

# Introduction to Galaxy

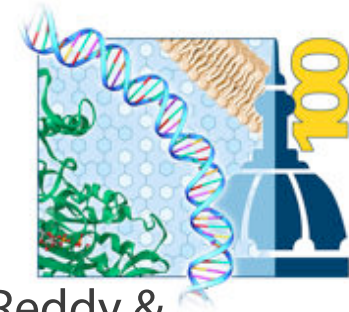
---

Johns Hopkins University  
12 April 2012

Dave Clements, Emory University  
<http://galaxyproject.org/>



Salzberg Lab



Reddy &  
Sollner-Webb Labs



The Galaxy logo, which features a stylized icon of three horizontal bars of increasing height to the left of the word "Galaxy" in a bold, white sans-serif font.

# Agenda

- 9:00 **Welcome, Intro**
- 9:20 Basic Analysis
- 10:20 Break
- 10:45 NGS Analysis I
- 12:00 Lunch
- 12:45 NGS Analysis II
- 1:45 Visualization and Visual Analytics
- 2:05 Persistence, Workflows, Sharing & Publishing
- 2:30 Break
- 3:00 Do your own analysis
- 4:30 Done

# Goals for this workshop

1. Introduce Galaxy
2. Hands-on experience:
  - Load and integrate data from online resources
  - Perform bioinformatics analysis with Galaxy
  - Save, share, describe and publish your analysis
  - Visualize your results

**This workshop will not cover** details of how the tools are implemented or new algorithm designs or which assembler or mapper or ... is best for you.

# The Motivation Slide



# Next Generation Genomics: World Map of High-throughput Sequencers

Nick Loman, James Hadfield

<http://pathogenomics.bham.ac.uk/hts/>

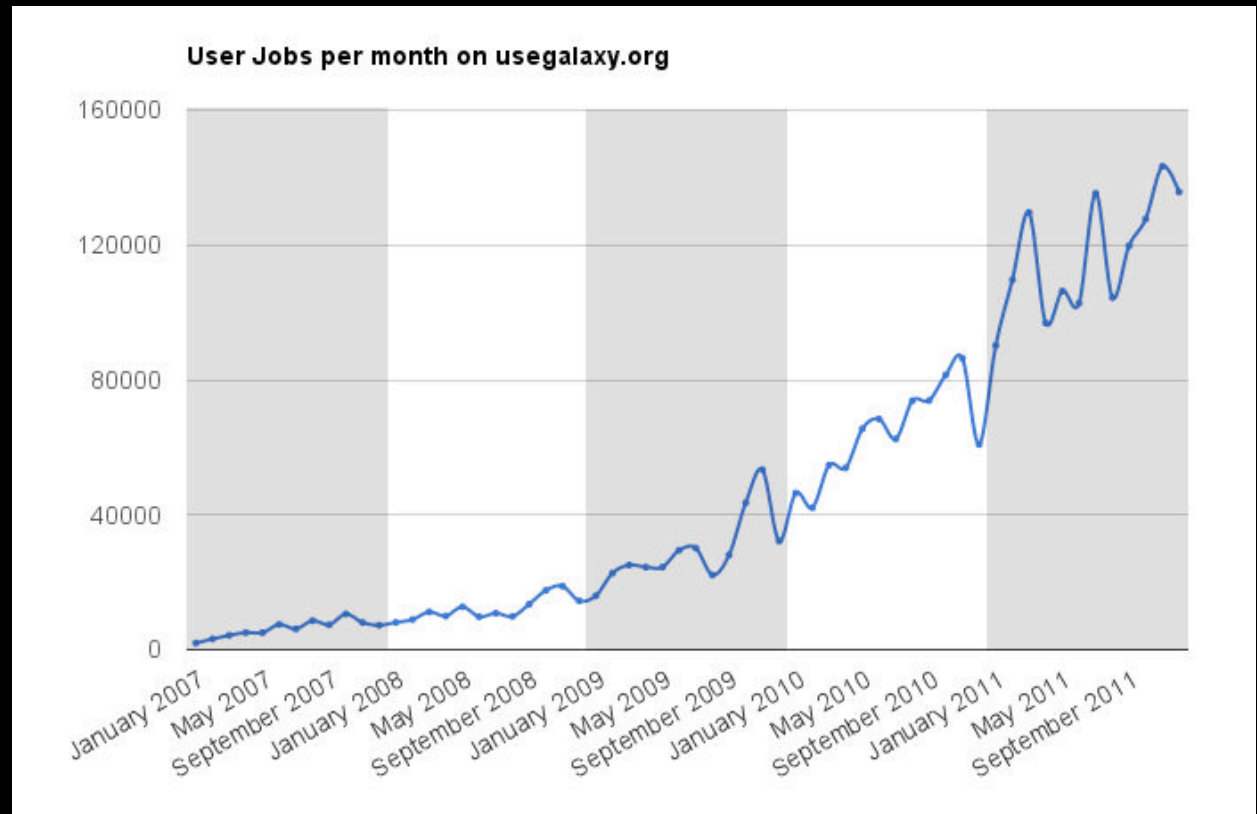
# What is Galaxy?

