# Transparent, accessible, reproducible analysis with Galaxy

Dave Clements
Emory University
12 September 2012

http://galaxyproject.org/

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

SASBCB
SOUTH AFRICAN SOCIETY FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

SAGS
South African Genetics Society

# Acknowledgements

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

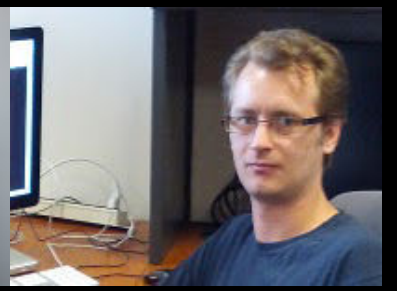
University of Pretoria
Stellenbosch University

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

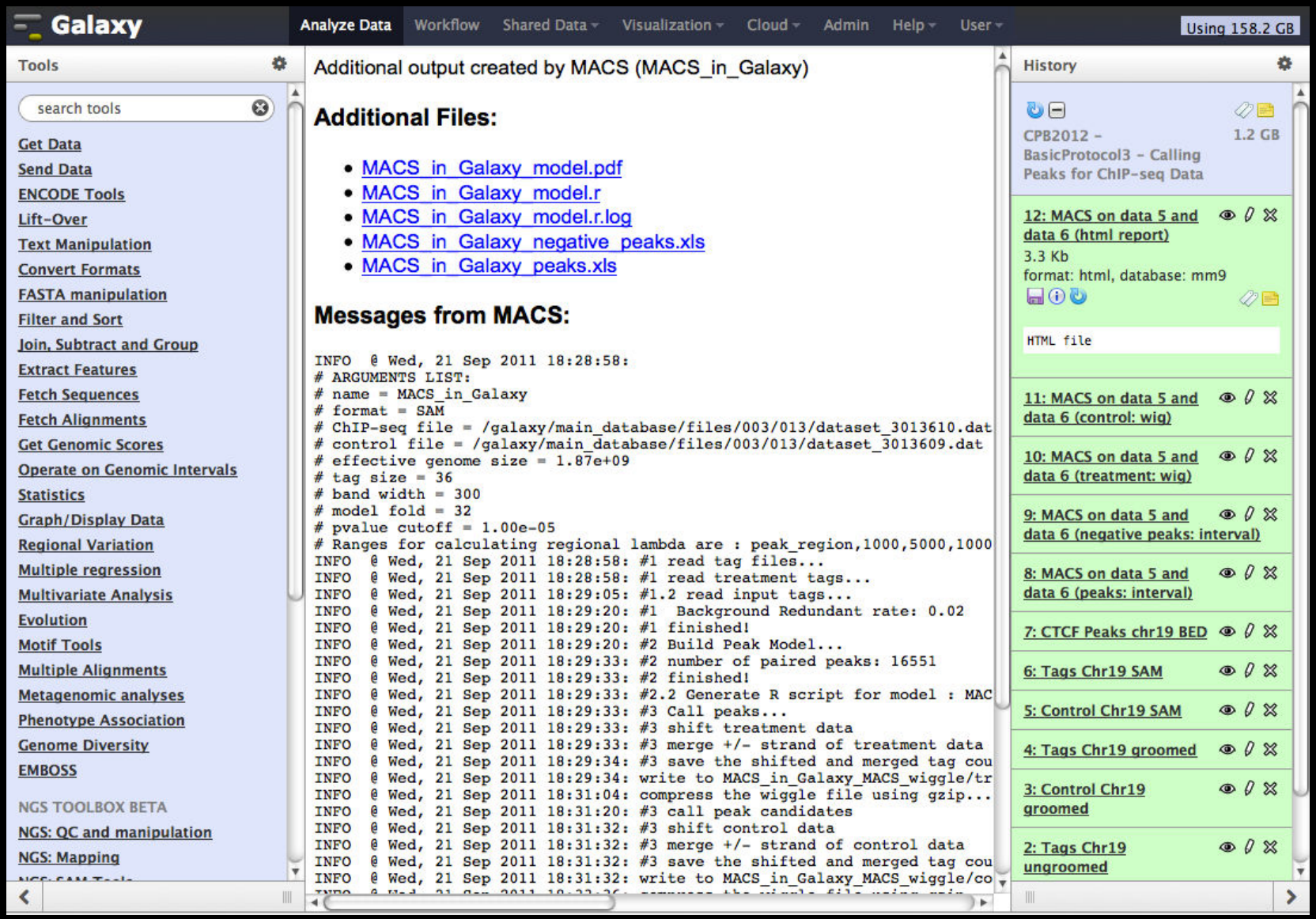
- 18 Nat. Genetics microarray gene expression experiments
- Less than 50% reproducible
- Problems
  - missing data (38%)
  - missing software, hardware details (50%)
  - missing method, processing details (66%)

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

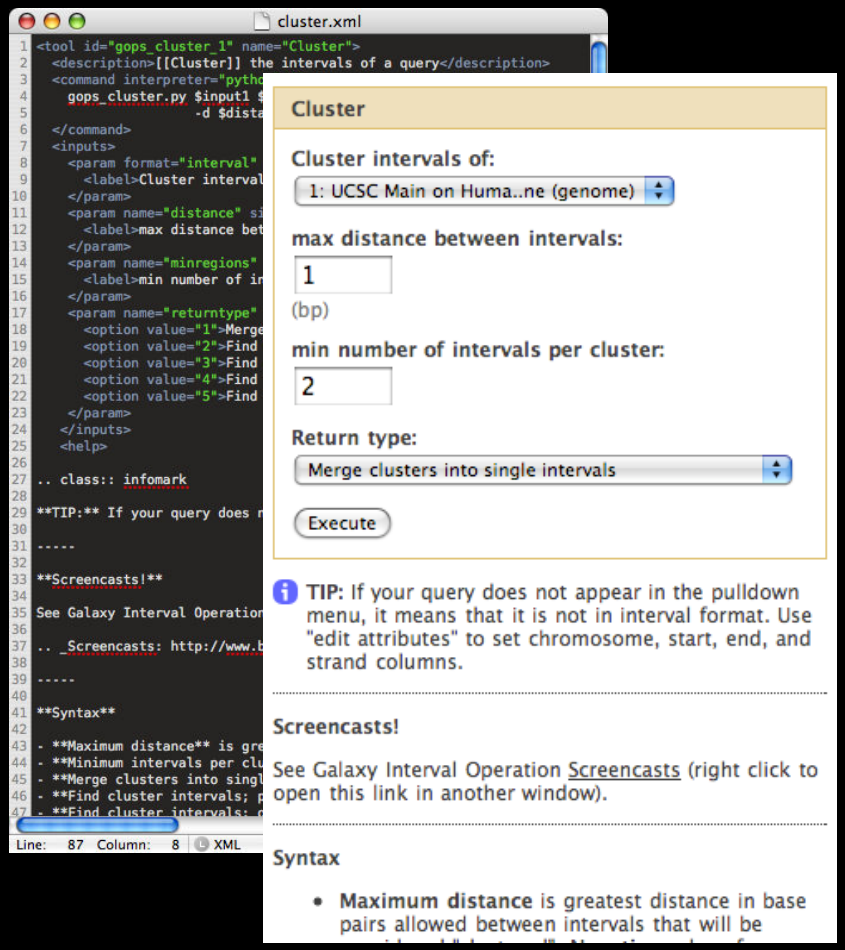
# NGS Re-sequencing Experiment Reproducibility

