Galaxy Workshop

Purdue University
22 October 2012

Dave Clements
Emory University

http://galaxyproject.org/

PURDUE UNIVERSITY
Discovery Park

PURDUE
UNIVERSITY

amazon
webservices™

Galaxy

Acknowledgements 1

Jyothi Thimmapuram
Radhika Khetani

Purdue University
Bioinformatics Core
Cyber Center
Discovery Park

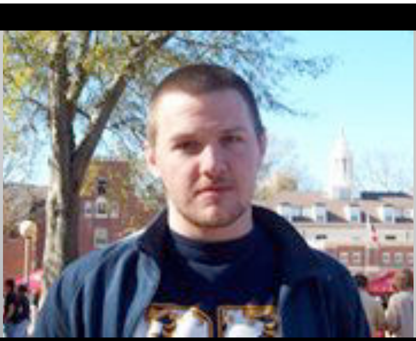
AWS Education Grant

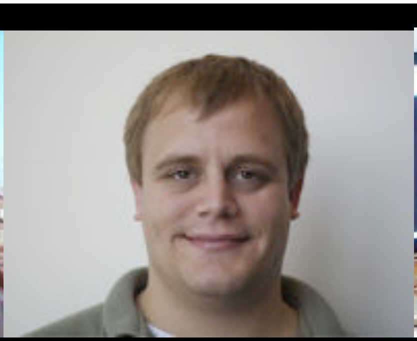
NIH
NSF
Huck Institute

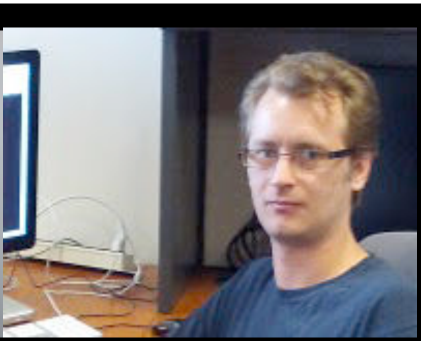
Penn StateUniversity
Emory University

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

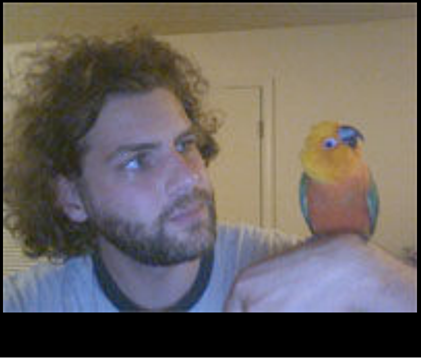
Scott McManus

Anton
Nekrutenko

The Galaxy Team

http://galaxyproject.org/wiki/GalaxyTeam

James
Taylor



Agenda

Welcome, Basic Analysis

Basic analyses into Reusable Workflows

NGS Analysis I: Through Tophat

Galaxy Project Overview

NGS Analysis II: Cufflinks, Visualization

Manage, Reuse, and Share your Analyses

Setting up your own Galaxy on the Cloud

Coffee and lunch breaks throughout the day

Slides at galaxyproject.org/wiki/Events/Purdue2012

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/gxygold

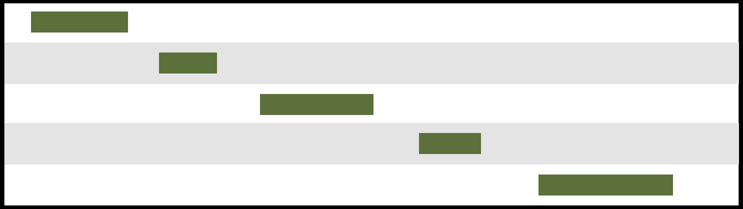
http://bit.ly/gxyblack

http://bit.ly/gxyold

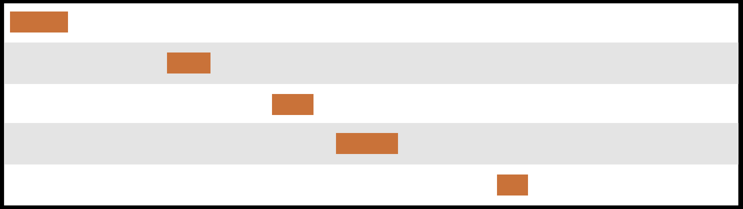
Repetitious Pigs: A Rough Plan

- 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

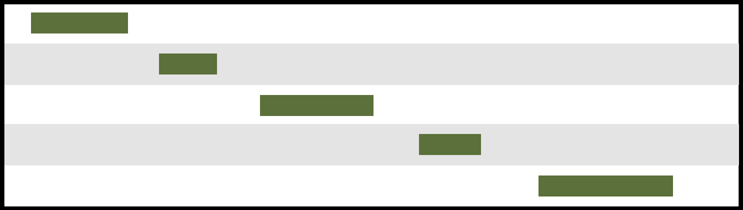
(~ http://usegalaxy.org/galaxy101)



