

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

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?

The screenshot displays the Galaxy CPT web interface. At the top, a navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', and 'Help', along with a user profile icon and a grid icon. The top right corner indicates 'Using 170.2 GB' of storage.

The main content area is titled 'News' and features a 'Galaxy Updates' section. Below this, a workflow is shown for analyzing a genome. The workflow includes several tracks: 'Gene Calls', 'Reference sequence', 'GeneMarkS', 'Glimmer3', 'MetaGeneAnnotator', and 'ShineFind from MGA'. The 'Gene Calls' track shows gene models for genes 27, 28, 29, 30, and 31. The 'Reference sequence' track shows the reference sequence for the region. The 'GeneMarkS' track shows gene predictions. The 'Glimmer3' track shows gene predictions. The 'MetaGeneAnnotator' track shows gene predictions. The 'ShineFind from MGA' track shows gene predictions.

On the left side, there is a 'Tools' panel with a search bar and a list of tools categorized by 'Get Data', 'CPT2: 464 Tools', 'CPT2: Utilities', 'CPT2: ABIF/AB1', 'CPT2: Blast', 'CPT2: Fasta Tools', 'CPT2: GFF3', 'CPT2: Genbank Tools', 'CPT2: Comparative Genomics', 'CPT2: Phage Analysis Tools', 'CPT2: NGS', 'CPT2: PAUSE3', 'CPT2: JBrowse', 'CPT:Scripts and Analysis', 'CPT:Genbank', 'CPT:Blast', 'CPT:Admin', 'CPT:External Software', 'CPT:Oneoff/Custom', 'CPT:Circos Tools', and 'CPT:PHANTASM v1'.

On the right side, there is a 'History' panel showing a list of datasets. The top dataset is 'Copy of 'PAP of Mt0425'' shared by 'ryland@tamu.edu' (active items only), with 16 shown, 12 deleted, and 25 hidden. Below it are several other datasets with links to their respective tools, such as '61: NCBI EFetch Results', '40: Concatenate datasets on data 39 and data 37', '39: Filter sequences by length on data 38', '38: Genbank Genome Sequence Export on data 36', '37: JBrowse on Mt0425DNA.fasta', '36: NCBI Entrez EFetch on data 33', '35: Rebase GFF3 features on data 32 and data 14', and '34: Report on top blast hits'.

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

The screenshot displays the Galaxy CPT web interface. The top navigation bar includes links for 'Galaxy / CPT', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', and 'Help', along with a user profile icon and a grid icon. The right side of the top bar indicates 'Using 170.2 GB'.

Left Panel: Tools

- Search tools:
- Get Data
- Get Data: NCBI
- Get Data: MORE
- CPT TOOLS
- CPT2: 464 Tools
- CPT2: Utilities
- CPT2: ABIF/AB1
- CPT2: Blast
- CPT2: Fasta Tools
- CPT2: GFF3
- CPT2: Genbank Tools
- CPT2: Comparative Genomics
- CPT2: Phage Analysis Tools
- CPT2: NGS
- CPT2: PAUSE3
- CPT2: JBrowse
- CPT: Scripts and Analysis
- CPT: Genbank
- CPT: Blast
- CPT: Admin
- CPT: External Software
- CPT: Oneoff/Custom
- CPT: Circos Tools
- CPT: PHANTASM v1

Center Panel: News

Galaxy Updates

Available Tracks

- ☒ filter tracks
- Gene Calls** (4)
- ☒ GeneMarkS
- ☒ Glimmer3
- ☒ MetaGeneAnnotator
- ☒ ShineFind from MGA
- Reference sequence** (1)
- ☒ Reference sequence

Genome Track View

Genome: 0 20,000 40,000 60,000 80,000 100,000 120,000

Track: NC_005880:9611..13120 (3.51 Kb)

View: 10,000 11,250 12,500

Help: Zoom in to see sequence, Zoom in to see sequence

GeneMarkS

- gene_27
- gene_28
- gene_29
- gene_30
- gene_31

Glimmer3

- cds_orf00027
- cds_orf00028
- cds_orf00029
- cds_orf00030
- cds_orf00031
- cds_orf00032

MetaGeneAnnotator

- cds_gene_27
- cds_gene_28
- cds_gene_29
- cds_gene_30
- cds_gene_31
- cds_gene_32

Right Panel: History

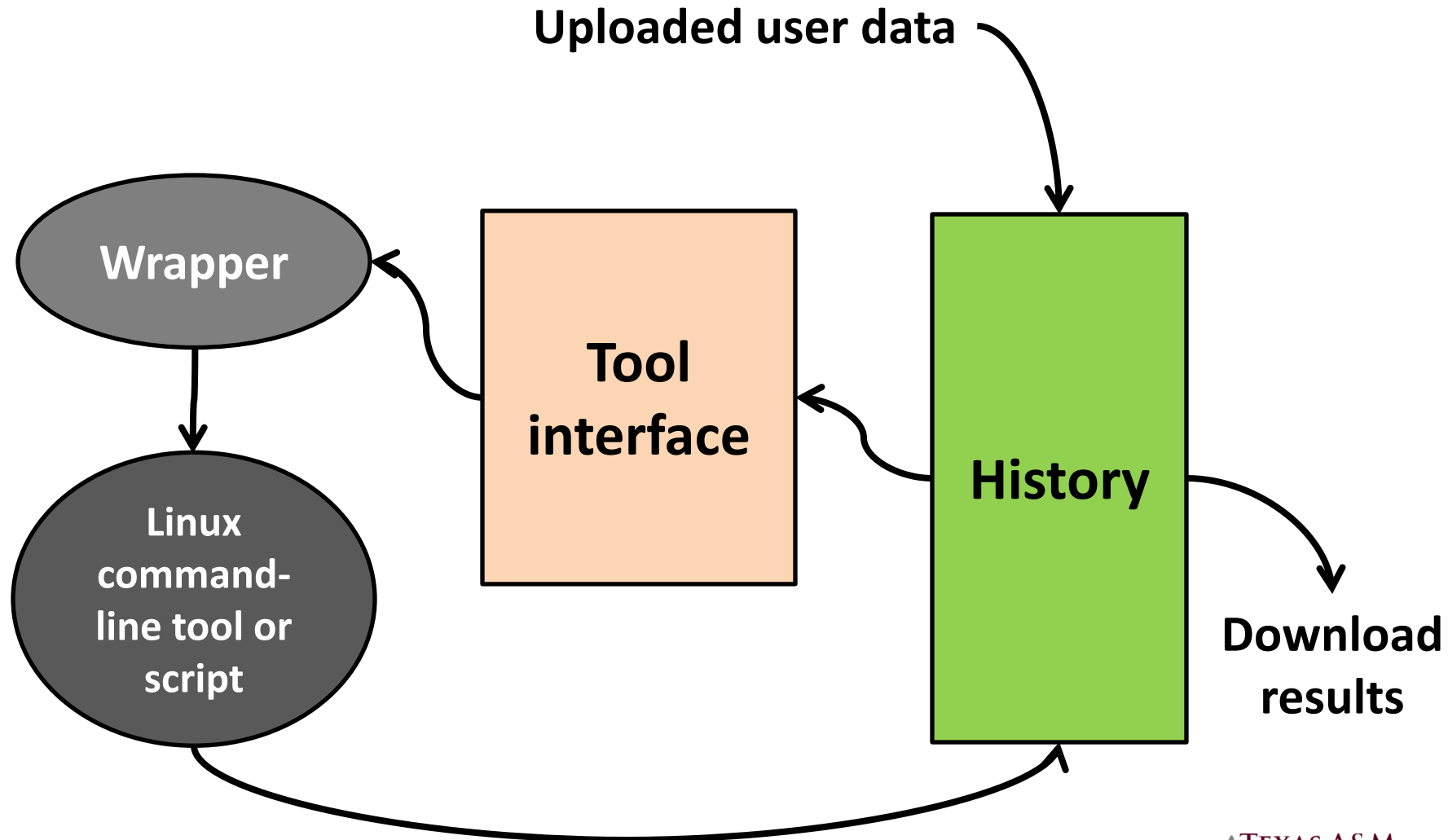
search datasets

Copy of 'PAP of Mt0425' shared by 'ryland@tamu.edu' (active items only)
16 shown, 12 deleted, 25 hidden

84.04 MB

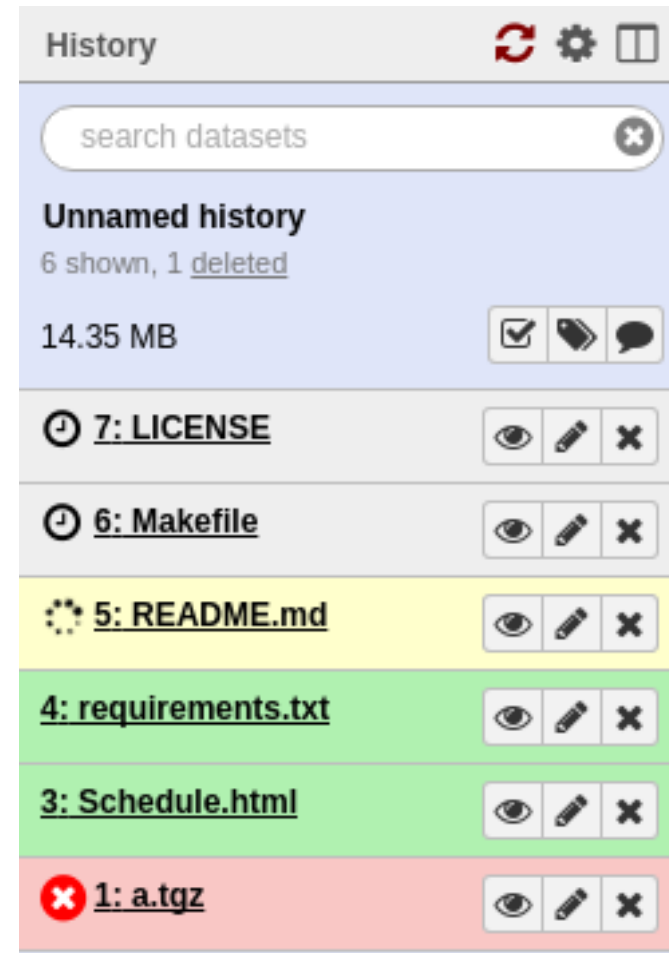
- 61: NCBI EFetch Results
- 40: Concatenate datasets on data 39 and data 37
- 39: Filter sequences by length on data 38
- 38: Genbank Genome Sequence Export on data 36
- 37: JBrowse on Mt0425DNA.fasta
- 36: NCBI Entrez EFetch on data 33
- 35: Rebase GFF3 features on data 32 and data 14
- 34: Report on top blast hits
- 33: Top accession

General order of operations



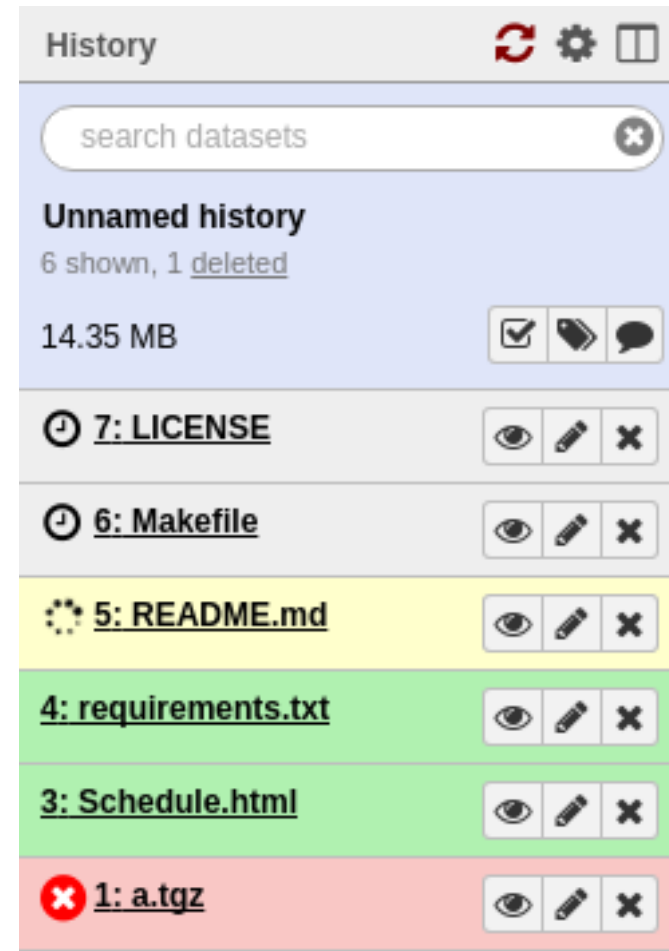
The history

- 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

The screenshot displays the Galaxy / CPT web interface. At the top, a dark navigation bar contains the 'Galaxy / CPT' logo, a series of menu items (Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User), a grid icon, and a storage indicator 'Using 33.7 GB'. Below this is a light gray header bar with search fields: 'search histories' and 'search all datasets', and a 'Create new' button. The main content area is divided into four vertical panels, each representing a different history. Each panel has a title, a summary of items (shown, deleted, hidden), a size, and a list of datasets. The 'Current History' panel on the left shows 'Margaery stitching' with 27 shown, 6 deleted, and 17 hidden items, totaling 16.8 MB. It lists three datasets: '50: JBrowse on Margaery 150805', '49: Convert XMFA to gapped GFF3 on data 13, data 43, and data 45', and '48: MIST v3 on data 44'. The other three panels ('Papaya PAUSE', 'Pierogi PAUSE', and 'Barrett PAUSE') follow a similar layout with their respective dataset lists. Each dataset entry includes a name, a small icon, and a set of control icons (eye, pencil, and X).

