

# Transparent, accessible, reproducible analysis with Galaxy

---

Indiana University  
19 October 2012

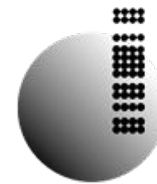
Dave Clements  
Emory University

<http://galaxyproject.org/>



NATIONAL CENTER FOR  
GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY



THE CENTER FOR  
GENOMICS AND  
BIOINFORMATICS



INDIANA UNIVERSITY

The Galaxy logo consists of a stylized 'G' made of three horizontal bars (two grey, one yellow) followed by the word 'Galaxy' in a bold, white, sans-serif font, all set against a dark blue rectangular background.

# Acknowledgements

Richard LeDuc  
William Barnett  
Scott Michaels  
Radhika Khetani

National Center for Genome Analysis Support (NCGAS)  
The Center for Genomics and Bioinformatics  
Indiana University



Enis Afgan



Guru Ananda



Dannon Baker



Dan Blankenberg



Dave Bouvier



Dave Clements



Nate Coraor



Carl Eberhard



Jeremy Goecks



Nuwan Goonasekera



Jen Jackson



Greg von Kuster



Ross Lazarus



Rémi Marenco



Scott McManus



Anton  
Nekrutenko



James  
Taylor

# The Galaxy Team

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

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)*

# 50 papers citing bwa

31 provide **no** version and **no** settings

8 lists versions

4 list settings

7 lists versions **and** settings

26 do not provide access to data

*Nekrutenko & Taylor, "Next-generation sequencing data interpretation: enhancing reproducibility and accessibility" Nature Reviews Genetics 13, 667-672 (September 2012)  
doi:10.1038/nrg3305*



# Galaxy: accessible analysis system

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

- Top Navigation Bar:** Includes tabs for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Admin', 'Help', and 'User'. A status indicator on the right shows 'Using 158.2 GB'.
- Left Panel (Tools):** 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'.
- Main Panel:**
  - Header:** 'Additional output created by MACS (MACS\_in\_Galaxy)'
  - Additional Files:** A list of five downloadable files:
    - [MACS in Galaxy model.pdf](#)
    - [MACS in Galaxy model.r](#)
    - [MACS in Galaxy model.r.log](#)
    - [MACS in Galaxy negative peaks.xls](#)
    - [MACS in Galaxy peaks.xls](#)
  - Messages from MACS:** A log of system messages starting with 'INFO @ Wed, 21 Sep 2011 18:28:58:'. The log details the execution of the MACS workflow, including argument lists, file paths, genome size (1.87e+09), band width (300), model fold (32), p-value cutoff (1.00e-05), and the generation of peak regions and wiggle files.
- Right Panel (History):** A list of workflow history items, each with a thumbnail, name, size, and format. The items include:
  - 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 Galaxy tool interface for a tool named 'Cluster'. On the left, a code editor displays the tool's XML definition. The XML includes a description, command interpreter (python), command (gops\_cluster.py), inputs (format, distance, minregions, returntype), and help text. The main panel shows the tool's GUI with a dropdown for 'Cluster intervals of:' set to '1: UCSC Main on Human genome', input fields for 'max distance between intervals:' (1) and 'min number of intervals per cluster:' (2), and a 'Return type:' dropdown set to 'Merge clusters into single intervals'. An 'Execute' button is at the bottom. A tip box states: '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.' Below the tip are sections for 'Screencasts!' and 'Syntax'. The 'Syntax' section lists: 'Maximum distance' is greatest distance in base pairs allowed between intervals that will be included in the cluster; 'Minimum intervals per cluster' is the minimum number of intervals that must be present in a cluster; 'Merge clusters into single intervals' is the default; 'Find cluster intervals' is the default; 'Find cluster intervals' is the default.

```
<?xml version="1.0" encoding="UTF-8"?>
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="python">
    gops_cluster.py $input1 $
    -d $distance
  </command>
  <inputs>
    <param format="interval"
      <label>Cluster interval
    </param>
    <param name="distance" si
      <label>max distance bet
    </param>
    <param name="minregions"
      <label>min number of in
    </param>
    <param name="returntype"
      <option value="1">Merge
      <option value="2">Find
      <option value="3">Find
      <option value="4">Find
      <option value="5">Find
    </param>
  </inputs>
  <help>
    .. class:: infomark
    **TIP:** If your query does n
    ....
    **Screencasts!**
    See Galaxy Interval Operation
    .. _Screencasts: http://www.b
    ....
    **Syntax**
    - **Maximum distance** is gre
    - **Minimum intervals per clu
    - **Merge clusters into singl
    - **Find cluster intervals; p
    - **Find cluster intervals; c
  </help>
</tool>
```

Cluster

Cluster intervals of:  
1: UCSC Main on Human genome

max distance between intervals:  
1  
(bp)

min number of intervals per cluster:  
2

Return type:  
Merge clusters into single intervals

Execute

**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 included in the cluster.
- Minimum intervals per cluster is the minimum number of intervals that must be present in a cluster.
- Merge clusters into single intervals is the default.
- Find cluster intervals; p is the default.
- Find cluster intervals; c is the default.

- 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. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Cloud, Admin, Help, and User. The left sidebar lists various tool categories such as 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, NGS: Mapping, NGS: SAM Tools, and NGS: GATK Tools (beta).




The main panel shows the configuration for the MACS (version 1.0.1) tool. The Experiment Name is "MACS in Galaxy". The Paired End Sequencing is set to "Single End". The ChIP-Seq Tag File is "6: Tags Chr19 SAM" and the ChIP-Seq Control File is "5: Control Chr19 SAM". The Effective genome size is "1870000000.0" (default: 2.7e+9). The Tag size is "36". The Band width is "300". The Pvalue cutoff for peak detection is "1e-05" (default: 1e-5). The Select the regions with MFOLD high-confidence enrichment ratio against background to build model is "32". The Parse xls files into into distinct interval files checkbox is checked. The Save shifted raw tag count at every bp into a wiggle file dropdown is set to "Save". The Extend tag from its middle point to a wigextend size fragment dropdown is set to "-1" (Use value less than 0 for default (modeled d)). The Resolution for saving wiggle files is set to "1".


The right panel shows the History system, listing previous analyses. The top entry is "CPB2012 - BasicProtocol3 - Calling Peaks for ChIP-seq Data" (1.2 GB). Below it are several MACS analyses on data 5 and data 6, including html reports, control wigs, treatment wigs, negative peaks intervals, and peaks intervals. The bottom entry is "7: CTCF Peaks chr19 BED" (720 regions, 1 comments, format: bed, database: mm9), with links to display at UCSC main, view in GeneTrack, display in IGB Local Web, and display at Ensembl Current. Below the history list is a table with columns 1.Chrom, 2.Start, 3.End, and 4.Name, showing track names and coordinates for MACS peaks.





