

Towards automating and publishing workflow analyses in Galaxy

Andrea Bagnacani

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

Bielefeld, 24th October 2017

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

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.

Life Sciences have become more and more data-driven.

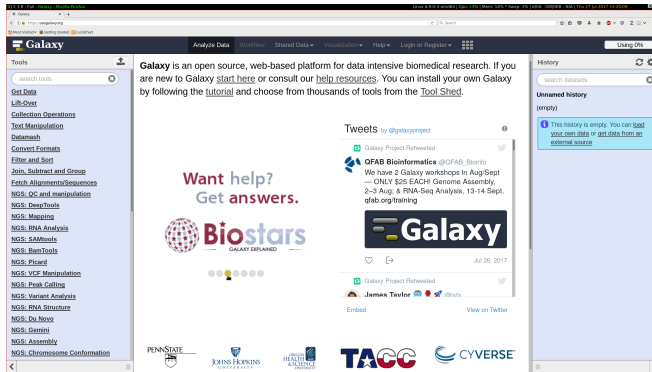
Insights on biological problems are gained leveraging on computational approaches:

- 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

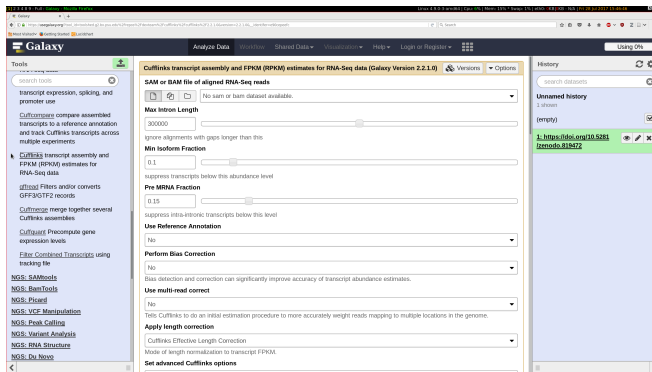
a web-based framework



The screenshot shows the Galaxy web interface. On the left is a 'Tools' sidebar with 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', and 'NGS: Chromosome Conformation'. The main content area features a welcome message, a 'Want help? Get answers.' section with a Biostars logo, and a tweet from @galaxyproject about Galaxy workshops. The bottom of the main area displays logos for Penn State, Johns Hopkins University, EMBL, TACCG, and CYVERSE. On the right is a 'History' sidebar showing a search bar and a message that the history is empty.

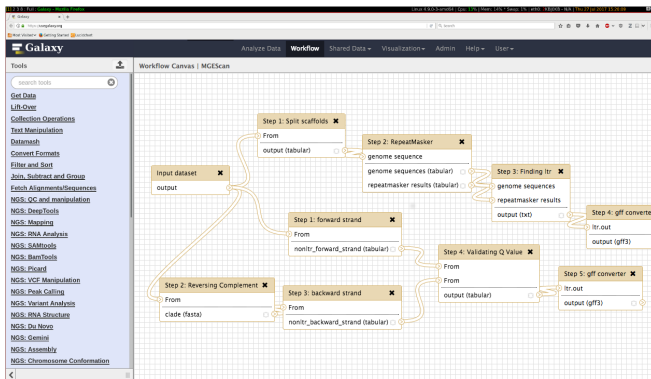
Galaxy is a web framework for computational bio/medical research

- public/private online/offline instances
- domain specific



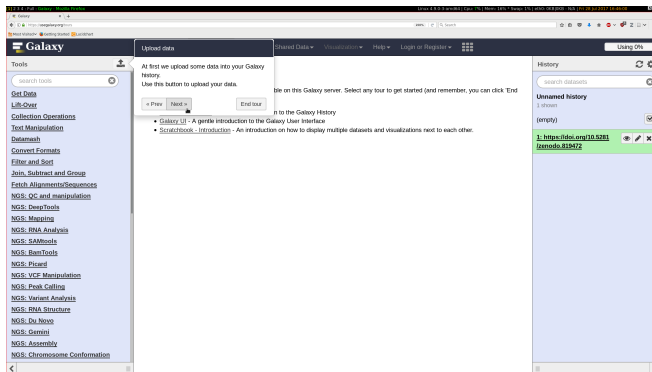
Galaxy is a web framework for computational bio/medical research

