EMBnet.journal
Bioinformatics in Action

Klingström et al. (2017) EMBnet.journal 23, e897
http://dx.doi.org/10.14806/ej.23.0.897

Technical Notes

Galaksio, a user friendly workflow-centric front end for Galaxy

Tomas Klingström[1*✉], **Rafael Hernández-de-Diego**[1*], **Théo Collard**[1], **Erik Bongcam-Rudloff**[1]

[1]SLU-Global Bioinformatics Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Sweden

[*]Both authors contributed equally (TK and RHD)

Competing interests: TK none; RHD none; TC none; EBR none

Abstract

There is a severe shortage of statisticians and bioinformaticians available in research. As universities fail to cover the increasing need of graduates with the necessary skills, ad hoc training and workshops have become commonplace but are insufficient to cover the needs. Technical solutions that distribute the workload more efficiently between researchers with a different education background (*e.g.*, computer scientists and biologists) are therefore necessary to cover some of this shortage.

Galaksio provides a workflow-centric graphical user interface for the Galaxy Workflow Management system that is easy to use for biologists and medical researchers who need to run routine tasks in bioinformatics. Combined with back end tools such as BioBlend, CloudMan and Pulsar, Galaksio provides a novel, layered approach to Galaxy making it easier to divide research tasks to researchers depending on their skills in interdisciplinary subjects such as bioinformatics and computational science.

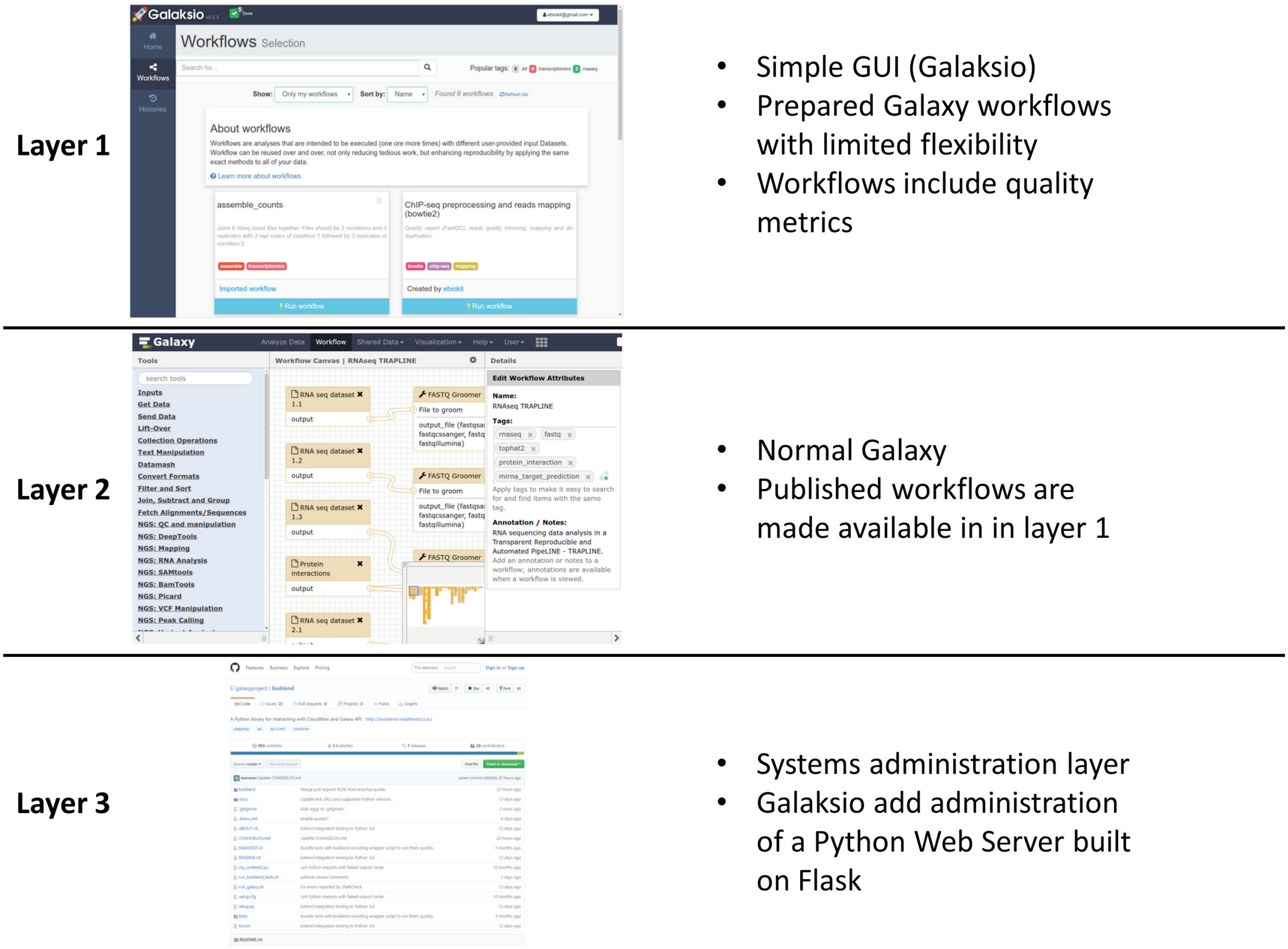
Galaksio is developed by the B3Africa project for the eB3Kit but can easily be installed independently using docker and configured to provide access to workflows on any Galaxy server using the Galaxy API. Galaksio can be downloaded at: **https://github.com/fikipollo/galaksio**.

