<sup>i</sup>

Loss of resistance to C.higginsianum. [1 Publication](#) ▼

## PTM / Processing<sup>i</sup>

### Molecule processing

Feature key	Position(s)	Description	Actions	Graphical view	Length
Chain <sup>i</sup> (PRO_0000431361)	<a href="#">1</a> – <a href="#">1373</a>	Disease resistance protein RRS1	<a href="#">Add</a> <a href="#">BLAST</a>		1373

### Proteomic databases

PaxDb<sup>i</sup> [C4B7M5](#).

## Interaction<sup>i</sup>

### Subunit structure<sup>i</sup>

Interacts with PopP2, a R.solanacearum type III effector. [By similarity](#) ▼

### Protein-protein interaction databases

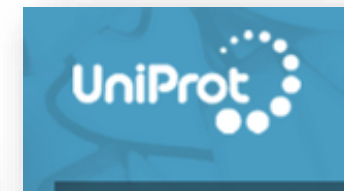
STRING<sup>i</sup> [3702.AT5G45260.1](#).

## Structure<sup>i</sup>

### 3D structure databases

ProteinModelPortal<sup>i</sup> [C4B7M5](#).  
ModBase<sup>i</sup> [Search...](#)  
MobiDB<sup>i</sup> [Search...](#)

# Bioinformatics Databases



Family & Domains<sup>1</sup>

Domains and Repeats				
Feature key	Position(s)	Description	Actions	Graphical view
Domain <sup>1</sup>	8 – 145	TIR <a href="#">Curated</a>	<a href="#">Add</a> <a href="#">BLAST</a>	
Domain <sup>1</sup>	170 – 421	NB-ARC <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	
Repeat <sup>1</sup>	498 – 522	LRR 1 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	
Repeat <sup>1</sup>	535 – 553	LRR 2 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	
Repeat <sup>1</sup>	554 – 575	LRR 3 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	
Repeat <sup>1</sup>	577 – 598	LRR 4 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	
Repeat <sup>1</sup>	621 – 646	LRR 5 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	
Repeat <sup>1</sup>	665 – 688	LRR 6 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	
Repeat <sup>1</sup>	742 – 766	LRR 7 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	
Repeat <sup>1</sup>	768 – 793	LRR 8 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	
Repeat <sup>1</sup>	831 – 854	LRR 9 <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	

Motif				
Feature key	Position(s)	Description	Actions	Graphical view
Motif <sup>1</sup>	988 – 1005	Nuclear localization signal <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>	

Compositional bias				
Feature key	Position(s)	Description	Actions	Graphical view
Compositional bias <sup>1</sup>	463 – 466	Poly-Arg <a href="#">Sequence analysis</a>		

Sequence<sup>1</sup>

Sequence status<sup>1</sup>: Complete.

C4B7M5-1 [UniParc] [FASTA](#) [Add to basket](#)

« Hide

10	20	30	40	50
MTNCEKDEEF	VCISCVVEVR	YSFVSHLSEA	LRRKGINNVV	VGVDSDLLF
60	70	80	90	100
KESQAKIEKA	GVSMVLPGN	CDPSDVWLDK	FAKVLQCRN	NKDQAVVPVL
110	120	130	140	150
YGDILLRDQW	LSELDKGLS	RIHQSRKECS	DSILVEEIVR	DVYETHFYVG
160	170	180	190	200
RIGIYSKLL	IENMVNKQPI	GIRCVGIWGM	PGIGKTLAK	AVFDQMSSAF
210	220	230	240	250
DASCFIEDYD	KSIHEKGLYC	LLEEQLLPGN	DATIMKLSSL	RDRLNSKRVL
260	270	280	290	300
VVLDDVRNAL	VGESFLEGFD	WLGPGSLIII	TSRDKQVFCL	CGINQIYEVQ
310	320	330	340	350
GLNEKEARQL	FLLSASIKED	MGEQNLQELS	VRVINYANGN	PLAINVYGRE
360	370	380	390	400
LKGKKLSEM	ETAFLLKRR	PPFKIVDAFK	STYDLSLSDNE	KNIFLDIACF
410	420	430	440	450
FQGENVNYVI	QLLEGCGFFP	HVEIDLVDK	CLVTISENRV	WLHKLTDIG
460	470	480	490	500
REIINGETVQ	IERRRLWEP	WSIKYLLEYN	EHKANGPKT	TFKRAQGSSE
510	520	530	540	550
IEGLFLDTSN	LRFDLQPSAF	KNMLNRLRL	IYCSNPEVHP	VINFPTGSLH
560	570	580	590	600
SLPNELRLH	WENYPLKSLP	QNFDPRHLVE	INMPYSQLQK	LWGGTKNLEM
610	620	630	640	650

