A. *thaliana*, Apollo and You:

Collaborative Genome Annotation Editing

With many thanks to Moni Munoz Torres, Apollo Expert extraordinaire

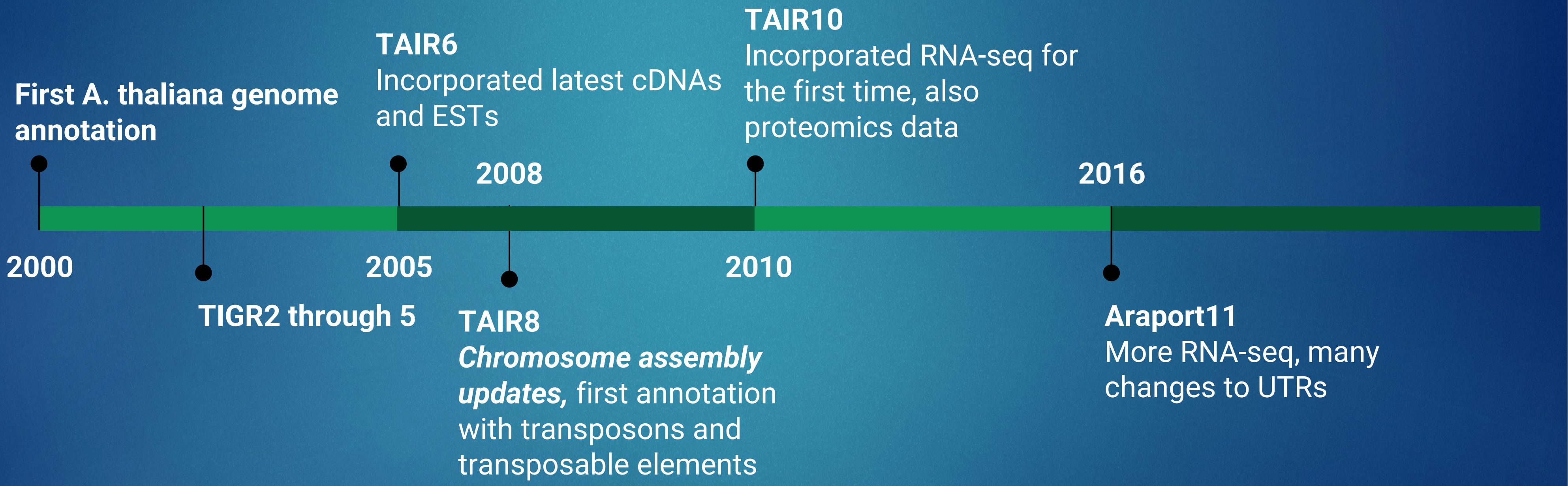
Today's team:

- Tanya Berardini
  - Screen sharing, leading hands on
- Shabari Subramaniam
  - Technical issues
  - Monitoring chat for questions
  - Sharing links and other info in the chat

Today:

- Brief overview of the Arabidopsis reannotation project
- Intro to Apollo and manual review
- Hands on exercises
- What’s next

Rough Timeline



First A. thaliana genome annotation

TIGR2 through 5

TAIR6
Incorporated latest cDNAs and ESTs

TAIR8
Chromosome assembly updates, first annotation with transposons and transposable elements

TAIR10
Incorporated RNA-seq for the first time, also proteomics data

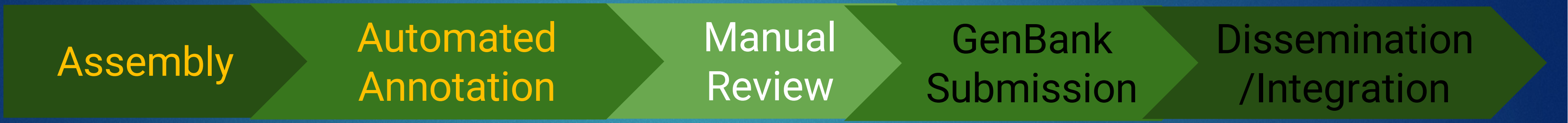
Araport11
More RNA-seq, many changes to UTRs

What has changed over the past 20 years?

- Greater amounts of supporting data
- Increased kinds of supporting data
- Improved sequencing technology, longer reads
- Improved genome assembly software
- Improved automated annotation pipelines

TAIR: coordinator, community hub

Reannotation project phases



Assembly

Automated Annotation

Manual Review

GenBank Submission

Dissemination /Integration

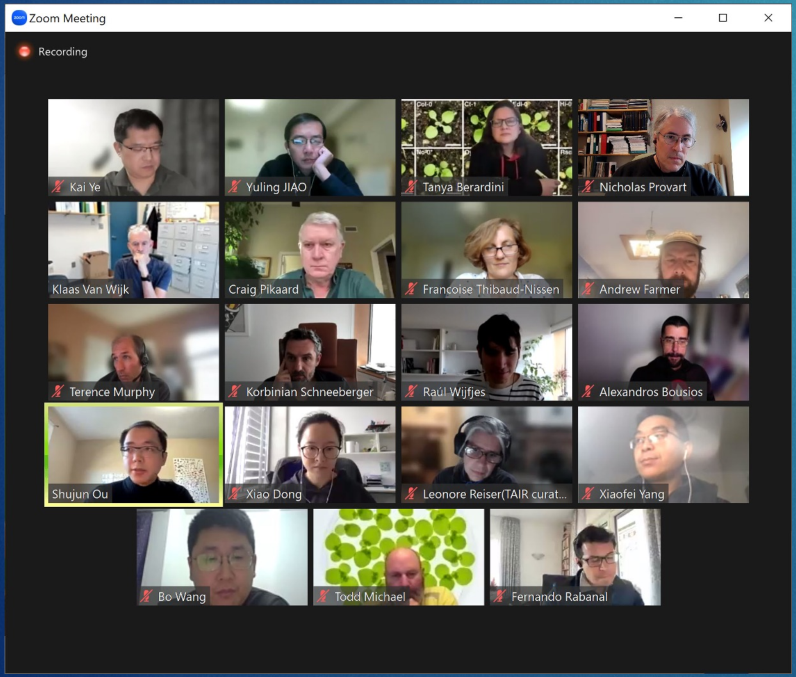
Who: Schneeberger lab (MPI)