- 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

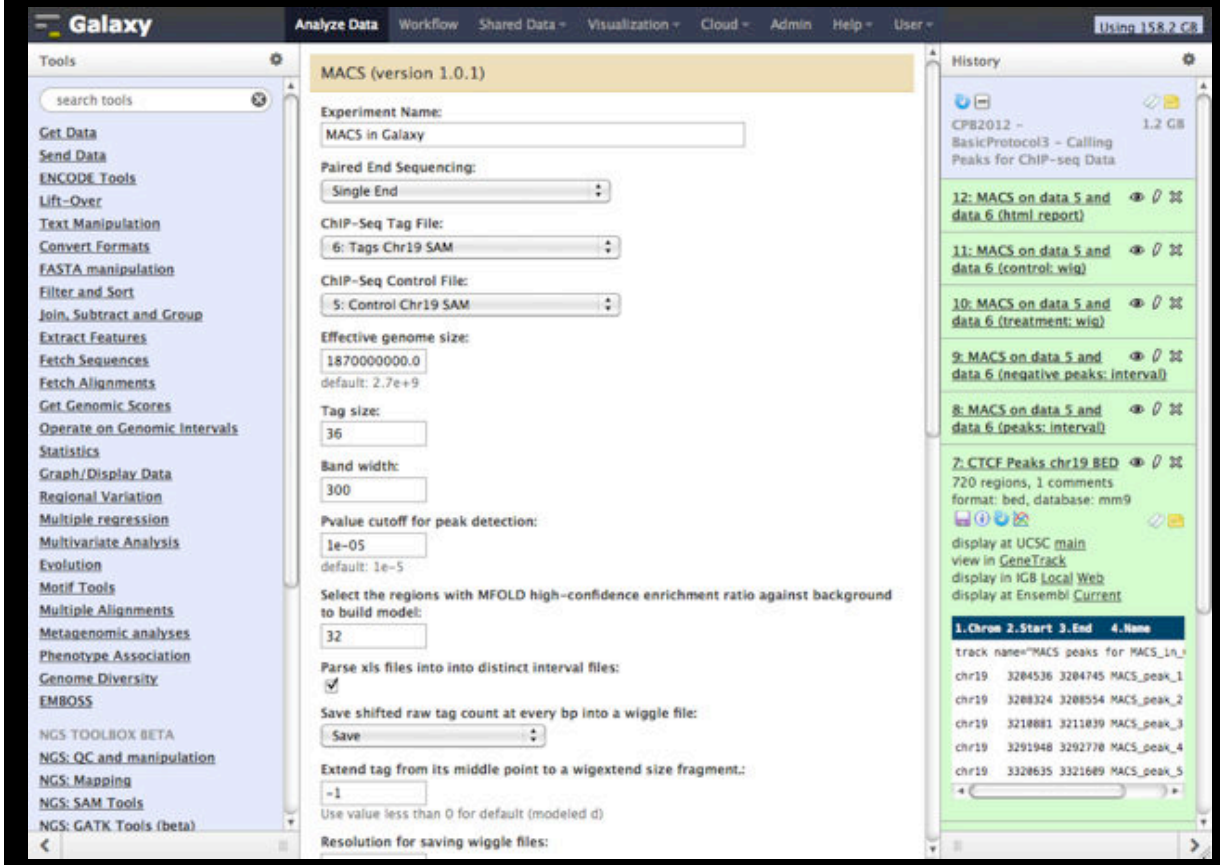


# Integrating existing tools into a uniform framework



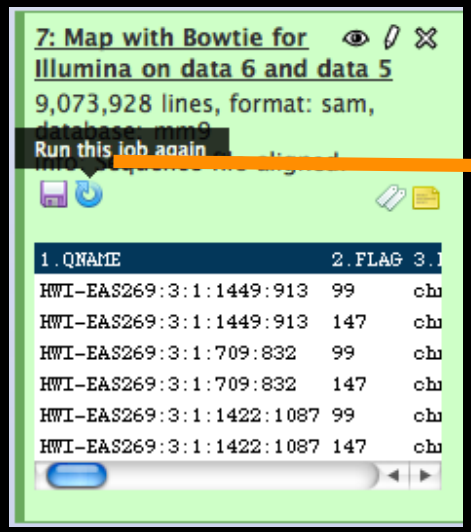

- Defined in terms of an abstract interface (inputs and outputs)
  - In practice, mostly command line tools, a declarative XML description of the interface, how to generate a command line
- Designed to be as easy as possible for tool authors, while still allowing rigorous reasoning

