Scaling Galaxy for Big Data

Dave Clements, Galaxy Team

Galaxy Project, Johns Hopkins University, Baltimore, USA

Clements D (2014) *EMBnet.journal* **20**(Suppl A), e764. http://dx.doi.org/10.14806/ej.20.A.764

Galaxy[1] is a widely-used, web-based platform for data integration and analysis in the life sciences (Goecks *et al.*, 2010; Blenkenberg *et al.*, 2010; Giardine *et al.*, 2005). It is available as a free public server[2] on the web, and as open-source software that can be installed locally and on the cloud[3]. Galaxy enables life scientists to perform bioinformatics analysis using the large and varied datasets now being generated in biomedical research. It does this without requiring researchers to learn Linux system management, scripting, or command line interfaces.

In addition to making these methods accessible to bench researchers, Galaxy also enables sharing, reproducibility and transparency in research. Galaxy features a robust history mechanism that automatically and unobtrusively records all data, metadata, and analysis steps, allowing the analysis to be shared and published, and run again with the same or different data. The platform also supports creation of reusable pipelines, either *de novo*, or by extracting them from existing analyses.

This talk will introduce Galaxy and then focus on what the project is doing to scale to support complex analysis in experiments with hundreds or even thousands of samples and datasets. It also includes a discussion on the challenges faced, and how they are being addressed

Acknowledgements

The Galaxy Project is an open source project with core team members on three continents, and across the United States. It has a very active community, contributing support, code, documentations, tools, and training back to the project. Principal Investigators are located at Penn State University, Johns Hopkins University, and George Washington University. The Galaxy Project is primarily funded through NIH NHGRI grant 5U41HG006620.

Dave Clements would like to thank The Genome Analysis Center for making it possible for him to present at this meeting.

References

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, et al. (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **89**: 19.10:19.10.1–19.10.21. http://dx.doi.org/10.1002/0471142727.mb1910s89

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**(10): 1451-1455. http://dx.doi.org/10.1101/gr.4086505

Goecks, J, Nekrutenko A, Taylor, J, and The Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**(8): R86. http://dx.doi.org/10.1186/gb-2010-11-8-r86

[1] galaxyproject.org
[2] usegalaxy.org
[3] getgalaxy.org