Using Apollo at the i5k Workspace@NAL

Monica Poelchau, USDA-ARS NAL
June 16th, 2020
Agenda

• Manual annotation general overview
• 15k Workspace tools for manual annotation
  – BLAST, Clustal, HMMER
  – Apollo2
• Manual annotation example: preparation
• Manual annotation live example
Other resources

• Monica Munoz-Torres from the Apollo group has a number of comprehensive tutorials:
  – https://www.slideshare.net/MonicaMunozTorres/presentations
    • I recommend these slides if you need more background:
      – https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015
    – If you are new to Apollo, or need a refresher, I highly recommend that you review one of her presentations
• The official Apollo annotation guide:
  – http://genomearchitect.org/users-guide/
• I5k Workspace manual annotation landing page:
• Other manual curation tutorials:
  http://genomencuration.github.io/genometrain/d-feature-curation-crossing/
• VEuPathDB Apollo training webinar:
  https://eupathdb.org/eupathdb/webinars.jsp#apollo
MANUAL ANNOTATION GENERAL OVERVIEW
What is manual annotation?

- Manual review and improvement of an existing gene prediction
- Draw on external evidence (e.g. RNA-Seq, cDNA, genes from other species) to improve a computationally predicted gene model
Why manually annotate?

• Automated gene predictions are not always correct
• “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
General process of manual annotation

1. Select a chromosomal region of interest (e.g. scaffold)
   1. E.g. find sequence of interest from one or several other species, and align against proteins or genome sequence from your species

2. Select appropriate evidence (tracks in Apollo, or your own files)

3. Determine whether a feature in your evidence provides a reasonable starting gene model
   1. If yes: select and drag the feature to the ‘user-created annotations’ area, creating an initial gene model. If necessary use editing functions to adjust the model.
   2. If not – get in touch with us!

4. Edit model if necessary

5. Check your edited gene model for integrity and accuracy by comparing it with available homologs
   1. Verify that the gene model is the best representation of the underlying biology

6. Repeat steps 1 through 5 as needed to refine model

7. Add annotation details in the “Information Editor”
   1. Name, symbol, other comments

Adapted from https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015
MANUAL ANNOTATION: 15K WORKSPACE TOOLS
First, some conventions

• HSP – High scoring pair in BLAST/BLAT alignments
  – The ‘Hits’ in an alignment result set
  – A subsection of a pair of sequences with sufficient score
  – HSPs can change based on the alignment parameters

• Five prime end and three prime end
  – Based on direction of transcription
  – Initiation site is at the five prime end
  – Stop codon is at the three prime end

• In the genome browser, arrowheads indicate direction
JBrowse and Apollo2

- **File**: Add your own files
- **View**: Change coloring scheme
- **Tools**: Search using BLAT
- **Locate where you are on the scaffold**: Apollo2 Track selector
- **Search for a gene or location**: Revert to ‘old’ track selector
- **Zoom in/out**: Log out
- **User-created annotations track**: Find information about tracks

**JBrowse is a web-based genome browser**
- Visualize features that are mapped to a genome
- These features are displayed as tracks
- Many different types of data may be displayed

**Apollo adds editing functions to JBrowse**
- Manual gene curation
- Changes automatically saved back to server
- Edits are visible to other annotators in real-time
- Editing history is tracked
Apollo2 – Annotations Panel

- Annotations panel
- Filter annotations
- View annotation overview
- Click on arrow to jump to annotation
- View functional annotation details
- Can modify individual features via ‘Coding’ Panel
Apollo2 – Ref Sequence Panel

- Reference sequence panel
- Filter sequences
- Export sequences/annotation gff3
- View reference sequence list
i5k Workspace BLAST: one way to access Apollo

URL: https://i5k.nal.usda.gov/webapp/blast/
i5k Workspace BLAST: one way to access Apollo

Click on blue blastdb icon next to your favorite HSP

Blast results are displayed in Apollo

BLAST result page with 4 panels
HMMER and Clustal

- Use HMMER to detect remote protein homologs
  - https://i5k.nal.usda.gov/webapp/hmmer/

- Use Clustal to perform multiple sequence alignments
  - https://i5k.nal.usda.gov/webapp/clustal/
Tips and Tricks

• The i5k Workspace BLAST results persist for one week
  – You can bookmark and share searches
  – BLAST HSPs are ‘draggable’ and can be used in annotations

• Jbrowse/Apollo URLs can be shared
  – Allow you to share the exact view (including active tracks) with others

• In Apollo, you can pin tracks to the top

• If you know the name or ID of the gene that you’d like to annotate, you can paste it into the search box in Apollo to navigate to it
# O. brunneus and R. dominica information