Introduction

Galaxy is a widely supported workflow management system used in bioinformatics (Goecks *et al.*, 2010; Leipzig, 2016; Tastan Bishop *et al.*, 2015; Atwood *et al.*, 2015) to facilitate accessible and reproducible research. One of the main aims of Galaxy is to provide access to bioinformatic analysis tools for experimentalists with limited expertise in programming (Atwood *et al.*, 2015; Blankenberg *et al.*, 2010). Nevertheless, our experience with Galaxy, gained by implementing it in the eBiokit (Hernández-de-Diego *et al.*, 2017) and by using Galaxy in several training and capacity building projects (Fuxelius *et al.*, 2010; Atwood *et al.*, 2015; Mulder *et al.*, 2016) has shown us that many potential Galaxy users find themselves in a bit of a conundrum when trying to use Galaxy. Researchers skilled enough in bioinformatics to install and configure tools prefer command line tools, whereas less advanced users are left on their own struggling to find and combine tools using the user interface provided by Galaxy. Therefore, many research groups remain reliant on in-house scripts maintained by a small number of bioinformaticians spending significant time on providing ad hoc support to other researchers in the group. To provide an attractive technology platform for researchers it was therefore deemed necessary to provide a more simplified, workflow-centric model of operations. In the workflow-centric model researchers with limited bioinformatics training are provided with prepared workflows and default input parameters, while more advanced users can create and modify workflows using the normal Galaxy GUI. This allows research teams to work in a more efficient way. Trained bioinformaticians can adapt and develop tools and then provide the finished workflows for routine analysis to lab researchers.

In standard Galaxy all users rely on the same GUI, despite significantly different education background and expertise. Trained bioinformaticians often rely on a set of skills dependent on education decisions taken by students several years ahead of enrolling at a university (Wightman and Hark, 2012) while other researchers may have little or no formal training. Given the complexities of training needs, influential stakeholders such as the US National Research Council has therefore concluded that bioinformatics research is likely to be carried out by two disparate groups of researchers: quantitative biologists, who work at the interface of mathematical/computer science and biology, and research biologists, who need familiarity with a range of mathematical and computational concepts without necessarily being an

Article history
Received: 18 May 2017
Accepted: 3 October 2017
Published: 9 November 2017



Layer 1

- Simple GUI (Galaksio)
- Prepared Galaxy workflows with limited flexibility
- Workflows include quality metrics

Layer 2

- Normal Galaxy
- Published workflows are made available in in layer 1

Layer 3

- Systems administration layer
- Galaksio add administration of a Python Web Server built on Flask

Figure 1. This figure shows the layered approach used by Galaksio and implemented in the eB3Kit to divide labour more efficiently between researchers with different background.

expert (National Research Council (US) Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century, 2003).

We therefore present Galaksio, a solution based on the Galaxy API and a Python web server, that we have developed to provide a layered access to Galaxy functions that facilitate the work of research biologists through an easy-to-use web interface, while the default Galaxy interface is used by bioinformaticians to create new workflows and systems administration tasks that are facilitated by packages created by other researchers such as BioBlend (Sloggett *et al.*, 2013), CloudMan (Afgan *et al.*, 2010) and Pulsar (Afgan *et al.*, 2015). With Galaksio, all data is managed within the normal Galaxy workflow management system and user credentials are passed on to the Galaxy server to manage user privileges, meaning that Galaksio can be used to access all workflows created on a normal Galaxy server using the command line tools implemented on the server.

