

Transparent, accessible, reproducible analysis with Galaxy

Dave Clements
Emory University
12 September 2012

<http://galaxyproject.org/>



Acknowledgements

Fourie Joubert
Rouvay Roodt-Wilding
Oleg Reva
Anelda Van der Walt

University of Pretoria
Stellenbosch University



Enis Afgan



Guru Ananda



Dannon Baker



Dan Blankenberg



Dave Bouvier



Dave Clements



Nate Coraor



Carl Eberhard



Jeremy Goecks



Jen Jackson



Greg von Kuster



Ross Lazarus



Rémi Marenco



Scott McManus



Anton Nekrutenko

The Galaxy Team

<http://galaxyproject.org/wiki/Galaxy%20Team>

James Taylor



As science becomes increasingly dependent on computation:

How best to ensure that analysis are **reproducible**?

How can methods best be made **accessible** to scientists?

How to facilitate **transparent** communication of analyses?

A crisis in genomics research:
reproducibility

Key Reproducibility Problems

- **Datasets:** not all available, difficult to access
- **Tools:** inaccessible, hard to record details
- **Publication:** results, data, methods separate

Microarray Experiment Reproducibility

- 18 Nat. Genetics microarray gene expression experiments
- **Less than 50% reproducible**
- Problems
 - missing data (38%)
 - missing software, hardware details (50%)
 - missing method, processing details (66%)

Ioannidis, J.P.A. et al. Repeatability of published microarray gene expression analyses. Nat Genet 41, 149-155 (2009)

NGS Re-sequencing Experiment Reproducibility

- 14 re-sequencing experiments in Nat. Genetics, Nature, and Science (2010)
- 0% reproducible?
- Problems
 - limited access to primary data (50%)
 - some or all tools unavailable (50%)
 - settings & versions not provided (100%)

Galaxy: accessible analysis system

The screenshot displays the Galaxy web interface with the following components:

- Header:** Galaxy logo, navigation tabs (Analyze Data, Workflow, Shared Data, Visualization, Cloud, Admin, Help, User), and a status bar indicating "Using 158.2 GB".
- Tools Panel (Left):** A sidebar with a search bar and a list of tool categories including Get Data, Send Data, ENCODE Tools, 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, Phenotype Association, Genome Diversity, EMBOSS, NGS TOOLBOX BETA, NGS: QC and manipulation, and NGS: Mapping.
- Main Content Area:**
 - Additional output created by MACS (MACS_in_Galaxy)**
 - Additional Files:** A list of five files: [MACS in Galaxy model.pdf](#), [MACS in Galaxy model.r](#), [MACS in Galaxy model.r.log](#), [MACS in Galaxy negative peaks.xls](#), and [MACS in Galaxy peaks.xls](#).
 - Messages from MACS:** A log of execution messages starting with "INFO @ Wed, 21 Sep 2011 18:28:58: # ARGUMENTS LIST:" and detailing the workflow steps from reading tag files to writing the final wiggle file.
- History Panel (Right):** A list of workflow history items, including:
 - CPB2012 - BasicProtocol3 - Calling Peaks for CHIP-seq Data (1.2 GB)
 - 12: MACS on data 5 and data 6 (html report) (3.3 Kb, format: html, database: mm9)
 - 11: MACS on data 5 and data 6 (control: wig)
 - 10: MACS on data 5 and data 6 (treatment: wig)
 - 9: MACS on data 5 and data 6 (negative peaks: interval)
 - 8: MACS on data 5 and data 6 (peaks: interval)
 - 7: CTCF Peaks chr19 BED
 - 6: Tags Chr19 SAM
 - 5: Control Chr19 SAM
 - 4: Tags Chr19 groomed
 - 3: Control Chr19 groomed
 - 2: Tags Chr19 ungroomed

Integrating existing tools into a uniform framework

The image shows a screenshot of a Galaxy tool interface for a tool named "Cluster". The background is a code editor showing the XML definition of the tool, and the foreground is the user interface.

```
1 <tool id="gops_cluster_1" name="Cluster">
2   <description>[[Cluster]] the intervals of a query</description>
3   <command interpreter="python">
4     gops_cluster.py $input1 $
5     -d $dista
6   </command>
7   <inputs>
8     <param format="interval"
9       <label>Cluster interval
10    </param>
11    <param name="distance" s
12      <label>max distance bet
13    </param>
14    <param name="minregions"
15      <label>min number of in
16    </param>
17    <param name="returntype"
18      <option value="1">Merge
19      <option value="2">Find
20      <option value="3">Find
21      <option value="4">Find
22      <option value="5">Find
23    </param>
24  </inputs>
25  <help>
26
27  .. class:: infomark
28
29  **TIP:** If your query does r
30
31  ----
32
33  **Screencasts!**
34
35  See Galaxy Interval Operatio
36
37  .. _Screencasts: http://www.b
38
39  ----
40
41  **Syntax**
42
43  - **Maximum distance** is gre
44  - **Minimum intervals per clu
45  - **Merge clusters into singl
46  - **Find cluster intervals; p
47  - **Find cluster intervals; p
48
49  Line: 87 Column: 8 XML
```

The user interface for the "Cluster" tool includes the following elements:

- Cluster intervals of:** A dropdown menu showing "1: UCSC Main on Huma..ne (genome)".
- max distance between intervals:** A text input field containing "1" (bp).
- min number of intervals per cluster:** A text input field containing "2".
- Return type:** A dropdown menu showing "Merge clusters into single intervals".
- Execute** button.
- TIP:** If your query does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.
- Screencasts!** See Galaxy Interval Operation [Screencasts](#) (right click to open this link in another window).
- Syntax**
 - **Maximum distance** is greatest distance in base pairs allowed between intervals that will be

