I5k Workspace webinar: Naming and BLAST use for manual annotation

9/19/2023

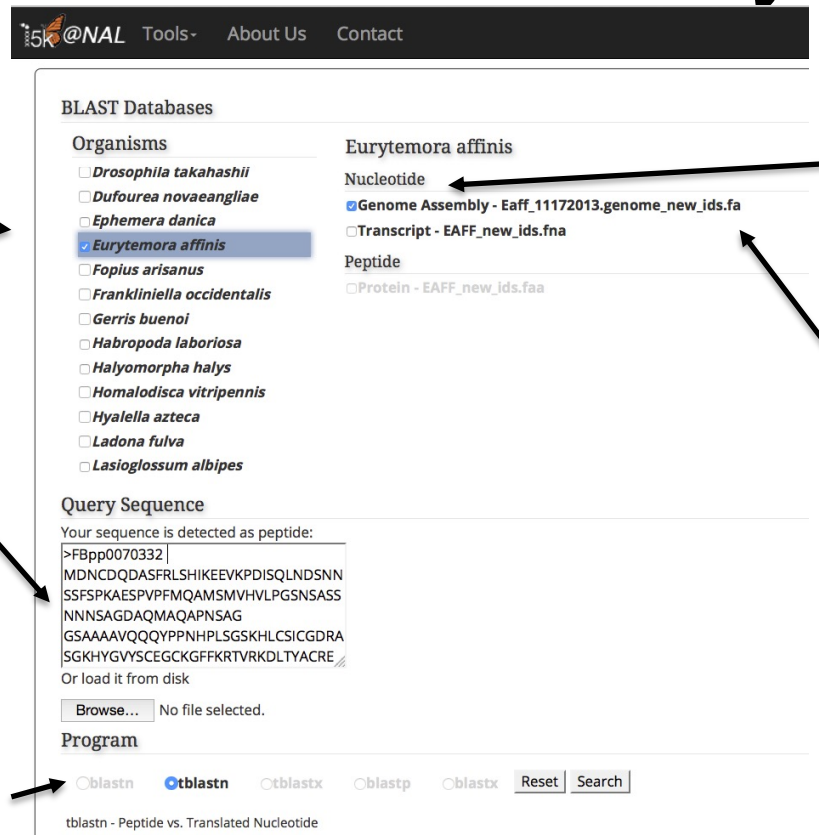
Agenda

1. i5k BLAST tool review: how to interpret BLAST results in JBrowse
2. I5k Workspace naming guidelines
3. Q&A

Other Apollo training resources

- Our previous Apollo tutorials go in-depth on how to use various Apollo functions.
- Using the Apollo2 manual annotation tool:
  - Slides: https://i5k.nal.usda.gov/sites/default/files/presentations/apollo_training_november_2021.pdf
  - Recording: https://www.zoomgov.com/rec/share/8ZxP3FhFE9ahBpVTUFDP8HwKNUS1prZlhAflINmrQaZMKeVoyo4F97RtmwJYdwQK.936mXuEkaOcjfCDW
    Passcode: hm3tu1A@
- In-depth annotation techniques:
  - Slides: https://i5k.nal.usda.gov/sites/default/files/presentations/Apollo_webinar_9-20-2022.pdf
  - Recording: https://www.zoomgov.com/rec/share/HckZOnI1po3QZL1Xh6OZCukps6IzKPPRBHfnTCM-UbefiwSi6x4eLcvIlqY1OPw.WfMjRxdDGA2qhd9v
    Passcode: 1%yC!#w#

