



# BPRI Annotation Training

3/30/2023

# BPRI annotation jamboree goals and outcomes

- Goals:
  - Educational goal: everyone is educated on how to look at a genome with a question in mind, make a tree, generate an phylogenetic insight and perhaps design some primers, RNAi or crispr reagent
  - Integration goal: integrate the genomes in a phylogenomic manner with all other BPRI research.
- Outcomes:
  - We hope that each group will have a list of 5-20 genes, for a total of  $> 100$ .
  - The aim will be a phylogenetic tree for each gene family, and a paragraph of text summarizing why you looked at those genes, what you expected based on other species, what was found.

# Agenda

1. The basics: Apollo registration; Schistocerca datasets available; NCBI annotation pages
2. Using NCBI and i5k resources to find genes in your species in Apollo
  1. Example 1: Orco
  2. Example 2: ebony
3. Naming
4. Annotation outcomes
5. 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](https://i5k.nal.usda.gov/sites/default/files/presentations/apollo_training_november_2021.pdf)
  - Recording: <https://www.zoomgov.com/rec/share/8ZxP3FhFE9ahBpVTUFDP8HwKNUS1prZIhAflINmrQaZMKeVoyo4F97RtmwJYdwQK.936mXuEkaOcjfCDW>  
Passcode: hm3tu1A@
- In-depth annotation techniques:
  - Slides: [https://i5k.nal.usda.gov/sites/default/files/presentations/Apollo\\_webinar\\_9-20-2022.pdf](https://i5k.nal.usda.gov/sites/default/files/presentations/Apollo_webinar_9-20-2022.pdf)
  - Recording: <https://www.zoomgov.com/rec/share/HckZOnl1po3QZL1Xh6OZCukps6IzKPPRBHfnTCM-UbefiwSi6x4eLcvllqY1OPw.WfMjRxdDGA2qhd9v>  
Passcode: 1%yC!#w#

# The Basics

- Ask new annotators to register here: <https://i5k.nal.usda.gov/web-apollo-registration>
  - Approval process: ideally, send me a pre-approved list of emails
- Datasets
  - Schistocerca americana example: [https://i5k.nal.usda.gov/bio\\_data/1394317](https://i5k.nal.usda.gov/bio_data/1394317)
  - RNA-Seq alignments in Jbrowse: <https://apollo.nal.usda.gov/apollo/4615055/jbrowse/index.html>
- The RefSeq gene predictions
  - Schistocerca nitens AR example: [https://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Schistocerca\\_nitens/100/](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Schistocerca_nitens/100/)

# Finding your gene in Apollo - Orco

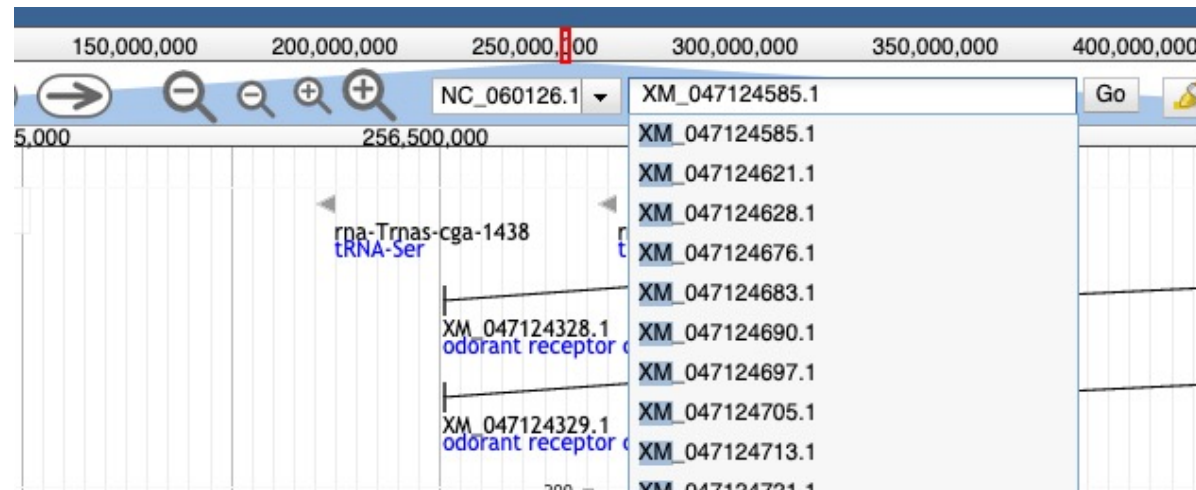
- Using NCBI's gene name search (I don't recommend it)
  - Can try using the advanced search in both protein and gene databases
  - I have contacted NCBI about some inconsistencies in their search results
  - That said, nomenclature inconsistencies between different sources (RefSeq vs. self-submitted) still make a comprehensive name-based search impractical