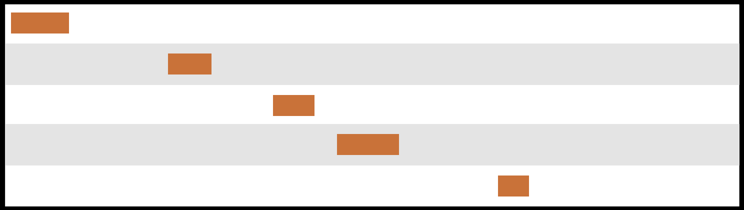
Exons, from UCSC



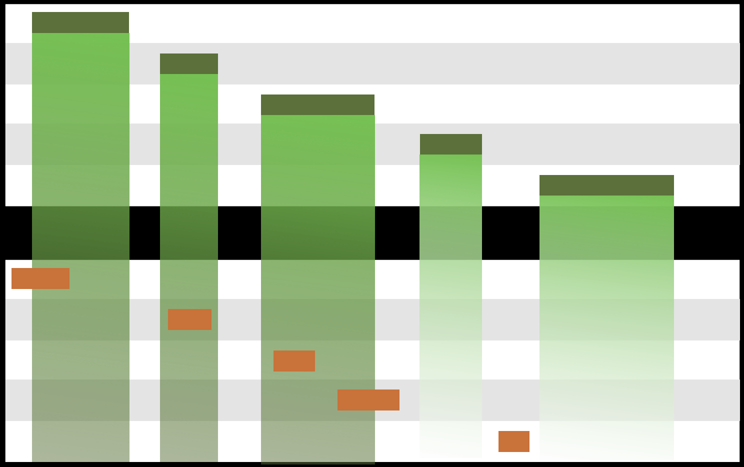
Repeats, from UCSC



Exons, from UCSC



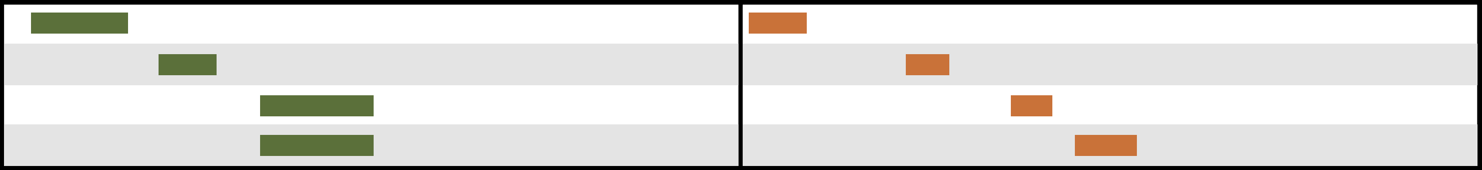
Repeats, from UCSC

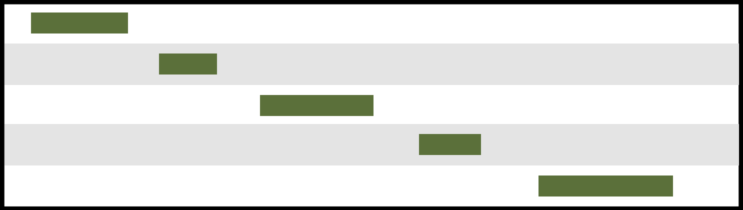


Exons, from UCSC

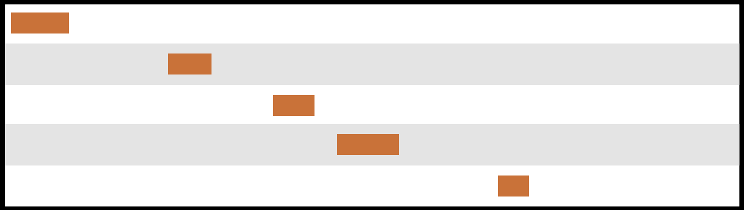
Repeats, from UCSC

Overlap pairings

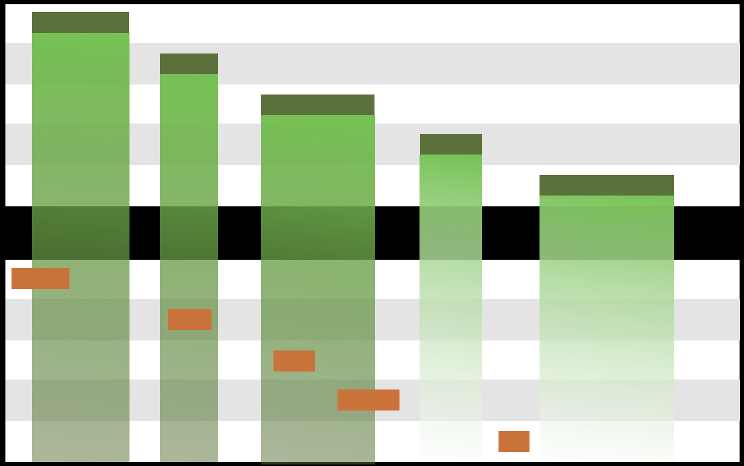




Exons, from UCSC



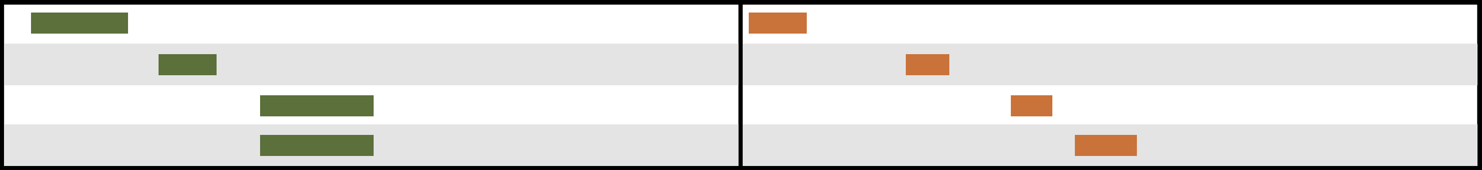
Repeats, from UCSC

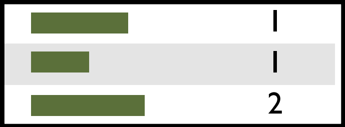


Exons, from UCSC

