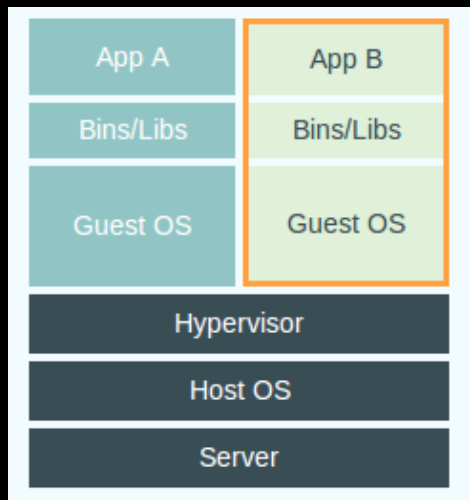


Rapidly Bringing Software to Biologists with Galaxy and Docker

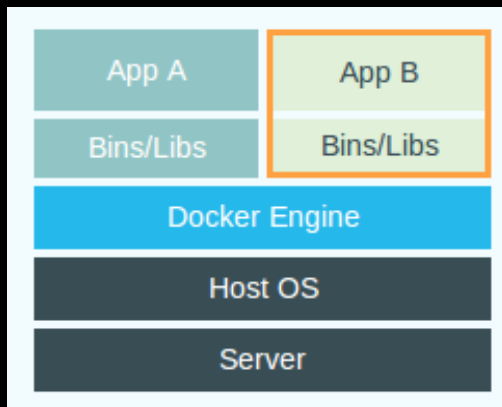
John Chilton, Björn Grüning, Eric Rasche,
the Galaxy Team

Slides @ <https://bit.ly/biodata14gx>

What is Docker?



Virtual Machines



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."

<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...



The screenshot shows a blog interface with a dark navigation bar containing links: misc, personal, python, science (highlighted in red), teaching, and testing. Below the navigation bar is the URL <http://ivory.idyll.org/blog/vms-considered-harmful.html> in red. The main content area has a white background with the title 'Virtual machines considered harmful for reproducibility' in a large, dark font. Below the title, it says 'In his paper, [Reproducible Research and Cloud Computing](#), Bill Howe asks:'. To the right of this text, it says 'Tue 06 November 2012', 'By [C. Titus Brown](#)', and 'In [science](#).'. At the bottom right, it says 'tags: [reproducibility should be obvious](#)'. Below the main text, there is a paragraph of text in italics: 'What happens if you do all your work on a virtual machine hosted in the cloud? When it came time to publish, you might make a snapshot of the VM, make it public, and cite it in your paper. Those who wish to'.

misc personal python **science** teaching testing

<http://ivory.idyll.org/blog/vms-considered-harmful.html>

Virtual machines considered harmful for reproducibility

In his paper, [Reproducible Research and Cloud Computing](#), Bill Howe asks:

What happens if you do all your work on a virtual machine hosted in the cloud? When it came time to publish, you might make a snapshot of the VM, make it public, and cite it in your paper. Those who wish to

Tue 06 November 2012
By [C. Titus Brown](#)
In [science](#).
tags: [reproducibility should be obvious](#)

- 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?

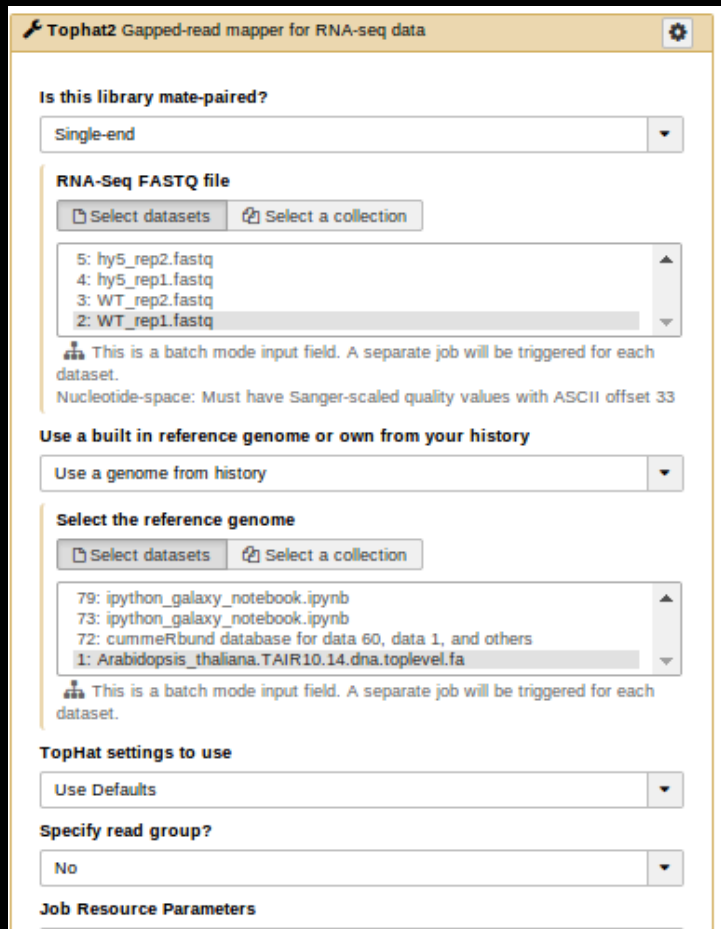
The screenshot displays the Galaxy web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Cloud, Help, and User. The left sidebar contains a 'Tools' panel with a search bar and a list of tool categories such as Get Data, Send Data, Lift-Over, Text Manipulation, Convert Formats, FASTA manipulation, Filter and Sort, Join, Subtract and Group, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Regional Variation, Multiple regression, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments, Metagenomic analyses, Genome Diversity, NGS TOOLBOX BETA, Phenotype Association, NGS: QC and manipulation, NGS: Mapping, NGS: SAM Tools, NGS: GATK Tools (beta), NGS: Peak Calling, and NGS: RNA-seq.

The main workspace shows the 'Unified Genotyper (version 0.0.6)' tool configuration. The 'Choose the source for the reference list:' section has a dropdown set to 'History'. The 'BAM files' section shows a single file: '56: (hidden) Map with BWA for Illumina on data 2, data 7, and data 3: mapped reads'. The 'Using reference file:' section has a dropdown set to '3: AgamP3 assembly'. The 'Binding for reference-ordered datas' section has a dropdown set to 'D:--dbsnp <dbsnp>'. The 'Genotype likelihoods calculation model to employ:' section has a dropdown set to 'BOTH'. The 'The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called:' section has a text input set to '30.0'. The 'The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be emitted (and filtered if less than the calling threshold):' section has a text input set to '30.0'. The 'Basic or Advanced GATK options:' section has a dropdown set to 'Basic'. The 'Basic or Advanced Analysis options:' section has a dropdown set to 'Basic'. An 'Execute' button is at the bottom of the configuration panel.

