The i5k Workspace@NAL - a Genome Database for Arthropods

Monica Poelchau[1], Li-Mei Chiang[2], Yi Hsiao[2], Yu-Yu Lin[2], Christopher Childers[1]

[1]USDA/Agricultural Research Service/National Agricultural Library, Beltsville, MD,

[2]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

What is the i5k Workspace@NAL?

- https://i5k.nal.usda.gov
- A workspace for **genomic data access, dissemination, and curation** for **any 'orphaned' arthropod genome project**, hosted by the USDA's National Agricultural Library (NAL)[1].
- We provide a central **organism page** for each project, **gene pages** for projects with an Official Gene Set, **data downloads**, a **BLAST**[2] search engine, the **JBrowse**[3] genome browser, and the **Apollo**[4] manual curation tool.

I5k Workspace by the numbers

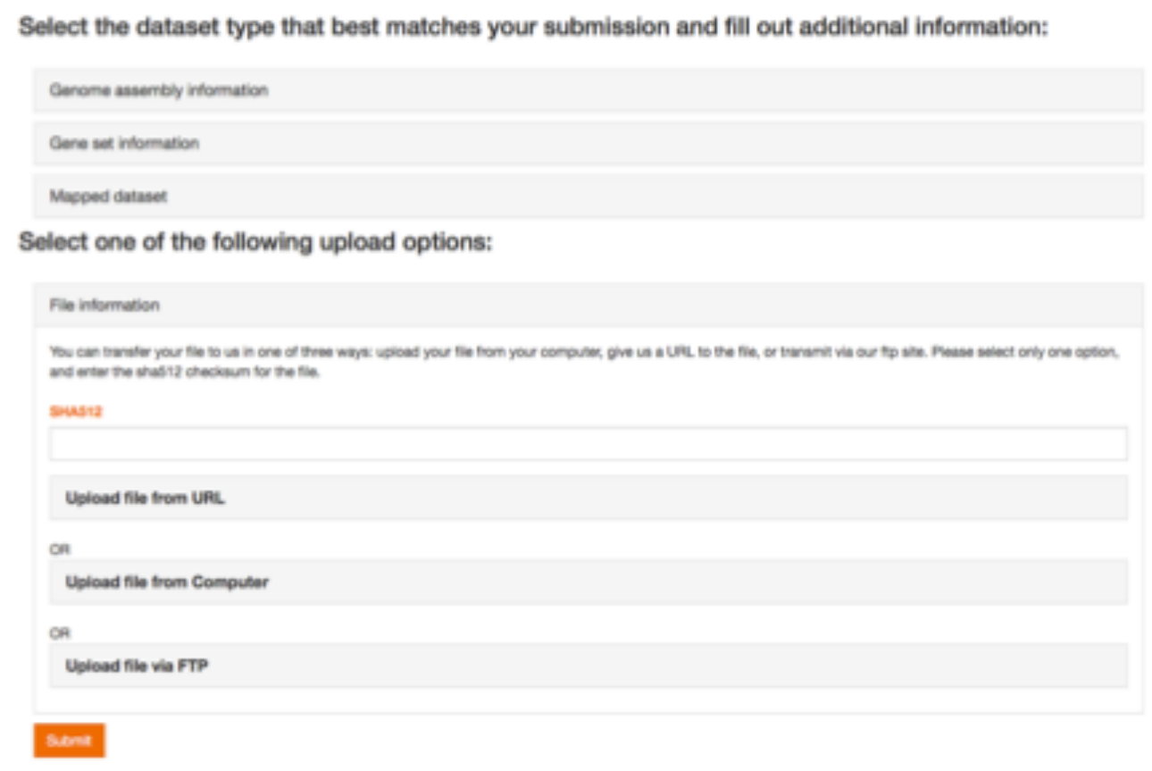
Metric Name	FY2018 - Q1	FY2018 - Q2
# of organisms hosted (cumulative)	61	63
# of registered users (cumulative)	490	499
# of pageviews	27,582	23,781
# of annotations created	884	1,682
# of active annotators	19	30

Table 1. Metrics describing the usage of the i5k Workspace@NAL.

