The i5k Workspace@NAL - Updates and new developments of an arthropod genome portal

Monica Poelchau[1], Christopher Childers[1], Gary Moore[1], Vijaya Tsavatapalli[1], **Ursula Pieper**[1], Mei-Ju May Chen[2], Yu-Yu Lin[3]

[1]USDA/Agricultural Research Service/National Agricultural Library, Beltsville, MD, [2]Genome and Systems Biology Degree Program, National Taiwan University and Academia Sinica, Taipei, Taiwan

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

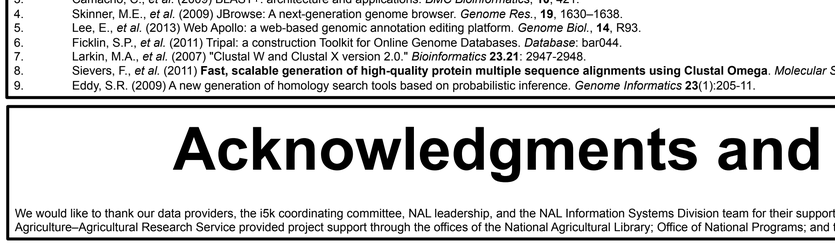
What is the i5k initiative?

- The **5,000 arthropod genomes initiative (i5k)** coordinates the sequencing of 5,000 insect or related arthropod genomes[1].
- International effort to seek funding from academia, governments, industry, and private sources; prioritize insect genomes for sequencing; develop best practices for genome sequencing and curation.

What is the i5k Workspace@NAL?

- A workspace for **genomic data access, dissemination, and curation** for **any 'orphaned' arthropod genome project**, hosted by the USDA's National Agricultural Library (NAL)[2].
- Any orphaned arthropod genome project in need of manual curation or other genome portal resources can submit their data.
- We provide a central **organism page** for each project, **gene pages** for projects with an Official Gene Set, **data downloads**, a **BLAST**[3] search engine, the **Jbrowse**[4] genome browser, and the **Web Apollo**[5] manual curation tool.
- We currently host genome project data from **43 arthropod species.**
- The i5k Workspace is built on a customized version of **Tripal**[6].
- URL: **https:/i5k.nal.usda.gov**

Updates: new user interface, search engine, gene pages



New user interface with easier navigation

New SOLR search engine for better result retrieval

Customized gene pages that display functional and sequence information

References

1. i5K Consortium (2013) The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. *J. Hered.*, **104**, 595–600.
2. 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
3. Camacho, C., *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
4. Skinner, M.E., *et al.* (2009) JBrowse: A next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
5. Lee, E., *et al.* (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*, **14**, R93.
6. Ficklin, S.P., *et al.* (2011) Tripal: a construction Toolkit for Online Genome Databases. *Database*: bar044.
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. Eddy, S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Informatics* **23**(1):205-11.