- Defined in terms of an abstract interface (inputs and outputs)
- In practice, mostly command line tools, a declarative XML description of the interface, how to generate a command line
- Designed to be as easy as possible for tool authors, while still allowing rigorous reasoning




Galaxy analysis interface

The screenshot displays the Galaxy web interface for running the MACS (version 1.0.1) tool. The interface is divided into several sections:

- Tools Panel (Left):** A sidebar with a search bar and a list of tool categories including Get Data, Send Data, ENCODE Tools, 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, Phenotype Association, Genome Diversity, EMBOSS, and NGS TOOLBOX BETA.
- Tool Configuration (Center):** The MACS (version 1.0.1) tool configuration page. It includes fields for:
 - Experiment Name: MACS in Galaxy
 - Paired End Sequencing: Single End
 - ChIP-Seq Tag File: 6: Tags Chr19 SAM
 - ChIP-Seq Control File: 5: Control Chr19 SAM
 - Effective genome size: 1870000000.0 (default: 2.7e+9)
 - Tag size: 36
 - Band width: 300
 - Pvalue cutoff for peak detection: 1e-05 (default: 1e-5)
 - Select the regions with MFOLD high-confidence enrichment ratio against background to build model: 32
 - Parse xis files into into distinct interval files:
 - Save shifted raw tag count at every bp into a wiggle file: Save
 - Extend tag from its middle point to a wigextend size fragment: -1 (Use value less than 0 for default (modeled d))
 - Resolution for saving wiggle files:
- History Panel (Right):** A list of previous tool runs. The most recent run is highlighted in green and includes:
 - 12: MACS on data 5 and data 6 (html report)
 - 11: MACS on data 5 and data 6 (control: wig)
 - 10: MACS on data 5 and data 6 (treatment: wig)
 - 9: MACS on data 5 and data 6 (negative peaks: interval)
 - 8: MACS on data 5 and data 6 (peaks: interval)
 - 7: CTCF Peaks chr19 BED (720 regions, 1 comments format: bed, database: mm9) with links to UCSC main, GeneTrack, IGB Local Web, and Ensembl Current.

- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses
- Exact parameters of a step can always be inspected, and easily rerun

Automatically tracks every step of every analysis

7: Map with Bowtie for Illumina on data 6 and data 5   

9,073,928 lines, format: sam,
database: mm9
Run this job again

1. QNAME	2. FLAG	3. I
HWI-EAS269:3:1:1449:913	99	chr
HWI-EAS269:3:1:1449:913	147	chr
HWI-EAS269:3:1:709:832	99	chr
HWI-EAS269:3:1:709:832	147	chr
HWI-EAS269:3:1:1422:1087	99	chr
HWI-EAS269:3:1:1422:1087	147	chr

Map with Bowtie for Illumina

Will you select a reference genome from your history or use a built-in index?

Built-ins were indexed using default options

Select a reference genome:

if your genome of interest is not listed – contact Galaxy team

Is this library mate-paired?:

Forward FASTQ file:

Must have Sanger-scaled quality values with ASCII offset 33

Reverse FASTQ file:

Must have Sanger-scaled quality values with ASCII offset 33

Maximum insert size for valid paired-end alignments (-X):

The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff):

Bowtie settings to use:

For most mapping needs use Commonly used settings. If you want full control use Full parameter list

Suppress the header in the output SAM file:

Bowtie produces SAM with several lines of header information by default

As well as user-generated metadata and annotation...

History Options ▾

Variant Analysis for Sample E18

Tags:

snp × pileup × bowtie ×
demo × sample:e18 ×

Annotation / Notes:
Perform a variant analysis with default parameters to identify variants in sample E18 that lie in annotated genes.

10: Variants from sample E18 👁️ ✎️ ✕

26,742 regions, format: interval, database: mm9

Info: 📄 🔄

Tags:

pileup × sample:e18 ×
snps ×

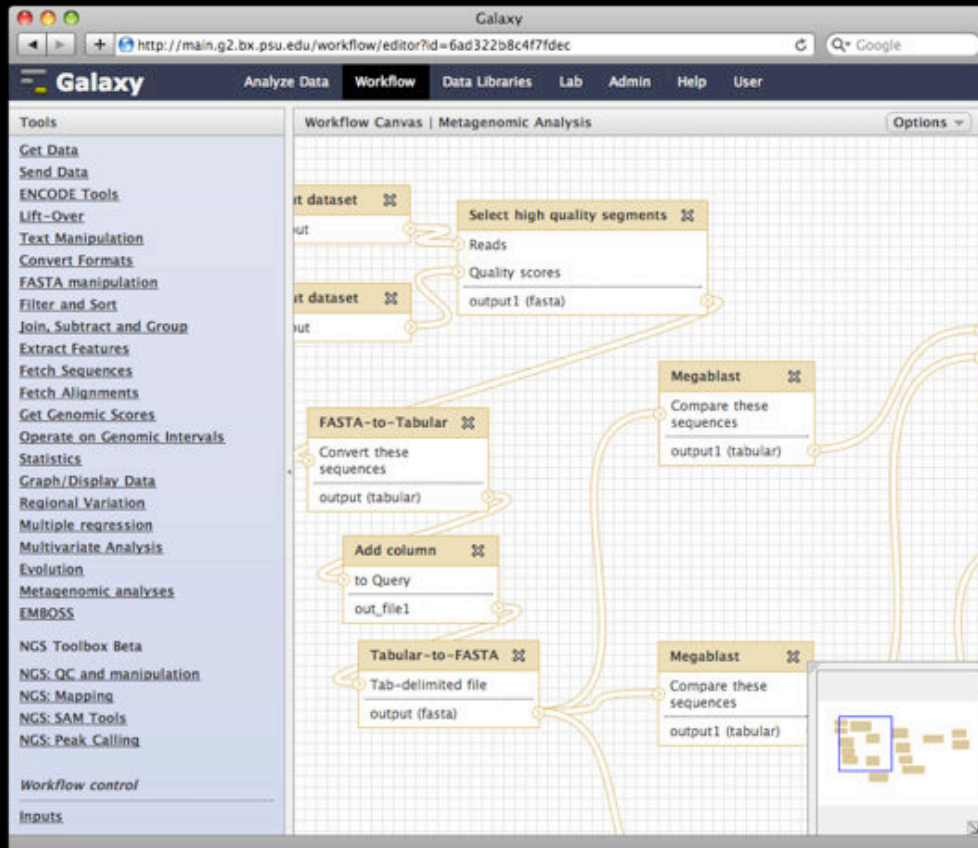
Annotation:

Find variants with coverage ≥ 30 and quality score ≥ 20 .

| display at UCSC [main](#) | view in [GeneTrack](#) | display at Ensembl [Current](#)

1. Chrom	2. Start	3. End	4	5	6
chr10	6882036	6882037	A	A	107
chr10	14243075	14243076	G	G	96
chr10	14243079	14243080	C	C	106
chr10	14465082	14465083	T	K	173
chr10	14465083	14465084	G	K	144
chr10	14465084	14465085	T	T	117

Galaxy workflow system



- **Workflows** can be constructed from scratch or extracted from existing analysis histories
- Facilitate reuse, as well as providing precise reproducibility of a complex analysis

Transparency: Sharing and publishing

The screenshot shows a web browser window displaying a Galaxy page. The browser's address bar shows the URL: <http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter>. The page title is "Galaxy | Published Page | Windshield Splatter". The main content area features the title "Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement" and lists authors: SERGEI KOSAKOVSKY POND^{1,2*}, SAMIR WADHAWAN^{3,6*}, FRANCESCA CHIAROMONTE⁴, GURUPRASAD ANANDA^{1,3}, WEN-YU CHUNG^{1,3,7}, JAMES TAYLOR^{1,5}, ANTON NEKRUTENKO¹⁻³ and THE GALAXY TEAM^{1*}. Below the authors, there is a section titled "How to use this document" which explains that the document is a live copy of supplementary materials for a manuscript, providing access to analyses and workflows. It includes three interactive elements: a "Galaxy History | Galaxy vs MEGAN" comparison, a "Galaxy History | metagenomic analysis", and a "Galaxy Workflow | metagenomic analysis". The page also has a "Supplemental Analysis" section with a link to "Comparison between Galaxy pipeline and Megan".

- All analysis components (datasets, histories, workflows) can be shared among Galaxy users and published
- Pages and annotation allow analysis to be augmented with textual content and provided in the form of an integrated document

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement



HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP

Institution: PENN STATE UNIV Sign In via User Name/Password

Search for Keyword: Go
Advanced Search

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond^{1,2,6,9}, Samir Wadhawan^{3,6,7},
Francesca Chiaromonte⁴, Guruprasad Ananda^{1,3}, Wen-Yu Chung^{1,3,8},
James Taylor^{1,5,9}, Anton Nekrutenko^{1,3,9} and The Galaxy Team¹

OPEN ACCESS ARTICLE

This Article

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109
Copyright © 2009 by Cold Spring Harbor Laboratory Press

» Abstract **Free**
» Full Text (PDF) **Free**

Current Issue

October 2010, 20 (10)



1,3,7, JAMES TAYLOR^{1,5}, ANTON

es and workflows discussed data. Specifically, we provide em. You can also import workflows you must create e and password.

Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript.)

Footnotes

[Supplemental material is available online at <http://www.genome.org>. All data and tools described in this manuscript can be downloaded or used directly at <http://galaxyproject.org>. Exact analyses and workflows used in this paper are available at <http://usegalaxy.org/u/aun1/p/windshield-splatter>.]

Supplemental Analysis

Comparison between Galaxy pipeline and Megan

Loading "http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter", completed 5 of 6 items

Galaxy is available ...

- **As a free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

<http://usegalaxy.org>

usegalaxy.org: a wealth of tools

NGS: QC and manipulation

ILLUMINA DATA

- [FASTQ Groomer](#) convert between various FASTQ qual formats
- [FASTQ splitter](#) on joined paired end reads
- [FASTQ joiner](#) on paired end reads
- [FASTQ Summary Statistics](#) by column

ROCHE-454 DATA

- [Build base quality distribution](#)
- [Select high quality segments](#)
- [Combine FASTA and QUAL](#) in FASTQ

AB-SOLID DATA

- [Convert SOLiD output to fastq](#)
- [Compute quality statistics](#) for SOLiD data
- [Draw quality score boxplot](#) for SOLiD data

GENERIC FASTQ MANIPULATION

- [Filter FASTQ reads](#) by quality score and length
- [FASTQ Trimmer](#) by column
- [FASTQ Quality Trimmer](#) by sliding window
- [FASTQ Masker](#) by quality score

