I5k Workspace webinar
Changes to gene and protein data access

Monica Poelchau

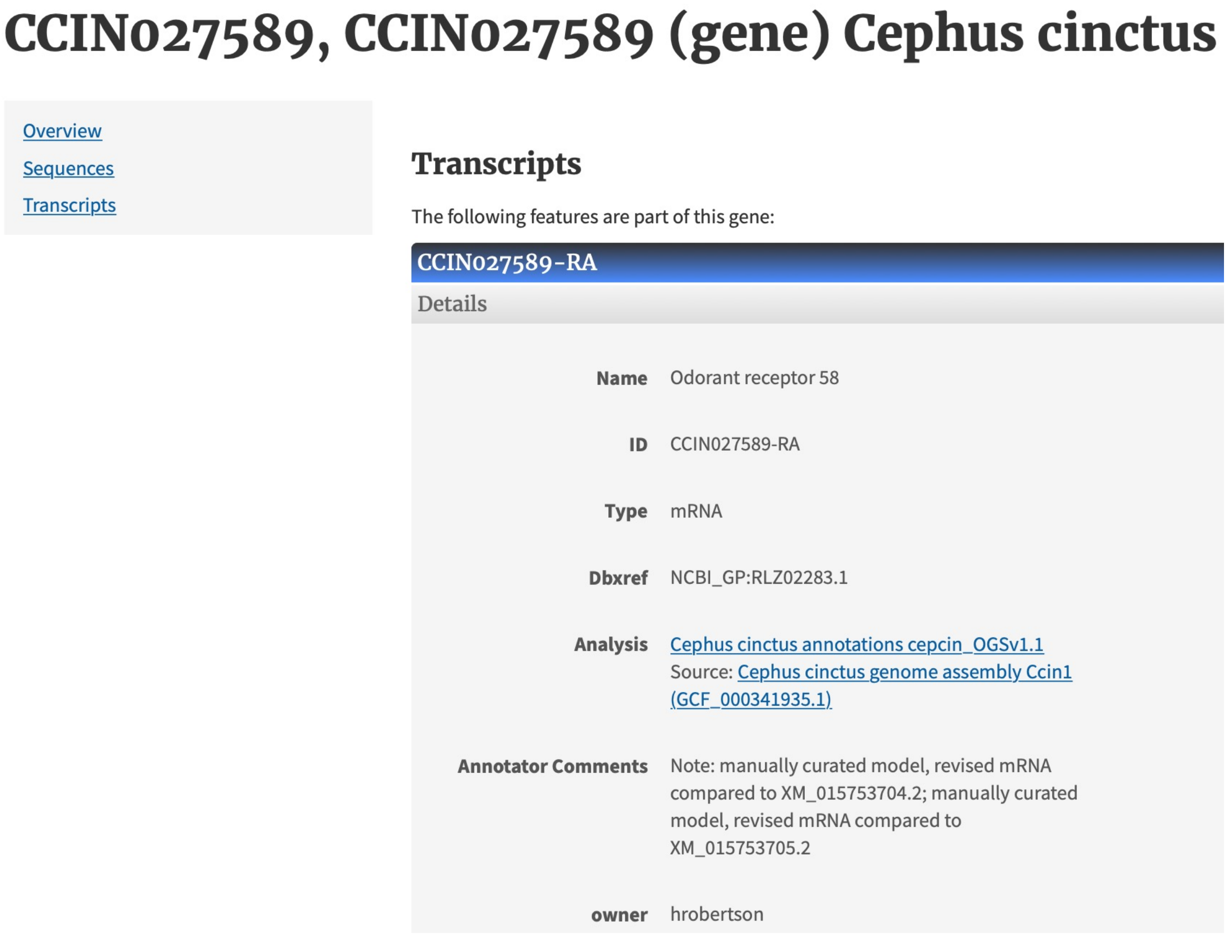
7/19/2022

Agenda

- Why did we remove the i5k Workspace gene and protein pages?
  - What exactly did we remove?
  - Why?
  - A brief digression on how we keep i5k Workspace data in sync with NCBI
- The data are still accessible – how can you find what you need?
  - Gene and protein information
  - Dataset-level information
- Q&A to address any questions or concerns.

Gene and protein page removal

- What was removed?
  - Pages that showed information at the individual gene and protein level
  - Ability to search for gene and protein information (e.g. “Cimex lectularius heat shock protein”)



CCIN027589, CCIN027589 (gene) Cephus cinctus

Overview
Sequences
Transcripts

Transcripts

The following features are part of this gene:

CCIN027589-RA

Details

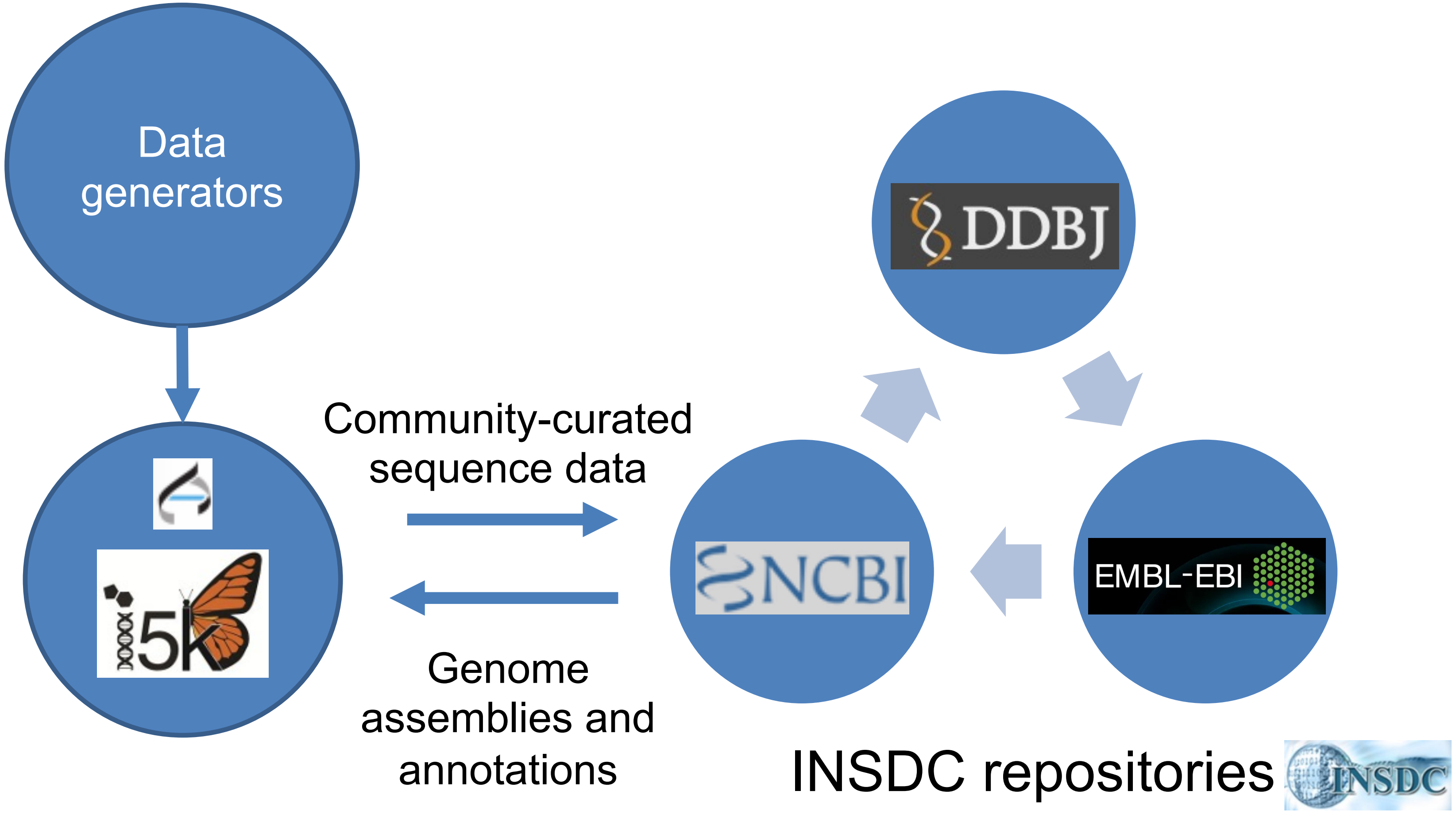
Name	Odorant receptor 58
ID	CCIN027589-RA
Type	mRNA
Dbxref	NCBI_GP:RLZ02283.1
Analysis	Cephus cinctus annotations cepcin_OGSv1.1 Source: Cephus cinctus genome assembly Ccin1 (GCF_000341935.1)
Annotator Comments	Note: manually curated model, revised mRNA compared to XM_015753704.2; manually curated model, revised mRNA compared to XM_015753705.2
owner	hrobertson

Why did we remove the pages?

