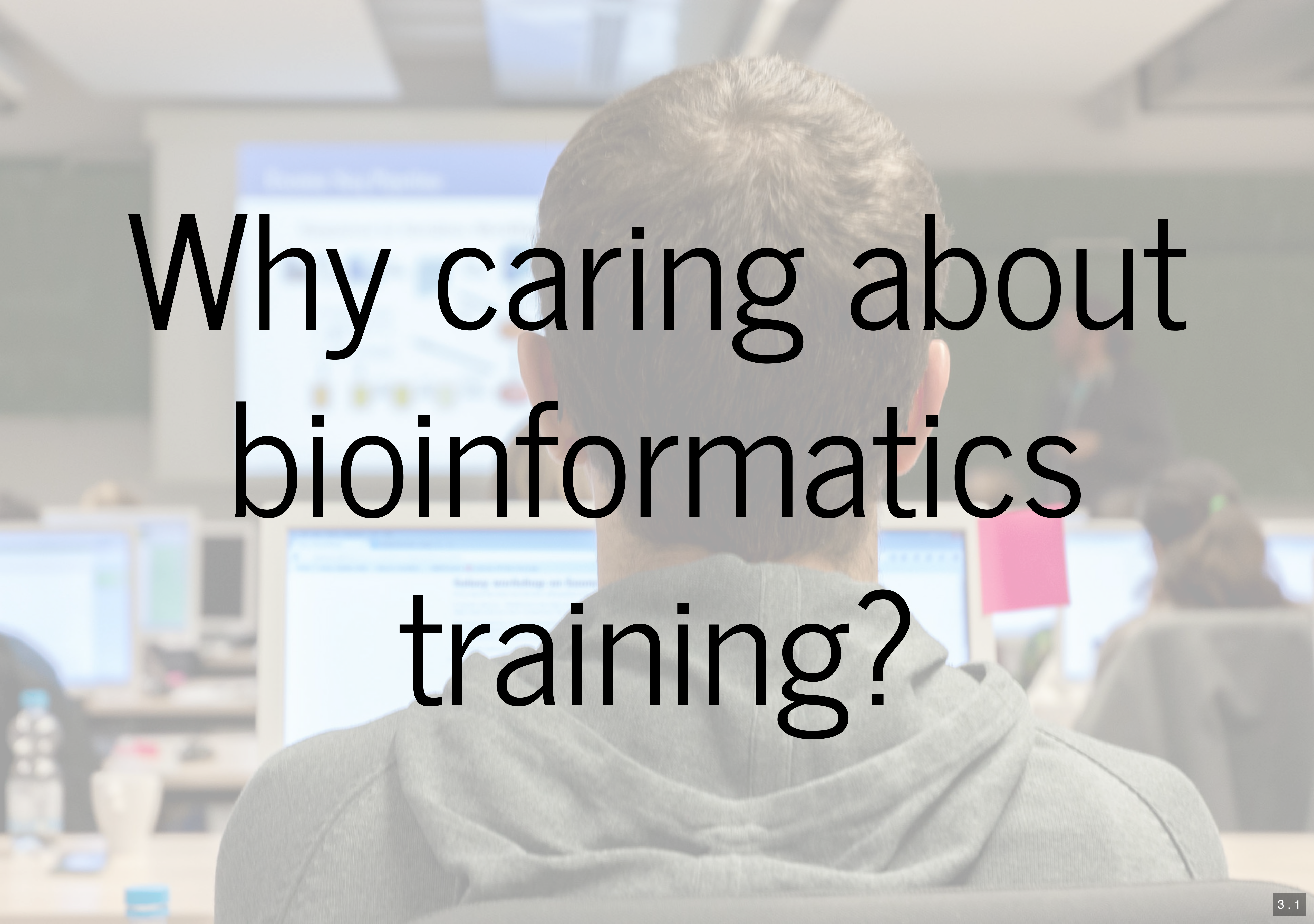
Building an open, collaborative, online infrastructure for bioinformatics training

Bérénice Batut

33rd TBI Winterseminar
Februar 2018



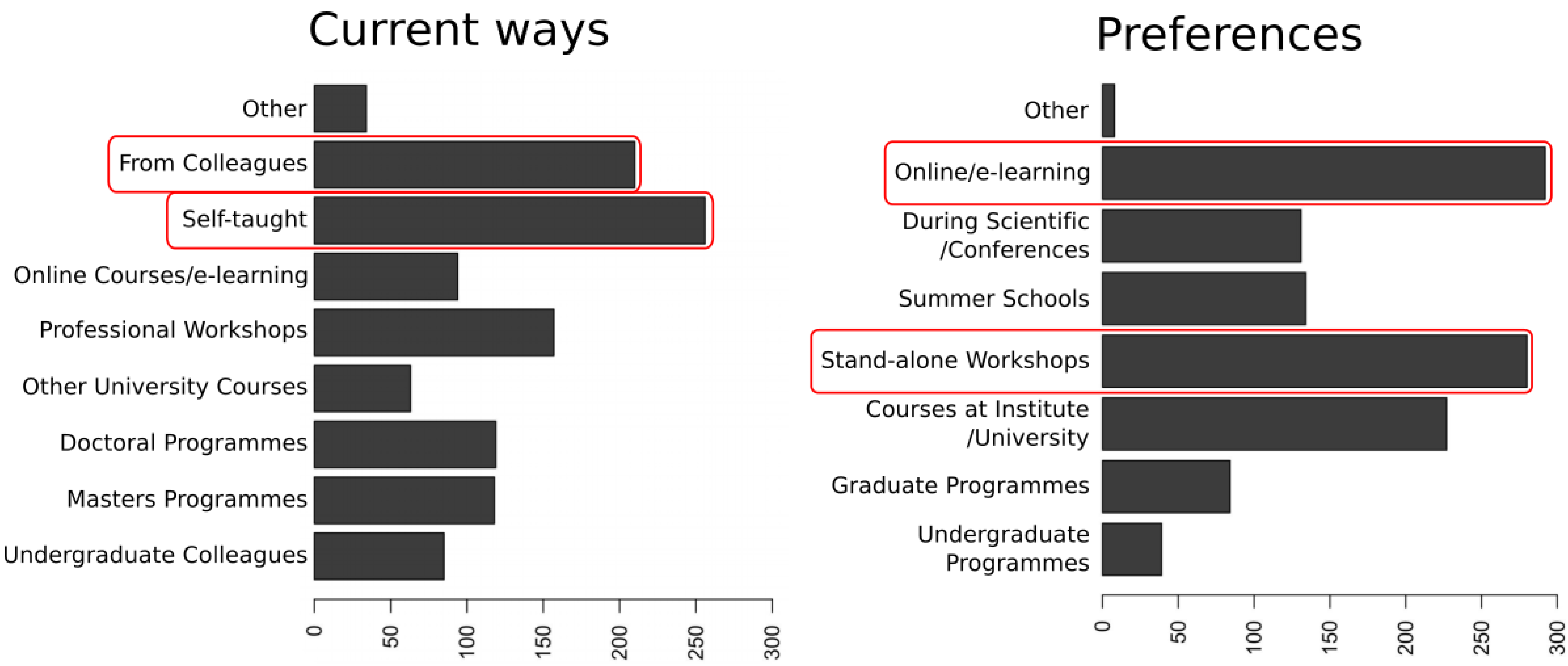
Why caring about bioinformatics training?

