

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

Eric Rasche, Björn Grüning, Nathan Dunn, Anthony Bretaudeau

2017-06-29T13:50:00Z



Galaxy for Genome Annotation

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who
What
Why

Annotation

Solutions
Galaxy-Apollo

Teaching

Current
Expansion

Databases

Tooling

Q&A

Galaxy is great for:

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

But we are missing the Annotation step.

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

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



What are we building?

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA
Who
What
Why

Annotation
Solutions
Galaxy-Apollo

Teaching
Current
Expansion

Databases
Tooling

Q&A

- **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 for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who

What

Why

Annotation

Solutions

Galaxy-Apollo

Teaching

Current

Expansion

Databases

Tooling

Q&A

- Fun Challenges
- Useful to real-life people, solve real problems
- Project longevity



Annotation Infrastructure

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who
What
Why

Annotation

Solutions
Galaxy-Apollo

Teaching

Current
Expansion

Databases

Tooling

Q&A

- Easy to deploy
- Shareable
- Configurable, pick-and-choose subcomponents



Annotation Solutions

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who
What
Why

Annotation
Solutions
Galaxy-Apollo

Teaching
Current
Expansion

Databases
Tooling

Q&A

- Galaxy Images: Suites of annotation tools pre-installed in docker images
- GMOD Containers: Containers for the big projects
- Glue Code: Easy to deploy connections to annotation backends (e.g. Apollo)
- `docker-compose.yml`: describe complete infrastructure stack
- Peripherals: parsec, arrow, chakin



Galaxy / Apollo Bridge

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who
What
Why

Annotation

Solutions
Galaxy-Apollo

Teaching

Current
Expansion

Databases

Tooling

Q&A

- Grew out of use in an undergraduate course
- Initially quite simple, tool to add organism (JBrowse instance) to Apollo
- Expanded to include automation, tools for creating & removing “human-created” annotations.
- Jenkins job now re-runs functional annotation workflows regularly
- Improved with student help, finding unexpected bugs



Bacteriophage Annotation Course

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who
What
Why

Annotation

Solutions
Galaxy-Apollo

Teaching

Current
Expansion

Databases

Tooling

Q&A

- CPT runs a GGA deployment, Galaxy+Apollo+Chado
- Students run tools in Galaxy, annotate in Apollo, “published” data is stored in Chado
- Converging implementations; updates to our phage course + GGA suite



Current Applications of GGA

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who
What
Why

Annotation

Solutions
Galaxy-Apollo

Teaching

Current
Expansion

Databases

Tooling

Q&A

- Undergraduate Phage annotation course
- Genome sequence to publication
- (Parallel track for environmental sample to isolated phage)
- Gigantic workflows using many of the tools pre-installed in image

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who
What
Why

Annotation

Solutions
Galaxy-Apollo

Teaching

Current
Expansion

Databases

Tooling

Q&A

CPT's use case:

- Adding on Tripal + JBrowse-based public sequence display
- Building a sample tracking database, can leverage GGA Apollo / Chado, and Galaxy integrations

GGA:

- More tutorials / training resources
- More GMOD projects



Genomic Databases

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who
What
Why

Annotation
Solutions
Galaxy-Apollo

Teaching
Current
Expansion

Databases

Tooling

Q&A

- Galaxy tools for Genomic Access
- Reproducibility for generally unreproducible external databases
- Democratization of resources
- Galaxy Tools handle CRUD operations & updating permissions



Genomic Databases

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who
What
Why

Annotation

Solutions
Galaxy-Apollo

Teaching

Current
Expansion

Databases

Tooling

Q&A

- Tools for querying annotation resources, answering specific questions.
- Tools for fetching remote data into Galaxy
- Galaxy is used for Analysis
- Tools for creating new annotations from analysis results

Galaxy for
Genome
Annotation,
Teaching,
Databases

ER, BG, ND,
AB

GGA

Who
What
Why

Annotation
Solutions
Galaxy-Apollo

Teaching

Current
Expansion

Databases

Tooling

Q&A

Thank you

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