

# How to use functional annotations at the i5k Workspace@NAL

Monica Poelchau

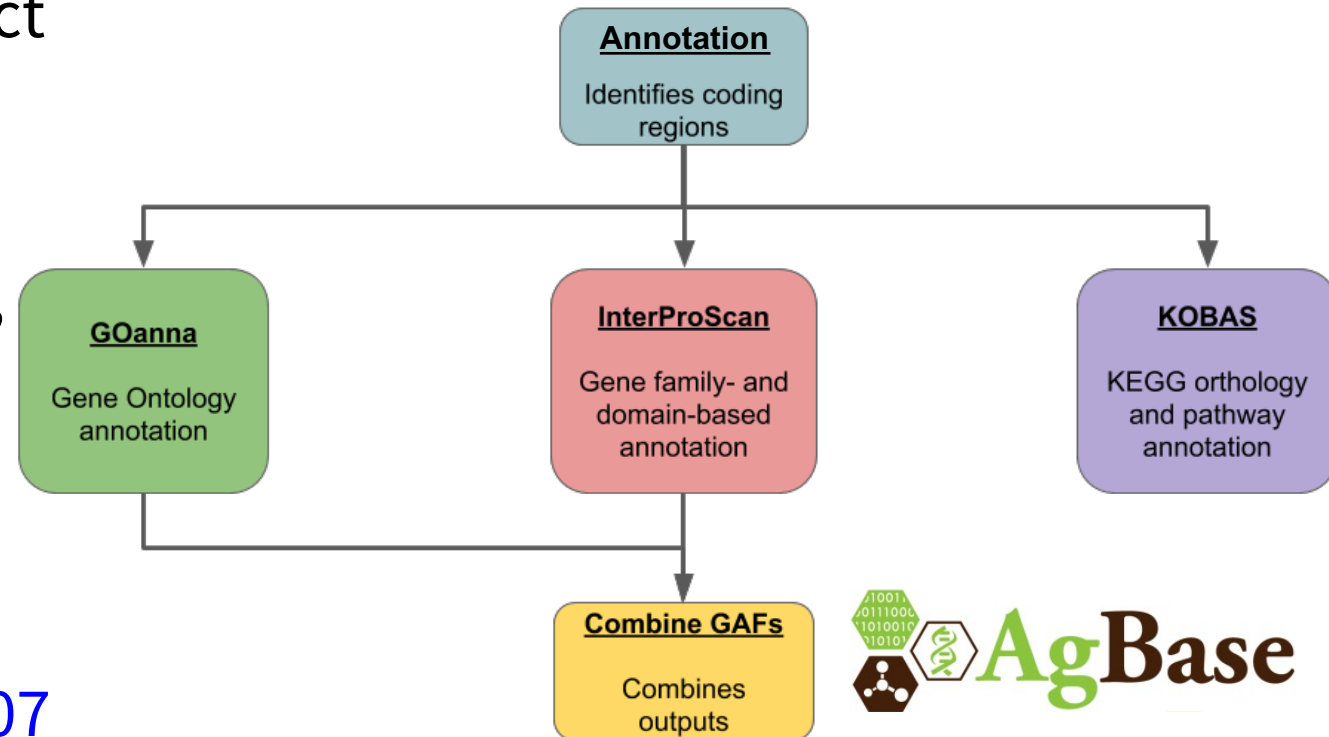
12/13/2022

# Agenda

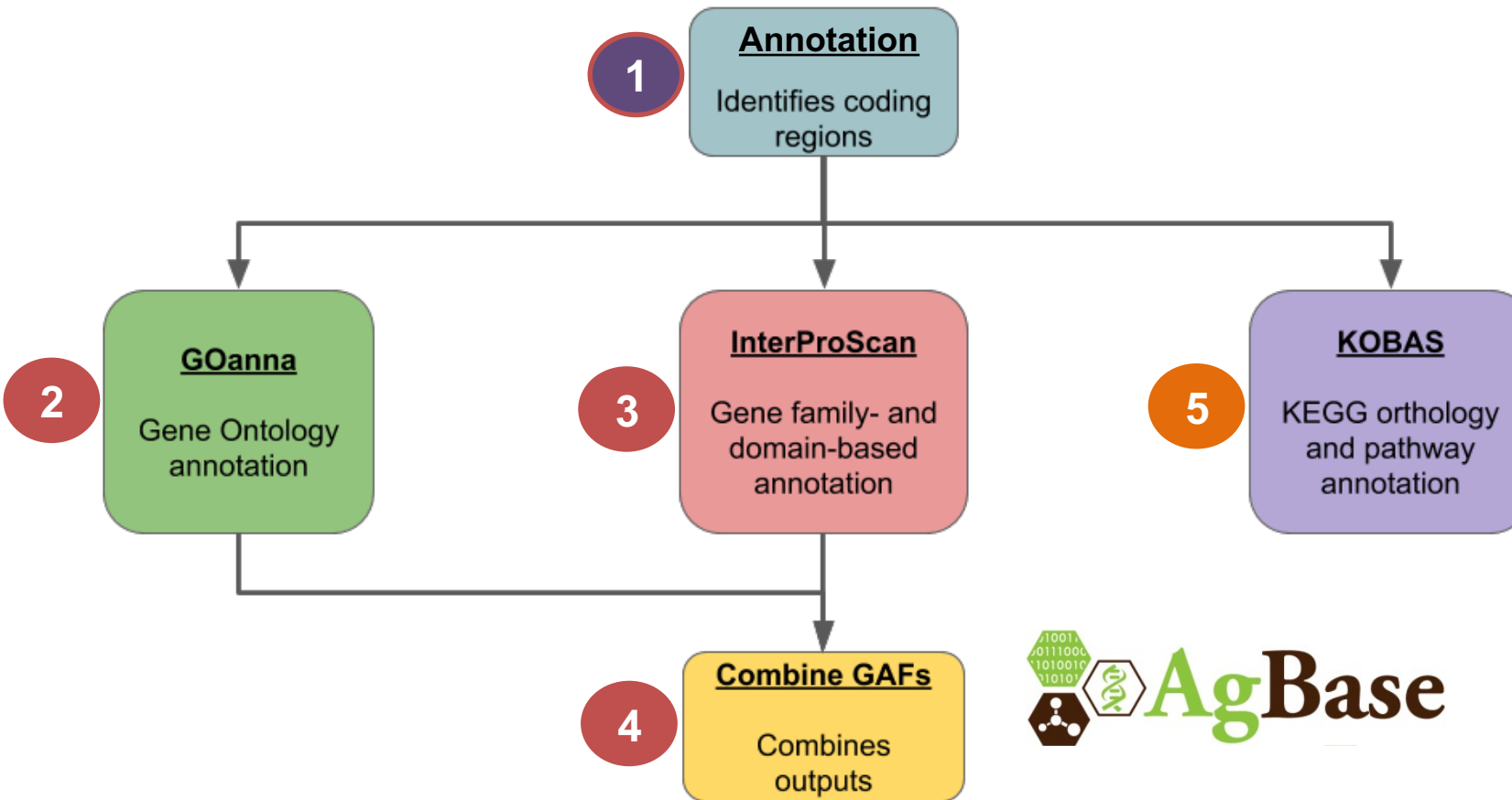
- Background
  - What is functional annotation?
  - What is the AgBase functional annotation pipeline?
  - Why do we use it?
- How to find and use the functional annotations at the i5k Workspace
  - 1. I want to know what functional annotations my protein.
  - 2. I want to find my protein in the functional annotation files.
  - 3. I want to get all the protein accession numbers for a GO category.
  - 4. I want to find all the *Schistocerca americana* proteins annotated to the ubiquitination pathway.

# How do we move from sequence to biology?

- ARS-University of Arizona joint project to develop common workflows and practices for functionally annotating invertebrate genomes.
- Credits: Fiona McCarthy, Surya Saha, Amanda Cooksey, Anna Childers
- Workflows for Rapid Functional Annotation of Diverse Arthropod Genomes. Saha et al., Insects 2021, 12(8), 748;  
<https://doi.org/10.3390/insects12080748>



# Functional annotation tools



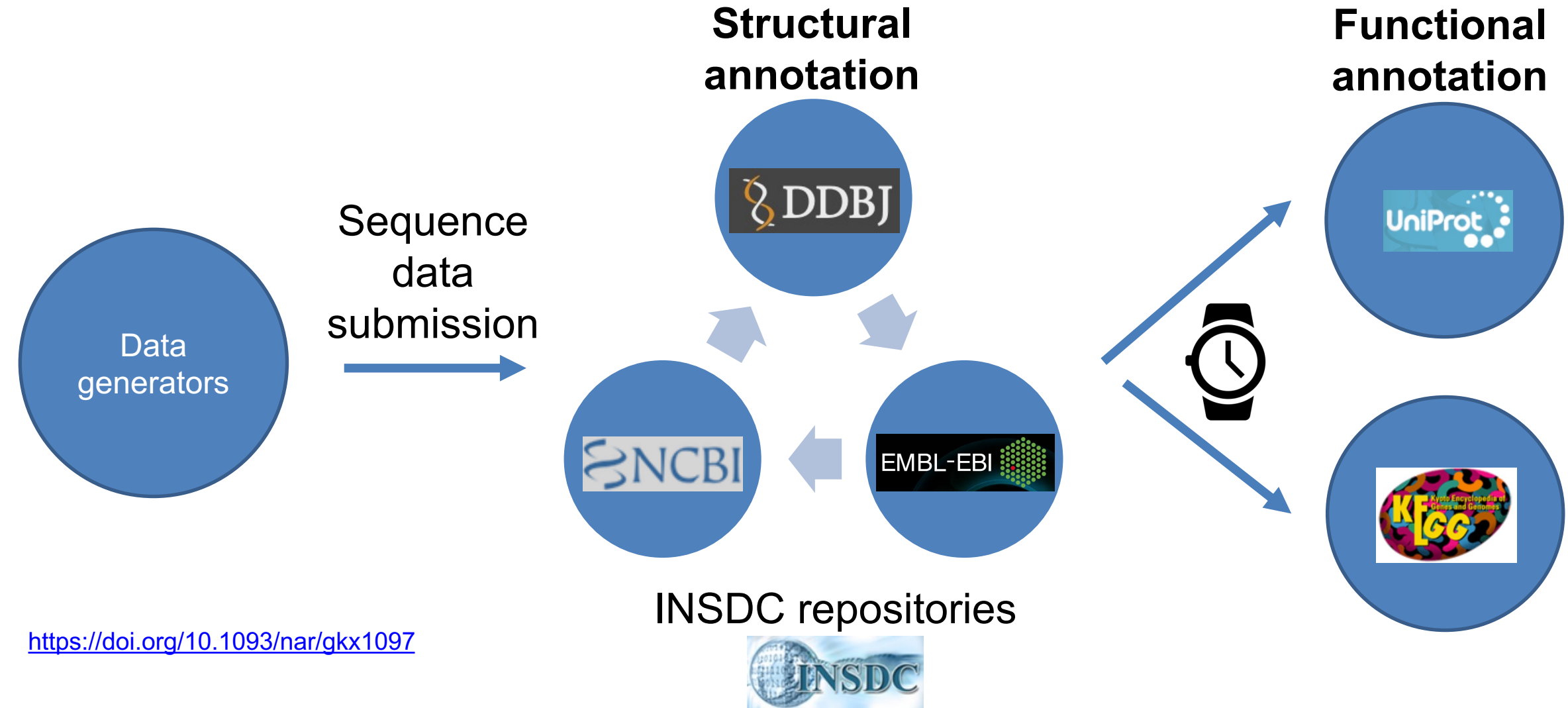
1. Identify proteins
2. Transfer function based upon sequence homology
3. Assign function based upon functional motifs/domains
4. Combine GO, QC, formatting for use
5. Pathway information