Thanks to Galaksio, the Galaxy user's experience can be managed at three different levels: 1) a layer suited to research biologists (*i.e.*, users using tools); 2) a layer suited to bioinformaticians (*i.e.*, users developing tools); 3) a layer suited to computer scientists (*i.e.*, users developing the environment tools work in) (Figure 1).

This approach is currently being implemented in the B3Africa project using the eB3Kit which includes Galaksio and relies on these resources to connect the relatively light weight Mac Pro Server, commonly hosting the eB3Kit, to external computing resources (Klingstrom *et al.*, 2016).

Materials, Methodologies and Techniques

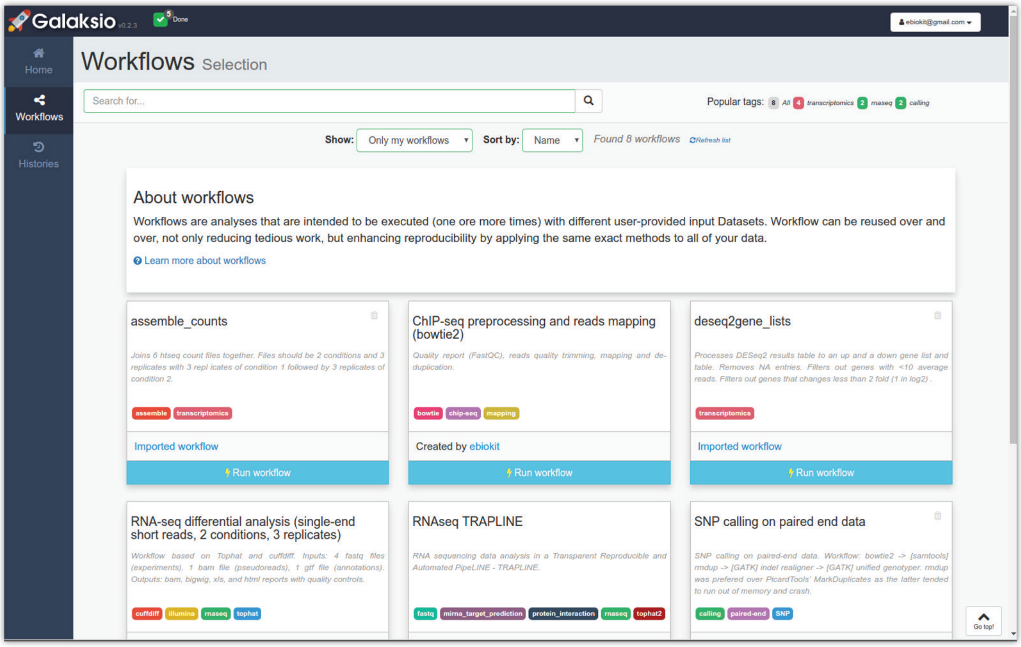
Galaksio has been designed as a multiuser web application and is divided in two components: the server side application and the web interface for users.

The server side, which is built on Python Flask server[1], is responsible for accessing the Galaxy data using the tools provided by the Galaxy application programming interface (API) (Blankenberg *et al.*, 2010; Goecks *et al.*, 2010). The Galaksio web interface has been developed using AngularJS[2] and Bootstrap[3], both popular HTML, CSS, and JavaScript cross-browser frameworks for developing responsive and user-friendly

[1] http://flask.pocoo.org/

[2] https://angular.io

[3] http://getbootstrap.com



Workflows Selection

About workflows

Workflows are analyses that are intended to be executed (one ore more times) with different user-provided input Datasets. Workflow can be reused over and over, not only reducing tedious work, but enhancing reproducibility by applying the same exact methods to all of your data.

assemble_counts

Joins 6 htseq count files together. Files should be 2 conditions and 3 replicates with 3 repl icates of condition 1 followed by 3 replicates of condition 2.

assemble transcriptomics

Imported workflow

ChIP-seq preprocessing and reads mapping (bowtie2)

Quality report (FastQC), reads quality trimming, mapping and de-duplication.

bowtie chip-seq mapping

Created by ebiokit

deseq2gene_lists

Processes DESeq2 results table to an up and a down gene list and table. Removes NA entries. Filters out genes with <10 average reads. Filters out genes that changes less than 2 fold (1 in log2) .

transcriptomics

Imported workflow

RNA-seq differential analysis (single-end short reads, 2 conditions, 3 replicates)

Workflow based on Tophat and cuffdiff. Inputs: 4 fastq files (experiments), 1 bam file (pseudoreads), 1 gtf file (annotations). Outputs: bam, bigwig, xls, and html reports with quality controls.

cuffdiff illumina rnaseq tophat

RNAseq TRAPLINE

RNA sequencing data analysis in a Transparent Reproducible and Automated PipeLINE - TRAPLINE.