Submit your data

- **Any orphaned arthropod genome project** in need of manual curation or other genome portal resources can start an i5k Workspace project.
- Users can also submit data to existing projects.
- Our main requirements are: 1) your assembly is accessioned by NCBI/INSDC; 2) datasets need to be mapped to the genome assembly

1. Register for an account
2. Request a new organism
3. Submit (meta)data (Fig. 1)
4. Upload files
5. Most file formats accepted



Select the dataset type that best matches your submission and fill out additional information:

Genome assembly information

Gene set information

Mapped dataset

Select one of the following upload options:

File information

You can transfer your file to us in one of three ways: upload your file from your computer, give us a URL to the file, or transmit via our ftp site. Please select only one option, and enter the sha512 checksum for the file.

SHA512

Upload file from URL

OR

Upload file from Computer

OR

Upload file via FTP

Submit

Figure 1. Screenshot of the i5k Workspace data submission portal.

Visit us online

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

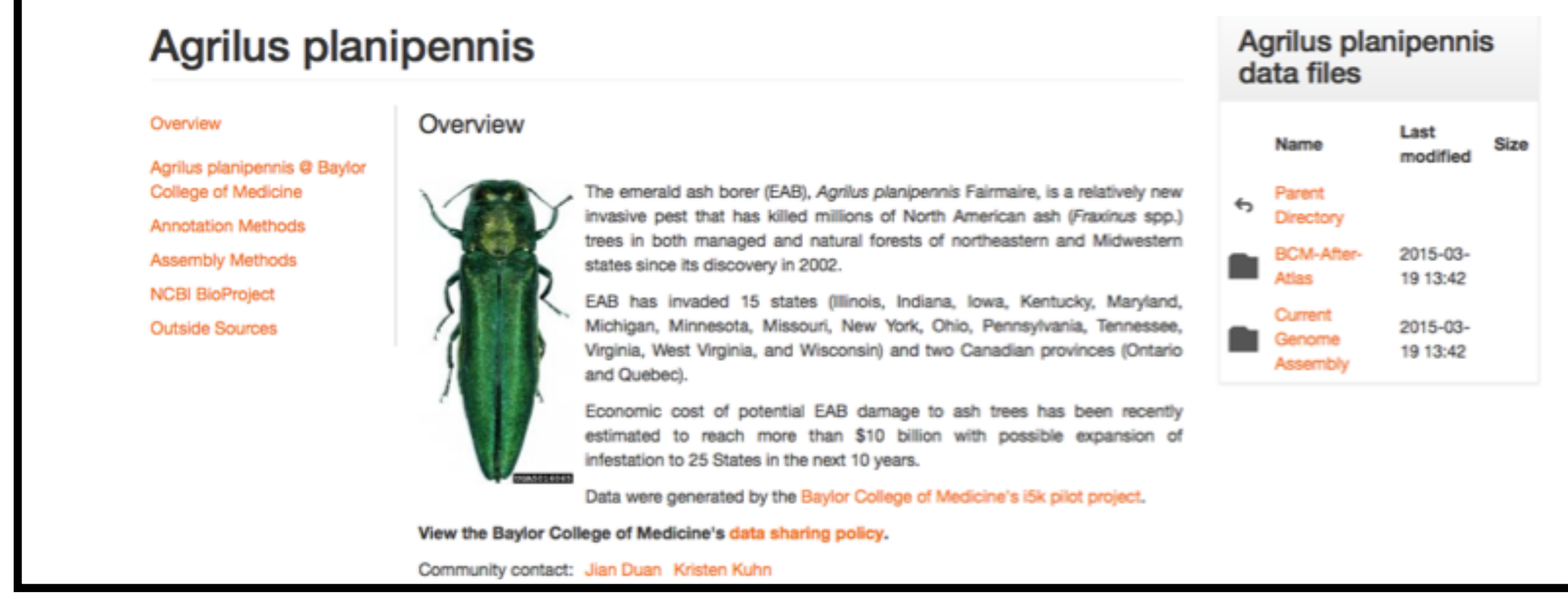
Email: i5k@ars.usda.gov

Acknowledgments and Funding

We would like to thank our data providers, the i5k coordinating committee, NAL leadership, and the NAL Information Systems Division team for their support and encouragement of this project. United States Department of Agriculture–Agricultural Research Service provided project support through the offices of the National Agricultural Library; Office of National Programs; and the Bee Research Laboratory.