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



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



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

Interactive tours - behind the scenes

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

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

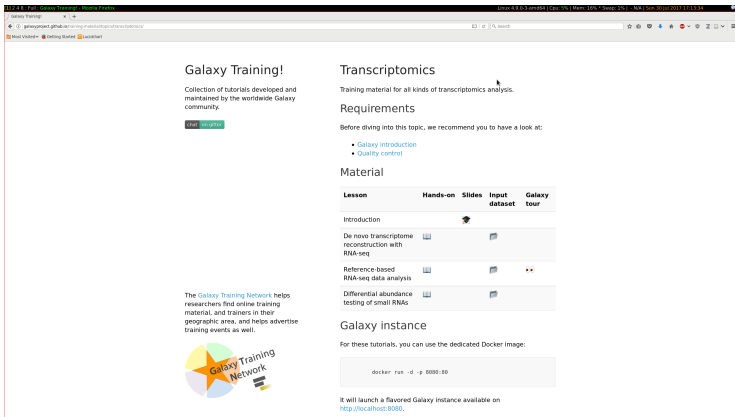
# A tour is made of several steps, each of them beginning with a dash '-'
steps:
  # 'title's will be displayed in the header of each step-container
  # If you don't specify any title, a default title is used, defined above.
  - title: Welcome to Galaxy
    # 'content' is the actual text that is shown 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
```



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




The screenshot shows the Galaxy Training! website. On the left, under 'Galaxy Training!', it says 'Collection of tutorials developed and maintained by the worldwide Galaxy community.' and has a 'Get started' button. Below this is a paragraph about the Galaxy Training Network and a logo. On the right, under 'Transcriptomics', it says 'Training material for all kinds of transcriptomics analysis.' and 'Requirements'. It then lists 'Galaxy introduction' and 'Quality control' as recommended topics. Below this is a 'Material' section with a table of lessons. The table has columns: Lesson, Hands-on, Slides, Input dataset, and Galaxy tour. The lessons listed are: Introduction, De novo transcriptome reconstruction with RNA-seq, Reference-based RNA-seq data analysis, and Differential abundance testing of small RNAs. Below the table is a 'Galaxy instance' section with a code block for a Docker command and a link to a Galaxy instance.

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

[Get started](#)

The Galaxy Training Network helps researchers find online training material, and trainers in their geographic area, and helps advertise training events as well.



Transcriptomics
Training material for all kinds of transcriptomics analysis.

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

- Galaxy introduction
- Quality control

Material

Lesson	Hands-on	Slides	Input dataset	Galaxy tour
Introduction				
De novo transcriptome reconstruction with RNA-seq				
Reference-based RNA-seq data analysis				
Differential abundance testing of small RNAs				

Galaxy instance
For these tutorials, you can use the dedicated Docker image:

```
docker run -d -p 8080:80
```

It will launch a flavored Galaxy instance available on <http://localhost:8080>.

Each training topic contains multiple examples, available as:

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

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 ✓

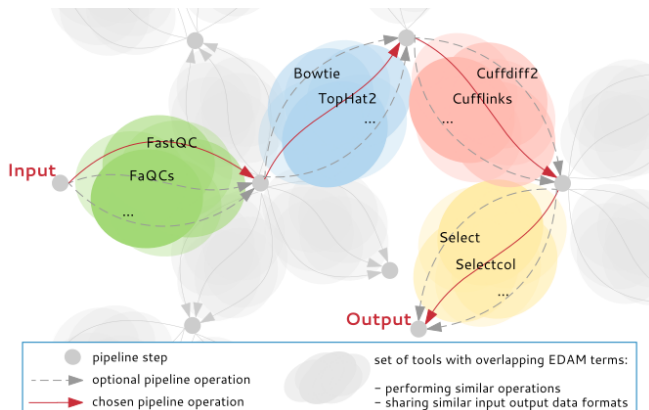
Galaxy is able to cover all these requirements for carrying out reproducible best-practices in data-driven research.

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.

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



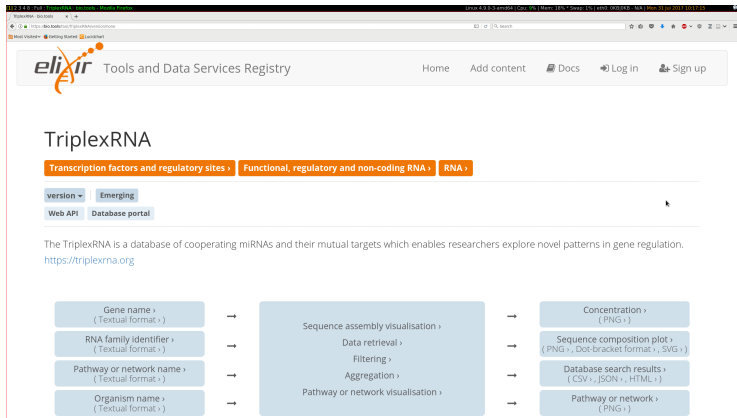
Behind the interface, the idea is that of enabling users to decide which *path* to walk towards the completion of their own analysis.

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 aforementioned recommendations to propose solutions within the available set of tools

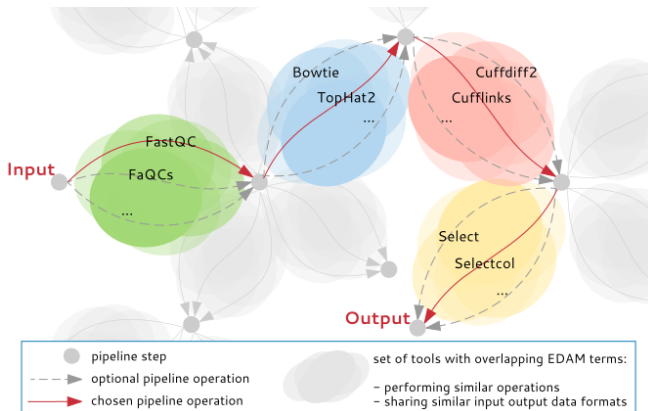
Recommendation system

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](#).



The screenshot shows the elixir Tools and Data Services Registry website. The main heading is "TriplexRNA". Below it, there are three orange buttons: "Transcription factors and regulatory sites", "Functional, regulatory and non-coding RNA", and "RNA". Underneath these buttons, there are two tabs: "version" (selected) and "Emerging". Below the tabs, there are two more tabs: "Web API" and "Database portal". A paragraph of text describes the TriplexRNA database: "The TriplexRNA is a database of cooperating miRNAs and their mutual targets which enables researchers explore novel patterns in gene regulation." followed by the URL "https://triplexrna.org". Below this text, there is a diagram showing a flow of data. On the left, there are four input boxes: "Gene name > (Textual format >)", "RNA family identifier > (Textual format >)", "Pathway or network name > (Textual format >)", and "Organism name > (Textual format >)". Arrows point from these boxes to a central box labeled "Sequence assembly visualisation >". This central box has three sub-items: "Data retrieval > Filtering > Aggregation >" and "Pathway or network visualisation >". Arrows point from the central box to a final box on the right labeled "Concentration > (PNG >)". This final box has three sub-items: "Sequence composition plot > (PNG >, Dot-bracket format >, SVG >)", "Database search results > (CSV >, JSON >, HTML >)", and "Pathway or network > (PNG >)".

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



Bundesministerium
für Bildung
und Forschung



Markus Wolfien

