GGA: Galaxy for Genome Annotation, Teaching, and Genomic Databases

Eric Rasche, Björn Grüning, Nathan Dunn, Anthony Bretaudeau

2017-06-29T09:50:00Z

Galaxy for Genome Annotation

Galaxy is great for:

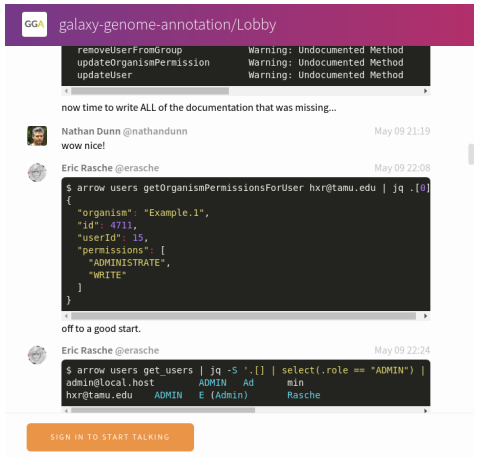
- NGS Analysis
- Assembly
- ...
- Annotation Analysis (Tabular processing, etc)
- *omics

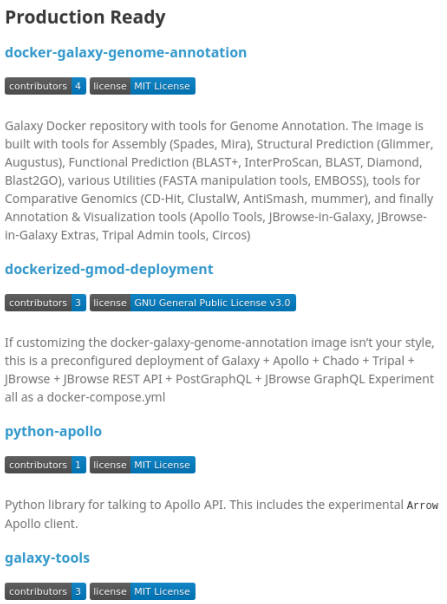
But we are missing the Annotation step.

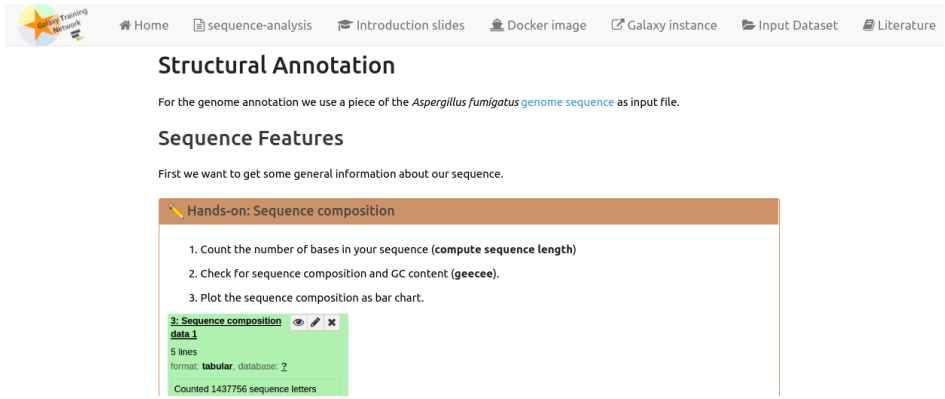
We are missing the tooling, the trainings, and the community for genome annotation.

What are we building?

- Galaxy Flavour(s)
- GMOD Containers
- Glue Code
- Training Materials
- Peripherals
- Community

