Establishing a National Genomics Virtual Laboratory with Galaxy CloudMan

Enis Afgan, Clare Sloggett, Andrew Lonie, Michael Pheasant

GCC 2012, Chicago

3 things

- New CloudMan features
- Blend library
- Creating a National Genomics Virtual Lab

CloudMan Features

- Start/launch through a web browser or the command line
- Choose between four cluster types
- Multiple cloud support
- Terminate/restart
- Scale
- Auto-scale
- Spot support
- Persist changes
- Share-an-instance (customized one too)
- Expand the file system
- Customize (via CLI or the Tool Shed): tools, data, references
- Access through ssh
- File system access for any S3 bucket -> data library
- Control the Galaxy process
- API (via Blend)
- Build your own image (via mi-deployment)

CloudMan Features

- Start/launch through a web browser or the command line
- Choose between four cluster types
- Multiple cloud support
- Terminate/restart
- Scale
- Auto-scale
- Spot support
- Persist changes
- Share-an-instance (customized one too)
- Expand the file system
- Customize (via CLI or the Tool Shed): tools, data, references
- Access through ssh
- File system access for any S3 bucket -> data library
- Control the Galaxy process
- API (via Blend)
- Build your own image (via mi-deployment)

Multiple Cloud Support

- Support for AWS, OpenNebula (thanks to Mattias de Hollander!), and OpenStack cloud middleware
- Run the same setup on a private or public cloud
  - Infrastructure setup is still automated

Spot Support

- Support for AWS Spot instances
  - Generally at ~30% cost of a full instance
  - Depend on demand and can terminate at any moment

Use S3 Buckets as File Systems

- Mount any S3 bucket as a local file system
- Gain fast access to AWS Public Datasets (e.g., 1000genomes)
- Also works on non-AWS clouds

APIs & Blend

- CloudMan has an API

- Blend

  A Python library for interacting with Galaxy's and CloudMan's APIs

```
$ [sudo] pip install blend-lib
```

Requesting the infrastructure

```
# Request an instance of CloudMan
from blend.cloudman.launch import CloudManLaunch
cml = CloudManLaunch('access_key', 'secret_key',
cloud)
cml.launch('cluster_name', 'img_id', 'inst_type', 'pwd')
cml.get_status()
```

Manipulating the infrastructure

```
# Manage an instance of CloudMan
from blend.cloudman import CloudMan
cm = CloudMan('http://115.146.92.145', 'pwd')
cm.initialize(type="Galaxy")
cm.get_status()
cm.add_nodes(3)
cm.enable_autoscaling(0, 5)
cm.get_galaxy_state()
```

Working with Galaxy

```
# Import data into a Data Library and execute a workflow with it
from blend.galaxy import GalaxyInstance
gi = GalaxyInstance('http://usegalaxy.org', 'your API key')
lib = gi.libraries.create_library('Set 13')
d1 = gi.libraries.upload_file_from_url(lib['id'], 'http://
tinyurl.com/gcc-exons')
d2 = gi.libraries.upload_file_from_url(lib['id'], 'http://
tinyurl.com/gcc-snps')
gi.histories.create_history('Run 13')
w = gc.workflows.get_workflows()[0]
gi.workflows.show_workflow(w['id'])
dataset_map = {'25': {'id':d1['id'], 'src':'ld'},
                     '27':  {'id':d2['id'], 'src':'ls'}}
gi.workflows.run(lib['id'], dataset_map, 'F'+lib['id'])
```

Docs and examples included
http://blend.readthedocs.org/

Project Versions
latest

RTD Search
Go
Full-text doc search.

Table Of Contents
Blend
About
Installation
Usage
Development
API Documentation
BioCloudCentral.org AP
CloudMan API
Galaxy API
Testing
Getting help
Indices and tables

Next topic
API documentation for interacti

This Page
Show Source

Blend

About

Blend is a Python (2.6 or higher) library for interacting with BioCloudCentral.org, CloudMan, and Galaxy's API. Conceptually, it makes it possible to script and automate the process of cloud infrastrucutre provisioning and scaling, as well as running of analyses within Galaxy. In reality, it makes it possible to do things like this:

- Create a CloudMan compute cluster, via an API and directly from your local machine:

```
from blend.cloudman.launch import CloudManLaunch
cml = CloudManLaunch('<your cloud access key>', '<your cloud secret key')
cml.launch('Blend CloudMan', 'ami-<ID>', 'm1.small', 'password')
cml.get_status()
```

- Manipulate your CloudMan instance and react to the current needs:

```
from blend.cloudman import CloudMan
cm = CloudMan("instance IP", "password")
cm.initialize(type="Galaxy")
cm.add_nodes(3)
cluster_status = cm.get_status()
cm.remove_nodes(2)
```