- An **data analysis and integration** tool
- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
- These options result in several **ways to use Galaxy**

<http://galaxyproject.org>

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

- **Public web site**
- **Anybody can use it**
- Hundreds of tools
- **Persistent**
- + 500 users / month
- ~100 TB of user data
- ~140,000 analysis 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**

# Scaling Galaxy

- Encourage local Galaxy instances and Galaxy on the cloud
- 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



# Local Galaxy Instances

<http://getgalaxy.org>

Galaxy is designed for local installation and customization

- Easily integrate new tools
- Easy to deploy and manage on nearly any (Unix) system

# 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

# Got your own cluster?

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



# 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 are using this today**



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

# Galaxy Community

Annual Community Meeting

Tool Shed

Mailing Lists (very active)

Screencasts

Events Calendar, News Feed

Community Wiki

Local Public Installs

CiteULike group, Mendeley mirror

<http://galaxyproject.org/wiki>



<http://galaxyproject.org/GCC2012>



New **Training Day** added July 25  
7 topics, 3 parallel tracks, 12 sessions

1. Intro
2. Installing
3. CloudMan
4. Integrating Tools & Sources
5. API
6. Tool Shed
7. Ion Torrent SDK



Key Dates

**April 16: Abstracts due**

June 11: Early registration ends  
(early reg is *cheap*)

# Galaxy URLs to Remember

<http://galaxyproject.org>

<http://usegalaxy.org>

<http://getgalaxy.org>

# Agenda

9:00 Welcome, Intro

9:20 **Basic Analysis**

10:20 Break

10:45 NGS Analysis I

12:00 Lunch

12:45 NGS Analysis II

1:45 Visualization and Visual Analytics

2:05 Persistence, Workflows, Sharing & Publishing

2:30 Break

3:00 Do your own analysis

4:30 Done



# Hands On: Basic Analysis

On pig chromosome 18,  
which coding exons have the most  
repeats in them?

<http://cloud1.galaxyproject.org> (gold)

<http://cloud2.galaxyproject.org> (sable)

# Repetitious Pigs: A Rough Plan

- Get some data
  - Coding exons on chromosome 18
  - Repeats on chromosome 18
- Mess with it
  - Identify which exons have repeats
  - Count repeats per exon
  - Save, download, ... exons with most repeats.

(~ <http://usegalaxy.org/galaxy101> )

# Agenda

- 9:00 Welcome, Intro
- 9:20 Basic Analysis
- 10:20 **Break**
- 10:45 NGS Analysis I
- 12:00 Lunch
- 12:45 NGS Analysis II
- 1:45 Visualization and Visual Analytics
- 2:05 Persistence, Workflows, Sharing & Publishing
- 2:30 Break
- 3:00 Do your own analysis
- 4:30 Done



# Agenda

- 9:00 Welcome, Intro
- 9:20 Basic Analysis
- 10:20 Break
- 10:45 **NGS Analysis I**
- 12:00 Lunch
- 12:45 NGS Analysis II
- 1:45 Visualization and Visual Analytics
- 2:05 Persistence, Workflows, Sharing & Publishing
- 2:30 Break
- 3:00 Do your own analysis
- 4:30 Done

# RNA-seq Exercise

<http://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise>

<http://bit.ly/gxyRNASEX>

<http://cloud1.galaxyproject.org> (gold)

<http://cloud2.galaxyproject.org> (sable)

# RNA-seq Exercise: A Plan

- Get input datasets; hg18, will mostly map to chr19
- Look at quality
- Trim as we see fit.
- Map the reads to the human reference using Tophat
- Run Cufflinks on Tophat output to assemble reads into transcripts
- Maybe run Cuffcompare and Cuffdiff

<http://bit.ly/gxyRNASEX>

# Agenda

- 9:00 Welcome, Intro
- 9:20 Basic Analysis
- 10:20 Break
- 10:45 NGS Analysis I
- 12:00 **Lunch**
- 12:45 NGS Analysis II
- 1:45 Visualization and Visual Analytics
- 2:05 Persistence, Workflows, Sharing & Publishing
- 2:30 Break
- 3:00 Do your own analysis
- 4:30 Done

# Agenda

- 9:00 Welcome, Intro
- 9:20 Basic Analysis
- 10:20 Break
- 10:45 NGS Analysis I
- 12:00 Lunch
- 12:45 **NGS Analysis II**
- 1:45 Visualization and Visual Analytics
- 2:05 Persistence, Workflows, Sharing & Publishing
- 2:30 Break
- 3:00 Do your own analysis
- 4:30 Done



# Two RNA-seq Papers

*NATURE METHODS* | REVIEW

## Computational methods for transcriptome annotation and quantification using RNA-seq

**Manuel Garber, Manfred G Grabherr, Mitchell Guttman & Cole Trapnell**

**Affiliations | Corresponding author**

*Nature Methods* **8**, 469–477 (2011) | doi:10.1038/nmeth.1613

Published online 27 May 2011 | Corrected online **15 June 2011**

*NATURE PROTOCOLS* | PROTOCOL

## Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

**Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter**

**Affiliations | Contributions | Corresponding author**

*Nature Protocols* **7**, 562–578 (2012) | doi:10.1038/nprot.2012.016

Published online 01 March 2012

# Agenda

- 9:00 Welcome, Intro
- 9:20 Basic Analysis
- 10:20 Break
- 10:45 NGS Analysis I
- 12:00 Lunch
- 12:45 NGS Analysis II
- 1:45 **Visualization and Visual Analytics**
- 2:05 Persistence, Workflows, Sharing & Publishing
- 2:30 Break
- 3:00 Do your own analysis
- 4:30 Done

