Community annotation across 26 non-model arthropod species

Monica Poelchau[1], Mei-Ju May Chen[2], Yu-Yu Lin[2], Chien-Yueh Lee[2], Monica Munoz-Torres[3], Stephen Richards[4], 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; [3]Lawrence Berkeley National Laboratory, Berkeley, CA; [4]Baylor College of Medicine, Houston, TX

Community annotation in non-model organisms

- Manual annotation can improve the value and accuracy of computationally predicted gene models
- Non-model genomes with small research communities usually lack resources to hire dedicated curators for manual annotation
- Community annotation can harness the expertise of the scientific community for genomes with fewer curation resources
- **Can community manual annotation of non-model genomes be performed across many genomes?**
- **What level of manual annotation can a distributed community achieve?**

Methods: Community annotation in the I5k pilot project

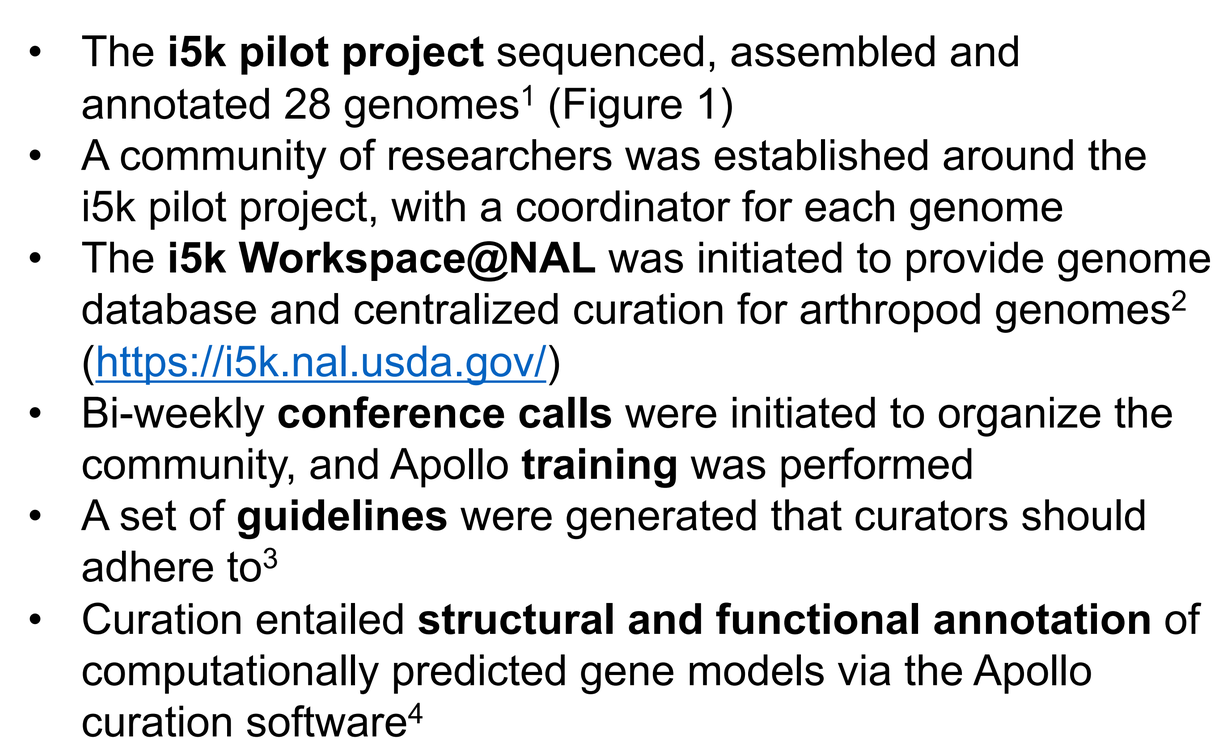

Figure 1. General community manual annotation workflow for the i5k pilot project.

- The **i5k pilot project** sequenced, assembled and annotated 28 genomes[1] (Figure 1)
- A community of researchers was established around the i5k pilot project, with a coordinator for each genome
- The **i5k Workspace@NAL** was initiated to provide genome database and centralized curation for arthropod genomes[2] (https://i5k.nal.usda.gov/)
- Bi-weekly **conference calls** were initiated to organize the community, and Apollo **training** was performed
- A set of **guidelines** were generated that curators should adhere to[3]
- Curation entailed **structural and functional annotation** of computationally predicted gene models via the Apollo curation software[4]
- **Manual annotation results from 26** out of 28 **i5k pilot genomes** from Apollo were evaluated
- The types of changes that occurred during the manual curation process in three completed projects were evaluated with the gff-cmp-cat software[5]

Acknowledgements

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. Apollo is supported by NIH (NIGMS, NHGRI), and by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under contract with University of California, Berkeley.

