

Building an open, collaborative, online infrastructure for bioinformatics training



**Bérénice Batut, Galaxy Training Network,
Dave Clements, Björn Grüning**

Galaxy Community Conference
June 2017

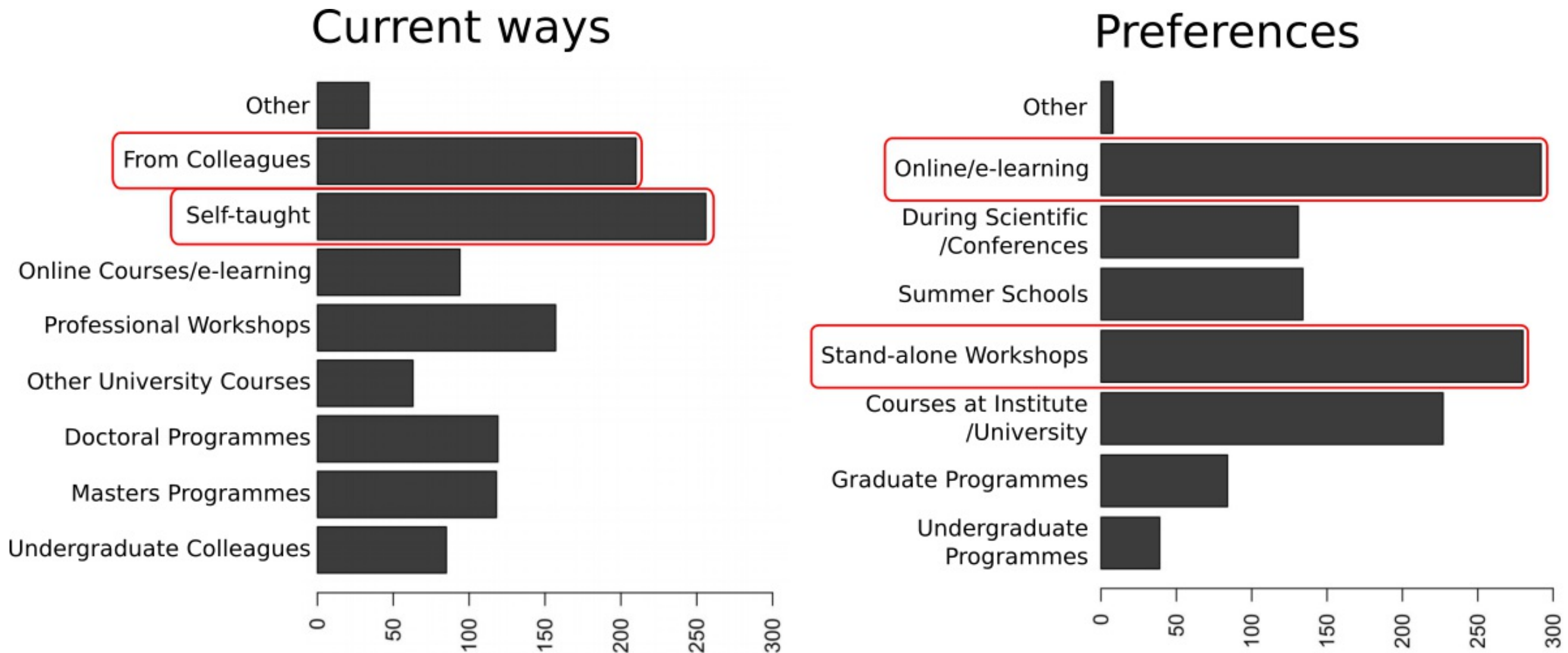
Why caring about bioinformatics training?

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!

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-hand navigation pane, a central tool configuration area, and a right-hand history pane.

Left-hand navigation pane: Contains a search bar and a list of tool categories 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 tool configuration area: Titled 'Diamond alignment tool for short sequences against a protein database (Galaxy Version 0.8.24)', it includes several configuration options:

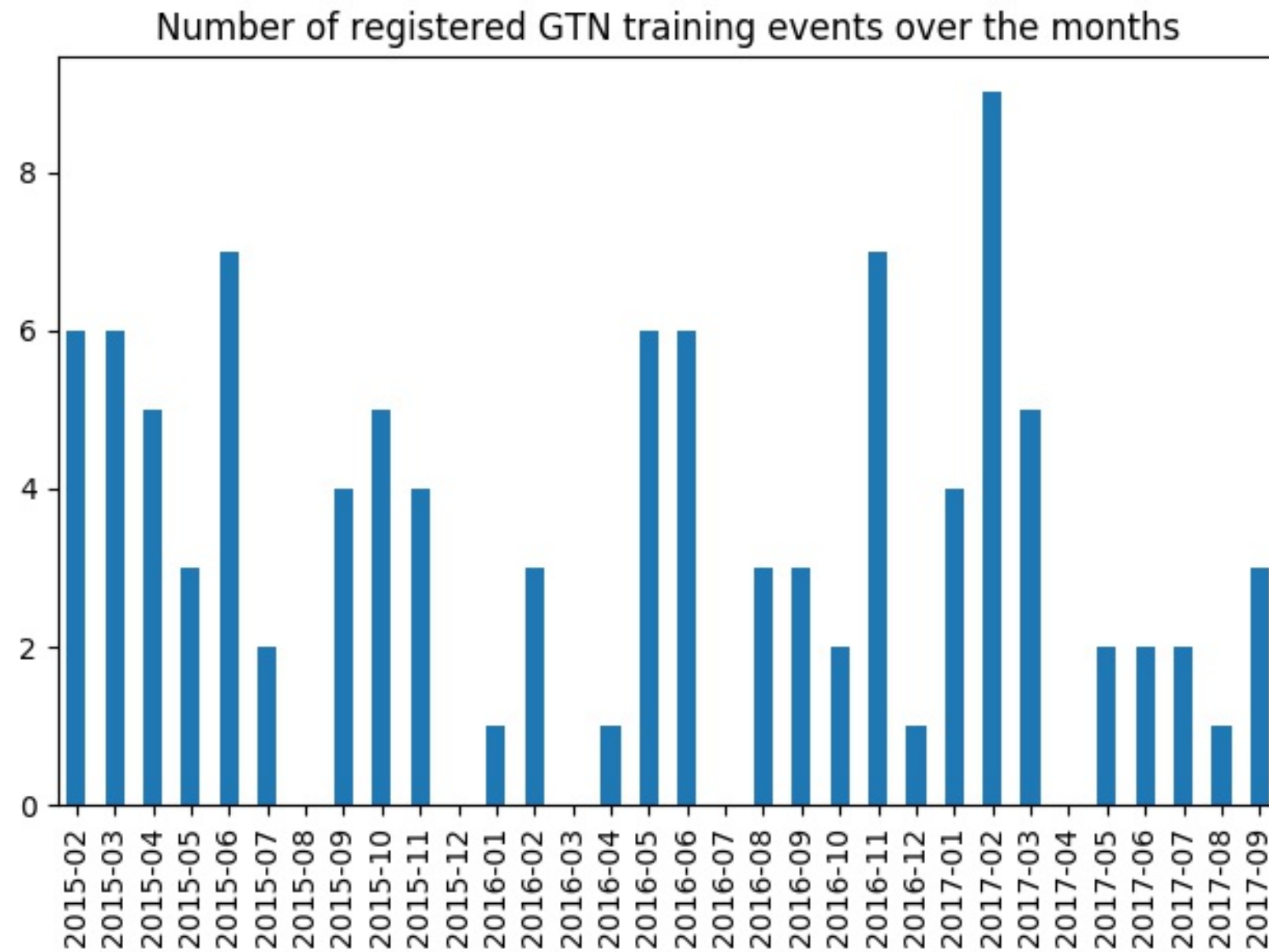
- What do you want to align?**: A dropdown menu set to 'Align amino acid query sequences (blastp)'. Below it, the command line option is shown as '(--blastp/--blastx)'.
- Input query file in FASTA or FASTQ format**: A text input field with a file upload icon. Below it, the command line option is shown as '(--query)'.
- 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'. Below it, the command line option is shown as '(--use-built-in-index)'.
- Select a reference genome**: A dropdown menu set to 'No options available'. Below it, a note states: 'If your genome of interest is not listed, contact your Galaxy admin'.
- Genetic code used for translation of query in BLASTX mode**: A dropdown menu set to 'The Standard Code'. Below it, the command line option is shown as '(--query-gencode)'.
- Format of output file**: A dropdown menu set to 'BLAST XML'. Below it, the command line option is shown as '(--outfmt)'.
- Include full length subject titles in output?**: A toggle switch set to 'Yes'. Below it, the command line option is shown as '(--salltitles)'.
- Trigger the sensitive alignment mode with a 16x9 seed shape configuration?**: A toggle switch set to 'Yes'. Below it, the command line option is shown as '(--sensitive)'.
- Trigger the more sensitive mode?**: A toggle switch set to 'Yes'. Below it, a note states: 'This mode provides some additional sensitivity compared to the sensitive mode. (--more-sensitive)'.
- Gap open penalty**: A text input field set to '11'. Below it, the command line option is shown as '(--gapopen)'.
- Gap extension penalty**: A text input field set to '1'. Below it, the command line option is shown as '(--gapextend)'.
- Scoring matrix**: A dropdown menu set to 'BLOSUM62 ((6-11)/2; (9-13)/1)'. Below it, a note states: 'In brackets are the supported values for (gap open)/(gap extend) (--matrix)'.
- Enable SEG masking of low complexity segments in the query?**: A toggle switch set to 'Yes'. Below it, the command line option is shown as '(--seg)'.
- Method to restrict the number of hits?**: A text input field.

Right-hand history pane: Titled 'History', it contains a search bar and a section for 'Unnamed history (empty)'. A blue box with an information icon 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



Quite a lot of events...



Worldwide!