Galaxy / CPT

Analyze Data Workflow Shared Data Visualization Admin Help User

Using 33.7 GB

Done search histories search all datasets Create new

Current History

Margaery stitching
27 shown, 6 deleted, 17 hidden
16.8 MB

search datasets

Drag datasets here to copy them to the current history

50: JBrowse on Margaery 150805

49: Convert XMFA to gapped GFF3 on data 13, data 43, and data 45

48: MIST v3 on data 44

Papaya PAUSE
8 shown, 9 deleted, 4 hidden
1.7 GB

search datasets

21: Fix Gene Boundaries on Percy

19: Caulobacter phage Percy.gbkl

16: Analyse TerL Sequences on data 15

15: Pasted Entry

Pierogi PAUSE
11 shown, 8 deleted, 36 hidden
2.0 GB

search datasets

55: JBrowse on Pierogi.fa

53: MIST v3 on data 49

36: Phage QC on data 14 and data 31

34: Start Codon Usage

Barrett PAUSE
4 shown, 22 deleted, 19 hidden
1.9 GB

search datasets

8: JBrowse on

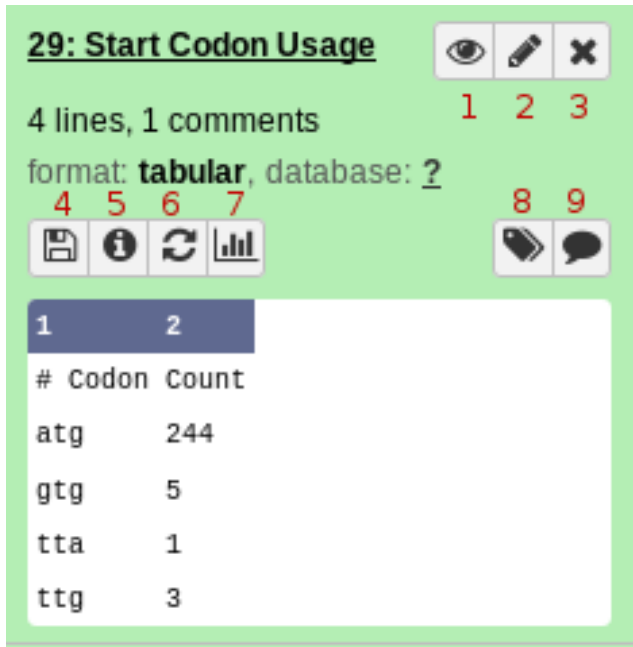
NODE2 Barrett150730 143268 cov 23.5 fa

3:

NODE2 Barrett150730 143268 cov 23.5 fa

2: GSAF Download (Sample2_S3_L001_R2_001.fastq)

Datasets in the history



29: Start Codon Usage

4 lines, 1 comments

format: **tabular**, database: ?

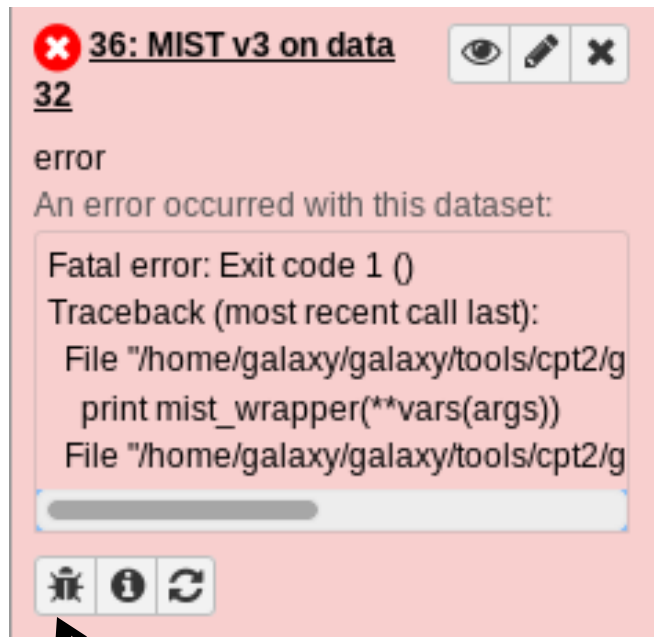
4 5 6 7 8 9

# Codon	Count
atg	244
gtg	5
tta	1
ttg	3

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

- 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






Bug report button



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




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



Analyzing data

NCBI BLAST+ blastp Search protein database with protein query sequence(s) 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



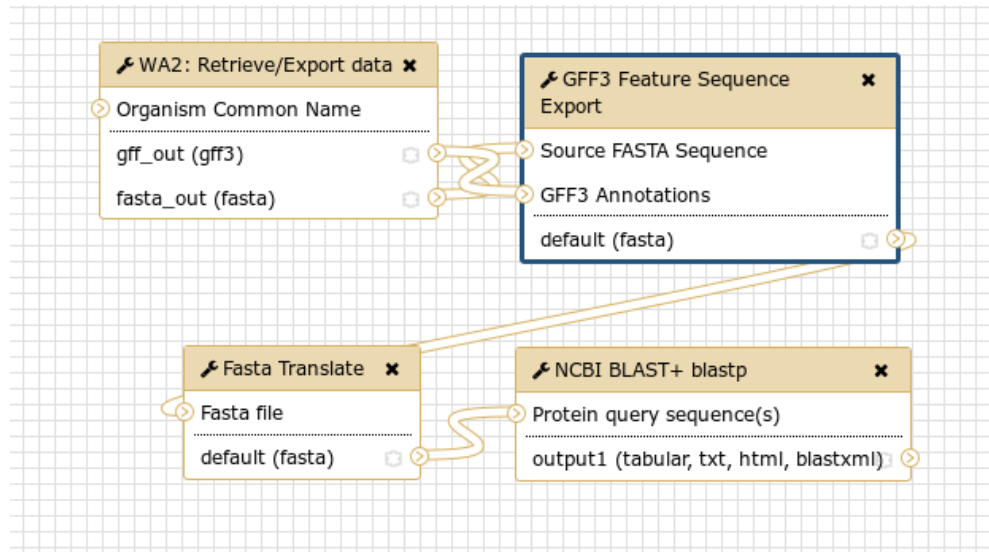
 **Note.** Database searches may take a substantial amount of time. For large input datasets it is advisable to allow overnight processing.

- 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

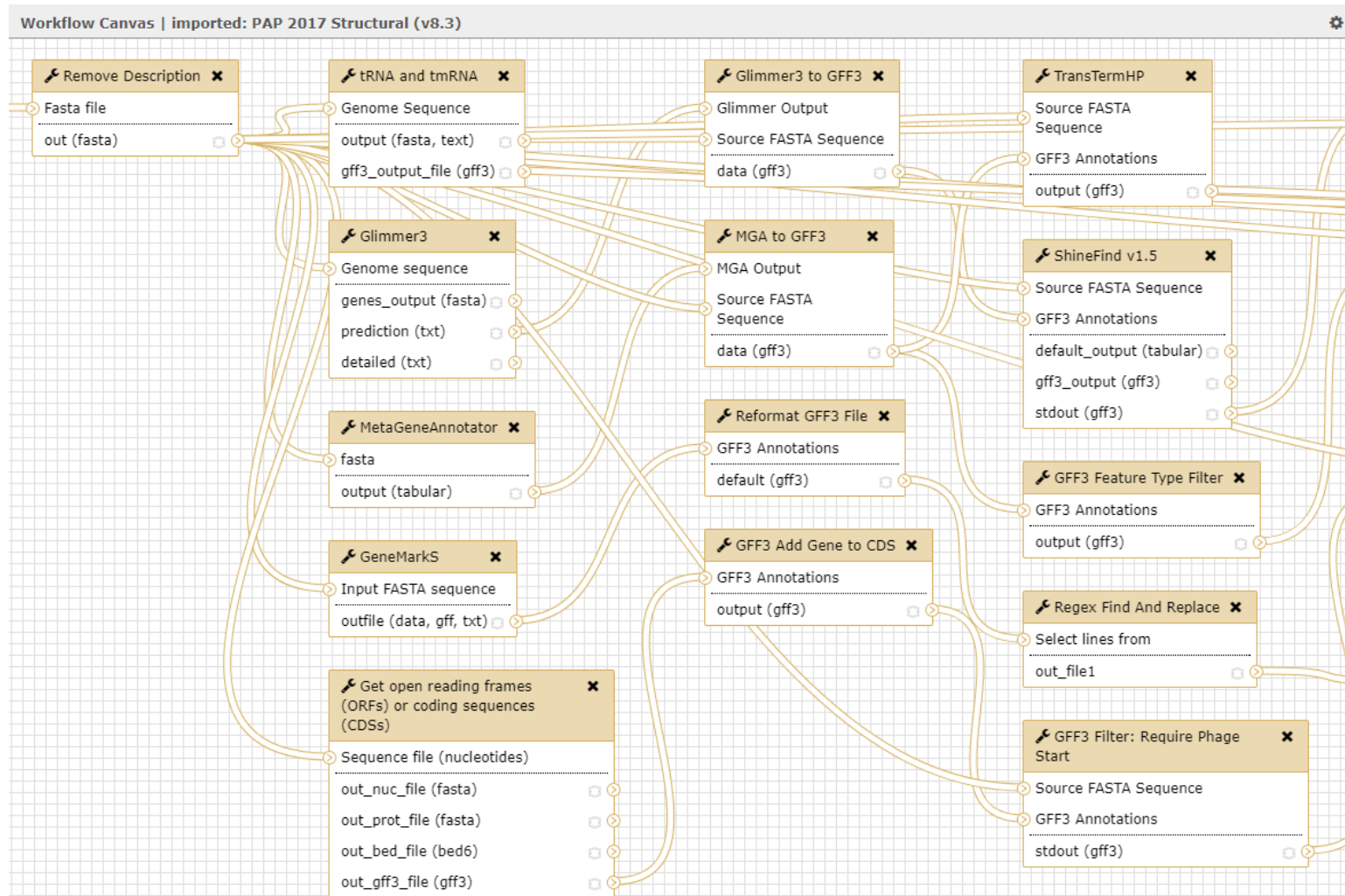
- One of Galaxy's most powerful features is the ability to connect jobs in workflows
- Some analyses take a long time; output from one job will automatically be handed off to the next when it finishes



