The 5k Workspace@NAL: Introduction, overview and examples

Chris Childers and Monica Poelchau
USDA-ARS, National Agricultural Library
July 14, 2017
Overview

1. Background: What is the i5k Workspace?

2. Finding data at the i5k Workspace
   1. General search/Content types
   2. Data downloads
   3. BLAST
   4. Clustal(s)
   5. HMMER
   6. Jbrowse/Apollo

3. Improving data at the i5k Workspace via community annotation
The 5K Workspace@NAL

Our focus:

• We support any arthropod genome project:
  • Genome assembly needs to be in GenBank/ENA/DDBJ
  • Data should be open access (no private repositories)
• We enable and support community curation.
• We enable content search and retrieval
Overview

1. Background: What is the i5k Workspace?

2. Finding data at the i5k Workspace
   1. General search/Content types
   2. Data downloads
   3. BLAST
   4. Clustal(s)
   5. HMMER
   6. JBrowse/Apollo

3. Improving data at the i5k Workspace via community annotation
Finding Data at the i5k Workspace

• We have different kinds of information to search for:
  • Information about each i5k Workspace project (project metadata)
  • For Official Gene Sets: Gene names, gene metadata ("Feature")
  • Sequence data
  • Flat files (bulk data downloads)
Organism pages

Megachile rotundata

Overview

The alfalfa leafcutting bee, *Megachile rotundata* F. (Megachilidae) is a Eurasian solitary bee species that was inadvertently introduced to North America sometime before the 1940s. By the mid 1950s, *M. rotundata* had become established in the farming regions of western United States. With the discovery of *M. rotundata*’s pollination impact on alfalfa seed production, early efforts to increase its populations near alfalfa fields were undertaken a few years later. Currently, *M. rotundata* is the most intensely managed solitary bee species in the world and is surpassed only by the honey bee for its economic impact.

Females are gregarious cavity nesters constructing nests composed of leaf pieces in a linear series of cells in naturally occurring cavities or in artificial nesting boards. In most North American latitudes, *M. rotundata* emerge in late June and early July. Females provision each cell with nectar and pollen, lay a single egg and seal the cell before starting the construction of the next cell. The larvae will develop through five larval instars, spin a cocoon and enter a prepupal diapause and overwinter. A portion of the larvae laid in early spring will avert diapause and produce a second generation of bees. The second generation is problematic to farmers. Depending on the length of the growing season the larvae of the summer generation of females may enter diapause and overwinter. Therefore, in some years the second generation will result in an increased number of bees entering diapause and available for the next growing season. But if the growing season is too short, the larvae will not have sufficient time to complete development and will not be able to enter diapause. Besides influencing the total number of bees entering diapause each year, the second
Organism pages

latitudes, *M. rotundata* emerge in late June and early July. Females provision each cell with nectar and pollen, lay a single egg and seal the cell before starting the construction of the next cell. The larvae will develop through five larval instars, spin a cocoon and enter a prepupal diapause and overwinter. A portion of the larvae laid in early spring will avert diapause and produce a second generation of bees. The second generation is problematic to farmers. Depending on the length of the growing season the larvae of the summer generation of females may enter diapause and overwinter. Therefore, in some years the second generation will result in an increased number of bees entering diapause and available for the next growing season. But if the growing season is too short, the larvae will not have sufficient time to complete development and will not be able to enter diapause. Besides influencing the total number of bees entering diapause each year, the second generation has been implicated as a major factor in the spread of chalk brood, the primary disease of *M. rotundata*. The development of a *M. rotundata* genome database is an important advancement for understanding basic physiology and disease management of this important pollinator.

Community contact: George Yocum, Karen Kapaheim, Hailin Pan

### Assembly Information

<table>
<thead>
<tr>
<th>Analysis Name</th>
<th>Megachile rotundata genome assembly MROT_1.0 (GCF_000220095.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software</td>
<td>SOAPdenovo Assembler (1.05)</td>
</tr>
<tr>
<td>Source</td>
<td>BioProject PRJNA66515</td>
</tr>
<tr>
<td>Date performed</td>
<td>2016-03-23</td>
</tr>
<tr>
<td>Materials &amp; Methods</td>
<td></td>
</tr>
</tbody>
</table>

### Statistics

<table>
<thead>
<tr>
<th>Assembly Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contig N50</td>
</tr>
<tr>
<td>Scaffold N50</td>
</tr>
<tr>
<td>GC Content</td>
</tr>
</tbody>
</table>

| Manual Annotations |
Gene pages (Official gene sets only)
Finding Data at the i5k Workspace

- Website search for metadata (e.g. search term “Anoplophora glabripennis”)
Bulk data downloads for full files

• From menu, select ‘Data -> Data Downloads’
  • https://i5k.nal.usda.gov/content/data-downloads, or
  • https://i5k.nal.usda.gov/data/
Sequence Search – BLAST+