# Finding your gene in Apollo - Orco

- BLAST-base sequence retrieval
  - D. melanogaster record in RefSeq: NP\_524235.2
  - NCBI BLAST: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
    - Gives you a more comprehensive search, allows you to find proteins beyond the RefSeq predictions
  - I5k BLAST: <https://i5k.nal.usda.gov/webapp/blast>
    - Allows you to limit your BLAST search only to the models in Apollo, find the identifiers you need

# Finding your gene in Apollo - Orco

- To find your gene in Apollo, you need the name and/or the NCBI accession number
  - starts with XM - nucleotide - or XP – protein
- Search for either of these in the search box





# Finding your gene in Apollo - ebony

- No search results for 'Schistocerca americana' and 'ebony' in NCBI
- BLAST approach with Fly ebony protein sequence (NP\_524431.2)
- Best S. americana hit: [XP\\_046995316.1/mycosubtilin synthase subunit C isoform X2 \[Schistocerca americana\]](#)
- Search for XP\_046995316.1 in Apollo

# Naming

# Names vs. symbols vs. accession numbers

- **Name:** Describes the function of a gene or protein, e.g. “odorant receptor coreceptor”.
- **Symbol:** A short form of the Name, e.g. ‘Orco’. Only applies to the gene (not protein).
- **Accession number:** The permanent, unique identifier for a feature, e.g. XM\_050086732.1. These are provided and maintained by NCBI.

name	Seq	type	Length	Updated
Odorant receptor coreceptor	NC_064646.1	gene	340,278	Mar 29, 2023
Odorant receptor coreceptor		mRNA	340,278	Mar 29, 2023

gene: Odorant receptor coreceptor [Link to annotation](#) [Close\(x\)](#)

Details GO Gene Product Provenance DbXref Comment Attributes

☐ Obsolete

Type: **gene** ← Status: No status selected

Name: Odorant receptor coreceptor ← Name goes here

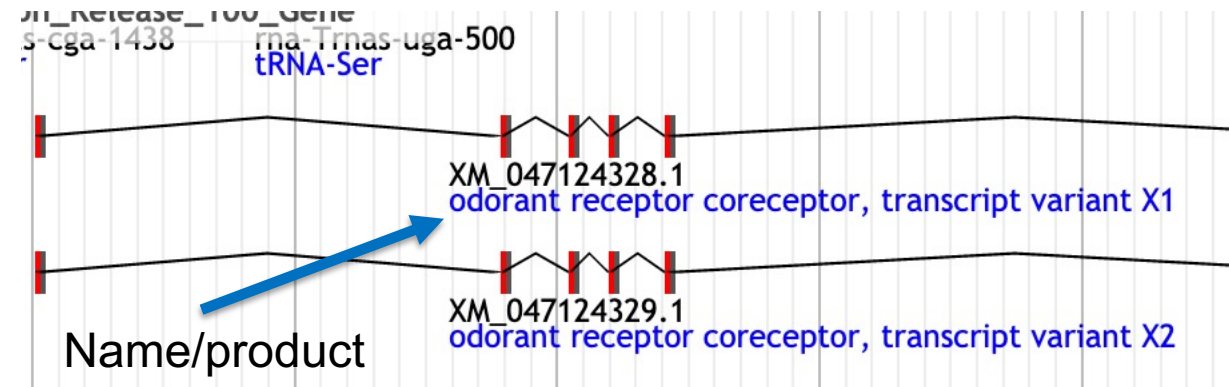
Symbol: Orco ← Symbol goes here

Aliases ('|' separated)

Description

# Names vs. symbols vs. accession numbers

- The protein name (aka product in NCBI terms) is in blue.
- The accession number is in black.
- Change the NCBI name in Apollo if it is incorrect.



