CPT Annotation Infrastructure

Eric Rasche
C527 B0FC
0AF6 3592

Data Analysis
Galaxy
Apollo
Summary
Q&A

CPT Annotation Infrastructure

Eric Rasche
Software Application Developer III

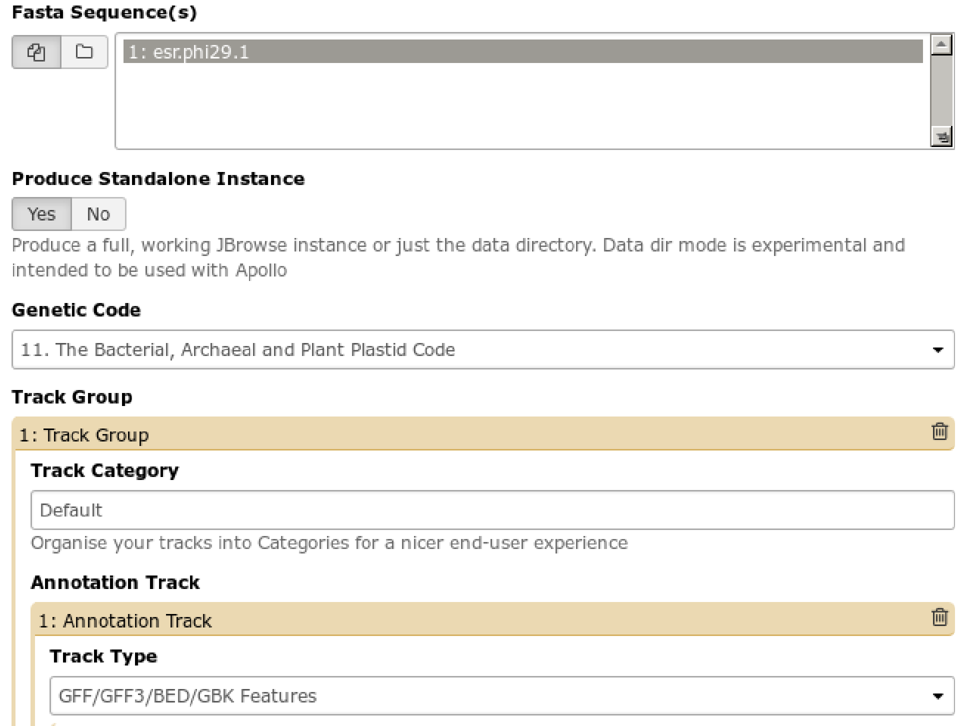
June 9, 2017

Data Analysis for Genome Annotation

- Sequencing Data
- Assembly to Contigs
- Structural Prediction
- Functional Prediction
- Publishing

Galaxy for Reproducible Genomics

- Standard interface to huge variety of tools
- “Histories” (audit logs) for later reference
- Workflows for sharing complex, multi-step analyses
- Collaboration between developers and end users



Fasta Sequence(s)

1: esr.phi29.1

Produce Standalone Instance

[✓ Yes] No

Produce a full, working JBrowse instance or just the data directory. Data dir mode is experimental and intended to be used with Apollo

