Towards automating and publishing workflow analyses in Galaxy

Andrea Bagnacani

Dept. of Systems Biology & Bioinformatics
University of Rostock, Rostock (Germany)

Bielefeld, 24th October 2017

Summary

- From data to information
- The Galaxy framework
- A platform for training
- de.STAIR
- Conclusions

From data to information

Life Sciences have become more and more data-driven.

Insights on biological problems are gained leveraging on computational approaches:

- collecting data through experiments or simulations
- structuring data into data-sets
- analyzing data leveraging on multidisciplinary approaches
- sharing protocols and best practices to reproduce results

These approaches enable researchers to put data into context, and obtain workflows for further investigation and reproducibility.

From data to information

Life Sciences have become more and more data-driven.

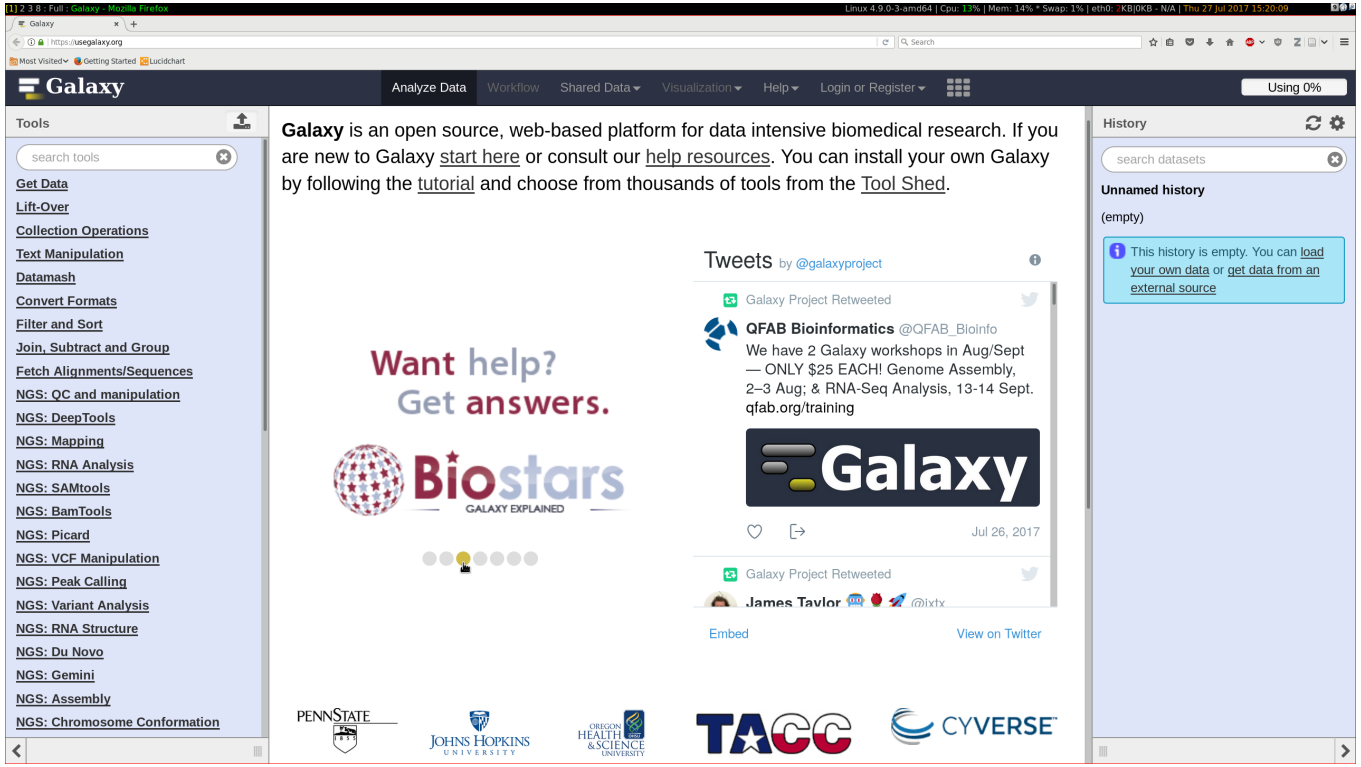
Insights on biological problems are gained leveraging on computational approaches:

- collecting data through experiments or simulations ✓
- structuring data into data-sets ✓
- analyzing data leveraging on multidisciplinary approaches ✓
- sharing protocols and best practices to reproduce results ✗

However, the novelty of such approaches requires an effort for the promotion of standards, and ease of availability and reproducibility.