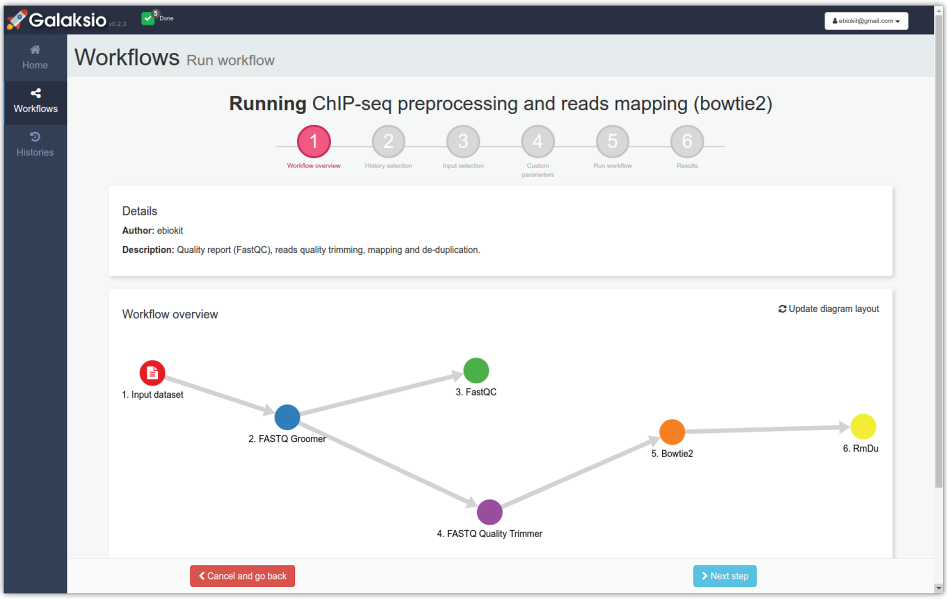
fastq mirna_target_prediction protein_interaction rnaseq tophat2

SNP calling on paired end data

SNP calling on paired-end data. Workflow: bowtie2 -> [samtools] rmdup -> [GATK] indel realigner -> [GATK] unified genotyper. rmdup was prefered over PicardTools' MarkDuplicates as the latter tended to run out of memory and crash.

calling paired-end SNP

Figure 2. The figure shows the graphical interface for the workflow selection in Galaksio.



Workflows Run workflow

Running ChIP-seq preprocessing and reads mapping (bowtie2)

Details

Author: ebiokit

Description: Quality report (FastQC), reads quality trimming, mapping and de-duplication.

Workflow overview

Figure 3. The figure shows the Galaksio web interface that is presented to the user after the selection of a workflow.

web applications. The exchange of the data between clients and the server is handled using asynchronous JavaScript and XML (AJAX) communication.

Results

Galaksio is free to use and is distributed under the GNU General Public License, Version 3. A public copy of the application is hosted at the SLU facilities as part of the eBioKit platform[4] and source code is available at GitHub[5], allowing other laboratories to browse, propose code reviews, and download the code in order to set up their own instance of the application. Additionally, Galaksio can easily be installed using Docker[6], an open-source virtualisation software that provides a lightweight, stand-alone, portable, and ready-to-execute package that includes the software and all the dependencies necessary to run the application independently of the operating system installed on the server. Documentation for the project can be found at the ReadTheDocs platform[7].

Figure 2 shows the Galaksio's GUI for biologists. Using this interface users can run any workflow implemented in the associated Galaxy instance in just a few clicks and get a clear image of the analysis steps

[4] http://ebiokit.eu/

[5] https://github.com/fikipollo/galaksio

[6] https://www.docker.com

[7] https://galaksio.readthedocs.io

included in the selected workflow (Figure 3). The user interface allows the user to customise the execution of pre-selected tools, the uploading of the necessary files, the downloading of the results, and the execution of several workflows simultaneously in the background.

Table 1 provides an overview of all the developed features in the current Galaksio version. As all interactions with Galaxy are managed through the Galaxy API, the Galaksio implementation can be hosted independently as a separate server sending commands to any available Galaxy server. This includes public servers such as the popular usegalaxy.org website. Information on the connected server is provided when logging in via the Galaksio interface. It should however be noted that Galaksio, while light-weight in itself, is completely dependent on the speed of the Galaxy server when returning workflows and any user restrictions defined by the Galaxy server such as the amount of storage available.