<table>
<thead>
<tr>
<th></th>
<th><strong>Odontomachus brunneus</strong></th>
<th><strong>Rhyzopertha dominica</strong></th>
<th><strong>General</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Apollo registration</td>
<td><a href="https://i5k.nal.usda.gov/web-apollo-registration">https://i5k.nal.usda.gov/web-apollo-registration</a></td>
<td>Email <a href="mailto:monica.poelchau@usda.gov">monica.poelchau@usda.gov</a></td>
<td><a href="https://i5k.nal.usda.gov/web-apollo-registration">https://i5k.nal.usda.gov/web-apollo-registration</a></td>
</tr>
<tr>
<td>Organism page</td>
<td><a href="https://i5k.nal.usda.gov/odontomachus-brunneus">https://i5k.nal.usda.gov/odontomachus-brunneus</a></td>
<td>NA</td>
<td><a href="https://i5k.nal.usda.gov/species">https://i5k.nal.usda.gov/species</a></td>
</tr>
<tr>
<td>Gene pages</td>
<td><a href="https://i5k.nal.usda.gov/search/site/Odontomachus%2520brunneus%2520AND%2520gene">https://i5k.nal.usda.gov/search/site/Odontomachus%2520brunneus%2520AND%2520gene</a></td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>RNA-Seq (NAL)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Functional annotation (NAL)</td>
<td>Not yet</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>BLAST URL</td>
<td><a href="https://i5k.nal.usda.gov/blast">https://i5k.nal.usda.gov/blast</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**O. brunneus and R. dominica information – RNA-Seq**

- From NAL: Odobru_Obru_v1_RNA-Seq-alignments_2020-06-02; Rhydom_RdoDt3_RNA-Seq-alignments_2020-06-09
- From data providers: RNA-Seq by tissue or developmental stage
- Includes coverage plots, splice junctions, and read mappings
- [https://github.com/NAL-i5K/NAL_RNA_seq_annotatio_n_pipeline/](https://github.com/NAL-i5K/NAL_RNA_seq_annotatio_n_pipeline/)
O. brunneus and R. dominica

information – functional annotation

- Provided in collaboration with AgBase
- GO, Kegg, Reactome
- Tab-delimited files available soon!
- We would love any kind of feedback on these new resources!

Reactome pathway information
GO information (IPRscan)
MANUAL ANNOTATION EXAMPLE:
PREPARATION
Annotation Example

• Glycerate kinase-like in the trap-jaw ant *Odontomachus brunneus*

• More information about the trap-jaw ant genome assembly: https://i5k.nal.usda.gov/odontomachus-brunneus

• *Odontomachus brunneus* Apollo URL: https://apollo.nal.usda.gov/apollo/4006447/jbrowse/index.html?
  – For an Apollo login, please register here: https://i5k.nal.usda.gov/web-apollo-registration
Choosing reference proteins: *D. melanogaster* glycerate kinase in UniProt

Annotation score is a heuristic for annotation quality

Flybase is another great resource

Retrieve FASTA from ‘sequence’ tab

Source: [https://www.uniprot.org/uniprot/Q9VQC4](https://www.uniprot.org/uniprot/Q9VQC4)
Choosing reference proteins: *Apis mellifera* glycerate kinase


Lots of additional information on function

FASTA available under ‘Transcript’ tab
Resources for learning about insect gene/protein structure and function

- UniProt: [https://www.uniprot.org/](https://www.uniprot.org/)
- OrthoDB: [https://www.orthodbd.org/](https://www.orthodbd.org/)
- FlyBase: [http://flybase.org/](http://flybase.org/)
- VectorBase: [https://www.vectorbase.org/](https://www.vectorbase.org/)
- AphidBase/BIPAA: [https://bipaa.genouest.org/is/](https://bipaa.genouest.org/is/)
MANUAL ANNOTATION LIVE EXAMPLE
BLAST dmel, amel proteins against *O. brunneus* genome

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

Results are filtered by e-value; we’re going to focus on the second-best match here
BLAST dmel, amel proteins against *O. brunneus* genome

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

BLAST results are displayed as glyphs in browser; can be used as annotation starting points if the alignment is high quality
Modify *O. brunneus* model sequence in Apollo

• Questions:
  – What evidence do you choose to check the integrity of the model?
  – Do you need additional evidence?
  – How do you evaluate whether the protein sequence is as complete as it can be?
  – Should you add/modify UTRs?
View available evidence

Very different coverage between UTR and CDS
View available evidence

Junction reads don’t support connection between the two expressed regions
Create new model in user-created annotations track

Drag evidence to UcA track (or right-click and select “create annotation”)

![Image of DNA sequence and annotations]
Split model

Select exons on which to split the model using the ‘shift’ key.

