Rapidly Bringing Software to Biologists with Galaxy and Docker

John Chilton, Björn Grüning, Eric Rasche, the Galaxy Team

Slides @ https://bit.ly/biodata14gx
What is Docker?

“It runs as an isolated process in userspace on the host operating system, sharing the kernel with other containers. Thus, it enjoys the resource isolation and allocation benefits of VMs but is much more portable and efficient.”

Virtual Machines

Docker

https://docker.com/whatisdocker/
Why care about Docker?

Docker provides **reproducibility** of virtual machines - but with superior performance, easier development, and greater scalability.
Not just performance...

- A Dockerfile is a recipe, not a black box.
- More compact than VMs with rich CLI and API - makes it is easy to build pipelines spanning many Docker containers.
What is **Galaxy**?

Showing the Infravec history published by Dan Lawson @ usegalaxy.org
What is a Galaxy Tool?

- Declaratively describe UI and command line.
- “Job” executed either locally or on a cluster.
- Provide a consistent, high-level graphical interface to a wide variety of bioinformatics software.
The Galaxy ToolShed

2700 tools in 1200 repositories
Docker and Galaxy Tools

- Tools can be annotated with Docker image ids.
- This provides an even greater level of recomputability than the ToolShed alone.
- More secure tool execution.
- Potentially write tool wrappers with all of their dependencies bundled much more quickly.
Galaxy enables Docker over shared clusters
**Pulsar and Docker**

Galaxy’s **remote job runner**.

- Docker support for local execution, traditional clusters, or **Apache Mesos**.

- Staging jobs allows Pulsar to enforce security in ways Galaxy cannot when used with Docker - **prevents tools from reading or modifying files not related to the job**.

... **benefits of Galaxy job running without needing a Galaxy database or object store**.
Community-Driven Documentation

A couple of thorough community developed tutorials including introduction to Docker. Complete examples for STACKS, SMALT, OBItools, and LoRDEC.

More than tools...

A couple more ways Galaxy can leverage Docker for greater accessibility.
Galaxy has traditionally been more about... software for biologists... than programmers!

No programming or command lines needed but...
Golden age for learning bioinformatics programming?

IP[y]: IPython Interactive Computing

Bioconductor
Open Source Software for Bioinformatics

Teaching Lab Skills for Scientific Computing

DATA CARPENTRY
Making Data Science More Efficient

RStudio

GOBLET
Global Organisation for Bioinformatics Learning, Education & Training

knitr
Docker enables Galaxy Interactive Environments Framework for spinning up secure, isolated environments that can interact with Galaxy’s history.

First integration with IPython and RStudio is on the way.
Setup: A Cufflinks-based RNA-seq workflow
Welcome to the interactive Galaxy IPython Notebook.

You can access your data via the dataset number. For example, handle = get(42). To save data, write your data to a file, and then call put('filename.txt'). The dataset will then be available in your galaxy history. Notebooks can be saved to Galaxy by clicking the large green button at the top right of the IPython interface.

More help and informations can be found on the project [website](#).
In [75]:  # Load IPython magic for R integration.
    %load_ext rpy2.ipython
    %R library(cummeRbund)

    The rpy2.ipython extension is already loaded. To reload it, use:
    %reload_ext rpy2.ipython

Out[75]: array(['mgcv', 'nlme', 'cummeRbund', 'Gviz', 'grid', 'rtracklayer',
               'GenomicRanges', 'GenomeInfoDb', 'IRanges', 'S4Vectors', 'stats4',
               'fastcluster', 'reshape', 'ggplot2', 'RSQML', 'DBI',
               'BiocGenerics', 'parallel', 'tools', 'stats', 'graphics',
               'grDevices', 'utils', 'datasets', 'methods', 'base'],
              dtype='|S13')

In [72]: get(72, True)  # Download Galaxy history id number 72 as file in current directory with name '72'
Out[72]: '/import/72'