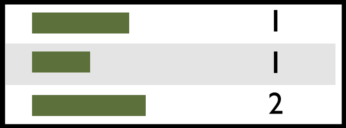
Repeats, from UCSC

Overlap pairings

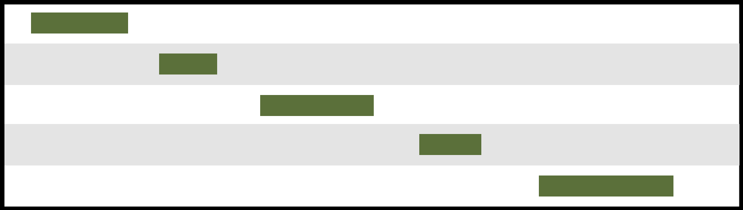




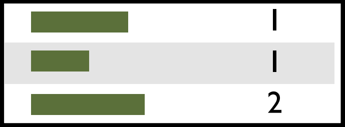
Exon overlap counts



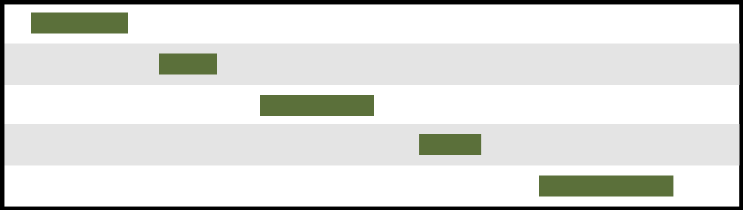
Exon overlap counts



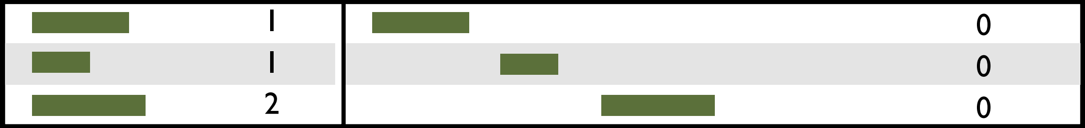
Exons, from UCSC



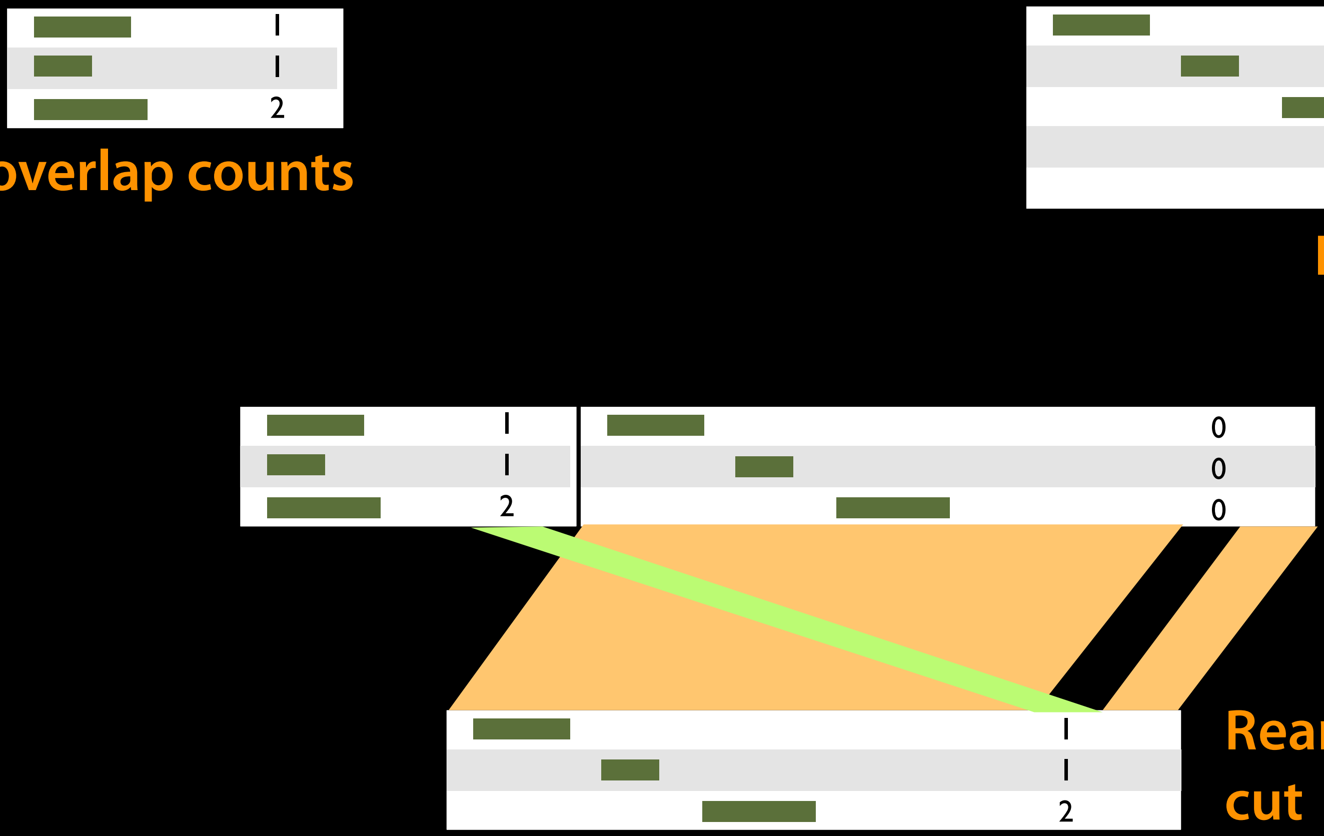
Exon overlap counts



Exons, from UCSC



Join on exon name



Exon overlap counts

Exons, from UCSC

Join on exon name

Rearrange columns w/ cut

Agenda

Welcome, Basic Analysis

Basic analyses into Reusable Workflows

NGS Analysis I: Through Tophat

Galaxy Project Overview

NGS Analysis II: Cufflinks, Visualization

Manage, Reuse, and Share your Analyses

Setting up your own Galaxy on the Cloud

Some Galaxy Terminology

Dataset:

Any input, output or intermediate set of data + metadata

Datasets from previous histories can be reused

History:

A series of inputs, analysis steps, intermediate datasets, and outputs

Current history can be cloned

Resume any previous history

Workflow:

