The i5k Workspace@NAL AGS 2018 Workshop

Christopher Childers and Monica Poelchau

June 7, 2018

USDA

Agenda

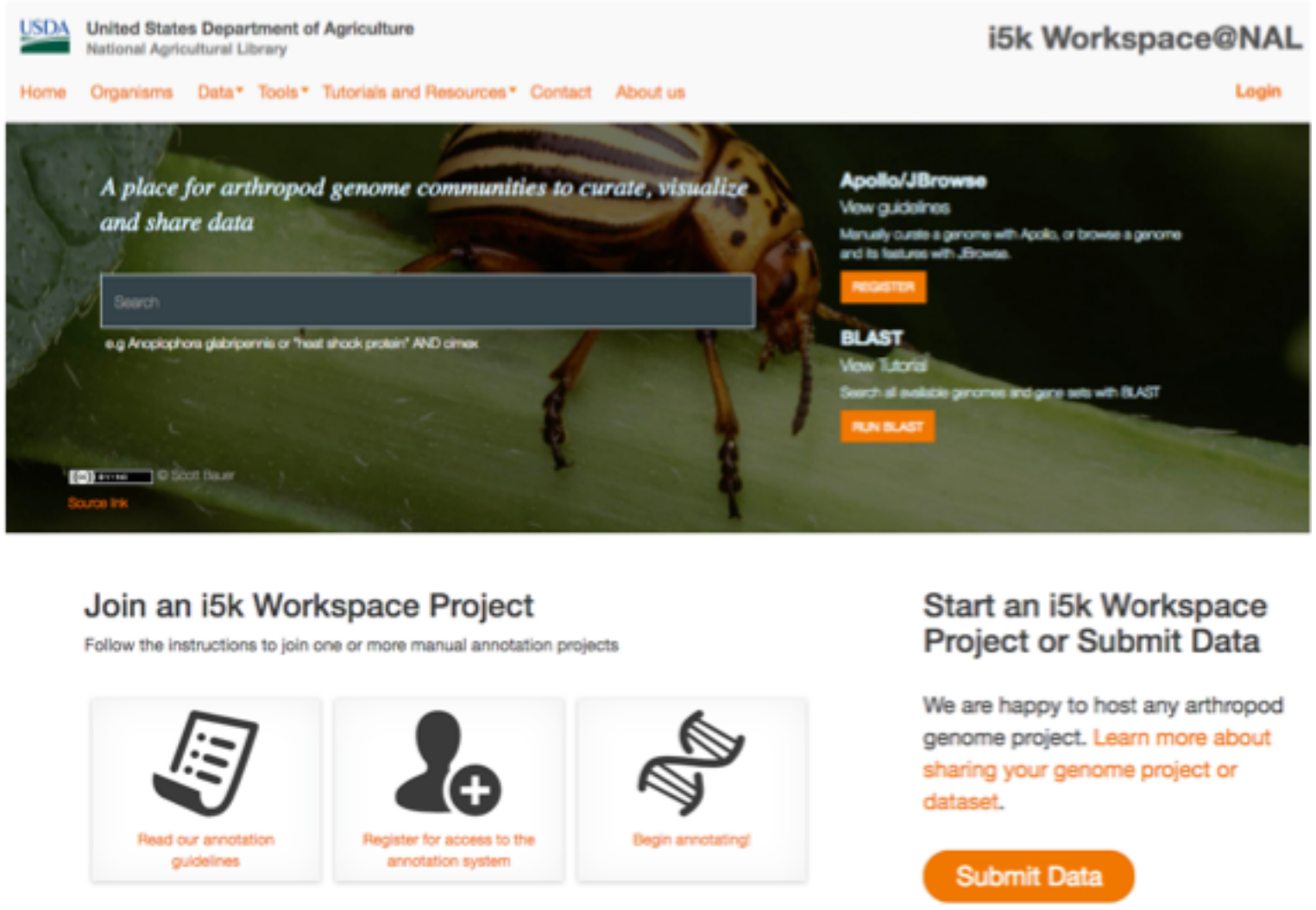
- Introduction and background
- Workspace Overview
- Tutorials and documents to help you get started
- Starting a project
- Submitting your genome and associated data
- Tools and example workflows

The i5k initiative

"This project is aimed at sequencing and analyzing the genomes of all species known to be important to worldwide **agriculture and food safety, medicine, and energy production**; all species used as **models in biology**; the **most abundant insects in world ecosystems**; and, to achieve a deep understanding of arthropod evolution, **representatives of insect relatives in every major branch of arthropod phylogeny**. The i5k initiative will be **broad and inclusive**, seeking to involve scientists from around the world and obtain funding from academia, governments, industry, and private sources. We also aim to **encourage new collaborative research** by computer scientists, bioinformaticians, and biologists to overcome the challenges of handling this unprecedented volume of data and derive meaning from these genomes."