1. The data are already available elsewhere at the gene/protein level: :
    1. NCBI
    2. Other insect databases, e.g. HymenopteraMine, VEuPathDB
2. The data are still available at the dataset level:
    1. I5k Workspace@NAL
    2. Ag Data Commons
    3. NCBI
    4. Other insect databases, e.g. HymenopteraMine, VEuPathDB
3. Not a high priority for users
    1. Structured interview results
    2. Fewer page views
4. Challenging to maintain the code
5. Challenging to keep the content up-to-date and comprehensive

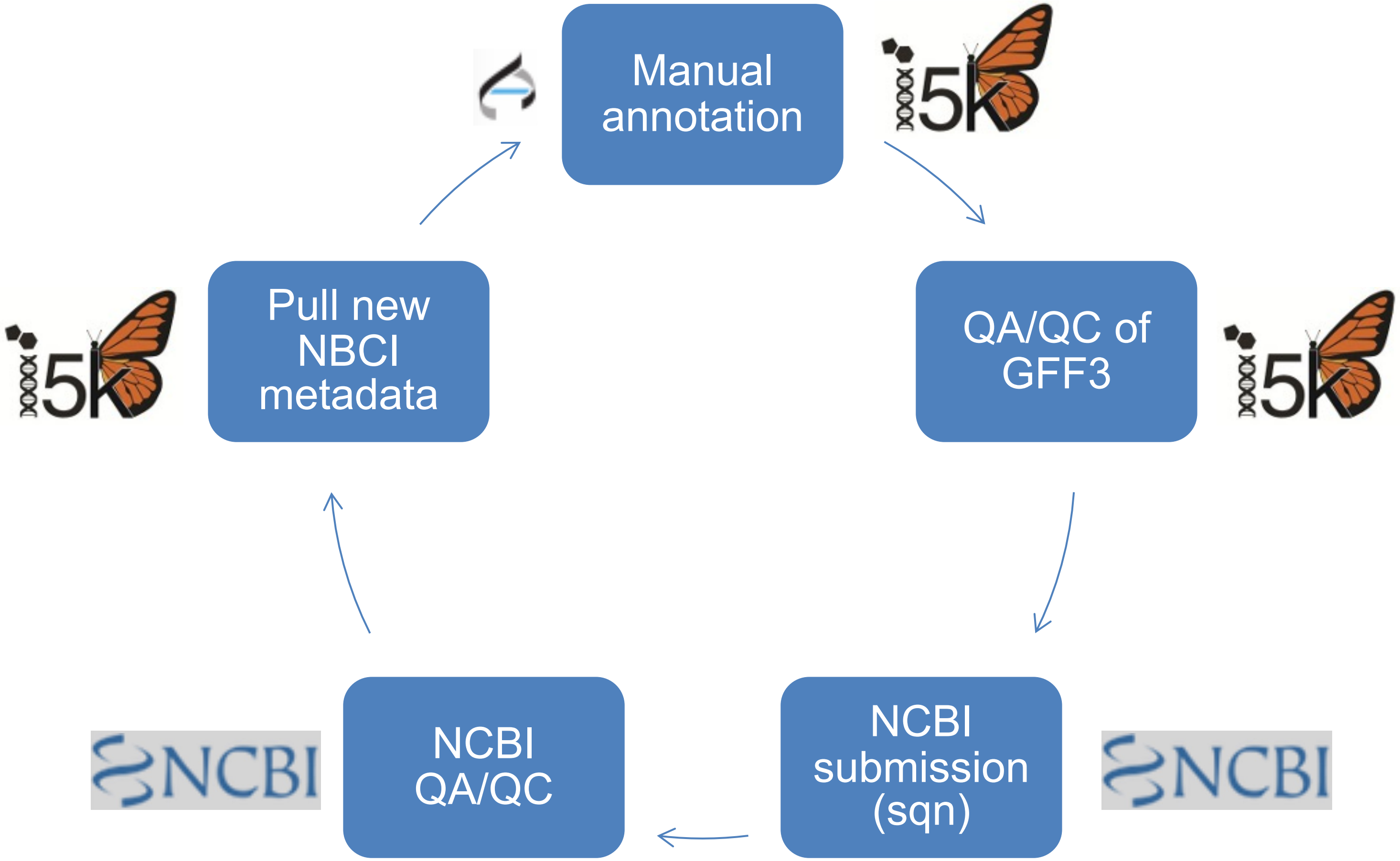
Why are the data at NCBI?

- NCBI, as part of the INSDC, is the primary archive for all sequence data
- Our policy is to submit all sequence data and metadata that can go to NCBI, to NCBI


Data generators
5k
Community-curated sequence data
Genome assemblies and annotations
DDBJ
NCBI
EMBL-EBI
INSDC repositories
INSDC

https://doi.org/10.1093/nar/gkx1097

Manual annotation QA/QC and submission



Manual annotation

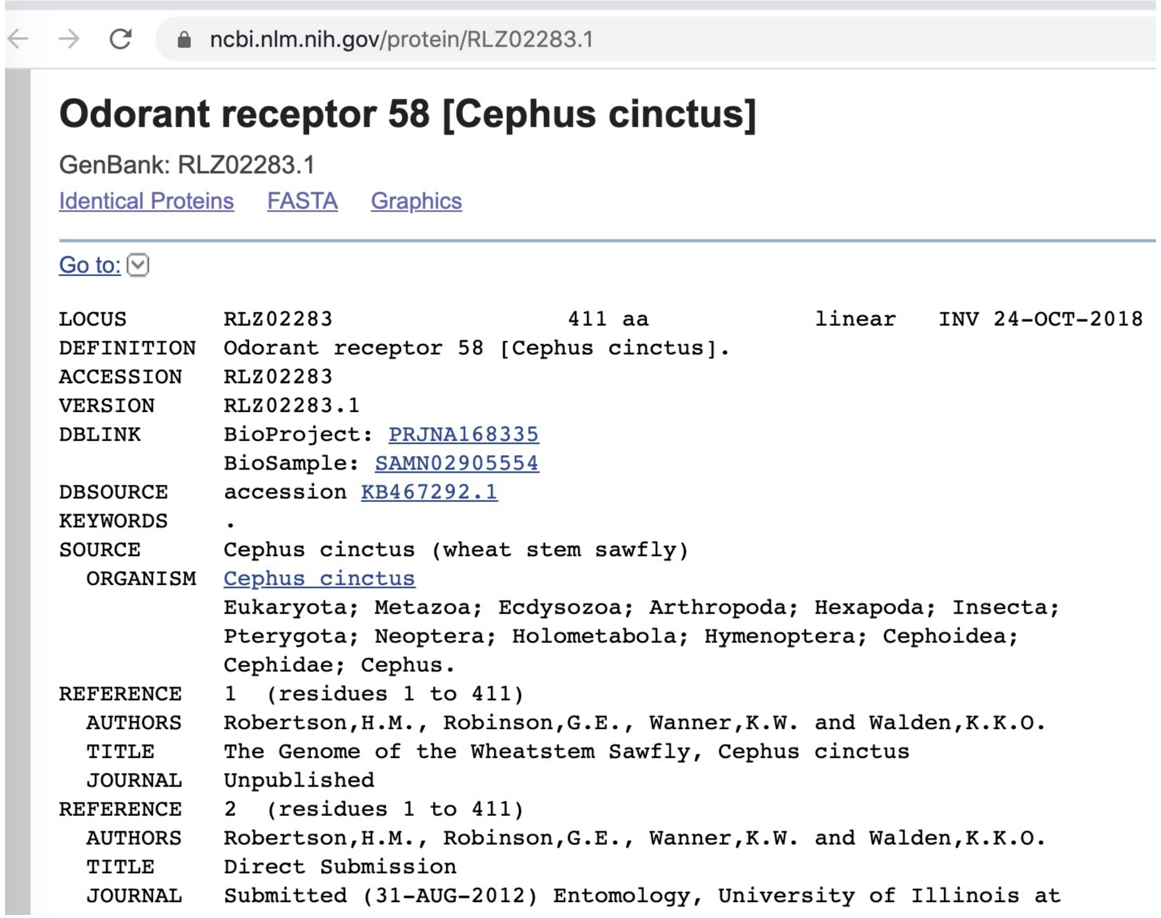
QA/QC of GFF3

NCBI submission (sqn)