# Visualize

Send data results to **external** genome browsers

**Trackster:** Galaxy's genome browser

# External Genome Browsers

UCSC

Ensembl

GBrowse

IGV

# UCSC Genome Browser on Mouse July 2007 (NCBI37)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out

position/search chr12:57,795,963-57,815,592

gene

jump

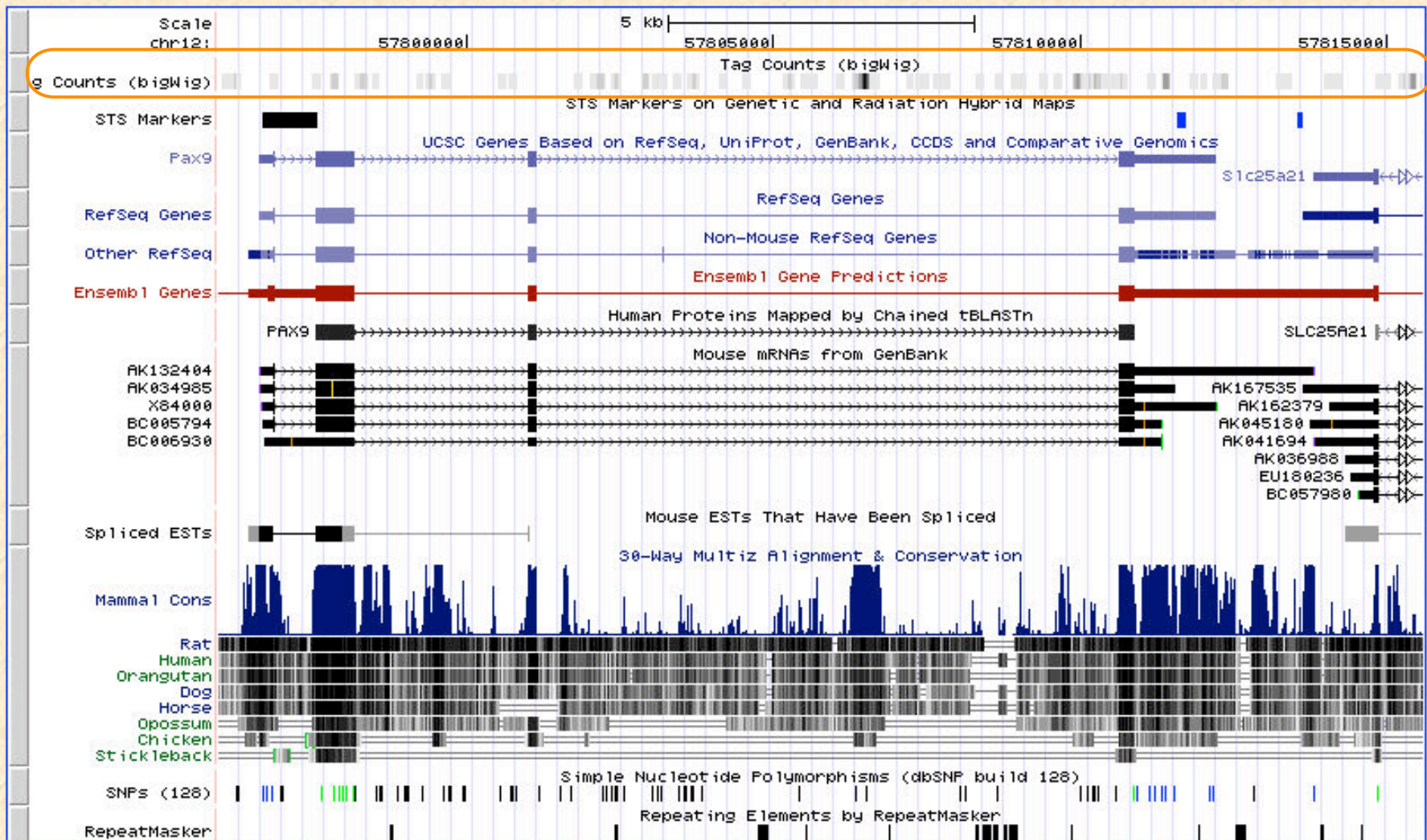
clear

size

17,000 bp

configure

chr12 (qC1) 12qA1.1 qA2 12qA3 qB1 12qB3 12qC1 12qC2 12qC3 qD1 D2 12qD3 12qE 12qF1 qF2



14: Tag Counts (bigWig)

Info

2.4 Gb, format: bigwig, database: mm9




Info:








display at UCSC main

Binary UCSC BigWig file

# Integrative Genomics Viewer (IGV)

**1: Sample data**   

1.2 Gb  
format: bam, database: mm9  
Info: uploaded bam file

display at UCSC [main](#) [test](#)  
display at Ensembl [Current](#)  
display with IGV [web](#) [local](#)