The right sidebar shows the 'History' panel. It lists several jobs, including 'Infravec: imported from Dan Lawson' (76.3 GB) and '89: AgamP3.7 GTF for mat'. The job '69: Unified Genotyper on data 3 and data 56 (log)' is highlighted, showing 340 lines and a format of 'txt, database: ?'. Below the job list, there is a section for 'Picked up JAVA_OPTIONS:' with a list of system paths and options.

Showing the Infravec history published by Dan Lawson @ usegalaxy.org

What is a Galaxy Tool?



The screenshot shows the TopHat2 tool interface in Galaxy. The title bar reads 'TopHat2 Gapped-read mapper for RNA-seq data'. The interface is organized into several sections: 1. 'Is this library mate-paired?' with a dropdown menu set to 'Single-end'. 2. 'RNA-Seq FASTQ file' section containing two buttons: 'Select datasets' and 'Select a collection'. Below these is a list of four datasets: '5: hy5_rep2.fastq', '4: hy5_rep1.fastq', '3: WT_rep2.fastq', and '2: WT_rep1.fastq', with the second dataset selected. A note below the list states: 'This is a batch mode input field. A separate job will be triggered for each dataset. Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33'. 3. 'Use a built in reference genome or own from your history' section with a dropdown menu set to 'Use a genome from history'. 4. 'Select the reference genome' section with 'Select datasets' and 'Select a collection' buttons. Below is a list of four items: '79: ipython_galaxy_notebook.ipynb', '73: ipython_galaxy_notebook.ipynb', '72: cummeRbund database for data 60, data 1, and others', and '1: Arabidopsis_thaliana.TAIR10.14.dna.toplevel.fa', with the last item selected. A similar note about batch mode is present. 5. 'TopHat settings to use' section with a dropdown menu set to 'Use Defaults'. 6. 'Specify read group?' section with a dropdown menu set to 'No'. 7. 'Job Resource Parameters' section at the bottom.

TopHat2 Gapped-read mapper for RNA-seq data

Is this library mate-paired?

Single-end

RNA-Seq FASTQ file

Select datasets Select a collection

5: hy5_rep2.fastq
4: hy5_rep1.fastq
3: WT_rep2.fastq
2: WT_rep1.fastq

This is a batch mode input field. A separate job will be triggered for each dataset.
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Use a built in reference genome or own from your history

Use a genome from history

Select the reference genome

Select datasets Select a collection

79: ipython_galaxy_notebook.ipynb
73: ipython_galaxy_notebook.ipynb
72: cummeRbund database for data 60, data 1, and others
1: Arabidopsis_thaliana.TAIR10.14.dna.toplevel.fa

This is a batch mode input field. A separate job will be triggered for each dataset.

TopHat settings to use

Use Defaults

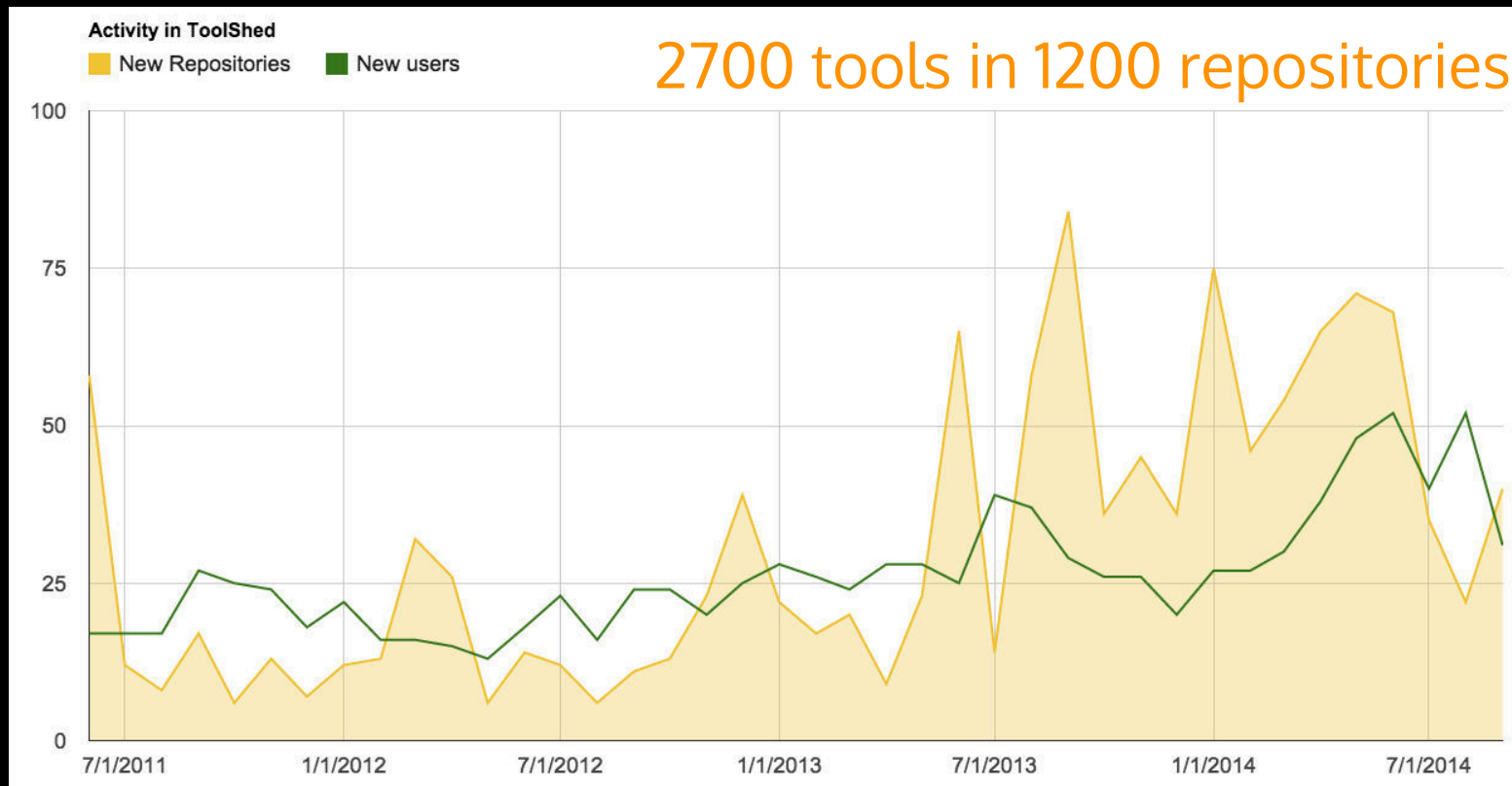
Specify read group?

