National Agricultural Library
U.S. DEPARTMENT OF AGRICULTURE

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

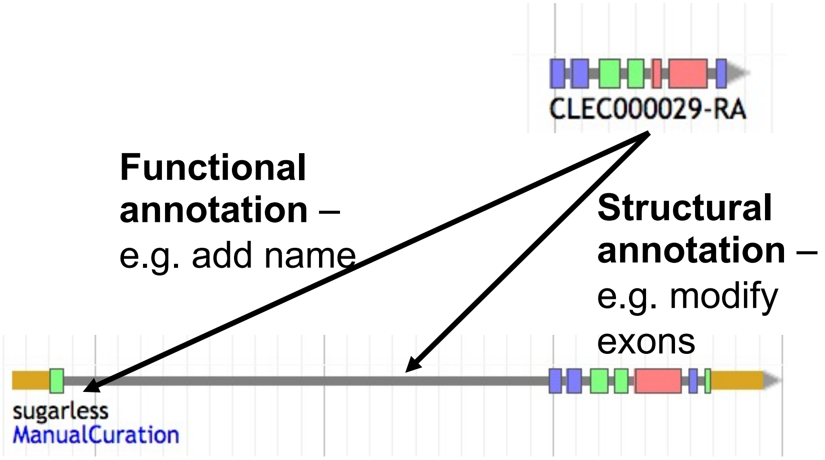
Other resources

- 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:
  - https://genomearchitect.readthedocs.io/en/latest/UsersGuide.html
- 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



Functional annotation –
e.g. add name

Structural annotation –
e.g. modify exons

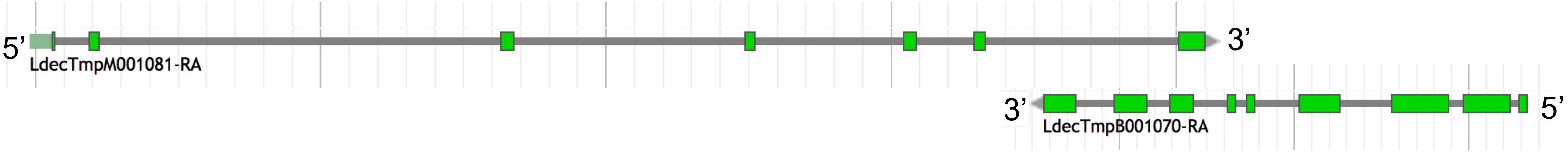
Why manually annotate?

- 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

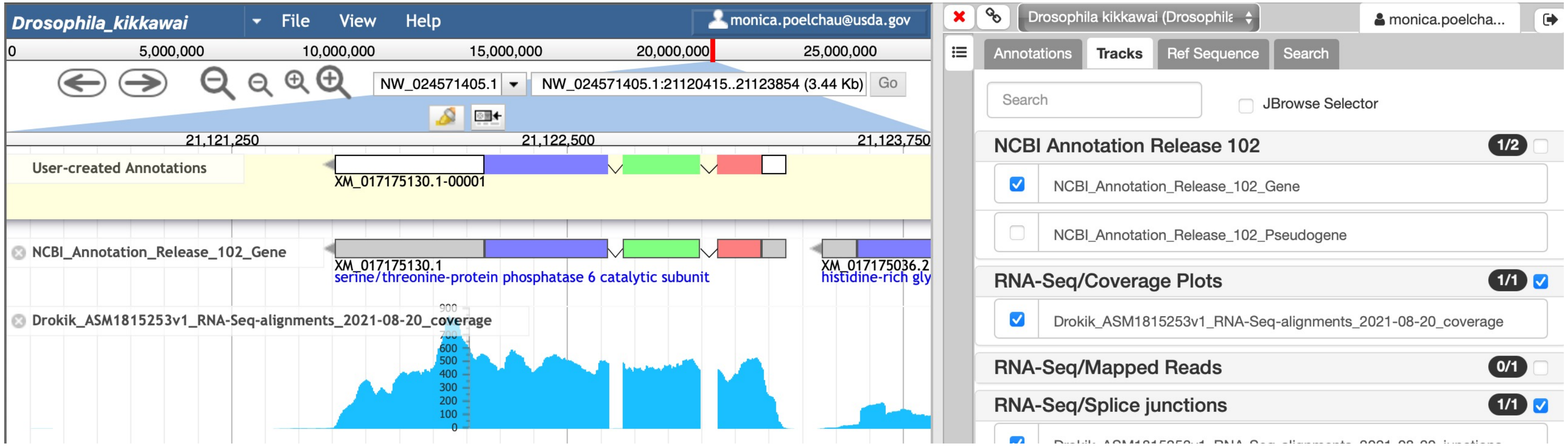
MANUAL ANNOTATION: I5K WORKSPACE TOOLS

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

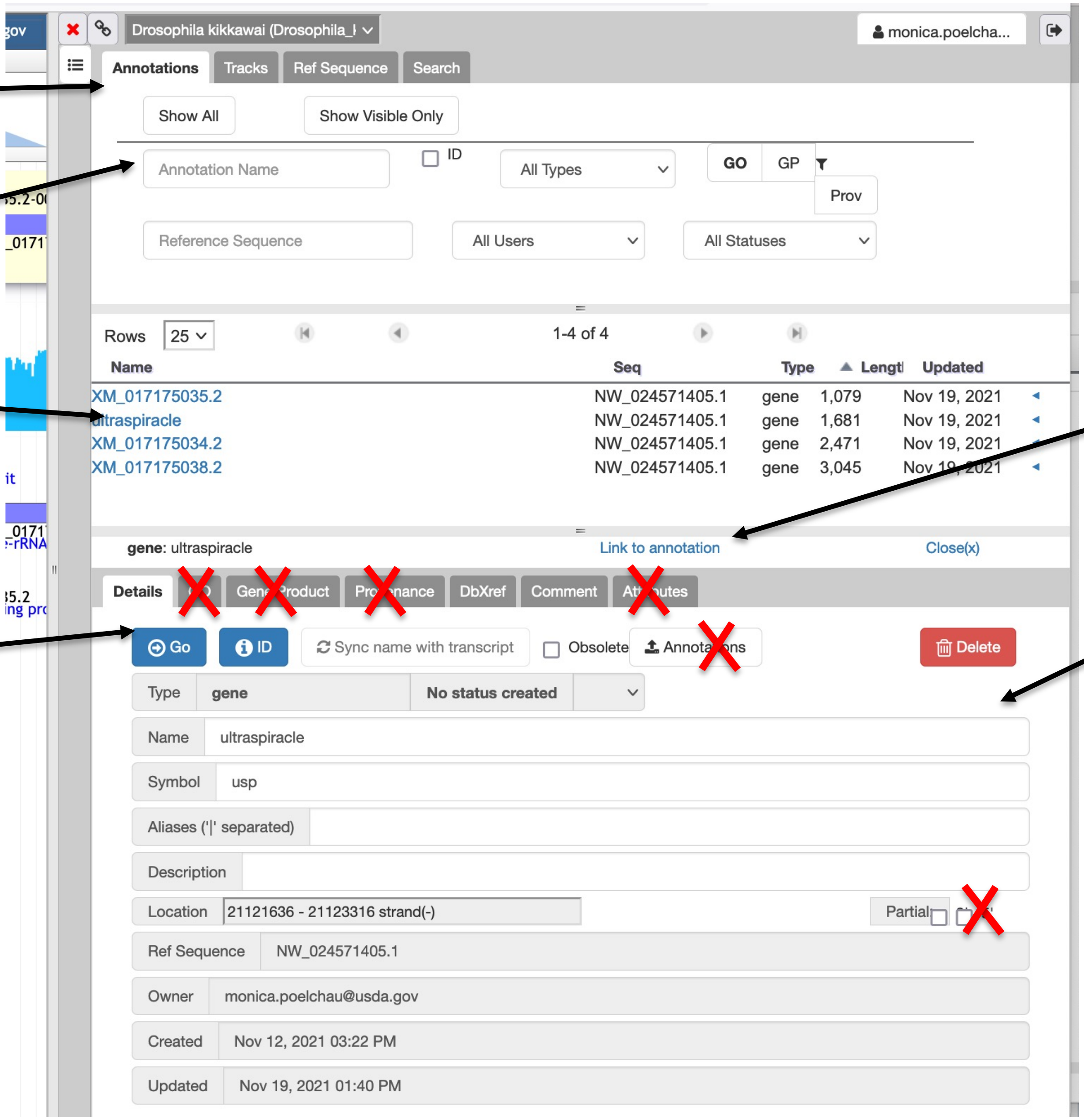
Revert to 'old' track selector

Log out

User-created annotations track

Find information about tracks

Apollo2 – Annotations Panel



Name	Seq	Type	Length	Updated
XM_017175035.2	NW_024571405.1	gene	1,079	Nov 19, 2021
ultraspiracle	NW_024571405.1	gene	1,681	Nov 19, 2021
XM_017175034.2	NW_024571405.1	gene	2,471	Nov 19, 2021
XM_017175038.2	NW_024571405.1	gene	3,045	Nov 19, 2021

gene: ultraspiracle

Type: gene
No status created
Name: ultraspiracle
Symbol: usp
Aliases ('|' separated):
Description:
Location: 21121636 - 21123316 strand(-)
Partial ☐ ☐
Ref Sequence: NW_024571405.1
Owner: monica.poelchau@usda.gov
Created: Nov 12, 2021 03:22 PM
Updated: Nov 19, 2021 01:40 PM

