

# The i5k Workspace@NAL: a pan-Arthropoda Genome Database

Chris Childers and Monica Poelchau  
USDA-ARS, National Agricultural Library



# Outline

- Background and overview
- Why join the i5k Workspace?
- What do we need for a project?
- What we do with your data?
- What don't we do with your data?
- Our new system for submitting projects and data

# Background

- The i5k initiative tasked itself with coordinating the sequencing and assembly of 5000 insect or related arthropod genomes
- International effort to **prioritize** insect genomes for sequencing; provide **guidelines for genome sequencing and curation**; and seek **funding**.
- The i5k Workspace@NAL is available to help any i5k (arthropod) project with genome hosting needs

- Research plan
- Generate material for sequencing
- Genome sequencing
- Genome assembly
- Automated annotation of genome assembly

- **Manual Curation**
- **Official gene set (OGS) generation**
- **Genome project maintenance**

- Biological insights/Publication

Genome Project Trajectory

# Workspace Project Basics

- The i5k Workspace centers around *projects*.
  - A project is a collection of data based on the genome assembly of an arthropod
  - All data is used in the context of the genome assembly
- Each project has a *project coordinator*.
  - Serves as the point of contact for questions about the project
  - Main responsibility: approve or reject new Apollo users
- All of our data is user-submitted

# Why join the i5k Workspace?

- Gain access to a large diverse community
  - A diversity of organisms
    - 58 species and counting
    - 20% of the arthropods with genome assemblies at NCBI
  - Large user community with many different interests
    - People versed in the biology of specific systems
    - Experts in a species or group of species
- A common interface for accessing data, tools and search
- Detailed policies on data and project management
  - Helpful if you have data management requirements
  - Data management
    - <https://i5k.nal.usda.gov/data-management-policy>
  - Long-term project management
    - <https://i5k.nal.usda.gov/long-term-i5k-workspace-project-management>



# What do we need for a project?

- Your project metadata
  - Information about your organism
  - Metadata for submitted data files (the more the better)
    - What tools or methods were used
    - Software versions and options set
    - When and where the data were generated
    - Other information (location collected, life-stage, etc.)
- Your data files
  - Genome assembly needs to be in GenBank/ENA/DDBJ
  - Data should be open access (no private repositories)
  - Additional datasets need to be mapped to the same assembly

# What do we do with your data?

- Create resources
  - Organism and gene pages
  - Data downloads
- Integrate your data with our tools
  - Genome browser
  - BLAST, Clustal, HMMer
  - Apollo for gene curation
- Offer post curation services
  - Annotation QC and Official Gene Set (OGS) Creation
  - Update gene pages, Apollo, BLAST with OGS



'Frozen' genome assembly

Automated annotations

Ancillary datafiles (e.g. RNA-Seq alignments)

Submission



Workspace@NAL  
<https://i5k.nal.usda.gov/>

## Resources

### Organism Information Page

Anoplophora glabripennis

The Asian long-horned beetle (Anoplophora glabripennis) is an invasive pest from Asia that is threatening forests in the United States and Europe. It is a serious threat to deciduous hardwoods, suburban, and forested parts of the country. It has already established itself in the United States and Europe. It is a serious threat to deciduous hardwoods, suburban, and forested parts of the country. It has already established itself in the United States and Europe. It is a serious threat to deciduous hardwoods, suburban, and forested parts of the country. It has already established itself in the United States and Europe.

most destructive and costly invasive species ever (USDA Pest Alert No. 165). This is a USDA eradication effort.

### Bulk data downloads

Data downloads

Conditions of use

Genomes, predicted gene sets, and RNA-Seq data hosted on the i5k Workspace@NAL publication. These data are covered by the F1 Lauterbach and Tonella agreement permission to use the prepublication data and cite the appropriate source, and the each organism page, if you wish to use these sequences in published analyses.

Data DOWNLOADS

- expand all: collapse all
- Arthropoda
- zygophylloptera, planipennia
- seq
- contig
- cds
- analyses
- blastp
- blastn
- blastx
- blastp+blastn
- NOTE: contamination

### Tutorials

BLAST tutorial

The BLAST resources at the i5k Workspace@NAL have been completely re-engineered to provide support for many new features. Most of the new suite of BLAST+ programs, a completely redesigned user interface that incorporates intelligent option selection, a query-a-multi-panel result viewing page that allows for many different ways to view and filter the BLAST results.

In this tutorial, we explain some of the features of the BLAST site.