Need for bioinformatic training

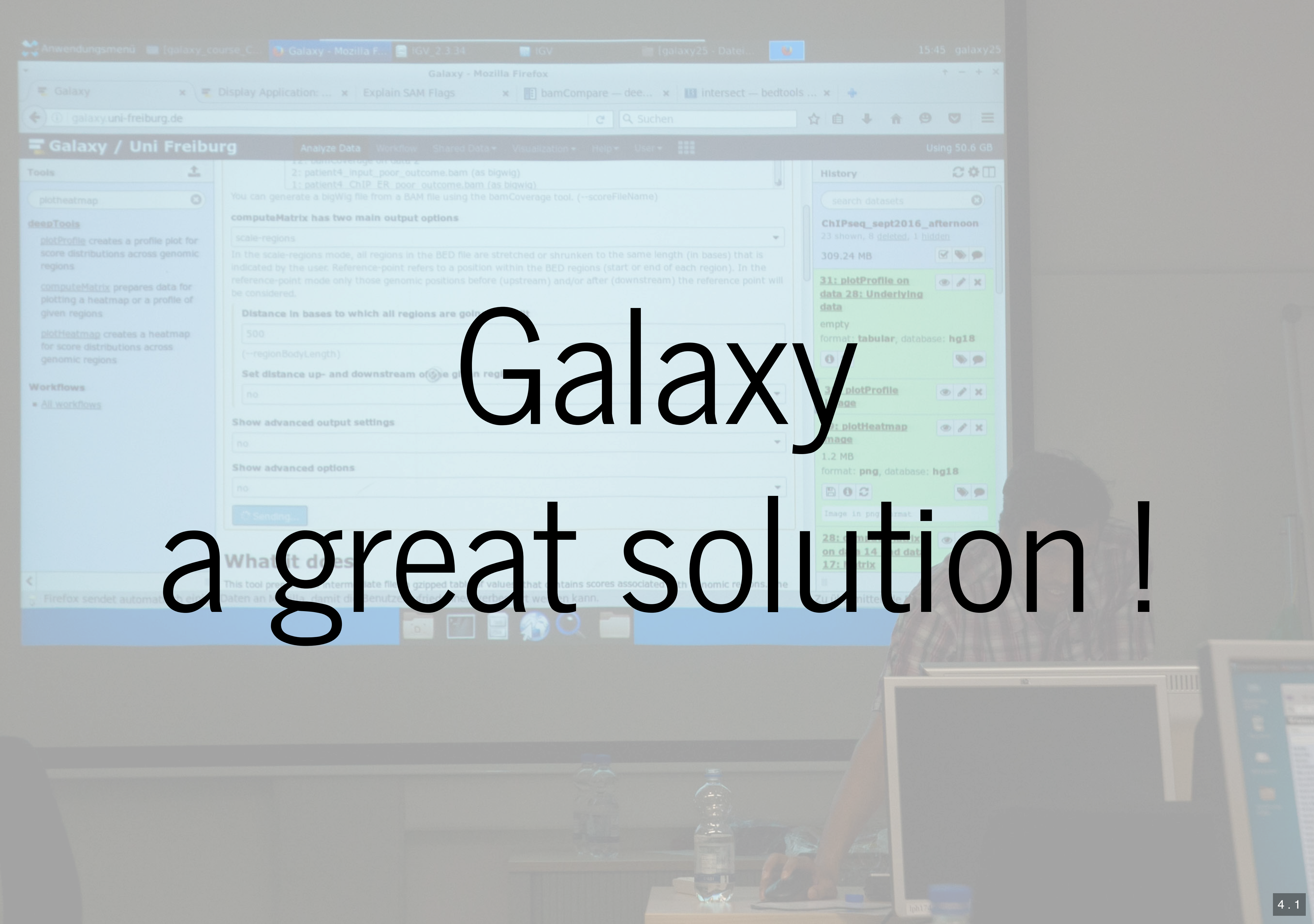
> *Bioinformatics has become too central to biology to be left to specialist bioinformaticians*

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

An increasing demand for learning bioinformatics

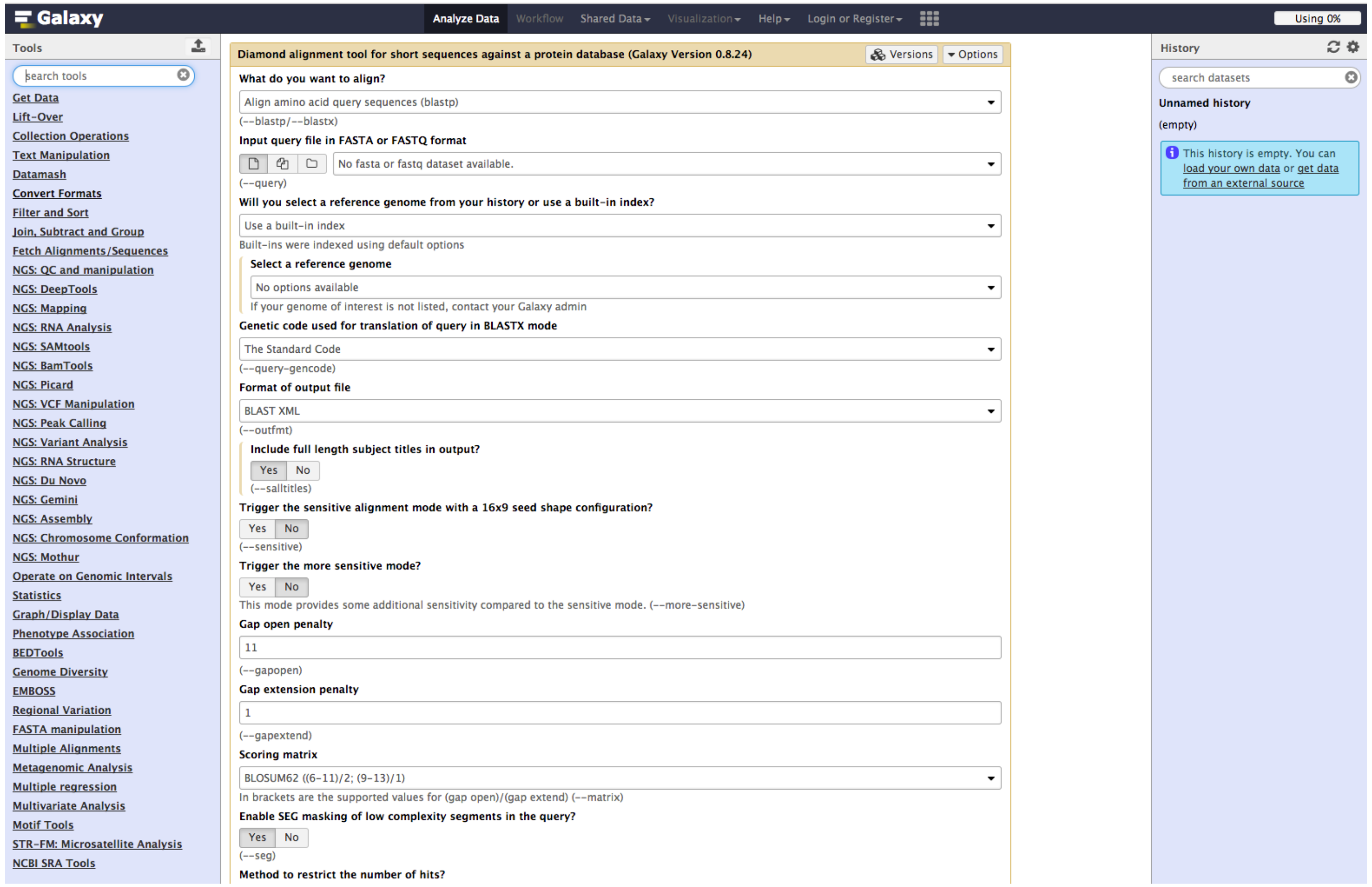


Graphs of Brazas et al, 2017



Galaxy
a great solution !

Computational knowledge: Not required!



Diamond alignment tool for short sequences against a protein database (Galaxy Version 0.8.24)

What do you want to align?
Align amino acid query sequences (blastp)
(--blastp/--blastx)

Input query file in FASTA or FASTQ format
No fasta or fastq dataset available.
(--query)

Will you select a reference genome from your history or use a built-in index?
Use a built-in index
Built-ins were indexed using default options

Select a reference genome
No options available
If your genome of interest is not listed, contact your Galaxy admin

Genetic code used for translation of query in BLASTX mode
The Standard Code
(--query-gencode)

Format of output file
BLAST XML
(--outfmt)

Include full length subject titles in output?
[✓ Yes] No
(--salltitles)

Trigger the sensitive alignment mode with a 16x9 seed shape configuration?
Yes [✓ No]
(--sensitive)

Trigger the more sensitive mode?
Yes [✓ No]
This mode provides some additional sensitivity compared to the sensitive mode. (--more-sensitive)

Gap open penalty
11
(--gapopen)

Gap extension penalty
1
(--gapextend)

Scoring matrix
BLOSUM62 ((6-11)/2; (9-13)/1)
In brackets are the supported values for (gap open)/(gap extend) (--matrix)

Enable SEG masking of low complexity segments in the query?
[✓ Yes] No
(--seg)

Method to restrict the number of hits?