The Galaxy framework

a web-based framework

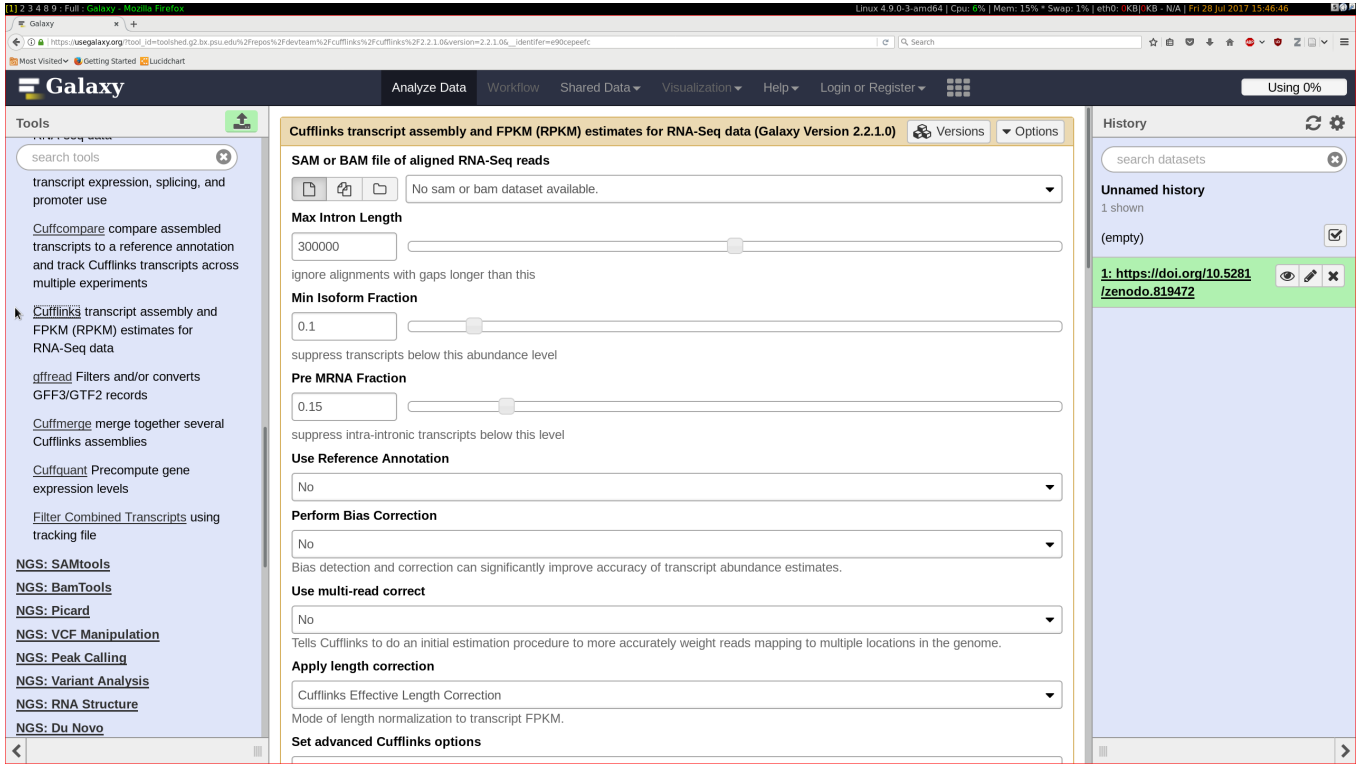


Galaxy is a web framework for computational bio/medical research

- public/private online/offline instances
- domain specific

The Galaxy framework