Who: NCBI (National Center for Bioinformatics)

Who: Community experts for review, TAIR for coordination, tool hosting

Who: TAIR + NCBI

Who: BAR, TAIR, EnsemblPlants, NCGR GCV, AtPeptide Atlas, many more



After this session, you will be able to:

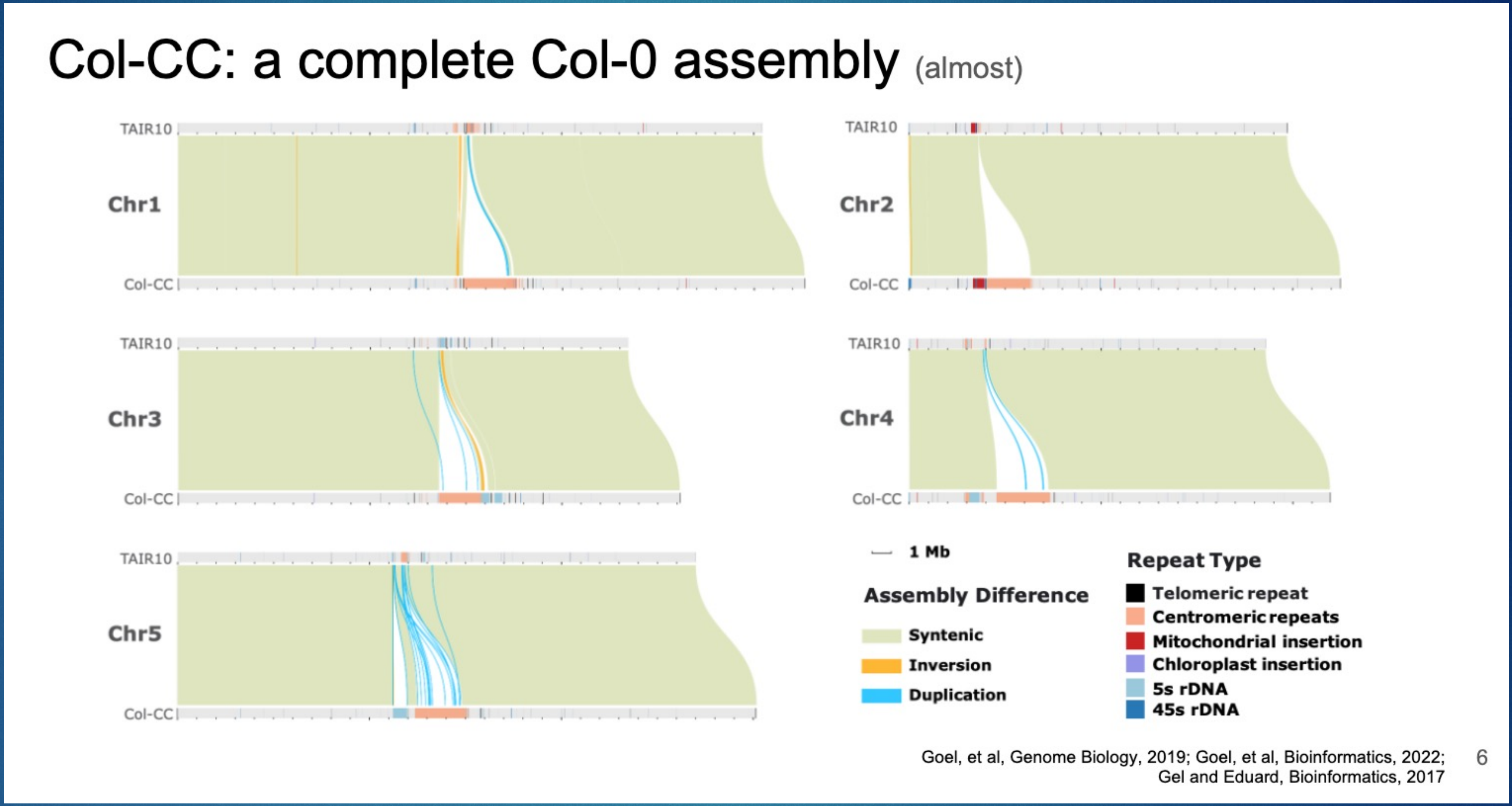
- Log into Apollo and view annotation and evidence tracks
- Navigate through the interface, find genes, zoom in to sequence level
- Perform basic gene model manipulation with ease
- Save comments and status
- Access resources in case you need help remembering what we covered

Known issues: Patience appreciated

- Mt and Cp annotations not yet visible
- Load time especially the first time can be slow
- Some evidence tracks are still missing (working hard on this)