- [Manipulate FASTQ reads](#) on various attributes

- [FASTQ to FASTA](#) converter
- [FASTQ to Tabular](#) converter
- [Tabular to FASTQ](#) converter

FASTX-TOOLKIT FOR FASTQ DATA

- [Quality format converter](#) (ASCII Numeric)
- [Compute quality statistics](#)
- [Draw quality score boxplot](#)
- [Draw nucleotides distribution chart](#)

- [FASTQ to FASTA](#) converter
- [Filter by quality](#)
- [Remove sequencing artifacts](#)

- [Barcode Splitter](#)
- [Clip adapter sequences](#)
- [Collapse sequences](#)
- [Rename sequences](#)
- [Reverse-Complement](#)
- [Trim sequences](#)

FASTQ QC

- [FastQC:Read QC](#) reports using FastQC

NGS: Mapping

ILLUMINA

- [Map with Bowtie for Illumina](#)

- [Map with BWA for Illumina ROCHE-454](#)

- [Lastz](#) map short reads against reference sequence

- [Megablast](#) compare short reads against htgs, nt, and wgs databases

- [Parse blast XML output](#)

AB-SOLID

- [Map with Bowtie for SOLiD](#)

- [Map with BWA for SOLiD](#)

NGS: SAM Tools

- [Filter SAM](#) on bitwise flag values

- [Convert SAM](#) to interval

- [SAM-to-BAM](#) converts SAM format to BAM format

- [BAM-to-SAM](#) converts BAM format to SAM format

- [Merge BAM Files](#) merges BAM files together

- [Generate pileup](#) from BAM dataset

- [Filter pileup](#) on coverage and SNPs

- [Pileup-to-Interval](#) condenses pileup format into ranges of bases

- [flagstat](#) provides simple stats on BAM files

- [rmdup](#) remove PCR duplicates

- [MPileup](#) SNP and indel caller

- [Slice BAM](#) by provided regions

NGS: GATK Tools (beta)

ALIGNMENT UTILITIES

- [Depth of Coverage](#) on BAM files

- [Print Reads](#) from BAM files

REALIGNMENT

- [Realigner Target Creator](#) for use in local realignment

- [Indel Realigner](#) - perform local realignment

BASE RECALIBRATION

- [Count Covariates](#) on BAM files

- [Table Recalibration](#) on BAM files

- [Analyze Covariates](#) - draw plots

GENOTYPING

- [Unified Genotyper](#) SNP and indel caller

ANNOTATION

- [Variant Annotator](#)

FILTRATION

- [Variant Filtration](#) on VCF files

- [Select Variants](#) from VCF files

VARIANT QUALITY SCORE RECALIBRATION

- [Variant Recalibrator](#)

- [Apply Variant Recalibration](#)

VARIANT UTILITIES

- [Validate Variants](#)

- [Eval Variants](#)

- [Combine Variants](#)

NGS: Indel Analysis

- [Filter Indels](#) for SAM

- [Extract indels](#) from SAM

- [Indel Analysis](#)

NGS: Peak Calling

- [MACS Model-based Analysis](#) of ChIP-Seq

- [SICER](#) Statistical approach for the Identification of ChIP-Enriched Regions

- [GeneTrack indexer](#) on a BED file

- [Peak predictor](#) on GeneTrack index

NGS: RNA Analysis

RNA-SEQ

- [Tophat for Illumina](#) Find splice junctions using RNA-seq data

- [Cufflinks](#) transcript assembly and FPKM (RPKM) estimates for RNA-Seq data

- [Cuffcompare](#) compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments

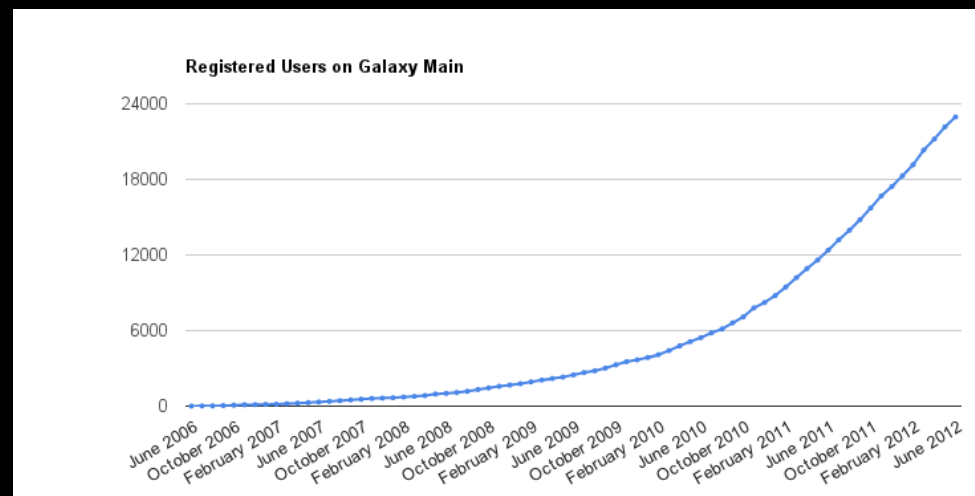
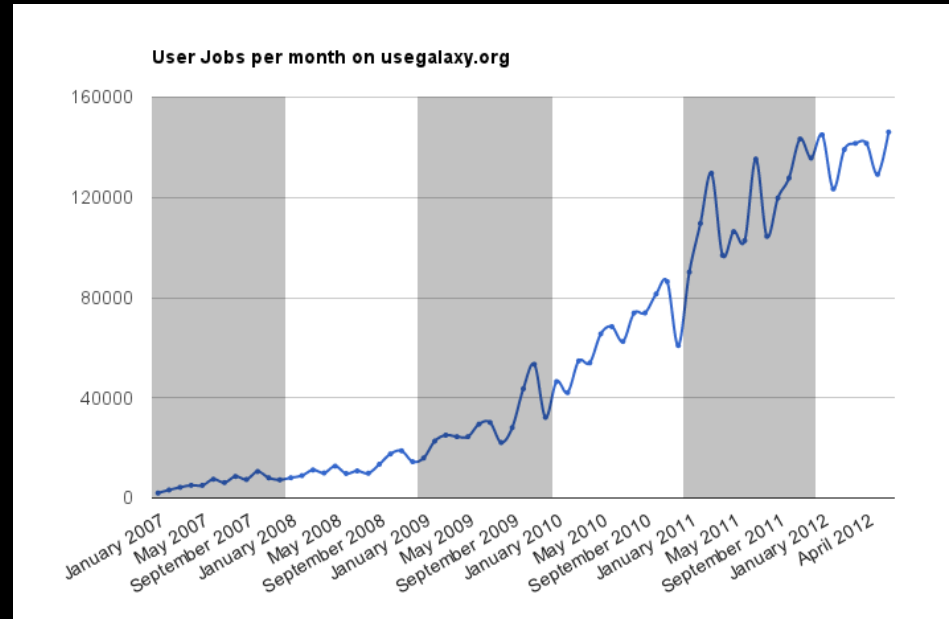
- [Cuffmerge](#) merge together several Cufflinks assemblies