## Sequence databases

### NUCLEOTIDE SEQUENCE

AB470471

[EMBL](#) [GenBank](#) [DDBJ](#)

### PROTEIN SEQUENCE

BAH59424.1

[EMBL](#) [GenBank](#) [DDBJ](#)

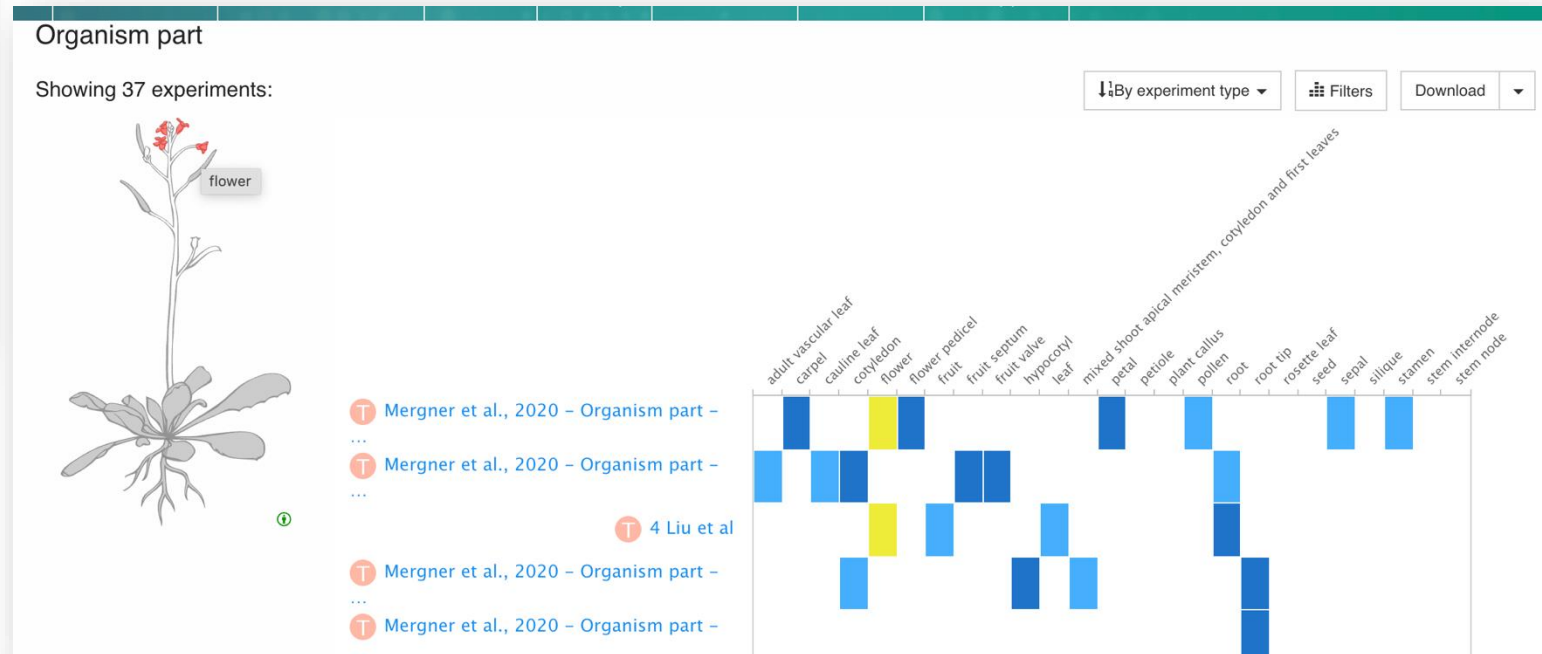
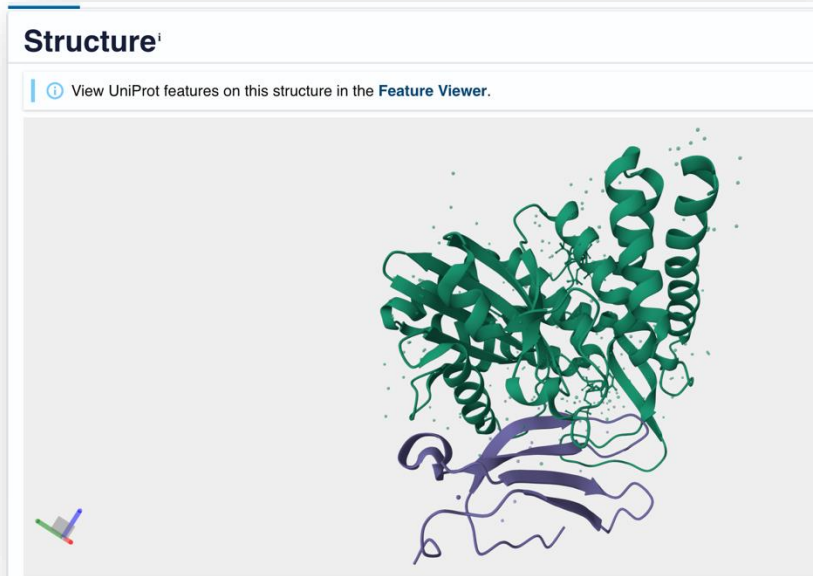
### MOLECULE TYPE

