Schatz *Genome Biology* 2010, **11**:128
http://genomebiology.com/2010/11/8/128

Genome Biology

RESEARCH HIGHLIGHT

The missing graphical user interface for genomics

Michael C Schatz*

See research article http://genomebiology.com/2010/11/8/R86

Abstract

The Galaxy package empowers regular users to perform rich DNA sequence analysis through a much-needed and user-friendly graphical web interface.

With the advent of affordable and high-throughput DNA sequencing, sequencing is becoming an essential component in nearly every genetics lab. These data are being generated to probe sequence variations, to understand transcribed, regulated or methylated DNA elements, and to explore a host of other biological features across the tree of life and across a range of environments and conditions. Given this deluge of data, novices and experts alike are facing the daunting challenge of trying to analyze the raw sequence data computationally. With so many tools available and so many assays to analyze, how can one be expected to stay current with the state of the art? How can one be expected to learn to use each tool and construct robust end-to-end analysis pipelines, all while ensuring that input formats, command-line options, sequence databases and program libraries are set correctly? Finally, once the analysis is complete, how does one ensure the results are reproducible and transparent for others to scrutinize and study?

In an article published in *Genome Biology*, Jeremy Goecks, Anton Nekrutenko, James Taylor and the rest of the Galaxy Team (Goecks *et al.* [1]) make a great advance towards resolving these critical questions with the latest update to their Galaxy Project. The ambitious goal of Galaxy is to empower regular users to carry out their own computational analysis without having to be an expert in computational biology or computer science. Galaxy adds a desperately needed graphical user interface to genomics research, making data analysis universally accessible in a web browser, and freeing users from the minutiae of archaic command-line parameters, data formats and scripting languages. Data inputs and computational steps are selected from dynamic graphical menus, and the results are displayed in intuitive plots and summaries that encourage interactive workflows and the exploration of hypotheses. The underlying data analysis tools can be almost any piece of software, written in any language, but all their complexity is neatly hidden inside of Galaxy, allowing users to focus on scientific rather than technical questions.

What Galaxy can do for you

For most users, this high level of accessibility is the most welcome and immediate benefit of Galaxy, but this is just the beginning. Just as letting untrained people loose with construction tools does not lead to well-built houses, empowering users to run analysis tools does not in itself lead to sound results. The deeper goal of computational robustness demands that the results and methods of an analysis can stand scrutiny, and Galaxy provides its most significant capabilities in this domain. To start, Galaxy automatically records the inputs, tools, parameters and settings used for each step in an analysis, thereby ensuring that each result can be exactly reproduced and reviewed later.

This record has important short- and long-term consequences. In the short term, different parameters and thresholds can be explored, and once the analysis is done, the Galaxy record will eliminate any ambiguity as to which result used which settings. In the long term, the Galaxy history is invaluable if an unforeseen follow-up analysis is performed. For example, I have had the all too common experience of mistakenly trying to analyze targeted sequencing results by mapping the reads to build 37 of the human genome, when the coordinates for the design referenced an earlier build, leading to subtle changes and confusing results. If I had been working inside Galaxy, the exact history would have been automatically recorded, and this mistake could have been easily avoided, saving hours of wasted effort.

Beyond automatically providing provenance, Galaxy makes it easy for users to annotate each step with a human-readable description on interactive web documents called Galaxy Pages. Galaxy Pages enhance transparency far beyond the raw command list, as they can be

*Correspondence: mschatz@cshl.edu
Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

used to communicate the intent of each step with written descriptions, figures and even embedded videos and screencasts. Transparency, more so than reproducibility, is essential for verifying computational analysis, because in the extreme case a programmatic or logical error will lead to exactly the same erroneous result time after time. A well-annotated Galaxy Page helps the analyzer to catch such errors by enabling them to narrate the logical process of the pipeline, potentially with the same rigor as a mathematical proof. Users can then publish Galaxy Pages as supplementary material for a publication to document the exact stages of the analysis.

After an analysis has been carefully customized and debugged for one dataset, Galaxy users can repeatedly apply that command history on different data. Each time the workflow is run, the same sequence of tools will be executed with the same parameters as before but with the new data. This way a Galaxy user can develop a rich, organized catalog of reusable workflows rather than starting from scratch each time or trying to navigate a collection of *ad hoc* analysis scripts. In addition, users can share their workflows and Galaxy Pages on the central Galaxy website, tapping into the collective intelligence of the Galaxy community and improving the field for everyone.