- Interact with Galaxy via a straighforward API:

```
from blend.galaxy import GalaxyInstance
gi = GalaxyInstance('<Galaxy IP>', key='your API key')
libs = gi.libraries.get_libraries()
gi.workflows.show_workflow('workflow ID')
gi.workflows.run_workflow('workflow ID', input_dataset_map)
```

Note

Although this library allows you to blend these three services into a cohesive unit, the library itself can be used with

CREATING A NATIONAL GENOMICS VIRTUAL LABORATORY

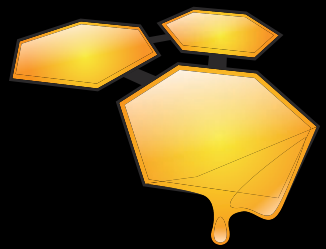
Univ. of Queensland and Victorian Life Sciences Compute Initiative (VLSCI)

Australian National Research Cloud

Provide computational infrastructure to support research

Compute and Storage

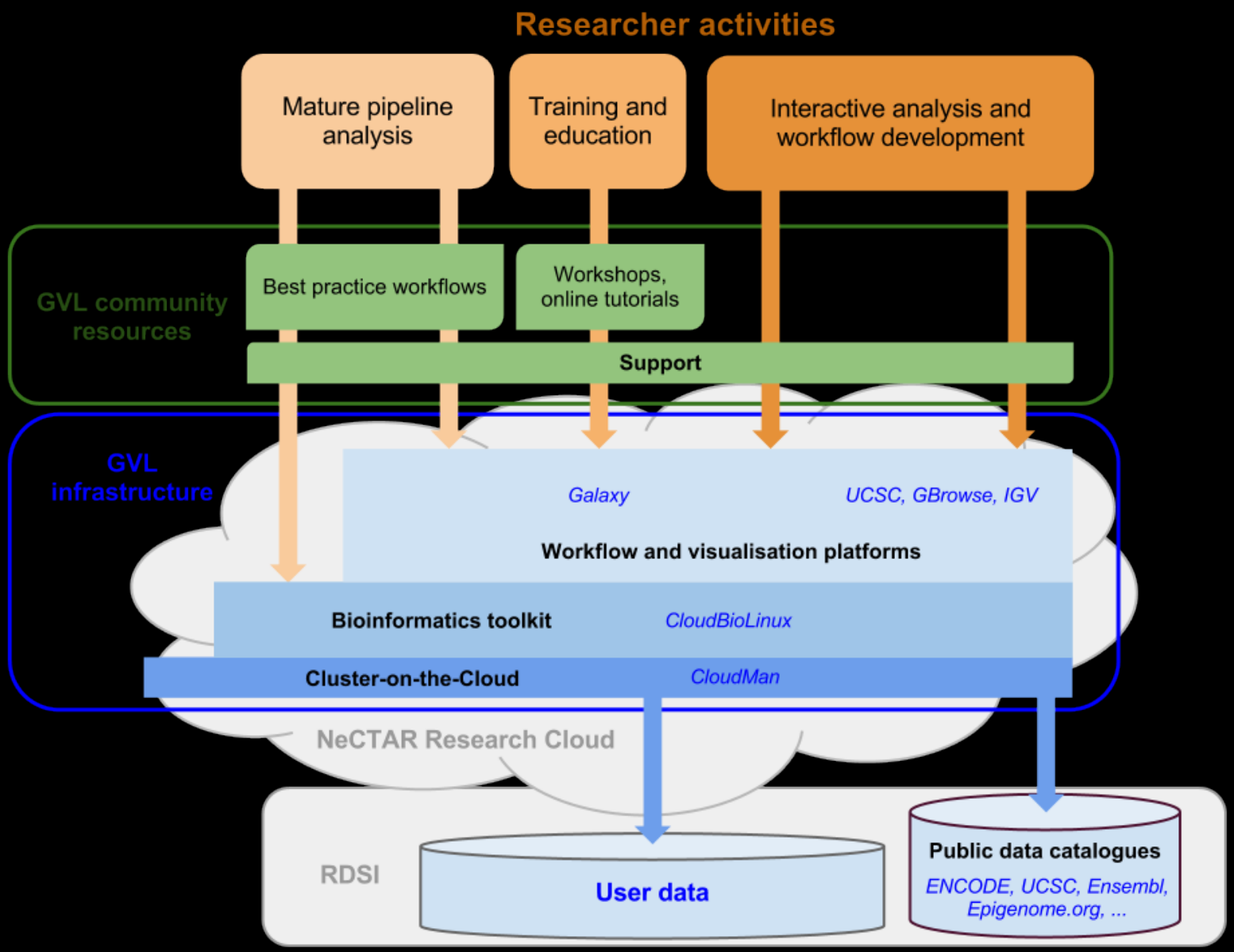
(~25,000 cores + multiple PBs)



