Towards automating and publishing workflow analyses in Galaxy

Andrea Bagnacani, Markus Wolfien, Martin Scharm, Olaf Wolkenhauer

Research gaps and objectives for the present work

The Galaxy community is promoting RNA-Seq protocols and best practices through the reuse of existing tools, and the consolidation of a Training Network to provide guidance to researchers through example datasets, tutorials, and interactive tours. However, the more tools and techniques are showcased, the more complex the options for tool chaining and parametrization become. And as more material is gathered to illustrate approaches on the analysis of Life Science data, the higher the risk to overwhelm users. To assist researchers in carrying out their investigations, we propose to integrate the Galaxy framework with an interactive recommendation system that leverages on community consolidated best practices as well as EDAM-annotated tools. Such a system would promote the adoption of well established pipelines, while at the same time allow room for testing further experimental tools, easing reuse, and consolidating protocols and reproducibility.

Bridging the gap of tool chaining and parametrization

- Dockerized Galaxy flavor
- Integration within Galaxy RNA Workbench
- Provide interactive guidance on designing RNA-Seq experiments
- Test of novel and experimental tools
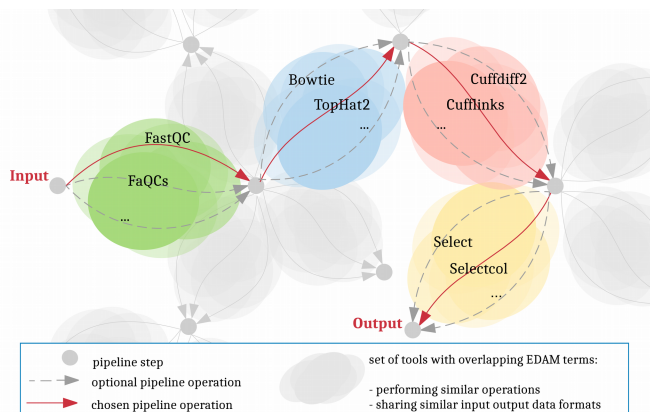
The de.STAIR Galaxy interactive recommendation system. The recommendation system traces a path from the beginning of a Life Science data analysis to its end. Here, researchers can decide which path to walk towards the completion of the desired analysis by providing an input dataset and an analysis goal.

Tools are recommended step by step, based on both tool's pertinence within the scope of the requested analysis goal and existing best practices within the current analysis step.

- Best practice approaches are shared among the scientific community through the Galaxy platform as reusable workflows.
- Tool pertinence is inferred from the provided EDAM ontology terms, associated to its operations and input/output data formats.

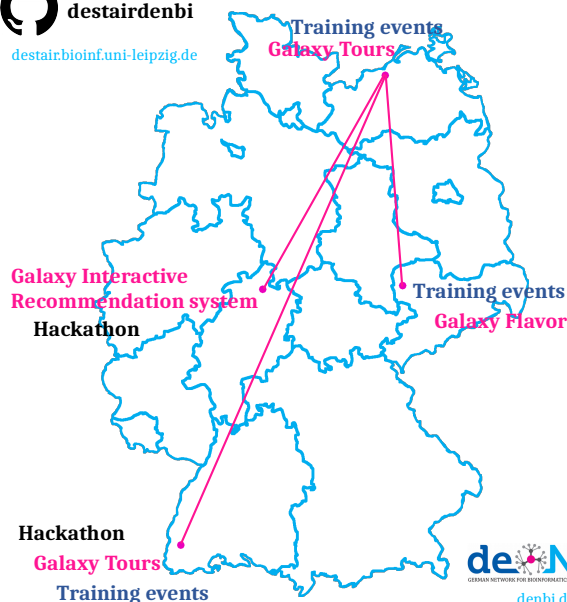## Towards the automation of workflow construction

Leverage on the Galaxy interactive tours to provide users interchangeable tools

Galaxy
galaxyproject.org

Gather EDAM ontology terms describing each tool in terms of operations, input, and output formats

EDAM
edamontology.org

Mine Galaxy instances for user-tracked data: tool chaining, parametrization, formats

RNA Bioinformatics Center R C B
denbi.de/rbc

destairdenbi
destair.bioinf.uni-leipzig.de

de.NBI
denbi.de

References

Wolfien M *et al*. TRAPLINE: A standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. BMC Bioinformatics, 2016. *10.1186/s12859-015-0873-9*

Lott SC *et al*. Customized workflow development and data modularization concepts for RNA-Sequencing and metatranscriptome experiments. Journal of Biotechnology, 2017. *10.1016/j.jbiotec.2017.06.1203*

Grüning BA *et al*. The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy. Nucleic Acids Research, 2017. *10.1093/nar/gkx409*

Andrea Bagnacani
andrea.bagnacani@uni-rostock.de
Systems Biology and Bioinformatics
University of Rostock
www.sbi.uni-rostock.de