A series of analysis steps

Can be repeated with different data

Can be extracted from any history or created from scratch

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

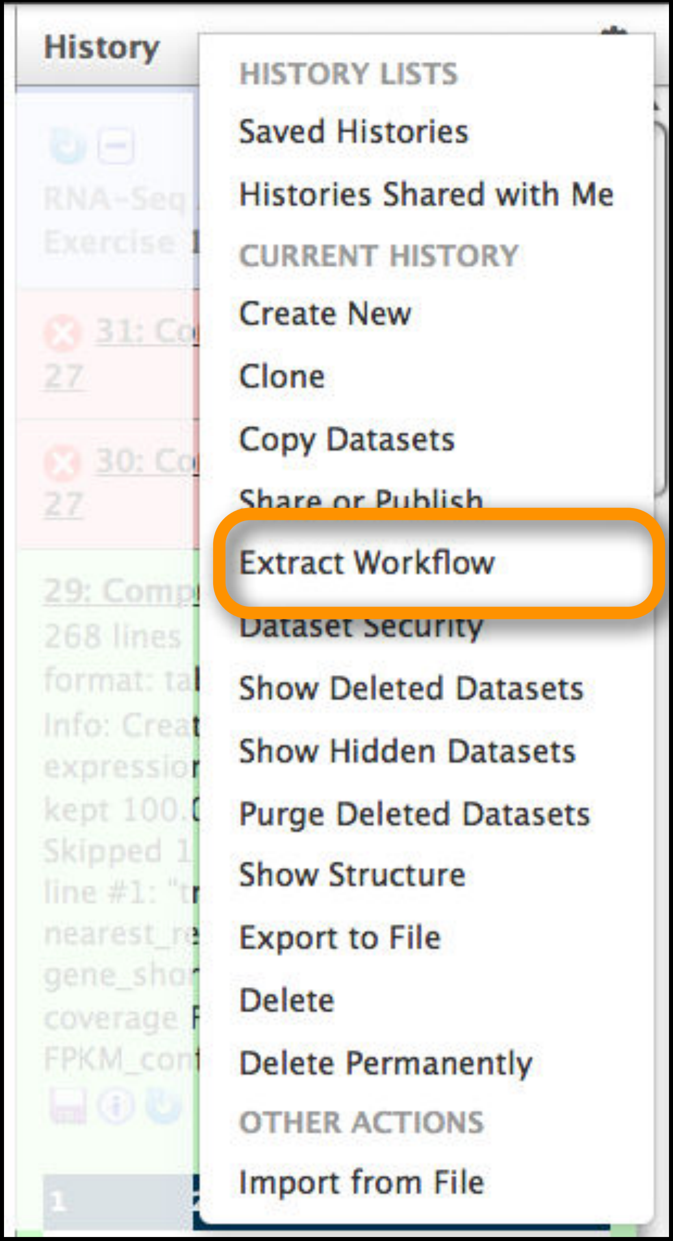
Did that work?

On your own:

Count # of exons in each repeat

Did that work? *Why not?*

Edit workflow: doc assumptions



Agenda

Welcome, Basic Analysis

Basic analyses into Reusable Workflows

NGS Analysis I: Through Tophat

Galaxy Project Overview

NGS Analysis II: Cufflinks, Visualization

Manage, Reuse, and Share your Analyses

Setting up your own Galaxy on the Cloud

RNA-seq Exercise

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

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
- Run Cufflinks on Tophat output to assemble reads into transcripts

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
    - Shared → Data Libraries → RNA-Seq Datasets
    - All datasets are FASTQ and from the Body Map 2.0 project
    - You will often receive data in FASTQ format.

http://bit.ly/gxyRNASEX

FASTQ Format

Specifies sequence (FASTA) and quality scores (PHRED)

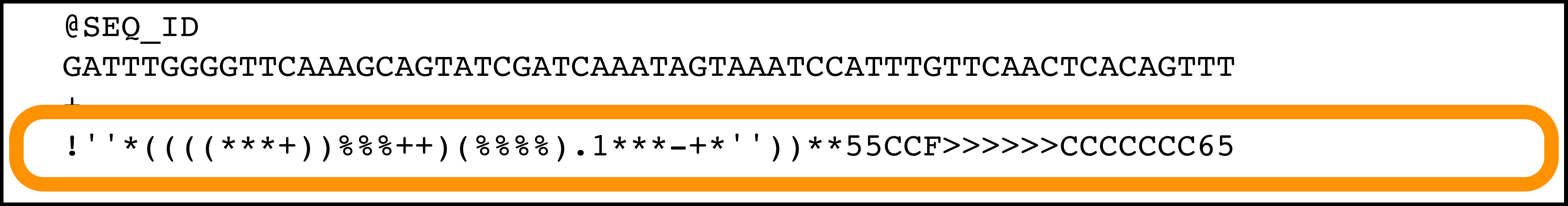
Text format, 4 lines per entry:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

