The i5k Workspace@NAL

A Tripal based Arthropod genome portal

Christopher Childers
USDA/ARS/NAL
i5k.nal.usda.gov

Background

- The i5k initiative tasked itself with coordinating the sequencing and assembly of 5000 insect or related arthropod genomes

- The i5k Workspace@NAL is available to help any i5k project with genome hosting needs
  - Support community needs
  - Connect researchers to the data
  - Create standardized tools for accessing the data in useful ways
  - Provide resources to facilitate annotation projects

The i5k Workspace@NAL

- Diverse representation across Arthropoda
  - 55 species and counting
  - 2 had OGS when they came to us
  - 8 have completed an annotation phase and have OGS
    - 3 from external collaborator
    - 5 from our workflow
    - 2 in process
  - 47 species are accepting applications for annotators
- Facilitate collaboration between groups
  - Annotators work on related taxa
  - Domain specialists that focus on gene families across multiple species

USDA
United States Department of Agriculture

'Frozen' genome assembly | Automated annotations | Ancillary datafiles (e.g. RNA-Seq alignments)

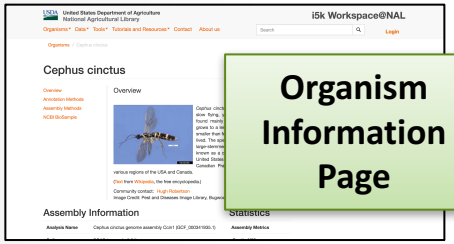
↓

Submission

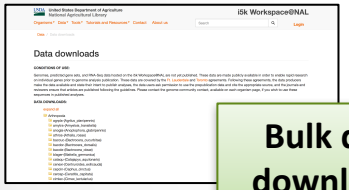
i5k Workspace@NAL

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

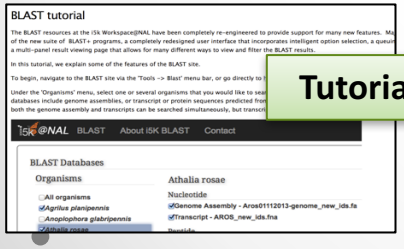
Resources



Organism Information Page

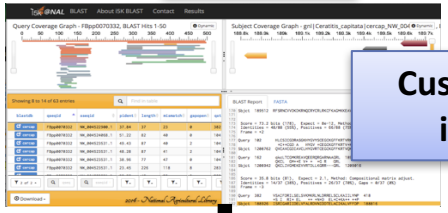


Bulk data downloads

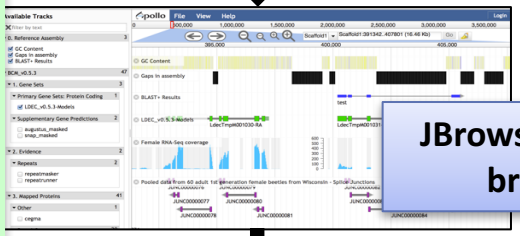


Tutorials

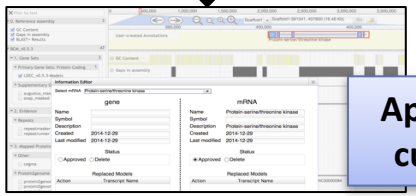
Tools



Custom BLAST interface



JBrowse genome browser



Apollo manual curation tool

HMMer

Clustal

Services

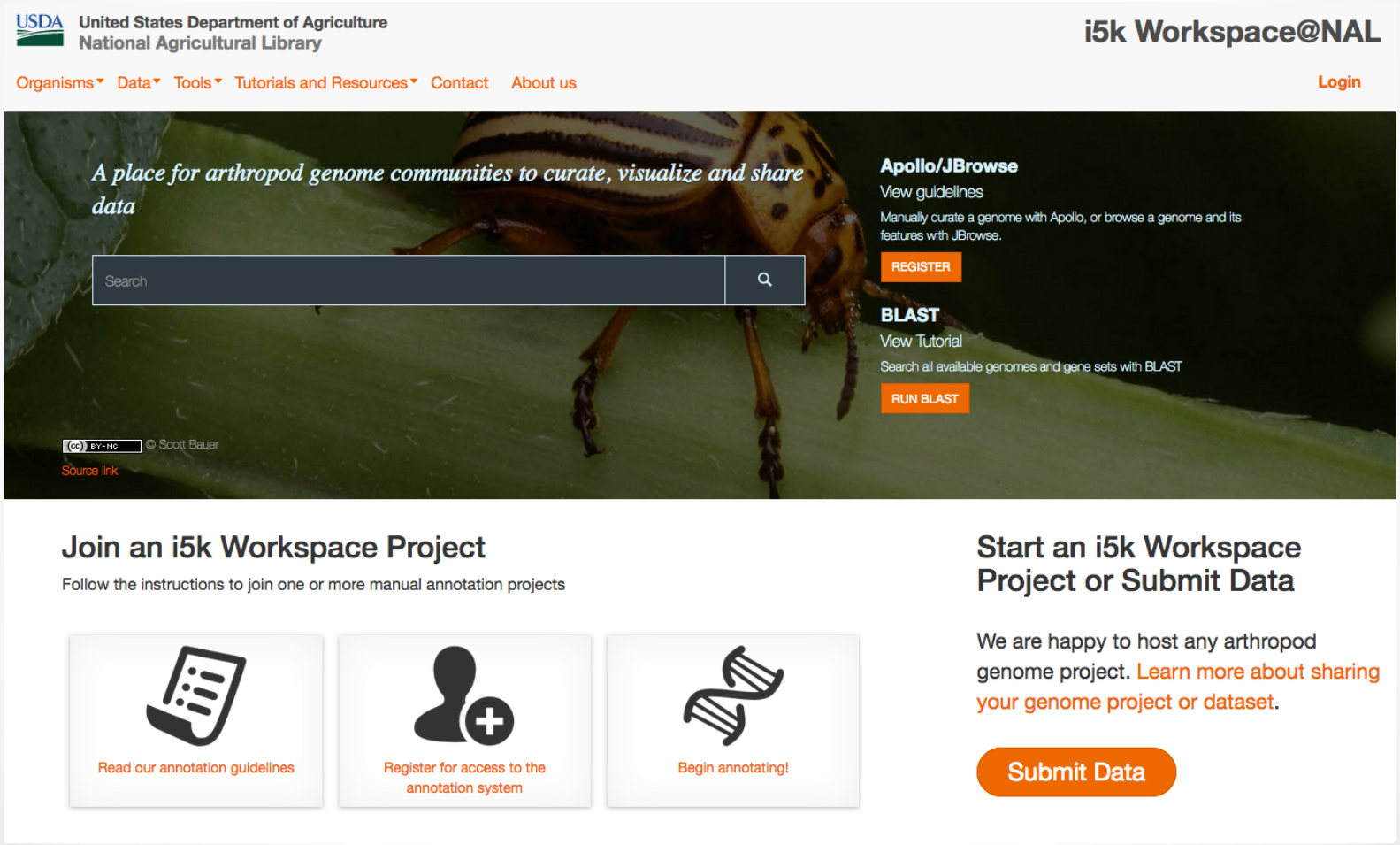
Manual annotation quality control

Official gene set generation

Challenges

Non-standard data formatting

Failure to submit all metadata (ex: sample origin; analysis methods)