- 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 x

pileup x

bowtie x

demo x




sample:e18 x



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 x

sample:e18 x

snps x




Annotation:

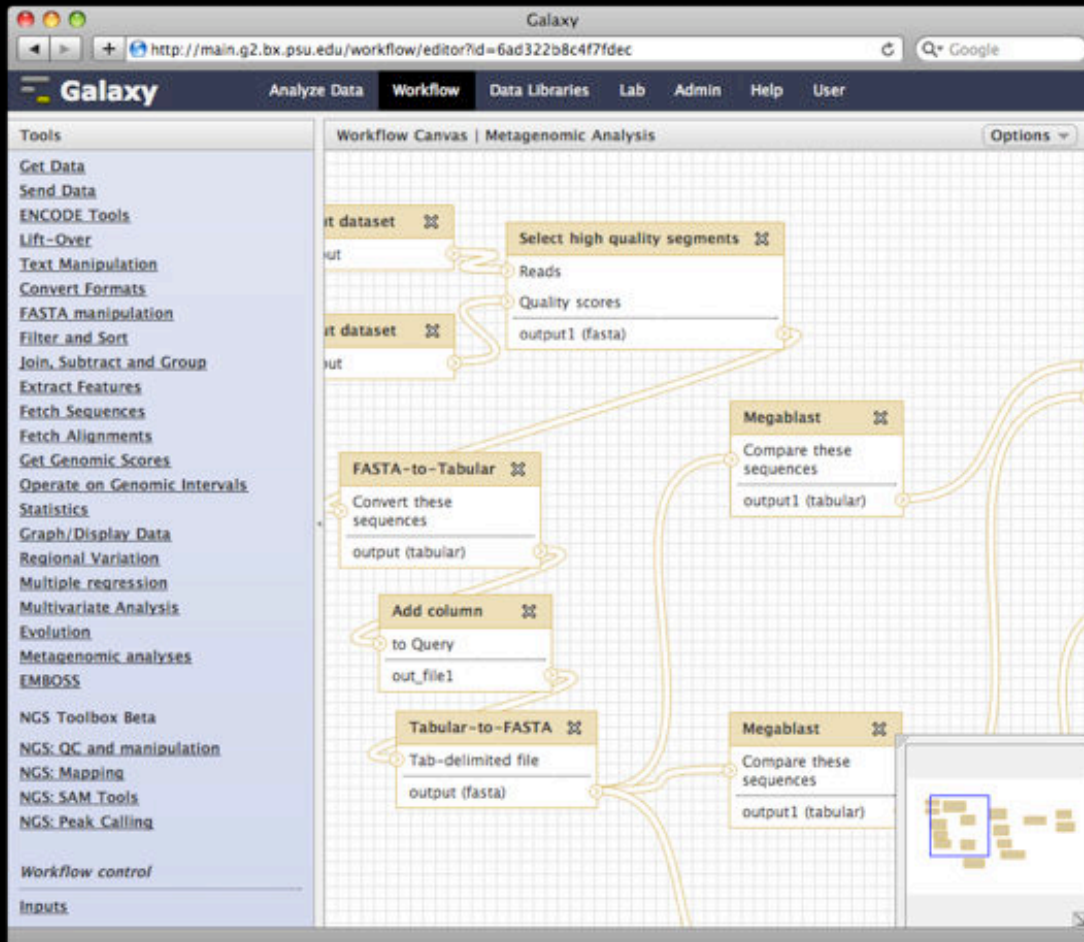
Find variants with coverage  $\geq 30$  and quality score  $\geq 20$ .

| 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 Galaxy header includes navigation links: Analyze Data, Workflow, Data Libraries, Lab, Admin, Help, and User. The page title is "Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement". Below the title, the authors are listed: SERGEI KOSAKOVSKY POND<sup>1,2\*</sup>, SAMIR WADHAWAN<sup>3,6\*</sup>, FRANCESCA CHIAROMONTE<sup>4</sup>, GURUPRASAD ANANDA<sup>1,3</sup>, WEN-YU CHUNG<sup>1,3,7</sup>, JAMES TAYLOR<sup>1,5</sup>, ANTON NEKRUTENKO<sup>1,3</sup> and THE GALAXY TEAM<sup>1\*</sup>. A note indicates correspondence should be addressed to SKP, IT, or AN. The section "How to use this document" explains that the page is a live copy of supplementary materials for a manuscript, providing access to exact analyses and workflows. It describes how to interact with the content, such as re-running analyses, changing parameters, or applying them to new data. It also mentions the ability to import items into a Galaxy workspace and start using them. A note states that to import workflows, one must create a Galaxy account (unless they already have one), which is a hassle-free procedure where only a username and password are required. The page then presents three interactive elements, each with a plus icon and a link to expand: 1. "Galaxy History | Galaxy vs MEGAN Comparison of Galaxy vs. MEGAN pipeline." 2. "Galaxy History | metagenomic analysis" 3. "Galaxy Workflow | metagenomic analysis" The "Supplemental Analysis" section is partially visible at the bottom, with a link to "Comparison between Galaxy pipeline and Megan". The footer shows the loading status: "Loading 'http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter', completed 5 of 6 items".

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



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

**GENOME  
RESEARCH**

illumina®

Apply today for the  
Cancer GWAS Grant.

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<sup>1,2,6,9</sup>, Samir Wadhawan<sup>3,6,7</sup>,  
Francesca Chiaromonte<sup>4</sup>, Guruprasad Ananda<sup>1,3</sup>, Wen-Yu Chung<sup>1,3,8</sup>,  
James Taylor<sup>1,5,9</sup>, Anton Nekrutenko<sup>1,3,9</sup> and The Galaxy Team<sup>1</sup>

### 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)



### 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.q2.bx.psu.edu/u/aun1/p/windshield-splatter", completed 5 of 6 items



# Give it a spin: [usegalaxy.org/galaxy101](https://usegalaxy.org/galaxy101)

The screenshot shows the Galaxy 101 tutorial page. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Cloud, Help, and User. The main content area is titled "Galaxy 101: The first thing you should try" and includes a list of topics to be covered: Getting data from UCSC, Performing simple data manipulation, Understanding Galaxy's History system, Creating and editing workflows, and Applying workflows to your data. A right-click context menu is open over the "Open Link in New Window" option. The right sidebar shows the author "aun1", related pages, a rating of 5 stars, and tags for "tutorial", "exons", and "snps".