File conversion tools

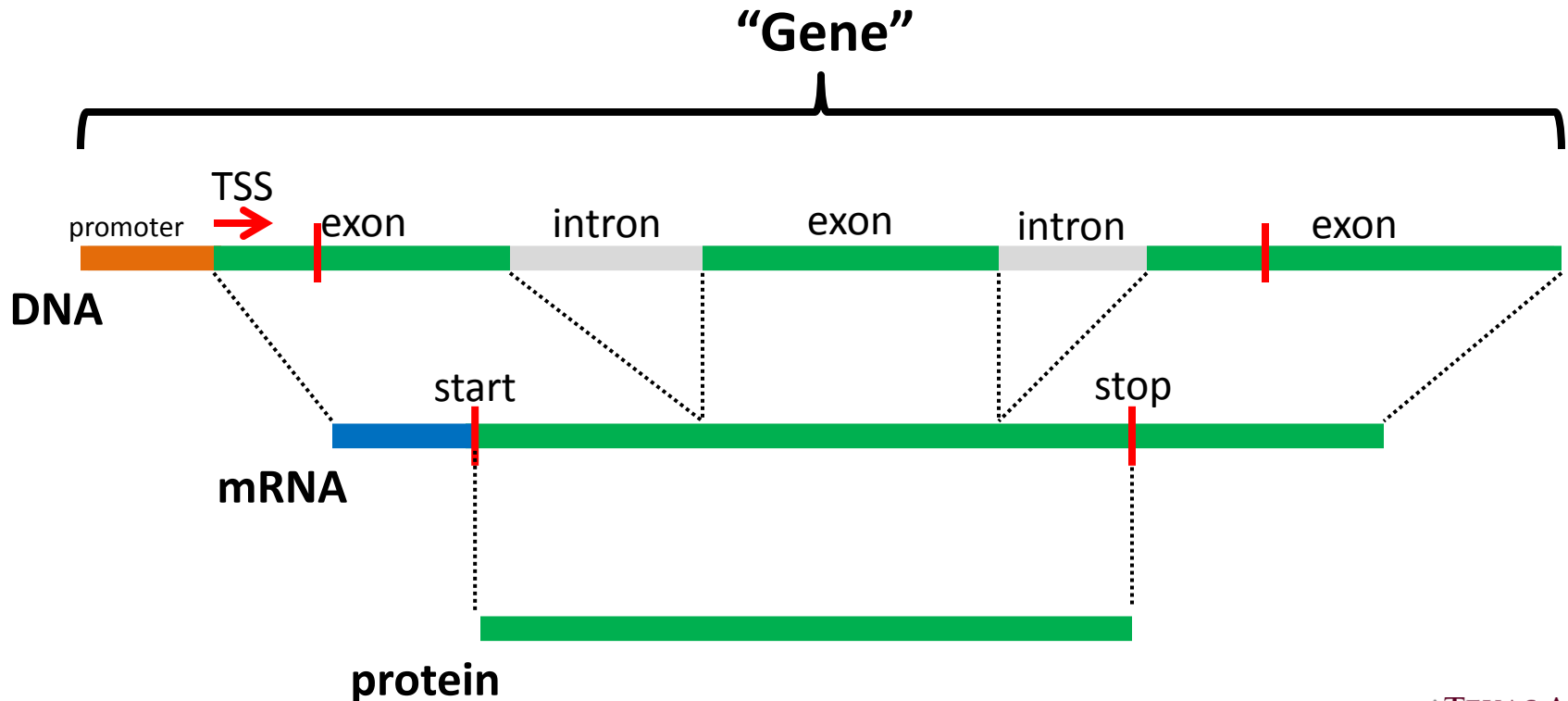
- Almost every analysis tool has its own unique output format
- An often invisible but vital part of Galaxy workflows is reformatting the output of one tool to serve as input for the next step
- Part of the workflows are simple tools that parse, reformat or extract data from outputs
- The GFF3 format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

Part of the structural annotation workflow



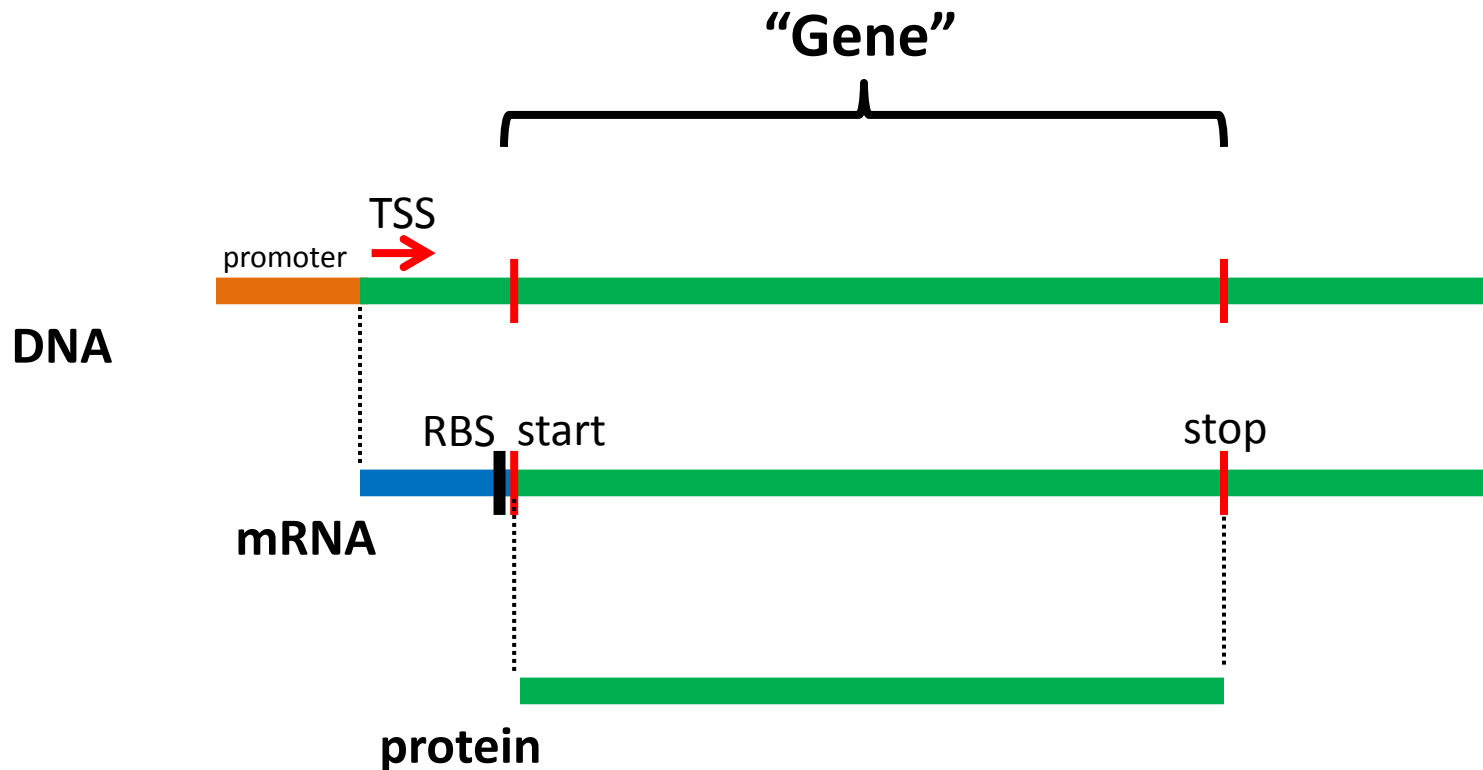
Eukaryotic gene

- 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.

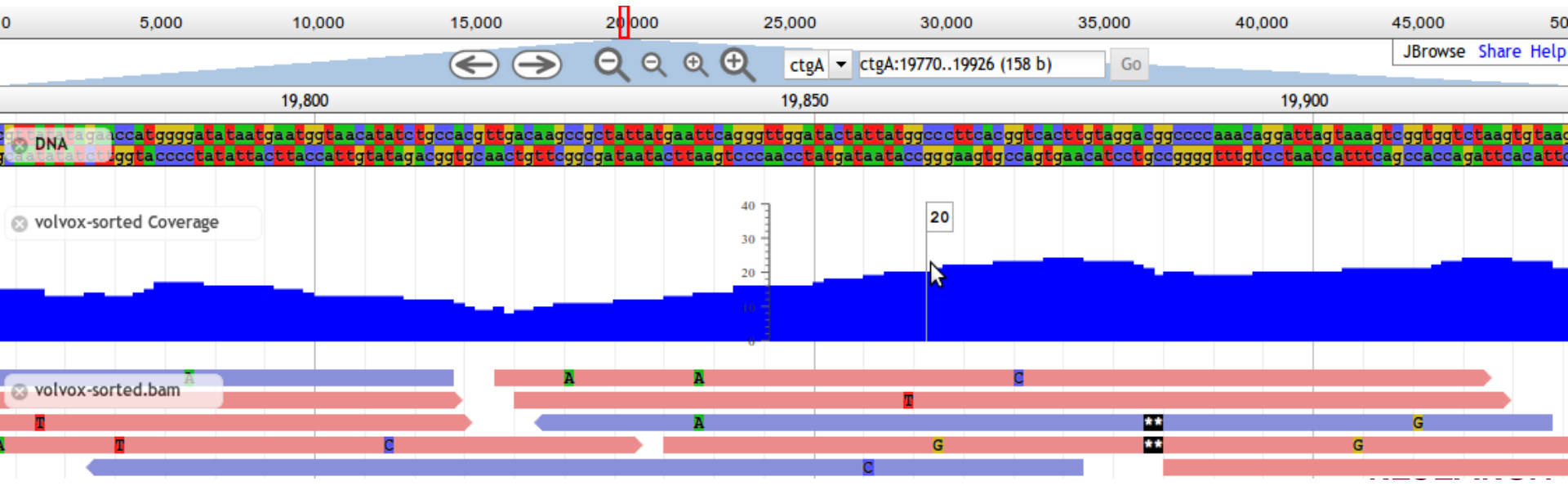
attributes: Feature attributes listed as tag=value

