# Multi-omic data analysis using Galaxy

Boekel, Jorrit; Chilton, John M; Cooke, Ira R; Horvatovich, Peter L; Jagtap, Pratik D; Käll, Lukas; Lehtiö, Janne; Lukasse, Pieter; Moerland, Perry D; Griffin, Timothy J

*Published in:*
Nature Biotechnology

*DOI:*
10.1038/nbt.3134



*Publication date:*
2015

factor (IF) (http://wokinfo.com/products_tools/analytical/jcr/) ranging from 1 to 3 (**Fig. 1b**). Nine reports (1.3%) were published in journals with an IF higher than ten, whereas 77 reports (11.3%) appeared in journals with an IF <1. Additionally, there were 51 reports (7.3%) published in journals without an IF (**Fig. 1b**). Generally speaking, the IF of journals reporting GM food/feed safety research carried out in agriculture is noticeably lower than IFs of journals associated with high-profile areas of basic or clinical research.

In conclusion, GM food/feed safety issues have been and continue to be extensively studied. The cumulative number of original research reports has dramatically increased over the past years, and publication levels remain high. Different aspects of GM food/feed safety have been addressed from a scientific perspective, and animal health is the most frequently studied topic.

My analysis indicates that only approximately one-quarter of all reports investigated here have COIs related to author affiliation and/or declared funding source, with 15% not reporting funding information. We confirmed that the majority of reports have no conflict from author affiliation and funding source. In other words, at least 58.3% have no COI.

Overall, the analysis of all 698 reports collected here makes it clear that GM crops have been extensively evaluated for potential risks and that genetic modification technologies based on recombinant DNA do not carry a greater risk than other types of genetic modification. Claims either that there is not sufficient peer-reviewed literature evaluating GM food/feed safety issues or that COIs prevail in the published literature are not supported by this analysis.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

ACKNOWLEDGMENTS

M.A.S. thanks Drs. H. Campos and W. Parrott for their comments and insights on drafts of this manuscript.

COMPETING FINANCIAL INTERESTS

The author declares competing financial interests: details are available in the online version of the paper.

*Miguel A Sanchez*

*ChileBio, Santiago, Chile.*
*e-mail: masanchez@chilebio.cl.*

1. Vain, P. *Nat. Biotechnol.* **25**, 624–626 (2007).
2. Nicolia, A. *et al. Crit. Rev. Biotechnol.* **34**, 77–88 (2014).
3. Diels, J. *et al. Food Policy* **36**, 197–203 (2011).
4. Robinson, C. *et al. J. Epidemiol. Community Health* **67**, 717–720 (2013).
5. Nielsen, K. *PLoS Biol.* **11**, e1001499 (2013).

# Multi-omic data analysis using Galaxy

**To the Editor:**

Comprehensive multi-omic data acquisition has become a reality, largely driven by the availability of high-throughput sequencing technologies for genomes and transcriptomes[1], and high-resolution mass spectrometry (MS)[2,3] for the in-depth characterization of proteomes and metabolomes. Integrating genomic and proteomic data enables proteogenomic[4] and metaproteomic[5] approaches, whereas integrating metabolomic and transcriptomic or proteomic data links biochemical activity profiles to expressed genes and proteins[6]. Despite the potential for new discoveries, integrated analysis of raw multi-omic data is an often overlooked challenge[7], demanding the use of disparate software programs and requiring computational resources beyond the capacity of most biological research laboratories. For these reasons, multi-omic approaches remain out of reach for many. Here, we describe how Galaxy[8] can be used as one solution to this problem.

A scalable software framework in which disparate omics software could be effectively combined into workflows in an environment accessible to biological researchers would catalyze increased usage of multi-omic approaches. However, there are specific requirements (**Table 1**) for the success of such a framework, making its development far from simple. Although the requirements in **Table 1** are all important, some are crucial for success including the flexibility to accommodate constantly evolving data types and emerging software across omics domains, reproducibility, open and free access, and long-term sustainability.

Fortunately, some frameworks (also known as workflow management systems) already have the potential to meet these requirements. Most prominent among these are the well-established Galaxy[8] and Taverna[9] frameworks. More recently the KNIME (Konstanz Information Miner) platform has been extended for bioinformatics applications[10]; Yabi has emerged as a framework for life sciences computation in high-performance computing environments[11]; the p-GRADE/gUSE framework has been developed for general scientific workflow applications, including biological data[12]; and bioKepler (http://www.biokepler.org/) has been described as an option for large-scale biological data workflow development.

All of these frameworks have been designed with capabilities that meet most of the requirements listed for multi-omic data analysis. Therefore, in principle, one could argue for any of these as an effective choice for multi-omic data analysis workflow development and dissemination. However, in practice, two factors make the Galaxy framework stand out as an excellent, practical choice.