**Galaxy 101: The first thing you should try**

In this very simple example we will introduce you to bare basics of Galaxy:

- Getting data from UCSC
- Performing simple data manipulation
- Understanding Galaxy's History system
- Creating and editing workflows
- Applying workflows to your data

You can watch a step-by-step explanation of this entire tutorial [here](#).

**What are we trying to do?**

Suppose you get the following question: "Mom (or Dad) ... Which coding exon has the highest number of single nucleotide polymorphisms on chromosome 22?". You think to yourself "Wow! This is a simple question ... I know exactly where the data is (at UCSC) but how do I actually compute this?" The truth is, there is really no straightforward way of answering this question in a time frame comparable to the attention span of a 7-year-old. Well ... actually there is and it is called Galaxy. So let's try it...

**0. Organizing your windows and setting up Galaxy account**

**0.0. Getting your display sorted out**

To get the most of this tutorial open two browser windows. One you already have (it is this page). To open the other, right click [this link](#) and choose "Open in a New Window" (or something similar depending on your operating system and browser):

Open Link in New Window  
Open Link in New Tab  
Download Linked File  
Download Linked File As...  
Add Link to Bookmarks...  
Copy Link

Then organize your windows as something like this (depending on the size of your monitor you may or may not be able to organize things this way, but you get the idea):

The screenshot shows two browser windows. The left window displays the "Galaxy Pages" section, and the right window displays the "Galaxy 101: The first thing you should try" page.

**0.1. Setting up Galaxy account**

Go to the **User** link at the top of Galaxy interface and choose **Register** (unless of course you already have an account):

User  
Login  
Register

Galaxy 101, is a hands-on exercise that demonstrates many Galaxy basics.

Galaxy 101 includes histories, datasets, and workflows, and is itself a *Galaxy Page*.

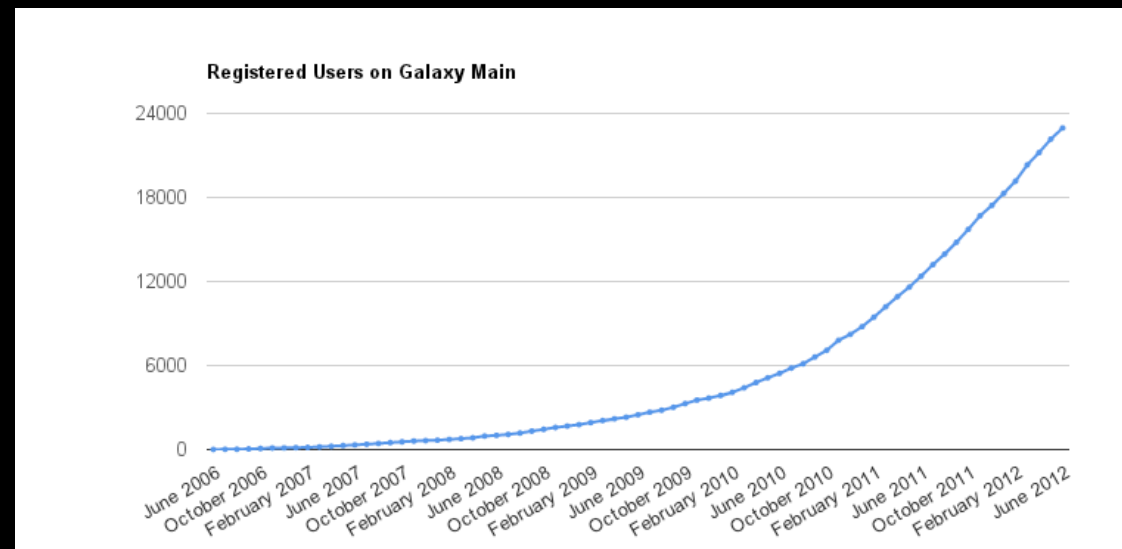
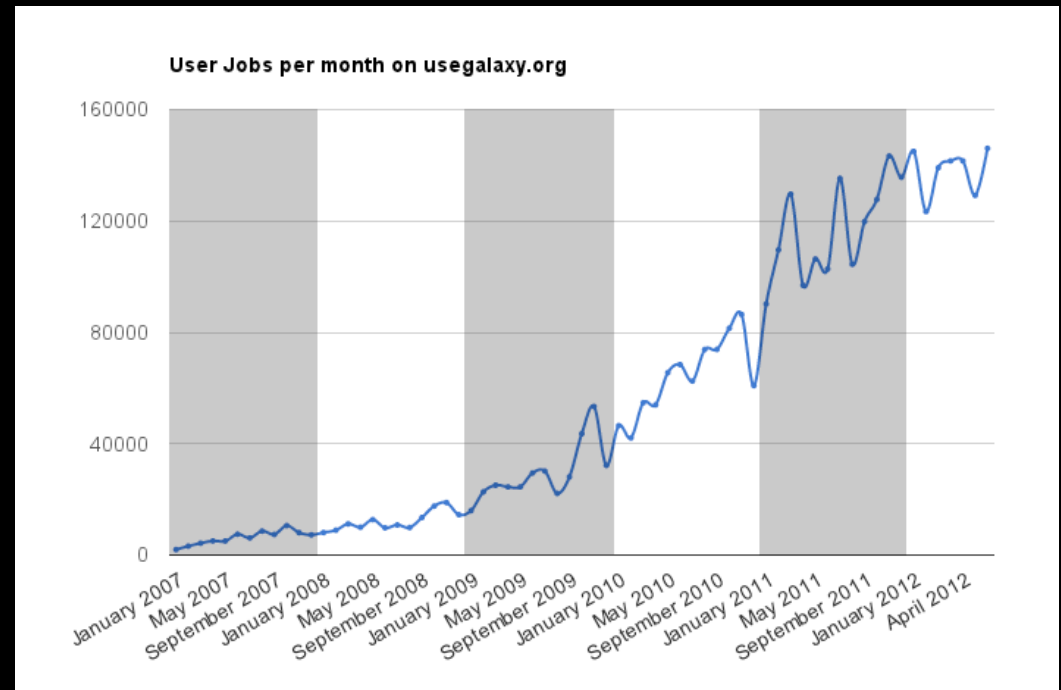
# Galaxy is available ...

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

<http://usegalaxy.org>

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

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



<http://bit.ly/gxystats>

# 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

# 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](https://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>**

# Got your own cluster?

- Move tool execution to other systems
- Galaxy works with DRMAA compliant cluster job schedulers (which is most of them).
- Galaxy is just another client to your scheduler.