To begin, navigate to the BLAST site via the 'Tools' -> 'BLAST' menu bar, or go directly to [Under the 'Organisms' menu, select one or several organisms that you would like to see. Data includes genome assemblies, or transcript or protein sequences predicted from both the genome assembly and transcripts can be searched simultaneously, but transcripts](#).

i5k@NAL BLAST About i5k BLAST Contact

BLAST Databases

Organisms

- All organisms
- Anoplophora glabripennis
- Anoplophora glabripennis
- Atthalia rosea
- Nucleotide
- Genome Assembly - Aros01112013-genome\_new\_ids.fa
- Transcript - AROS\_new\_ids.fna
- blastsite

## Tools

### Custom BLAST interface

Available Tracks

- 0. Reference Assembly
- 1. GC Content
- 2. Coverage Assembly of BLAST+ Results
- 3. BLAST v0.5.3
- 4. Gene Sets
- 5. Primary Gene Sets: Protein Coding
- 6. LDC v0.5.3 Models
- 7. Supplementary Gene Predictions
- 8. Evidence
- 9. Reports
- 10. Mapped Proteins
- 11. Other

### JBrowse genome browser

Available Tracks

- 0. Reference Assembly
- 1. GC Content
- 2. Coverage Assembly of BLAST+ Results
- 3. BLAST v0.5.3
- 4. Gene Sets
- 5. Primary Gene Sets: Protein Coding
- 6. LDC v0.5.3 Models
- 7. Supplementary Gene Predictions
- 8. Evidence
- 9. Reports
- 10. Mapped Proteins
- 11. Other

### Apollo manual curation tool

Available Tracks

- 0. Reference Assembly
- 1. GC Content
- 2. Coverage Assembly of BLAST+ Results
- 3. BLAST v0.5.3
- 4. Gene Sets
- 5. Primary Gene Sets: Protein Coding
- 6. LDC v0.5.3 Models
- 7. Supplementary Gene Predictions
- 8. Evidence
- 9. Reports
- 10. Mapped Proteins
- 11. Other

### 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)

# What don't we do with your data?

- Computationally intense analyses such as
  - Gene prediction
  - Raw RNAseq mapping
- We are not a long-term archive or repository
  - NCBI
  - Ag Data Commons
  - Dryad Digital Repository
  - CyVerse Data commons
  - Many other options available



# Criteria for starting a project

- You need to have an **arthropod** genome assembly, **accessioned by NCBI** (or another INSDC member)
  - Using GenBank's accession numbers avoids confusion about assembly version
  - The GenBank contamination screen improves the assembly quality
  - Using a stable assembly is beneficial for the labor-intensive community annotation process

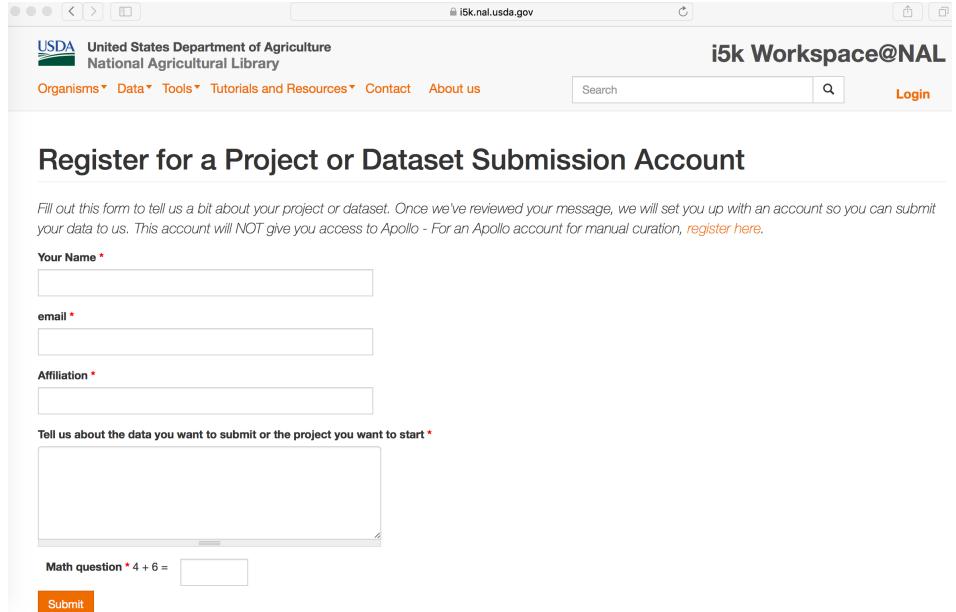
# Other things to consider before submitting

- *All data submitted to the i5k Workspace is public.*
  - However, we do state whether Ft. Lauderdale/Toronto agreements of data sharing should apply
- Is your genome an ‘orphan’, or is there another suitable database?
  - We can host genomes that are already hosted elsewhere, and actively communicate with other database providers
  - All manual annotation efforts need to be at one database



# Getting an account

- Apply for a dataset submission account:  
<https://i5k.nal.usda.gov/register/project-dataset/account>
- Once your account is approved, you can submit projects, assemblies or other datasets



The screenshot shows a web browser window for the URL [i5k.nal.usda.gov](https://i5k.nal.usda.gov). The page title is "i5k Workspace@NAL". The header includes the USDA logo and links for "Organisms", "Data", "Tools", "Tutorials and Resources", "Contact", and "About us". A search bar and a "Login" button are also present. The main content area is titled "Register for a Project or Dataset Submission Account". It contains instructions: "Fill out this form to tell us a bit about your project or dataset. Once we've reviewed your message, we will set you up with an account so you can submit your data to us. This account will NOT give you access to Apollo - For an Apollo account for manual curation, [register here](#)". There are four required fields: "Your Name", "email", "Affiliation", and a large text area for "Tell us about the data you want to submit or the project you want to start". Below the text area is a "Math question" field asking "4 + 6 =". A "Submit" button is at the bottom.