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

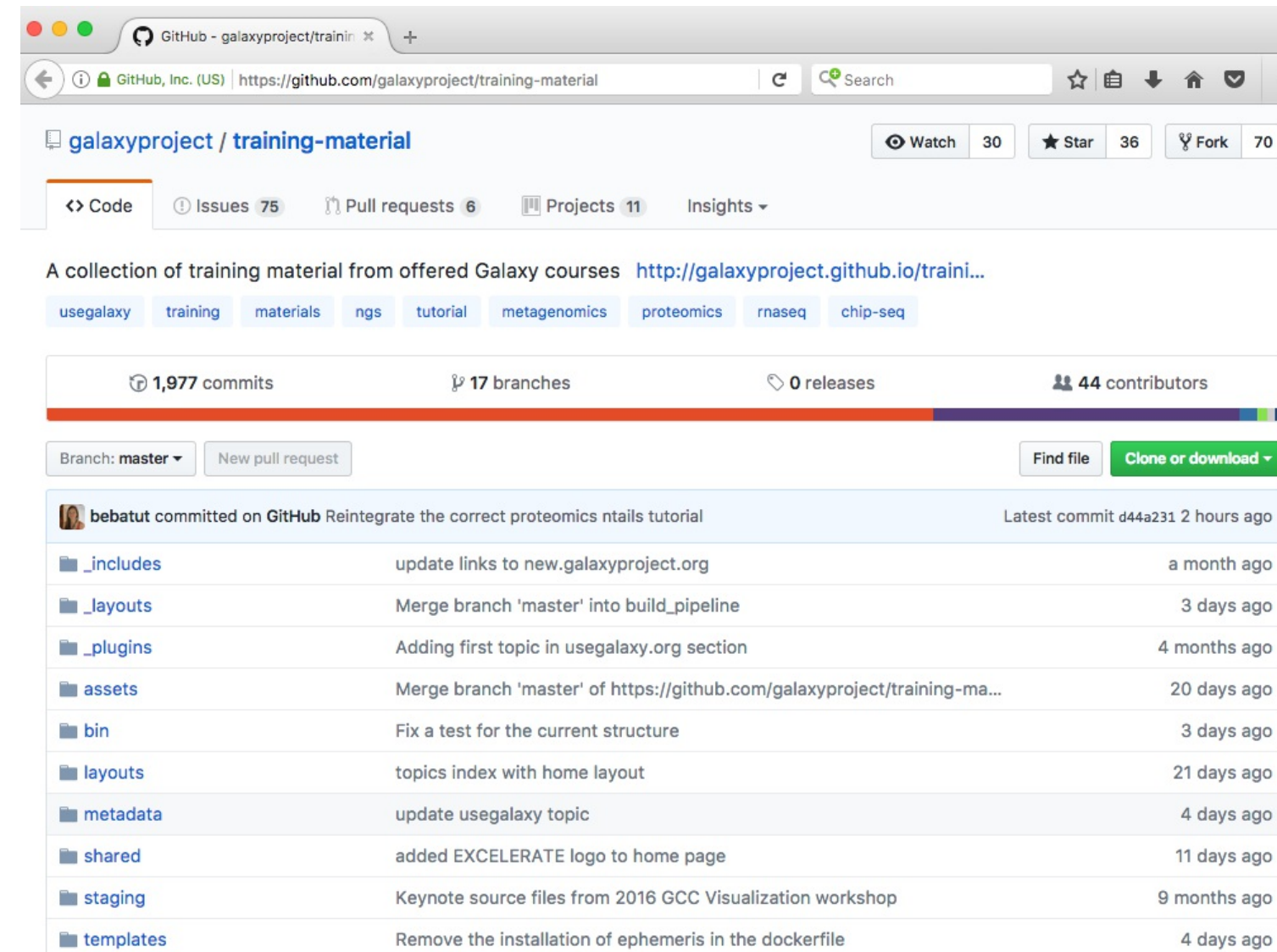
Requirements

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



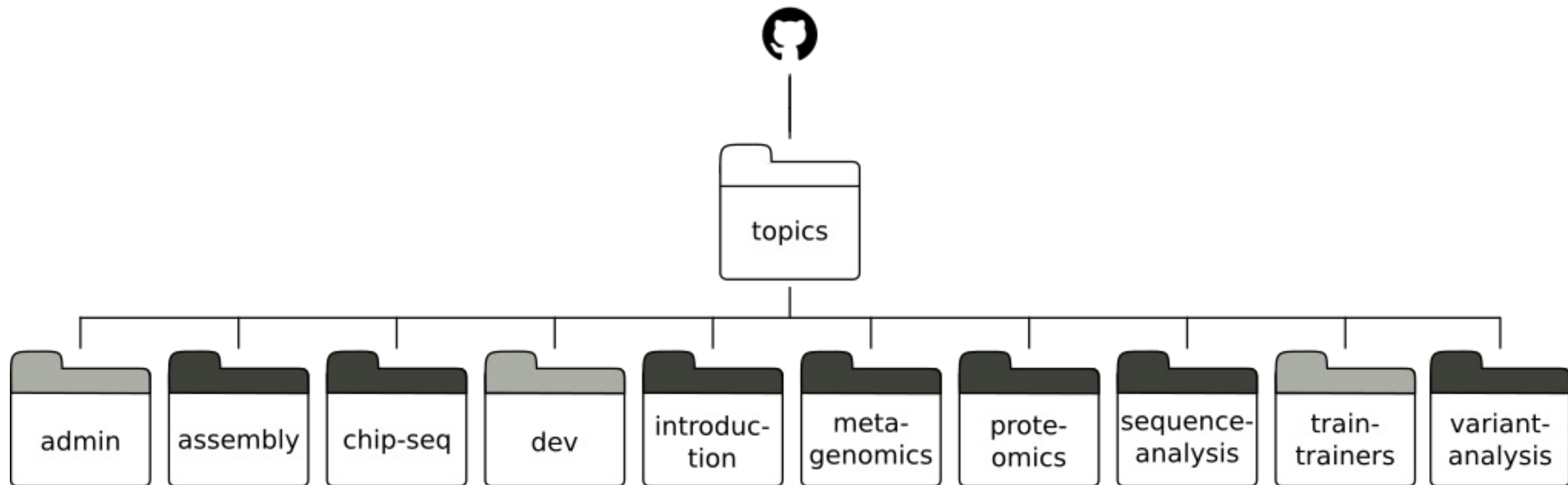
The model

One repository to collect everything

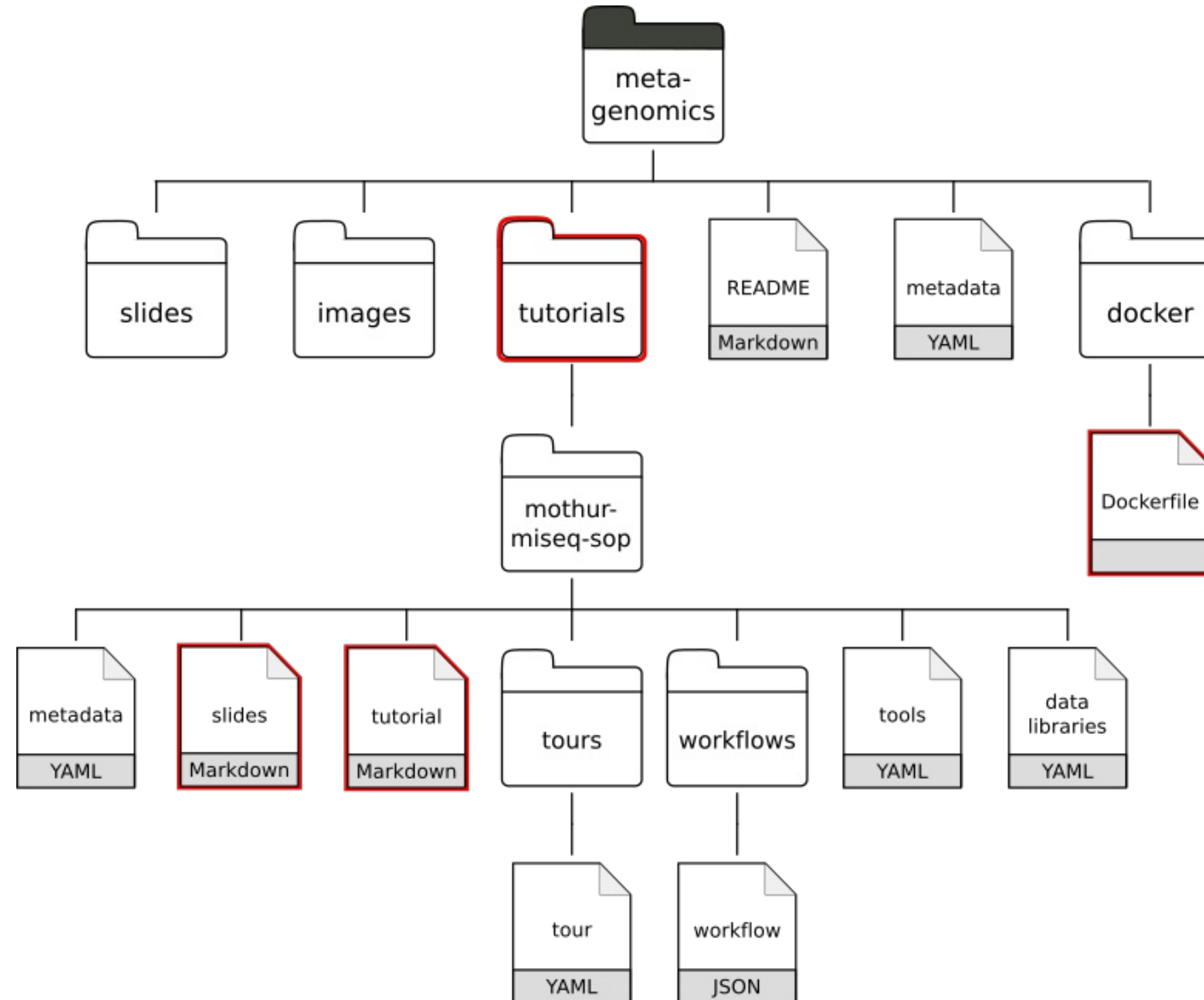


GitHub: [galaxyproject/training-material](https://github.com/galaxyproject/training-material)

Topics for different targeted users



Similar structure, content and formats

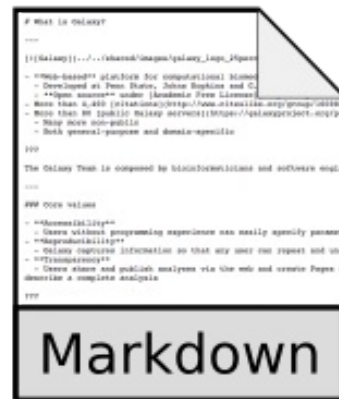


Separation between content and formatting

tutorials



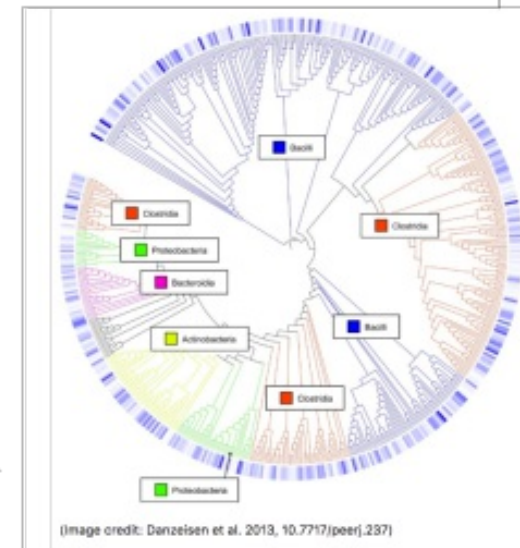
slides



metadata



Templating system



Hands-on: Cluster mock sequences into OTUs

First we calculate the pairwise distances between our sequences

- `Dist.seqs` with the following parameters
 - "fasta" to the fasta from Get.groups
 - "cutoff" to 0.20

Next we group sequences into OTUs

- `Cluster` with the following parameters
 - "column" to the dist output from Dist.seqs
 - "count" to the count table from Get.groups

Now we make a shared file that summarizes all our data into one handy table

- `Make.shared` with the following parameters
 - "list" to the OTU list from Cluster
 - "count" to the count table from Get.groups

Galaxy Training!