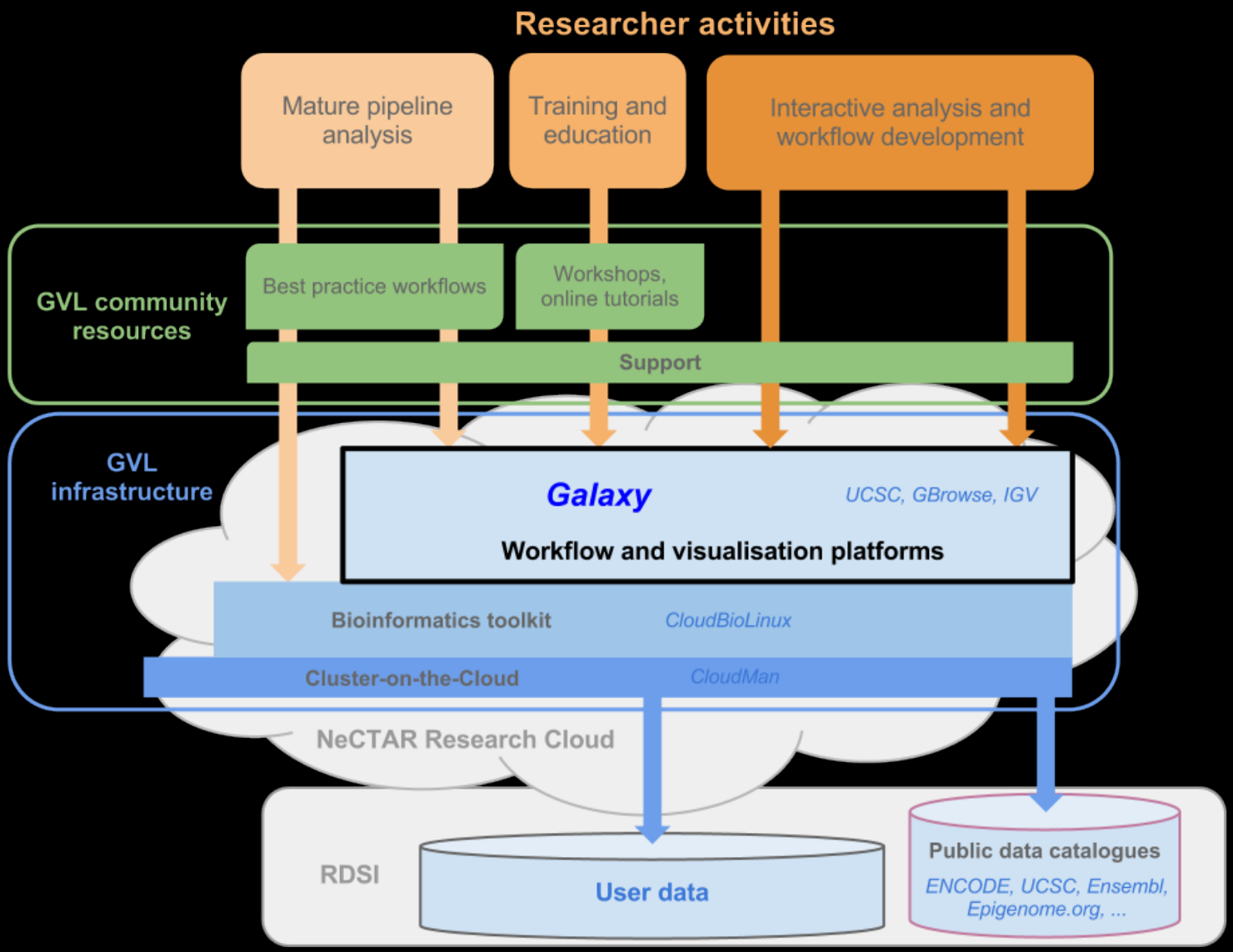
nectar

shell vs. IDE

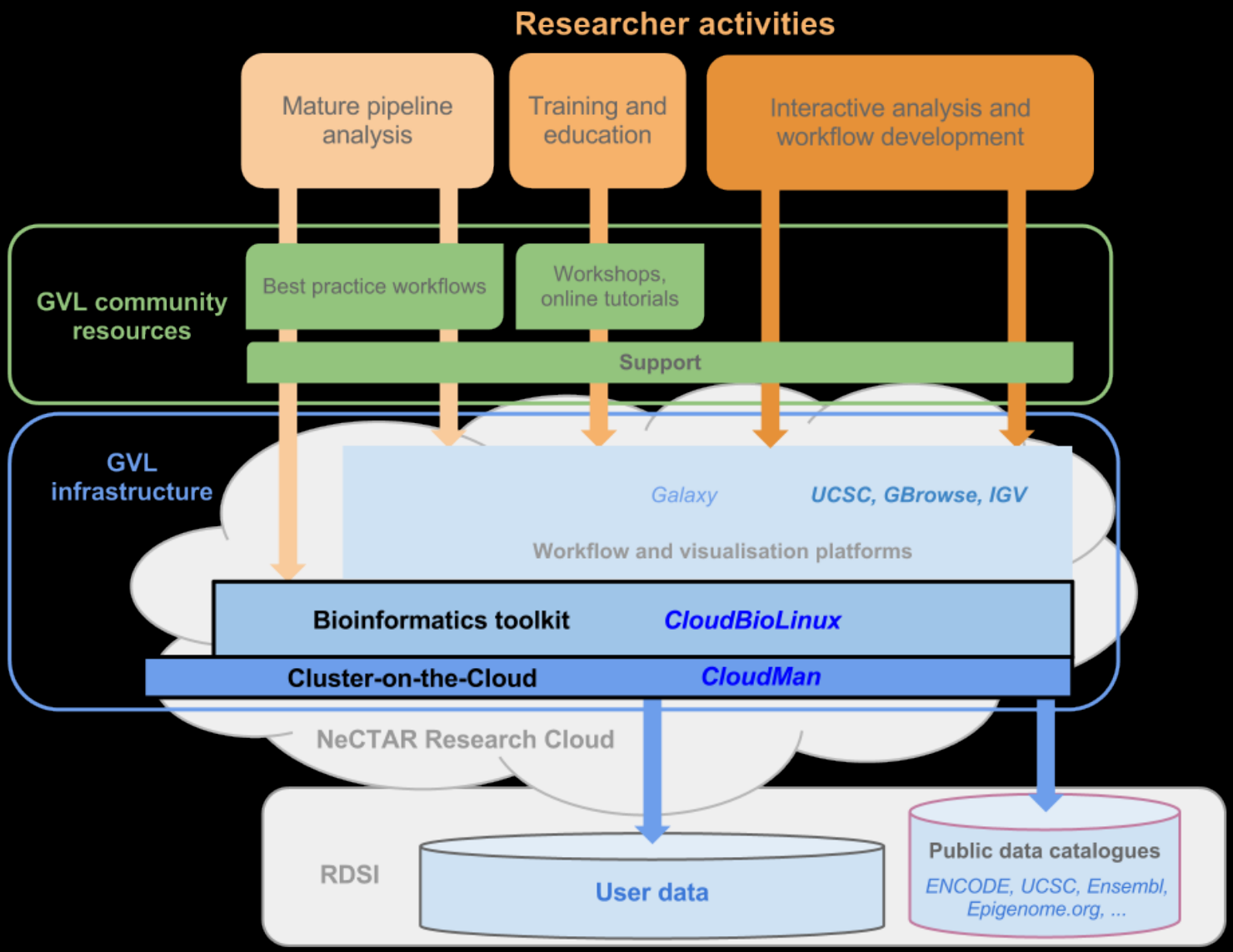
Genomics Virtual Lab



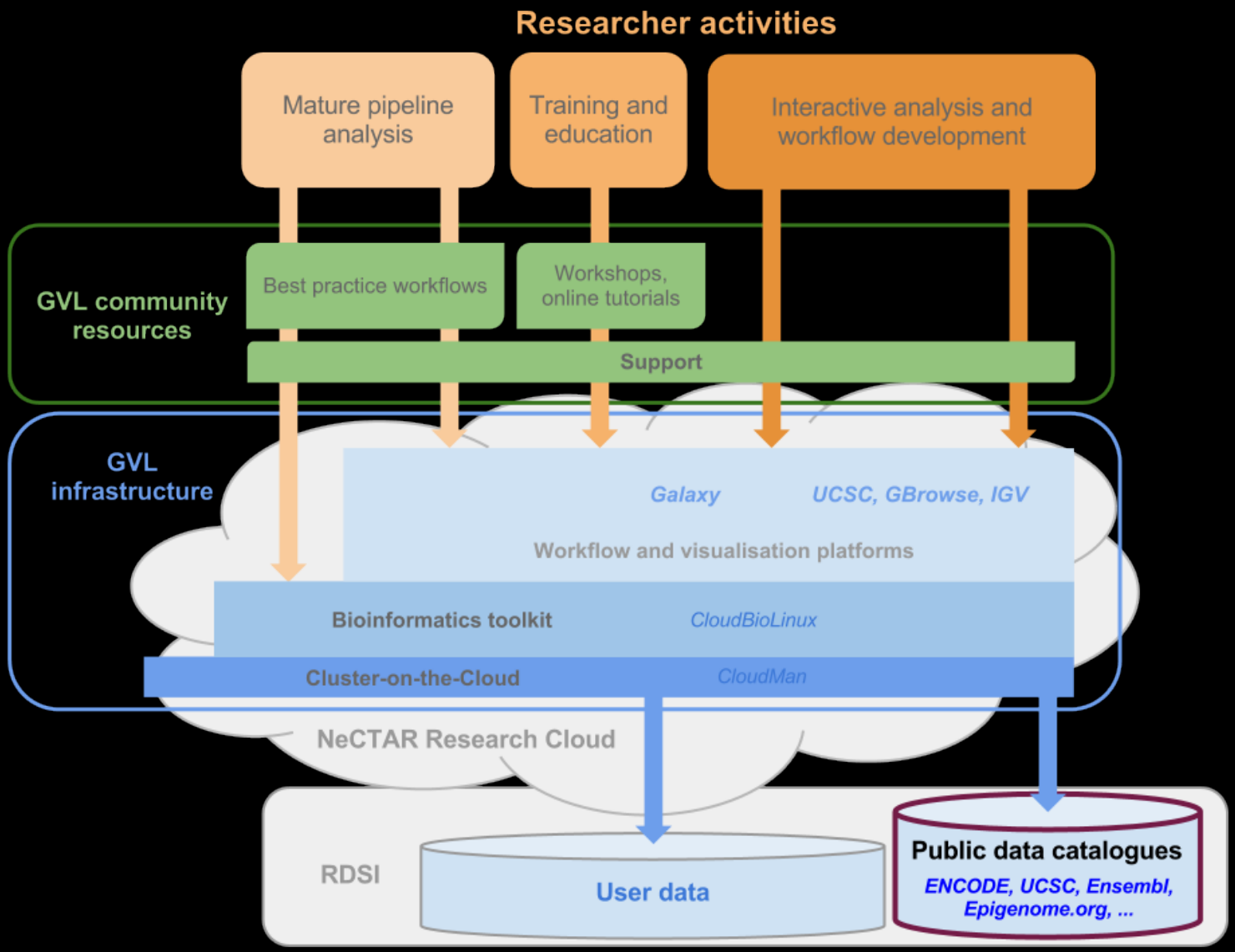
Workflow platforms

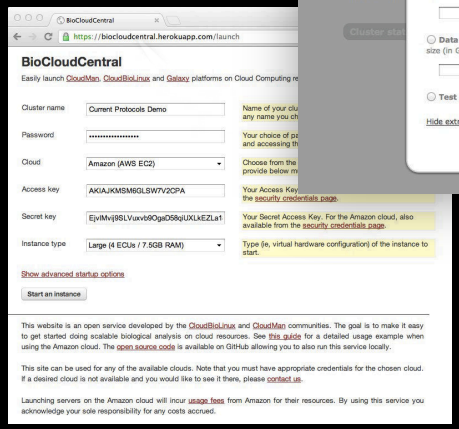


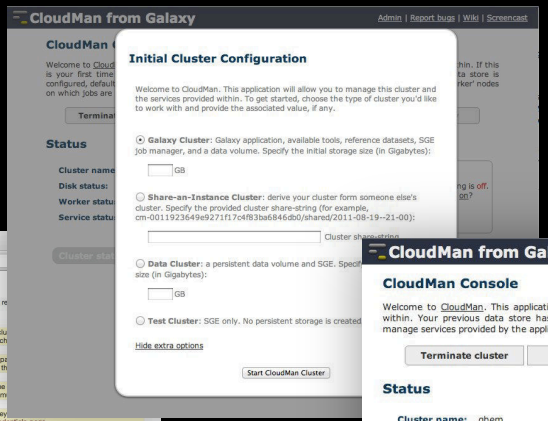
Cluster-on-the-cloud