# Start an i5k Workspace Project

- Log in
  - <https://i5k.nal.usda.gov/user>
- From menu, select 'Data -> Submit data -> Request a new i5k Workspace Project'
  - <https://i5k.nal.usda.gov/datasets/request-project>
- We'll review your submission and will get in touch with you

The image shows two screenshots of the i5k Workspace@NAL website. The top screenshot is a desktop view of the homepage, featuring the USDA logo, a search bar, and links for Apollo/JBrowse, BLAST, and other resources. A red arrow points to the 'Login' button in the top right corner. The bottom screenshot shows a detailed view of the 'Request a new i5k Workspace Project' form. It includes fields for Genus\*, Species\*, NCBI Taxonomy ID\*, and Common Name\*. Below the form is a question: 'Is the genome assembly already hosted at another genome portal, or is there another genome portal that would also be appropriate to host your dataset (e.g. VectorBase, HGID)?\*'.

# Submit your genome assembly

- All information submitted through this form will be reformatted for display at the i5k Workspace (except for email address and file checksum)
- From menu, select ‘Data -> Submit data -> Submit a genome assembly’
  - <https://i5k.nal.usda.gov/datasets/assembly-data>

The screenshot shows the 'Submit a Genome Assembly' page. At the top, there's a navigation bar with the USDA logo, 'United States Department of Agriculture National Agricultural Library', and links for 'Organisms', 'Data', 'Tools', 'Tutorials and Resources', 'Contact', and 'About us'. On the right, it says 'i5k Workspace@NAL' with a search bar and a 'My Account' link. Below the navigation, the page title 'Submit a Genome Assembly' is centered. A note below the title states: 'Please note that this dataset will be visible to the public in the JBrowse genome browser. Contact us if this dataset needs to remain private. Refer to our [data management](#) and [long-term management policy](#) documents for information about the data types that we store and our long-term data management policy.' There are two main sections: 'Genome coordinator information' which includes fields for 'Full Name' and 'Email Address'; and 'Project Background' which includes a dropdown for 'Organism' and a text area for 'Project description to display in your organism page'.

# Submit gene predictions

- All information submitted through this form will be re-formatted for display at the i5k Workspace (except for email address and file checksum)
- Under menu bar, select ‘Data -> Submit data -> Submit Gene Predictions’
  - <https://i5k.nal.usda.gov/datasets/gene-prediction>

The screenshot shows the 'Submit Gene Predictions' page from the i5k Workspace@NAL website. At the top, there's a navigation bar with links for USDA National Agricultural Library, Organisms, Data, Tools, Tutorials and Resources, Contact, About us, a search bar, and a 'My Account' dropdown. Below the header, the URL 'Data / Submit Data / Submit Gene Predictions' is visible. The main section is titled 'Submit Gene Predictions'. A note below the title states: 'Please note that this dataset will be visible to the public in the JBrowse genome browser. Contact us if this dataset needs to remain private. Refer to our [data management](#) and [long-term management policy](#) documents for information about the data types that we store and our long-term data management policy.' There are several input fields: 'Organism \*' with a dropdown menu, 'Analysis Method' with sections for 'Program \*' and 'version \*', and a large 'Additional Information' text area. At the bottom left of the form, there's a link 'Other Methods'.

# Submit mapped datasets

- All information submitted through this form will be re-formatted for display at the i5k Workspace (except for email address and file checksum)
- Under menu bar, select ‘Data -> Submit data -> Submit a Mapped Dataset’
  - <https://i5k.nal.usda.gov/datasets/mapped>

The screenshot shows the 'Submit a Mapped Dataset' page from the i5k Workspace@NAL website. At the top, there's a header with the USDA logo, the text 'United States Department of Agriculture National Agricultural Library', and a search bar. Below the header, the page title 'Submit a Mapped Dataset' is displayed. A note below the title states: 'Please note that this dataset will be visible to the public in the JBrowse genome browser. Contact us if this dataset needs to remain private. Refer to our [data management](#) and [long-term management policy](#) documents for information about the data types that we store and our long-term data management policy.' The main form area contains fields for 'Organism \*' (with a dropdown menu), 'Data provider' (with fields for 'Full Name \*', 'Email Address \*', and 'Affiliation \*'), 'Geo location', and 'Tissues/Life stage included'. A note at the bottom of the form area says '(Whole individual/ antenna / pooled larva / pooled adult female/ etc.)'



# Send us your files

- There are currently **five** ways to share files with us:
  1. Use our data submission forms
  2. Transmit the file via **ftp** (only for files < 2 Gb)
  3. Email it to us (for files < 25 Mb only)
  4. Provide us with a **URL**, if available
  5. Upload the file to **CyVerse** and share with our organization, “NAL Bioinformatics”
- We prefer that you share your files with us via our data submission forms.
- For more information, see  
<https://i5k.nal.usda.gov/content/sharing-files-us>

# Other resources at the NAL: the Ag Data Commons

- Hosts any dataset funded by the USDA
- Landing page
- Citable DOI
- <https://data.nal.usda.gov/>
- 9 i5k datasets already available

The screenshot shows the homepage of the Ag Data Commons Beta. At the top, there is a navigation bar with the USDA logo, the text "Ag Data Commons Beta", "National Agricultural Library", and links for "Datasets", "About", "News", "Log in", and "Register". Below the navigation bar is a large banner featuring a close-up image of a plant stem with small brown spots. Overlaid on the banner is the text "Featured program: The Veterinary Pest Genomics Center" and a description: "This program uses big data to evaluate risk from and develop mitigations for invasive and other economically important veterinary pests." To the right of the banner is a search bar with the placeholder "Search Ag Data Commons" and a magnifying glass icon. Below the banner, there are two sections: "Topics" (with icons for Agricultural Products and Agroecosystems) and "Highlighted Datasets" (with a thumbnail image of a landscape and a descriptive text box).