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

[Get an offer](#)

Metagenomics

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

Requirements

Before diving into this topic, we recommend you to have a look at:

- [Galaxy introduction](#)

Material

Lesson	Hands-on	Slides	Input dataset	Galaxy tour
16S Microbial Analysis with Mothur				
Analyzing whole genome sequencing data				

Contributors

This material is maintained by:

- [Béatrice Batut](#) (Beatrice.batut@gmail.com)
- [Saskia Hillemann](#) (saskia@gmail.com)

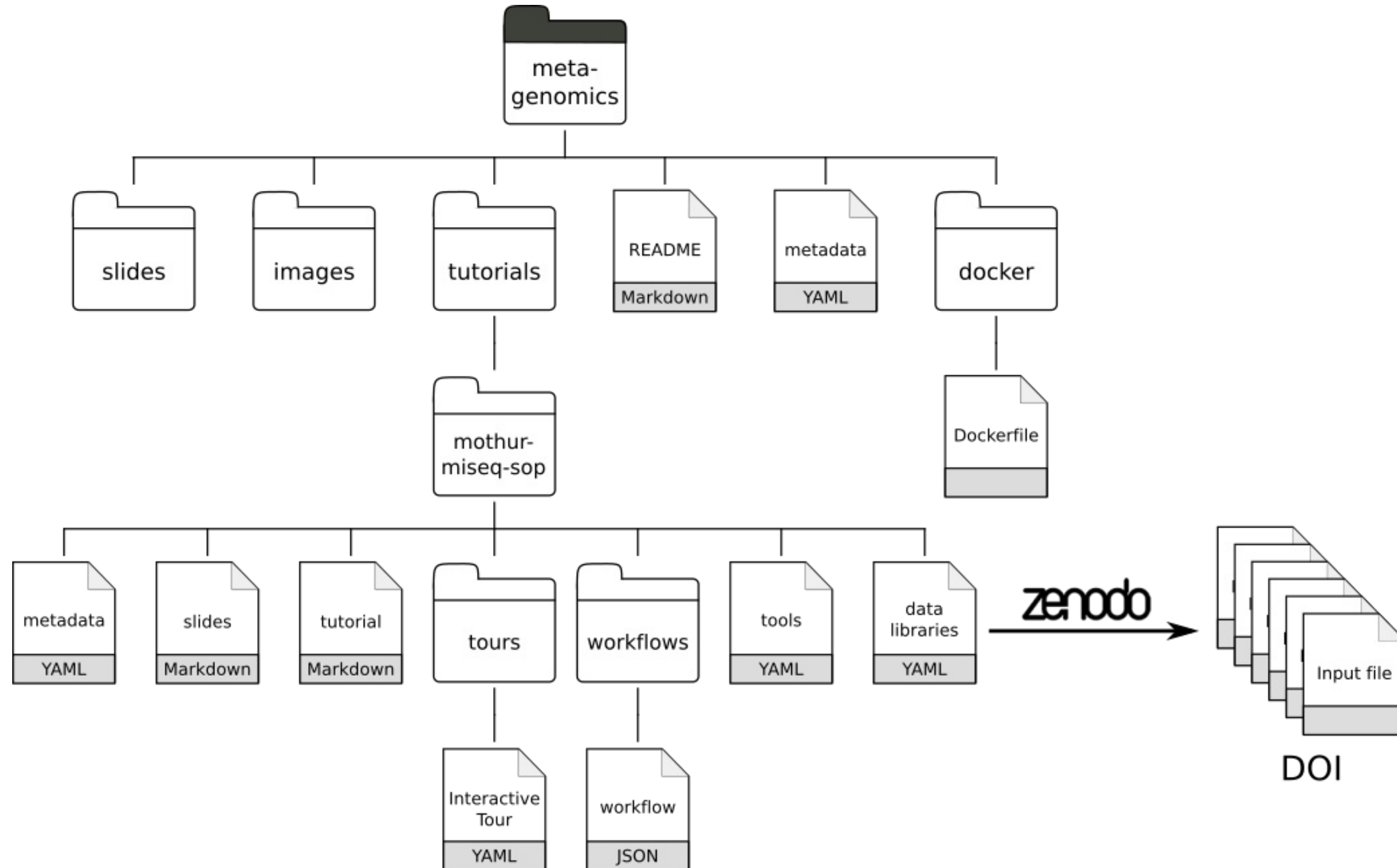
For any question related to this topic and the content, you can contact them.