# 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'll use Amazon for the *Galaxy for Biologists* workshop later today.**



<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](#).

### 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](#)

### AWS

- [Get Started](#)
- [Capacity Planning](#)
- [AMIs](#)
- [↑ CloudMan](#)

## 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

### Step 1 Screenshots



# Instant CloudMan

The image shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. A 'Using 0%' status bar is on the right. The left sidebar has a 'Tools' section with a search bar and a list of data sources under 'Get Data'. The main content area displays 'Managing Data' with the text 'Store, Manage, and Share data with Libraries' and 'An in-depth tutorial'. A 'Live Quickies' section is visible below. The right sidebar shows '0 bytes' and a message: 'Your history is empty. Click 'Get Data' on the left pane to start'. A 'New Cloud Cluster' dropdown menu is open from the 'Cloud' tab. Below this, a modal window titled 'Launch a Galaxy Cloud Instance' is shown, containing a form with fields for Cluster Name, Password, Key ID, Secret Key, and Instance Share String (optional). The 'Instance Type' is set to 'Large'. A 'Submit' button is at the bottom. A note at the bottom of the modal states: 'Requesting the instance may take a moment, please be patient. Do not refresh your browser or navigate away from the page'.

**Galaxy** Analyze Data Workflow Shared Data Visualization Cloud Help User Using 0%

Tools

search tools

**Get Data**

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- BX main browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server

**Managing Data**  
Store, Manage, and Share data with Libraries  
An in-depth tutorial

Live Quickies

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

**Launch a Galaxy Cloud Instance**

Cluster Name

Password

Key ID

Secret Key

Instance Share String (optional)

Instance Type

Large

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

Submit

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



# Galaxy Community & Resources

Mailing Lists (very active)

Screencasts

Events Calendar, News Feed

Community Wiki

CiteULike group, Mendeley mirror

Local Public Installs

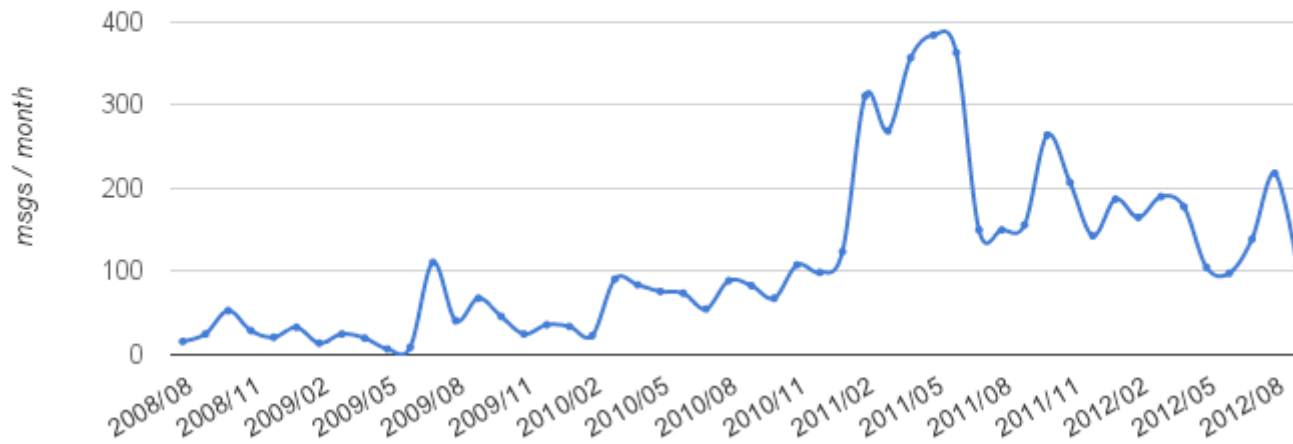
Tool Shed

Annual Community Meeting

<http://galaxyproject.org/wiki>

# Mailing Lists

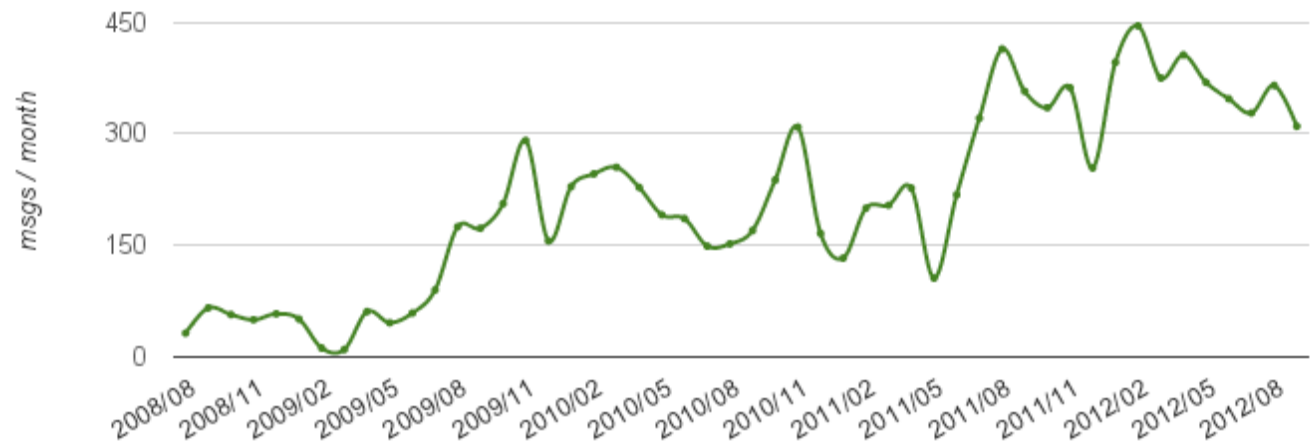
Galaxy-User monthly messages since 2008/08



Galaxy-User  
messages / month


Galaxy-Dev  
messages / month

Galaxy-Dev monthly messages since 2008/08




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

# Galaxy Search: <http://galaxyproject.org/search>

 **Galaxy Web Search**

Google™ Custom Search

Search 

Search the entire set of Galaxy web sites and mailing lists using Google.

[Run this search at Google.com \(useful for bookmarking\)](#)

Want a [different search](#)?

[Project home](#)

**Find**

Everything on ...

Tools for ...

Email about ...


Source code for ...

Published Histories, Pages, Workflows, about ...

Documentation on ...

Papers using Galaxy for ...