# Need more information?

i5k Workspace@NAL:

- <https://i5k.nal.usda.gov/>
- <https://github.com/NAL-i5K/>

The i5k initiative:

- New website: <http://i5k.github.io/>



# Official Gene Set creation at the i5k Workspace

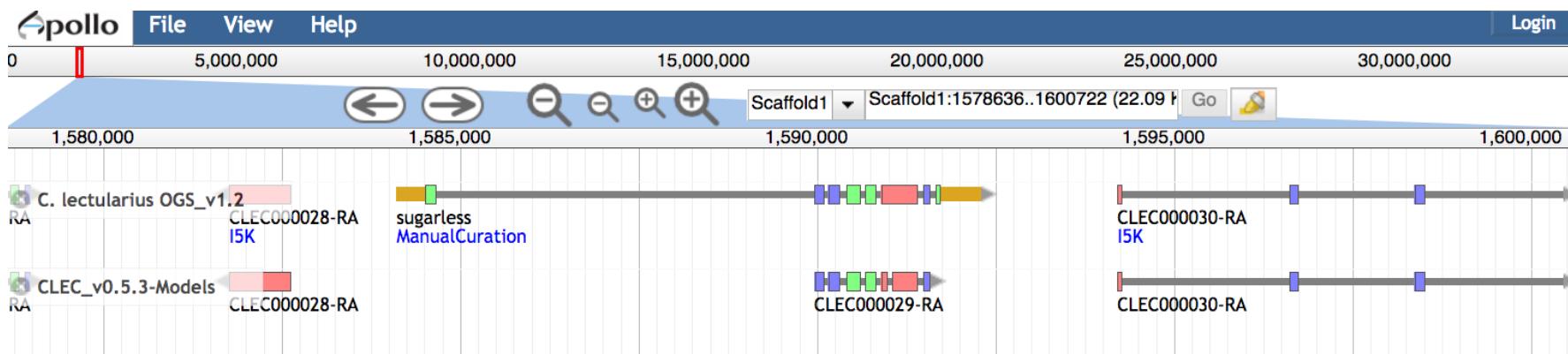


# Official Gene Set creation at the i5k Workspace

- Official Gene Set definition
- Our OGS generation process
  - Manual and community annotation
  - Quality control
  - Merge
  - Release
- Examples and future directions of the OGS generation process

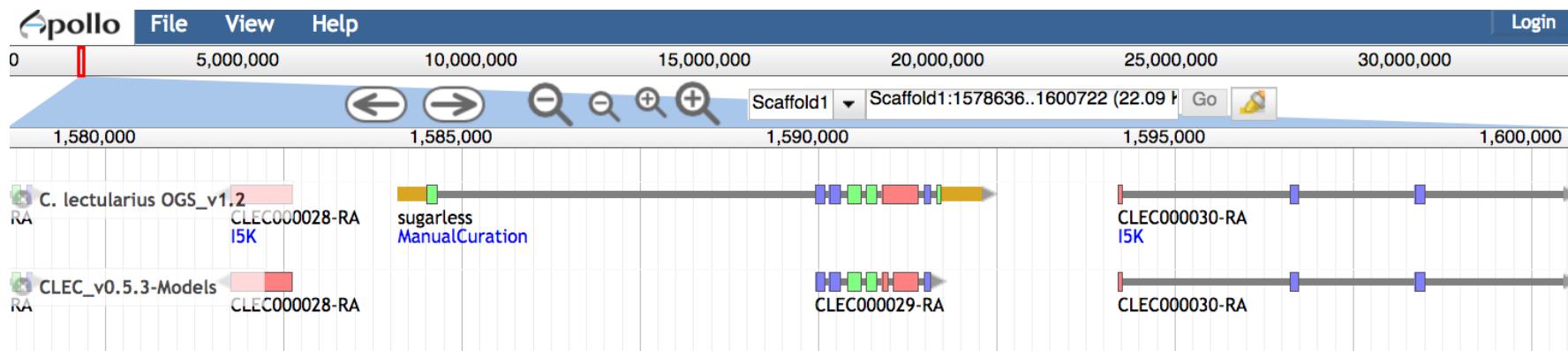
# The Official Gene Set – what is it?

