

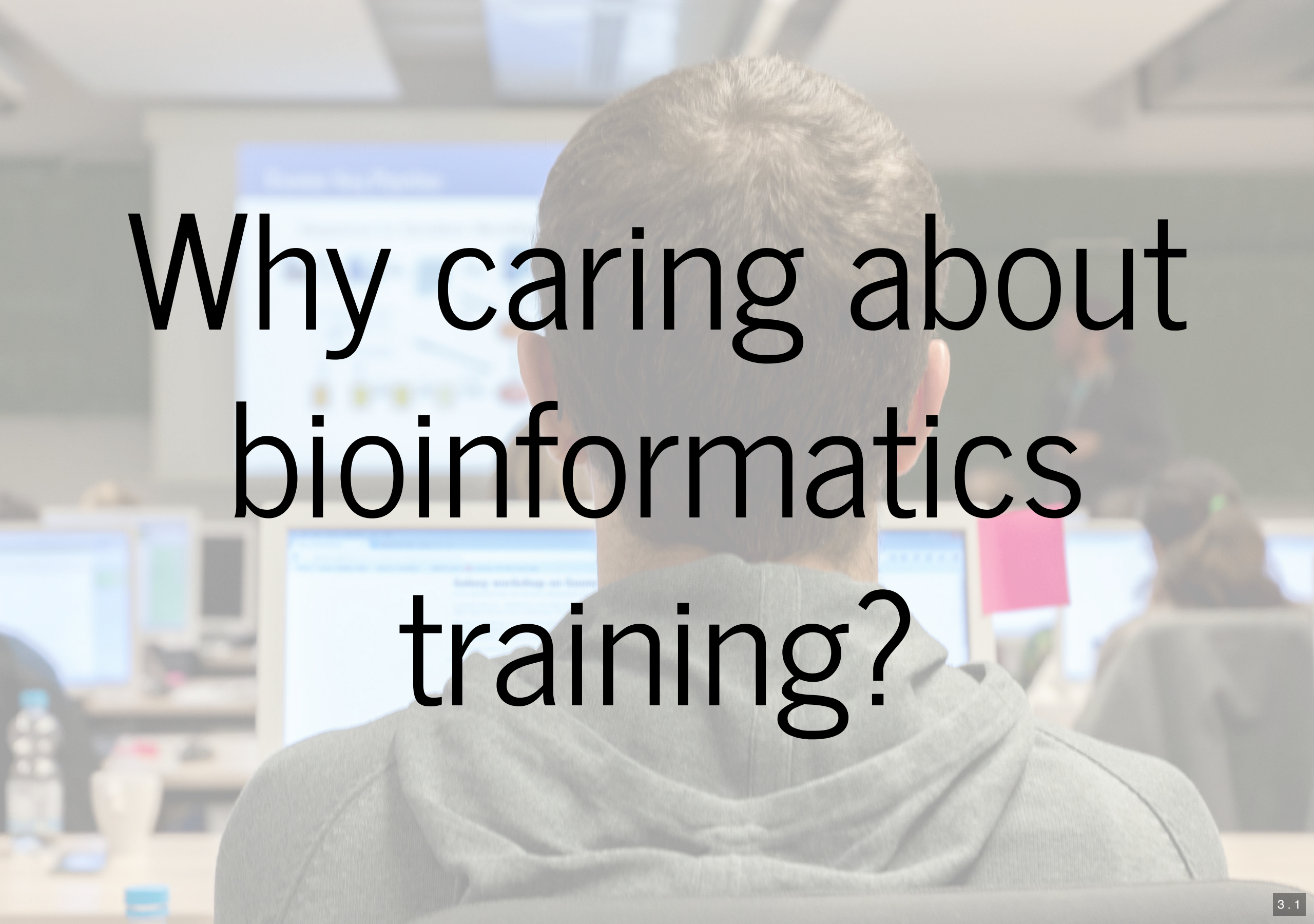
# Community-driven training for biological data analysis with the Galaxy Training Network



Picture from Bérénice Batut - Icons from the Noun Project and FlatIcon

Bérénice Batut

Galaxy Africa - April 2018

A person with dark hair, seen from behind, is sitting at a desk in what appears to be a computer lab or office. They are wearing a grey hoodie. In front of them are several computer monitors displaying various data visualizations, including bar charts and line graphs. The background is slightly blurred, showing other people and more monitors, suggesting a busy, collaborative work environment. The text "Why caring about bioinformatics training?" is overlaid in a large, black, sans-serif font across the center of the image.

Why caring about  
bioinformatics  
training?