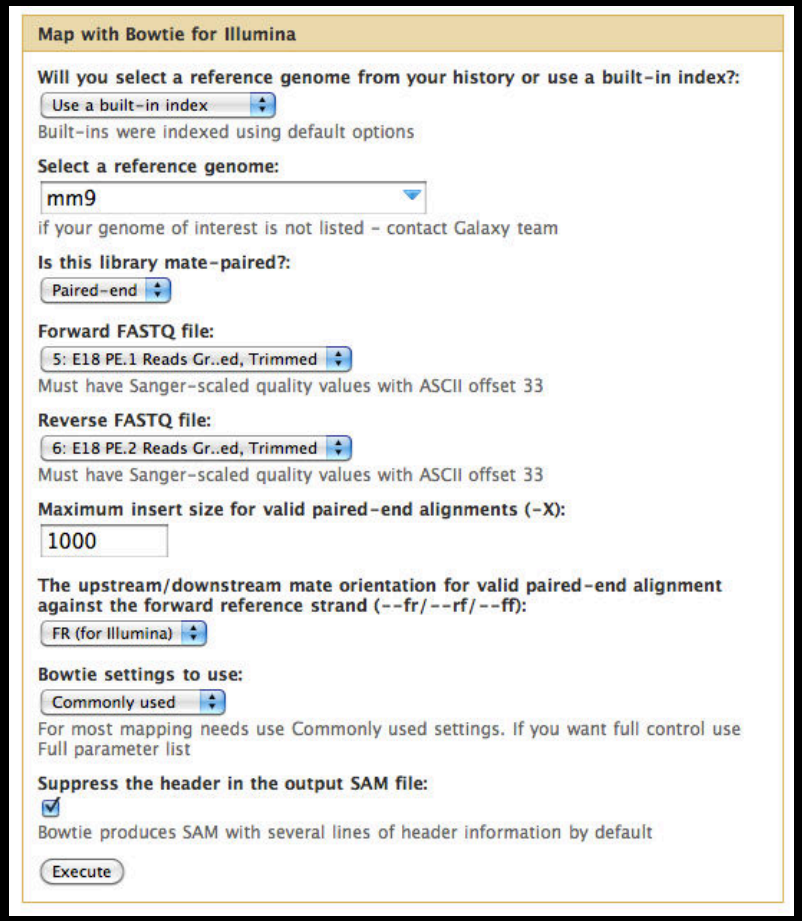
# Galaxy analysis interface



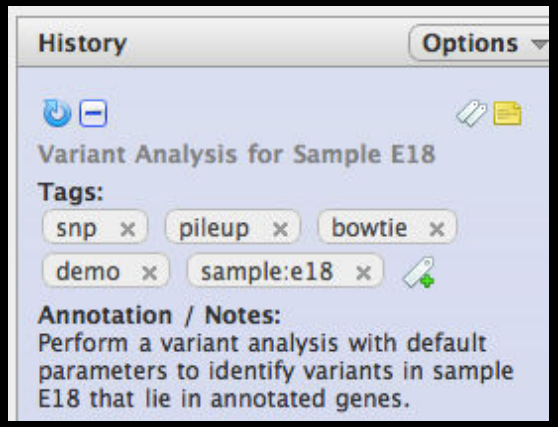

- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses
- Exact parameters of a step can always be inspected, and easily rerun

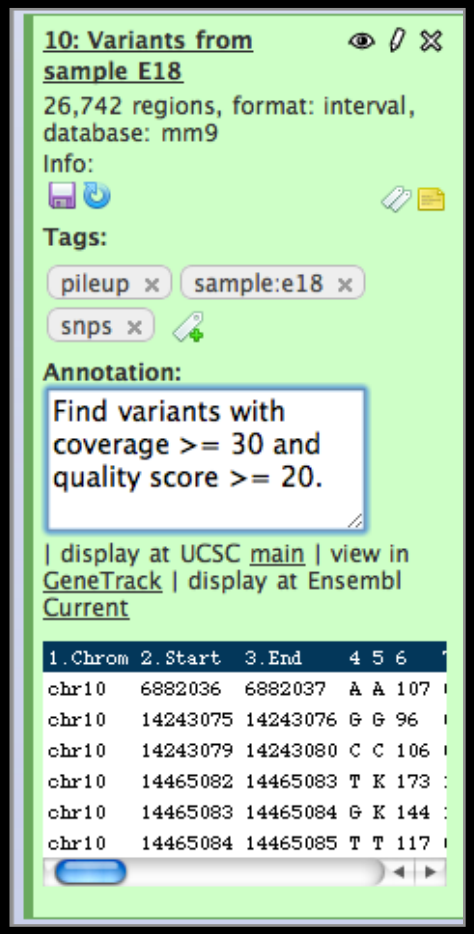
# Automatically tracks every step of every analysis