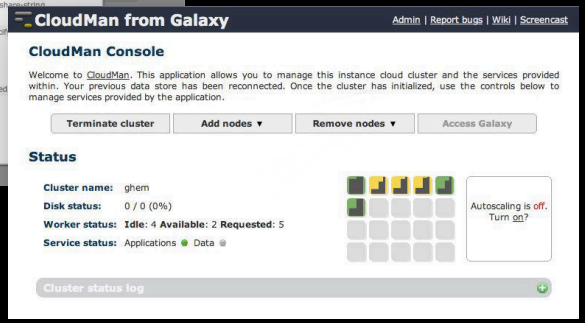


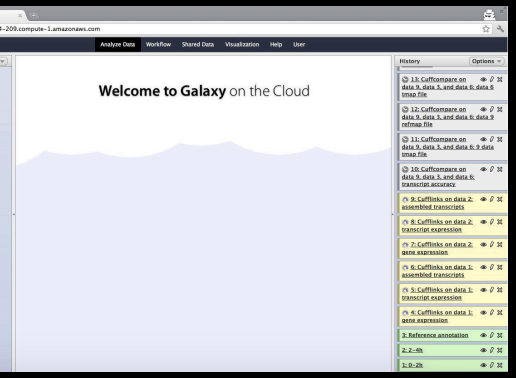
Data catalogs

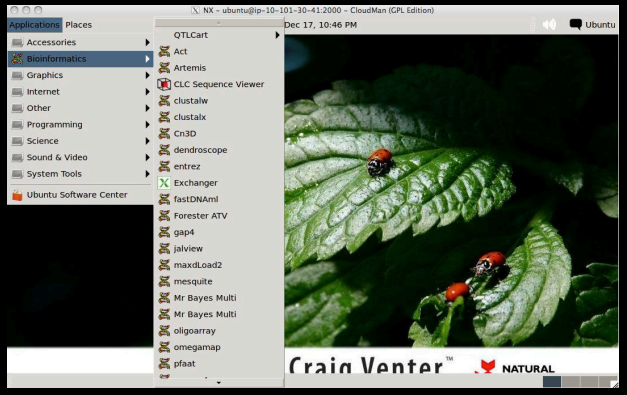


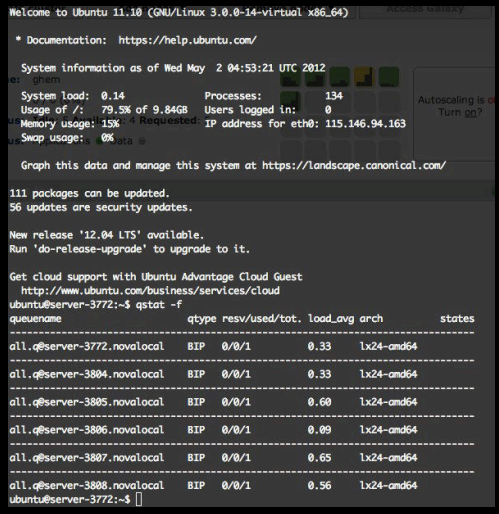

BioCloudCentral
https://biocloudcentral.herokuapp.com/launch
BioCloudCentral
Easily launch CloudMan, CloudBioLinux and Galaxy platforms on Cloud Computing re
Cluster name
Current Protocols Demo
Password
Cloud
Amazon (AWS EC2)
Access key
AKIAJKMSM6QLSW7V2CPA
Secret key
EjvIMvj9SLVuxvb9OgaDS8qIUXLkEZLa1
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.


CloudMan from Galaxy
Admin | Report bugs | Wiki | Screencast
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-0011923649e9271f17c4f83ba6846db0/shared/2011-08-19-21-00):
Cluster share-string
Data Cluster: a persistent data volume and SGE. Specify the provided size (in Gigabytes):
GB
Test Cluster: SGE only. No persistent storage is created
Hide extra options
Start CloudMan Cluster
Status