- Loose definition: The best known representation of gene models for a genome assembly
- When the i5k Workspace generates an OGS, this is a merge between one gene set (usually computationally predicted), and a set of manually validated annotations (usually from the Apollo software)

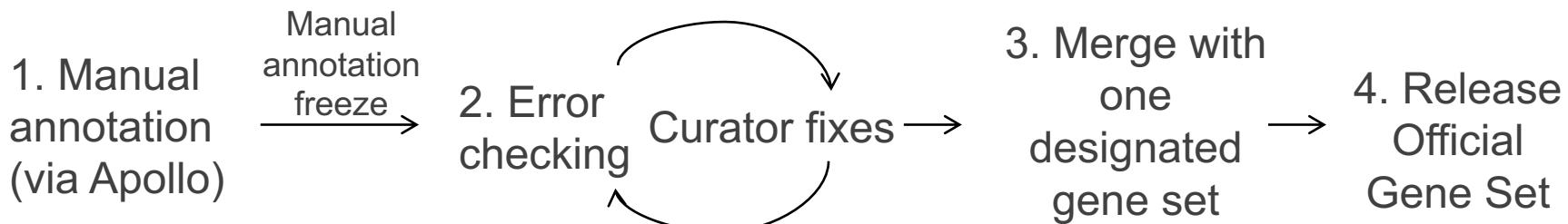


# Why generate an Official Gene Set?

- This depends on your genome community's needs.
- If several groups want to perform downstream analyses, it helps to have an authoritative 'reference gene set' for your community, rather than multiple competing gene sets



# Our OGS generation process

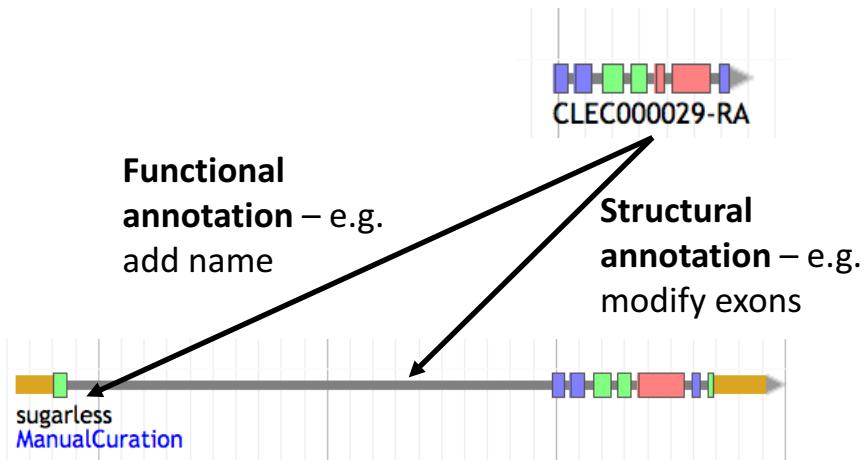


- New public version of program is available:  
<https://github.com/NAL-i5K/GFF3toolkit> (Mei-Ju Chen, Li-Mei Chiang)
- The full process is time-consuming, but we are generally available to perform OGS generation for i5k Workspace projects

# 1. Manual and community annotation

## What is manual annotation?

- Manual review and improvement of an existing gene prediction
- Often, but not always: drawing on external evidence (e.g. RNA-Seq, cDNA, genes from other species) to improve a computationally predicted gene model



# 1. Manual and community annotation

## Why manually annotate?

- “Incorrect annotations poison every experiment that makes use of them ... Worse still, the poison spreads because incorrect annotations from one organism are often unknowingly used by other projects to help annotate their own genomes.”
  - Yandell and Ence 2012, doi:10.1038/nrg3174
- Link gene models to existing literature and ontologies, providing richer data
- One current ‘model’ of the genome paper often draws heavily from insights confirmed by manual annotation