USDA United States Department of Agriculture
National Agricultural Library
i5k Workspace@NAL
Organisms Data Tools Tutorials and Resources Contact About us
Login
A place for arthropod genome communities to curate, visualize and share data
Search
Apollo/JBrowse
View guidelines
Manually curate a genome with Apollo, or browse a genome and its features with JBrowse.
REGISTER
BLAST
View Tutorial
Search all available genomes and gene sets with BLAST
RUN BLAST
© Scott Bauer
Source link
Join an i5k Workspace Project
Follow the instructions to join one or more manual annotation projects
Read our annotation guidelines
Register for access to the annotation system
Begin annotating!
Start an i5k Workspace Project or Submit Data
We are happy to host any arthropod genome project. Learn more about sharing your genome project or dataset.
Submit Data

Organism Pages

USDA United States Department of Agriculture
National Agricultural Library

i5k Workspace@NAL

Organisms Data Tools Tutorials and Resources Contact About us

Search

Login

Organisms / Catajapyx aquilonaris

Catajapyx aquilonaris

Overview
Annotation Methods
Assembly Methods
Catajapyx aquilonaris @ Baylor College of Medicine
NCBI BioProject

Overview

The japygid *Catajapyx aquilonaris* is a blind predator of the soil. Like Protura (*Acerentomon maius*) and Collembola (*Sminthurus viridis*), Diplura lack wings, mirroring the wingless insect ancestor. Like in all primarily wingless hexapods, sperms are not transferred directly during copulation. Males rather deposit a spermatophore on the ground and females subsequently take the spermatophore up.

Diplura are critical for understanding the evolutionary origin of Hexapoda (e.g., terrestrialization), the evolutionary origin of wings (ancestral condition in Diplura), and the evolution of direct sperm transfer (ancestral condition in Diplura).

Data were generated by the Baylor College of Medicine's i5k pilot project.

View the Baylor College of Medicine's data sharing policy.

Image Credit: Copyright Nikola Szucsich

Catajapyx aquilonaris data files

	Name	Last modified
↩	Parent Directory	
📁	Current Genome Assembly	2015-03-19 13:42

Assembly Information

Statistics

Organism Pages

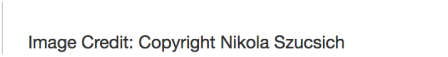


Image Credit: Copyright Nikola Szucsich

Assembly Information

Analysis Name	Whole genome assembly of Catajapyx aquilonaris
Software	Baylor College of Medicine genome assembly pipeline (NA)
Source	forcepstail.consistent.scaffolds
Date performed	2014-10-07
Materials & Methods	Sequence generation for assembly. For this project we are generating fairly high coverage in a number of different insert sized libraries. The assembly strategy is based around a seed allpaths assembly (the Broad Allpaths assembler) followed by seed assembly improvement using homegrown tools, Atlas-link and Atlas-GapFill, which can significantly improve the results. Thus we generate sequence data to enable the Allpaths assembly. As of Nov 2011 this is: - 40X genome coverage in 180bp insert library (100bp reads forward and reverse); and 40X 3kb insert data. To enable better scaffolding and local gap filling we additionally generate 500bp, 1kb, 2kb, and 8kb insert sizes at > 20X coverage.

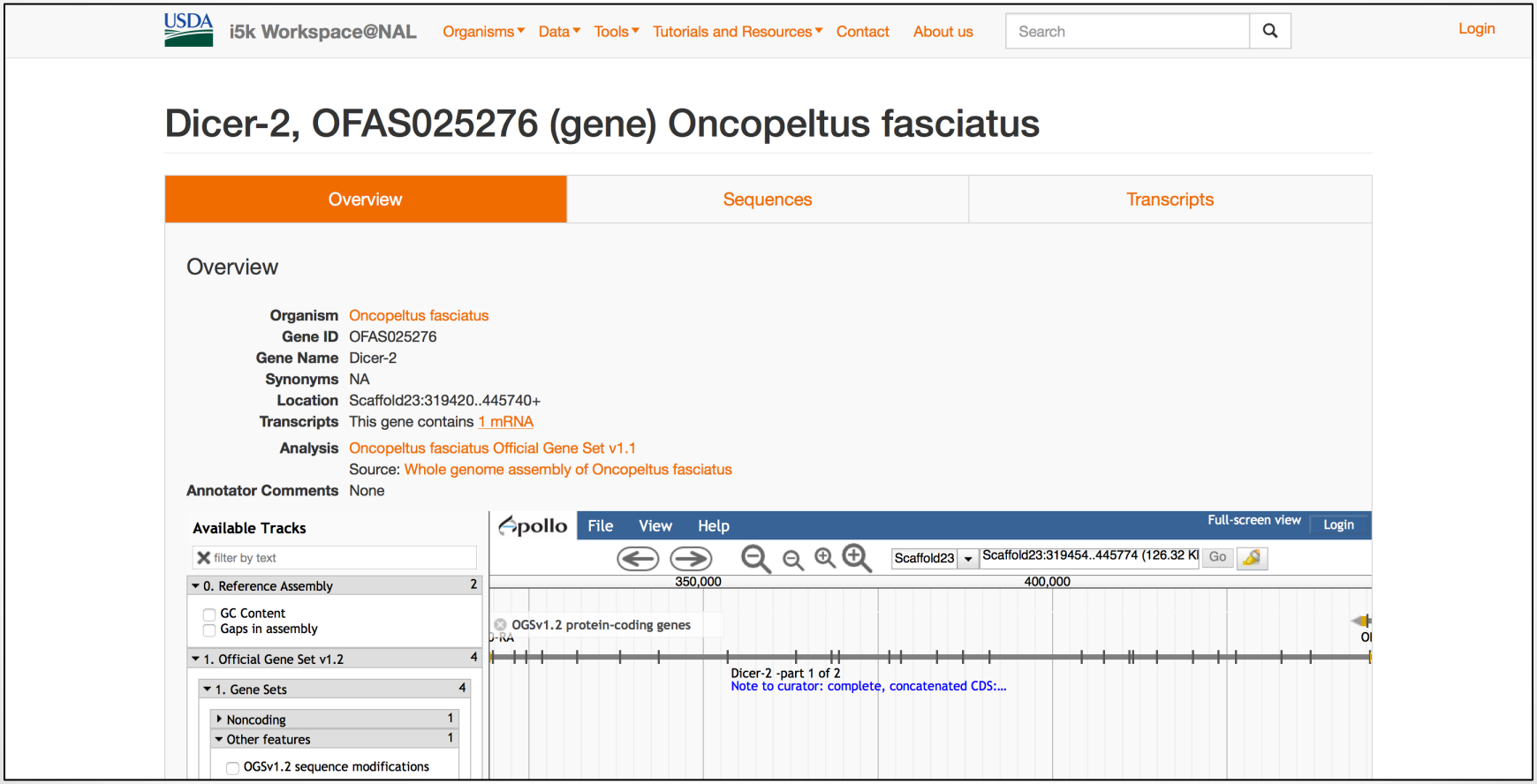
Source: Baylor College of Medicine i5K Project Summary