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

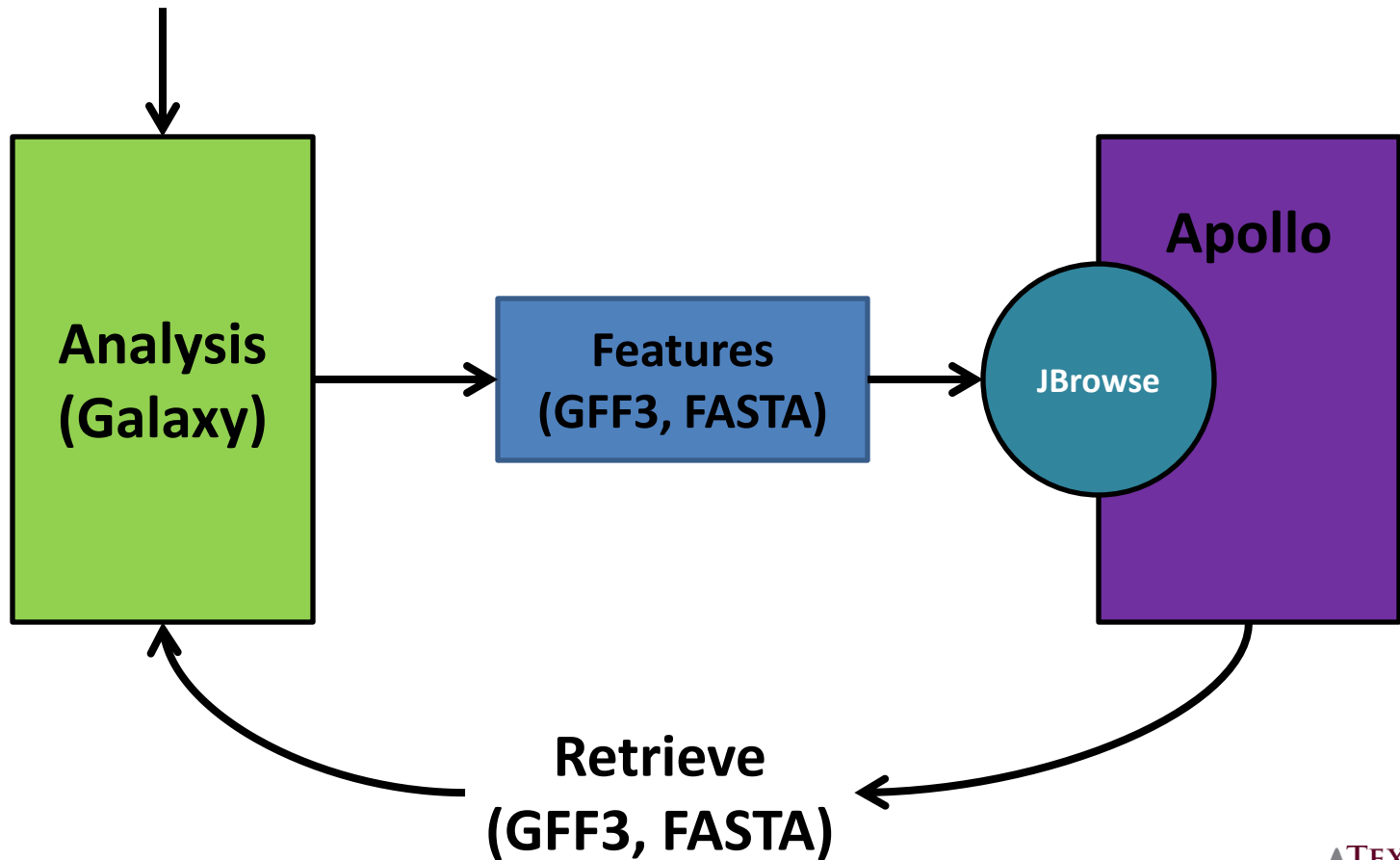
JBrowse

- 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

User uploaded data



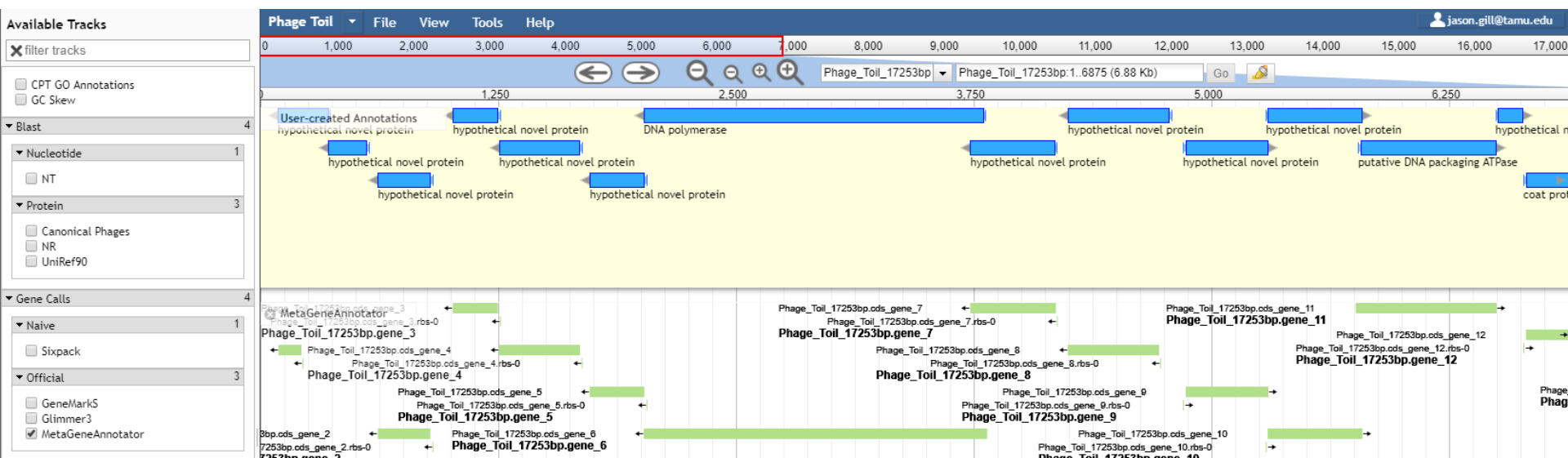
Annotations in Apollo

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

The screenshot displays the Apollo genome browser interface. On the left, the 'Available Tracks' panel is visible, showing a search bar for 'filter tracks' and several track categories: 'Blast' (4 tracks), 'Nucleotide' (1 track), 'Protein' (3 tracks), 'Gene Calls' (4 tracks), and 'Naive' (1 track). The 'Protein' section is expanded, showing checkboxes for 'Canonical Phages', 'NR', and 'UniRef90'. The 'Gene Calls' section is also expanded, showing checkboxes for 'GeneMarkS', 'Glimmer3', and 'MetaGeneAnnotator'. The main display area shows a genomic track for 'Phage Toil' with a scale from 0 to 11,000. The top track, 'User-created Annotations', contains several blue bars representing protein annotations, including 'hypothetical novel protein', 'putative DNA packaging ATPase', 'LysM domain protein', and 'coat protein'. Below this, the 'Gene Calls' track shows annotations from various tools, including 'DNA polymerase' and several 'hypothetical novel protein' entries. The interface includes a menu bar with 'File', 'View', 'Tools', and 'Help', and a search bar at the top right.

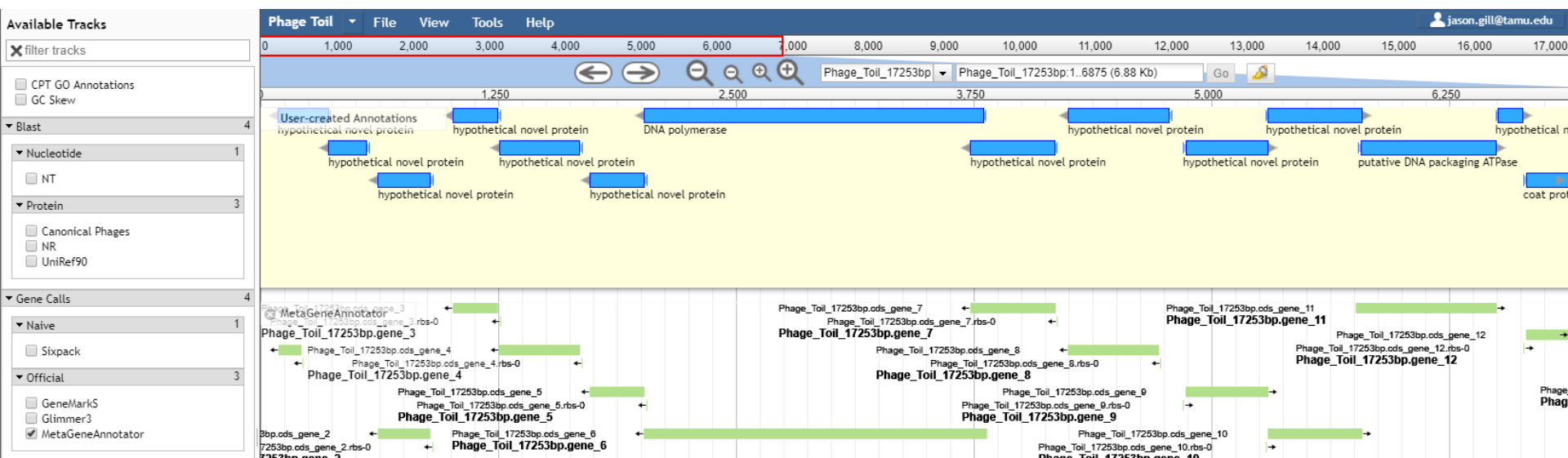
Evidence tracks in Apollo

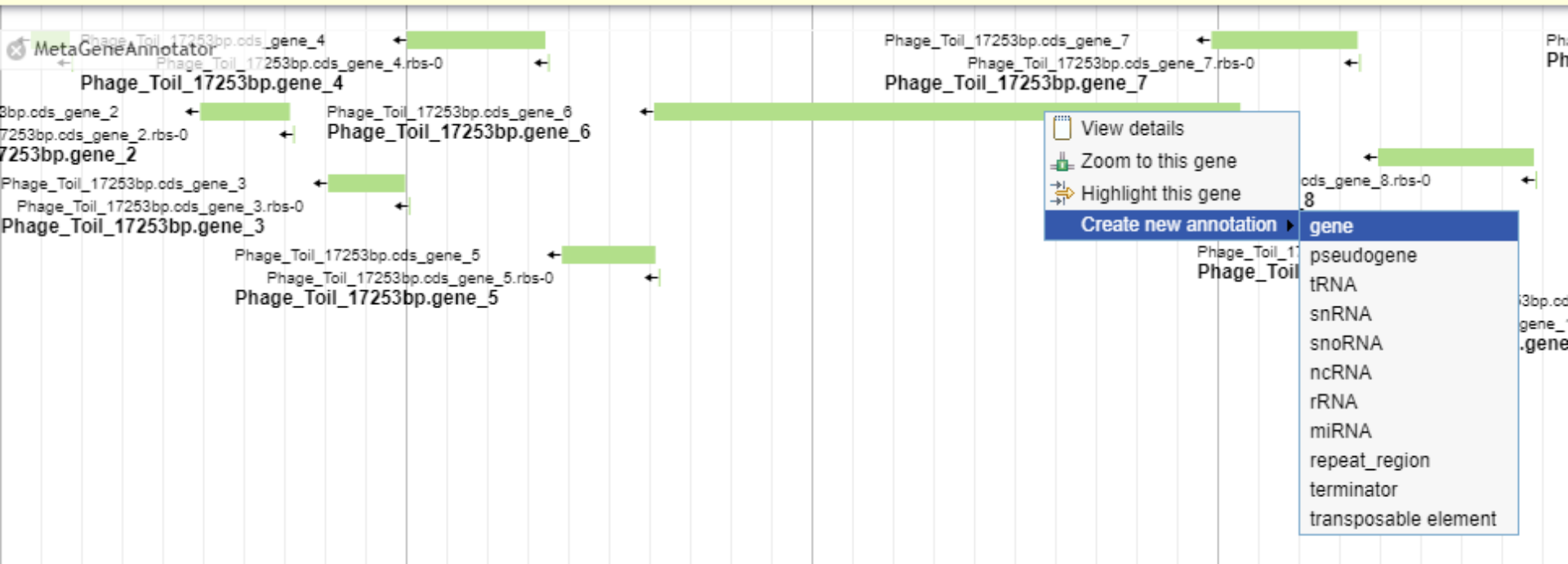
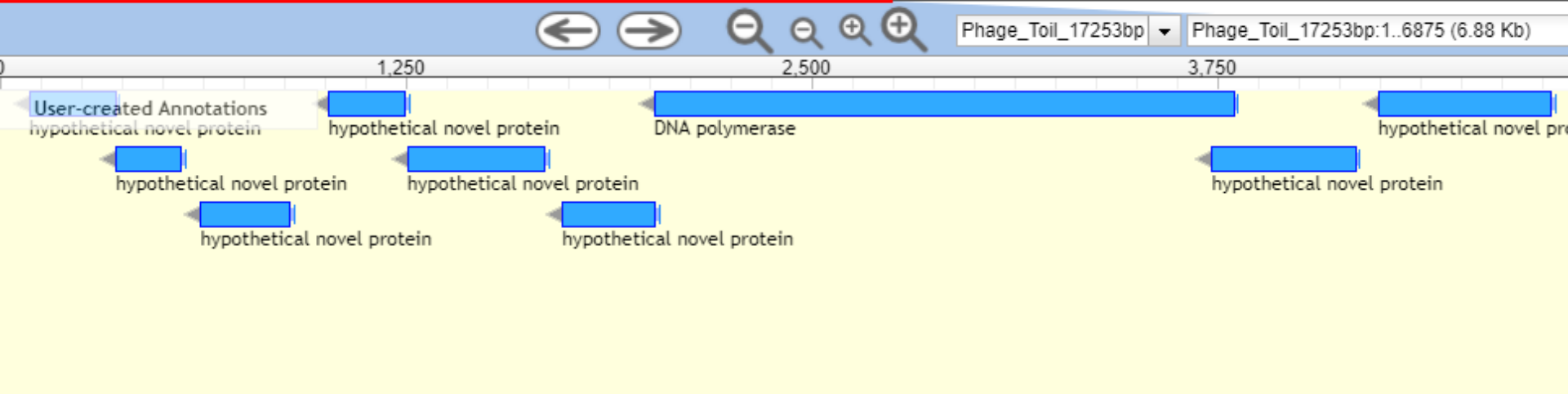
- Features in Apollo can **only** be created from evidence tracks
- Unlike purely manual editors like Artemis, the user cannot select sequence and create a feature *de novo*
- This is part of a philosophical decision by Apollo, that features are **only created with evidence**