• From menu, select ‘Tools -> BLAST’
  • https://i5k.nal.usda.gov/webapp/blast/

• Tutorial:
  • https://i5k.nal.usda.gov/content/blast-tutorial

• Example query:
  • http://flybase.org/cgi-bin/getseq.html?source=dmel&id=FBpp0070037&chr=3L&dump=PrecompiledFasta&targetset=translation
Sequence Search – BLAST+ Result

Result URL: https://i5k.nal.usda.gov/webapp/blast/0b0f5690f06b446f86005bcc9f1bd3d7
Sequence Search – BLAST+ Result
Sequence Search – BLAST+ Result

Result URL: https://i5k.nal.usda.gov/webapp/blast/0b0f5690f06b446f86005bcc9f1bd3d7
Sequence Search – links to JBrowse
Sequence Search – links to JBrowse
Sequence alignment – ClustalW and Clustal Omega

- From menu, select ‘Tools -> Clustal (beta)’
  - [https://i5k.nal.usda.gov/webapp/clustal/](https://i5k.nal.usda.gov/webapp/clustal/)

- Example query sequences:
Sequence alignment – ClustalW and Clustal Omega

Result URL:
https://i5k.nal.usda.gov/webapp/clustal/ed00819ab40441ca959eacdcbb78c0f5
Sequence search – HMMER

• From menu, select ‘Tools -> HMMER (beta)’
  • https://i5k.nal.usda.gov/webapp/hmmer/

• Example query sequences (fasta, restrict to <10):
  • http://www.orthodb.org/fasta?query=EOG091906CT&level=&species=&universal=&singlecopy=
Sequence search – HMMER Result

HMMER Success

Report Details

Jump To Dataset BDOR_v1-blast.proteins.fasta

<table>
<thead>
<tr>
<th>E-value</th>
<th>Score</th>
<th>Bias</th>
<th>E-value</th>
<th>Score</th>
<th>Bias</th>
<th>exp</th>
<th>N</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4e-220</td>
<td>731.0</td>
<td>47.0</td>
<td>5.4e-220</td>
<td>731.8</td>
<td>42.5</td>
<td>1.9</td>
<td>1</td>
<td>gnl</td>
<td>Bactrocera_dorsalis_protein_v1</td>
</tr>
<tr>
<td>9.4e-14</td>
<td>50.7</td>
<td>5.6</td>
<td>1.1e-09</td>
<td>37.2</td>
<td>0.3</td>
<td>2.1</td>
<td>1</td>
<td>gnl</td>
<td>Bactrocera_dorsalis_protein_v1</td>
</tr>
<tr>
<td>1.7e-09</td>
<td>36.6</td>
<td>0.8</td>
<td>4.9e-05</td>
<td>22.0</td>
<td>0.2</td>
<td>2.1</td>
<td>1</td>
<td>gnl</td>
<td>Bactrocera_dorsalis_protein_v1</td>
</tr>
<tr>
<td>1.7e-09</td>
<td>36.6</td>
<td>0.8</td>
<td>4.9e-05</td>
<td>22.0</td>
<td>0.2</td>
<td>2.1</td>
<td>1</td>
<td>gnl</td>
<td>Bactrocera_dorsalis_protein_v1</td>
</tr>
<tr>
<td>3.0e-08</td>
<td>32.5</td>
<td>0.8</td>
<td>3.9e-06</td>
<td>25.5</td>
<td>0.2</td>
<td>2.1</td>
<td>1</td>
<td>gnl</td>
<td>Bactrocera_dorsalis_protein_v1</td>
</tr>
<tr>
<td>4.7e-08</td>
<td>21.9</td>
<td>1.6</td>
<td>4.9e-05</td>
<td>21.9</td>
<td>0.2</td>
<td>2.3</td>
<td>2</td>
<td>gnl</td>
<td>Bactrocera_dorsalis_protein_v1</td>
</tr>
<tr>
<td>1.6e-07</td>
<td>30.1</td>
<td>4.5</td>
<td>2.5e-07</td>
<td>29.5</td>
<td>0.7</td>
<td>2.0</td>
<td>2</td>
<td>gnl</td>
<td>Bactrocera_dorsalis_protein_v1</td>
</tr>
</tbody>
</table>
Genome browser (JBrowse)

• From menu, select ‘Tools -> JBrowse/Apollo -> JBrowse/Apollo Organisms’
  • [link](https://i5k.nal.usda.gov/available-genome-browsers)
Genome browser (JBrowse)

- If you know the gene ID of your gene of interest, you can paste it into the JBrowse ‘Search’ bar
Overview

1. Background: What is the i5k Workspace?
2. Submitting data
3. Finding data at the i5k Workspace
   1. General search
   2. Data downloads
   3. BLAST
   4. Clustal(s)
   5. HMMER
   6. Jbrowse/Apollo
4. Improving data at the i5k Workspace via community annotation
Improving Data at the i5k Workspace via Manual Annotation

• What is manual annotation?
  • Verify or improve the biological validity of computationally predicted gene models
  • Assign function to gene models via comparative analysis

• Why manually annotate?
  • Automated gene predictions often contain errors
  • Improve gene models for specific analyses