Genetic Code

11. The Bacterial, Archaeal and Plant Plastid Code

Track Group

1: Track Group

Track Category

Default

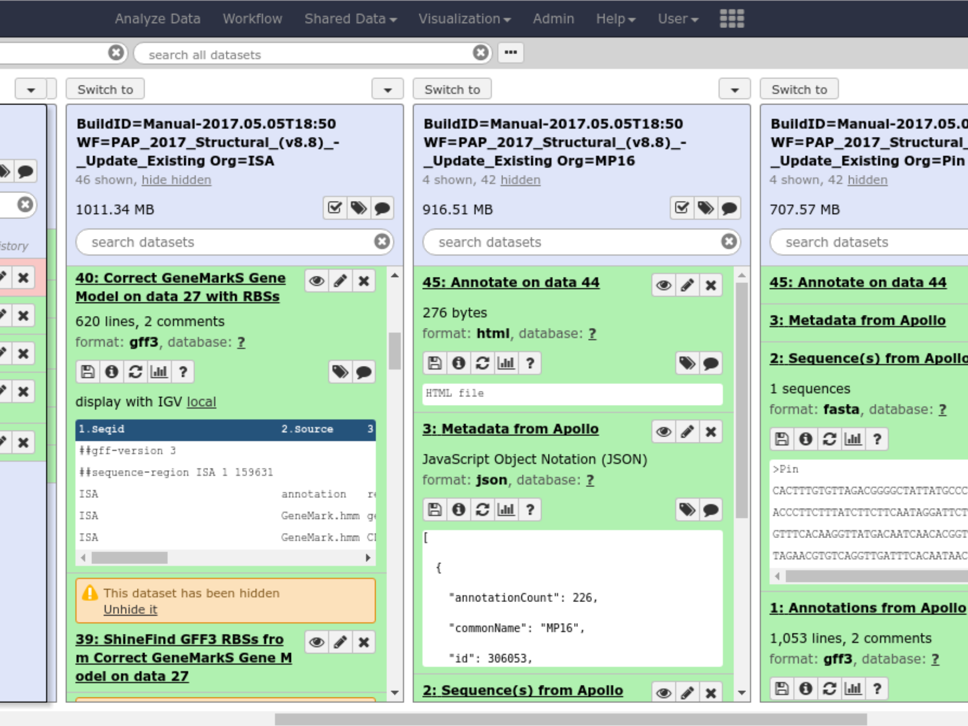
Organise your tracks into Categories for a nicer end-user experience

Annotation Track

1: Annotation Track

Track Type

GFF/GFF3/BED/GBK Features



BuildID=Manual-2017.05.05T18:50 WF=PAP_2017_Structural_(v8.8)_-_Update_Existing Org=ISA

46 shown, hide hidden

1011.34 MB

search datasets

40: Correct GeneMarkS Gene Model on data 27 with RBSs

620 lines, 2 comments

format: **gff3**, database: ?

display with IGV local

```
1.Seqid                    2.Source       3
##gff-version 3
##sequence-region ISA 1 159631
ISA                        annotation     r
ISA                        GeneMark.hmm   g
ISA                        GeneMark.hmm   C
```

This dataset has been hidden
Unhide it

39: ShineFind GFF3 RBSs from Correct GeneMarkS Gene Model on data 27

BuildID=Manual-2017.05.05T18:50 WF=PAP_2017_Structural_(v8.8)_-_Update_Existing Org=MP16

4 shown, 42 hidden

916.51 MB

search datasets

45: Annotate on data 44

276 bytes

format: **html**, database: ?

```
HTML file
```

3: Metadata from Apollo

JavaScript Object Notation (JSON)

format: **json**, database: ?

```
[
  {
    "annotationCount": 226,
    "commonName": "MP16",
    "id": 306053,
```

