GALAXY-P: RECENT DEVELOPMENTS AND EMERGING APPLICATIONS

Pratik D. Jagtap[1], James E. Johnson[1], Thomas McGowan[1], Innocent Onsongo[1], Benjamin Lynch[1], Candace R. Guerrero[1], Kevin Murray[1], Lloyd M. Smith[2], Michael R. Shortreed[2], Anthony J. Cesnik[2], Lennart Martens[3], Adrian D. Hegeman[1] and Timothy J. Griffin[1]

[1]University of Minnesota, Minneapolis, MN; [2]University of Wisconsin-Madison, Madison, WI; [3]Ghent University/VIB, Ghent, Belgium

Introduction. The Galaxy for Proteomics (Galaxy-P) project was launched several years ago, with the objective of extending the genomics-centric Galaxy bioinformatics framework (*Genome Biol.* **11**: R86) to employ proteomics informatics tools. Since its inception, the Galaxy-P project has moved its focus from proteomics tools to integrative analysis across different 'omic domains (i.e. multi-omics). The Galaxy software framework offers numerous advantages as a platform for multi-omics data analysis and informatics (*Nat Biotechnol* **33**: 137). These include the flexibility to implement and integrate disparate software programs that cross 'omic domains, scalability for large data volumes and compute-intensive operations, and easy sharing of tools and complete workflows, even those comprised of complicated, multiple step processes.

Results. Here we outline the current state of the Galaxy-P project, summarizing recent developments and emerging and future plans in multi-omic data analysis and informatics. Areas of active development described include:

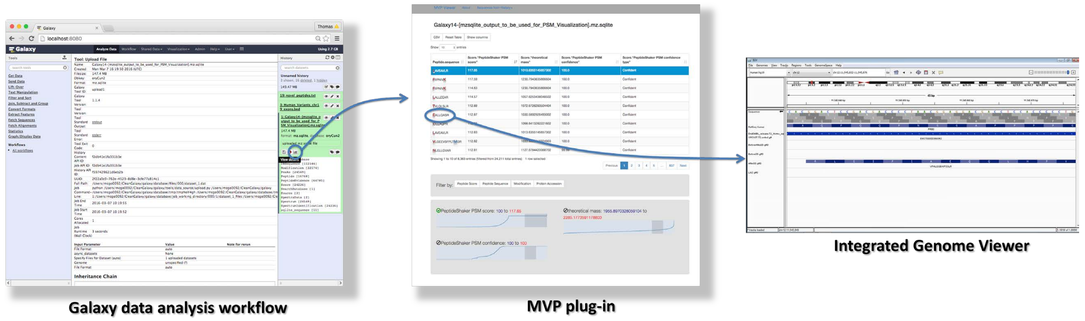
- Results visualization and interpretation
- MS-based metabolomics data analysis and informatics
- Integrative genomic-proteomic data analysis and informatics
- Collaborations and outreach activities

Conclusions. Galaxy-P has enabled numerous research studies in multi-omics. Emerging and future developments will focus on enhancing its capabilities, especially in the realm of visualization and interpretation of results, and dissemination of the high-value workflows and tools to the community.

Multi-omics visualization platform (MVP)

Current status. MVP has been developed as a Galaxy-compatible plug-in for visualization and interpretation of multi-omics results generated in Galaxy data analysis workflows. MVP utilizes standard JavaScript and open-source libraries, receiving data from a Galaxy SQLite data provider API. The plugin integrates with the Galaxy visualizations registry, such that any registered data of type mz.sqlite will be viewable from the MVP tool. Functions within MVP currently include:

- Sorting and organization of data by peptide sequence, protein accession or other annotation
- Visualization of MS/MS data via the Lorikeet viewer
- Integration with the Integrated Genome Viewer (IGV) utilizing the IGV JavaScript package for mapping and viewing peptide sequences against genomes and transcriptomes



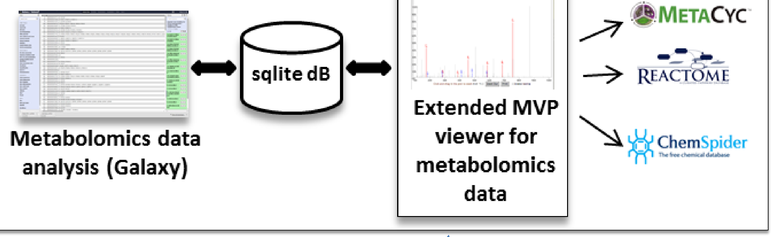
Emerging and future work. MVP has been developed with an eye toward extensibility, enabling not only visualization of data and results, but also connectivity to informatics resources to aid in interpretation. Future extensions will include:

- Enhanced functionalities for filtering peptide sequence matches and post-translational modifications
- Added functionalities for viewing and interpreting MS-based metabolomics data
- Connectivity to web-based informatics resources to enable results interpretation (e.g. CBioPortal for cancer informatics, NDEx for pathway analysis)

Metabolomics

Current status. The Galaxy platform has seen some recent developments implementing tools for MS-based metabolomics data analysis (*Gigascience* **5**:10; *Bioinformatics* **31**:1493). Metabolomics informatics is still maturing, especially in the area of identification of metabolites detected by MS. Galaxy is well-suited for the needs of metabolomics, wherein numerous software tools are needed to interpret the multi-faceted MS data for metabolite quantification and identification. Currently we are exploring the state of tools already implemented in Galaxy, with eye towards developing extensions that offer users new capabilities.