Structural annotation

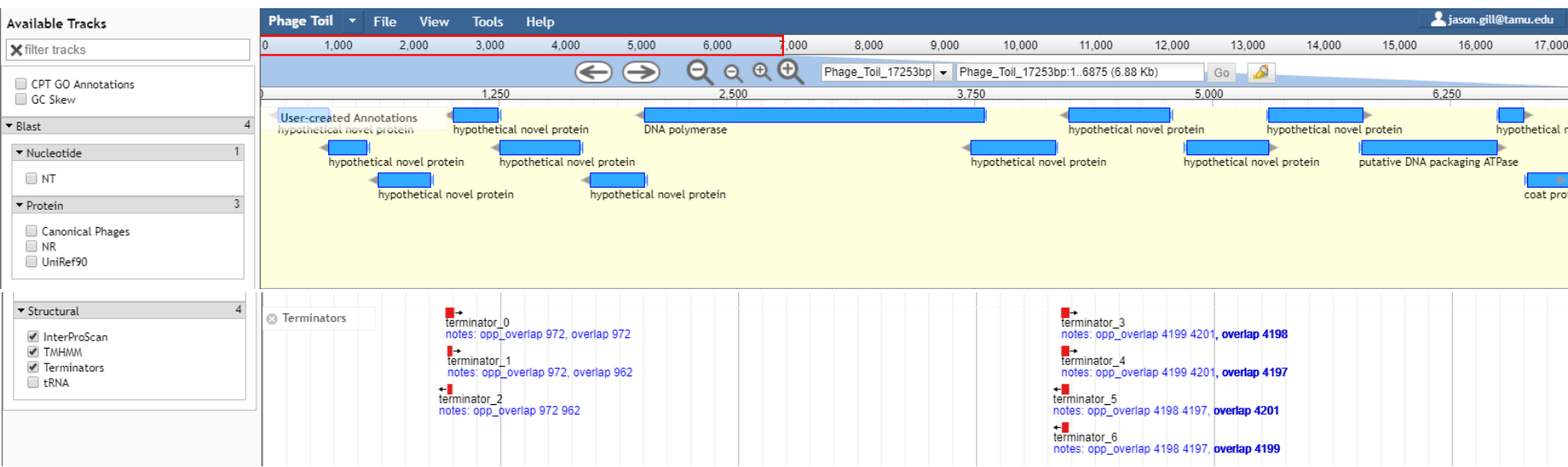
- The output from gene-calling programs appears in the evidence track
 - Output is first run through a program that adds RBS sequences to each gene
 - Genes are manually called and moved into the annotation track
- Gene callers sometimes miss genes, so there is a track generated by a naive ORF caller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





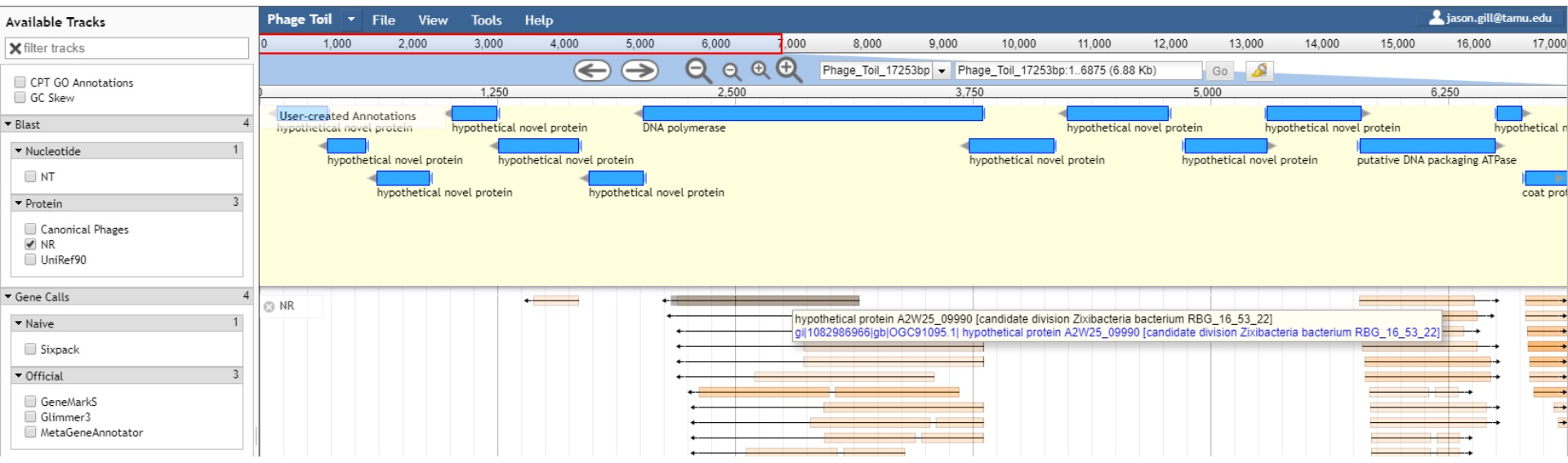
Structural annotation: terminators

- 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



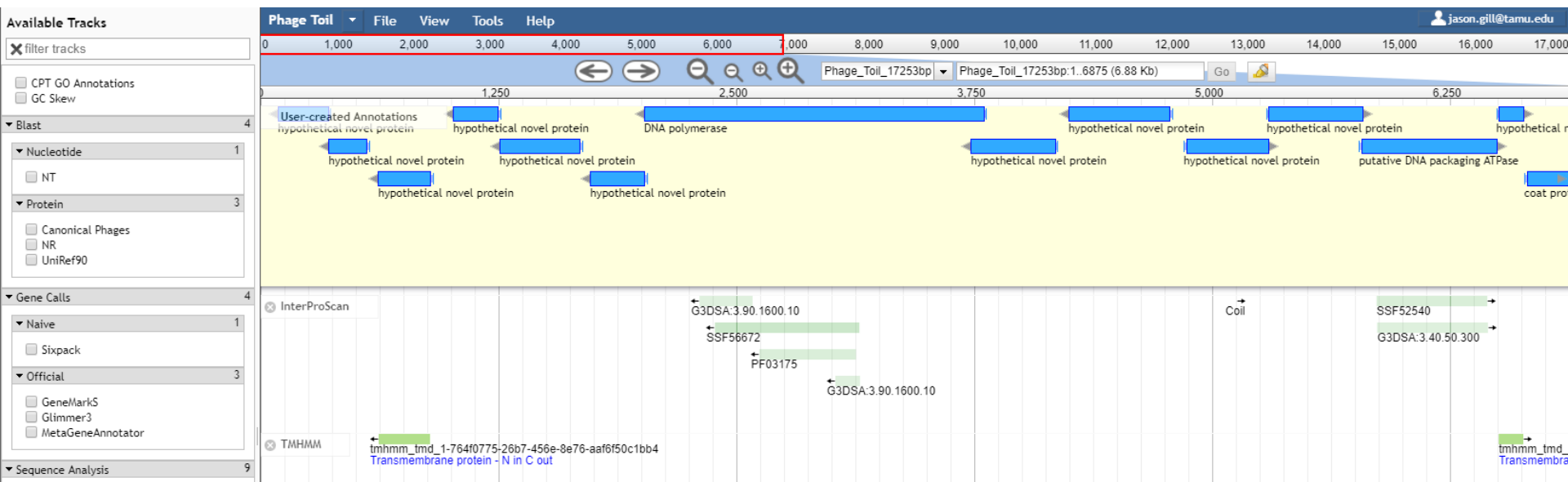
Functional annotation: BLASTp

- 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

The screenshot displays the Phage Toil software interface. A window titled "protein_match PF03175" is open, showing details for a protein match. The background shows a genomic map with various annotations like "User-created Annotations", "hypothetical novel protein", and "DNA polymerase".

Primary Data

Name	PF03175
Type	protein_match
Score	6e-13
Position	Phage_Toil_17253bp:2621..3127 (- strand)
Length	507 bp

Attributes

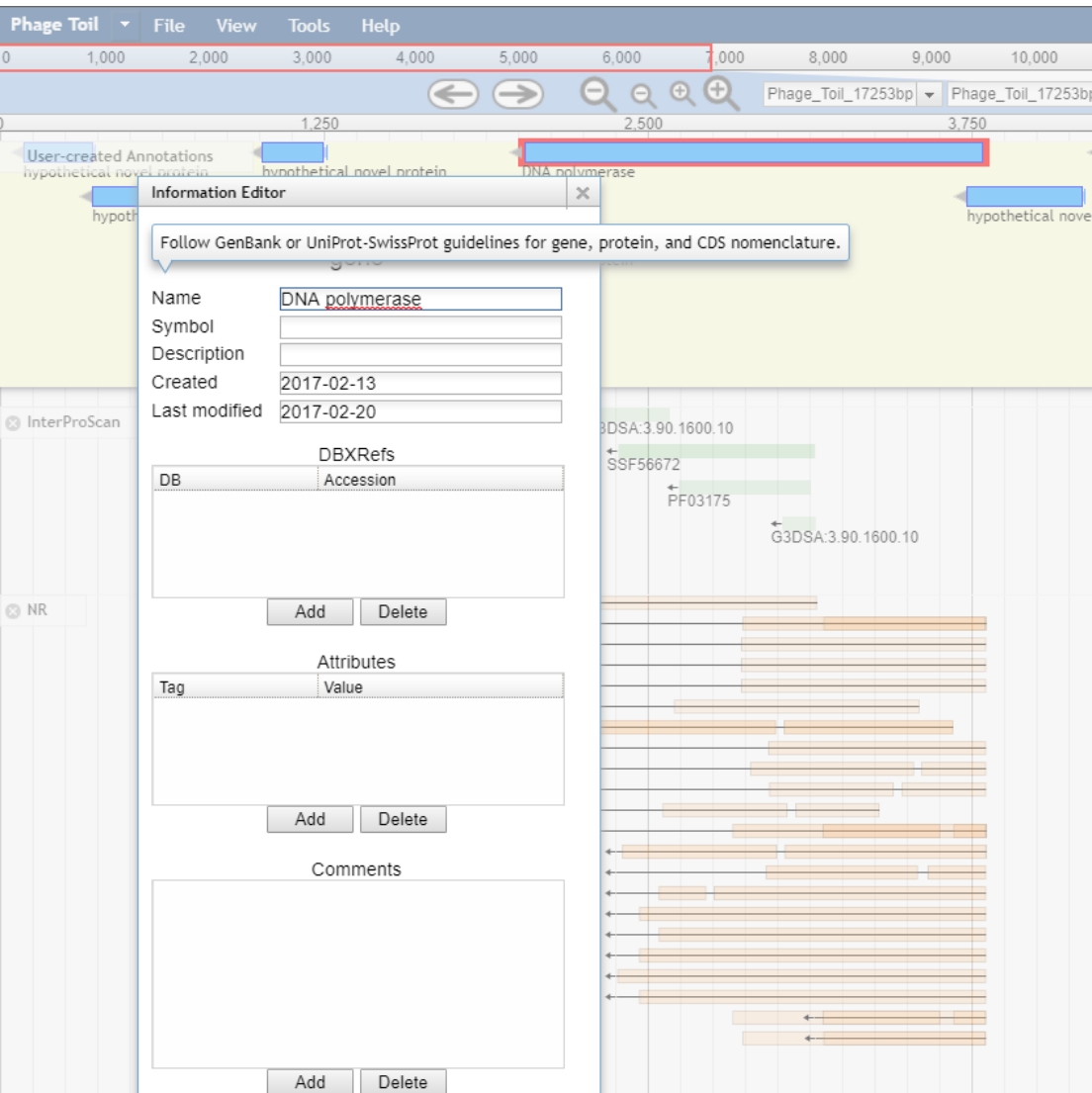
Date	19-02-2017
Dbxref	InterPro:IPR004868", "KEGG:00230+2.7.7.7", "KEGG:00240+2.7.7.7
Id	match\$14_227_395
Ontology_term	GO:0000166", "GO:0003677", "GO:0003887", "GO:0006260", "GO:0008408
Seq_id	Phage_Toil_17253bp
Signature_desc	DNA polymerase type B organellar and viral
Source	Pfam

