Lessons from developing the Genomics Virtual Lab

Ron Horst¹, Mike Pheasant¹, Andrew Lonie²  
¹University of Queensland, r.horst@uq.edu.au  
²University of Queensland, mike@pheasant.co.nz  
³Victorian Life Sciences Computation Initiative, University of Melbourne, Australia, alonie@unimelb.edu.au

OVERVIEW
The Genomics Virtual Laboratory (GVL) is a combination of scalable genomics infrastructure, analysis platforms, resources, and support being implemented on the Australian Research Cloud. Researchers can access the GVL to directly perform genomic analyses and visualisation, or use the GVL training resources. Alternatively, research groups or institutes may instantiate, manage and tailor their own workflow platform on their own Research Cloud allocation or other infrastructure.

THE CHALLENGE
Genome research comprises the gathering, indexing and analysis of vast amounts of data, and its comparison to large public catalogues of genomic knowledge. Analyses typically involve multiple stages of data transformation using a selection from myriad bioinformatics tools, and results can be viewed with the UCSC [1], IGV [2] or other genome browsers. Biologists can develop scientific workflows (or ‘pipelines’) using web based workflow platforms such as Galaxy [3] or GenePattern [4].

These tools and data resources can be complicated to install and customise, may require significant hardware investment, and involve significant ongoing maintenance.

Furthermore, best practice and knowledge continues to rapidly evolve, typically requiring considerable expertise, up-to-date resources, and community participation to remain current.

THE GENOMICS VIRTUAL LAB
GVL leverages a number of public community driven genomics research resources:
• The Galaxy Project [5] provides an open, web-based platform for data intensive biomedical research with an emphasis on reproducible complete analyses.
• The CloudBiolinux project [6] provides a Linux command line based implementation of many popular bioinformatics tools.
• The CloudMan project [7] has extended the Galaxy and CloudBiolinux implementations to the cloud, especially to Amazon Web Services.
• The UCSC genome browser [1], provides reference sequences and working draft assemblies for a large collection of genomes.
• Science Collaboration Framework [8], for effective collaboration between researchers, such as publishing and discussing on-line content.

The GVL takes the IT out of Bioinformatics. It lets Biologists use this suite of genomics analysis tools, which currently often require specialist assistance. The tools will be available on the Australian Research Cloud, accompanied with a rich set of Researcher Resources such as:
• Protocols describing bioinformatics analysis techniques
• Tutorials to demonstrate specific uses of Galaxy tools and workflows. This includes tutorial documentation, (possibly new) tools, and Galaxy sample data and histories.
• Tools required for Protocols, Tutorials and general use.
• Workshops for accelerated learning of the tools and techniques.

PROJECT PROGRESS
Development of the GVL is funded under the NeCTAR Virtual Laboratories program [9], with prototype nodes being developed at the University of Qld and the University of Melbourne. The GVL is designed to scale to multiple locations and arbitrary cluster sizes on the Research Cloud through the CloudMan platform.

The GVL project is producing:
1. Infrastructure (tools, scripts, guides, images) for building, deploying and managing tailored GVLs including workflow (galaxy), visualisation (UCSC), collaboration (SCF) and data (framework datasets).
Users of GVL Infrastructure would typically have enough IT knowhow or local support to be able to launch and manage computer servers and storage on a Cloud allocation or own dedicated hardware.

2. **Community resources** for best practice bioinformatics including tutorials, protocols, tools and workshops. These resources are aimed squarely at practising researchers.

3. **Hosted GVL service(s),** such as our Galaxy-tut tutorial server used for tutorials. Hosted services are aimed at biologists with limited IT knowledge.

GVL Infrastructure and Community Resources are by far the most mature. Development of the Hosted GVL Service(s) on the Research Cloud Infrastructure has experienced a number of challenges.

**FIGURES**

![The Genomics Virtual Laboratory](image)

**REFERENCES**


ABOUT THE AUTHOR(S)

**Mr Ron Horst** is the Project Manager of the NeCTAR Genomics Virtual Lab project, working within the Research Computing Centre at the University of Queensland. Ron coordinates the efforts of members of the project from the University of Queensland, the Life Sciences Computation Centre at the VLSCI, Garvan Institute of Medical Research, CSIRO and others. Ron has previously been an IT Manager and Project Manager for CSIRO; and before that held a number of IT related roles within BHP.

**Dr Michael Pheasant** was until recently the manager for Genome Research Computing at UQ. He worked as a project scientist and software engineer on the UCSC Genome Browser in 2008 and 2009. He has published papers in genomics and received his PhD in Bioinformatics from the UQ in 2005 with Professor John Mattick’s lab at the IMB. He is currently involved in some private consulting.

**A/Prof Andrew Lonie** is a faculty member of the department of Computing and Information Systems at the University of Melbourne and foundation head of the Life Sciences Computation Centre, a cross-institutional centre of bioinformatics and computational biology expertise and infrastructure support within the Victorian Life Sciences Computation Initiative. Genomics, proteomics and biomedical image analysis experts within the LSCC collaborate with and support life sciences researchers in a variety of research projects across Victoria; the centre is also responsible for implementing and disseminating best practice methods and techniques, advising on experimental design and interpretation, and resourcing and maintaining informatics analysis platforms. A/Prof Lonie is also academic coordinator of the MSc (Bioinformatics) at the University of Melbourne.