Statistics

Assembly Metrics	
Contig N50	11472
Scaffold N50	30909
GC Content	44.41
Manual Annotations	

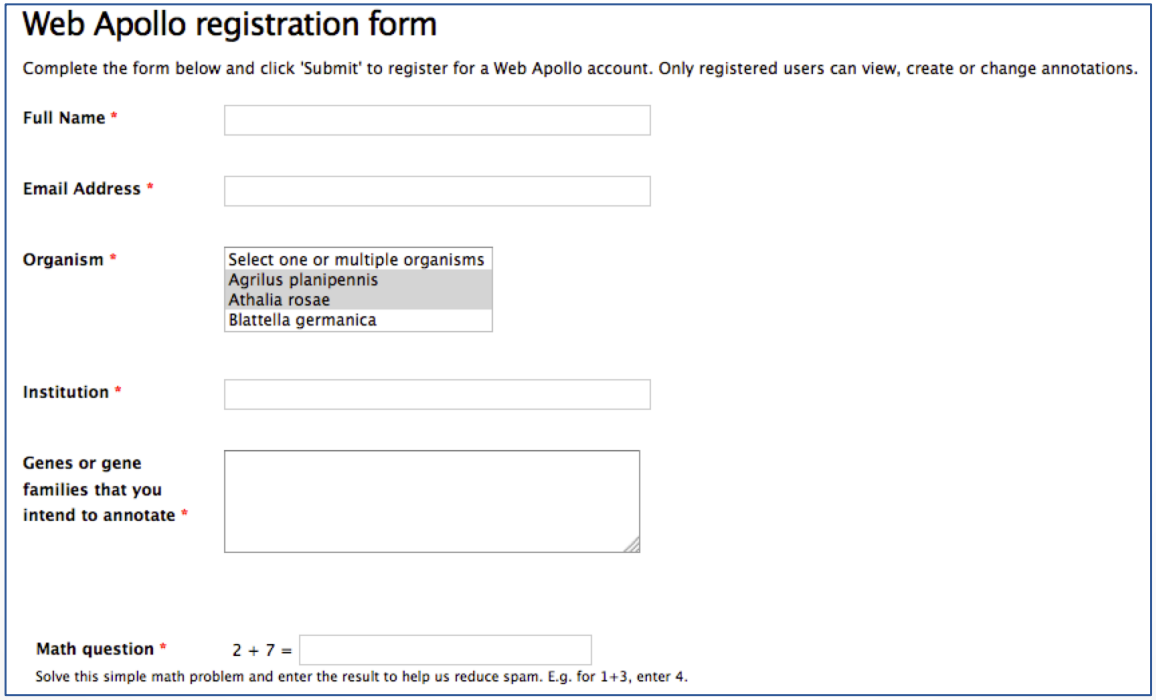
Please cite the use of our resources: doi: 10.1093/nar/gku983

Gene Pages


USDA
i5k Workspace@NAL
Organisms
Data
Tools
Tutorials and Resources
Contact
About us
Search
Login
Dicer-2, OFAS025276 (gene) Oncopeltus fasciatus
Overview
Sequences
Transcripts
Overview
Organism Oncopeltus fasciatus
Gene ID OFAS025276
Gene Name Dicer-2
Synonyms NA
Location Scaffold23:319420..445740+
Transcripts This gene contains 1 mRNA
Analysis Oncopeltus fasciatus Official Gene Set v1.1
Source: Whole genome assembly of Oncopeltus fasciatus
Annotator Comments None
Available Tracks
filter by text
0. Reference Assembly 2
GC Content
Gaps in assembly
1. Official Gene Set v1.2 4
1. Gene Sets 4
Noncoding 1
Other features 1
OGSv1.2 sequence modifications
Apollo
File
View
Help
Full-screen view
Login
Scaffold23
Scaffold23:319454..445774 (126.32 Kl
Go
350,000
400,000
OGSv1.2 protein-coding genes
Dicer-2 -part 1 of 2
Note to curator: complete, concatenated CDS:...

Annotator Registration

- Registration through a custom in-house module
- All registrants are reviewed before approval
- Permissions are at the organism level



Web Apollo registration form

Complete the form below and click 'Submit' to register for a Web Apollo account. Only registered users can view, create or change annotations.

Full Name *

Email Address *

Organism * — Select one or multiple organisms: Agrilus planipennis, Athalia rosae, Blattella germanica

Institution *

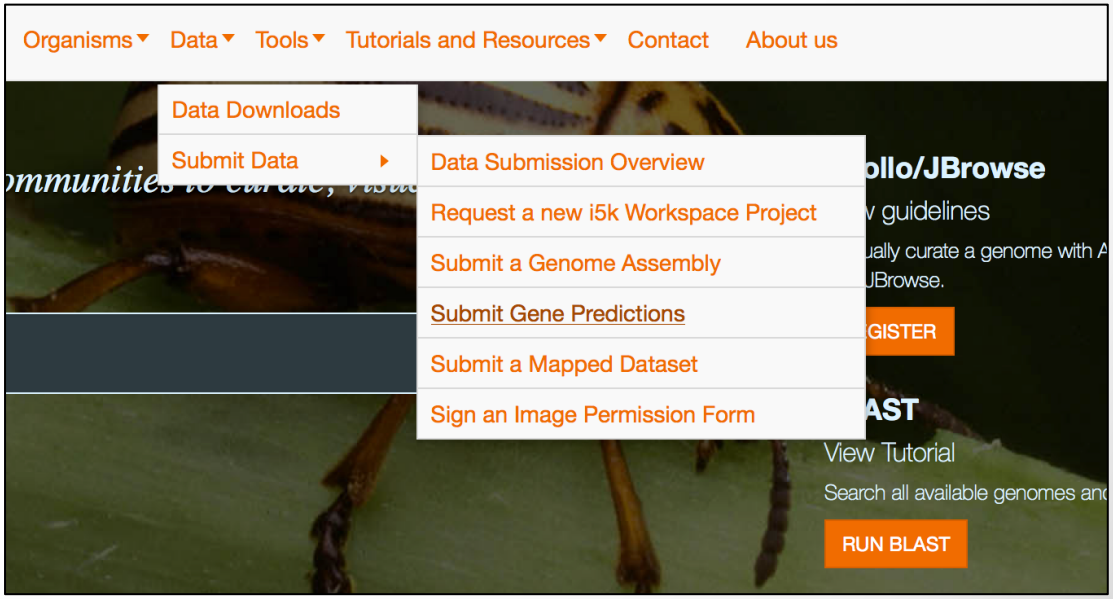
Genes or gene families that you intend to annotate *

Math question * 2 + 7 =

Solve this simple math problem and enter the result to help us reduce spam. E.g. for 1+3, enter 4.