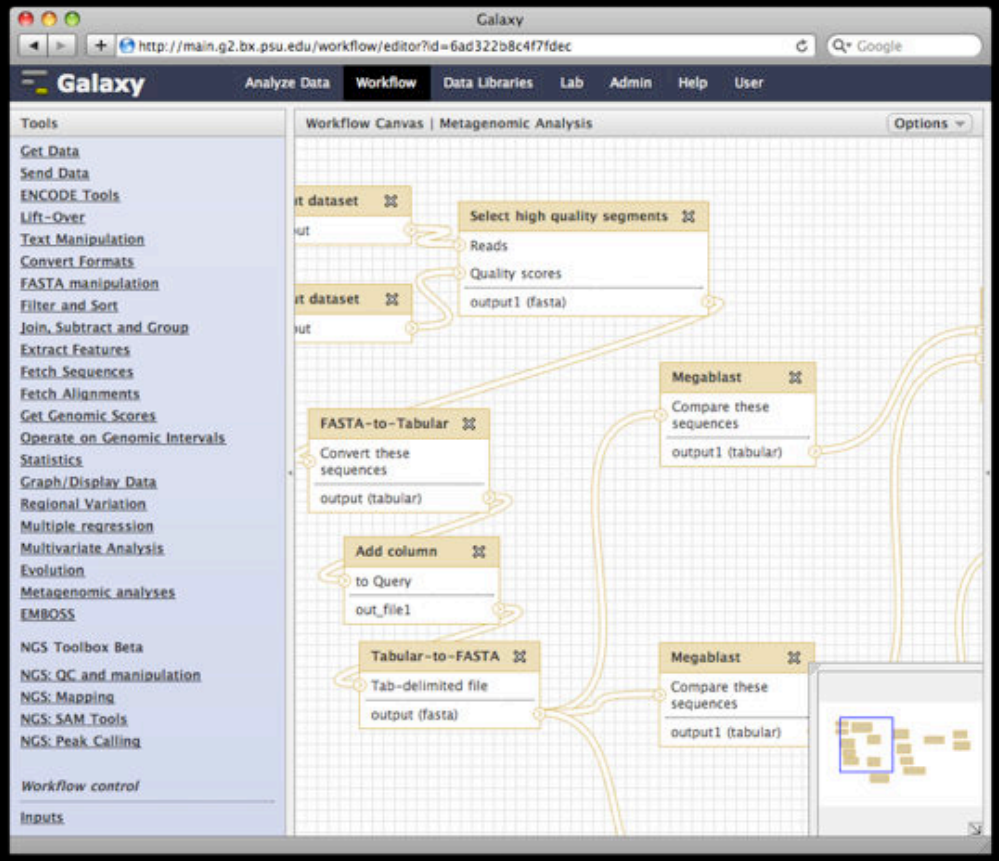



# As well as user-generated metadata and annotation...



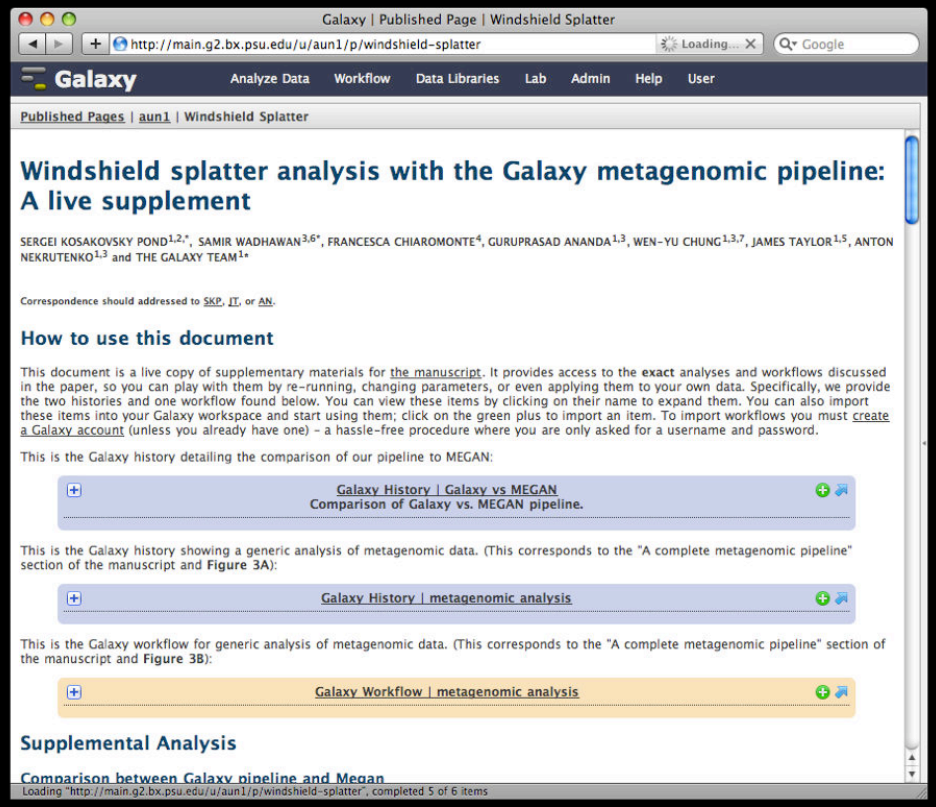



# Galaxy workflow system




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

# Tranparency: Sharing and publishing




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

Galaxy | Published Page | Windshield Splatter

http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter

Galaxy | Analyze Data | Workflow | Data Libraries | Lab | Admin | Help | User

Published Pages | aun1 | Windshield Splatter

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



## Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

OPEN ACCESS ARTICLE

This Article

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



» Abstract *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 manus

This is the Galaxy wo the manuscript and F

## 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 pai end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics b column

ROCHE-454 DATA

- Build base quality distributio
- Select high quality segments
- Combine FASTA and QUAL i FASTQ

AB-SOLID DATA

- Convert SOLiD output to fas
- Compute quality statistics f SOLiD data
- Draw quality score boxplot f 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 sc

- 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 us in local realignment
- Indel Realigner - perform local realignment

BASE RECALIBRATION

- Count Covariates on BAM files
- Table Recalibration on BAM file
- Analyze Covariates - draw plots

GENOTYPING

- Unified Genotyper SNP and inde 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**

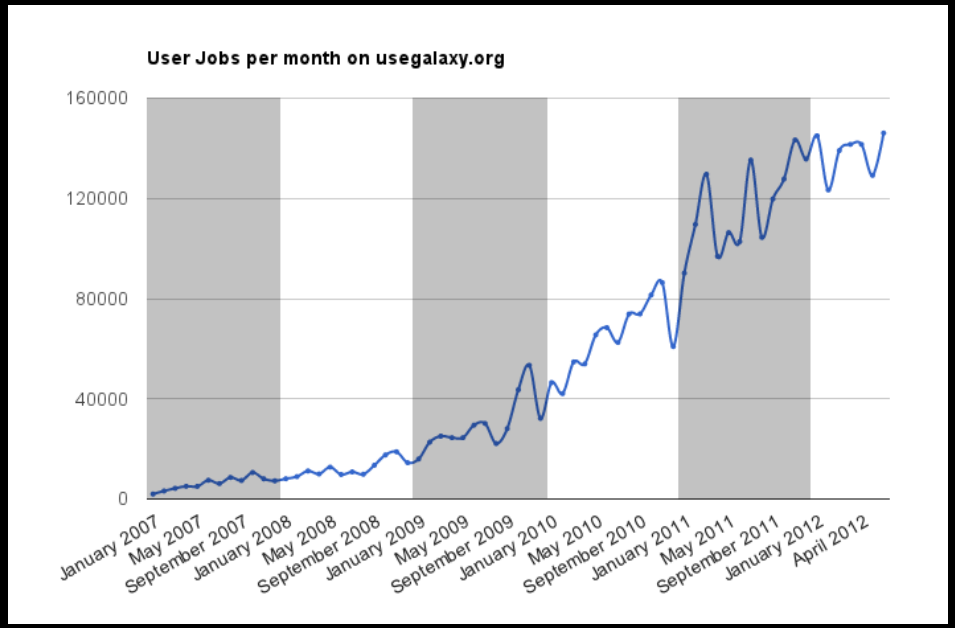
RNA-SEQ

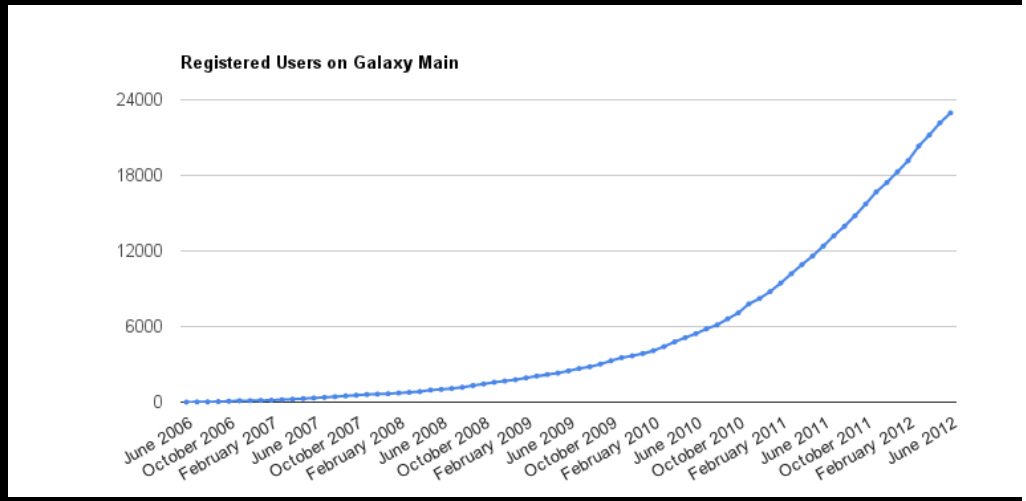
- Tophat for Illumina Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffmerge merge together several Cufflinks assemblies
- Cuffdiff find significant changes