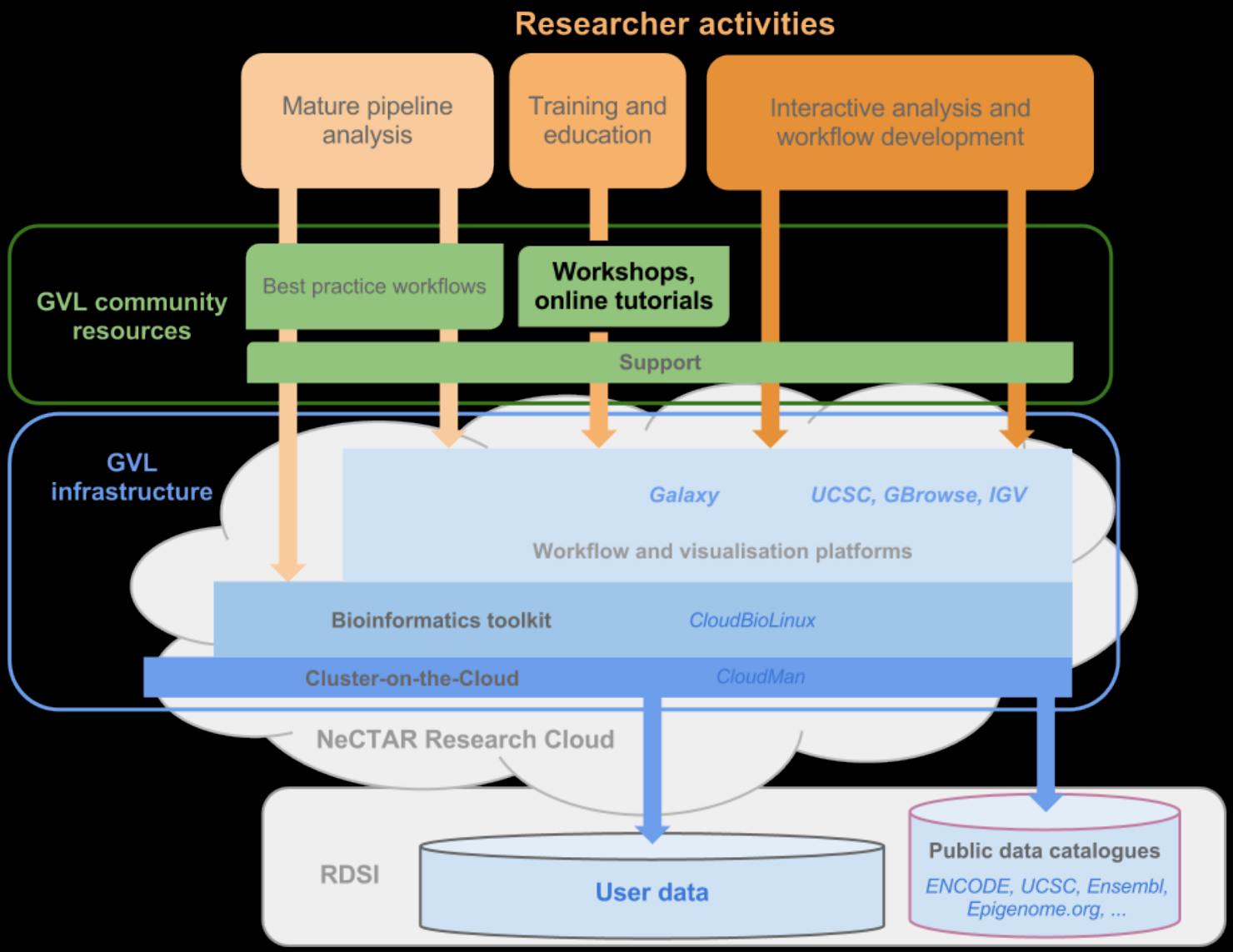

CloudMan from Galaxy
Admin | Report bugs | Wiki | Screencast
CloudMan Console
Welcome to CloudMan. This application allows you to manage this instance cloud cluster and the services provided within. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below 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?
Cluster status log


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Welcome to Galaxy on the Cloud