# So What Does this Process Get Us?

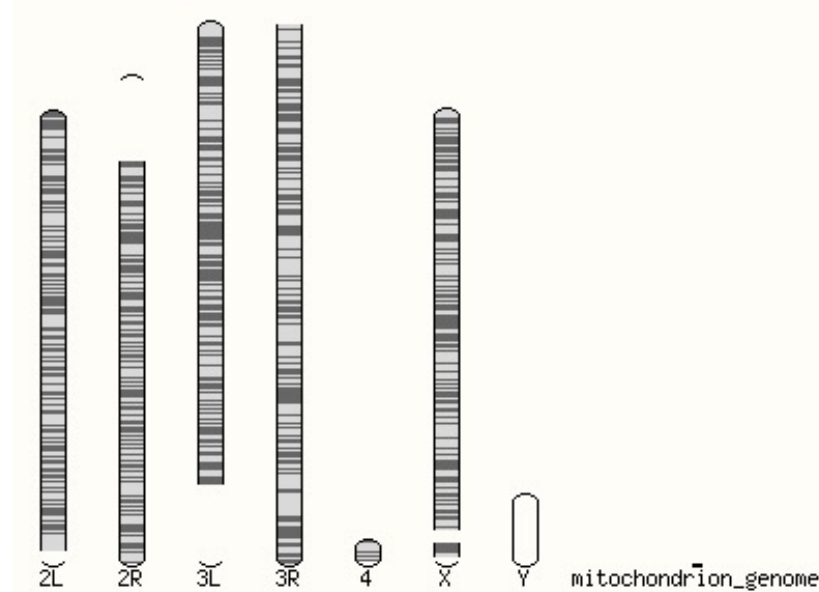
- ***Support for comparative genomics***
  - Motif/domain information for comparative & evolutionary studies -> Evolution of gene families
- ***Support for functional genomics***
  - GO information for GO enrichment
  - Pathway information for pathways enrichment
- ***Targets for genome annotation***

# Typical flow of sequence data

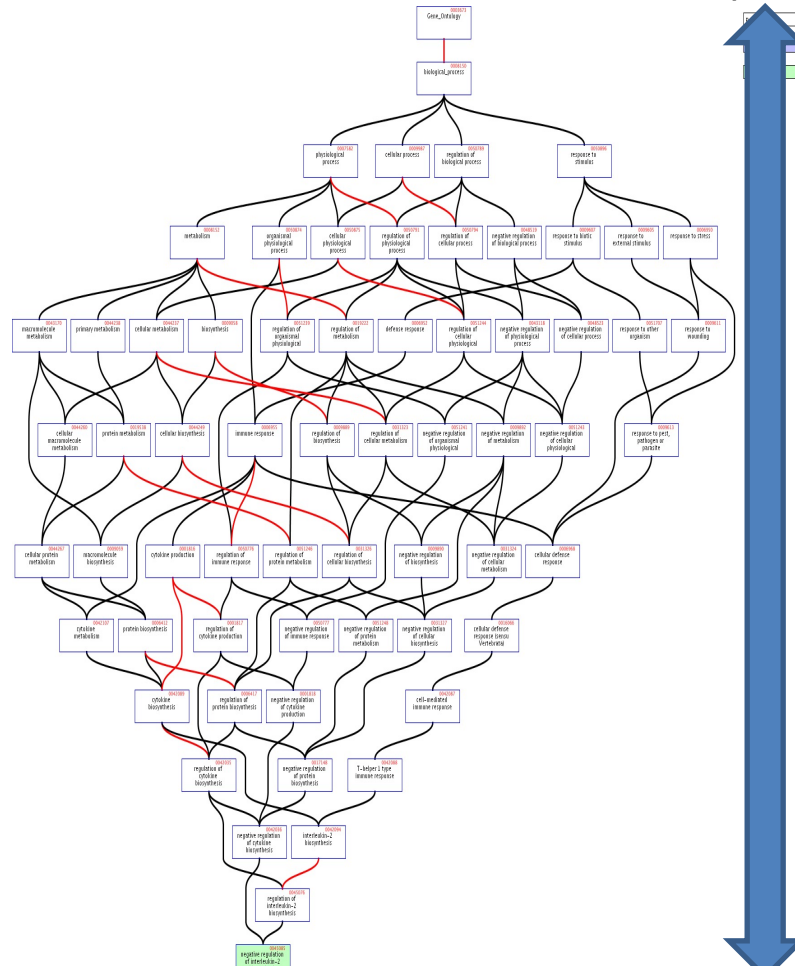


# How do we Measure GO Quality?

**BREADTH:** all gene products should have GO annotation (for CC, MF, BP).



**DEPTH:** function should be as detailed as possible.



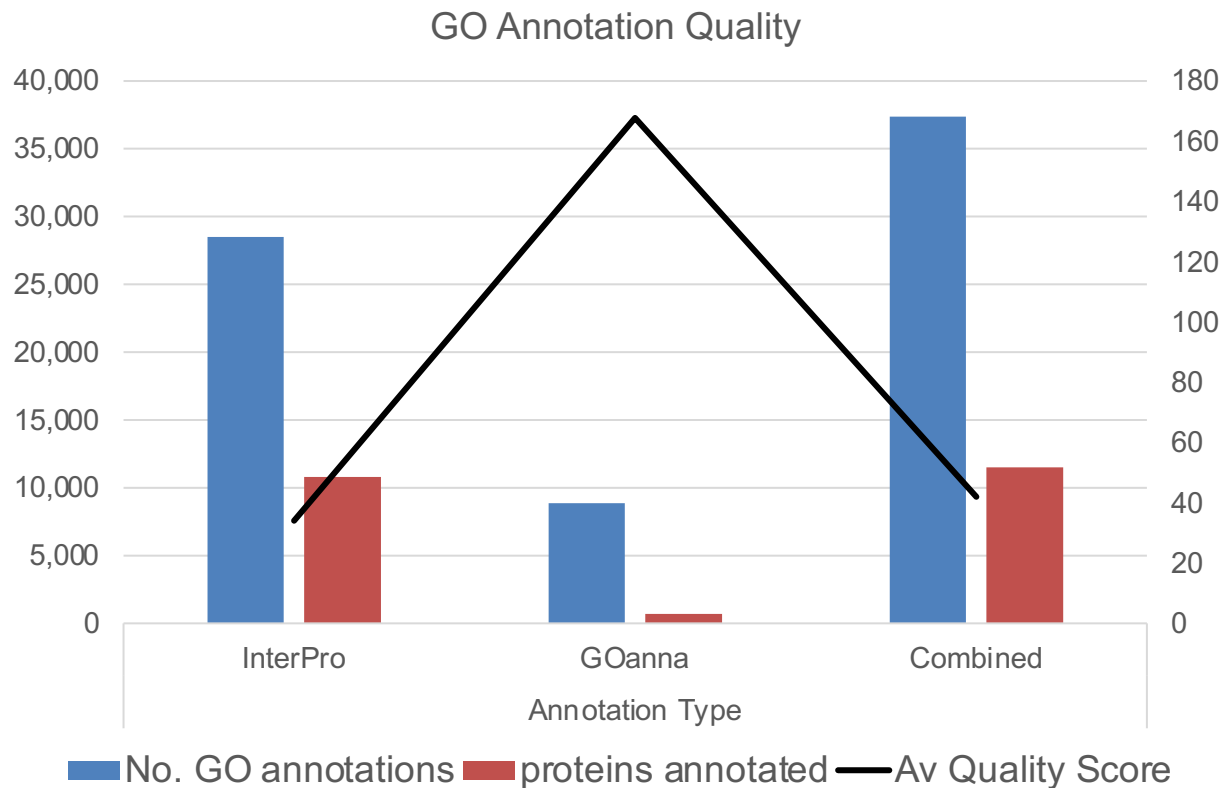
**EVIDENCE:** Published experiments provide direct evidence of function in that species.

Code	Code definition	Evidence code rank
IDA	Inferred from Direct Assay	5
IGI	Inferred from Genetic Interaction	5
IMP	Inferred from Mutant Phenotype	5
IPI	Inferred from Physical Interaction	5
IC	Inferred by Curator	4
TAS	Traceable Author Statement	4
IEP	Inferred from Expression Pattern	3
RCA	Inferred from Reviewed Computational Analysis	3
IGC	Inferred from Genomic Context	3
ISS	Inferred from Sequence or Structural Similarity	2
IEA	Inferred from Electronic Annotation	2
NAS	Non-traceable Author Statement	2
NR	Not Recorded	1
ND	No Biological data available	0

**GOanna**  
Gene Ontology annotation

# Adding Details to InterProScan GO: GOanna

Interpro & GOanna are **complementary** approaches.  
InterProScan provides "breadth" (some GO annotation for most proteins)  
GOanna provides "depth" (more detailed GO terms)




GO Quality Analysis	Annotation Type		
	InterPro	GOanna	Combined
No. GO annotations	28,494	8,866	37,360
proteins annotated	10,810	691	11,500
Av Quality Score	34.064	167.751	42.085

**GOanna**

Gene Ontology  
annotation

# Accessing functional annotation tools


AgBase  
latest

AGBASE HOME  
Functional Annotation Workflow

GOANNA  
Intro  
GOanna on CyVerse  
GOanna on the Command Line  
GOanna on the ARS Ceres HPC

INTERPROSCAN  
Intro  
InterProScan on CyVerse  
InterProScan on the Command Line  
InterProScan on the ARS Ceres HPC

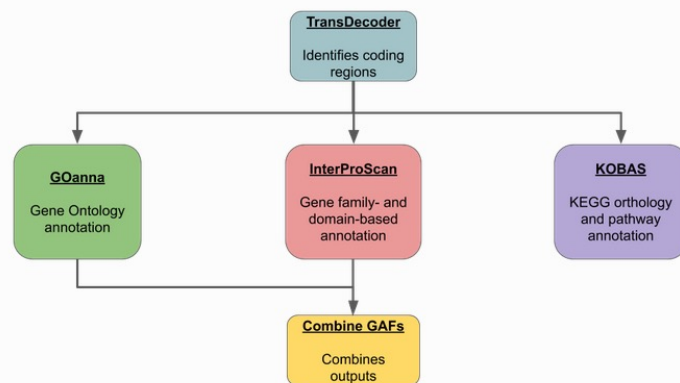
COMBINE GAFs  
Intro  
Combine GAFs on CyVerse  
Combine GAFs on the Command Line  
Combine GAFs on the ARS Ceres HPC