Galaxy

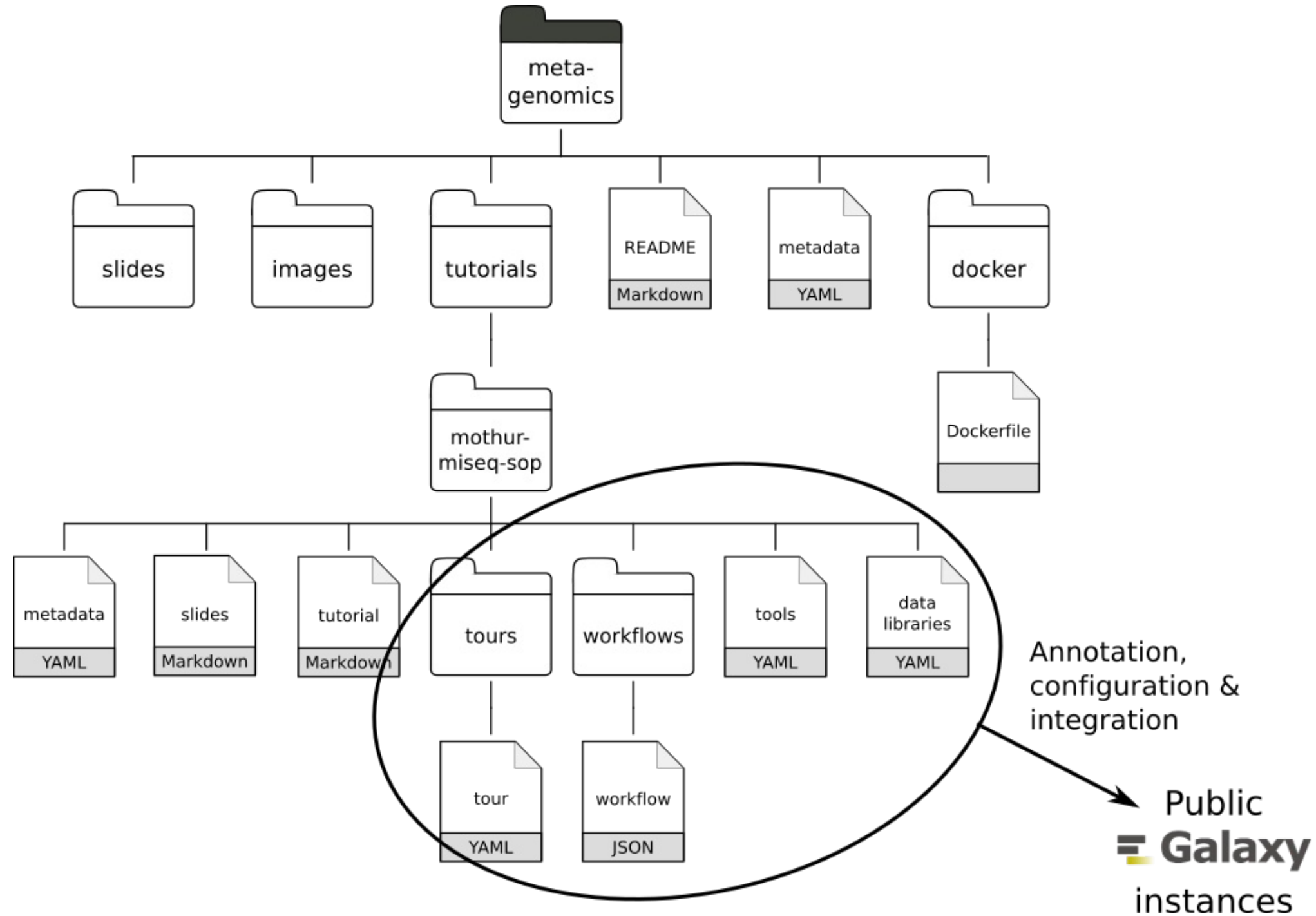
- **Web-based platform** for computational biomedical research
 - Developed at Penn State, Johns Hopkins and G. Washington universities with substantial outside contributions
 - **Open source** under [Academic Free License](#)
- **More than 4,400 citations**
- **More than 80 public Galaxy servers**
 - Many more non-public
 - Both general-purpose and domain-specific

