BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 9 2013, pages 1233–1234
doi:10.1093/bioinformatics/btt115

Data and text mining Advance Access publication March 11, 2013

Galaxy LIMS for next-generation sequencing

Jelle Scholtalbers*, Jasmin Rößler, Patrick Sorn, Jos de Graaf, Valesca Boisguérin, John Castle† and Ugur Sahin†

Computational Medicine, TRON - Translational Oncology at the Johannes Gutenberg University of Mainz Medicine, 55131 Mainz, Germany

Associate Editor: Inanc Birol

ABSTRACT

Summary: We have developed a laboratory information management system (LIMS) for a next-generation sequencing (NGS) laboratory within the existing Galaxy platform. The system provides lab technicians standard and customizable sample information forms, barcoded submission forms, tracking of input sample quality, multiplex-capable automatic flow cell design and automatically generated sample sheets to aid physical flow cell preparation. In addition, the platform provides the researcher with a user-friendly interface to create a request, submit accompanying samples, upload sample quality measurements and access to the sequencing results. As the LIMS is within the Galaxy platform, the researcher has access to all Galaxy analysis tools and workflows. The system reports requests and associated information to a message queuing system, such that information can be posted and stored in external systems, such as a wiki. Through an API, raw sequencing results can be automatically pre-processed and uploaded to the appropriate request folder. Developed for the Illumina HiSeq 2000 instrument, many features are directly applicable to other instruments.

Availability and implementation: The code and documentation are available at http://tron-mainz.de/tron-facilities/computational-medicine/galaxy-lims/

Contact: jelle.scholtalbers@tron-mainz.de

Received on October 2, 2012; revised on February 28, 2013; accepted on March 1, 2013

1 INTRODUCTION

Next-generation sequencing (NGS) has enabled researchers to sequence large numbers of samples. For example, one flow cell on the Illumina HiSeq 2000 sequencer can sequence 192 samples using the 24 standard Illumina multiplexing indexes or more with alternative barcoding methods. Challenges for sequencing facilities and small labs include tracking sequencing requests, individual samples, sample status throughout the sequencing process and the resultant data, including the interpretation of the results.

Public and commercial laboratory information management systems (LIMSs) exist to track NGS samples. GnomEx (Nix *et al*., 2010), for example, is an excellent open-source resource for NGS and microarray LIMS. Both GnomEx and openBIS (Bauch *et al*., 2011) provide complete solutions for more than NGS data organization, such as imaging, microarray and proteomics data. GnomEx includes its own data processing and analysis tools for use within the platform. In contrast to these platforms, our aim was to build a light-weight yet effective NGS LIMS within an established data processing and analysis platform.

The Galaxy platform (Blankenberg *et al*., 2010; Giardine *et al*., 2005; Goecks *et al*., 2010) is a widely used easy-to-setup platform for the analysis of NGS and genomic data. Here, we extended Galaxy's basic sample-tracking capability into a user-friendly and flexible LIMS, able to handle multiple samples per request. Furthermore, we enabled more reporting capabilities by extending the existing Galaxy notification system to update external systems (e.g. wikis) with sequencing request and status information. Our approach is in part similar to the extension by B. Chapman (http://bcbio.wordpress.com/2011/01/11/next-generation-sequencing-information-management-and-analysis-system-for-galaxy/), but greatly differs in its implementation for creating requests, samples and data processing.

2 USER WEB INTERFACE

2.1 Request submission

The researcher is presented with a customized request creation form, which is pre-programmed to support common sequencing protocols, including genome and exome sequencing, ChIP-Seq and transcriptome sequencing. Added fields allow input at predefined sample information points and the user is assisted in the input of appropriate parameters. In all cases, user input is validated. For example, submissions including total RNA samples are checked for bad RNA Integrity Numbers and will be marked accordingly. After completing the required information boxes, the user is able to upload quality control files, such as Agilent Bioanalyzer trace images. After upload, the user and lab administrator can view and analyze the information like any other Galaxy dataset. To complete the new request, the user clicks 'submit' and is then able to print a new request submission form, which is barcoded with the request id, lists all user-provided information and has a predefined boilerplate, which can be signed and sent to the lab with the samples.