mRNA

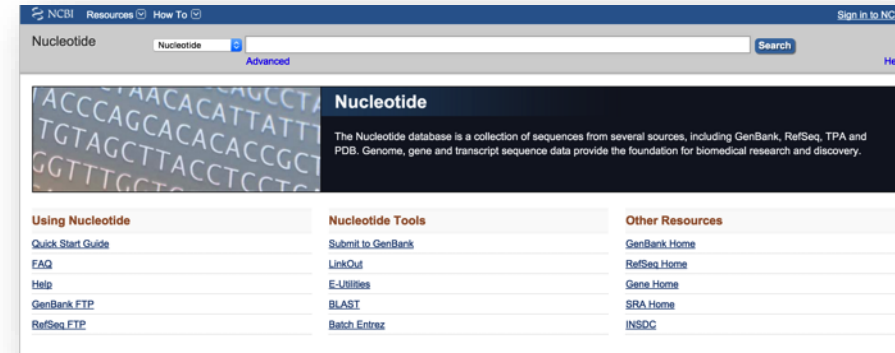
### STATUS



# Bioinformatics Databases



# Bioinformatics Databases





# Bioinformatics Databases

**OrthoDB**

the hierarchical catalogue of orthologs

**OrthoDB**

**OrthoDB** v12.1

[About](#) [Documentation](#) [SparQL](#) [API](#) [Data](#) [Soft](#) [Charts](#) [Upload](#) [Login](#)