Annotations panel

Filter annotations

Select Name to view ***and edit*** details

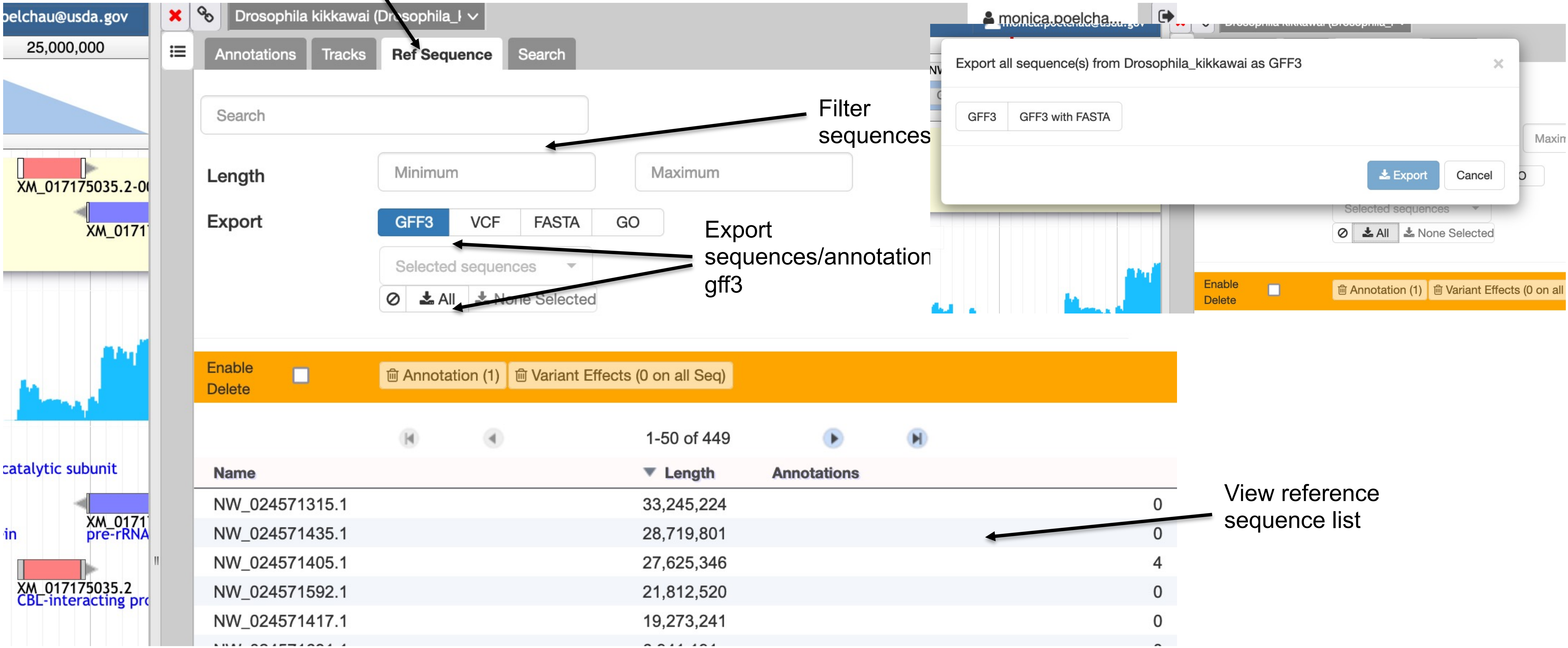
Click on arrow to jump to annotation in browser

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



Export all sequence(s) from Drosophila_kikkawai as GFF3

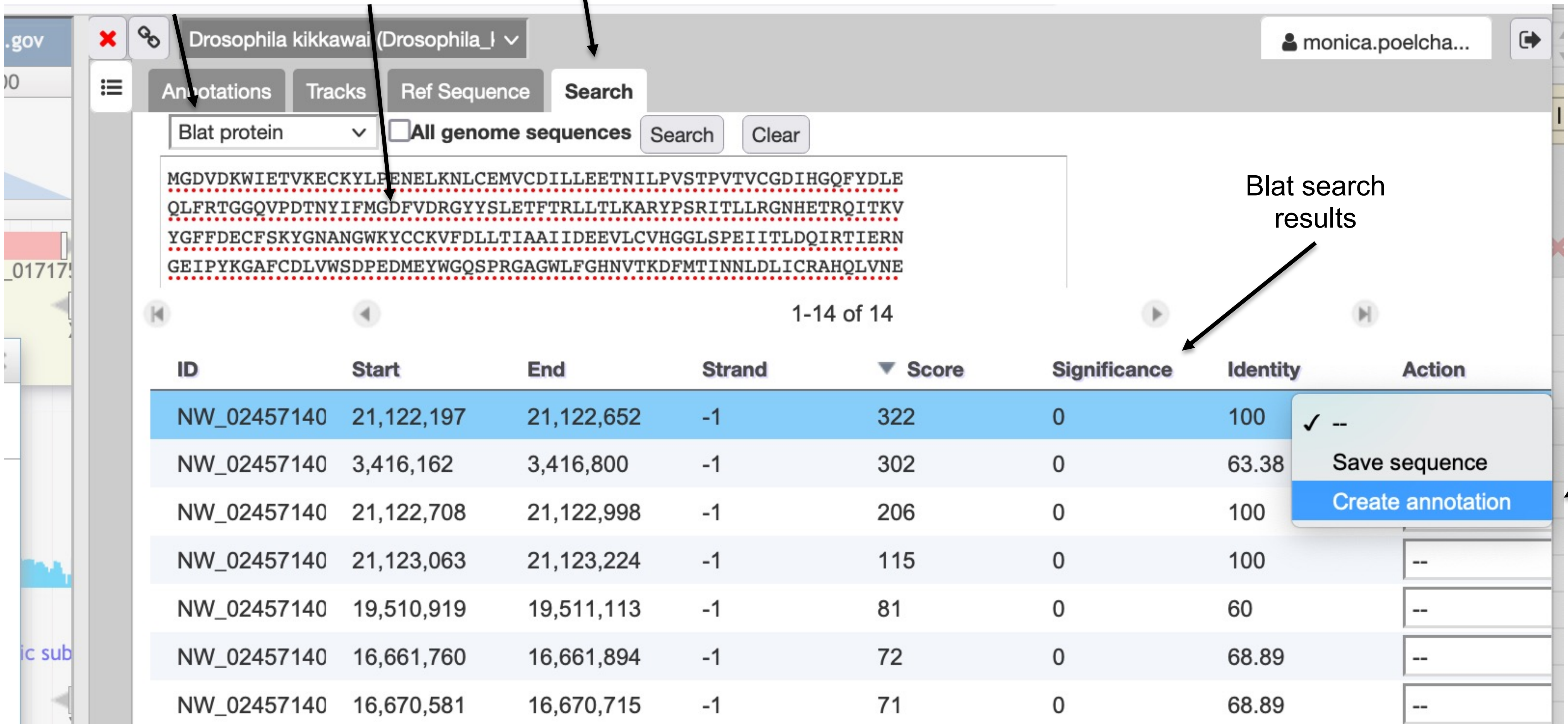
Name	Length	Annotations
NW_024571315.1	33,245,224	0
NW_024571435.1	28,719,801	0
NW_024571405.1	27,625,346	4
NW_024571592.1	21,812,520	0
NW_024571417.1	19,273,241	0

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

```
MGDVDKWIETVKECKYLPENELKNLCEMVCDILLEETNILPVSTPVTVCGDIHGQFYDLE
QLFRTGGQVPDTNYIFMGDFVDRGYYSLETFTRLLTLKARYPSRITLLRGNHETRQITKV
YGFFDECFSKYGNANGWKYCCKVFDLLTIAAIIDEEVLCVHGGLSPEIITLDQIRTIERN
GEIPYKGAFCDLVWSDPEDMEYWGQSPRGAGWLFGHNVTKDFMTINNLDLICRAHQLVNE
```

1-14 of 14

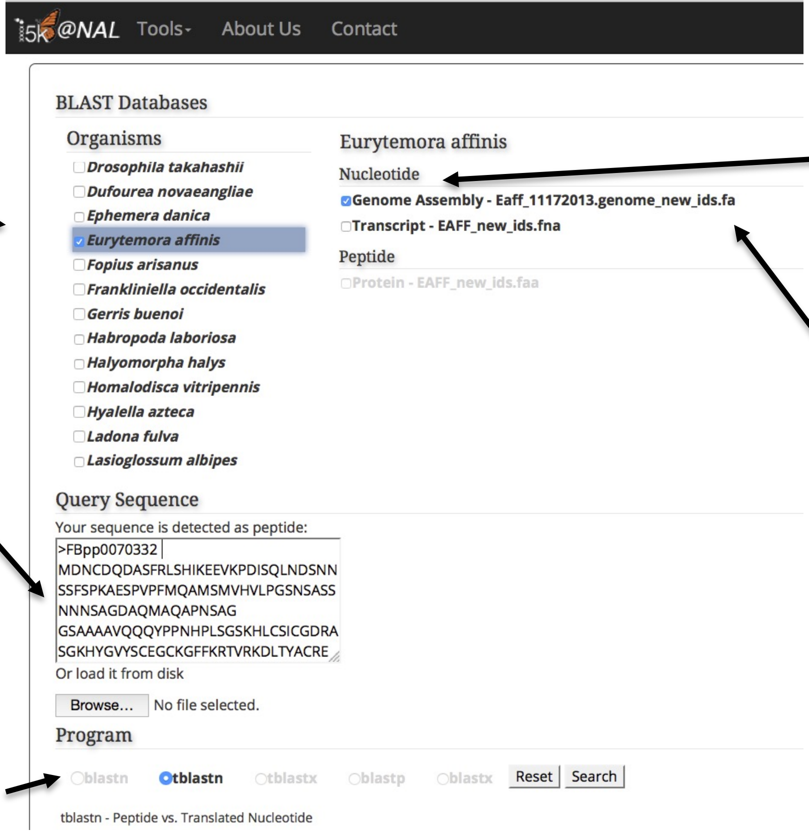
ID	Start	End	Strand	Score	Significance	Identity	Action
NW_02457140	21,122,197	21,122,652	-1	322	0	100	
NW_02457140	3,416,162	3,416,800	-1	302	0	63.38	
NW_02457140	21,122,708	21,122,998	-1	206	0	100	
NW_02457140	21,123,063	21,123,224	-1	115	0	100	--
NW_02457140	19,510,919	19,511,113	-1	81	0	60	--
NW_02457140	16,661,760	16,661,894	-1	72	0	68.89	--
NW_02457140	16,670,581	16,670,715	-1	71	0	68.89	--

i5k Workspace BLAST: one way to access Apollo

Select organism

Paste or upload query sequence(s)

Program is automatically selected



BLAST Databases

Organisms

- ☐ *Drosophila takahashii*
- ☐ *Dufourea novaeangliae*
- ☐ *Ephemera danica*
- ☑ *Eurytemora affinis*
- ☐ *Fopius arisanus*
- ☐ *Frankliniella occidentalis*
- ☐ *Gerris buenoi*
- ☐ *Habropoda laboriosa*
- ☐ *Halyomorpha halys*
- ☐ *Homalodisca vitripennis*
- ☐ *Hyalella azteca*
- ☐ *Ladona fulva*
- ☐ *Lasioglossum albipes*

Eurytemora affinis

Nucleotide

- ☑ Genome Assembly - Eaff_11172013.genome_new_ids.fa
- ☐ Transcript - EAFF_new_ids.fna

Peptide

- ☐ Protein - EAFF_new_ids.faa

Query Sequence

Your sequence is detected as peptide:

```
>FBpp0070332
MDNCDQDASFRLSHIKEEVKPDISQLNDSNN
SSFSPKAESPVPFMQAMSMVHVLPGSNSASS
NNNSAGDAQMAQAPNSAG
GSAAAAVQQQYPPNHPLSGSKHLCSICGDRA
SGKHYGVYSCEGCKGFFKRTVRKDLTYACRE
```

Or load it from disk

Browse… No file selected.