Galaxy's goals are ambitious, and the project is not without limitations, but it is now the leading platform for computational analysis of DNA sequence data. The standard installation is loaded with analysis tools for trimming and preparing raw sequences [2], mapping sequences to reference genomes [3,4], cataloging variations [5] and statistically analyzing the results. I've heard of Galaxy users developing and running new analyses in hours that would have previously taken weeks of effort at the command line. Already, several papers [6-8] have been published in leading journals in which the analysis was completed within Galaxy and augmented with detailed Galaxy Pages, allowing other researchers to study and understand the methods used in greater detail than before. The public repository of Galaxy Pages, workflows, and datasets is poised to become one of the most valuable bioinformatics resources online and the first stop for analysts facing new challenges.

Use with care

Multiple studies have shown that software developers are much more productive when using higher-level abstractions such as modern programming languages, sophisticated software libraries and richer development environments [9]. However, these abstractions sometimes also cause new problems because they hide potentially important details of when they are suitable. Similarly, Galaxy users will become more productive working at a higher level, but also face new dangers of this kind.

Consider the case of a casual user discovering and running a workflow in the Galaxy repository for analyzing differential expression within an RNA-seq experiment. Even if the workflow was scrutinized and published for one dataset, the user could reach a disastrous conclusion if they failed to realize that the workflow depends on a particular library preparation or requires a certain type of technical replicate that their experiment did not use. Galaxy verifies that file formats are compatible and makes analysis accessible, but until systems for analyzing semantic dependencies of this kind are available, Galaxy cannot make analysis fully automatic and intelligent. The very popular R/Bioconductor package [10] recognizes and addresses this issue by deliberately not offering a single prepackaged analysis 'wizard' for common tasks, but instead offers a selection of choices and requires users to consider their options carefully. This is the most practical approach for Galaxy as well, but creates its own usability problems, especially the additional burden placed on the user to select the appropriate tool or workflow.

Power users may find Galaxy too restrictive because not every software package is available within it, especially cutting-edge software for novel analyses, and the graphical interface does not offer the same flexibility as a scripting environment or R/Bioconductor. However, the other benefits of Galaxy, especially its productivity, provenance tracking and transparency, may outweigh these limitations for analysis tasks leading to publication. Until massively parallel and powerful computational resources are available, all users face the frustration of working with very large datasets, where computation can run for days or weeks. Galaxy users would do best to install it on their own servers or utilize the new cloud-computing-based version that can be dynamically provisioned on demand.

A final problem with any computation-based project is whether it can enable long-term reproducibility. For example, none of the software packages I purchased for my first computer in the 1980s works today, and it is not clear if any package I use today will work in 20 years. Galaxy mitigates this problem by using open standards and building a community of users and developers beyond a single funding source, but no one knows whether future web browsers and operating systems will work on today's standards. This challenge is beyond the scope of Galaxy alone, and journals and the publication archives need to actively research how to maintain legacy software accessibility in the future, perhaps through the use of virtualized machine images for interactive or enhanced media supplementary material.

Published: 25 August 2010

References

1. Goecks J, Nekrutenko A, Taylor J; The Galaxy Team: **Galaxy: a comprehensible approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, 11:R86.
2. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A; Galaxy Team: **Manipulation of FASTQ data with Galaxy.** *Bioinformatics* 2010, **26:**1783-1785.
3. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.
4. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
6. Peleg S, Sananbenesi F, Zovoilis A, Burkhardt S, Bahari-Javan S, Agis-Balboa RC, Cota P, Wittnam JL, Gogol-Doering A, Opitz L, Salinas-Riester G, Dettenhofer M, Kang H, Farinelli L, Chen W, Fischer A: **Altered histone acetylation is associated with age-dependent memory impairment in mice.** *Science* 2010, **328:**753-756.
7. Kosakovsky Pond S, Wadhawan S, Chiaromonte F, Ananda G, Chung WY, Taylor J, Nekrutenko A; Galaxy Team: **Windshield splatter analysis with the Galaxy metagenomic pipeline.** *Genome Res* 2009, **19:**2144-2153.
8. Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, Panhuis TM, Mieczkowski P, Secchi A, Bosco D, Berney T, Montanya E, Mohlke KL, Lieb JD, Ferrer J: **A map of open chromatin in human pancreatic islets.** *Nat Genet* 2010, **42:**255-259.
9. Brooks FP: **No silver bullet: essence and accidents of software engineering.** *IEEE Computer* 1987, **20:**10-19.
10. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5:**R80.

doi:10.1186/gb-2010-11-8-128
Cite this article as: Schatz MC: **The missing graphical user interface for genomics.** *Genome Biology* 2010, **11:**128.