i5k Workspace BLAST: one way to access Apollo



Select organism

Select organism-specific database

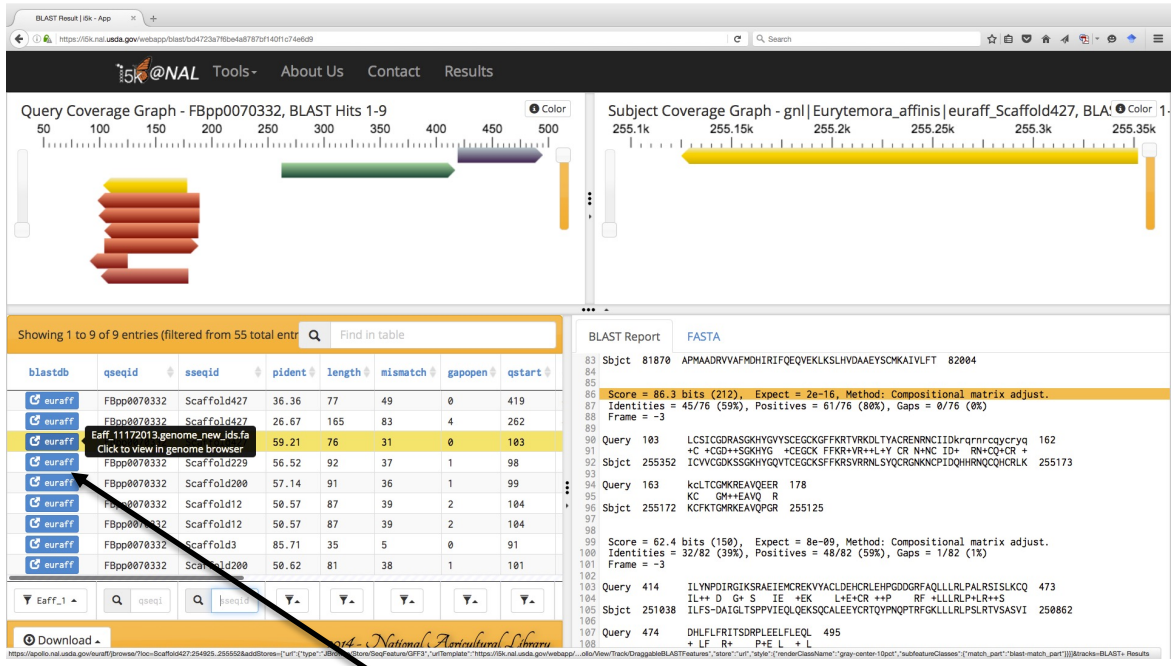
BLAST against the genome assembly to view HSPs in Jbrowse

Paste or upload query sequence(s)

