

Community-Driven Training for Biological Data Analysis with the Galaxy Training Network

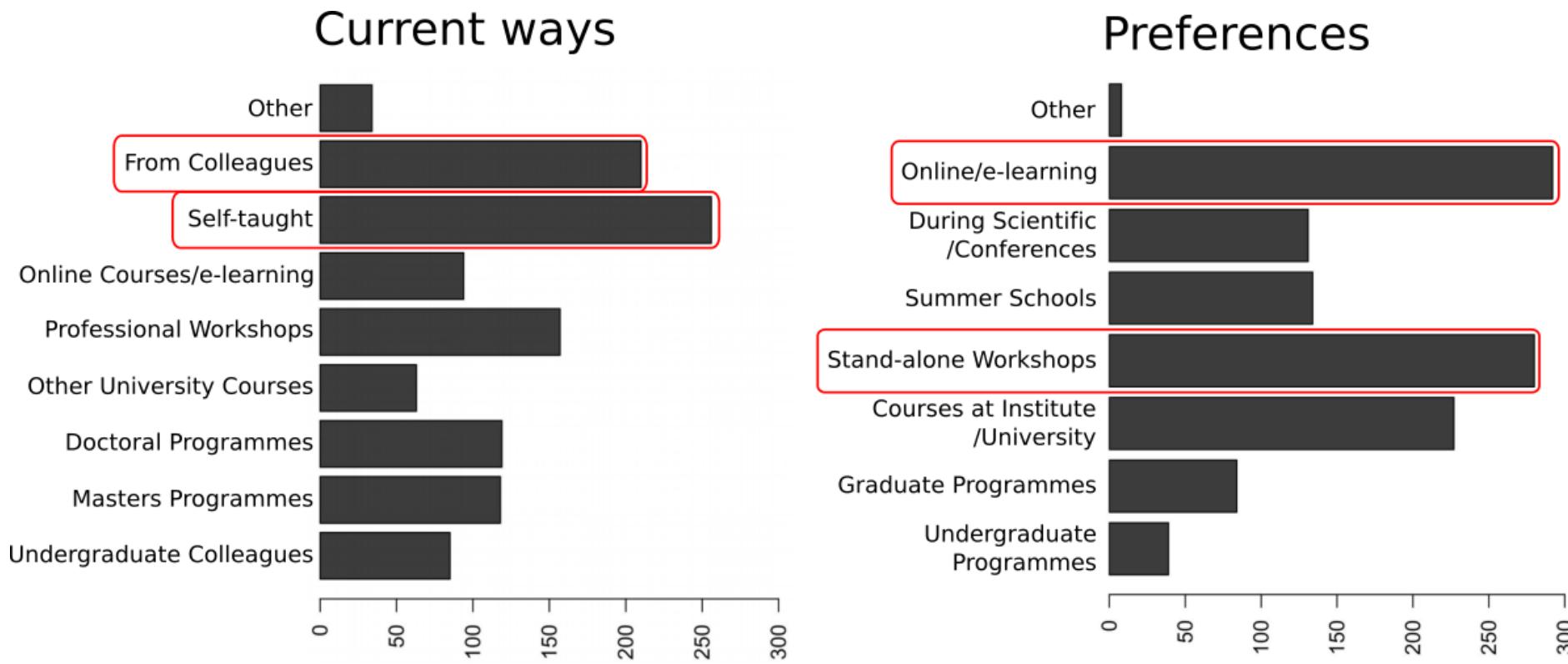


Bérénice Batut

CarpentryCon - May 2018

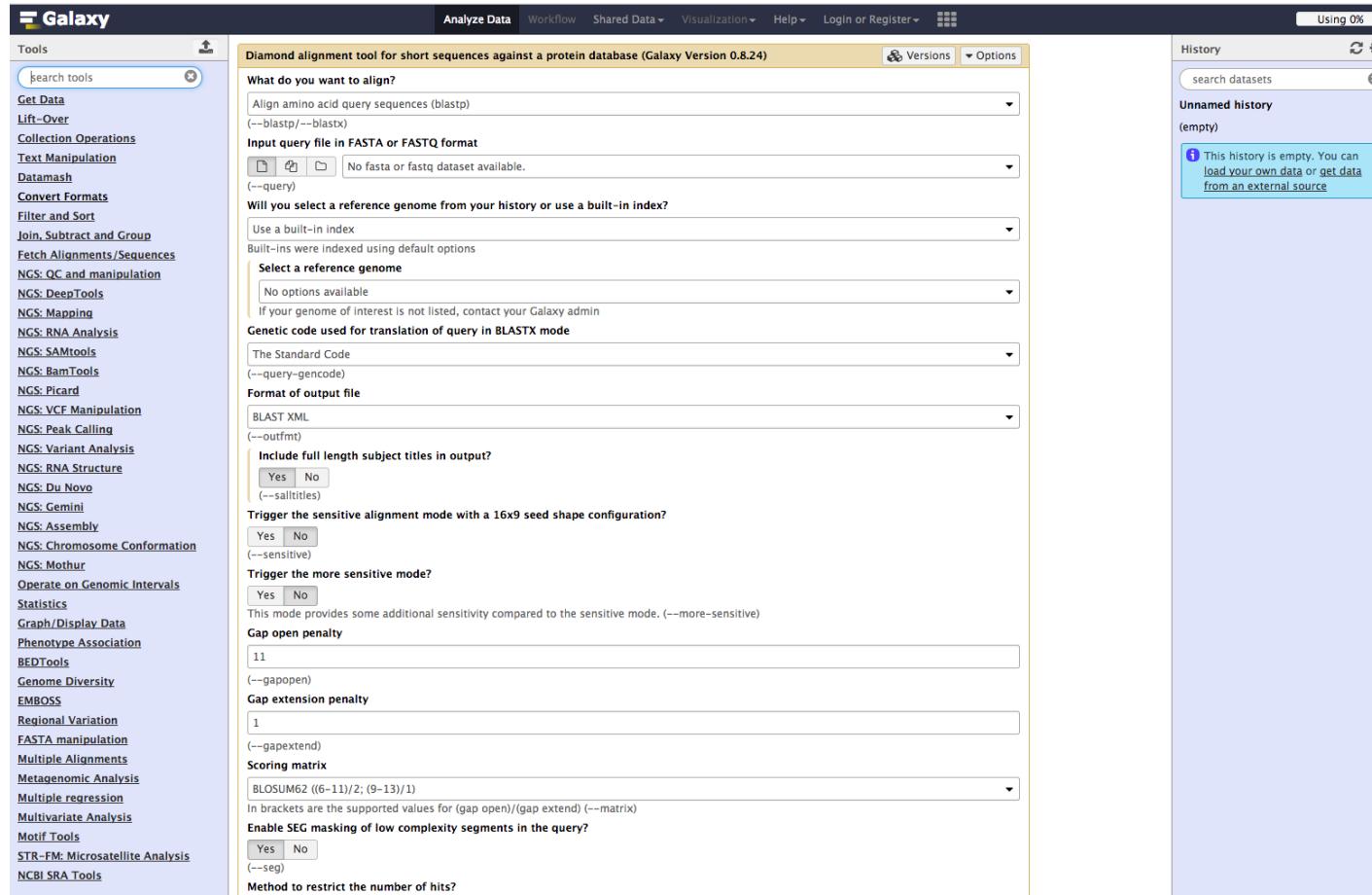
Need and demand for bioinformatic training

*Bioinformatics has become too central to biology
to be left to specialist bioinformaticians*



Graphs of Brazas et al, 2017

Galaxy: a great solution !