Related feature requests

 **Galaxy Web Search**

chip-seq

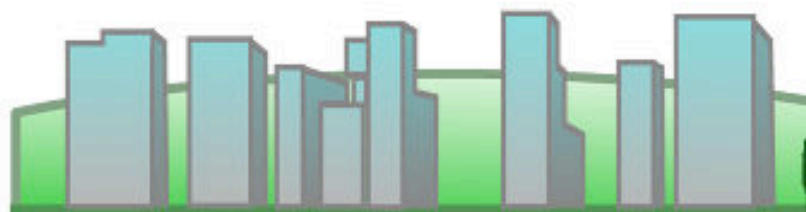
All Tools Email Source code Shared Documentation Abstracts Requests

About 444 results (0.06 seconds)

[Galaxy | Accessible Page | ChIP-seq exercise](#)

# Galaxy

Community  
Conference



OSLO



30 June  
- 2 July

2013



UiO : University of Oslo

<http://galaxyproject.org/GCC2013>

# Other Upcoming Galaxy Events



Date	Topic/Event	Venue/Location	Contact
October 15-17	<i>Advanced NGS Course: RNA-seq data analysis</i>	Amsterdam Medical Centre (AMC), The Netherlands	Patrick Koks
October 18-30	<i>Advanced Sequencing Technologies and Applications Course</i>	Cold Spring Harbor Laboratory, New York, United States	Anton Nekrutenko
October 31 - November 6	<i>Computational &amp; Comparative Genomics Course</i>	Cold Spring Harbor Laboratory, New York, United States	William Pearson, James Taylor
October 28 - November 2	<i>Genomic Virtual Laboratory Workshop</i>	eResearch Australasia, Sydney, Australia	Enis Afgan
November 6-10	<i>Galaxy 101: Data Integration, Analysis and Sharing</i>	<b>American Society of Human Genetics (ASHG)</b> , San Francisco, California, United States	Jennifer Jackson, Jeremy Goecks
	Sold out		
	<i>Working with High-Throughput Data and Data Visualization</i>		
November 12-14	<i>The Genome Access Course</i>	Cold Spring Harbor Laboratory, New York, United States	Assaf Gordon
November 13-15	<i>Analyse des données RNA-seq et ChIP-seq (séquençage haut-débit), à l'aide d'outils orientés vers un public de biologistes</i>	PRABI (Pôle Rhône-Alpes de Bioinformatique), Doua de l'Université Claude Bernard - Lyon, Lyon, France	Guy Perrière
January 14-18	<b>Plant and Animal Genome (PAG 2013)</b>	San Diego, California, United States	Dave Clements
March 2-5	<i>W6: Community Resource Solutions to Analyzing</i>	<b>ABRF 2013</b>	Dave Clements

