A Genomics Virtual Laboratory

Enis Afgan[1], Clare Sloggett[2], Nuwan Goonasekera[2], Michael Pheasant[3], Ron Horst[3], Mark Crowe[4], Igor Manukin[3], Simon Gladman[2], Yousef Kowsar[2], Derek Benson[3], Andrew Lonie[2,5]

[1]Johns Hopkins University, USA; [2]University of Melbourne, Australia; [3]University of Queensland, Australia; [4]Queensland Facility for Advanced Bioinformatics, Australia; [5]alonie@unimelb.edu.au

Project URL: http://genome.edu.au
Code URL: https://github.com/galaxyproject/galaxy-cloudman-playbook

Background
Analyzing high throughput genomics data is a complex and compute intensive task, generally requiring numerous software tools and large reference data sets, tied together in successive stages of data transformation and visualization. A computational platform enabling best practice genomics analysis ideally meets a number of requirements, including: a wide range of analysis and visualisation tools, closely linked to large user and reference data sets; workflow platform(s) enabling accessible, reproducible, portable analyses, through a flexible set of interfaces; highly available, scalable computational resources; and flexibility and versatility in the use of these resources to meet demands and expertise of a variety of users. Access to an appropriate computational platform can be a significant barrier to researchers, as establishing such a platform requires a large upfront investment in hardware, experience, and expertise.

Results
We designed and implemented the Genomics Virtual Laboratory (GVL) as a middleware layer of machine images, cloud management tools, and online services that enable researchers to build arbitrary sized compute clusters on demand, pre-populated with fully configured bioinformatics tools, reference datasets and workflow and visualisation options. The platform is flexible in that users can conduct analyses through web-based (Galaxy, RStudio, IPython Notebook) or command-line interfaces, and add/remove compute nodes and data resources as required. Best practice tutorials and protocols provide a path from introductory training to practice. The GVL is available on the OpenStack-based Australian Research Cloud (http://nectar.org.au) and the Amazon Web Services cloud. The principles, implementation and build process are designed to be cloud agnostic.

Conclusion
We provide a blueprint for the design and implementation of a cloud-based Genomics Virtual Laboratory. We discuss scope, design considerations and technical and logistical constraints, and explore the value added to the research community through the suite of services and resources provided by our implementation.