Text ▼ e.g. hsp70, sex-lethal, "cytochrome c", kinase -serine ▶ Advanced Submit

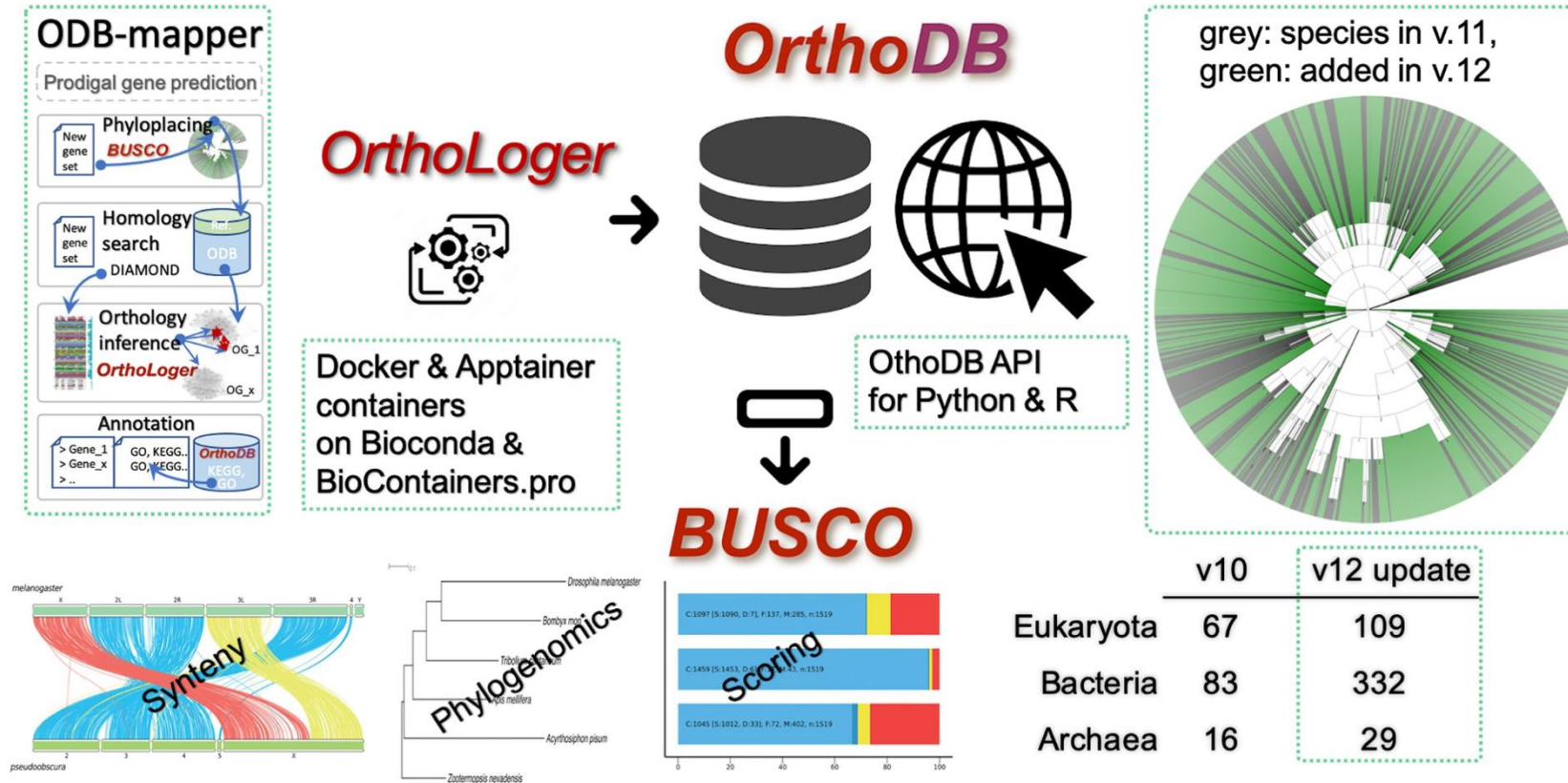
**The hierarchical catalog of orthologs**  
mapping genomics to functional data

Eukaryotes	Prokaryotes	Viruses	Genes
<b>5,827</b>	<b>18,158</b>	<b>7,962</b>	<b>162M</b>