**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**

amazon web services™

**http://aws.amazon.com/education**

# Step by Step Instructions on the Wiki for Amazon

Galaxy Wiki

Login | Search:

CloudMan/AWS/GettingStarted

## 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

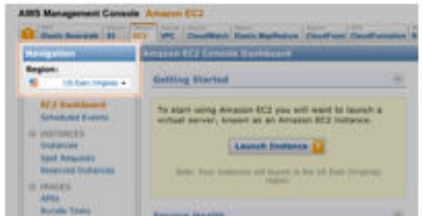
Contents



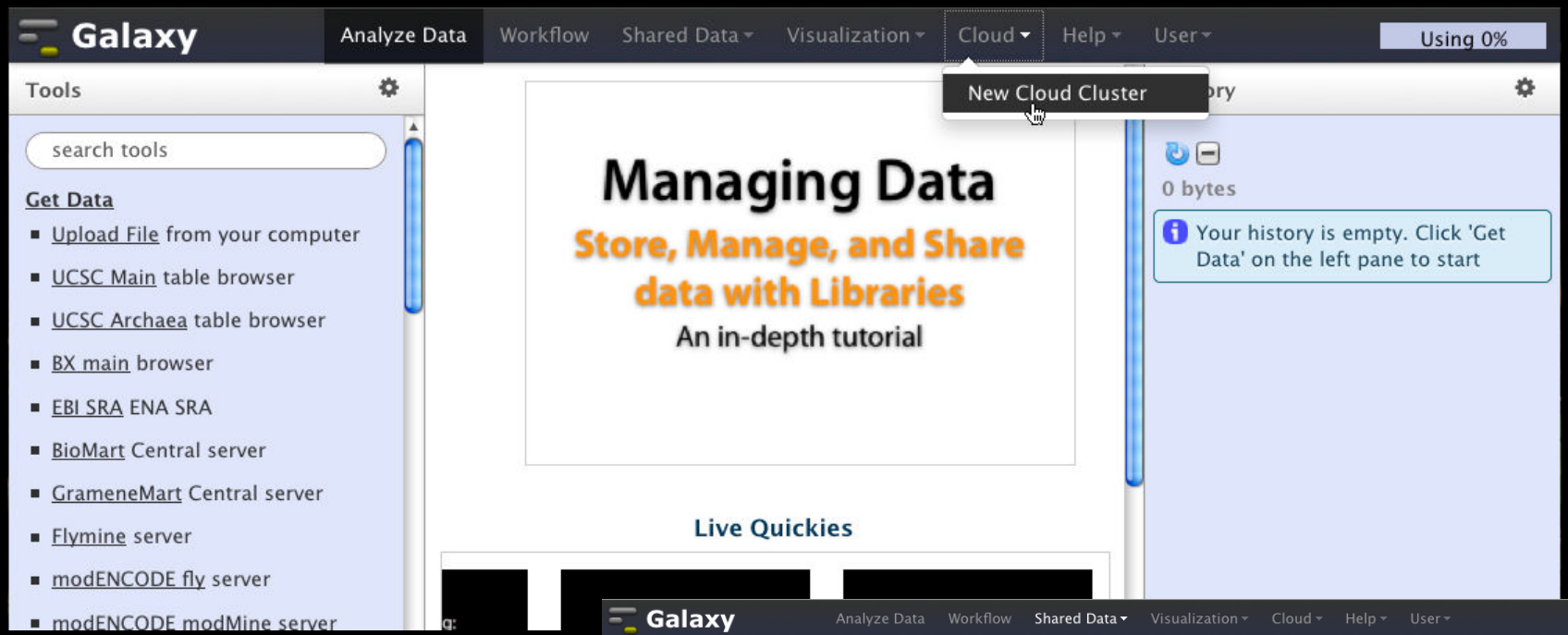
### Step 1: One Time Amazon Setup

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

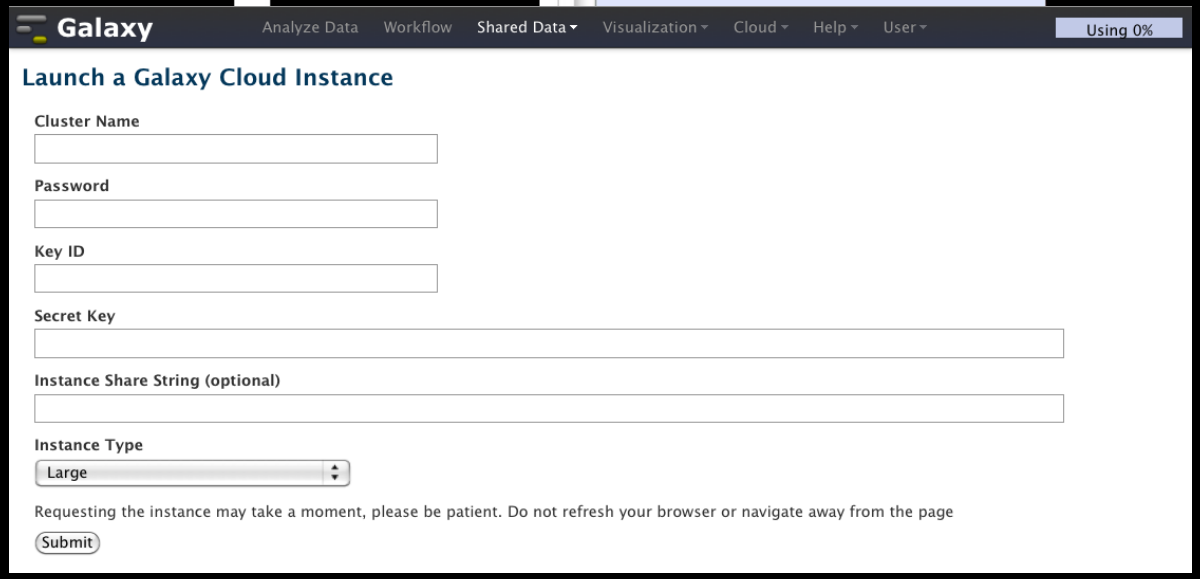
**Step 1 Screenshots**



# Instant CloudMan




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



# Visualize

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

UCSC, Ensembl, GBrowse, IGV

**Trackster:** Galaxy's genome browser

# Trackster

View your data from within Galaxy

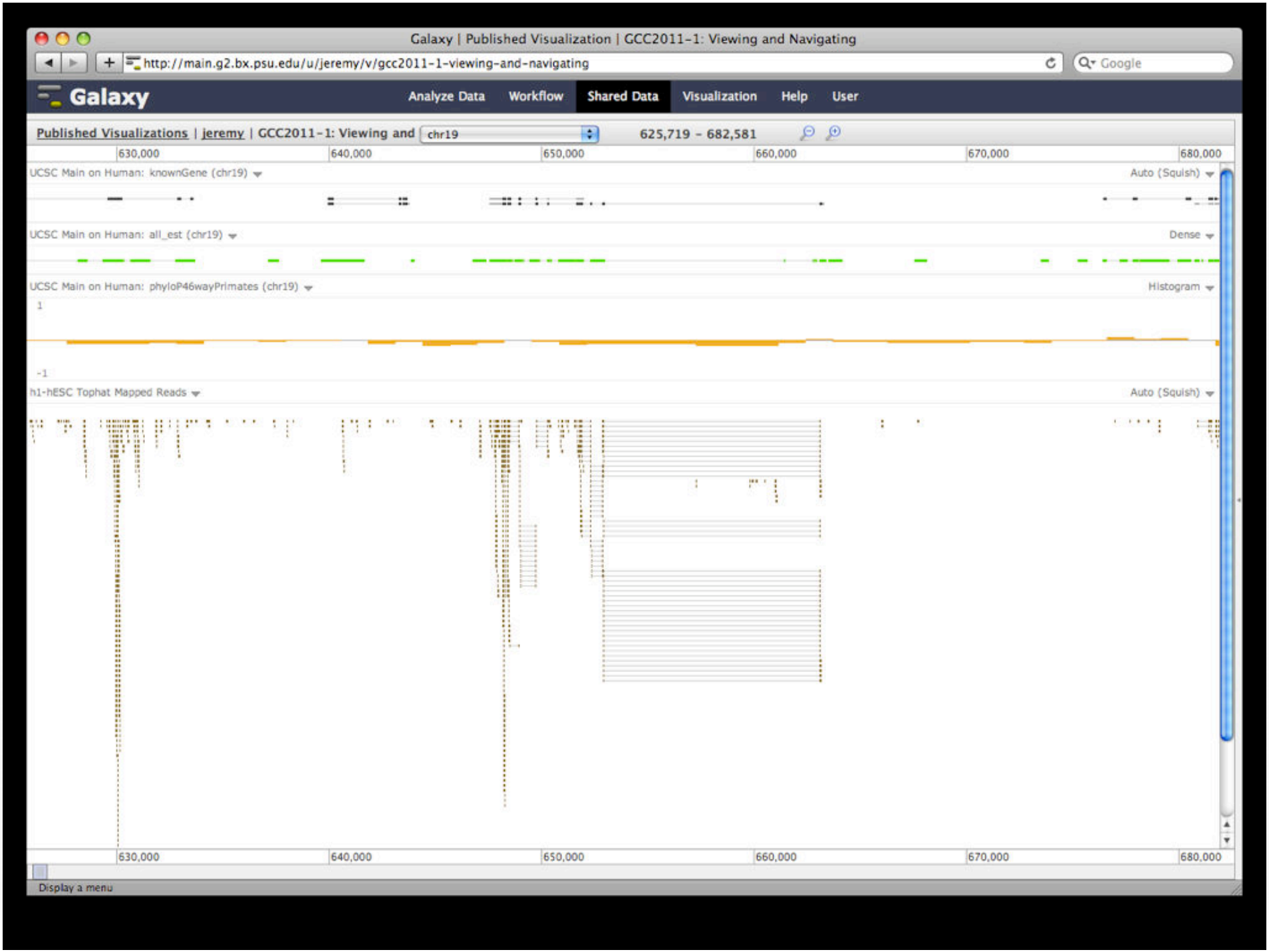
- No data transfers to external site
- Use it locally, even without internet access

