Connecting Galaxy with the NIH Sequence Read Archive (SRA)

Wednesday, June 24

Marius van den Beek
Daniel Blankenberg
Dave Clements

@galaxyproject #UseGalaxy

bit.ly/galaxy-sra-slides

Agenda

- SRA?
- Galaxy?
- **SRA + Galaxy!**
  - A live demo

Please ask questions using the Zoom Q&A window, as we go.

“Is there anything you would like to specifically learn about in this webinar?”

Today:

- How to import SRA fastq files to galaxy online
- Benefits of the Galaxy/NCBI partnership!
- SRA data integration in Galaxy!
- how to fetch multiple SRA data sets to perform a bioinformatic analysis in the Galaxy platform
- how to import SRA fastq files to galaxy online
- are there are limits to how many datasets can be imported at once?

Not Today:

- assess QC metrics before analyses
- Using all features in Galaxy
- Expression analysis
- Bacterial whole sequences submission
- Submission of RNA seq files (transcriptome) data to sra database.
- BLAST SRA

Um, maybe?

- The meaning of life

Sequence Read Archive (SRA)

- Poll
- SRA is NIH's primary archive of *unassembled reads*
- SRA is a great place to get the sequencing data that underlie publications and studies
  - All of SRA now on AWS, GCP clouds

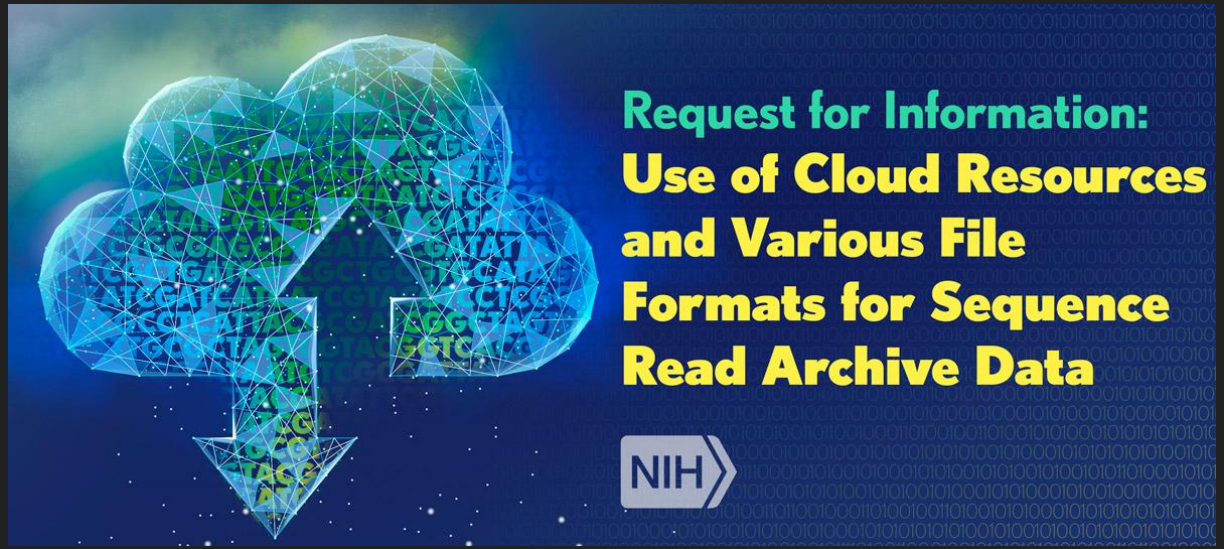
You will also hear it referred to as the *Short Read Archive*, its former name.

https://www.ncbi.nlm.nih.gov/sra

@NCBI

Entrez and SRA Run Selector

- Two interfaces to SRA data that complement each other
- Today you will see both.



NIH has released a request for information (RFI) to solicit community feedback on a **proposed Sequence Read Archive (SRA) data formats**.

Learn more and share your thoughts at https://go.usa.gov/xvhdr.

The response deadline is July 17th, 2020. We encourage you all to share with your colleagues and networks, and respond if you are an SRA submitter or data user.

