An Introduction to the CPT Galaxy and WebApollo for Phage Whole Genome Annotation

Jason J. Gill

Department of Animal Science
Center for Phage Technology
Texas A&M University

Center for Phage Technology

Texas A&M AgriLife Research

Genome annotation tools

Fully automated annotation

- RAST/myRAST
  - http://rast.nmpdr.org/
- Prokka
  - http://www.vicbioinformatics.com/software.prokka.shtml
- NCBI Prokaryotic Pipeline
  - https://www.ncbi.nlm.nih.gov/genome/annotation_prok/

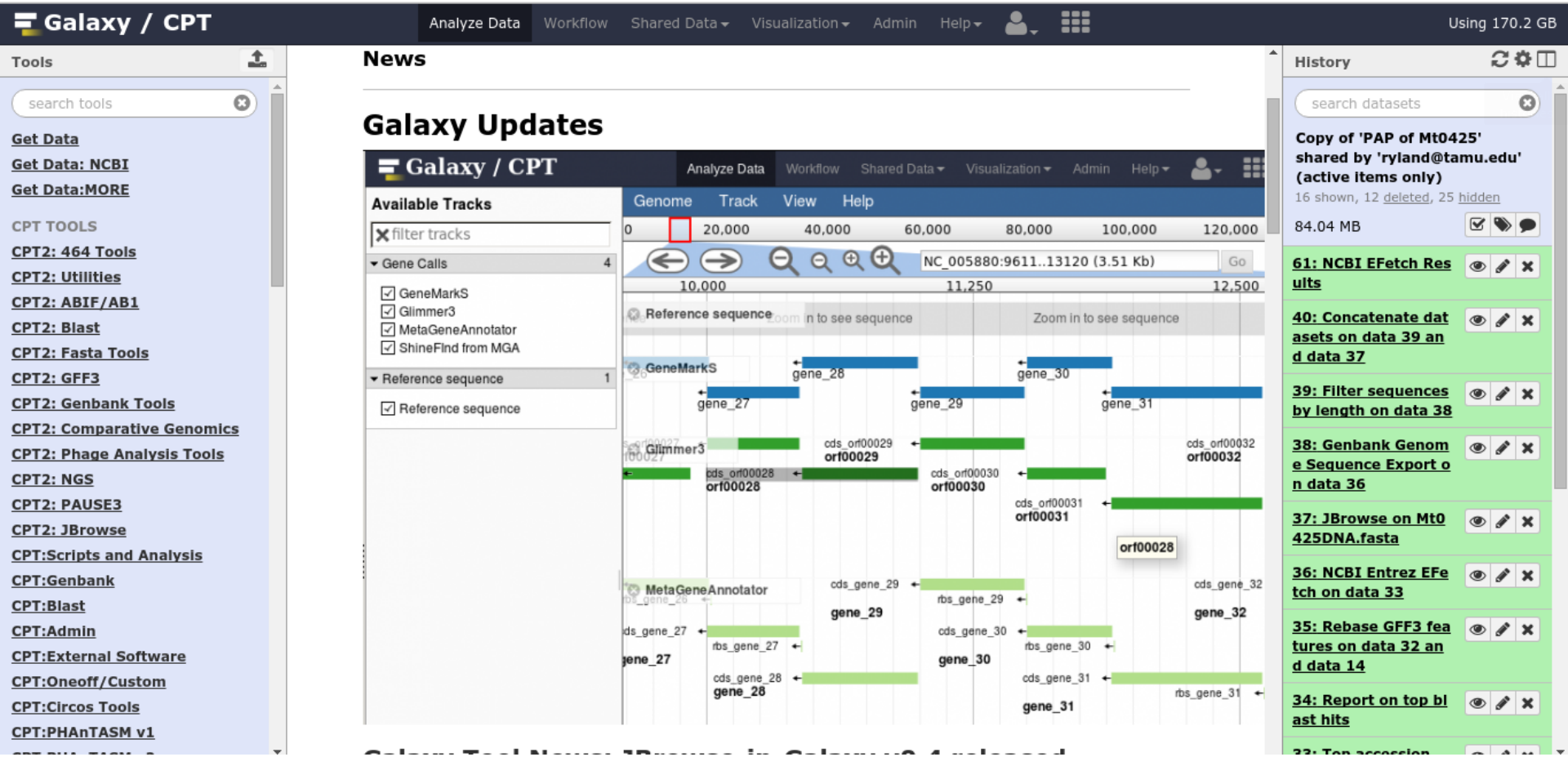
Semi-automated annotation

- DNA Master
  - http://cobamide2.bio.pitt.edu/
- CPT Galaxy/Apollo
  - https://cpt.tamu.edu/galaxy-pub/

Manual annotation / genome editors

- Sanger Artemis
  - http://www.sanger.ac.uk/science/tools/artemis
- Broad Argo
  - https://archive.broadinstitute.org/annotation/argo/

What is Galaxy?



Center for Phage Technology

Texas A&M AgriLife Research

What is Galaxy?

- Galaxy is not an analysis tool itself
- Galaxy provides a **platform** for performing reproducible bioinformatics research
- Provides a Web-browser-based **user interface** for other command-line tools
- Provides a **history** of actions performed and tool outputs
- Allows users to chain operations together in **workflows** to perform complex analyses
- Galaxy is **open-source** (free) with an active user community

What is Galaxy?

- Galaxy can interface with any Linux command-line program via a short script called a "wrapper"
  - The wrapper presents input options to the user and passes these back to the invoked program (e.g., BLASTp, Glimmer3, etc.)
- Galaxy then keeps a record of that job, its inputs and outputs in a history
- Wrappers are available for many common tools or can be written if needed
- The Galaxy **toolshed** contains available wrappers and these can be installed from within Galaxy

Why use Galaxy?

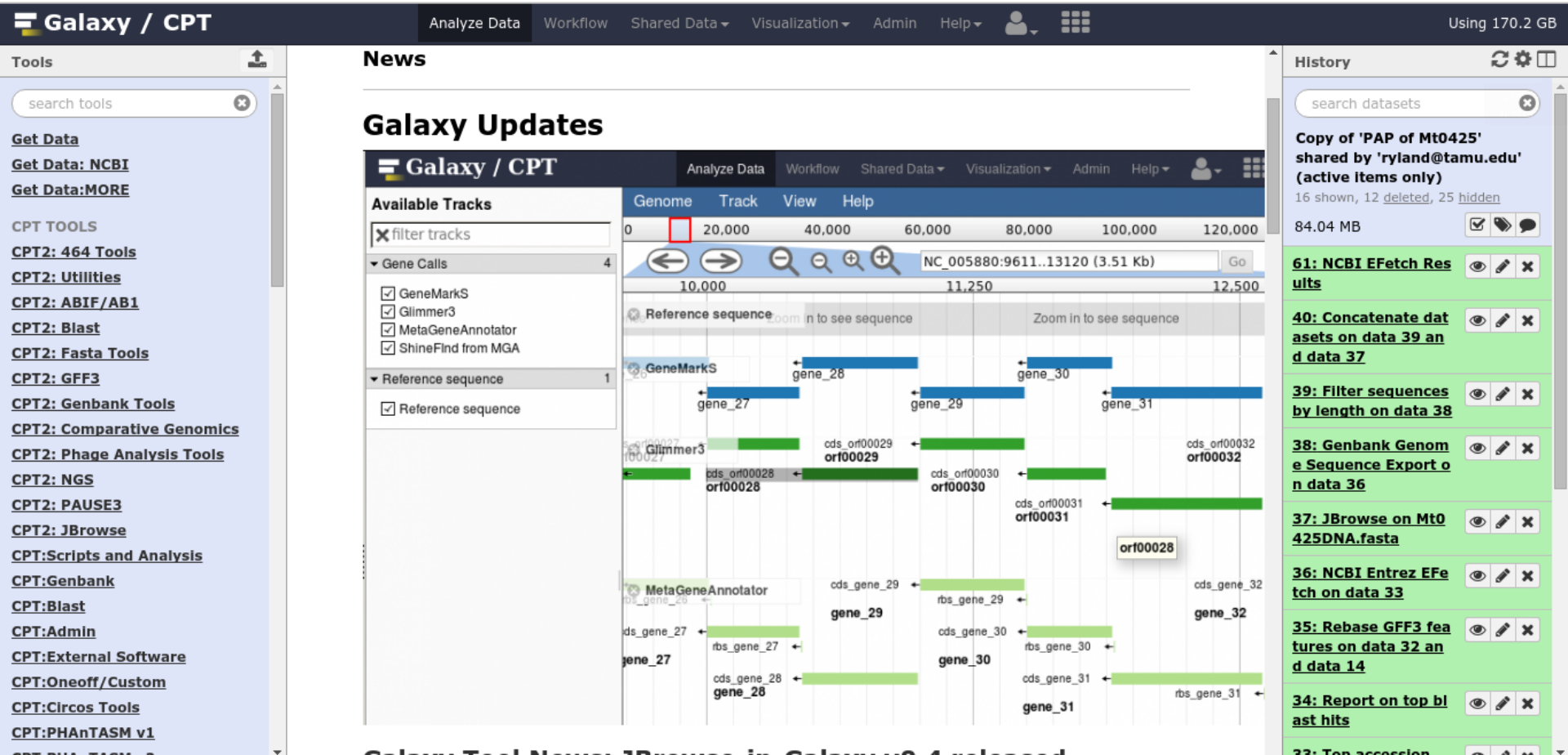
- Ultimately, Galaxy offers the power and flexibility of command-line Linux data analysis to the average biologist
- Good for teaching principles of bioinformatics without getting bogged down in command-line instructions
  - Galaxy/WebApollo is a major component of our phage genomics course
- Maintains a record of work you've done, even years later
  - Better than trying to read an old notebook!

The Galaxy interface

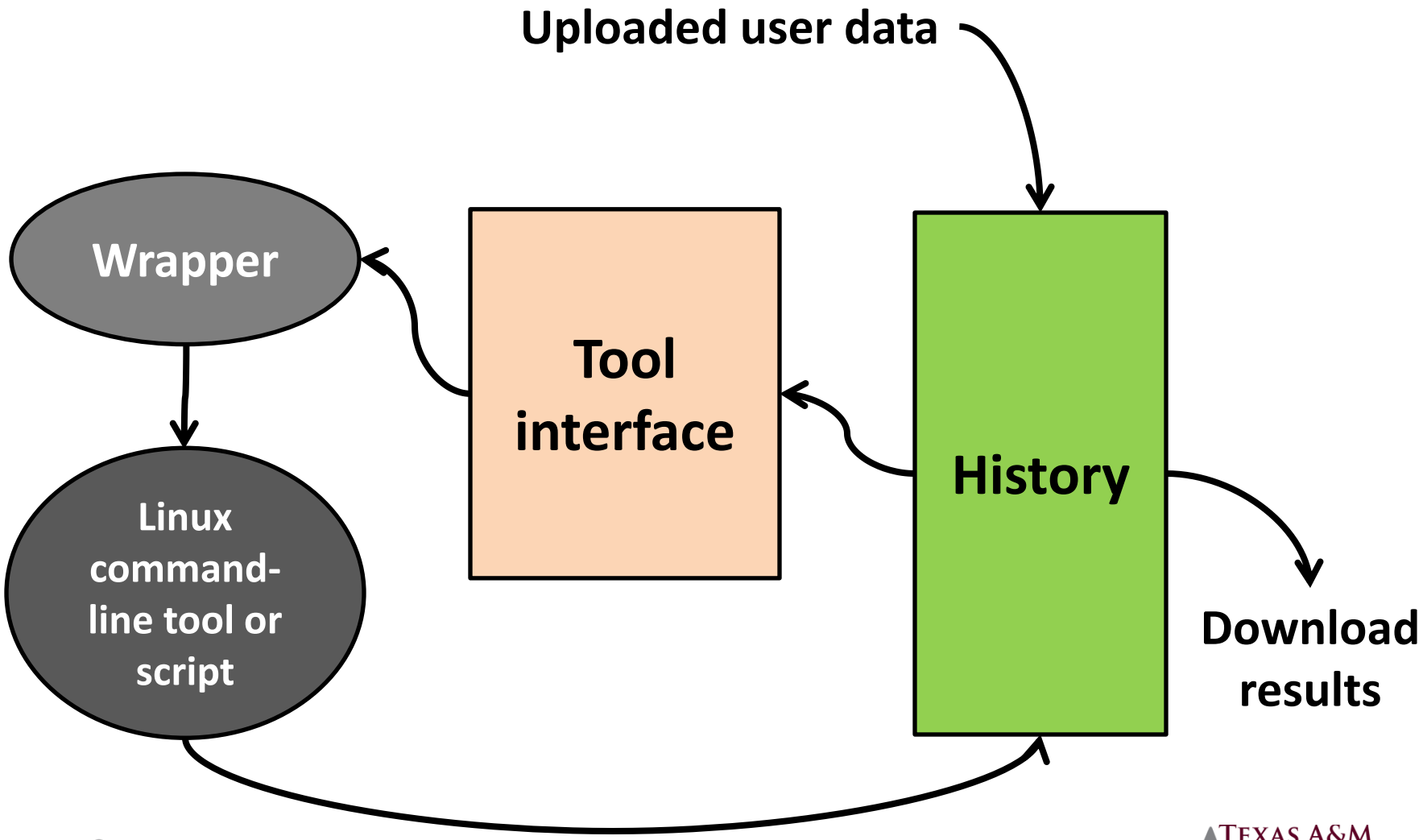
Left panel: tools

Center panel: analysis and results

Right panel: history

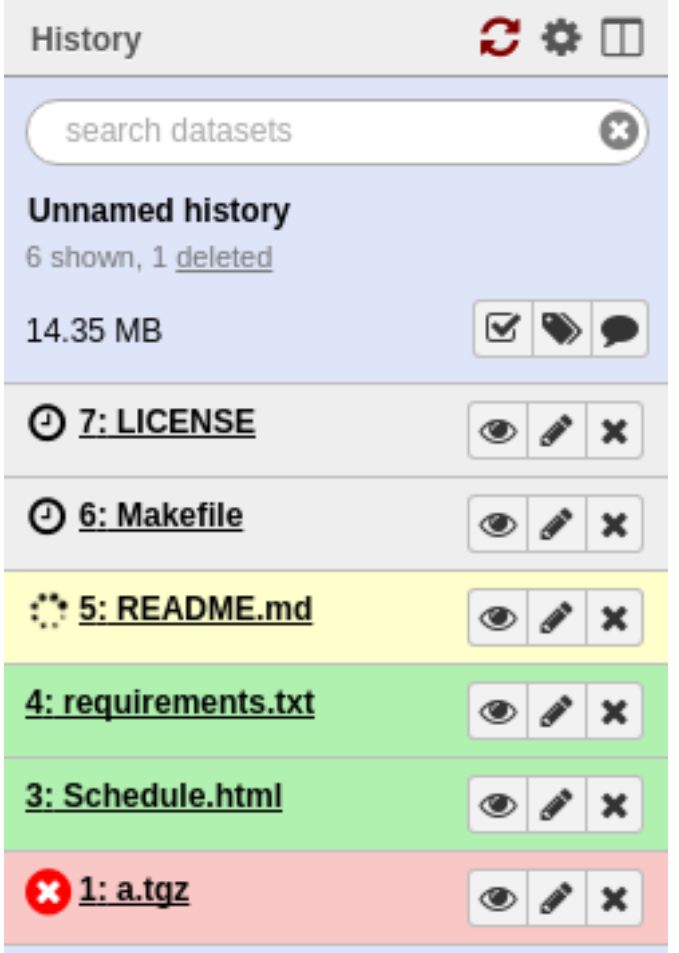


General order of operations



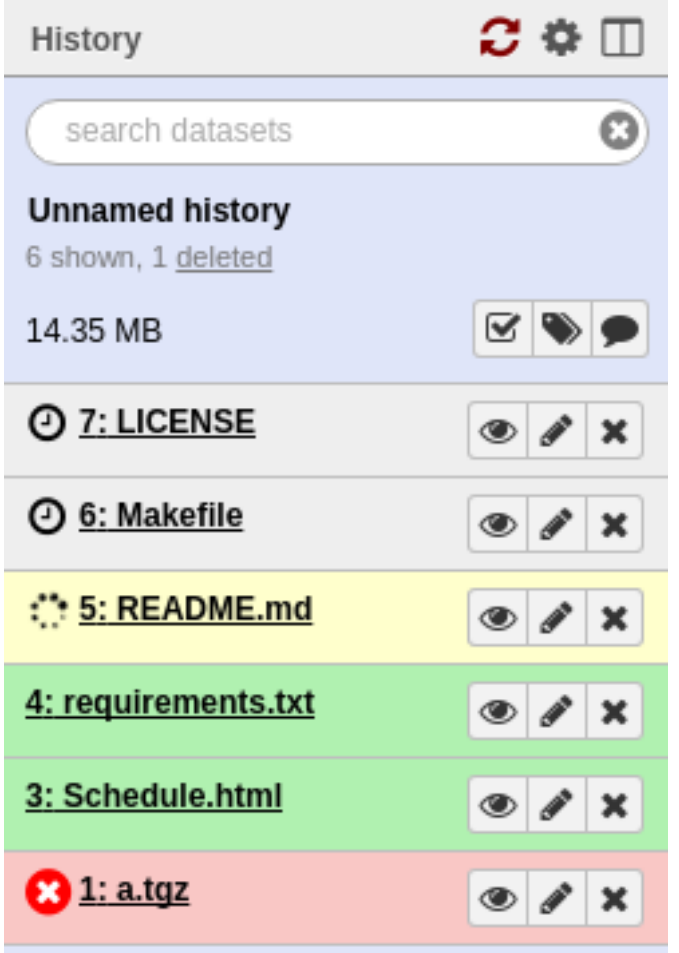
The history

- The history panel contains all input and output data for your analysis
- All input data for tools (sequences, Genbank files, etc.) must be uploaded to the history to be passed through to the tool
- All running jobs will appear in the history in the order they were entered
- All tool output data will appear in the history