NCBI QA/QC

Pull new NBCI metadata

Result of a successful GenBank submission



Odorant receptor 58 [Cephus cinctus]

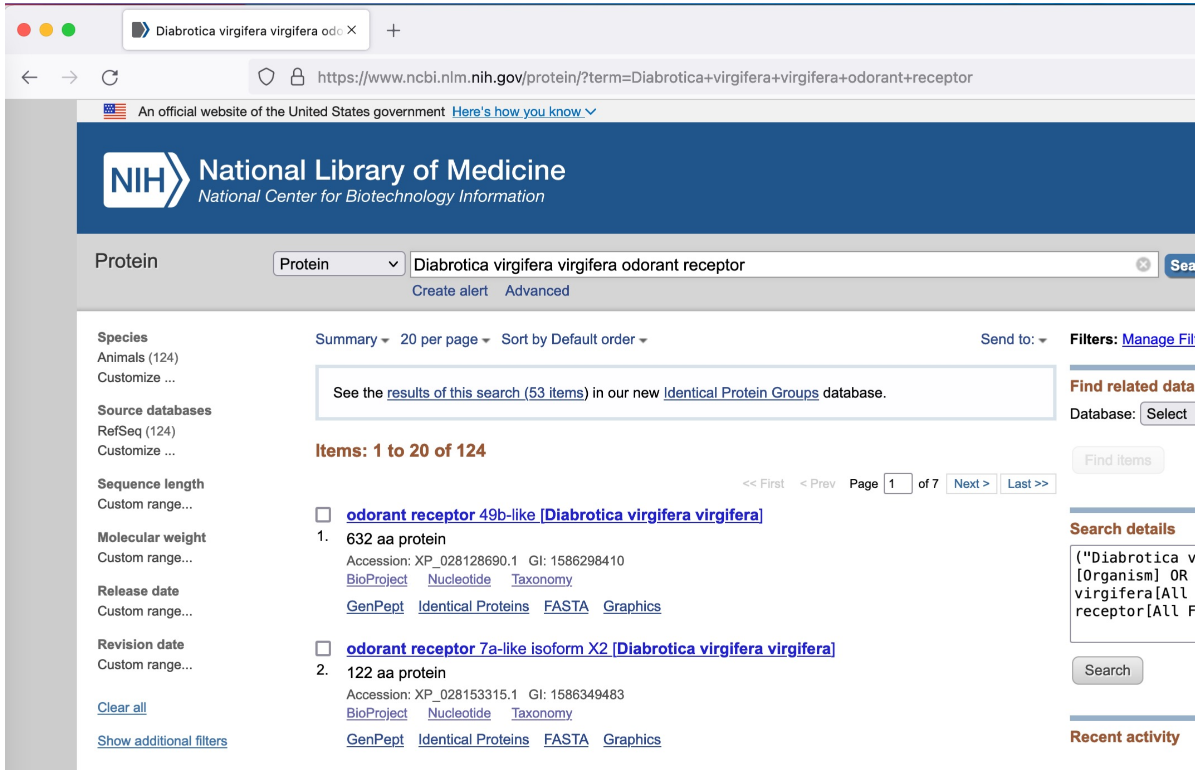
GenBank: RLZ02283.1

```
LOCUS       RLZ02283                 411 aa            linear   INV 24-OCT-2018
DEFINITION  Odorant receptor 58 [Cephus cinctus].
ACCESSION   RLZ02283
VERSION     RLZ02283.1
DBLINK      BioProject: PRJNA168335
            BioSample: SAMN02905554
DBSOURCE    accession KB467292.1
KEYWORDS    .
SOURCE      Cephus cinctus (wheat stem sawfly)
  ORGANISM  Cephus cinctus
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Holometabola; Hymenoptera; Cephoidea;
            Cephidae; Cephus.
REFERENCE   1  (residues 1 to 411)
  AUTHORS   Robertson,H.M., Robinson,G.E., Wanner,K.W. and Walden,K.K.O.
  TITLE     The Genome of the Wheatstem Sawfly, Cephus cinctus
  JOURNAL   Unpublished
REFERENCE   2  (residues 1 to 411)
  AUTHORS   Robertson,H.M., Robinson,G.E., Wanner,K.W. and Walden,K.K.O.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-AUG-2012) Entomology, University of Illinois at
```

https://www.ncbi.nlm.nih.gov/protein/RLZ02282.1

How can I find a specific gene?

NCBI Protein search



Diabrotica virgifera virgifera odo

https://www.ncbi.nlm.nih.gov/protein/?term=Diabrotica+virgifera+virgifera+odorant+receptor

An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Protein | Protein | Diabrotica virgifera virgifera odorant receptor | Sea

Create alert Advanced

Species
Animals (124)
Customize ...

Source databases
RefSeq (124)
Customize ...

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Summary 20 per page Sort by Default order Send to:

See the results of this search (53 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 124

<< First < Prev Page 1 of 7 Next > Last >>

1. □ odorant receptor 49b-like [Diabrotica virgifera virgifera]
632 aa protein
Accession: XP_028128690.1 GI: 1586298410
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

2. □ odorant receptor 7a-like isoform X2 [Diabrotica virgifera virgifera]
122 aa protein
Accession: XP_028153315.1 GI: 1586349483
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

Filters: Manage Fil

Find related data
Database: Select
Find items

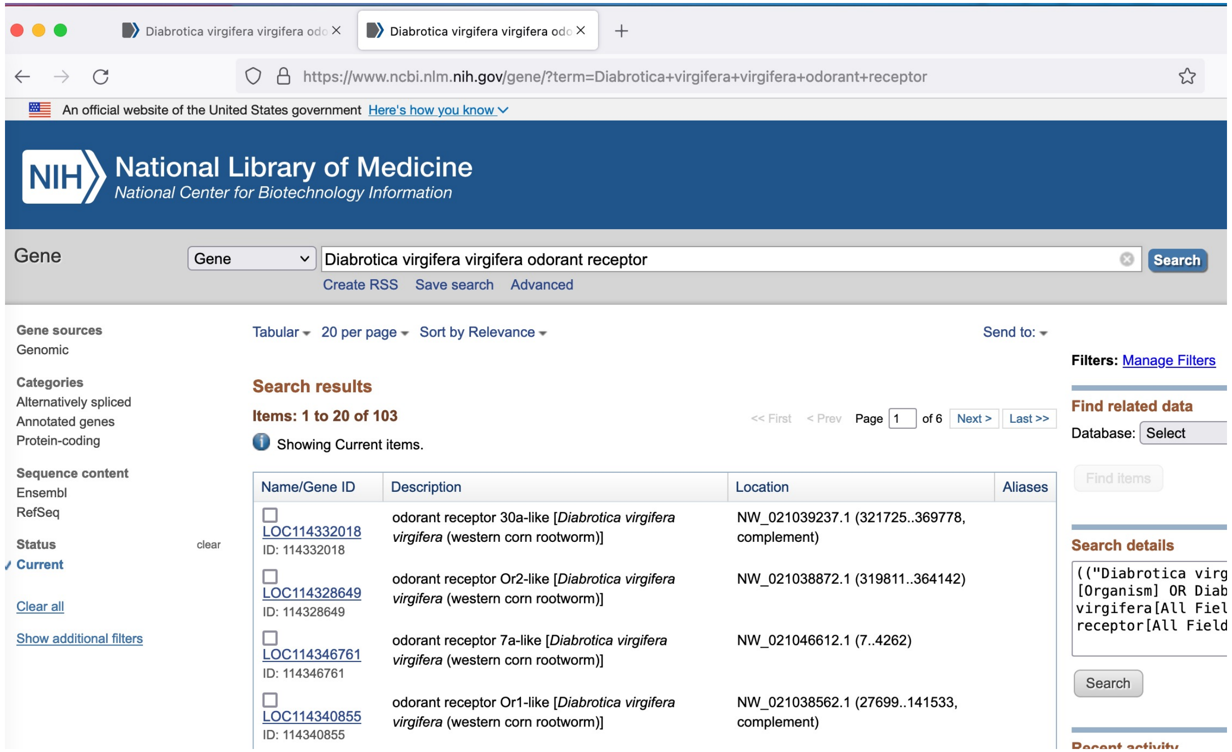
Search details

```
("Diabrotica v
[Organism] OR
virgifera[All
receptor[All F
```

Search

Recent activity

NCBI Gene search



Diabrotica virgifera virgifera odo | Diabrotica virgifera virgifera odo

https://www.ncbi.nlm.nih.gov/gene/?term=Diabrotica+virgifera+virgifera+odorant+receptor

An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Gene | Gene | Diabrotica virgifera virgifera odorant receptor | Search

Create RSS Save search Advanced

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Protein-coding

Sequence content
Ensembl
RefSeq

Status clear
✓ Current

Clear all

Show additional filters

Tabular 20 per page Sort by Relevance Send to:

Search results

Items: 1 to 20 of 103

<< First < Prev Page 1 of 6 Next > Last >>

Showing Current items.