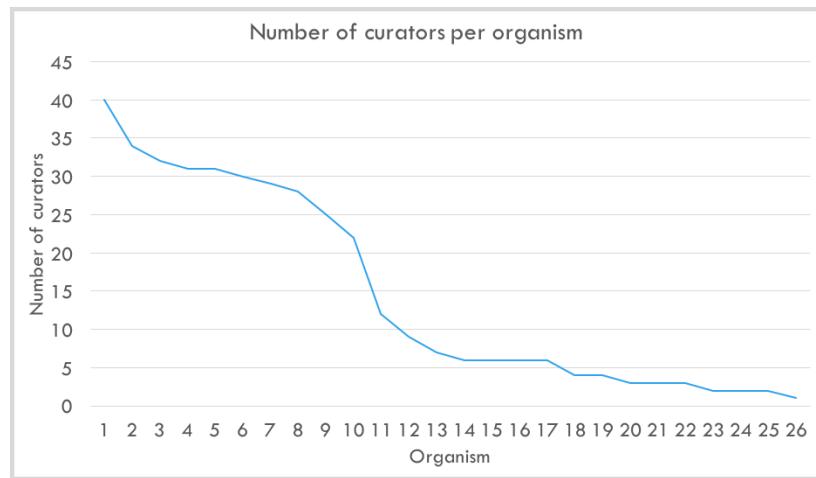


# 1. Manual and community annotation

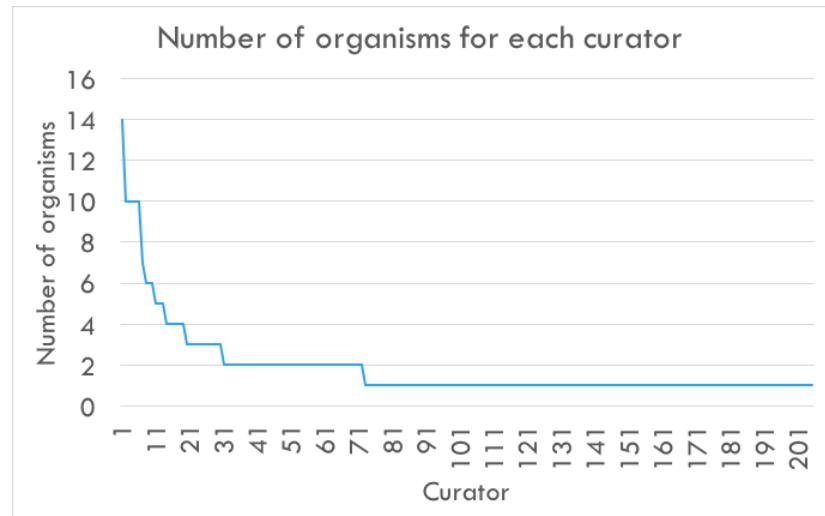
- What is community annotation?
  - Scientists collectively examine and improve gene models (usually computationally predicted)
- Community annotation at the i5k Workspace:
  - Access to a large community of curators
  - Tutorials, guidelines, webinars
  - Registration mechanism for new annotators
  - One-on-one support
  - Over 400 registered annotators have curated over 10,000 gene models using the Apollo software

# 1. Manual and community annotation – i5k pilot example

**Number of curators per organism. Community size varies among organisms.**



**Number of organisms per curator. 35% of curators worked on more than one organism**



# 1. Manual and community annotation – i5k pilot example

organism	Total number of manually annotated models	Proportion of manually annotated models with structural changes
Anoplophora glabripennis <sup>6</sup>	1144	0.75
Cimex lectularius <sup>7</sup>	1354	0.76
Oncopeltus fasciatus	1518	0.76

- Three organisms that completed the manual annotation process had to perform similar amounts of structural annotations to computationally predicted gene annotations
  - Computationally predicted genes often have inaccurate gene structures
  - Community annotation can effectively improve gene sets

## 2. OGS generation – Quality Control

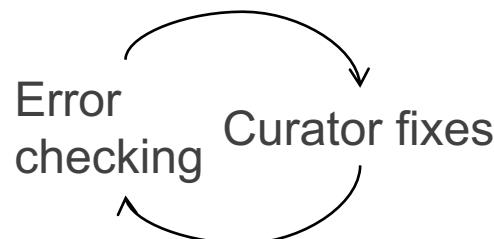
- Manual curation can introduce many errors, even using standard software packages (e.g. Apollo)
- QC program identifies common *formatting* errors from the manual curation process
  - Github repo: <https://github.com/NAL-i5K/GFF3toolkit>
- Identifies over 50 error types
- Another in-house pipeline corrects many of these errors

# 2. OGS generation – Quality Control

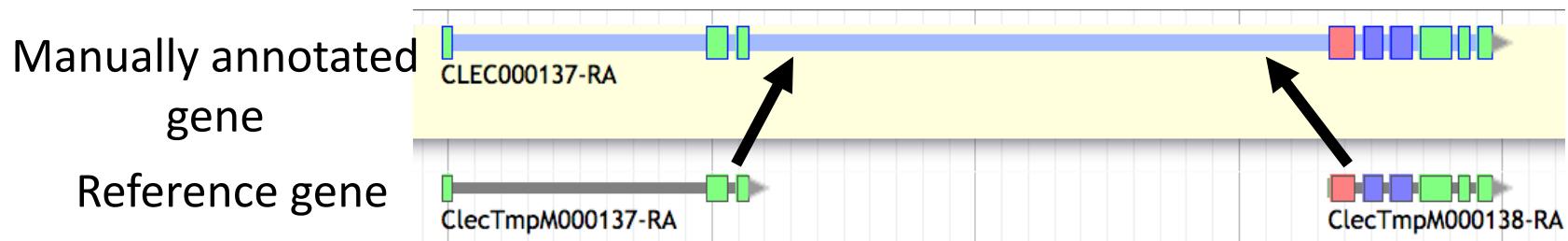
- Requires some manual review – can't be completely automated
  - e.g. did you name your gene model 'test' or 'Contig277'?
- Note that i5k Workspace staff aren't 'curators' in the traditional sense – we **do not** review the biological validity of any of the community-annotated models.
- The degree of manual review of community annotations is higher if Official Gene Sets are to be submitted to NCBI

# 2. OGS generation – Quality Control

- *Diaphorina citri* example (Database, doi: [10.1093/database/bax032](https://doi.org/10.1093/database/bax032))
  - First round of corrections for community curation:
    - 513 errors in 587 manually annotated models
    - 397 of these errors needed curator feedback
  - Second round of corrections :
    - 15 errors needed annotator feedback