## Orco Odorant receptor co-receptor [ *Drosophila melanogaster* (fruit fly)

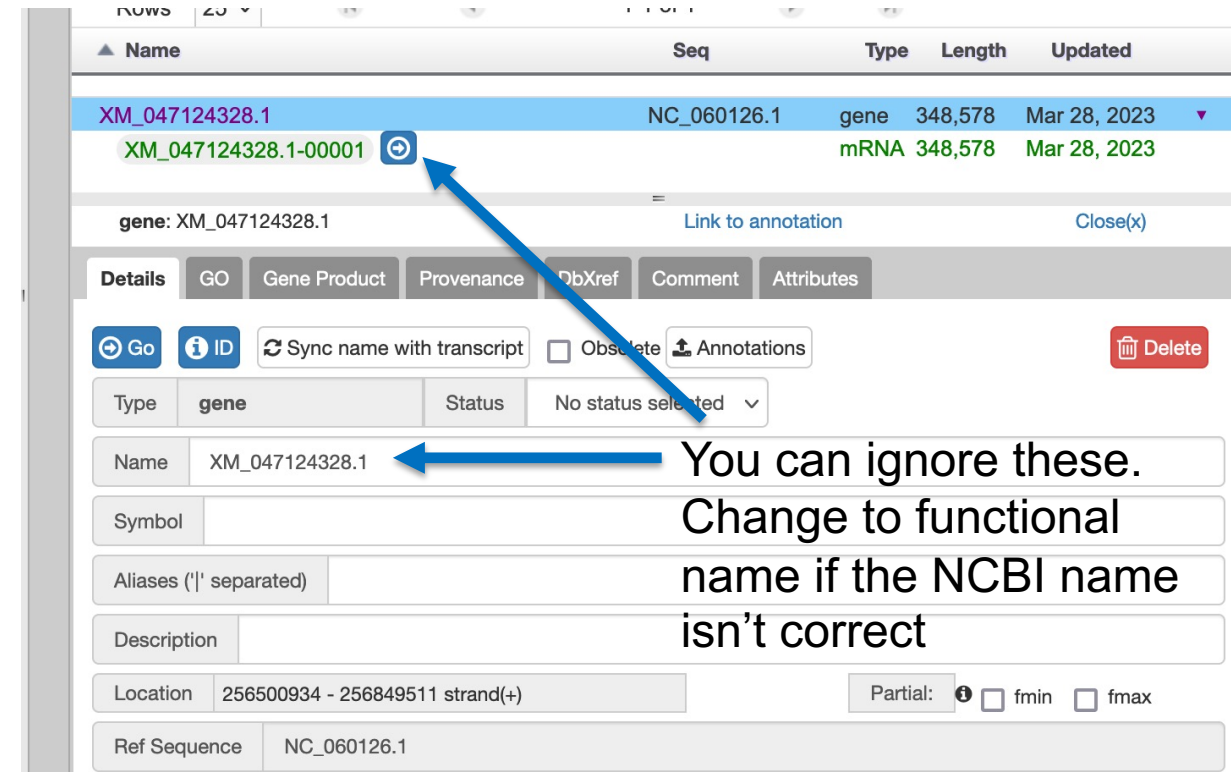
Gene ID: 40650, updated on 9-Mar-2023

### Summary

Official Symbol	Orco provided by <a href="#">FlyBase</a>
Official Full Name	Odorant receptor co-receptor provided by <a href="#">FlyBase</a>
Primary source	<a href="#">FLYBASE:FBgn0037324</a>
Locus tag	Dmel_CG10609
See related	<a href="#">AllianceGenome:FB:FBgn0037324</a>
Gene type	protein coding
RefSeq status	REVIEWED
Organism	<i>Drosophila melanogaster</i>