KOBAS  
Intro  
KOBAS on CyVerse  
KOBAS on the Command Line  
KOBAS on the ARS Ceres HPC

Docs » Functional Annotation Workflow

[Edit on GitHub](#)

## Functional Annotation Workflow



This functional annotation workflow employs three annotation tools:

1. **GOanna**: performs a BLAST search and transfers gene ontology (GO) annotations from BLAST matches to the query gene products.
2. **InterProScan**: InterPro is a database which integrates together predictive information about proteins' function from a number of partner resources, giving an overview of the families that a protein belongs to and the domains and sites it contains. InterProScan can also provide GO and pathway annotations.
3. **KOBAS**: uses BLAST to annotate the input with KEGG Orthology terms and KEGG pathways

**Note**



[agbase-docs.readthedocs.io](http://agbase-docs.readthedocs.io)



[de.cyverse.org](http://de.cyverse.org)

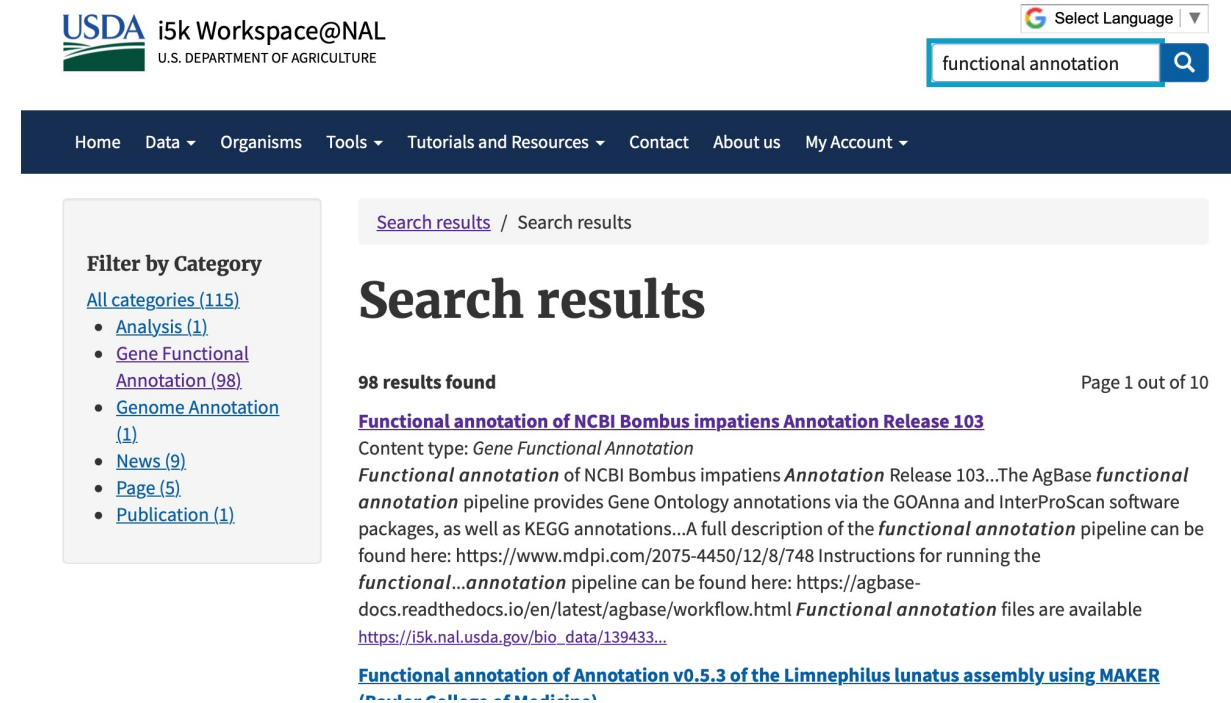


[hub.docker.com/u/agbase](https://hub.docker.com/u/agbase)

# How to find and use the functional annotations at the i5k Workspace

# 15k Workspace functional annotations

- Functional annotations are available for 98 datasets;
- All new i5k Workspace organisms and assemblies will be functionally annotated.



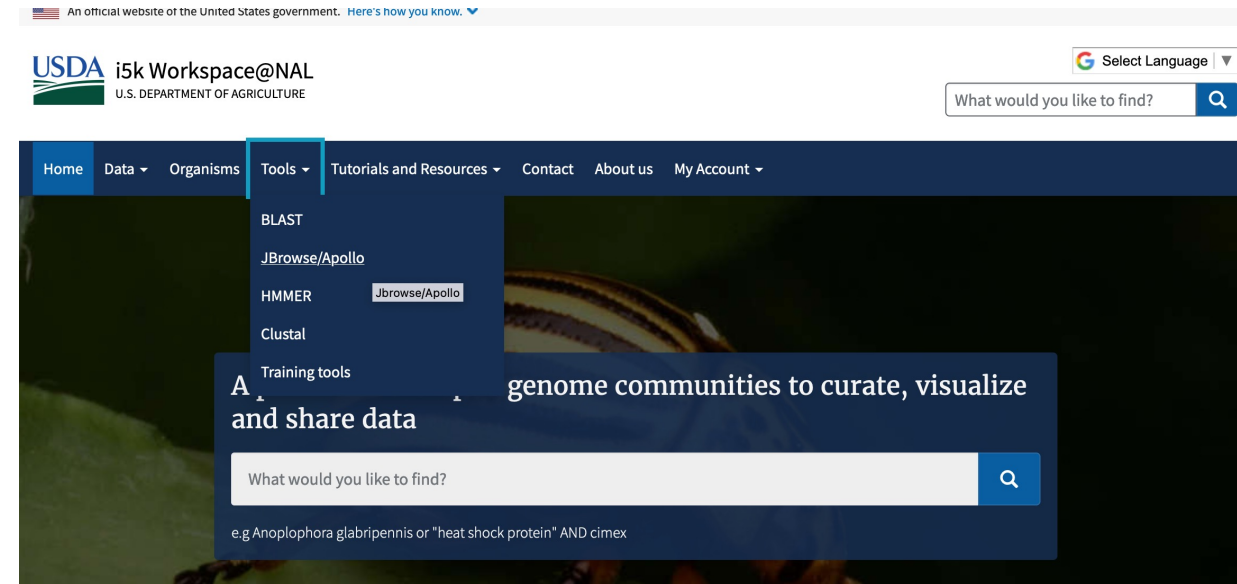
The screenshot displays the i5k Workspace@NAL website interface. At the top, the USDA logo and 'i5k Workspace@NAL U.S. DEPARTMENT OF AGRICULTURE' are visible. A search bar on the right contains the text 'functional annotation'. Below the search bar is a navigation menu with links: Home, Data, Organisms, Tools, Tutorials and Resources, Contact, About us, and My Account. The main content area shows 'Search results' for the query. On the left, a 'Filter by Category' sidebar lists: All categories (115), Analysis (1), Gene Functional Annotation (98), Genome Annotation (1), News (9), Page (5), and Publication (1). The search results section on the right shows '98 results found' and 'Page 1 out of 10'. The first result is titled 'Functional annotation of NCBI Bombus impatiens Annotation Release 103'. The content type is 'Gene Functional Annotation'. The description states: 'Functional annotation of NCBI Bombus impatiens Annotation Release 103...The AgBase functional annotation pipeline provides Gene Ontology annotations via the GOAnna and InterProScan software packages, as well as KEGG annotations...A full description of the functional annotation pipeline can be found here: https://www.mdpi.com/2075-4450/12/8/748 Instructions for running the functional...annotation pipeline can be found here: https://agbase-docs.readthedocs.io/en/latest/agbase/workflow.html Functional annotation files are available https://i5k.nal.usda.gov/bio\_data/139433...'. A second link is provided: 'Functional annotation of Annotation v0.5.3 of the Limnephilus lunatus assembly using MAKER (Routledge College of Medicine)'.

# 15k Workspace functional annotation long-term storage and citation

- The functional annotation datasets are intended to be ephemeral, as the underlying GO databases update regularly. Therefore, we do not provide long-term storage of these datasets.
- If you plan on using a particular functional annotation dataset in a publication, we can archive this dataset for you at the Ag Data Commons.
- If you use the functional annotations in a publication, please cite the following:
  - The i5k Workspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes. Poelchau et al., Nucleic Acids Research, Volume 43, Issue D1, 28 January 2015, Pages D714–D719, <https://doi.org/10.1093/nar/gku983>
  - Workflows for Rapid Functional Annotation of Diverse Arthropod Genomes. Saha et al., Insects 2021, 12(8), 748; <https://doi.org/10.3390/insects12080748>

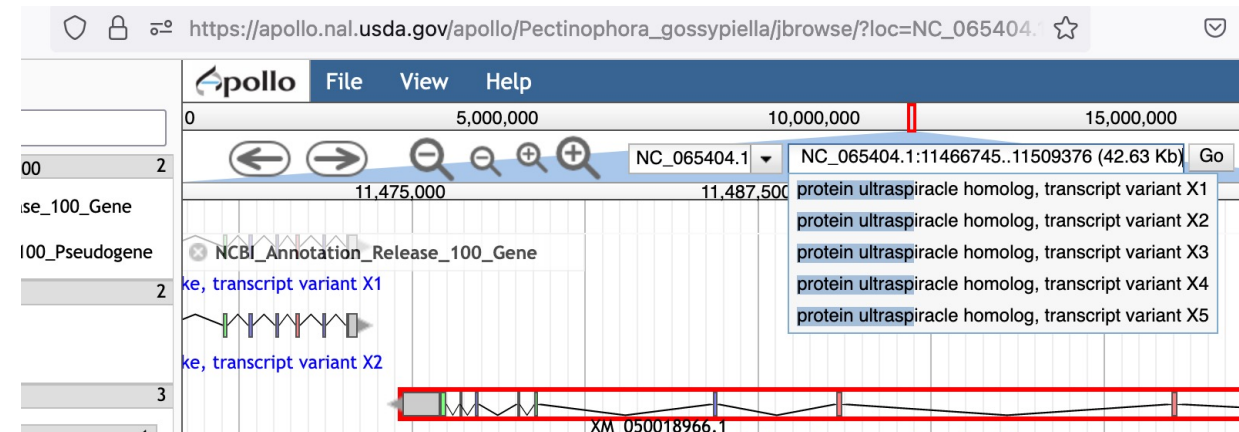
# 1. I want to know what functional annotations my protein has

- Example protein: ultraspiracle in the pink bollworm, *Pectinophora gossypiella*
- Go to <https://i5k.nal.usda.gov/> -> Tools -> Jbrowse/Apollo -> Find a genome browser: Available genome browsers
- Find organism (*Pectinophora gossypiella*) in list