5/26

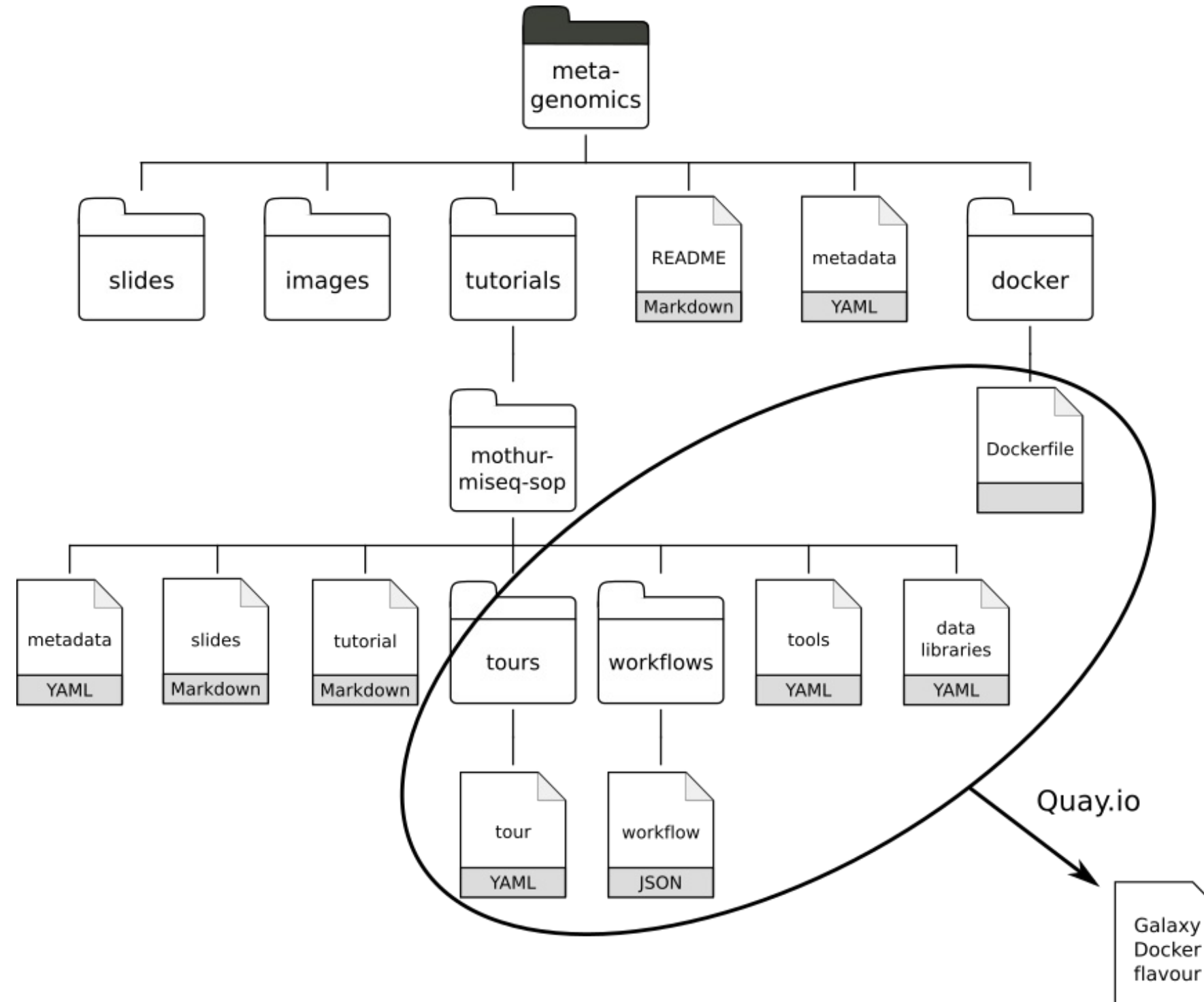
Citable data & Credit



Definition of the technical infrastructure



Definition of the technical infrastructure

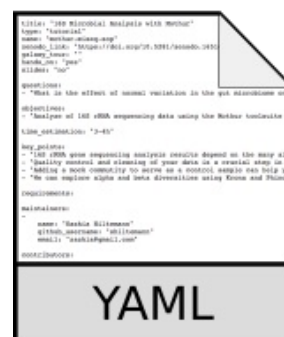


Findable **A**ccessible **I**nteroperable **R**eusable



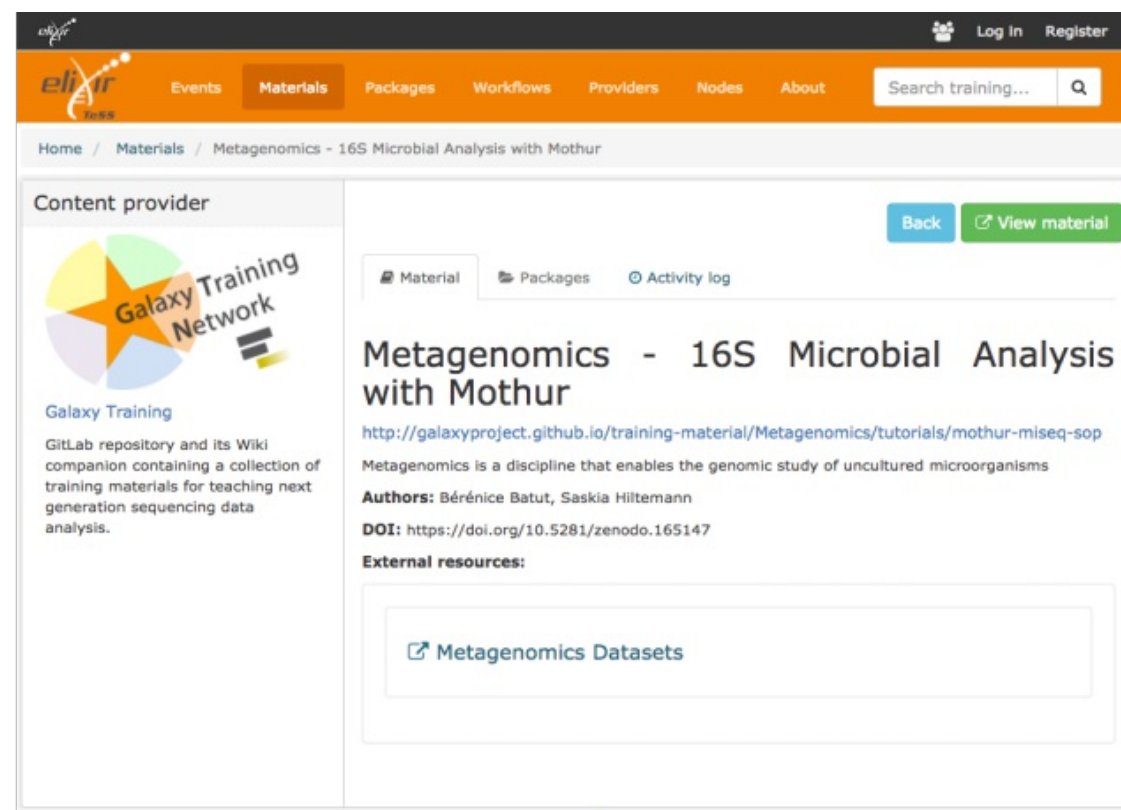
model

Findable



metadata

Automatically
added



TeSS:  Training Portal

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



Accessible

- Online

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

- Technical support
 - Self-training boxes with Galaxy Docker flavor
 - Annotated public Galaxy instances



Interoperable

- Metadata description in YAML and integrating EDAM ontology
- Content for different targets (workshops and self-training)
- Technical support for different platforms

