Genomics Virtual Laboratory Workshop

Enis Afgan (VLSCI)
Michael Pheasant (UQ)
Andrew Lonie (VLSCI)
Clare Sloggett (VLSCI)

@ eResearch Australasia 2012

http://genome.edu.au
Workshop outline

1. Overview of the GVL *(15 mins)*

2. Instantiation of a scalable GVL instance *(30 mins)*

3. 4. Simple genomic analysis on the GVL instance *(90 mins)*

4. 3. Tailoring of a GVL instance with new tools from the Galaxy Toolbox *(30 mins)*
Genomics Virtual Laboratory

Best practice tools, workflows and visualisation
+ Accessibility
+ Scalability
+ Collaboration
+ Reproducibility

Research Cloud

Public Instance
- UCSC Browser Server
- Galaxy Server
- User data
- Galaxy Workers
- Public data catalogues

Private Instance(s)
- UCSC Browser Server
- Galaxy Server
- User data
- Galaxy Workers

Development instance(s)
The big picture

A. Users in different labs
B. Isolated Galaxy instance(s)
C. Dense data center

Galaxy
CloudMan
IaaS
What is CloudMan?

A cloud manager that orchestrates all of the steps required to provision, manage, and share a cluster compute platform on a cloud infrastructure, all through a web browser.
Instantiating a CloudMan Platform

1. An account on the supported cloud
2. Start a master instance via BioCloudCentral.org or the cloud web console
3. Use the CloudMan web interface on the master instance to manage the platform
Instantiation of a small scalable GVL instance on the Research Cloud AWS
Start an Instance

BioCloudCentral

Easily launch CloudMan, CloudBioLinux and Galaxy platforms on Cloud Computing resources (including Amazon Web Services).

- **Cluster name**: Current Protocols Demo
- **Password**: **************
- **Cloud**: Amazon (AWS EC2)
- **Access key**: AKIAJKMSM6GLSW7V2CPA
- **Secret key**: EjvIMvij9SLVuxvb9OgaD58qiUXLkEZLa1
- **Instance type**: Large (4 ECUs / 7.5GB RAM)

Show advanced startup options

Start an instance

This website is an open service developed by the CloudBioLinux and CloudMan communities. The goal is to make it easy to get started doing scalable biological analysis on cloud resources. See this guide for a detailed usage example when using the Amazon cloud. The open source code is available on GitHub allowing you to also run this service locally.

This site can be used for any of the available clouds. Note that you must have appropriate credentials for the chosen cloud. If a desired cloud is not available and you would like to see it there, please contact us.

Launching servers on the Amazon cloud will incur usage fees from Amazon for their resources. By using this service you acknowledge your sole responsibility for any costs accrued.
Configure Your Cluster

Initial Cluster Configuration

Welcome to CloudMan. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and provide the associated value, if any.

- **Galaxy Cluster**: Galaxy application, available tools, reference datasets, SGE job manager, and a data volume. Specify the initial storage size (in Gigabytes):
  
  [ ] GB

- **Share-an-Instance Cluster**: Derive your cluster form someone else's cluster. Specify the provided cluster share-string (for example, cm-0011923649e9271f17c4f83da6846db0/shared/2011-08-19--21-00):

  [ ] Cluster share-string

- **Data Cluster**: A persistent data volume and SGE. Specify the initial storage size (in Gigabytes):

  [ ] GB

- **Test Cluster**: SGE only. No persistent storage is created.

Hide extra options

Start CloudMan Cluster
Manage Your Cluster

CloudMan from Galaxy

CloudMan Console

Welcome to CloudMan. This application allows you to manage this instance cloud cluster and the services provided within. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to manage services provided by the application.

- Terminate cluster
- Add nodes
- Remove nodes
- Access Galaxy

Status

Cluster name: ghem
Disk status: 0 / 0 (0%)
Worker status: Idle: 4 Available: 2 Requested: 5
Service status: Applications • Data •

Autoscaling is off. Turn on?
Customizing, Sharing, Scaling

Value Added Features
Deploying your own CloudMan platform

- CloudMan provides a general-purpose *cluster-in-the-cloud* solution
- It works on Amazon, OpenStack, OpenNebula, and Eucalyptus
- It can *work with any application*
- Can be deployed on any (Ubuntu) machine image

More @ http://usecloudman.org
Simple genomic analysis on the GVL instance
Goals for this part of the workshop

1. Introduce Galaxy
2. Introduce BED file format
3. Hands-on experience:
   • Load and integrate data from online resources
   • Perform bioinformatics analysis with Galaxy
   • Save, share, describe and publish your analysis
   • Visualize your results

This workshop will not cover details of how the tools are implemented or new algorithm designs or which assembler or mapper or ... is best for you.
What is Galaxy?

- A **data analysis and integration** framework
- **Any UNIX command line tool can be wrapped** and automatically exposed as an available tool, allowing it to be integrated into an open-ended analysis
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
- These options result in several **ways to use Galaxy**

http://galaxyproject.org
Galaxy Vision

Galaxy is an open, Web-based platform for accessible, reproducible, and transparent computational biomedical research
Join two Datasets (version 2.0.2)

Join:
9: Generate pileup converted pileup
using column:
  c1
with:
  4: NA12878.chr22.exome.22_filtered
and column:
  c1
Keep lines of first input that do not join with second input:
  No
Keep lines of first input that are incomplete:
  No
Fill empty columns:
  No

Execute

This tool will attempt to reuse the metadata from your first input. To change metadata assignments click on the "edit attributes" link of the history item generated by this tool.