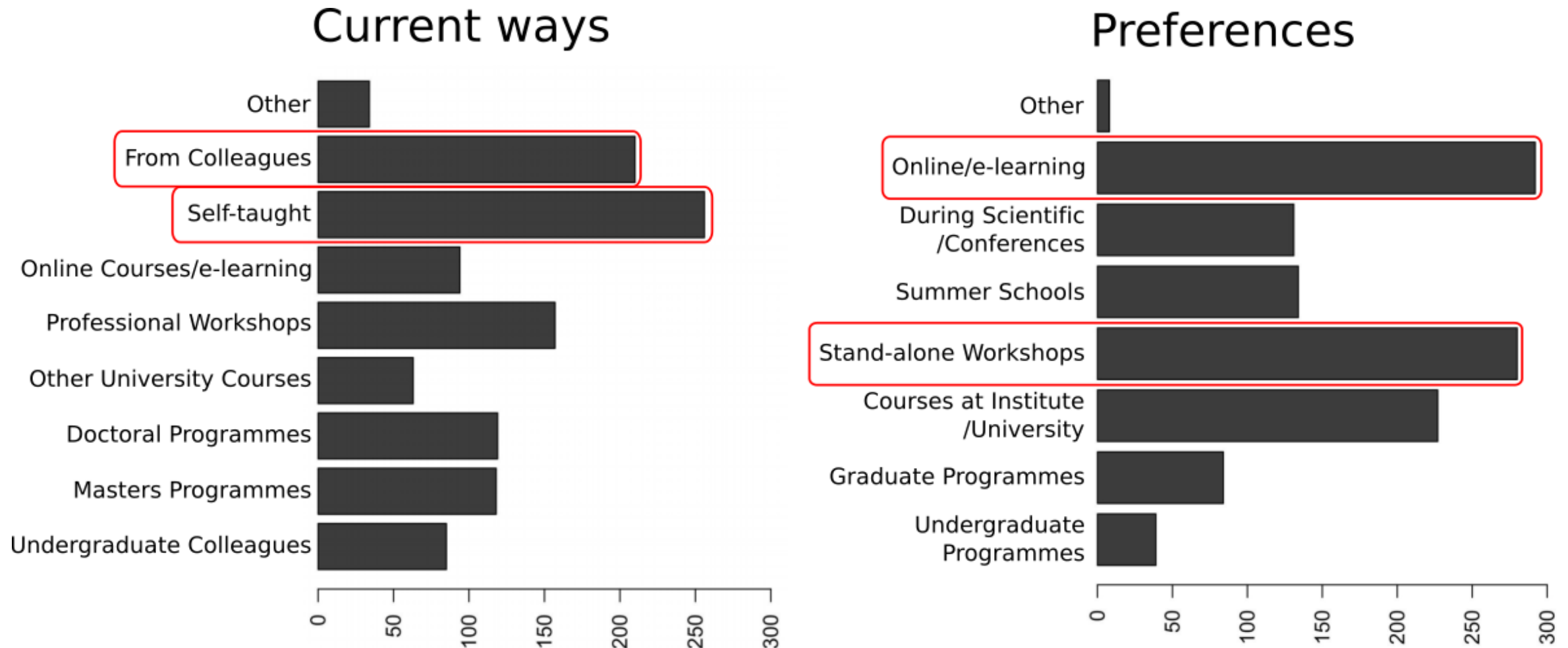
# Need for bioinformatic training

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

- Explosion of data to analyze
- Access to computational power
- Thousand of possible tools for specialized analyses



# An increasing demand for learning bioinformatics



Graphs of Brazas et al, 2017



# Galaxy

## a great solution!

Anwendungsmenü [galaxy\_course\_C... Galaxy - Mozilla F... IGV\_2.3.34 IGV [galaxy25 - Datei... 15:45 galaxy25

Galaxy - Mozilla Firefox

Galaxy x Display Application: ... x Explain SAM Flags x bamCompare — dee... x intersect — bedtools ... x

galaxy.uni-freiburg.de Suchen

Galaxy / Uni Freiburg Analyze Data Workflow Shared Data Visualization Help User Using 50.6 GB

Tools

plotheatmap

deepTools

plotProfile creates a profile plot for score distributions across genomic regions

computeMatrix prepares data for plotting a heatmap or a profile of given regions

plotHeatmap creates a heatmap for score distributions across genomic regions

Workflows

All workflows

12: bamCoverage on data 2

2: patient4\_input\_poor\_outcome.bam (as bigwig)

1: patient4\_ChIP\_ER\_poor\_outcome.bam (as bigwig)

You can generate a bigWig file from a BAM file using the bamCoverage tool. (--scoreFileName)

computeMatrix has two main output options

scale-regions

In the scale-regions mode, all regions in the BED file are stretched or shrunk to the same length (in bases) that is indicated by the user. Reference-point refers to a position within the BED regions (start or end of each region). In the reference-point mode only those genomic positions before (upstream) and/or after (downstream) the reference point will be considered.

Distance in bases to which all regions are going to be scaled

500

(--regionBodyLength)

Set distance up- and downstream of the given region

no

Show advanced output settings

no

Show advanced options

no

Sending...

History

search datasets

ChIPseq\_sept2016\_afternoon

23 shown, 8 deleted, 1 hidden

309.24 MB

31: plotProfile on data 28: Underlying data

empty

format: tabular, database: hg18

3: plotProfile on data 31: Underlying data

1.2 MB

format: png, database: hg18

28: bamCoverage on data 14: Underlying data

17: bamCoverage on data 14: Underlying data

What it does

This tool produces an intermediate file (a gzipped tabular file) that contains scores associated with genomic regions. The scores are calculated by the bamCoverage tool. The resulting file can be used by the plotProfile tool to generate a profile plot or by the plotHeatmap tool to generate a heatmap.

Firefox sendet automatisch eine kleine Menge an Daten an Google, damit die Suche schneller und präziser sein kann. Zu den Einstellungen

# Computational knowledge: Not required!

The screenshot displays the Galaxy web interface for the 'Diamond alignment tool for short sequences against a protein database (Galaxy Version 0.8.24)'. The interface is divided into three main sections: a left sidebar with a 'Tools' menu, a central configuration area, and a right sidebar with a 'History' panel.

**Tools Sidebar:** A search bar is at the top. Below it, a list of tool categories is provided, including 'Get Data', 'Lift-Over', 'Collection Operations', 'Text Manipulation', 'Datamash', 'Convert Formats', 'Filter and Sort', 'Join, Subtract and Group', 'Fetch Alignments/Sequences', 'NGS: QC and manipulation', 'NGS: DeepTools', 'NGS: Mapping', 'NGS: RNA Analysis', 'NGS: SAMtools', 'NGS: BamTools', 'NGS: Picard', 'NGS: VCF Manipulation', 'NGS: Peak Calling', 'NGS: Variant Analysis', 'NGS: RNA Structure', 'NGS: Du Novo', 'NGS: Gemini', 'NGS: Assembly', 'NGS: Chromosome Conformation', 'NGS: Mothur', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Phenotype Association', 'BEDTools', 'Genome Diversity', 'EMBOSS', 'Regional Variation', 'FASTA manipulation', 'Multiple Alignments', 'Metagenomic Analysis', 'Multiple regression', 'Multivariate Analysis', 'Motif Tools', 'STR-FM: Microsatellite Analysis', and 'NCBI SRA Tools'.

**Central Configuration Area:** The tool title is 'Diamond alignment tool for short sequences against a protein database (Galaxy Version 0.8.24)'. The configuration options include:

- What do you want to align?**: A dropdown menu set to 'Align amino acid query sequences (blastp)'.
- Input query file in FASTA or FASTQ format**: A text input field with a file upload icon and a button labeled 'No fasta or fastq dataset available'.
- Will you select a reference genome from your history or use a built-in index?**: A dropdown menu set to 'Use a built-in index'.
- Select a reference genome**: A dropdown menu set to 'No options available'.
- Genetic code used for translation of query in BLASTX mode**: A dropdown menu set to 'The Standard Code'.
- Format of output file**: A dropdown menu set to 'BLAST XML'.
- Include full length subject titles in output?**: Radio buttons for 'Yes' and 'No', with 'No' selected.
- Trigger the sensitive alignment mode with a 16x9 seed shape configuration?**: Radio buttons for 'Yes' and 'No', with 'No' selected.
- Trigger the more sensitive mode?**: Radio buttons for 'Yes' and 'No', with 'No' selected.
- Gap open penalty**: A text input field set to '11'.
- Gap extension penalty**: A text input field set to '1'.
- Scoring matrix**: A dropdown menu set to 'BLOSUM62 ((6-11)/2; (9-13)/1)'.
- Enable SEG masking of low complexity segments in the query?**: Radio buttons for 'Yes' and 'No', with 'No' selected.
- Method to restrict the number of hits?**: A text input field.