The story so far

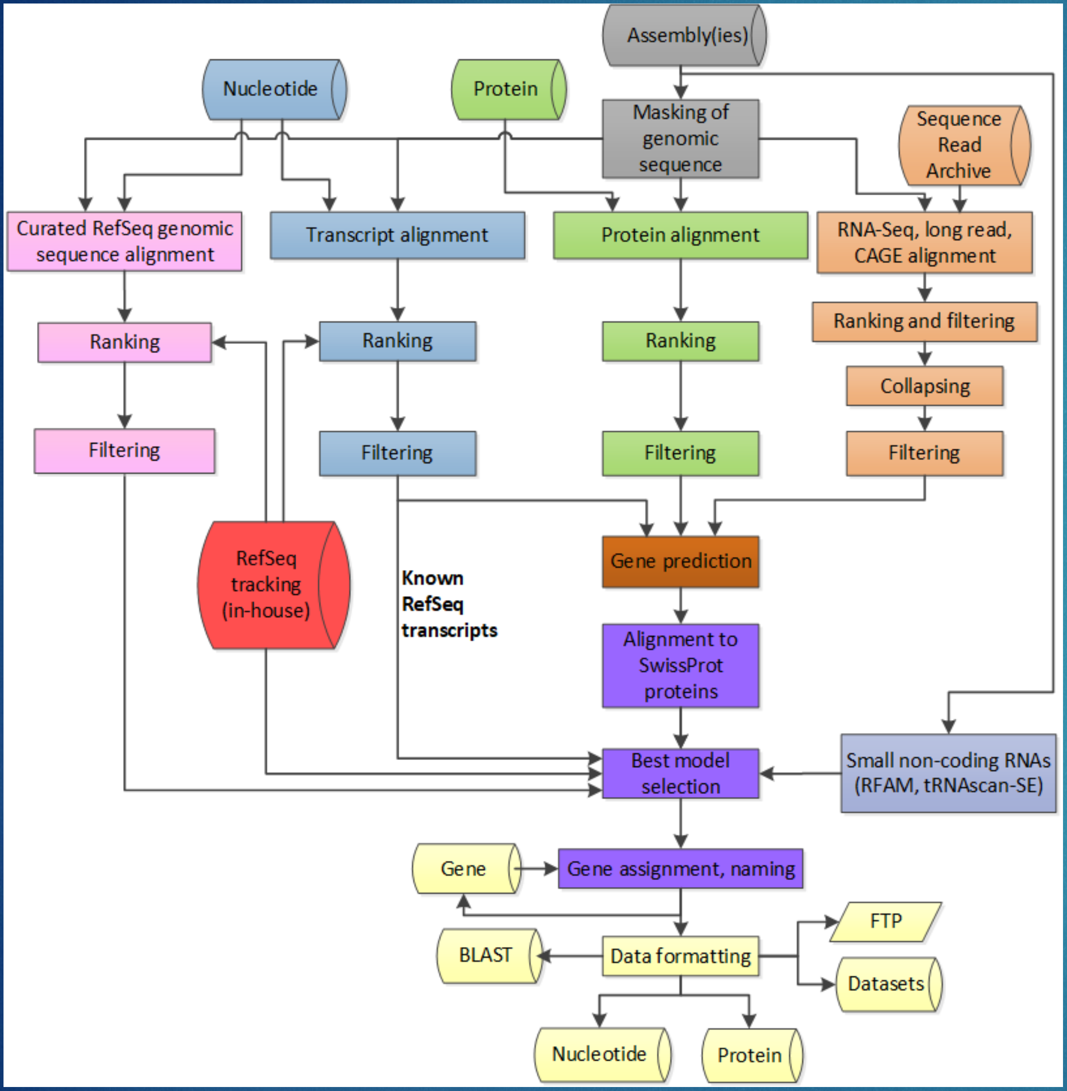
Col-CC = 13 Col-0 assemblies



Col-CC: a complete Col-0 assembly (almost)

Goel, et al, Genome Biology, 2019; Goel, et al, Bioinformatics, 2022; Gel and Eduard, Bioinformatics, 2017

The NCBI Eukaryotic Annotation Pipeline



- Automated
- Highly dependent on experimental data
  - Proteins
  - cDNA
  - RNA-Seq
  - IsoSeq, ONT
  - CAGE

Changes between Araport11 and Col-CC

- New genes
- Deleted genes
- Split genes
- Merged genes
- Changes in CDS
- Fewer alternative transcripts

Why do manual review?

- Remove elements reflecting errors in automated analyses
- To accurately annotate gene families
- To verify novel genes and isoforms
- To efficiently take advantage of transcriptomic analyses
- Achieve the best representation of the genome for translational use in other organisms

Today:

- Brief overview of the Arabidopsis reannotation project
- Intro to Apollo and manual review
- Hands on exercises
- What’s next

JBrowse

Apollo

Collaborative, instantaneous,
web-based, built on top of JBrowse.

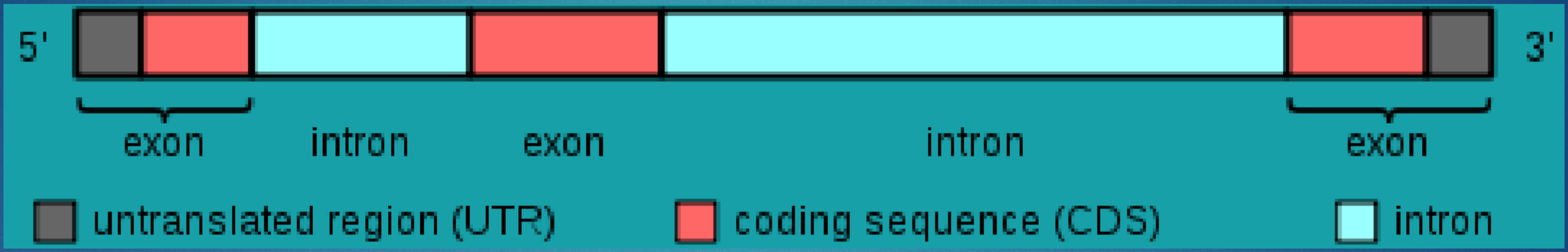
GMOD

General process of manual review

1. Select or find a **region of interest** (e.g., gene or coordinate range).
2. Select appropriate **evidence** tracks to review the genome element to annotate (e.g., gene model).
3. If necessary, **adjust** the gene model.
4. Check your edited gene model for **integrity and accuracy** by comparing it with available homologs.
5. **Comment, change status,** and finish.

A brief refresher: focus on protein-coding genes

mRNA structure



"Gene structure" by Daycd- Wikimedia Commons

Splice sites

Splicing “signals” (from the point of view of an intron):

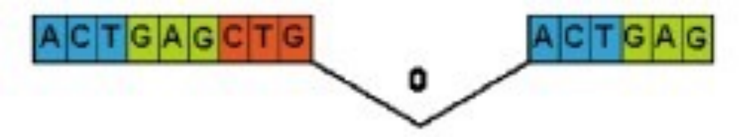
- 5’ end splice “signal” (site): usually GT (less common: GC)
- 3’ end splice site: usually AG

…]5’ - GT / AG - 3’[…

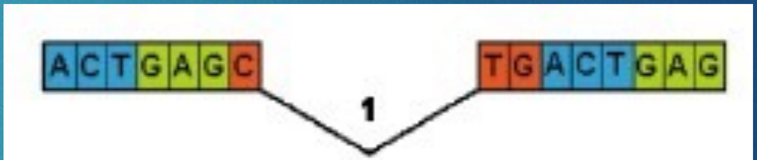
Alternatively bringing exons together produces more than one protein from the same genic region: isoforms.

Exons and Introns

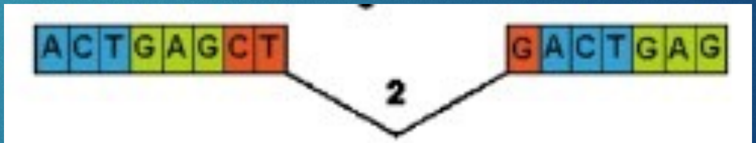
- Introns can interrupt the reading frame of a gene by inserting a sequence between two consecutive codons



- Between the first and second nucleotide of a codon



- Or between the second and third nucleotide of a codon



Today:

- Brief overview of the Arabidopsis reannotation project
- Intro to Apollo and manual review
- Hands on exercises
- What’s next

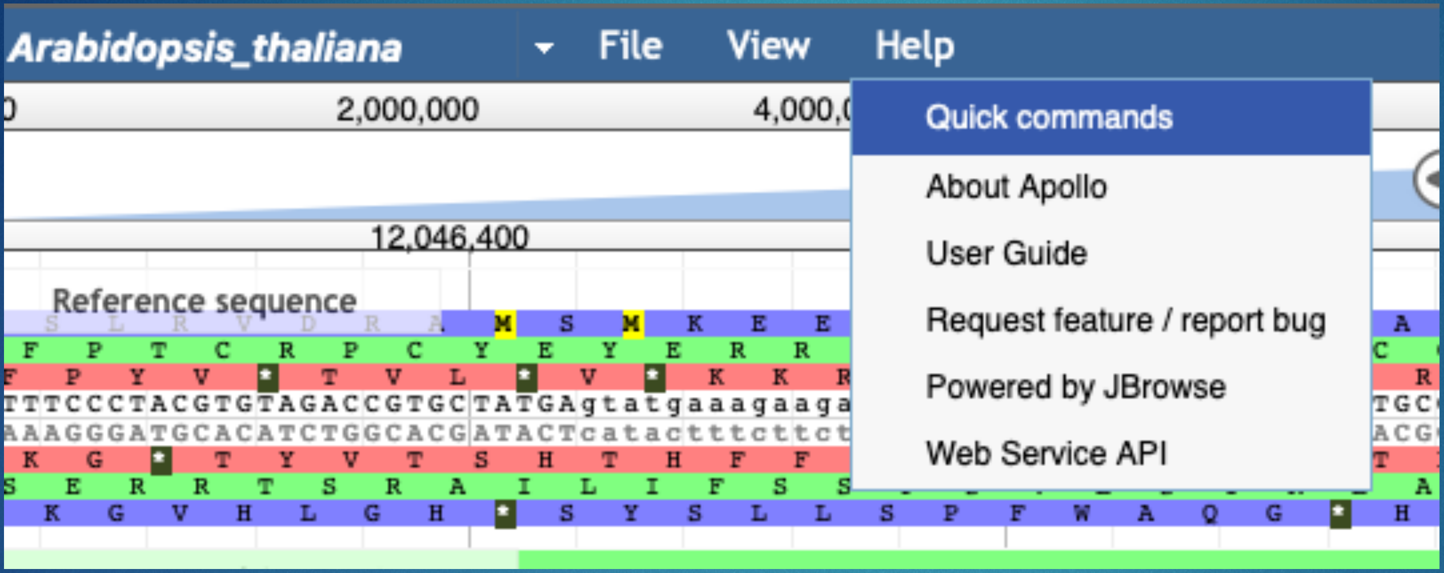
Set up for success: Have these tabs ready to go in your browser window

- Apollo: http://ec2-34-221-77-232.us-west-2.compute.amazonaws.com:8080/apollo/annotator/index
- Apollo User Guide: https://genomearchitect.readthedocs.io/en/latest/UsersGuide.html
- Jbrowse:https://jbrowse.arabidopsis.org/index.html?data=Araport11&loc=Chr5%3A8946356..8953040&tracks=TAIR10_genome%2CA11-GL-Jan23%2CA11-PC-Jan23&highlight=
- BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Activity

- Set up tabs
- Log into Apollo
- Set up tracks

Tips and tricks: HELP!



Quick commands

About Apollo

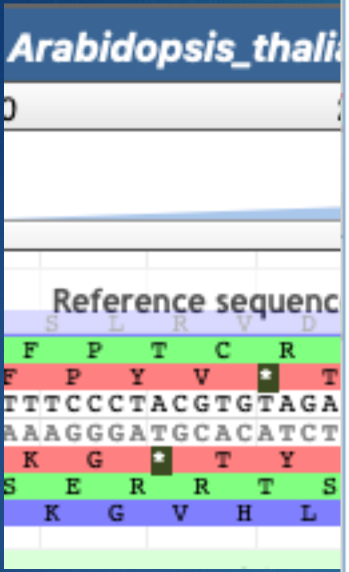
User Guide

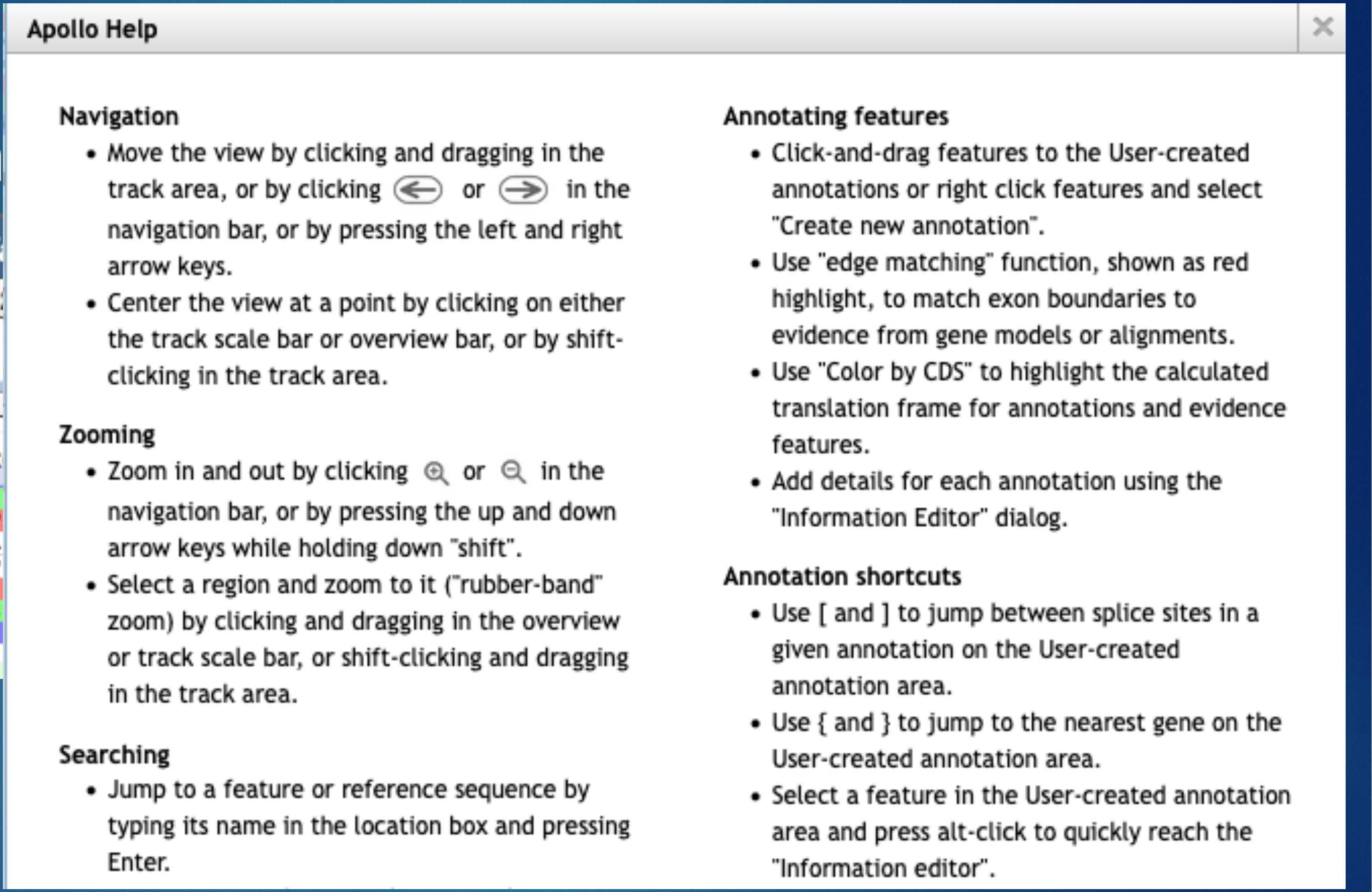
Request feature / report bug

Powered by JBrowse

Web Service API

Tips an





Apollo Help

Navigation

- Move the view by clicking and dragging in the track area, or by clicking ⇐ or ⇒ in the navigation bar, or by pressing the left and right arrow keys.
- Center the view at a point by clicking on either the track scale bar or overview bar, or by shift-clicking in the track area.

Zooming

- Zoom in and out by clicking ⊕ or ⊖ in the navigation bar, or by pressing the up and down arrow keys while holding down "shift".
- Select a region and zoom to it ("rubber-band" zoom) by clicking and dragging in the overview or track scale bar, or shift-clicking and dragging in the track area.

Searching

- Jump to a feature or reference sequence by typing its name in the location box and pressing Enter.

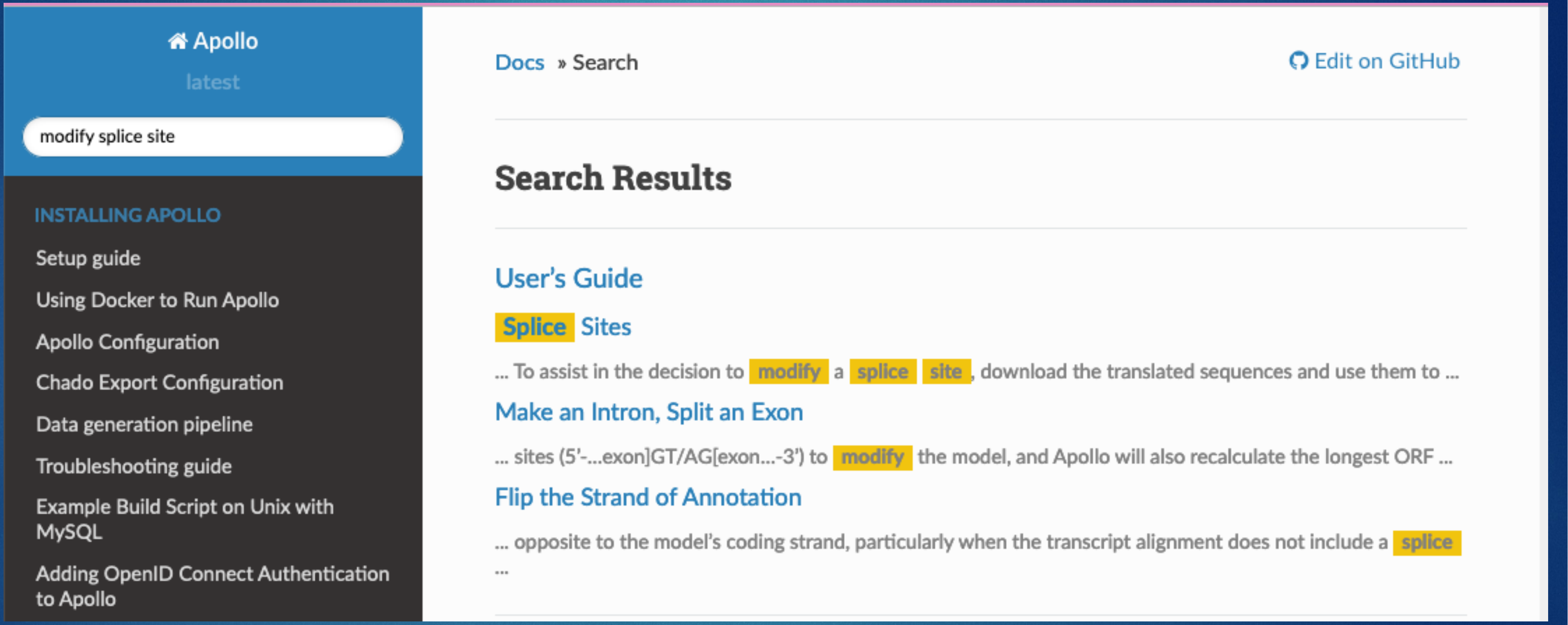
Annotating features

- Click-and-drag features to the User-created annotations or right click features and select "Create new annotation".
- Use "edge matching" function, shown as red highlight, to match exon boundaries to evidence from gene models or alignments.
- Use "Color by CDS" to highlight the calculated translation frame for annotations and evidence features.
- Add details for each annotation using the "Information Editor" dialog.

Annotation shortcuts

- Use [and] to jump between splice sites in a given annotation on the User-created annotation area.
- Use { and } to jump to the nearest gene on the User-created annotation area.
- Select a feature in the User-created annotation area and press alt-click to quickly reach the "Information editor".

Tips and tricks: Apollo Help Docs



Apollo

latest

modify splice site

INSTALLING APOLLO

Setup guide

Using Docker to Run Apollo

Apollo Configuration

Chado Export Configuration

Data generation pipeline

Troubleshooting guide

Example Build Script on Unix with MySQL

Adding OpenID Connect Authentication to Apollo

Docs » Search

Edit on GitHub

Search Results

User's Guide

Splice Sites

... To assist in the decision to modify a splice site, download the translated sequences and use them to ...

Make an Intron, Split an Exon

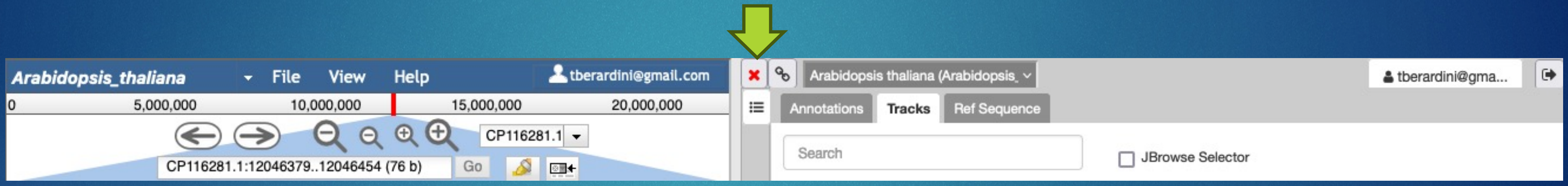
... sites (5'-...exon]GT/AG[exon...-3') to modify the model, and Apollo will also recalculate the longest ORF ...

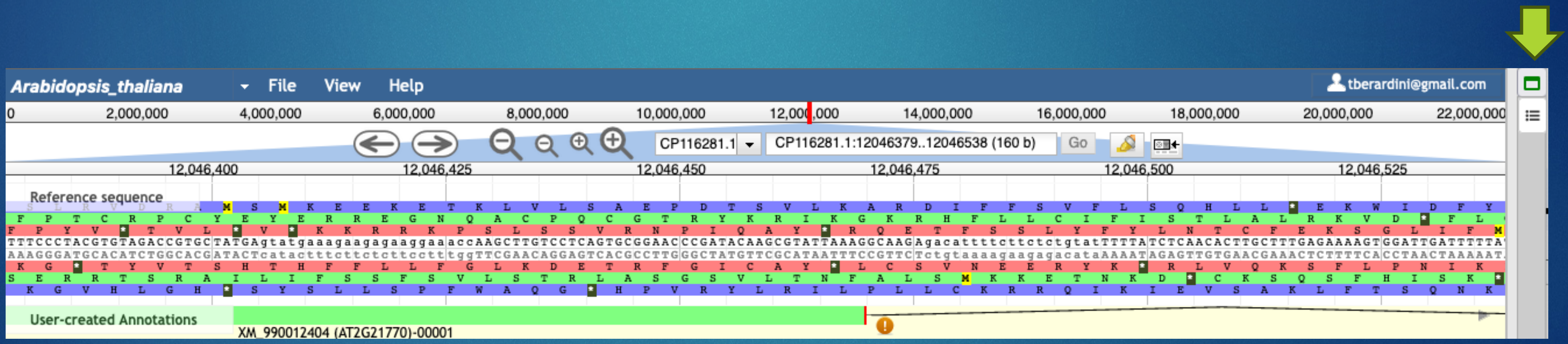
Flip the Strand of Annotation

... opposite to the model's coding strand, particularly when the transcript alignment does not include a splice ...

https://genomearchitect.readthedocs.io/en/latest/search.html

Tips and tricks: Show/hide sidebar





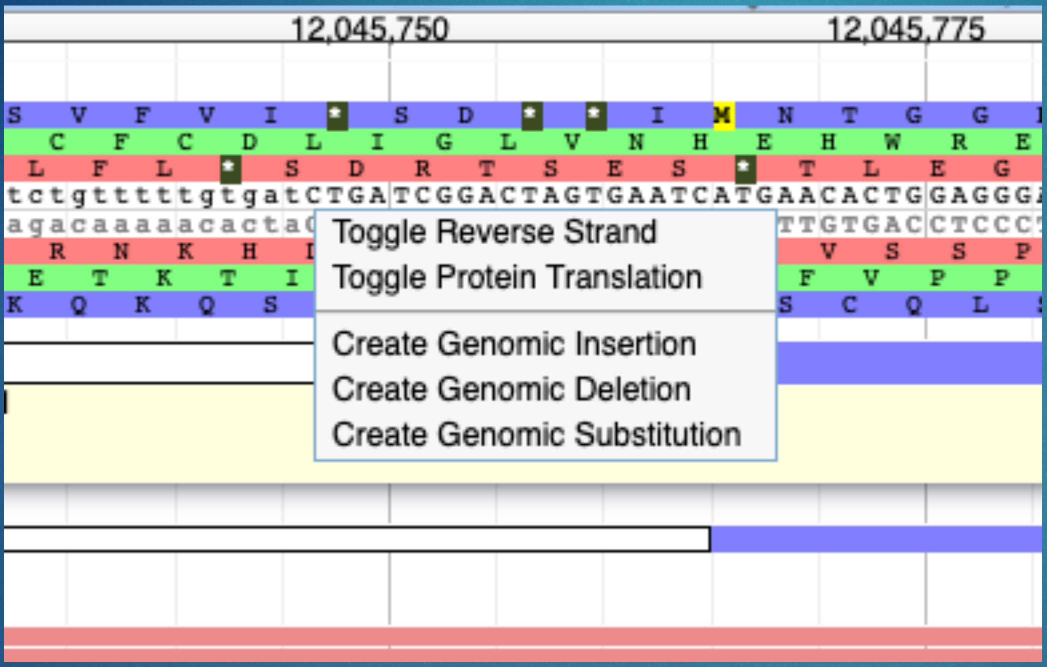
Exercise 1: Finding regions of interest

- Working with protein coding genes for now
- Search by AGI: AT1G69120
  - Right click menu
- Search by chromosome + coordinates: CP116282.1:19803781..19806663
- What are we looking at ?
  - Colored boxes = exons, coding sequence, reading frames
  - Clear boxes = exons, UTRs
  - Arrows = direction of transcription/translation
  - Lines = introns
- Sanity check: JBrowse view of same AGI/region

Exercise 2: Creating your own gene model

- Groups
  - 1: AT1G45545
  - 2: AT2G21850
  - 3: AT3G46510
  - 4: AT4G20060
  - 5: AT5G25640
- Click on intron, highlight whole gene annotation
- Drag from Col-CC Annotation track (or Gnomon track) into user-created Annotation band (Yellow)
- Rename
- Delete
- Zoom to Base level
- Toggle sequences