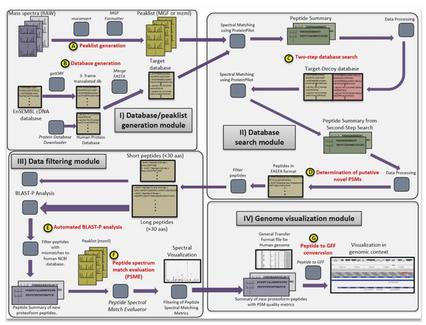
Emerging and future work. Galaxy can serve as a valuable tool for storing and interpreting multi-faceted information on metabolites of interest. We are pursuing development of a platform wherein different attributes of MS metabolite data can be captured, visualized and interpreted. Disparate information would be captured in a database, visualized via an extension of our MVP tool and analyzed via appropriate tools and informatics resources available to the community.



Multi-omic data analysis and informatics hub

Integrative analysis of genomic-proteomic data: proteogenomics and metaproteomics

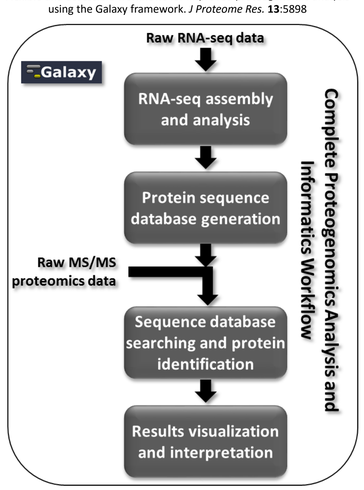
Current status. Given its already rich suite of tools for genomics and transcriptomics, coupled with ongoing efforts to implement MS-based proteomics tools, Galaxy is an obvious choice to enable integrative analysis across 'omic domains. The Galaxy-P team has published numerous papers demonstrating Galaxy's value in proteogenomics (e.g. *J Proteome Res.* **13**:5) and metaproteomics (e.g. *Proteomics* **15**:3553). These approaches require complex, multi-step and customized workflows (e.g. see schematic to right) that are well-suited for Galaxy's capabilities.



Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework. *J Proteome Res.* **13**:5898

Emerging and future work. Proteogenomics and metaproteomics both offer many analytical and informatics challenges. In ongoing work we are seeking to extend our capabilities in the following areas:

- Full integration of genomic/transcriptomic tools, proteomic tools and downstream filtering and visualization tools, building a novel environment for complete proteogenomics analysis (see workflow diagram to right)
- Improved mechanisms for interpreting the significance of identified protein variants from proteogenomic analyses via integration with informatic resources
- Deployment of additional tools to address challenges in metaproteomics such as large database searching and taxonomic and functional analysis
- Implementation of Global PTM (G-PTM) database searching (*J Proteome Res.* **15**:800) for comprehensive identification of PTMs as well as variant sequences from proteogenomics



Collaborations, outreach and training

Current status. From the outset, Galaxy-P development has been based on collaborations with biology and biomedical researchers with challenging project, such as those employing proteogenomics and metaproteomics. Numerous joint studies have been published through these collaborations (see z.umn.edu/galaxypreferences). We have also engaged in numerous collaborations with software developers (e.g. the Compomics team). We have also presented a number of workshops at national conferences, such as ASMS and ABRF, seeking to promote and train others in the use of Galaxy for multi-omic applications. Our Galaxy-based tools are made publically available through the Galaxy Tool Shed.

Emerging and future work. A main focus going forward is making the Galaxy-based multi-omics tools accessible by more researchers. Two main avenues for these dissemination include:

- We are actively working with Globus Genomics to implement our Galaxy-P instance in cloud infrastructure backed by Amazon Web Services. The Globus instance will be used for training purposes and as a scalable option for collaborative, large-scale studies
- We are leveraging the Docker technology, to create Galaxy-P "Flavours", which are customized instances that can be downloaded and easily installed on local infrastructure or implemented in cloud-based infrastructure

Acknowledgements

We thank Mark Vaudel, Harald Barsnes, Björn Grüning and Ira Cooke for assistance in optimizing and deploying software tools. We thank the Galaxy development community and core Galaxy team for continued innovation and improvement of the framework. We also acknowledge NSF grants 1147079 and 1458524, and NIH grant 1U24CA199347 for funding support.