Operations

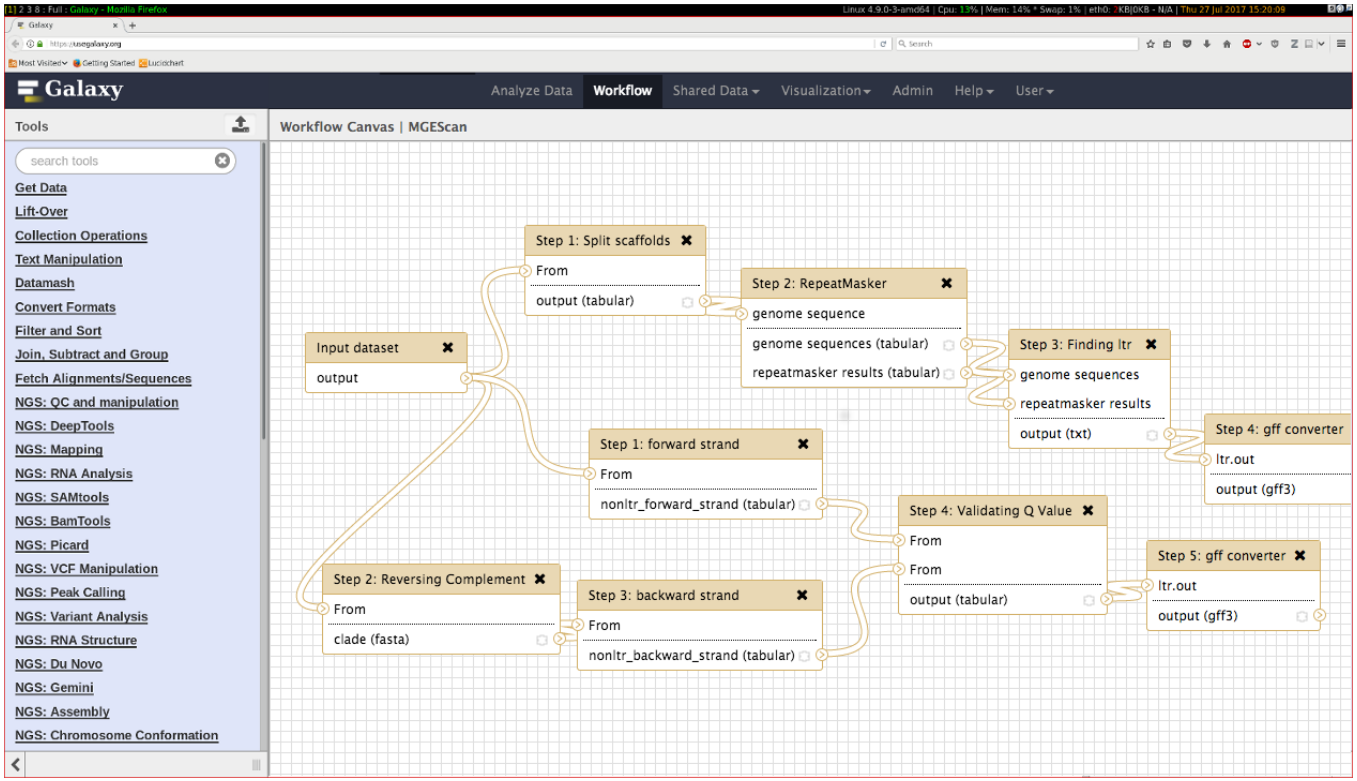


Galaxy is a web framework for computational bio/medical research

- tools can be searched, parametrized, and applied on a dataset
- a history of data and operations is kept for tracking/reuse