Region sequence

```
>Phage_Toil_17253bp Phage_Toil_17253bp:2621..3127 (- strand) class=protein_match length=507
TTGCGATACGCAGGAGGACGTTTCGAGTCCTACAAAACAGGACTGTATGAAAGCCCTGT
CTATCAGTACGACATTTCGTTCCGCATATCCTTACGCACCTCACACAGTGTCTGCACTCA
CCGAGGATTACGAGCGTGACGATAGCCCCACACAGGGCAGACCCGTCCTAGCTTTAGC
CTCTGTGGAATTCGATACTATGACAACACCATTTGACCGACGAGGGGATTAATCCATTTGC
TCTGCGTTCTGGGTCAGGAGCAATCTACTTCCCTAACTTTGTCGAAACATGGGTTTGGG
GAATTGAATACAAATGCAGCACTACGACACCGGGCGGAGTTTCTCGAGTTGGTTTCTACC
ATCACGTTTACGATGACGGTACTAGACCATTTTCATTTCATCGGTGATCTATACGATCA
GCGGGCGAAATGGAAGCGAGAAGGTAACCTGACACATTTGGCTTGTAAAGCTAGGGATGA
ATAGTTGCTATGGGAAGTTGGCGCAACGCGTCGCG
```



Adding annotations

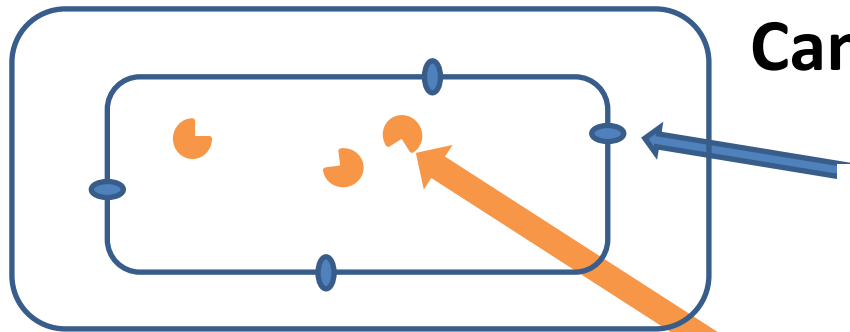


- Functional annotations are typed in by the user for each gene based on evidence
- Hoping to automate this step in the future

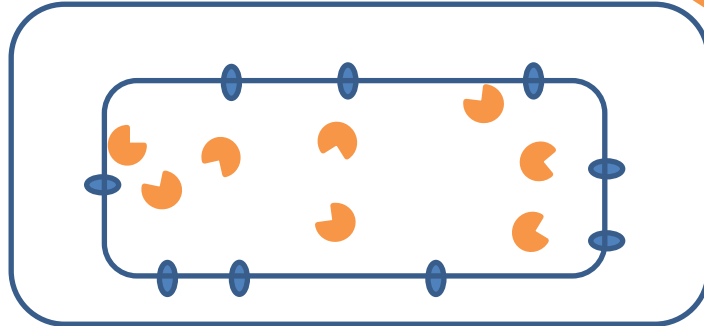
Solutions for “special” phage cases

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

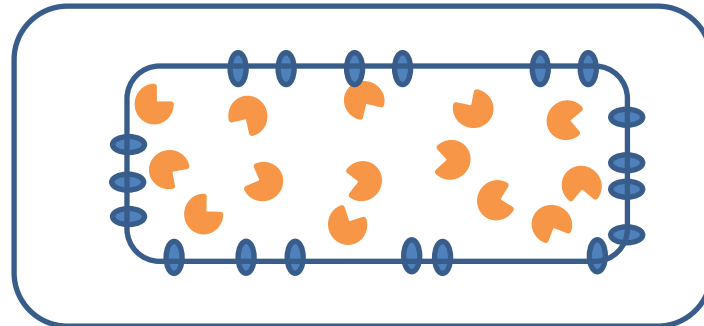
Canonical holin /endolysin lysis



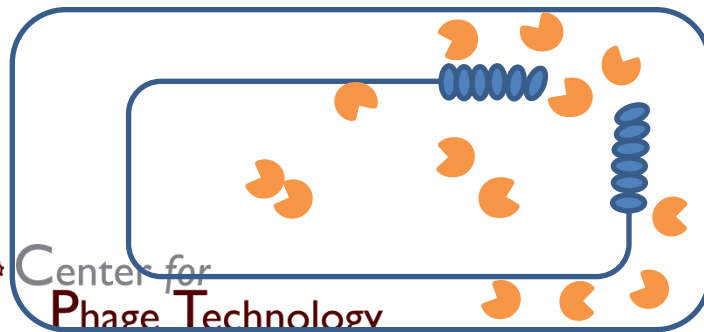
Holin accumulates in IM,
mobile and harmless



Endolysin accumulates in
cytoplasm, fully active



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

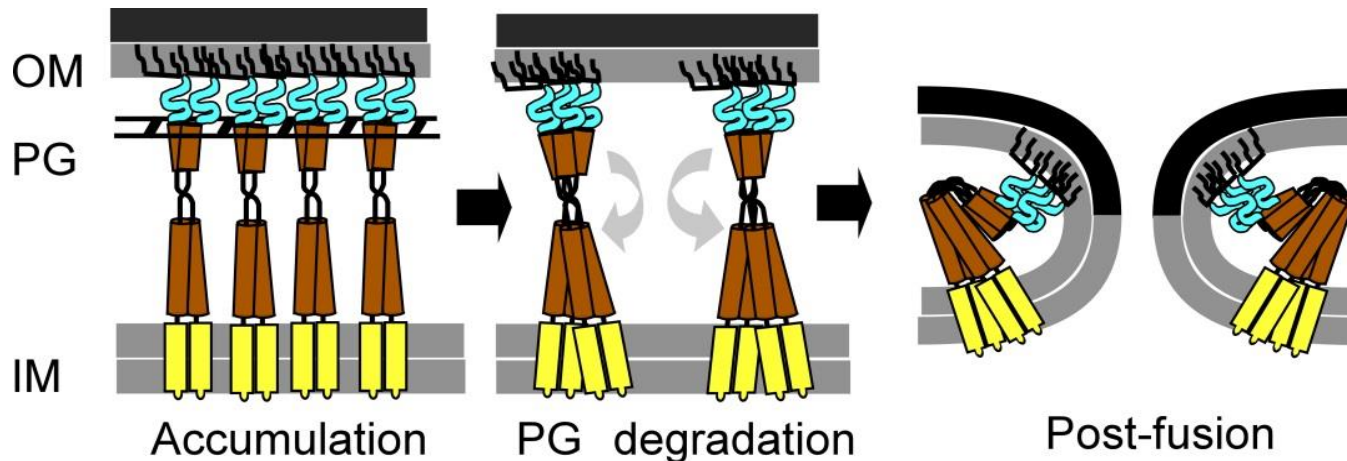


Endolysin escapes through holes
& attacks peptidoglycan



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

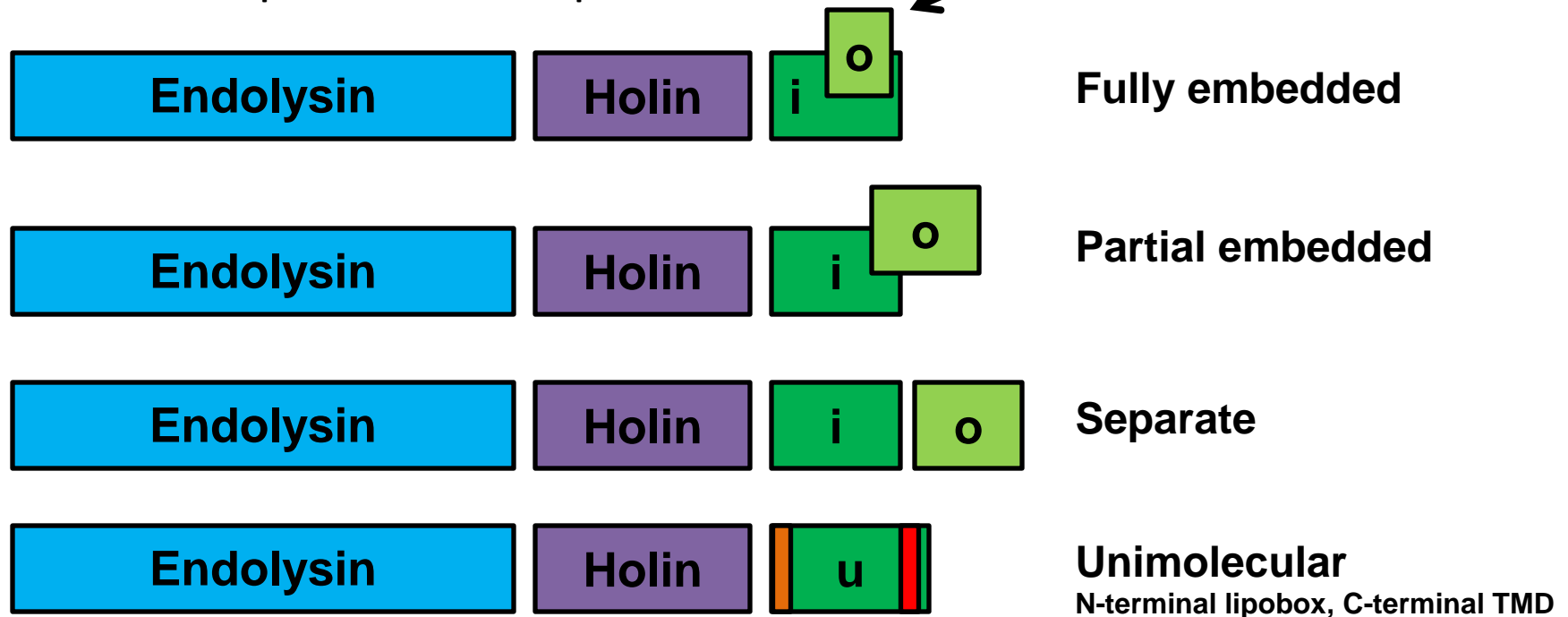


Rajaure et al. 2015, PNAS



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



Holin finding

▼ 2017-03-29 Functional Annotation 14

▼ Blast 4

▼ Nucleotide 1

☐ NT

▼ Protein 3

☐ Canonical Phages

☐ NR

☐ UniRef90

▼ Sequence Analysis 10

▼ Phage 2

☐ Possible Frame Shifts

☐ Possible Intron Locations

▼ Spanin 3

☐ Candidate ISPs

☐ Candidate ISPs and OSPs from BLAST

☐ Candidate OSPs

▼ Structural 5

☐ InterProScan

☐ TMHMM

☐ TMHMM Inside Probability

☐ TMHMM Membrane Probability

☐ TMHMM Outside Probability

- 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

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)

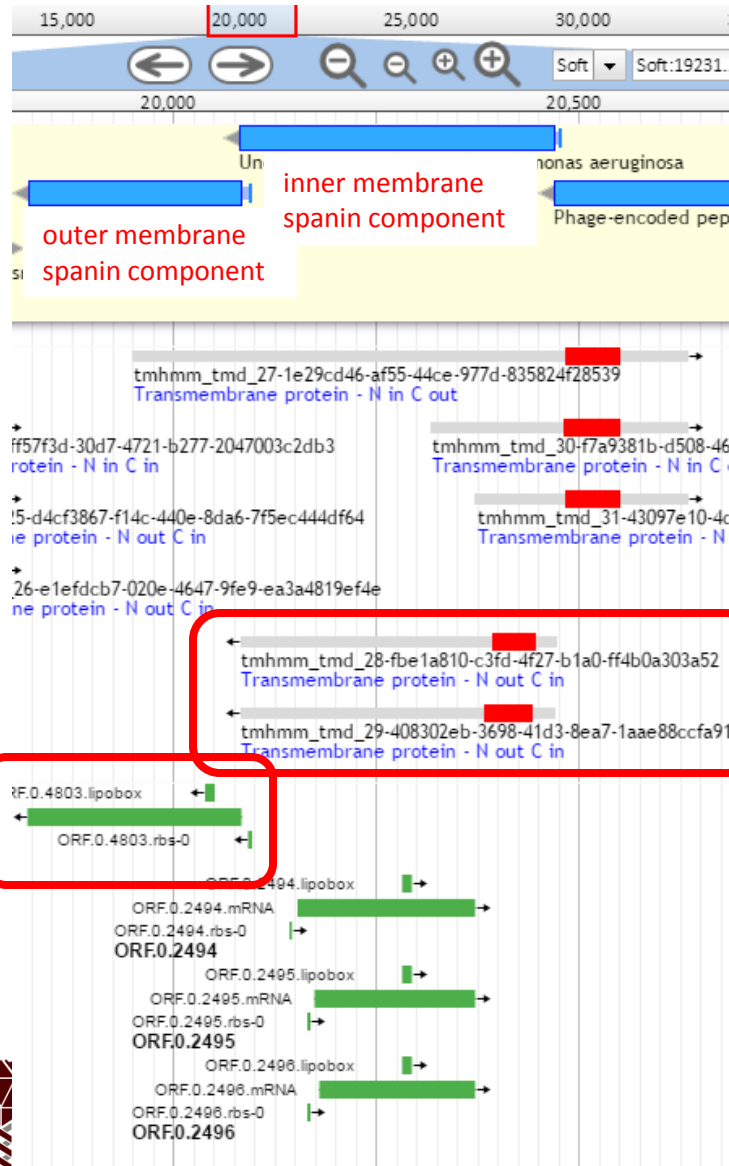
Spanin finding

2017-03-29 Functional Annotation 14

- Blast 4
 - Nucleotide 1
 - ☐ NT
 - Protein 3
 - ☐ Canonical Phages
 - ☐ NR
 - ☐ UniRef90
- Sequence Analysis 10
 - Phage 2
 - ☐ Possible Frame Shifts
 - ☐ Possible Intron Locations
 - Spanin 3
 - ☐ Candidate ISPs
 - ☐ Candidate ISPs and OSPs from BLAST
 - ☐ Candidate OSPs
 - Structural 5
 - ☐ InterProScan
 - ☐ TMHMM
 - ☐ TMHMM Inside Probability
 - ☐ TMHMM Membrane Probability
 - ☐ TMHMM Outside Probability

