The Genomics Virtual Laboratory

Andrew Lonie
Victorian Life Sciences Computation Initiative, University of Melbourne
What is the Genomics Virtual Lab?

NeCTAR funded nationally distributed platform for genomics, built on the Research Cloud and RDSI
What is the Genomics Virtual Lab?

NeCTAR funded nationally distributed platform for genomics, built on the Research Cloud and RDSI

1. Compute & workflow platforms
2. Tutorials, protocols
3. UCSC browser, datasets

http://genome.edu.au

http://nectar.org.au
What is the Genomics Virtual Lab?

In practice:

1. A way to build on-demand analysis environments for genomics

2. A set of prebuilt analysis and visualisation servers for tutorials and general use

3. Resources to teach genomics, and regular workshops using them

4. Development work exploring new capabilities
What is the Genomics Virtual Lab?

1. Build on-demand

2. Prebuilt servers for general use

3. Tutorials and workshops

4. R & D
2. Preconfigured GVL servers

Galaxy-tut  
http://galaxy-tut.genome.edu.au
- For GVL Galaxy-based tutorials
- Has tools, datasets and Galaxy histories pre-installed

Galaxy-Qld  
http://galaxy-qld.genome.edu.au
- For (Galaxy-based) research use
- Has lots of disk space, scalable as required

UCSC Browser  
http://ucsc.genome.edu.au
- Local mirror of US-based browser - quicker to upload your tracks, etc
3. Resources and tutorials

http://genome.edu.au → LEARN

Tutorials and Protocols based on Galaxy and command line

- RNA-Seq DGE
- Variant detection
- small RNA
- Microbial genome assembly
- others

Some of these are based on GVL-developed tools
1. Build an analysis environment

http://genome.edu.au → GET

Why would you want to build?

Flexibility
- Configure tools, ref data
- Command line interface

Availability

Scalability

Privacy

Most importantly, maybe: Control
GET a GVL

http://genome.edu.au → GET

Building (deploying and running) a GVL instance:

1. Create a CloudBioLinux server VM
2. Download and install a preconfigured Galaxy
3. Attach pre-populated indexed genomes data
4. Start Galaxy
5. Add extra compute nodes as required
Deploying and running a GVL

http://launch.genome.edu.au

Cloudman = Middleware for building, distributing and managing cloud-based platforms, especially Galaxy
# GET what GVL?

<table>
<thead>
<tr>
<th></th>
<th>Personal GVL</th>
<th>Server GVL</th>
<th>Cluster GVL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Suitable for</strong></td>
<td>Single user</td>
<td>Single user Small group/lab</td>
<td>Large groups Institutions</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>60GB</td>
<td>100-5000GB</td>
<td>TBs</td>
</tr>
<tr>
<td><strong>Compute</strong></td>
<td>2 cores</td>
<td>8-64* cores</td>
<td>&gt;50 cores</td>
</tr>
<tr>
<td><strong>Requires</strong></td>
<td>NeCTAR account</td>
<td>NeCTAR allocation: Compute and Volume storage</td>
<td>Large NeCTAR allocation of compute + user-provided fast storage</td>
</tr>
<tr>
<td><strong>Runs on</strong></td>
<td>Any Research Cloud node</td>
<td>RC nodes with volumes</td>
<td>RC nodes co-located with fast file system</td>
</tr>
<tr>
<td><strong>Setup</strong></td>
<td>Automatic via website</td>
<td>Automatic via website</td>
<td>Collaboration with GVL team</td>
</tr>
<tr>
<td><strong>Configuration</strong></td>
<td>No configuration required</td>
<td>Some configuration to tune analyses</td>
<td>Dedicated management</td>
</tr>
<tr>
<td></td>
<td>Personal GVL</td>
<td>Server GVL</td>
<td>Cluster GVL</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------------------------</td>
<td>---------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td><strong>Suitable for</strong></td>
<td>Single user</td>
<td>Single user</td>
<td>Large groups</td>
</tr>
<tr>
<td></td>
<td>Small group/lab</td>
<td></td>
<td>Institutions</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>60GB</td>
<td>100-5000GB</td>
<td>TBs</td>
</tr>
<tr>
<td><strong>Compute</strong></td>
<td>2 cores</td>
<td>8-64* cores</td>
<td>&gt;50 cores</td>
</tr>
<tr>
<td><strong>Requires</strong></td>
<td>NeCTAR account</td>
<td>NeCTAR allocation: Compute and Volume storage</td>
<td>Large NeCTAR allocation of compute + user-provided fast storage</td>
</tr>
<tr>
<td><strong>Runs on</strong></td>
<td>Any Research Cloud node</td>
<td>RC nodes with volumes</td>
<td>RC nodes co-located with fast file system</td>
</tr>
<tr>
<td><strong>Setup</strong></td>
<td>Automatic via website</td>
<td><strong>Automatic via website</strong></td>
<td>Collaboration with GVL team</td>
</tr>
<tr>
<td><strong>Configuration</strong></td>
<td>No configuration required</td>
<td>Some configuration to tune analyses</td>
<td>Dedicated management</td>
</tr>
</tbody>
</table>
Initiatives: NeCTAR & RDSI