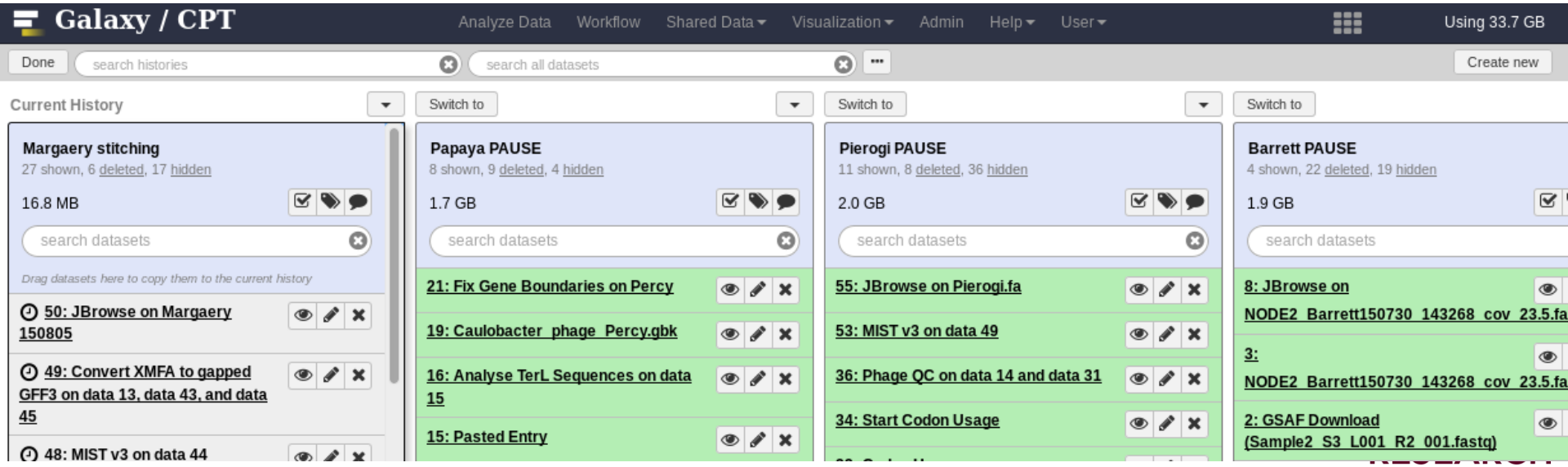
The history

- Each item in the history is a **dataset** and appears in the order it was entered
  - Each item is numbered
  - Numbers can't be changed but names can be edited by the user
- **Grey** items are queued to run
- **Yellow** items are running
- **Green** items are jobs that are completed and ready for viewing, download or input into the next tool
- **Red** items are jobs that failed or returned an error

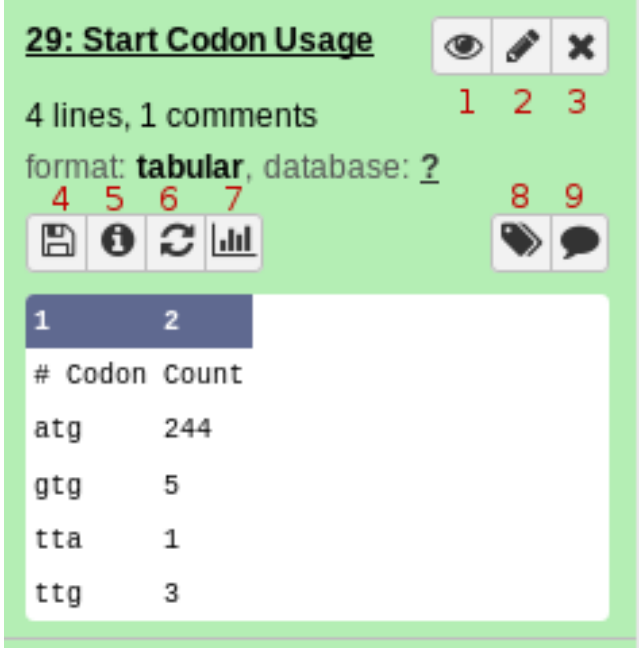


The history

- The user can create an unlimited number of new histories to keep track of related analyses
- Datasets can be copied or moved between histories
- Histories can be copied or shared between users

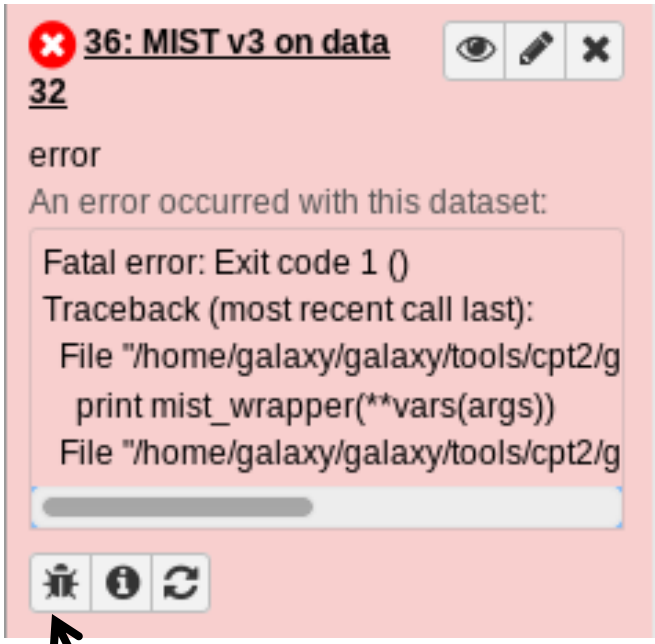


Datasets in the history



1. **Eyeball** views the dataset in the main panel
2. **Pencil** modifies metadata: name, data type, etc
3. **X** sends a dataset to the trash. You can recover deleted datasets (see below)
4. **Save** downloads the dataset to your hard-drive. You don't *need* to do this, as Galaxy will always have a copy for you
5. **Information** views details about the tool that was run and how it was configured
6. **Rerun** is a very commonly used button. This lets you re-run the tool, with the same parameters configured
   - Need to run the same tool with slightly different parameters? Don't waste time filling out the tool form; re-run it and tweak those.
   - Job failed? Try modifying the tool inputs and re-running it.
7. **Visualize** lets you visualize compatible datasets
8. **Tags** let you annotate datasets with tags
9. **Comments** let you comment on a dataset to remind yourself why you did it, or maybe to annotate some interesting results you found in the output

Failed jobs



Bug report button

- Not failures but learning opportunities!
- The CPT Galaxy is in a beta-test phase, we want to find out when/why things break
- Submitting but reports will help us improve the service

Analyzing data

NCBI BLAST+ blastp Search protein database with protein query sequence(s) (Galaxy Version 0.1.01) Options

Protein query sequence(s)

40: Phage K NO INTRONS all CDS

Subject database/sequences

Locally installed BLAST database

Protein BLAST database

NR 2017-9

Type of BLAST

- (●) blastp - Traditional BLASTP to compare a protein query to a protein database
- ( ) blastp-short - BLASTP optimized for queries shorter than 30 residues

Set expectation value cutoff

0.001

Output format

Tabular (extended 25 columns)

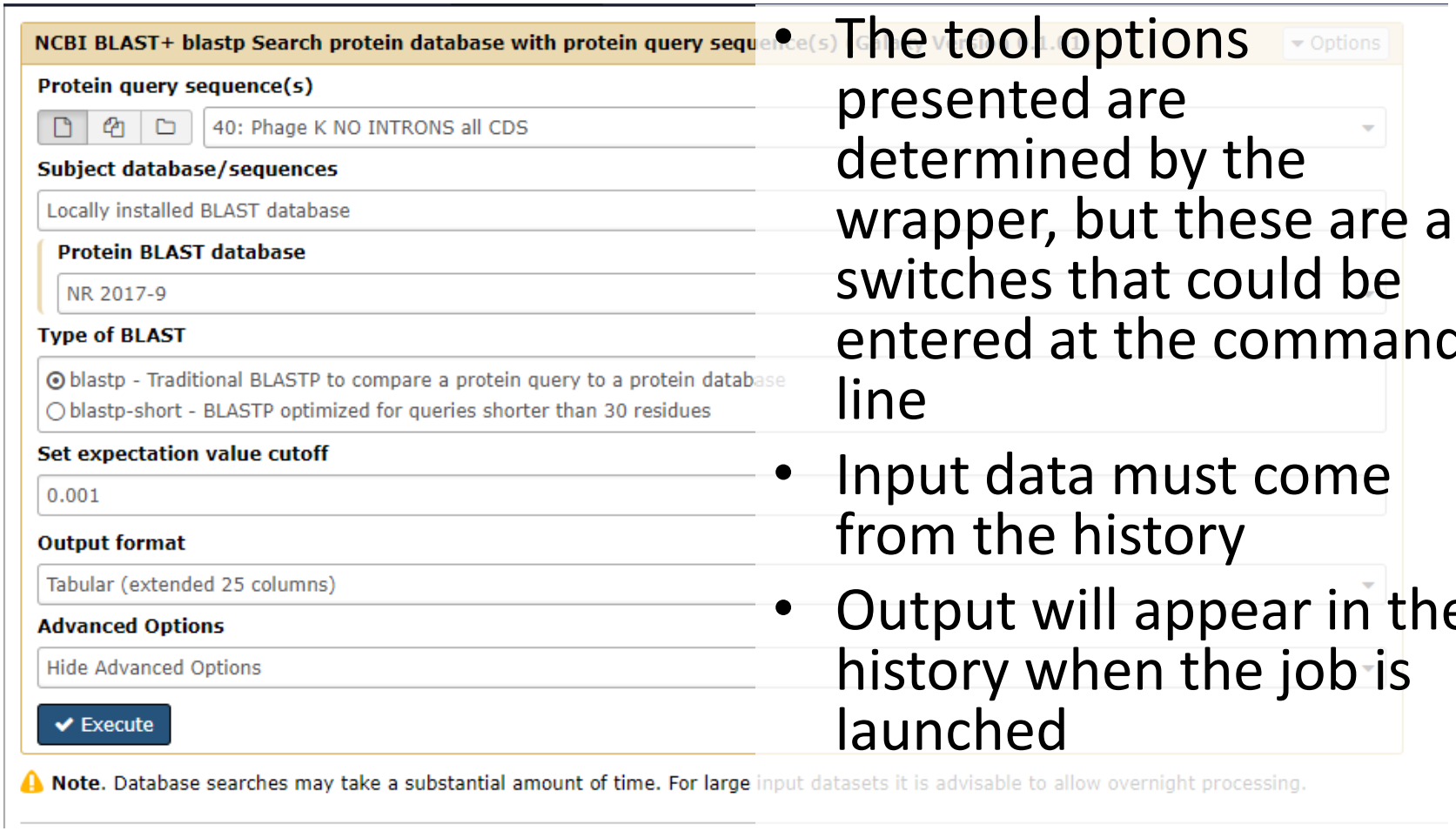
Advanced Options

Hide Advanced Options

Execute

Note. Database searches may take a substantial amount of time. For large input datasets it is advisable to allow overnight processing.

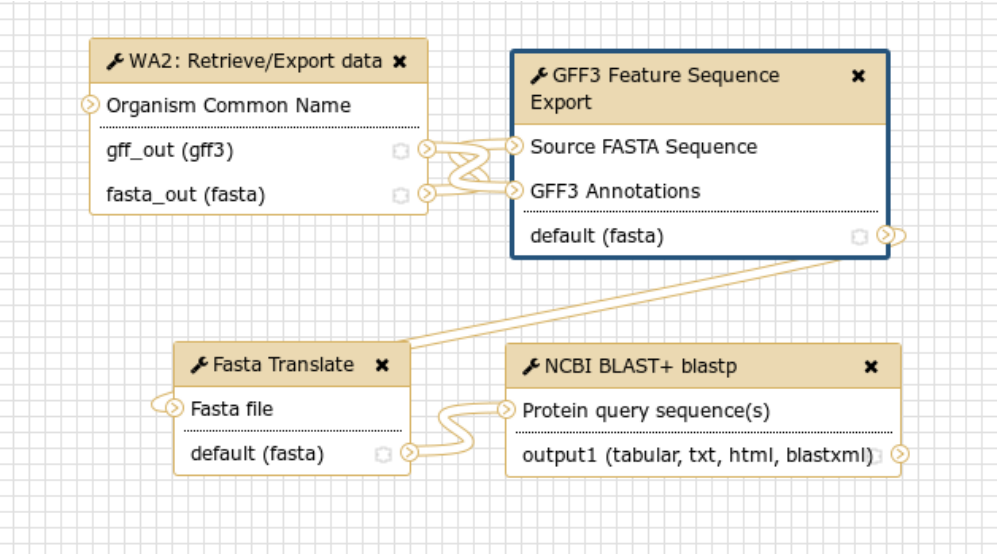
Analyzing data



- The tool options presented are determined by the wrapper, but these are all switches that could be entered at the command line
- Input data must come from the history
- Output will appear in the history when the job is launched

Workflows

- One of Galaxy's most powerful features is the ability to connect jobs in workflows
- Some analyses take a long time; output from one job will automatically be handed off to the next when it finishes