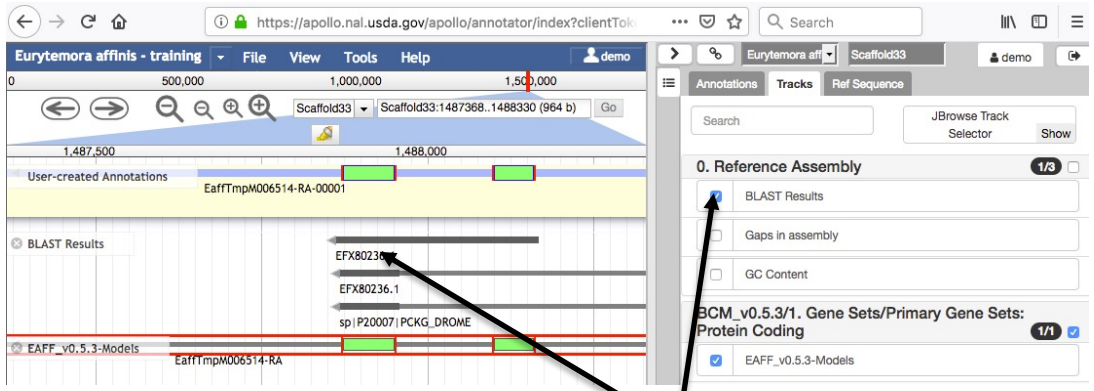
Program is automatically selected

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

i5k Workspace BLAST: one way to access Apollo



BLAST result page with 4 panels

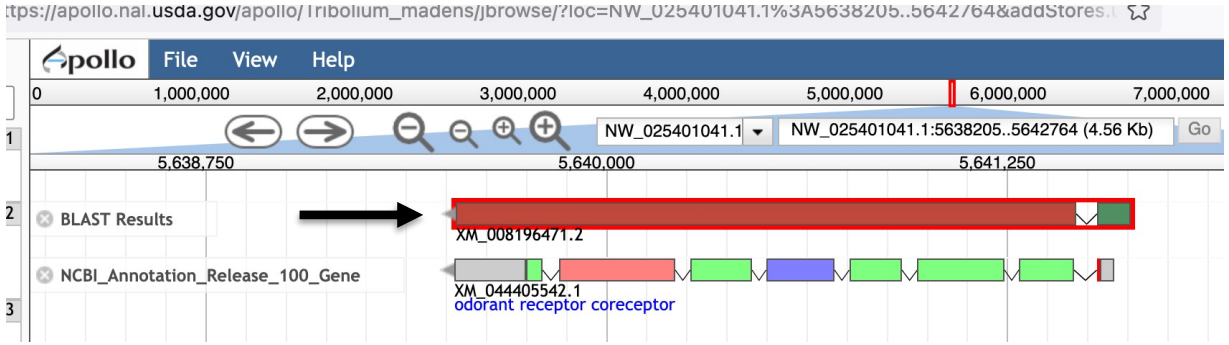


Click on blue blastdb icon next to your favorite HSP

Blast results are displayed in Apollo

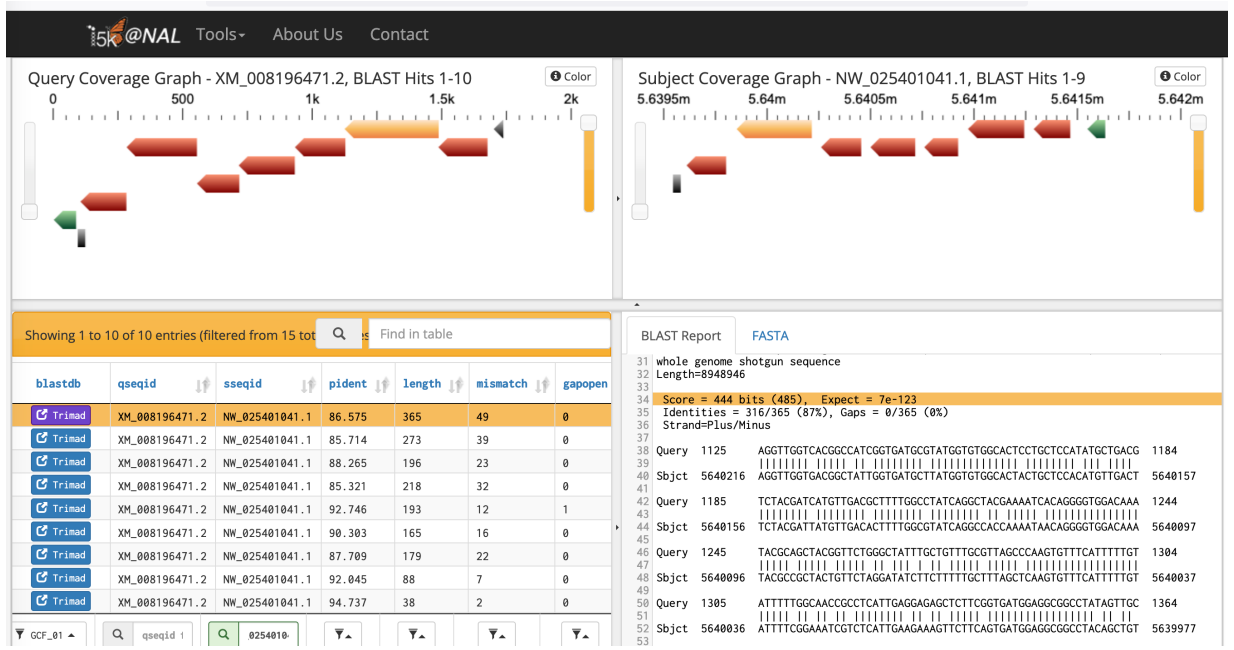
Blast example: T. castaneum Odorant receptor co-receptor vs. T. madens