Name/Gene ID	Description	Location	Aliases
□ LOC114332018 ID: 114332018	odorant receptor 30a-like [*Diabrotica virgifera virgifera* (western corn rootworm)]	NW_021039237.1 (321725..369778, complement)	
□ LOC114328649 ID: 114328649	odorant receptor Or2-like [*Diabrotica virgifera virgifera* (western corn rootworm)]	NW_021038872.1 (319811..364142)	
□ LOC114346761 ID: 114346761	odorant receptor 7a-like [*Diabrotica virgifera virgifera* (western corn rootworm)]	NW_021046612.1 (7..4262)	
□ LOC114340855 ID: 114340855	odorant receptor Or1-like [*Diabrotica virgifera virgifera* (western corn rootworm)]	NW_021038562.1 (27699..141533, complement)	

Filters: Manage Filters

Find related data
Database: Select
Find items

Search details

```
(("Diabrotica virg
[Organism] OR Diab
virgifera[All Fiel
receptor[All Field
```

Search

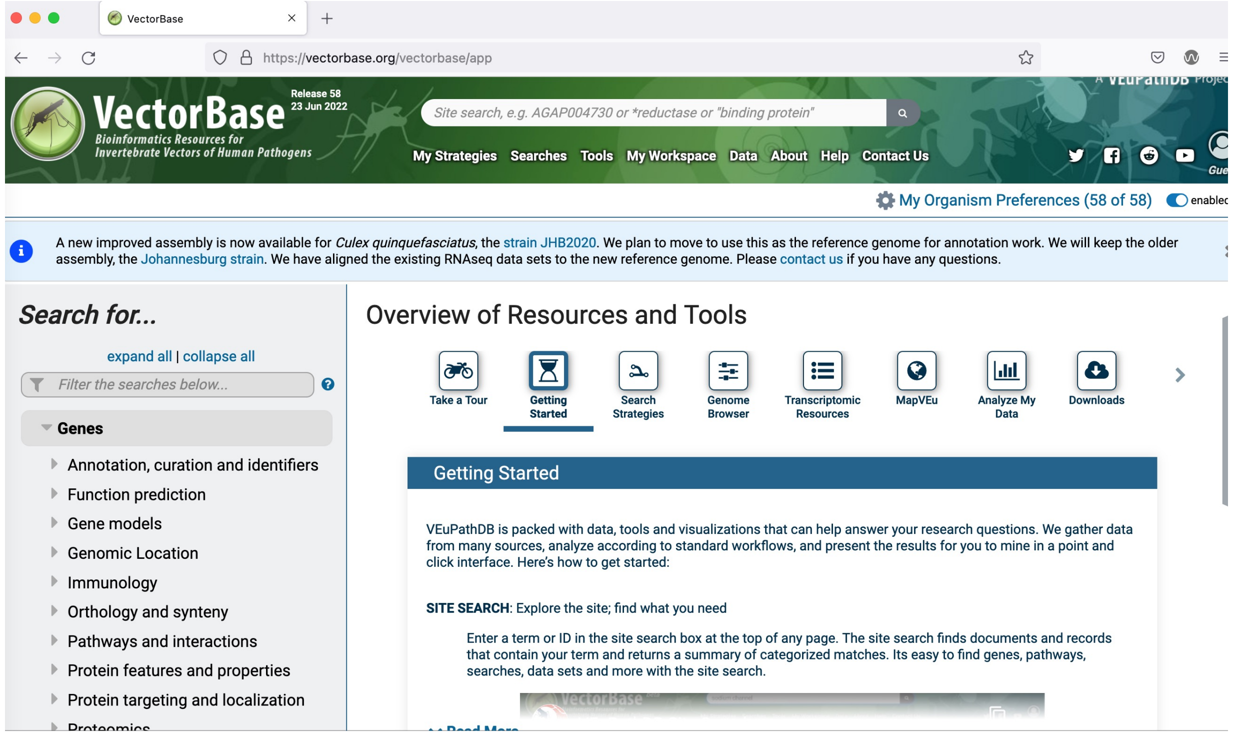
Recent activity

https://www.ncbi.nlm.nih.gov/protein/

https://www.ncbi.nlm.nih.gov/gene/

How can I find a specific gene?

Vectors


VectorBase
https://vectorbase.org/vectorbase/app
VectorBase
Release 58
23 Jun 2022
Bioinformatics Resources for Invertebrate Vectors of Human Pathogens
A VEuPathDB Project
Site search, e.g. AGAP004730 or *reductase or "binding protein"
My Strategies
Searches
Tools
My Workspace
Data
About
Help
Contact Us
My Organism Preferences (58 of 58)
enabled
Search for...
expand all | collapse all
Filter the searches below...
Genes
Annotation, curation and identifiers
Function prediction
Gene models
Genomic Location
Immunology
Orthology and synteny
Pathways and interactions
Protein features and properties
Protein targeting and localization
Proteomics
Overview of Resources and Tools
Take a Tour
Getting Started
Search Strategies
Genome Browser
Transcriptomic Resources
MapVEu
Analyze My Data
Downloads
Getting Started
Read More

A new improved assembly is now available for *Culex quinquefasciatus*, the strain JHB2020. We plan to move to use this as the reference genome for annotation work. We will keep the older assembly, the Johannesburg strain. We have aligned the existing RNAseq data sets to the new reference genome. Please contact us if you have any questions.

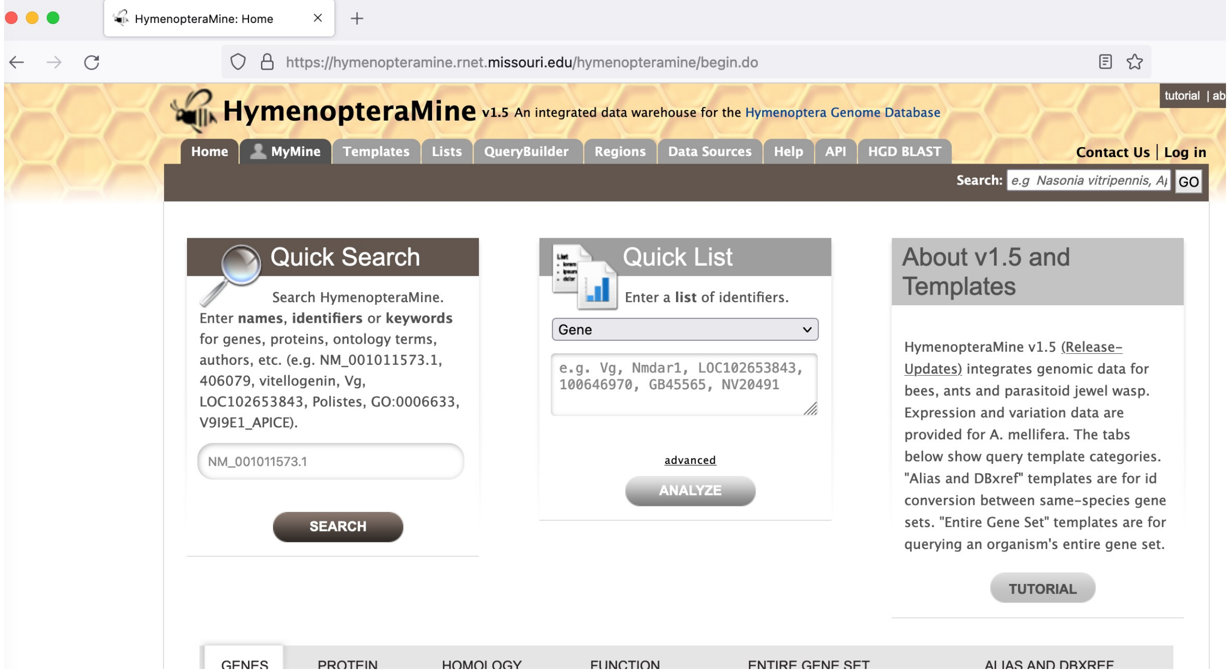
VEuPathDB is packed with data, tools and visualizations that can help answer your research questions. We gather data from many sources, analyze according to standard workflows, and present the results for you to mine in a point and click interface. Here's how to get started:

SITE SEARCH: Explore the site; find what you need

Enter a term or ID in the site search box at the top of any page. The site search finds documents and records that contain your term and returns a summary of categorized matches. Its easy to find genes, pathways, searches, data sets and more with the site search.

https://vectorbase.org/vectorbase/app

Hymenoptera



Search HymenopteraMine. Enter **names**, **identifiers** or **keywords** for genes, proteins, ontology terms, authors, etc. (e.g. NM_001011573.1, 406079, vitellogenin, Vg, LOC102653843, Polistes, GO:0006633, V9I9E1_APICE).