In [77]: %R cuff <- readCufflinks(dbFile='72')  # Load history downloaded item (cuffdiff for cummeRbund)
Out[77]: <RS4 - Python:8x26781568 / R:8x1dfb8f58>
Scale for 'x' is already present. Adding another scale for 'x', which will replace the existing scale. geom_smooth: method="auto" and size of largest group is >=1000, so using gam with formula: y ~ s(x, bs = "cs"). Use 'method = x' to change the smoothing method.
```r
# Find all differentially expressed genes at a given alpha
%R sig <- getSig(cuff, alpha=0.01, level='genes')
%R sigGenes <- getGenes(cuff,sig)
%R print(length(sig))

Getting gene information:
  FPKM
  Differential Expression Data
  Annotation Data
  Replicate FPKM
  Counts

Getting isoforms information:
  FPKM
  Differential Expression Data
  Annotation Data
  Replicate FPKM
  Counts

Getting CDS information:
  FPKM
  Differential Expression Data
  Annotation Data
  Replicate FPKM
  Counts

Getting TSS information:
  FPKM
  Differential Expression Data
  Annotation Data
  Replicate FPKM
  Counts

Getting promoter information:
  distData

Getting splicing information:
  distData

Getting relCDS information:
  distData

[1] 783
```
Using tracking id, sample name as id variables
No id variables; using all as measure variables
In [83]: # Describe gene density.
   %R dens <- csDensity(genes(cuff))
   %R print(dens)

In [84]: # Now pull that dens data structure out of R and make available as a numpy structure in Python
   %R pull dens

In []: # Iterate through it in Python and create a file 'gene_fpkm.tsv' and upload it to Galaxy history.
   with open('gene_fpkm.tsv', "w") as f:
       for val in zip(dens[0]['gene_id'], dens[0]['fpkm']):
           f.write("\t".join(map(str, val)) + "\n")
   put("gene_fpkm.tsv")
Save the current notebook in Galaxy
Simple Galaxy is Easy

% hg clone bitbucket.org/galaxy/galaxy-dist
% sh run.sh

Galaxy is up and running, but...

- Tool shed not configured
- No FTP server
- No external job manager
- Default SQLite database is easy to break
Galaxy on Docker Hub

Automated Build Repository
bgruening / galaxy-stable
No description set

# Galaxy - Stable
#
# VERSION 0.2.0

FROM ubuntu:14.04

MAINTAINER Bjoern A. Gruning, bjoern.gruening@gmail.com

# * Make sure the package repository is up to date
# * Install all requirements that are recommend by the Galaxy project
# * Enable the @matefoo magic
# Web server infrastructure matching usegalaxy.org - supervisor, uwsgi, and nginx.

ENV DEBIAN_FRONTEND noninteractive
Scalable Galaxy with **One Command**

% docker run -d -p 8080:80 bgruening/galaxy-stable

- Galaxy split into multiple processes, uwsgi configured for optimized web request processing.
- Nginx proxy for optimized data uploads.
- Pre-configured ToolShed integration.
- Postgres database.
- ProFTP server.
- SLURM job management.

*Same technology stack that allows usegalaxy.org to scale to thousands of users.*
A Step Further - Preconfigured Bundles

- Deep sequencing
- Exome sequencing
- Allele-specific mapping
- Cheminformatics
- Proteomics
Lots of Links

Docker
docker.com/
VMs Considered Harmful
ivory.idyll.org/blog/vms-considered-harmful.html
Bioinformatics Platforms Leveraging Docker
galaxyproject.org/
www.arvados.org/
www.sbgenomics.com/
bcbio-nextgen.readthedocs.org/
www.iplantcollaborative.org/
agaveapi.co/
www.nextflow.io/
IPython
ipython.org/ jupyter.org/
Example Notebook
Galaxy IPython Demo
Galaxy
usegalaxy.org/
getgalaxy.org/
Pulsar
pulsar.readthedocs.org