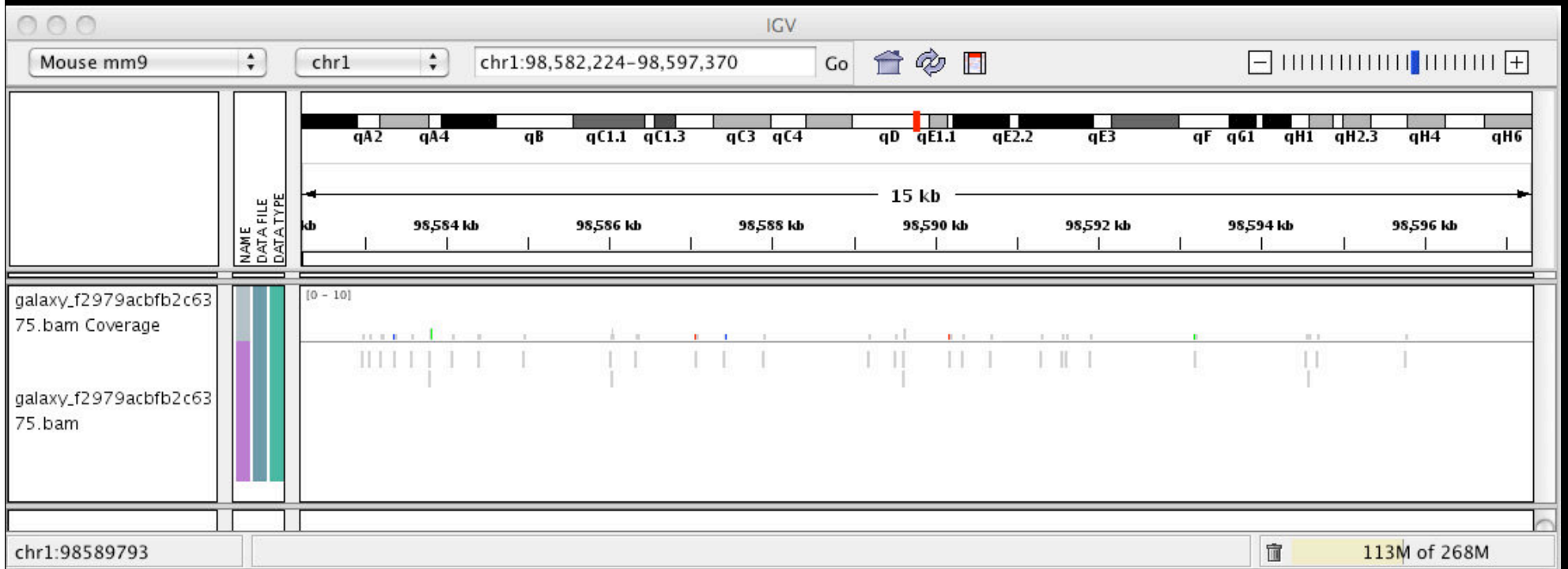
Binary bam alignments file

 **The application "IGV 1.5" from "www.broadinstitute.org" is requesting access to your computer.**

The digital signature could not be verified.

☐ Allow all applications from "www.broadinstitute.org" with this signature

 [Show Details...](#) [Deny](#) [Allow](#)



## **Galaxy**

- ✦ tool integration framework
- ✦ heavy focus on usability
- ✦ sharing, publication framework

## **Genome Browser**

- ✦ physical depiction of data
- ✦ visually identify correlations
- ✦ find interesting regions, features

## **Galaxy**

- ✦ tool integration framework
- ✦ heavy focus on usability
- ✦ sharing, publication framework

## **Genome Browser**

- ✦ physical depiction of data
- ✦ visually identify correlations
- ✦ find interesting regions, features

## **Trackster**



```
graph LR; Galaxy[Galaxy] --> Trackster[Trackster]; GenomeBrowser[Genome Browser] --> Trackster;
```

The diagram illustrates the relationship between three genomic data visualization tools. On the left, two light blue rounded rectangular boxes are stacked vertically. The top box is titled 'Galaxy' and lists three features: 'tool integration framework', 'heavy focus on usability', and 'sharing, publication framework'. The bottom box is titled 'Genome Browser' and lists three features: 'physical depiction of data', 'visually identify correlations', and 'find interesting regions, features'. On the right, a third light blue rounded rectangular box is titled 'Trackster'. Two orange curved arrows point from the right side of the 'Galaxy' box and the right side of the 'Genome Browser' box towards the 'Trackster' box, indicating that both tools feed into or are integrated with Trackster.