Program

○ blastn ● tblastn ○ tblastx ○ blastp ○ blastx Reset Search

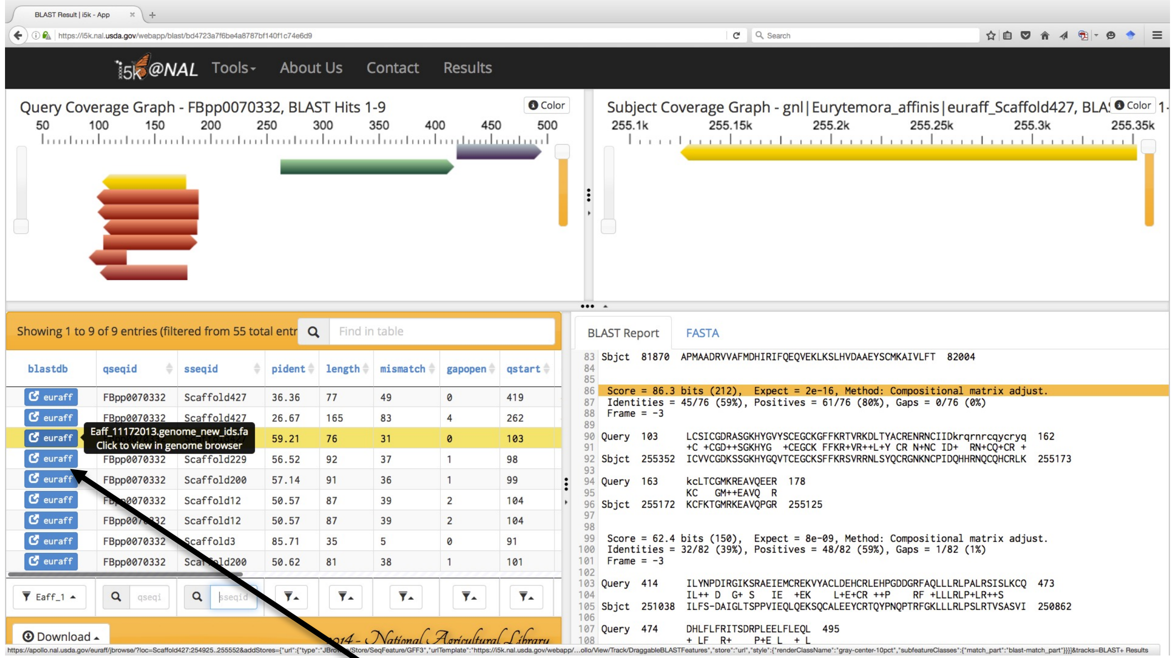
tblastn - Peptide vs. Translated Nucleotide

Select organism-specific database

BLAST against the genome assembly to view HSPs in Jbrowse

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

i5k Workspace BLAST: one way to access Apollo



Showing 1 to 9 of 9 entries (filtered from 55 total entr

blastdb	qseqid	sseqid	pident	length	mismatch	gapopen	qstart
euraff	FBpp0070332	Scaffold427	36.36	77	49	0	419
euraff	FBpp0070332	Scaffold427	26.67	165	83	4	262
euraff	FBpp0070332	Scaffold427	59.21	76	31	0	103
euraff	FBpp0070332	Scaffold229	56.52	92	37	1	98
euraff	FBpp0070332	Scaffold200	57.14	91	36	1	99
euraff	FBpp0070332	Scaffold12	50.57	87	39	2	104
euraff	FBpp0070332	Scaffold12	50.57	87	39	2	104
euraff	FBpp0070332	Scaffold3	85.71	35	5	0	91
euraff	FBpp0070332	Scaffold200	50.62	81	38	1	101

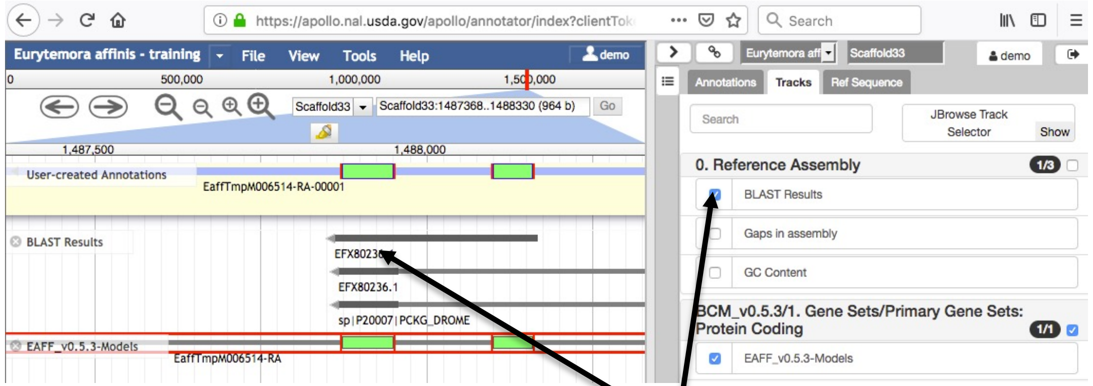
Eaff_11172013.genome_new_ids.fa
Click to view in genome browser

BLAST Report FASTA

```
Sbjct  81870  APMAADRVVAFMDHIRIFQEQVEKLKSLHVDAAEYSCMKAIVLFT  82004


 Score = 86.3 bits (212),  Expect = 2e-16, Method: Compositional matrix adjust.
 Identities = 45/76 (59%), Positives = 61/76 (80%), Gaps = 0/76 (0%)
 Frame = -3

Query  103     LCSICGDRASGKHYGVYSCEGCKGFFKRTVRKDLTYACRENRNCIIDkrqrnrcqycryq  162
               +C +CGD++SGKHYG  +CEGCK FFKR+VR++L+Y CR N+NC ID+  RN+CQ+CR +
Sbjct  255352  ICVVCGDKSSGKHYGQVTCEGCKSFFKRSVRRNLSYQCRGNKNCPIDQHHRNQCQHCRLK  255173

Query  163     kcLTCGMKREAVQEER  178
               KC   GM++EAVQ  R
Sbjct  255172  KCFKTGMRKEAVQPGR  255125


 Score = 62.4 bits (150),  Expect = 8e-09, Method: Compositional matrix adjust.
 Identities = 32/82 (39%), Positives = 48/82 (59%), Gaps = 1/82 (1%)
 Frame = -3

Query  414     ILYNPDIRGIKSRAEIEMCREKVYACLDEHCRLEHPGDDGRFAQLLLRLPALRSISLKCQ  473
               IL++ D  G+ S   IE  +EK    L+E+CR ++P    RF +LLLRLP+LR++S
Sbjct  251038  ILFS-DAIGLTSPPVIEQLQEKSQCALEEYCRTQYPNQPTRFGKLLLRLPSLRTVSASVI  250862

Query  474     DHLFLFRITSDRPLEELFLEQL  495
               + LF  R+    P+E L  + L
```

https://apollo.nal.usda.gov/euraff/jbrowse/?loc=Scaffold427:254925..255552&addStores={"url":{"type":"JBrowse/Store/SeqFeature/GFF3","urlTemplate":"https://i5k.nal.usda.gov/webapp/...ollo/View/Track/DraggableBLASTFeatures","store":"url","style":{"renderClassName":"gray-center-10pct","subfeatureClasses":{"match_part":"blast-match_part"}}}}&tracks=BLAST+ Results

BLAST result page with 4 panels



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/

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

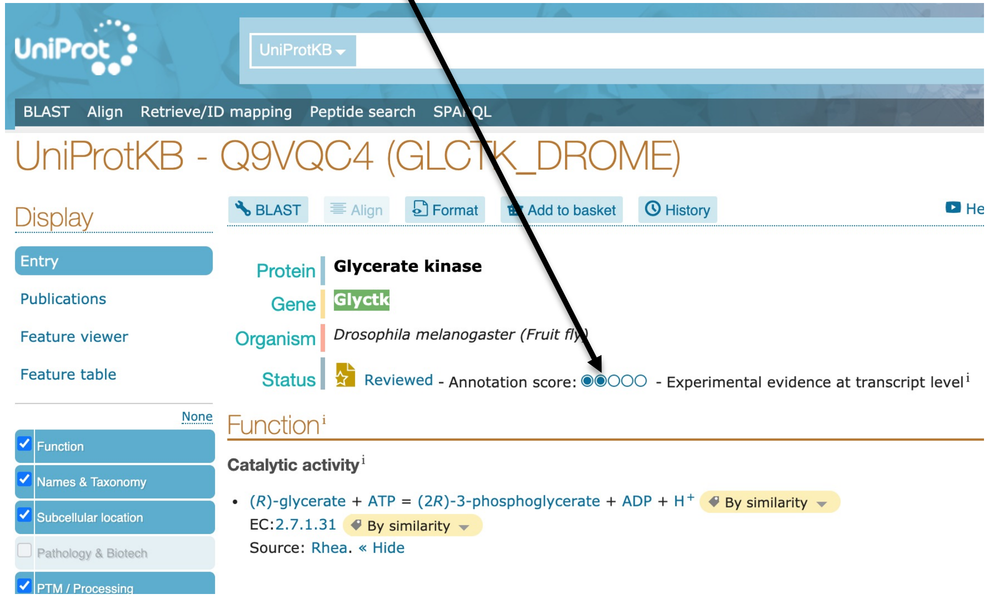
MANUAL ANNOTATION EXAMPLE: PREPARATION

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
- *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



UniProtKB - Q9VQC4 (GLCTK_DROME)

Protein Glycerate kinase

Gene Glyctk

Organism *Drosophila melanogaster (Fruit fly)*

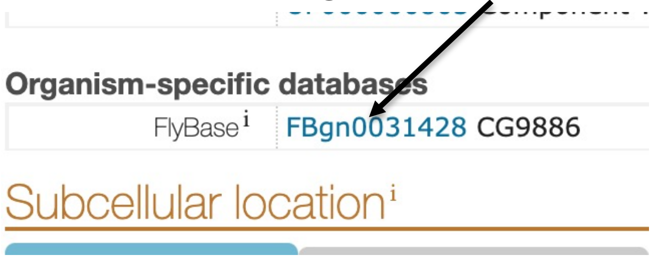
Status Reviewed - Annotation score: ●●○○○ - Experimental evidence at transcript level

Function

Catalytic activity

- (*R*)-glycerate + ATP = (2*R*)-3-phosphoglycerate + ADP + H^+ By similarity
  EC:2.7.1.31 By similarity
  Source: Rhea. « Hide

Flybase is another great resource

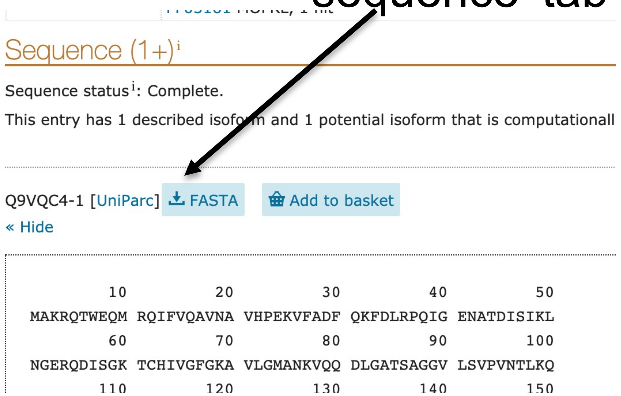


Organism-specific databases

FlyBase FBgn0031428 CG9886

Subcellular location

Retrieve FASTA from 'sequence' tab



Sequence (1+)

Sequence status: Complete.