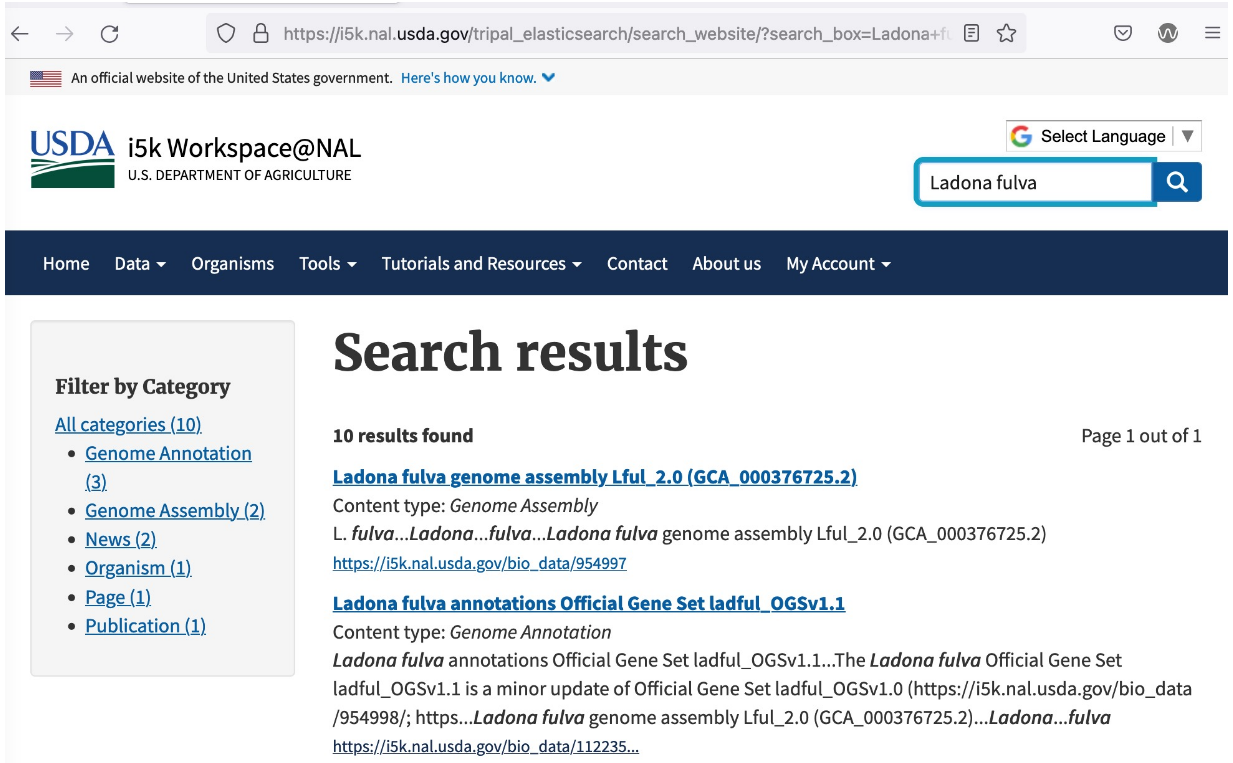
Enter a **list** of identifiers.

HymenopteraMine v1.5 (Release-Updates) integrates genomic data for bees, ants and parasitoid jewel wasp. Expression and variation data are provided for A. mellifera. The tabs below show query template categories. "Alias and DBxref" templates are for id conversion between same-species gene sets. "Entire Gene Set" templates are for querying an organism's entire gene set.

https://hymenopteramine.rnet.missouri.edu/hymenopteramine/begin.do

How can I find a gene set?

I5k Workspace@NAL search


https://i5k.nal.usda.gov/tripal_elasticsearch/search_website/?search_box=Ladona+f
An official website of the United States government. Here's how you know.
USDA i5k Workspace@NAL
U.S. DEPARTMENT OF AGRICULTURE
Select Language
Ladona fulva
Home
Data
Organisms
Tools
Tutorials and Resources
Contact
About us
My Account

Search results

Filter by Category

- All categories (10)
  - Genome Annotation (3)
  - Genome Assembly (2)
  - News (2)
  - Organism (1)
  - Page (1)
  - Publication (1)

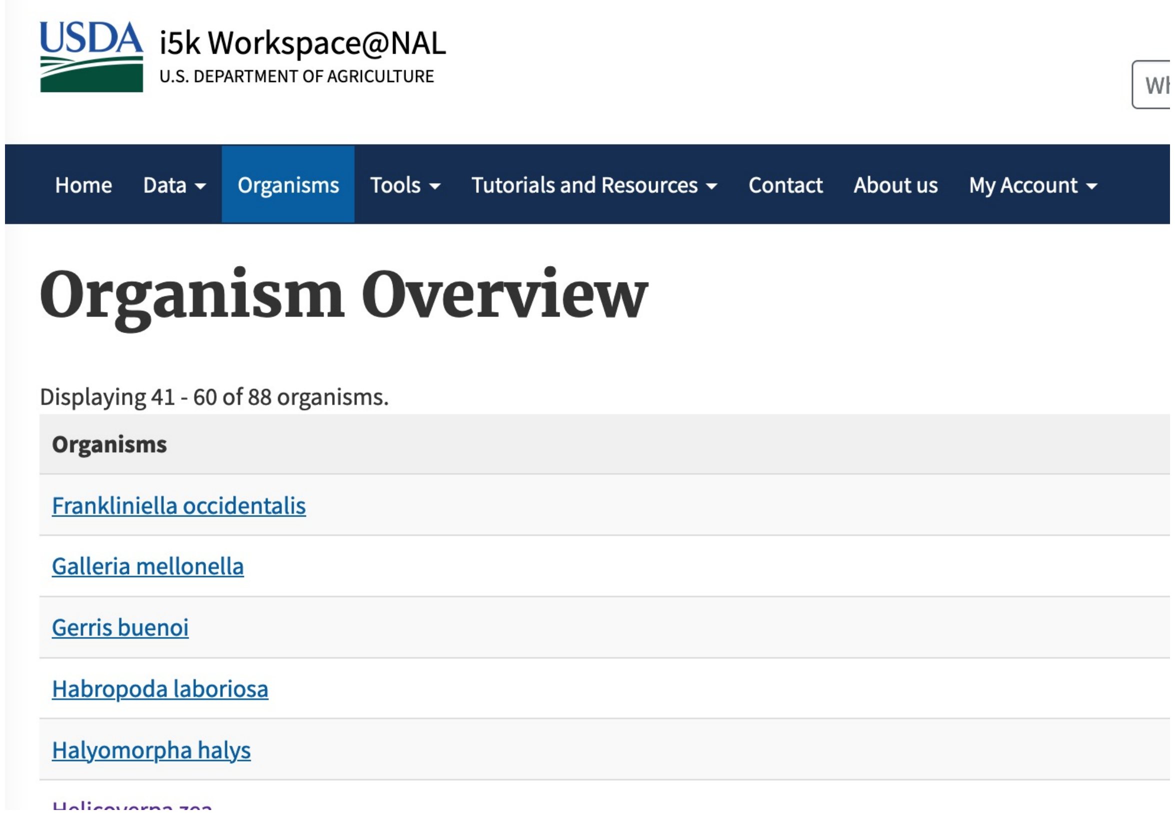
10 results found Page 1 out of 1

Ladona fulva genome assembly Lful_2.0 (GCA_000376725.2)
Content type: *Genome Assembly*
L. ***fulva***...***Ladona***...***fulva***...***Ladona fulva*** genome assembly Lful_2.0 (GCA_000376725.2)
https://i5k.nal.usda.gov/bio_data/954997