Galaxy

- Poll
- A data integration and analysis platform for life sciences data
- A worldwide community of users, trainers, developers, infrastructure providers, tool developers, and software engineers

https://galaxyproject.org/

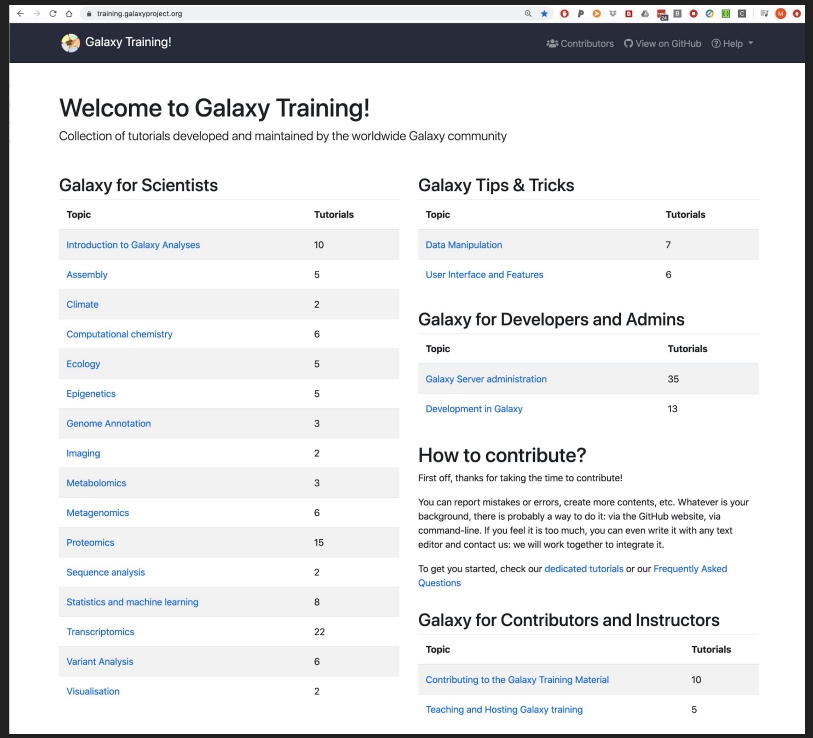
Galaxy is available

- At over 100 free, online web servers
- On commercial and academic clouds
- In containers and virtual machines
- As open source software that can be installed anywhere

https://galaxyproject.org/use/ https://getgalaxy.org/

Galaxy training materials

- Galaxy is used by scientists from many domains
- Detailed tutorials and workflows available
- Everyone can contribute



training.galaxyproject.org

Galaxy Training!

Contributors | View on GitHub | Help

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy Analyses	10
Assembly	5
Climate	2
Computational chemistry	6
Ecology	5
Epigenetics	5
Genome Annotation	3
Imaging	2
Metabolomics	3
Metagenomics	6
Proteomics	15
Sequence analysis	2
Statistics and machine learning	8
Transcriptomics	22
Variant Analysis	6
Visualisation	2

Galaxy Tips & Tricks

Topic	Tutorials
Data Manipulation	7
User Interface and Features	6

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	35
Development in Galaxy	13

How to contribute?

First off, thanks for taking the time to contribute!

You can report mistakes or errors, create more contents, etc. Whatever is your background, there is probably a way to do it: via the GitHub website, via command-line. If you feel it is too much, you can even write it with any text editor and contact us: we will work together to integrate it.

To get you started, check our dedicated tutorials or our Frequently Asked Questions

Galaxy for Contributors and Instructors

Topic	Tutorials
Contributing to the Galaxy Training Material	10
Teaching and Hosting Galaxy training	5

https://training.galaxyproject.org/

SRA + Galaxy: A live demo

- Our experiment
  - COVID-19 datasets
  - *But*, our domain does not actually matter
  - Today we are focused on the integration and this integration can be used with SRA data in any domain
- The plan
  - Go from Galaxy to SRA to Galaxy to get sequence metadata, including SRA accession numbers
  - Get the sequence data from SRA
  - Run a short analysis in Galaxy using the SRA data

usegalaxy.org bit.ly/galaxy-sra-tutorial

Some caveats

- Submitters often do not provide complete/correct metadata
- There is a discrepancy between SRR and ERR entries
- In some cases downloads fail

https://bit.ly/galaxy-sra-history

SRA Resources

Questions? Contact the NCBI team at sra@ncbi.nlm.nih.gov

Additional resources

- https://www.ncbi.nlm.nih.gov/sra
- https://www.ncbi.nlm.nih.gov/sars-cov-2/

Submitting data?

- https://submit.ncbi.nlm.nih.gov/

Galaxy Resources

- galaxyproject.org/
- help.galaxyproject.org/
- gitter.im/galaxyproject
- usegalaxy.{org|eu|org.au}
- bcc2020.github.io

Thank you!

NCBI
Yuriy Skripchenko
Lydia Fleischmann
Ravinder Eskandary
Kurt McDaniel
Sergiy Ponomarov

NIAID

NHGRI

NSF

Galaxy Community