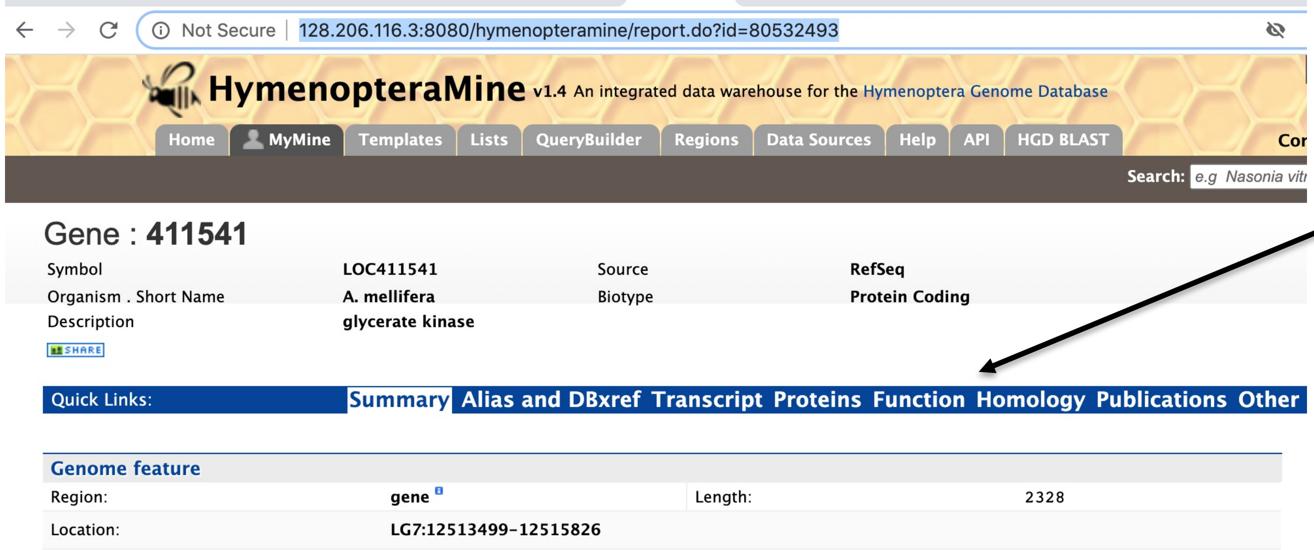
This entry has 1 described isoform and 1 potential isoform that is computationall

Q9VQC4-1 [UniParc]

```
        10         20         30         40         50
MAKRQTWEQM RQIFVQAVNA VHPEKVFADF QKFDLRPQIG ENATDISIKL
        60         70         80         90        100
NGERQDISGK TCHIVGFGKA VLGMANKVQQ DLGATSAGGV LSVPVNTLKQ
       110        120        130        140        150
```

Source: https://www.uniprot.org/uniprot/Q9VQC4

Choosing reference proteins: *Apis mellifera* glycerate kinase



Gene : **411541**

Symbol	**LOC411541**	Source	**RefSeq**
Organism . Short Name	**A. mellifera**	Biotype	**Protein Coding**
Description	**glycerate kinase**		

Genome feature

Region:	**gene**	Length:	2328
Location:	**LG7:12513499-12515826**		

Lots of additional information on function

Transcript

All Transcripts for Gene - LOC411541 411541
All Transcripts for Gene
Transcripts: 1 Exons: 2 CDSs: 1

Transcript	Exons		CDSs
LOC411541 XM_003251169.3 2328 FASTA...	exon230448	115 FASTA...	XM_003251169.3-CDS 2163 FASTA...
	exon230449	2126 FASTA...	

FASTA available under ‘Transcript’ tab

Source: http://128.206.116.35/hymenopteramine/report.do?id=92926973

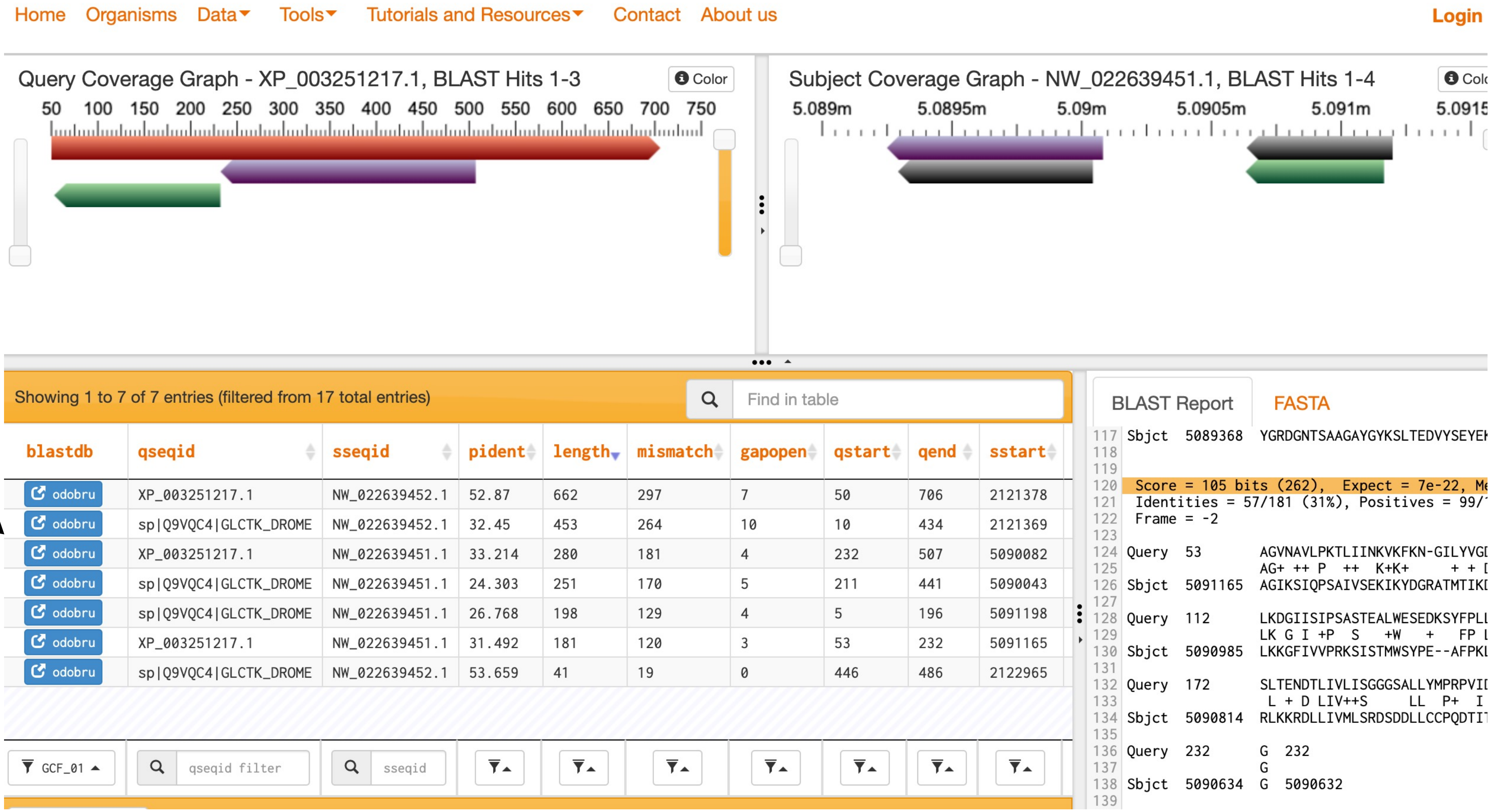
Resources for learning about insect gene/protein structure and function

- 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

BLAST dmel, amel proteins against *O. brunneus* genome

Results are filtered by e-value; we're going to focus on the second-best match here



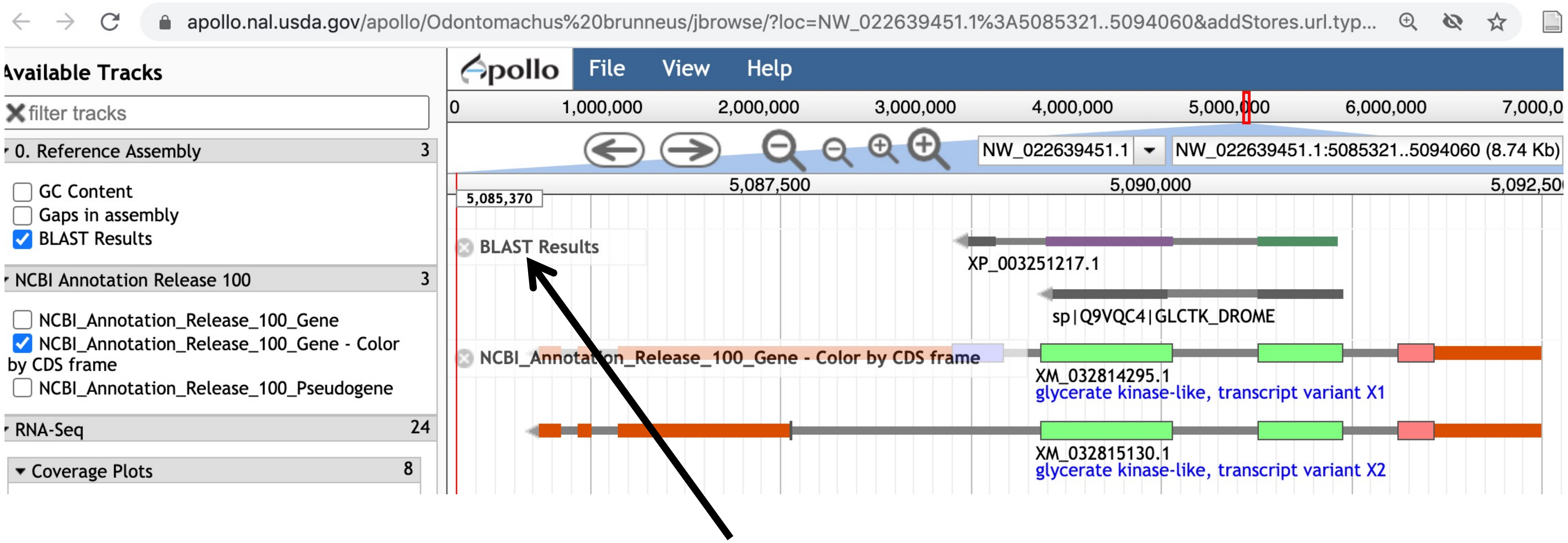
Showing 1 to 7 of 7 entries (filtered from 17 total entries)