Features

- **Exposure for your data.** All i5k Workspace projects are set up with an organism page and data downloads (Fig.2) using the Tripal[5] software
- For Official Gene Sets, we can set up gene pages



Agrilus planipennis

Overview

Agrilus planipennis @ Baylor College of Medicine

Annotation Methods

Assembly Methods

NCBI BioProject

Outside Sources

Overview

The emerald ash borer (EAB), *Agrilus planipennis* Fairmaire, is a relatively new invasive pest that has killed millions of North American ash (Fraxinus spp.) trees in both managed and natural forests of northeastern and Midwestern states since its discovery in 2002.

EAB has invaded 15 states (Illinois, Indiana, Iowa, Kentucky, Maryland, Michigan, Minnesota, Missouri, New York, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia, and Wisconsin) and two Canadian provinces (Ontario and Quebec).

Economic cost of potential EAB damage to ash trees has been recently estimated to reach more than $10 billion with possible expansion of infestation to 25 States in the next 10 years.

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

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

Community contact: Jian Duan Kristen Kuhn

Agrilus planipennis data files

Name	Last modified	Size
Parent Directory		
BCM-After-Atlas	2015-03-19 13:42	
Current Genome Assembly	2015-03-19 13:42	

Figure 2. Example organism page with data downloads.

Tools – Sequence search and alignment

- Your sequence datasets will be searchable via BLAST[2] (Fig. 3) and HMMER[6]; Clustal[7,8] is available for alignments
- Web applications are built using the Django framework (see our companion poster)