# 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

Published Visualizations | [Jeremy](#) | GCC2011-1: Viewing and

chr19

1,290 - 4,168,475

0 1,000,000 2,000,000 3,000,000 4,000,000

UCSC Main on Human: knownGene (chr19) ▾

Auto (Squish) ▾



UCSC Main on Human: all\_est (chr19) ▾

Auto (coverage histogram) ▾

11431



UCSC Main on Human: phyloP46wayPrimates (chr19) ▾

Histogram ▾

1



h1-hESC Tophat Mapped Reads ▾

Auto (coverage histogram) ▾

8732



h1-hESC Cufflinks assembled transcripts ▾

Auto (Squish) ▾



0 1,000,000 2,000,000 3,000,000 4,000,000

Published Visualizations | [Jeremy](#) | GCC2011-1: Viewing and

chr19

625,719 - 682,581

630,000

640,000

650,000

660,000

670,000

680,000

UCSC Main on Human: knownGene (chr19) ▾

Auto (Squish) ▾

UCSC Main on Human: all\_est (chr19) ▾

Dense ▾

UCSC Main on Human: phyloP46wayPrimates (chr19) ▾

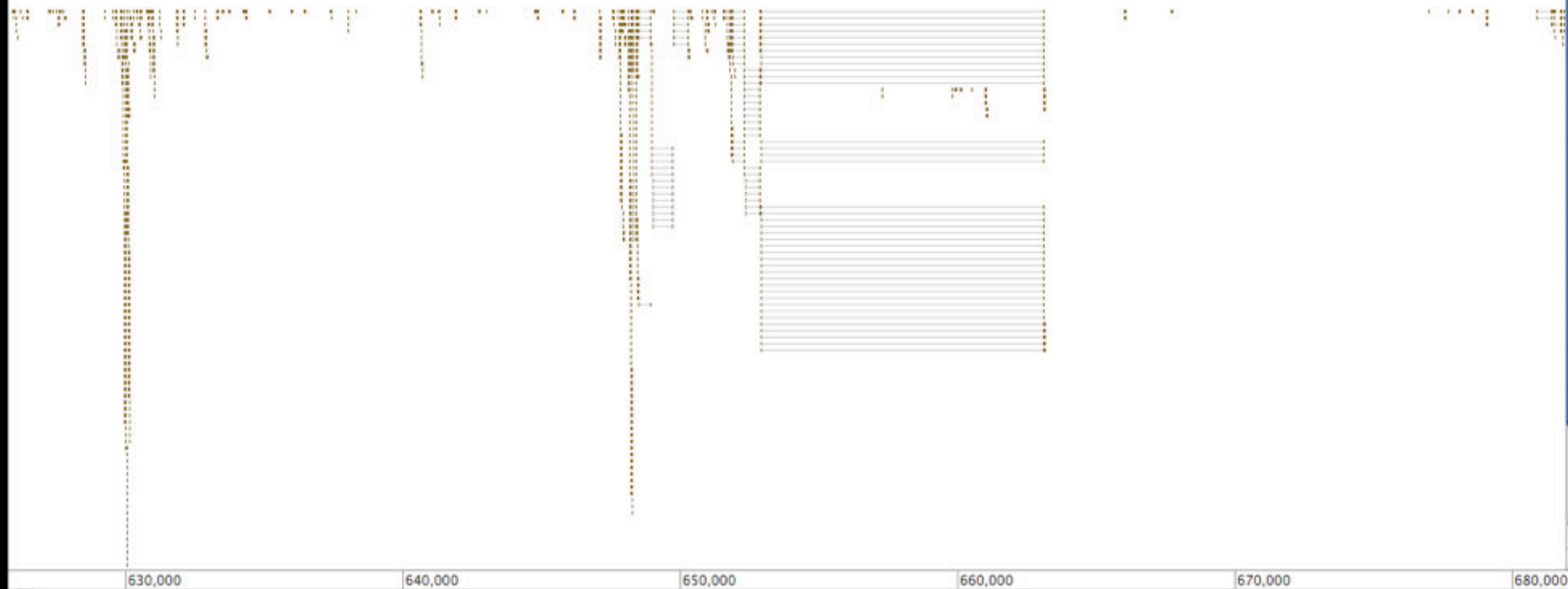
Histogram ▾

1

-1

h1-hESC Tophat Mapped Reads ▾

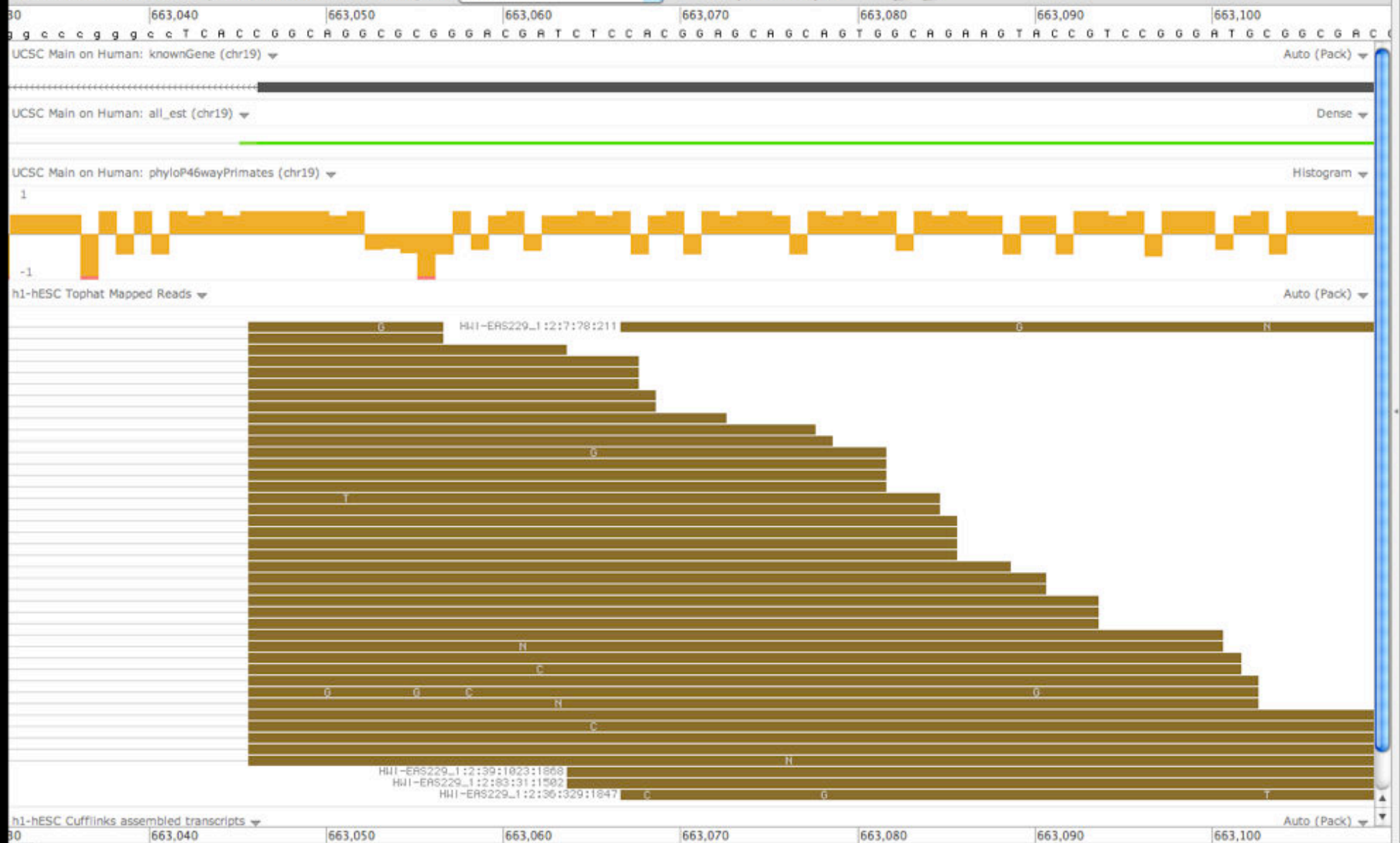
Auto (Squish) ▾



Published Visualizations | Jeremy | GCC2011-1: Viewing and

chr19

663,032 - 663,110



Canceled opening the page

# But really, why *another* genome browser

From static browsing to **visual analysis**

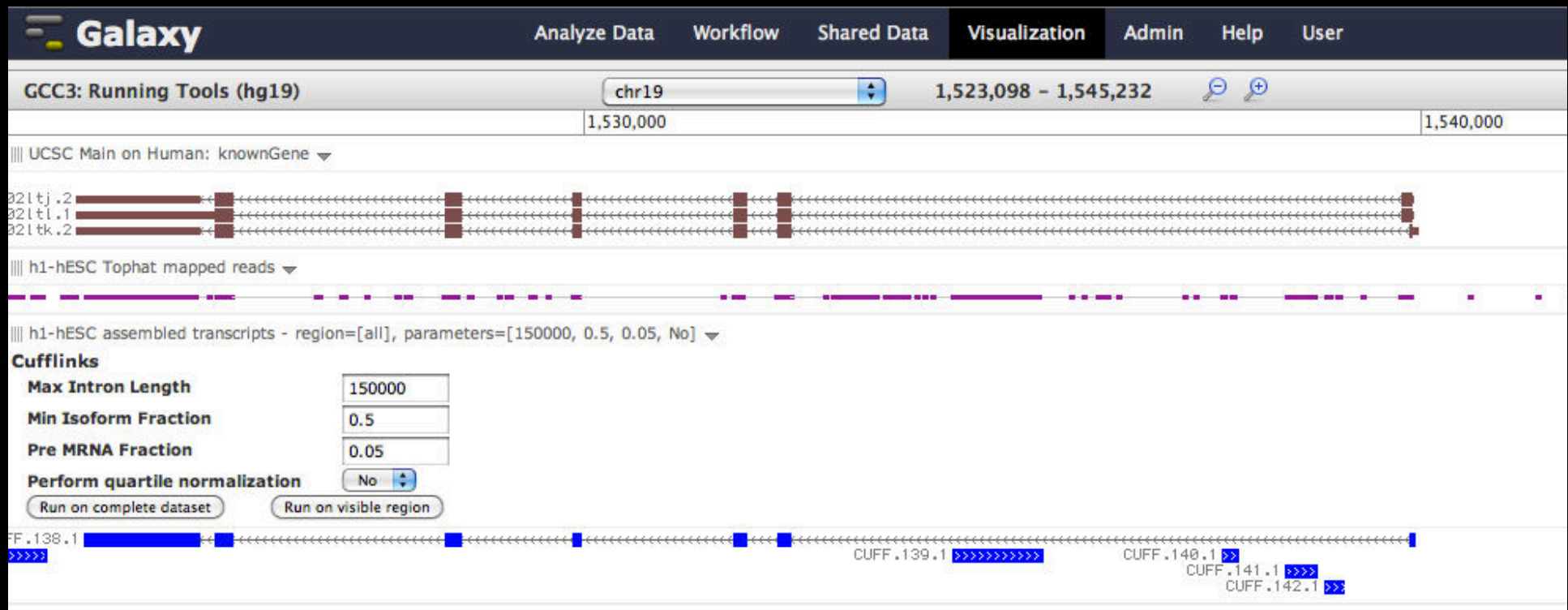
**Visual feedback and experimentation** needed for complex tools with many parameters

**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



# Agenda

- 9:00 Welcome, Intro
- 9:20 Basic Analysis
- 10:20 Break
- 10:45 NGS Analysis I
- 12:00 Lunch
- 12:45 NGS Analysis II
- 1:45 Visualization and Visual Analytics
- 2:05 **Persistence, Workflows, Sharing & Publishing**
- 2:30 Break
- 3:00 Do your own analysis
- 4:30 Done