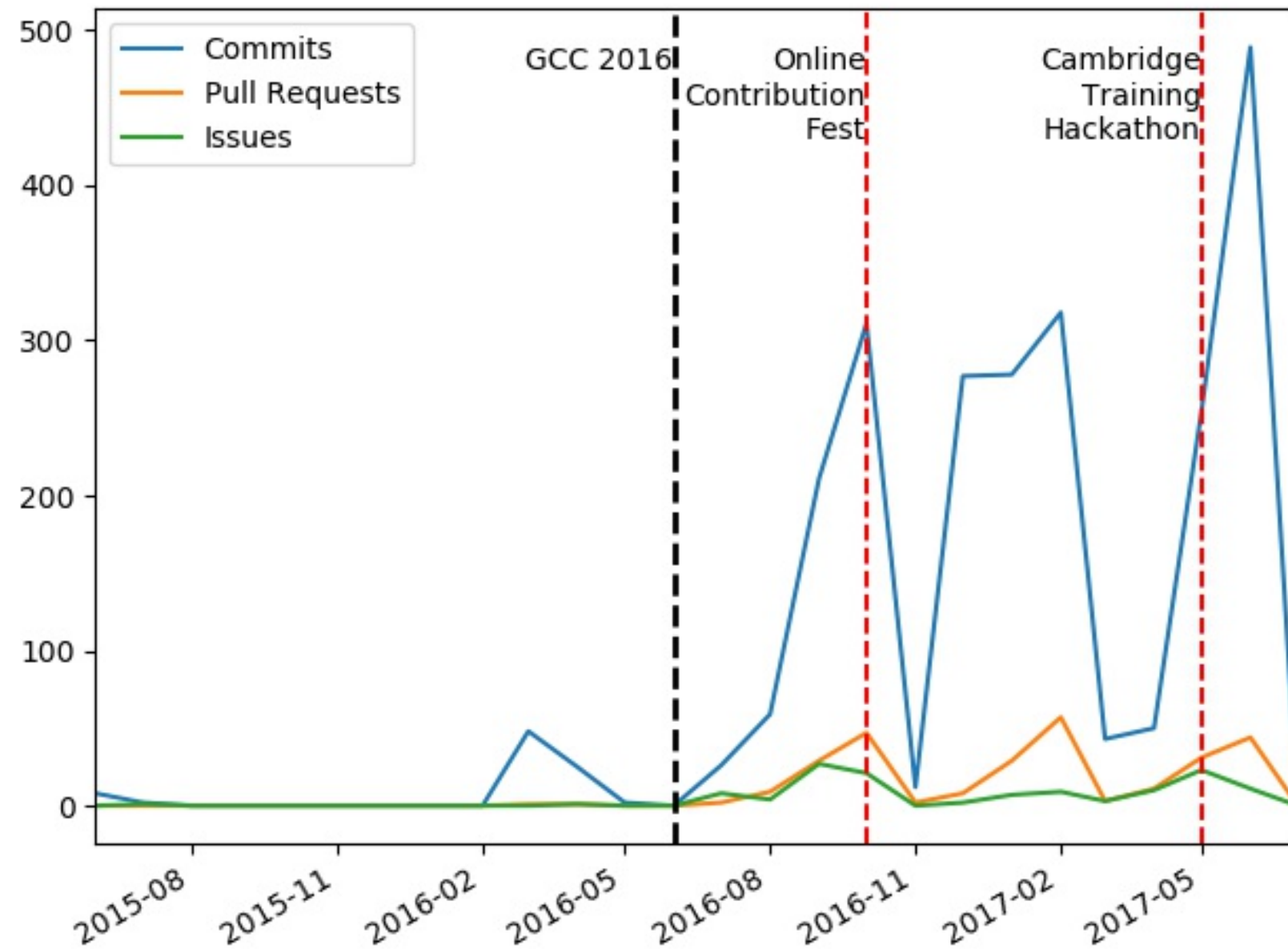


Reusable & Open

- [CC BY 4.0](#) license
- DOI for the input datasets
- Open development process on GitHub & via Gitter
- Open education movement


Community effort

Numerous contributions




And 2 successful hackathons!


Numerous discussions



Galaxy Training Network/Lobby


<https://new.galaxyproject.org/MailingLists/>






Gildas Le Corguillé @lecorguille

👍




John Chilton @jmchilton

I tried checking out build_pipeline and now only the usegalaxy tutorials show up in my training site
Not admin or dev tutorials at all - any ideas?




Björn Grüning @bgruening

This branch is not yet working ... @dannon needs to finish up his methal smith build magic




John Chilton @jmchilton

Are there slide content changes or just restructuring and metadata reorganization in that branch?




Dannon Baker @dannon

Yep, the build_ Blocked Plug-in contains a full reorganization of the content, which doesn't build quite yet. (but that's where we'll want new stuff for now)
I wouldn't guess there are likely conflicting slide content changes




Björn Grüning @bgruening

Afaik there are no content changes, just the build procedure and the organisation.




Victoria Dominguez del Angel @vdda

I'm with Berenice, please take care 🙏




Slugger70 @Slugger70

Hi all, Torsten Seemann came up with an idea after looking at the GTN website. He would like to see tags on the various tutorials for things like Eukaryotic vs prokaryotic specific, or virus etc etc... I reckon it's a good idea.




Björn Grüning @bgruening

👍
We need more tags, also for supported Galaxy instance etc ...




Slugger70 @Slugger70

I agree. Minimum Galaxy version at least.




Yvan Le Bras @yvanlebras

+1




Mallory Freeberg @malloryfreeberg

+1
Also willing to help with this 😊




Slugger70 @Slugger70

Could be something to add at the hackathon?




Mallory Freeberg @malloryfreeberg

Absolutely. I'll add it to suggested data hack topics



Mallory Freeberg @malloryfreeberg



Click here to type a chat message. Supports GitHub flavoured markdown.

May 31 16:01

Jun 01 21:11

Jun 01 21:12

Jun 01 21:18

Jun 01 21:18

Jun 01 21:22

Jun 02 16:15

Jun 05 12:29

Jun 05 12:29

Jun 05 12:32

Jun 05 12:50


Jun 05 16:04

Jun 05 16:07

Jun 05 16:08

Jun 05 16:16


PEOPLE



ADD


SEE ALL (50 PEOPLE)

ACTIVITY




shiltemann

 on general_metagenomics_tutorial update tutorial (compare) 01:04




nsoranzo

 commented #358 Jun 15




shiltemann

 on master Add authors of Introduction sli... Merge pull request #359 from ns... (compare) Jun 15




shiltemann

 closed #359 Jun 15




nsoranzo

 opened #359 Jun 15




shiltemann

 on fix-slides change slide deck type (compare) Jun 15




shiltemann

 synchronize #358 Jun 15




shiltemann

 opened #358 Jun 15




shiltemann

 on fix-slides change slide deck type (compare) Jun 15



nsoranzo

 commented #354 Jun 15



shiltemann

 on general_metagenomics_tutorial start updating amplicon part (compare) Jun 15

Gitter: Galaxy-Training-Network/Lobby

8 . 3

A constantly growing community