- 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

Spanin finding

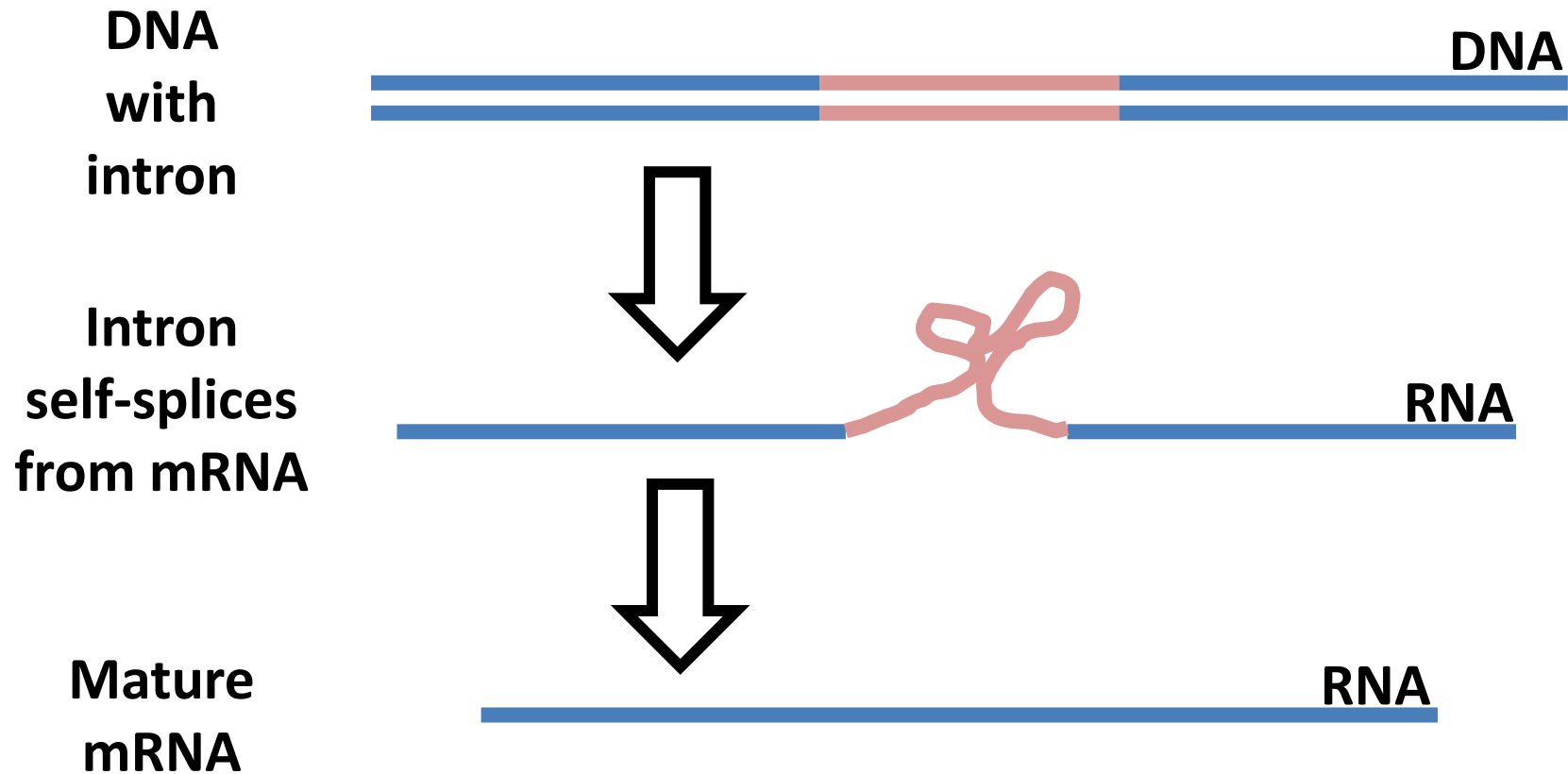


- Likely spanin gene pairs
 - 1 protein with N-terminal TMD (top)
 - 1 protein with N-terminal lipobox (bottom)
 - Adjacent or o-spanin embedded in i-spanin
 - i-spanin is never embedded in o-spanin
- OSP tool conducts naive ORF calls, should find embedded genes not found during structural annotation

Introns

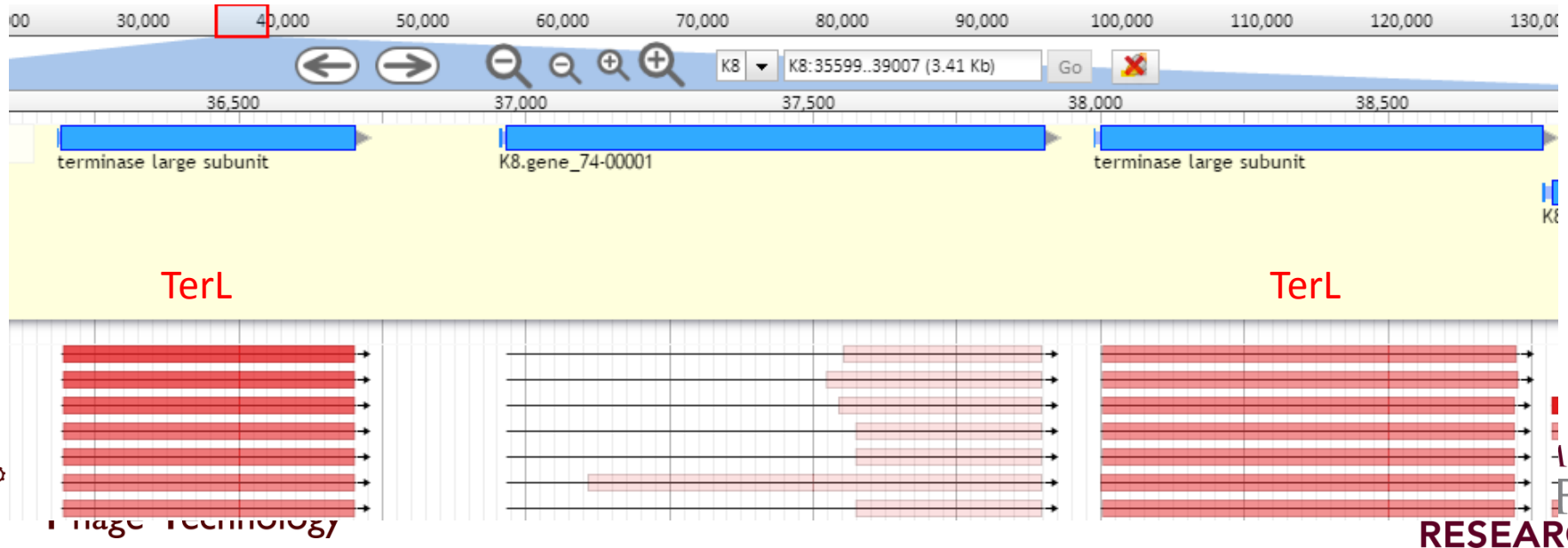
- An intron is an extra section of DNA that interrupts the protein-coding sequence of a gene
- This sequence has *ribozyme* activity and splices itself out of the mRNA, leaving an intact message and a free intron RNA
- Introns often (but not always) contain a homing endonuclease gene
- These are often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

Introns are self-splicing elements

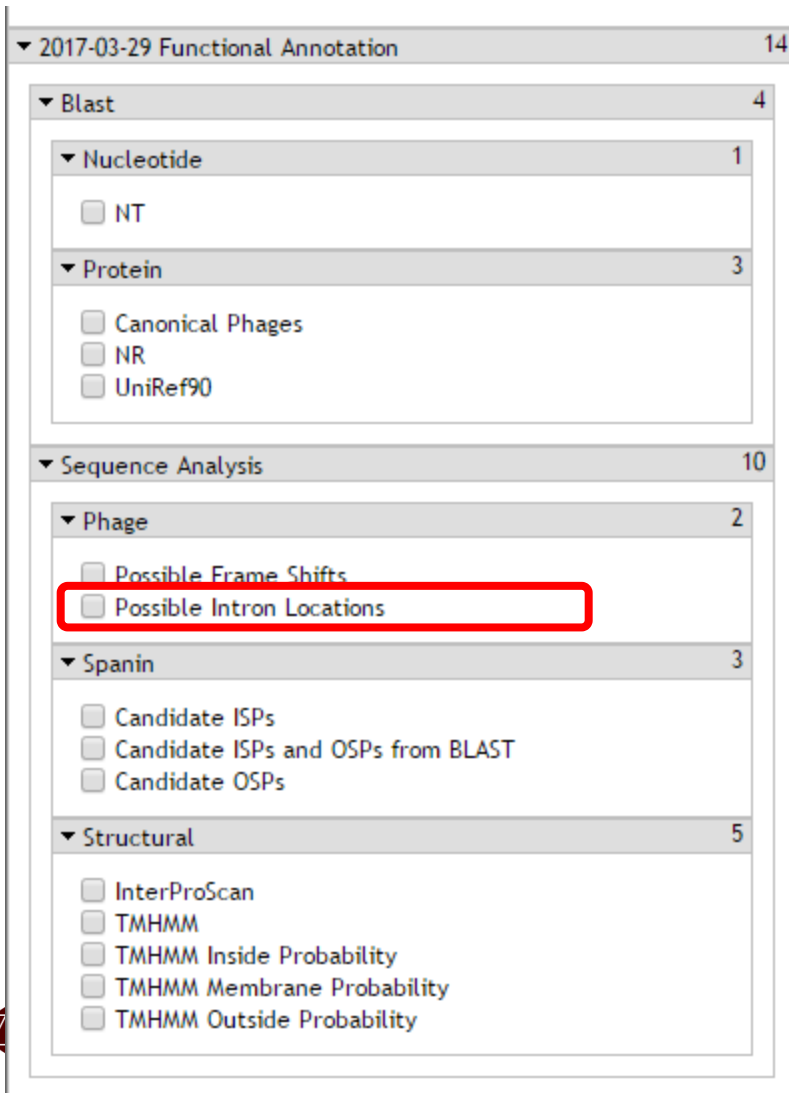


Introns

- 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?