Use case

Due to delays in achieving approval for tool wrappers created by the Galaksio team, an alternative use case has been created with much appreciated support from Marius van den Beek at the Institut Curie, Paris, France. The test dataset is available from the Zenodo data repository (Freeberg and Heydarian, 2016) but all data can also be imported from usegalaxy.org.

History containing dataset collections: https://usegalaxy.org/u/tomkl/h/galaksio-use-case-mouse-chip-seq-data.

Main workflow: https://usegalaxy.org/u/tomkl/w/copy-of-imported-parent-workflow-chipseq

Subworkflow: https://usegalaxy.org/u/tomkl/w/copy-of-imported-chipseqtutorialchild1

The workflows can be imported inside Galaksio by any users logged into a Galaksio server connected to usegalaxy.org. Other use cases will be added with the addition of "Galaksio use case" in the name of the workflow to make them easy to be identified in the Galaksio's repository. Issues are tracked using the Galaksio repository on GitHub[8] and external contributions are welcome.

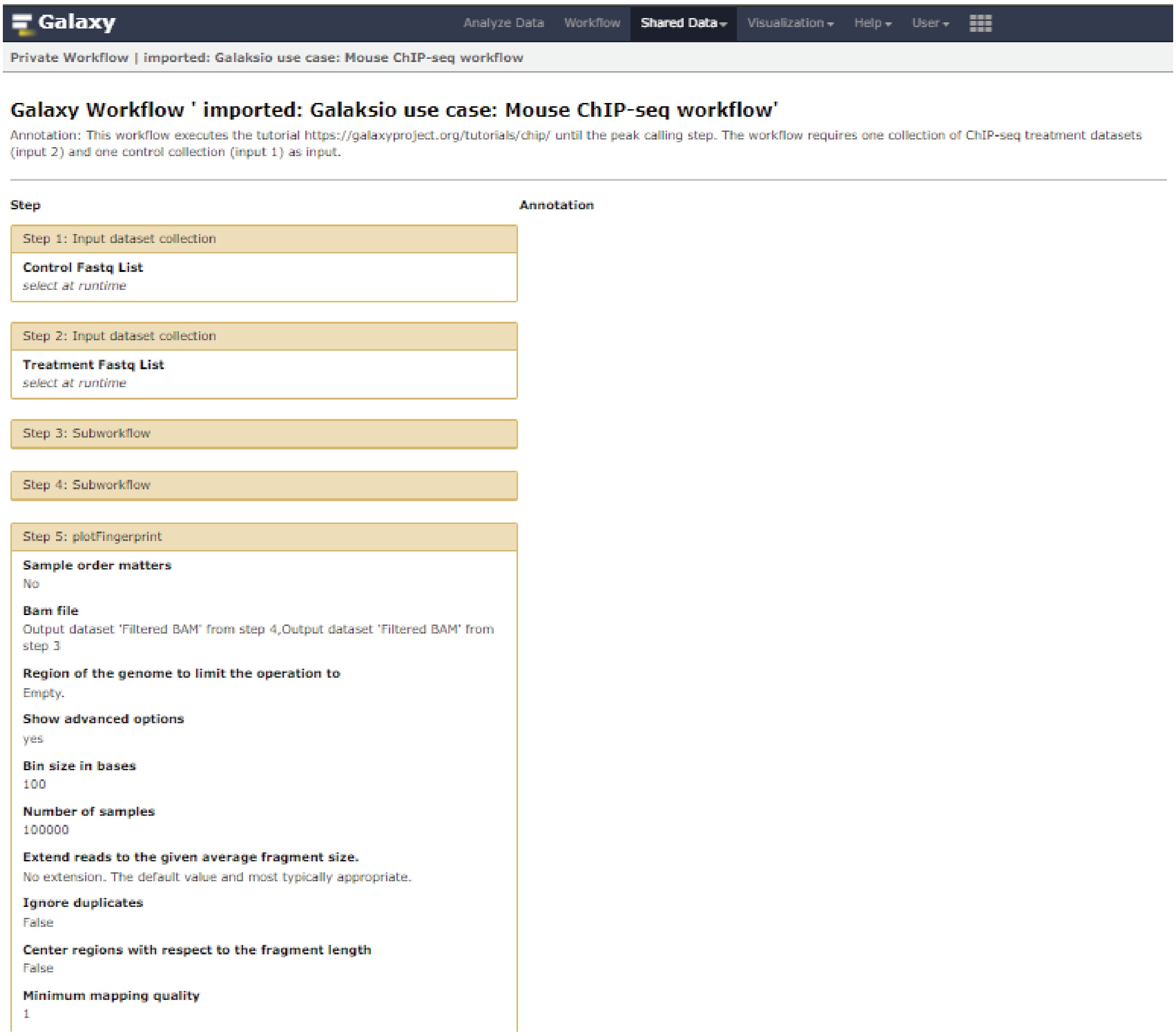
Discussion

Compared to the clearly defined classes of "research biologist" and "quantitative biologist", proposed by the US National Research Council, bioinformatics has developed into a field where its practitioners share a number of characteristics, but none of which are essential enough to characterise what a bioinformatician truly is (Vincent and Charette, 2015). Many people may therefore be highly skilled and productive researchers in bioinformatics, despite very limited skills in one or more of the core competencies associated with being a bioinformatician (Smith, 2015). Due to the shortage

Table 1. Implemented and planned features for Galaksio.

Feature	Category	Implemented	Planned
User sign-in/out	Users	X	
User sign-up	Users	X	
Workflow listing	Workflows	X	
Workflow importing	Workflows	X	
Workflow execution	Workflows	X	
Workflow creation	Workflows		X
Simultaneous execution of workflows	Workflows	X	
Recovering previous executions	Workflows	X	
Help and description for tools in workflow	Workflows	X	
Input selection and parameter configuration	Workflows	X	
History selection	History	X	
History creation	History		X
History deletion	History		X
Dataset uploading	Dataset manipulation	X	
Dataset downloading	Dataset manipulation	X	
Dataset deletion	Dataset manipulation	X	
Dataset collection creation	Dataset manipulation	X	
Dataset collection deletion	Dataset manipulation		X
Tool execution	Tools		X

[8] https://github.com/fikipollo/galaksio/issues



Private Workflow | imported: Galaksio use case: Mouse ChIP-seq workflow

Galaxy Workflow ' imported: Galaksio use case: Mouse ChIP-seq workflow'

Annotation: This workflow executes the tutorial https://galaxyproject.org/tutorials/chip/ until the peak calling step. The workflow requires one collection of ChIP-seq treatment datasets (input 2) and one control collection (input 1) as input.