# Some Galaxy Terminology

## **Dataset:**

Any input, output or intermediate set of data

## **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:**

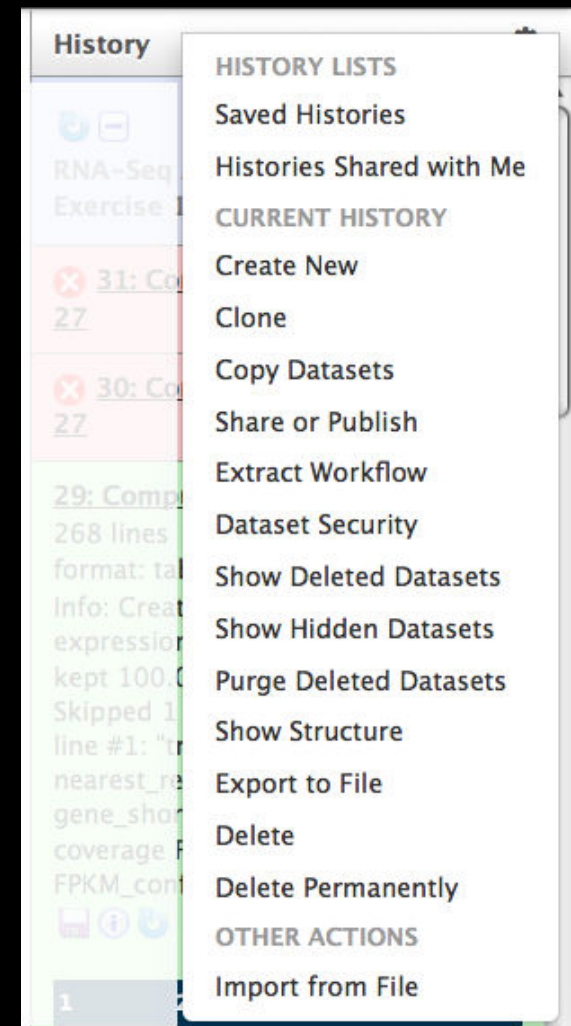
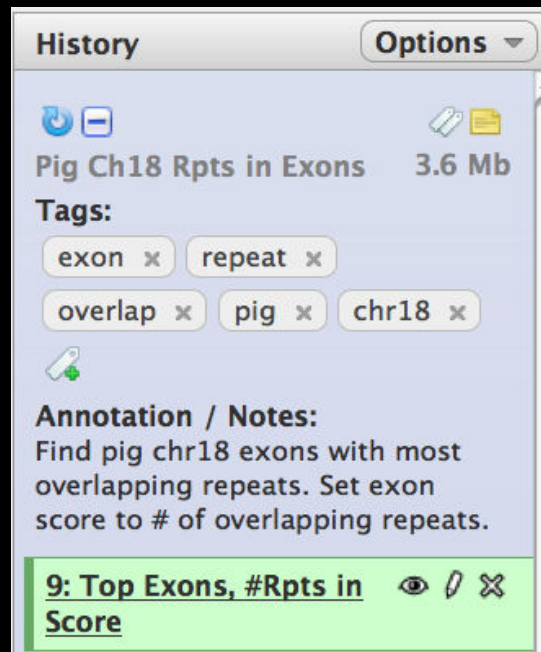
Make something available to everyone

# Managing Histories and Datasets

Give every **history**  
**and dataset**  
a **clear name**

**Datasets and**  
**histories** can also  
have annotation and tags

Each **history** has an options/actions list



# Reuse & Workflows

## Histories

Datasets from previous histories can be imported into current one.

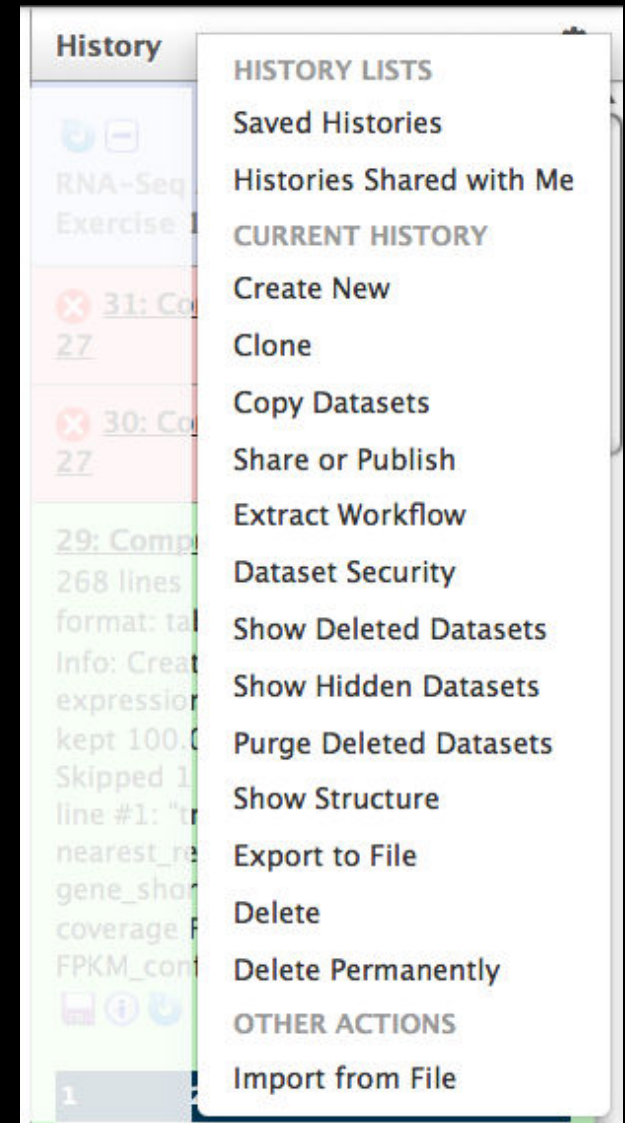
Resume any previous history

Current history can be cloned

## Workflows

Can be extracted from any history

Allows you rerun analysis with different inputs, settings



# Sharing and Publishing Your Work

The screenshot shows the Genome Research journal website. At the top, there are logos for CSH PRESS, GENOME RESEARCH, and a banner for the illumina Cancer GWAS Grant. Below the logos is a navigation bar with links: HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP. A blue bar below the navigation bar contains the text 'Institution: PENN STATE UNIV Sign In via User Name/Password' and a search bar with the text 'Search for Keyword: Go' and 'Advanced Search'. The main content area features the article title 'Windshield splatter analysis with the Galaxy metagenomic pipeline' by 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>. To the right of the article title is a section titled 'OPEN ACCESS ARTICLE' with the subheading 'This Article'. It contains the text: 'Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109', 'Copyright © 2009 by Cold Spring Harbor Laboratory Press', and two links: '» Abstract Free' and '» Full Text (PDF) Free'. To the right of the article is a section titled 'Current Issue' for October 2010, 20 (10), featuring a cover image of the journal.