Right-click on the model while continuing to hold shift to get the drop-down menu.

Select ‘split’.
Split model

You now have 2 models! Let’s start fixing the model on the right – it needs a 3’ exon.
Add an exon

Zoom in, select the missing exon, drag up to UcA track
Merge exons

Shift-select both exons, shift-right click, then select ‘merge’ from the dropdown menu
Adjust gene boundaries

Adjust the 3’ UTR to match the RNA-Seq evidence
Next, the ‘left’ model
Adjust 5’ UTR

Adjust the 5’ UTR to match the RNA-Seq evidence
Set translation start at Methionine
Evaluate new protein sequence

• Blast modified sequence to NCBI’s nr database
  – Make sure it doesn’t match a potential contaminant
  – Get an idea whether you have the right sequence
  – Blastp home:

• Once contamination is ruled out, it’s better to align your sequence against a smaller set of high-quality proteins

• If you notice that parts of the protein are missing, check the ‘Gaps in assembly’ track in the browser
Using the Information Editor

Review our naming guidelines before naming:

Use the mRNA/transcript side of the IE
Using the Information Editor

• Select the model in Apollo, then right-click, and select ‘Edit Information’ from the drop-down menu
  – Use the ‘mRNA’ section
    • If a naming convention exists, use it (e.g. for gene families)
    • Name should be unique and attributed to all orthologs (as far as possible)
    • Use name from an orthologous protein if you are sure that your gene model is an ortholog.
    • Document your justification for the name in the Comments field (e.g. “88% sequence similarity via blastp to D. melanogaster pepck P20007”)  
  – Comments – Document what changes you performed, and your justification for the name. These notes will be visible in the OGS, so make sure that others understand them
Checklist for accuracy and integrity

• Check start, stop and exon boundaries (splice sites)
  – Try to fix non-canonical splice sites if possible
• Check if you can annotate UTRs (e.g. using RNA-Seq data)
• Check for gaps in the genome
• If you change the genome sequence, add a justification comment to the corresponding gene model
• Use BLAST or a multiple sequence aligner
  – To look at completeness of model
  – To verify the appropriateness of the gene name
• In the Information editor **mRNA** field
  – Update the Name if appropriate
  – Add comments that describe
    • your evidence for the annotation
    • Modifications that you made to the gene model

cf. [https://www.slideshare.net/MonicaMunozTorres/editing-functionality-apollo-workshop](https://www.slideshare.net/MonicaMunozTorres/editing-functionality-apollo-workshop)
What happens to my annotation when I’m done?

• This depends on the genome project that you’re working on.
  – For *O. brunneus* and *R. dominica* – we will assist with generating an OGS and submitting to GenBank
• If the genome coordinator has asked us to generate an OGS (Official Gene Set), we will do so
  – Includes submission to GenBank, where they will be archived/accessioned
  – This takes some time, but we are working on expediting it
• Otherwise, don’t assume that your annotation will be archived.
  – If you need it to be, get in touch with us and we’ll figure out what to do.
• Get in touch with us and the genome project coordinator if you’re not sure about the status of a genome project.
I5k Workspace ‘Etiquette’

1. Use Apollo to improve a gene model in an i5k Workspace assembly.
   1. If you just want to practice – use one of our training instances.
      1. [https://i5k.nal.usda.gov/jbrowseapollo-training](https://i5k.nal.usda.gov/jbrowseapollo-training)
   2. If you just want to view the data – you probably can get what you want without using Apollo. All of the data that we host is public.

2. Your annotation work is a community effort.
   1. If you notice that someone else is working on your model of choice, get in touch with them (or us) and collaborate – don’t make a 2nd model or delete the other model.
   2. Keep in mind that your work may be used by the scientific community once you’re done.

3. If you publish any of your work generated in the i5k workspace:
   1. Get in touch with the genome contact first (you can find the contact info on the organism page; [https://i5k.nal.usda.gov/species](https://i5k.nal.usda.gov/species));
   2. Please cite the i5k Workspace paper! This helps us continue to exist.
      1. [https://doi.org/10.1093/nar/gku983](https://doi.org/10.1093/nar/gku983)
Thank you!

The NAL Team
- Chris Childers
- Gary Moore
- Susan McCarthy
- Min-Chen Hsu
- Chun-Hung Lin
- Chia-Tung Wu

i5k Workspace alumni
- Yu-yu Lin
- Chaitanya Gutta
- Li-Mei Chiang
- Yi Hsiao
- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli
- Mei-Ju Chen
- Chao-I Tuan

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

Contact us:
- https://i5k.nal.usda.gov/contact
- i5k@ars.usda.gov
- Monica.Poelchau@usda.gov
- Christopher.Childers@usda.gov