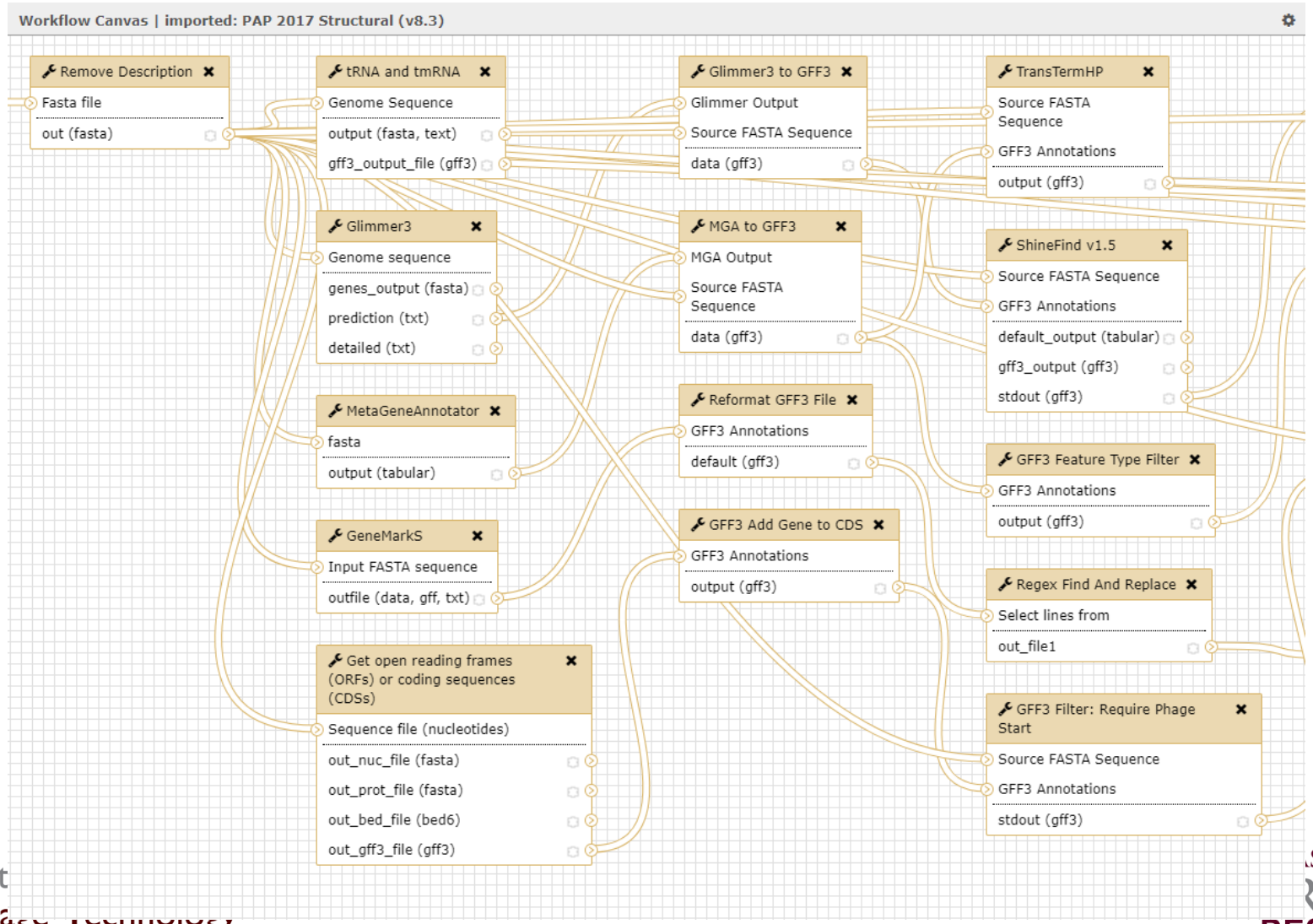
Center for Phage Technology

Texas A&M AgriLife Research

File conversion tools

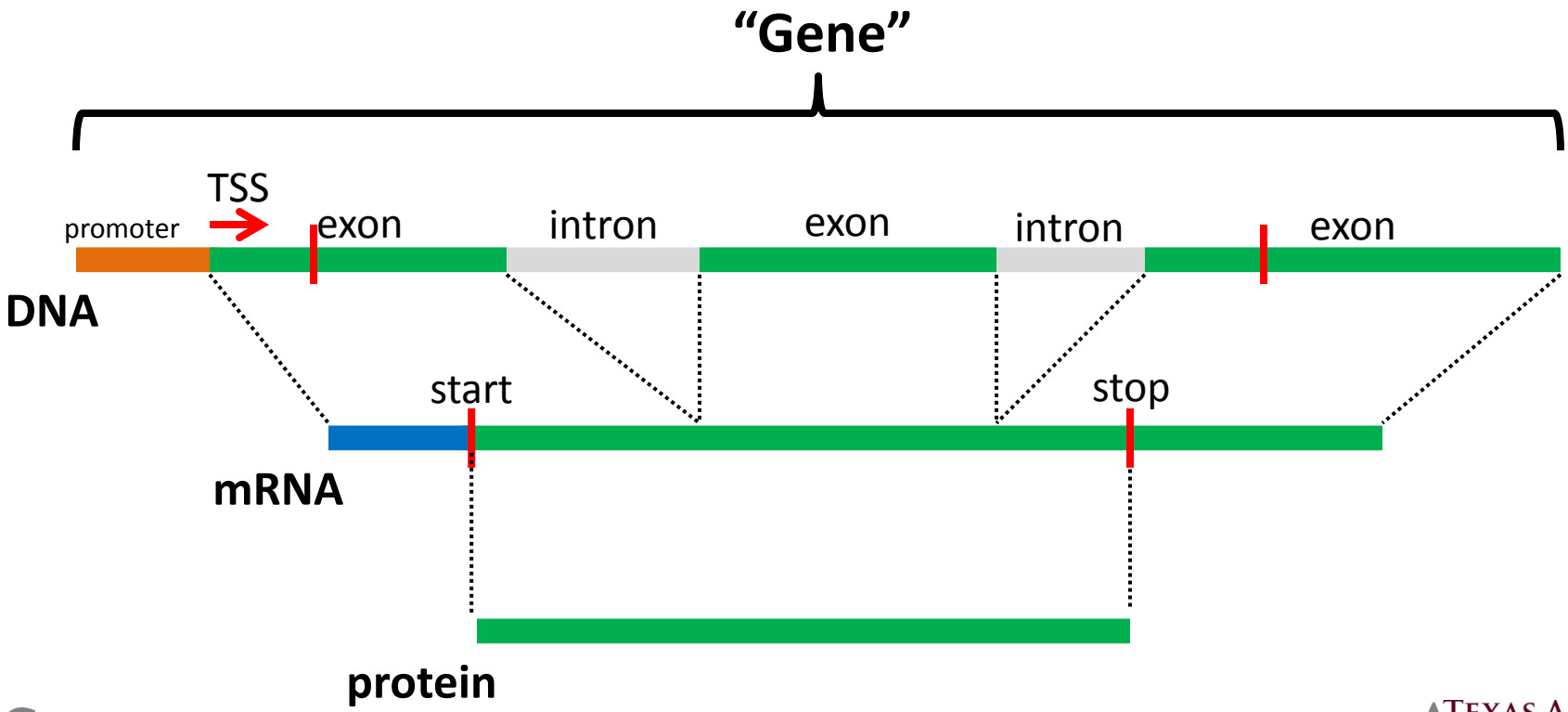
- Almost every analysis tool has its own unique output format
- An often invisible but vital part of Galaxy workflows is reformatting the output of one tool to serve as input for the next step
- Part of the workflows are simple tools that parse, reformat or extract data from outputs
- The GFF3 format is a common format and is recognized natively in Apollo

Part of the structural annotation workflow



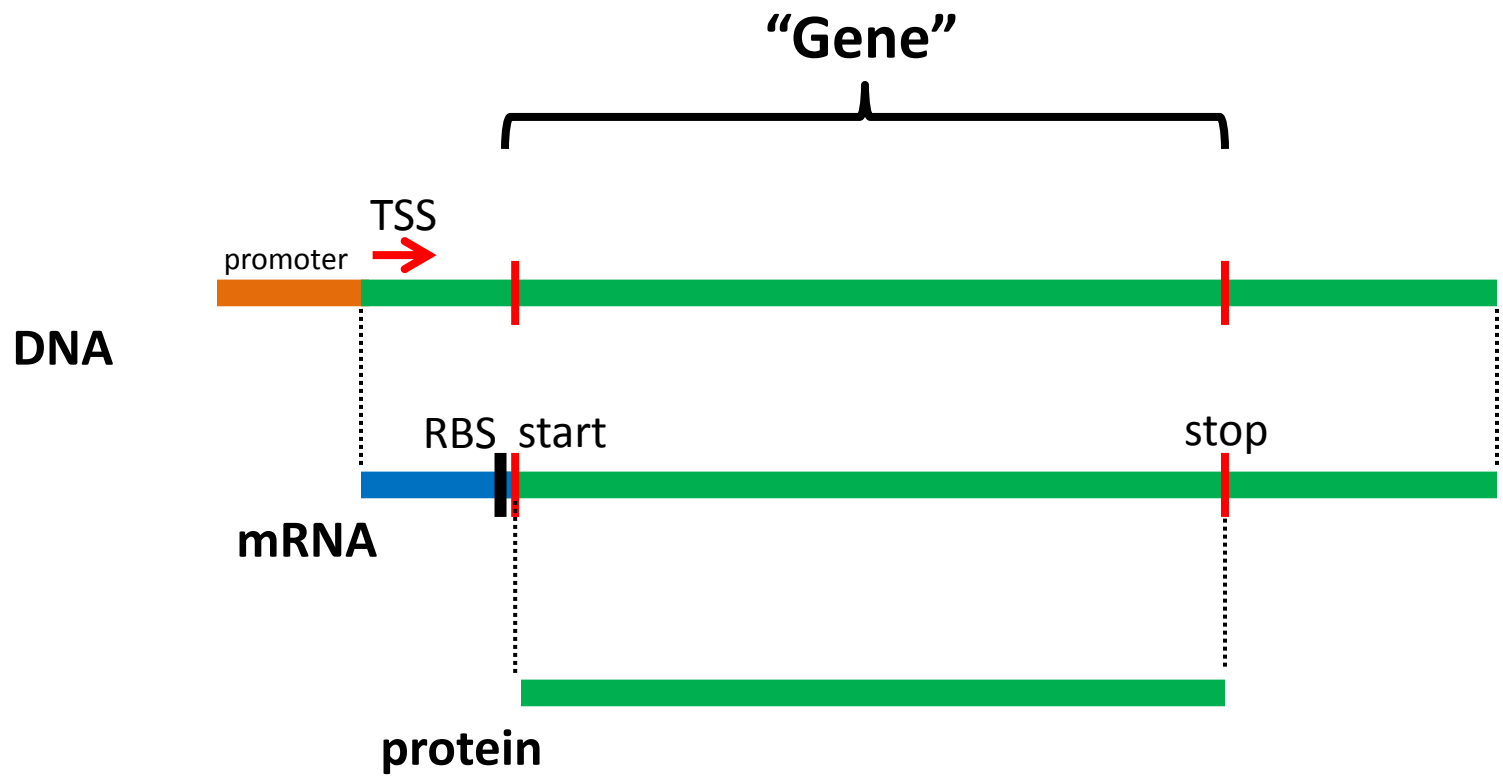
Eukaryotic gene

- Extensive mRNA processing for intron splicing, 5’ and 3’ modification
- Difficult to infer protein sequence directly from DNA sequence



Prokaryotic gene

- Introns rare, little mRNA processing
- Easy to infer protein sequence directly from DNA sequence



General Feature Format, version 3

```
##gff-version 3
ctg123 . mRNA            1300  9000  .  +  .  ID=mrna0001;Name=sonichedgehog
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mrna0001
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mrna0001
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mrna0001
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mrna0001
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mrna0001
```

- GFF3 is becoming a dominant format for storing sequence data
- One line per feature: compact, easier to search, parse, and process
- Can be used to store data other than genome annotations: BLAST alignments, conserved domains, etc.
  - **Most of the data in Apollo is stored in GFF3 format**
- The DNA sequence can be stored as part of the GFF3 file as a FASTA sequence, or can exist as a separate FASTA file

General Feature Format, version 3

```
##gff-version 3
ctg123 . mRNA          1300  9000  .  +  .  ID=mrna0001;Name=sonichedgehog
ctg123 . exon          1300  1500  .  +  .  ID=exon00001;Parent=mrna0001
ctg123 . exon          1050  1500  .  +  .  ID=exon00002;Parent=mrna0001
ctg123 . exon          3000  3902  .  +  .  ID=exon00003;Parent=mrna0001
ctg123 . exon          5000  5500  .  +  .  ID=exon00004;Parent=mrna0001
ctg123 . exon          7000  9000  .  +  .  ID=exon00005;Parent=mrna0001
```

type: Type of feature (gene, exon, CDS, etc.)

- Feature types are hierarchical with parent/child relationships
- mRNA > gene > exon = CDS

source: Name of the program that generated the feature

seqid: Name of the chromosome or contig the annotation refers to. This is usually a separate FASTA sequence file.

General Feature Format, version 3

```
##gff-version 3
ctg123 . mRNA            1300  9000  .  +  .  ID=mrna0001;Name=sonichedgehog
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mrna0001
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mrna0001
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mrna0001
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mrna0001
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mrna0001
```

start, end : Coordinates of the start and end of the feature, as base position of the sequence specified by **seqid**

score: Feature score (e.g., E-value, P value)

strand: Plus (+) or minus (-) DNA strand

phase: where the feature begins relative to the reading frame; 0, 1, or 2 base offset. CDS features must have a phase.

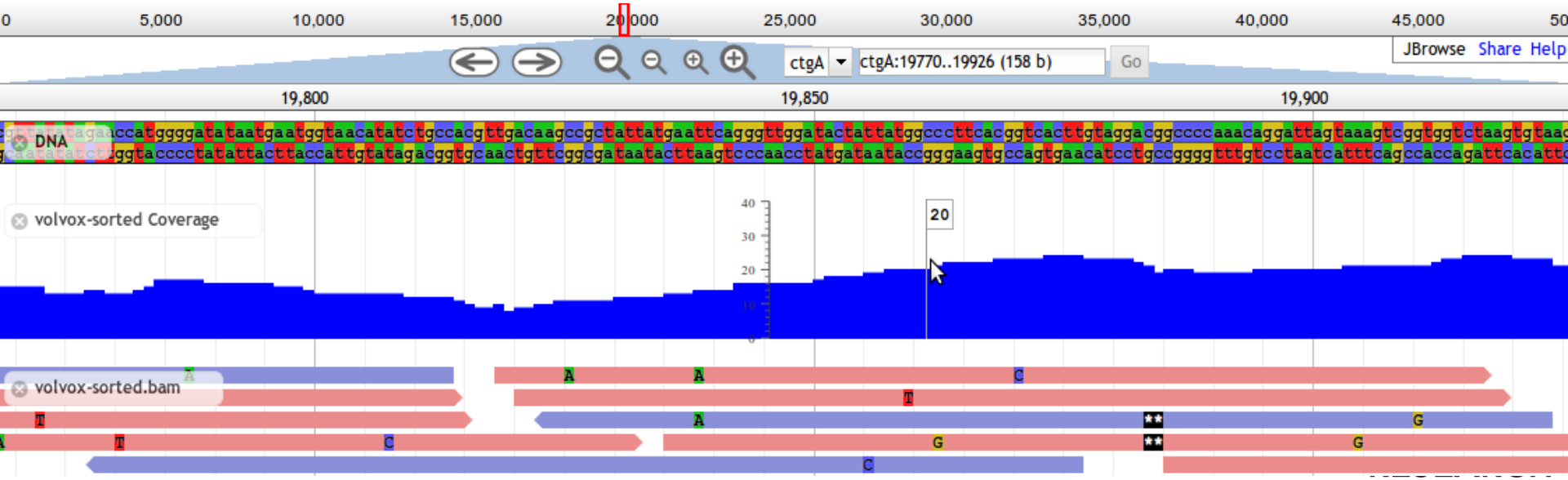
attributes: Feature attributes listed as tag=value

What is WebApollo?