**Histories, workflows, visualizations** and *pages* can be shared with others or published to the world.

<http://usegalaxy.org/u/aun1/p/windshield-splatter>

# Sharing and Publishing Your Work

The screenshot shows the top of a Genome Research article page. At the top left is the CSH PRESS logo and the 'GENOME RESEARCH' title. To the right is an Illumina banner with the text 'Apply today for the Cancer GWAS Grant.' Below the header is a navigation bar with links: HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP. A blue bar below the navigation contains the text 'Institution: PENN STATE UNIV Sign In via User Name/Password' and a search box with the text 'Search for Keyword: Go' and 'Advanced Search'.

The main content area features the article title 'Windshield splatter analysis with the Galaxy metagenomic pipeline' by 'Sergei Kosakovsky Pond<sup>1,2,6,9</sup>, Samir Wadhawan<sup>3,6,7</sup>, Fran James'. To the right of the title is a box labeled 'OPEN ACCESS ARTICLE' containing 'This Article' information: 'Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109 Copyright © 2009 by Cold Spring Harbor Laboratory Press'. Further right is a 'Current Issue' box for 'October 2010, 20 (10)' with a 'GENOME RESEARCH' journal cover image.

A large orange oval highlights a 'Footnotes' section. The first footnote reads: '[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>.]

**Histories, workflows, visualizations** and **pages** can be shared with others or published to the world.

<http://usegalaxy.org/u/aun1/p/windshield-splatter>

# 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



# Galaxy Tool Shed

- Allow users 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

<http://usegalaxy.org/community>

# Agenda

- 9:00 Welcome, Intro
- 9:20 Basic Analysis
- 10:20 Break
- 10:45 NGS Analysis I
- 12:00 Lunch
- 12:45 NGS Analysis II
- 1:45 Visualization and Visual Analytics
- 2:05 Persistence, Workflows, Sharing & Publishing
- 2:30 **Break, but first ...**
- 3:00 Do your own analysis
- 4:30 Done



# Workshop Feedback

Please help.

<http://bit.ly/jhufedback>

Try it now:  
<http://UseGalaxy.org>

Develop and deploy:  
<http://GetGalaxy.org>



Dannon Baker



Jeremy Goecks



Dave Clements



James Taylor



Enis Afgan  
(IRB)



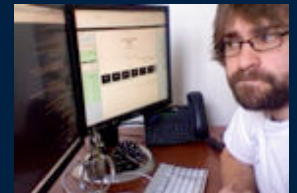
Ross Lazarus  
(Baker IDI, Harvard)



Guru Ananda



Dan Blankenberg



Nate Coraor



Jennifer Jackson



Greg von Kuster



Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

<http://GalaxyProject.org>

Try it now:  
<http://UseGalaxy.org>

Develop and deploy:  
<http://GetGalaxy.org>



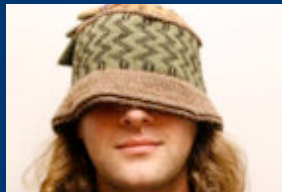
Dannon Baker



Jeremy Goecks



Ross Lazarus  
(Baker IDI, Harvard)



James Taylor



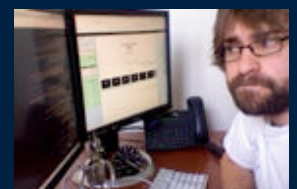
Enis Afgan  
(IRB)



Jennifer Jackson



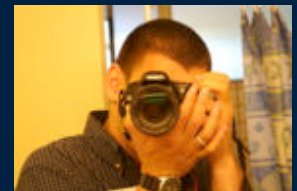
Dan Blankenberg



Nate Coraor



Greg von Kuster



Anton Nekrutenko



Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

<http://GalaxyProject.org>

# Thanks



**Mohammed Heydarian**

**Reddy Lab**

**Salzberg Lab**

**Sollner-Webb Lab**

**Mary Mangan**

(OpenHelix)

**+**

**You**

# Agenda

- 9:00 Welcome, Intro
- 9:20 Basic Analysis
- 10:20 Break
- 10:45 NGS Analysis I
- 12:00 Lunch
- 12:45 NGS Analysis II
- 1:45 Visualization and Visual Analytics
- 2:05 Persistence, Workflows, Sharing & Publishing
- 2:30 **Break, really**
- 3:00 Do your own analysis
- 4:30 Done



# Agenda

- 9:00 Welcome, Intro
- 9:20 Basic Analysis
- 10:20 Break
- 10:45 NGS Analysis I
- 12:00 Lunch
- 12:45 NGS Analysis II
- 1:45 Visualization and Visual Analytics
- 2:05 Persistence, Workflows, Sharing & Publishing
- 2:30 Break
- 3:00 **Do your own analysis**
- 4:30 Done

# Workshop Feedback

Please help.

<http://bit.ly/jhufedback>