Results

Annotator behavior

- We recruited over 200 curators (Fig. 2), 35% of whom worked on >1 organism (Fig. 3)

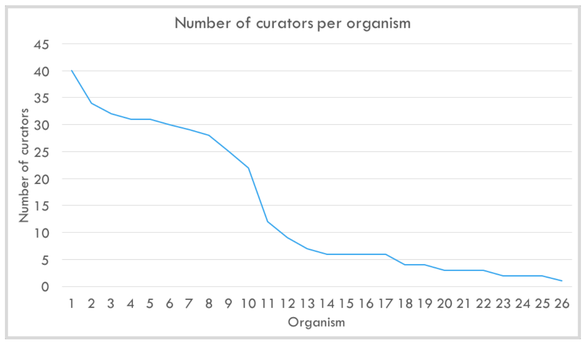

Figure 2. The number of curators per organism demonstrates a variety of community sizes, with some organisms having quite robust curation communities.

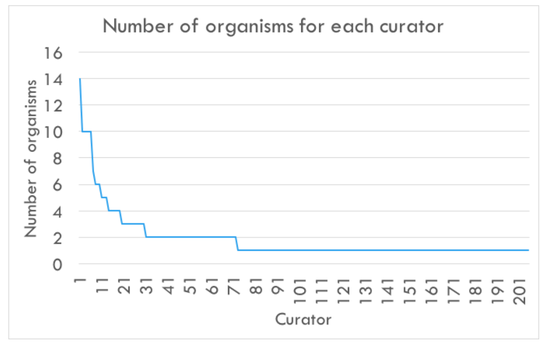

Figure 3. 35% of curators work across more than one organism, demonstrating the utility of hosting multiple arthropod genomes in one centralized database.

Manual annotation results

- Over a 4-year period, annotators for 26 organisms generated 16,647 annotations (annotator; 452/organism; Figure 4)
- In three organisms where one round of manual annotation is fully completed, 75% of curated models had structural modifications (Table 1)

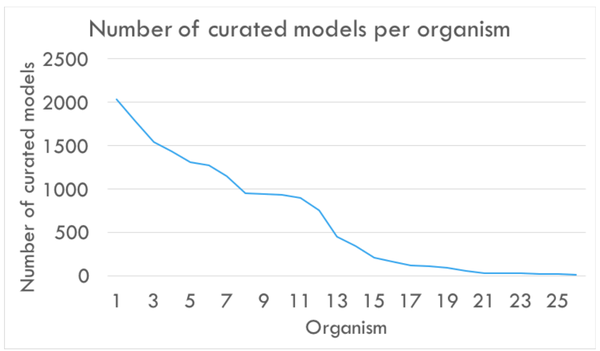

Figure 4. Number of manually annotated gene models per organism. Annotation activity differs greatly among communities.

organism	Models with structural changes	Total number of curated models	Proportion structurally changed models/total
Anoplophora glabripennis[6]	863	1144	0.75
Cimex lectularius[7]	1026	1354	0.76
Oncopeltus fasciatus	1159	1518	0.76

Table 1. The number of structurally changed models vs. total number of curated models for three organisms. The extent of structural modifications to computationally predicted gene models performed by the annotation community suggests that when manual annotation was performed, improvements of the basic gene structure were necessary 75% of the time.

Conclusions

- Large communities can manually annotate across non-model organisms, given the appropriate setup
- The nature of manual changes can be assessed computationally, but identifying the biological validity of the changes to the manually annotated models at scale is challenging

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. https://i5k.nal.usda.gov/content/rules-web-apollo-annotation-i5k-pilot-project
4. Lee, E., et al. (2013) Web Apollo: a web-based genomic annotation editing platform. Genome Biol., 14, R93.
5. https://github.com/chienyuehlee/gff-cmp-cat
6. McKenna, Duane D., et al. "Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle–plant interface." *Genome Biology* 17.1 (2016): 227.
7. Benoit, Joshua B., et al. "Unique features of a global human ectoparasite identified through sequencing of the bed bug genome." *Nature communications* 7 (2016).