- **Research Cloud node**
- **RDSI node**
- **Coming 2014-15**
Initiatives: Genomics Virtual Lab

- Research Cloud node
- RDSI node

Coming 2014-15
4. Research and Development

http://genome.edu.au

What’s next for GVL?

Genomics is characterised by:
- Large data
- Data parallel computation
- Moving best practice, tools

Case Study: Human exome analysis
Store SECURELY, LONG TERM
Analyse QUICKLY

Data 'home'

Raw data

Results

Fast-ish, 'permanent', reliable

Compute Facility eg VLSCI, NCI

Fast

Super fast

Very fast

Fast, med-term

Very fast, reliable

Ref data

Raw data to results
Analyse REPRODUCIBLY

**Data 'home'**

- **Fast**
- **Very fast**

**Raw data**

- **Fast-ish, 'permanent', reliable**

**Results**

**Workflows, Galaxy, Ruffus, PEOPLE**

**Compute Facility**

- **Super fast**
- **Very fast, med-term**

**Raw data to results**

**Ref data**
Share PUBLICLY, PRIVATELY

Data 'home'

Raw data

Results

Fast

Fast-lish, 'permanent', reliable

Very fast

Super fast

Very fast, med-term

Publishing, dissemination

Web-based, read-only, publishable

Slow

Med

Fast

Share PUBLICLY, PRIVATELY

Workflows, Galaxy, Ruffus, PEOPLE

Compute Facility
eg VLSCI, NCI

Raw data to results

Ref data

Slowly, 'permanent', reliable

Web-based, read-only, publishable

Fast, very fast, med-term

Share PUBLICLY, PRIVATELY

Workflows, Galaxy, Ruffus, PEOPLE

Compute Facility
eg VLSCI, NCI

Raw data to results

Ref data

Slowly, 'permanent', reliable

Web-based, read-only, publishable

Fast, very fast, med-term
Recommendations

1. Separate 'data' and 'compute'
   - physically and conceptually

2. Have a data 'home'
   - Secure, Long term

3. Have a HPC compute platform, well managed
   - High speed linked to 'Data Home'

4. Find a data sharing and publishing option
   **Don't buy/manage your own if you can avoid it!**
GVL 2015-

Wide user base
- students, researchers, tool developers

Broadening from genomics:
- Proteomics GVL, Metabolomics GVL

Platform for tools development and distribution

Training sessions at least fortnightly, closer ties with Galaxy folk internationally

Entire data-compute cycle on the cloud
Making the GVL possible

Go8 Universities
- The University of Queensland
- The University of Melbourne
- Monash University
- The University of Sydney
- The University of Western Australia

Medical Research Institutes
- The Garvan Institute of Medical Research
- Victor Chang Cardiac Research Institute
- Baker IDI Heart and Diabeties Institute
- Peter MacCallum Cancer Centre

eResearch Agencies
- Queensland Facility for Advanced Bioinformatics (QFAB)
- Queensland Cyber Infrastructure Foundation (QCIF)
- Life Sciences Computation Centre (LSCC) at the VLSCI
- Victorian eResearch Strategic Initiative (VeRSI)

National Agencies
- NeCTAR, DIISRTE
- CSIRO
- EMBL Australia
- Bioplatforms Australia (BPA)
- Australian Genome Research Facility (AGRF)
- Australian National Data Service (ANDS)