Step | Annotation

Step 1: Input dataset collection

Control Fastq List
select at runtime

Step 2: Input dataset collection

Treatment Fastq List
select at runtime

Step 3: Subworkflow

Step 4: Subworkflow

Step 5: plotFingerprint

Sample order matters
No

Bam file
Output dataset 'Filtered BAM' from step 4,Output dataset 'Filtered BAM' from step 3

Region of the genome to limit the operation to
Empty.

Show advanced options
yes

Bin size in bases
100

Number of samples
100000

Extend reads to the given average fragment size.
No extension. The default value and most typically appropriate.

Ignore duplicates
False

Center regions with respect to the fragment length
False

Minimum mapping quality
1

Figure 4. The figure displays a report generated by Galaxy by exporting a workflow after running a ChIP-seq use case.

of comprehensive university programmes in the field (Williams and Teal, 2017; Atwood *et al.*, 2015), most researchers currently active in bioinformatics have participated in a number of courses, workshops and self-learning sessions that, step by step, has taken them to a skill level where they may be considered qualified bioinformaticians or quantitative biologists. Such a self-organised curriculum encourages bioinformaticians to obtain exactly the skills necessary to complete their own projects but with limited consideration for auxiliary skills such as code documentation and a deeper understanding of computer science.

As a result of this self-motivated style of learning, significant delays occur when new technologies emerge if they require significant retraining of practitioners before becoming fully competitive with the new solution. This is perhaps most evident in the slow adoption of distributed computing systems such as Hadoop[9]. While significant investments in large Hadoop infrastructures has been made, the production of bioinformatics tools to use them has been delayed as bioinformatics tools are developed by bioinformaticians focused on high-level languages which, until recently, had limited support for Hadoop. Thereby delaying the adoption of distributed computing in bioinformatics (Oliphant, 2016).

The Galaksio interface itself is tailored towards enhancing user friendliness for biologists and medical researchers with limited IT-skills. The implementation of such a tool is a necessary step towards a multi-layered approach to Galaxy which allows distribution of labour not only between biologists and bioinformaticians, but also between "scripting" bioinformaticians and bioinformaticians with a strong background in computer science. Enabling researchers with the latter form of education background to provide access to more advanced computation tools by creating tools such as BioBlend (Sloggett *et al.*, 2013), CloudMan (Afgan *et al.*, 2010) and Pulsar (Afgan *et al.*, 2015) connect the Galaxy workflow management system to more powerful computation resources.