First, Galaxy has been in use for almost a decade and is the most established workflow framework for genomic and transcriptomic data analysis. Numerous reviews on the capabilities of Galaxy have described its flexibility, scalability and amenability to transparent sharing of complete, complex workflows[8]. Importantly, Galaxy contains hundreds of state-of-the-art tools covering two of the core domains (genomics/transcriptomics) that make up multi-omic data analysis applications. For example, numerous Galaxy tools exist for processing and assembling high-throughput sequencing data (e.g., RNA-seq data) and metagenomic data (e.g., whole genome shotgun sequencing or 16S rRNA data), important for proteogenomic and metaproteomic applications, respectively.

Second, Galaxy is poised for wide adoption in the life sciences community. As of June, 2014, some 50,000 users from around the world have registered at the public Galaxy website, and dozens of publicly available local versions of the framework are in use at institutions worldwide (https://wiki.galaxyproject.org/GalaxyProject/Statistics). As of January, 2015 >2,000 publications have cited the use of Galaxy (http://www.citeulike.org/group/16008/). Galaxy is also interoperable with other workflow systems, including Taverna, whose developers have taken steps to make their workflows operable within Galaxy (http://www.taverna.org.uk/documentation/taverna-galaxy/).

Given the practical benefits offered by Galaxy, researchers have recently begun extending the framework for applications beyond genomics and transcriptomics. The move toward multi-omic applications has begun relatively recently. A look at the software tools deposited in the Galaxy Tool Shed under the categories of 'Proteomics' and 'Metabolomics' indicates activity in these

domains beginning in 2013. This activity coincides with the maturation of proteomic and metabolomic technologies, making comprehensive data acquisition across these domains possible, and driving current needs for effective new data analysis options.

**Table 2** summarizes some of our ongoing Galaxy-based development efforts. These developments mainly focus on the implementation of open, freely available proteomic and metabolomic software that will complement Galaxy's toolbox. In our efforts to extend Galaxy, its flexibility has been key. Its well-designed application programming interface and software wrapping architecture has enabled tight integration with popular stand-alone omics software. Examples in MS-based proteomic software include the msconvert tool (http://proteowizard.sourceforge.net/tools/msconvert.html), used prominently for converting instrument-specific raw data to a standard format (mzML) that serves as a cross-platform compatible input for downstream analysis, as well as popular sequence database searching programs and tools for organizing and visualizing outputted protein identification results (a listing is provided at http://toolshed.g2.bx.psu.edu/ within the 'Proteomics' category). Many already provide outputs in standard formats conforming to community-accepted guidelines for data exchange.

For software publication and use, the Galaxy Tool Shed (http://toolshed.g2.bx.psu.edu/) is a central hub for implementing software and scripts, and facilitating easy public access. The Tool Shed community is host to a growing range of contributions from different omics fields. With this functionality, Galaxy encourages diverse software development and provides a means to publish and promote use and evaluation, and to comment on performance and suggest improvements. In addition, the Tool Shed promotes reproducible analyses by tracking software versions and providing specialized tools for managing versioned reference data.

Galaxy's flexibility extends to modifications to the underlying Galaxy framework that are crucial to accommodating the diverse software requirements of different omics software. In common with the genomics and transcriptomics fields, the MS-based proteomics and metabolomics fields have accumulated a rich selection of open and freely available software tools, in addition to a number of commercially available options. Unlike genomic and transcriptomic software, many of these options for proteomics and metabolomics are Windows-based. To meet this need, Galaxy now offers the ability to submit jobs to a remote Windows server (https://pulsar.readthedocs.org/en/latest/). Other modifications have been made to enable Galaxy to handle multiple-file data sets, necessary for many omics workflows, especially those used in MS-based proteomics applications.

**Table 1 Needs, requirements and enabled applications and outcomes for a multi-omic software framework**

| General need | Specific requirements | Enabled applications and outcomes |
|---|---|---|
| Flexible | Amenable to heterogeneous computing environments<br>Open and extendable | Integration of Linux-based software (e.g., genomics) with Windows-based software (e.g., proteomics) |
| Complete | Automate complex, multistep workflows using disparate software<br>Capture all specifications for each software in a workflow<br>Quality control methods to assess the tool quality and integration efficiency | Complex applications using diverse software (e.g., proteogenomics, metaproteomics) made more routine |
| Scalable | Compatible with high-performance computing and/or cloud environments<br>Large-memory allocation integration with diverse storage infrastructures | Processing of high-throughput nucleic acid sequencing data; sequence database searching of large-scale MS data |
| Transparent and/or shareable | Publication and sharing of complete workflows, including all software specifications and data<br>Attention to data provenance | Improved reproducibility and dissemination of even complex workflows (e.g., proteogenomics, metaproteomics); collaborative analysis of multi-omic data sets |
| Widely adopted and/or sustainable | User-friendly interface (e.g., native or Web-based GUI)<br>Open and transparent<br>Sustained by community rather than a single laboratory or funding agency | Use by bench scientists with limited computational expertise<br>Easy publishing of new software by developers<br>Community evaluation of software options; consensus on best practices and definition of standards; increased adherence to standards |

GUI, graphical user interfaces.