- [Cuffdiff](#) find significant changes in transcript expression

For example, the first 5 pages of NGS tools

<http://usegalaxy.org> (a.k.a Main)

- Free public web site
- Anybody can use it
- Hundreds of tools
- Persistent
- 24,000 registered users
- 300+ TB of user data
- 140,000+ jobs / month



<http://bit.ly/gxystats>

But, it's a big world

- Main has lots of tools, storage, processor, users, ...
 - But **not all tools** - there are thousands and adding new tools is not taken lightly
 - But **not infinite storage and processors** - Main now has job limits and storage quotas
- **A centralized solution cannot scale to meet data analysis demands of the whole world**

Galaxy is available ...

- As a free (for everyone) web service integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **As open source software** that makes integrating your own tools and data and customizing for your own site simple

<http://getgalaxy.org>

Local Galaxy Instances

- Galaxy is designed for local installation and customization
 - Easily integrate new tools
 - Easy to deploy and manage on nearly any (unix) system
 - Run jobs on existing compute clusters
- Requires an existing computational resource on which to be deployed

<http://getgalaxy.org>

Encourage Local Galaxy Instances

- Support **increasingly decentralized model** and improve access to existing resources
- Focus on building **infrastructure to enable the community to integrate and share** tools, workflows, and best practices

Galaxy Tool Shed

- Allow sites to share “suites” containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies
- Integration with Galaxy instances to automate tool installation and updates

toolshed.g2.bx.psu.edu

Public Galaxy Servers

<http://galaxyproject.org/wiki/PublicGalaxyServers>

Interested in:

- ChIP-chip and ChIP-seq?
 - ✓ Cistrome
- Statistical Analysis?
 - ✓ Genomic Hyperbrowser
- Sequence and tiling arrays?
 - ✓ Oqtans
- Text Mining?
 - ✓ DBCLS Galaxy
- Reasoning with ontologies?
 - ✓ GO Galaxy
- Internally symmetric protein structures?
 - ✓ SymD

Local Galaxy Instances

- Galaxy is designed for local installation and customization
 - Easily integrate new tools
 - Easy to deploy and manage on nearly any (unix) system
 - Run jobs on existing compute clusters
- Requires an **existing computational resource** on which to be deployed

<http://getgalaxy.org>

Galaxy is available ...

- As a free (for everyone) web service integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- As open source software that makes integrating your own tools and data and customizing for your own site simple
- **On the Cloud**

<http://usegalaxy.org/cloud>

Galaxy CloudMan

<http://usegalaxy.org/cloud>

- Start with a **fully configured and populated** (tools and data) Galaxy instance.
- Allows you to scale up and down your compute assets as needed.
- Someone else manages the data center.
- **We used Amazon for the Pretoria and Stellenbosch workshops**



<http://aws.amazon.com/education>

Step by Step Instructions on the Wiki for Amazon

Getting Started with Galaxy CloudMan

This page provides a step-by-step instructions on how to start your own instance of Galaxy on [Amazon Web Services \(AWS\) Elastic Compute Cloud \(EC2\)](#). More general information and instructions about Galaxy CloudMan (GC) can be found [here](#).