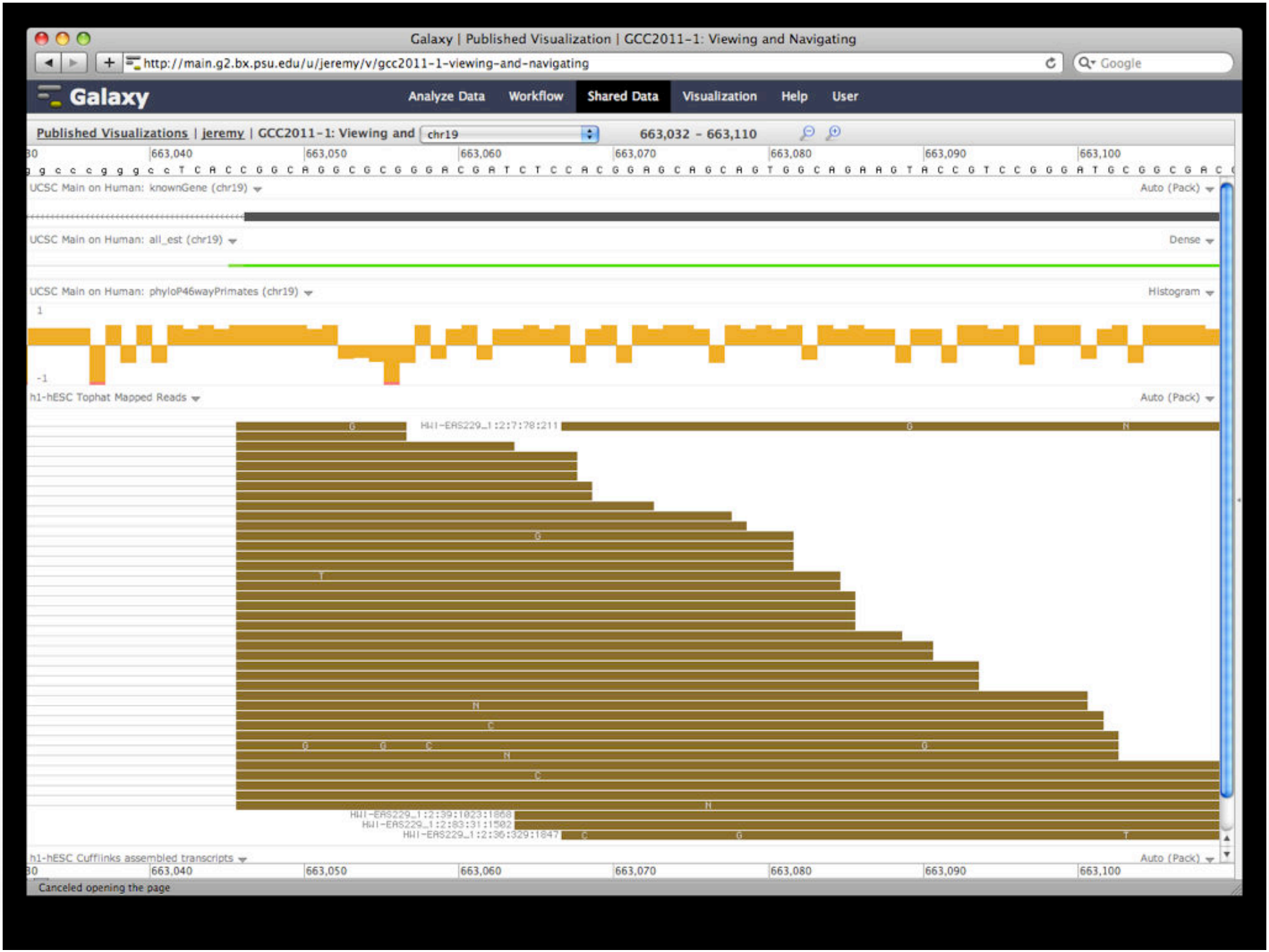
Supports common filetypes

- BAM, BED, GFF/GTF, WIG

Unique features

- custom genomes
- highly interactive


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)
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
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-1: Viewing and
chr19
663,032 - 663,110
663,040
663,050
663,060
663,070
663,080
663,090
663,100
UCSC Main on Human: knownGene (chr19)
Auto (Pack)