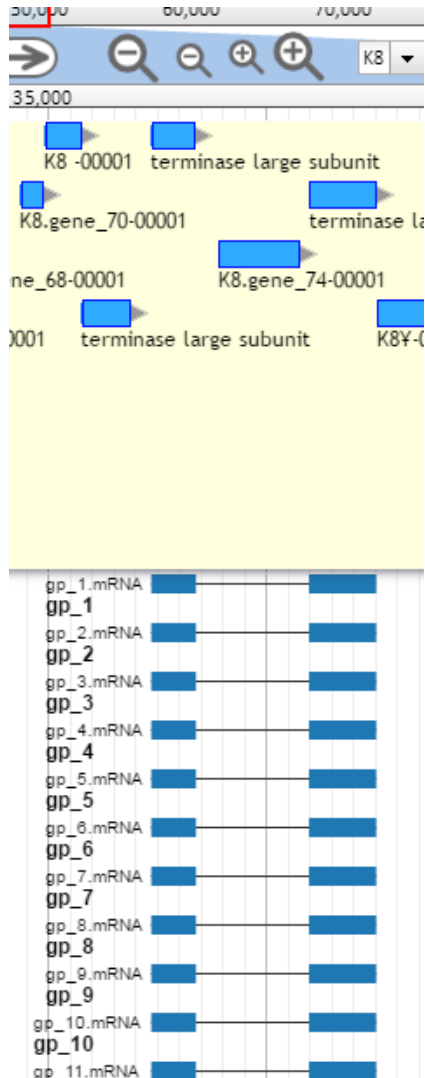


Intron finding track



- 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

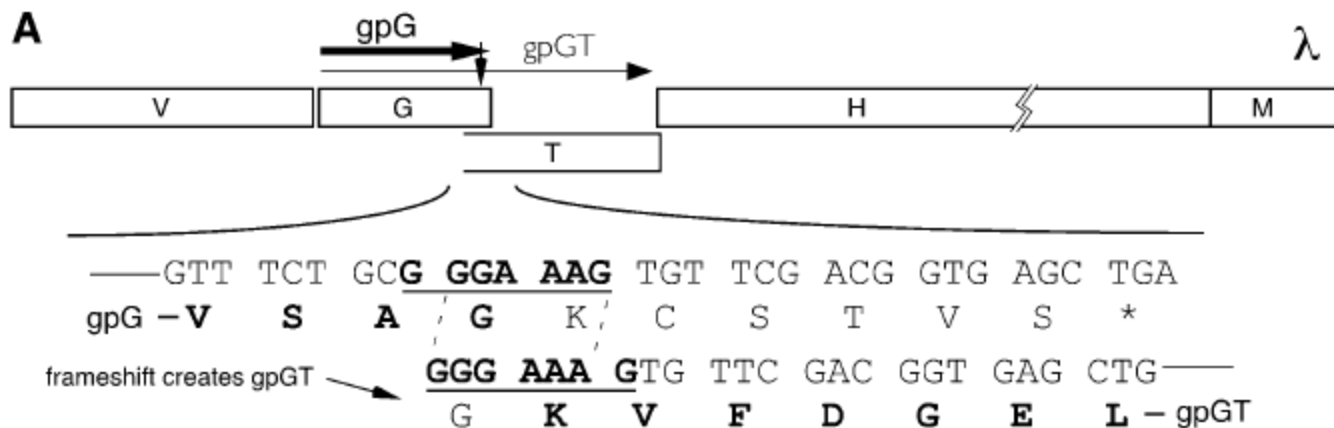
Intron finding



- 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

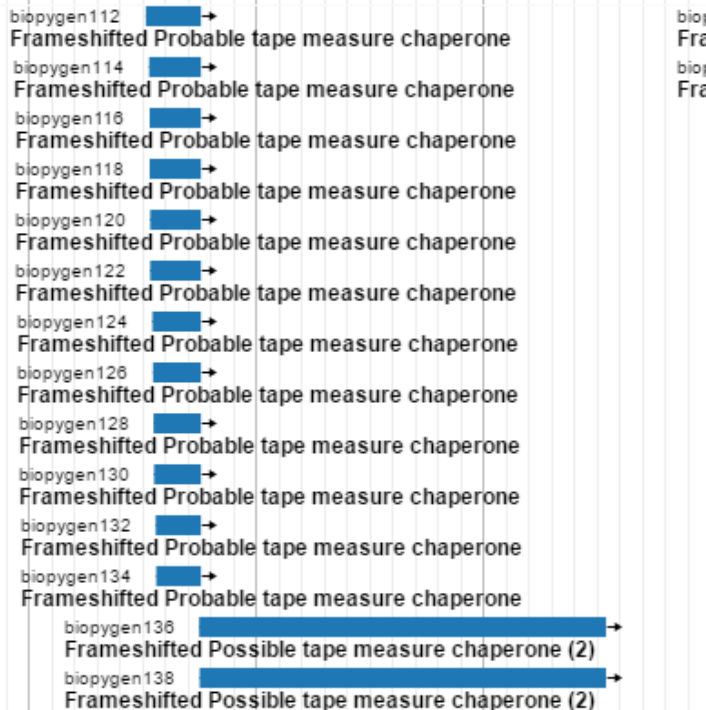
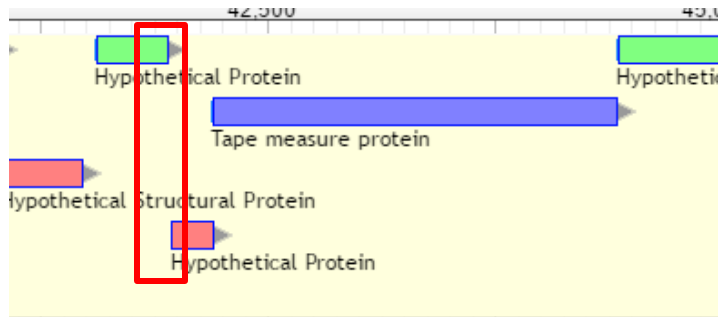


Phage	Slippery Sequence
λ^a	TGTTTCTGCGGGAAAGTGTTCGACG C F C <u>G K</u> V F D → V S A <u>G K</u> C S T
P2 ^a	GGTGGTCGGTTTTTTGTTCGCCGAAC G G R <u>F F</u> V A E → V V G <u>F L</u> S P N
L5 ^a	GACGCAACTGGGGGAAGCCGCGCCC D A T <u>G G</u> S R A → T Q L <u>G E</u> A A P
ϕ C31 ^a	GAACCTGAAGGGGAAGCGAAGCCG E P E <u>G E</u> G E A → N L K <u>G K</u> A K P
TM4 ^a	GCTGATCGAGGGAAAATCTCGCAGG A D R <u>G K</u> I S Q → L I E <u>G K</u> S R R
Mu ^a	ACACAAGAACGGGGGCGAGTGGCTG H K N <u>G G</u> E W L → T Q E R <u>G R</u> V A
HK97 ^a	CGAAGCGCGGGAAAAGTCTCAACCC R S A <u>G K</u> V S T → E A R <u>E K</u> S Q P
HK022 ^a	TAACGATCTGGGAAAGACTTCGAGC * R S <u>G K</u> D F E → N D L <u>G K</u> T S S
SPP1	TAACTCAGGTTTTTTCAAACGAGCT * L R <u>F F</u> Q T S → N S G <u>F F</u> K R A
PBSX	AGAAGCAGCAAAAACTAGTAAAAG R S S <u>K K</u> L V K → E A A <u>K N</u> *

Slippery sequences

- 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

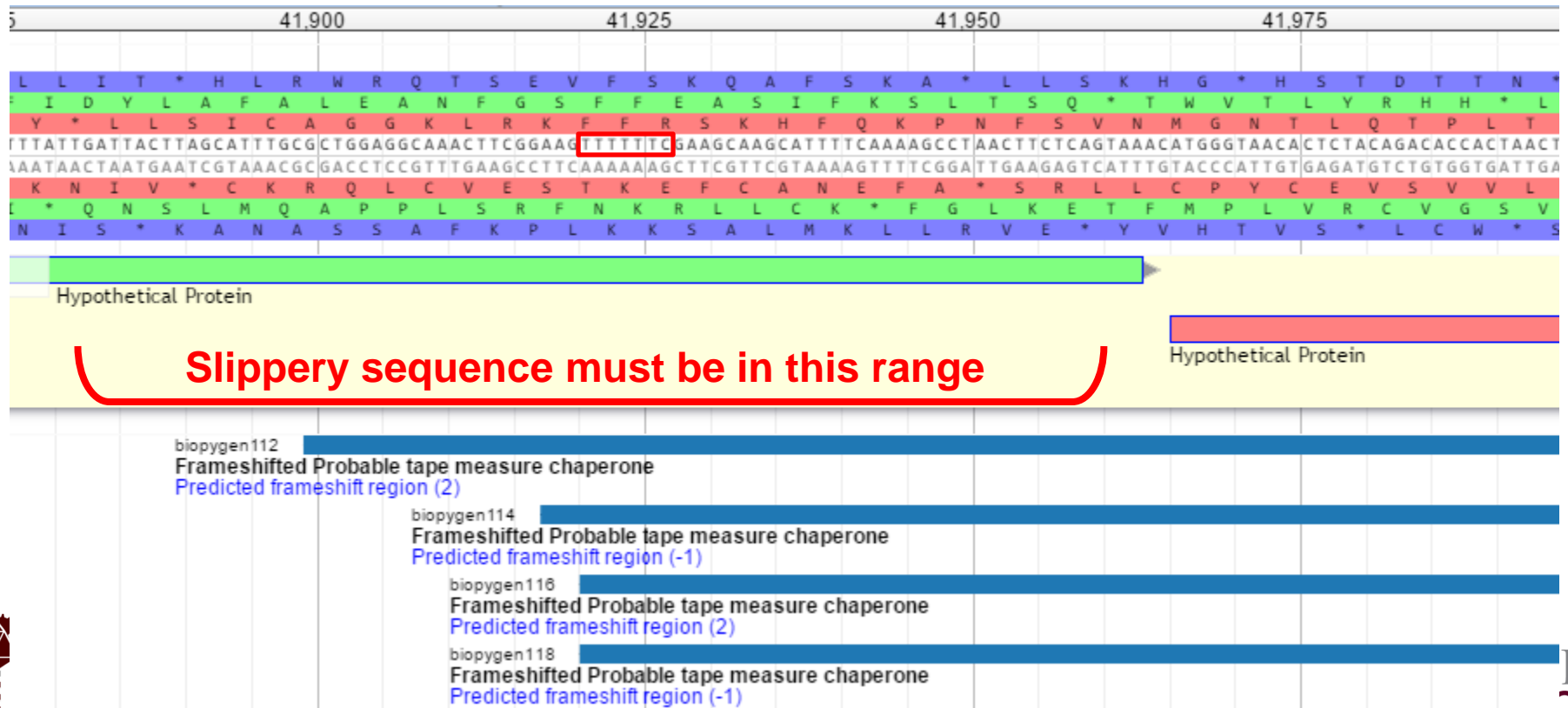
Finding the tape measure chaperones



- 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

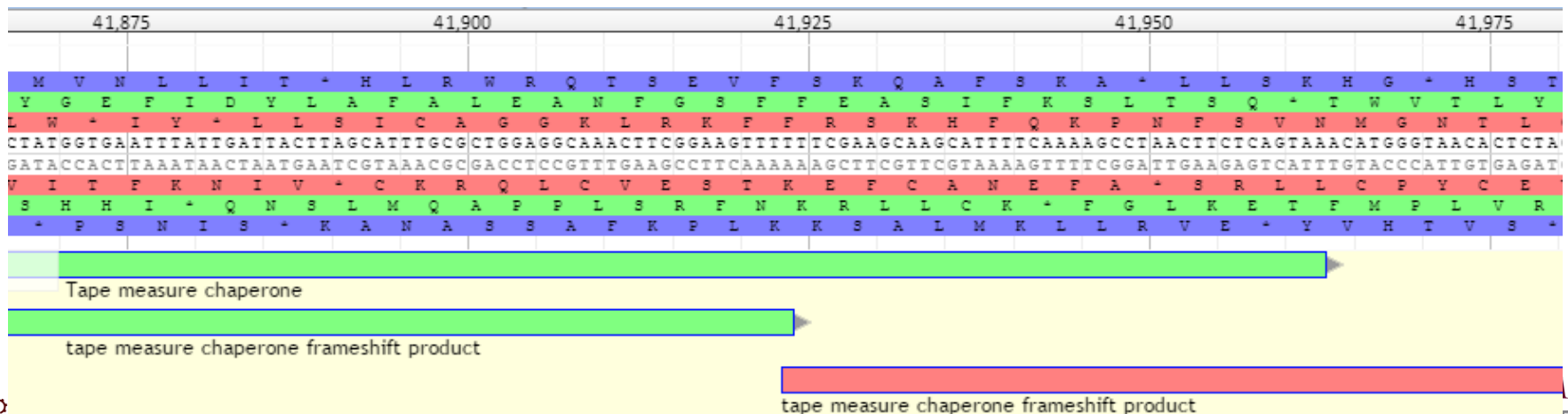
Finding the frameshift location

- 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