As a concrete demonstration of Galaxy's strengths in multi-omic analysis, we provide three examples from ongoing projects. Each shows the flexibility of Galaxy in enabling diverse software integration and making complex workflows accessible and usable for different applications. **Supplementary Figure 1** details a Galaxy-based proteogenomics workflow. Representative workflows for metaproteomics (**Supplementary Fig. 2**) and metabolo-proteomics (**Supplementary Fig. 3**) are also described.

Challenges to the widespread adoption of Galaxy include alternative, competing multi-omic analysis platforms, such as the semi-automated pipeline described by Castenella and colleagues[4] and the PEPPY software[13]. Galaxy promotes collaboration with these efforts, as most of the software modules composing these platforms could be deployed in the framework, and combined with other complementary, Galaxy-based software tools. Deposition in Galaxy of newly developed software across all omics fields will hopefully become common practice. Given its strong foothold in the bioinformatics community, Galaxy is poised to become the standard repository for multi-omic software. Such widespread acceptance by developers may present new challenges, such as ensuring that software in the Tool Shed remains accessible and operational over the long term. Fortunately, the Galaxy Team has outlined a number of steps to meet these framework maintenance challenges as they arise[8], including mechanisms to ensure the quality of the software published in the Tool Shed (https://wiki.galaxyproject.org/ReviewingToolShedRepositories)

Like any software, Galaxy requires some training to master, particularly by nonexpert researchers. The Galaxy framework provides excellent online training resources (https://wiki.galaxyproject.org/Learn and https://wiki.galaxyproject.org/Teach). The web-based interface used by Galaxy is easy to use especially when compared with the alternative of command-line interfaces normally required to operate different software tools across the omics domains. Command-line tools can be wrapped into Galaxy, providing a new user-friendly platform that increases their usability.

As multi-omic data analysis demands continue to grow, workflow management frameworks offer a way to streamline such analyses. Galaxy's flexibility should lend itself to more types of systems-level molecular data in addition to those we discuss here, for example, tools for NMR-based metabolomics or high-throughput imaging data, providing a platform for comprehensive systems biology

**Table 2 Galaxy development projects**

| Contributing institution(s) | Hosting URL | Applications emphasized |
|---|---|---|
| Netherlands Proteomics Centre (Utrecht, The Netherlands); Netherlands Bioinformatics Centre (Nijmegen, The Netherlands); University of Groningen (Groningen, The Netherlands); Academic Medical Center (Amsterdam, The Netherlands) | http://galaxy.nbic.nl/ | MS-based proteomic and metabolomic software integration; interactomics, proteogenomics |
| La Trobe University (Melbourne, Australia) | Galaxy Tool Shed under 'Proteomics' http://toolshed.g2.bx.psu.edu | Tools for general analysis and visualization of MS-proteomic data |
| University of Minnesota (Minneapolis) | https://usegalaxyp.org/ | Tools for general analysis and visualization of MS-based proteomic data; integration for metaproteomic and proteogenomic applications |
| Plant Research International, Wageningen University and Research Center (Wageningen, The Netherlands) | http://galaxy.wur.nl/ | Tools for MS-based proteomics and metabolomics; software integration for metabolo-proteomic applications |

studies. Given the one constant across all the omics fields—that technologies and data analysis needs continually change—this flexibility toward new software and data types should prove beneficial.

Galaxy's transparency and shareability also facilitates reproducible and publicly available analyses of the 'Big Data' produced in omics studies. Coupled with emerging efforts to make workflow frameworks interoperable[13,14], the sharing functions inherent to frameworks such as Galaxy could transform the way in which large-scale molecular data are exchanged, wherein raw data along with the complete workflow used for its analysis would be deposited and made publically available. With this vision in mind, we hope that this article will stimulate a much-needed discussion on the best ways to meet the challenges of multi-omic data analysis and move us closer to realizing its potential for biological discovery.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nbt.3134).*