Unnamed history

(empty)

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

Galaxy Training Network

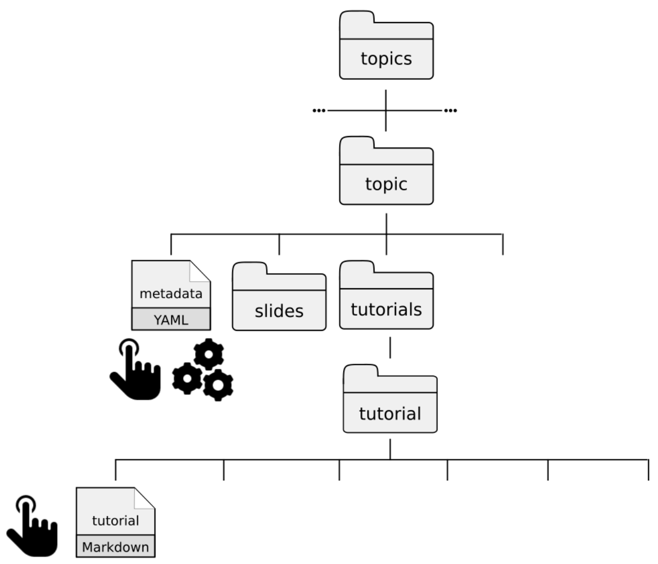
Building a new **open**, **collaborative** and **FAIR** model for bioinformatics training

Requirements

- Easy to use
- Support for effective training for
  - Individual users
  - Instructors
- Definition of technological infrastructure
- Limited redundancy