### 3. OGS generation – Merge



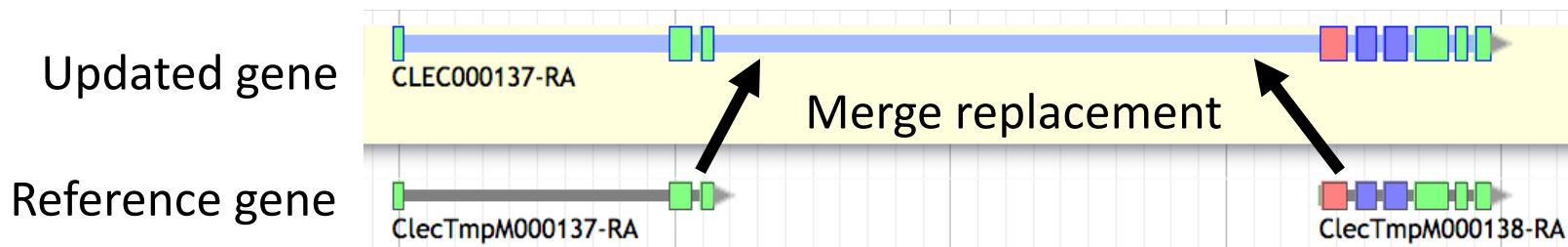
The GFF3 toolkit Merge program can identify which gene models in the ‘reference’ gene set should be replaced by gene models in a second gene set (i.e. the manually annotated models) via ‘auto-assignment’)

# 3. OGS generation – Merge

- Auto-assignment uses both sequence similarity and coordinate overlap
  - Extract CDS and pre-mRNA sequences from mRNA features from both gene sets.
  - Use blastn to determine which sequences from the modified and reference gene set align to each other **in their coding sequence.**
    - These parameters are used: -eval 1e-10 -penalty -15 -ungapped
  - If two models pass the alignment step, check that matched models also have coordinate overlap
  - Add a 'Replace Tag' with the ID of each overlapping model to the modified gene set.
  - If no reference model overlaps with a new model, then the program will add 'replace=NA'.

# 3. OGS generation – Merge

- The program determines merge actions based on the Replace Tags:
  1. deletion
  2. simple replacement
  3. new addition
  4. split replacement
  5. merge replacement
- Models from modified manual annotations replace models from reference annotations based on merge actions in step 2.



# 3. OGS generation – Merge

- *Diaphorina citri* example (Database, doi: [10.1093/database/bax032](https://doi.org/10.1093/database/bax032))
  1. # genes deleted: 1
  2. # genes with simple replacement: 437
  3. # genes added: 72
  4. # genes split: 38
  5. # genes merged: 31
  6. Total number of genes in OGS: 20,217

# 3. OGS generation – Merge

- Other software tools can be used to merge gene sets
  - Combiner tools that use ‘weights’ for different input annotations, e.g.
    - EVidenceModeler (EVM, <https://evidencemodeлер.github.io/>)
    - Glean (<https://sourceforge.net/projects/glean-gene/>)
  - Other overlap-based replacement tools, e.g Bedtools intersect (<http://bedtools.readthedocs.io/en/latest/>)

# 4. OGS generation – Release OGS

- Generate new or maintain old gene model IDs
- Establish release date with genome coordinator
- Generate fasta files
- Add to i5k Workspace@NAL database
- \*Submit to NCBI if requested by genome coordinator\*

# Completed OGS projects using i5k Workspace's pipeline

- *Diaphorina citri* OGSv1.0
- *Frankliniella occidentalis* OGSv1.0
- *Hyalella azteca* OGSv1.0
- *Oncopeltus fasciatus* OGSv1.2
- *Athalia rosae* OGSv1.0
- *Orussus abietinus* OGSv1.0
- *Leptinotarsa decemlineata* OGSv1.0

# Future updates

- Current improvements:
  - GFF3 toolkit support for QC and merge of non-coding transcripts (Li-Mei Chiang)
- Future work:
  - Improve methods for merging multi-isoform models
  - Improve QC process – how to improve communications about errors with annotators



# Questions?

i5k Workspace@NAL:

- <https://i5k.nal.usda.gov/>
- <https://github.com/NAL-i5K/>
- GFF3 toolkit issue tracker: <https://github.com/NAL-i5K/GFF3toolkit/issues>
- Email: i5k@ars.usda.gov