blastdb	qseqid	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart
odobru	XP_003251217.1	NW_022639452.1	52.87	662	297	7	50	706	2121378
odobru	sp|Q9VQC4|GLCTK_DROME	NW_022639452.1	32.45	453	264	10	10	434	2121369
odobru	XP_003251217.1	NW_022639451.1	33.214	280	181	4	232	507	5090082
odobru	sp|Q9VQC4|GLCTK_DROME	NW_022639451.1	24.303	251	170	5	211	441	5090043
odobru	sp|Q9VQC4|GLCTK_DROME	NW_022639451.1	26.768	198	129	4	5	196	5091198
odobru	XP_003251217.1	NW_022639451.1	31.492	181	120	3	53	232	5091165
odobru	sp|Q9VQC4|GLCTK_DROME	NW_022639452.1	53.659	41	19	0	446	486	2122965

```
Sbjct  5089368  YGRDGNTSAAGAYGYKSLTEDVYSEYE


 Score = 105 bits (262),  Expect = 7e-22, M
 Identities = 57/181 (31%), Positives = 99/
 Frame = -2

Query  53       AGVNAVLPKTLIINKVKFKN-GILYVG
                AG+ ++ P  ++  K+K+     + +
Sbjct  5091165  AGIKSIQPSAIVSEKIKYDGRATMTIK

Query  112      LKDGIISIPSASTEALWESEDKSYFPL
                LK G I +P  S   +W   +   FP
Sbjct  5090985  LKKGFIVVPRKSISTMWSYPE--AFPK

Query  172      SLTENDTLIVLISGGGSALLYMPRPVI
                 L + D LIV++S      LL  P+  I
Sbjct  5090814  RLKKRDLLIVMLSRDSDDLLCCPQDTI

Query  232      G  232
                G
Sbjct  5090634  G  5090632

```

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

BLAST dmel, amel proteins against *O. brunneus* genome



apollo.nal.usda.gov/apollo/Odontomachus%20brunneus/jbrowse/?loc=NW_022639451.1%3A5085321..5094060&addStores.url.typ...

Available Tracks

- 0. Reference Assembly 3
  - ☐ GC Content
  - ☐ Gaps in assembly
  - ☑ BLAST Results
- NCBI Annotation Release 100 3
  - ☐ NCBI_Annotation_Release_100_Gene
  - ☑ NCBI_Annotation_Release_100_Gene - Color by CDS frame
  - ☐ NCBI_Annotation_Release_100_Pseudogene
- RNA-Seq 24
- Coverage Plots 8

NW_022639451.1 NW_022639451.1:5085321..5094060 (8.74 Kb)

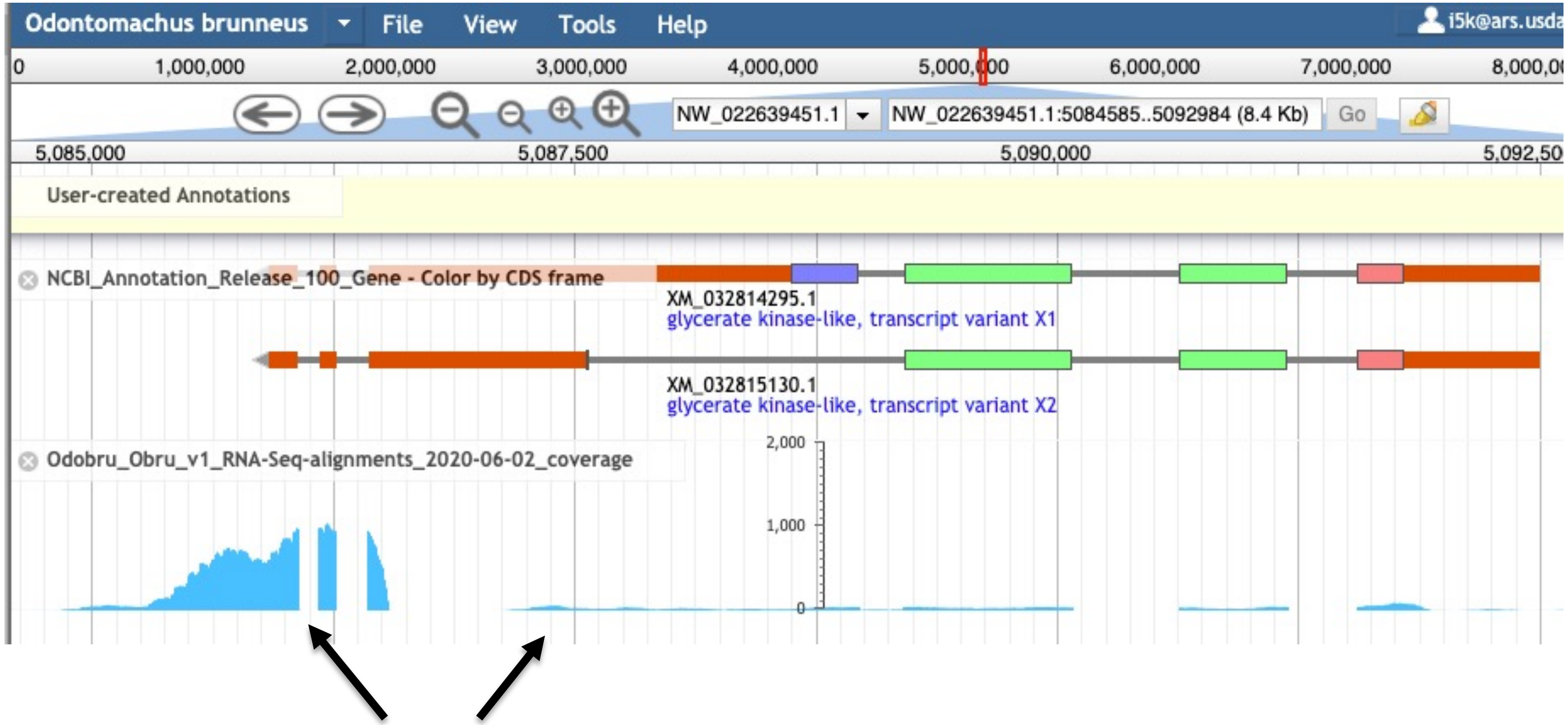
BLAST results are displayed as glyphs in browser; can be used as annotation starting points if the alignment is high quality

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?