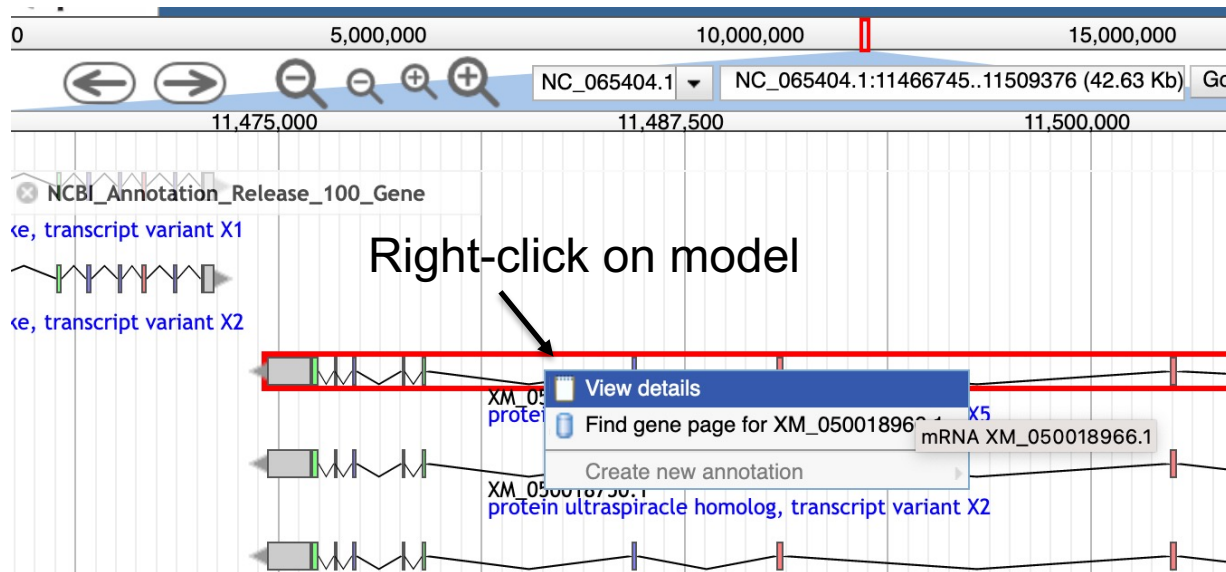


# 1. I want to know what functional annotations my protein has

- search for ‘ultraspiracle’ in the Jbrowse search bar (tip – if you can’t find it, add ‘protein’ before the name)
- Protein or mRNA accession numbers are also typically searchable



# 1. I want to know what functional annotations my protein has



mRNA XM\_050018966.1

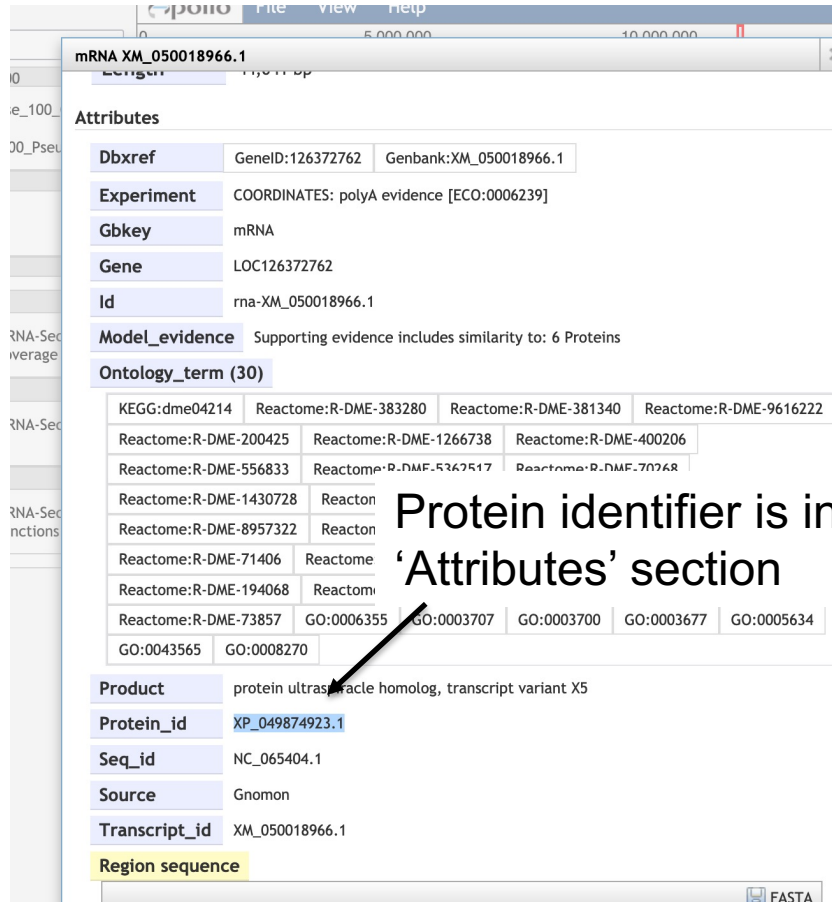
Primary Data	
Name	XM_050018966.1
Type	mRNA
Description	protein ultraspiracle homolog, transcript variant X5
Position	NC_065404.1:11474671..11518711 (- strand)
Length	44,041 bp

Attributes	
Dbxref	GeneID:126372762 Genbank:XM_050018966.1
Experiment	COORDINATES: polyA evidence [ECO:0006239]
Gbkey	mRNA
Gene	LOC126372762
Id	rna-XM_050018966.1
Model_evidence	Supporting evidence includes similarity to: 6 Proteins
Ontology_term (30)	<div>           KEGG:dme04214           Reactome:R-DME-383280           Reactome:R-DME-381340           Reactome:R-DME-9616222         </div> <div>           Reactome:R-DME-200425           Reactome:R-DME-1266738           Reactome:R-DME-400206         </div> <div>           Reactome:R-DME-556833           Reactome:R-DME-5362517           Reactome:R-DME-70268         </div> <div>           Reactome:R-DME-1430728           Reactome:R-DME-74160           Reactome:R-DME-1428517         </div> <div>           Reactome:R-DME-8957322           Reactome:R-DME-9006931           Reactome:R-DME-162582         </div> <div>           Reactome:R-DME-71406           Reactome:R-DME-159418           Reactome:R-DME-212436         </div> <div>           Reactome:R-DME-194068           Reactome:R-DME-204174           Reactome:R-DME-8978868         </div> <div>           Reactome:R-DME-73857           GO:0006355           GO:0003707           GO:0003700           GO:0003677           GO:0005634         </div> <div>           GO:0043565           GO:0008270         </div>
Product	protein ultraspiracle homolog, transcript variant X5

Annotations are in Ontology\_term section

## 2. I want to find my protein in the functional annotation files.



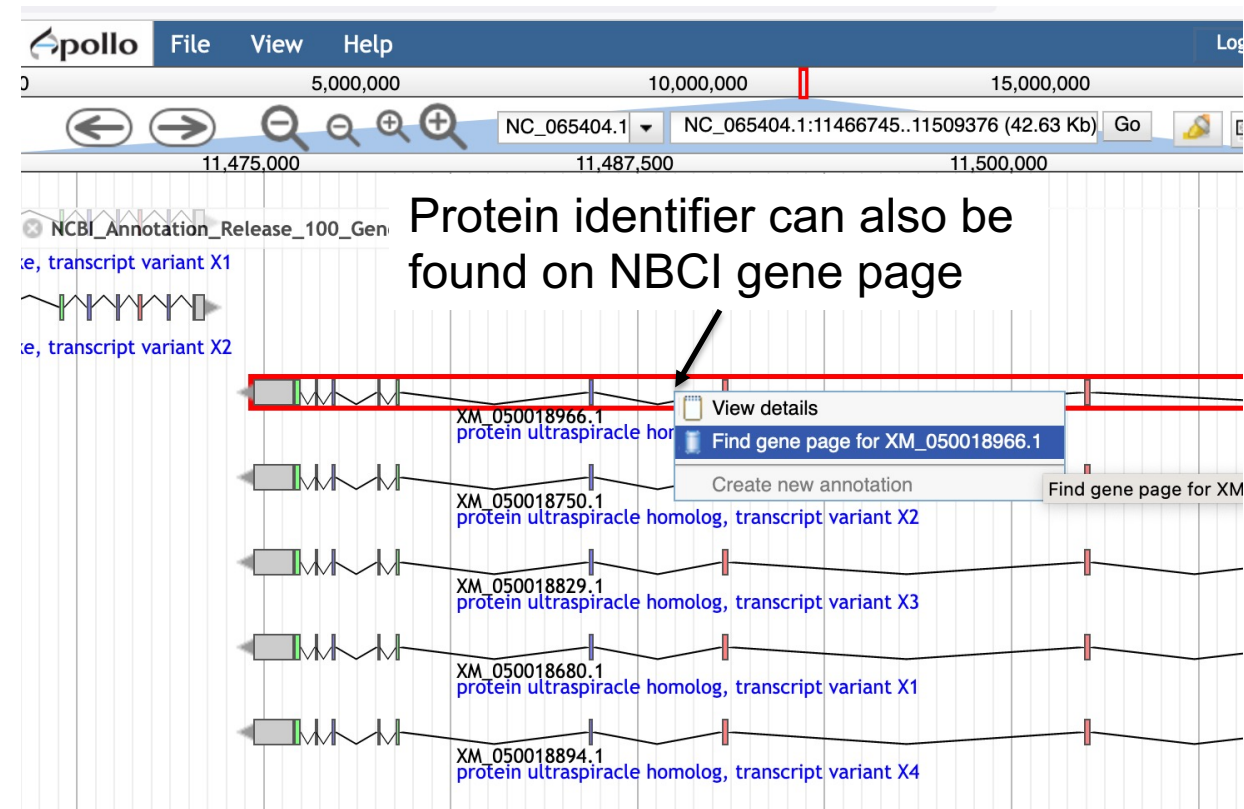
**mRNA XM\_050018966.1**

**Attributes**

<b>Dbxref</b>	GenelD:126372762	Genbank:XM_050018966.1
<b>Experiment</b>	COORDINATES: polyA evidence [ECO:0006239]	
<b>Gbkey</b>	mRNA	
<b>Gene</b>	LOC126372762	
<b>Id</b>	rna-XM_050018966.1	
<b>Model_evidence</b>	Supporting evidence includes similarity to: 6 Proteins	
<b>Ontology_term (30)</b>	KEGG:dme04214   Reactome:R-DME-383280   Reactome:R-DME-381340   Reactome:R-DME-9616222 Reactome:R-DME-200425   Reactome:R-DME-1266738   Reactome:R-DME-400206 Reactome:R-DME-556833   Reactome:R-DME-5367517   Reactome:R-DME-70768 Reactome:R-DME-1430728   Reactome: Reactome:R-DME-8957322   Reactome: Reactome:R-DME-71406   Reactome: Reactome:R-DME-194068   Reactome: Reactome:R-DME-73857   GO:0006355   GO:0003707   GO:0003700   GO:0003677   GO:0005634 GO:0043565   GO:0008270	
<b>Product</b>	protein ultraspiracle homolog, transcript variant X5	
<b>Protein_id</b>	XP_049874923.1	
<b>Seq_id</b>	NC_065404.1	
<b>Source</b>	Gnomon	
<b>Transcript_id</b>	XM_050018966.1	
<b>Region sequence</b>		

FASTA

Protein identifier is in 'Attributes' section



**Protein identifier can also be found on NBCI gene page**

NCBI\_Annotation\_Release\_100\_Gen

NC\_065404.1   NC\_065404.1:11466745..11509376 (42.63 Kb)   Go

11,475,000   11,487,500   11,500,000

ie, transcript variant X1

ie, transcript variant X2

XM\_050018966.1  
protein ultraspiracle homolog, transcript variant X5

XM\_050018750.1  
protein ultraspiracle homolog, transcript variant X2


XM\_050018829.1  
protein ultraspiracle homolog, transcript variant X3

XM\_050018680.1  
protein ultraspiracle homolog, transcript variant X1


XM\_050018894.1  
protein ultraspiracle homolog, transcript variant X4

View details  
Find gene page for XM\_050018966.1  
Create new annotation  
Find gene page for XM

## 2. I want to find my protein in the functional annotation files.



→ ↻ 🔒 https://i5k.nal.usda.gov/tripal\_elasticsearch/search\_website/?search\_box=Pectinophora+gossypiella+functional+anno 📄 ☆ 📧 🌐

 i5k Workspace@NAL  
U.S. DEPARTMENT OF AGRICULTURE

Search for “Pectinophora gossypiella functional annotations”

Select Language ▼

Pectinophora gossypiella functional annotations 🔍

Home Data ▼ Organisms Tools ▼ Tutorials and Resources ▼ Contact About us My Account ▼

### Search results

2 results found

Page 1 out of 1

#### Functional annotation of NCBI *Pectinophora gossypiella* Annotation Release 100

Content type: *Gene Functional Annotation*

*Functional* annotation of NCBI *Pectinophora gossypiella* Annotation Release 100...The AgBase *functional* annotation pipeline provides Gene Ontology *annotations* via the GOAnna and InterProScan software packages, as well as KEGG *annotations*...P.  
*gossypiella*...*Pectinophora*...*gossypiella*  
[https://i5k.nal.usda.gov/bio\\_data/139441...](https://i5k.nal.usda.gov/bio_data/139441...)

#### New Ag100Pest assemblies, and full resources for four Ag100Pest/BPRI locust assemblies

Content type: *News*

We are excited to announce three new Ag100Pest species at the i5k Workspace@NAL : the boll weevil, *Anthonomus grandis*...  
...*Pectinophora gossypiella*...*Pectinophora gossypiella*...*Pectinophora gossypiella*...





#### Filter by Category

[All categories \(2\)](#)

- [Gene Functional Annotation \(1\)](#)
- [News \(1\)](#)

Search result

## 2. I want to find my protein in the functional annotation files.

← → ↻  [https://i5k.nal.usda.gov/bio\\_data/1394410](https://i5k.nal.usda.gov/bio_data/1394410)   

[/workflow.html](#)

Functional annotation files are available under <https://i5k.nal.usda.gov/content/data-downloads>

<b>Program, Pipeline, Workflow or Method Name</b>	AgBase functional annotation pipeline
<b>Program Version</b>	NA
<b>Organism</b>	<a href="#">P. gossypiella (pink bollworm)</a>
<b>Data Source</b>	<div><b>Source Name</b> : GCF_024362695.1_ilPecGoss1.1_protein.faa</div> <div><b>Source URI</b> : <a href="https://i5k.nal.usda.gov/data/Arthropoda/pecgos-%28Pectinophora_gossypiella%29/ilPecGoss1.1/2.Official%20or%20Primary%20Gene%20Set/NCBI%20Pectinophora%20gossypiella%20Annotation%20Release%20100%20functional%20annotation/">https://i5k.nal.usda.gov/data/Arthropoda/pecgos-%28Pectinophora_gossypiella%29/ilPecGoss1.1/2.Official%20or%20Primary%20Gene%20Set/NCBI%20Pectinophora%20gossypiella%20Annotation%20Release%20100%20functional%20annotation/</a></div>
<b>Publication</b>	There are no publications associated with this record.
<b>Algorithm</b>	

Link to data  
downloads



## 2. I want to find my protein in the functional annotation files.

- We'll download the 'complete GAF' file, and search for protein accession XP\_049874923.1
- GAF file format specification:  
<http://geneontology.org/docs/go-annotation-file-gaf-format-2.2/>

### Index of /data/Arthropoda/pecgos-(Pectinophora\_gossy Primary Gene Set/NCBI Pectinophora gossypiella Ann annotation/

../	17-Oct-2022 22:24	-
GOanna/	17-Oct-2022 22:24	-
Interproscan/	17-Oct-2022 22:24	-
KOBAS/	17-Oct-2022 22:24	-
GCF_024362695.1_complete.gaf.tsv	17-Oct-2022 22:24	6667116
README.txt	19-Oct-2022 19:49	2039
README.txt~	17-Oct-2022 22:24	2032

# The GAF format.

Protein source database (DB)

DB protein ID

DB Protein Symbol

Qualifier

GO ID (the functional annotation)

DB Reference (annotation method)

Evidence code for annotation

With/From (what evidence was used for the annotation)

Aspect (e.g. CC, MF, BP)

DB Object Name

DB Object Synonym

DB Object Type

Taxon (taxonomic identifier)

Assigned By

Date

!gaf-version: 2.0														
RefSeq	XP_049864889.1	XP_049864889.1		GO:0000266	GO_REF:0000024	ECO:0000247	UniprotKB:Q9VQE0	P	XP_049864889.1		protein	taxon:13191	20220909	AgBase
RefSeq	XP_049864889.1	XP_049864889.1		GO:0000281	GO_REF:0000024	ECO:0000247	UniprotKB:Q8IHG0	P	XP_049864889.1		protein	taxon:13191	20220909	AgBase
Refseq	XP_049876919.1	XP_049876919.1		GO:0016485	GO_REF:0000002	ECO:0000501	InterPro:IPR039245	P	XP_049876919.1		protein	taxon:13191	20220910	AgBase
Refseq	XP_049876920.1	XP_049876920.1		GO:0004252	GO_REF:0000002	ECO:0000501	InterPro:IPR039245	F	XP_049876920.1		protein	taxon:13191	20220910	AgBase

GO\_REF:0000024 – GOAnna

GO\_REF:0000002 - InterProScan

ECO:0000247 - GOAnna

ECO:0000501 - InterProScan

GOAnna: accession # of the exp. characterized sequences(s) that match the query

InterProScan: individual seqs, seq objects, methods, keyword mapping files, etc. that underlie the annotation.

## 2. I want to find my protein in the functional annotation files.

AutoSave OFF GCF\_024362695.1\_complete.gaf

Home Insert Draw Page Layout Formulas Data Review View Acrobat Tell me

Paste Calibri (Body) 12 GO annotations

General

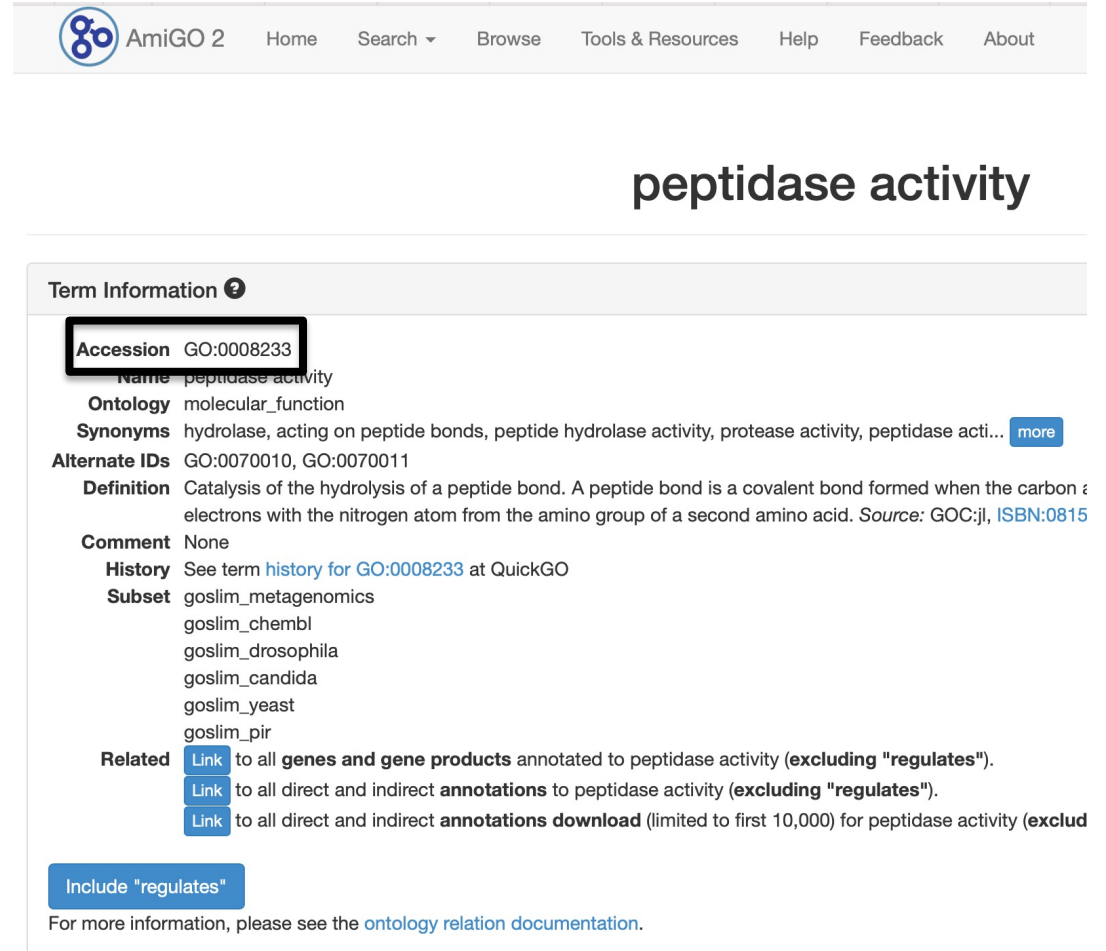
Conditional Formatting Format as Table Cell Styles