The Galaxy framework

Sharable workflows

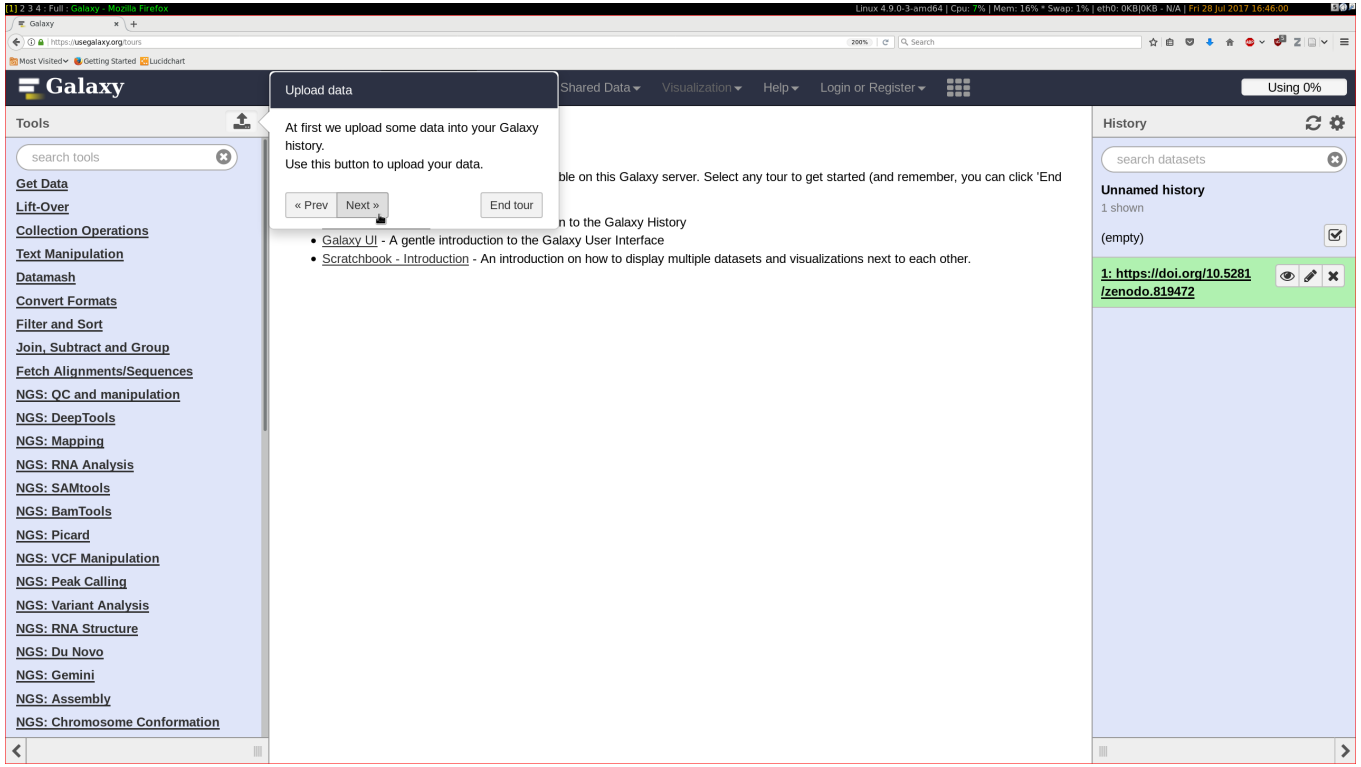


Galaxy is a web framework for computational bio/medical research

- histories can be exported to graphical workflows
- workflows can be shared, edited, and run for replication

The Galaxy framework

Interactive tours



Galaxy is a web framework for computational bio/medical research

- interactive tours can guide users through its interface
- tours can be created for any topic, for showcase or guidance