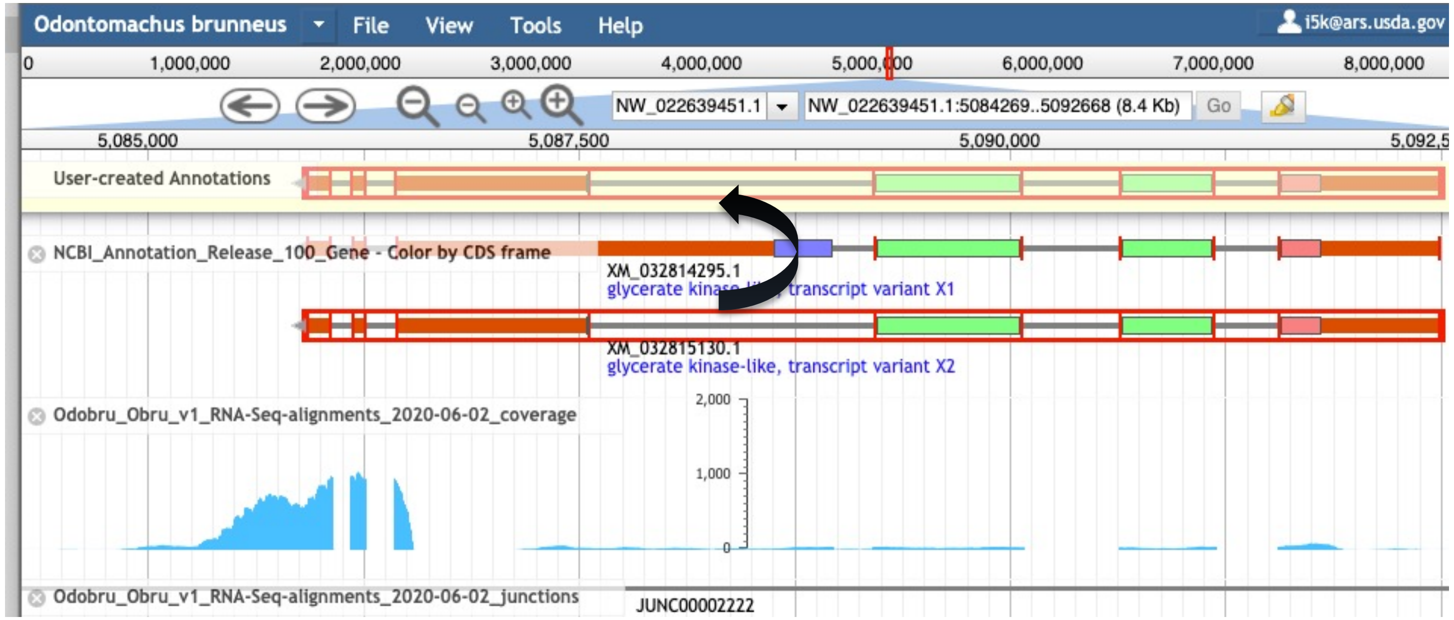
View available evidence



Very different coverage
between UTR and CDS

Create new model in user-created annotations track

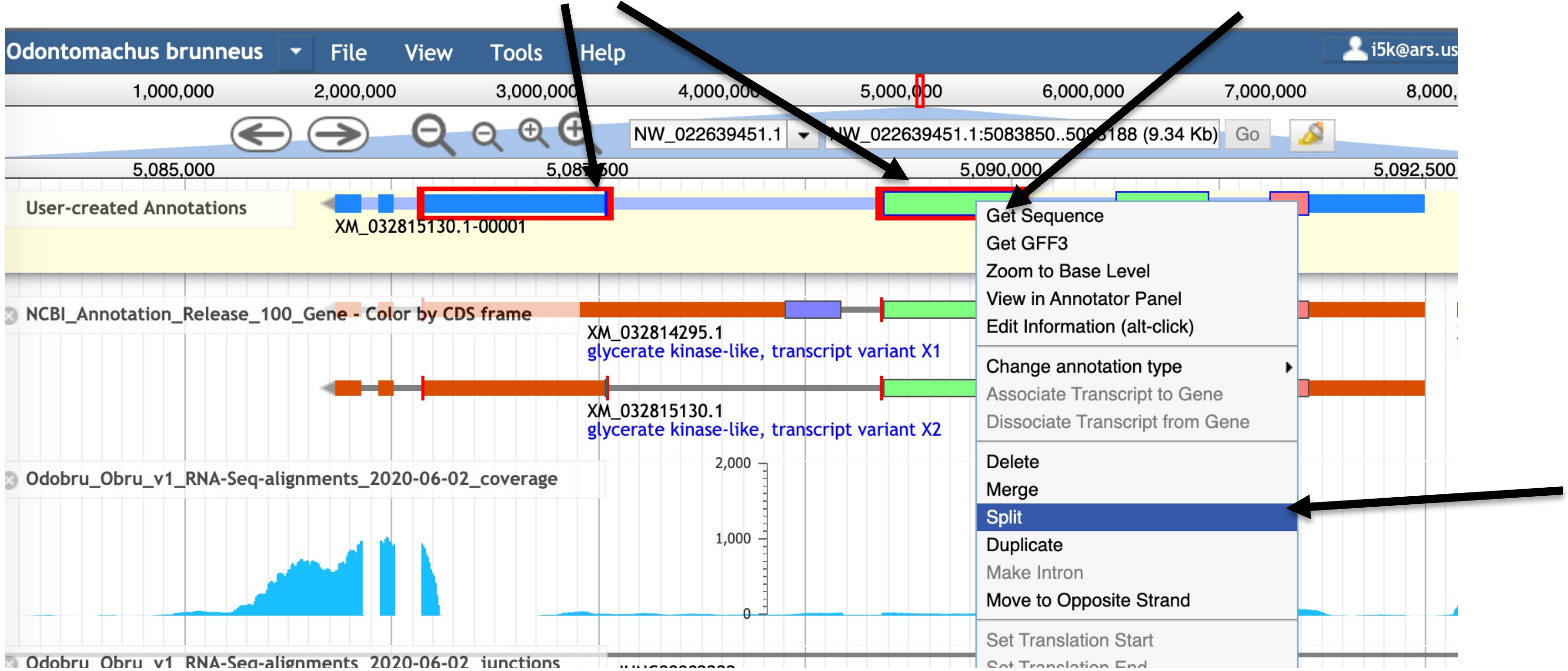
Drag evidence to UcA track (or right-click and select “create annotation”)



Split model

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

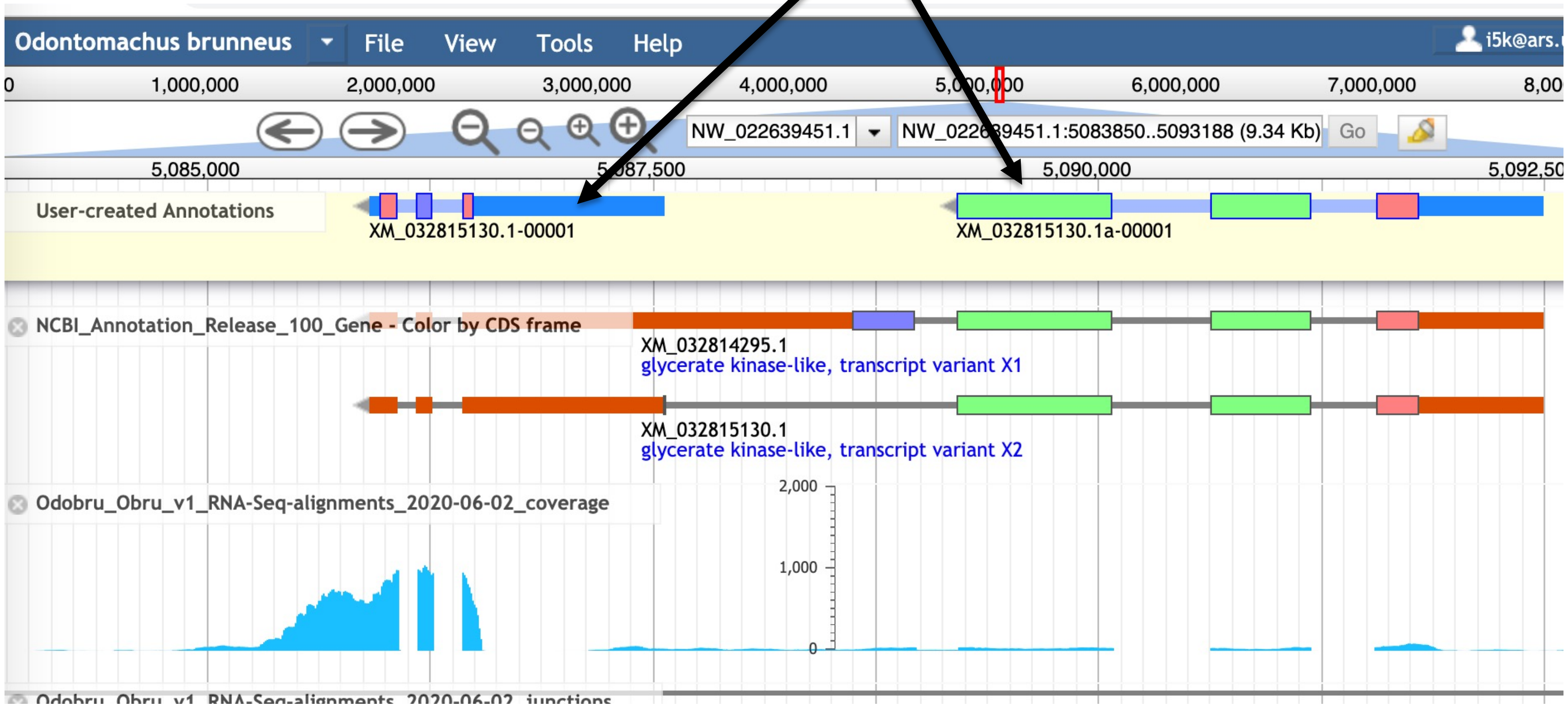
Right-click on the model while continuing to hold shift to get the drop-down menu



Select ‘split’