A20004 X ✓ fx Refseq

	A	B	C	D	E	F	G	H	I	J	K
0000	Refseq	XP_049874922.1	XP_049874922.1		GO:0005839	GO_REF:0000002	ECO:0000501	InterPro:IPR001353 InterPro:IPR037555	C	XP_049874922.1	
0001	Refseq	XP_049874922.1	XP_049874922.1		GO:0006511	GO_REF:0000002	ECO:0000501	InterPro:IPR000426	P	XP_049874922.1	
0002	Refseq	XP_049874922.1	XP_049874922.1		GO:0019773	GO_REF:0000002	ECO:0000501	InterPro:IPR000426	C	XP_049874922.1	
0003	Refseq	XP_049874922.1	XP_049874922.1		GO:0051603	GO_REF:0000002	ECO:0000501	InterPro:IPR001353	P	XP_049874922.1	
0004	Refseq	XP_049874923.1	XP_049874923.1		GO:0003677	GO_REF:0000002	ECO:0000501	InterPro:IPR000003	F	XP_049874923.1	
0005	Refseq	XP_049874923.1	XP_049874923.1		GO:0003700	GO_REF:0000002	ECO:0000501	InterPro:IPR001628	F	XP_049874923.1	
0006	Refseq	XP_049874923.1	XP_049874923.1		GO:0003707	GO_REF:0000002	ECO:0000501	InterPro:IPR000003	F	XP_049874923.1	
0007	Refseq	XP_049874923.1	XP_049874923.1		GO:0005634	GO_REF:0000002	ECO:0000501	InterPro:IPR000003	C	XP_049874923.1	
0008	Refseq	XP_049874923.1	XP_049874923.1		GO:0006355	GO_REF:0000002	ECO:0000501	InterPro:IPR001628 InterPro:IPR013088	P	XP_049874923.1	
0009	Refseq	XP_049874923.1	XP_049874923.1		GO:0008270	GO_REF:0000002	ECO:0000501	InterPro:IPR001628 InterPro:IPR013088	F	XP_049874923.1	
0010	Refseq	XP_049874923.1	XP_049874923.1		GO:0043565	GO_REF:0000002	ECO:0000501	InterPro:IPR001628	F	XP_049874923.1	
0011	Refseq	XP_049874924.1	XP_049874924.1		GO:0003676	GO_REF:0000002	ECO:0000501	InterPro:IPR036397 InterPro:IPR001878	F	XP_049874924.1	
0012	Refseq	XP_049874924.1	XP_049874924.1		GO:0008270	GO_REF:0000002	ECO:0000501	InterPro:IPR001878	F	XP_049874924.1	
0013	Refseq	XP_049874924.1	XP_049874924.1		GO:0015074	GO_REF:0000002	ECO:0000501	InterPro:IPR001584	P	XP_049874924.1	

# 3. I want to get all the protein accession numbers for a GO category.

- The Amigo site can be used to search for relevant GO categories (<http://amigo.geneontology.org/>)
- Here, I searched for “peptidase activity” (GO:0008233)



The screenshot shows the AmiGO 2 interface. At the top, there is a navigation bar with links: Home, Search, Browse, Tools & Resources, Help, Feedback, and About. The main heading is "peptidase activity". Below this, the "Term Information" section is displayed. The "Accession" field is highlighted with a red box and contains "GO:0008233". Other fields include "Name" (peptidase activity), "Ontology" (molecular\_function), "Synonyms" (hydrolase, acting on peptide bonds, peptide hydrolase activity, protease activity, peptidase acti...), "Alternate IDs" (GO:0070010, GO:0070011), "Definition" (Catalysis of the hydrolysis of a peptide bond. A peptide bond is a covalent bond formed when the carbon : electrons with the nitrogen atom from the amino group of a second amino acid. Source: GOC:jl, ISBN:0815), "Comment" (None), "History" (See term history for GO:0008233 at QuickGO), "Subset" (goslim\_metagenomics, goslim\_chembl, goslim\_drosophila, goslim\_candida, goslim\_yeast, goslim\_pir), and "Related" (Link to all genes and gene products annotated to peptidase activity (excluding "regulates")., Link to all direct and indirect annotations to peptidase activity (excluding "regulates")., Link to all direct and indirect annotations download (limited to first 10,000) for peptidase activity (exclud). At the bottom, there is a button "Include 'regulates'" and a note "For more information, please see the ontology relation documentation."

AmiGO 2 Home Search Browse Tools & Resources Help Feedback About

## peptidase activity

**Term Information**

**Accession** GO:0008233

**Name** peptidase activity

**Ontology** molecular\_function

**Synonyms** hydrolase, acting on peptide bonds, peptide hydrolase activity, protease activity, peptidase acti... [more](#)

**Alternate IDs** GO:0070010, GO:0070011

**Definition** Catalysis of the hydrolysis of a peptide bond. A peptide bond is a covalent bond formed when the carbon : electrons with the nitrogen atom from the amino group of a second amino acid. Source: GOC:jl, ISBN:0815

**Comment** None

**History** See term history for GO:0008233 at QuickGO

**Subset** goslim\_metagenomics  
goslim\_chembl  
goslim\_drosophila  
goslim\_candida  
goslim\_yeast  
goslim\_pir

**Related** [Link](#) to all genes and gene products annotated to peptidase activity (excluding "regulates").  
[Link](#) to all direct and indirect annotations to peptidase activity (excluding "regulates").  
[Link](#) to all direct and indirect annotations download (limited to first 10,000) for peptidase activity (exclud

[Include "regulates"](#)


For more information, please see the [ontology relation documentation](#).

# 3. I want to get all the protein accession numbers for a GO category.

Protein  
accessions

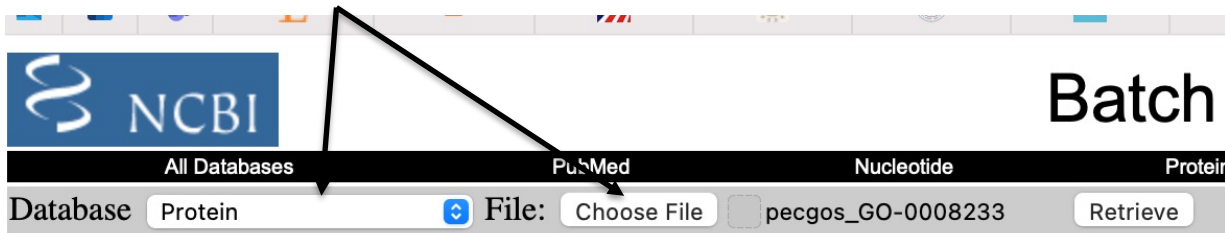
GO  
annotations

752	Refseq	XP_049875033.1	XP_049875033.1	GO:0008203	GO_REF:000	ECO:0000501	InterPro:IPRC P	XP_049
753	Refseq	XP_049875034.1	XP_049875034.1	GO:0008203	GO_REF:000	ECO:0000501	InterPro:IPRC P	XP_049
754	RefSeq	XP_049877931.1	XP_049877931.1	GO:0008233	GO_REF:000	ECO:0000247	UniprotKB:Q F	XP_049
755	RefSeq	XP_049877931.1	XP_049877931.1	GO:0008233	GO_REF:000	ECO:0000247	UniprotKB:Q F	XP_049
756	Refseq	XP_049870258.1	XP_049870258.1	GO:0008233	GO_REF:000	ECO:0000501	InterPro:IPRC F	XP_049
757	Refseq	XP_049872517.1	XP_049872517.1	GO:0008233	GO_REF:000	ECO:0000501	InterPro:IPRC F	XP_049
758	Refseq	XP_049877431.1	XP_049877431.1	GO:0008233	GO_REF:000	ECO:0000501	InterPro:IPRC F	XP_049
759	Refseq	XP_049877445.1	XP_049877445.1	GO:0008233	GO_REF:000	ECO:0000501	InterPro:IPRC F	XP_049
760	Refseq	XP_049865743.1	XP_049865743.1	GO:0008234	GO_REF:000	ECO:0000501	InterPro:IPRC F	XP_049
761	Refseq	XP_049866023.1	XP_049866023.1	GO:0008234	GO_REF:000	ECO:0000501	InterPro:IPRC F	XP_049
762	Refseq	XP_049866656.1	XP_049866656.1	GO:0008234	GO_REF:000	ECO:0000501	InterPro:IPRC F	XP_049
763	Refseq	XP_049866955.1	XP_049866955.1	GO:0008234	GO_REF:000	ECO:0000501	InterPro:IPRC F	XP_049
764	Refseq	XP_049867066.1	XP_049867066.1	GO:0008234	GO_REF:000	ECO:0000501	InterPro:IPRC F	XP_049


 Copy all protein  
 accessions to a text  
 file

# 3. I want to get all the protein accession numbers for a GO category.

Select protein database and upload file with protein accessions



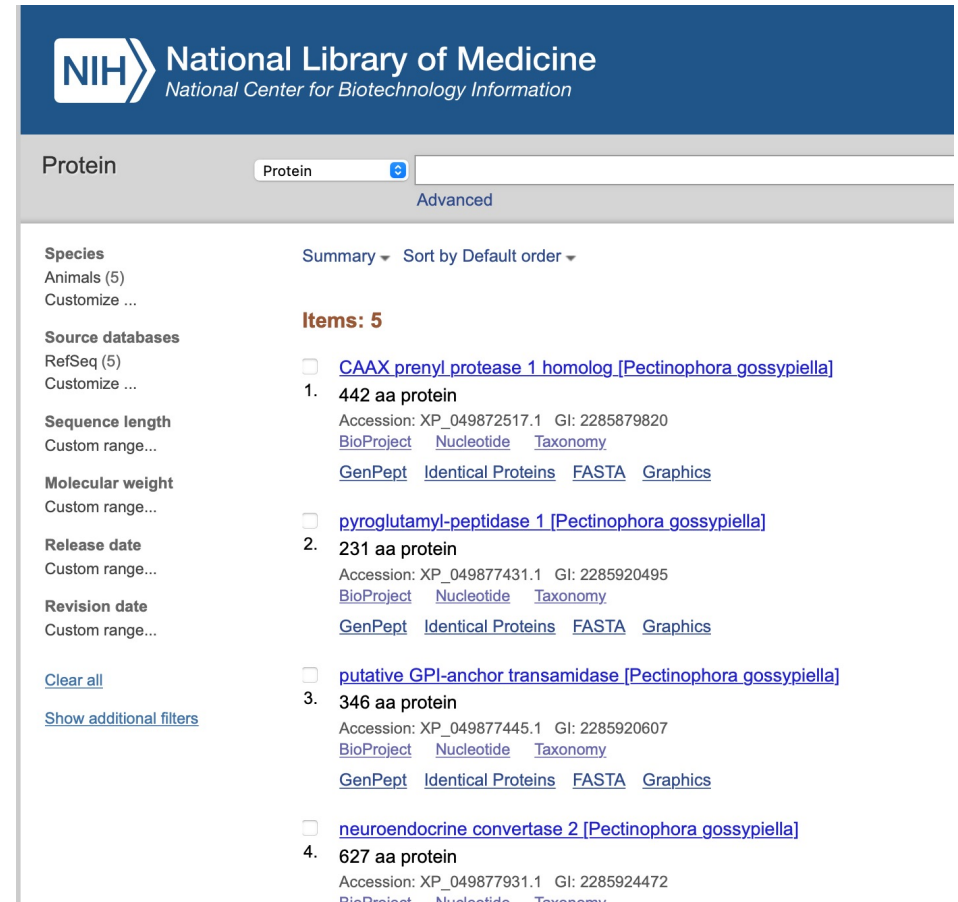
NCBI Batch Entrez interface showing the 'Database' dropdown set to 'Protein' and the 'File' field containing 'pecgos\_GO-0008233'.