# A word on Apollo default 'names'

- Apollo adds placeholder names when you create a model
  - the NCBI accession for gene names
  - the NCBI accession plus a number for the mRNA.
- This is just Apollo's default behavior – this is NOT the new name of the model.
- I will remove placeholder names before submitting to NCBI.
- Please change these if you need to make a change to the actual name (e.g. odorant receptor co-receptor)





The screenshot shows the Apollo genome browser interface. At the top, there is a table of sequences with columns: Name, Seq, Type, Length, and Updated. The table lists two entries: XM\_047124328.1 (gene, 348,578 bp, Mar 28, 2023) and XM\_047124328.1-00001 (mRNA, 348,578 bp, Mar 28, 2023). Below the table, there is a details panel for the gene XM\_047124328.1. The panel includes tabs for Details, GO, Gene Product, Provenance, ObXref, Comment, and Attributes. The Details tab is active, showing fields for Name, Symbol, Aliases, Description, Location, and Ref Sequence. The Name field contains the placeholder name XM\_047124328.1. A blue arrow points from the text 'You can ignore these. Change to functional name if the NCBI name isn't correct' to the Name field.

You can ignore these.  
Change to functional  
name if the NCBI name  
isn't correct

# 15k 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>

# 15k 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>




# Symbols

- Symbol:
  - a short form of the name, e.g. Orco
  - Is assigned to the gene (not protein)
  - We don't recommend coining new symbols – okay to adopt existing ones, though
- You can add a symbol at the gene level in Apollo if there is an existing one in an ortholog. This may help with searching in NCBI in the long run.

**Orco** Odorant receptor co-receptor [ *Drosophila melanogaster* (fruit fly) ]

Gene ID: 40650, updated on 9-Mar-2023


**Summary**

<b>Official Symbol</b>	Orco provided by <a href="#">FlyBase</a>	 Symbol in NCBI gene page
<b>Official Full Name</b>	Odorant receptor co-receptor provided by <a href="#">FlyBase</a>	
<b>Primary source</b>	<a href="#">FLYBASE:FBgn0037324</a>	
<b>Locus tag</b>	Dmel_CG10609	
<b>See related</b>	<a href="#">AllianceGenome:FB:FBgn0037324</a>	

Details GO Gene Product Provenance DbXref Comment Attributes

☐ Obsolete

Type	gene	Status	No status selected	<input type="button" value="Delete"/>	
Name	Odorant receptor coreceptor				
Symbol	Orco	 Symbol goes here in Apollo			
Aliases (' ' separated)					

# 15k Workspace Guidelines - Symbols

- 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 in Apollo (e.g. Clec-Pepck). Okay to use in publications to distinguish between species, though.
- Examples: Orco, Pepck, Ser12