Ladona fulva annotations Official Gene Set ladful_OGSv1.1
Content type: *Genome Annotation*
Ladona fulva annotations Official Gene Set ladful_OGSv1.1...The ***Ladona fulva*** Official Gene Set ladful_OGSv1.1 is a minor update of Official Gene Set ladful_OGSv1.0 (https://i5k.nal.usda.gov/bio_data /954998/; https...***Ladona fulva*** genome assembly Lful_2.0 (GCA_000376725.2)...***Ladona***...***fulva***
https://i5k.nal.usda.gov/bio_data/112235...

https://i5k.nal.usda.gov/

I5k Workspace@NAL browse



Organism Overview

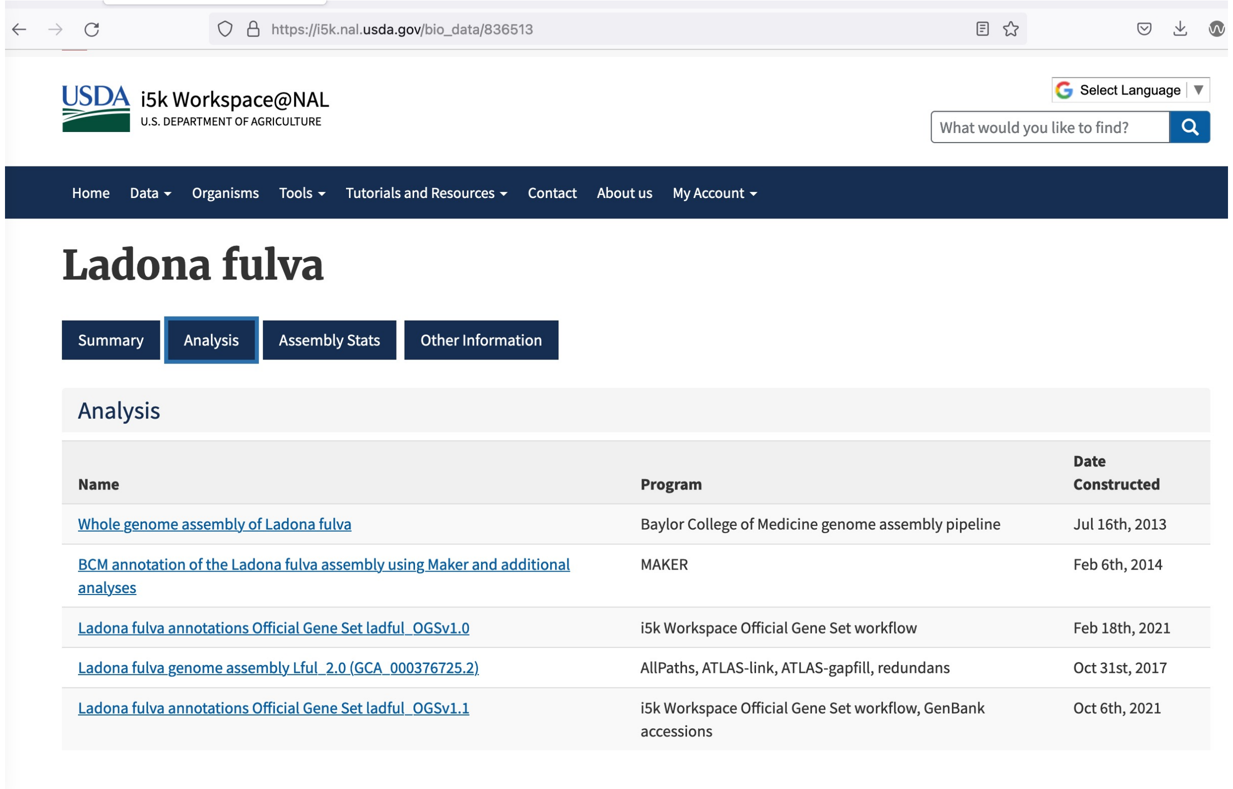
Displaying 41 - 60 of 88 organisms.

Organisms
Frankliniella occidentalis
Galleria mellonella
Gerris buenoi
Habropoda laboriosa
Halyomorpha halys
Helicoverpa zea

https://i5k.nal.usda.gov/organisms

How can I find a gene set?

I5k Workspace@NAL browse

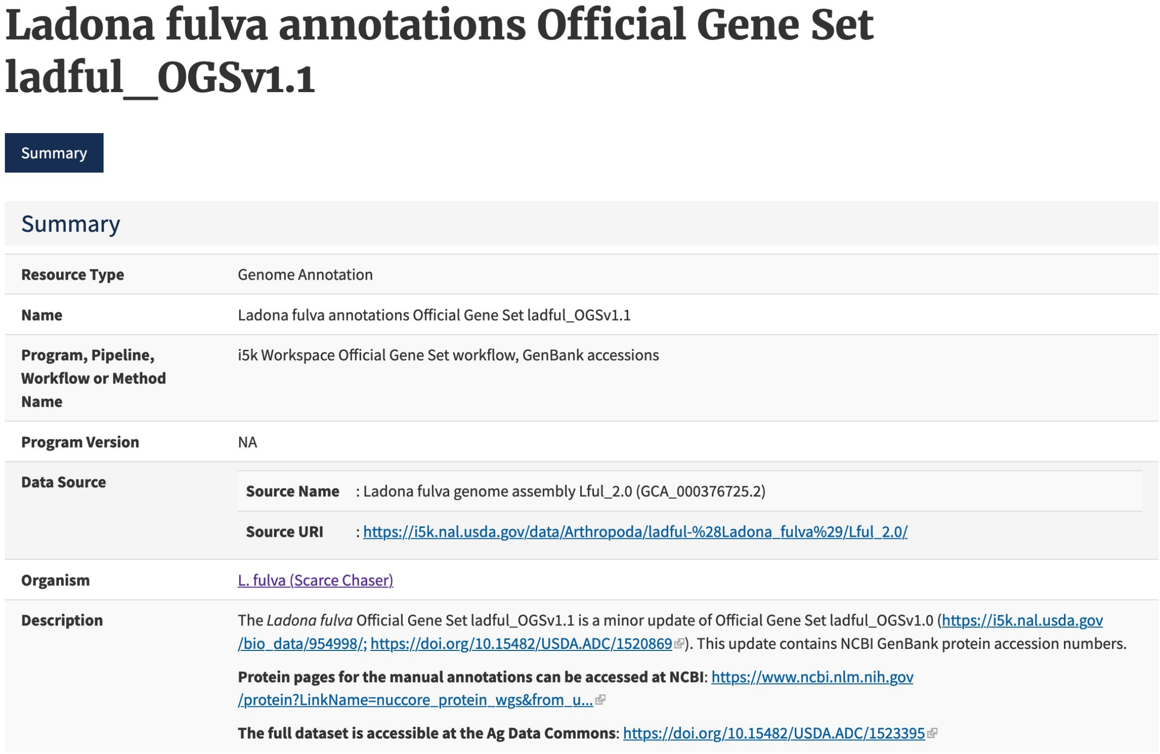


Ladona fulva

Analysis

Name	Program	Date Constructed
Whole genome assembly of Ladona fulva	Baylor College of Medicine genome assembly pipeline	Jul 16th, 2013
BCM annotation of the Ladona fulva assembly using Maker and additional analyses	MAKER	Feb 6th, 2014
Ladona fulva annotations Official Gene Set ladful_OGSv1.0	i5k Workspace Official Gene Set workflow	Feb 18th, 2021
Ladona fulva genome assembly Lful_2.0 (GCA_000376725.2)	AllPaths, ATLAS-link, ATLAS-gapfill, redundans	Oct 31st, 2017
Ladona fulva annotations Official Gene Set ladful_OGSv1.1	i5k Workspace Official Gene Set workflow, GenBank accessions	Oct 6th, 2021