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding for P.L.H. from the Netherlands Bioinformatics Centre SP5.12.2.1 Bioassist and 2.2.3 Biorange and the Netherlands Proteomics Centre NPC II E4.2 programs. P.D.M. acknowledges support from the Netherlands Bioinformatics Centre and Netherlands Proteomics Centre (NPC-GM WP3.2) J.B. and J.L. are supported by grants from the Swedish Research Council, Bioinformatics Infrastructure for Life Sciences (BILS) Sweden, Swedish Cancer Foundation and EU FP7 GlycoHit Project. P.L. acknowledges support from the Consortium for Improving Plant Yield (CIPY) and the 7th Framework Program FUEL4ME (FP7-ENERGY-2012-1-2stage grant number 308983). P.D.J., J.M.C. and T.J.G. acknowledge support from US National Science Foundation grant 1147079.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

*Jorrit Boekel[1,2], John M Chilton[3,11], Ira R Cooke[4,5], Peter L Horvatovich[6], Pratik D Jagtap[3,7], Lukas Käll[8], Janne Lehtiö[1], Pieter Lukasse[9], Perry D Moerland[10] & Timothy J Griffin[7]*

*[1]Department of Oncology-Pathology, Science for Life Laboratory, Karolinska Institute, Stockholm, Sweden. [2]Bioinformatics Infrastructure for Life Sciences (BILS), Stockholm, Sweden. [3]Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, USA. [4]Department of Biochemistry, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Australia. [5]Life Sciences Computation Centre, Victorian Life Sciences Computation Initiative, Melbourne, Australia. [6]Analytical Biochemistry, Department of Pharmacy, University of Groningen, Groningen, The Netherlands. [7]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, USA. [8]School of Biotechnology, Science for Life Laboratory, Royal Institute of Technology - KTH, Stockholm, Sweden. [9]Plant Research International, Wageningen University and Research Center, Wageningen, The Netherlands. [10]Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. [11]Present address: Department of Biochemistry and Molecular Biology, Pennsylvania State University, State College, Pennsylvania, USA. e-mail: tgriffin@umn.edu*

1. Lappalainen, T. *et al. Nature* **501**, 506–511 (2013).
2. Junot, C., Fenaille, F., Colsch, B. & Becher, F. *Mass Spectrom. Rev.* **33**, 471–500 (2013).
3. Low, T.Y. *et al. Cell Reports* **5**, 1469–1478 (2013).
4. Castellana, N.E. *et al. Mol. Cell. Proteomics* **13**, 157–167 (2014).
5. Armengaud, J. *Environ. Microbiol.* **15**, 12–23 (2013).
6. Martinez-Outschoorn, U.E. *et al. Cell Cycle* **10**, 1271–1286 (2011).
7. Palsson, B. & Zengler, K. *Nat. Chem. Biol.* **6**, 787–789 (2010).
8. Goecks, J., Nekrutenko, A. & Taylor, J. *Genome Biol.* **11**, R86 (2010).
9. Hull, D. *et al. Nucleic Acids Res.* **34**, W729–W732 (2006).
10. Jagla, B., Wiswedel, B. & Coppee, J.Y. *Bioinformatics* **27**, 2907–2909 (2011).
11. Hunter, A.A., Macgregor, A.B., Szabo, T.O., Wellington, C.A. & Bellgard, M.I. *Source Code Biol. Med.* **7**, 1 (2012).
12. Kacsuk, P. *et al. J. Grid Comput.* **10**, 601–630 (2012).
13. Abouelhoda, M., Issa, S.A. & Ghanem, M. *BMC Bioinformatics* **13**, 77 (2012).
14. Goble, C.A. *et al. Nucleic Acids Res.* **38**, W677–W682 (2010).

# A split-Cas9 architecture for inducible genome editing and transcription modulation

**To the Editor:**

The RNA-guided CRISPR-associated (Cas) endonuclease Cas9 has been harnessed as a tool for genome editing in mammalian cells[1,2]. In addition, strategies employing catalytically inactive Cas9 can direct effector proteins to genomic targets[3–5] to modulate transcription. Here, we demonstrate that Cas9 can be split into two fragments and rendered chemically inducible by rapamycin-binding dimerization domains for controlled reassembly to mediate genome editing and transcription modulation.

To develop a split-Cas9 system, we identified 11 potential split sites based on a crystal structure of Cas9 in complex with a single guide RNA (sgRNA) and complementary target DNA[6] (**Fig. 1a** and **Supplementary Fig. 1a**). The resulting C-terminal Cas9 fragment Cas9(C) and N-terminal Cas9 fragment Cas9(N) were fused to FK506 binding protein 12 (FKBP) and FKBP rapamycin binding (FRB) domains[7] of the mammalian target of rapamycin (mTOR), respectively, to make 11 split-Cas9 sets (split-1 through split-11)