An open, collaborative, FAIR, online infrastructure

One GitHub repository to collect everything



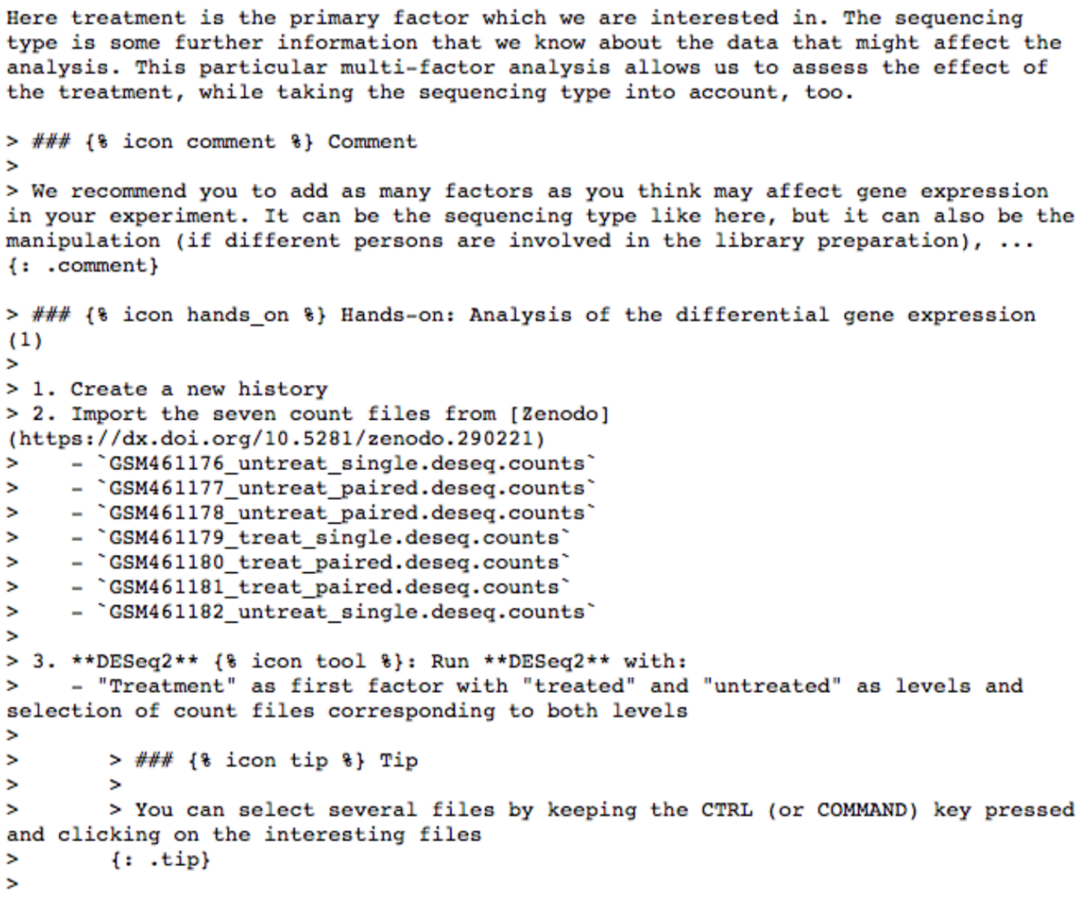
🔍 Findable

Accessible

Interoperable

♻ Reusable

Separation between content and format



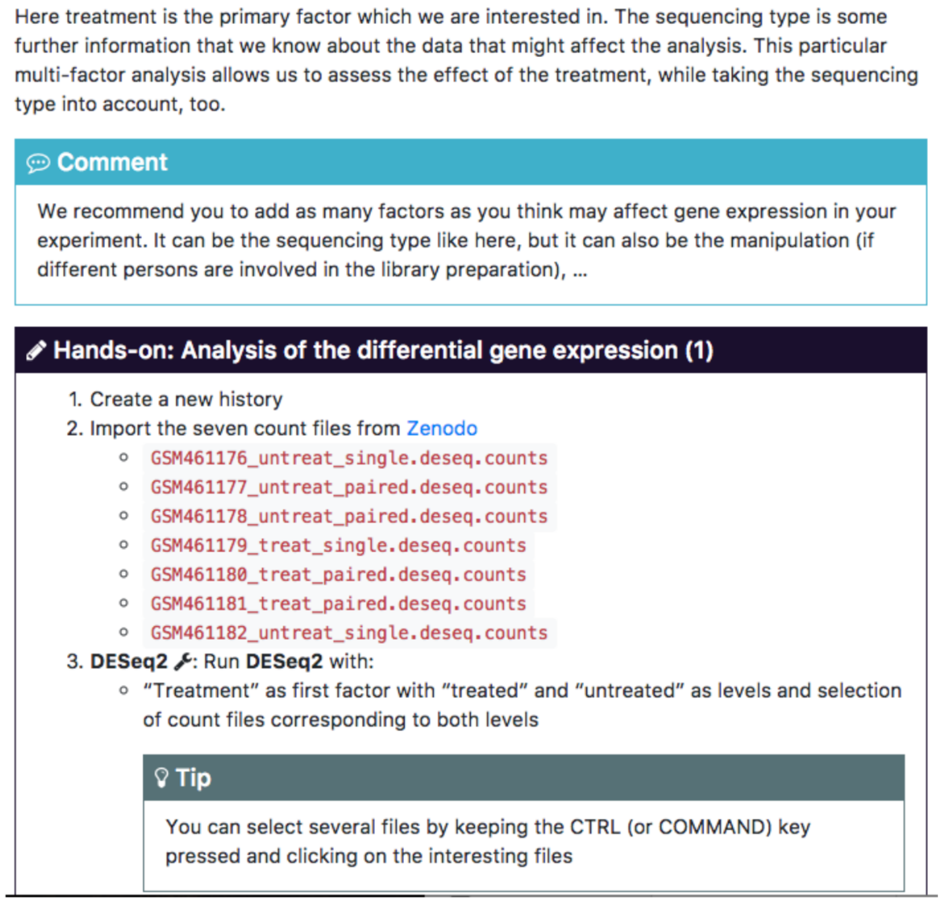
```
Here treatment is the primary factor which we are interested in. The sequencing
type is some further information that we know about the data that might affect the
analysis. This particular multi-factor analysis allows us to assess the effect of
the treatment, while taking the sequencing type into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

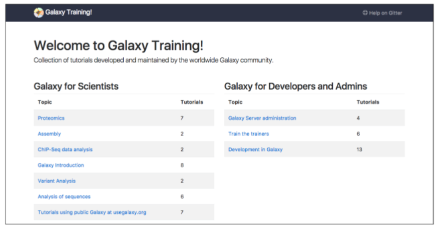
   Tip

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

User-friendly HTML

An open, collaborative, FAIR, online infrastructure

Separation between content and form



Galaxy Training!

Help on Gitter

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community.

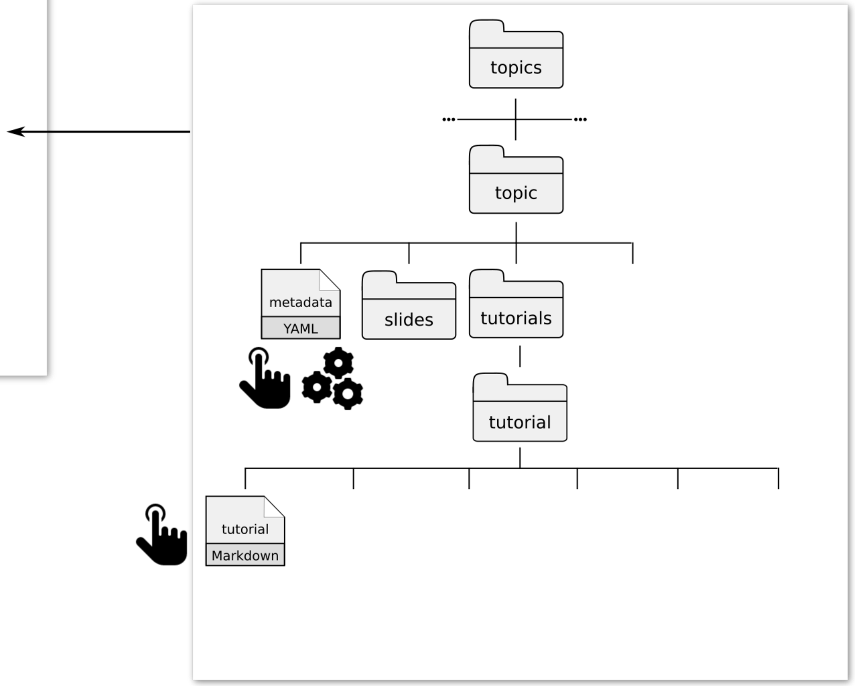
Galaxy for Scientists

Topic	Tutorials
Proteomics	7
Assembly	2
ChIP-Seq data analysis	2
Galaxy Introduction	8
Variant Analysis	2
Analysis of sequences	6
Tutorials using public Galaxy at usegalaxy.org	7

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	4
Train the trainers	6
Development in Galaxy	13

One GitHub repository to collect everything



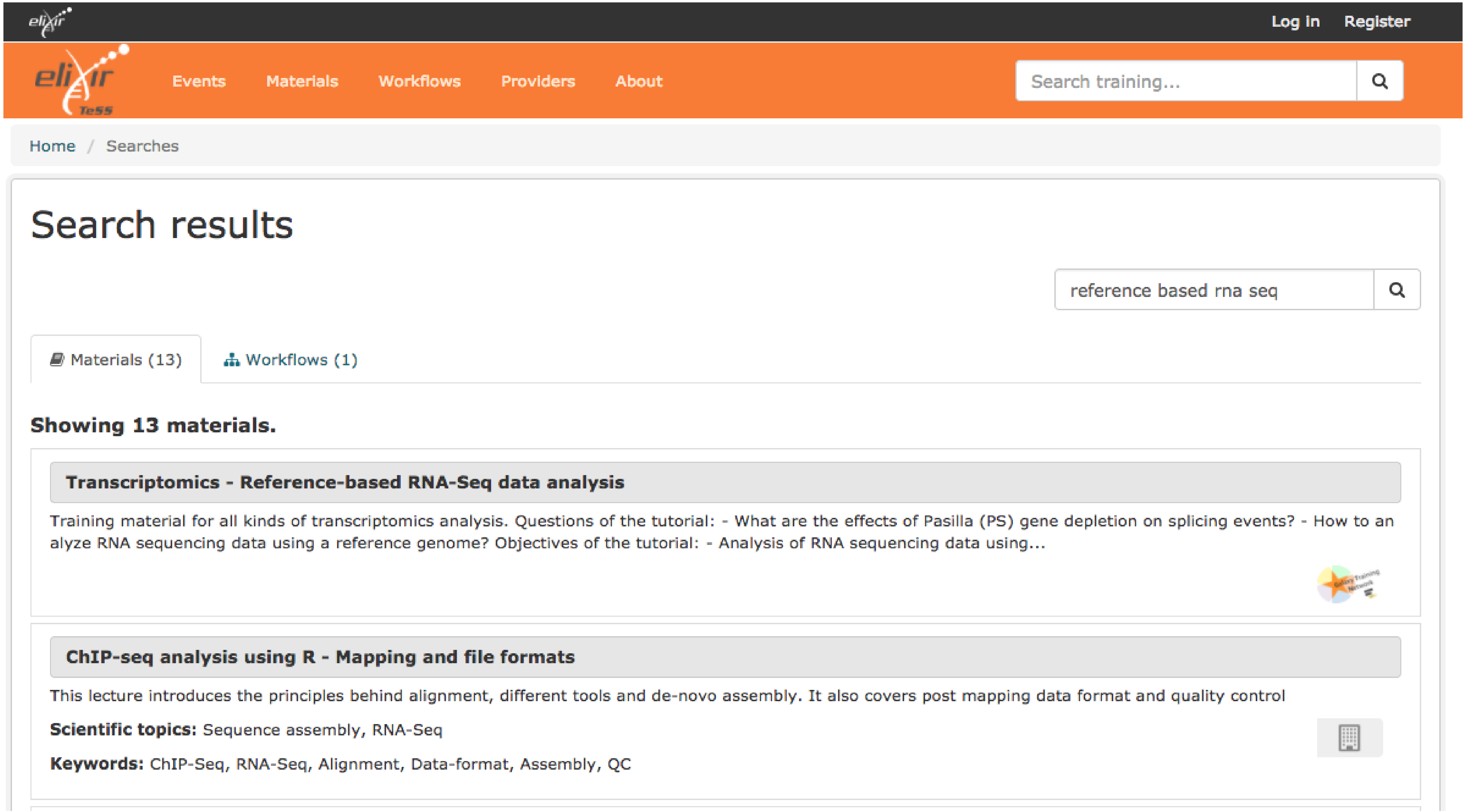
Findable

Accessible

Interoperable

Reusable

Findable material via TeSS



Home / Searches

Search results

reference based rna seq

Materials (13) Workflows (1)

Showing 13 materials.

Transcriptomics - Reference-based RNA-Seq data analysis

Training material for all kinds of transcriptomics analysis. Questions of the tutorial: - What are the effects of Pasilla (PS) gene depletion on splicing events? - How to analyze RNA sequencing data using a reference genome? Objectives of the tutorial: - Analysis of RNA sequencing data using...

ChIP-seq analysis using R - Mapping and file formats

This lecture introduces the principles behind alignment, different tools and de-novo assembly. It also covers post mapping data format and quality control

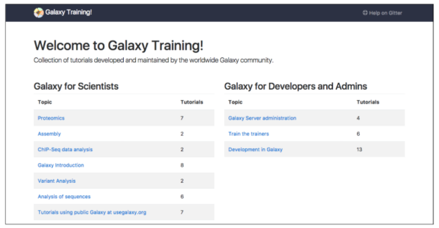
Scientific topics: Sequence assembly, RNA-Seq