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

# Visualization

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

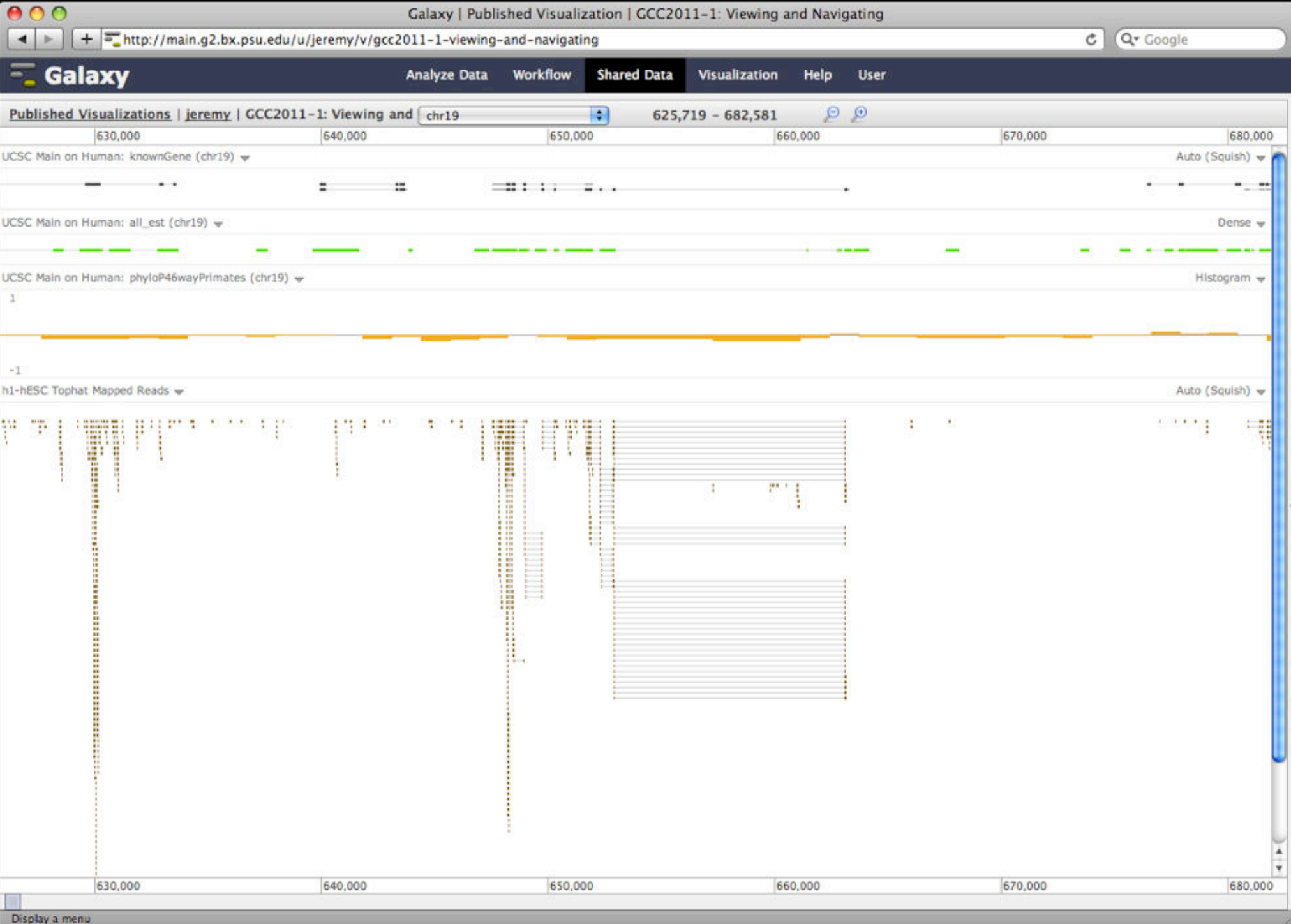
## Supports common filetypes

- ✦ BAM, BED, GFF/GTF, WIG

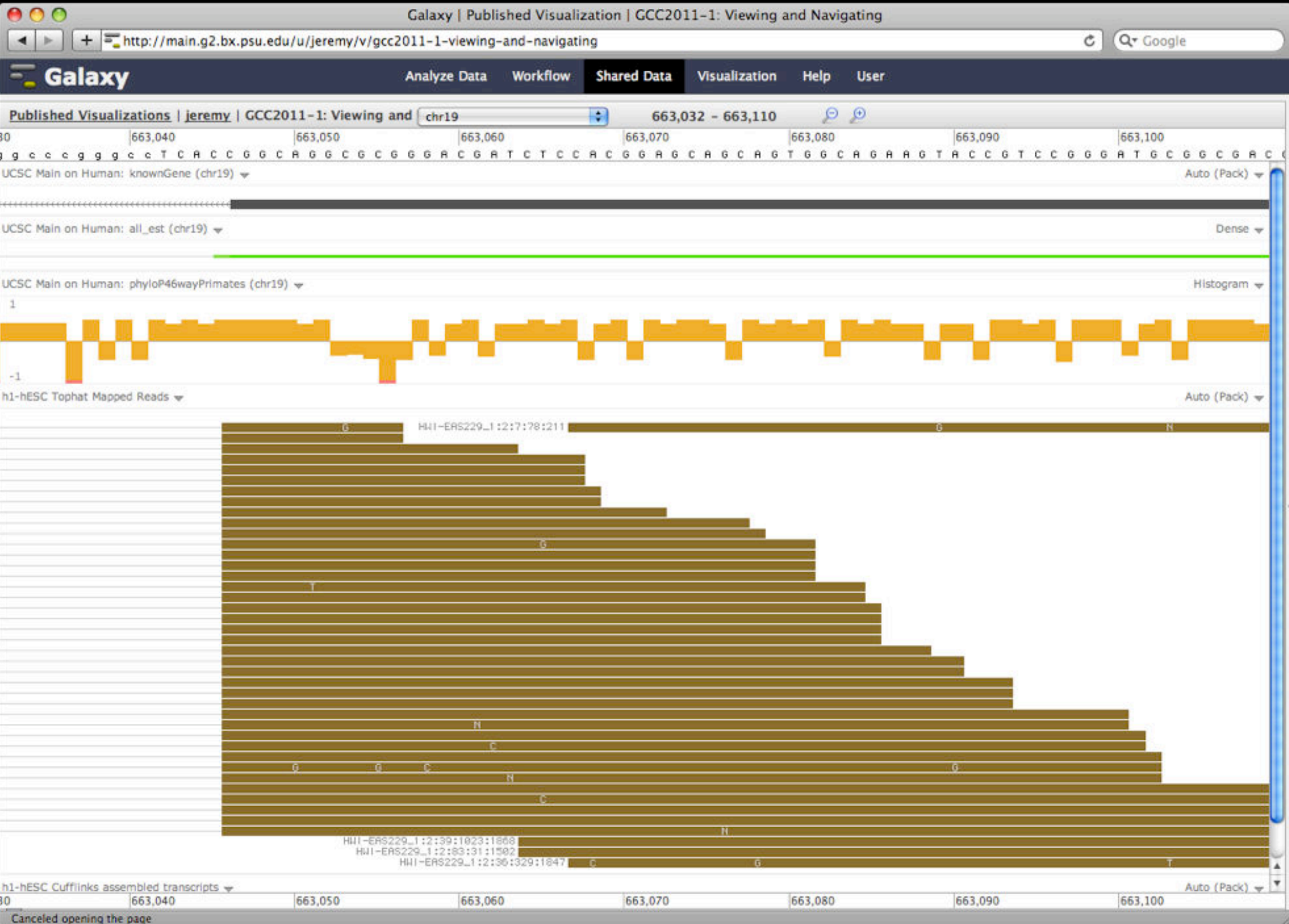
## Unique features

- ✦ custom genomes
- ✦ highly interactive

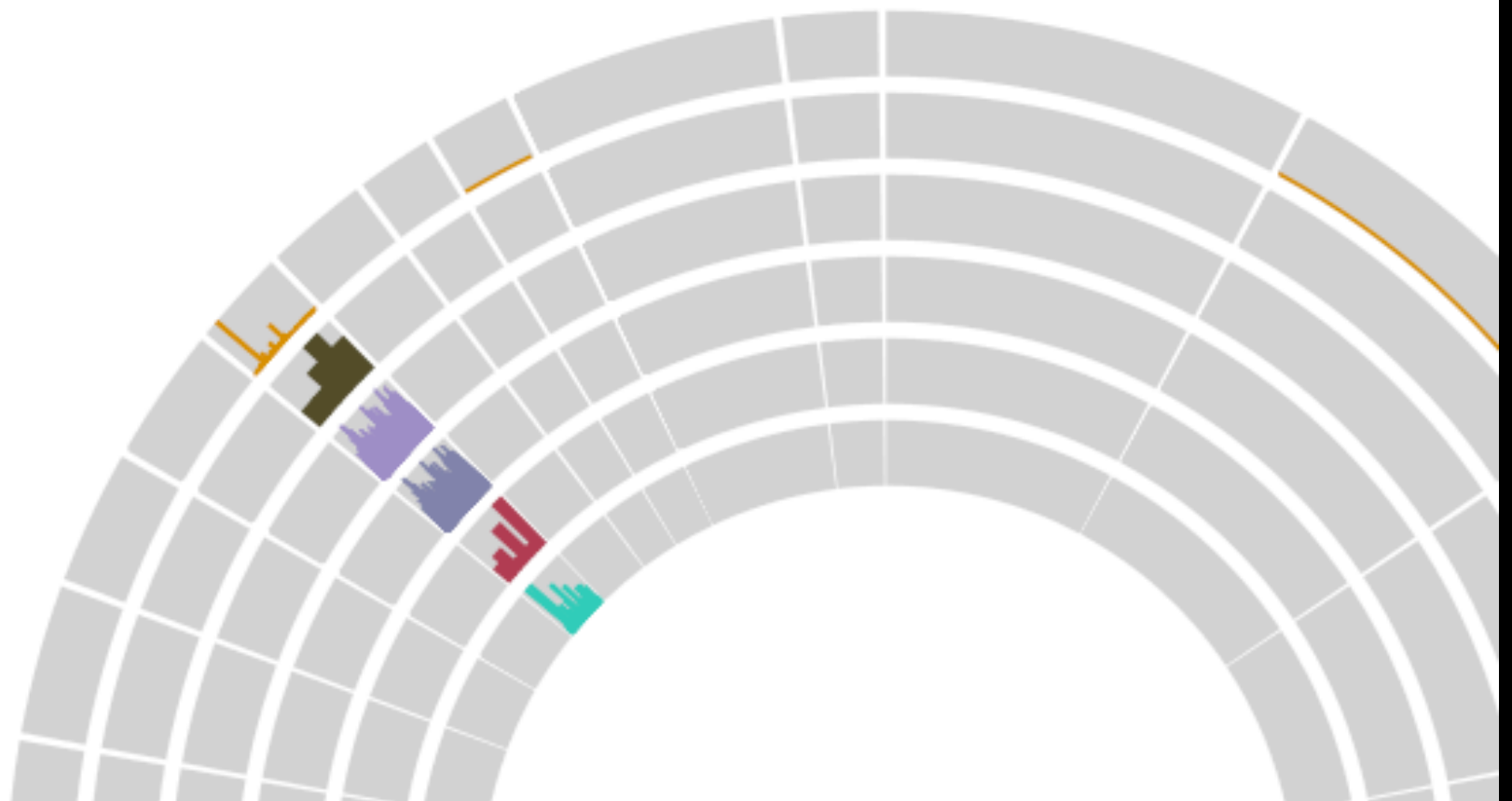