** [(**Science, 2011**)](http://science.sciencemag.org/content/331/6023/1386)*

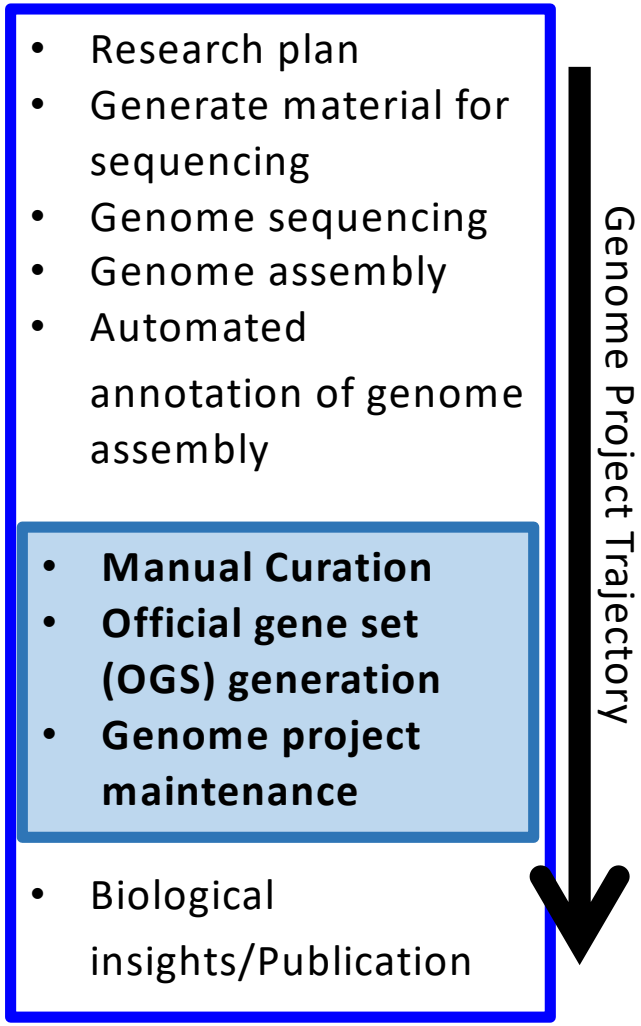
The i5k Workspace@NAL



USDA

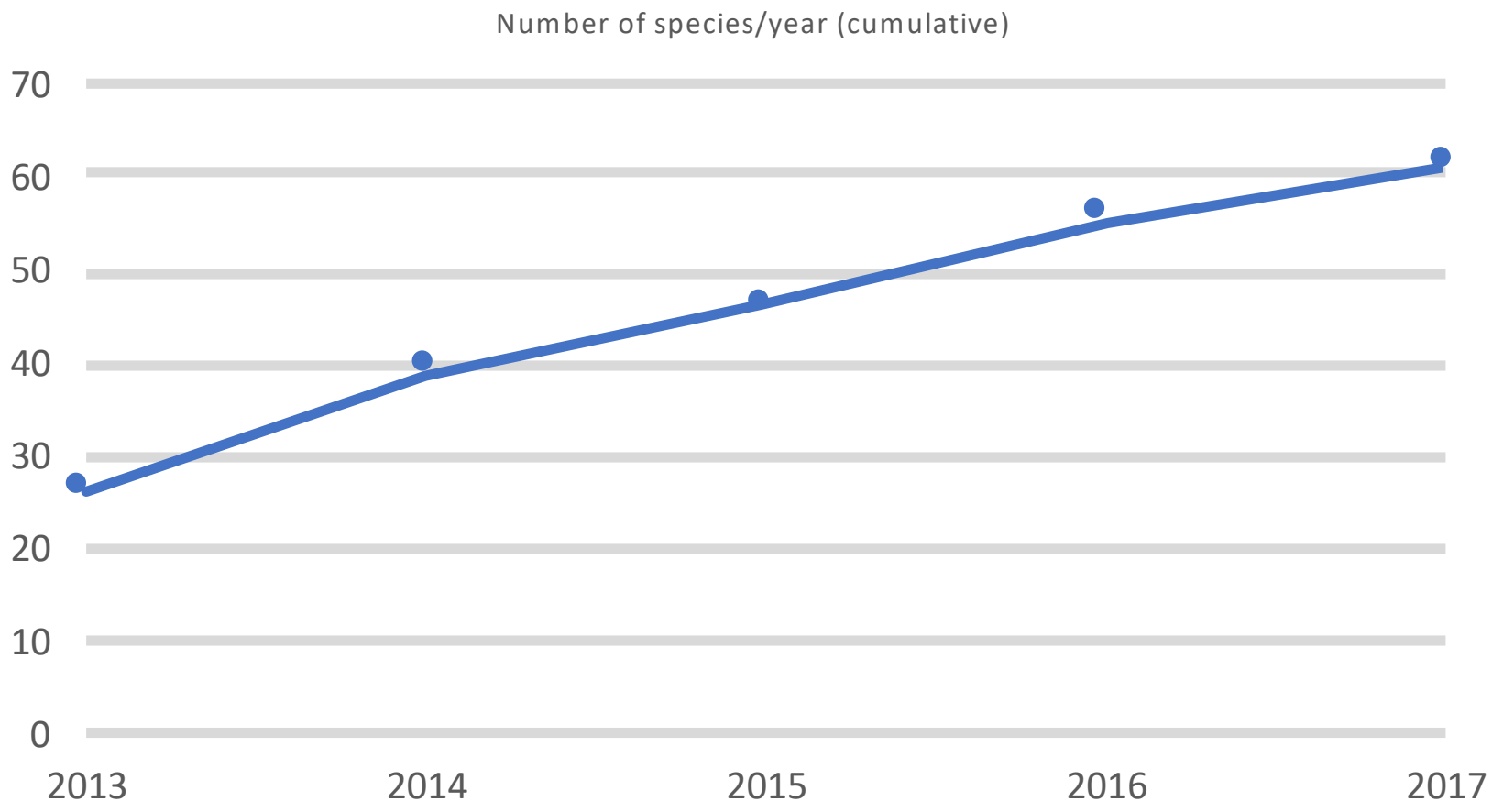
The i5k Workspace@NAL

- The i5k Workspace@NAL was launched in 2013 to help any i5k (arthropod) project with genome hosting needs
- Projects are owned by the users
- Our tools are free and open source
- Our data is all open access



USDA

Number of i5k Workspace species/year



USDA

i5k Workspace content 2018– 61 species and counting

Order	Quantity	Order	Quantity
Amphipoda	1	Hemiptera	8
Araneae	3	Hymenoptera	15
Blattodea	1	Ixodida	1
Calanoida	1	Lepidoptera	3
Coleoptera	7	Odonata	1
Diplura	1	Orthoptera	1
Diptera	13	Scorpiones	1
Ephemeroptera	1	Thysanoptera	1
Harpacticoida	1	Trichoptera	1

- Many other datasets mapped to, or predicted from each genome assembly (gene predictions, transcriptomes, RNA-Seq, etc.)

Genome database and community resources

- Organism pages
- Gene pages
- Project creation/data submission tools
- Training resources
  - Guides and tutorials
  - Training tools
- Additional documents
  - Long term management plan
  - Data management policy

Organism pages

United States Department of Agriculture
National Agricultural Library

i5k Workspace@NAL

Home Organisms Data Tools Tutorials and Resources Contact About us Search Login

Organisms / Aethina tumida

Aethina tumida

Overview
Annotation Methods
Assembly Methods
NCBI BioSample

Overview

The small hive beetle is a widespread parasite of honey bee colonies. Originally from southern Africa, this beetle has followed honey bees to several continents and is a nuisance pest throughout much of its introduced range. Genomic and transcriptomic data can lead to basic insights into beetle biology, and to potential methods of control for this parasite.
Community contact: Jay D Evans
Image Credit: James D. Ellis, University of Florida. View Source. CC-BY-3.0-US

Aethina tumida data files

	Name	Last modified
	Parent Directory	
	Current Genome Assembly	2017-03-24 18:51
	GCF_001937115.1	2017-03-24 18:51