No

Job Resource Parameters

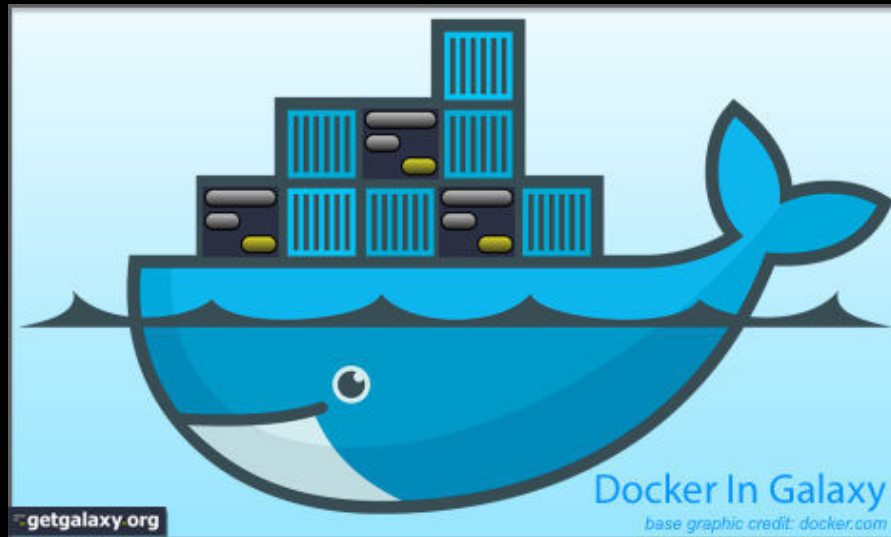
- Declaratively describe UI and command line.
- "Job" executed either locally or on a cluster.
- Provide a **consistent, high-level graphical interface** to a wide variety of **bioinformatics software**.

The Galaxy ToolShed



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

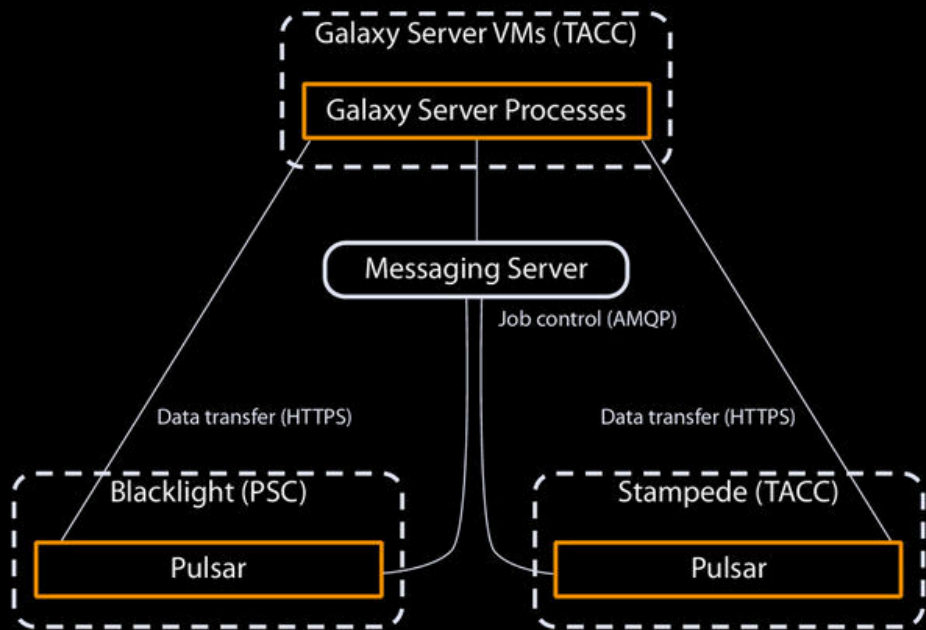


Pulsar and Docker


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.

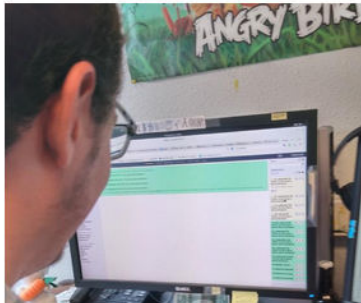
... benefits of Galaxy job running without needing a Galaxy database or object store.



Community-Driven Documentation



#GUGGO w/ @cmonjeaud :
@galaxyproject @docker STACKS image
test to analyse RADseq : validated!
@jcachen #usegalaxy



First GenQuest

Article

As we test several Docker

Galaxy et Docker

- Docker installation :
 - We here use a
 - Install Kernel 3.14 from wheezy-backports

```
#Add the following line to your /etc/apt/sources.list  
  
deb http://http.debian.net/debian wheezy-backports main  
  
#then install the linux-image-amd64 package (note the use of -t wheezy-backports)  
sudo apt-get update  
sudo apt-get install -t wheezy-backports linux-image-amd64  
#Install Docker using the get.docker.com script:  
sudo apt-get install curl  
curl -sSL https://get.docker.com/ | sh
```

apetkau / galaxy-hackathon-2014

19 commits 1 branch

branch: master galaxy-hackathon-2014 / +

Update README.md

apetkau authored on Jun 29

smalt	Update README.md
README.md	spelling
catDocker.xml	More details on readme + catDocker.xml

README.md

Integrating Docker-based tools within Galaxy

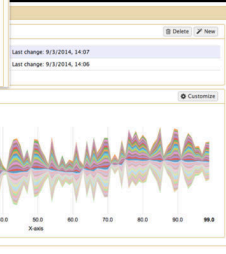
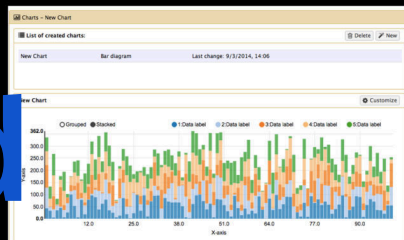
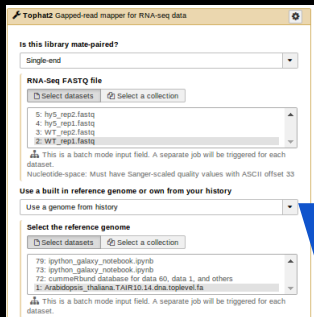
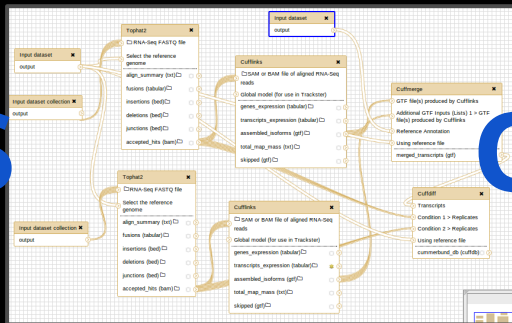


A couple of thorough community developed tutorials including introduction to Docker. Complete examples for STACKS, SMALT, OBIttools, and LoRDEC.