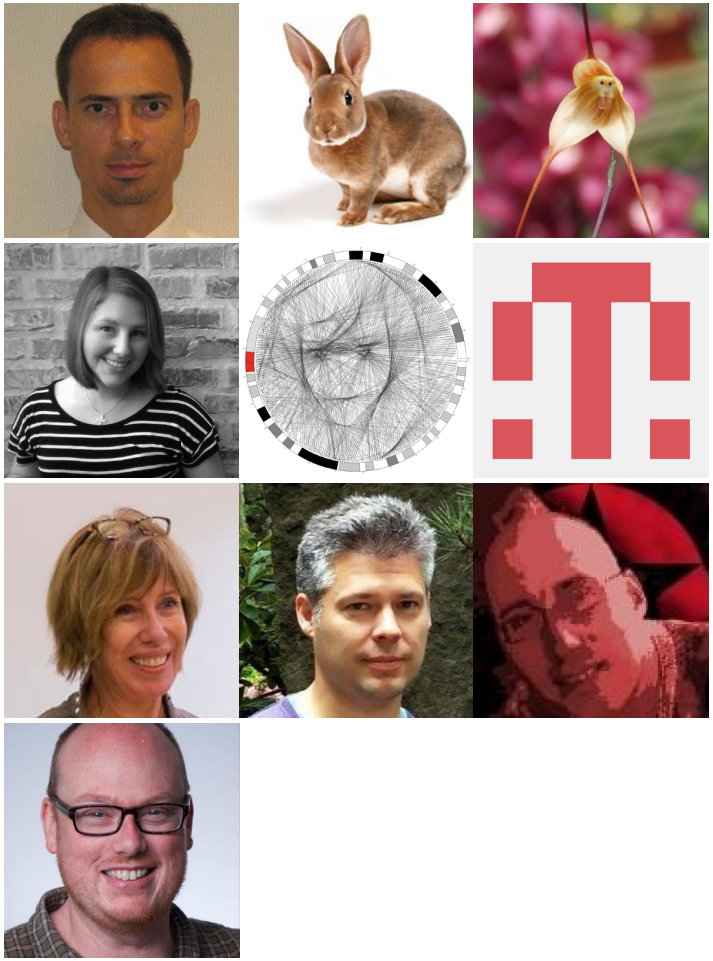






Who are we?

- CPT Phage Team (Eric Rasche, Eleni Mijalis, Cory Maughmer)
- Anthony Bretaudeau
- Nathan Dunn
- Björn Grüning
- Peter van Heusen
- Suzanna Lewis
- Eduardo de Paiva Alves
- Torsten Seemann
- (. . . you!)



Why?

- GMOD at its best
- Annotation still requires humans
- Powerful Analysis + Interactive Annotations
- Useful to real-life people, solve real problems
- Project longevity

#InfrastructureGoals

Launch Galaxy, Apollo, Chado, Tripal, ...

Duplicate this Galaxy

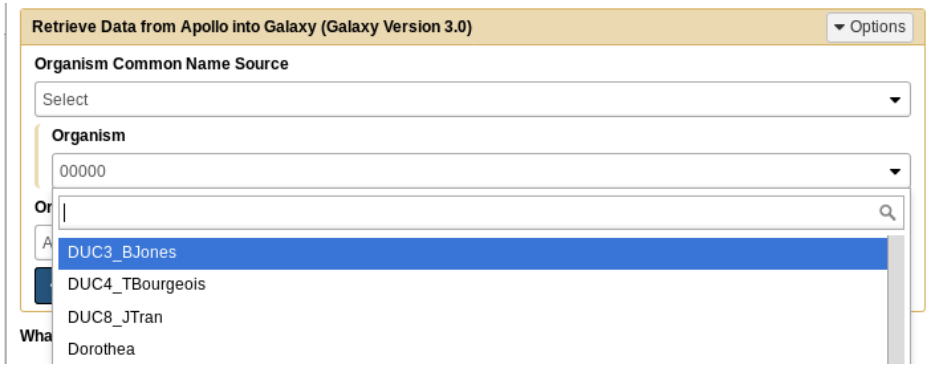
Customize Deployment

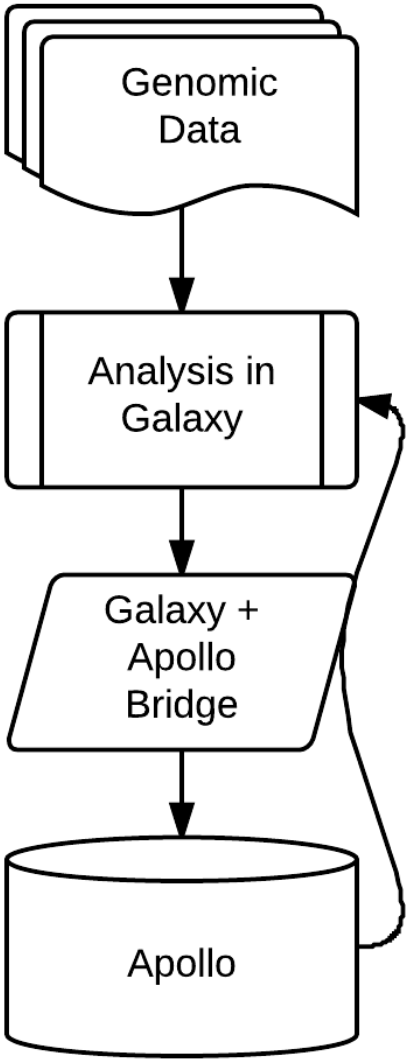
Infrastructure Solutions

- Docker Image: Galaxy + Annotation Tools (Apollo Tools, Tripal Admin Tools, Circos, JBrowse, BLAST+, InterProScan, Glimmer, Augustus, FASTA manipulation tools, Spades, Mira, CD-Hit, ClustalW, AntiSmash, mummer, EMBOSS, BLAST, Diamond, Blast2GO, . . .)
- Dockerized GMOD Deployment (Galaxy, JBrowse, Apollo, Chado, Chado APIs, Tripal pre-configured to work together seamlessly)
- Apollo, Chado python libraries (+parsec like tools, "Arrow" and "Chakin")
- Various Apollo support projects (git-backup, experimental google docs integration)