2: Sequence(s) from Apollo

BuildID=Manual-2017.05.0 WF=PAP_2017_Structural _Update_Existing Org=Pin

4 shown, 42 hidden

707.57 MB

search datasets

45: Annotate on data 44

3: Metadata from Apollo

2: Sequence(s) from Apollo

1 sequences

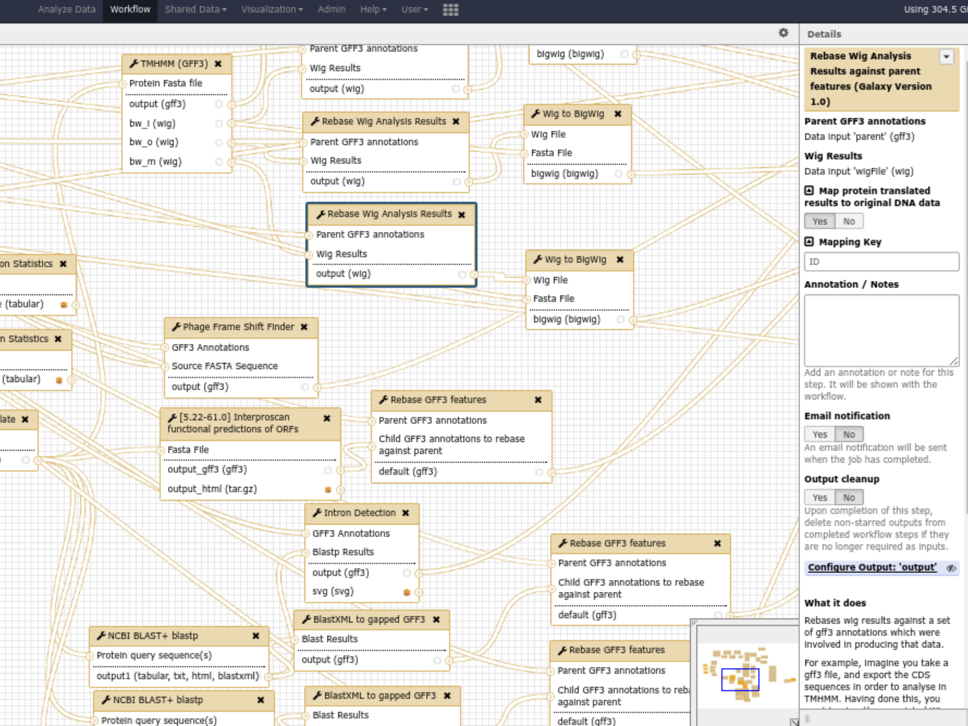
format: **fasta**, database: ?

```
>Pin
CACTTTGTGTTAGACGGGGCTATTATGCCC
ACCCTTCTTTATCTTCTTCAATAGGATTCT
GTTTCACAAGGTTATGACAATCAACACGGT
TAGAACGTGTCAGGTTGATTTCACAATAAC
```

1: Annotations from Apollo

1,053 lines, 2 comments

format: **gff3**, database: ?



Rebase Wig Analysis Results against parent features (Galaxy Version 1.0)

Parent GFF3 annotations

Data input 'parent' (gff3)

Wig Results

Data input 'wigFile' (wig)

Map protein translated results to original DNA data

[✓ Yes] No

Mapping Key

ID

Annotation / Notes

Add an annotation or note for this step. It will be shown with the workflow.

Email notification

Yes [✓ No]

An email notification will be sent when the job has completed.

Output cleanup

Yes [✓ No]

Upon completion of this step, delete non-starred outputs from completed workflow steps if they are no longer required as inputs.

Configure Output: 'output'