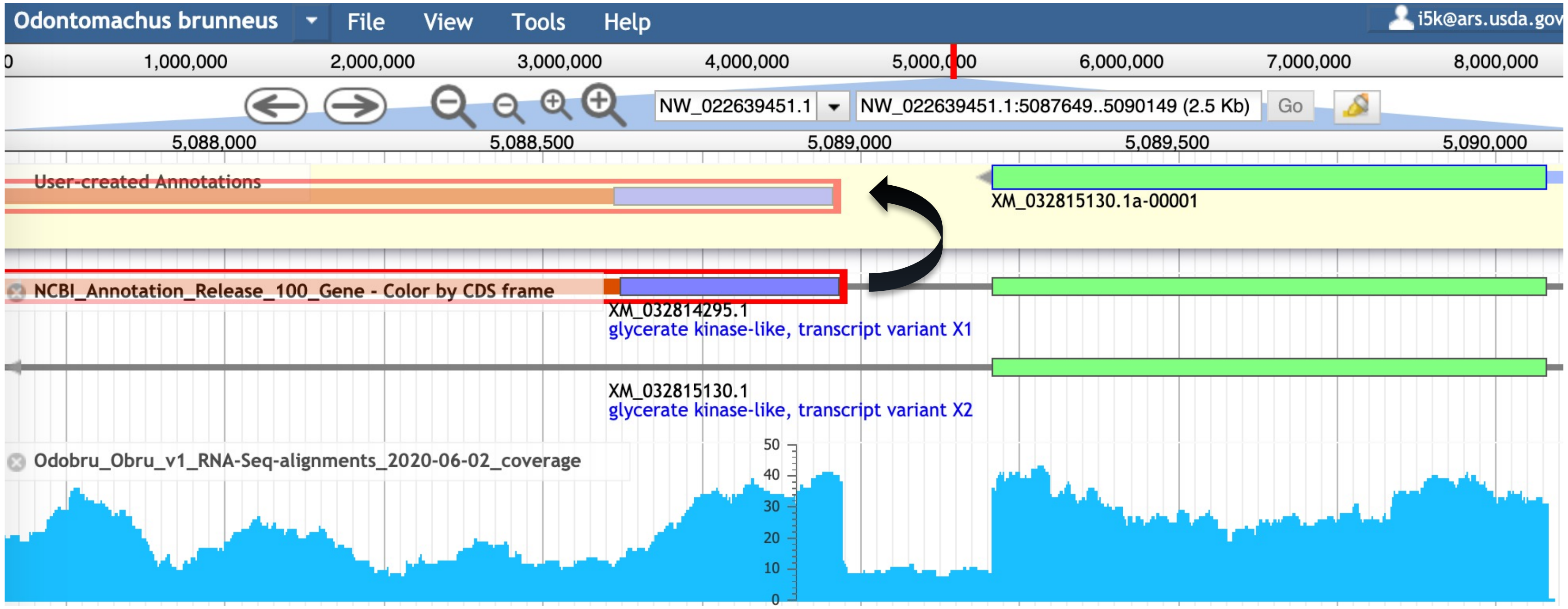
Split model

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



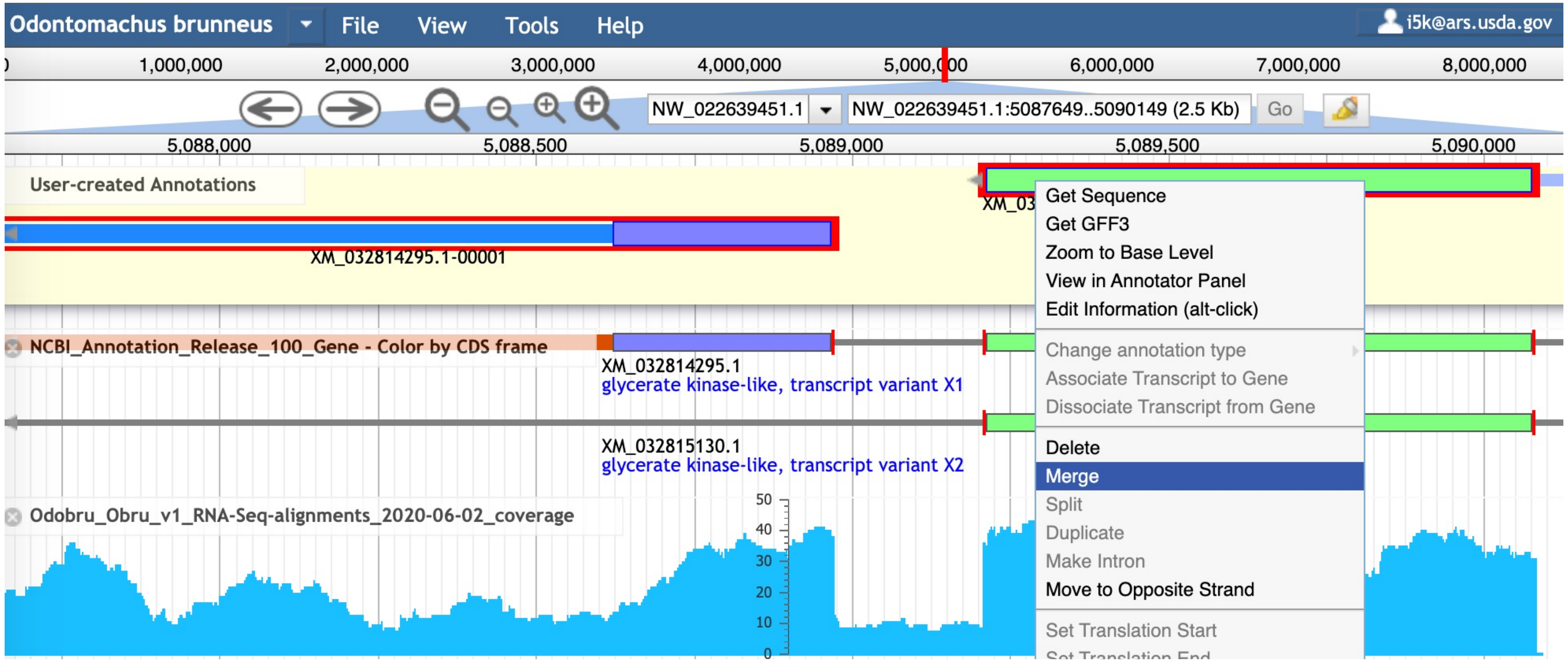
Add an exon

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



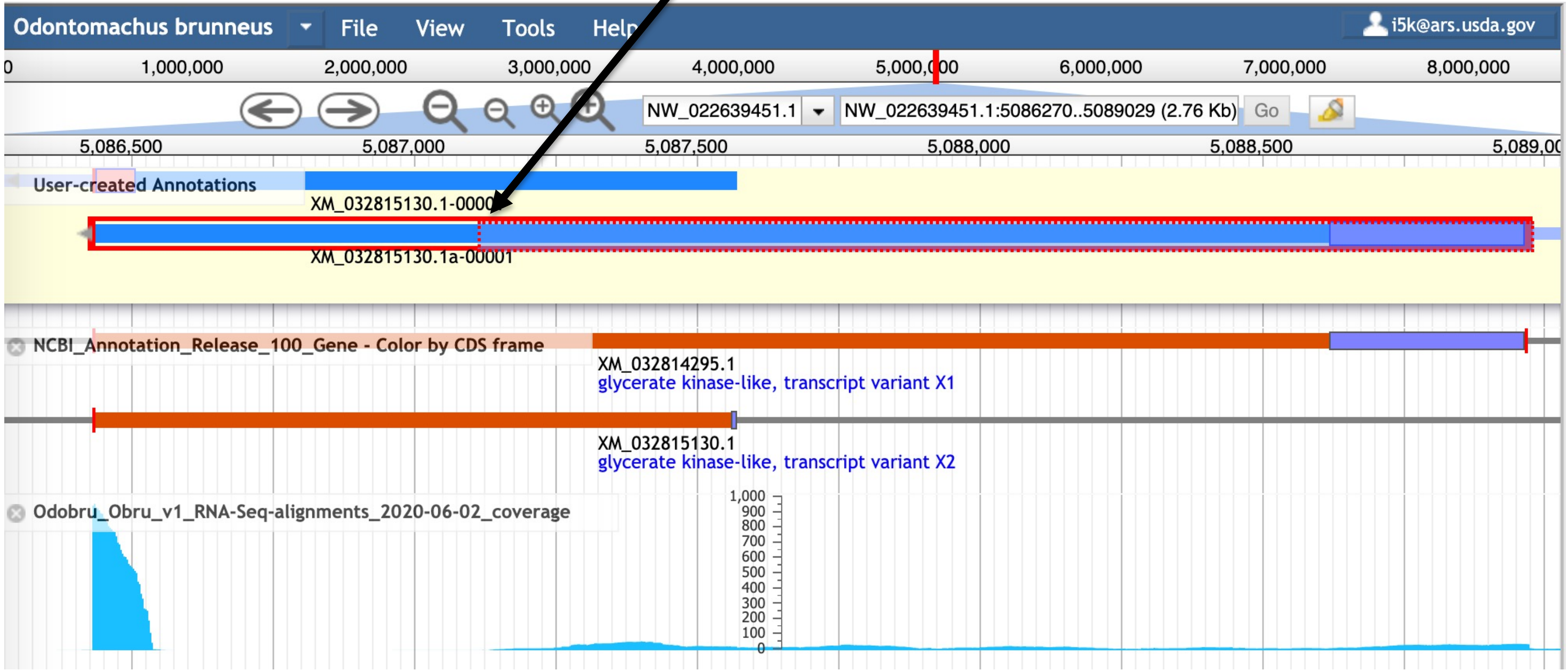
Merge exons

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

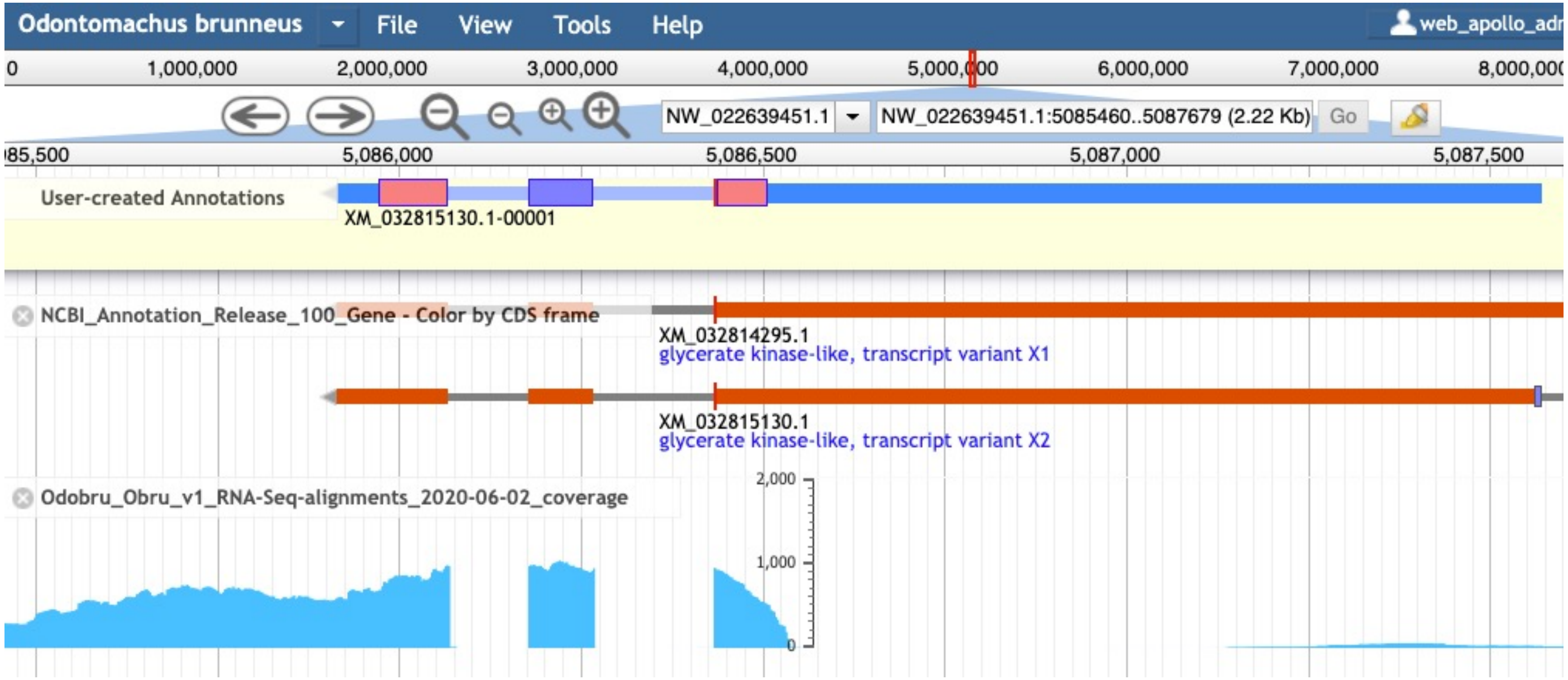


Adjust gene boundaries

Adjust the 3’ UTR to match the RNA-Seq evidence

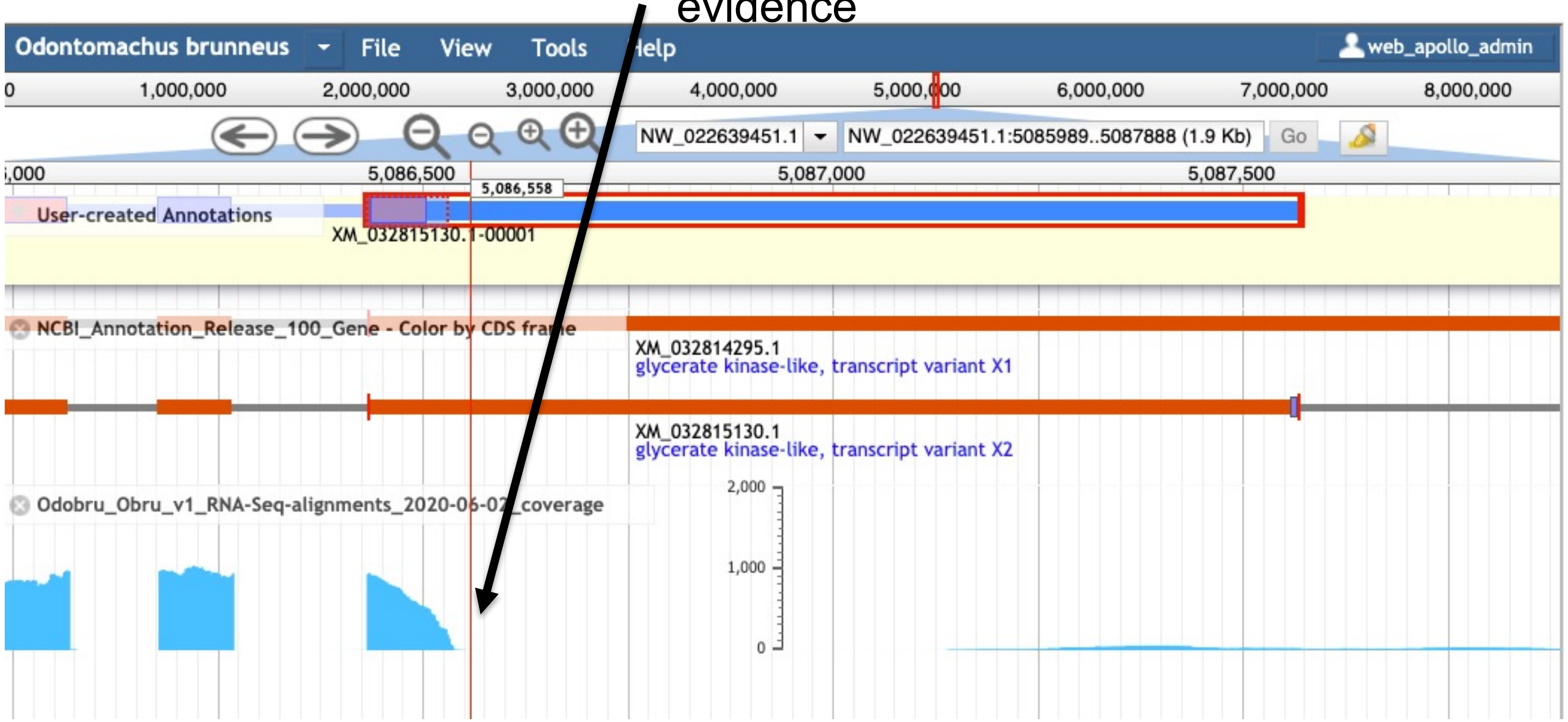


Next, the ‘left’ model



Adjust 5’ UTR

Adjust the 5’ UTR to match the RNA-Seq evidence



Set translation start

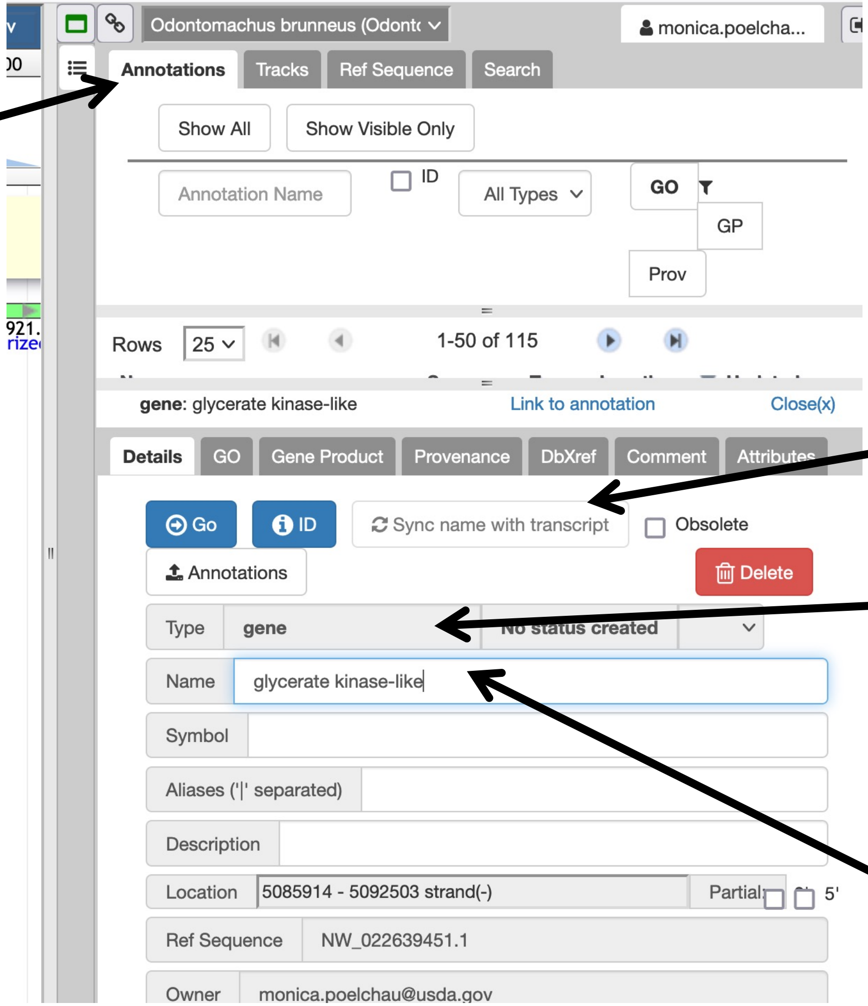
Set translation start at Methionine

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:
    - https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
- 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



Rows 25 1-50 of 115

gene: glycerate kinase-like

Type gene No status created

Name glycerate kinase-like

Symbol

Aliases ('|' separated)

Description

Location 5085914 - 5092503 strand(-) Partial ☐ ☐ 5'

Ref Sequence NW_022639451.1

Owner monica.poelchau@usda.gov

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

Using the Information Editor

- Add information about the model in the 'Annotations' panel.