Brain / Adrenal Chr19 (hg19)



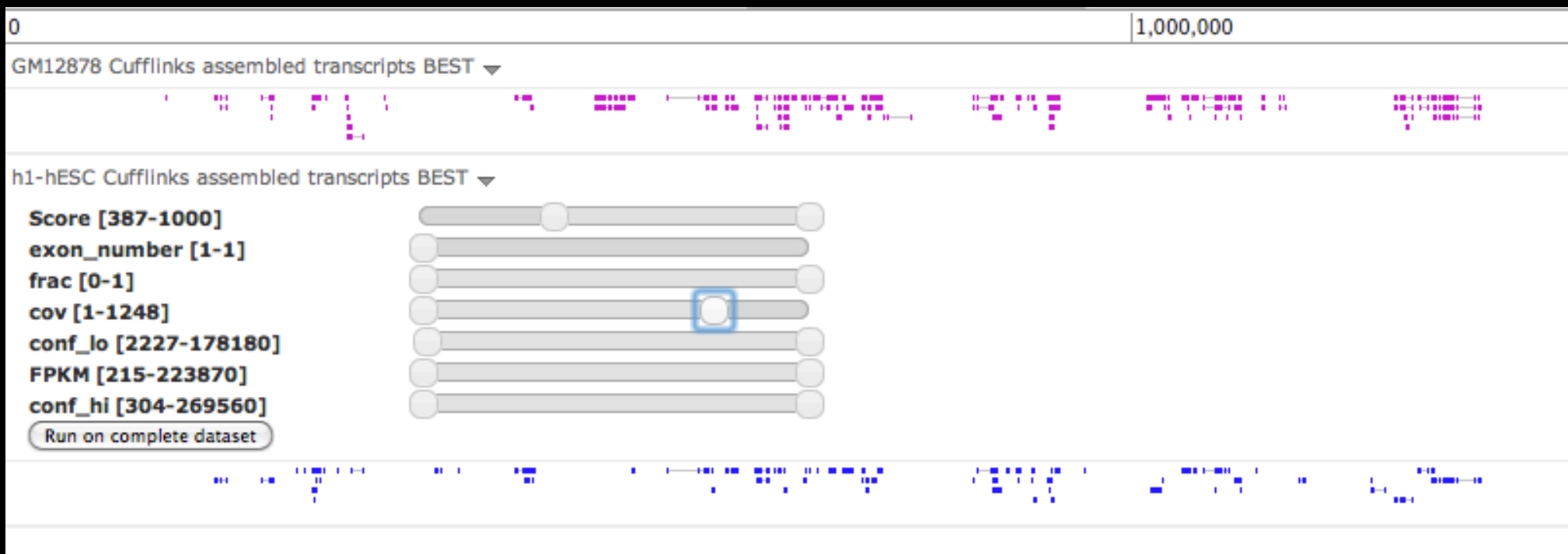
# But really, why *another* genome browser

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

Brain / Adrenal Chr19 (hg19) chr19 3,165,571 – 3,337,978 3,200,000

Tool

|||| Cufflinks assembled transcripts for Brain - region=[all], parameters=[300000, 0.1, 0.15, No] [v] [–] [↓] [⚙] [↕] [↗] [✕]

