MULTI-OMIC INFORMATICS IN THE CLOUD: GALAXY-P TAKES A RIDE ON JETSTREAM

Timothy J. Griffin[1], Matthew C. Chambers[1], James E. Johnson[2], Thomas McGowan[2], Thomas G. Doak[3], Jeremy Fischer[4], Praveen Kumar[1,5], and Pratik Jagtap[1]

[1]University of Minnesota, Minneapolis, MN; [2]Minnesota Supercomputing Institute, Minneapolis, MN; [3]National Center for Genome Analysis Support, Indiana University, Bloomington, IN; [4]UITS Research Technologies, Indiana University, Bloomington, IN; [5]Biomedical Informatics and Computational Biology, University of Minnesota, Minneapolis, MN

Introduction. The Galaxy platform (Nucleic Acids Res. **44**(W1):W3-W10) has proven a powerful solution to meet the multi-faceted requirements of multi-omic informatics. The collaborative Galaxy for proteomics project (Galaxy-P, galaxyp.org) has demonstrated Galaxy's value in integrating genomic and MS-based proteomic software for next generation applications such as proteogenomic and metaproteomic analysis (see z.umn.edu/galaxypreferences). To date, the tools and workflows developed have been most readily available to other users of Galaxy who are operating local instances of the platform. To increase accessibility, the Galaxy-P team has partnered with Jetstream, a cloud-based cyberinfrastructure aimed at supporting research computing needs. Here we describe the process of utilizing the Jetstream resource, as well as details on the multi-omic tools that are being made available on this scalable and extensible infrastructure.

Results. Working with the Jetstream team (see below for more information), initial work has focused on two main multi-omic applications, proteogenomics and metaproteomics. For both of these research areas, we have developed "gateway" Galaxy instances supported by cloud-based, cyberinfrastructure maintained by the Jetstream resource. The proteogenomics resource (see top right) offers workflows for carrying out core operations required for integrative analysis of RNA-Seq and mass spectrometry (MS)-based proteomics data. The metaproteomics resource (see middle right) offers tools for integrating metagenomics-guided data on bacterial community composition with MS-based proteomics data, as well as taxonomic and functional annotation tools, to better understand the functional dynamics of complex communities. We have also generated an educational gateway (see lower right), which serves as an instance for training researchers via workshops and other venues.

Conclusions. Jetstream provides a powerful cyberinfrastructure resource, accessible to all researchers, and supporting Galaxy-based multi-omic platforms such as those being developed by the Galaxy-P project. The resource provides scalable infrastructure needed to support big data applications in multi-omics. In addition to those gateways described here we anticipate Jetstream to provide access for more application areas in the future (e.g. metabolomics, data-independent acquisition).

What is Jetstream?

jetstream-cloud.org

Background. Jetstream (*Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*. ACM: 2792774. 1-8.) is a cloud-based cyberinfrastructure resource, supported primarily through a grant from the National Science Foundation (ACI-1445604). Development and maintenance of the resource is led by the Indiana University Pervasive Technology Institute. The resource offers the opportunity for researchers to create remote virtual machines, accessible via the web from their own computer workstation, but with orders of magnitude more computing power. More information on Jetstream is available at **jetstream-cloud.org**

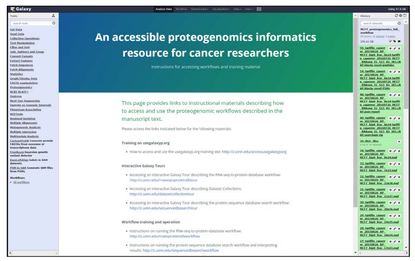
Accessing Jetstream. Resources are accessed through a web interface, based on the Atmosphere cloud computing environment customized to support science and engineering applications. The operating software environment is based on OpenStack.

Jetstream is supported by the Extreme Science and Engineering Discovery Environment (XSEDE) digital resource (www.xsede.org). U.S.-based researchers and collaborators can request an allocation, free of charge, on Jetstream through XSEDE.