<http://bit.ly/dockergx>

More than tools...

A couple more ways Galaxy can leverage Docker for greater **accessibility**.



workflows



No programming or command lines needed **but...**

Golden age for learning bioinformatics programming?

IP[y]: IPython
Interactive Computing



TEACHING LAB SKILLS FOR
SCIENTIFIC COMPUTING

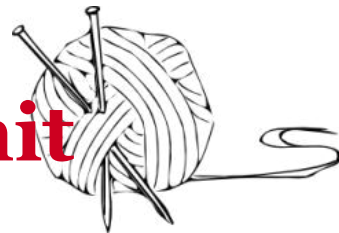


DATA CARPENTRY

MAKING DATA SCIENCE MORE EFFICIENT



knitr
r



GOBLET

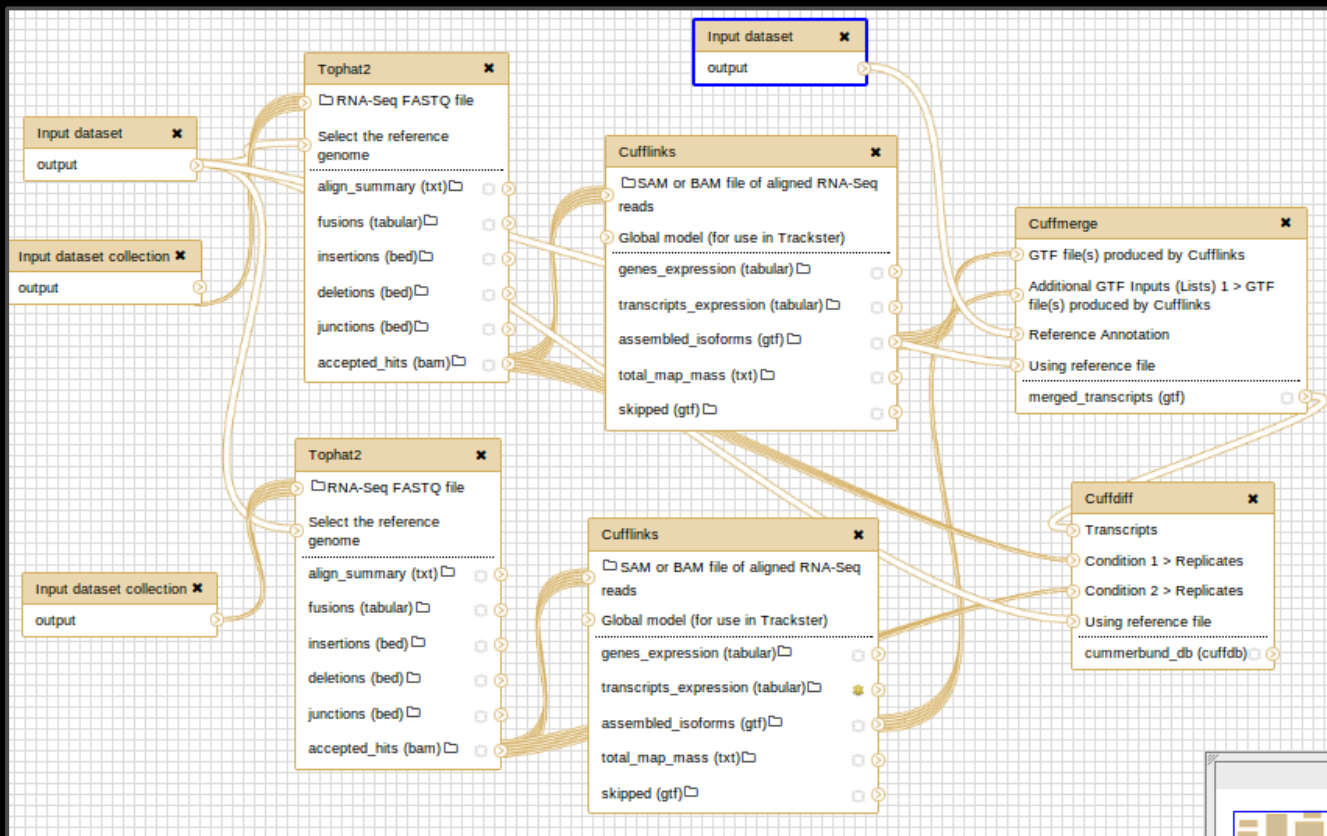
Global Organisation for Bioinformatics
Learning, Education & Training

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



t:8080



Analyze Data

Workflow

Shared Data

Visualization

Admin

Help

User



Using 1.6 GT



IP[y]: Notebook ipython_galaxy_notebook (autosaved)

File

Edit

View

Insert

Cell

Kernel

Help



Heading 1

Cell Toolbar:

None



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 [1]:

History



search datasets



RNAseqTry2

61 shown, 11 [deleted](#)

857.9 MB



72: cummeRbund database for data 60, data 1, and others



897,201 lines

format: **cuffdb**, database: ?

[23:00:48] Loading reference annotation and sequence.

[23:00:52] Inspecting maps and determining fragment length distributions.

[23:01:39] Modeling fragment count overdispersion.

[23:03:12] Modeling fragment count overdispersion.