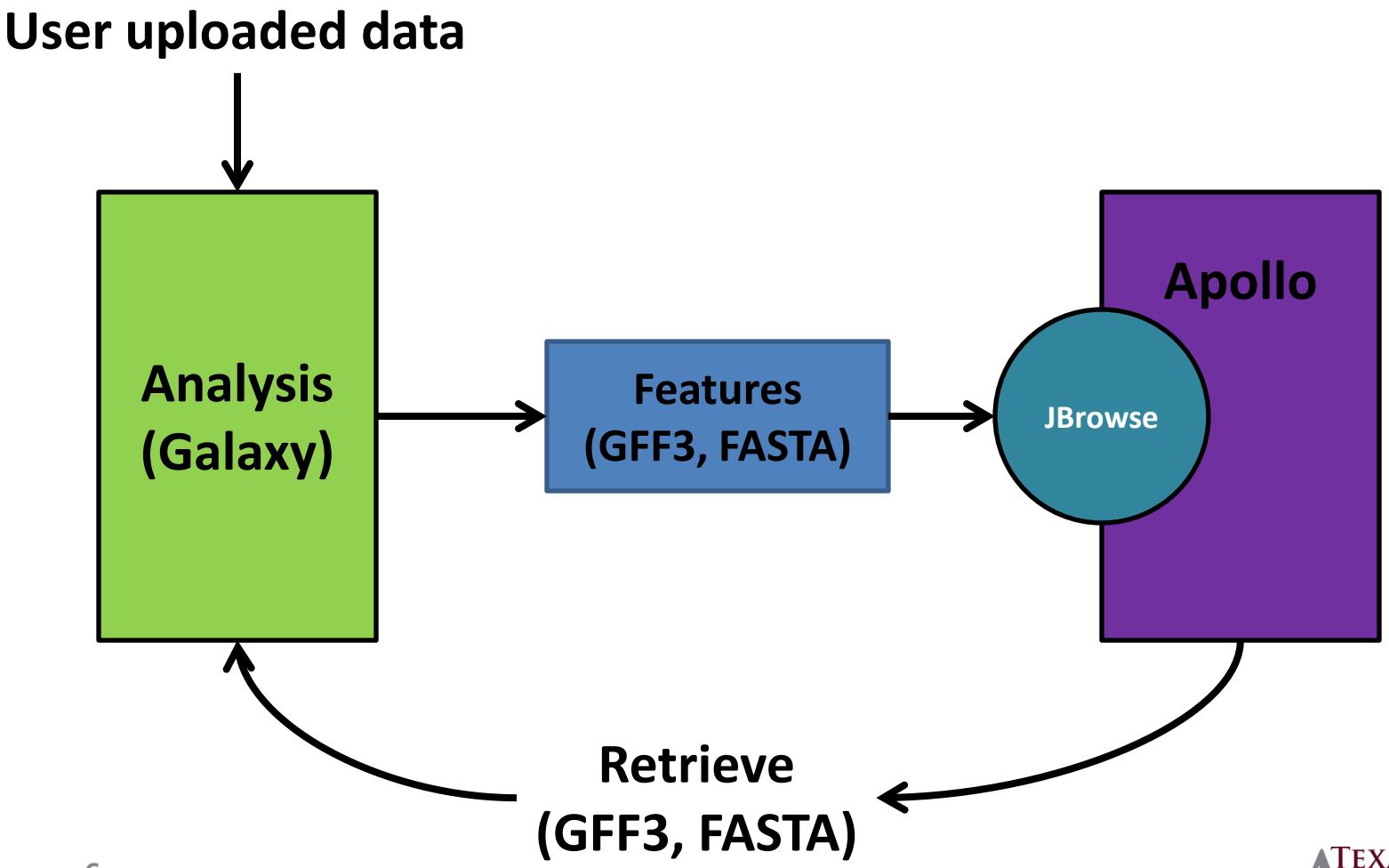
- WebApollo is an interactive genome visualizer that supports collaborative genome annotation
- "Google Docs, but for genomes"
- Still in development, Apollo is less robust than Galaxy and new features continue to be added
- Maintains genome annotations and multiple evidence "tracks" to guide annotations
- The CPT has developed tools that bridge Galaxy ←→ Apollo

JBrowse

- Jbrowse is a genome viewer implemented in many online tools (including RAST and PATRIC)
- Viewer only, no editing function
- Apollo is an addition to JBrowse
  - To work with data in Apollo, it is first used to generate a JBrowse instance that is then loaded into Apollo



Galaxy/Apollo order of operations

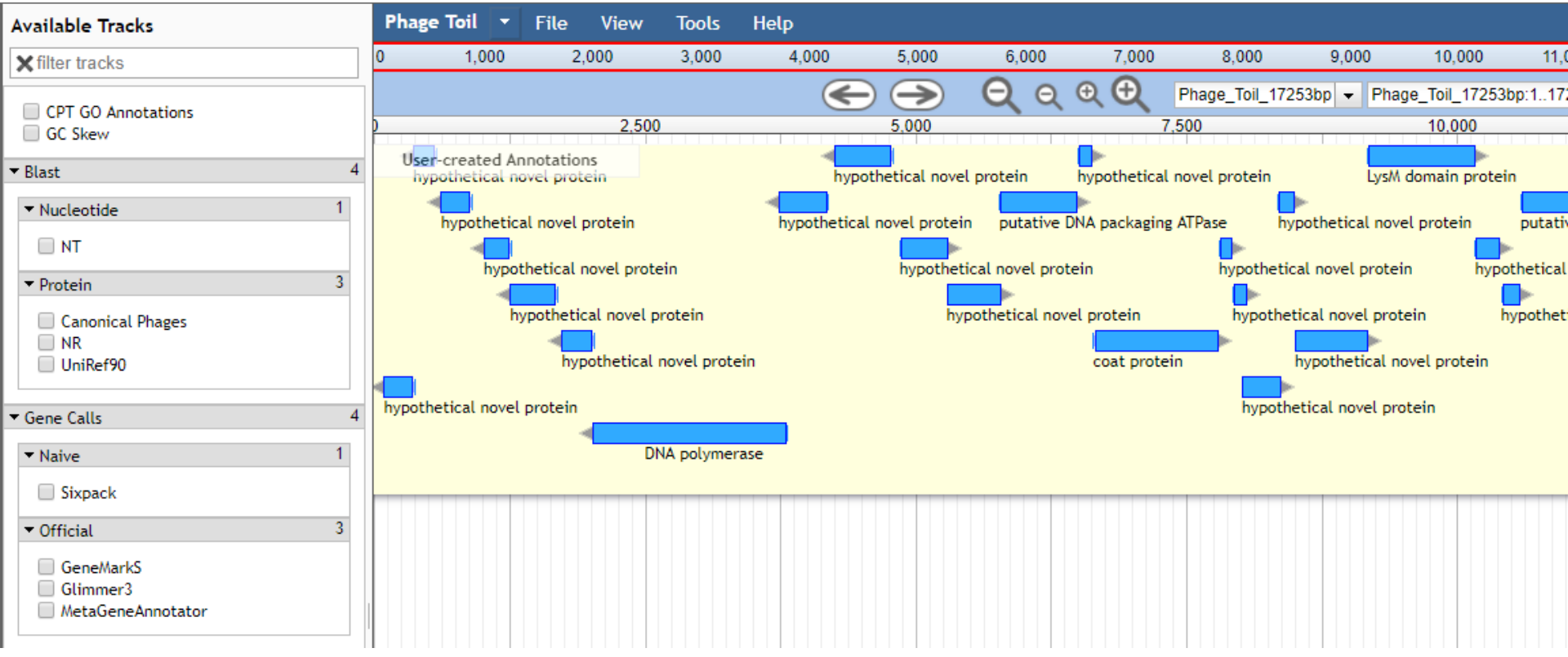


Center for Phage Technology

Texas A&M AgriLife Research

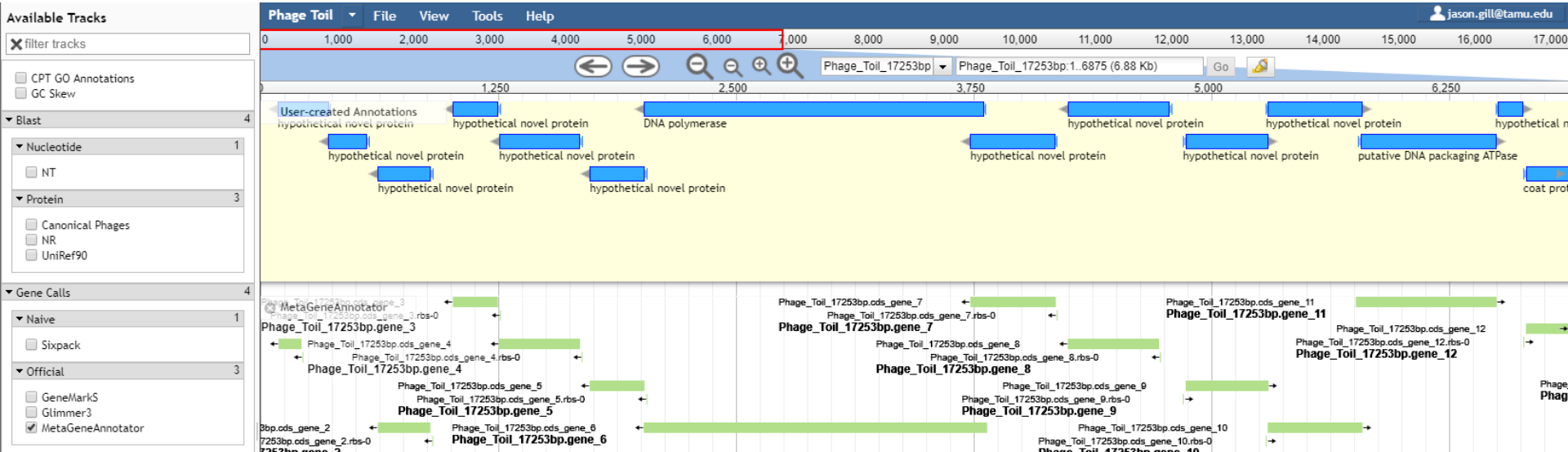
Annotations in Apollo

- User annotations appear in the topmost track of the display
- Tracks below this are generated by various tools in Galaxy and imported to Apollo



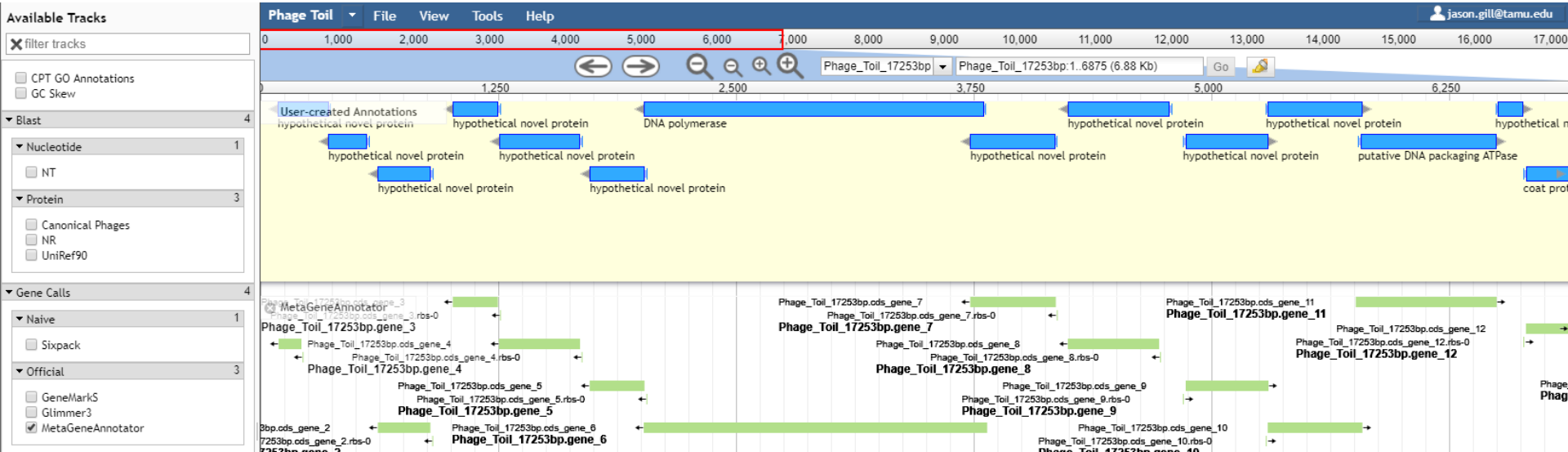
Evidence tracks in Apollo

- Features in Apollo can **only** be created from evidence tracks
- Unlike purely manual editors like Artemis, the user cannot select sequence and create a feature *de novo*
- This is part of a philosophical decision by Apollo, that features are **only created with evidence**



Structural annotation

- The output from gene-calling programs appears in the evidence track
  - Output is first run through a program that adds RBS sequences to each gene
  - Genes are manually called and moved into the annotation track
- Gene callers sometimes miss genes, so there is a track generated by a naive ORF caller (modified Sixpack) to call all possible genes
- tRNAScan/ARAGORN and TransTermHP also run at this step



Phage Toil File View Tools Help

0 1,000 2,000 3,000 4,000 5,000 6,000 7,000 8,000 9,000 10,000 11,000 12,00

Phage_Toil_17253bp Phage_Toil_17253bp:1..6875 (6.88 Kb)