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

# Annotation outcomes

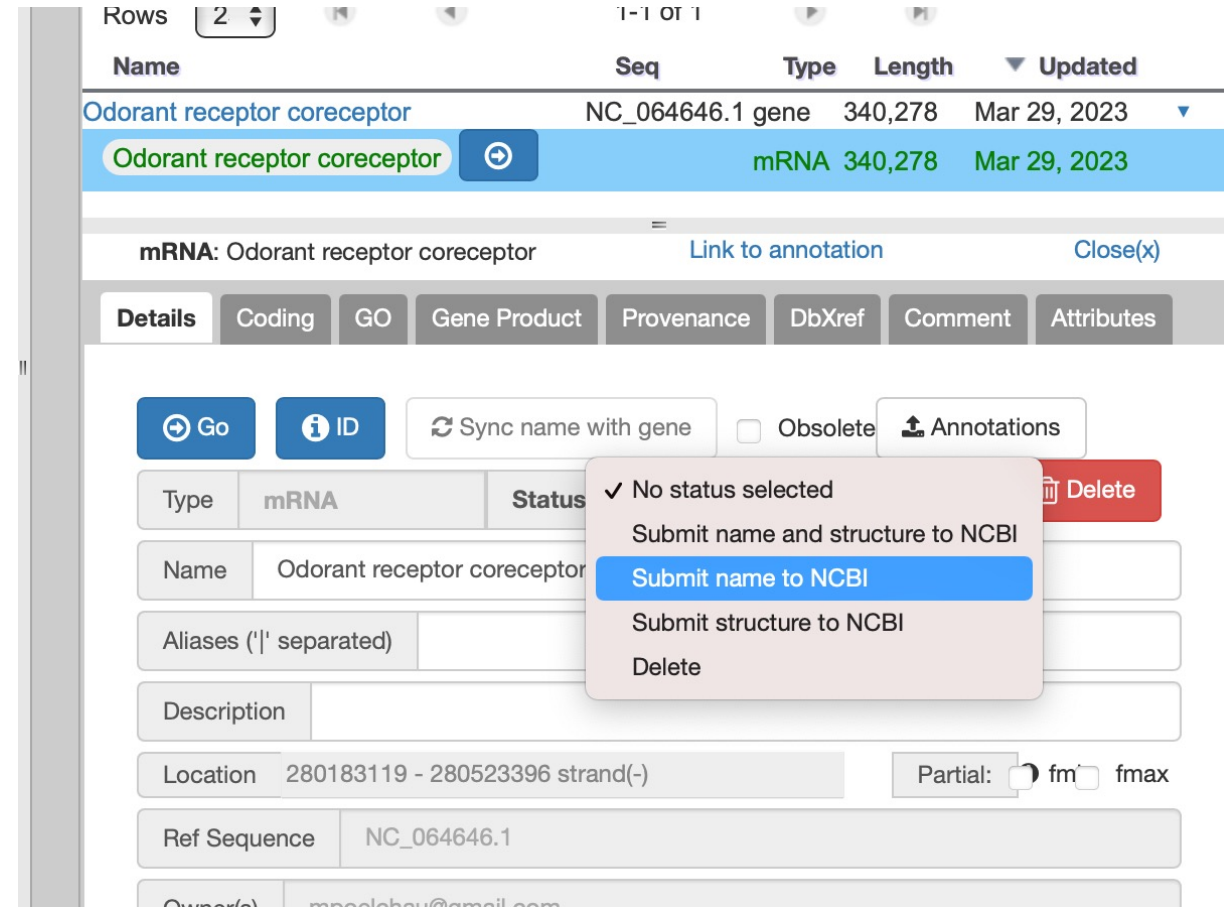
How to handle annotations with 1) different types of changes and  
2) different publication outcomes

# Annotation outcomes

- Will you refer to a changed gene model in a published paper, or should others use your changes?
- If yes, then the i5k Workspace should submit any changes to the gene model back to NCBI.
- If we don't submit your changes to NCBI, they can be submitted to a generic repository such as Ag Data Commons or Dryad – but it's much harder to re-use sequence data from these locations.

# The 'Status' field

- We will use the 'Status' field in Apollo to indicate what should be submitted to NCBI.
  - Submit name and structure to NCBI
  - Submit name to NCBI
  - Submit structure to NCBI



The screenshot displays the Apollo genome browser interface. At the top, a table lists genomic features. The feature 'Odorant receptor coreceptor' (NC\_064646.1 gene) is highlighted. Below the table, a detailed view of the 'mRNA: Odorant receptor coreceptor' is shown. The 'Status' dropdown menu is open, revealing options: 'No status selected', 'Submit name and structure to NCBI', 'Submit name to NCBI' (highlighted in blue), 'Submit structure to NCBI', and 'Delete'. The form includes fields for Name, Aliases, Description, Location, Ref Sequence, and Owner(s).

Name	Seq	Type	Length	Updated
Odorant receptor coreceptor	NC_064646.1 gene		340,278	Mar 29, 2023
Odorant receptor coreceptor		mRNA	340,278	Mar 29, 2023

mRNA: Odorant receptor coreceptor [Link to annotation](#) [Close\(x\)](#)

[Go](#) [ID](#) [Sync name with gene](#) ☐ Obsolete [Annotations](#)

Type: mRNA Status: **Submit name to NCBI**

Name: Odorant receptor coreceptor

Aliases ('|' separated):