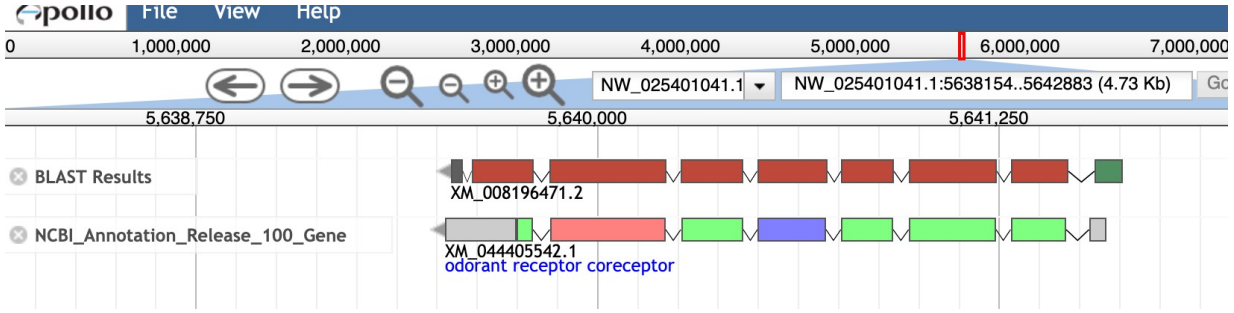
- Query sequence: https://www.ncbi.nlm.nih.gov/nuccore/XM_008196471.2?report=fasta
- First, search with default Blast parameters
- HSP (high scoring pair) with lowest e-value has many gaps relative to the subject sequence
- BLAST result glyphs in Jbrowse show a continuous alignment



Blast example: T. castaneum Odorant receptor co-receptor vs. T. madens

- BLAST search with same query and subject – increased gap opening and extension penalties
- More, shorter HSPs as a a result
- BLAST result glyphs in Jbrowse now match better (but not perfectly) to gene prediction
- Upshot: BLAST can be useful for quickly finding matches between query and subject sequence, but is not the best tool for finding exon boundaries





A few notes on BLAST and finding your gene in JBrowse/Apollo

- The JBrowse search box has a type-ahead function that is useful for finding genes by name or accession number
- You can use Blat in Apollo – please only search on the same scaffold
- We are developing a new BLAST interface, and are expecting to release it next year. We will ask you for feedback!

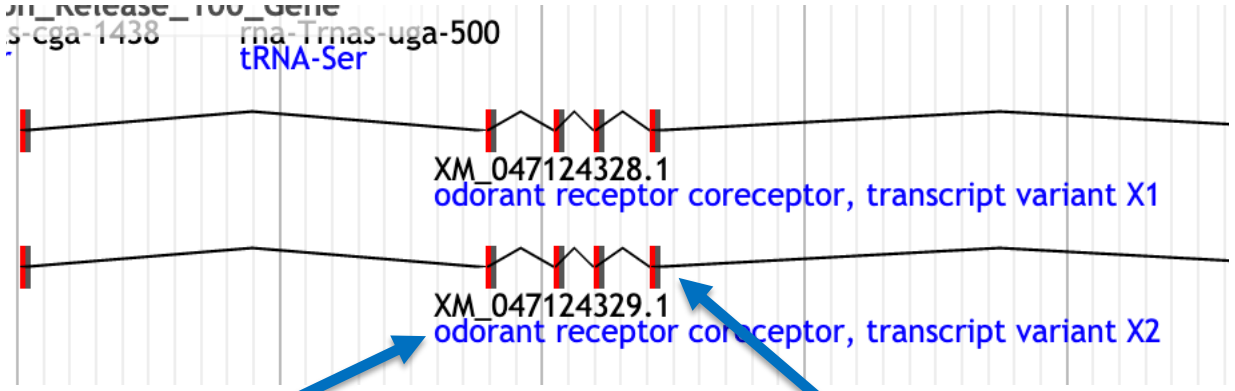
Naming

Names vs. Symbols vs. Accessions

Name	Symbol	Accession
• A description of a gene or protein's function • *Example*: odorant receptor co-receptor	• A short form of the Name • Only applies to the gene (not protein) • *Example*: Orco	• A locally unique identifier for a feature • *Example:* XM_044405542.1