UCSC Main on Human: all_est (chr19)
Dense
UCSC Main on Human: phyloP46wayPrimates (chr19)
Histogram
h1-hESC Tophat Mapped Reads
Auto (Pack)
h1-hESC Cufflinks assembled transcripts
Auto (Pack)
Canceled opening the page

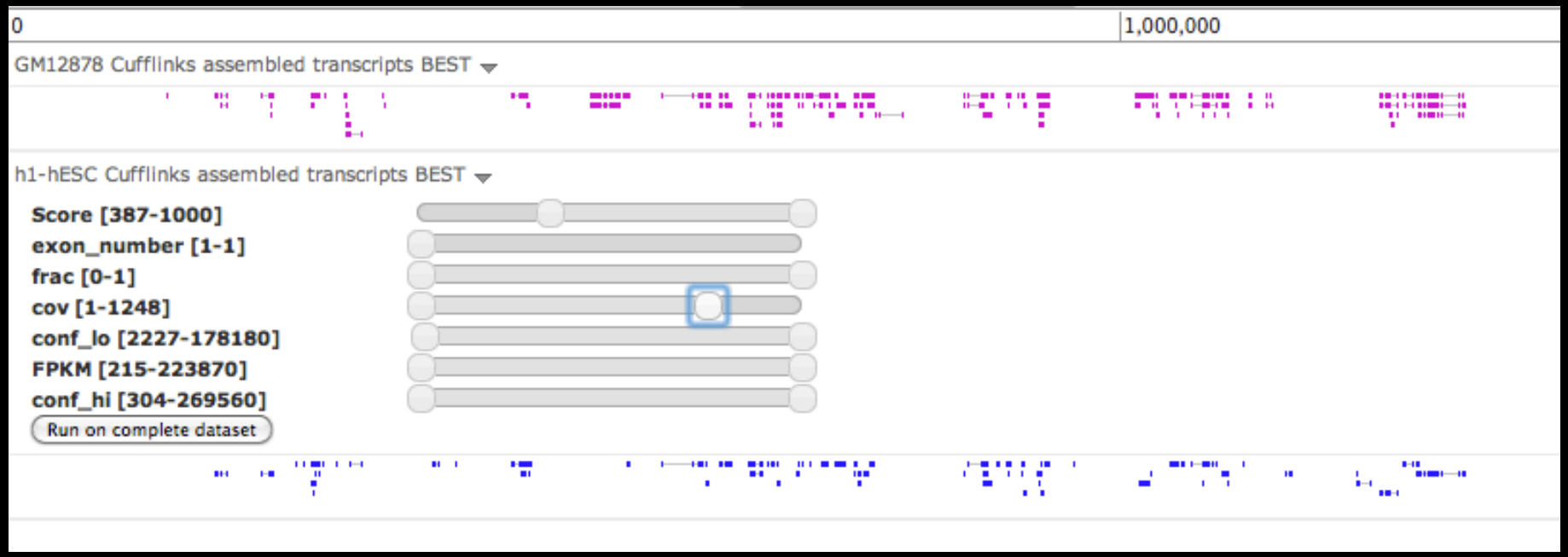
## But really, why *another* genome browser

From static browsing to visual analysis

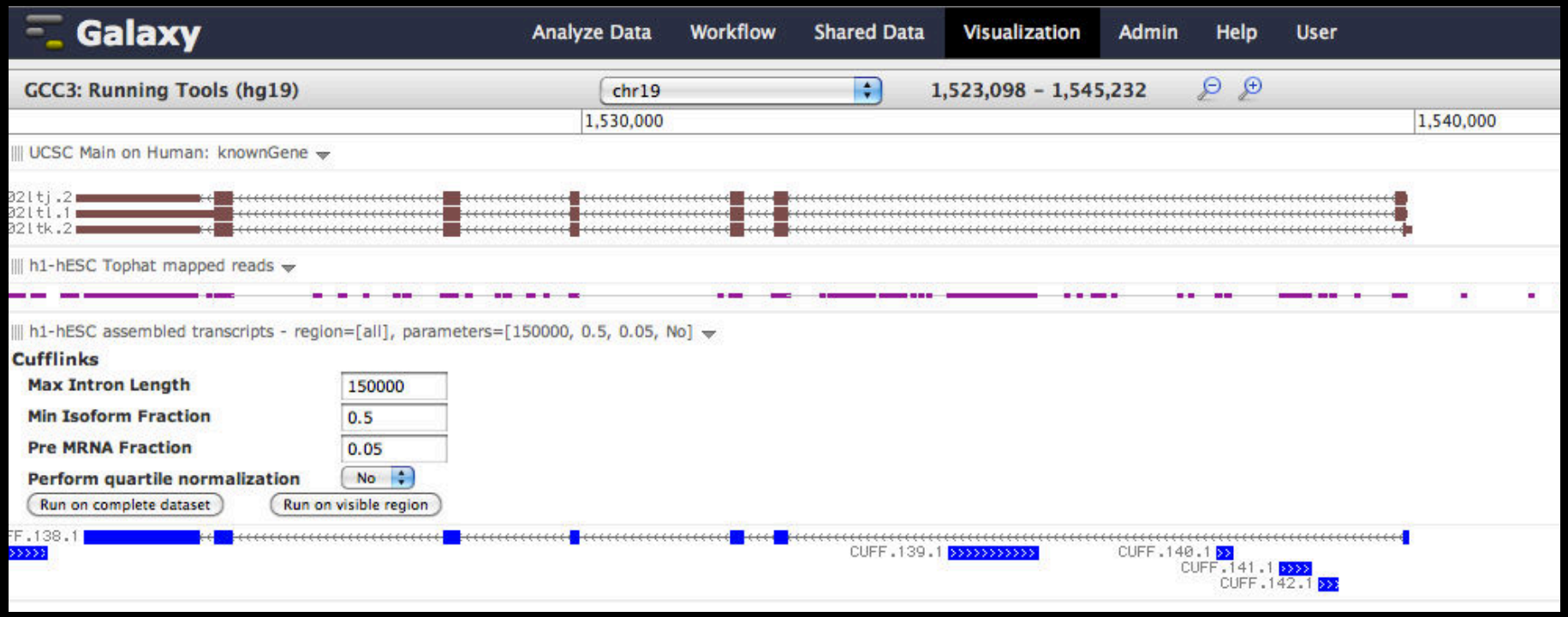
Visual feedback and experimentation needed for complex tools with many parameters

