Genome Biology

MEETING REPORT

Turning pipe dreams into reality

Peter Li[1,*], Jeremy Goecks[2] and Tin-Lap Lee[3,4]

Abstract

A report on the Galaxy Community Conference at the University of Illinois, Chicago, July 25-27, 2012.

Keywords Galaxy, workflows, data integration, tool integration, next generation sequencing

Introduction

The 2012 Galaxy Community Conference returned to the United States after its sojourn last year in the Netherlands (abstracts and other related materials are freely available online [http://wiki.g2.bx.psu.edu/Events/GCC2012]). This two day event was held at the University of Illinois at Chicago, hosted by the University of Illinois-Champaign (UIUC) and the University of Illinois-Chicago (UIC), and sponsored by Ion Torrent, EMC Isilon, UIUC, and UIC. The conference provided a forum for users and developers of Galaxy [http://galaxyproject.org], a Web-based platform for undertaking all facets of genomic analysis, including data retrieval and integration, multi-step analysis, reproducible analyses via workflows, visualization, collaboration, and publication. Galaxy's growing popularity was clearly evident with over 200 delegates attending from Europe, the Far East and Australia, as well as North America. Equally impressive was that the participants were a mix of software developers, bioinformaticians and biologists, underlining the diversity of the community using Galaxy.

Inaugural Galaxy training day

In a change to the previous two meetings, delegates were invited to a day of Galaxy tutorials preceding the main conference. This was a welcome addition and positively received by delegates who were given the opportunity to learn about new and updated features of Galaxy. A selection of hands-on sessions were provided by the core Galaxy team and collaborators on a range of technical topics, such as the deployment of Galaxy, defining and installing tools and data resources, and the use of Galaxy on cloud computing platforms. For those interested in extending Galaxy, tutorials were provided to learn about the Galaxy code architecture and the Galaxy Application Programming Interface (API). Applied sessions included learning how to use Galaxy for variant and single nucleotide polymorphism (SNP) analysis, and analyzing RNA-Seq data. All teaching materials are available on the conference web site.

Galaxy developments

The continued adoption of Galaxy by the life sciences community depends on the enhancement of features and development of new functionality. A number of new features were highlighted by members of the Galaxy development team, including Greg von Kuster (Penn State University, USA), who provided an update on the Galaxy tool shed, a framework built using the Mercurial distributed version control system, which enables the sharing of tools wrapped for use in Galaxy as well as workflows and data. The tool shed now has the ability to track the build version information of Galaxy tools and this will help towards supporting the reproducibility of results in computational analyses. The main public tool shed contains many tools for processing next generation sequencing (NGS) data, meaning that Galaxy has had to address issues associated with big data, an aspect of NGS that was frequently referred to in many of the conference presentations. Enis Afgan (University of Melbourne, Australia) described new features in Galaxy CloudMan, which provides researchers with access to computational and storage infrastructure for processing massive NGS data sets from the Australian National Genomics Virtual Laboratory project. The scale of data generated by NGS presents new challenges for visualizing the millions of reads. To this end, the Galaxy team presented a Web-based visual analysis framework providing track browsers, Circos plots, and the display of phylogenetic trees for viewing the results of NGS workflows.

Integration of tools and data

It is not possible for the core development team to integrate all of the bioinformatics tools into Galaxy that are required for research in the life sciences. Galaxy is reliant on its developer community to make this happen and a number of presentations detailed such integration efforts. For example, Yufei Luo (Unité de recherche en

*Correspondence: peter@gigasciencejournal.com
[1]GigaScience, BGI-Hong Kong Co. Ltd, 16 Dai Fu Street, Tai Po Industrial Estate, NT, Hong Kong SAR, China
Full list of author information is available at the end of the article

Génomique-Info, France) spoke about how the S-MART toolbox was exposed as a Galaxy tool for use in RNA-Seq analysis pipelines. Seqware is a toolkit for processing NGS data and this was incorporated into Galaxy by Zhibin Lu (Ontario Institute for Cancer Research, Canada). The large scale of NGS data can be problematic in certain situations, for example, when there is a need to transfer gigabyte volumes of data across network locations. A more efficient solution to do this compared to using standard file transfer protocols was provided by Bo Liu (University of Chicago, USA), who described work on using the Globus Online suite of tools to move data in and out of Galaxy.