1,250 2,500 3,750

User-created Annotations

hypothetical novel protein

hypothetical novel protein

DNA polymerase

hypothetical novel pr

hypothetical novel protein

hypothetical novel protein

hypothetical novel protein

hypothetical novel protein

hypothetical novel protein

MetaGeneAnnotator

Phage_Toil_17253bp.cds_gene_4

Phage_Toil_17253bp.cds_gene_4.rbs-0

Phage_Toil_17253bp.gene_4

Phage_Toil_17253bp.cds_gene_7

Phage_Toil_17253bp.cds_gene_7.rbs-0

Phage_Toil_17253bp.gene_7

3bp.cds_gene_2

7253bp.cds_gene_2.rbs-0

7253bp.gene_2

Phage_Toil_17253bp.cds_gene_6

Phage_Toil_17253bp.gene_6

View details

Zoom to this gene

Highlight this gene

Create new annotation

gene

pseudogene

tRNA

snRNA

snoRNA

ncRNA

rRNA

miRNA

repeat_region

terminator

transposable element

Phage_Toil_17253bp.cds_gene_3

Phage_Toil_17253bp.cds_gene_3.rbs-0

Phage_Toil_17253bp.gene_3

cds_gene_8.rbs-0

8

Phage_Toil_17253bp.cds_gene_5

Phage_Toil_17253bp.cds_gene_5.rbs-0

Phage_Toil_17253bp.gene_5

Phage_Toil

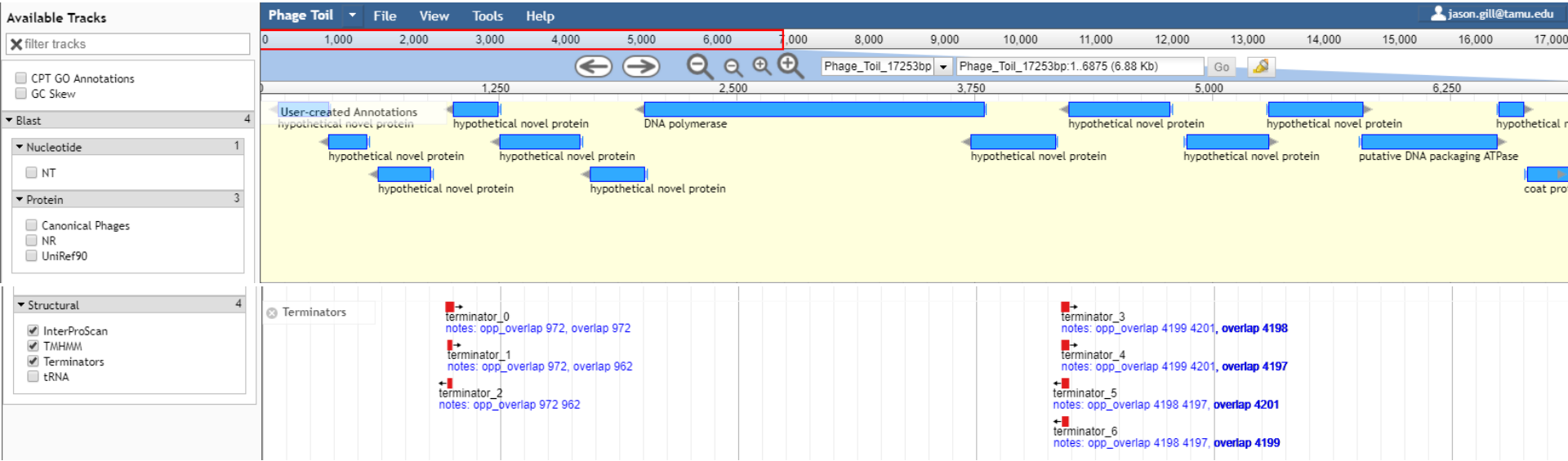
.gene

Center for Phage Technology

AGRILIFE RESEARCH

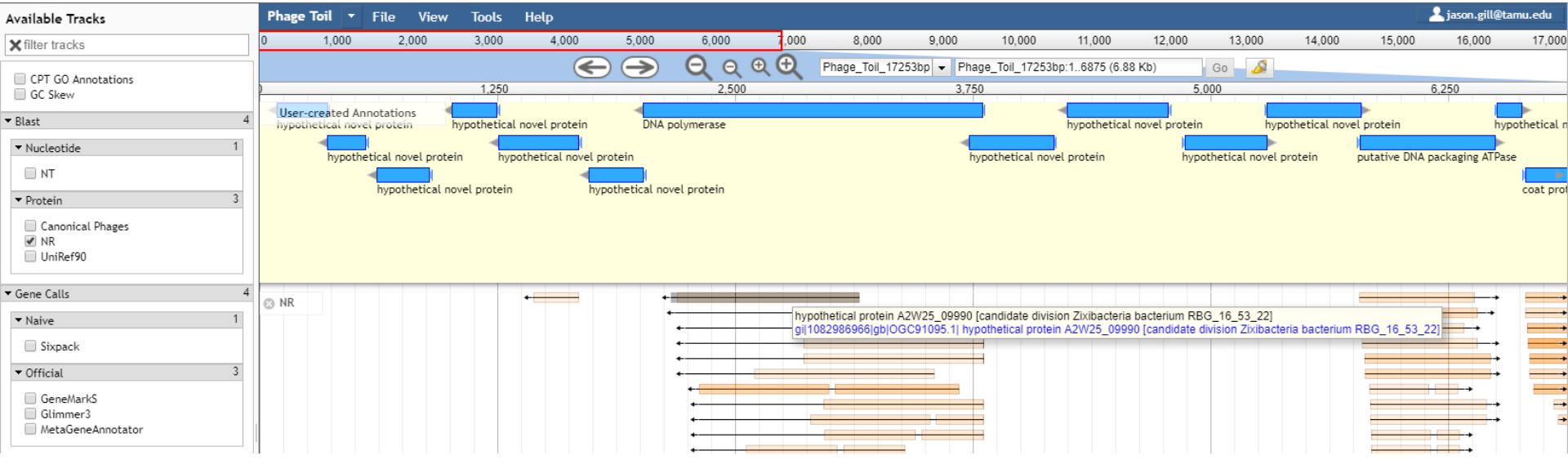
Structural annotation: terminators

- TransTermHP detects potential rho-independent terminator but is noisy
- Many false-positive results
- Terminators must be called after genes to put them in proper context



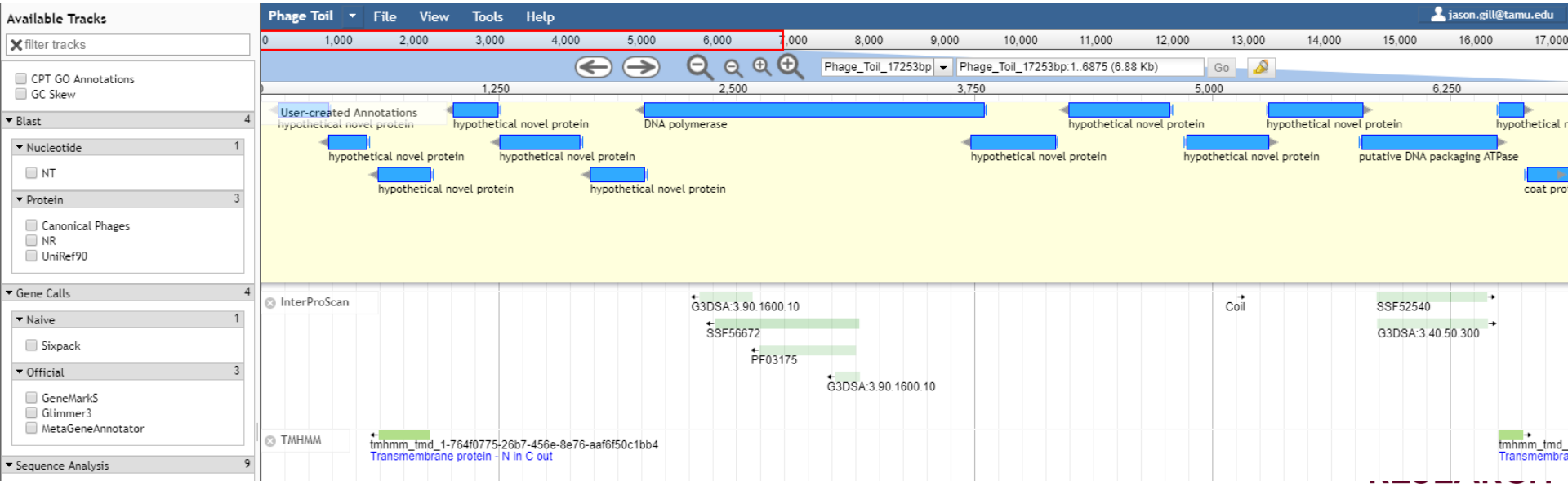
Functional annotation: BLASTp

- All CDS's called in the structural annotation phase are exported to Galaxy, searched in BLASTp and the results converted to GFF3 format for display
- Hits, E-values and alignments are available in Apollo
- BLAST against other more useful databases also possible

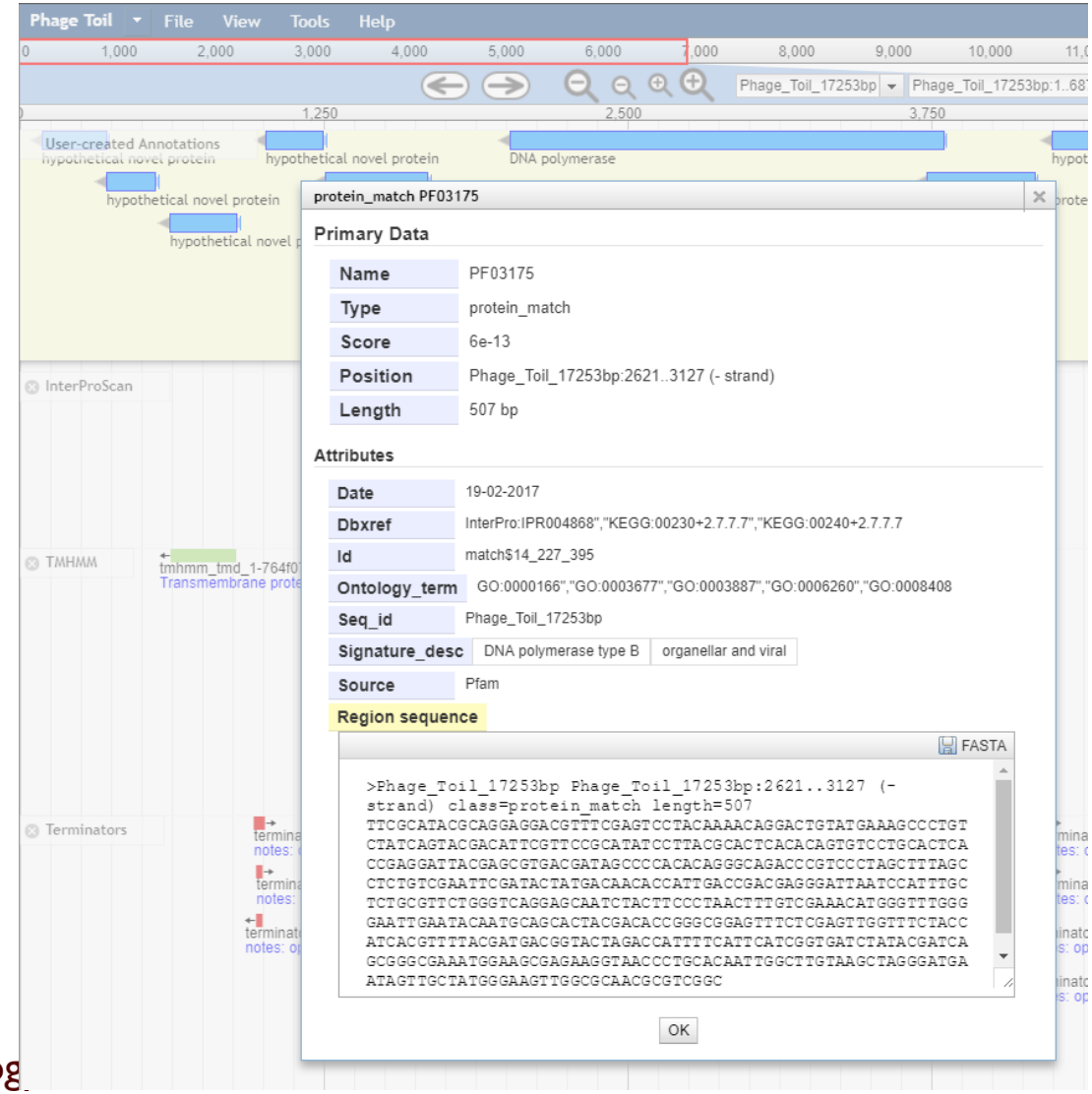


Functional annotation: conserved domains

- All proteins are run through InterProScan at the same time as BLASTp
- Outputs are converted to GFF3 for display
- Conserved domains can give evidence for gene function that simple homology (BLAST) will not



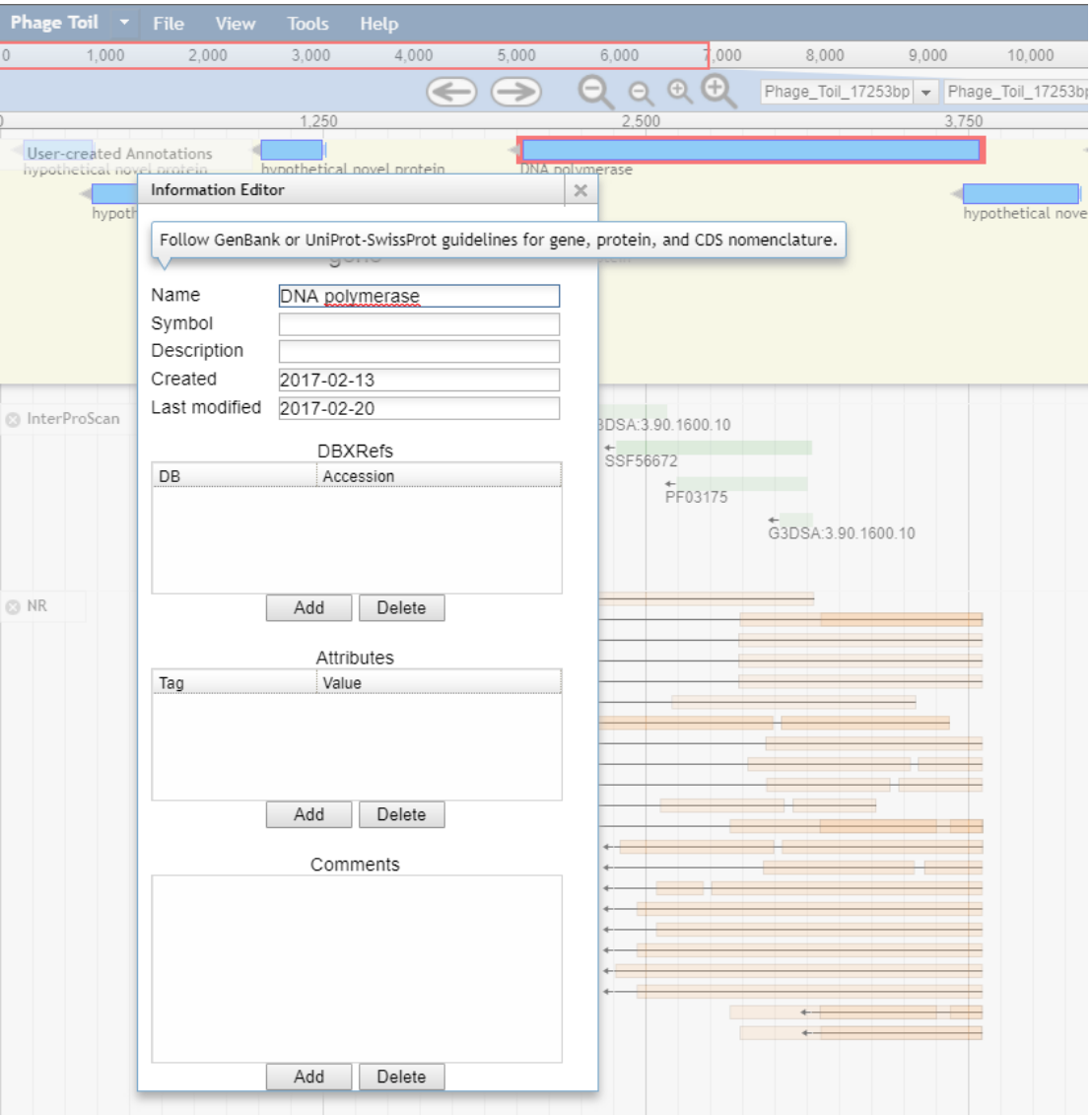
Conserved domain hit information



Center for Phage Technolog

Texas A&M AgriLife Research

Adding annotations



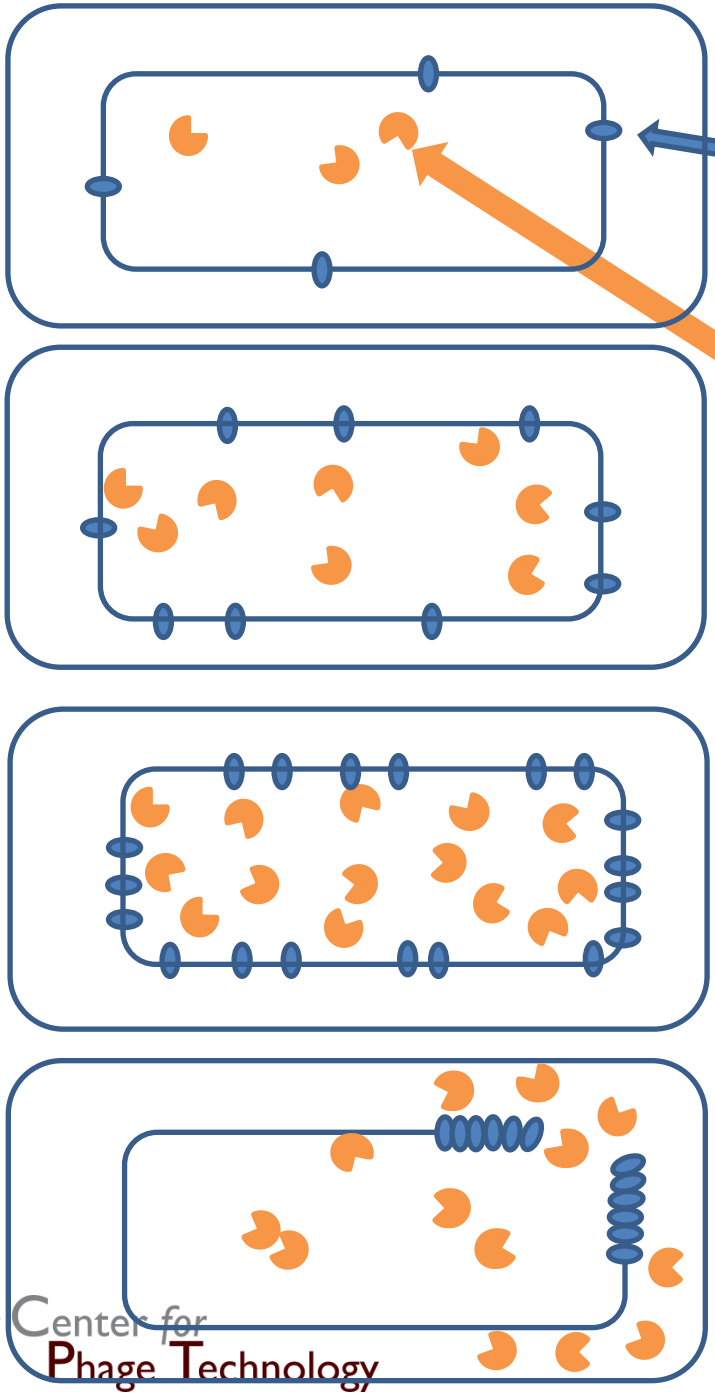
- Functional annotations are typed in by the user for each gene based on evidence
- Hoping to automate this step in the future