> Map Properties:

> Norm



bin Visualize in IPython

60: Cuffmerge on data 1, data 28, and others:



```
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', 'reshape2', 'ggplot2', 'RSQLite', '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:0x26781560 / R:0x1dfb8f50>
```

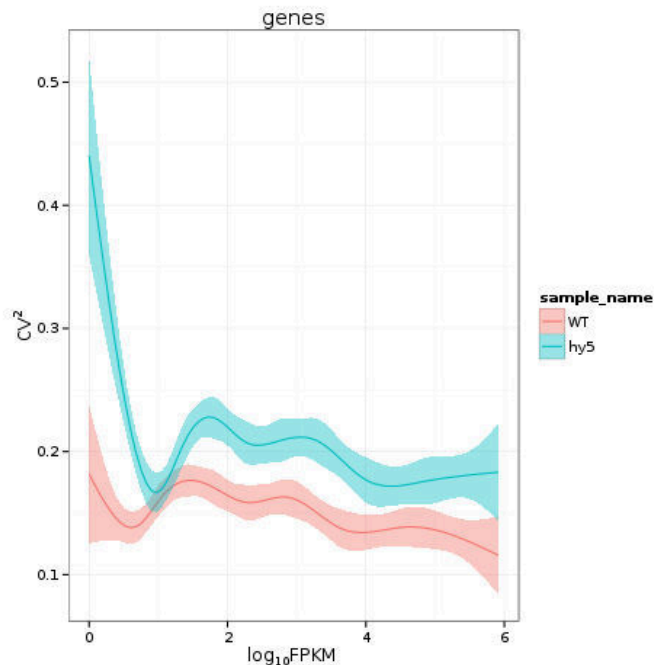
```
In [77]: %R cuff <- readCufflinks(dbFile='72') # Load history downloaded item (cuffdiff for cummeRbund)
```

```
Out[77]: <RS4 - Python:0x26781560 / R:0x1dfb8f50>
```

```
In [78]: %R print(fpkmSCVPlot(genes(cuff))) # Demonstrate plotting with R
```

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: $y \sim s(x, bs = "cs")$. Use 'method = x' to change the smoothing method.



```
In [80]: # 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 FPKMs
- Counts

Getting isoforms information:

- FPKM
- Differential Expression Data
- Annotation Data
- Replicate FPKMs
- Counts

Getting CDS information:

- FPKM
- Differential Expression Data
- Annotation Data
- Replicate FPKMs
- Counts

Getting TSS information:

- FPKM
- Differential Expression Data
- Annotation Data
- Replicate FPKMs
- Counts

Getting promoter information:

- distData

Getting splicing information:

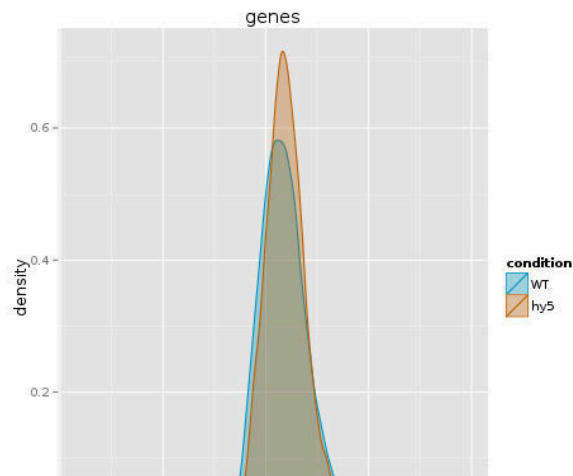
- distData

Getting relCDS information:

- distData

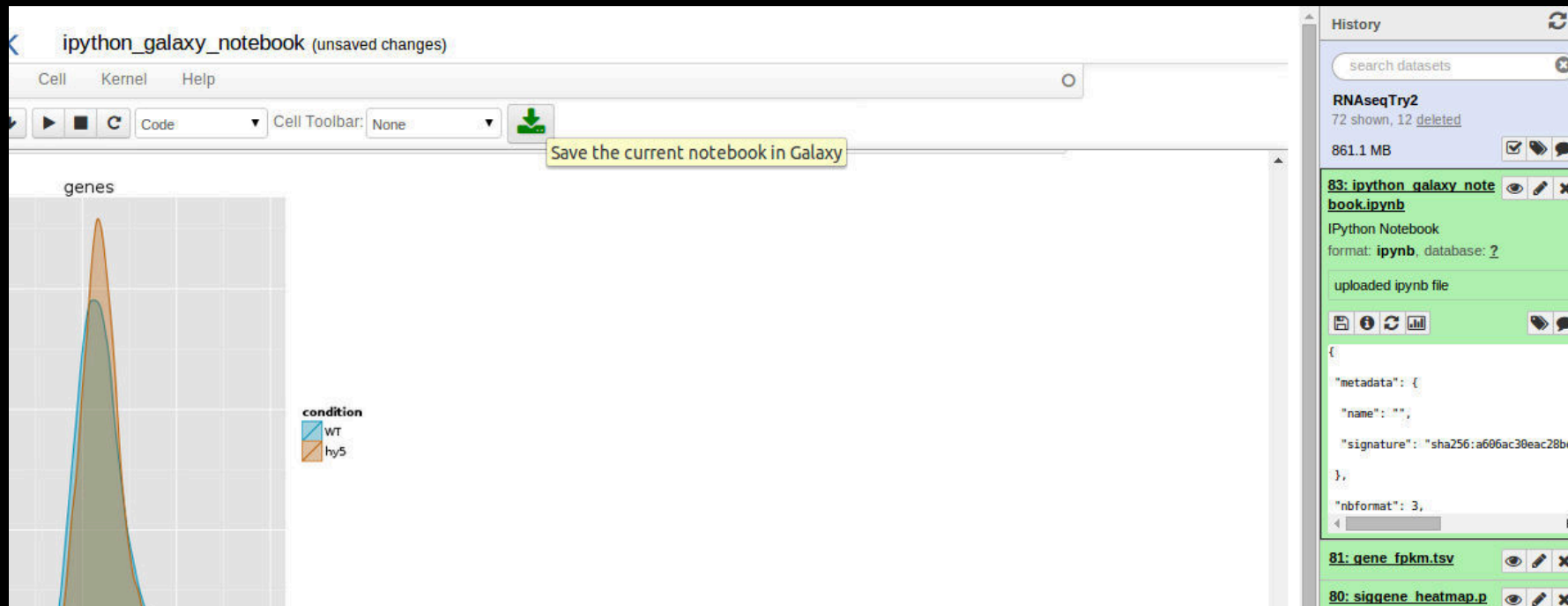
[1] 783


```
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  
%Rpull 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")
```




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

 **docker**

[What is Docker?](#) [Use Cases](#) [Try It!](#) [Browse](#) [Install & Docs](#) [Log In](#) [Sign Up](#)




 **AUTOMATED BUILD REPOSITORY**

bgruening / galaxy-stable Updated 3 days, 2 hours ago

Pull this repository

docker pull bgruening/galaxy-stable

No description set

 **1**  **0**  **180**

Information

Dockerfile

Build Details

Tags

```
# Galaxy - Stable
#
# VERSION      0.2.0


FROM ubuntu:14.04


MAINTAINER Björn A. Grüning, 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 @natefoo magic
# Web server infrastructure matching usegalaxy.org - supervisor, uwsgi, and nginx.