Data Submission

- Dynamic web forms
  - Standardized and validated metadata inputs
- Replaces the spreadsheet submission system
- Under the new system:
  - 4 new species
  - 4 assemblies
  - 4 gene prediction sets
  - 2 mapped datasets



USDA
United States Department of Agriculture

'Frozen' genome assembly | Automated annotations | Ancillary datafiles (e.g. RNA-Seq alignments)

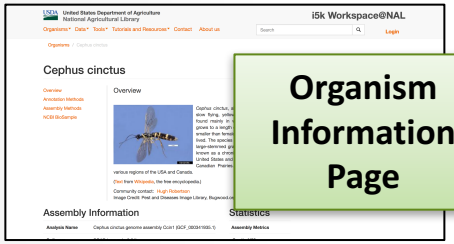
↓

Submission

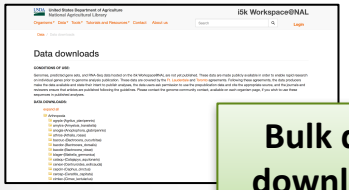
i5k Workspace@NAL

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

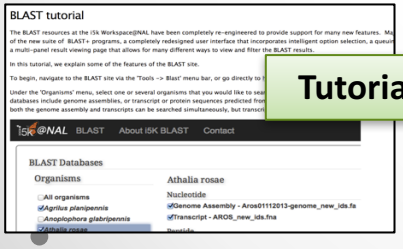
Resources



Organism Information Page

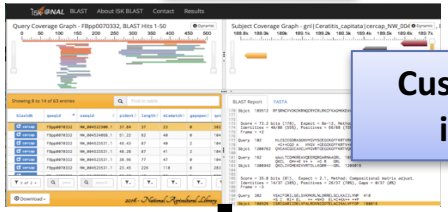


Bulk data downloads

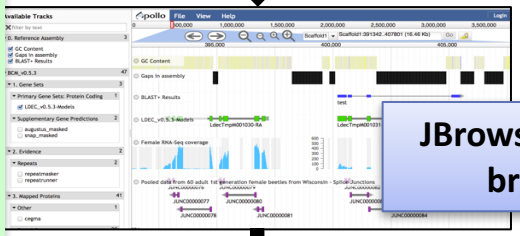


Tutorials

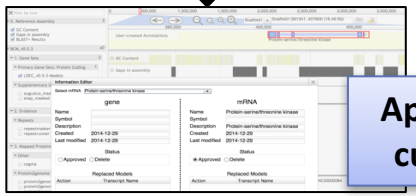
Tools



Custom BLAST interface



JBrowse genome browser



Apollo manual curation tool

HMMer

Clustal

Services

Manual annotation quality control

Official gene set generation

Challenges

Non-standard data formatting

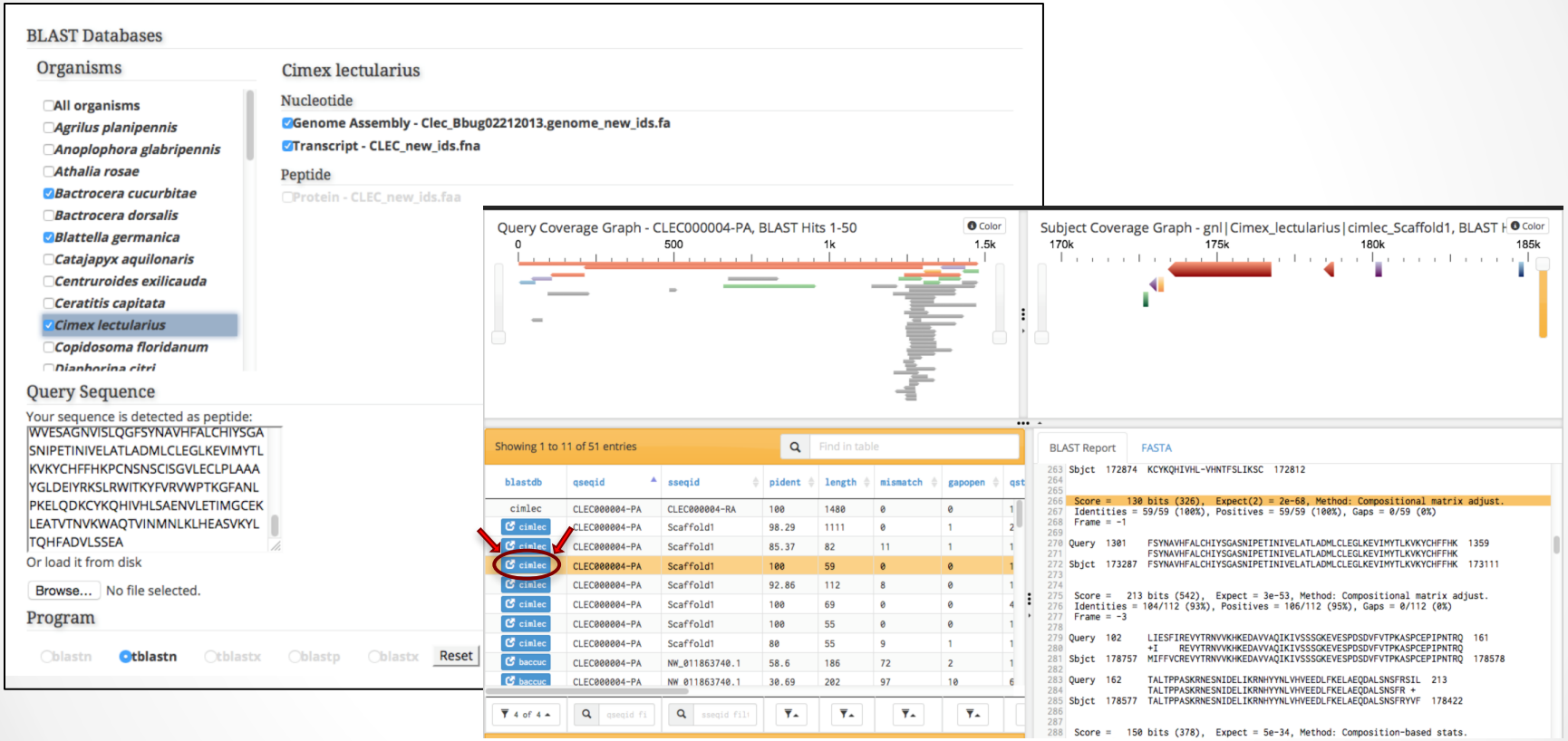
Failure to submit all metadata (ex: sample origin; analysis methods)