The Galaxy framework

Interactive tours - behind the scenes

Tours are YAML files, containing dialogues' text, position, and click actions

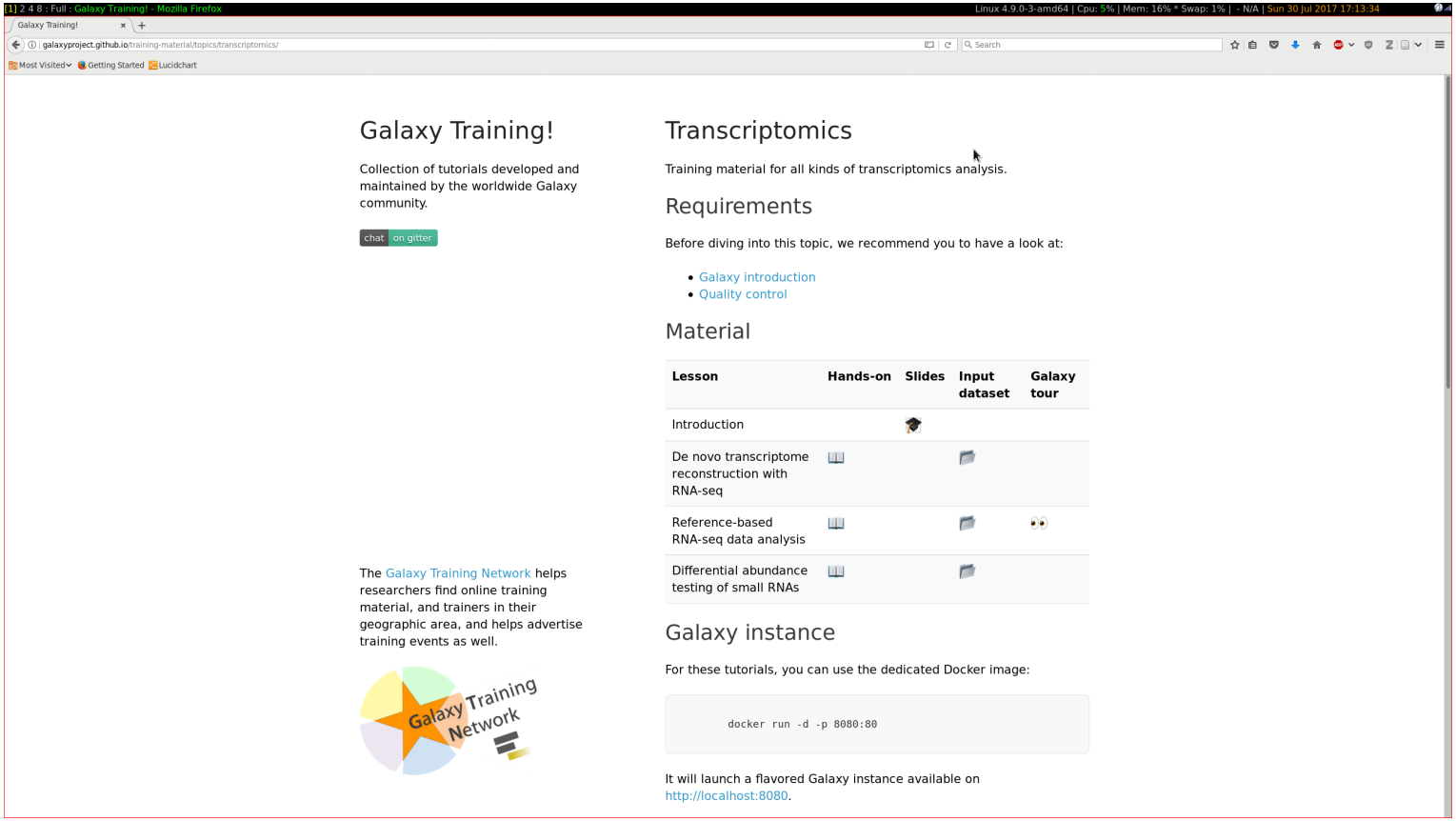
```yaml
id: galaxy_ui
name: Galaxy UI
description: A gentle introduction to the Galaxy User Interface
title_default: Welcome to Galaxy

# A tour is made of several steps, each of them beginning with a dash '-'
steps:
  # 'title's will be displayed in the header of each step-container
  # If you don't specify any title, a default title is used, defined above.
  - title: Welcome to Galaxy
    # 'content' is the actual text that is shown to the user
    content: This short tour guides you through the Galaxy user interface.<br/>
    backdrop: true

  # 'element' is the JQuery Selector of the element you want to describe
  - title: Upload your data
    element: ".upload-button"
    intro: Galaxy supports many ways to get in your data.<br>
           Use this button to upload your data.
    # position of the text box relative to the selected element
    position: right
    # You can trigger click() events on arbitrary elements before (preclick)
    # or after (postclick) the element is shown
    postclick:
      - .upload-button
```

A platform for training

The Galaxy Training Network is a collection of tutorials for researchers, developers, and admins leveraging on the Galaxy framework for deliver or provide answers to bio/medical research.

- Galaxy interface
- Proteomics
- Assembly
- ChIP-Seq data analysis
- Variant analysis
- Sequence analysis
- Transcriptomics
- Epigenetics
- Metagenomics
- Server administration
- Training the trainers
- Galaxy development

A platform for training



Each training topic contains multiple examples, available as:

- hands-on material with example datasets
- slides
- interactive tours

From where we start

Life Sciences have become more and more data-driven.

Insights on biological problems are gained leveraging on computational approaches:

- collecting data through experiments or simulations ✓
- structuring data into data-sets ✓
- analyzing data leveraging on multidisciplinary approaches ✓
- sharing protocols and best practices to reproduce results ✓

Galaxy is able to cover all these requirements for carrying out reproducuble best-practices in data-driven research.

However, new problems require new workflows, and the workflow collection might be too restrictive to address them.

