CSIRO Galaxy Bioinformatics Portal
User Friendly Bioinformatics on HPC

Joel Ludbey | Technical Solutions Engineer
May 2018
Overview

• What is the Galaxy portal?

• How does it tie into CSIRO HPC Infrastructure?

• What can I do with it?

• Why would I want to use it?
Some History

- Galaxy pilot (GVL) made available in September 2012
  - Multiple introductory workshops
  - 101 users as of May 2013
  - Single machine

- Galaxy production service plan started June/July 2013
  - Enterprise infrastructure & redundancy
  - HPC infrastructure integration
  - .......?
Galaxy Production Service

• The current production Galaxy service
  • https://galaxy-bio.csiro.au
  • Enterprise infrastructure
  • Full HPC integration
    – Provenance and consistency across Galaxy & HPC infrastructure
  • Managed deployment (Ansible & Bitbucket)
    – Reproducible
    – Software defined service
    – Simplified updates
Galaxy Production Service...

sudo -i
apt-get install software-properties-common
apt-add-repository ppa:ansible/ansible
apt-get update
apt-get install ansible git

Clone the playbook and install requirements

```
git clone https://bitbucket.csiro.au/scm/gal/gal-csiro-ansible-galaxy.git
cd csiro-ansible-galaxy
```

You will need to make the required changes for your configuration before executing the playbook. Inside the `inventory_files` folder, you will find a file called `all`. This will define your galaxy hosts and your cluster virtual environment hosts for their respective playbooks.

Execute the playbook for the galaxy install first.

```
sudo ansible-playbook --inventory-file inventory_files/all galaxy.yml
```

Execute the playbook for the HPC cluster virtual environment. You will need permissions to impersonate the galaxy user on the HPC clusters and configured Passwordless SSH Keys for the playbook to succeed.

```
anible-playbook --inventory-file inventory_files/all galaxy-cluster-venv.yml --private-key=/ssh/keys/$PRIVATE_KEYS --user=$USER_ID$ --extra-vars "ansible_sudo_pass=$PASSWORD$"
```
What is Galaxy?

• A scientific workflow, data integration and data analysis platform focused around computational biology.

• Accessibility!
  • Simplified, reasonably intuitive GUI
  • Abstraction of HPC infrastructure

• Reproducibility!
  • Detailed provenance information for each step of an analysis
  • Historical record of input/output datasets
Galaxy Tools

• Wrapped in a script and exposed via xml files describing:
  • Input/Output datasets
  • Commands and parameters
  • Hep data, tests, citations

• Thousands of tools currently available to install
  • https://toolshed.g2.bx.psu.edu/
  • Multiple version of tools can exist on a given system
Galaxy Histories

• Individual and independent analysis
  • Multiple histories/multiple analyses
  • Exportable to visual workflows
  • Shareable

• Reproducibility and provenance
  • Name, format, creation date, data specific meta-data
  • Tool ID’s, versions, parameters, stdout, stderr
  • State: Waiting Running Success Failed
Galaxy Workflows

- Exported from history item or create manually
  - Drag and drop interface
  - Tools are added and configured manually on the workflow canvas

- Why?
  - Re-run same analysis on different dataset or using different parameters
  - Create sub-workflows
  - Collaboration
Galaxy Workflows...
Galaxy & HPC Infrastructure

• Where does HPC come in?
  • Compute via Bracewell & Pearcey HPC Clusters
  • Data storage on Bowen Storage Appliances
Galaxy & HP Infrastructure...

- No HPC account or command line
- 340+ Nodes on Pearcey (July 2018)
- 114 Nodes on Bracewell
  - 456 GPU’s (P100)
Galaxy & HP Infrastructure...

- March 302 jobs
- April 138 jobs
- May 275 jobs
- 2018 1000+ jobs
Now Everything Goes Wrong!

- Live demo time!
- https://galaxy-bio.csiro.au
Summary

• Accessible and user friendly GUI for bioinformatics
• Access to 450+ compute nodes 8000+ CPU cores
• HPC layer is abstracted from the user
  • No CLI, no batch jobs, no scripting
• Reproducible workflows
• Chain of provenance for individual analyses
• Shareable/Collaborative workflows
• https://galaxy-bio.csiro.au
Thank you

IM&T Scientific Computing
Joel Ludbey
Technical Solutions Engineer

IM&T Scientific Computing
Ondrej Hlinka
Technical Solutions Consultant

Defense Science & Technology
Stephen McMahon
Engineering Manager

IM&T Scientific Computing
Tim Ho
Technical Solutions Manager