ENV DEBIAN_FRONTEND noninteractive
```

Properties

 2014-08-13 04:06:57

 **bgruening**

Build Details

[Source Project Page](#)

[Source Repository](#)

Scalable Galaxy with **One Command**

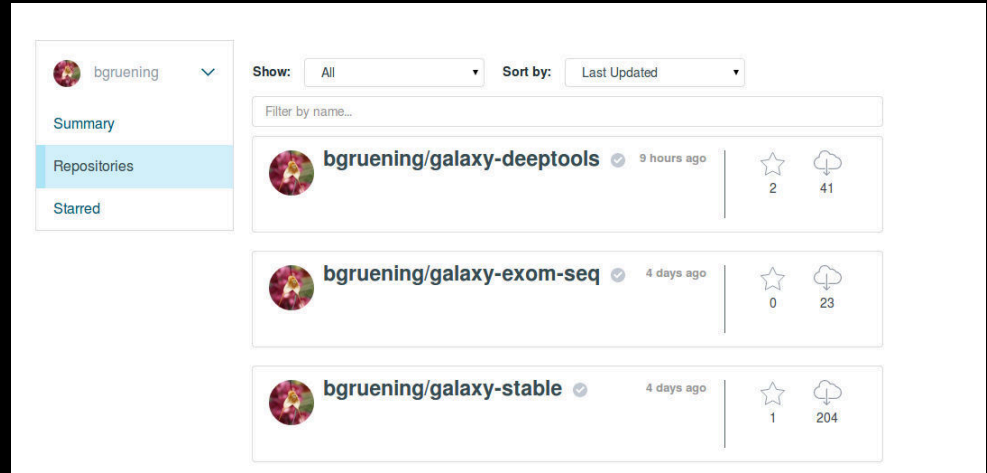
```
% docker run -d -p 8080:80 bgruening/galaxy-stable
```

- 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/

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

activecomputing.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



Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Marten Cech

John Chilton



Dave Clements

Nate Coraor

Carl Eberhard

Dorine Francheteau

Jeremy Goecks

Sam Guerler



Jen Jackson

Greg von Kuster

Ross Lazarus

Anton Nekrutenko

Nick Stoler

James Taylor

Eric Rasche



Björn Grüning

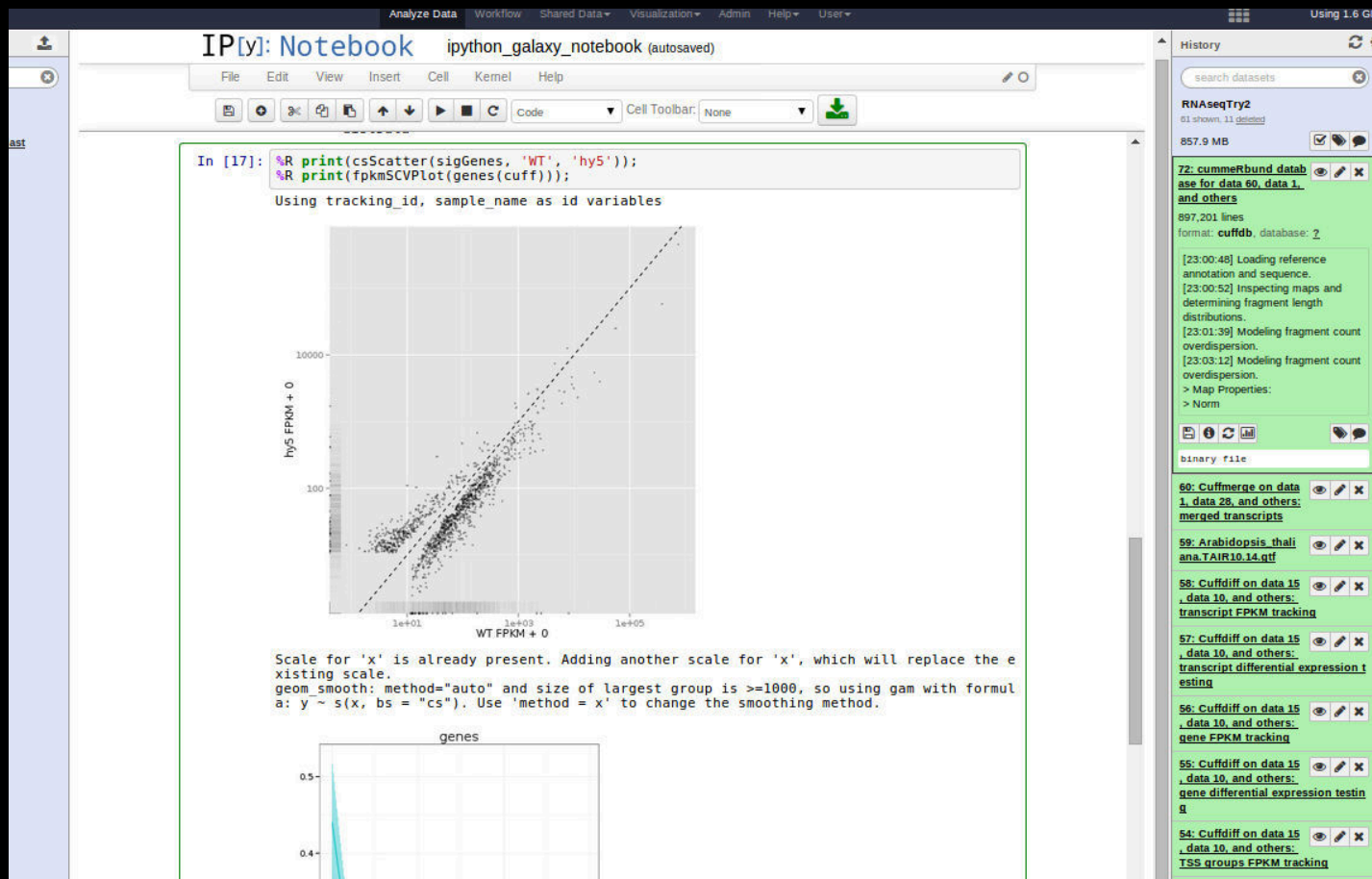


Future Plans

- 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

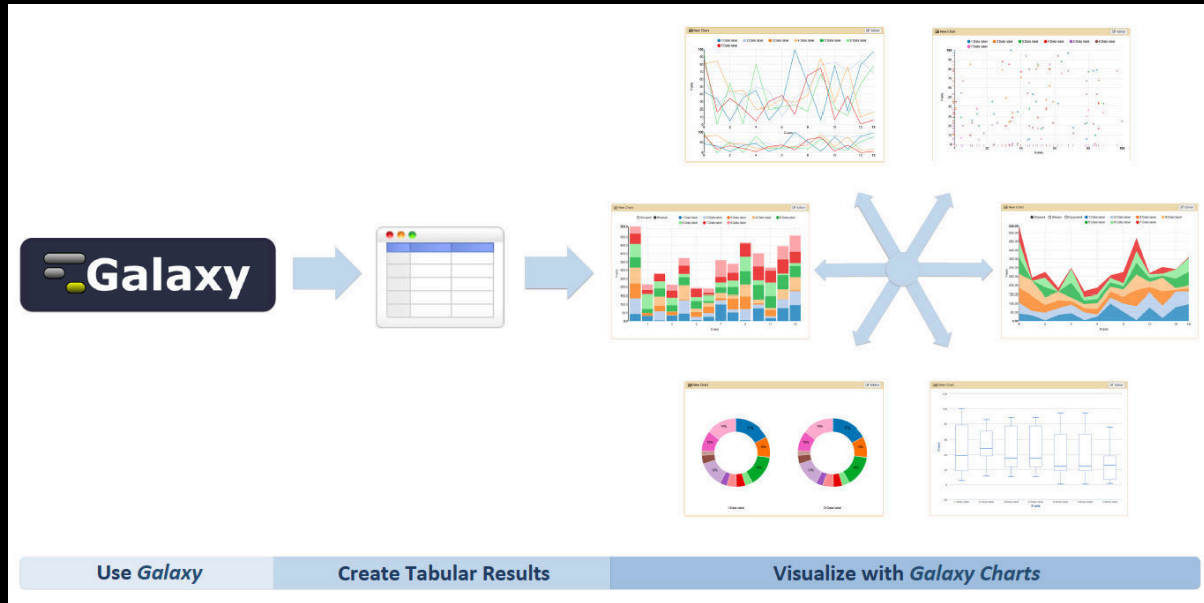
IPython

The Notebook



Charts

Charts enables users to quickly visualize tabular data.

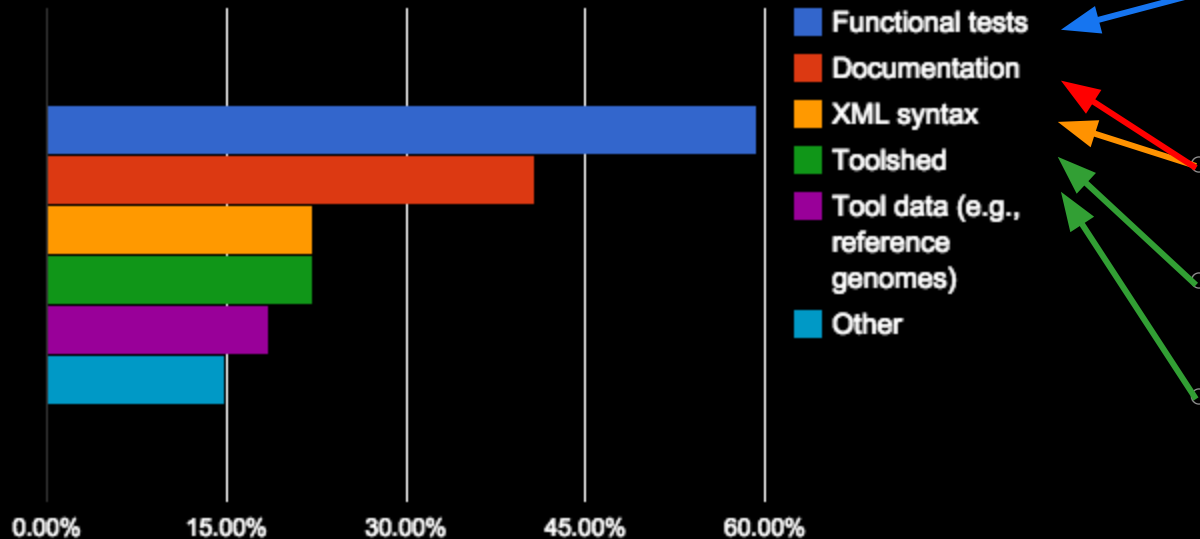


[Watch Screencast](#)

Tool Development

Planemo

What are the main hurdles in the tool development process?



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.

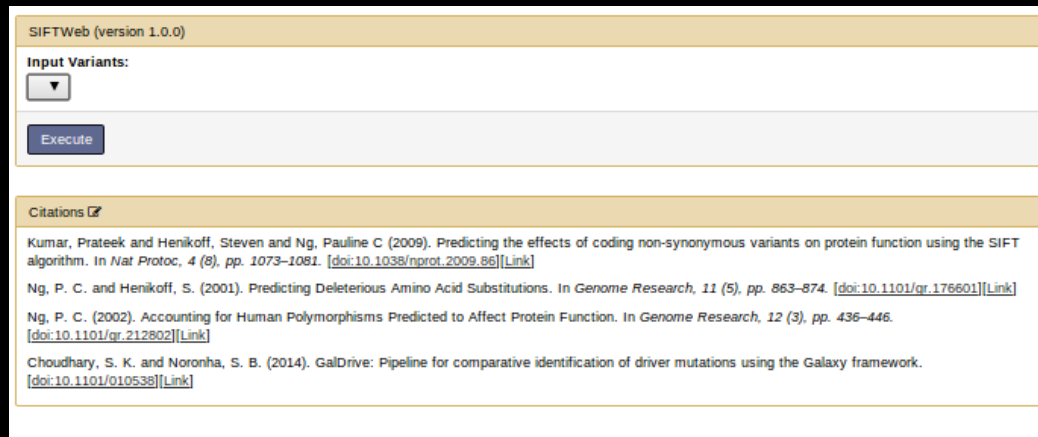
Testbed for new developments such as Docker and Homebrew.

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.



SIFTWeb (version 1.0.0)

Input Variants:

Execute

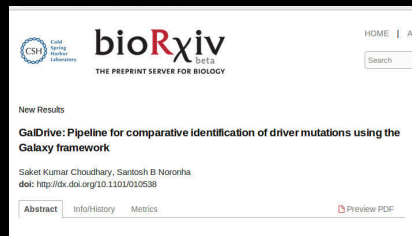
Citations 

Kumar, Prateek and Henikoff, Steven and Ng, Pauline C (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. In *Nat Protoc*, 4 (8), pp. 1073–1081. [[doi:10.1038/nprot.2009.86](https://doi.org/10.1038/nprot.2009.86)][Link]

Ng, P. C. and Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. In *Genome Research*, 11 (5), pp. 863–874. [[doi:10.1101/gr.176601](https://doi.org/10.1101/gr.176601)][Link]

Ng, P. C. (2002). Accounting for Human Polymorphisms Predicted to Affect Protein Function. In *Genome Research*, 12 (3), pp. 436–446. [[doi:10.1101/gr.212802](https://doi.org/10.1101/gr.212802)][Link]