A common objection to user-friendly and automated systems such as Galaksio is the fear that automation can increase the error rate or can reduce the willingness of

[9]http://hadoop.apache.org/

researchers to learn bioinformatics properly. Automation is however one of the core concepts of advanced research ever since the introduction of the automated sequencing (Smith *et al.*, 1986). Indeed, without the automation of routine tasks even the sequencing and analysis of a single genome would be an impossible task (Ewing *et al.*, 1998). The relevance of automation within specific research tasks is perhaps best demonstrated by the common reliance on FASTQ files, with automatically assigned phred-quality scores, rather than the more expansive sequence read format (SRF) when working with large volumes of data (Clarke *et al.*, 2012; Van der Auwera *et al.*, 2013). With Galaksio automation is moved from a per-tool basis to a per-workflow basis and it is therefore appropriate to not only look at the risks that a further automation of tasks can bring, but also to evaluate how the current state of automation is facilitated in bioinformatics and other IT heavy fields. As an example, in healthcare the data management is seen as a way to reduce error rates and three key factors to success have been proposed for automation to be beneficial (Nolan, 2000):

- the system should prevent errors;
- procedures must be transparent so that they may be intercepted;
- procedures should be designed to mitigate the adverse effects of errors when they are not detected and intercepted.

Current practices in research are far from optimal when considering these three criteria for automation of bioinformatics. When dealing with bioinformatics tasks beyond their expertise, biologists may prefer commercial software that provides a more comprehensive, but also expensive platform with a dependency on proprietary software (Pabinger *et al.*, 2014; Smith, 2015b). As an alternative they may rely on outsourcing computing tasks to collaborators. Other biologists take the course of establishing their own curriculum of training as previously discussed. Some of these researchers may, over time, become proficient bioinformaticians but even in the best case scenario researchers are likely to produce a number of papers based on ad-hoc scripting with low transparency and potentially serious errors, unlikely to be caught by reviewers. In comparison, prepared workflows accessed in Galaxy or Galaksio limits the time spent on ad-hoc scripting and provide a comprehensive file history with source data and the individual steps used to generate the final results that greatly improve the reproducibility of the results (see Figure 4).

The downside of Galaksio is that it does not provide a natural exposure to the command line environment. However, Galaksio provides a comprehensive overview of any workflow available in the Galaxy system. If used properly Galaksio can therefore also serve as a training tool to explain theoretical concepts prior to coding exercises and function as a road map for researchers aiming to improve their skills in bioinformatics and build their own workflows step-by-step using the command line.

Conclusions

Galaksio does not replace the role of trained bioinformaticians in a research environment. It does however allow bioinformaticians to automate routine tasks and promote transparency in research as researchers with limited, or no, bioinformatics training can run best practice procedures and automatically generate the data necessary for others to evaluate their work. Such automation of routine tasks have contributed positively to the productivity and to the reduction of error rates in other information heavy fields (Horsfall, 1992; Leek and Peng, 2015; Nolan, 2000). Automation can thereby reduce the work load of expert bioinformaticians and provide them with the freedom to target more challenging tasks as well as to develop a curriculum for the evaluation and training of colleagues with basic or intermediate training (Peng, 2015).

Key Points

- Galaksio is built to provide a more layered approach to Galaxy, providing a simplified user interface based on workflows.
- Galaksio reduces the workload of bioinformaticians as routine tasks can be performed with minimal training. The presentation of workflows also provides a comprehensive overview of necessary input data as well as methodological changes to the end user.
- Galaksio can be used to rapidly deploy new services. Public Galaxy servers are a powerful tool to support collaborative research and Galaksio provides a more lightweight user interface for researchers who wish to make a specific project or workflow available.

Acknowledements

This work was financed by the B3Africa project. B3Africa is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 654404.

We would also like to acknowledge the much appreciated help provided by Marius van den Beek at the Institut Curie, Paris, France and other active users in the Galaxy chat.