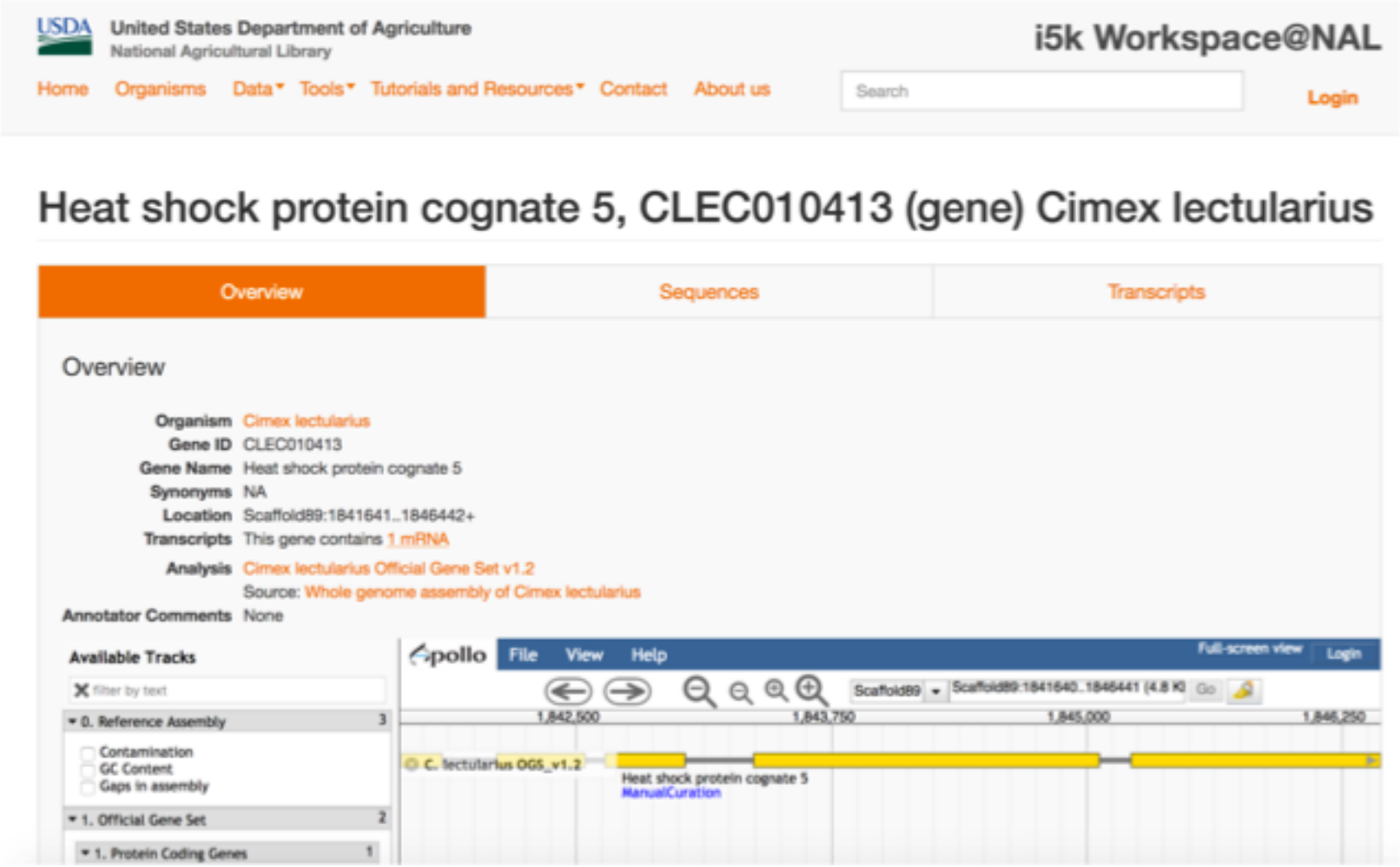
Assembly Information

Analysis Name	Aethina tumida genome assembly Atum_1.0 (GCF_001937115.1)
Software	Sparse assembler (ILLUMINA reads) followed by SPAR C (error-corrected PacBio reads)
Source	SAMN06036204
Date performed	2017-01-05
	Data source: Geo location: USA, Beltsville, MD, 39.0391019,-76.8815947 Tissues/Life stage included: larvae Sex: Pooled Strain: BRL

Statistics

Assembly Metrics	
Contig N50	NA
Scaffold N50	298879
GC Content	30.55

Gene Pages


United States Department of Agriculture
National Agricultural Library
i5k Workspace@NAL
Home
Organisms
Data
Tools
Tutorials and Resources
Contact
About us
Search
Login
Heat shock protein cognate 5, CLEC010413 (gene) Cimex lectularius
Overview
Sequences
Transcripts
Overview
Organism Cimex lectularius
Gene ID CLEC010413
Gene Name Heat shock protein cognate 5
Synonyms NA
Location Scaffold89:1841641..1846442+
Transcripts This gene contains 1 mRNA
Analysis Cimex lectularius Official Gene Set v1.2
Source: Whole genome assembly of Cimex lectularius
Annotator Comments None
Available Tracks
filter by text
0. Reference Assembly
Contamination
GC Content
Gaps in assembly
1. Official Gene Set
1. Protein Coding Genes
Apollo
File
View
Help
Full-screen view
Login
Scaffold89
Go
1,842,500
1,843,750
1,845,000
1,846,250
C. lectularius OGS_v1.2
Heat shock protein cognate 5
ManualCuration

Submitting Data

See https://i5k.nal.usda.gov/data-submission-overview

- There are several types of data you can submit to us
- The genome must be available from NCBI or another INSDC member
- All data sets should be mapped back to the same assembly
- Collect as much information about the data as possible

Your Project: Part 1 – Create a user account

- https://i5k.nal.usda.gov/register/project-dataset/account
- Fill out the form with
  - Name
  - Email
  - Affiliation
  - Some description of your project or the data you want to submit
- After approved, you can proceed