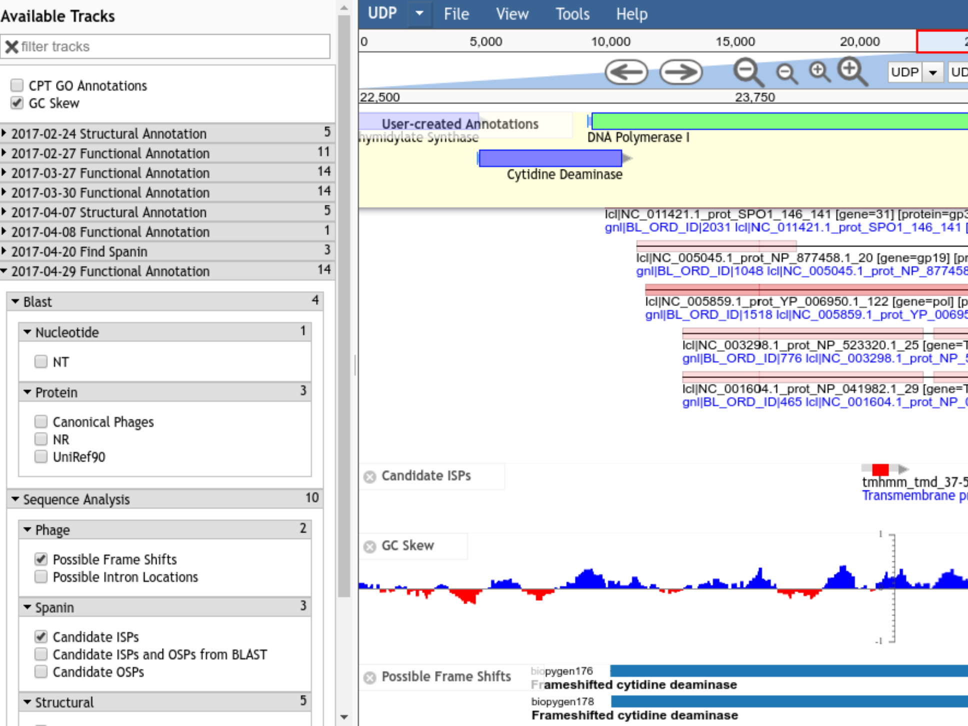
What it does

Rebases wig results against a set of gff3 annotations which were involved in producing that data.

For example, imagine you take a gff3 file, and export the CDS sequences in order to analyse in TMHMM. Having done this, you

Apollo for Interactive Annotation

- “Google Docs” for genome annotation
- Standard interface to analysis data
- Rapidly evolving service with a bright future



Available Tracks

filter tracks

☐ CPT GO Annotations
☑ GC Skew

- 2017-02-24 Structural Annotation 5
- 2017-02-27 Functional Annotation 11
- 2017-03-27 Functional Annotation 14
- 2017-03-30 Functional Annotation 14
- 2017-04-07 Structural Annotation 5
- 2017-04-08 Functional Annotation 1
- 2017-04-20 Find Spanin 3
- 2017-04-29 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

User-created Annotations

hymidylate Synthase

DNA Polymerase I

Cytidine Deaminase

lcl|NC_011421.1_prot_SPO1_146_141 [gene=31] [protein=gp3
gnl|BL_ORD_ID|2031 lcl|NC_011421.1_prot_SPO1_146_141 [

lcl|NC_005045.1_prot_NP_877458.1_20 [gene=gp19] [pr
gnl|BL_ORD_ID|1048 lcl|NC_005045.1_prot_NP_877458

lcl|NC_005859.1_prot_YP_006950.1_122 [gene=pol] [p
gnl|BL_ORD_ID|1518 lcl|NC_005859.1_prot_YP_00695

lcl|NC_003298.1_prot_NP_523320.1_25 [gene=T
gnl|BL_ORD_ID|776 lcl|NC_003298.1_prot_NP_5

lcl|NC_001604.1_prot_NP_041982.1_29 [gene=T
gnl|BL_ORD_ID|465 lcl|NC_001604.1_prot_NP_0

Candidate ISPs

tmhmm_tmd_37-5
Transmembrane pr

GC Skew

Possible Frame Shifts

biopygen176
Frameshifted cytidine deaminase

biopygen178
Frameshifted cytidine deaminase

Phage Genomics with CPT Galaxy & Apollo

- Full-spectrum platform, *sequencing to publishing*
- *Collaboration*, genome annotation and analysis
- *Reproducible* science

Q&A

Thank you

GitHub	github.com/@erasche
Work Email	esr@tamu.edu
GPG Fingerprint	F063 D331 6E63 E7B5 23FD
	B9EA C527 B0FC 0AF6 3592

CENTER FOR PHAGE TECHNOLOGY
TEXAS A&M UNIVERSITY

Center for PHAGE TECHNOLOGY

This material is based upon work supported by the National Science Foundation under Grant Number (Award 1565146)