The Genomics Virtual Laboratory

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

Nationally distributed platform for genomics, built on the federal Research Cloud

http://nectar.org.au
Cloud node: 3-6000 cores

Data node: 1-5 PB

Coming 2014-15

The Australian Research Cloud
GVL: Drivers

To provide a genomics analysis platform with:

1. Reproducibility
2. Accessibility
3. Performance
4. Flexibility
5. Consistency
6. Functionality

for as many researchers as possible
# GVL: Design principles

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Design Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessible</td>
<td>Minimal client-side requirements</td>
</tr>
<tr>
<td>Reproducible</td>
<td>Workflow support + software &amp; tool management process</td>
</tr>
<tr>
<td>Performance</td>
<td>User-managed scaling of compute resources + high availability resources</td>
</tr>
<tr>
<td>Flexible</td>
<td>User configurable + administrable Multiple interaction modes</td>
</tr>
<tr>
<td>Consistent</td>
<td>Single platform from training to analysis</td>
</tr>
<tr>
<td>Functional</td>
<td>Pre-populated with suite of tools for common use cases + required reference data + visualisation options</td>
</tr>
</tbody>
</table>
# GVL: Design implications

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Design Implication</th>
<th>Technical implication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessible</td>
<td>Minimal client-side requirements</td>
<td><strong>Web based</strong> tool and management interfaces</td>
</tr>
<tr>
<td>Reproducible</td>
<td>Workflow support + software &amp; tool management process</td>
<td><strong>Workflow platforms</strong> + automated process for <strong>deployable underlying environment</strong></td>
</tr>
<tr>
<td>Performance</td>
<td>User-managed scaling of compute resources + high availability resources</td>
<td><strong>Cloud-based architecture</strong> + interface for managing resources</td>
</tr>
<tr>
<td>Flexible</td>
<td>User configurable + administrable</td>
<td><strong>Per-user instances</strong> accessible through web and command line; user-administrable environment</td>
</tr>
<tr>
<td>Consistent</td>
<td>Single platform from training to analysis</td>
<td><strong>Tutorials and guides</strong> for training using best practice tools + <strong>scalability</strong></td>
</tr>
<tr>
<td>Functional</td>
<td>Pre-populated with suite of tools for common use cases + required reference data + visualisation options</td>
<td>Process for <strong>building underlying images</strong> Automated configuration of <strong>reference datasets</strong></td>
</tr>
</tbody>
</table>
For as many researchers as possible...

Galaxy Main
GVL: Philosophy

Genomics Virtual Lab

Galaxy Main
GVL: In practice

Online tutorials and protocols
GVL: In practice

Managed services

- Galaxy
- Genome Browser

Online tutorials and protocols
GVL: http://genome.edu.au

**GET**

*GVL Workbench*

Workflow & analysis platforms
Cloud based compute

**USE**

*Managed services*

Galaxy
Genome Browser

**LEARN**

Online tutorials and protocols
GVL: Developer’s perspective

GET

GVL Workbench
Workflow & analysis platforms
Scalable compute
Fast reference data

Managed services
Galaxy
Genome Browser

USE

LEARN

Online tutorials and protocols
What characterises genomic analysis?

eg: Differential Gene Expression

“What genes are turned on in blue cells and turned off in red cells?”
Differential Gene Expression

2E7 observations over 1E5 genes for 6 samples

(30GB raw data)

“What genes are turned on in blue cells and turned off in red cells?”
Differential Gene Expression

DATA REDUCTION

- Workflows
- Large reference data
- High compute

2E7 observations over 1E5 genes for 2x3 samples
Differential Gene Expression

**DATA REDUCTION**
- Workflows
- Large reference data
- High compute

**DATA ANALYSIS**
- Interactive
- Flexibility

2E7 observations over 1E5 genes for 2x3 samples

Test difference: [1,2,3] vs [4,5,6]
Differential Gene Expression

DATA REDUCTION
- Workflows
- Large reference data
- High compute

DATA ANALYSIS
- Interactive
- Flexibility

2E7 observations over 1E5 genes for 2x3 samples

Test difference: [1,2,3] vs [4,5,6]
Differential Gene Expression

DATA REDUCTION
- Workflows
- Large reference data
- High compute

DATA ANALYSIS
- Interactive
- Flexibility

DATA INTERPRETATION
- External services
- Visualisation tools

2E7 observations over 1E5 genes for 2x3 samples

Test difference: [1,2,3] vs [4,5,6]

Biological pathways affected? Mechanism?