Choudhary, S. K. and Noronha, S. B. (2014). GalDrive: Pipeline for comparative identification of driver mutations using the Galaxy framework. [[doi:10.1101/010538](https://doi.org/10.1101/010538)][Link]



bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

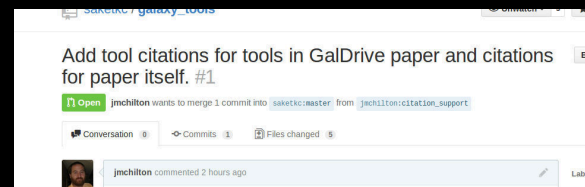
HOME | ABOUT | SEARCH

New Results

GalDrive: Pipeline for comparative identification of driver mutations using the Galaxy framework

Saket Kumar Choudhary, Santosh B Noronha.
doi: <http://dx.doi.org/10.1101/010538>

Abstract InfoHistory Metrics Preview PDF



Saket/galaxy_tools

Add tool citations for tools in GalDrive paper and citations for paper itself. #1

Open jmc Hilton wants to merge 1 commit into [saket:master](#) from [jmc Hilton:citation_support](#)

Conversation Commits Files changed

jmc Hilton commented 2 hours ago

ToolShed - A developers perspective

The vision

WITH

REPLACE

Create Repository

Name:

Repository type:

Select the repository type based on the following criteria.
Unrestricted - contents can be any set of valid Galaxy utilities or files
Repository suite definition - contents will always be restricted to one file named repository_dependencies.xml
Tool dependency definition - contents will always be restricted to one file named tool_dependencies.xml

Synopsis:

Detailed description:

Categories

☐ Assembly
☐ ChIP-seq
☐ Combinatorial Selections
☐ Computational Chemistry

Multi-select list - hold the appropriate key while clicking to select multiple categories.

Contents:

☐ suite_linkyx_bundle_0.1
☐ repository_dependencies.xml

Click on a file to display its contents below. You may delete files from the repository by clicking the check box next to each file and clicking the Delete selected files button.

Message:

Deleted selected files

This is the commit message for the mercurial change set that will be created if you delete selected files.