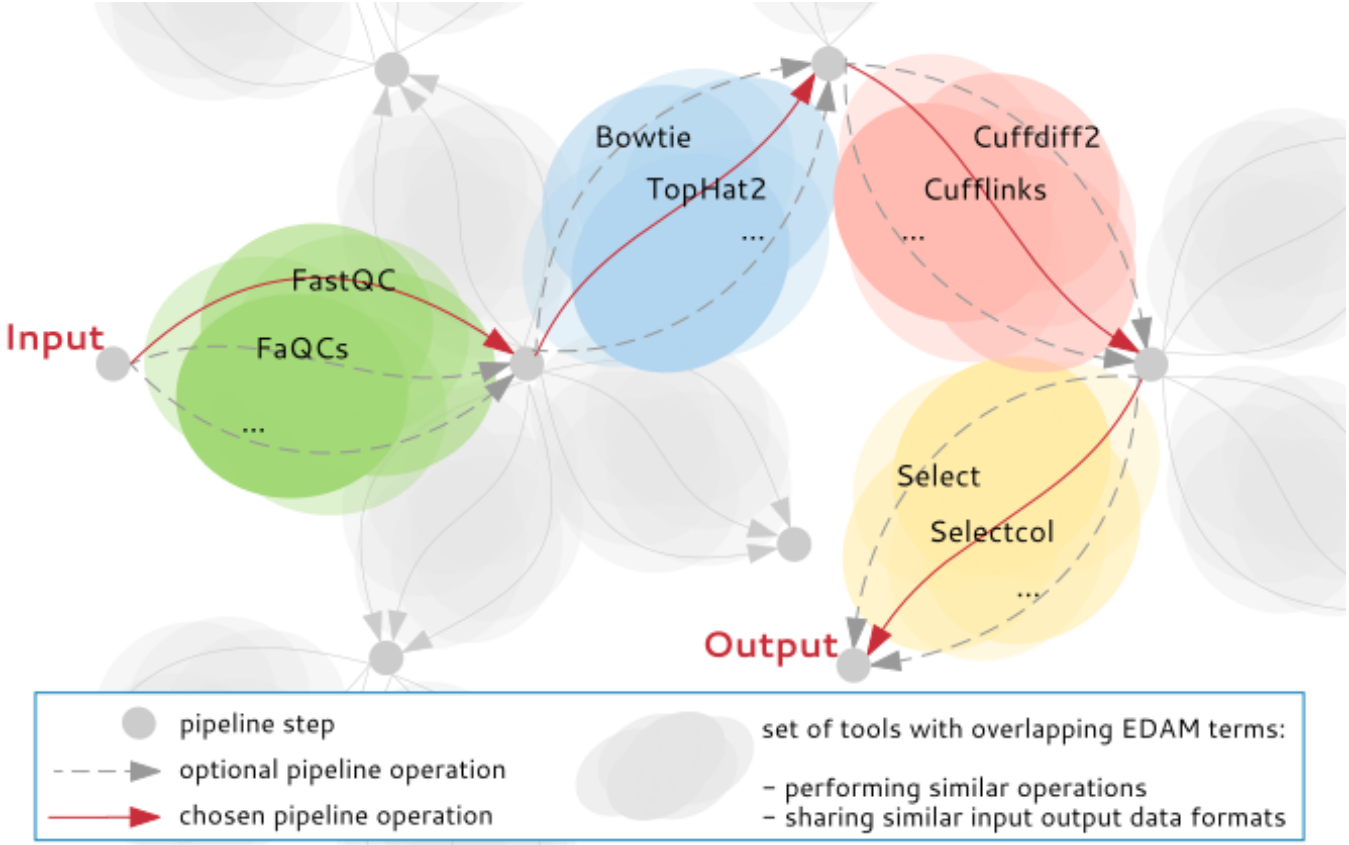
- Reusing the idea of the interactive tours, we can provide the user alternative tools to carry out their analyses.
- This approach allows for modular workflows, enabling the inclusion of alternative/experimental tools to carry out similar operations within new workflows.

To do so, Galaxy needs a *recommendation system* able to suggest tools to its users, and load them on-demand to build and allow flexible workflows.

de.STAIR

Recommendation system

Behind the interface, the idea is that of enabling users to decide which *path* to walk towards the completion of their own analysis.



Behind the interface, the idea is that of enabling users to decide which *path* to walk towards the completion of their own analysis.

