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

Galaxy

Acknowledgements

Richard LeDuc
William Barnett
Scott Michaels
Radhika Khetani

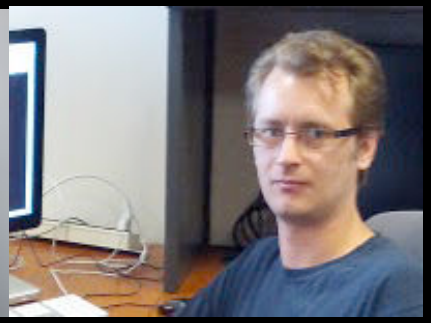
National Center for Genome Analysis Support (NCGAS)
The Center for Genomics and Bioinformatics
Indiana University


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

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Admin | Help | User | Using 158.2 GB

Tools

search tools

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

Additional output created by MACS (MACS_in_Galaxy)

Additional 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:

```
INFO  @ Wed, 21 Sep 2011 18:28:58:
# ARGUMENTS LIST:
# name = MACS_in_Galaxy
# format = SAM
# ChIP-seq file = /galaxy/main_database/files/003/013/dataset_3013610.dat
# control file = /galaxy/main_database/files/003/013/dataset_3013609.dat
# effective genome size = 1.87e+09
# tag size = 36
# band width = 300
# model fold = 32
# pvalue cutoff = 1.00e-05
# Ranges for calculating regional lambda are : peak_region,1000,5000,1000
INFO  @ Wed, 21 Sep 2011 18:28:58: #1 read tag files...
INFO  @ Wed, 21 Sep 2011 18:28:58: #1 read treatment tags...
INFO  @ Wed, 21 Sep 2011 18:29:05: #1.2 read input tags...
INFO  @ Wed, 21 Sep 2011 18:29:20: #1  Background Redundant rate: 0.02
INFO  @ Wed, 21 Sep 2011 18:29:20: #1 finished!
INFO  @ Wed, 21 Sep 2011 18:29:20: #2 Build Peak Model...
INFO  @ Wed, 21 Sep 2011 18:29:33: #2 number of paired peaks: 16551
INFO  @ Wed, 21 Sep 2011 18:29:33: #2 finished!
INFO  @ Wed, 21 Sep 2011 18:29:33: #2.2 Generate R script for model : MAC
INFO  @ Wed, 21 Sep 2011 18:29:33: #3 Call peaks...
INFO  @ Wed, 21 Sep 2011 18:29:33: #3 shift treatment data
INFO  @ Wed, 21 Sep 2011 18:29:33: #3 merge +/- strand of treatment data
INFO  @ Wed, 21 Sep 2011 18:29:34: #3 save the shifted and merged tag cou
INFO  @ Wed, 21 Sep 2011 18:29:34: write to MACS_in_Galaxy_MACS_wiggle/tr
INFO  @ Wed, 21 Sep 2011 18:31:04: compress the wiggle file using gzip...
INFO  @ Wed, 21 Sep 2011 18:31:20: #3 call peak candidates
INFO  @ Wed, 21 Sep 2011 18:31:32: #3 shift control data
INFO  @ Wed, 21 Sep 2011 18:31:32: #3 merge +/- strand of control data
INFO  @ Wed, 21 Sep 2011 18:31:32: #3 save the shifted and merged tag cou
INFO  @ Wed, 21 Sep 2011 18:31:32: write to MACS_in_Galaxy_MACS_wiggle/co
```

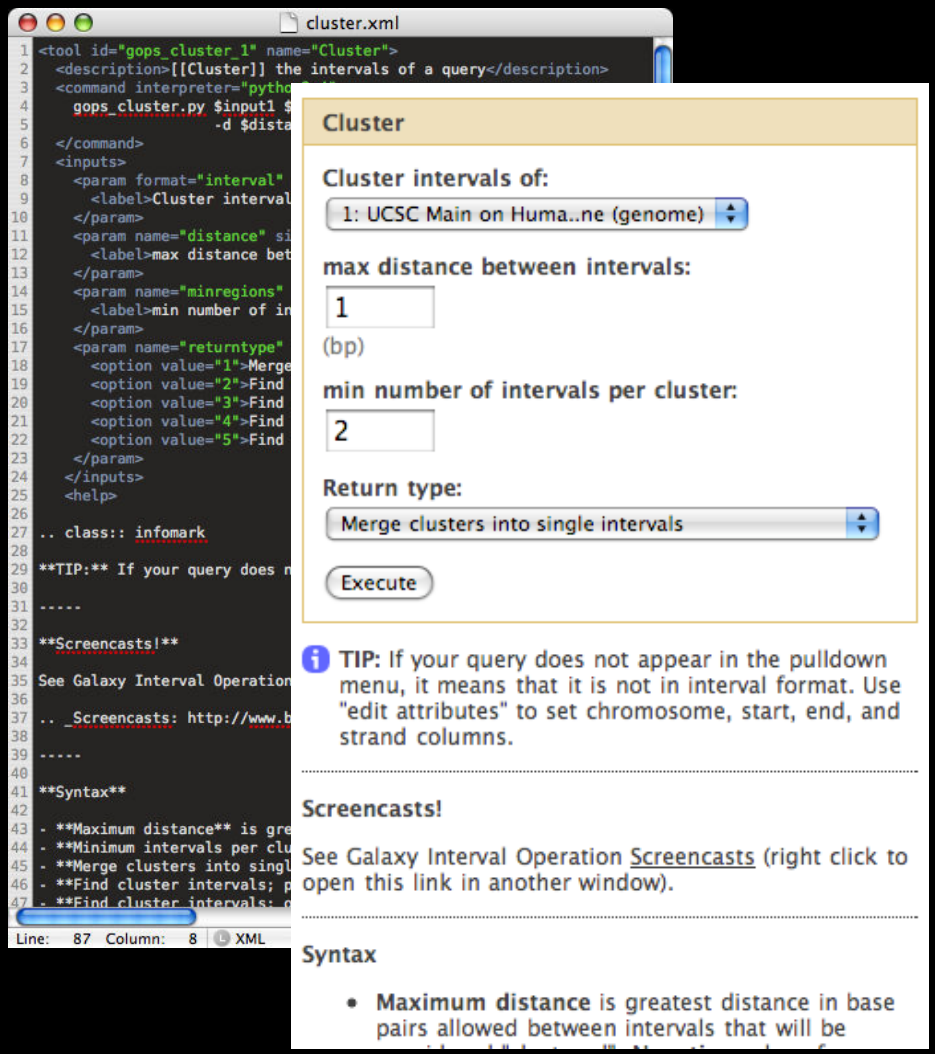
History

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
HTML file

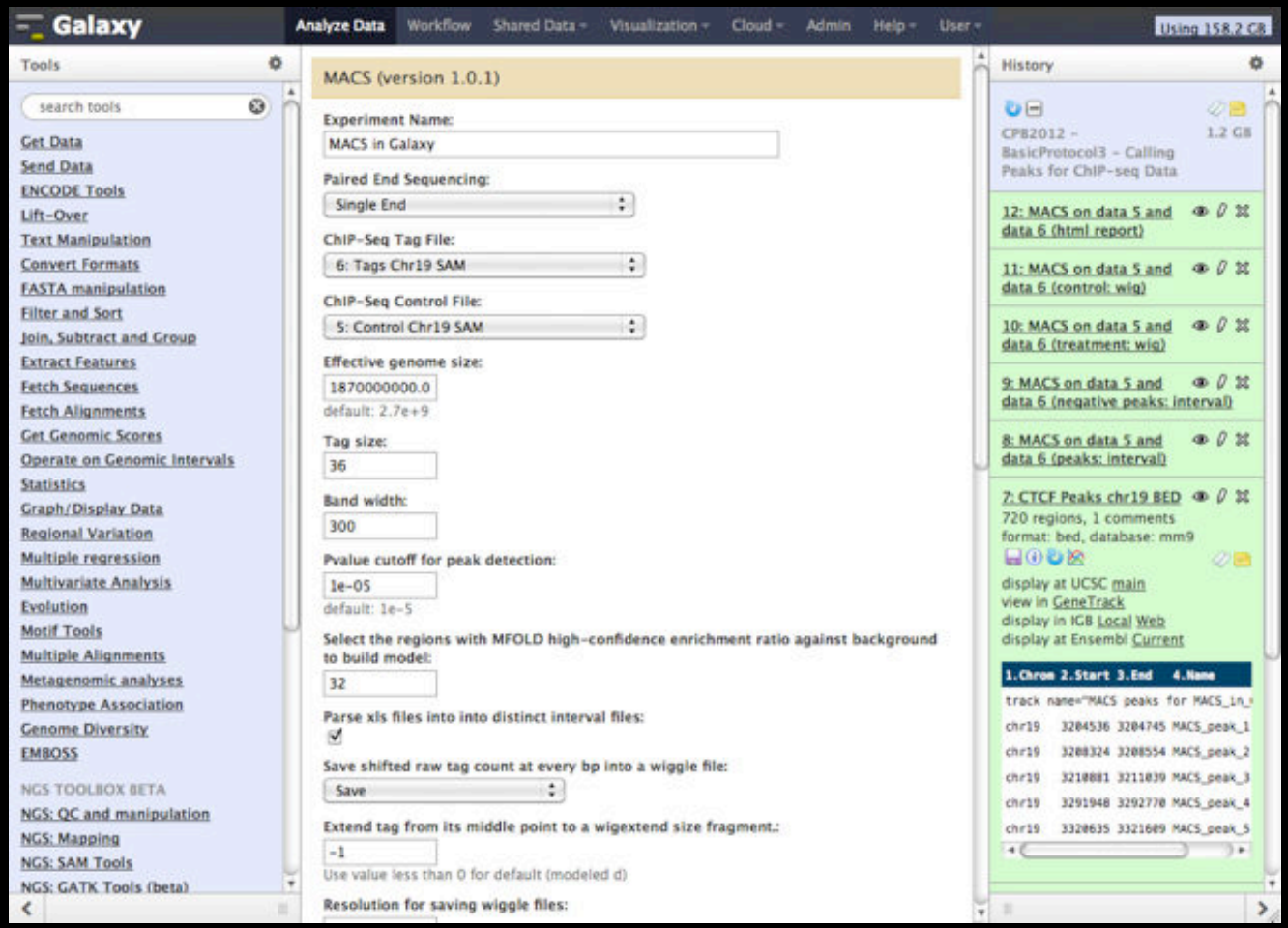
11: MACS on data 5 and data 6 (control: wig)
10: MACS on data 5 and data 6 (treatment: wig)
9: MACS on data 5 and data 6 (negative peaks: interval)
8: MACS on data 5 and data 6 (peaks: interval)
7: CTCF Peaks chr19 BED
6: Tags Chr19 SAM
5: Control Chr19 SAM
4: Tags Chr19 groomed
3: Control Chr19 groomed
2: Tags Chr19 ungroomed

Integrating existing tools into a uniform framework



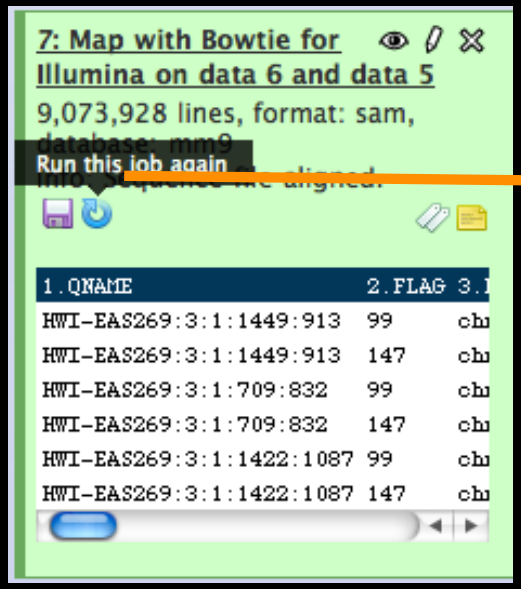
- Defined in terms of an abstract interface (inputs and outputs)
  - In practice, mostly command line tools, a declarative XML description of the interface, how to generate a command line
- Designed to be as easy as possible for tool authors, while still allowing rigorous reasoning

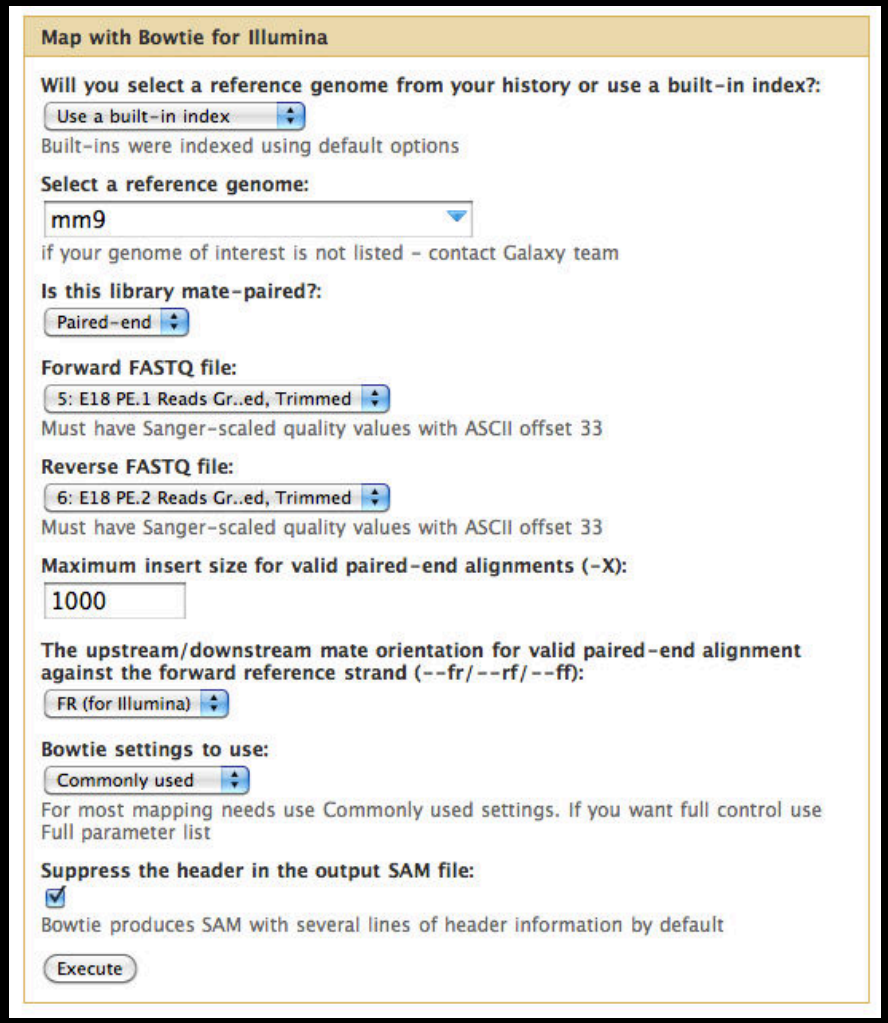
Galaxy analysis interface



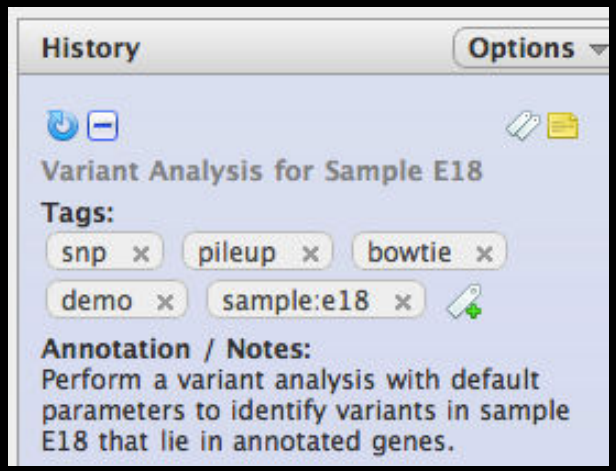
- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses
- Exact parameters of a step can always be inspected, and easily rerun

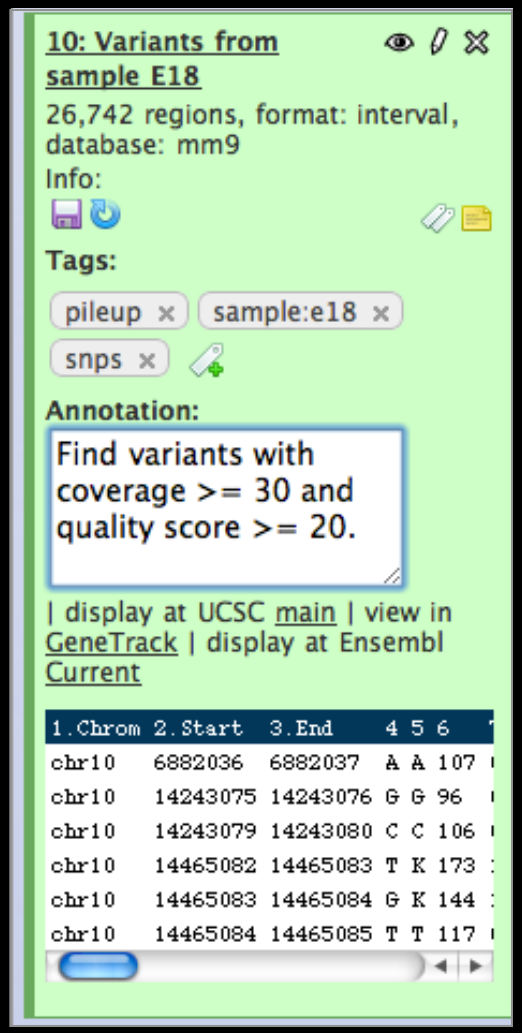
Automatically tracks every step of every analysis



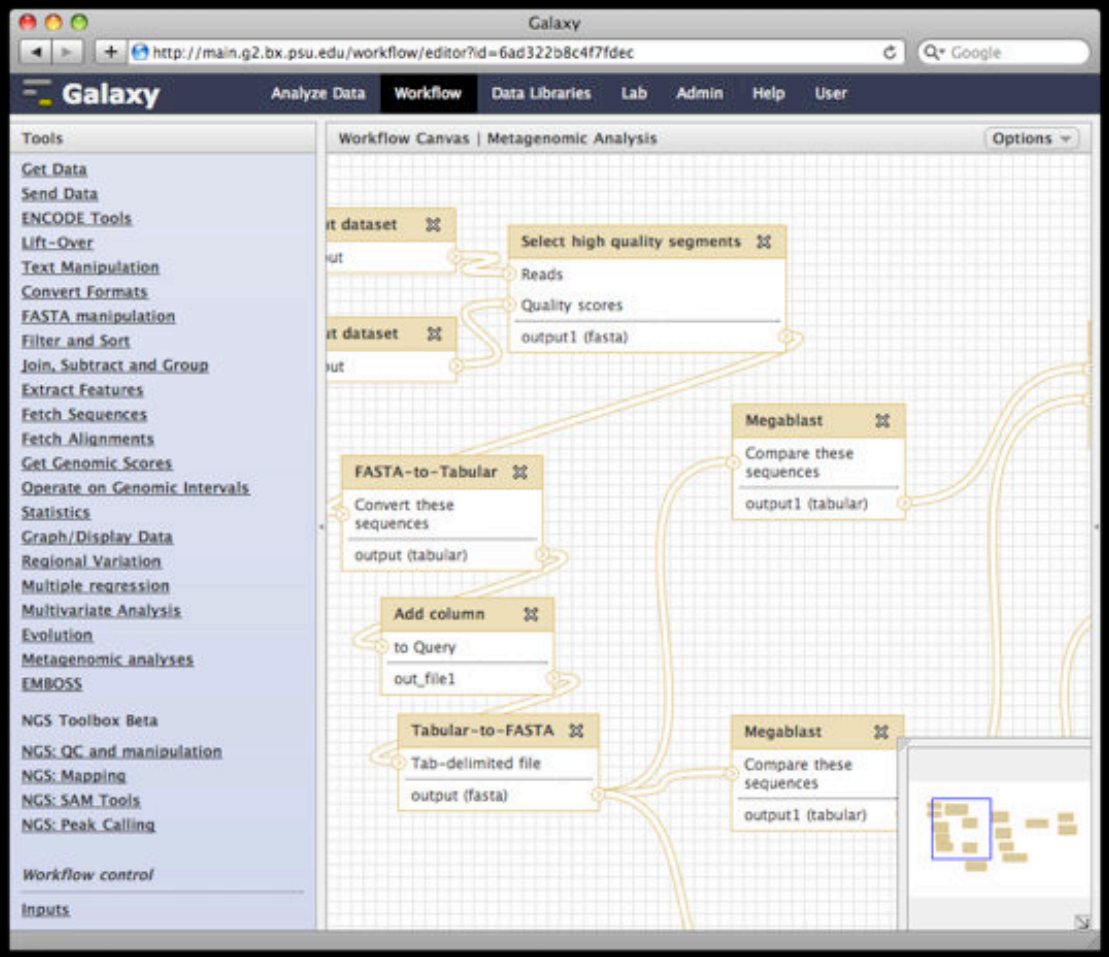


As well as user-generated metadata and annotation...



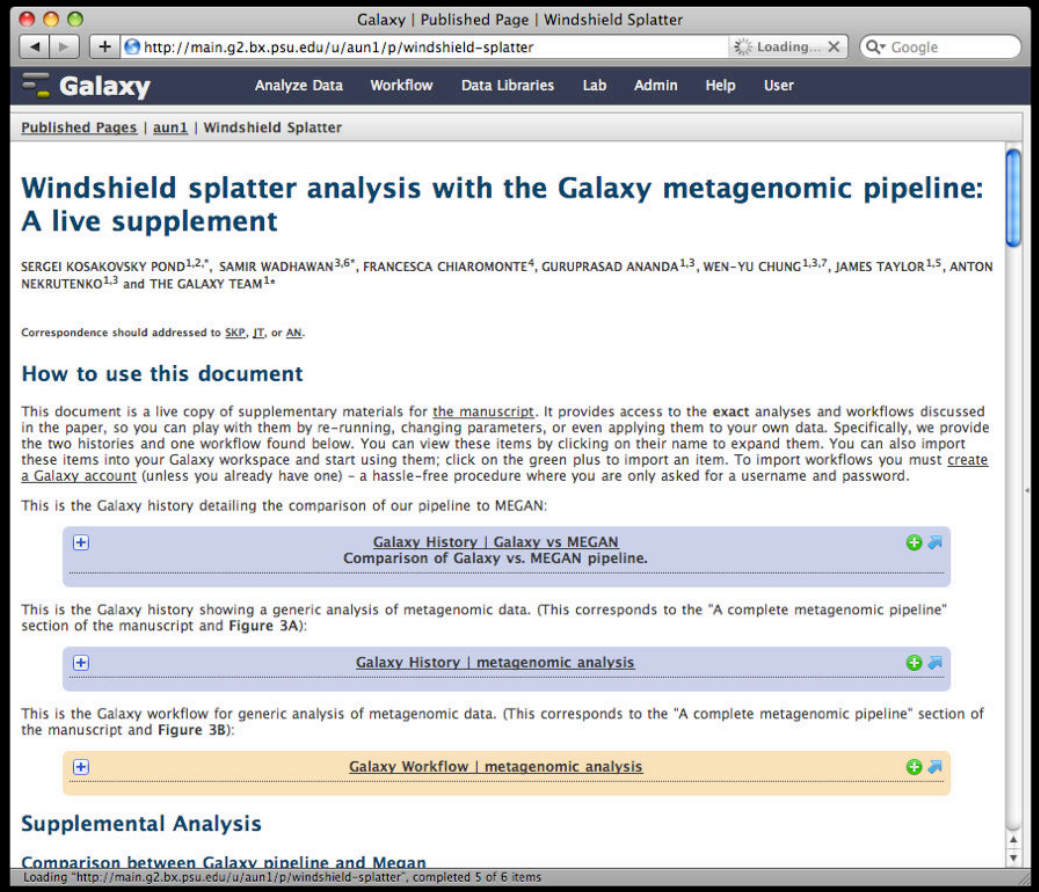


Galaxy workflow system



- 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*
- Annotation and Galaxy Pages allow analyses to be augmented with textual content and provided in the form of an integrated

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*

Full Text (PDF) *Free*

Current Issue

October 2010, 20 (10)

[1,3,7], JAMES TAYLOR[1,5], ANTON

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

Comparison of Galaxy vs. MEGAN pipeline.

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

This is the Galaxy wo the manuscript and F

Footnotes

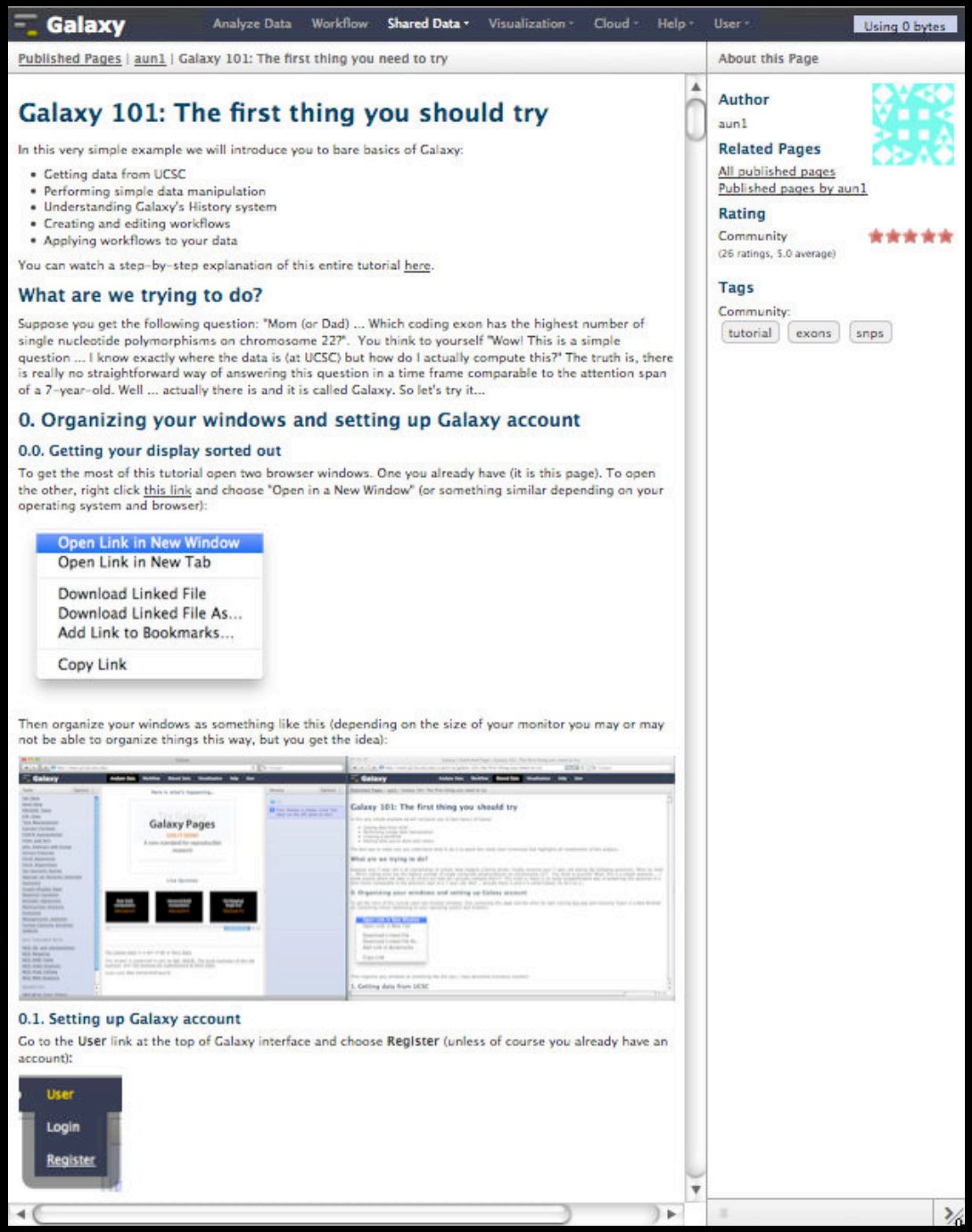
[Supplemental material is available online at http://www.genome.org. All data and tools described in this manuscript can be downloaded or used directly at http://galaxyproject.org. Exact analyses and workflows used in this paper are available at http://usegalaxy.org/u/aun1/p/windshield-splatter.]

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

Give it a spin: usegalaxy.org/galaxy101



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*.

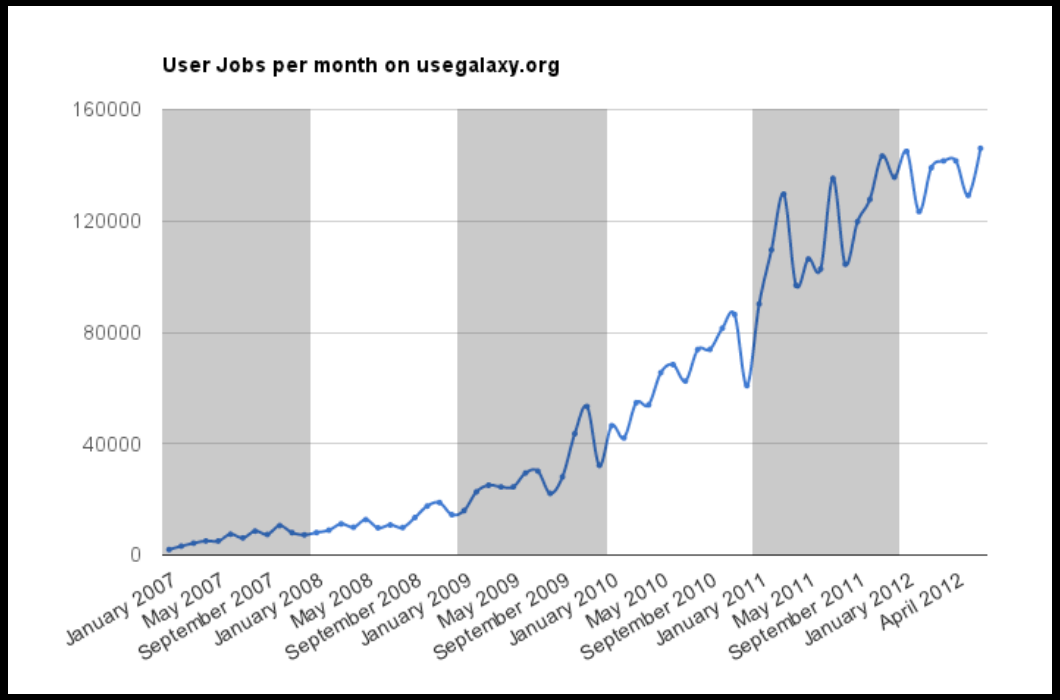
Galaxy is available ...

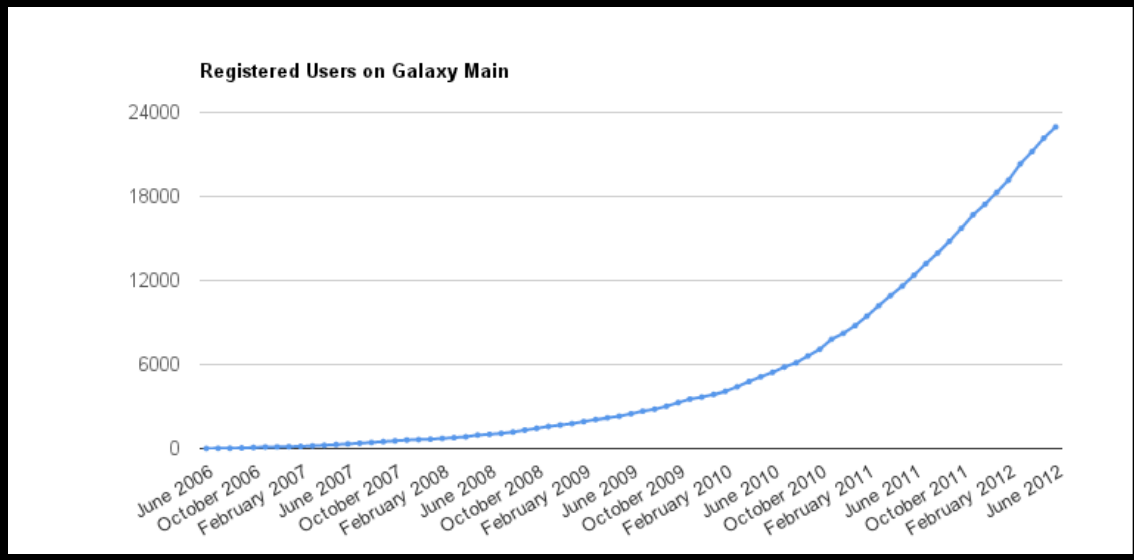
- **As a free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

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 pair end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

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

AB-SOLID DATA

- Convert SOLiD output to fast
- Compute quality statistics fo 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 sco
- 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

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

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.

CLUSTER RESOURCES TORQUE

GRID ENGINE

Platform Computing

DRMAA
Distributed Resource Management Application API — www.drmaa.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'll use Amazon for the *Galaxy for Biologists* workshop later today.**

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

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

Step 1: One Time Amazon Setup

1. Because AWS services implement pay-as-you-go access model for compute resources, it is necessary for every user of the service to *register with Amazon*. You will need a credit card to register. (You can apply for a AWS Education Grant after you register).
2. Once your account has been approved by Amazon (note that this may take up to

Step 1 Screenshots

Instant CloudMan

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

New Cloud Cluster

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

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

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

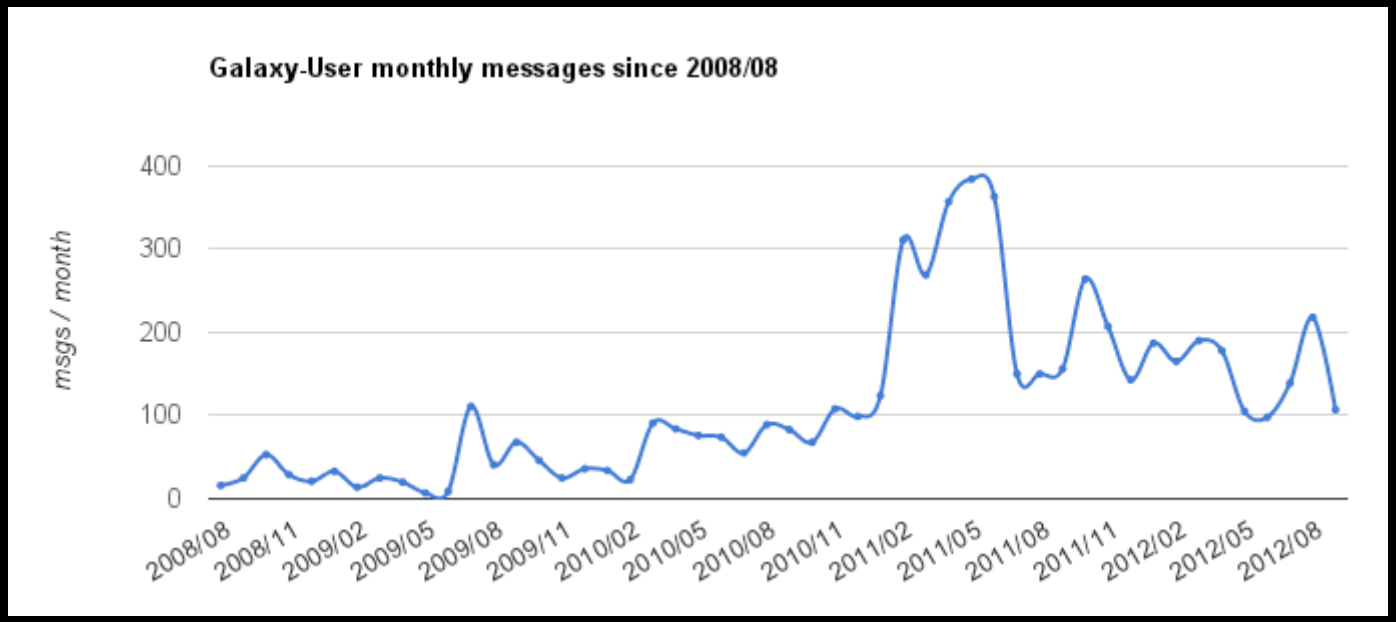
Submit

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 Meting

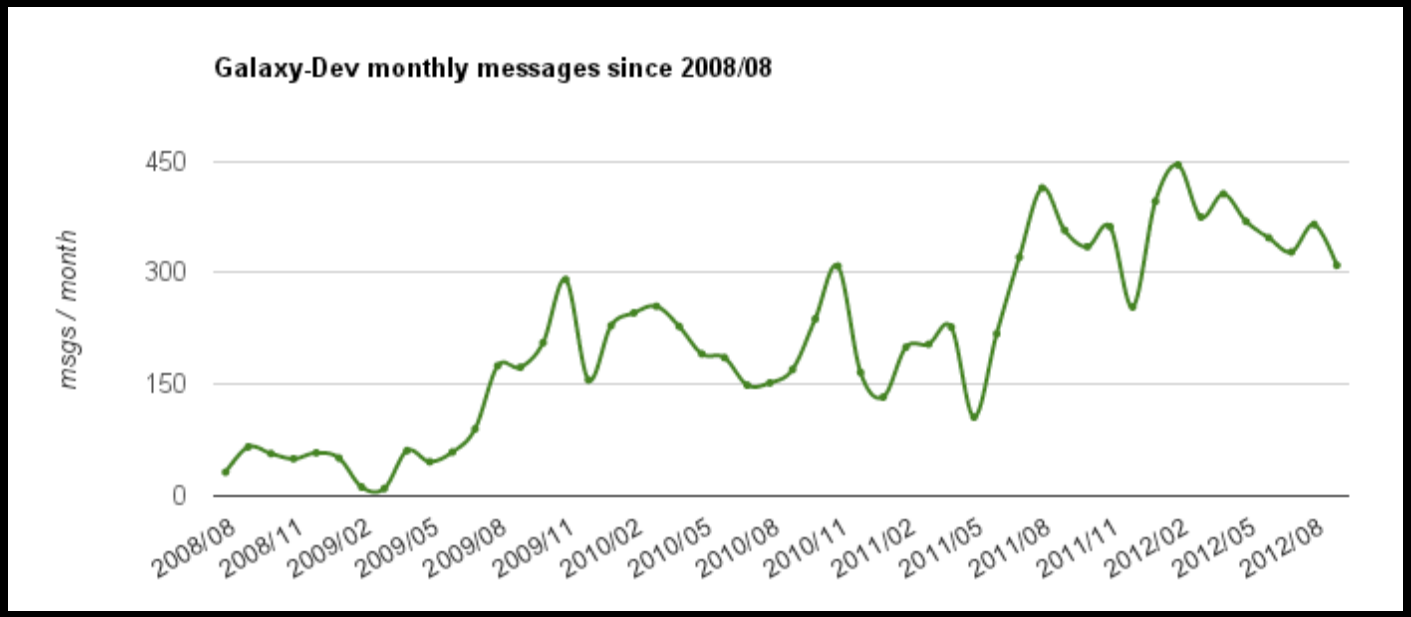
http://galaxyproject.org/wiki

Mailing Lists



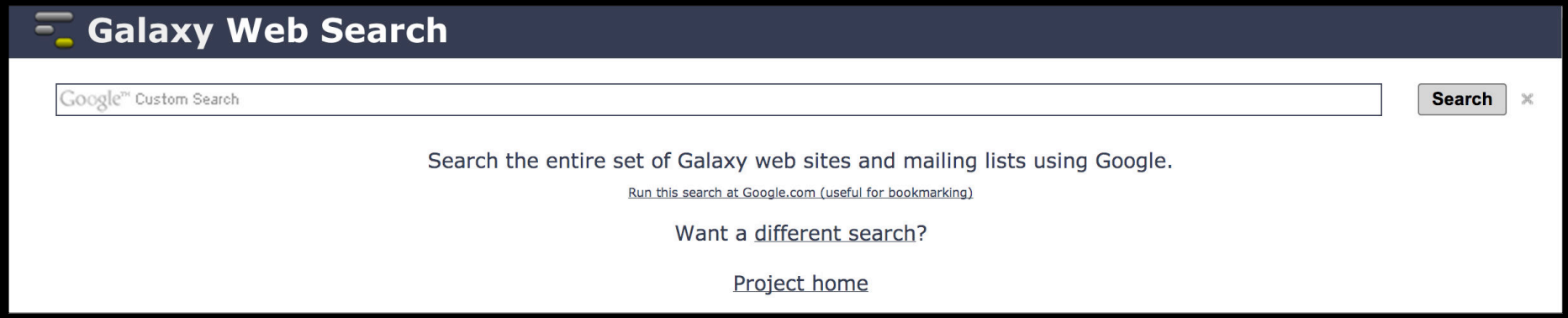
Galaxy-User
messages / month

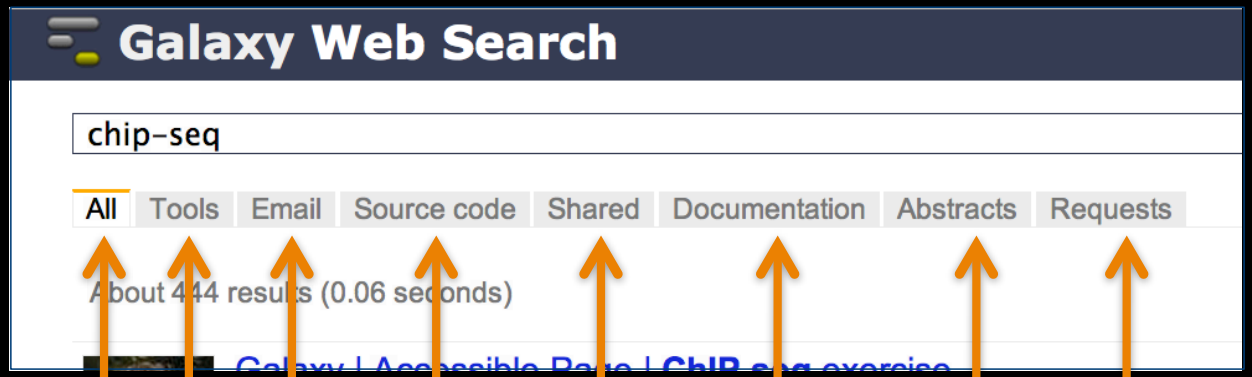
Galaxy-Dev
messages / month



http://galaxyproject.org/wiki/MailingLists

Galaxy Search: http://galaxyproject.org/search





Find

Everything on …

Tools for …

Email about …

Source code for …

Published Histories, Pages, Workflows, about …

Documentation on …

Papers using Galaxy for …

Related feature requests

http://galaxyproject.org/GCC2013

Other Upcoming Galaxy Events

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

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

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 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

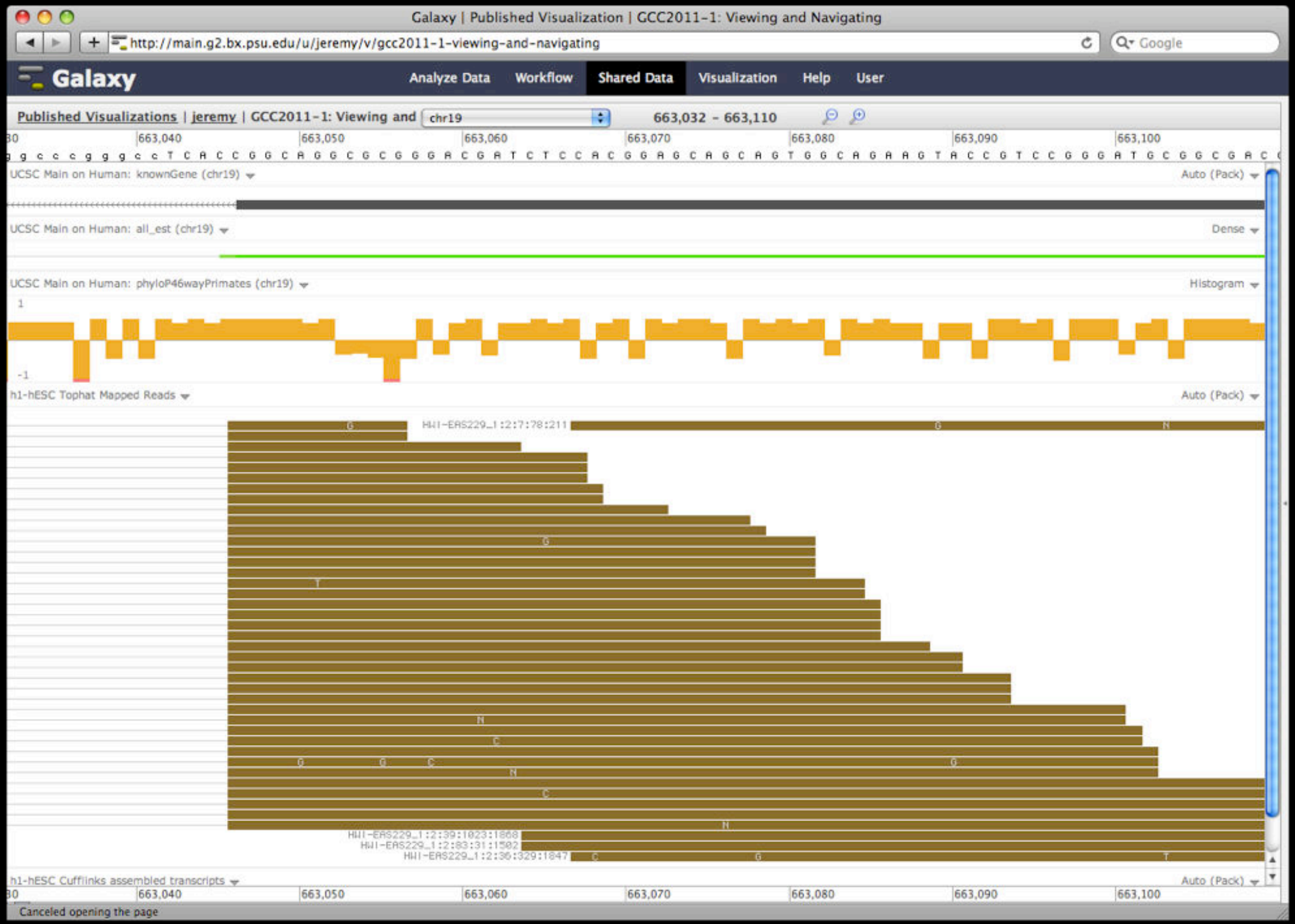
1

-1

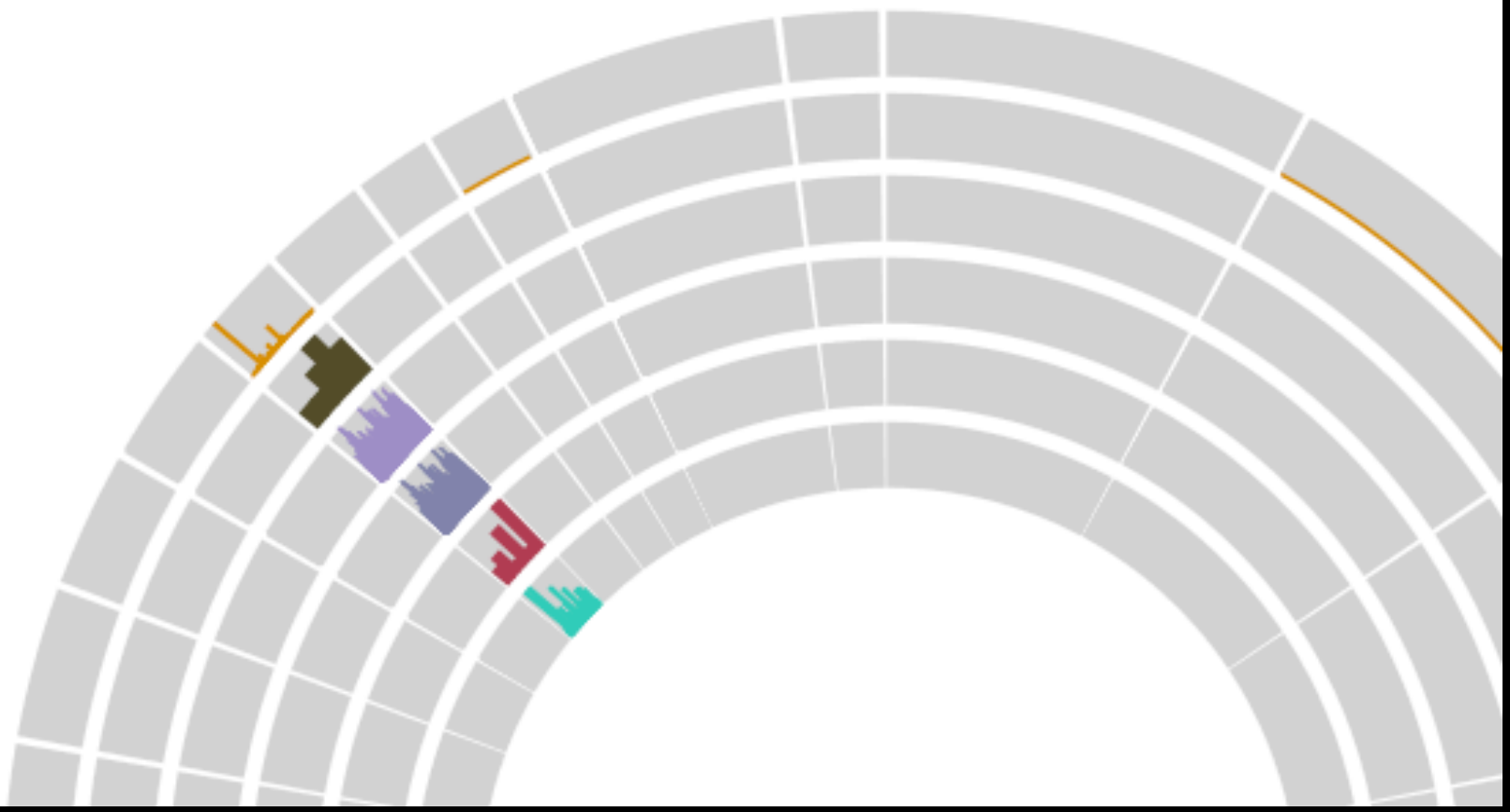
h1-hESC Tophat Mapped Reads Auto (Squish)

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
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

Brain / Adrenal Chr19 (hg19)



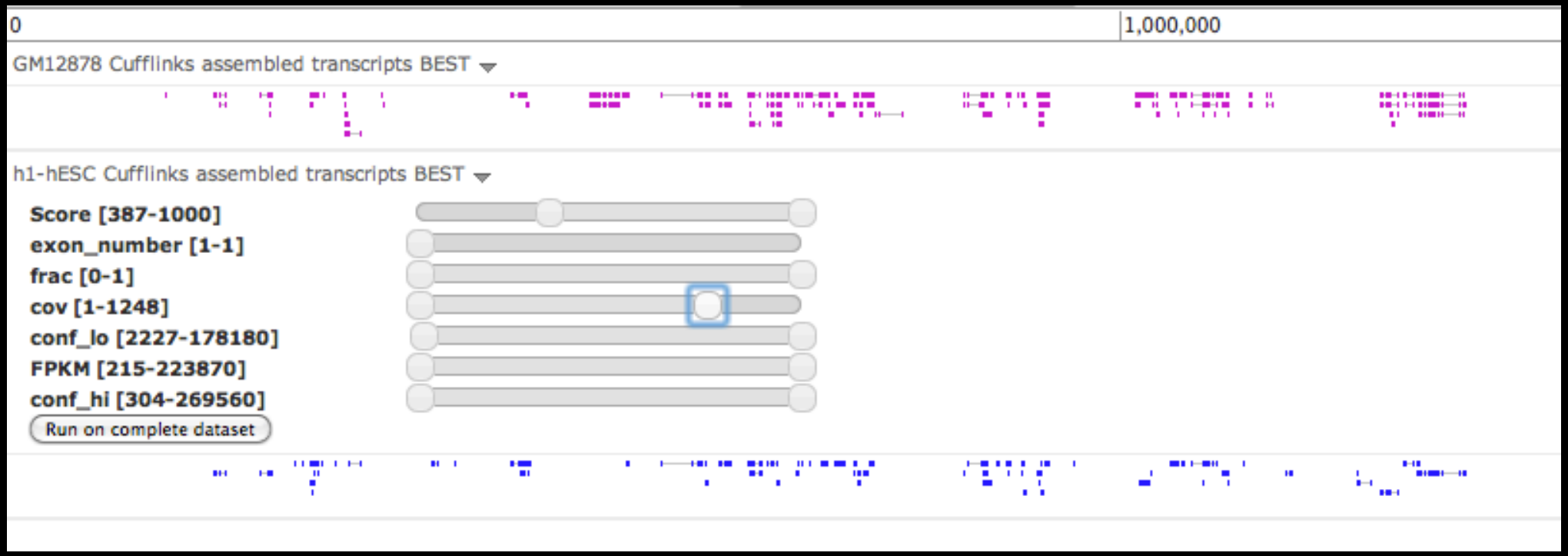
But really, why *another* genome browser

From static browsing to visual analysis

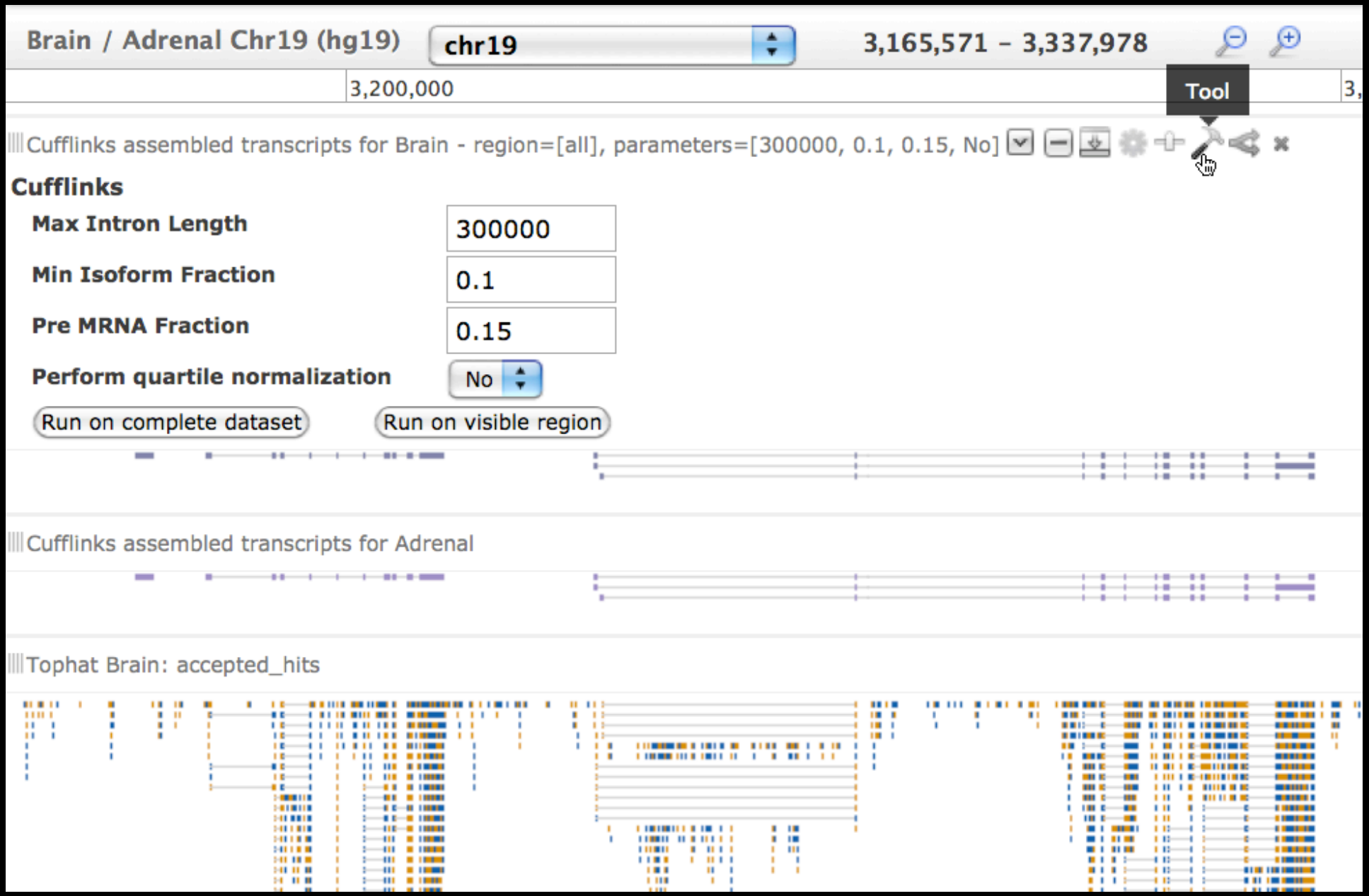
Visual feedback and experimentation needed for complex tools with many parameters