## Batch Entrez

Given a file of Entrez accession numbers or other identifiers, Batch Entrez downloads the corresponding records.

## Instructions

<https://www.ncbi.nlm.nih.gov/sites/batchentrez>



NIH National Library of Medicine  
National Center for Biotechnology Information

Protein

Species  
Animals (5)  
Customize ...

Source databases  
RefSeq (5)  
Customize ...

Sequence length  
Custom range...

Molecular weight  
Custom range...

Release date  
Custom range...

Revision date  
Custom range...

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

Summary ▾ Sort by Default order ▾

Items: 5

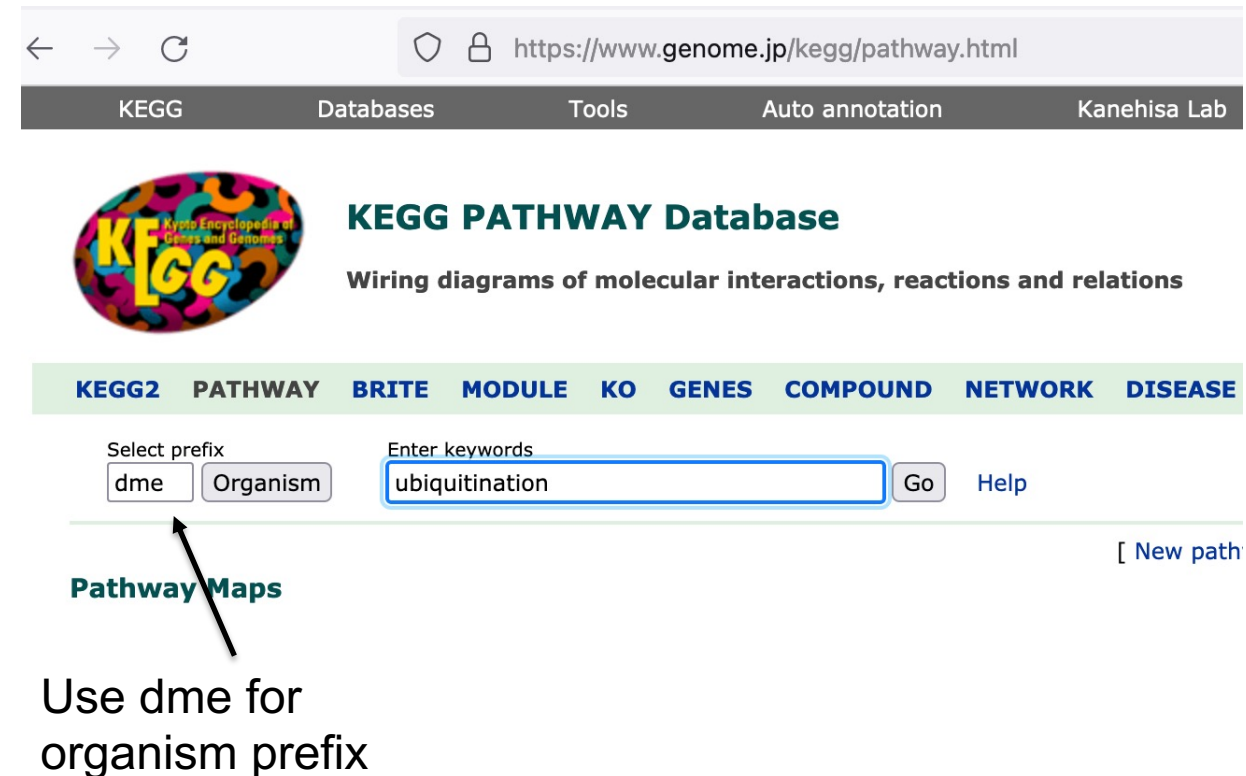
- ☐ [CAAX prenyl protease 1 homolog \[Pectinophora gossypiella\]](#)  
442 aa protein  
Accession: XP\_049872517.1 GI: 2285879820  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)
- ☐ [pyroglutamyl-peptidase 1 \[Pectinophora gossypiella\]](#)  
231 aa protein  
Accession: XP\_049877431.1 GI: 2285920495  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)
- ☐ [putative GPI-anchor transamidase \[Pectinophora gossypiella\]](#)  
346 aa protein  
Accession: XP\_049877445.1 GI: 2285920607  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)
- ☐ [neuroendocrine convertase 2 \[Pectinophora gossypiella\]](#)  
627 aa protein  
Accession: XP\_049877931.1 GI: 2285924472  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)

### 3. I want to get all the protein accession numbers for a GO category.

The screenshot displays the Jbrowse genome browser interface. The top navigation bar includes the 'pollo' logo and menu items: File, View, Help, and a 'Logi' button. A scale bar at the top shows genomic coordinates from 0 to 14,000,000. Below the scale bar are navigation controls (left and right arrows, zoom in/out) and a search bar containing 'NC\_065421.1'. A 'Go' button is next to the search bar. The main display area shows a genomic track for 'NC\_065421.1'. A red box highlights the gene track for 'neuroendocrine convertase 2' (XM\_050021974.1). Below the gene track, several transcript variants are listed: 'ceptor, transcript variant X2', 'ceptor, transcript variant X4', 'ceptor, transcript variant X1', and 'ceptor, transcript variant X2'. An arrow points from the text 'Find individual protein accessions in Jbrowse' to the protein track entry 'XP\_049877931.1'.

## 4. I want to find all the *Schistocerca americana* proteins annotated to ubiquitination pathways.

- Start with KEGG2:  
<https://www.genome.jp/kegg/pathway.html>
- Search for ubiquitination
- Choose 'dme' for prefix – *Drosophila melanogaster*
- Comparison paper:  
<https://www.biorxiv.org/content/10.1101/2021.10.11.464014v1.abstract>



The screenshot shows the KEGG PATHWAY Database homepage. The browser address bar displays <https://www.genome.jp/kegg/pathway.html>. The navigation bar includes links for KEGG, Databases, Tools, Auto annotation, and Kanehisa Lab. The main header features the KEGG logo and the text "KEGG PATHWAY Database" and "Wiring diagrams of molecular interactions, reactions and relations". Below this is a horizontal menu with tabs: KEGG2, PATHWAY, BRITE, MODULE, KO, GENES, COMPOUND, NETWORK, and DISEASE. The search section contains a "Select prefix" dropdown menu with "dme" selected, an "Organism" button, a text input field with "ubiquitination", a "Go" button, and a "Help" link. A "Pathway Maps" link is visible below the search area. An arrow points from the text "Use dme for organism prefix" to the "dme" selection in the dropdown menu.

Use dme for organism prefix

# 4. I want to find all the *Schistocerca americana* proteins annotated to the ubiquitination pathway.

← → ↺

🔒 [https://www.kegg.jp/kegg-bin/search\\_pathway\\_text?map=dme&keyword=ubiq](https://www.kegg.jp/kegg-bin/search_pathway_text?map=dme&keyword=ubiq)


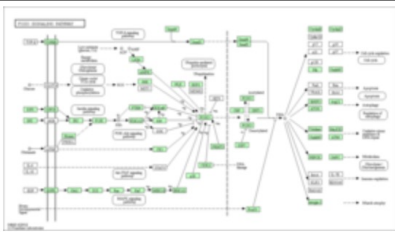
Pathway Text Search

Number of entries in a page 20 ▾

Hide thumbnail

Items : 1 - 2 of 2


Select dme04120

	Thumbnail Image	Name	Description	
<div>dme04120</div>		Ubiquitin mediated proteolysis	Protein <b>ubiquitination</b> plays an important role in eukaryotic cellular processes. It mainly functions...	Dmel_CG: Dmel_CG: Dmel_CG:
<div>dme04068</div>		FoxO signaling pathway	The forkhead box O (FOXO) family of transcription factors regulates the expression of genes in cellu...	C00031 (Phosphat C00008 (

Items : 1 - 2 of 2

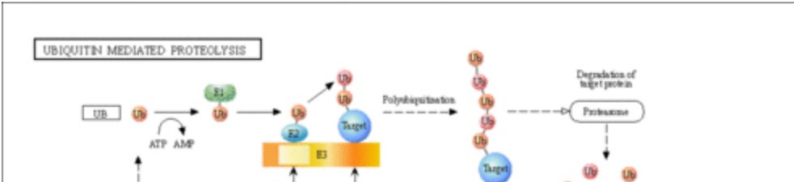
← → ↺

🔒 <https://www.kegg.jp/entry/dme04120>



PATHWAY: dme04120


Help

Entry	dme04120	Pathway
Name	Ubiquitin mediated proteolysis - Drosophila melanogaster (fruit fly)	
Description	Protein ubiquitination plays an important role in eukaryotic cellular processes. It mainly functions as a signal for 26S proteasome dependent protein degradation. The addition of ubiquitin to proteins being degraded is performed by a reaction cascade consisting of three enzymes, named E1 (ubiquitin activating enzyme), E2 (ubiquitin conjugating enzyme), and E3 (ubiquitin ligase). Each E3 has specificity to its substrate, or proteins to be targeted by ubiquitination. Many E3s are discovered in eukaryotes and they are classified into four types: HECT type, U-box type, single RING-finger type, and multi-subunit RING-finger type. Multi-subunit RING-finger E3s are exemplified by cullin-Rbx E3s and APC/C. They consist of a RING-finger-containing subunit (RBX1 or RBX2) that functions to bind E2s, a scaffold-like cullin molecule, adaptor proteins, and a target recognizing subunit that binds substrates.	
Class	Genetic Information Processing; Folding, sorting and degradation <a href="#">BRITE hierarchy</a>	
Pathway map	<div><div>dme04120</div> Ubiquitin mediated proteolysis</div> 	

# 4. I want to find all the *Schistocerca americana* proteins annotated to the ubiquitination pathway.

← → ↺

https://reactome.org/content/query?q=ubiquitination&species=Drosophila+melanogaster&types=Pathway&cluster=t



About ▾Content ▾Docs ▾Tools ▾Con

ubiquitination

Go!

Search results for **ubiquitination**

Showing 8 results out of 8

Species

☒ Drosophila melanogaster (8)

☐ Homo sapiens (102)

☐ Gallus gallus (8)

☐ Bos taurus (5)

☐ Canis familiaris (5)

☐ Danio rerio (5)

More...