**History Sidebar:** A search bar is at the top. Below it, the 'Unnamed history' panel is shown, which is currently empty. A message box states: 'This history is empty. You can load your own data or get data from an external source'.

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







Building an infrastructure facilitating data  
analysis training in life sciences

# Requirements for a 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
- Open

# Interactive learning via hands-on tutorials

**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

**Tip**

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

- 2. Inspect on the generated webpage for `GSM461177_1` sample**

**Questions**

What is the read length?  
▶ Click to view answers

- 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**

**Questions**

What is the quality for the sequences for the different files?  
▶ Click to view answers

- 5. Trim Galore** 🛠️: Treat for the quality of sequences by running Trim Galore! with
  - "Is this library paired- or single-end?" to **Paired-end**

**Galaxy Interface**

**Tools**

- fastqc
- NGS: QC and manipulation
  - Manipulate FASTQ reads on various attributes
  - Combine FASTA and QUAL into FASTQ
  - FastQC Read Quality reports
- Workflows
  - All workflows

**FastQC Read Quality reports (Galaxy)** Version 0.69

**Short read data from your current history**

No fastq, fastq.gz, fastq.bz2, bam or sam...

**Contaminant list**

Nothing selected

**Submodule and Limit specifying file**

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

**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

**History**

search datasets

**Unnamed history**


4 shown

(empty)

- 4: [https://zenodo.org/record/1185122/files/GSM461180\\_2.fastqsanger](https://zenodo.org/record/1185122/files/GSM461180_2.fastqsanger)
- 3: [https://zenodo.org/record/1185122/files/GSM461180\\_1.fastqsanger](https://zenodo.org/record/1185122/files/GSM461180_1.fastqsanger)
- 2: [https://zenodo.org/record/1185122/files/GSM461177\\_2.fastqsanger](https://zenodo.org/record/1185122/files/GSM461177_2.fastqsanger)
- 1: [https://zenodo.org/record/1185122/files/GSM461177\\_1.fastqsanger](https://zenodo.org/record/1185122/files/GSM461177_1.fastqsanger)



# Hands-on tutorials built around a "research story"

 Galaxy Training!

Transcriptomics Introduction slides ▾ Input Dataset Literature Help ▾ Edit

## Reference-based RNA-Seq data analysis

### Overview

**❓ Questions**

- What are the effects of Pasilla (PS) gene depletion on splicing events?
- How to analyze RNA sequencing data using a reference genome?

**🎯 Objectives**

- Analysis of RNA sequencing data using a reference genome
- Analysis of differentially expressed genes
- Identification of functional enrichment among differentially expressed genes

**✔ Requirements**

- [Galaxy introduction](#)
- [Quality control](#)

**🕒 Time estimation:** 1d

## Introduction

Transcriptomics - Reference-based RNA-Seq data analysis

# Hands-on also supported by Interactive Tours

The screenshot displays the Galaxy web interface at Uni Freiburg. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The left sidebar lists various tool categories such as Get Data, Send Data, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Regional Variation, Multiple regression, Evolution, Motif Tools, Multiple Alignments, Metagenomic analyses, FASTA/FASTQ manipulation, NGS: QC and manipulation, NGS: Assembly, NGS: Mapping, NGS: RNA Analysis, NGS: Peak Calling, NGS: Simulation, DNA methylation, SNP/WGA: Data; Filters, SNP/WGA: Statistical Models, Phenotype Association, Annotation, Proteomics, ChemicalToolBox, OBO Ontology manipulation, CellOrganizer, Test Tools, NGS: Peak Calling, NGS: Picard, EMBOSS Tools, NCBI Blast, NGS: Indel Analysis, and VCF Tools.


The main content area is titled "Galaxy Tours" and contains a list of interactive tours available on the server. The tours listed are:

- [Exome Sequencing](#) – Sequencing all the protein-coding genes in a genome, the EXOME
- [History Introduction](#) – A detailed introduction to the Galaxy History
- [Galaxy UI](#) – A gentle introduction to the Galaxy User Interface
- [Scratchbook – Introduction](#) – An introduction on how to display multiple datasets and visualizations next to each other.
- [Exome Sequencing](#) – Sequencing all the protein-coding genes in a genome, the EXOME

A "Welcome to Galaxy" dialog box is open in the center of the screen, providing instructions on how to navigate the tour. The dialog box includes buttons for "« Prev", "Next »", and "End tour".

The right sidebar shows the "History" section, which is currently empty. It includes a search bar for datasets and a message indicating that the history is empty and suggesting to load data from an external source.

# A collection of materials covering many topics

 Galaxy Training!

Fork me on GitHub Help ▾

## Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

### Galaxy for Scientists

Topic	Tutorials
<a href="#">Introduction to Galaxy</a>	13
<a href="#">Assembly</a>	3
<a href="#">ChIP-Seq data analysis</a>	2
<a href="#">Epigenetics</a>	2
<a href="#">Metagenomics</a>	2