References

1. Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, *et al.* (2010) Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics **11** (Suppl 12), S4. http://dx.doi.org/10.1186/1471-2105-11-S12-S4
2. Afgan E, Coraor N, Chilton J, Baker D, Taylor J, *et al.* (2015) Enabling cloud bursting for life sciences within Galaxy: Enabling Cloud Bursting for Life Sciences within Galaxy. Concurr. Comput. Pract. Exp. **27** (16), 4330–4343. http://dx.doi.org/10.1002/cpe.3536

3. Atwood TK, Bongcam-Rudloff E, Brazas ME, Corpas M, Gaudet P, *et al.* (2015) GOBLET: The Global Organisation for Bioinformatics Learning, Education and Training. PLOS Comput. Biol. **11** (4), e1004143. http://dx.doi.org/10.1371/journal.pcbi.1004143
4. Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, *et al.* (2010) Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. Current Protocols in Molecular Biology. John Wiley & Sons, Inc., Hoboken, NJ, USA, Hoboken, NJ, USA,
5. Freeberg M and Heydarian M (2016) Training Material For Chip-Seq Analysis. http://dx.doi.org/10.5281/zenodo.197100
6. Fuxelius H, Bongcam E, and Jaufeerally Y (2010) The contribution of the eBioKit to Bioinformatics Education in Southern Africa. EMBnet.journal **16** (1), 29. http://dx.doi.org/10.14806/ej.16.1.173
7. Goecks J, Nekrutenko A, Taylor J, and Galaxy Team T (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. **11** (8), R86. http://dx.doi.org/10.1186/gb-2010-11-8-r86
8. Hernández-de-Diego R, de Villiers EP, Klingström T, Gourlé H, Conesa A, *et al.* (2017) The eBioKit, a stand-alone educational platform for bioinformatics. PLOS Comput. Biol. **13** (9), e1005616. http://dx.doi.org/10.1371/journal.pcbi.1005616
9. Horsfall K (1992) The human impact of library automation University of South Australia Library,.
10. Klingstrom T, Mendy M, Meunier D, Berger A, Reichel J, *et al.* (2016) Supporting the development of biobanks in low and medium income countries. IEEE, pp. 1–10
11. Leek JT and Peng RD (2015) Opinion: Reproducible research can still be wrong: Adopting a prevention approach: Fig. 1. Proc. Natl. Acad. Sci. **112** (6), 1645–1646. http://dx.doi.org/10.1073/pnas.1421412111
12. Leipzig J (2016) A review of bioinformatic pipeline frameworks. Brief. Bioinform. **18** (3), 530–536. http://dx.doi.org/10.1093/bib/bbw020
13. Mulder NJ, Adebiyi E, Alami R, Benkahla A, Brandful J, *et al.* (2016) H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. Genome Res. **26** (2), 271–277. http://dx.doi.org/10.1101/gr.196295.115
14. National Research Council (US) Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century (2003) Bio2010: Transforming Undergraduate Education for Future Research Biologists National Academies Press (US), Washington (DC),.
15. Nolan TW (2000) System changes to improve patient safety. BMJ **320** (7237), 771–773.
16. Oliphant T (2016) Anaconda and Hadoop --- a story of the journey and where we are now. http://technicaldiscovery.blogspot.se/2016/03/anaconda-and-hadoop-story-of-journey.html (accessed 7 April 2017).
17. Peng R (2015) The reproducibility crisis in science: A statistical counterattack. Significance **12** (3), 30–32. http://dx.doi.org/10.1111/j.1740-9713.2015.00827.x
18. Sloggett C, Goonasekera N, and Afgan E (2013) BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics **29** (13), 1685–1686. http://dx.doi.org/10.1093/bioinformatics/btt199
19. Smith DR (2015) Broadening the definition of a bioinformatician. Front. Genet. **6**, 258. http://dx.doi.org/10.3389/fgene.2015.00258
20. Tastan Bishop O, Adebiyi EF, Alzohairy AM, Everett D, Ghedira K, *et al.* (2015) Bioinformatics Education--Perspectives and Challenges out of Africa. Brief. Bioinform. **16** (2), 355–364. http://dx.doi.org/10.1093/bib/bbu022
21. Vincent AT and Charette SJ (2015) Who qualifies to be a bioinformatician? Front. Genet. **6**, 164. http://dx.doi.org/10.3389/fgene.2015.00164
22. Wightman B and Hark AT (2012) Integration of bioinformatics into an undergraduate biology curriculum and the impact on development of mathematical skills. Biochem. Mol. Biol. Educ. **40** (5), 310–319. http://dx.doi.org/10.1002/bmb.20637
23. Williams JJ and Teal TK (2017) A vision for collaborative training infrastructure for bioinformatics: Training infrastructure for bioinformatics. Ann. N. Y. Acad. Sci. **1387** (1), 54–60. http://dx.doi.org/10.1111/nyas.13207