Texas A&M AgriLife Research

Solutions for “special” phage cases

- Many *Caudovirales* phages contain “special case” genes that are not often detected by annotation tools or automated workflows
  - Lysis genes
  - Introns
  - Tapemeasure frameshift chaperones

Canonical holin /endolysin lysis



Holin accumulates in IM, mobile and harmless

Endolysin accumulates in cytoplasm, fully active

At a the programmed time, holin ***triggers*** to form massive **"holes"** (average 350 nm for lambda)

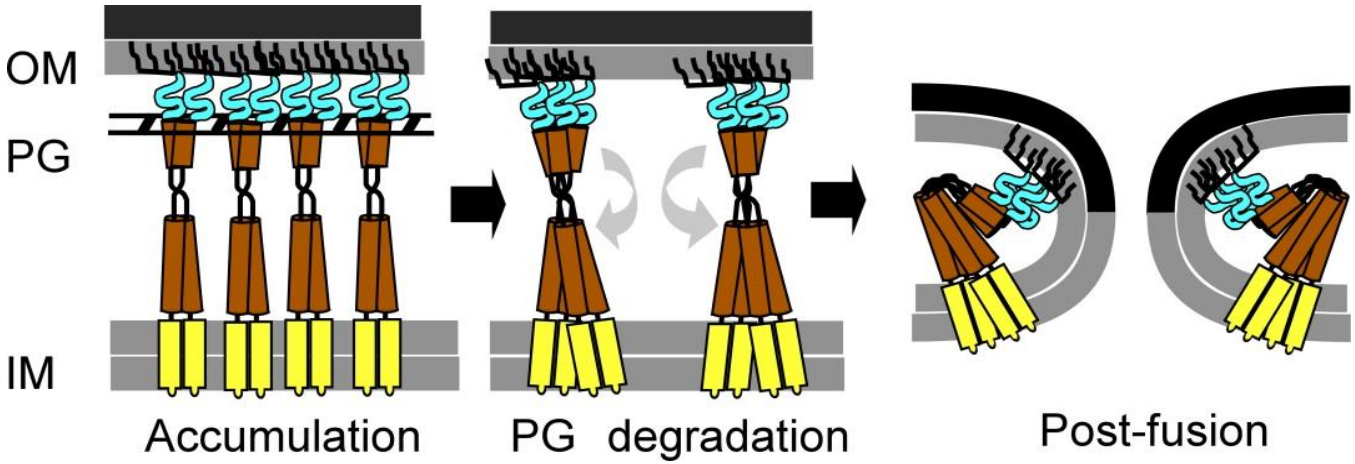
Endolysin escapes through holes & attacks peptidoglycan

Center for Phage Technology

TEXAS A&M AGRILIFE RESEARCH

Spanin complex

- After holin triggering, endolysin is released to degrade peptidoglycan
- In Gram –ve hosts, a third component, the spanin complex, disrupts the outer membrane
- The canonical spanin is 2 components: an **inner membrane** protein with an N-terminal TMD, and an **outer membrane** lipoprotein tethered to the inner leaflet of the OM by a lipid anchor

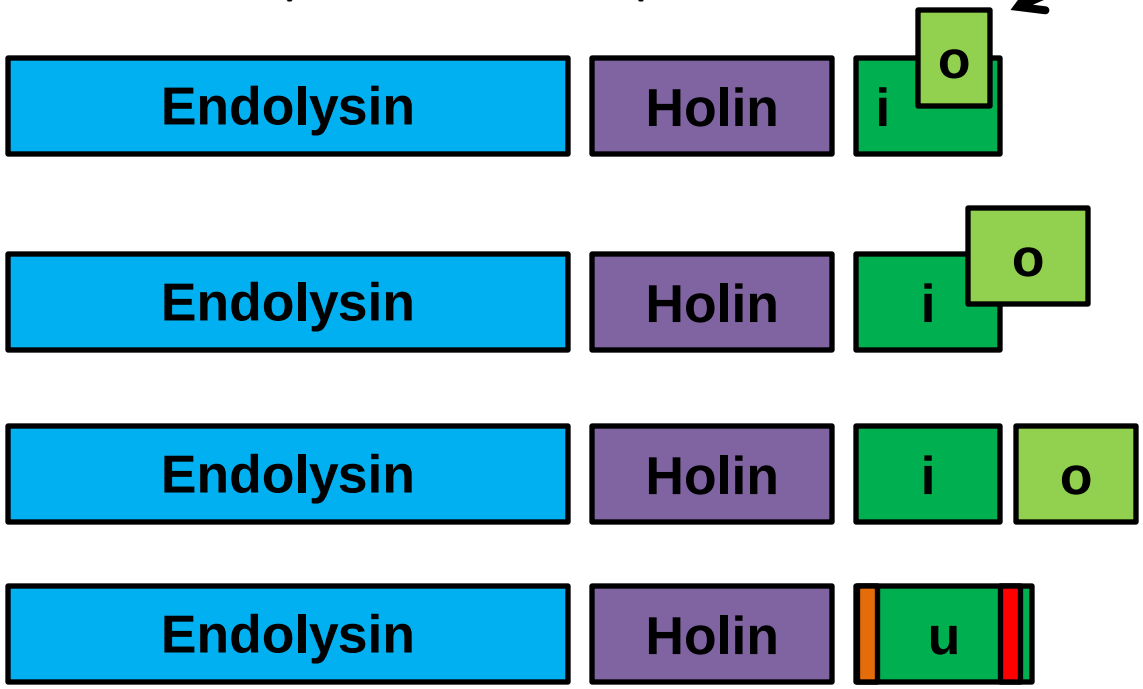


Rajaure et al. 2015, PNAS

Lysis genes

- Lysis genes are often found co-localized in **lysis cassettes**
  - **Endolysin:** conserved domains or BLAST homology
  - **Holins:** small, 1 or more TMD's
  - **Spanins:** Adjacent, partially or fully embedded
    - **i-spanin: 1 N-terminal TMD**
    - **o-spanin: 1 N-terminal lipobox**

Embedded genes not found in normal structural annotation!



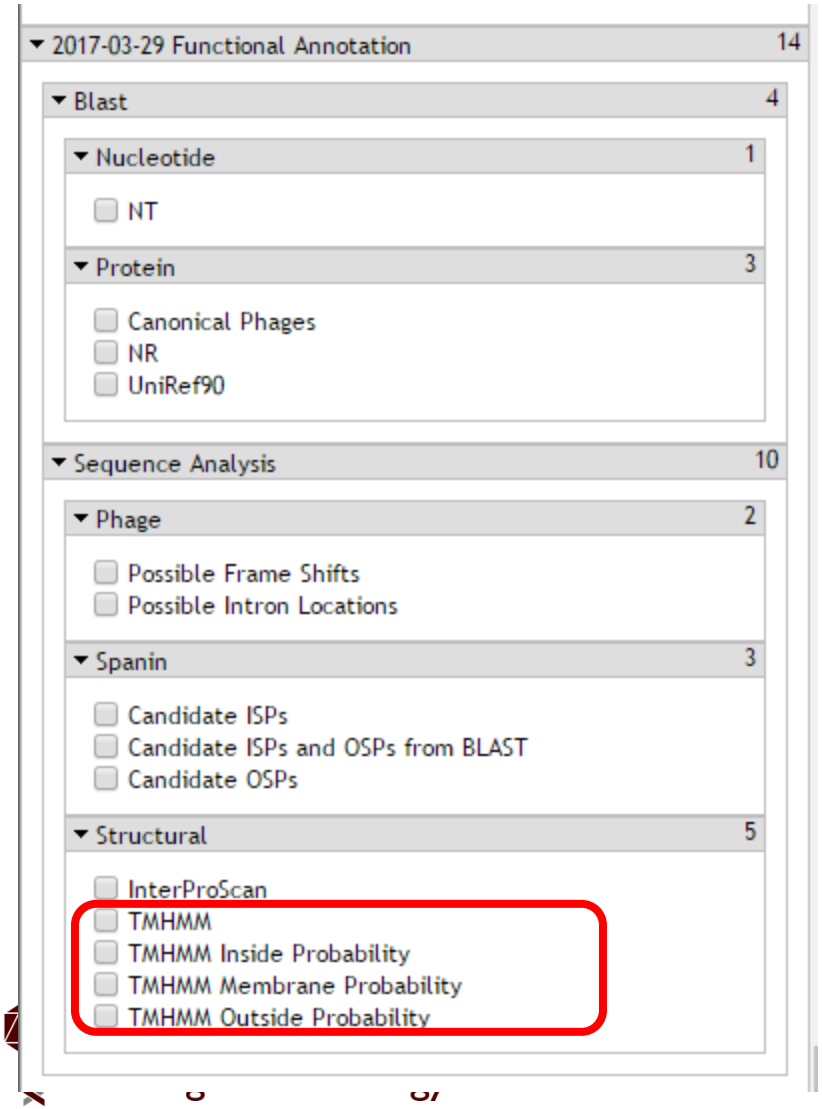
Fully embedded

Partial embedded

Separate

Unimolecular
N-terminal lipobox, C-terminal TMD

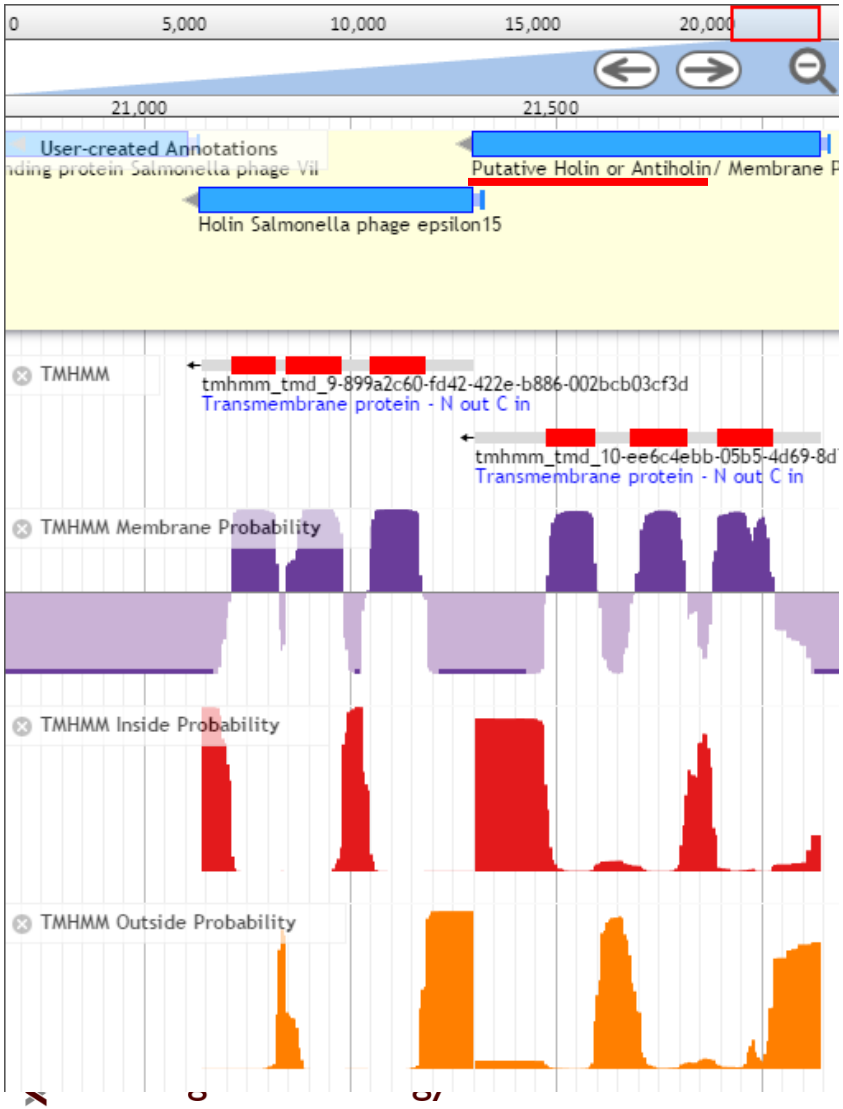
Holin finding



- Small, TMD-containing proteins
- Look for holins first next to the endolysin gene
- Must have 1 or more TMDs
- May have homology to another holin by BLAST but unlikely