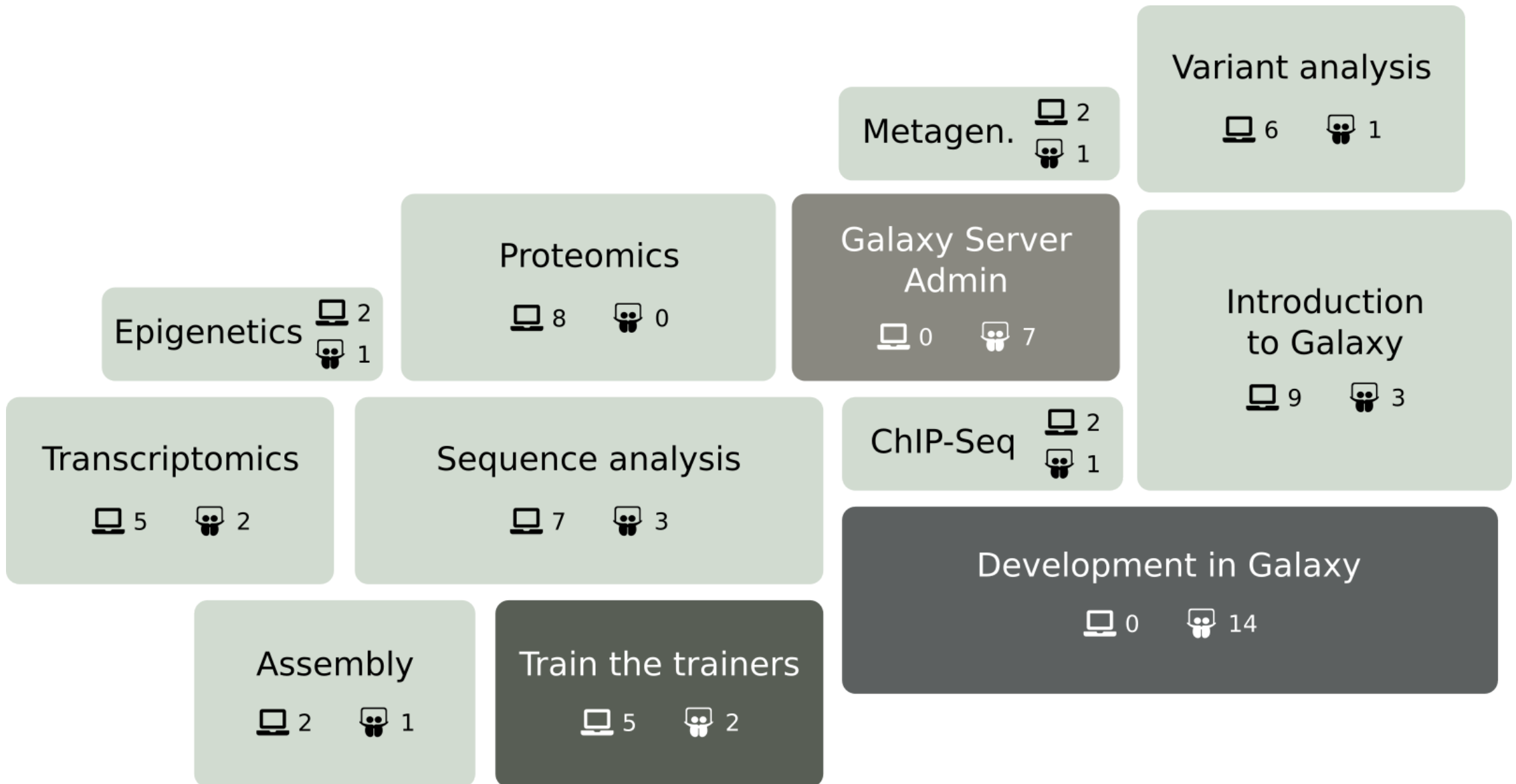
### Galaxy for Developers and Admins

Topic	Tutorials
<a href="#">Galaxy Server administration</a>	8
<a href="#">Development in Galaxy</a>	14
<a href="#">Train the trainers</a>	6

<https://training.galaxyproject.org>



# A collection of materials covering many topics



More than 80 tutorials!

# Used both by individual users

question about tutorial "Analyses of metagenomics data - The global picture"  
To: berenice.batut@gmail.com

11. March 2018 at 19:45

Hi

I am doing the tutorial "Analyses of metagenomics data - The global picture" and I have a question.

How I execute the following instruction in galaxy:

## Hands-on: Krona

1. **Visualize with Krona** with the following parameters
  - "Input file" to taxonomy output from **Classify.otu** (collection)
  - Set **Is this output from mothur?** to **Yes**

I can't find it in menu.

Thanks.

--  
[redacted]

Query regarding sequence analysis tutorial on Galaxy  
To: berenice.batut@gmail.com

29. November 2017 at 12:55

Dear sir/mam,  
I am following the RNA sequence analysis tutorial on galaxy (<https://galaxyproject.github.io/training-material/topics/sequence-analysis/>). I am unable to find **Augustus** in the Gene predication tutorial.  
Please help

-- Thanks and Regards

[redacted]  
Near Sola Bridge, Thaltej  
Ahmedabad-380054  
Gujrat, INDIA

Having trouble with the 16S Microbial Analysis with Mothur tutorial  
To: berenice.batut@gmail.com

Training 7. June 2017 at 06:54

NM

Dear Bérénice,

I recently found this tutorial on the galaxy site <https://galaxyproject.github.io/training-material//Metagenomics/tutorials/mothur-miseq-sop> and decided to contact you. I went through it till the "Hands-on: Combine forward and reverse reads into contigs." Unfortunately I'm not sure what dataset to specify on the **rfastq - Reverse Fastq Sequence file**. I made the previous dataset pair but i haven't successfully made any progress going through the workflow. I tried selecting the files manually for the forward and reverse using the multiple datasets icon but it didn't yield the required 6 new collections, trim.contig.fasta and scrap.contig.fasta. Instead i got a log file, contigs.matched file, contigs.qual and contigs.fasta file. In addition, i tried specifying the dataset collection with the paired ataset as input for both **ffastq** and **rfastq** and i got a lot of errors. What could i be doing wrong?

Please advice.

Regards,  
[redacted]

Make.contigs Aligns paired forward and reverse fastq files to contigs as fasta and quality (Galaxy Options)  
Version 1.27.0)

