Galaxy-P: Transforming MS-based proteomic informatics via innovative workflow development, dissemination, standardization and transparency

Timothy Griffin[1]*, John Chilton[2], James Johnson[2], Ebbing de Jong[1], Getiria Onsongo[2], Pratik Jagtap[2]

[1]Department Biochemistry, Molecular Biology and Biophysics and [2]Minnesota Supercomputing Institute

University of Minnesota, Minneapolis, MN 55455 (*contact: tgriffin@umn.edu)

UNIVERSITY OF MINNESOTA
Driven to Discover℠

What is Galaxy?

Galaxy is a web-based genomic and bioinformatics platform or workbench(1-3), akin to a life sciences core cyber infrastructure element. Galaxy was designed to address issues in genomic informatics similar to those currently faced in mass spectrometry-based informatics: software and workflow accessibility, usability, analytical transparency, reproducibility, scalability and share-ability. With Galaxy, because of its thriving community of users and developers, the burden of maintenance, as well as the benefits of innovation and sustainability, is distributed across many laboratories and researchers. Via internet and through a consistent and simple interface users have at their fingertips access to series of analytical programs and on demand tutorials guiding them through the process of multiples computational analyses and bioinformatics processing tasks. Using data provenance information and user activity tracking within the Galaxy space, history logs are recorded in stepwise increments that can be saved for future reference and shared with any or all Galaxy users or exported for publications. Galaxy also works as an invisible guide on what next steps are possible in an analytical process or pipeline development thus eliminating much of the guess work and the dreaded "then what?" or "where is that file" questions.

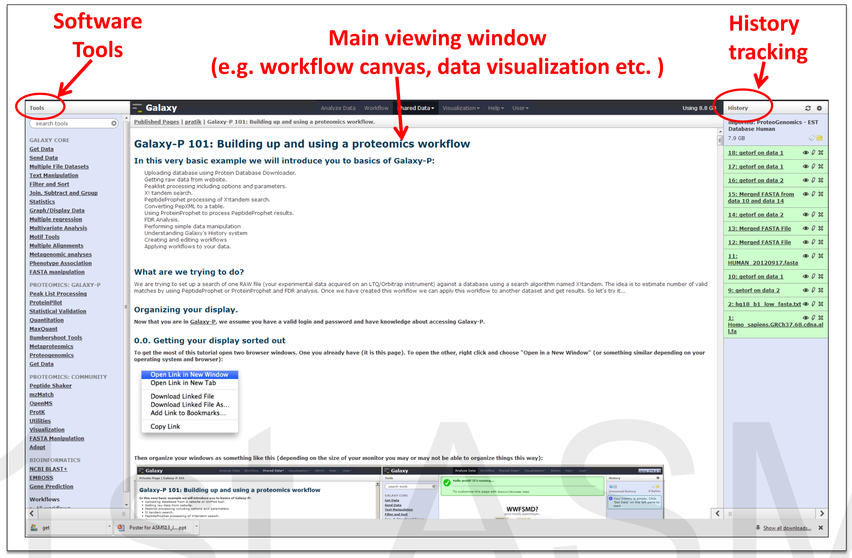
Why Galaxy for MS-based proteomic informatics?

- **It's already been built.** The open framework is already in place with highly desirable functionality for metadata capture and sharing, and straightforward integration of existing and newly developed software applications
- **It makes software integration easy.** Galaxy provides a central framework solution to answer the challenge of integrating into useful workflows the plethora of software programs continuously developed for MS-based proteomics, but lacking an easy means for their integration
- **It enables transparency and reproducibility.** Architecture facilitating complete sharing of even complex workflows, including all settings and parameters for each software program used, addressing current challenges in transparency, reproducibility and adherence to community standards
- **It enables sophisticated systems biology applications.** Because Galaxy already deploys genomics software, when extended for proteomics it is uniquely poised to enable systems biology applications (e.g. proteogenomics and metaproteomics)
- **It encourages community development and sustainment.** The open framework has an established history of community development, distributing the load in terms of advancement and sustainability

These collective advantages of Galaxy have motivated ***our extension of this framework, called Galaxy-P***. For its development we are employing a project-based strategy, using ongoing and challenging MS-based proteomic studies to direct extensions to the framework, deployment of existing and useful software programs, as well as development of new and innovative programs where necessary. Here we present a snapshot of Galaxy-P and this ongoing work.

The Galaxy-P environment

The user interface: Galaxy-P builds upon the easy and intuitive web-based interface of the Galaxy framework, shown in the screenshot below.

