Using Apollo at the i5k Workspace@NAL

Monica Poelchau, USDA-ARS NAL
November 23rd, 2021
Agenda

• Manual annotation general overview
• I5k 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:

• I5k Workspace manual annotation landing page: https://i5k.nal.usda.gov/manual-annotation-and-apollo

• Other manual curation tutorials: http://genomecuration.github.io/genometrain/d-feature-curation-crossing/

• VEuPathDB Apollo training webinar: https://veupathdb.org/veupathdb/app/static-content/webinars.html#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
MANUAL ANNOTATION: I5K 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

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
- Changes automatically saved back to server
- Edits are visible to other annotators in real-time
- Editing history is tracked
JBrowse and Apollo2

- **File:** Add your own files
- **View:** Change coloring scheme
- **Zoom in/out**
- **Locate where you are on the scaffold**
- **Search for a gene or location**
- **Apollo2 Track selector**
- **Revert to ‘old’ track selector**

**User-created annotations track**

**Find information about tracks**

**Log out**
Apollo2 – Annotations Panel

Annotations panel

Filter annotations

Select Name to view and edit details

Click on arrow to jump to annotation in browser

Create a link to the annotation to share with collaborators

Edit information – this section replaces the ‘Information Editor’ in previous Apollo versions
Apollo2 – Ref Sequence Panel

- Reference sequence panel
- Filter sequences
- Export sequences/annotation gff3
- View reference sequence list
Apollo2 – Blat Search

Select protein or nucleotide

Paste sequence

Blat search panel

Blat search results

Create Annotation from Blat results
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
HMMER and Clustal

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

- Use Clustal to perform multiple sequence alignments
  - [https://i5k.nal.usda.gov/webapp/clustal/](https://i5k.nal.usda.gov/webapp/clustal/)
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](https://i5k.nal.usda.gov/odontomachus-brunneus)
  - For an Apollo login, please register here: [https://i5k.nal.usda.gov/web-apollo-registration](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

Source: 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/
• OrthoDB: https://www.orthodb.org/
• FlyBase: http://flybase.org/
• VectorBase: https://www.vectorbase.org/
• Hymenoptera Genome Database: http://hymenopteragenome.org/
• AphidBase/BIPAA: https://bipaa.genouest.org/is/
MANUAL ANNOTATION EXAMPLE
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/webapp/blast/
BLAST dmel, amel proteins against *O. brunneus* genome

BLAST results are displayed as glyphs in browser; can be used as annotation starting points if the alignment is high quality

https://i5k.nal.usda.gov/webapp/blast/
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
Create new model in user-created annotations track

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

The information editor is now in the ‘Annotations’ panel.

Sync name with transcript to propagate gene name.

All information now pertains to gene (not mRNA).

Add name to gene. Use the i5k Workspace naming guidelines.

Using the Information Editor

• Add information about the model in the ‘Annotations’ panel.
    • 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
I5k Workspace Guidelines - Names

Are you adopting a name from a homolog?

- You can re-use existing, established names (e.g. from *Drosophila melanogaster*).
- Don’t add a species prefix (although okay to use in your manuscript for clarity).
- If you want to imply uncertainty, you can append ‘-like’ to the name.
  - Good: “Ultraspiracle”
  - Okay: “Ultraspiracle-like”
  - Bad: “Clec-ultraspiracle” or “similar to ultraspiracle”

I5k Workspace Guidelines - Names

• Are you naming an isoform?
  – use the suffix “isoform A”, “isoform B”, etc.

• Are you naming a fragmented gene?
  – include a comment 'Part X of Y', where Y is the total number of fragments, and X is the ordinal number for that gene.
  – Don’t add ‘partial’ or ‘part of’ to the name.

I5k Workspace Guidelines - Names

• Are you naming a ‘new’ gene?
  – Choose a name that could be propagated to all orthologous proteins; try not to make it species- or tissue-specific
    • **Good:** “magnesium transporter”
    • **Bad:** “diapause-associated protein”

• Are you naming a gene from a gene family?
  – Check if a naming system already exists: [http://www.uniprot.org/docs/nomlist.txt](http://www.uniprot.org/docs/nomlist.txt)
  – Use Arabic numbers to specify the different members encoded by a multigene family.
    
I5k Workspace Guidelines - Symbols

- Are abbreviations of the descriptive gene name.
- We do not recommend coining new symbols for newly named genes.
- However, if a name from an orthologous gene was adopted, you may use this gene’s symbol, as well.
- Don’t use species prefixes (e.g. Clec-Pepck)
- Examples: Pepck, Ser12

Using the Information Editor

Comments are still welcome

Edit here, then add
Using the Information Editor

Please *don’t* use the new GO, Gene product, provenance, and attributes tabs.

Cross-references are okay, but only for the same transcript in the same species.
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
  – 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
What happens to my annotation when I’m done?

• This depends on the genome project that you’re working on.
• 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!

**I5k Workspace team:**
- Sean Buehler
- Chris Childers
- Vern Chapman
- Ming Chan
- Amanda Cooksey
- Chin Dai
- Zhi-Xuan Lai
- Monica Poelchau

**Contact us:**
- [https://i5k.nal.usda.gov/contact](https://i5k.nal.usda.gov/contact)
- i5k@ars.usda.gov