TEXAS A&M AGRILIFE RESEARCH

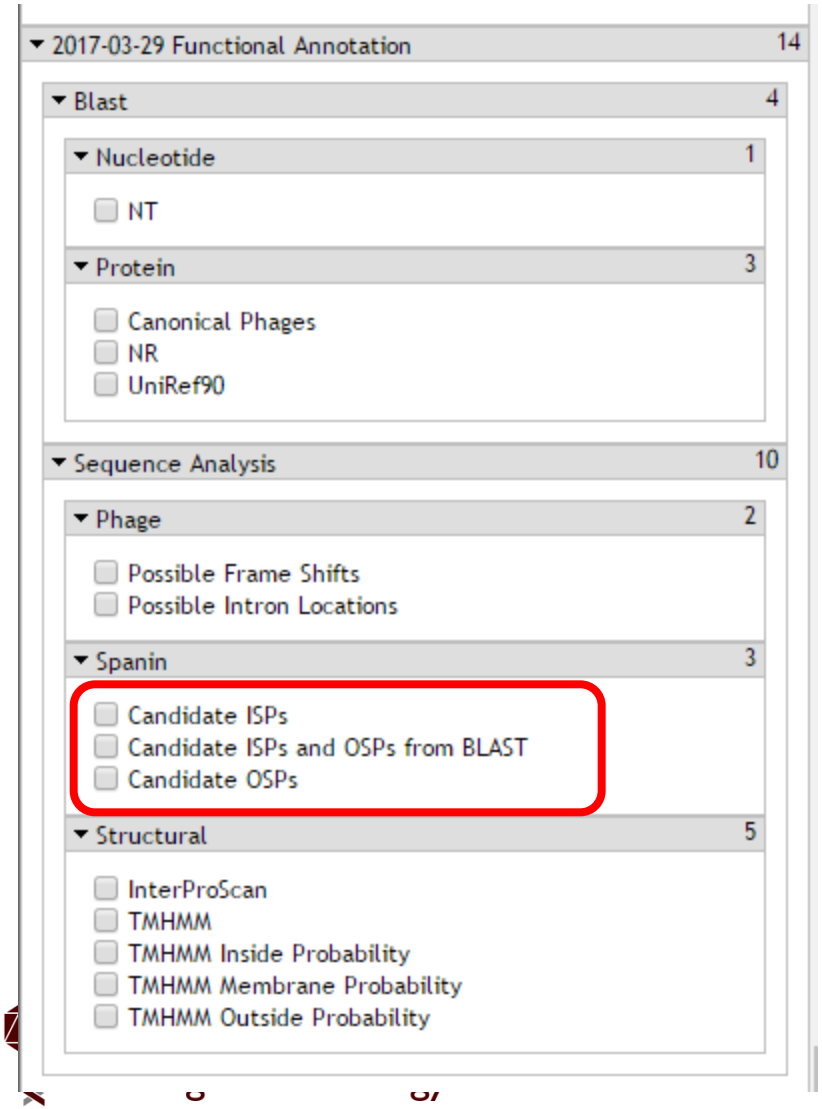
Holin finding



- **New display tracks for TMHMM**
  - TMHMM: number and location of predicted TMD's
  - Membrane: actual scores from TMHMM plotted to genome
  - Inside: Probability this region is in the cytoplasm
  - Outside: Probability this region is in the periplasm/extracellular
- **Most likely** location is adjacent to the endolysin
  - If no holin candidate near endolysin, lysis genes may be distributed
  - Probably will not be identifiable unless there is *only one* small TMD-containing protein in the whole genome (unlikely), or BLAST homology to a known holin (also unlikely)

TEXAS A&M AGRILIFE RESEARCH

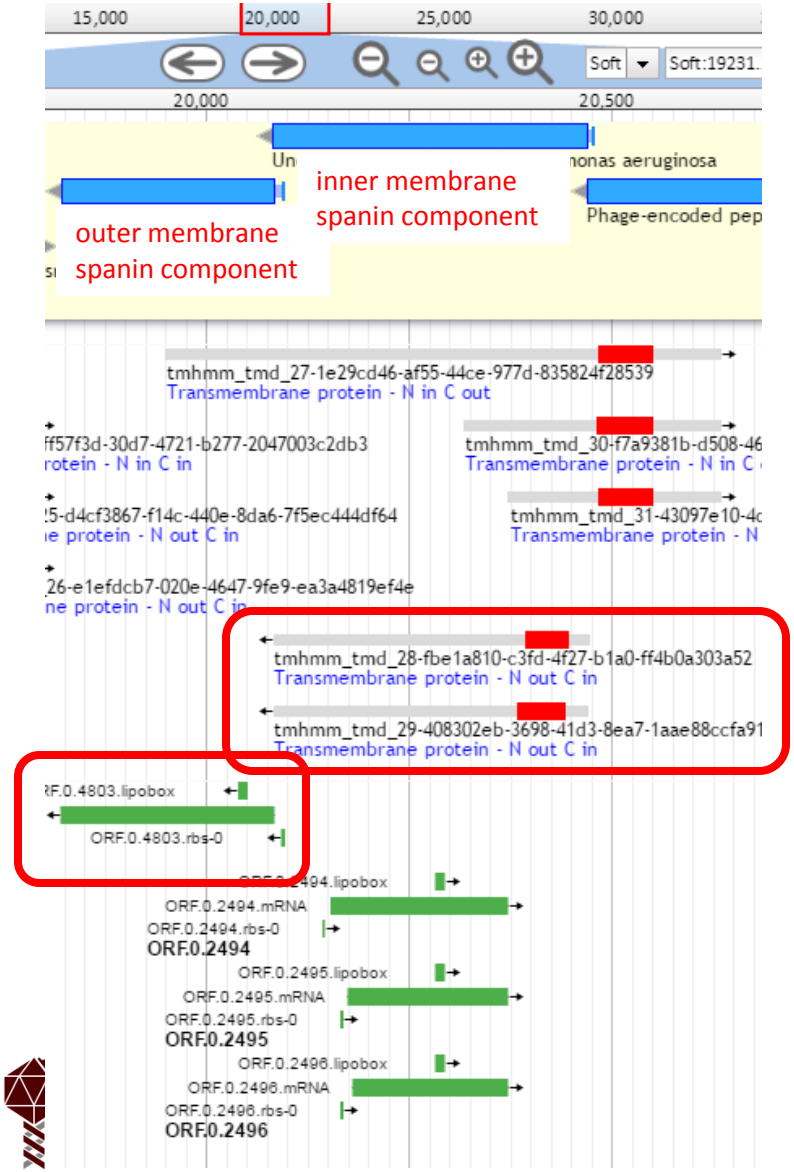
Spanin finding



- Candidates from BLAST
  - Low sequence conservation in spanins
- Candidate ISPs (i-spanin)
  - Naive ORF calls analyzed by TMHMM
- Candidate OSPs (o-spanin)
  - Naive ORF calls analyzed for N-terminal lipobox signals

Texas A&M AgriLife Research

Spanin finding



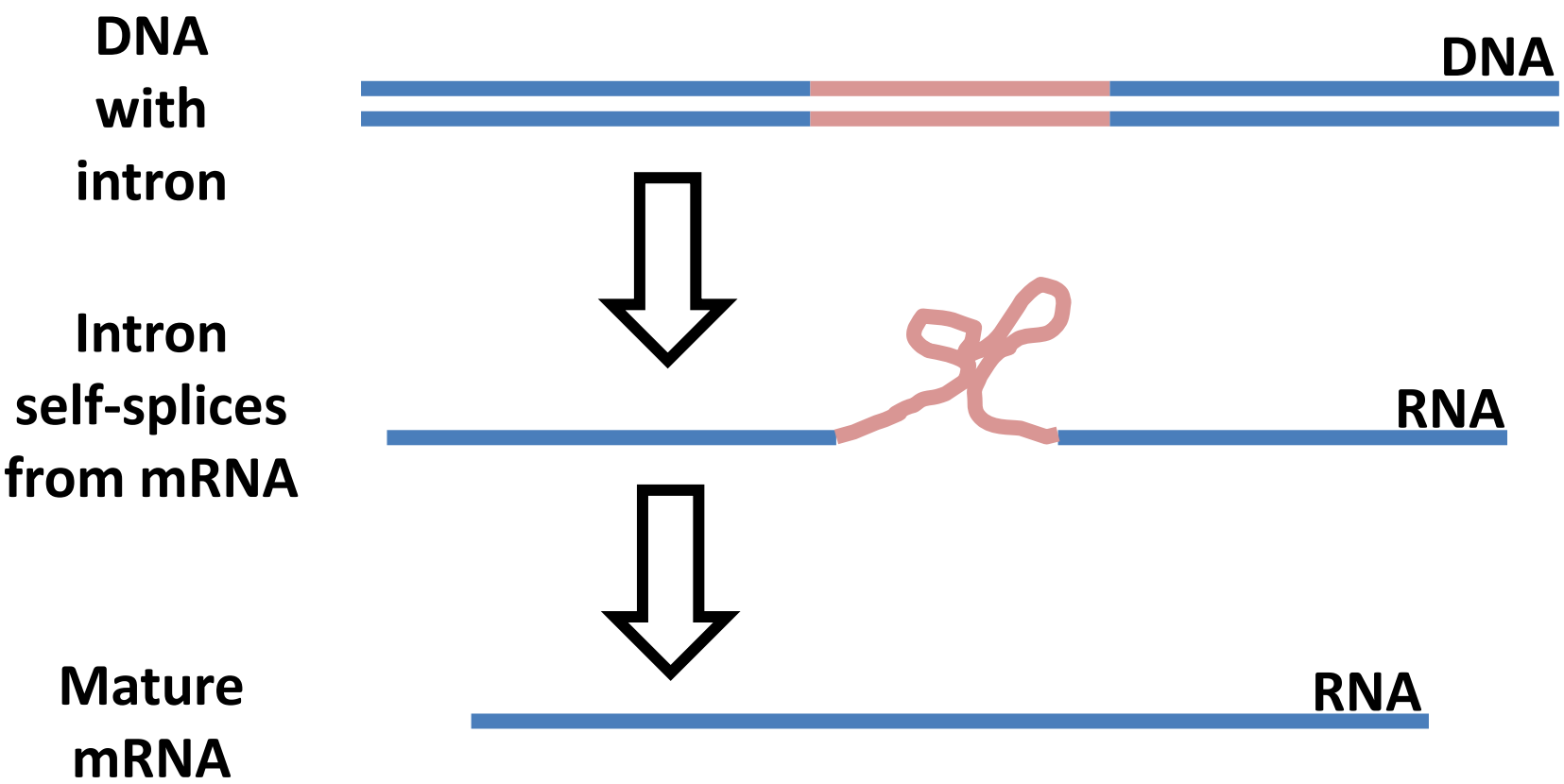
- Likely spanin gene pairs
  - 1 protein with N-terminal TMD (top)
  - 1 protein with N-terminal lipobox (bottom)
  - Adjacent or o-spanin embedded in i-spanin
  - i-spanin is never embedded in o-spanin
- OSP tool conducts naive ORF calls, should find embedded genes not found during structural annotation

TEXAS A&M AGRILIFE RESEARCH

Introns

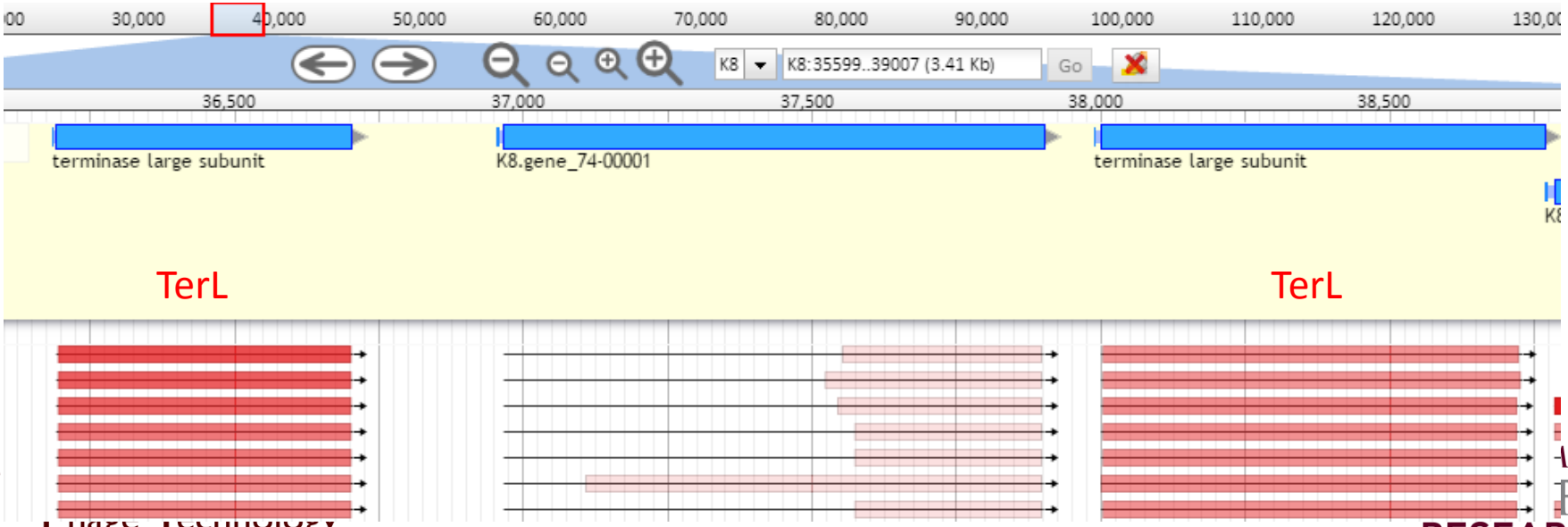
- An intron is an extra section of DNA that interrupts the protein-coding sequence of a gene
- This sequence has *ribozyme* activity and splices itself out of the mRNA, leaving an intact message and a free intron RNA
- Introns often (but not always) contain a homing endonuclease gene
- These are often found in *essential* genes, and often in genes involved in DNA metabolism

Introns are self-splicing elements

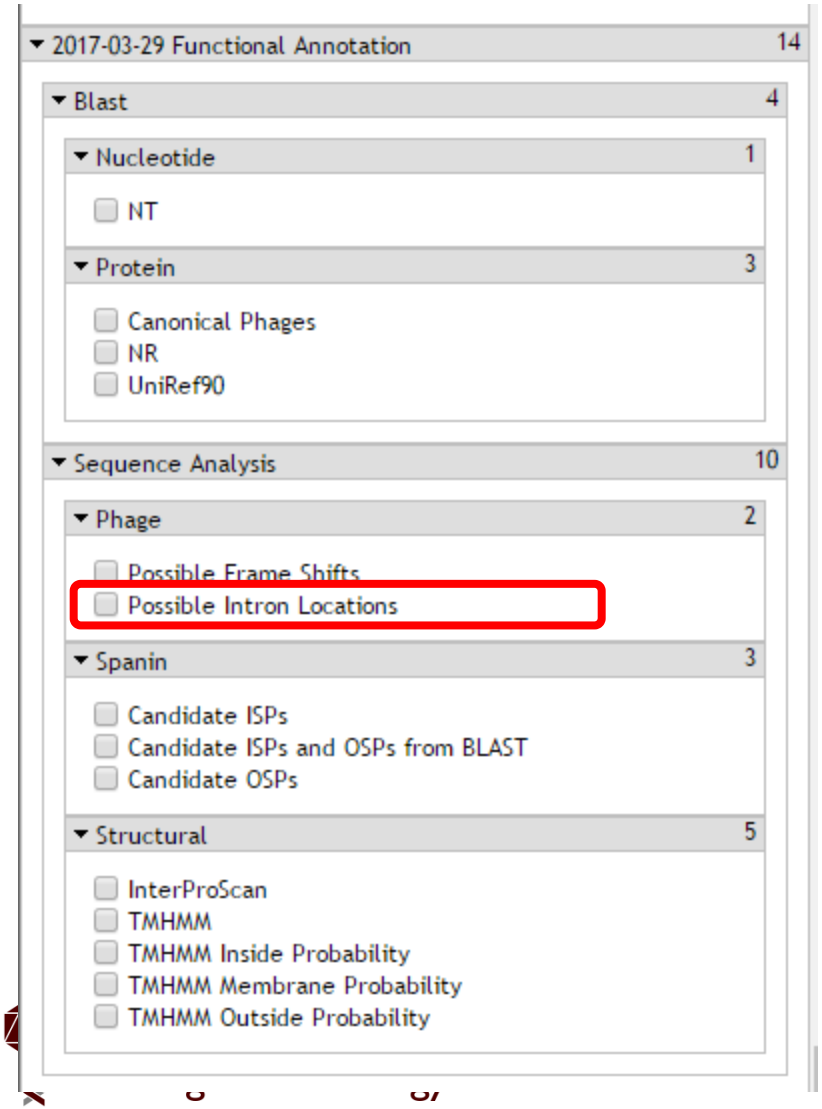


Introns

- Two or more genes that BLAST to the same protein could indicate an intron
- You can look at the BLAST results: do the two genes in your genome align to different portions of the same protein?