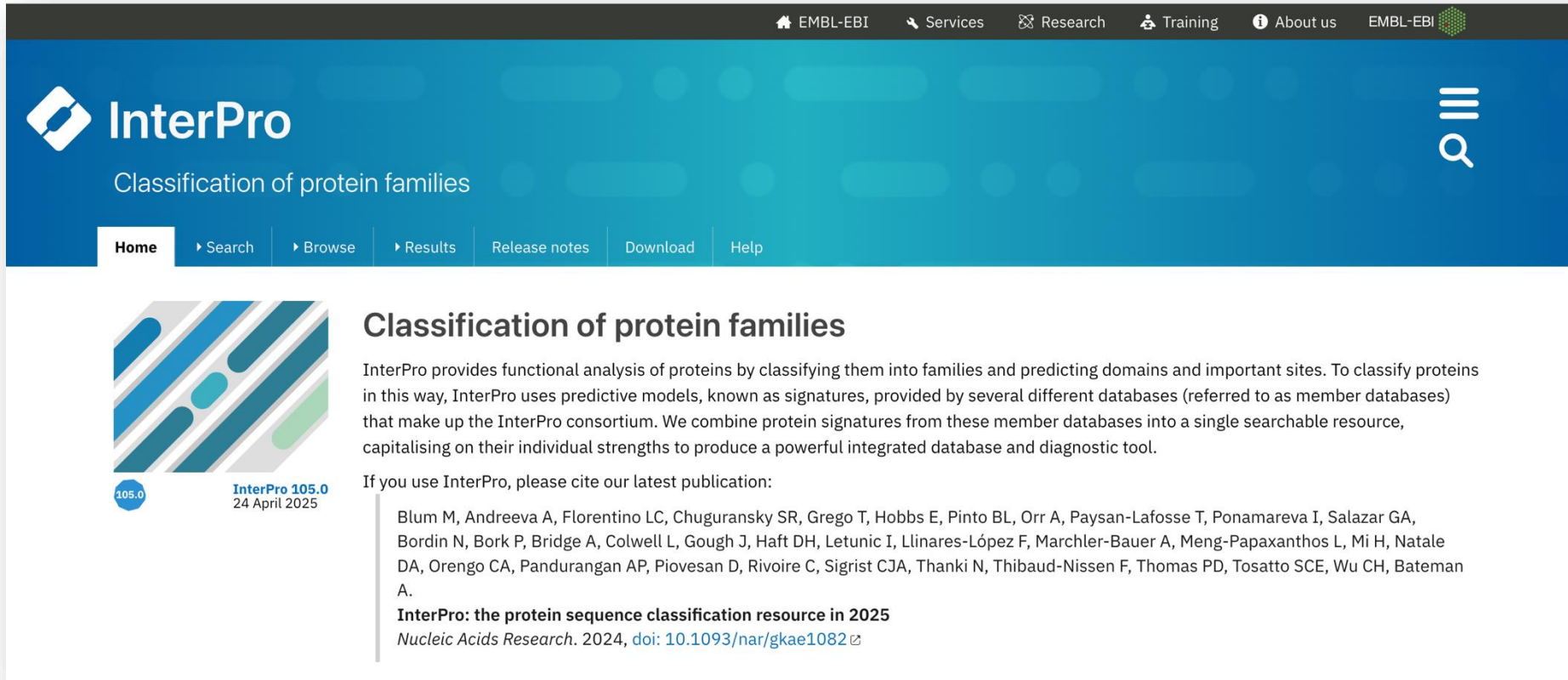
# Bioinformatics Databases

**OrthoDB**

## OrthoDB graphical abstract



# Bioinformatics Databases



The screenshot shows the InterPro website homepage. At the top, there is a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, and EMBL-EBI. Below this is a blue header with the InterPro logo and the text 'Classification of protein families'. A search icon is visible on the right. Below the header is a navigation menu with links to Home, Search, Browse, Results, Release notes, Download, and Help. The main content area features a graphic with diagonal stripes and the text 'InterPro 105.0 24 April 2025'. To the right of the graphic is the title 'Classification of protein families' and a paragraph describing the database. Below this is a section titled 'If you use InterPro, please cite our latest publication:' followed by a list of authors and the publication details.

EMBL-EBI Services Research Training About us EMBL-EBI

**InterPro**  
Classification of protein families

Home Search Browse Results Release notes Download Help

**InterPro 105.0**  
24 April 2025

## Classification of protein families

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.

If you use InterPro, please cite our latest publication:

Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, Pinto BL, Orr A, Paysan-Lafosse T, Ponamareva I, Salazar GA, Bordin N, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunic I, Llinares-López F, Marchler-Bauer A, Meng-Papaxanthos L, Mi H, Natale DA, Orengo CA, Pandurangan AP, Piovesan D, Rivoire C, Sigrist CJA, Thanki N, Thibaud-Nissen F, Thomas PD, Tosatto SCE, Wu CH, Bateman A.

**InterPro: the protein sequence classification resource in 2025**  
*Nucleic Acids Research*. 2024, doi: [10.1093/nar/gkae1082](https://doi.org/10.1093/nar/gkae1082)

# Bioinformatics Databases



 <b>CATH-Gene3D</b> 4.3.0 7k entries	 <b>CDD</b> 3.21 20k entries	 <b>HAMAP</b> 2025_01 2k entries	 <b>NCBIFAM</b> 17.0 32k entries
 <b>PANTHER</b> 19.0 16k entries	 <b>Pfam</b> 37.3 24k entries	 <b>PIRSF</b> 3.10 3k entries	 <b>PRINTS</b> 42.0 2k entries
 <b>PROSITE profiles</b> 2025_01 1k entries	 <b>PROSITE patterns</b> 2025_01 1k entries	 <b>SFLD</b> 4 303 entries	 <b>SMART</b> 9.0 1k entries
 <b>SUPERFAMILY</b> 1.75 2k entries			

## InterProScan

InterProScan is the software package that allows sequences to be scanned against InterPro's member database signatures.

Users who have novel nucleotide or protein sequences that they wish to functionally characterise can use InterProScan to run the scanning algorithms against the InterPro database in an integrated way.

### Download the latest version (5.74-105.0)







Linux (64-bit) - System requirements

- There are no versions planned for Windows or Apple operating systems. This is due to constraints in the various third-party binaries that InterProScan runs.
- Older versions of InterProScan are not supported anymore. We highly recommend you to update to the latest version.