Search/Alignment tools

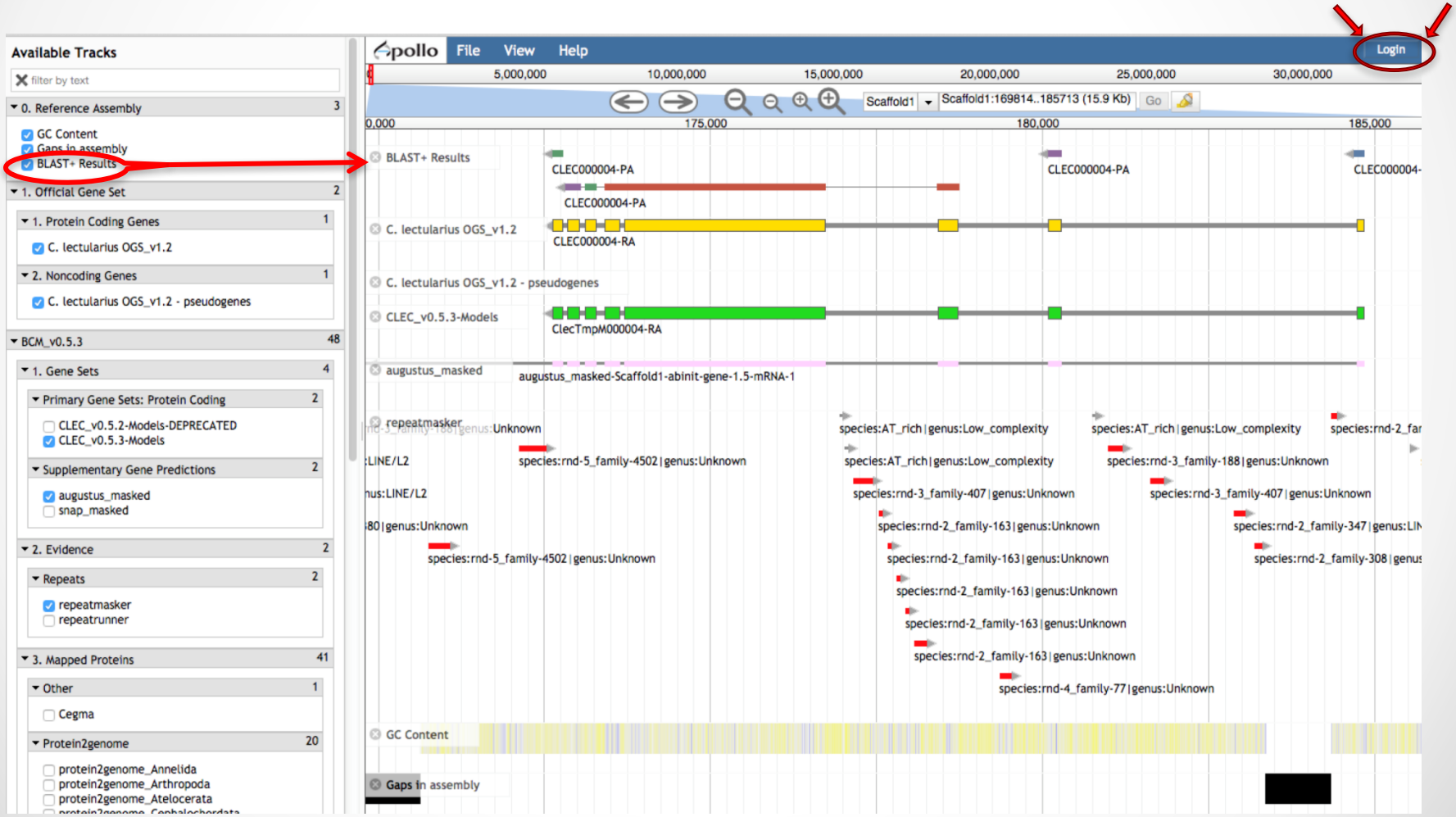
- Web interface is built in Python/Django
  - Interface dynamically updates options based on inputs
  - Results persist for one week
- Clustal
  - Multiple sequence alignment package
  - We provide both Clustalw and Clustal-Omega
  - Results may be passed directly to HMMer
- HMMer
  - Sequence search tool more sensitive than BLAST
  - Uses probabilistic models to find distant homologs
  - Can accept single sequences (FASTA) or Multiple alignments (MSA)
  - Available for use with our protein sets
- BLAST
  - Sequence alignment tool
  - Search across genome, CDS or peptide databases
  - Submit multiple sequences at once
  - Search multiple databases with one submission