- Web interface for numerous bioinformatics tools
- No issue with computer configuration during training

Requirements for an effective training infrastructure

- Interactive learning platform
- Support for current research problems
- Usable for effective training for individual users & instructors
- Community driven (content creation and maintenance)
- FAIR (Findable, Accessible, Interoperable, Reusable) and Open

Interactive learning via hands-on tutorials built around a "research story"

The image displays two side-by-side screenshots of a Galaxy web interface. The left screenshot shows a 'Hands-on: Quality control' tutorial with steps 1 through 5. Step 1 details how to run FastQC on FASTQ files. Step 2 instructs inspecting a generated webpage for sample 'GSM461177_1'. Step 3 covers MultiQC aggregation. Step 4 involves inspecting MultiQC output. Step 5 discusses Trim Galore! treatment. A 'Tip' box states: 'You can select several files by keeping the CTRL (or COMMAND) key pressed and clicking on the interesting files'. A 'Questions' section asks 'What is the read length?' with an option to 'Click to view answers'. The right screenshot shows the 'FastQC Read Quality reports (Galaxy)' tool details. It includes sections for 'Short read data from your current history' (with a dropdown menu for file types like fastq, fastq.gz, etc.), 'Contaminant list' (with a dropdown menu for 'Nothing selected'), and 'Submodule and Limit specifying file' (with a dropdown menu for 'Nothing selected'). A large 'Execute' button is present. Below these are sections for 'Purpose' (describing FastQC's aim to provide quality control checks) and 'FastQC' (noting it's a Galaxy wrapper for the FastQC package). The Galaxy header at the top shows the title 'Galaxy' and various navigation tabs.

Hands-on: Quality control

1. **FastQC**: Run FastQC on the FASTQ files to control the quality of the reads
 - "Short read data from your current history"
 - Click on "Multiple datasets"
 - Select all raw datasets
2. Inspect on the generated webpage for [GSM461177_1](#) sample
3. **MultiQC**: Aggregate the FastQC reports with
 - "Which tool was used generate logs?" to [FastQC](#)
 - "Type of FastQC output?" to [Raw data](#)
 - "FastQC output" to the generated [Raw data](#) files (multiple datasets)
4. Inspect the webpage output from MultiQC
5. **Trim Galore!**: Treat for the quality of sequences by running Trim Galore! with
 - "Is this library paired- or single-end?" to [Paired-end](#)

Tip
You can select several files by keeping the CTRL (or COMMAND) key pressed and clicking on the interesting files

Questions
What is the read length?
▶ Click to view answers

FastQC Read Quality reports (Galaxy)
Version 0.69

Short read data from your current history
No fastq, fastq.gz, fastq.bz2, bam or sam...
Manipulate FASTQ reads on various attributes
Combine FASTA and QUAL into FASTQ
FastQC Read Quality reports

Workflows
All workflows

Submodule and Limit specifying file
Nothing selected
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Execute