I5k Workspace@NAL browse



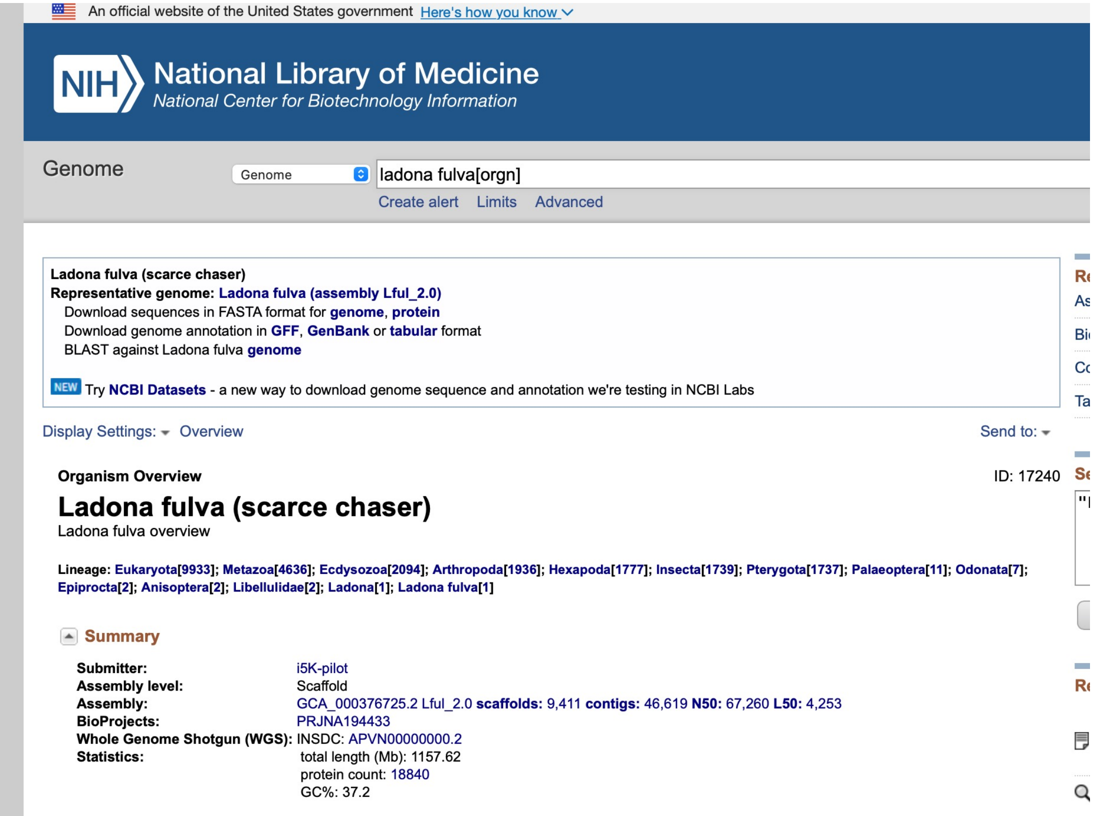
Ladona fulva annotations Official Gene Set ladful_OGSv1.1

Summary

Resource Type	Genome Annotation
Name	Ladona fulva annotations Official Gene Set ladful_OGSv1.1
Program, Pipeline, Workflow or Method Name	i5k Workspace Official Gene Set workflow, GenBank accessions
Program Version	NA
Data Source	**Source Name** : Ladona fulva genome assembly Lful_2.0 (GCA_000376725.2) **Source URI** : https://i5k.nal.usda.gov/data/Arthropoda/ladful-%28Ladona_fulva%29/Lful_2.0/
Organism	L. fulva (Scarce Chaser)
Description	The *Ladona fulva* Official Gene Set ladful_OGSv1.1 is a minor update of Official Gene Set ladful_OGSv1.0 (https://i5k.nal.usda.gov/bio_data/954998/; https://doi.org/10.15482/USDA.ADC/1520869). This update contains NCBI GenBank protein accession numbers. **Protein pages for the manual annotations can be accessed at NCBI:** https://www.ncbi.nlm.nih.gov/protein?LinkName=nuccore_protein_wgs&from_u... **The full dataset is accessible at the Ag Data Commons:** https://doi.org/10.15482/USDA.ADC/1523395

How can I find a gene set?

NCBI search



An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

Genome | Genome | ladona fulva[orgn]

Create alert Limits Advanced

Ladona fulva (scarce chaser)
Representative genome: Ladona fulva (assembly Lful_2.0)
Download sequences in FASTA format for **genome**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Ladona fulva **genome**

NEW Try **NCBI Datasets** - a new way to download genome sequence and annotation we're testing in NCBI Labs

Display Settings: Overview Send to:

Organism Overview ID: 17240

Ladona fulva (scarce chaser)
Ladona fulva overview

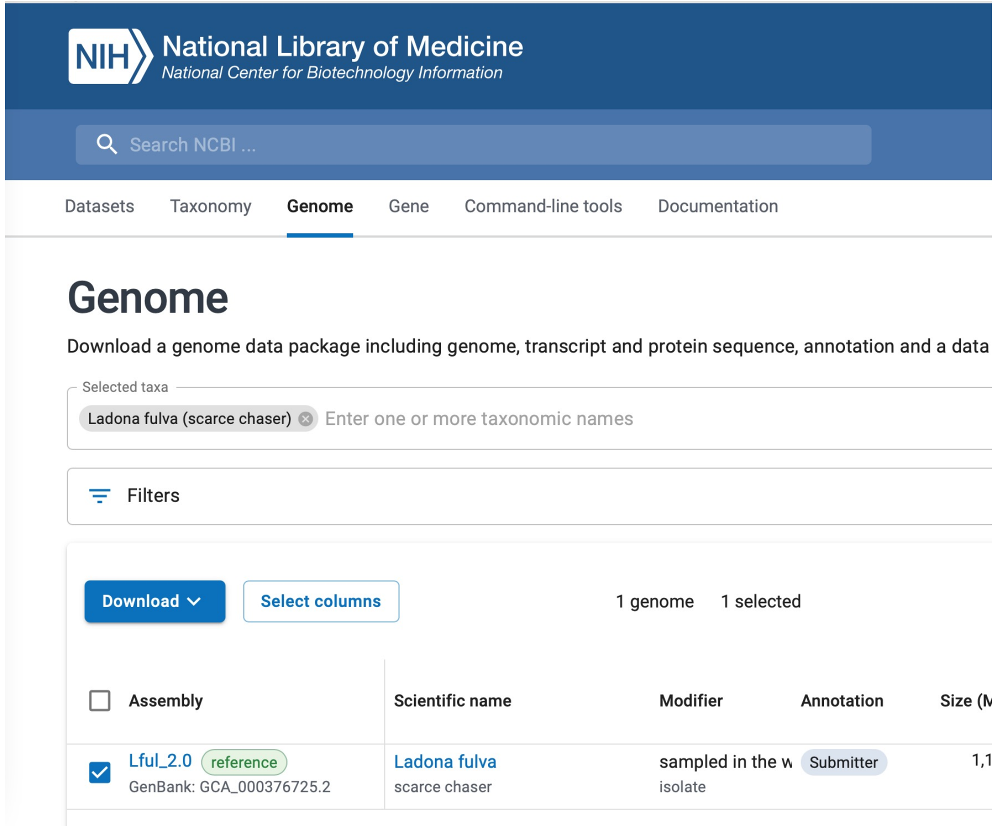
Lineage: Eukaryota[9933]; Metazoa[4636]; Ecdysozoa[2094]; Arthropoda[1936]; Hexapoda[1777]; Insecta[1739]; Pterygota[1737]; Palaeoptera[11]; Odonata[7]; Epiprocta[2]; Anisoptera[2]; Libellulidae[2]; Ladona[1]; Ladona fulva[1]

Summary

Submitter:	i5K-pilot
Assembly level:	Scaffold
Assembly:	GCA_000376725.2 Lful_2.0 **scaffolds:** 9,411 **contigs:** 46,619 **N50:** 67,260 **L50:** 4,253
BioProjects:	PRJNA194433
Whole Genome Shotgun (WGS):	INSDC: APVN00000000.2
Statistics:	total length (Mb): 1157.62 protein count: 18840 GC%: 37.2

https://www.ncbi.nlm.nih.gov/genome/

NCBI datasets



National Library of Medicine
National Center for Biotechnology Information

Search NCBI ...