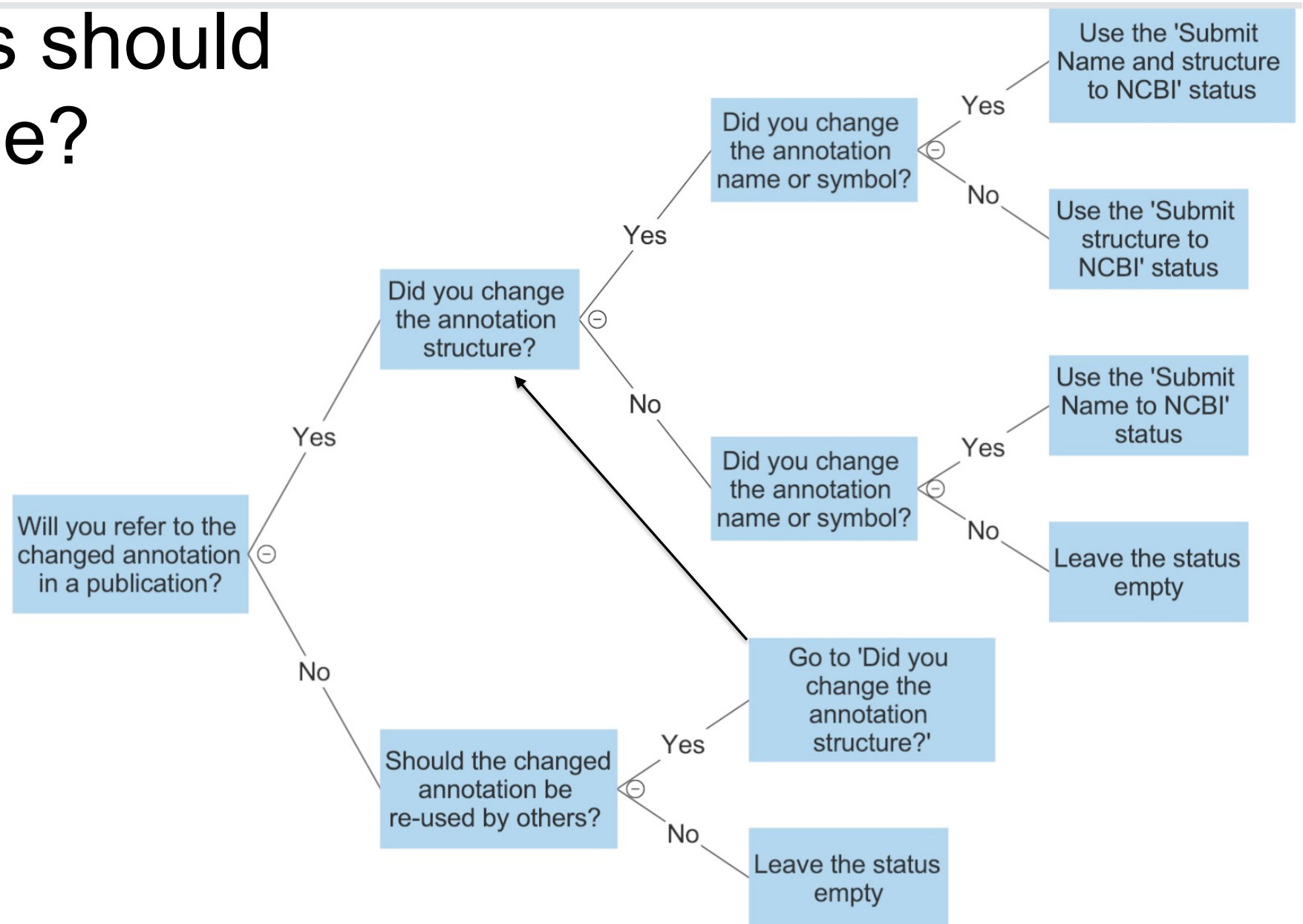
Description:

Location: 280183119 - 280523396 strand(-) Partial: ☐ fm ☐ fmax

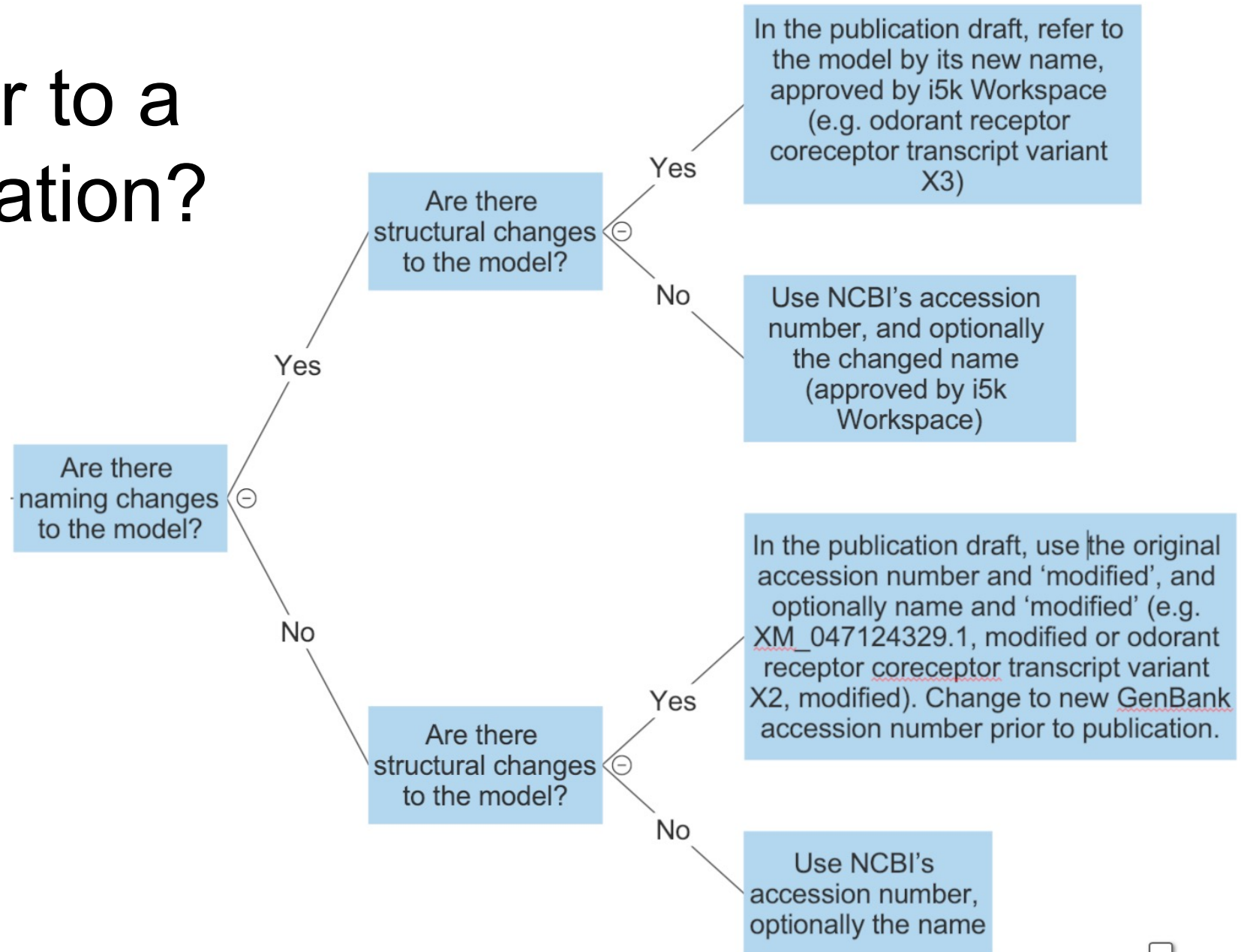
Ref Sequence: NC\_064646.1

Owner(s): mncolchau@gmail.com

# What Status should you use?



# How do you refer to a model in a publication?



# Other procedural notes

- It would help to send me a list of annotator emails
- We're missing *S. gregaria* RNA-Seq, and protein alignments
- Keep in touch with me if you have any questions
- Let me know when you'd like me to review your annotations that should be submitted to NCBI – this can take some time.
- Keep me up to date on your publication timelines, so I can communicate with NCBI



# Thank you!

- Questions?