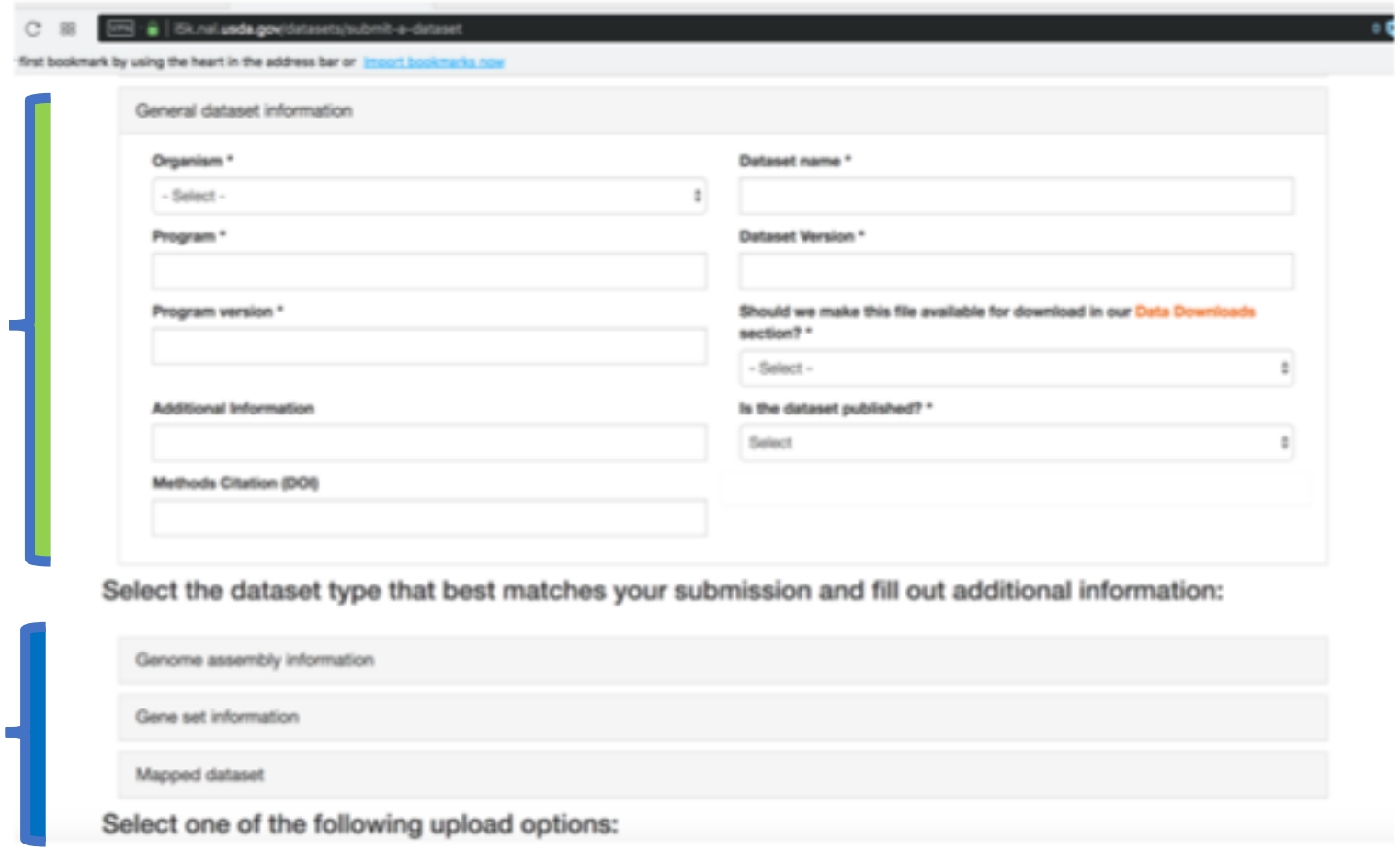
Your Project: Part 2 – Project request

- https://i5k.nal.usda.gov/datasets/request-project
- We need some general information about your project, such as:
- Your organism:
  - Species
  - NCBI taxonomic ID
- Your genome assembly:
  - Is it already hosted somewhere?
  - Is it published?
- General plans for the project (e.g. interested in gene curation?)
- Your name and email

Your Project: Part 3 – The data/metadata

- https://i5k.nal.usda.gov/datasets/submit-a-dataset
- Once the project is approved, you can start adding data
- One form to submit your:
  - Genome assembly
  - Gene predictions
  - Other mapped data
- Metadata is important!
  - Provide as much information as you can
  - Helps for reuse and downstream analyses

Data submission



USDA

Data submission – Genome Assembly

Genome assembly information

Project Background

Project description to display in your organism page

Image file name for your organism page (image submission form)

Will you manually curate this assembly using i5k workspace tools?

-Select-

Data source information

Geographic location (latitude and longitude)

Tissues/Life stage included

Sex

USDA

Data submission – Gene Set

Gene set information

Descriptive track name for JBrowse and Apollo

Is this an Official Gene Set?

-Select-

We will generate gene pages for Official Gene Sets - specify yes if the gene set is viewed as the community standard for this genome assembly

Mapped dataset

Data submission – Other Mapped Data Set

Mapped dataset

Geographic location (latitude and longitude)

Tissues/Life stage included

Sex

Select

Sequencing method

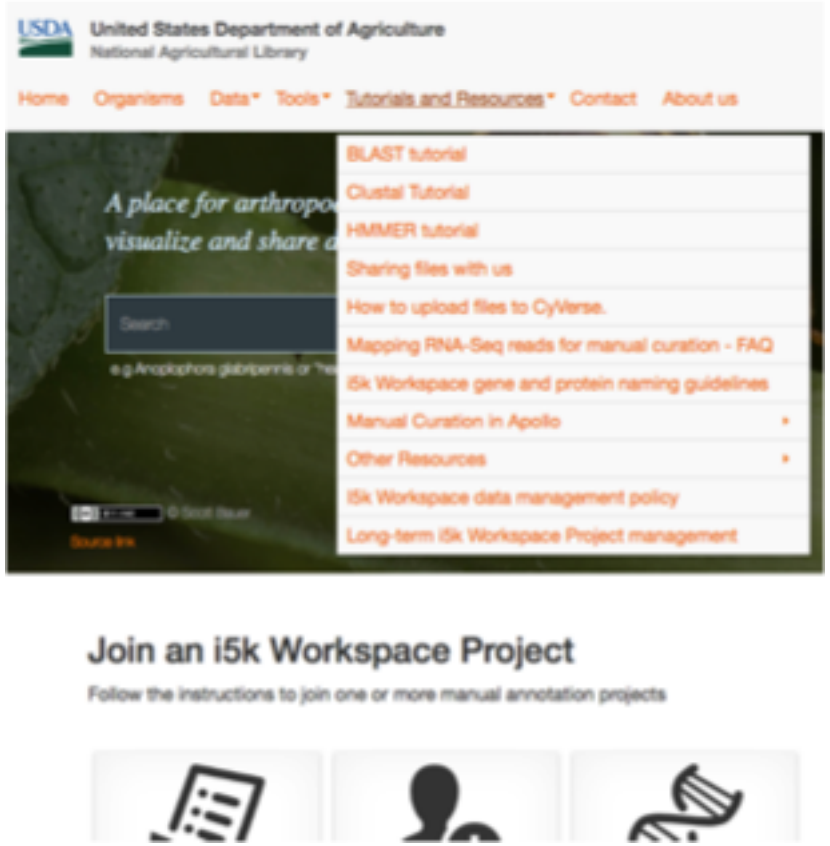
Descriptive track name for JBrowse and Apollo

NCBI SRA accession number(s)

USDA

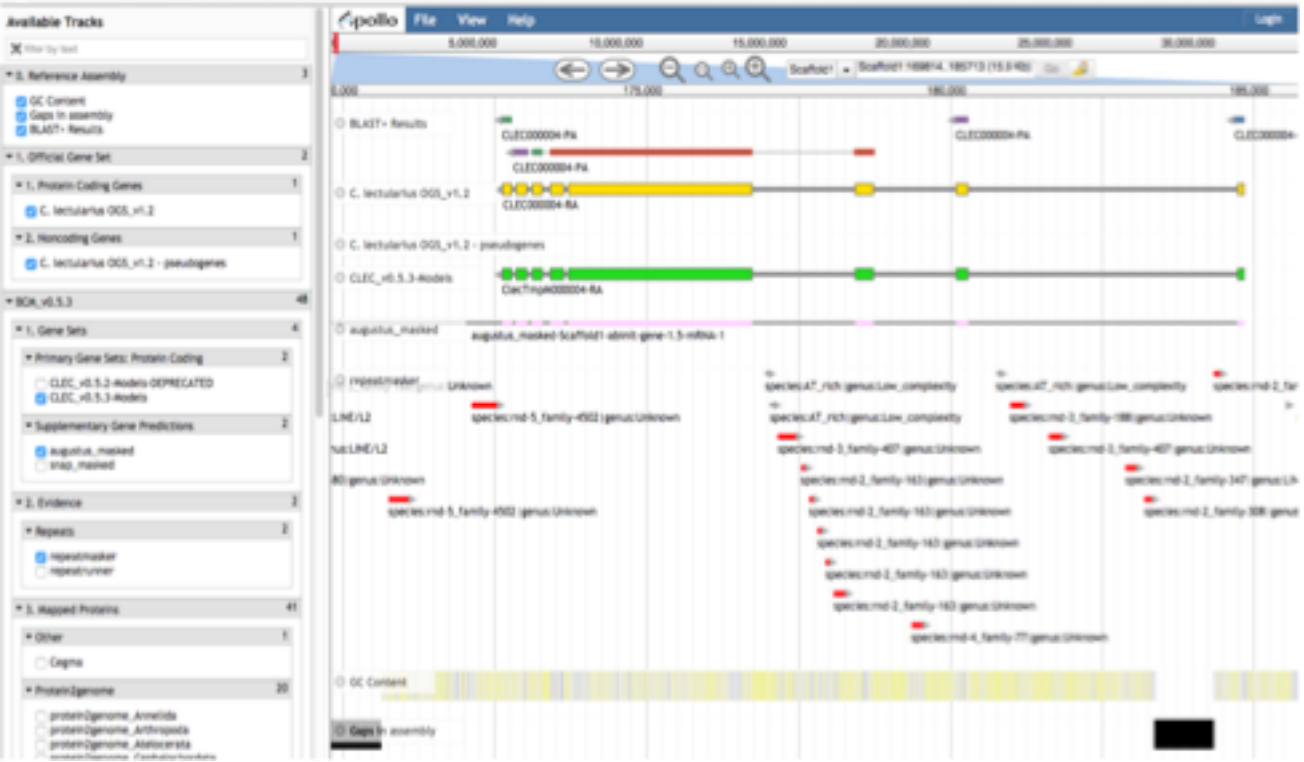
Documentation, tutorials, and outreach

- Tutorials for our tools
- Manual annotation
- Naming guidelines
- RNA-Seq mapping
- Webinars, talks, posters: https://i5k.nal.usda.gov/talks-and-presentations



Tools at the i5k Workspace@NAL

- Search
  - BLAST
  - HMMER
  - Clustal
- Visualization
  - Apollo
- Other tools for data processing: https://github.com/NAL-i5K/



USDA

Thank you!