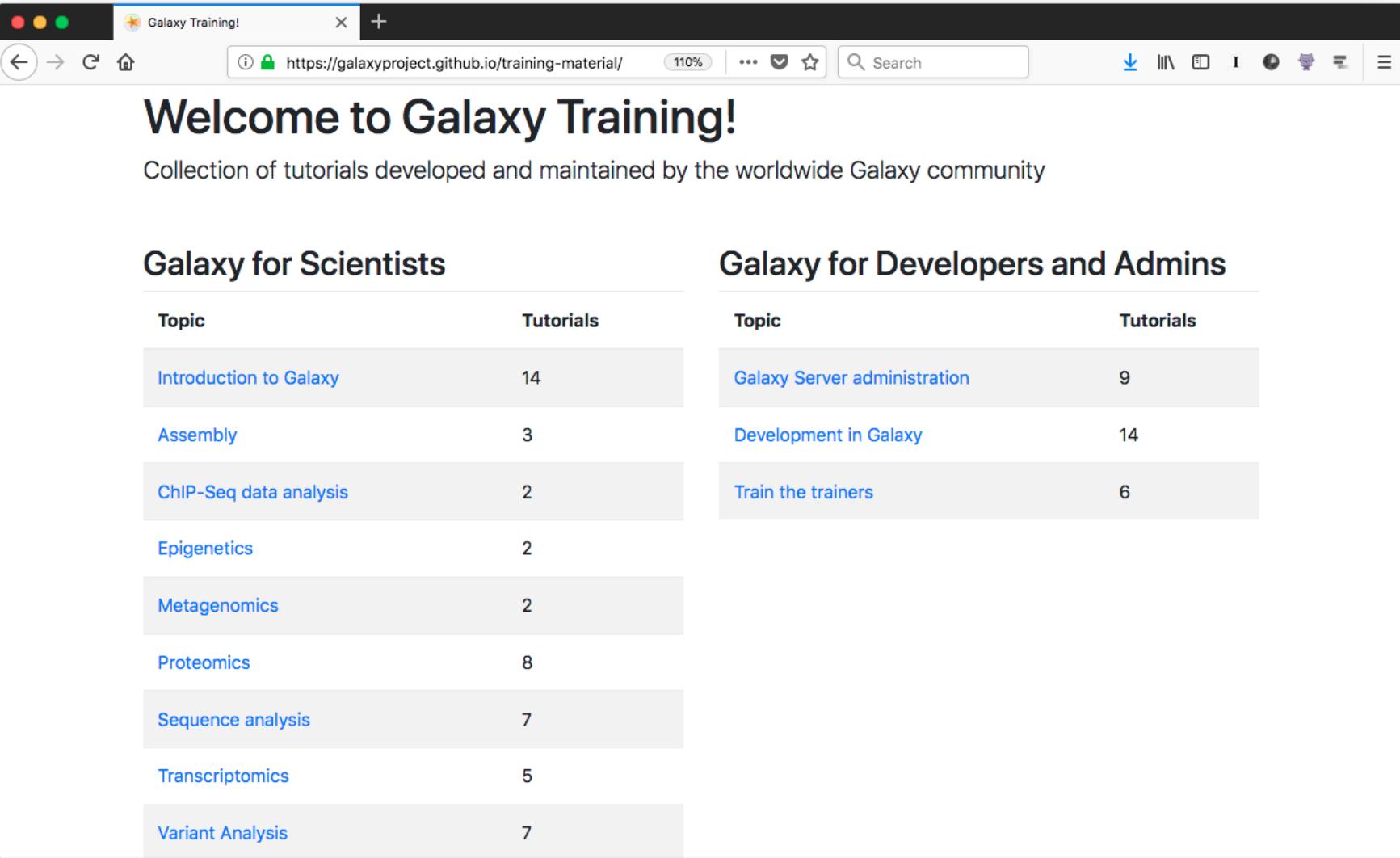
Purpose
FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ/FastQ.gz files (any variant),
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

FastQC
This is a Galaxy wrapper. It merely exposes the external package [FastQC](#) which is documented at [FastQC](#). Kindly acknowledge it as well as this

Online training material covering many topics



The screenshot shows a web browser window titled "Galaxy Training!" with the URL <https://galaxyproject.github.io/training-material/>. The page features a "Welcome to Galaxy Training!" header and a subtitle "Collection of tutorials developed and maintained by the worldwide Galaxy community". Below this, there are two main sections: "Galaxy for Scientists" and "Galaxy for Developers and Admins", each containing a list of topics and their corresponding tutorial counts.

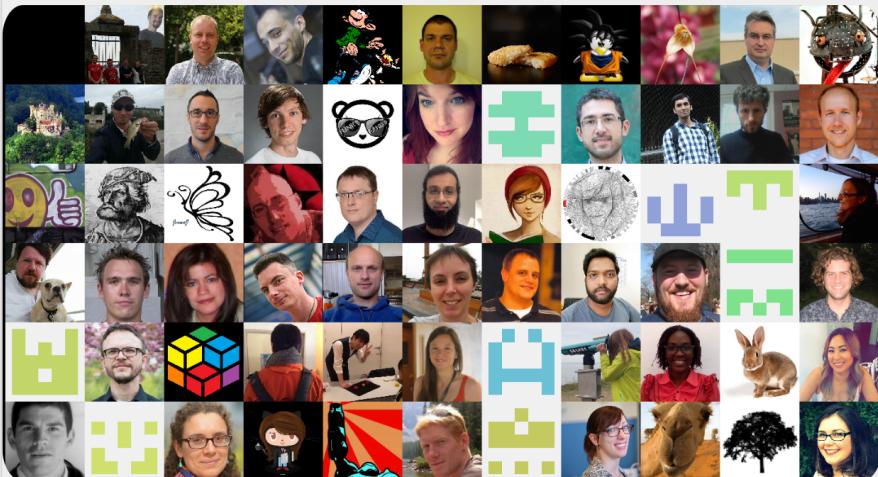
Topic	Tutorials
Introduction to Galaxy	14
Assembly	3
ChIP-Seq data analysis	2
Epigenetics	2
Metagenomics	2
Proteomics	8
Sequence analysis	7
Transcriptomics	5
Variant Analysis	7

Topic	Tutorials
Galaxy Server administration	9
Development in Galaxy	14
Train the trainers	6

<https://training.galaxyproject.org>

Community-driven

A growing community



>70 contributors!