**ffastq - Forward Fastq Sequence file**

[file icon] [copy icon] [folder icon] 40: Mock\_R2.fastq

**rfastq - Reverse Fastq Sequence file**

[file icon] [copy icon] [folder icon] 40: Mock\_R2.fastq

galaxy reference based rnaseq analysis  
To: berenice.batut@gmail.com

28. February 2018 at 19:43

SC

Hi,

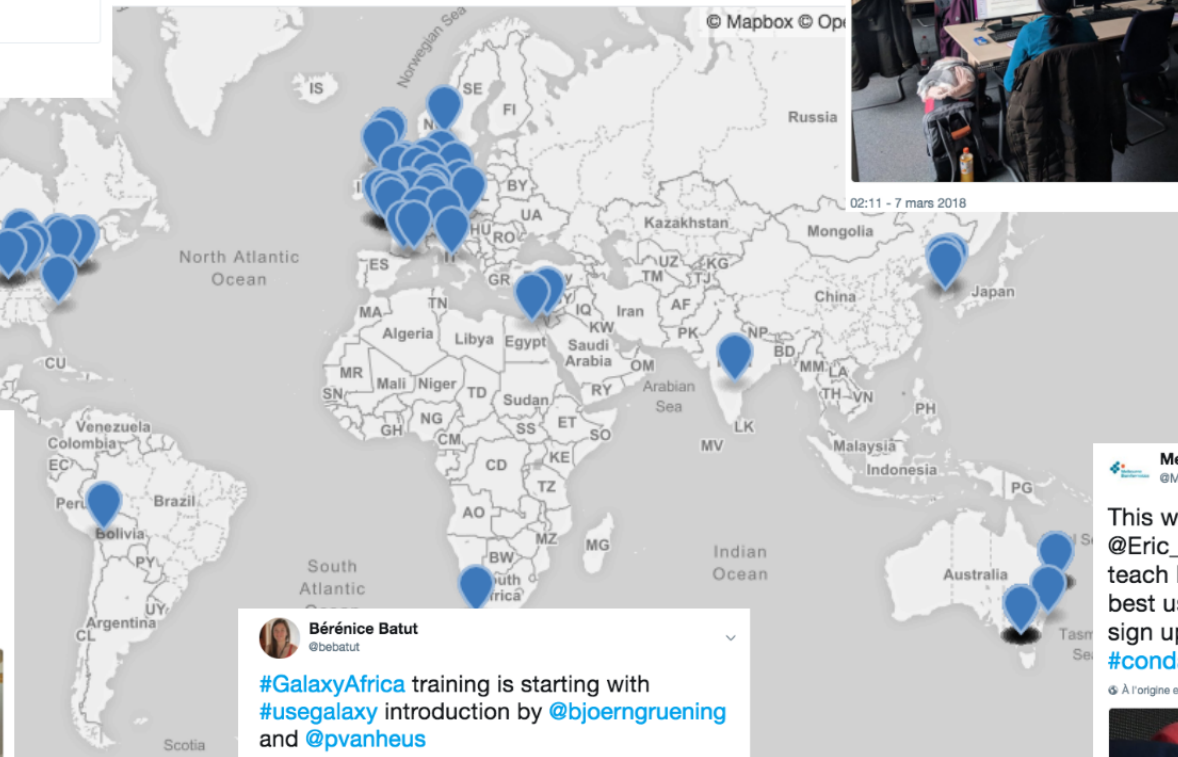
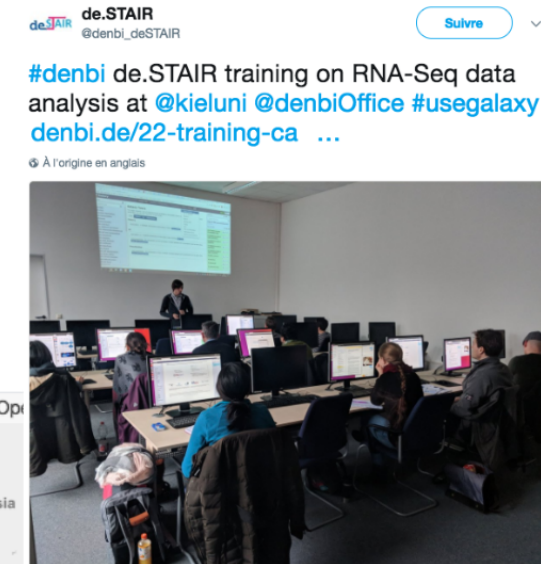
Last week, I analyze my data following your reference based rnaseq analysis using HISAT2 with HTseq. But now it seems like it changed to STAR alignment based analysis. Is there a way that I can find other tutorial. If so please provide me the link.

Thanks  
Sri

Used both by individual users & instructors



# Used both by individual users & instructors



# Requirements for a 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
- ☐ Open

# Building an infrastructure to facilitate community-led content development

- Makes tutorial creation a convenient, hassle-free process
- Enables transparent peer-review and curation to guarantee high-quality and current content

# Separation between content and format


Here treatment is the primary factor which we are interested in. The sequencing type is some further information that we know about the data that might affect the analysis. This particular multi-factor analysis allows us to assess the effect of the treatment, while taking the sequencing type into account, too.

```
> ### {% icon comment %} Comment
>
> We recommend you to add as many factors as you think may affect gene expression
in your experiment. It can be the sequencing type like here, but it can also be the
manipulation (if different persons are involved in the library preparation), ...
{: .comment}


> ### {% icon hands_on %} Hands-on: Analysis of the differential gene expression
(1)
>
> 1. Create a new history
> 2. Import the seven count files from [Zenodo]
(https://dx.doi.org/10.5281/zenodo.290221)
> - `GSM461176_untreat_single.deseq.counts`
> - `GSM461177_untreat_paired.deseq.counts`
> - `GSM461178_untreat_paired.deseq.counts`
> - `GSM461179_treat_single.deseq.counts`
> - `GSM461180_treat_paired.deseq.counts`
> - `GSM461181_treat_paired.deseq.counts`
> - `GSM461182_untreat_single.deseq.counts`
>
> 3. **DESeq2** {% icon tool %}: Run **DESeq2** with:
> - "Treatment" as first factor with "treated" and "untreated" as levels and
selection of count files corresponding to both levels
>
> > ### {% icon tip %} Tip
> >
> > You can select several files by keeping the CTRL (or COMMAND) key pressed
and clicking on the interesting files
> {: .tip}
>
```


Markdown


Here treatment is the primary factor which we are interested in. The sequencing type is some further information that we know about the data that might affect the analysis. This particular multi-factor analysis allows us to assess the effect of the treatment, while taking the sequencing type into account, too.

 **Comment**

We recommend you to add as many factors as you think may affect gene expression in your experiment. It can be the sequencing type like here, but it can also be the manipulation (if different persons are involved in the library preparation), ...

 **Hands-on: Analysis of the differential gene expression (1)**

1. Create a new history
2. Import the seven count files from [Zenodo](https://dx.doi.org/10.5281/zenodo.290221)
  - `GSM461176_untreat_single.deseq.counts`
  - `GSM461177_untreat_paired.deseq.counts`
  - `GSM461178_untreat_paired.deseq.counts`
  - `GSM461179_treat_single.deseq.counts`
  - `GSM461180_treat_paired.deseq.counts`
  - `GSM461181_treat_paired.deseq.counts`
  - `GSM461182_untreat_single.deseq.counts`
3. **DESeq2** : Run **DESeq2** with:
  - "Treatment" as first factor with "treated" and "untreated" as levels and selection of count files corresponding to both levels


 **Tip**

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

User-friendly HTML

<https://training.galaxyproject.org>

# Creating a tutorial







 Galaxy Training!