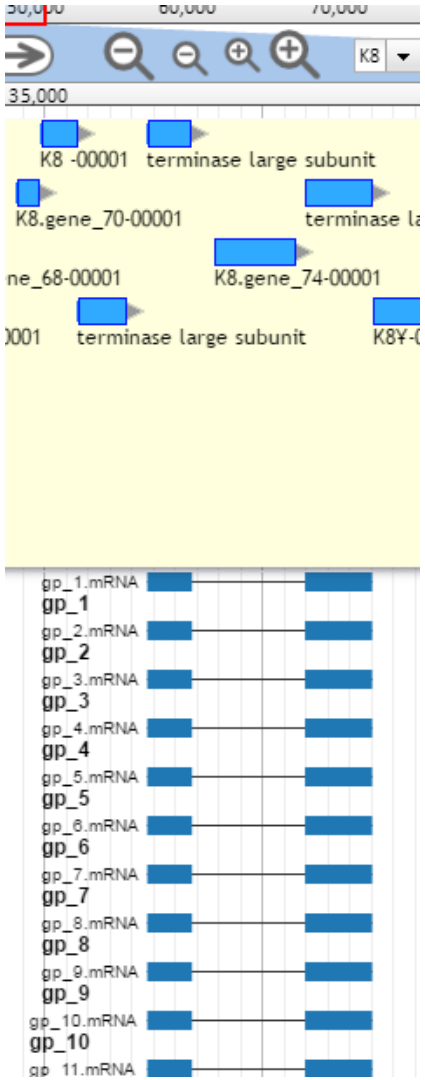


Intron finding track



- Will highlight genes that may have been disrupted by introns
  - Difficult to check all BLAST hits manually
- Searches BLAST results for nearby genes that BLAST to the same proteins
- Only works if the intron-disrupted gene has non-disrupted homologs in the database

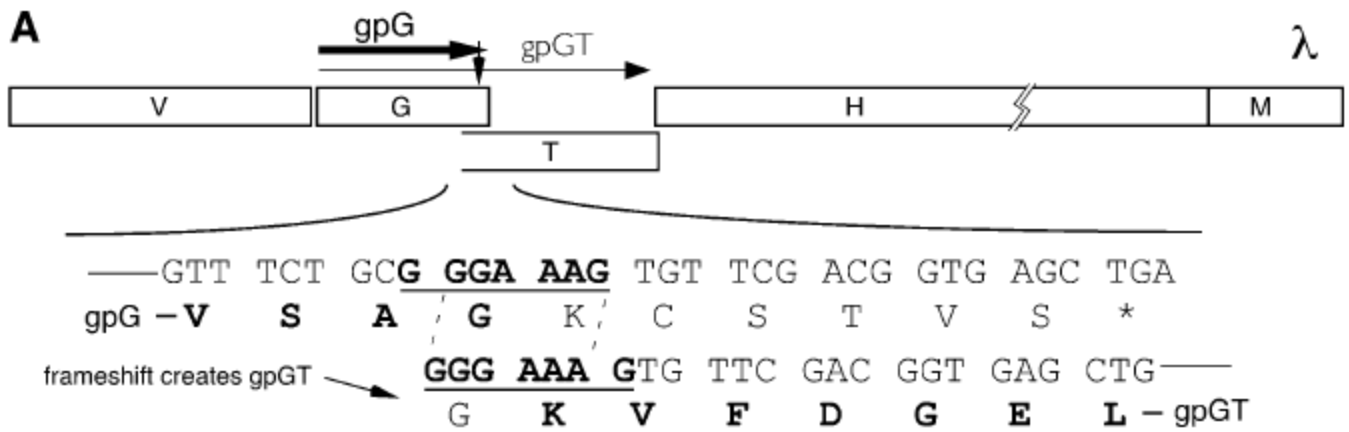
Intron finding



- Searches BLAST results for nearby genes that match to the same protein
- Will highlight possible genes that have been disrupted by introns
- May have a CDS within the intron, or not
- CDS may have a HNH or GIY-YIG domain, or not

Tail tape measure chaperones

- If you have a myophage or siphophage, you will have a tail tape measure protein that determines tail length
- Upstream of this protein 99% of the time will be the tape measure chaperones
- These often will be encoded by a **programmed translational frameshift**, a pair of genes upstream of the tape measure
  - The genes may have been called, but the frameshift is not



Xu *et al.* Mol. Cell 2004

Slippery sequences

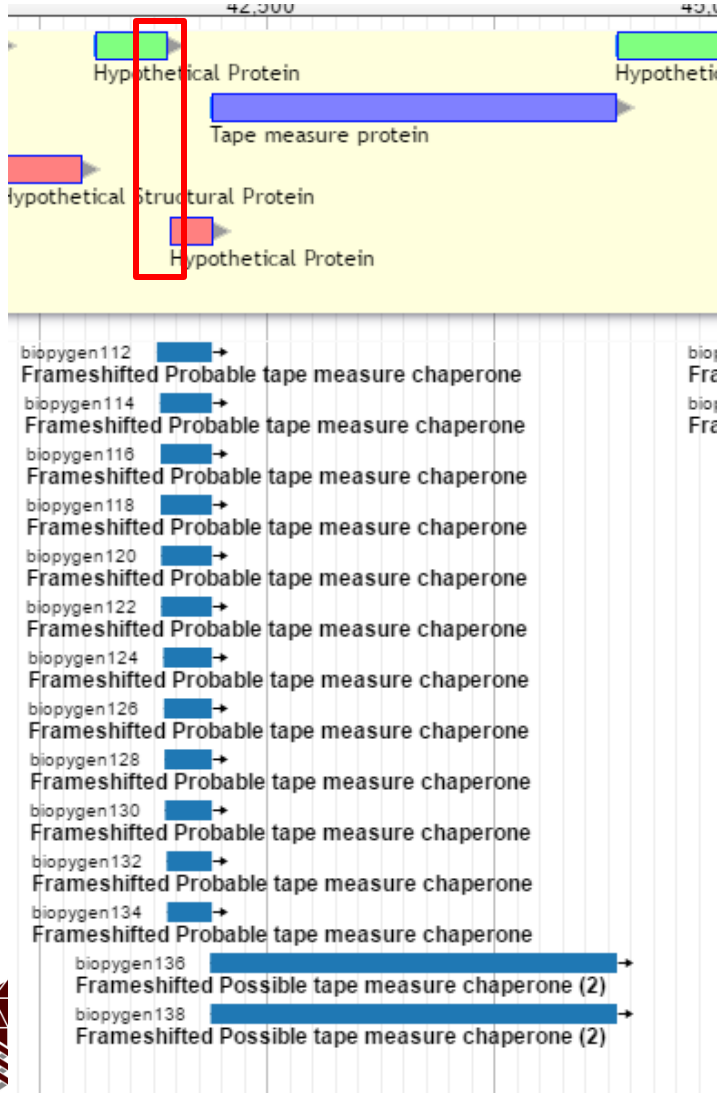
Phage	Slippery Sequence
λ[a]	TGTTTCTGCGGGAAAGTGTTCGACG
	C F C G K V F D →
	V S A G K C S T
P2[a]	GGTGGTCGGTTTTTTGTCGCCGAAC
	G G R F F V A E →
	V V G F L S P N
L5[a]	GACGCAACTGGGGGAAGCCGCGCCC
	D A T G G S R A →
	T Q L G E A A P
ϕC31[a]	GAACCTGAAGGGGAAGGCGAAGCCG
	E P E G E G E A →
	N L K G K A K P
TM4[a]	GCTGATCGAGGGAAAATCTCGCAGG
	A D R G K I S Q →
	L I E G K S R R
Mu[a]	ACACAAGAACGGGGGCGAGTGGCTG
	H K N G G E W L →
	T Q E R G R V A
HK97[a]	CGAAGCGCGGGAAAAGTCTCAACCC
	R S A G K V S T →
	E A R E K S Q P
HK022[a]	TAACGATCTGGGAAAGACTTCGAGC
	* R S G K D F E →
	N D L G K T S S
SPP1	TAACTCAGGTTTTTTCAAACGAGCT
	* L R F F Q T S →
	N S G F F K R A
PBSX	AGAAGCAGCAAAAAACTAGTAAAAG
	R S S K K L V K →
	E A A K N *

- Many tape measure chaperones will have a **slippery sequence** that allows the ribosome to shift into a different frame at low frequency
- Appears to be a conserved feature in many phages
- Follows the canonical motif **XXXYYYZ**
- Some phages do not appear to have this frameshift

Xu *et al.* Mol. Cell 2004

TEXAS A&M AGRILIFE RESEARCH

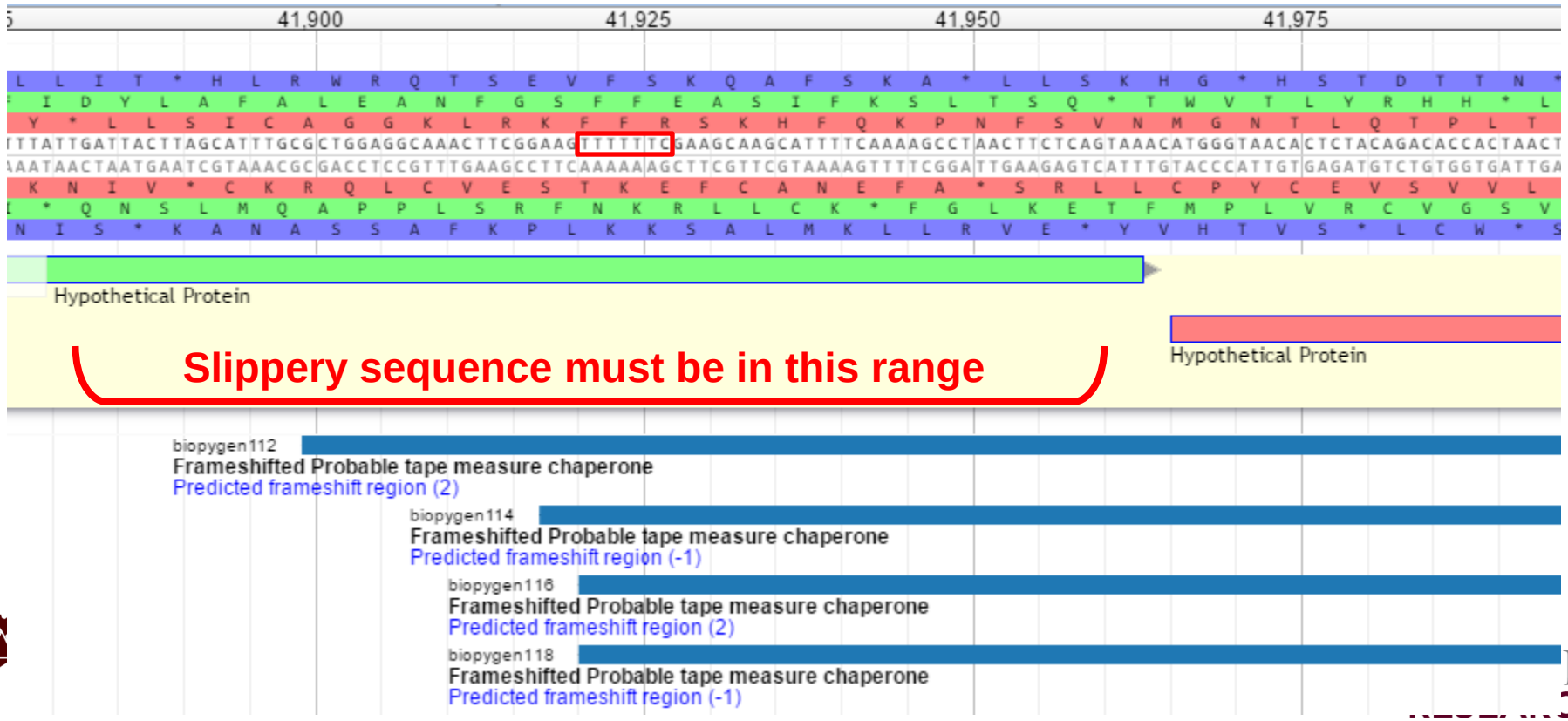
Finding the tape measure chaperones



- Locate the tape measure
  - Usually the longest gene in the genome, except for tail fibers
- Upstream will be the chaperones
- Turn on the “Possible Frame Shifts” track
- The frameshift will join the two genes upstream
  - Often the second gene was called but has a bad RBS or no RBS

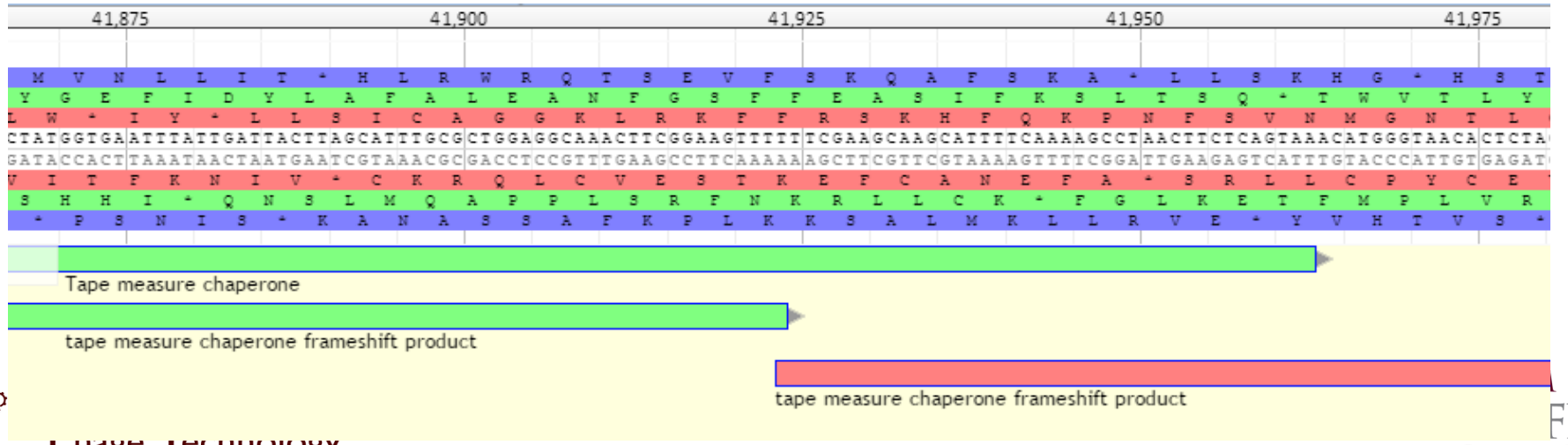
Finding the frameshift location

- Zoom in on the end of first chaperone component, look for frameshift sites
- Shifting into the -1 frame is most common but -2 shifts are possible



Annotating the frameshift proteins

- When complete, you will have **three** genes
  - Tape measure chaperone
  - 2 x Tape measure chaperone frameshift product
- The two frameshift product genes have to be merged when the genome is exported for submission
  - Apollo does not currently support overlapping exons



Handy tools and utilities

- Features detected in Genbank, Glimmer, MGA, TransTerm HP, BLAST, TMHMM, ARAGORN, InterProScan → GFF3 format
- BLAST XML format → human-readable table
- GFF3 annotations → Genbank or 5-column Sequin format
- Feature export/translate from Genbank or GFF3 format
- ShineFind: detects and annotates Shine-Dalgarno sequences
- Coding density, codon usage,
- PHACTS (Phantome)
- Circos
- PhageTerm
- Reopen/edit phage genome
- ...and more

Next steps

- We need (patient!) people to use the system with their real data
  - Locate bugs or inefficiencies
  - Find shortcomings in documentation
  - Determine needs for new tools or workflows
- Community annotation of canonical phage genomes (T-phages, lambda, Mu, N4, etc.)
- Provide options to automate structural and functional annotation
- https://cpt.tamu.edu/galaxy-pub
  - Accounts are free
  - Honor system to not kill our server with 500 InterProScan jobs

Acknowledgements

- CPT Software development
  - Cory Maughmer, lead developer and support
  - Eric Rasche, former lead developer
  - Eleni Mijalis, former assistant developer
- CPT faculty and staff
  - Ry Young
  - James Hu
  - Mei Liu