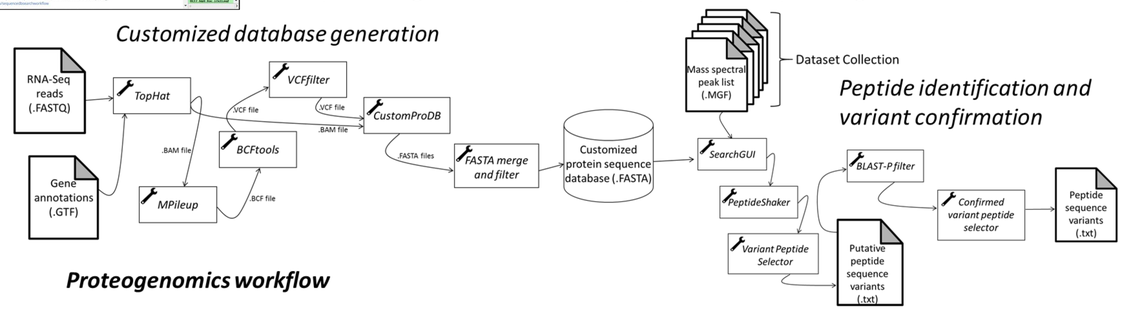
Services and capabilities. Jetstream offers pre-configured images with many popular software packages used in science and engineering. The popular Galaxy bioinformatics platform is one featured image. The resource also offers science "gateways" which include pre-configured virtual machines of workflow engines such as Galaxy, which can be deployed as persistently accessible web services and made publically available to any researcher through a URL.

Virtual machines can be configured to meet compute and storage needs of different applications. For the gateways described here, the VMs are utilizing the large capacity configurations (e.g. 44 CPUs, 120GB RAM and 4 TB storage). Extra volumes of storage can be added if additional storage is required.

Proteogenomics gateway (tiny.cc/galaxyp-proteogenomics)

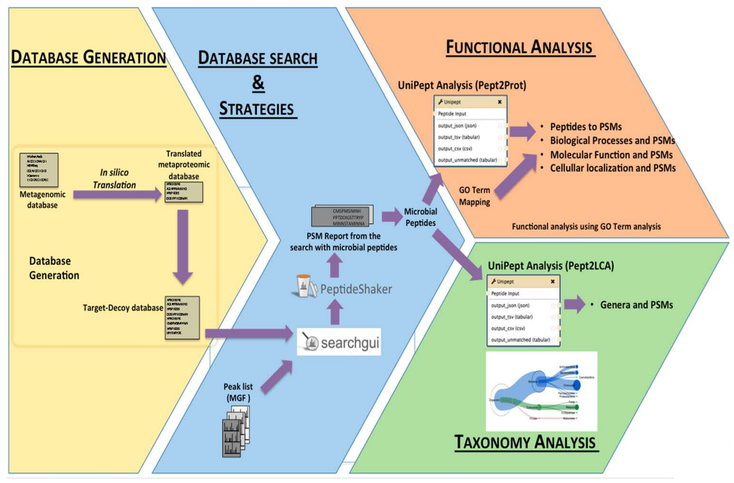


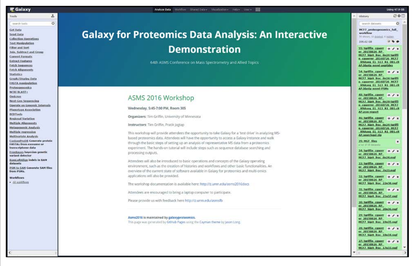
We have established a Galaxy-based gateway on Jetstream which offers tools and workflows (see below) for the core operations of proteogenomics: 1) Generation of protein sequence database from RNA-Seq data (reference proteins and selected variants); 2) Matching MS/MS spectra to sequences and confirming novelty of putative variants. This gateway provides instructions on how to use this resource, emphasizing cancer research applications of proteogenomics.



Metaproteomics gateway (z.umn.edu/metaproteomicsgateway)

Metaproteomics utilizes metagenomics data (e.g. 16sRNA, whole genome sequencing) coupled with MS-based proteomics from the same sample, to characterize proteins expressed by the community. Metaproteomics analysis can be used to capture the functional signature from bacterial communities contained in a host sample (e.g. humans) or from environmental samples. Our Galaxy-based gateway provides instructions on using our metaproteomics workflows and how to visualize the results.





Educational gateway

Jetstream also allows for special allocations focusing on educational activities. Since its inception, the Galaxy-P team has been active in conducting hands-on, training workshops for researchers interested in using bioinformatics tools for multi-omic applications. Jetstream offers an ideal resource for such training workshops, providing required computing power to enable simultaneous use of a single Galaxy instance by dozens of researchers. This provides the necessary infrastructure for interactive and hands-on training activities for researchers seeking to learn about the use of these tools.

Future Plans

- Additional functionalities to the proteogenomics gateway, including implementation of a multi-omics visualization platform
- Additional functionalities to the metaproteomics gateway, including visualization and statistical tools for comparative studies
- New gateway possibilities: Galaxy-based metabolomics and data-independent acquisition tools

Acknowledgements

We thank Subina Mehta for help in testing and documenting the Jetstream gateways. We thank the Galaxy development community and core Galaxy team for continued innovation and improvement of the framework. We also acknowledge NSF grant 1458524, and NIH grant 1U24CA199347 for funding support to the Galaxy-P team at the University of Minnesota, and NSF grant ACI-1445604 to the Jetstream team.