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

# Dynamic Filtering

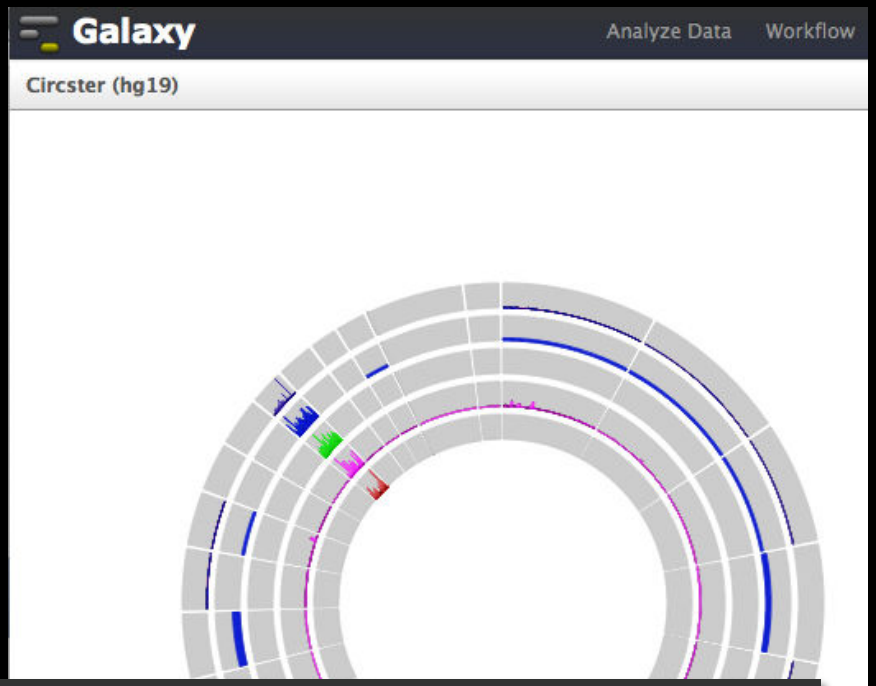


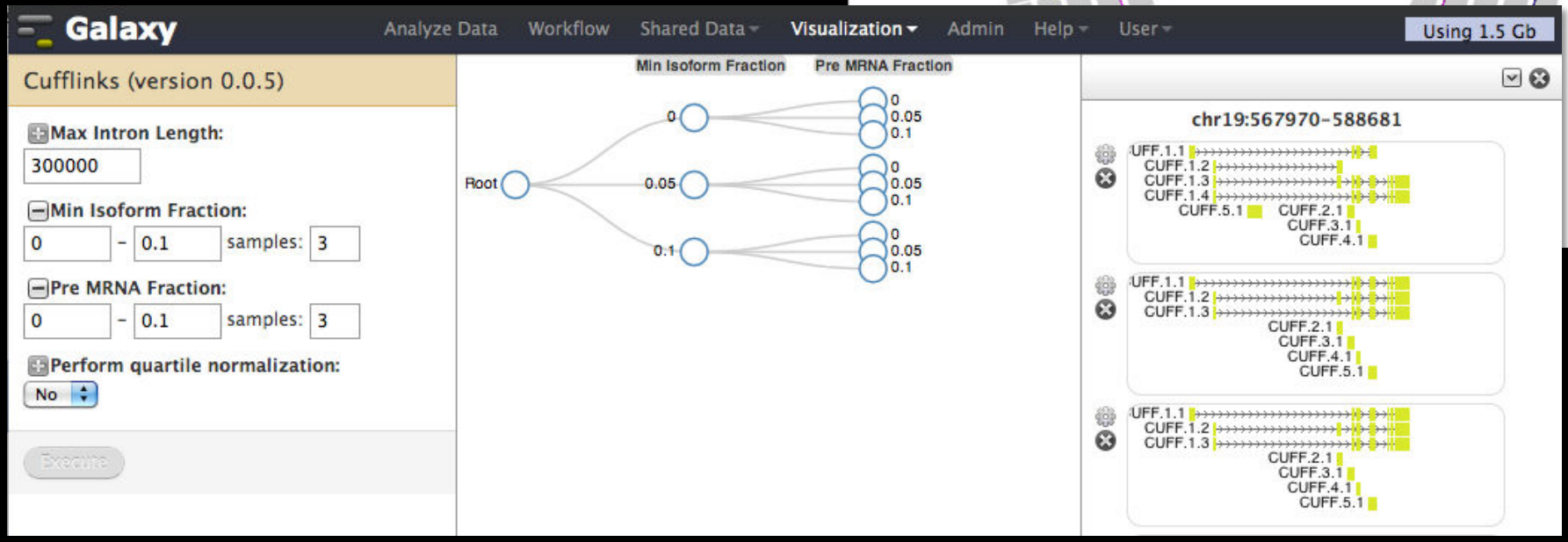
# Integrating Tools and Visualization



Trackster enables other integrations and visualizations:

Parameter sweeps
Circster



# Galaxy Community

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

http://galaxyproject.org/wiki


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