Support & Discussions

The screenshot shows a GitHub repository page for 'Galaxy Training Network/Lobby'. It displays a list of issues and pull requests. One issue is about Eukaryotic vs prokaryotic content, another is about build procedures, and others are about data hack topics and pipeline updates.

Real time chat
Galaxy-Training-Network/Lobby

Community events

10.16



05.17 06.17

ELIXIR/
GOLET/
GTN

Galaxy
Community
Conference

05.18 06.18 08.18 11.18

ELIXIR/
GTN

Galaxy
Community
Conference



Thank you!

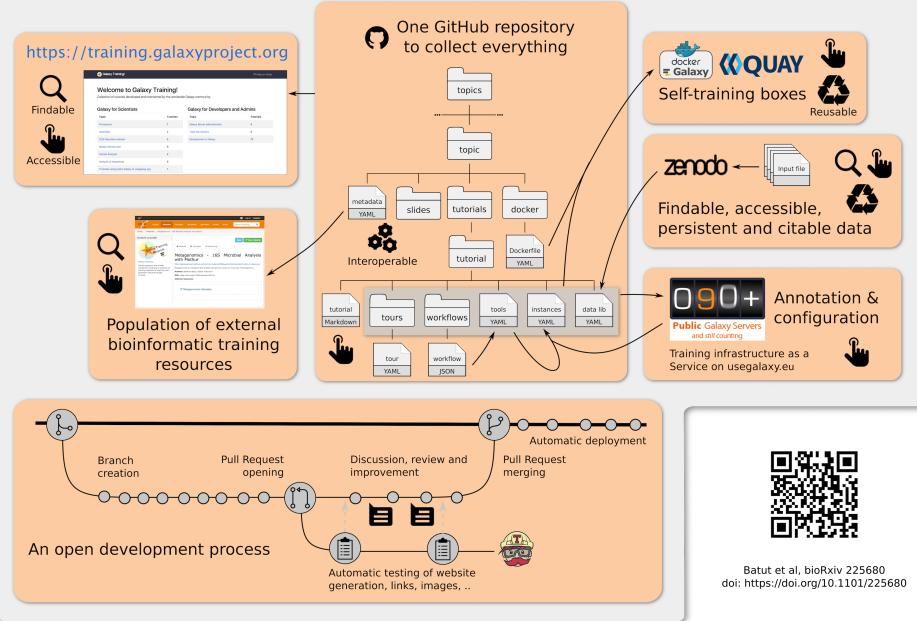
Community-Driven Training for Biological Data Analysis with the Galaxy Training Network

Bérénice Batut¹ (berenice.batut@gmail.com)
Saskia Hiltemann², Galaxy Training Network³, Björn Grüning¹

The primary problem with the explosion of biomedical datasets is not the data itself, nor computational resources, nor the required storage space, but the general lack of trained and skilled researchers to manipulate and analyze this data. Eliminating this problem requires development of comprehensive educational resources. Here we present a community-driven framework that enables modern, interactive teaching of data analytics in life sciences and facilitates the development of training materials.



A newly developed and already used Galaxy Training infrastructure



Interactive training

Quality control

Community driven

Support & Discussions

Community events

ELIXIR/ GOBLET/GTN, Galaxy Community Conference, ELIXIR/ Galaxy Community Conference

¹ University of Freiburg, Germany
² Erasmus University Rotterdam, The Netherlands
³ <https://galaxyproject.org/teach/gtn>

Icons by The Noun Project, Flaticon and icon8

Rx Community-driven data analysis training for biology

 training.galaxyproject.org



github.com/galaxyproject/training-material