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

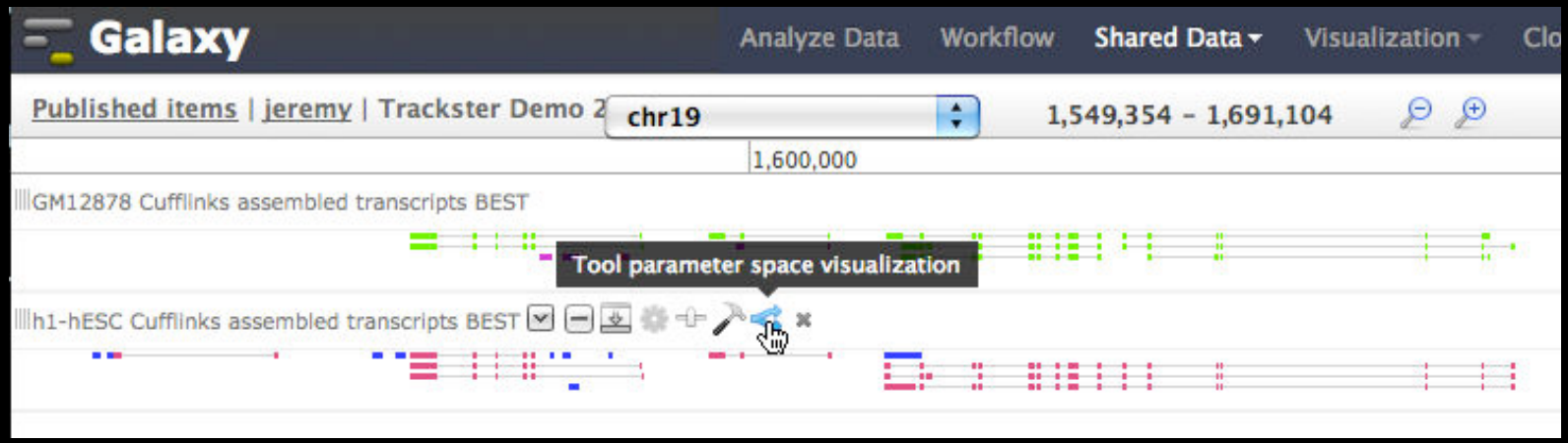
Dynamic Filtering

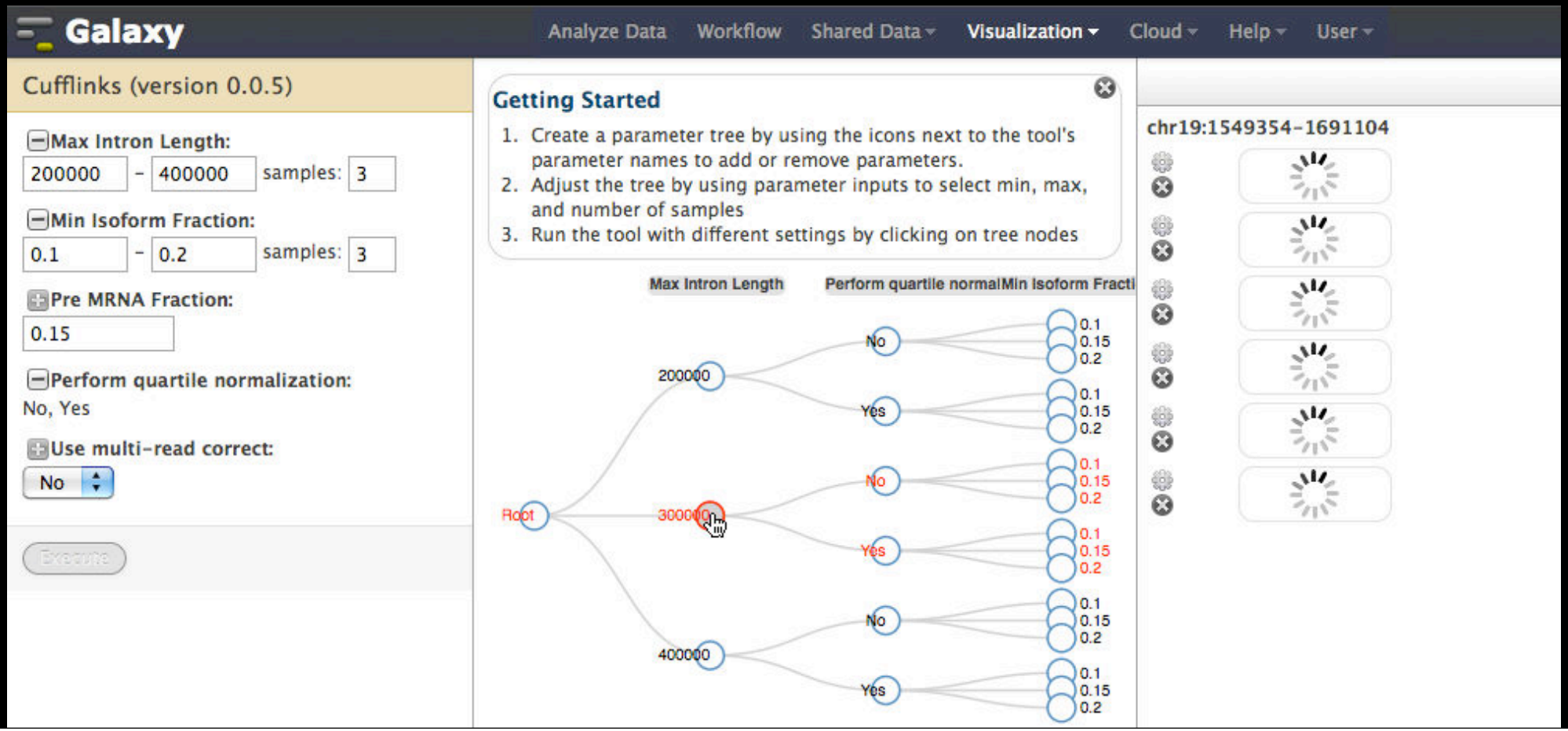


Integrating Tools and Visualization



Exploring Parameter Space with Trackster





Galaxy URLs to Remember

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

http://usegalaxy.org/galaxy101

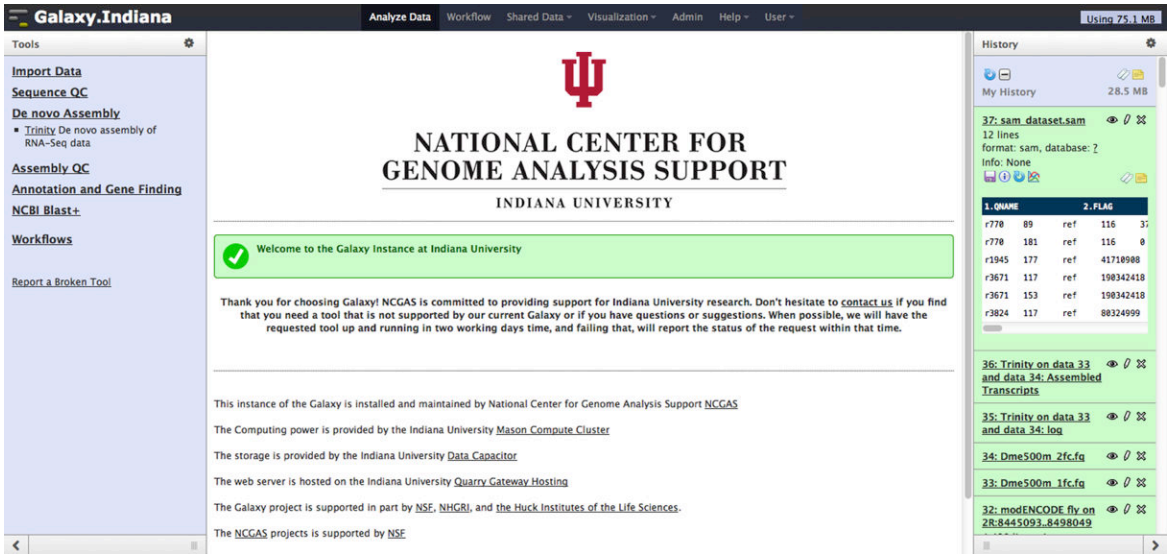
and

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



Thank you.

http://galaxyproject.org/GCC2013