Fork me on GitHub Help ▾

## Train the trainers

Train the trainers

### Material

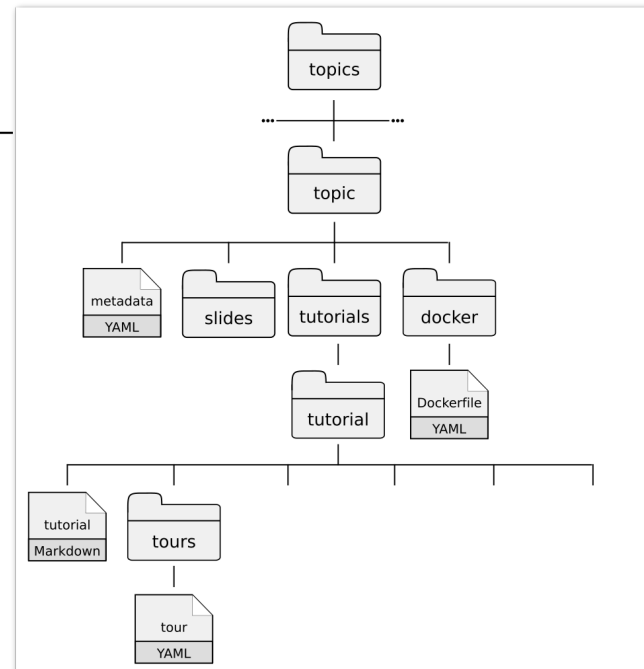
Lesson	Slides	Hands-on
Introduction to training with Galaxy		
Creating a new tutorial - Writing content in Markdown		
Creating a new tutorial - Defining metadata		
Creating a new tutorial - Setting up the infrastructure		
Creating a new tutorial - Creating Interactive Galaxy Tours		
		

<http://galaxyproject.github.io/training-material/topics/training/>

# One GitHub repository to collect everything

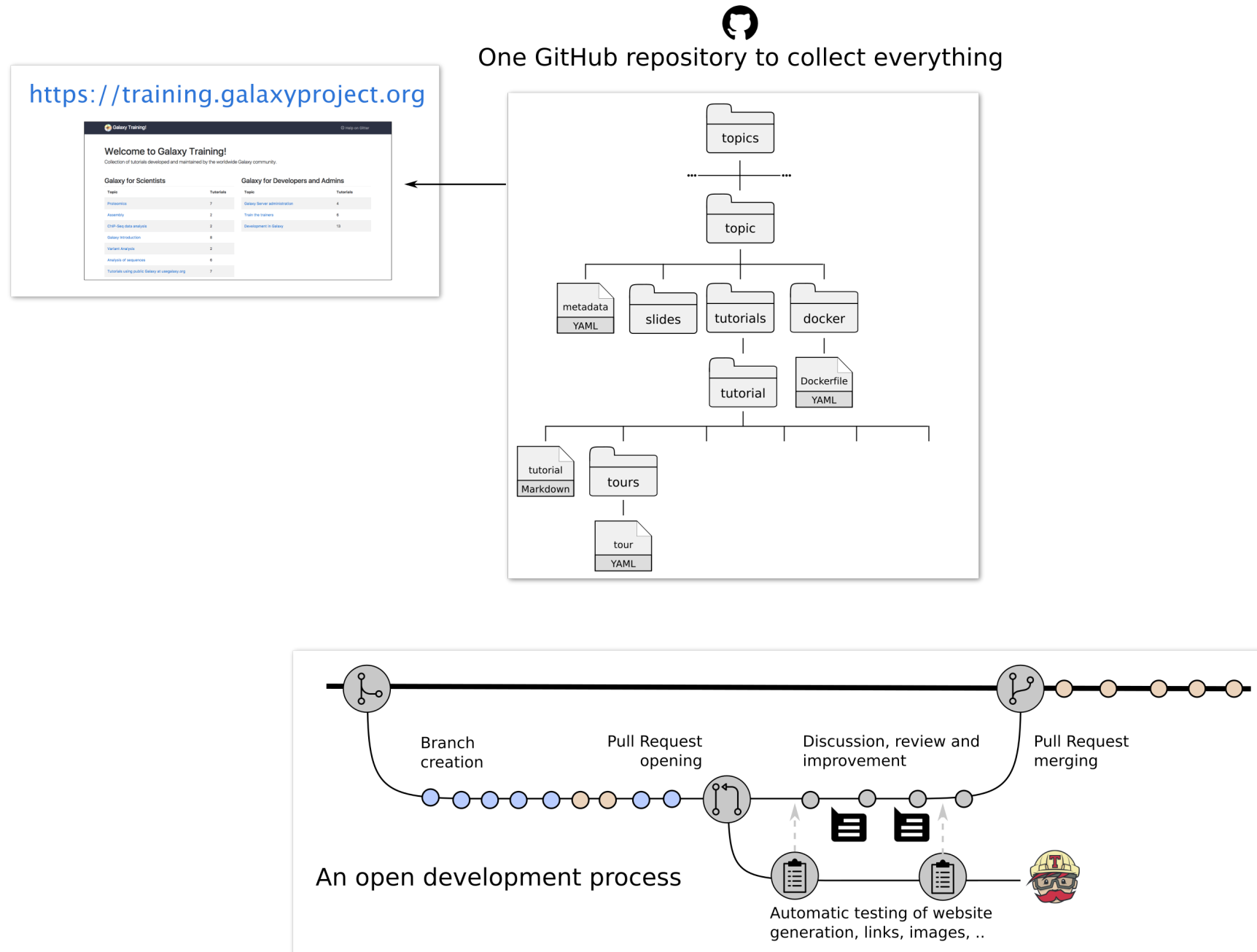


One GitHub repository to collect everything





# An open and accessible development process



# Community-driven

# Community-driven

# A constant support

The screenshot displays the Gitter chat interface for the Galaxy Training Network/Lobby. The header bar is purple and contains the project logo, name, and URL. The main chat area on the left shows a list of messages from various users, including Gildas Le Corguillé, John Chilton, Björn Grüning, Dannon Baker, and Victoria Dominguez del Angel. The messages discuss technical issues like build pipelines, slide content changes, and the need for more tags. On the right, there are two sidebars: 'PEOPLE' showing a grid of user avatars and 'ACTIVITY' showing a list of recent repository events such as pull requests and commits. The bottom of the interface includes a text input field for sending messages and a status bar.