  - Name: Use the i5k Workspace naming guidelines. https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines
    - 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” 🚫

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

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.

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

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
  - Use Arabic numbers to specify the different members encoded by a multigene family.

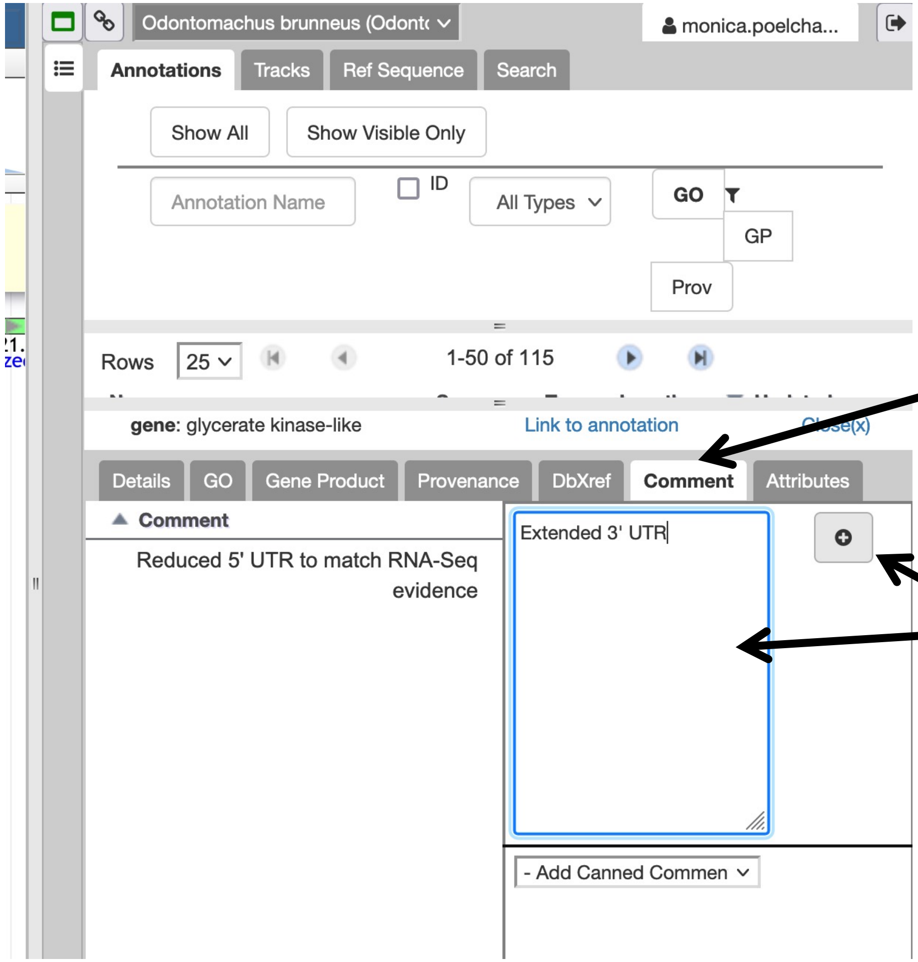
https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines

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

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

Using the Information Editor



gene: glycerate kinase-like

Reduced 5' UTR to match RNA-Seq evidence

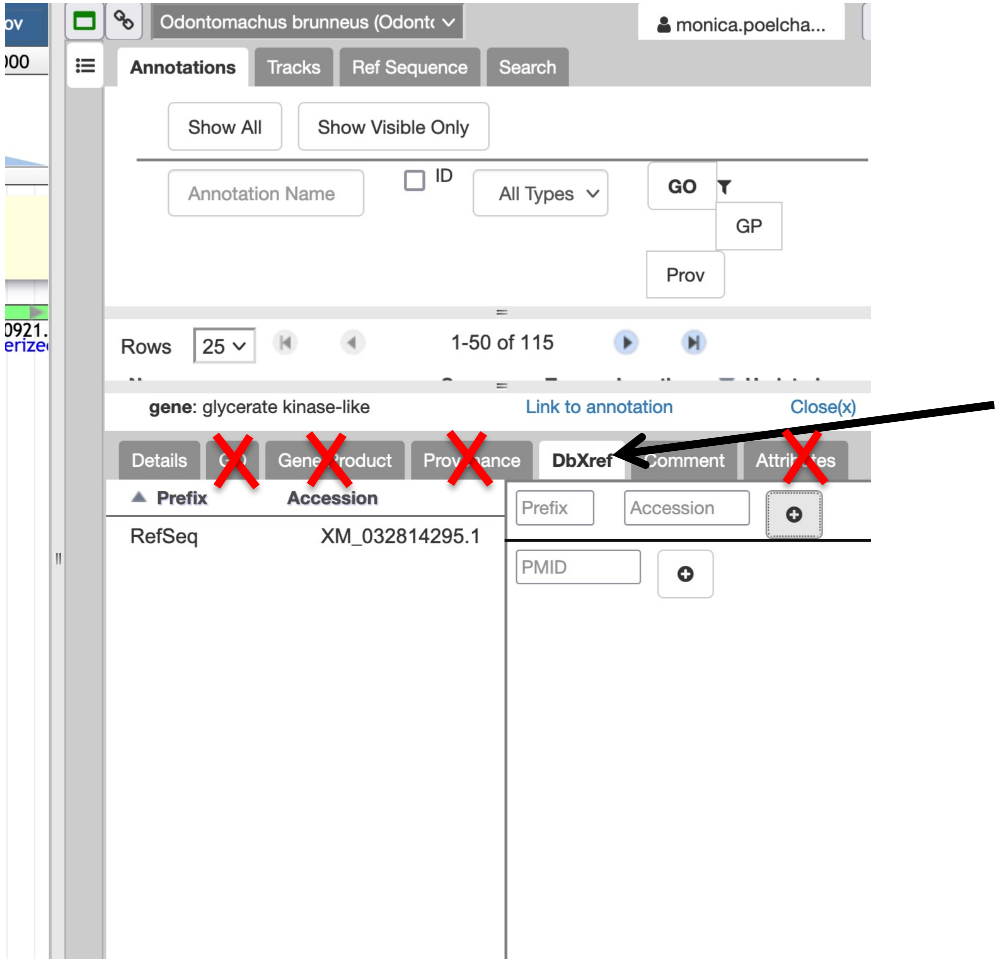
Extended 3' UTR

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



gene: glycerate kinase-like

Prefix	Accession
RefSeq	XM_032814295.1

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.
- https://i5k.nal.usda.gov/data-management-policy

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
    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);
    2. Please cite the i5k Workspace paper! This helps us continue to exist.
        1. 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
- i5k@ars.usda.gov