The NAL Team

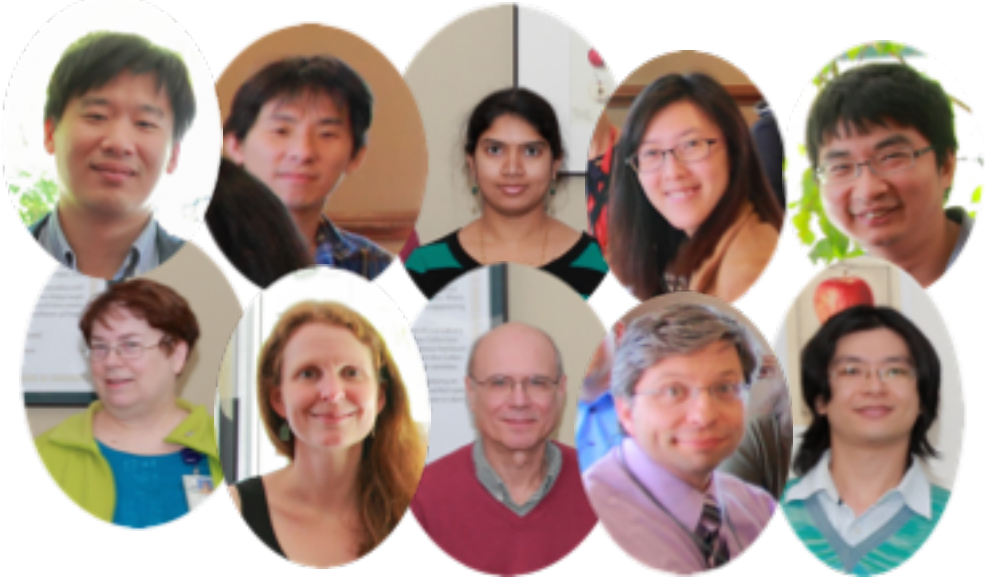
- Yu-yu Lin
- Chaitanya Gutta
- Li-Mei Chiang
- Yi Hsiao
- Gary Moore
- Susan McCarthy

I5k Workspace alumni

- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli
- Mei-Ju Chen
- Chao-I Tuan

i5k Workspace@NAL advisory committee

- i5k Coordinating Committee
- i5k Pilot Project
- Apollo & JBrowse Development Teams
- GMOD/Tripal community
- All of our users and contributors!



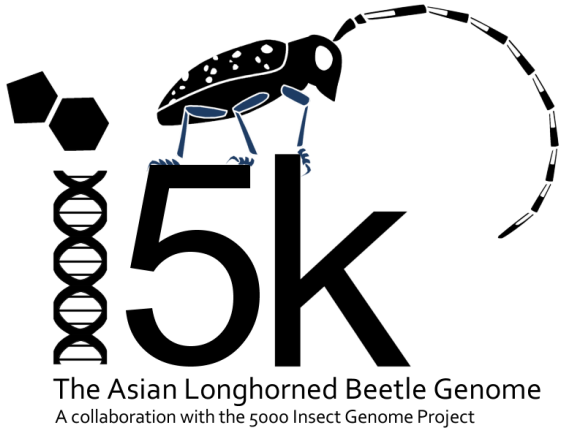
USDA

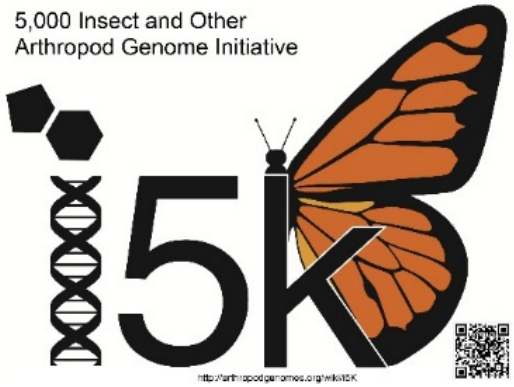
Would you like to know more?