• Apollo documentation:
  • [https://i5k.nal.usda.gov/content/rules-web-apollo-annotation-i5k-pilot-project](https://i5k.nal.usda.gov/content/rules-web-apollo-annotation-i5k-pilot-project)
  • [https://i5k.nal.usda.gov/manual-curation-example](https://i5k.nal.usda.gov/manual-curation-example)
Community curation at the i5k Workspace

Our support for community curation includes:

• Access to a large community of curators
• Tutorials, guidelines, webinars
• Registration mechanism for new annotators
• One-on-one support
• Software to evaluate changes between curated and original annotations (Chien-Yueh Lee, https://github.com/chienyuehlee/gff-cmp-cat)
Principles of community annotation

• Collaborative effort across many individuals, often in different time zones and countries
• We encourage annotators to work together to find the best solution
• We work with each project coordinator to facilitate communication and collaboration whenever possible.
Manual annotation life cycle (end goal: OGS)

Genome sequencing, assembly and annotation

Official Gene Set generation (Merge of manual annotations and reference gene set)

Community building: Conference calls and training

Manual annotation via Apollo

General QC (NAL)

Manual annotation ‘freeze’
Some Apollo notes

• We’re still using Apollo1 – Apollo2 has a slightly different interface
• Here, we’ll use our ‘Training’ applications
• Apollo credentials for training applications:
  • Username: demo
  • Password: demo
• To annotate on an actual project, you’ll need to register first:
  • From menu, select ‘Tools -> JBrowse/Apollo -> Apollo registration form’
  • https://i5k.nal.usda.gov/web-apollo-registration
  • Registration is only for the organisms that you select
Example workflow: alpha-catenin in the Colorado Potato Beetle

• From menu, select ‘Tools -> Training tools -> Training BLAST’
  • [URL] https://i5k.nal.usda.gov/training/webapp/blast

• Query sequence:
Example workflow: alpha-catenin in the Colorado Potato Beetle

Result URL: https://i5k.nal.usda.gov/training/webapp/blast/613bde5948cb450da1b1d2d891995c9d
Example workflow: alpha-catenin in the Colorado Potato Beetle

- To view HSP in the genome browser:
  - Go to result table on bottom left
  - Click on the blue box to the left of the best HSP result in the ‘blastdb’ column

Result URL: https://i5k.nal.usda.gov/training/webapp/blast/613bde5948cb450da1b1d2d891995c9d
Example workflow: alpha-catenin in the Colorado Potato Beetle

URL: https://tinyurl.com/ybf4ehld
Example workflow: alpha-catenin in the Colorado Potato Beetle

URL: https://tinyurl.com/y8688kgt
Example workflow: alpha-catenin in the Colorado Potato Beetle

URL: https://tinyurl.com/y8688kgt
Example workflow: alpha-catenin in the Colorado Potato Beetle

URL: https://tinyurl.com/y8688kgt
Example workflow: alpha-catenin in the Colorado Potato Beetle

URL: https://tinyurl.com/y8688kgt
Post-Annotation QC

• Manual annotations are run through our Quality Control pipeline

• Some issues need manual intervention
  • Missing required fields
  • Complex splits/merges
  • Incomplete models and those abandoned in process

• Some issues can be automatically corrected

• Iterative process
  • Models requiring inspection are referred back to curators
  • After resolution models are screened again to screen for additional issues
OGS (Official Gene Set) Generation

• An Official Gene Set is the gene set chosen by the community to be the representative set of gene models for that organism
• Our system takes a single existing gene set and incorporates the validated manual annotations
• The gene set may be a previous OGS or other gene set (e.g. Maker models)
• Manual curations are used to
  • Update models
  • Flag models for removal from the final set
• The resulting set is then tested for errors and once approved, disseminated to the community
OGS (Official Gene Set) Generation

• Requirements:
  • Designate a ‘reference gene set’ prior to the start of the annotation period
  • Use the ‘Replaced Models field’ during the manual annotation process
The i5k Workspace ‘Replaced Models’ field

- Accessible via the Information Editor
- Enter the name or ID of the reference gene model that your manually curated model replaces.
- Information is used to merge your annotation with reference gene set to make an OGS (Official Gene Set)
- More information:
Need more information?

i5k Workspace@NAL:
• [https://i5k.nal.usda.gov/](https://i5k.nal.usda.gov/)

Check us out on GitHub
• [https://github.com/NAL-i5K/](https://github.com/NAL-i5K/)

Questions?
• Email us [i5k@ars.usda.gov](mailto:i5k@ars.usda.gov)
• Leave us a note: [https://i5k.nal.usda.gov/contact](https://i5k.nal.usda.gov/contact)
Acknowledgements

The NAL Team
• Christopher Childers*
• Monica Poelchau*
• Gary Moore
• Susan McCarthy
• Chaitanya Gutta
• Yu-yu Lin

Workspace alumni
• Mei-Ju Chen
• Chien-Yueh Lee
• Han Lin
• Jun-Wei Lin
• Vijaya Tsavatapalli

i5k Workspace@NAL advisory committee
• Jay Evans
• Kevin Hackett
• Simon Liu
• Ursula Pieper

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