2.2 Request processing

The administration interface allows the lab administrator to modify requests and sample forms, control user requests, add and manage datasets and design and submit flow cells for sequencing. An administrator can define new request types and

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

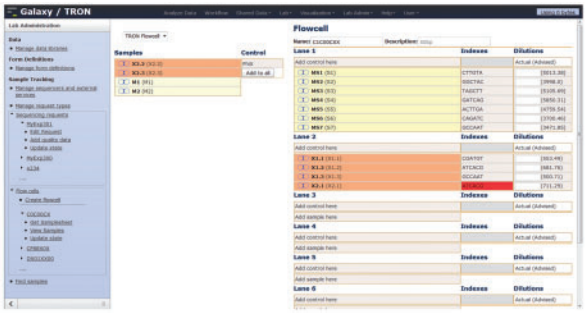

Fig. 1. Screenshot of Galaxy during flow cell design. The menu on the left is only accessible to a lab administrator to manage requests and flow cells. In the main window samples can be dragged-and-dropped into the flow cell lanes; duplicate indexes are marked red

can define the associated new sample information forms to capture the information necessary to process the new request. Sample quality and library preparation information can be added to the system and can be made visible to the customer. Furthermore, an administrator can change, correct and reject all information.

The system supports lab administrators with an interactive flow cell layout (see Fig. 1). During layout, samples to be added can be searched based on any form value, state or read length. Once selected, these samples can be manually dragged-and-dropped into a flow cell layout or automatically placed based on the request and compatible multiplexing indexes. On the flow cell, samples are displayed with their index, if present, and the suggested dilution for correct clustering is shown if sample quality information has been provided. After saving the layout, an administrator can edit the flow cell to add or rearrange samples. Once complete, the flow cell can be submitted, putting all samples in the 'flow cell prepared' state. After sequencing, the flow cell is set to 'completed', which subsequently updates all sample states. For each flow cell, the administrator can retrieve and print a sample sheet from the web interface, which lists the flow cell overview and the suggested dilutions.

2.3 Data handling and analysis

On creation of a request, a private data library is created within the Galaxy system, owned by the submitter and accessible by administrators. All data related to this request, quality data uploaded by the customer or administrator, raw and analyzed sequencing results, will be stored within this data library. This provides users and lab personal quick access and minimizes the need to send the resultant large datasets. In addition to request-specific data, a library for each flow cell is created to which data, like scanned lab protocols, can be uploaded.

A script watches the directory into which raw sequencing data are written. The script detects completed sequencing runs and retrieves the correct sample sheet from the Galaxy system. With the sample sheet, Illumina's CASAVA software is automatically started and the resulting fastq files are uploaded to the corresponding submitter data library. Depending on configuration, samples will be automatically analyzed by predefined workflows.

The Galaxy platform integrates many published and publicly available tools, including both genomic and NGS processing, analysis and interpretation tools. Because the datasets are derived within the Galaxy framework, researchers can directly analyze the data with the Galaxy tools. Researchers can upload additional datasets into their private workspace, for analysis and integration with existing sequencing data. Finally, as provided by the Galaxy framework, analysis steps, workflows and datasets are tracked, saved and can be easily exported from and imported to other Galaxy instances.

3 CONCLUSION

The presented work provides a lightweight sample-tracking system aimed at NGS centers with one or more next-generation sequencers. The system has been optimized for the Illumina HiSeq 2000 system. With barcoding, multiplexing support and assisted flow cell design, this system can be directly deployed at facilities that provide sequencing to internal or external clients. In addition, due to its flexibility and open access code, the system can be easily adapted for other core facility services. Developed inside the Galaxy framework, the user has access to a huge set of evolving and improving processing, analysis and interpretation tools.

Conflict of Interest: none declared.

REFERENCES

Bauch,A. *et al.* (2011) openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*, **12**, 468.

Blankenberg,D. *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.*, **Chapter 19**, Unit 19.10.1–21.

Giardine,B. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Nix,D.A. *et al.* (2010) Next generation tools for genomic data generation, distribution, and visualization. *BMC Bioinformatics*, **11**, 455.