Keywords: ChIP-Seq, RNA-Seq, Alignment, Data-format, Assembly, QC

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

An open, collaborative, FAIR, online infrastructure

Separation between content and form



Galaxy Training!

Help on Gitter

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community.

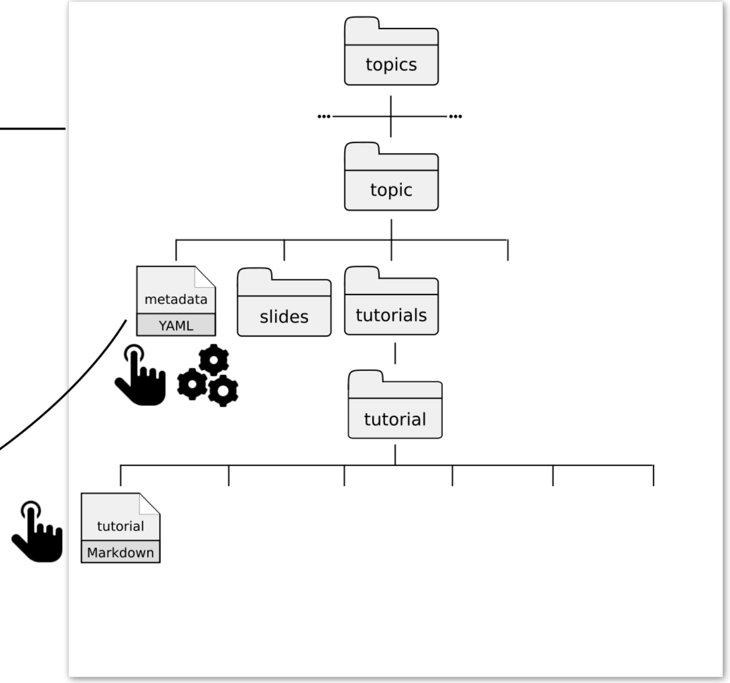
Galaxy for Scientists

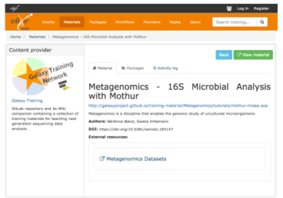
Topic	Tutorials
Proteomics	7
Assembly	2
ChIP-Seq data analysis	2
Galaxy Introduction	8
Variant Analysis	2
Analysis of sequences	6
Tutorials using public Galaxy at usegalaxy.org	7

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	4
Train the trainers	6
Development in Galaxy	13

One GitHub repository to collect everything





Content provider

Galaxy Training

GitLab repository and its Wiki companion containing a collection of training materials for teaching next generation sequencing data analysis.

Metagenomics - 16S Microbial Analysis with Mothur

http://galaxyproject.github.io/training-material/Metagenomics/tutorials/mothur-miseq-sop

Metagenomics is a discipline that enables the genomic study of uncultured microorganisms

Authors: Bérénice Batut, Saskia Hiltemann

DOI: https://doi.org/10.5281/zenodo.165147

External resources:

Metagenomics Datasets

Population of external bioinformatic training resources

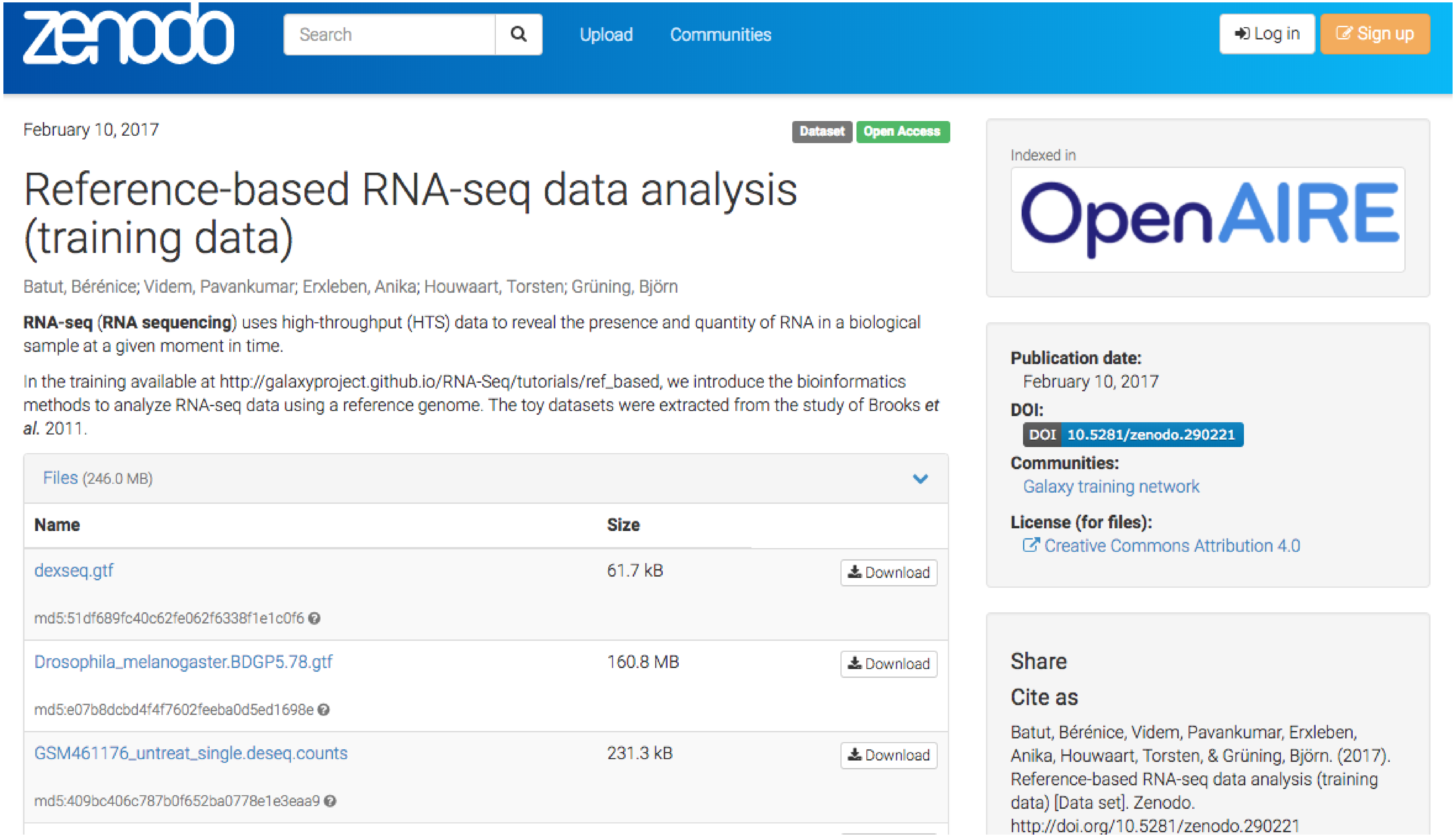
Findable

Accessible

Interoperable

Reusable

Storage of training datasets on Zenodo



February 10, 2017

Reference-based RNA-seq data analysis (training data)

Batut, Bérénice; Videm, Pavankumar; Erxleben, Anika; Houwaart, Torsten; Grüning, Björn

RNA-seq (**RNA sequencing**) uses high-throughput (HTS) data to reveal the presence and quantity of RNA in a biological sample at a given moment in time.

In the training available at http://galaxyproject.github.io/RNA-Seq/tutorials/ref_based, we introduce the bioinformatics methods to analyze RNA-seq data using a reference genome. The toy datasets were extracted from the study of Brooks *et al.* 2011.

Files (246.0 MB)

Name	Size
dexseq.gtf md5:51df689fc40c62fe062f6338f1e1c0f6	61.7 kB
Drosophila_melanogaster.BDGP5.78.gtf md5:e07b8dcbd4f4f7602feeba0d5ed1698e	160.8 MB