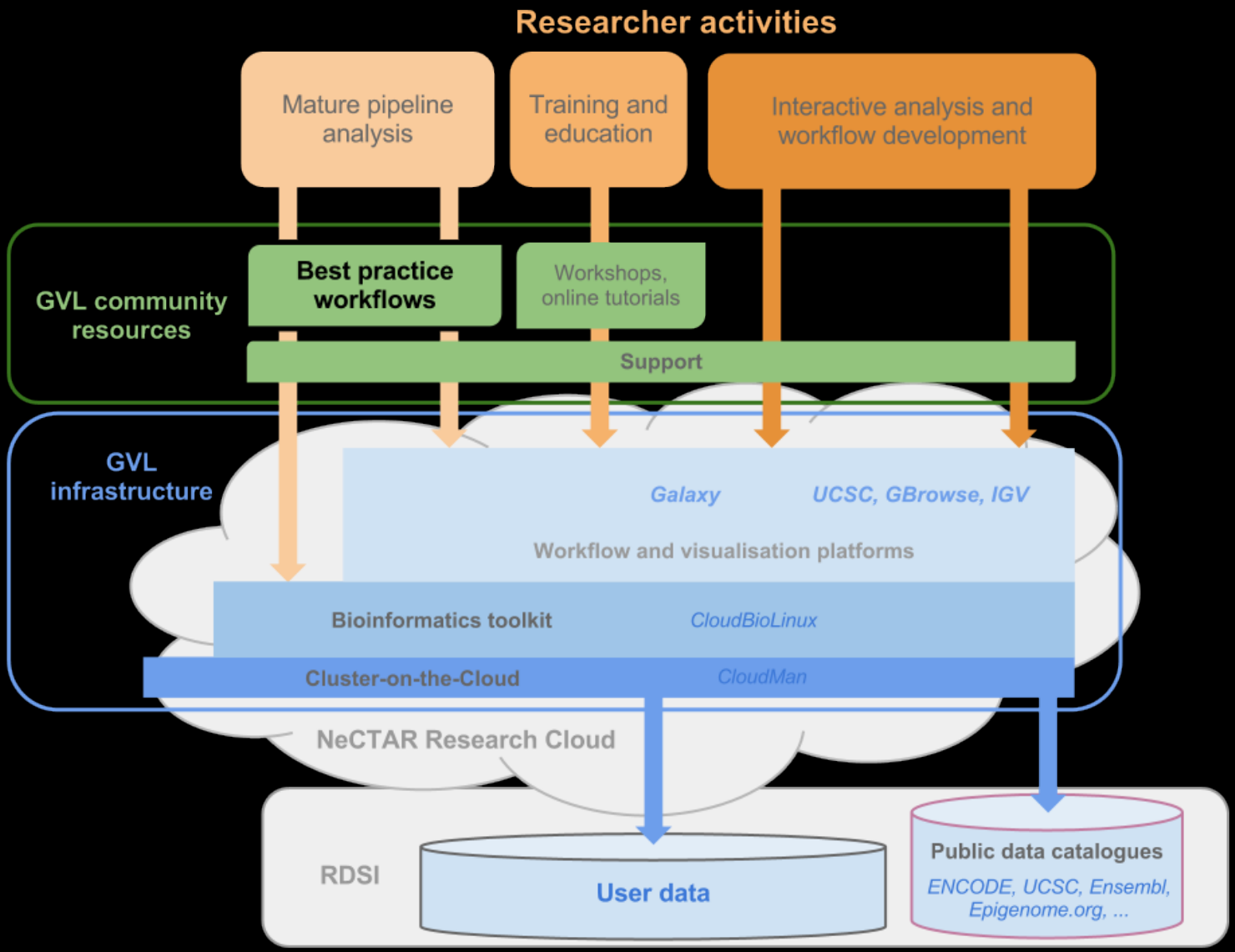

Applications
Places
Dec 17, 10:46 PM
Ubuntu
Accessories
Bioinformatics
Graphics
Internet
Other
Programming
Science
Sound & Video
System Tools
Ubuntu Software Center
QTLCart
Act
Artemis
CLC Sequence Viewer
clustalw
clustalx
Cn3D
dendroscope
entrez
Exchanger
fastDNAml
Forester ATV
gap4
jalview
maxdLoad2
mesquite
Mr Bayes Multi
Mr Bayes Multi
oligoarray
omegamap
pfaat
Craig Venter
NATURAL


Welcome to Ubuntu 11.10 (GNU/Linux 3.0.0-14-virtual x86_64)
* Documentation: https://help.ubuntu.com/
System information as of Wed May 2 04:53:21 UTC 2012
System load: 0.14 Processes: 134
Usage of /: 79.5% of 9.84GB Users logged in: 0
Memory usage: 15% IP address for eth0: 115.146.94.163
Swap usage: 0%
Graph this data and manage this system at https://landscape.canonical.com/
111 packages can be updated.
56 updates are security updates.
New release '12.04 LTS' available.
Run 'do-release-upgrade' to upgrade to it.
Get cloud support with Ubuntu Advantage Cloud Guest
http://www.ubuntu.com/business/services/cloud
ubuntu@server-3772:~$ qstat -f
queuename qtype resv/used/tot. load_avg arch states
all.q@server-3772.novalocal BIP 0/0/1 0.33 lx24-amd64
all.q@server-3804.novalocal BIP 0/0/1 0.33 lx24-amd64
all.q@server-3805.novalocal BIP 0/0/1 0.60 lx24-amd64
all.q@server-3806.novalocal BIP 0/0/1 0.09 lx24-amd64
all.q@server-3807.novalocal BIP 0/0/1 0.65 lx24-amd64
all.q@server-3808.novalocal BIP 0/0/1 0.56 lx24-amd64
ubuntu@server-3772:~$

Tutorials and workshops



Exemplar best practice workflows



Exemplar workflows

- Variant calling:
  - GATK best-practice
  - microbial
  - cancer-optimised
- RNA-seq differential expression
- Fusion gene discovery from RNA-seq
- MicroRNA analysis
- De novo genome and transcriptome assembly
- Metagenomics
- ChIP-seq
- Variant annotation
- Pathway analysis
- Methylation

Scalable, reproducible, accessible

- New approaches for dealing with the data
  - Provide an illusion of *unlimited* data storage
  - New model for data accessibility
- Scale and replicate across multiple nodes around the country (and the world)
  - Integrate with public and other academic clouds
- Enable sharing and provide consistency
- Ensure ‘upgradeability’
  - While permitting customization

Acknowledgments

Univ. of Queensland

- *Xin-Yi Chua*
- *Michael Pheasant*

VLSCI

- *Enis Afgan*
- *Clare Sloggett*
- Andrew Lonie

Garvan Institute

- Derrick Lin
- *Warren Kaplan*

Galaxy Team