```
<?xml version="1.0"?>
<repositories description="Metapackage for the installation of linkyx suite of tools.">
  <repository changeset_revision="09c1e388c20c" name="trinitymaseq" owner="jjohnson" toolshed="https://testtoolshed.g2.bx.psu.edu" />
  <repository changeset_revision="06c42572d7c0" name="fastq_groomer" owner="devteam" toolshed="https://testtoolshed.g2.bx.psu.edu" />
  <repository changeset_revision="3365c42b076d" name="bwa_mappers" owner="devteam" toolshed="https://testtoolshed.g2.bx.psu.edu" />
  <repository changeset_revision="7095d651c95f" name="sam_to_bam" owner="devteam" toolshed="https://testtoolshed.g2.bx.psu.edu" />
  <repository changeset_revision="93ace7e49295" name="picard" owner="devteam" toolshed="https://testtoolshed.g2.bx.psu.edu" />
  <repository changeset_revision="5e12c7427fa3" name="samtools_sort" owner="devteam" toolshed="https://testtoolshed.g2.bx.psu.edu" />
</repositories>
```

Galaxy Tool Shed

<https://toolshed.galaxyproject.org/>

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.

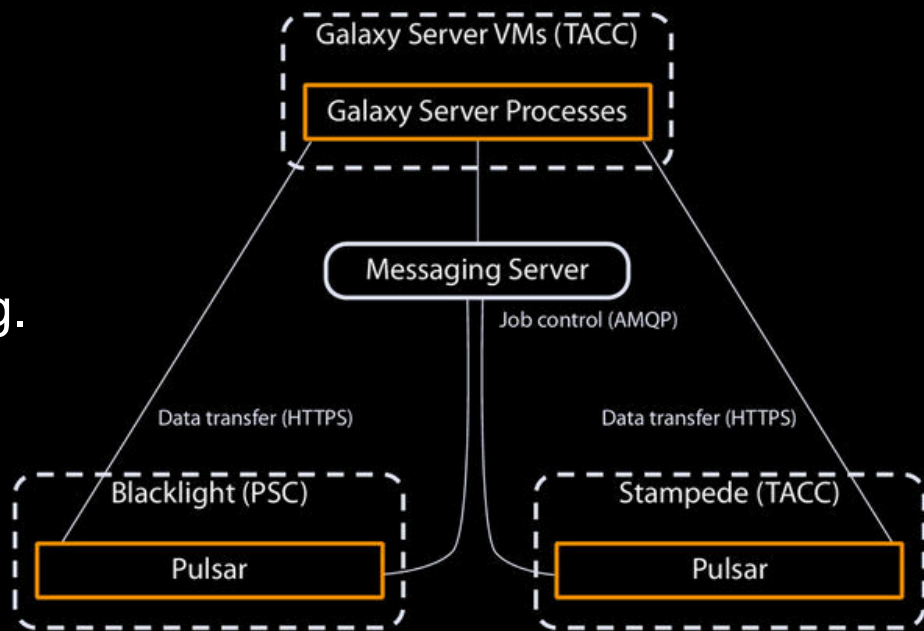
Watch Screencast

Reference iReport: a generalised Galaxy solution for integrated experimental reporting, Hiltemann et al., GigaScience 3:19, 2014

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"

	B	I	N	G	O
B	Hadoop	 <anyword>-Seq	Global Alliance	Peta	CRAM
iO	 Docker	github address 	The NHGRI graph on cost of sequencing	I'm the last person between you and [beer food]	Galaxy 
D	distributed computing	MongoDB	DATA	ICGC	Amazon
A	web services	 Cloud	Google	scalable	metadata
TA	"tweet me"	 Python Screenshot	wet lab experiment	Exa	TCGA

Abstract Session

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