AWS

- Get Started
- Capacity Planning
- AMIs
- ↑ CloudMan

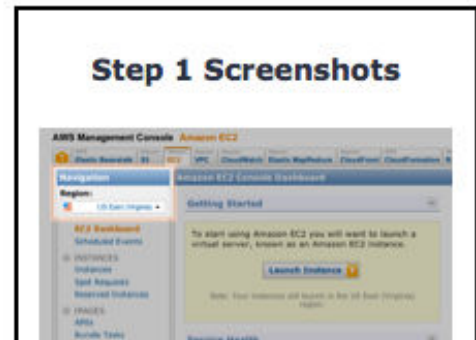
Contents

1. [Step 1: One Time Amazon Setup](#)
2. [Step 2: Starting a Master Instance](#)
3. [Step 3: Galaxy CloudMan Web Interface](#)
4. [Step 4: Use Galaxy as you normally would](#)
5. [Step 5: Shutting Down](#)

Step 1: One Time Amazon Setup

1. Because AWS services implement pay-as-you-go access model for compute resources, it is necessary for every user of the service to *register with Amazon*. **You will need a credit card to register.** (You can apply for a [AWS Education Grant](#) after you register).
2. Once your account has been approved by Amazon (note that this may take up to one business day), *log into the EC2 AWS Management Console* and set your AWS Region to *US East (Virginia)*. This is the only region Galaxy CloudMan is fully

Step 1 Screenshots



Instant CloudMan

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. A 'Using 0%' indicator is on the right. The left sidebar is titled 'Tools' and contains a search bar and a list of data sources under 'Get Data'. The main content area displays 'Managing Data: Store, Manage, and Share data with Libraries' with a link to 'An in-depth tutorial'. A 'Live Quickies' section is visible below. A 'New Cloud Cluster' dialog box is open over the 'Cloud' menu, showing a '0 bytes' status and a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.

Launch a CloudMan instance directly from Main, and transfer your current history.

The screenshot shows the 'Launch a Galaxy Cloud Instance' form. It includes the following fields and options:

- Cluster Name:
- Password:
- Key ID:
- Secret Key:
- Instance Share String (optional):
- Instance Type:

Requesting the instance may take a moment, please be patient. Do not refresh your browser or navigate away from the page

Visualize

Send data results to **external** genome browsers:

UCSC, Ensembl, GBrowse, IGV

Trackster: Galaxy's genome browser

Trackster

View your data from within Galaxy

- ✦ No data transfers to external site
- ✦ Use it locally, even without internet access

Supports common filetypes

- ✦ BAM, BED, GFF/GTF, WIG

Unique features

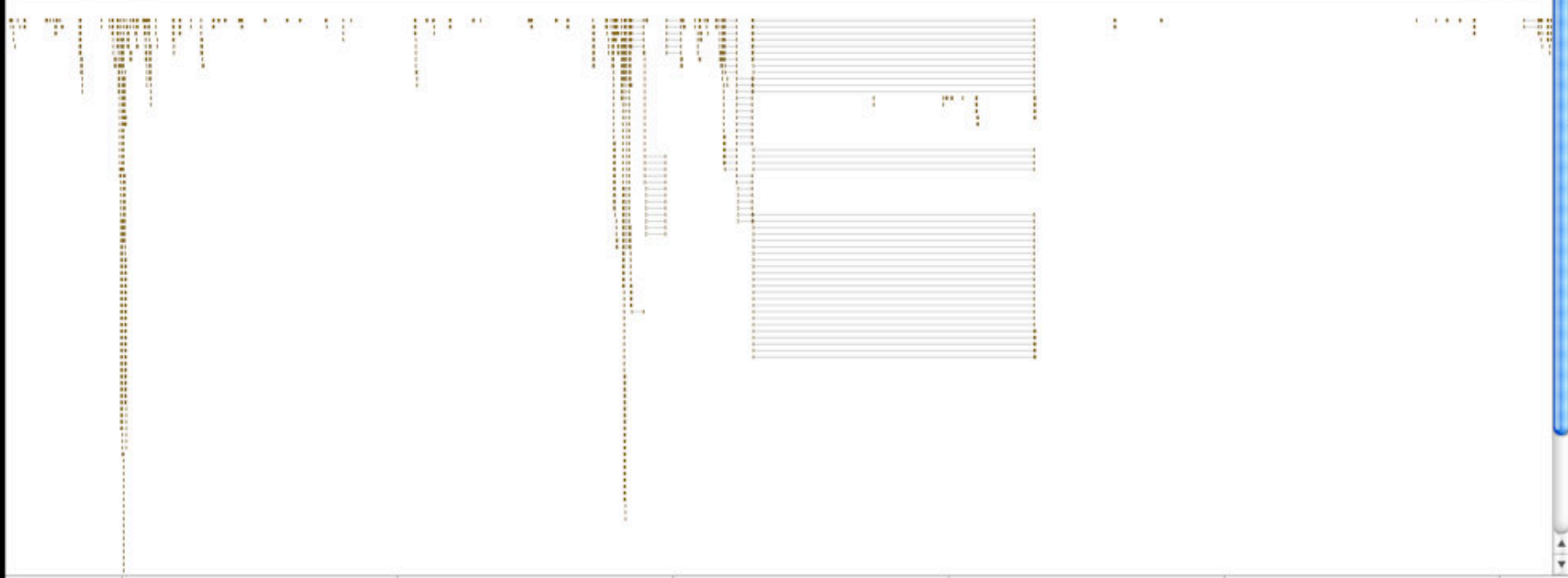
- ✦ custom genomes
- ✦ highly interactive



630,000 640,000 650,000 660,000 670,000 680,000



-1



630,000 640,000 650,000 660,000 670,000 680,000

Published Visualizations | jeremy | GCC2011-1: Viewing and chr19 663,032 - 663,110

g g c c e g g g c c T C A C C G G C A G G C G C G G G R C G A T C T C C A C G G A G C A G C A G T G G C A G A R G T A C C G T C C G G G A T G C G G C G A C C

UCSC Main on Human: knownGene (chr19) Auto (Pack)