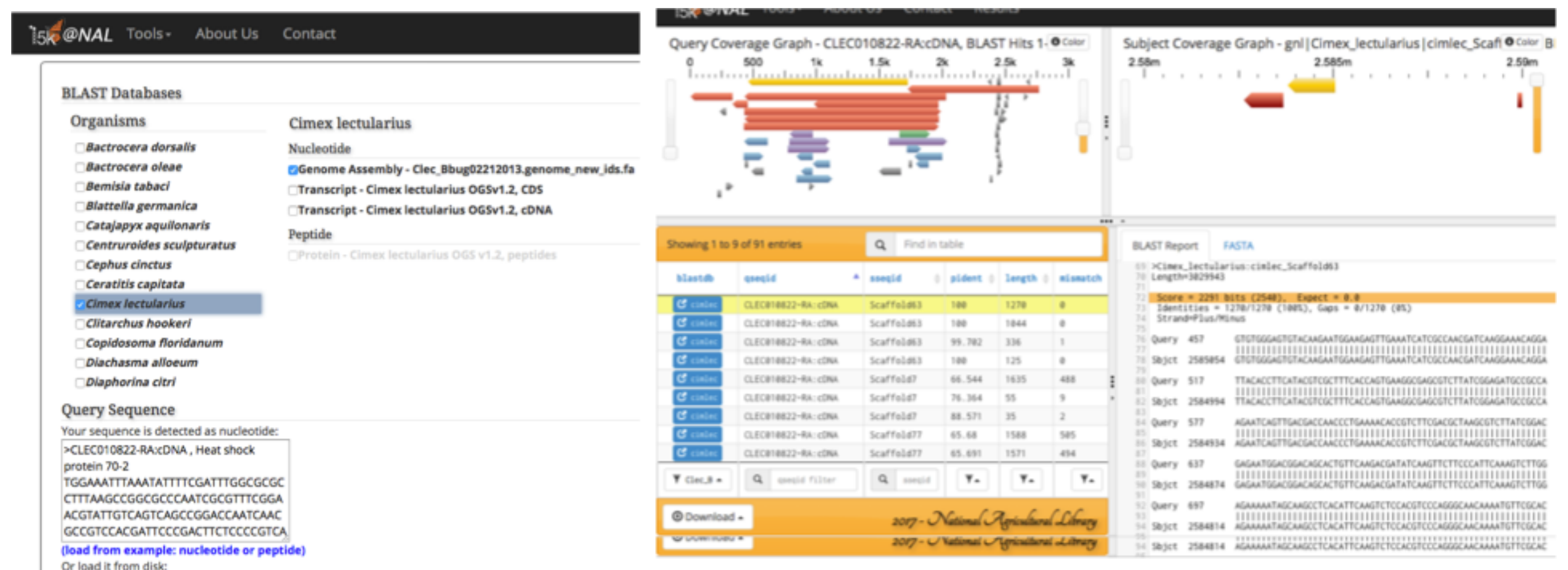


Figure 3. BLAST query and result pages.

Tools – Jbrowse and Apollo

- All I5k Workspace projects are set up with the JBrowse genome browser[3] and the Apollo manual annotation tool[4] (Fig. 4). We have tutorials, webinars, naming guidelines, and other resources to support manual annotation.

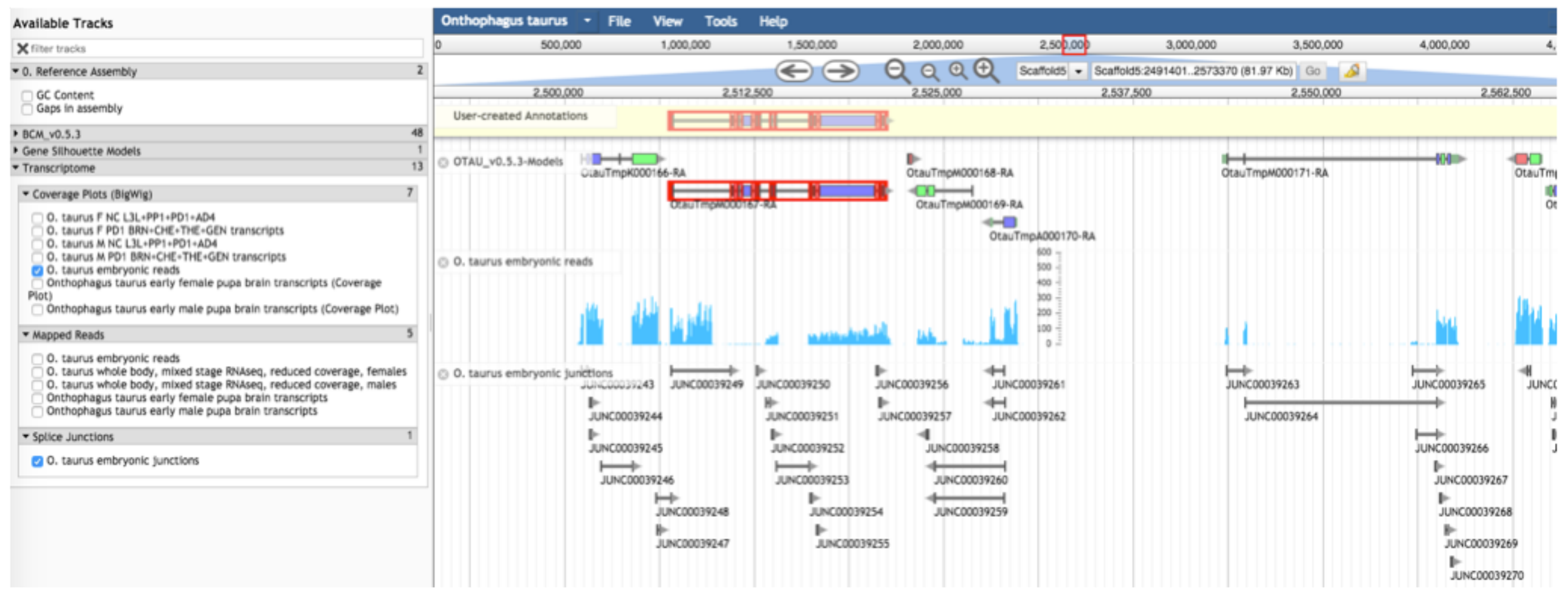
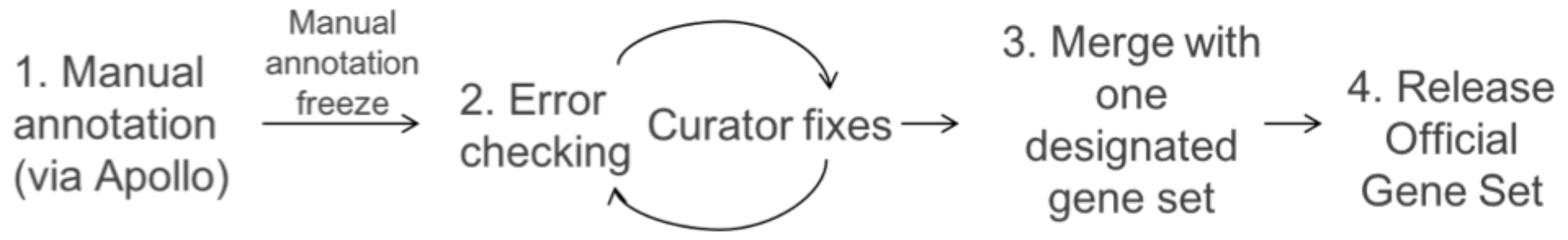