United States Department of Agriculture

BLAST



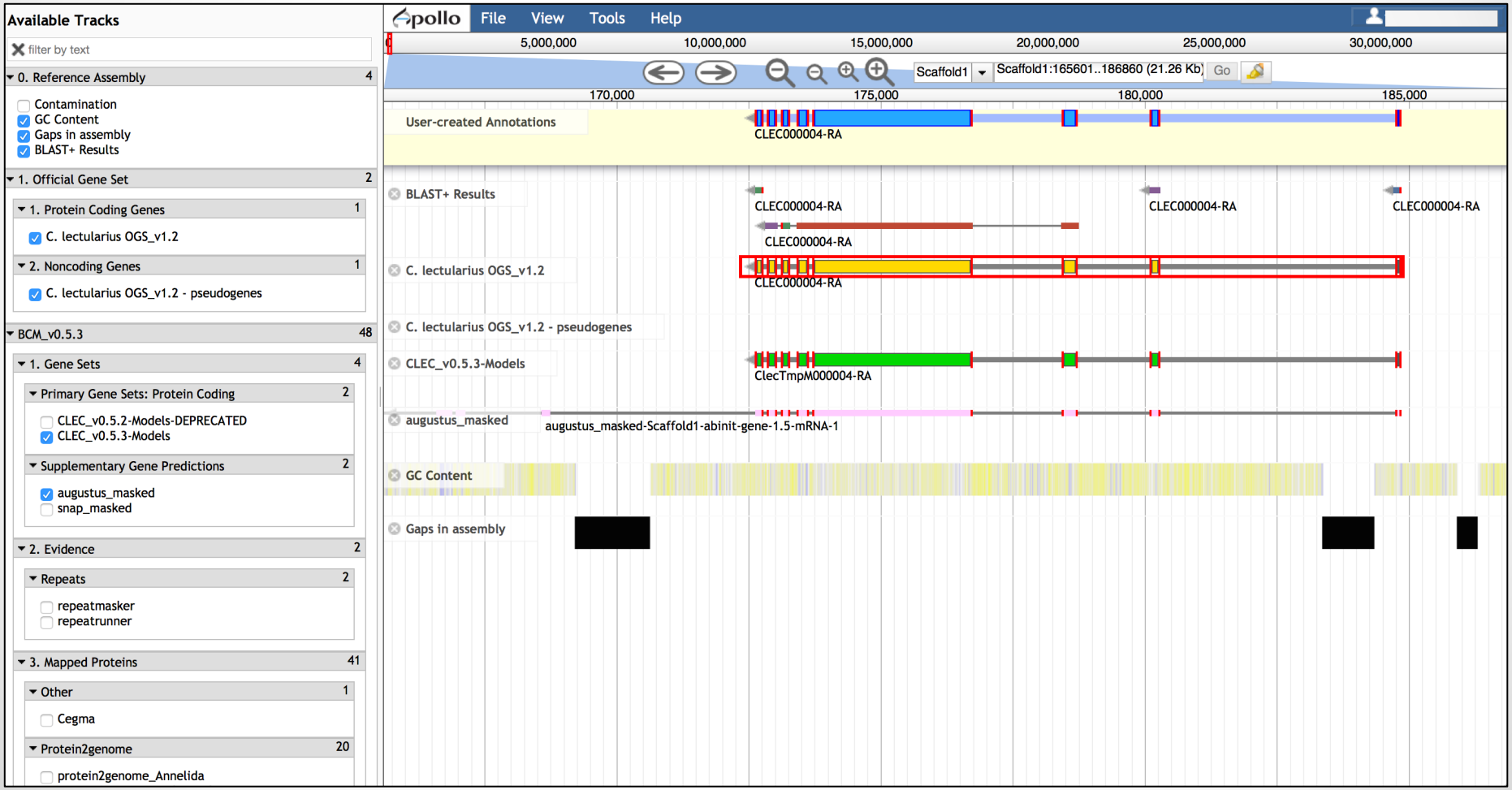
Feature Visualization



USDA

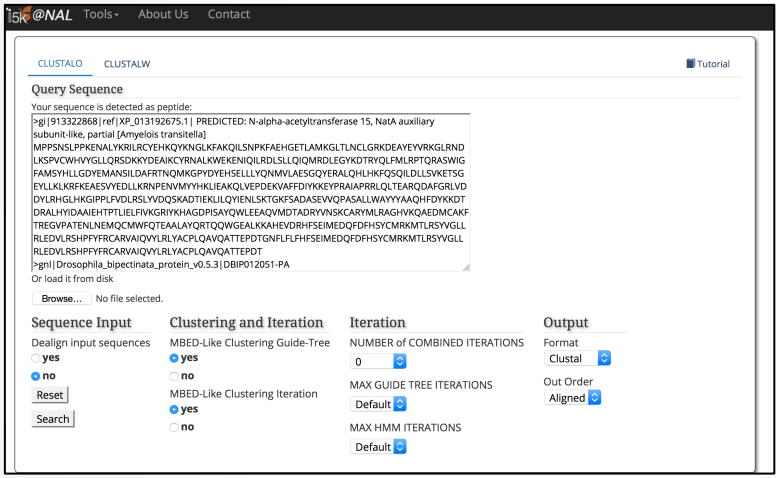
United States Department of Agriculture

Annotation via Apollo



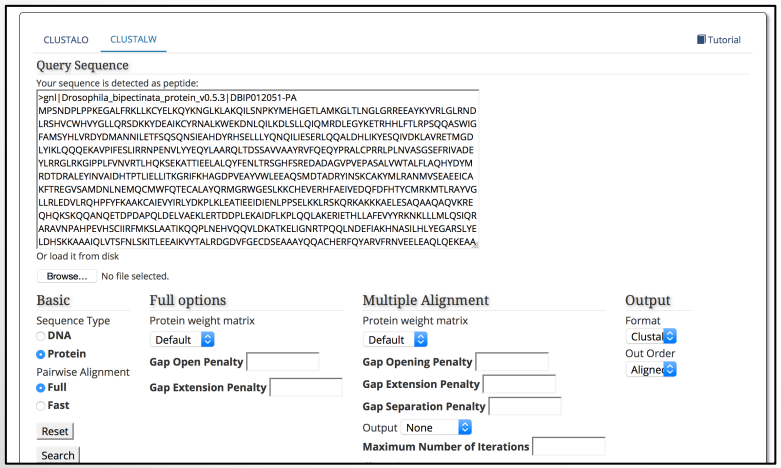
Clustal

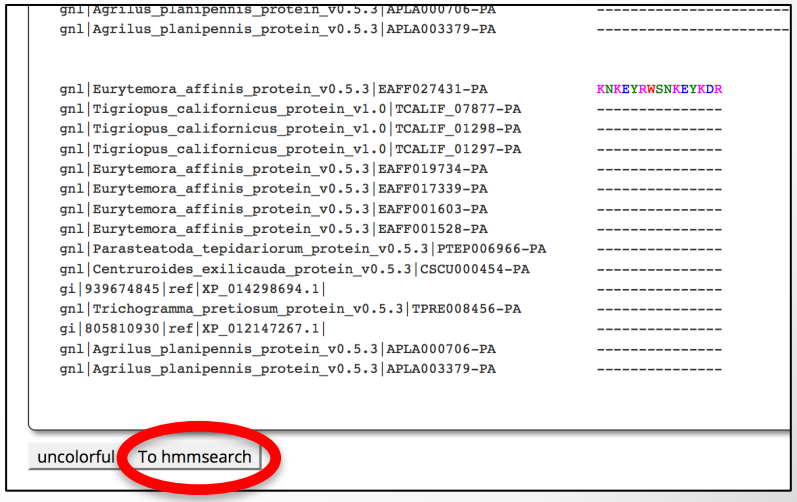
Clustal Omega



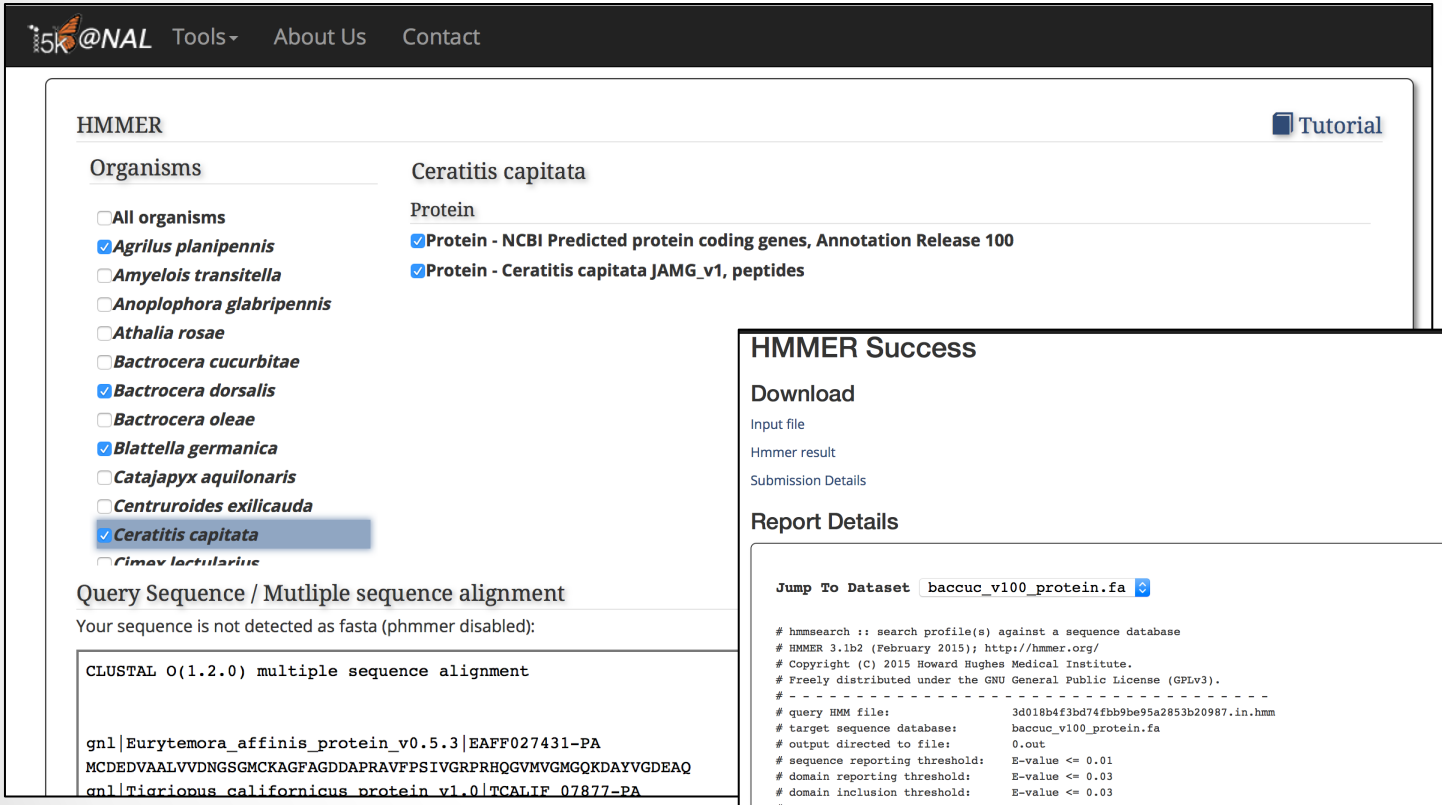
OR

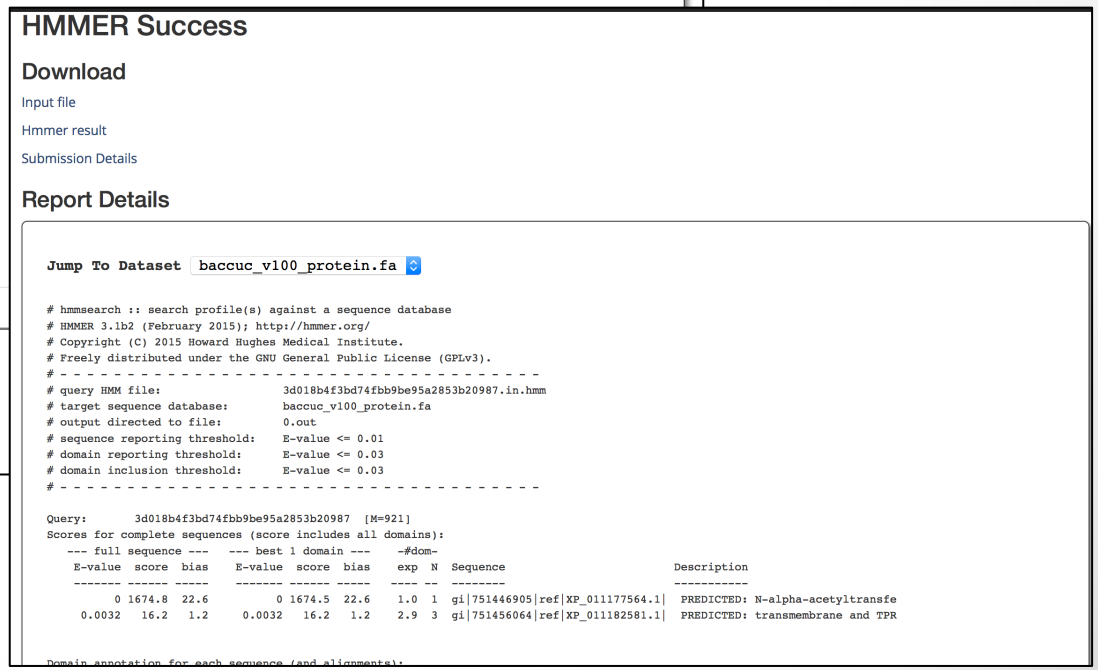
ClustalW





HMMer


i5k @NAL
Tools
About Us
Contact
HMMER
Tutorial
Organisms
All organisms
Agrilus planipennis
Amyelois transitella
Anoplophora glabripennis
Athalia rosae
Bactrocera cucurbitae
Bactrocera dorsalis
Bactrocera oleae
Blattella germanica
Catajapyx aquilonaris
Centruroides exilicauda
Ceratitis capitata
Ceratitis capitata
Protein
Protein - NCBI Predicted protein coding genes, Annotation Release 100
Protein - Ceratitis capitata JAMG_v1, peptides
Query Sequence / Mutliple sequence alignment
Your sequence is not detected as fasta (phmmer disabled):
CLUSTAL O(1.2.0) multiple sequence alignment
gnl|Eurytemora_affinis_protein_v0.5.3|EAFF027431-PA
MCDEDVAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDAYVGDEAQ
gnl|Tigriopus_californicus_protein_v1.0|TCALIF_07877-PA