The majority of bioinformatics software is designed to be used in a Linux-based environment, but it should not be forgotten that useful Windows-based applications also exist, such as those associated with mass spectrometers and microarray scanners purchased from commercial vendors. There is a need for these applications to be integrated into Galaxy pipelines so that they can, for example, process files containing data stored in a vendor's proprietary format. Since there is no support for running Galaxy in Windows, Liram Vardi (Agilent Labs, Israel) presented a project called Windows2Galaxy that can bridge the gap between a Galaxy client running on a Linux-platform calling a Windows-based tool using virtual machine technology.

Databases can also be integrated into Galaxy so that they can be queried from within workflows. Richard Park (Harvard Medical School, USA) showed how multiple data sets can be selected from a repository developed as part of a project at the Harvard Stem Cell Institute and sent to be processed by Galaxy workflows using a custom Web-based user interface. We also presented how Galaxy has been integrated with myExperiment, a repository that enables workflows to be shared in a more controlled manner, in work led by David De Roure at the University of Oxford, UK, as well as initial work on exposing the SOAP suite of NGS tools developed by the Beijing Genomics Institute (BGI, China).

Real-life applications using Galaxy

Several presentations reported on how Galaxy has been employed as a framework for users within an organization to perform data analyses. For example, Vincent Maillol and Jean-Francois Dufayard (INRA and CIRAD, France) spoke about the Bacchus pipeline they had developed for investigating clonal diversity in grapevine genomes that is used by plant breeding research teams in their country. David van Enckevort (NBIC, Netherlands) showed how Galaxy has been deployed on computational clusters at the Netherlands Bioinformatics Centre, and used in workflows for studying cancer and cardiovascular disease. Not all research groups have access to a centralized bioinformatics support unit, nor the computational infrastructure for processing NGS data, and this was highlighted in a captivating talk by Karen Reddy (Johns Hopkins University, USA). With assistance from the Galaxy team, biologists in Karen's research group were able to use Galaxy CloudMan, enabling them to process their NGS data using the Amazon cloud for studying how organization of the nucleus affects gene regulation.

Galaxy has primarily been employed for processing genomic data, so it was interesting to learn that it was now beginning to be applied in other areas of post-genomic science. For example, Ira Cooke (La Trobe University, Australia) reported on impressive work involving the integration of proteomics tools into Galaxy, whilst James Ireland (5AM Solutions, USA) presented how Galaxy can assist with the design of probes for nucleic acid targets in genotyping applications.

There was a key emphasis on community in this year's Galaxy conference, with Anton Nekrutenko (Penn State University, USA) and James Taylor (Emory University, USA) acknowledging that the work being done by the core development team is reciprocated and complemented by efforts in other organizations around the world. There was a good balance of presentations of different topics at the conference, several reporting ingenious customizations of Galaxy to meet the needs of science. This demonstrates the flexibility of the Galaxy software and shows why it has rapidly become a leading tool for developing and automating data analysis pipelines in the life sciences. Preparations for the 2013 Galaxy Community Conference, to be held in Oslo University, Norway are already well underway. With medium-term funding having been acquired for the Galaxy project, there is every reason to expect that the current momentum in the adoption of Galaxy in biology and other related sciences will continue into next year and beyond.

Competing interests
PL and TLL are both users of the Galaxy software. JG is a member of the core Galaxy development team.

Acknowledgments
The authors thank Scott Edmunds for reviewing the manuscript of this meeting report.

Author details
[1]GigaScience, BGI-Hong Kong Co. Ltd, 16 Dai Fu Street, Tai Po Industrial Estate, NT, Hong Kong SAR, China. [2]Departments of Biology, and Mathematics & Computer Science, Emory University, 1510 Clifton Road, Atlanta, GA, USA. [3]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China. [4]The CUHK–BGI Innovation Institute of Trans-omics, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China.

Published: 30 August 2012

doi:10.1186/gb-2012-13-8-318
Cite this article as: Li P, *et al*.: Turning pipe dreams into reality. *Genome Biology* 2012, **13**:318.