Galaxy and Docker
bit.ly/dockergx
github.com/bgruening/docker-galaxy-stable
github.com/bgruening/docker-recipes
Tutorial on Github by Aaron Petkau
Tutorial from GUGGO
Cluster Management
adaptivecomputing.com/products/open-source/torque/
ibm.com/systems/platformcomputing/
mesos.apache.org/
slurm.schedmd.com/
research.cs.wisc.edu/htcondor/
gridscheduler.sourceforge.net/
Tuxedo Suite
ccb.jhu.edu/software/tophat
cufflinks.cbcb.umd.edu/gff.html
compbio.mit.edu/cummeRbund/
R
r-project.org/ rstudio.com/
bioconductor.org/
yihui.name/knitr/

Find everything at bit.ly/biodata14gx
Thanks!

The Galaxy Team

Eric Rasche

Björn Grüning
Future Plans

● Homebrew for easier, more broadly useful dependency management.
● Docker images for everything in the tool shed.
● Cloud Bursting
● IPython/Jupyter notebooks as first class citizens.
  ○ Enhanced provenance tracking, usable within workflows.
  ○ Cluster support.
● Easier tool development - deep github integration.
● More interactive environments - e.g. Web Apollo
The Notebook
Charts

Charts enables users to quickly visualize tabular data.

Watch Screencast
Planemo

Command-line tools to aid development.

- Test tools quickly without worrying about configuration files.
- Check tools for common bugs and best practices.
- Optimized publishing to the ToolShed.
- Testbed for new developments such as Docker and Homebrew.

What are the main hurdles in the tool development process?

- Functional tests
- Documentation
- XML syntax
- Toolshed
- Tool data (e.g., reference genomes)
- Other
Tool Development

Tool Citations
Started at BOSC 2014 Codefest

Embed DOIs into tools, Galaxy resolves these per tools or for a full analysis histories into exportable citation list.
ToolShed - A developers perspective

The vision

WITH

REPLACE

Khmer

This Tool Shed Tool is connected to the development repository at https://github.com/ged-lab/khmer.

We checked 18 minutes ago and there are updates AVAILABLE.

We tested 13 hours ago and the tool PASSED.
Enhancing Homebrew for Reproducibility

TODO
Build-Your-Own Galaxy Docker

TODO
Tool Forms

Previously Tool Forms had to reload entirely in response to user interactions. This limited the UI’s responsiveness and resulted in cumbersome tool configurations.

Watch Screencast

We redesigned our Tool Forms such that reloads became unnecessary. UI elements are now refreshed instantly upon user interactions.

Watch Screencast

Pulsar and Docker

Galaxy’s remote job runner.

- RESTful API or Message Queue.
- Pluggable file actions allowing **optimized data transfer** for a deployment and remote job staging.
- Run on traditional clusters like Galaxy, locally, or with Apache Mesos, Docker support with any of these deployment options.
- Galaxy dependency resolution, job customizations, metrics, etc...

... benefits of Galaxy job running without needing a Galaxy database or object store.
#BioData14 BINGO

When ever term is mentioned, you can fill card -- fill a line, call out “BINGO”

<table>
<thead>
<tr>
<th>B</th>
<th>I</th>
<th>N</th>
<th>G</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Hadoop</td>
<td>&lt;anyword&gt;-Seq</td>
<td>Global Alliance</td>
<td>Peta</td>
</tr>
<tr>
<td>IO</td>
<td>Docker</td>
<td>github address</td>
<td>The NHGRI graph on cost of sequencing</td>
<td>I’m the last person between you and [beer</td>
</tr>
<tr>
<td>D</td>
<td>distributed computing</td>
<td>MongoDB</td>
<td>DATA</td>
<td>ICGC</td>
</tr>
<tr>
<td>A</td>
<td>web services</td>
<td>Cloud</td>
<td>Google</td>
<td>scalable</td>
</tr>
<tr>
<td>TA</td>
<td>“tweet me”</td>
<td>Python Screenshot</td>
<td>wet lab experiment</td>
<td>Exa</td>
</tr>
</tbody>
</table>
Abstract Session

Software for Biologists: creating, disseminating, training, and funding such software