HMMER Success
Download
Input file
Hmmer result
Submission Details
Report Details
Jump To Dataset baccuc_v100_protein.fa
hmmsearch :: search profile(s) against a sequence database
HMMER 3.1b2 (February 2015); http://hmmer.org/
Copyright (C) 2015 Howard Hughes Medical Institute.
Freely distributed under the GNU General Public License (GPLv3).
query HMM file: 3d018b4f3bd74fbb9be95a2853b20987.in.hmm
target sequence database: baccuc_v100_protein.fa
output directed to file: 0.out
sequence reporting threshold: E-value <= 0.01
domain reporting threshold: E-value <= 0.03
domain inclusion threshold: E-value <= 0.03
Query: 3d018b4f3bd74fbb9be95a2853b20987 [M=921]
Scores for complete sequences (score includes all domains):
E-value score bias E-value score bias exp N Sequence Description
0 1674.8 22.6 0 1674.5 22.6 1.0 1 gi|751446905|ref|XP_011177564.1| PREDICTED: N-alpha-acetyltransfe
0.0032 16.2 1.2 0.0032 16.2 1.2 2.9 3 gi|751456064|ref|XP_011182581.1| PREDICTED: transmembrane and TPR

USDA
United States Department of Agriculture

'Frozen' genome assembly | Automated annotations | Ancillary datafiles (e.g. RNA-Seq alignments)

↓

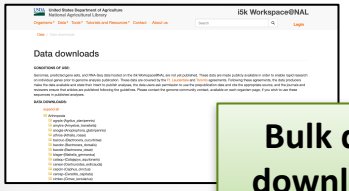
Submission

i5k Workspace@NAL

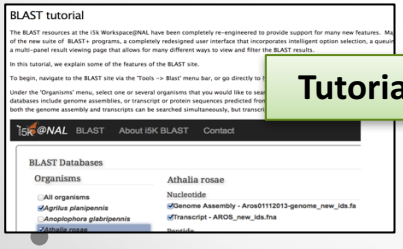
https://i5k.nal.usda.gov/

Resources

Organism Information Page

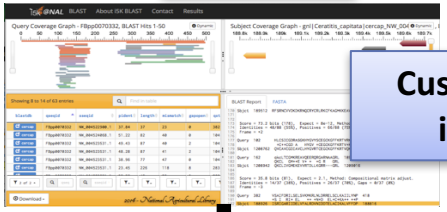


Bulk data downloads

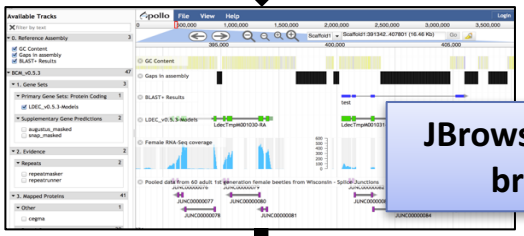


Tutorials

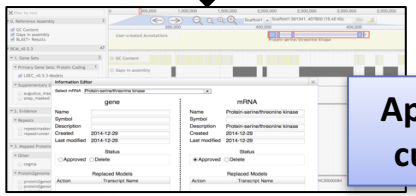
Tools



Custom BLAST interface



JBrowse genome browser



Apollo manual curation tool

HMMer

Clustal

Services

Manual annotation quality control

Official gene set generation

Challenges

Non-standard data formatting

Failure to submit all metadata (ex: sample origin; analysis methods)

Post-Annotation QC

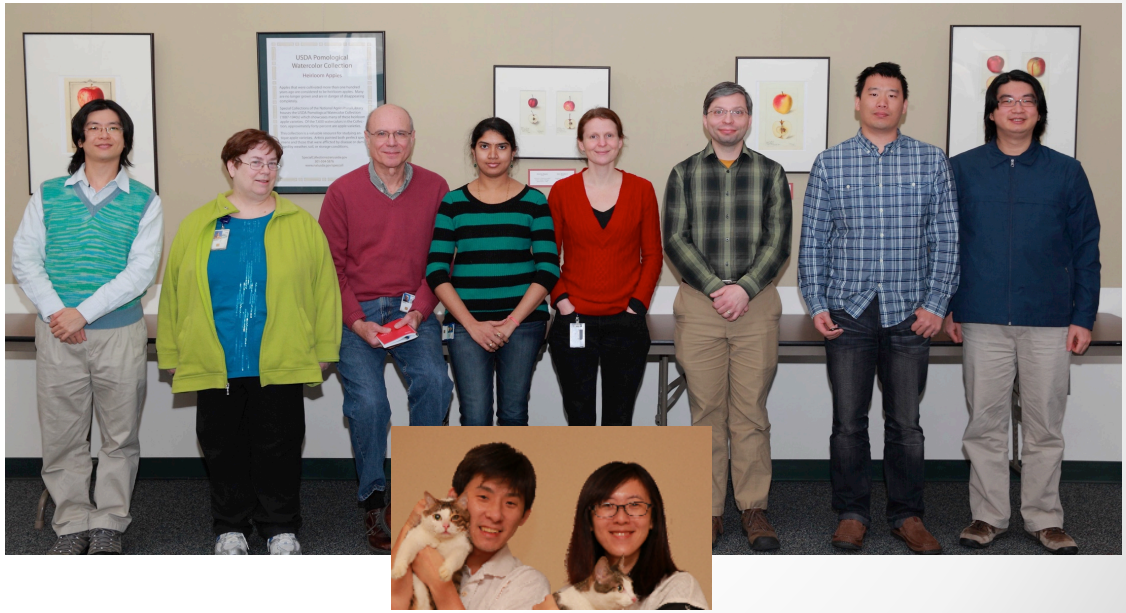
- Manual annotations are run through our Quality Control pipeline
- Some issues need manual intervention
  - Missing required fields
  - Complex splits/merges
  - Incomplete models and those abandoned in process
- Some issues can be automatically corrected
- Iterative process
  - Models requiring inspection are referred back to curators
  - After resolution models are screened again to screen for additional issues

OGS Generation

- An Official Gene Set is the gene set chosen by the community to be the representative set of gene models for that organism
- Our system takes a single existing gene set and incorporates the validated manual annotations
- The gene set may be a previous OGS or other gene set (e.g. Maker models)
- Manual curations are used to
  - Update models
  - Flag models for removal from the final set
- The resulting set is then tested for errors and once approved, disseminated to the community

Acknowledgements

- **The NAL team**
  - Monica Poelchau
  - Gary Moore
  - Susan McCarthy
  - Mei-Ju Chen
  - Yu-Yu (Fish) Lin
  - Limei Chiang
  - Chao-I Tuan
  - Chiatanya Gutta
- **i5k Workspace@NAL advisory committee**
  - Jay Evans
  - Don Gourley
  - Kevin Hackett
  - Simon Liu
  - Ursula Pieper
  - Paul Wester
- Team Alumni
  - Chien-Yueh Lee
  - Han Lin
  - Jun-Wei Lin
  - Vijaya Tsavatapalli



For More Information

Visit us!

http://i5k.nal.usda.gov

Contact us!

i5k@ars.usda.gov

Check out our code on GitHub!

https://github.com/NAL-i5K/