### Cufflinks

Max Intron Length 300000

Min Isoform Fraction 0.1

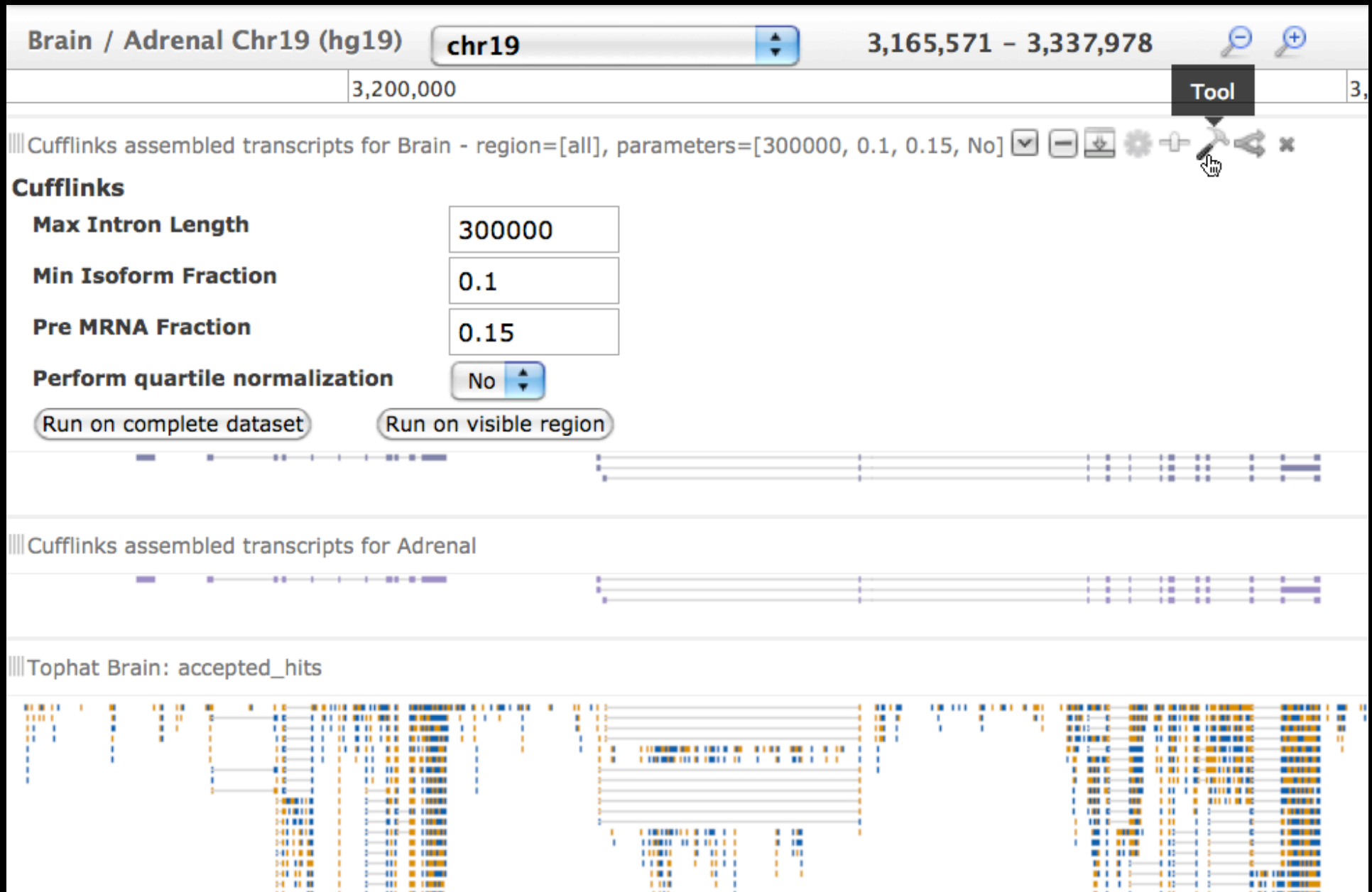
Pre mRNA Fraction 0.15

Perform quartile normalization No [v]

Run on complete dataset Run on visible region

|||| Cufflinks assembled transcripts for Adrenal

|||| Tophat Brain: accepted\_hits



Galaxy Analyze Data Workflow Shared Data Visualization Close

Published items | jeremy | Trackster Demo 2 chr19 1,549,354 - 1,691,104 1,600,000

GM12878 Cufflinks assembled transcripts BEST

h1-hESC Cufflinks assembled transcripts BEST

Tool parameter space visualization

**Galaxy**

Analyze Data Workflow Shared Data Visualization Cloud Help User

---

### Cufflinks (version 0.0.5)

**Max Intron Length:**  
 200000 -  400000 samples:  3

**Min Isoform Fraction:**  
 0.1 -  0.2 samples:  3

**Pre MRNA Fraction:**  
 0.15

**Perform quartile normalization:**  
No, Yes

**Use multi-read correct:**

---

**Getting Started**

- Create a parameter tree by using the icons next to the tool's parameter names to add or remove parameters.
- Adjust the tree by using parameter inputs to select min, max, and number of samples
- Run the tool with different settings by clicking on tree nodes

```

graph LR
    Root((Root)) --- M1(200000)
    Root --- M2(300000)
    Root --- M3(400000)
    M1 --- N1_1((No))
    M1 --- N1_2((Yes))
    M2 --- N2_1((No))
    M2 --- N2_2((Yes))
    M3 --- N3_1((No))
    M3 --- N3_2((Yes))
    N1_1 --- P1_1((0.1))
    N1_1 --- P1_2((0.15))
    N1_1 --- P1_3((0.2))
    N1_2 --- P2_1((0.1))
    N1_2 --- P2_2((0.15))
    N1_2 --- P2_3((0.2))
    N2_1 --- P3_1((0.1))
    N2_1 --- P3_2((0.15))
    N2_1 --- P3_3((0.2))
    N2_2 --- P4_1((0.1))
    N2_2 --- P4_2((0.15))
    N2_2 --- P4_3((0.2))
    N3_1 --- P5_1((0.1))
    N3_1 --- P5_2((0.15))
    N3_1 --- P5_3((0.2))
    N3_2 --- P6_1((0.1))
    N3_2 --- P6_2((0.15))
    N3_2 --- P6_3((0.2))
  
```

chr19:1549354-1691104

# Galaxy URLs to Remember

<http://galaxyproject.org>

<http://usegalaxy.org>

<http://getgalaxy.org>

<http://usegalaxy.org/galaxy101>

and

<https://galaxy.indiana.edu/>



# Galaxy at Indiana University

- Backend
  - Runs on UITS Supercomputers
  - Supported by the National Center for Genome Analysis Support
- Requests for new bioinformatic tools addressed within 48 hours
- Use your IU user name and password

**Galaxy.Indiana**

Analyze Data Workflow Shared Data Visualization Admin Help User Using 75.1 MB

**Tools**

- Import Data
- Sequence QC
- De novo Assembly
  - Trinity De novo assembly of RNA-Seq data
- Assembly QC
- Annotation and Gene Finding
- NCBI Blast+
- Workflows
- Report a Broken Tool

**History**

1. QNAME	2. FLAG	3.
r770 89 ref	116 3:	
r770 181 ref	116 0	
r1945 177 ref	41718988	
r3671 117 ref	198342418	
r3671 153 ref	198342418	
r3824 117 ref	88324999	

37: sam\_dataset.sam 12 lines format: sam, database: 2 Info: None

36: Trinity on data 33 and data 34: Assembled Transcripts

35: Trinity on data 33 and data 34: loc

34: DmeS00m\_2fc.fq

33: DmeS00m\_1fc.fq

32: modENCODE fly on 28-8445093..8498049

**WELCOME**

Welcome to the Galaxy Instance at Indiana University

Thank you for choosing Galaxy! NCGAS is committed to providing support for Indiana University research. Don't hesitate to [contact us](#) if you find that you need a tool that is not supported by our current Galaxy or if you have questions or suggestions. When possible, we will have the requested tool up and running in two working days time, and failing that, will report the status of the request within that time.

This instance of the Galaxy is installed and maintained by National Center for Genome Analysis Support [NCGAS](#)

The Computing power is provided by the Indiana University [Mason Compute Cluster](#)

The storage is provided by the Indiana University [Data Capacitor](#)

The web server is hosted on the Indiana University [Quarry Gateway Hosting](#)

The Galaxy project is supported in part by [NSF](#), [NHGRI](#), and the [Huck Institutes of the Life Sciences](#).

The [NCGAS](#) projects is supported by [NSF](#)



PERVASIVE TECHNOLOGY  
INSTITUTE

INDIANA UNIVERSITY



Thank you.



# Galaxy Community Conference

30 June  
- 2 July

2013



OSLO



UiO • University of Oslo

<http://galaxyproject.org/GCC2013>