GSM461176_untreat_single.deseq.counts md5:409bc406c787b0f652ba0778e1e3eaa9	231.3 kB

Publication date:
February 10, 2017

DOI:
DOI 10.5281/zenodo.290221

Communities:
Galaxy training network

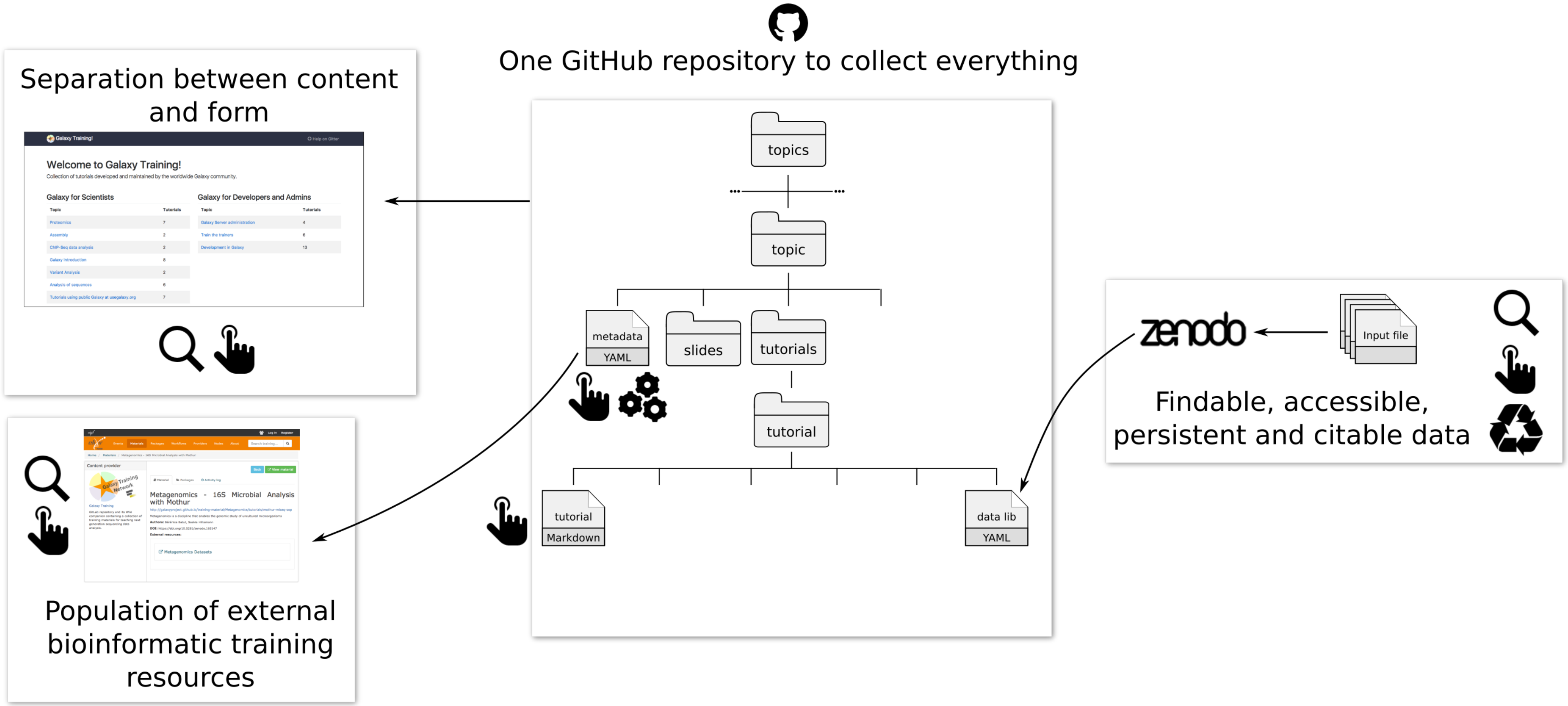
License (for files):
Creative Commons Attribution 4.0

Cite as

Batut, Bérénice, Videm, Pavankumar, Erxleben, Anika, Houwaart, Torsten, & Grüning, Björn. (2017). Reference-based RNA-seq data analysis (training data) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.290221

https://zenodo.org/

An open, collaborative, FAIR, online infrastructure



Separation between content and form

Galaxy Training!

Help on Gitter

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community.

Galaxy for Scientists

Topic	Tutorials
Proteomics	7
Assembly	2
ChIP-Seq data analysis	2
Galaxy Introduction	8
Variant Analysis	2
Analysis of sequences	6
Tutorials using public Galaxy at usegalaxy.org	7

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	4
Train the trainers	6
Development in Galaxy	13

Population of external bioinformatic training resources

Log in Register

Events Materials Packages Workflows Providers Nodes About Search training...

Home / Materials / Metagenomics - 16S Microbial Analysis with Mothur

Content provider

Galaxy Training

GitLab repository and its Wiki companion containing a collection of training materials for teaching next generation sequencing data analysis.

Back View material

Material Packages Activity log

Metagenomics - 16S Microbial Analysis with Mothur

http://galaxyproject.github.io/training-material/Metagenomics/tutorials/mothur-miseq-sop

Metagenomics is a discipline that enables the genomic study of uncultured microorganisms

Authors: Bérénice Batut, Saskia Hiltemann

DOI: https://doi.org/10.5281/zenodo.165147

External resources:

Metagenomics Datasets

One GitHub repository to collect everything

Findable, accessible, persistent and citable data

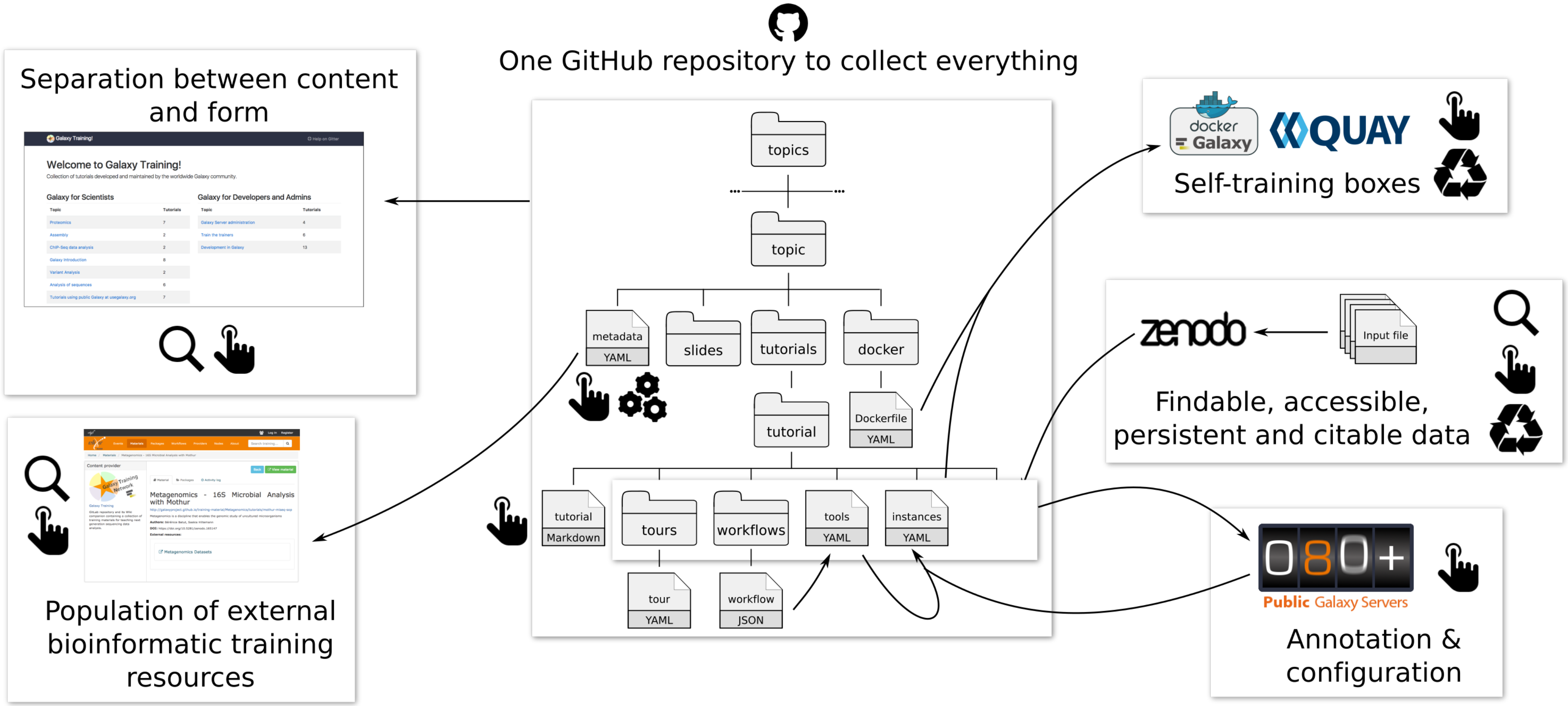
🔍 Findable

👆 Accessible

⚙ Interoperable

♻ Reusable

An open, collaborative, FAIR, online infrastructure



Separation between content and form

Galaxy Training! · Help on Gitter

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community.

Galaxy for Scientists

Topic	Tutorials