UCSC Main on Human: all_est (chr19) Dense

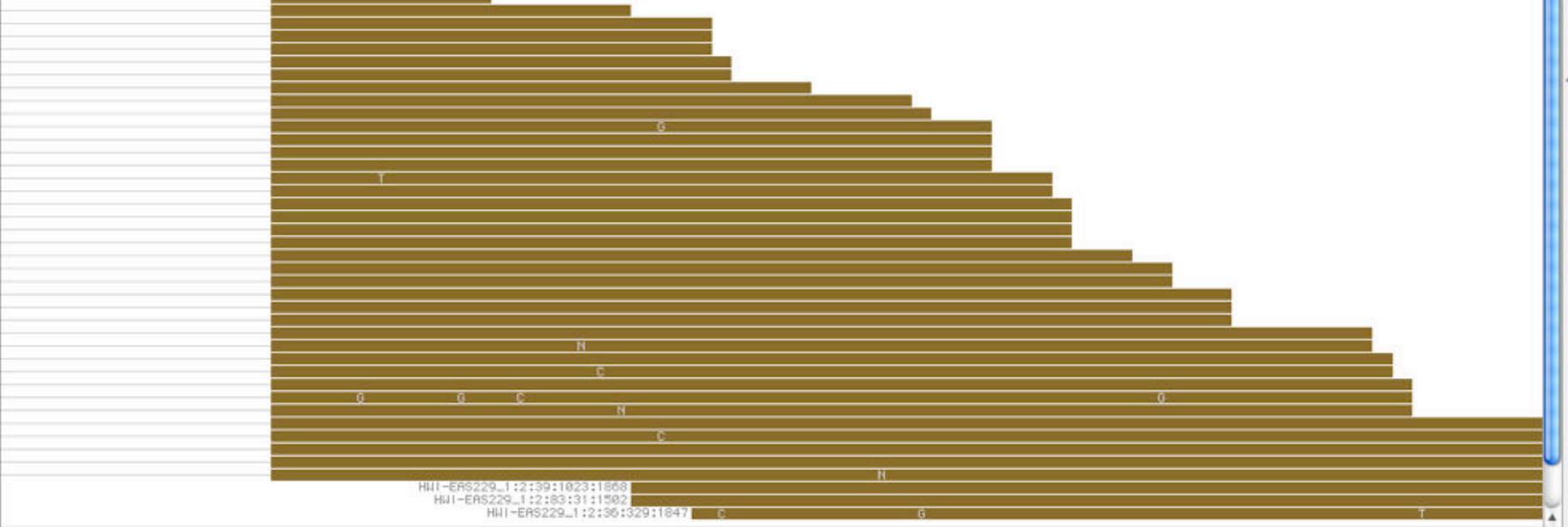
UCSC Main on Human: phyloP46wayPrimates (chr19) Histogram

1



-1

h1-hESC Tophat Mapped Reads Auto (Pack)



h1-hESC Cufflinks assembled transcripts Auto (Pack)

30 663,040 663,050 663,060 663,070 663,080 663,090 663,100

Canceled opening the page

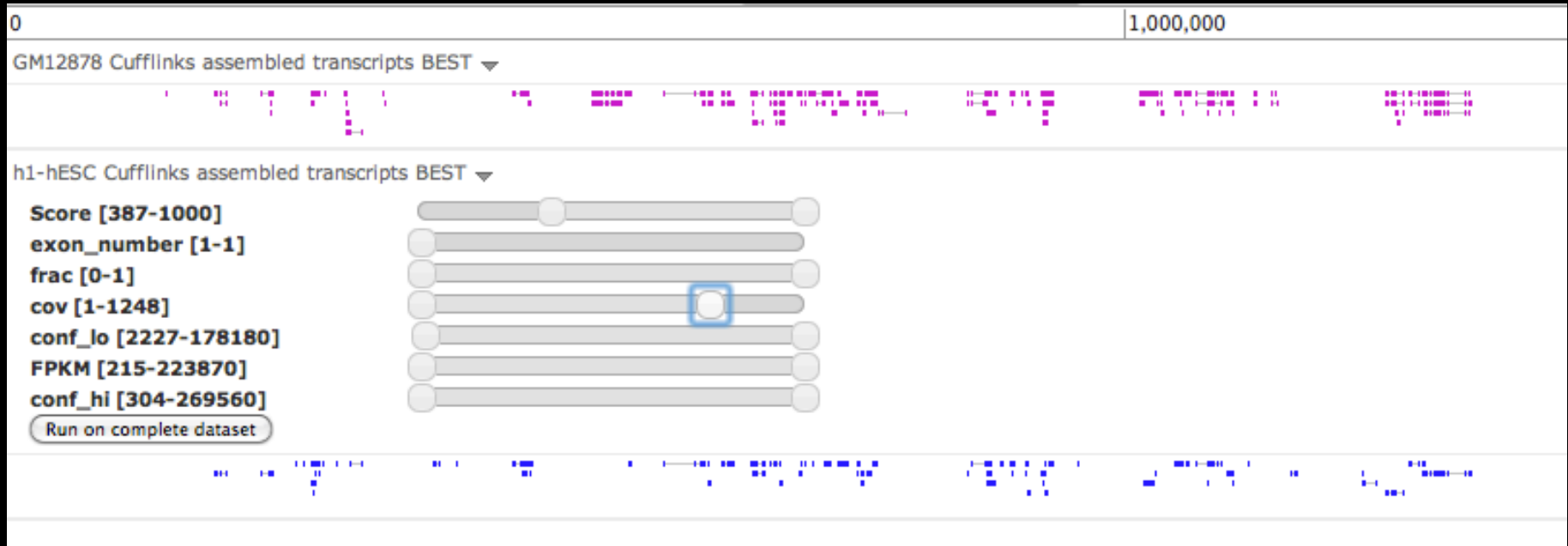
But really, why *another* genome browser