Galaxy / Apollo Bridge

- Initially quite simple, only a tool to add an organism (JBrowse instance) to Apollo
- Now includes automation (tools for creating/editing annotations)
- Tested and revised in collaboration with curators





Genomic Databases & Curators

- Annotator Independence & Agency
- Democratization of resources
- Enabled them to build powerful annotation and analysis pipelines

Reproducibility

- Reproducibility for generally unreproducible external databases

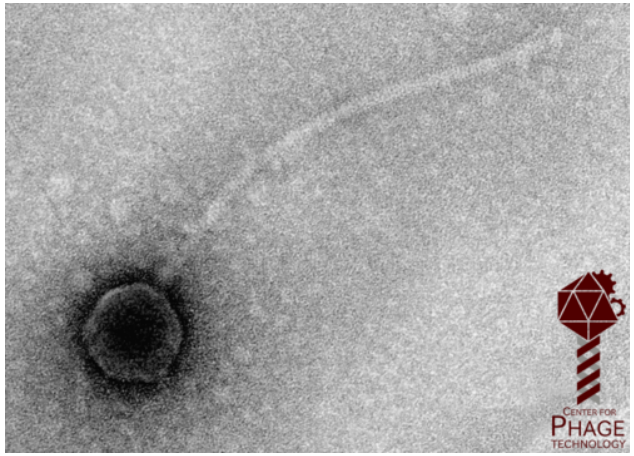
$$\text{Database} = \mathbf{f}_{\text{publication}}(\mathbf{g}_{\text{functional}}(\mathbf{h}_{\text{structural}}(\text{data})))$$

Tooling for Curators

- Tools for querying annotation resources, answering specific questions. (E.g. Find features with specific qualifier, GO terms)
- Tools for fetching data into Galaxy (Chado, Apollo)
- Tools for creating new annotations from analysis results

Bacteriophage Annotation Course

- Undergraduate Phage annotation course
- Genome sequence to publication
- Parallel track for environmental sample to isolated phage
- Novel genome, de novo annotation



genomeA
Journals.ASM.org

CrossMark
click for updates

Complete Genome Sequence of *Klebsiella pneumoniae* Carbapenemase-Producing *K. pneumoniae* Myophage Miro

Eleni M. Mijalis, Lauren E. Lessor, Jesse L. Cahill, Eric S. Rasche, Gabriel F. Kuty Everett

Center for Phage Technology, Texas A&M University, College Station, Texas, USA

Klebsiella pneumoniae **is a Gram-negative pathogen frequently associated with antibiotic-resistant nosocomial infections. Bacteriophage therapy against** ***K. pneumoniae*** **may be possible to combat these infections. The following describes the complete genome sequence and key features of the pseudo-T-even** ***K. pneumoniae*** **carbapenemase (KPC)-producing** ***K. pneumoniae*** **myophage Miro.**

Received 17 August 2015 **Accepted** 19 August 2015 **Published** 1 October 2015

Citation Mijalis EM, Lessor LE, Cahill JL, Rasche ES, Kuty Everett GF. 2015. Complete genome sequence of *Klebsiella pneumoniae* carbapenemase-producing *K. pneumoniae* myophage Miro. Genome Announc 3(5):e01137-15. doi:10.1128/genomeA.01137-15.

Address correspondence to Gabriel F. Kuty Everett, cpt@tamu.edu.

Klebsiella pneumoniae carbapenemase (KPC)-producing *K. pneumoniae* is a highly drug-resistant bacterium in the family *Enterobacteriaceae*. It can easily be spread in hospital settings, provoking deadly systemic infections (1, 2). This gives credibility to the prospect of bacteriophage-based therapy against the pathogen. Here, we describe the complete genome of pseudo-T-even myophage Miro.

opened to the *rIIb* gene, whose start codon overlaps with the stop codon of *rIIa*, such that the two genes cannot be separated, a common feature of pseudo-T-even phages (11). Miro is closely related to *Klebsiella* phage KP15 (accession no. NC_014036), with which it shares 94.5% nucleotide sequence identity across the genome. It also shares 92.9% nucleotide sequence identity across the genome with *Klebsiella* phage KP27 (accession no. NC_020080),

Current Applications of GGA

Good for Community and Us

- CPT can leverage the GGA infrastructure
- GGA tools help bring community best practices to the CPT
- CPT can contribute back well tested workflows

From students to experienced curators:

- Collaboration is easy
- High level: Apollo is the "Google Docs" of genome annotation
- Low level: Optimised genomic database queries as tools

GGA Going Forward

- Sharing of genomic database querying tools
- More tutorials / training resources
- More GMOD projects

Q&A

Thank you and join us at:

GGA GitHub	galaxy-genome-annotation.github.io
GGA Gitter	gitter.im/galaxy-genome-annotation/Lobby

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