Gitter: Galaxy-Training-Network/Lobby

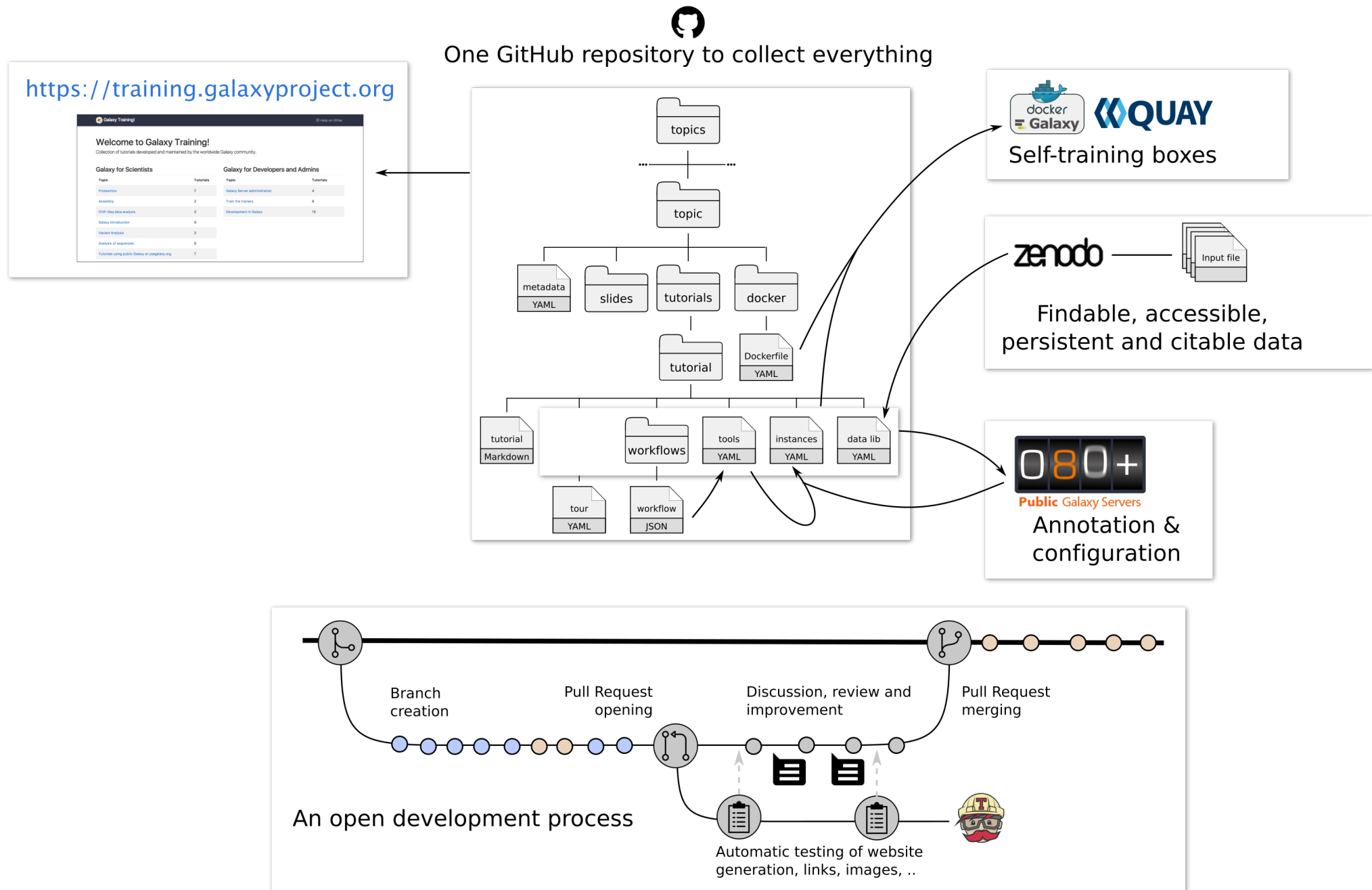
# Galaxy Tour Builder

A web extension to develop interactive tours

The screenshot displays the Galaxy web interface at <https://usegalaxy.org>. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'Login or Register'. The left sidebar lists various tools and categories, such as 'Get Data', 'Text Manipulation', 'NGS: QC and manipulation', and 'NGS: RNA Analysis'. The main content area features a welcome message: 'Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#). You can install your own Galaxy by following the [tutorial](#) and choose from thousands of tools from the [Tool Shed](#).' Below this is a promotional banner for the 'ISMB/ECCB 2017 Tutorial' with the text 'Making Galaxy work for you' and 'Register now'. The right sidebar shows a 'History' section with a search bar and a list of datasets, including one titled '7: <https://raw.githubusercontent.com/galaxyproject/galaxy/dev/test-data/1.fasta>' and another titled '6: <https://raw.githubusercontent.com/bgruenin/galaxytools/adf077b912ddeb97b07b947b855cdd2862ed8ef/tools/statistics/test-data/anderson.tabular>'. A tweet from the Galaxy Project (@galaxyproject) is also visible, mentioning a competition and a scoreboard.



<https://github.com/TailorDev/galaxy-tourbuilder>

# Ensuring accessibility of tutorials





# TeSS: the ELIXIR's training portal





EventsMaterialsWorkflowsProvidersAbout


Log InRegister

## Welcome to TeSS: ELIXIR's Training Portal


Browsing, discovering and organising life sciences training resources, aggregated from ELIXIR nodes and 3<sup>rd</sup>-party providers.




 Events




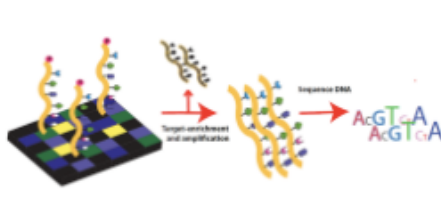
Discover the latest training events and news from ELIXIR nodes and 3<sup>rd</sup>-party providers.

 Materials





Browse the catalogue of training materials offered by ELIXIR nodes and 3<sup>rd</sup>-party providers.

 Workflows



Create training workflows to visualise learning steps and link to resources specific to your training needs.