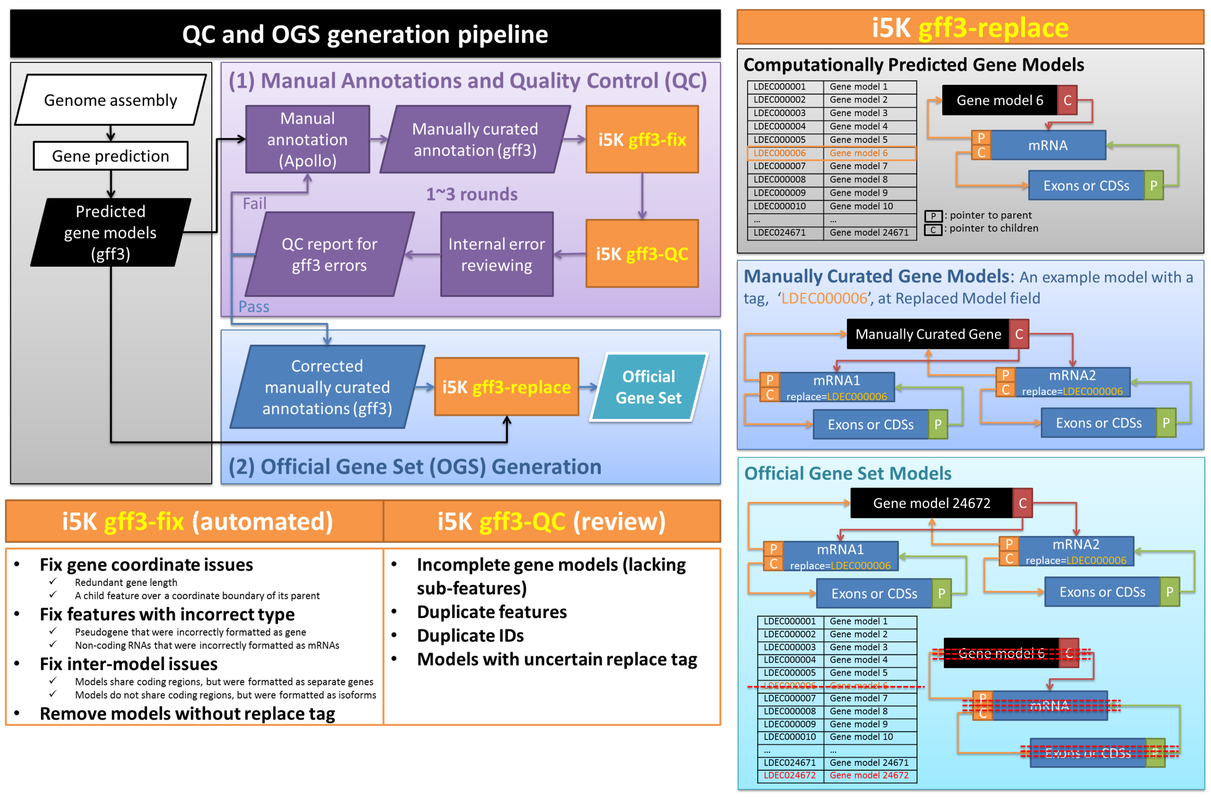
Acknowledgments and Funding

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.

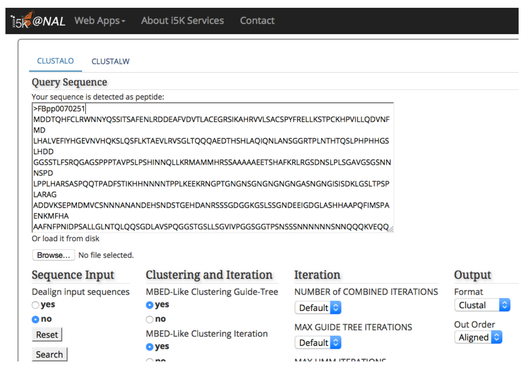
URL: https:/i5k.nal.usda.gov
Contact: i5k@ars.usda.gov

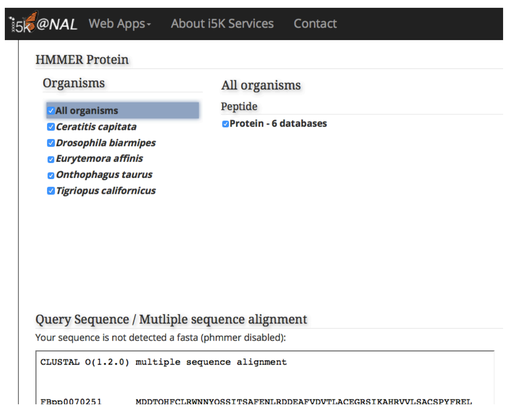
Coming soon: QC and OGS generation pipeline

- The pipeline incorporates 1) automated and manually-reviewed quality control (QC) of formatting errors caused by manual curation via the programs gff3-fix and gff3-QC; and 2) merging computationally predicted gene models with manually curated gene models into an Official Gene Set (OGS) via the program gff3-replace.
- The merge relies on curator-supplied information in a mandatory 'Replaced Models' Web Apollo[5] field that specifies which gene models from the computationally predicted gene set should be replaced by the manually curated model.



Coming soon: Clustal and Hmmer web applications





ClustalW[7], ClustalOmega[8], and Hmmer[9] web applications are under development to provide more sequence search options. Web applications are built in a Django framework. Features include: automatic input file format detection; query queuing system; and user accounts for result retrieval.