# Thank you!

## The NAL Team

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

## i5k Workspace alumni

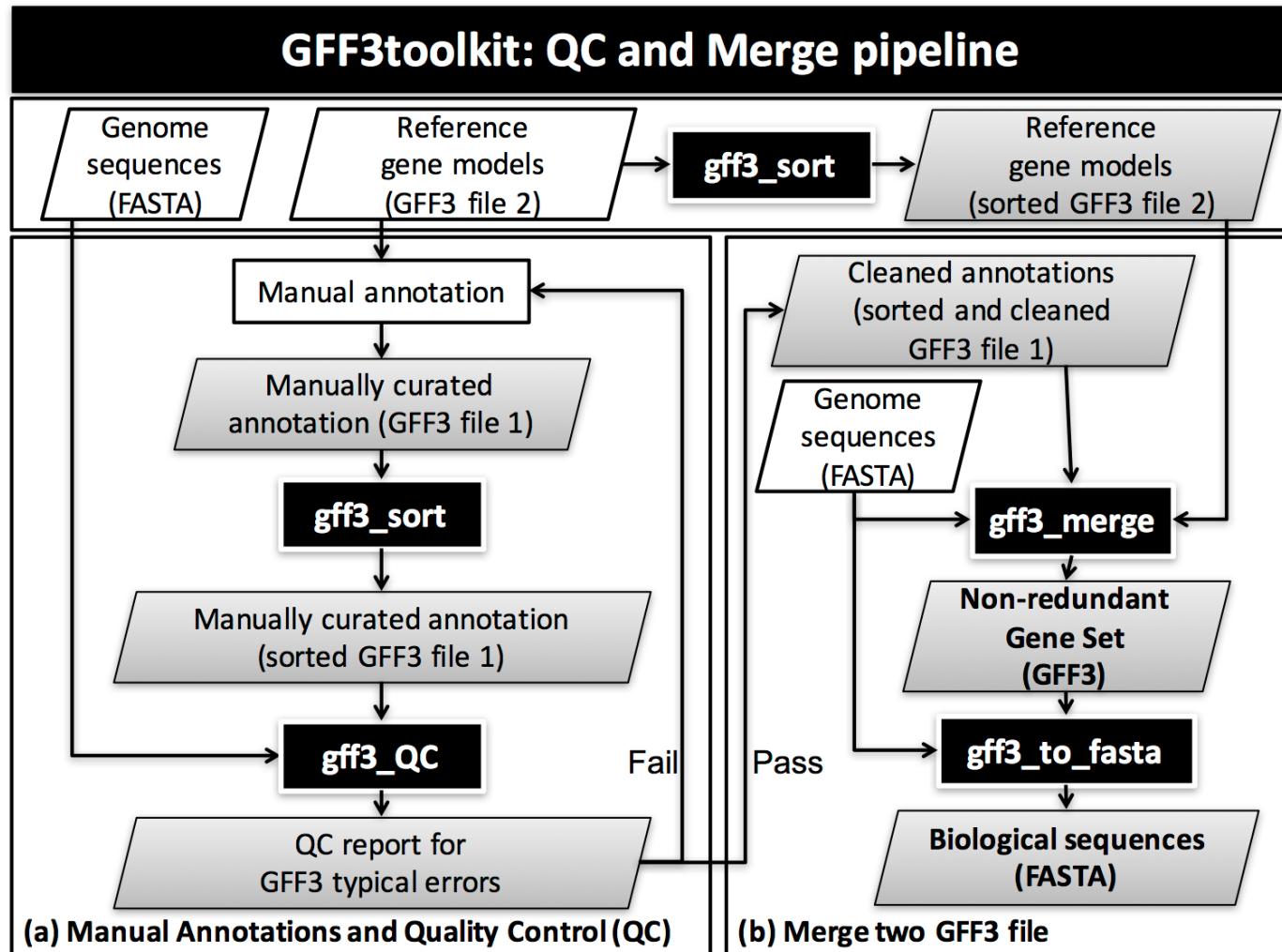
- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli
- Mei-Ju Chen
- Chao-I Tuan

## i5k Workspace@NAL advisory committee

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

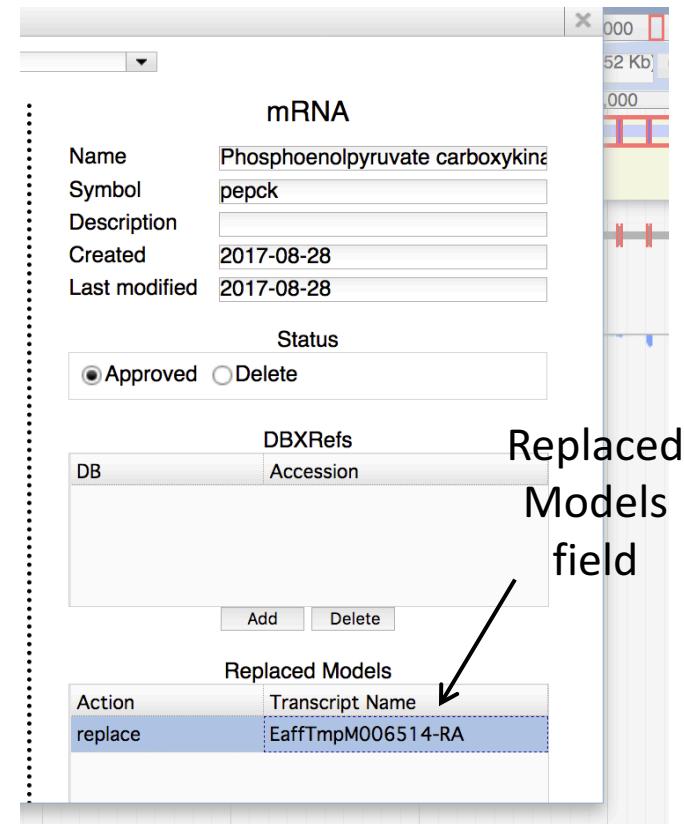


# OGS generation – the GFF3toolkit



# The Replaced Models field

- We use the information in this field to generate a merged, non-redundant gene set from the manually curated models and the official or primary gene set
- Your official or primary gene set is listed in the category field of the track selector
- If you don't know what your project's gene set is, contact us!



<https://i5k.nal.usda.gov/apollo-replaced-models-field-explanations-and-examples>

# Community annotation life cycle (end goal: OGS)