Figure 4. The Apollo manual annotation software.

Services – OGS generation

- Genome communities often want to generate a non-redundant "Official Gene Set" incorporating both manual and automated annotations.
- We developed the GFF3toolkit software to generate Official Gene Sets (OGS's). https://github.com/NAL-i5K/GFF3toolkit
- Updates to the GFF3toolkit, including a new function to fix errors in gff3 files, are in our companion poster
- We have facilitated the generation of 9 Official Gene Sets (OGS's) so far



1. Manual annotation (via Apollo) → Manual annotation freeze → 2. Error checking ⇄ Curator fixes → 3. Merge with one designated gene set → 4. Release Official Gene Set

Services – Annotation updates (in progress)

- Many of the i5k Workspace genome assemblies are being updated. We have developed a workflow to easily update annotation datasets in gff3 format to new assembly coordinates.
- The workflow uses NCBI's whole-genome alignment service and the CrossMap software[9]. The novel feature of this workflow is the easy reconstruction of gene models in gff3 format, which typically break using CrossMap
- https://github.com/NAL-i5K/remap-gff3

	Gerris buenoi[10]	*Leptinotarsa decemlineata*[11]
# of original gene models	21,105	24,837
# of scaffolds on assembly v1	20,259	24,393
# of scaffolds on assembly v2	18,844	26,908
% of whole-genome alignments used	93.84	95.60
% of gene models retained after crossmap	37.51	48.41
% of gene models retained after remap-gff3	90.62	75.11

Table 2. Genome annotation metrics before and after coordinate conversion between genome assembly versions.

References

1. Poelchau, MF, *et al.* (2014) The i5k Workspace@NAL – enabling genomic data access, visualization, and curation of arthropod genomes. *Nucl. Acids Res.* doi:10.1093/nar/gku983
2. Camacho, C., *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
3. Skinner, M.E., *et al.* (2009) JBrowse: A next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
4. Lee, E., *et al.* (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*, **14**, R93.
5. Ficklin, S.P., *et al.* (2011) Tripal: a construction Toolkit for Online Genome Databases. *Database*: bar044.
6. Eddy, S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Informatics* **23**(1):205-11.
7. Larkin, M.A., *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23.21**: 2947-2948.
8. Sievers, F., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**:539
9. Zhao, H., *et al.* (2013) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* btt730.
10. Armisen, David, *et al.* (2018) The genome of the water strider *Gerris buenoi* reveals expansions of gene repertoires associated with adaptations to life on the water. *bioRxiv* 242230
11. Schoville, Sean D., *et al*. (2018) A model species for agricultural pest genomics: the genome of the Colorado potato beetle, *Leptinotarsa decemlineata* (Coleoptera: Chrysomelidae). *Scientific reports* 8.1: 1931.