Types

☒ Pathway (8)

☐ Reaction (32)

Compartments


Pathway (8 results from a total of 8)

 **Protein ubiquitination**

Identifier: R-DME-8852135

Species: Drosophila melanogaster


This event has been computationally inferred from an event that has been demonstrated in another based on the homology mapping from PANTHER. Briefly, reactions for which all involved... [Read more](#)

 **Ubiquitination and proteolysis of phosphorylated CI**

Identifier: R-DME-209360

Species: Drosophila melanogaster

Spatzle (SPZ) dimer binding leads to Toll (TL) receptor homodimerisation and activation.

 **Ubiquitination and degradation of phosphorylated ARM**

Identifier: R-DME-209461


Species: Drosophila melanogaster

Compartment: cytosol

Spatzle (SPZ) dimer binding leads to Toll (TL) receptor homodimerisation and activation.

← → ↺


https://reactome.org/content/detail/R-DME-8852135



Pathway identifier

About ▾Content ▾

e.g. O95631, NTN1, signaling by EGFR, glucose, GO:0043293

 **Protein ubiquitination**

Stable Identifier

Type


Species

R-DME-8852135


Pathway


Drosophila melanogaster


Locations in the PathwayBrowser

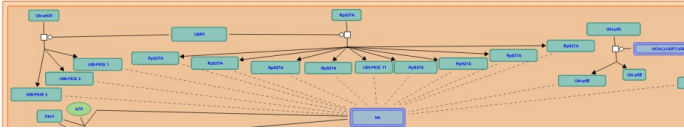
 **Metabolism of proteins (Drosophila melanogaster)**

General

 SBML

 BioPAX

 PDF



## 4. I want to find all the *Schistocerca americana* proteins annotated to the ubiquitination pathway.

- KEGG annotations (e.g. *Drosophila melanogaster* annotations computed by KEGG) are primarily based on sequence similarity to ortholog groups
- *D. melanogaster* Reactome annotations are based primarily on sequence similarity to human proteins
- FlyBase has additional pathway annotations.
- Delineations of the pathways themselves, and what proteins are assigned to a pathway, differ between all three resources.

# 4. I want to find all the *Schistocerca americana* proteins annotated to the ubiquitination pathway.

[←](#) [→](#) [↻](#) <https://i5k.nal.usda.gov/data/Arthropoda/schan>

## Index of /data/Arthropoda/scheme-(Schis Primary Gene Set/NCBI Schistocerca am annotation/

---

<a href="#">../</a>	23-Sep-2022 23:08
<a href="#">GOanna/</a>	23-Sep-2022 23:20
<a href="#">Interproscan/</a>	23-Sep-2022 23:20
<a href="#">KOBAS/</a>	27-Apr-2022 16:12
<a href="#">GCF_021461395.2_complete.gaf.tsv</a>	21-Jul-2022 19:46
<a href="#">README.txt</a>	27-Apr-2022 16:12
<a href="#">README.txt~</a>	

[←](#) [→](#) [↻](#) <https://i5k.nal.usda.gov/data/Arthropoda/schan>

## Index of /data/Arthropoda/scheme-(Schis Primary Gene Set/NCBI Schistocerca am annotation/KOBAS/

---

<a href="#">../</a>	
<a href="#">GCF_021461395.2</a>	27-Apr-2022 16:18
<a href="#">GCF_021461395.2_KOBAS_acc_pathways.tsv</a>	27-Apr-2022 16:18
<a href="#">GCF_021461395.2_KOBAS_pathways_acc.tsv</a>	27-Apr-2022 16:18

Select file  
 organized by  
 pathway

# 4. I want to find all the *Schistocerca americana* proteins annotated to the ubiquitination pathway.

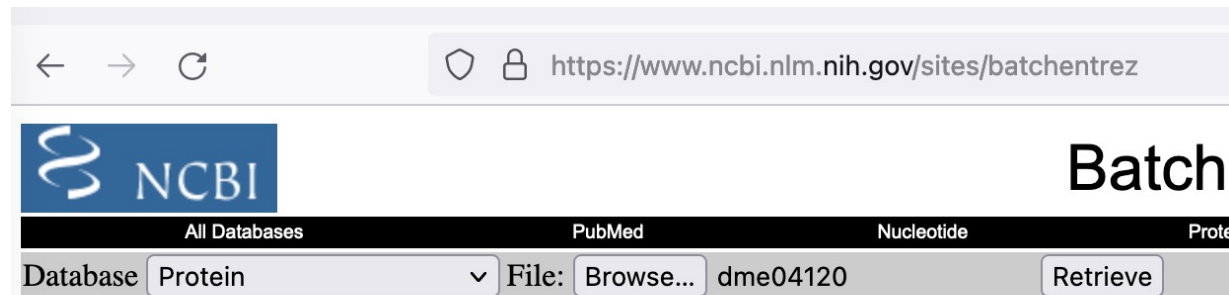
Q dme04120

0.1,XP\_046991197.1,XP\_046999343.1,XP\_046981228.1,XP\_046984995.1,XP\_0469  
 EGG:dme04120  
 P\_046993085.1,XP\_046981059.1,XP\_046994848.1,XP\_046999012.1,XP\_047000354  
 8.1,XP\_047003904.1,XP\_046990850.1,XP\_046995563.1,XP\_046999469.1,XP\_0469  
 005094.1,XP\_047000014.1,XP\_046996878.1,XP\_046979735.1,XP\_046994184.1,XP  
 P\_046985348.1,XP\_046999038.1,XP\_046993261.1,XP\_046983758.1,XP\_047003419  
 2.1,XP\_046991008.1,XP\_046981103.1,XP\_046979724.1,XP\_047003078.1,XP\_0470  
 001181.1,XP\_046989151.1,XP\_046994604.1,XP\_047000356.1,XP\_046986523.1,XP  
 P\_046991457.1,XP\_047000596.1,XP\_046996528.1,XP\_046987740.1,XP\_046982906  
 9.1,XP\_047003237.1,XP\_046989148.1,XP\_046998163.1,XP\_046994023.1,XP\_0469  
 991456.1,XP\_046997023.1,XP\_046999467.1,XP\_046982218.1,XP\_046995883.1,XP  
 P\_046979614.1,XP\_046989383.1,XP\_046992382.1,XP\_046993078.1,XP\_046986560  
 7.1,XP\_046979293.1,XP\_047000604.1,XP\_046996409.1,XP\_047002875.1,XP\_0469  
 990929.1,XP\_046989149.1,XP\_046999010.1,XP\_046999684.1,XP\_046992896.1,XP  
 P\_046992967.1,XP\_047001201.1,XP\_046981442.1,XP\_047000355.1,XP\_046999825  
 7.1,XP\_046982722.1,XP\_046980112.1,XP\_046999468.1,XP\_046998185.1,XP\_0469  
 981105.1,XP\_046996813.1,XP\_046991011.1,XP\_046996696.1,XP\_047003446.1  
 Reactome:R-DME-209228 XP\_046985860.1,XP\_046997807.1,XP\_046997808.1,XP  
 EGG:dme00280

Q R-DME-8852135

35.1  
 Reactome:R-DME-2980766  
 XP\_046992852.1,XP\_046996561.1,XP\_046979463.1,XP\_046995355.1,XP  
 77.1,XP\_046999640.1,XP\_047001499.1,XP\_046996559.1,XP\_046999552  
 5991727.1,XP\_046989353.1,XP\_047003441.1,XP\_046981732.1,XP\_0469  
 XP\_046992232.1,XP\_046991726.1,XP\_046987701.1,XP\_046993150.1,XP  
 00.1,XP\_046993152.1,XP\_046989146.1,XP\_046999642.1,XP\_046992851  
 Reactome:R-DME-8852135  
 XP\_046993085.1,XP\_046998490.1,XP\_046982807.1,XP\_047001445.1,XP  
 56.1,XP\_046986652.1,XP\_046996407.1,XP\_047003680.1,XP\_047003419  
 5995846.1,XP\_046998163.1,XP\_046998504.1,XP\_047001590.1,XP\_0469  
 XP\_046996409.1,XP\_047002875.1,XP\_046989247.1,XP\_046998169.1,XP  
 98.1,XP\_046998185.1,XP\_046993575.1,XP\_046997621.1,XP\_046993071  
 Reactome:R-DME-69109  
 XP\_047001650.1,XP\_046997046.1,XP\_046981277.1,XP\_046992141.1,XP  
 57.1,XP\_046999463.1,XP\_047001215.1,XP\_046991794.1,XP\_046979774  
 Reactome:R-DME-8939246 XP\_046988142.1,XP\_046990669.1,XP\_046  
 Reactome:R-DME-9607240  
 XP\_046993897.1,XP\_046992770.1,XP\_047004650.1,XP\_046997292.1,XP

# 4. I want to find all the *Schistocerca americana* proteins annotated to the ubiquitination pathway.



## Batch Entrez

Given a file of Entrez accession numbers or other identifiers, Batch Entrez downloads the corresponding records.

## Instructions

Select protein database and upload file with protein accessions



# 4. I want to find all the *Schistocerca americana* proteins annotated to the ubiquitination pathway.

Protein

Advanced

Summary ▾ 20 per page ▾ Sort by Default order ▾

Items: 1 to 20 of 196

☐ [baculoviral IAP repeat-containing protein 6 \[Schistocerca americana\]](#)

1. 5021 aa protein  
Accession: XP\_046995696.1 GI: 2209610067  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

☐ [elongin-B isoform X1 \[Schistocerca americana\]](#)

2. 141 aa protein  
Accession: XP\_047003078.1 GI: 2209610304  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

☐ [elongin-B isoform X2 \[Schistocerca americana\]](#)

3. 123 aa protein  
Accession: XP\_047003084.1 GI: 2209610306  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

Choose Destination  
☒ File ☐ Clipboard  
☐ Collections

Download 196 items.

Format

- ✓ Summary
- GenPept
- GenPept (full)
- FASTA**
- ASN.1
- XML
- INSDSeq XML
- TinySeq XML
- Feature Table
- FASTA CDS
- Accession List
- GI List
- GFF3

Send to: ▾ Filters: [Manage Filters](#)

Find related data


Database:

Option:

Nucleotide sequence from coding region

Recent activity

[Turn Off](#) [Clear](#)

 [chitinase A-like \[Pectinophora gossypiella\]](#)  
Protein

[See more...](#)

## 4. I want to find all the *Schistocerca americana* proteins annotated to the ubiquitination pathway.

apollo File View Help

0 200,000,000 400,000,000 600,000,000 800,000,000 1,000,000,000 1,200,000,000

139,500,000 139,625,000

NC\_060119.1 XP\_046995696.1 Go

XP\_046995696.1

NCBI\_Annotation\_Release\_100\_Gene

XM\_047139740.1  
baculoviral IAP repeat-containing protein 6

rna-TrnaI-caa-7  
tRNA-Leu

Search for XP identifier or  
name in Jbrowse

# Questions?

- How are you planning to use the functional annotations?

# Acknowledgements



Anna Childers



- The i5k Workspace@NAL team
- All of our data contributors!



Fiona M McCarthy



Amanda Cooksey



Surya Saha