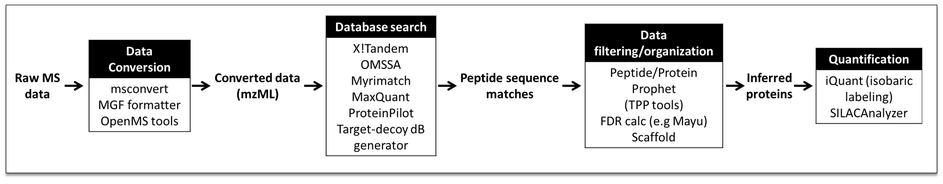


Enabling framework modifications for proteomic applications. Although highly sophisticated, the existing Galaxy framework built for genomic applications lacked a number of functionalities critical to MS-based proteomics software. We have modified the underlying framework in the following ways:

- **Windows applications.** Implemented a cross-platform server application and corresponding Galaxy job runner (client) that enables Galaxy to run jobs on remote Windows servers (**Poster 378, MP19**)
- **One sample, multiple raw files compatibility.** Enabled the framework to accept multiple files tied to a single sample (e.g. Mudpit-type applications) (**Poster 378, MP19**)
- **Cloud-enabled.** Made proteomics applications in Galaxy-P amenable to cloud computing environments (**Poster 362, MP18**)

Typical applications in Galaxy-P

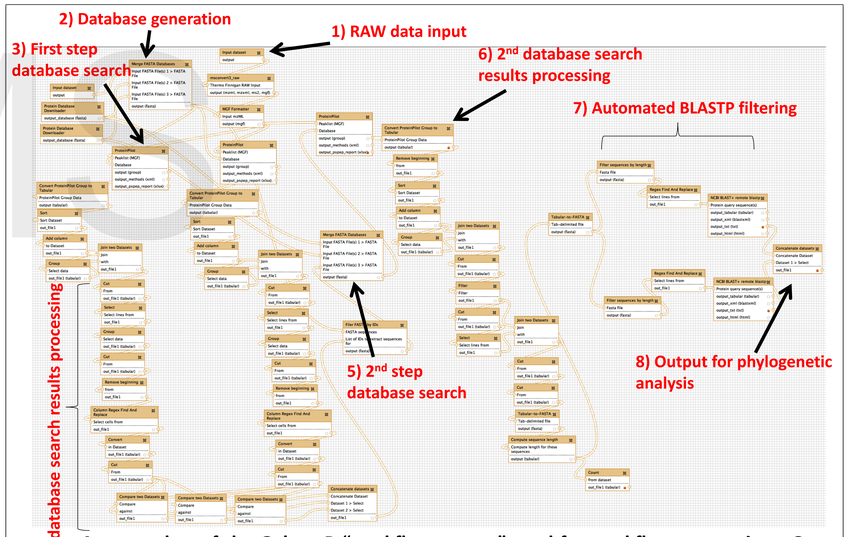
Numerous popular software programs for the most typical MS-based proteomics data analyses have been deployed in Galaxy, through our efforts building Galaxy-P as well as other community efforts. These include programs from popular software msconvert, OpenMS and MaxQuant. Shown below are a selection of these tools available in Galaxy that may be used for a typical quantitative proteomics workflow. The software shown only shows a fraction of the total software available.



Advanced applications in Galaxy-P

We are leveraging the unique features and existing genomic software in Galaxy to generate powerful workflows for more advanced applications that challenge existing informatic solutions. Some of these applications include:

- A novel, automated proteogenomics workflow (**Poster 248, TP17**). This includes simultaneous identification of novel peptide isoforms and non-host peptides, automated BLAST filtering to confirm novel protein products, a peptide sequence match evaluator/viewer, and a "Peptides to Genome" viewer for genomic mapping of newly discovered protein products.
- Quantification tools for dual isotopic/isobaric labeling (**Poster 514**, **TP28**), peptide-level quantification (**Poster 583, MP29**) and purely isobaric labeling (**Poster 474, ThP23** and **Poster 552, MP27**).
- A novel, automated workflow for metaproteomic applications (shown below), featuring our "Minnesota Two-Step"(4) database searching method addressing the challenges of large sequence databases used in such analysis, and connection with metagenomic tools available in Galaxy



A screenshot of the Galaxy-P "workflow canvas" used for workflow generation. Our "Minnesota Two-Step"(4) metaproteomics workflow is shown.

Want to use Galaxy-P?

A public server is up and running at: **usegalaxyp.org**

Instructions for building your own instance, including cloud installation, can be found at: **getgalaxyp.org**

Acknowledgements

- NSF grant 1147079
- The Center for Mass Spectrometry and Proteomics, University of Minnesota
- The Minnesota Partnership for Medical Genomics and Biotechnology
- The Minnesota Supercomputing Institute

REFERENCES CITED

1) Goecks J, Nekrutenko A, Taylor J. *Genome Biol*. 2010, **11**: R86.
2) Blankenberg D. et al. *Curr Protoc Mol Biol*. 2010, **Chap.19**: 1-21.
3) Giardine B. et al. *Genome Res*. 2005, **15**: 1451-5.
4) Jagtap P. et al. *Proteomics*, 2013, **13**: 1352-7.