Proteomics	7
Assembly	2
ChIP-Seq data analysis	2
Galaxy Introduction	8
Variant Analysis	2
Analysis of sequences	6
Tutorials using public Galaxy at usegalaxy.org	7

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	4
Train the trainers	6
Development in Galaxy	13

Population of external bioinformatic training resources

Home / Materials / Metagenomics - 16S Microbial Analysis with Mothur

Content provider

Galaxy Training

GitLab repository and its Wiki companion containing a collection of training materials for teaching next generation sequencing data analysis.

Metagenomics - 16S Microbial Analysis with Mothur

http://galaxyproject.github.io/training-material/Metagenomics/tutorials/mothur-miseq-sop

Metagenomics is a discipline that enables the genomic study of uncultured microorganisms

Authors: Bérénice Batut, Saskia Hiltemann

DOI: https://doi.org/10.5281/zenodo.165147

External resources:

Metagenomics Datasets

One GitHub repository to collect everything

Self-training boxes

Findable, accessible, persistent and citable data

Annotation & configuration

Findable

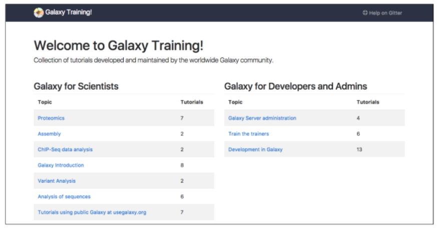
Accessible

Interoperable

Reusable

An open, collaborative, FAIR, online infrastructure

Separation between content and form



Galaxy Training!

Help on Gitter

Welcome to Galaxy Training!

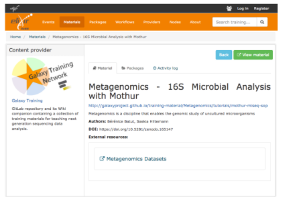
Collection of tutorials developed and maintained by the worldwide Galaxy community.

Galaxy for Scientists

Topic	Tutorials
Proteomics	7
Assembly	2
ChIP-Seq data analysis	2
Galaxy Introduction	8
Variant Analysis	2
Analysis of sequences	6
Tutorials using public Galaxy at usegalaxy.org	7

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	4
Train the trainers	6
Development in Galaxy	13



Home / Materials / Metagenomics - 16S Microbial Analysis with Mothur

Content provider

Galaxy Training

GitLab repository and its Wiki companion containing a collection of training materials for teaching next generation sequencing data analysis.

Metagenomics - 16S Microbial Analysis with Mothur

http://galaxyproject.github.io/training-material/Metagenomics/tutorials/mothur-miseq-sop

Metagenomics is a discipline that enables the genomic study of uncultured microorganisms