Tips and tricks: Toggle sequences



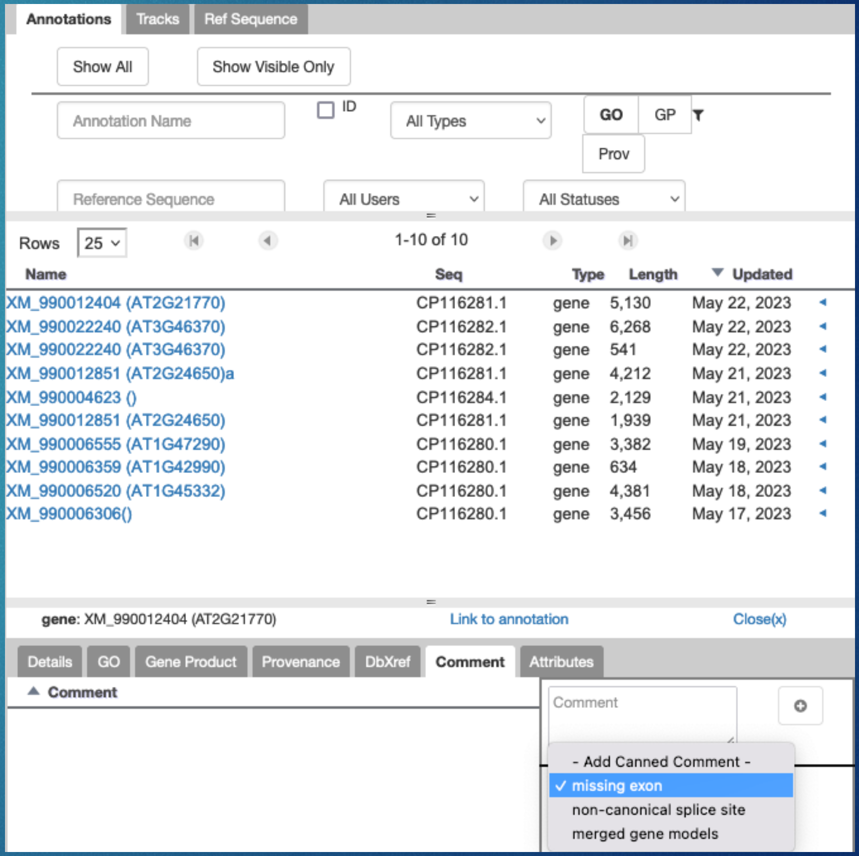
Click in DNA region

Exercise 3: Adjusting exons and introns

- Create new entry in user-created annotations again
- Verify direction of translation
- Check beginning of translation
- Adjust start of translation
- (Check RNA seq)
- Delete an exon
- Undo
- Delete an intron
- Undo

Tips and tricks: Saving comments/status

- Click out of the panel you've changed into another one



Name	Seq	Type	Length	Updated
XM_990012404 (AT2G21770)	CP116281.1	gene	5,130	May 22, 2023
XM_990022240 (AT3G46370)	CP116282.1	gene	6,268	May 22, 2023
XM_990022240 (AT3G46370)	CP116282.1	gene	541	May 22, 2023
XM_990012851 (AT2G24650)a	CP116281.1	gene	4,212	May 21, 2023
XM_990004623 ()	CP116284.1	gene	2,129	May 21, 2023
XM_990012851 (AT2G24650)	CP116281.1	gene	1,939	May 21, 2023
XM_990006555 (AT1G47290)	CP116280.1	gene	3,382	May 19, 2023
XM_990006359 (AT1G42990)	CP116280.1	gene	634	May 18, 2023
XM_990006520 (AT1G45332)	CP116280.1	gene	4,381	May 18, 2023
XM_990006306()	CP116280.1	gene	3,456	May 17, 2023

gene: XM_990012404 (AT2G21770)

2. Click here to save

1. Enter comment

Evidence tracks

- Col-CC annotation: end result of pipeline
- Gnomon models: *one* of the inputs into the pipeline
- TSA (transcript shotgun assembly): isoseq contigs + extra isoseq
- Protein alignments: alignments of protein sequences from Genbank records (multi-species) with Col-CC models
- (RNA seq) - A. thaliana
- (Long read RNA) – A. thaliana

Today:

- Brief overview of the Arabidopsis reannotation project
- Intro to Apollo and manual review
- Hands on exercises
- What’s next

What’s next?

- Second training session: more with Apollo
  - Examining evidence
  - Detailed manipulation/editing
  - Next week: Wed/Thu (May 31/June 1)
    - 7 – 8:30 am US PDT (UTC -7)

- ICAR 2023 – June 5 – 9 : who’s attending?

What's next?

- Gene set assignment
  1. split
  2. merged
  3. deleted
  4. novel
  5. locus type changed
  6. cds changed
  7. BUSCO gene disappeared
  8. desired gene family (may overlap with 1-7)

What's next? Further out

- Website: tinyurl.com/AthalianaV12
  - Updates, training material, video will be accessible from here
  - Tracking work and review
    - Google Sheet
    - Excel spreadsheet (no Google Drive access)
- Slack channel (#athalianav12-manual-review)
  - Bug reports, asynchronous feedback/questions, paste the link to the region and the issue
- When review starts in earnest: Regular call time: Zoom, Wed 7 – 7:30 am Pacific (proposed)

Thank you!

- **Col-CC Assembly:** Korbinian Schneeberger and lab team
- **NCBI Eukaryotic Genome Pipeline:** Françoise Thibaud-Nissen, Terence Murphy
- **Apollo setup @TAIR:** Shabari Subramaniam, Xingguo Chen, Trilok Prithvi, Chris Childers
- **Training materials:** Moni Munoz Torres, Marcela Tello Ruiz, Monica Poelchau, Jason Williams
- **The wider Arabidopsis community**
- **YOU**