TIP: If your data is not TAB delimited, use Text Manipulation->Convert

Syntax
This tool joins lines of two datasets on a common field. An empty string ("") is not a valid identifier. You may choose to include lines of your first input that do not join with your second input.

Columns are referenced with a number. For example, 3 refers to the 3rd column of a tab-delimited file.

Example
Dataset1:
  chrl 10 20 geneA
  chrl 50 80 geneB
  chr2 10 40 geneC

Dataset2:
  geneA tumor-suppressor
geneB Foxp2
geneC Gnas1
geneD INX4A

Joining the 4th column of Dataset1 with the 1st column of Dataset2 will yield:
  chrl 10 20 geneA tumor-suppressor
cr1 50 80 geneB Foxp2
Some Galaxy Terminology

Dataset:
Any input, output or intermediate set of data + metadata

History:
A series of inputs, analysis steps, intermediate datasets, and outputs

Workflow:
A series of analysis steps
Can be repeated with different data
On human chromosome 22, which coding exons have the most SNPs in them?
A Rough Plan

- Get some data (and explain BED)
  - Coding exons on chromosome 22
  - SNPs on chromosome 22
- Mess with it (and explain Galaxy operations)
  - Identify which exons have repeats
  - Count repeats per exon
- Visualize our results
Bioinformatics file formats

BED – Browser Extendible Display

In BED format each line represents a genomic feature that can be displayed in a genome browser. BED records have three required fields and nine additional optional fields. Most commonly the first 5-6 fields are used:

- **chrom** - The name of the chromosome (e.g. chr3, chrY)
- **chromStart** - The starting position of the feature in the chromosome. The first base in a chromosome is numbered 0.
- **chromEnd** - The ending position of the feature in the chromosome
- **name** - Defines the name of the feature
- **score** - A score between 0 and 1000, generally used to color the feature in the browser display
- **strand** - Defines the strand - either '+' or '-'.

```
chr22 16258185 16258303 uc002zlh.1_cds_1_0_chr22_16258186_r 0 -
chr22 16266928 16267095 uc002zlh.1_cds_2_0_chr22_16266929_r 0 -
chr22 16268136 16268181 uc002zlh.1_cds_3_0_chr22_16268137_r 0 -
chr22 16269872 16269943 uc002zlh.1_cds_4_0_chr22_16269873_r 0 -
chr22 16275206 16275277 uc002zlh.1_cds_5_0_chr22_16275207_r 0 -
```
Exons, from UCSC

SNPs, from UCSC
Exons, from UCSC

SNPs, from UCSC

Exons, from UCSC

SNPs, from UCSC

Joined pairings
Exons, from UCSC

SNPs, from UCSC

Exons, from UCSC

SNPs, from UCSC

Joined pairings

Exon SNPs counts

3

1

5

3

genomics
VIRTUAL LAB
Analysis in one picture
usegalaxy.org/galaxy101

YOUR TURN
Repetitious History → Reusable Workflow?

- The analysis we just finished was about
  - Human chromosome 22
  - Exons with most SNPs
- But, ...
  - There is nothing inherently in the analysis about human, chromosomes, exons or SNPs
  - It is a series of steps that sets the score of one set of features to the number of overlaps from another set of features.
Hands On: Basic Analysis, Take 2

On human chromosome 2218, which coding exons have the most SNPs (v135 instead of v131) in them?
Histories: Data
Datasets from previous histories can be imported into current one. Resume any previous history
Current history can be cloned

Workflows: Analyses
Can be extracted from any history
Allows you rerun analysis with different inputs, settings
Reproducible Workflow

Input dataset
output

Join
with
output (interval)

Sort Query
out_file1

Group
Select data
out_file1 (tabular)

Select first
from
out_file1

Compare two Datasets
Compare
against
out_file1
Persistence, Sharing, and Publishing
More Galaxy Terminology

Dataset:
Any input, output or intermediate set of data + metadata

History:
A series of inputs, analysis steps, intermediate datasets, and outputs

Workflow:
A series of analysis steps
Can be repeated with different data

Share:
Make something available to someone else

Publish:
Make something available to everyone
Managing Histories and Datasets

Give every history and dataset a clear name.

Datasets and histories can also have annotation and tags.

Each history has an options/actions list.
Histories, workflows, visualizations and pages can be shared with others or published to the world.  

http://usegalaxy.org/u/aun1/p/windshield-splatter
Sharing for Galaxy Administrators Too

Data Libraries
Make data easy to find

Genome Builds
Care about a particular subset of life?

Galaxy Tool Shed
Wrapping tools and datatypes
Tailoring of a GVL instance with new tools from the Galaxy Toolbox
Galaxy Tool Shed

- Allow users to share “suites” containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies

- Version controlled - multiple versions installed

- Community driven - rating, annotation, comments, review

- Integration with Galaxy instances to automate tool installation and updates
Galaxies on private clouds

Galaxies on public clouds

Galaxy Tool Shed

http://usegalaxy.org

http://usegalaxy.org/community

1 2 3 ... ∞

private Galaxy installations
OFF TO THE SHED
Acknowledgments