Authors: Bérénice Batut, Saskia Hiltemann

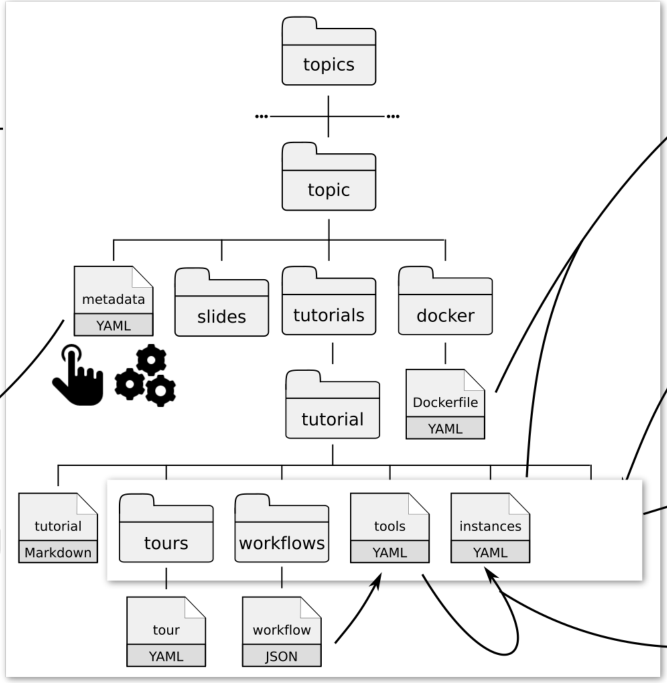
DOI: https://doi.org/10.5281/zenodo.165147

External resources:

Metagenomics Datasets

Population of external bioinformatic training resources

One GitHub repository to collect everything



docker Galaxy QUAY

Self-training boxes

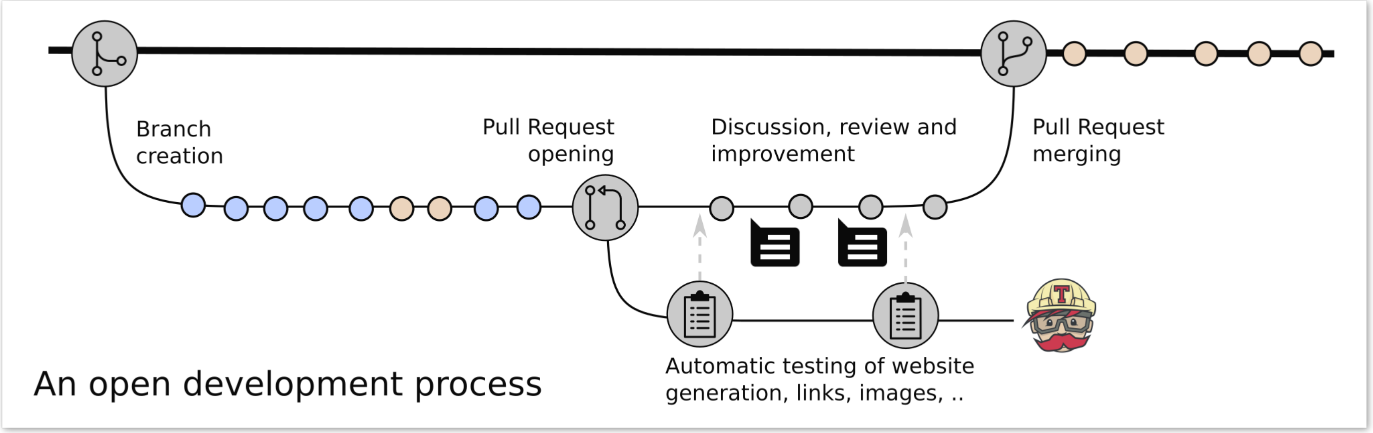
zenodo

Input file

Findable, accessible, persistent and citable data

080+ Public Galaxy Servers

Annotation & configuration

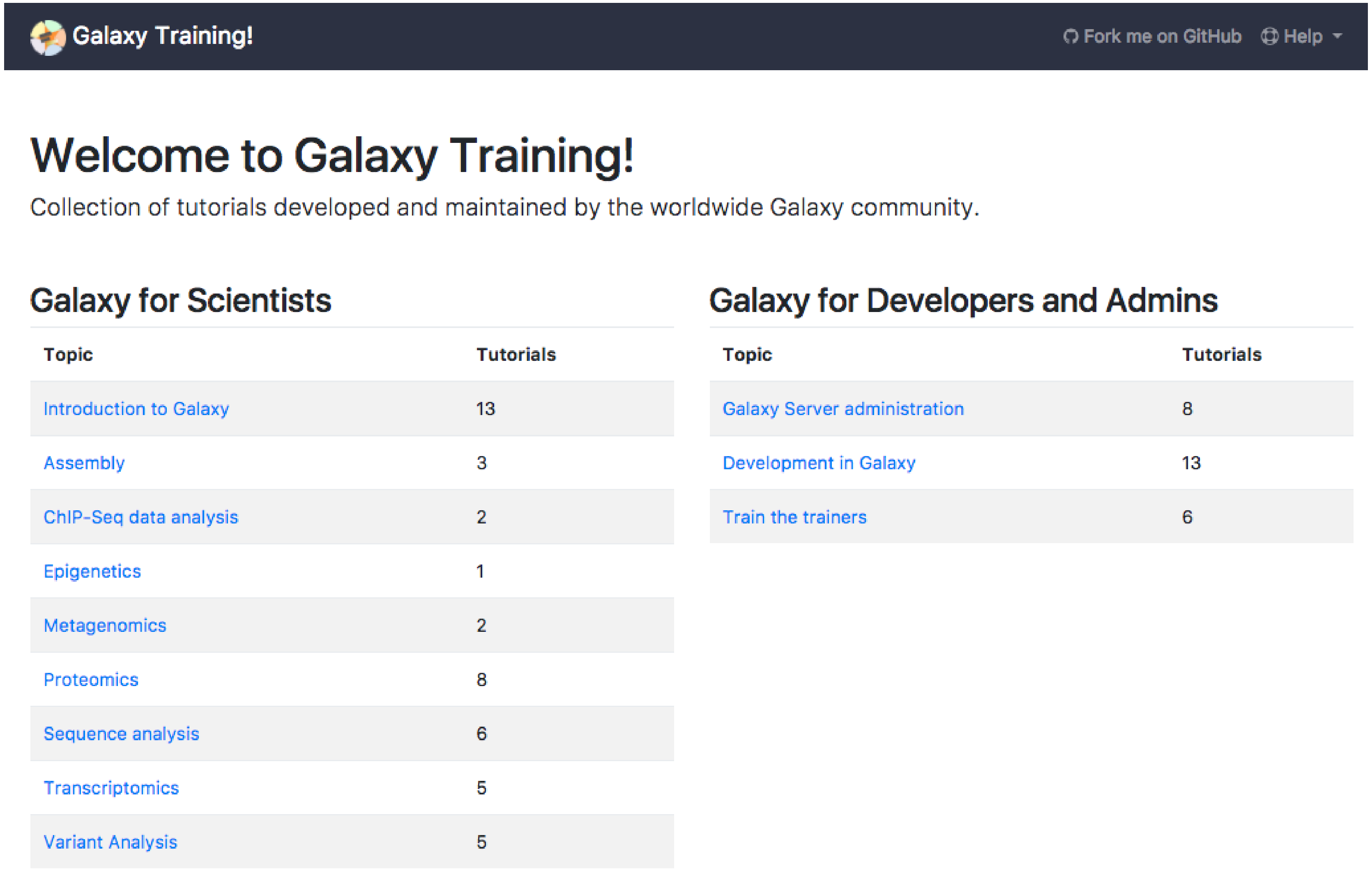


Branch creation
Pull Request opening
Discussion, review and improvement
Pull Request merging
Automatic testing of website generation, links, images, ..

An open development process

- Findable
- Accessible
- Interoperable
- Reusable

Galaxy Training materials



Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community.

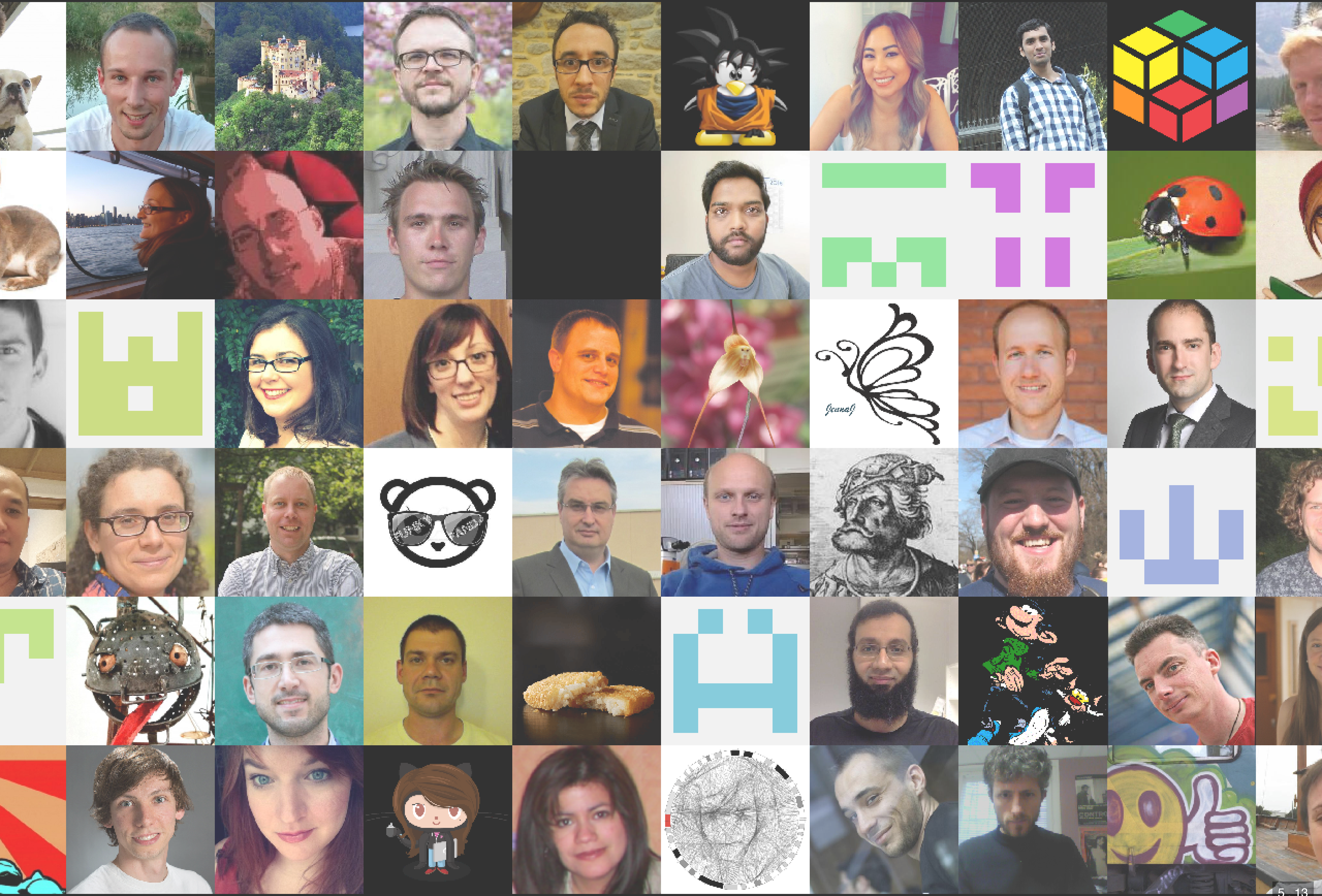
Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy	13
Assembly	3
ChIP-Seq data analysis	2
Epigenetics	1
Metagenomics	2
Proteomics	8
Sequence analysis	6
Transcriptomics	5
Variant Analysis	5

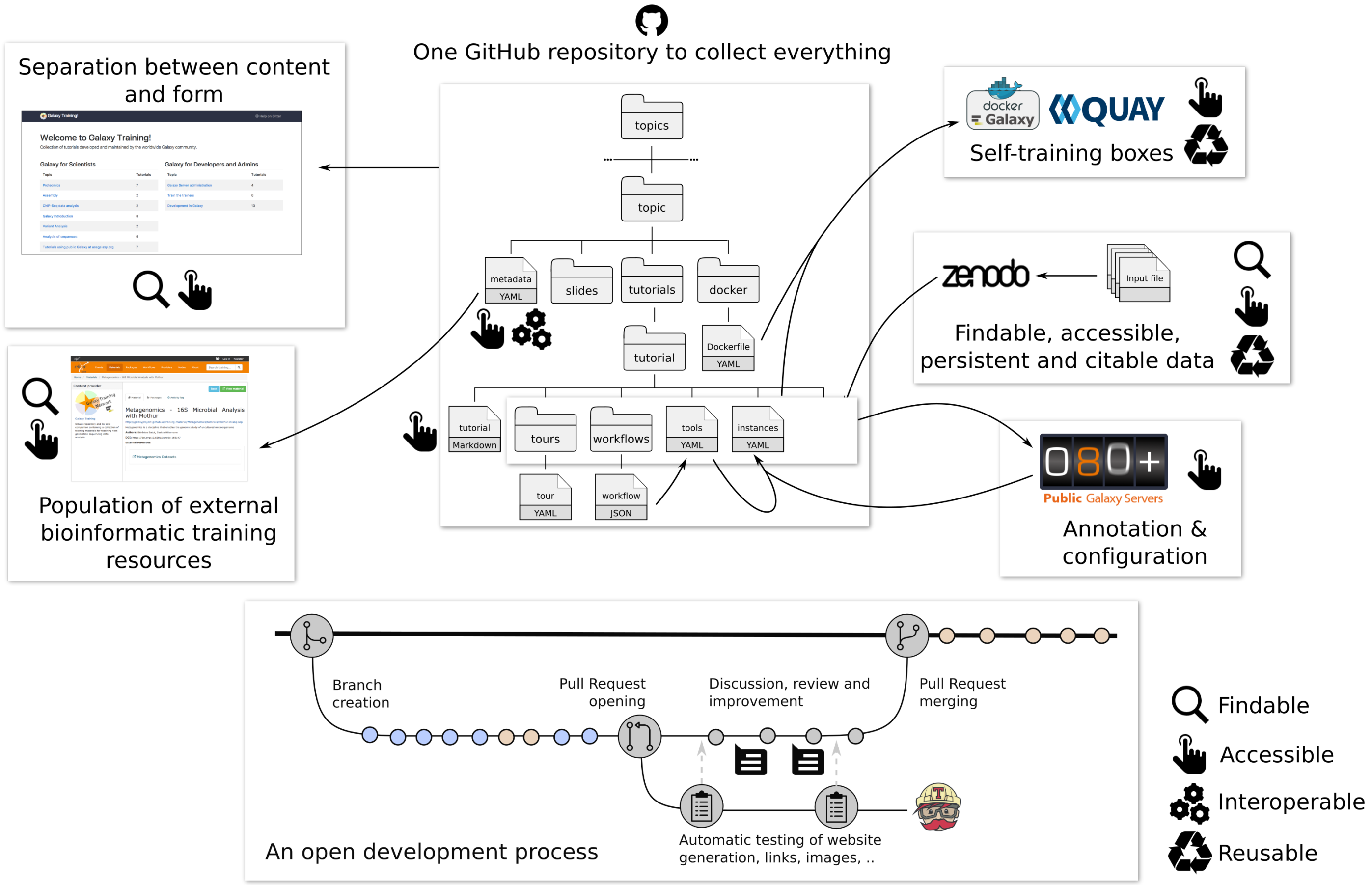
Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	8
Development in Galaxy	13
Train the trainers	6

http://training.galaxyproject.org



Thank you!


One GitHub repository to collect everything
Separation between content and form
Welcome to Galaxy Training!
Galaxy for Scientists
Galaxy for Developers and Admins
Population of external bioinformatic training resources
Metagenomics - 16S Microbial Analysis with Mothur
topics
topic
metadata
YAML
slides
tutorials
docker
tutorial
Dockerfile
YAML
tutorial
Markdown
tours
workflows
tools
YAML
instances
YAML
tour
YAML
workflow
JSON
docker
Galaxy
QUAY
Self-training boxes
zenodo
Input file
Findable, accessible, persistent and citable data
080+
Public Galaxy Servers
Annotation & configuration
Branch creation
Pull Request opening
Discussion, review and improvement
Pull Request merging
An open development process
Automatic testing of website generation, links, images, ..
Findable
Accessible
Interoperable
Reusable

An open, collaborative, FAIR, online infrastructure

training.galaxyproject.org

github.com/galaxyproject/training-material