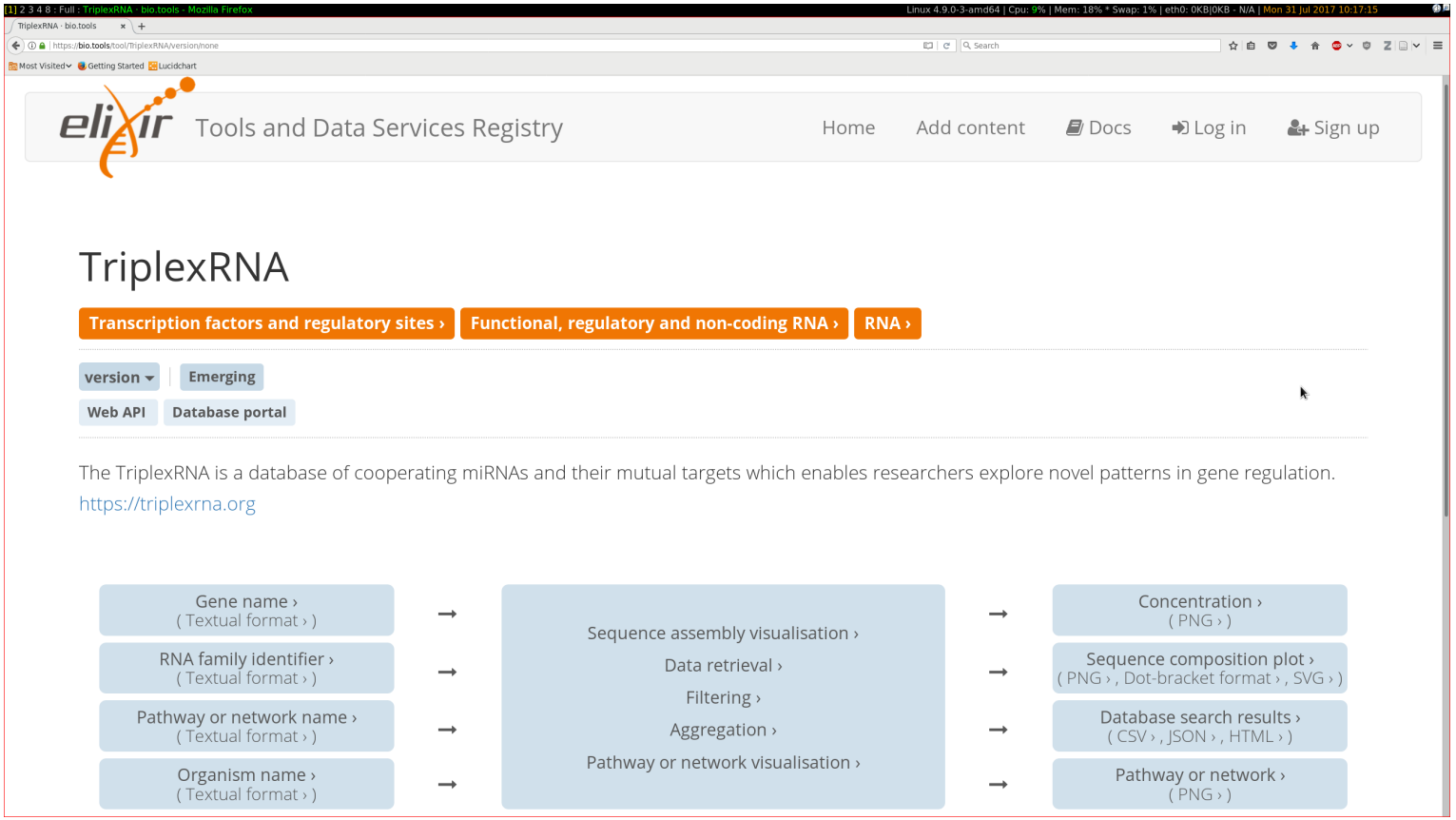
What is needed:

- group tools on the basis of their grade of overlap in carrying out a specific functionality
- define a set of input and output file formats at the immediate begin and end of each tool's operation
- define tags (trajectories throughout the possible steps) like RNA-Seq analysis, Epigenetic analysis, ...
- equip Galaxy with an interactive-tour-like dialog system that embeds the aforementionned recommendations to propose solutions within the available set of tools

de.STAIR

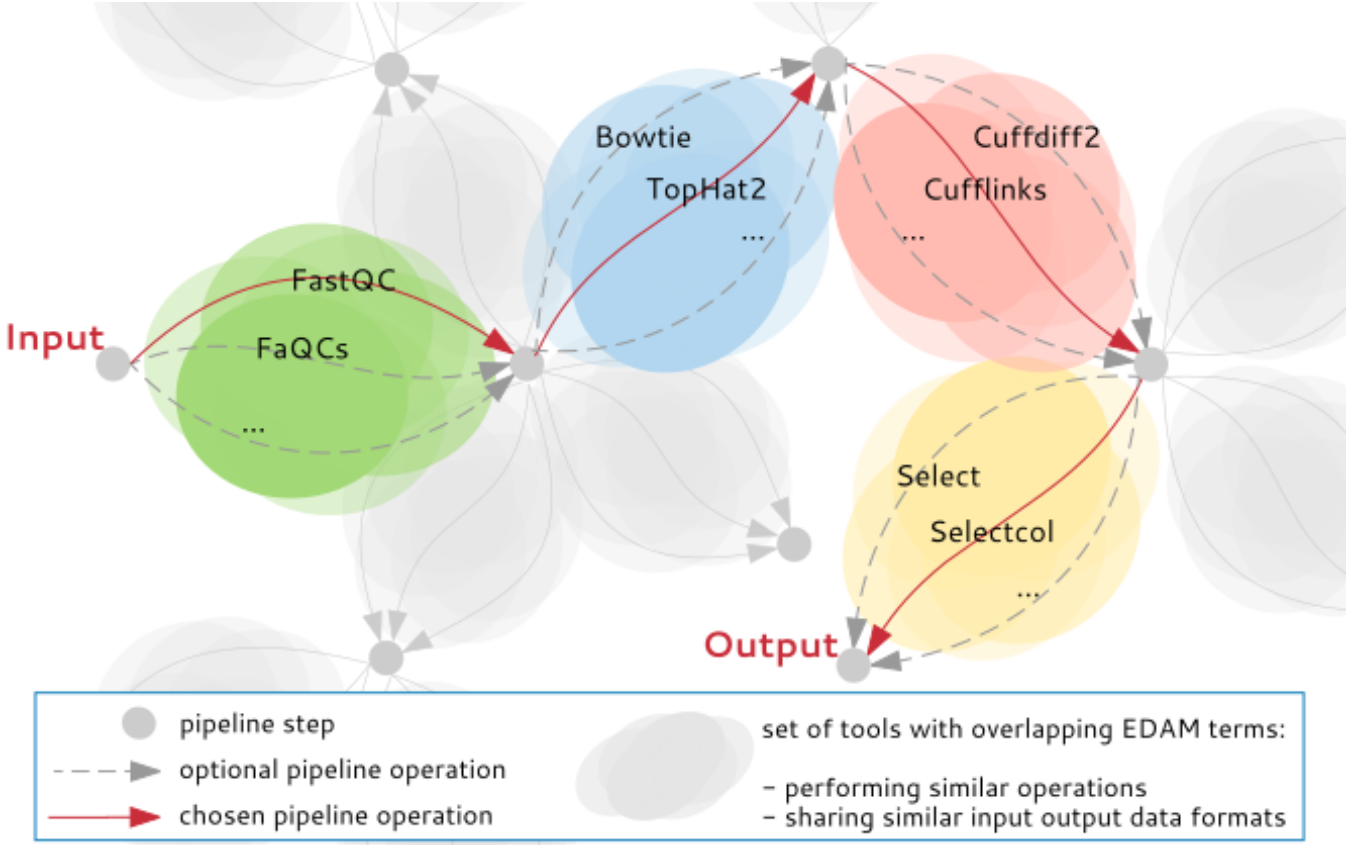
Recommendation system

To group tools in categories of operations and in/out file formats, we will rely on bio.tools, a registry of tools whose operation and formats are described by means of the EDAM Ontology.



Conclusions

These components will enable us to chain tools on the basis of their functionalities, in/out file formats, and pertinence within a specific kind of data-driven research topic, be it RNA-Seq, Epigenetic analysis, and so on



Acknowledgments

Olaf Wolkenhauer

Markus Wolfien

Bundesministerium
für Bildung
und Forschung

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE