Galaxy Workshop

Stellenbosch University
6-7 September 2012

Dave Clements
Emory University

http://galaxyproject.org/

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

amazon
webservices™

Galaxy

Acknowledgements 1

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

Stellenbosch University
University of Pretoria

Acknowledgements 2

Dannon Baker
Enis Afgan

and

the rest of the
Galaxy Team

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

Agenda: Day 1

Welcome, Basic Analysis
Basic analyses into Reusable Workflows
Galaxy Project Overview
NGS Analysis I: Through Tophat
A Simple Change ...
Persistence, Sharing, and Publishing
NGS Analysis II: Cufflinks, Cuffdiff
Visualization and visual analytics

Coffee and lunch breaks throughout the day

On Wiki: Documents/Presentations/2012_Stellenbosch...

Goals for this workshop

1. Introduce Galaxy
2. Introduce Common Bioinformatics Formats
3. 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.

Hands On: Basic Analysis

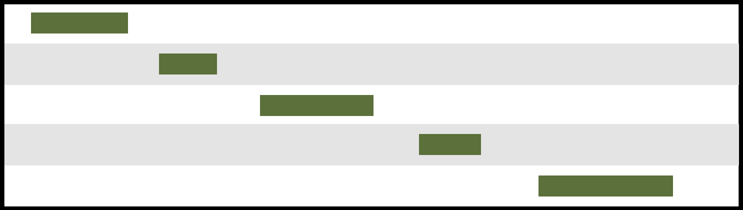
On pig chromosome 18,
which coding exons have the most
repeats in them?

http://bit.ly/SUNgray
http://bit.ly/SUNmaroon

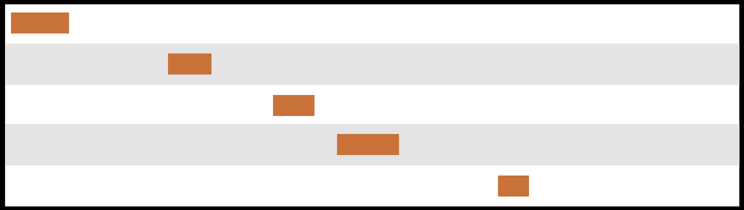
Repetitious Pigs: A Rough Plan

- Get some data (and explain BED)
  - Coding exons on chromosome 18
  - Repeats on chromosome 18
- Mess with it (and explain Galaxy operations)
  - Identify which exons have repeats
  - Count repeats per exon
- Visualize our results

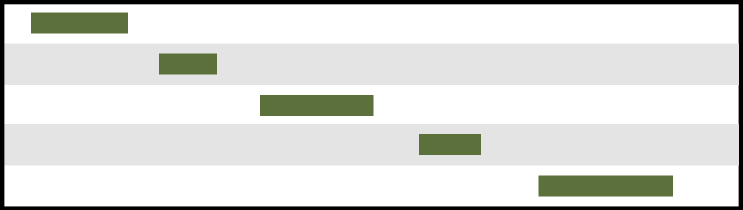
(~ http://usegalaxy.org/galaxy101)



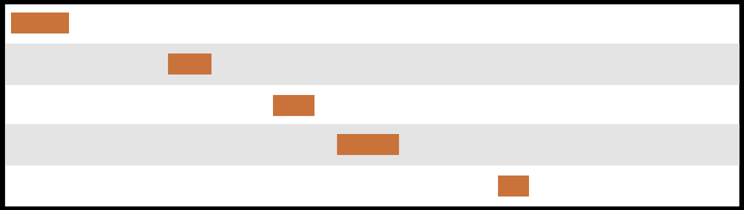
Exons, from UCSC



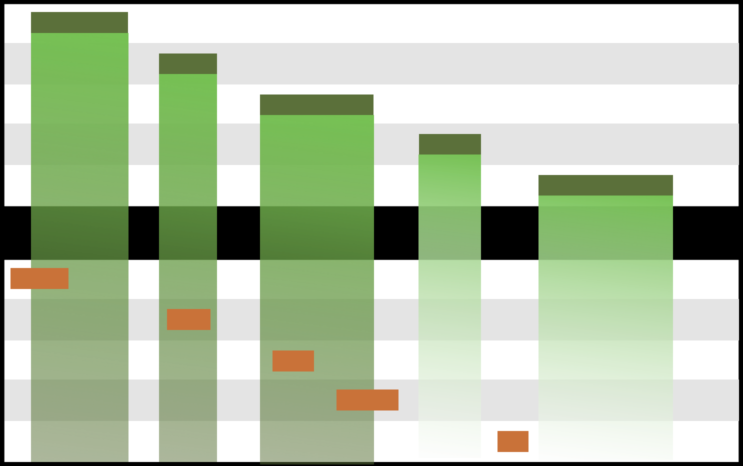
Repeats, from UCSC



Exons, from UCSC



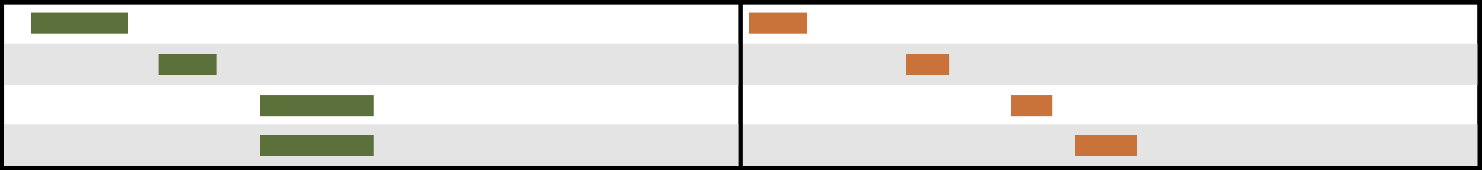
Repeats, from UCSC

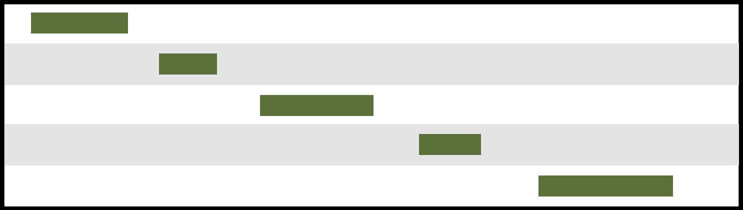


Exons, from UCSC

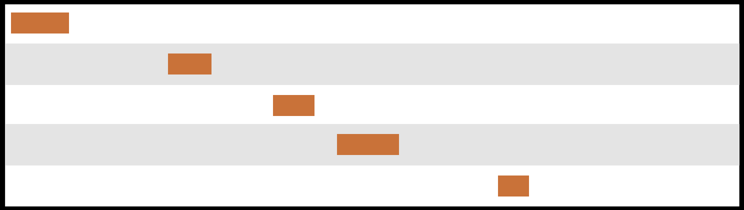
Repeats, from UCSC

Overlap pairings

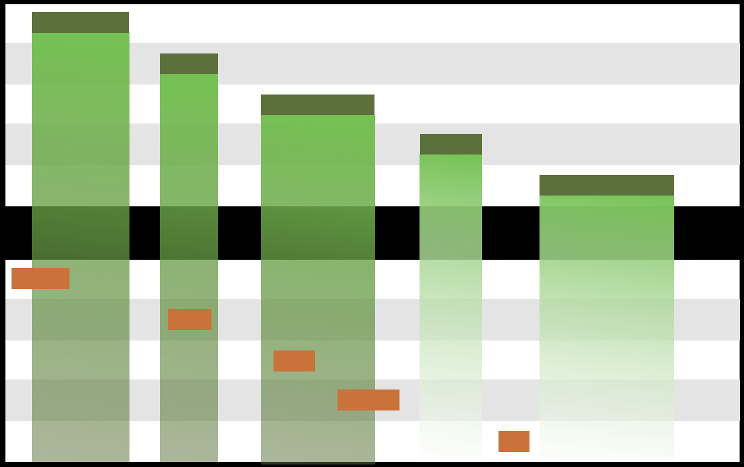




Exons, from UCSC



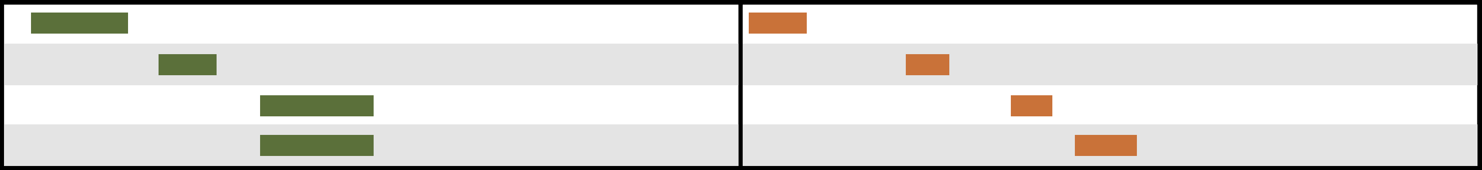
Repeats, from UCSC



Exons, from UCSC

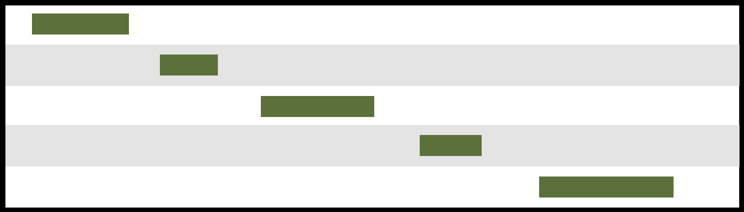
Repeats, from UCSC

Overlap pairings



Exon overlap counts

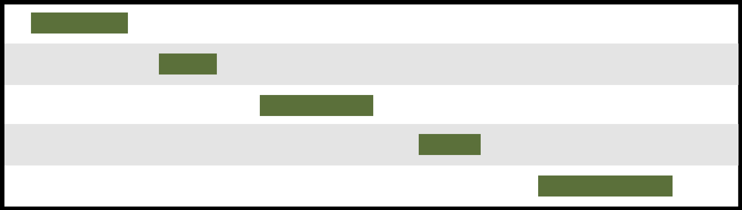
Exon overlap counts



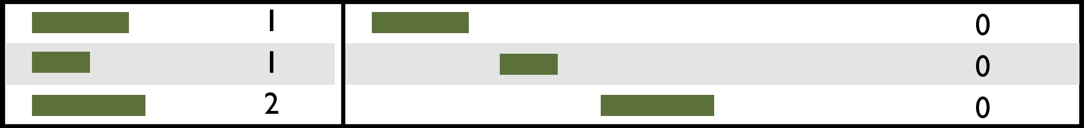
Exons, from UCSC

🟩	1
🟩	1
🟩	2

Exon overlap counts



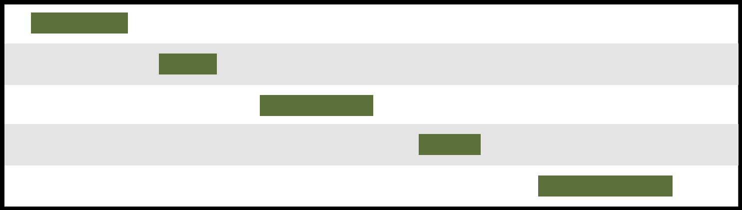
Exons, from UCSC



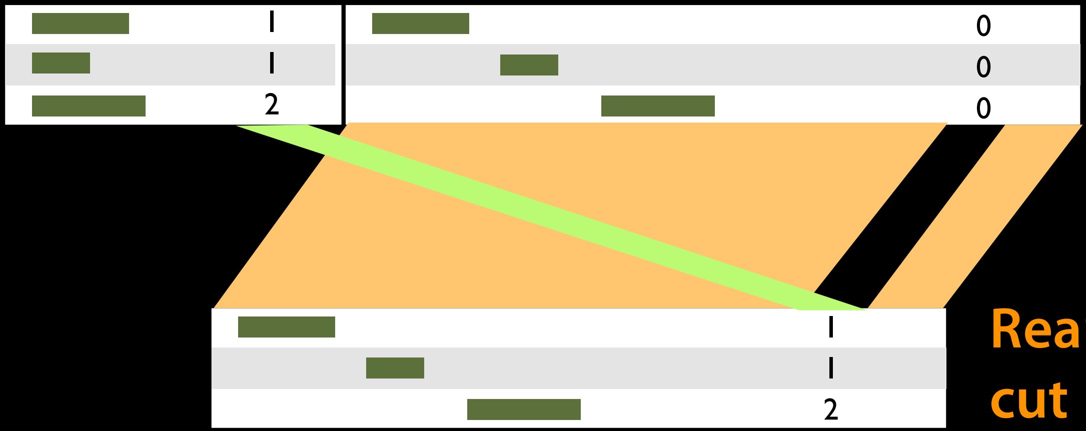
🟩	1	🟩	0
🟩	1	🟩	0
🟩	2	🟩	0

Join on exon name

Exon overlap counts



Exons, from UCSC



Join on exon name

Rearrange columns w/ cut

Agenda: Day 1

Welcome, Basic Analysis

Basic analyses into Reusable Workflows

Galaxy Project Overview

NGS Analysis I: Through Tophat

A Simple Change ...

Persistence, Sharing, and Publishing

NGS Analysis II: Cufflinks, Cuffdiff

Visualization and visual analytics

Coffee and lunch breaks throughout the day

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

Reuse: Data & Analyses

Histories: Data

Datasets from previous histories can be imported into current one.

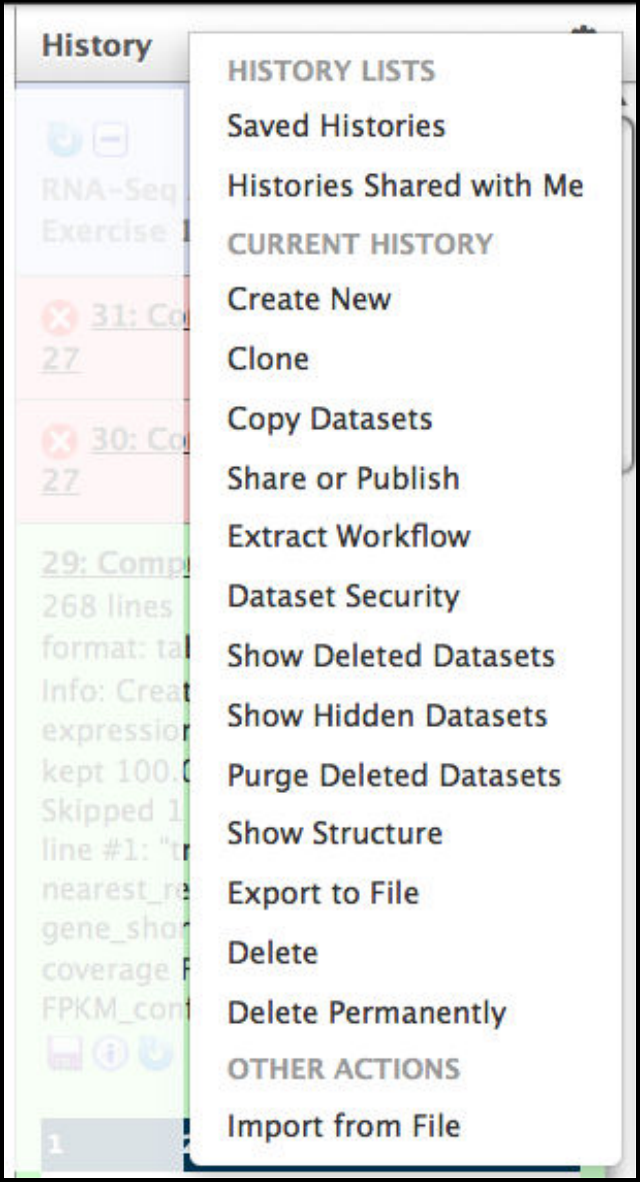
Resume any previous history

Current history can be cloned

Workflows: Analyses

Can be extracted from any history

Allows you rerun analysis with different inputs, settings



Repetitious Pigs *History* → Reusable *Workflow?*

- The analysis we just finished was about
  - Pig chromosome 18
  - Overlap between exons and repeats
- But, ...
  - there is nothing inherently in the analysis about pigs, chromosomes, exons or repeats
  - It is a series of steps that sets the score of one set of features to the number of overlaps from another set of features.

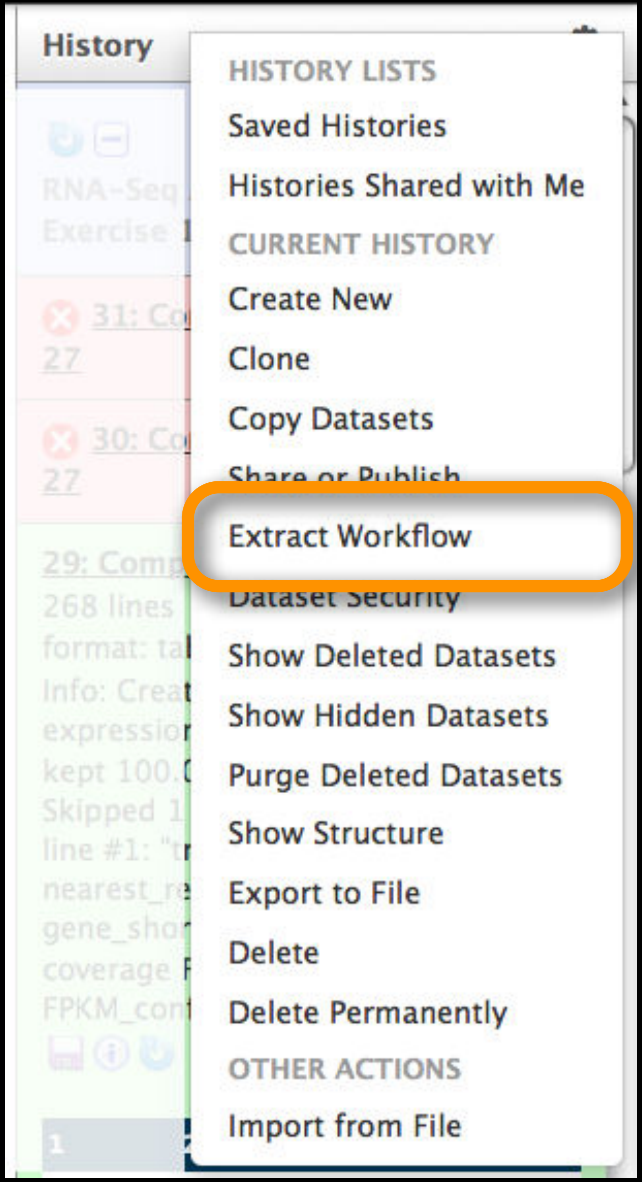
Reuse: Create a generic *Overlap* Workflow

Extract Workflow from history

- Create a workflow from this history.
- Edit it to make some things clearer.

Run / test it

- Guided: rerun with same inputs
- On your own:
  - Count # CpG islands in each exon
  - Did that work?
- On your own:
  - Count # of exons in each repeat
  - Did that work? *Why not?*
  - Edit workflow: doc assumptions

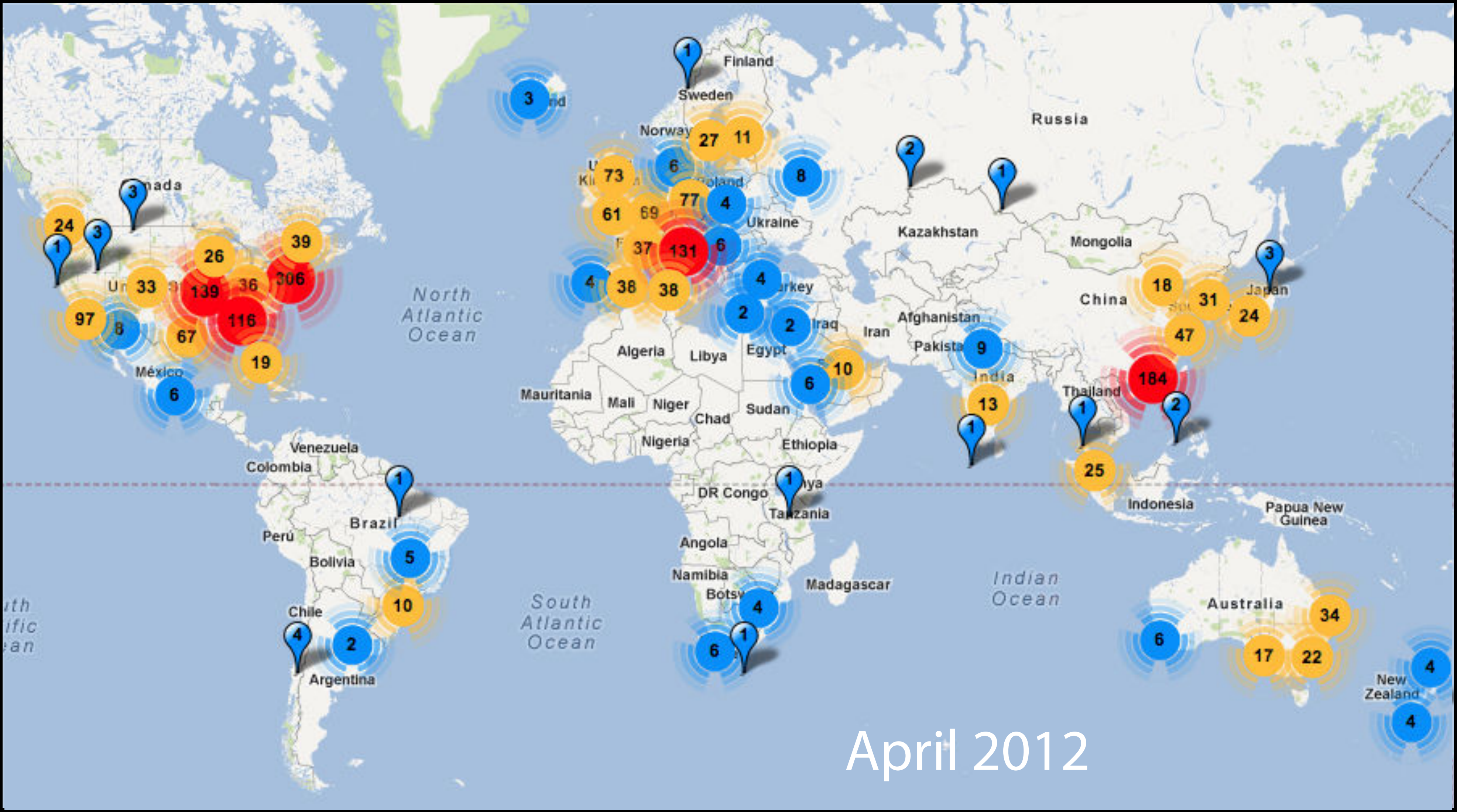


Agenda: Day 1

Welcome, Basic Analysis
Basic analyses into Reusable Workflows
Galaxy Project Overview
NGS Analysis I: Through Tophat
A Simple Change ...
Persistence, Sharing, and Publishing
NGS Analysis II: Cufflinks, Cuffdiff
Visualization and visual analytics

Coffee and lunch breaks throughout the day

The Motivation Slide



April 2012

Next Generation Genomics: World Map of High-throughput Sequencers
Nick Loman, James Hadfield

http://omicsmaps.com

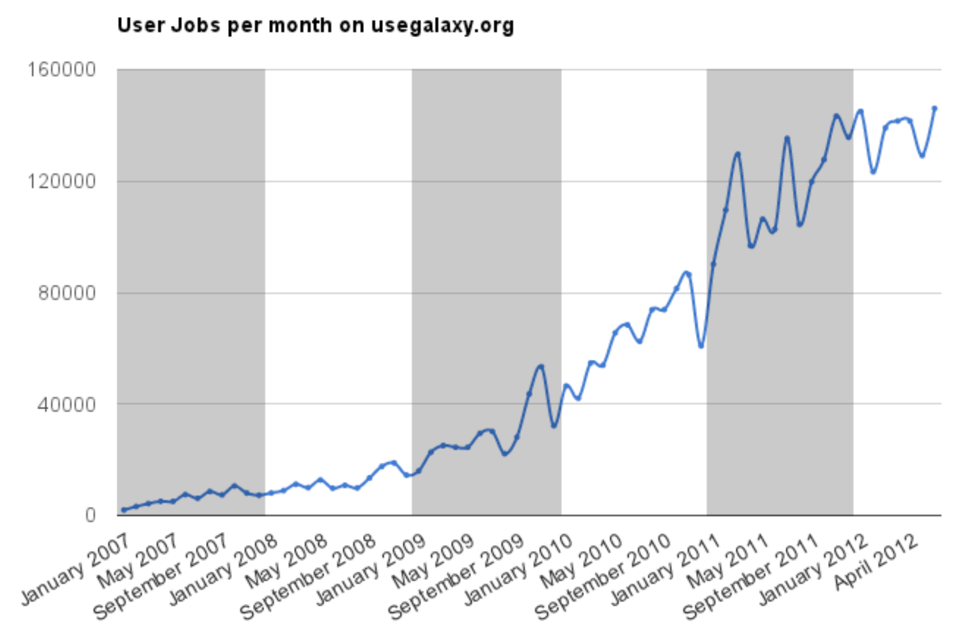
What is Galaxy?

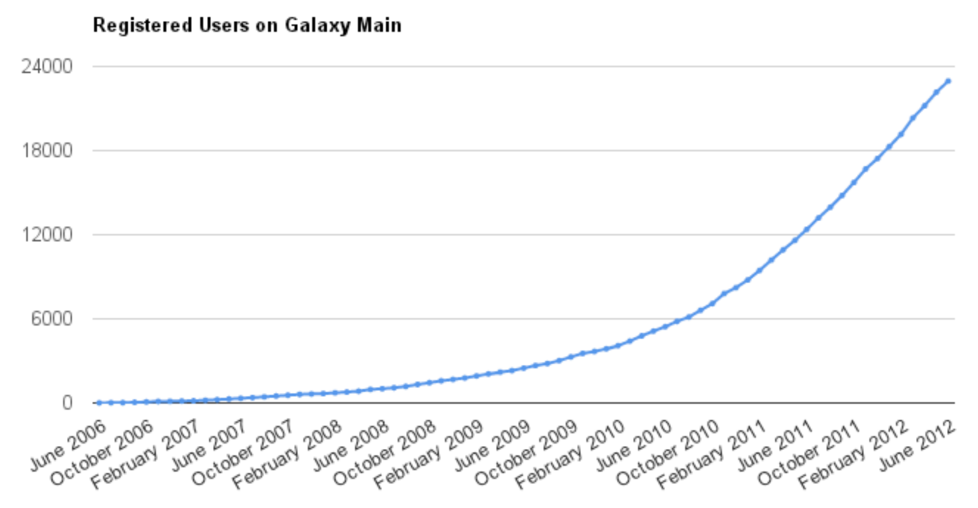
- A **data analysis and integration** tool
- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
- These 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 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.

CLUSTER RESOURCES
TORQUE

GRID ENGINE

Platform Computing

DRMAA
Distributed Resource Management Application API — www.drmaa.org

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**

amazon
web services™

http://aws.amazon.com/education

Galaxy Community

Tool Shed

Mailing Lists (very active)

Screencasts

Events Calendar, News Feed

Community Wiki

Local Public Installs

CiteULike group, Mendeley mirror

Annual Community Meting

http://galaxyproject.org/wiki

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

Other Upcoming Galaxy Events

Swiss Galaxy Workshop
Bern, 3 October 2012
http://bit.ly/gxyswiss

Date	Topic/Event	Venue/Location	Contact
September 3-4	*Galaxy Workshop*	University of Pretoria, Pretoria, South Africa	Dave Clements
September 6-7		Stellenbosch University, Stellenbosch, South Africa	
September 10-11	*Systems Bioinformatics Workshop*	Institute for Systems Biology Seattle, Washington, United States	James Taylor
September 10-12	*Transparent, accessible, reproducible analysis with Galaxy*	**South African Genetics & Bioinformatics Society Conference** University of Stellenbosch, Stellenbosch, South Africa	Dave Clements
	Assembling a cassava transcriptome using Galaxy on a high performance computing cluster		Aobakwe Matshidiso
September 11-13	*Facilitating Research on Heart Disease through SaaS*	Bio-IT World Cloud Summit, San Francisco, California, United States	Raimond Winslow
September 11-14	*Automated and reproducible analysis of NGS data (ARANGS12)*	Instituto Gulbenkian de Ciência, Oeiras, Portugal	Rutger Vos, Darin London
September 27-29	*Informatics Workshop*	**Beyond the Genome 2012**, Harvard Medical School, Boston, Massachusetts	James Taylor
October 3	*(first Swiss) Galaxy Workshop*	**SyBIT Tech Day**, Bern, Switzerland	Hans-Rudolf Hotz
October 8-12	*Extending High-Performance Computing Beyond its Traditional User Communities Workshop*	**8th IEEE International Conference on eScience (eScience 2012)**, Chicago Illinois, United States	James Taylor
October 9-11	*Tavaxy: A Workflow System with Taverna and Galaxy Capabilities and Cloud Computing Support*	Bio-IT World Europe, Vienna, Austria	Mohamed Abouelhoda
October 31 -	*Computaional & Comparative Genomics Course*	Cold Spring Harbor Laboratory, New York, United	William Pearson,

http://galaxyproject.org/wiki/Events

Galaxy URLs to Remember

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

Agenda: Day 1

Welcome, Basic Analysis
Basic analyses into Reusable Workflows
Galaxy Project Overview
NGS Analysis I: Through Tophat
A Simple Change ...
Persistence, Sharing, and Publishing
NGS Analysis II: Cufflinks, Cuffdiff
Visualization and visual analytics

Coffee and lunch breaks throughout the day

RNA-seq Exercise

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

http://bit.ly/gxyRNASEX

http://bit.ly/SUNgray
http://bit.ly/SUNmaroon

RNA-seq Exercise: A Plan

- Get input datasets; hg19, 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
- Run Cuffdiff on Cufflinks output to see differential expression

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
  - All datasets are FASTQ and from the Body Map 2.0 project
  - What is FASTQ?
    - http://en.wikipedia.org/wiki/FASTQ_format

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality: Option 1
    - **NGS QC and Manipulation → Compute Quality Statistics**
    - NGS QC and Manipulation → Draw quality score boxplot
    - Gives you no control over how it is calculated or presented.

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality: Option 2
  - NGS QC and Manipulation → **FastQ Summary Statistics**
  - Graph / Display Data →Boxplot of quality statistics
  - Gives you a lot of control over what the box plot looks like, but no additional information

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality: Option 3
  - NGS QC and Manipulation → **Fastqc**
  - Gives you a lot a lot more information but no control over how it is calculated or presented.

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality
- Trim as we see fit: Option 1
  - **NGS QC and Manipulation → FASTQ Trimmer by column**
  - Trim same number of columns from every record
  - Can specify different trim for 5′ and 3′ ends

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality
- ~~Trim~~ Filter as we see fit: Option 2
  - NGS QC and Manipulation → **Filter FASTQ reads by quality score and length**
  - Keep or discard whole reads at a time
  - Can have different thresholds for different regions of the reads.
  - Keeps original read length.

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality
- Trim as we see fit: Option 3
  - NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**
  - Trim from both ends, using sliding windows, until you hit a high-quality section.
  - Produces variable length reads

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality
- Trim as we see fit.
- Map the reads to the human reference using Tophat
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq mapping here.*
  - Visualize results

http://bit.ly/gxyRNASEX

Agenda: Day 1

Welcome, Basic Analysis
Basic analyses into Reusable Workflows
Galaxy Project Overview
NGS Analysis I: Through Tophat
A Simple Change ...
Persistence, Sharing, and Publishing
NGS Analysis II: Cufflinks, Cuffdiff
Visualization and visual analytics

Coffee and lunch breaks throughout the day

Hands On: Basic Analysis ...
until you go insane

On pig chromosome 18,
which coding exons (GTF format)
have the most repeats (BED format)
in them?

http://bit.ly/SUNgray
http://bit.ly/SUNmaroon

Repetitious Pigs: GTF and BED

- Get the GTF from UCSC
  - *Hmm*: There is no “coding exons” choice w/ GTF

- Points we will eventually ponder
  - Do we care about *coding exons* versus *exons?*
  - Do we care about *exon names, gene names, transcript names, or just coordinates?*
  - *Can the same approach even work with GTF?*

Agenda: Day 1

Welcome, Basic Analysis
Basic analyses into Reusable Workflows
Galaxy Project Overview
NGS Analysis I: Through Tophat
A Simple Change ...
Persistence, Sharing, and Publishing
NGS Analysis II: Cufflinks, Cuffdiff
Visualization and visual analytics

Coffee and lunch breaks throughout the day

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:

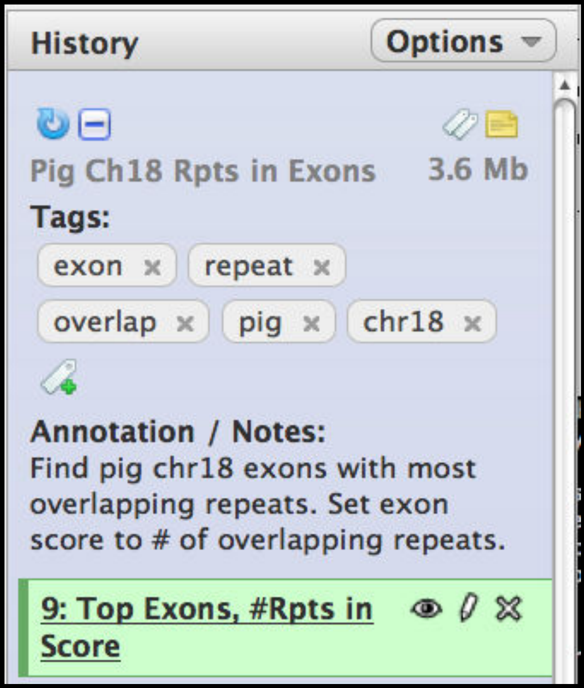
Make something available to everyone

Managing Histories and Datasets

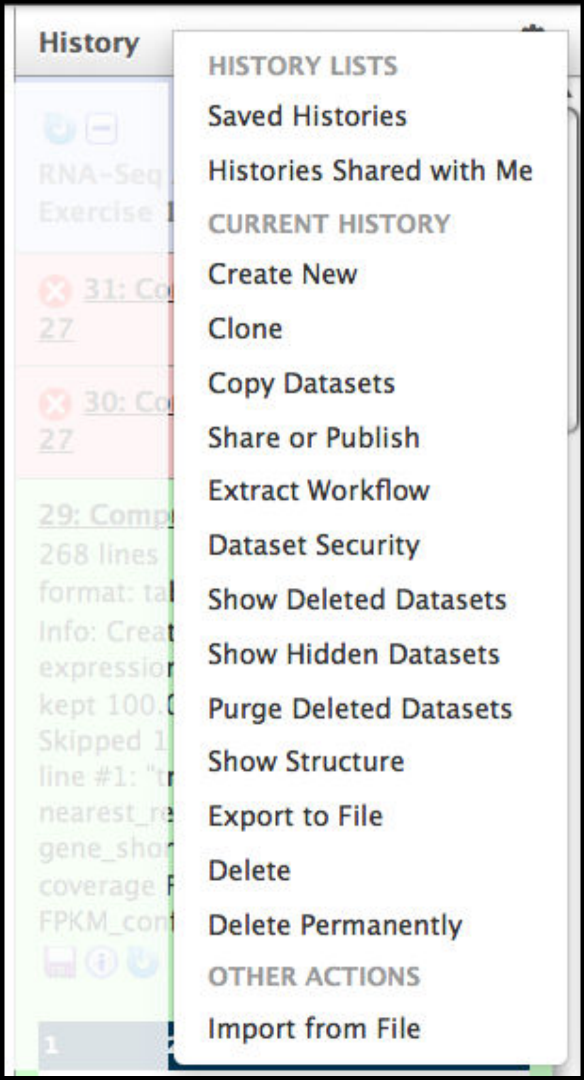
Give every **history and dataset** a **clear name**

Datasets and histories can also have annotation and tags

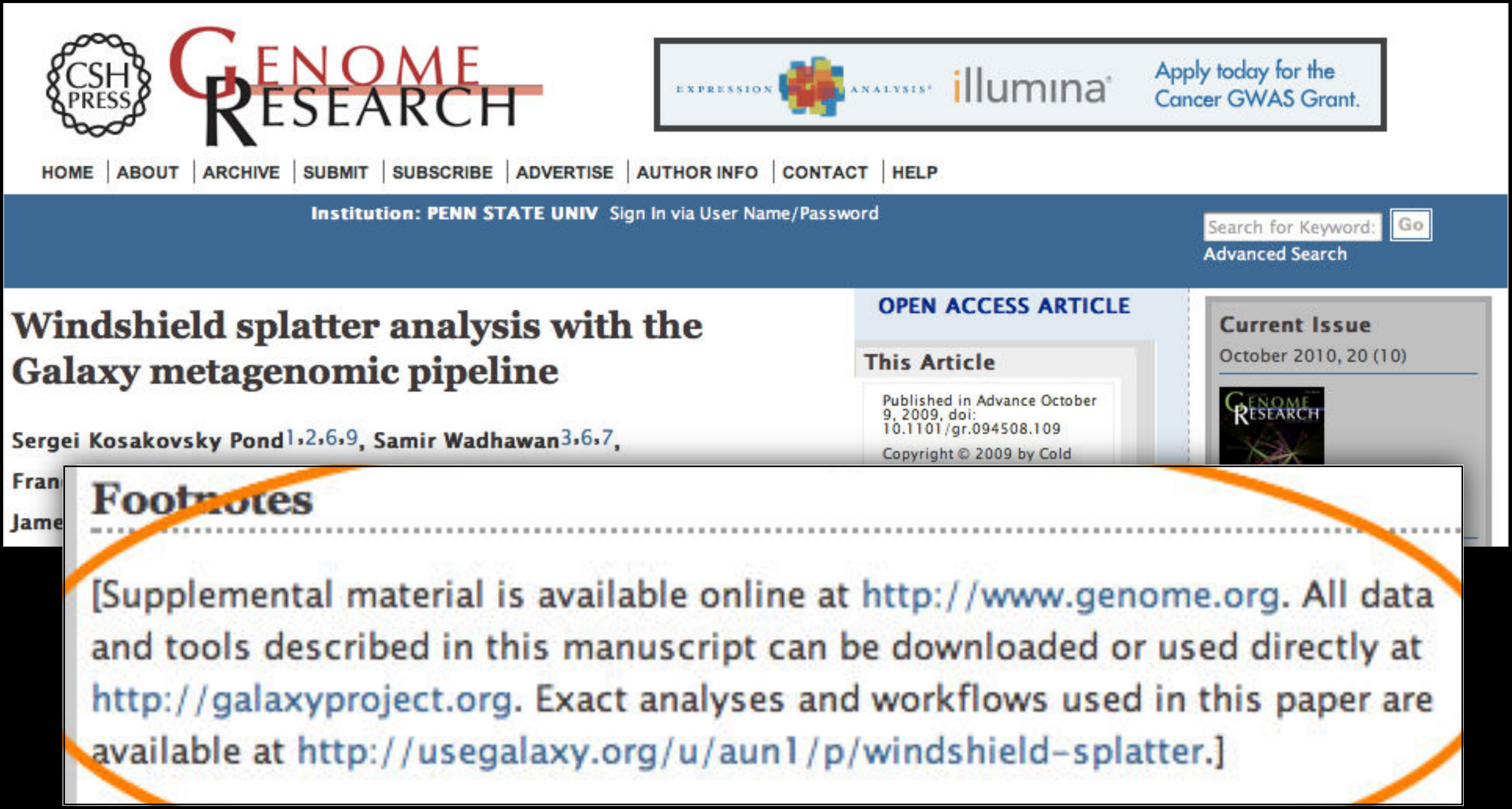
Each **history** has an options/actions list



Find pig chr18 exons with most overlapping repeats. Set exon score to # of overlapping repeats.



Sharing and Publishing Your Work



Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], **Samir Wadhawan**[3,6,7],

Fran

Jame

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

Copyright © 2009 by Cold

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.]

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

toolshed.g2.bx.psu.edu

Agenda: Day 1

Welcome, Basic Analysis
Basic analyses into Reusable Workflows
Galaxy Project Overview
NGS Analysis I: Through Tophat
A Simple Change ...
Persistence, Sharing, and Publishing
NGS Analysis II: Cufflinks, Cuffdiff
Visualization and visual analytics

Coffee and lunch breaks throughout the day

RNA-seq Exercise: A Plan

- ...
- Trim as we see fit.
- Map the reads to the human reference using Tophat
- Run Cufflinks on Tophat output to assemble reads into transcripts
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq transcript prediction here.*

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- ...
- Map the reads to the human reference using Tophat
- Run Cufflinks on Tophat output to assemble reads into transcripts
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq transcript prediction here.*
  - Visualize it

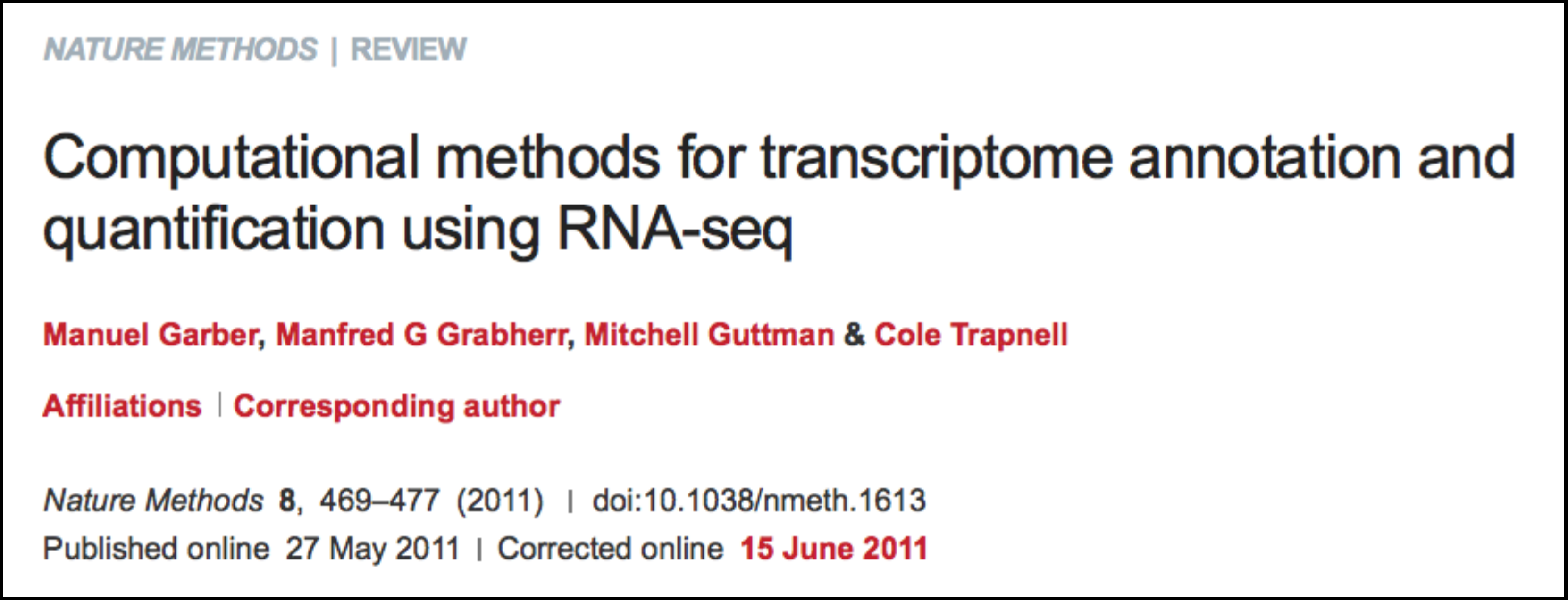
http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- ...
- Run Cufflinks on Tophat output to assemble reads into transcripts
- Run Cuffdiff on Cufflinks output to see differential expression
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq transcript prediction here.*

http://bit.ly/gxyRNASEX

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: Day 1

Welcome, Basic Analysis
Basic analyses into Reusable Workflows
Galaxy Project Overview
NGS Analysis I: Through Tophat
A Simple Change ...
Persistence, Sharing, and Publishing
NGS Analysis II: Cufflinks, Cuffdiff
Visualization and visual analytics

Coffee and lunch breaks throughout the day

Visualize

Send data results to **external** genome browsers

Trackster: Galaxy’s genome browser

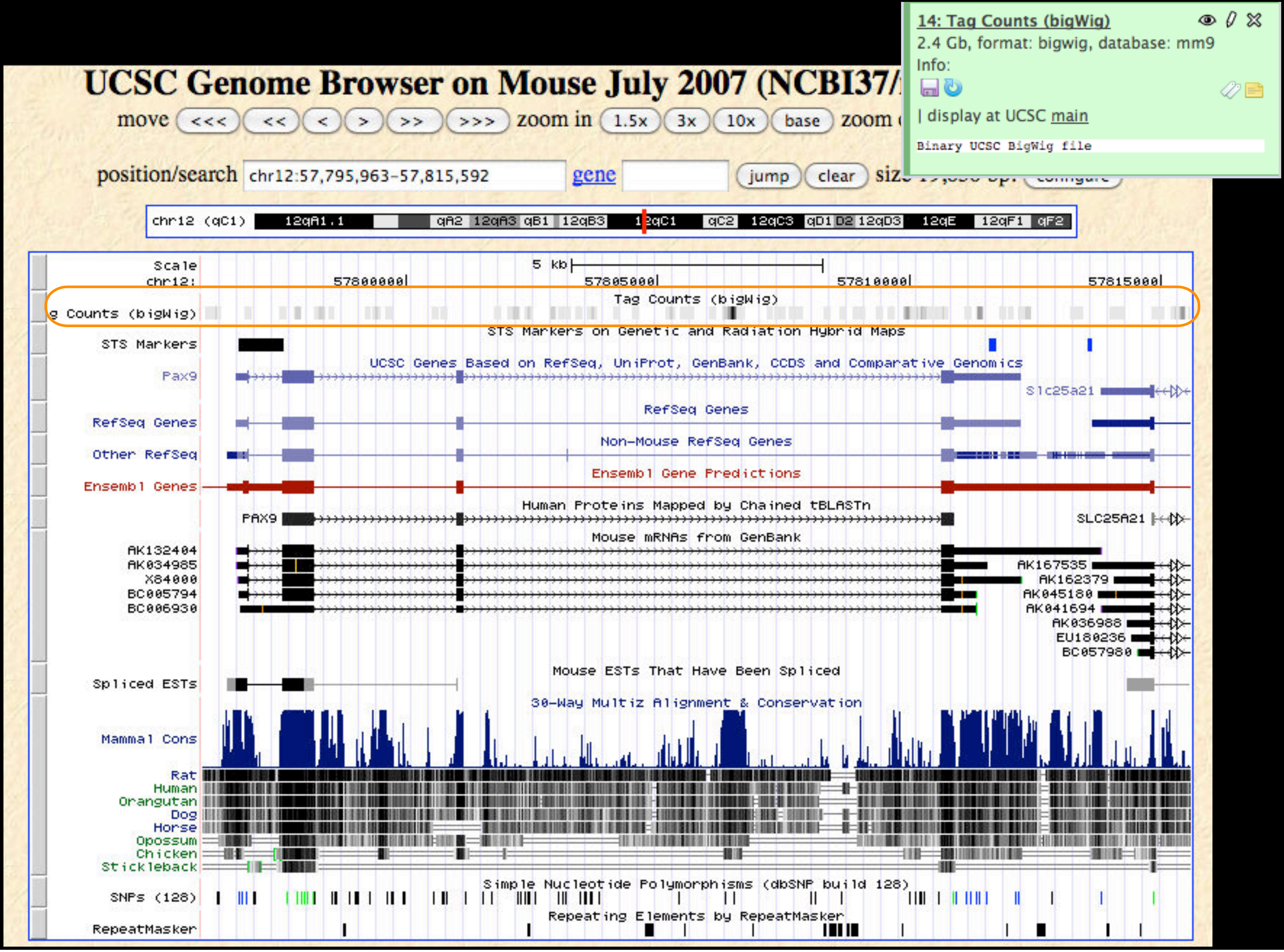
External Genome Browsers

UCSC

Ensembl

GBrowse

IGV



UCSC Genome Browser on Mouse July 2007 (NCBI37/

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

14: Tag Counts (bigWig)

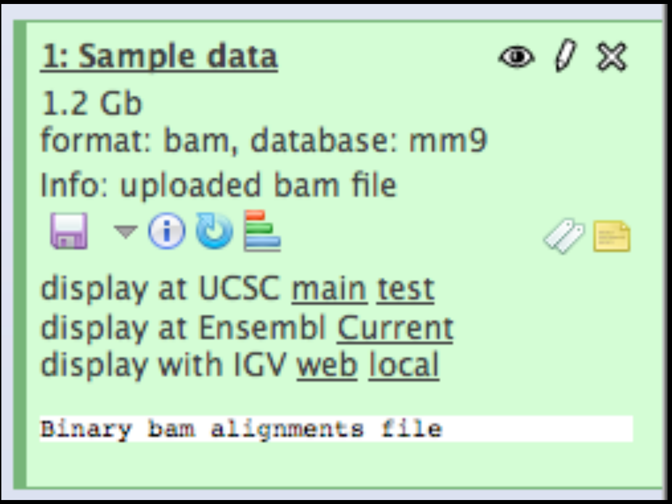
2.4 Gb, format: bigwig, database: mm9

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

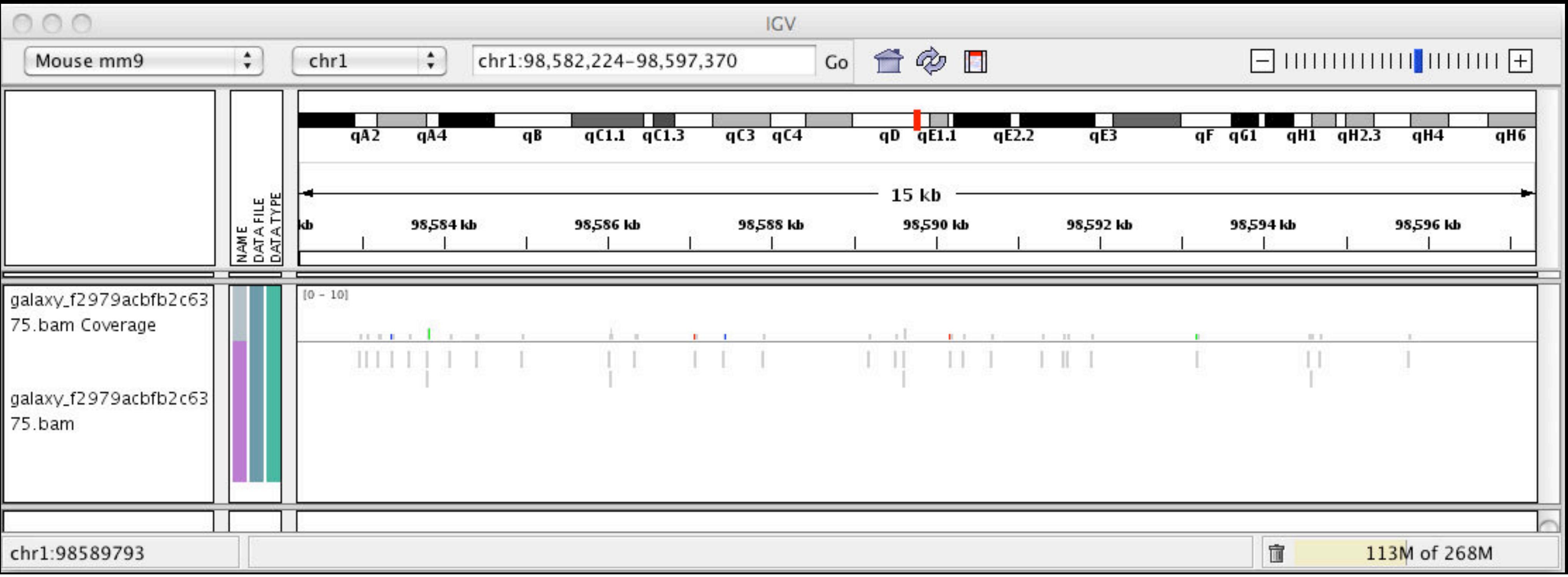
display with IGV web local

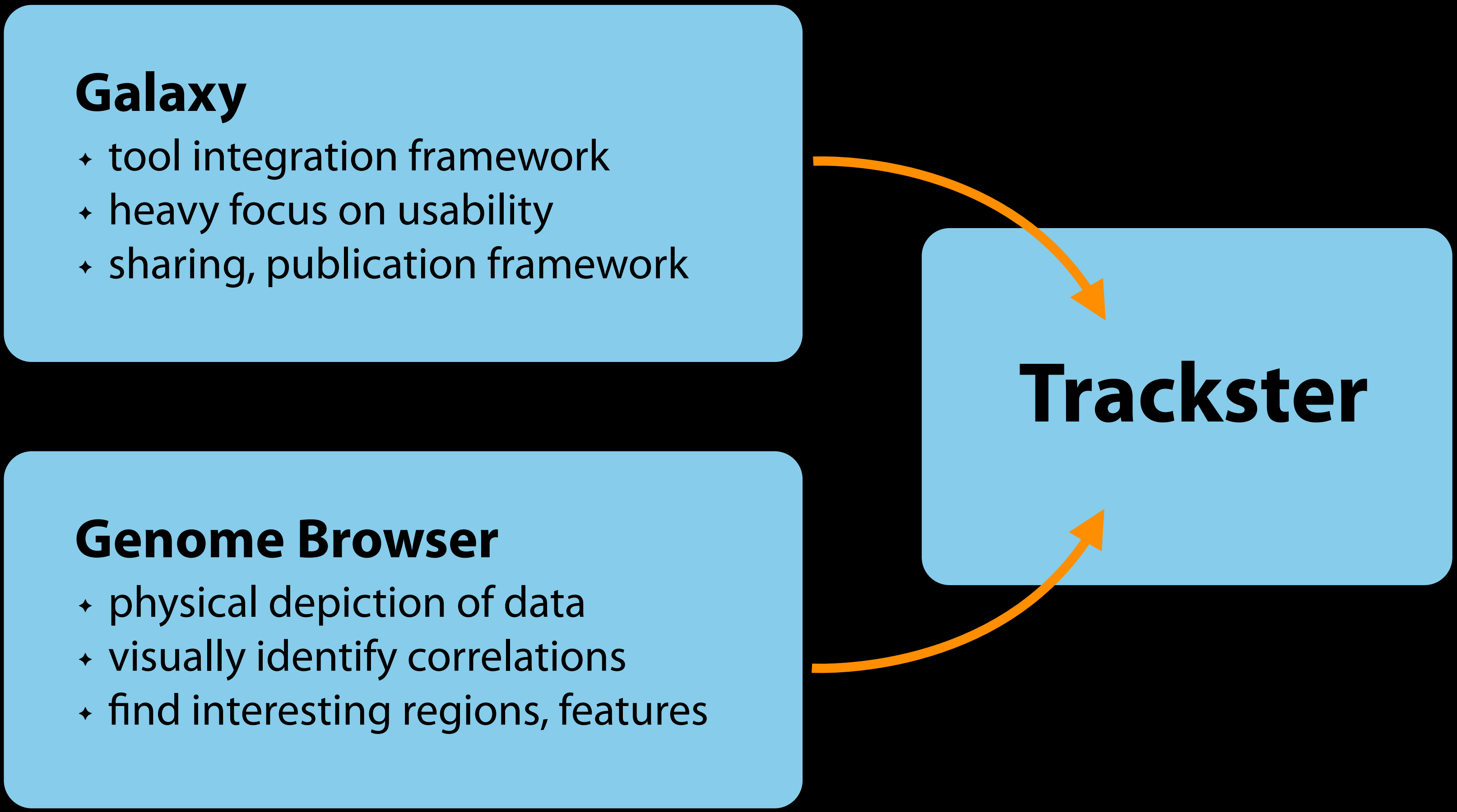
Binary bam alignments file

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





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

Trackster

View your data from within Galaxy

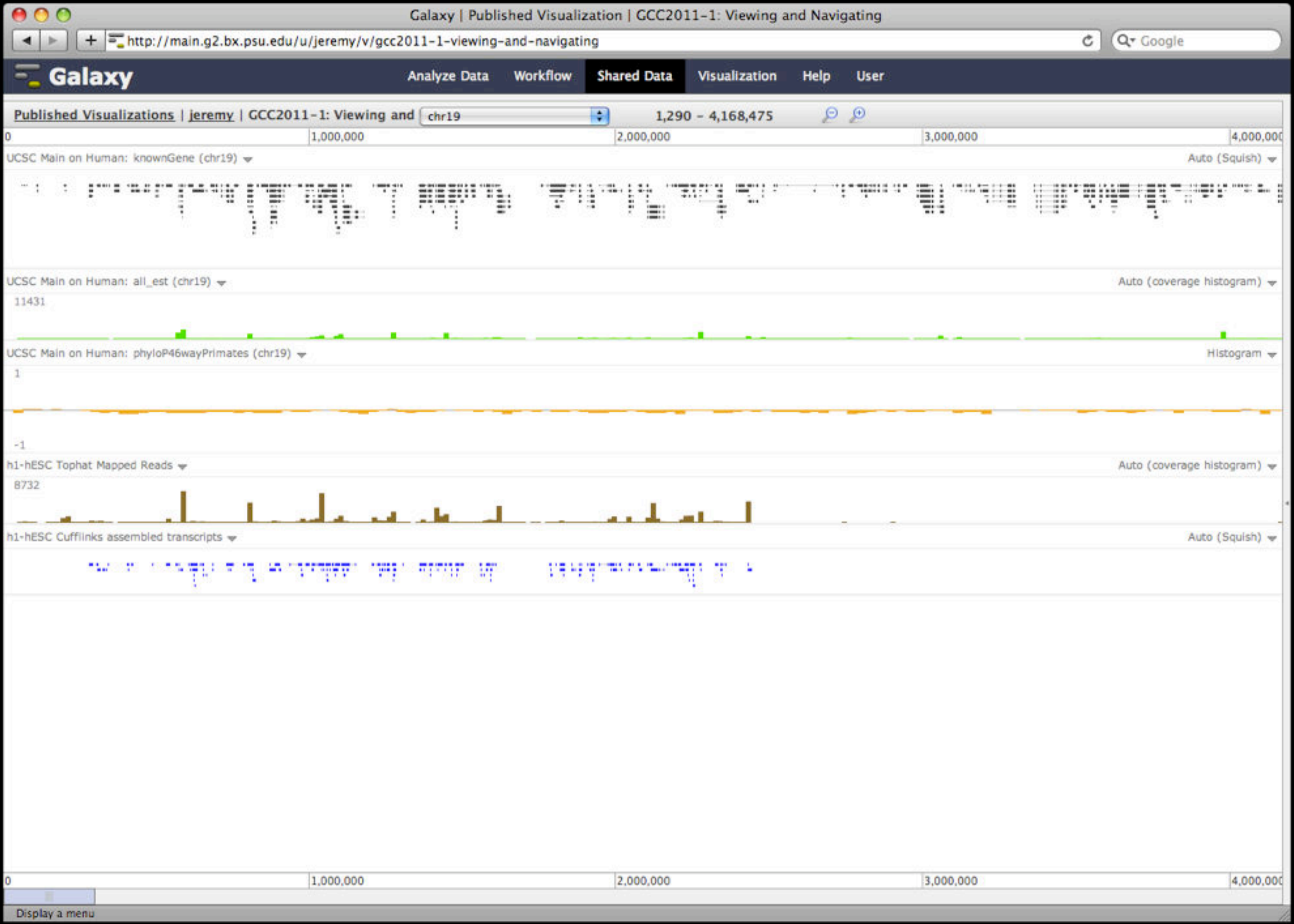
- No data transfers to external site
- Use it locally, even without internet access

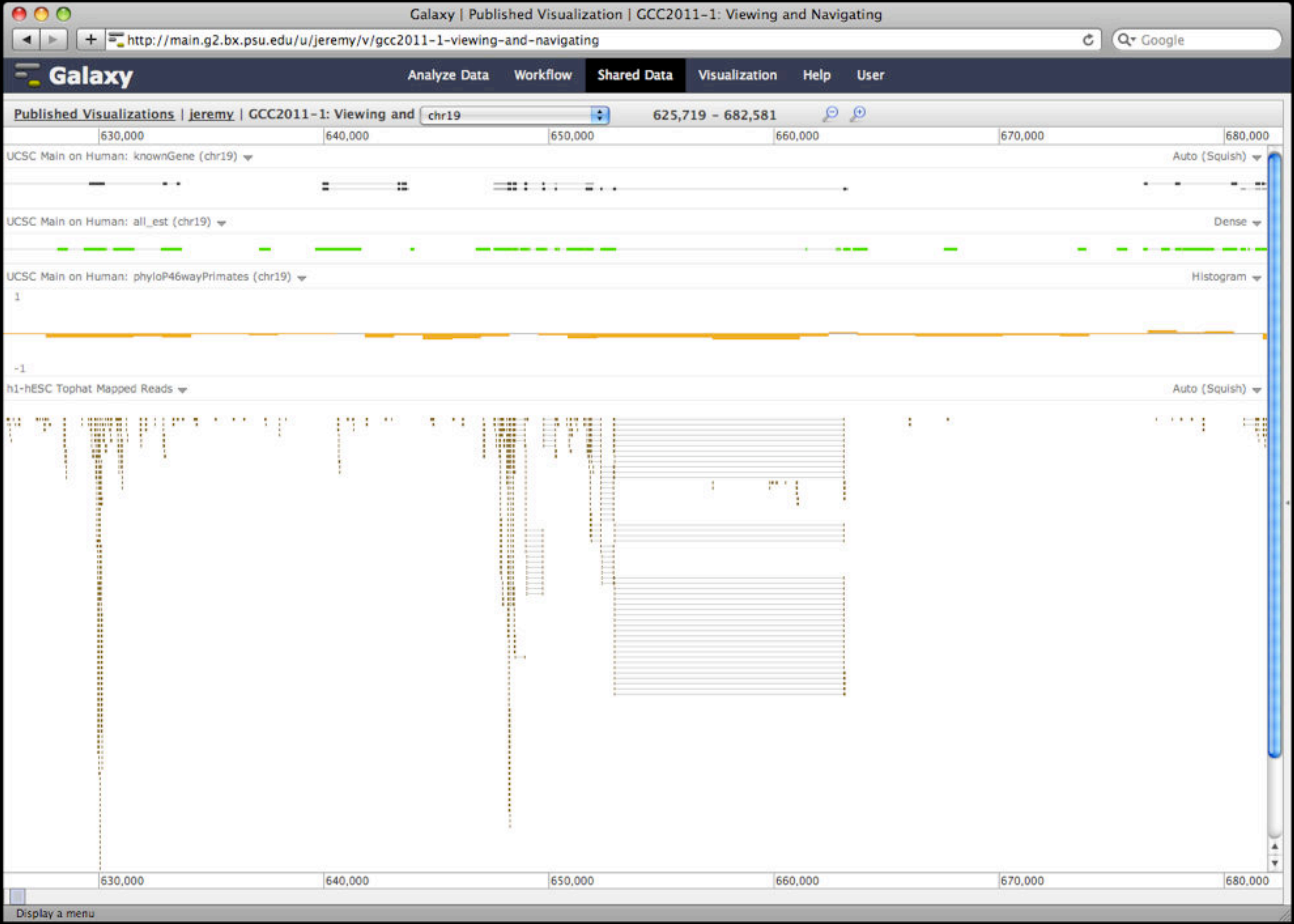
Supports common filetypes

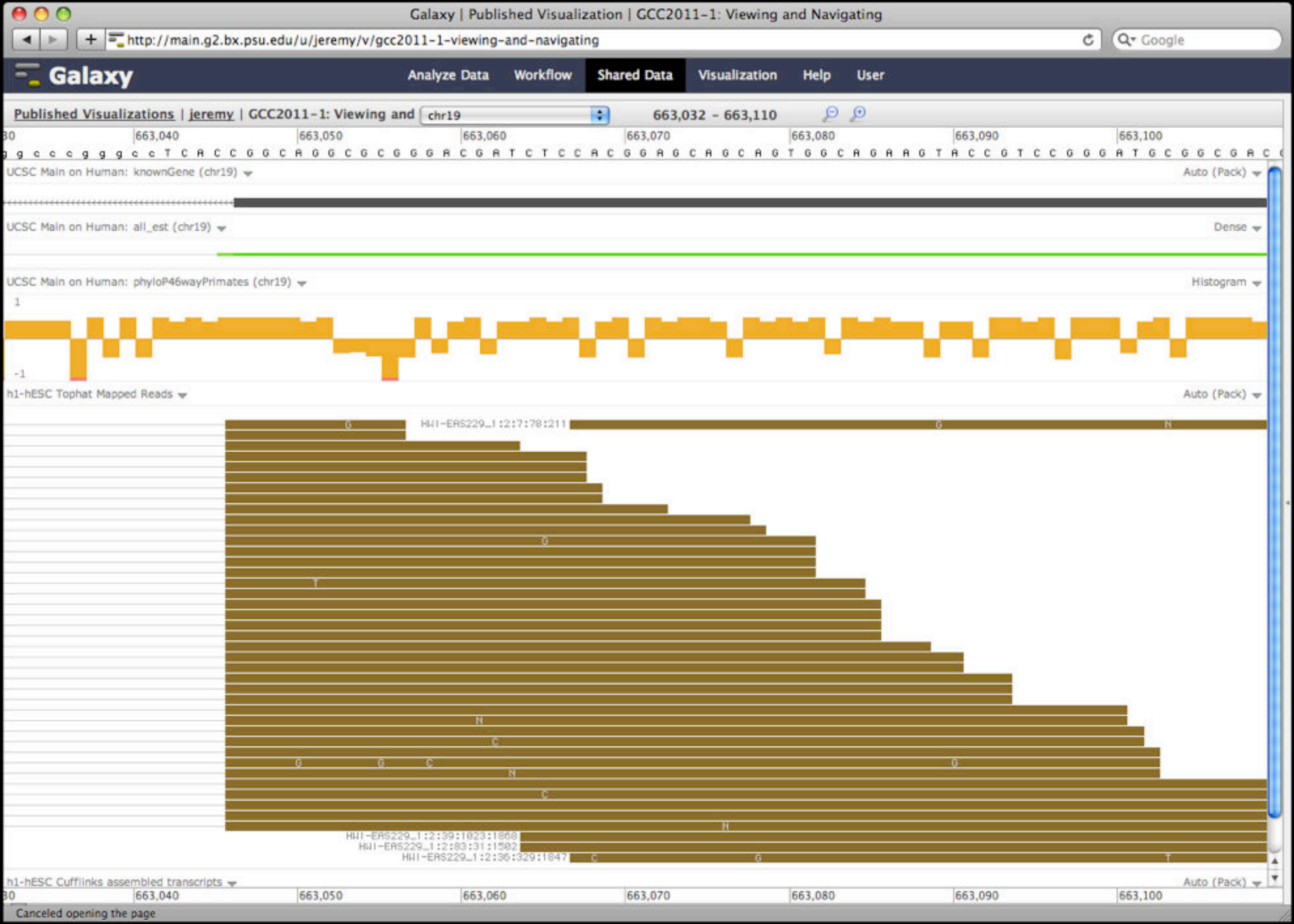
- BAM, BED, GFF/GTF, WIG

Unique features

- custom genomes
- highly interactive


Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating
http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
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
-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
Display a menu


Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating
http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
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)
630,000
640,000
650,000
660,000
670,000
680,000
Display a menu



Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating

http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating

Canceled opening the page

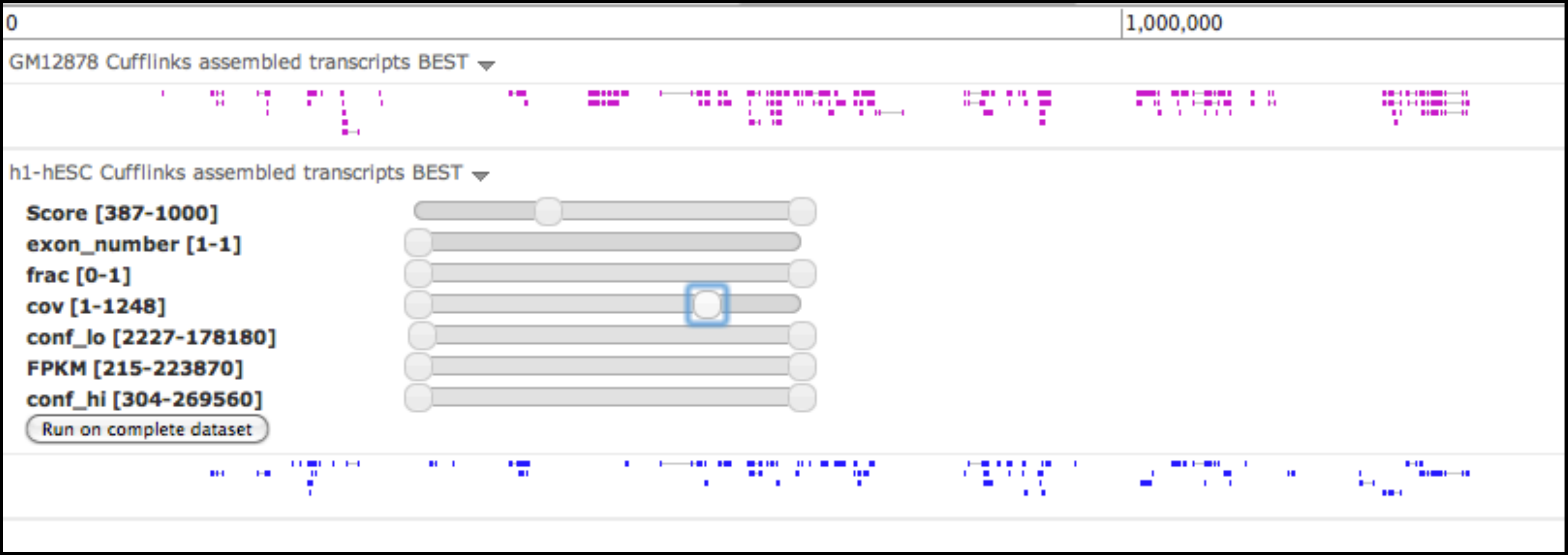
But really, why *another* genome browser

From static browsing to visual analysis

Visual feedback and experimentation needed for complex tools with many parameters

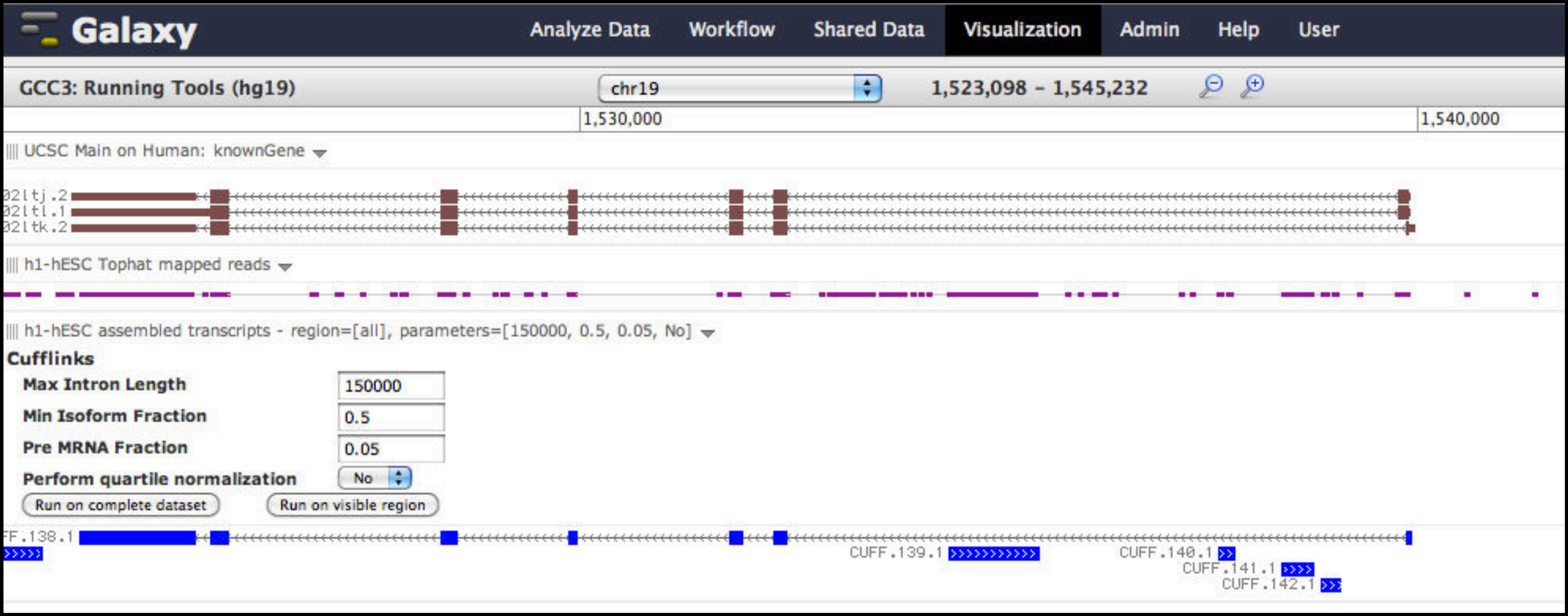
Leverage Galaxy strengths: a very sound model for abstracting interfaces to analysis tools and already integrates an enormous number

Dynamic Filtering



Score [387-1000]
exon_number [1-1]
frac [0-1]
cov [1-1248]
conf_lo [2227-178180]
FPKM [215-223870]
conf_hi [304-269560]

Integrating Tools and Visualization



GCC3: Running Tools (hg19)

chr19 1,523,098 – 1,545,232

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

Visualization: Even More

- usegalaxy.org → Shared Data → Published Visualizations
  - Don’t everyone do this!
- 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
  - Next @

Agenda: Day 1