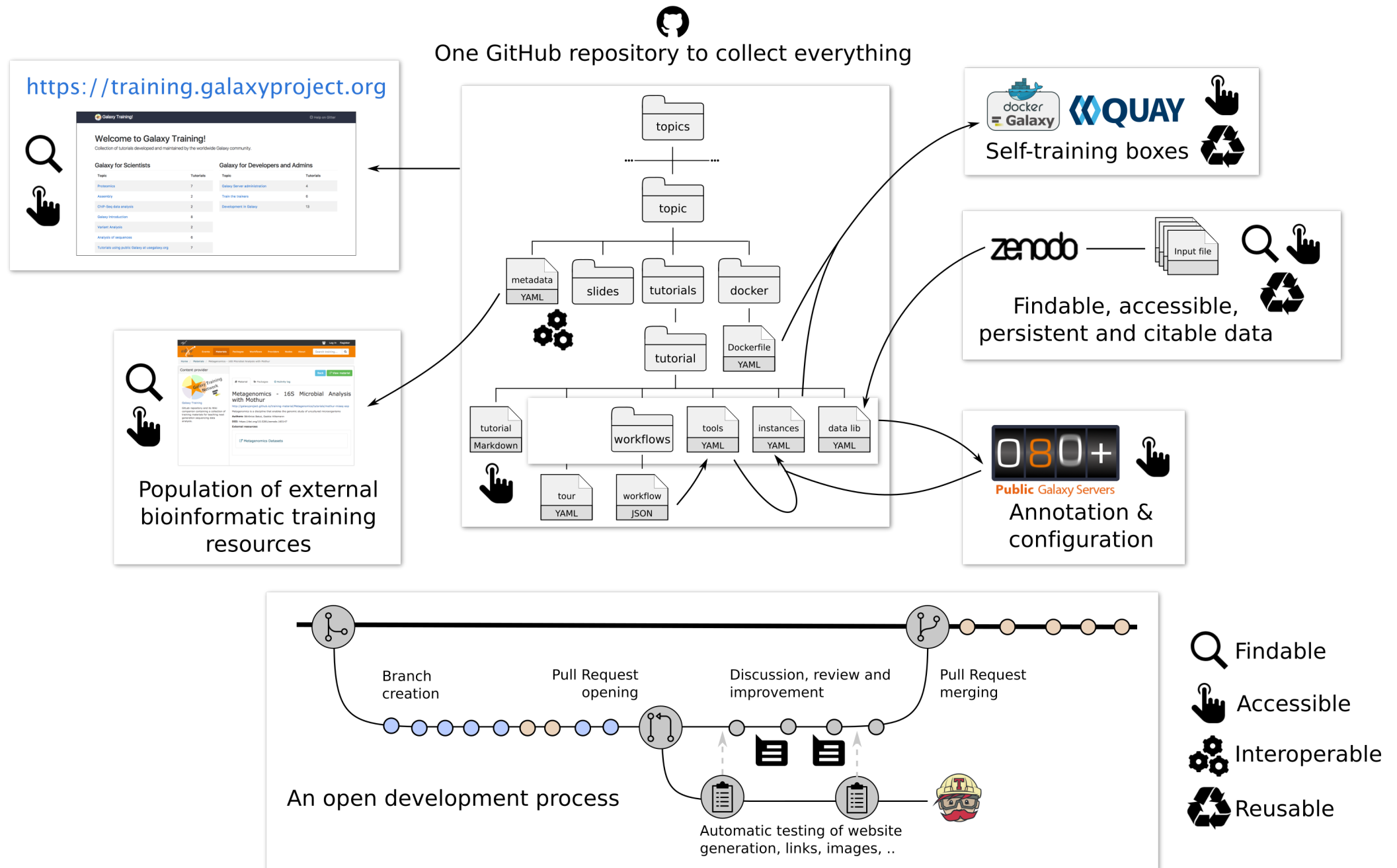
 Providers



Browse training providers to discover training resources they offer and follow links to their materials and courses.

<https://tess.elixir-europe.org/>

# Findable, Accessible, Interoperable, Reusable



# Requirements for a 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
- ☑ Open





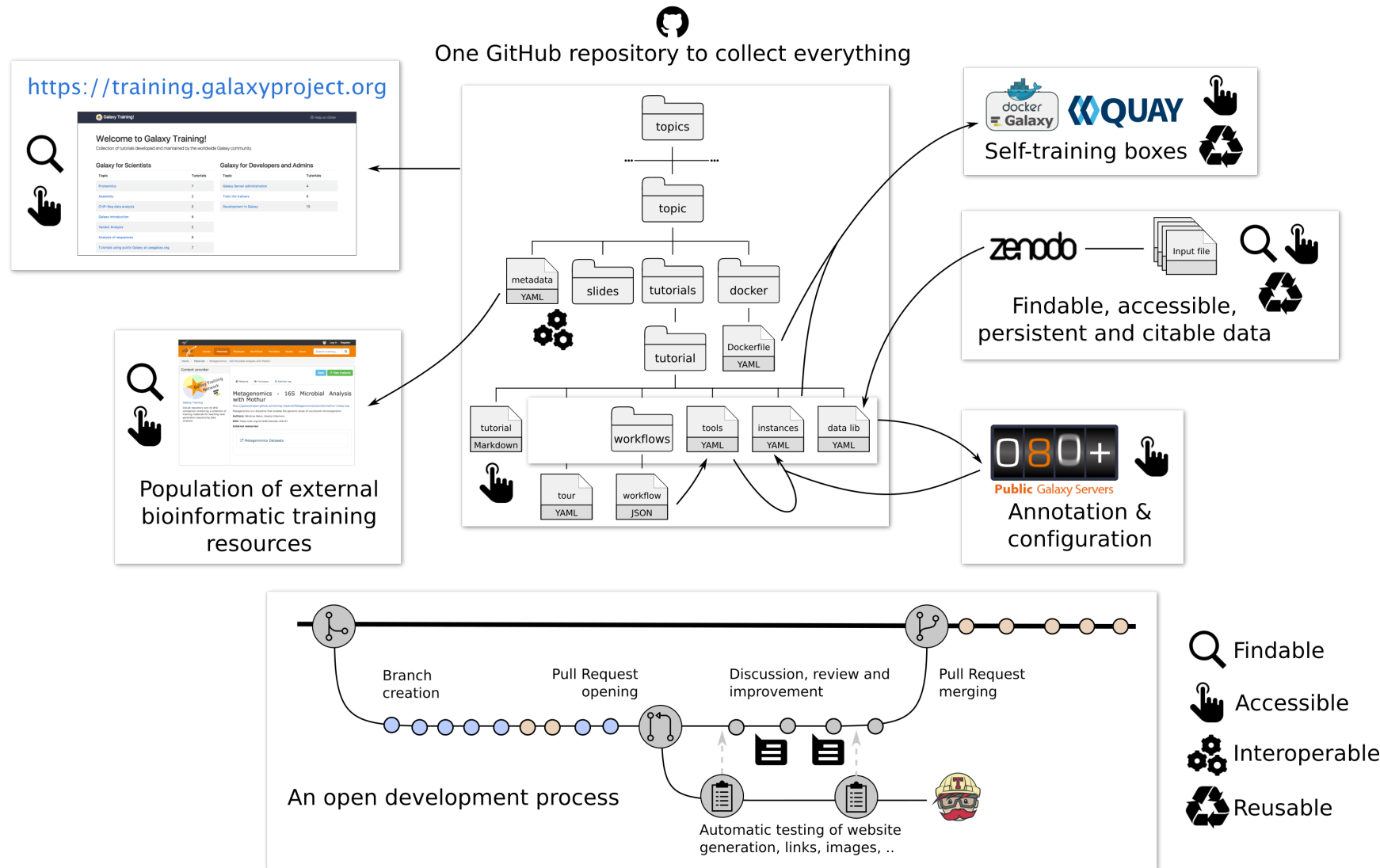


# Thank you!

## Sponsors







R<sub>X</sub> Community-driven data analysis training for biology

 [training.galaxyproject.org](https://training.galaxyproject.org)

 [github.com/galaxyproject/training-material](https://github.com/galaxyproject/training-material)