Datasets Taxonomy **Genome** Gene Command-line tools Documentation

Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data

Selected taxa: Ladona fulva (scarce chaser) Enter one or more taxonomic names

Filters

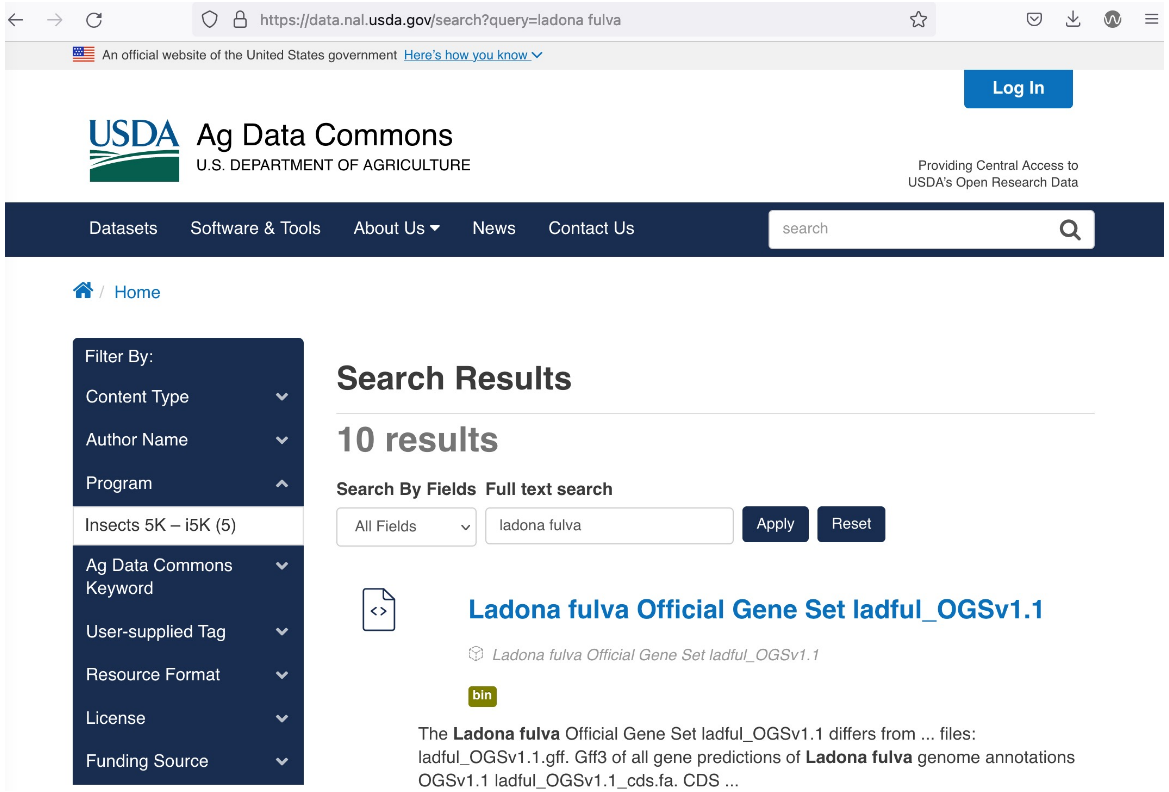
Download Select columns 1 genome 1 selected

□ Assembly	Scientific name	Modifier	Annotation	Size (M
☑ Lful_2.0 reference GenBank: GCA_000376725.2	Ladona fulva scarce chaser	sampled in the w isolate	Submitter	1,1

https://www.ncbi.nlm.nih.gov/datasets/
FYI, don't use Safari

How can I find a gene set?

Ag Data Commons search



Search Results

10 results

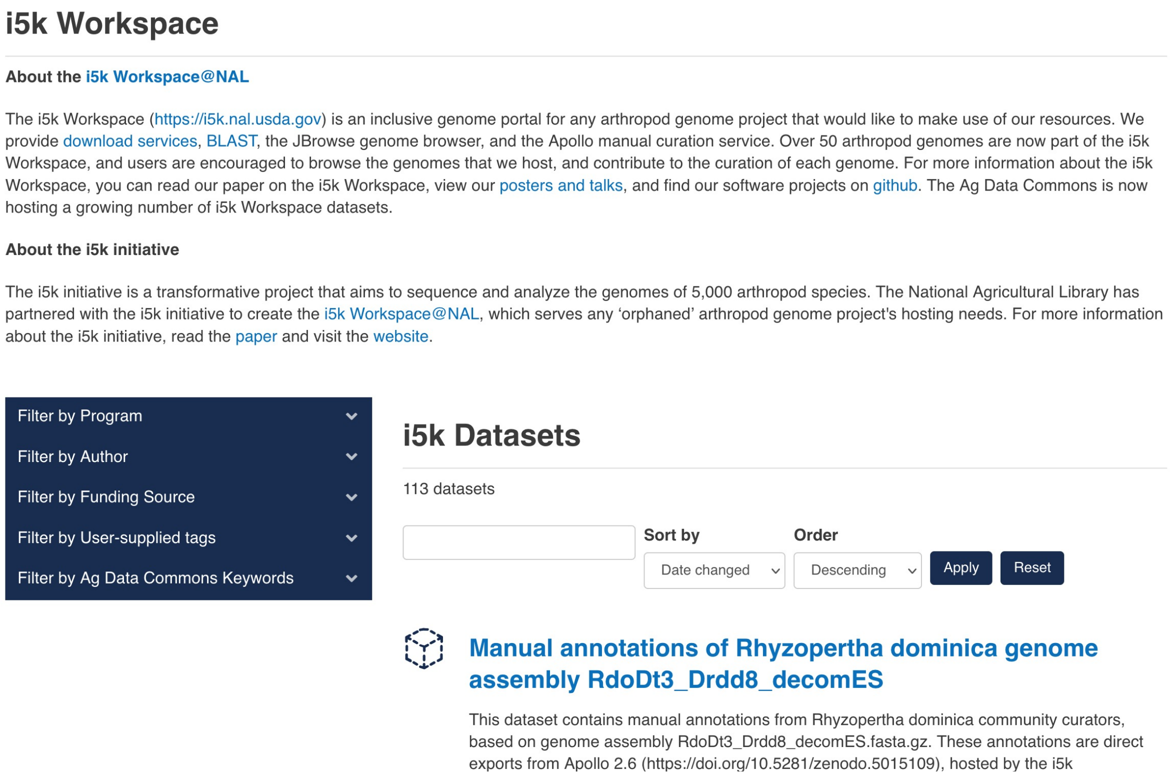
Ladona fulva Official Gene Set ladful_OGSv1.1

Ladona fulva Official Gene Set ladful_OGSv1.1

The **Ladona fulva** Official Gene Set ladful_OGSv1.1 differs from ... files: ladful_OGSv1.1.gff. Gff3 of all gene predictions of **Ladona fulva** genome annotations OGSv1.1 ladful_OGSv1.1_cds.fa. CDS ...

https://data.nal.usda.gov/

Ag Data Commons browse



i5k Workspace

About the i5k Workspace@NAL

The i5k Workspace (https://i5k.nal.usda.gov) is an inclusive genome portal for any arthropod genome project that would like to make use of our resources. We provide download services, BLAST, the JBrowse genome browser, and the Apollo manual curation service. Over 50 arthropod genomes are now part of the i5k Workspace, and users are encouraged to browse the genomes that we host, and contribute to the curation of each genome. For more information about the i5k Workspace, you can read our paper on the i5k Workspace, view our posters and talks, and find our software projects on github. The Ag Data Commons is now hosting a growing number of i5k Workspace datasets.

About the i5k initiative

The i5k initiative is a transformative project that aims to sequence and analyze the genomes of 5,000 arthropod species. The National Agricultural Library has partnered with the i5k initiative to create the i5k Workspace@NAL, which serves any 'orphaned' arthropod genome project's hosting needs. For more information about the i5k initiative, read the paper and visit the website.

i5k Datasets

113 datasets

Manual annotations of Rhyzopertha dominica genome assembly RdoDt3_Drdd8_decomES

This dataset contains manual annotations from Rhyzopertha dominica community curators, based on genome assembly RdoDt3_Drdd8_decomES.fasta.gz. These annotations are direct exports from Apollo 2.6 (https://doi.org/10.5281/zenodo.5015109), hosted by the i5k

https://data.nal.usda.gov/i5k

Questions?

Thank you!

- Contact us:
  - i5k@usda.gov
  - https://i5k.nal.usda.gov/contact

List of datasets with removed gene pages

- https://i5k.nal.usda.gov/news/i5k-workspace-removing-all-gene-and-mrna-pages