# Bioinformatics Databases

 **EggNOG** 5.0.0

Navigation

-  Home
-  Sequence search
-  **eggNOG-mapper v2**  
(Batch Functional Annotation)
-  Downloads
-  API
-  Methods

## EggNOG v5.0

A database of orthology relationships, functional annotation,  
and gene evolutionary histories.

Organisms	Viruses	Orthologous Groups	Tree & Algs
5,090	2,502	4.4M	4.4M

Citation

**eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.**

*Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, Peer Bork*

Nucleic Acids Res. 2019 Jan 8; 47(Database issue): D309–D314. doi: 10.1093/nar/gky1085

Previous versions

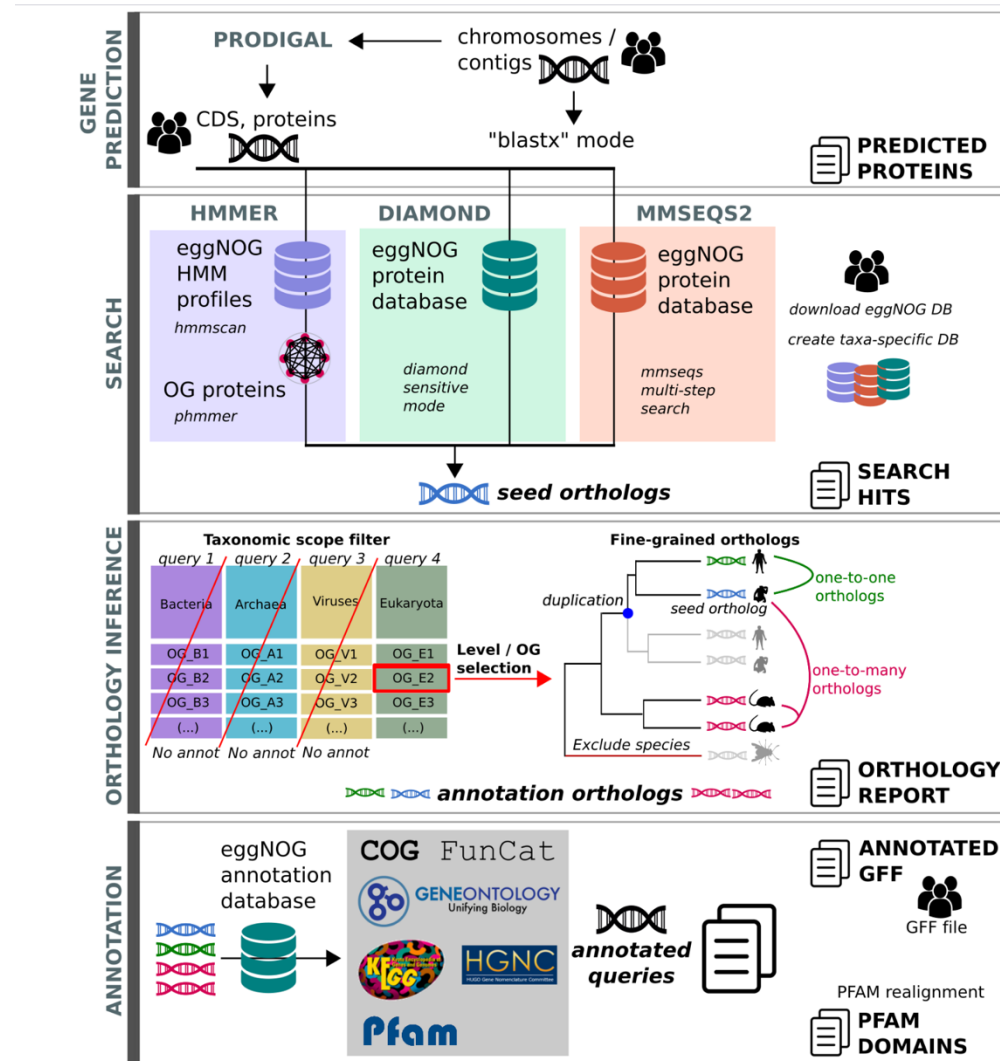
- Eggnog 4.5 (2016)
- Eggnog 4.0 (2014)
- Eggnog 3.0 (2012)



# Bioinformatics Databases



eggNOG-mapper v2 is a tool for functional annotation of large sets of sequences based on fast orthology assignments using precomputed **eggNOG v5.0** clusters and phylogenies.





# Bioinformatics Databases

# Bioinformatics Databases

# Bioinformatics Databases