Names vs. symbols vs. accession numbers

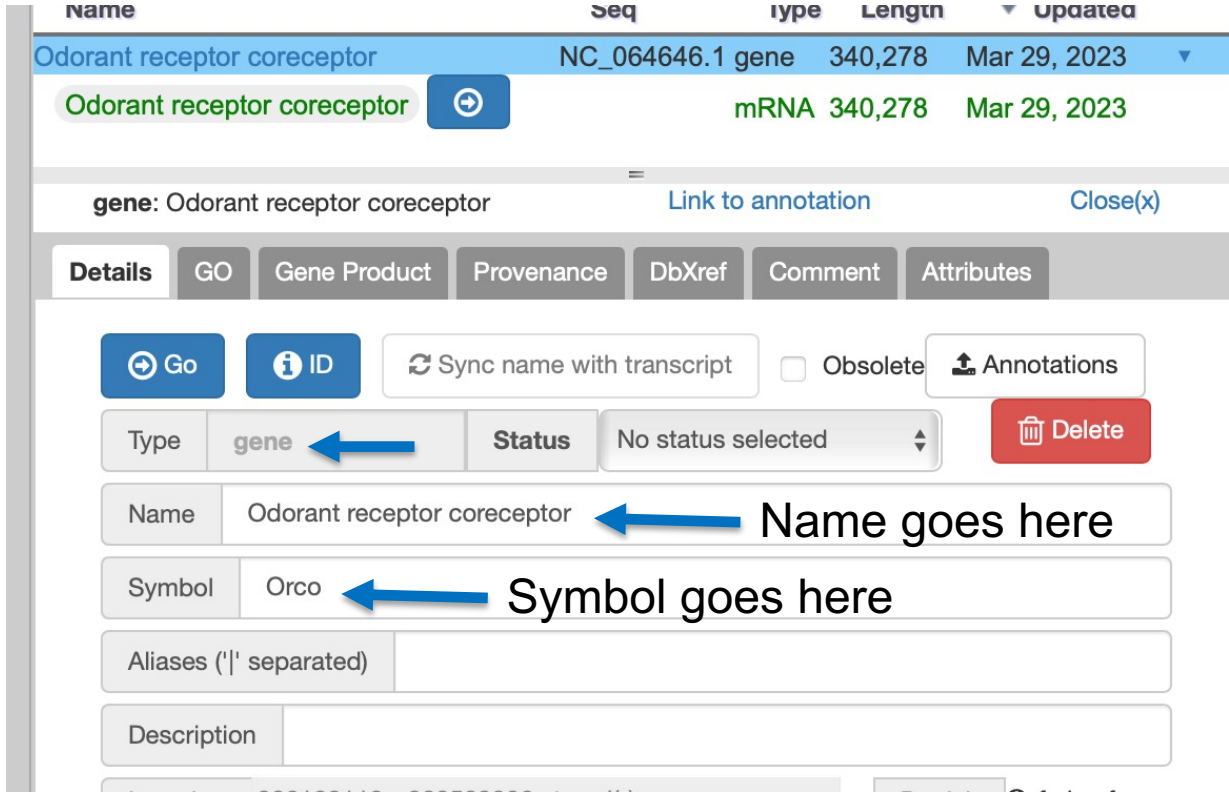
In JBrowse display



Name* in blue

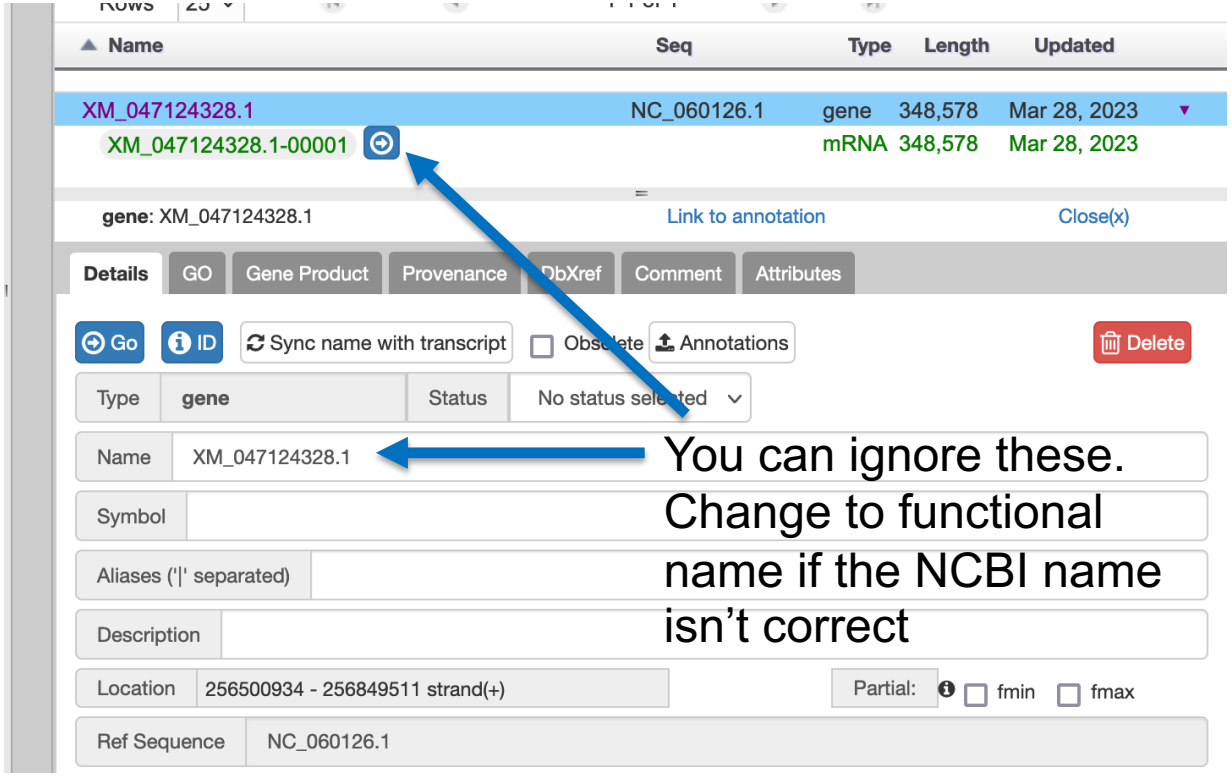
Accession number* in black

In Apollo Information Editor



A word on Apollo default ‘names’

- Apollo adds placeholder names when you create a model
- This is just Apollo’s default behavior – this is NOT the new name OR accession number of the model.
- If we submit your model to NCBI, I will remove the default accession number from the name field before submitting to NCBI.
- Please change these if you need to make a change to the actual name (e.g. odorant receptor co-receptor)



I5k Workspace Nomenclature Guidelines

- For a full review, visit https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines
- These guidelines are based on the International Protein Nomenclature Guidelines (https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/) – when in doubt, please consult these guidelines

Naming Use Cases

Are you adopting a name or symbol 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" 🚫

https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines

Naming Use Cases

- Are you naming several isoforms?
  - use the suffix “isoform A”, “isoform B”, etc. in the mRNA name if desired
- 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.

https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines

Naming Use Cases

- Are you creating a new name for a 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
  - Use Arabic numbers to specify the different members encoded by a multigene family.

https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines

A word on annotation goals

- Do you want your name change to be available outside of Apollo? Are you writing a publication with updated names?
  - Please get in touch with me so we can submit the names to GenBank and/or RefSeq for you! This does *not* happen automatically.
- Are you just changing the name for your own record-keeping?
  - In that case, you may not need to worry too much about the naming rules.

Questions?

Thank you for attending!