Pathway knowledge
Differential Gene Expression

**Characteristics**

**DATA REDUCTION**
- Workflows
- Large reference data
- High compute

**DATA ANALYSIS**
- Interactive
- Flexibility

**DATA INTERPRETATION**
- External services
- Visualisation tools

**Tools**

- 2E7 observations over 1E5 genes for 2x3 samples

**Domain Context**

- Test difference: [1,2,3] vs [4,5,6]
- Biological pathways affected? Mechanism?
What characterises genomics?

- Very large experimental datasets per user/group
- I/O intensive high compute initial analysis
  - ‘data reduction’: raw data to sample summaries
- Large suite of data analysis tools, interactive
  - a bit subjective
- Complex context for interpretation, external tools
  - more subjective, domain knowledge
- Little modelling/simulation
GVL Workbench: Requirements

A web-based per-user workbench providing:

- access to multiple tools
- on a scalable back end compute cluster
- with fast access to large reference data,
- user administrable and configurable
- with multiple modes of interaction
- and a mechanism for reproducible workflows

all highly available and accessible
i.e. with a minimal cost of entry to the user
Why per-user?

Managed service: objective

A short time later…
Why per-user?

Managed service: objective

A short time later…
GVL: Philosophical assertion
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
GVL Workbench: Architecture

GVL Workbench
Workflow & analysis platforms
Workbench: Architecture

GVL Workbench

Workflow & analysis platforms
Workbench: Architecture

GVL Workbench
Workflow & analysis platforms

BIOINF TOOLS

Galaxy

User Data

Research Cloud
Workbench: Architecture

GVL Workbench
Workflow & analysis platforms
Fast reference data

Indexed Genomes

BIOINF TOOLS

Galaxy

User Data

Research Cloud

Cloud BioLinux
Workbench: Architecture

GVL Workbench

- Workflow & analysis platforms
- Fast reference data

Indexed Genomes

BIOINF TOOLS

Galaxy

IP[y]: IPython Interactive Computing

Research Cloud

User Data

Cloud BioLinux
Workbench: Architecture

GVL Workbench
Workflow & analysis platforms
Fast reference data
Scalable compute
Workbench: Architecture

Indexed Genomes

BIOINF TOOLS

Galaxy

User Data

IPython: Interactive Computing

Cloud BioLinux

Research Cloud

Research

User Data
Engineering: Deploying and running a GVL

http://launch.genome.edu.au

Cloudman = Middleware for building, distributing and managing cloud-based platforms, especially Galaxy
Workbench: Engineering

GVL base image (OS, sys pkgs)

1

GVL Launcher

User Data
Workbench: Engineering

Galaxy FS snapshot (Galaxy + DB + tools)

CloudMan

GVL base image (OS, sys pkgs)

BIOINF TOOLS

Galaxy

User Data
Workbench: Engineering

- Centrally managed reference data (indexed genomes)
- Galaxy FS snapshot (Galaxy + DB + tools)
- GVL base image (OS, sys pkgs)

CloudMan

3

Indexed Genomes

BIOINF TOOLS

Galaxy

GlusterFS

User Data
Workbench: Engineering

Centrally managed reference data (indexed genomes)

Galaxy FS snapshot (Galaxy + DB + tools)

GVL base image (OS, sys pkgs)

Indexed Genomes

BIOINF TOOLS

Galaxy

User Data

CloudMan
Workbench: Engineering

- Centrally managed reference data (indexed genomes)
- Galaxy FS snapshot (Galaxy + DB + tools)
- GVL base image (OS, sys pkgs)
- Indexed Genomes
- BIOINF TOOLS
- Galaxy
- User Data
- IPython: Interactive Computing
- R Studio
- Cloud BioLinux
- Research Cloud
- 5
Workbench: All components

Management layer
- CloudMan
- GVL Launcher

Application layer
- Galaxy
- R Studio
- IP[y]: IPython Interactive Computing

System layer
- Cloud BioLinux
- GridEngine
- GlusterFS

Cloud layer
- openstack
- Amazon Web Services

Component build layer
- Base
- Galaxy FS
- Ref data
GVL: Does it work?

Technically?

Practically?
<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</td>
<td>Large groups</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Small group/lab</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>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>
</tbody>
</table>

[http://genome.edu.au → GET](http://genome.edu.au → GET)
Lessons?

Defining and maintaining a set of tools is challenging

Providing per-user performance is challenging

The cloud is only so scalable!

Not all cloud nodes are equal

Geography matters
Lessons?

Defining and maintaining a set of tools is challenging

Providing per-user performance is challenging

The cloud is only so scalable!

Not all cloud nodes are equal

Geography matters

Resourcing is key!
What’s next for GVL?

http://genome.edu.au

Moving data around is a problem

Whole genomes: 300GB raw data

We need to remove the desktop and USB sticks from the process!
Genomics lifecycle on the cloud: GenomeSpace
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 Diabetes 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)