From static browsing to **visual analysis**

Visual feedback and experimentation needed for complex tools with many parameters

Leverage Galaxy strengths: a very sound model for abstracting interfaces to analysis tools and already integrates an enormous number

Dynamic Filtering



Integrating Tools and Visualization

Galaxy Analyze Data Workflow Shared Data **Visualization** Admin Help User

GCC3: Running Tools (hg19) chr19 1,523,098 - 1,545,232 1,530,000 1,540,000

UCSC Main on Human: knownGene

h1-hESC Tophat mapped reads

h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No]

Cufflinks

Max Intron Length: 150000

Min Isoform Fraction: 0.5

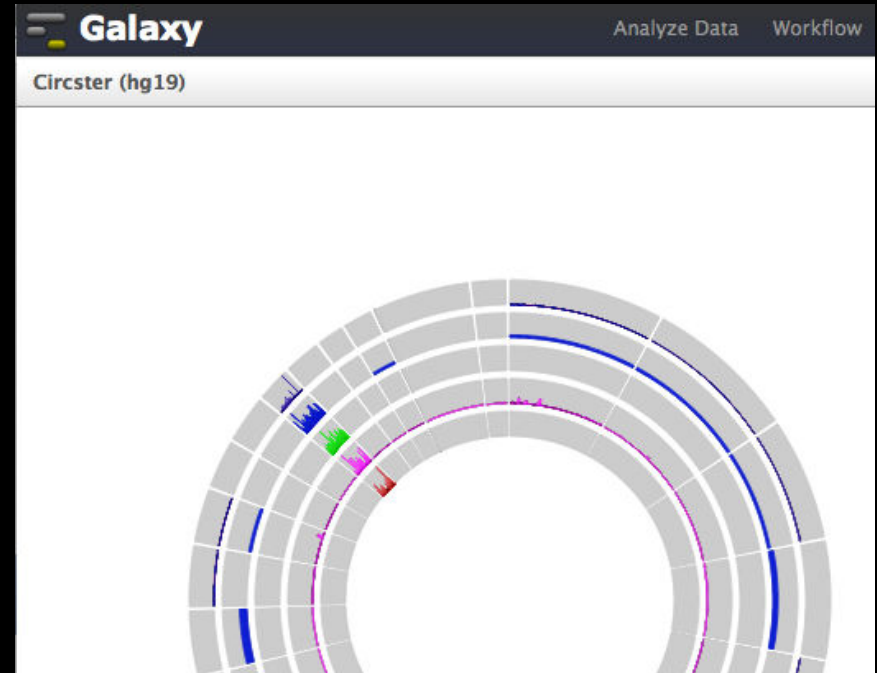
Pre MRNA Fraction: 0.05

Perform quartile normalization: No

Run on complete dataset Run on visible region

CUFF.139.1 CUFF.139.1 CUFF.140.1 CUFF.141.1 CUFF.142.1

Trackster enables other integrations and visualizations:
Parameter sweeps
Circster



The image shows a Galaxy interface window titled "Cufflinks (version 0.0.5)". The left panel contains the following parameters:

- Max Intron Length: 300000
- Min Isoform Fraction: 0 - 0.1 samples: 3
- Pre MRNA Fraction: 0 - 0.1 samples: 3
- Perform quartile normalization: No

The middle panel shows a visualization of the parameter space. It starts with a "Root" node on the left, which branches into three nodes representing different "Min Isoform Fraction" values: 0, 0.05, and 0.1. Each of these nodes further branches into three nodes representing different "Pre MRNA Fraction" values: 0, 0.05, and 0.1. This represents a 3x3 grid of parameter combinations.

The right panel shows a detailed view of the results for a specific region: "chr19:567970-588681". It displays three alternative splicing events, each with its own set of junctions and isoforms. The junctions are represented by yellow bars, and the isoforms are represented by horizontal lines with arrows indicating the direction of transcription. The events are labeled as CUFF.1.1, CUFF.1.2, CUFF.1.3, CUFF.2.1, CUFF.3.1, CUFF.4.1, and CUFF.5.1.

Galaxy Community

- Mailing Lists (very active)
- Screencasts
- Events Calendar, News Feed
- Community Wiki
- CiteULike group, Mendeley mirror
- Local Public Installs
- Tool Shed
- Annual Community Meeting

<http://galaxyproject.org/wiki>

Galaxy Community Conference

30 June
- 2 July

2013



OSLO



UiO : University of Oslo

<http://galaxyproject.org/GCC2013>

Galaxy URLs to Remember

<http://galaxyproject.org>

<http://usegalaxy.org>

<http://getgalaxy.org>

Thank you.





GCC2013

Annual gathering of the Galaxy Community will happen in Oslo Norway next summer

3 days of learning, best practices, and research

<http://galaxyproject.org/GCC2013>

Participants:

69 in 2010

148 in 2011

203 in 2012

??? in 2013



Visualization: Even More

- usegalaxy.org → Shared Data → Published Visualizations
- galaxyproject.org/wiki/Events/GCC2012/Program
→ Session 4 → The Galaxy Visualization Framework
 - Jeremy Goecks' GCC2012 presentation.
 - Basic Navigation Demo starts @ 10:40
 - Dynamic Filtering Demo starts @ 12:15
 - Circster Demo starts @ 14:10
 - Visual Analytics Demo starts @ 15:40

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

Share:

Make something available to someone else

Publish:

Make something available to everyone

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