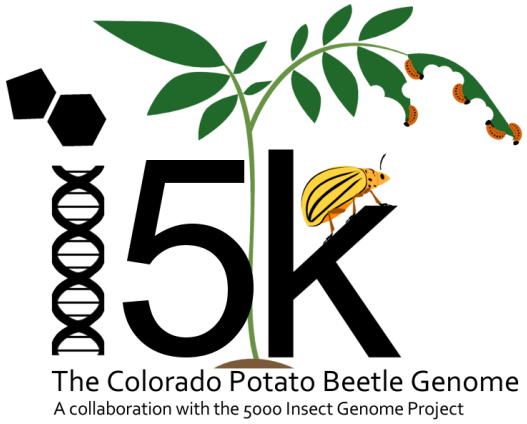
- Links to i5k resources
  - I5k website – i5k.github.io
  - Twitter – http://www.twitter.com/@Arthropod_i5k
  - I5k pilot project – http://www.hgsc.bcm.edu/arthropods/i5k
- The i5k Workspace@NAL:
  - Email us – i5k@ars.usda.gov
  - Visit the site – https://i5k.nal.usda.gov
  - Read all about us:
    - **The i5k Workspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes**
      - *doi: 10.1093/nar/gku983*
      - http://nar.oxfordjournals.org/content/43/D1/D714
  - I5k Workspace@NAL software repository – http://www.github.com/NAL-i5K

USDA

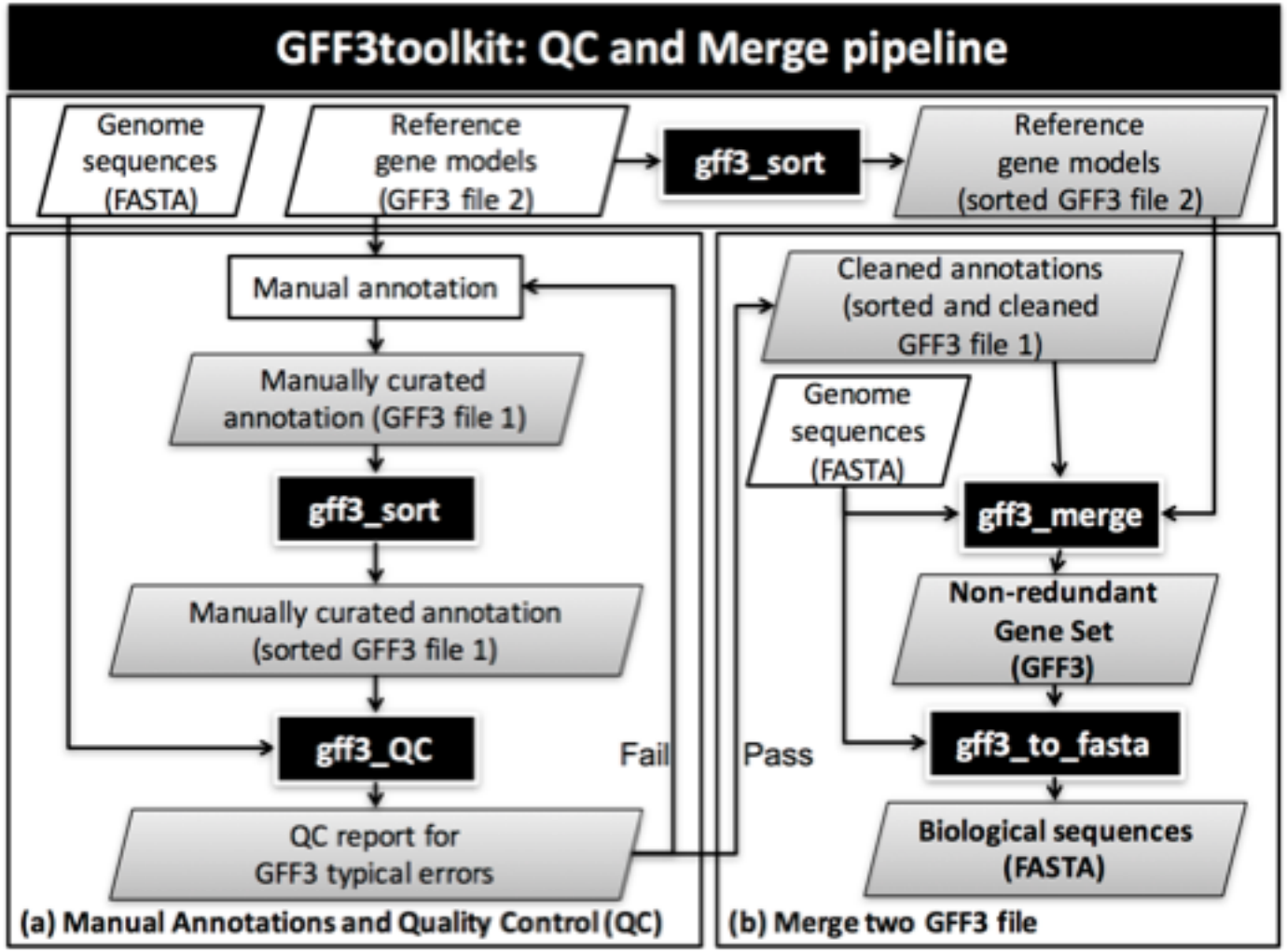
Community and Communities!







OGS generation – the GFF3toolkit



USDA