1. Sequence Identifier preceded by @
2. Called bases
3. “+” separator
4. PHRED scores for each called base

http://en.wikipedia.org/wiki/FASTQ_format

PHRED Scores



```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Encode confidence for each individual base call

Range from 0 to 93 theoretically, but practically from 0-40.

Are logarithmic:

0: No confidence at all!

10: 1 in 10 chance call is wrong

20: 1 in 100 chance call is wrong

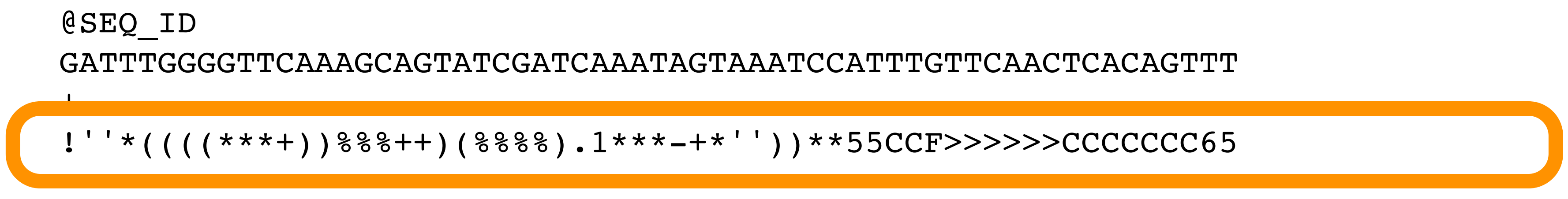
30: 1 in 1000 chance call is wrong

40: 1 in 10000 chance call is wrong

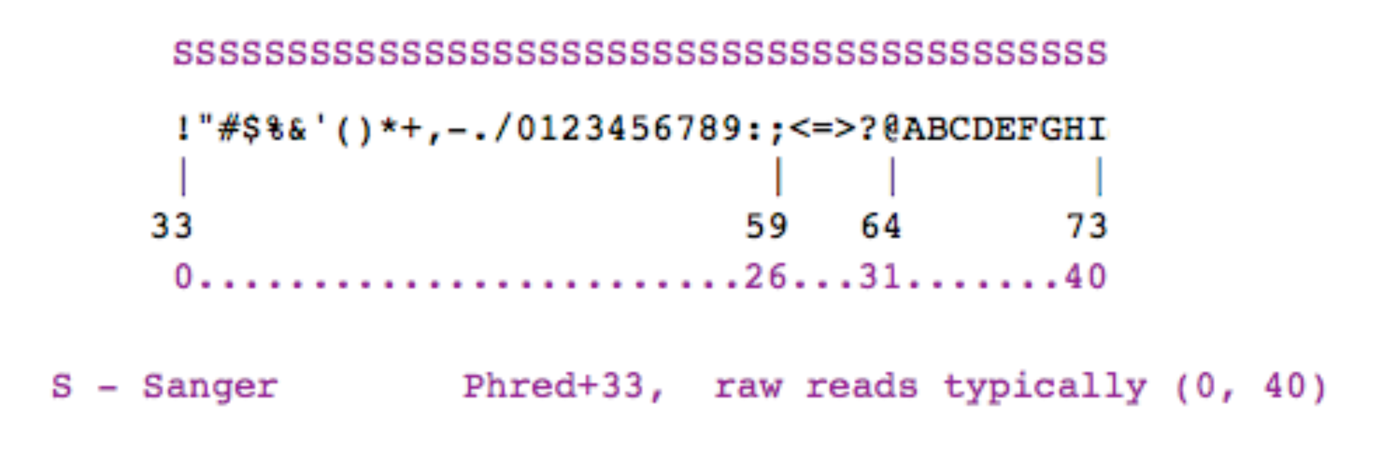
Where are the numbers?

http://en.wikipedia.org/wiki/FASTQ_format

PHRED Scores



```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```



```
     SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
     !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
     |                         |    |        |
    33                        59   64       73
     0........................26...31.......40

S - Sanger        Phred+33,  raw reads typically (0, 40)
```

Each different integer score is encoded as a single letter

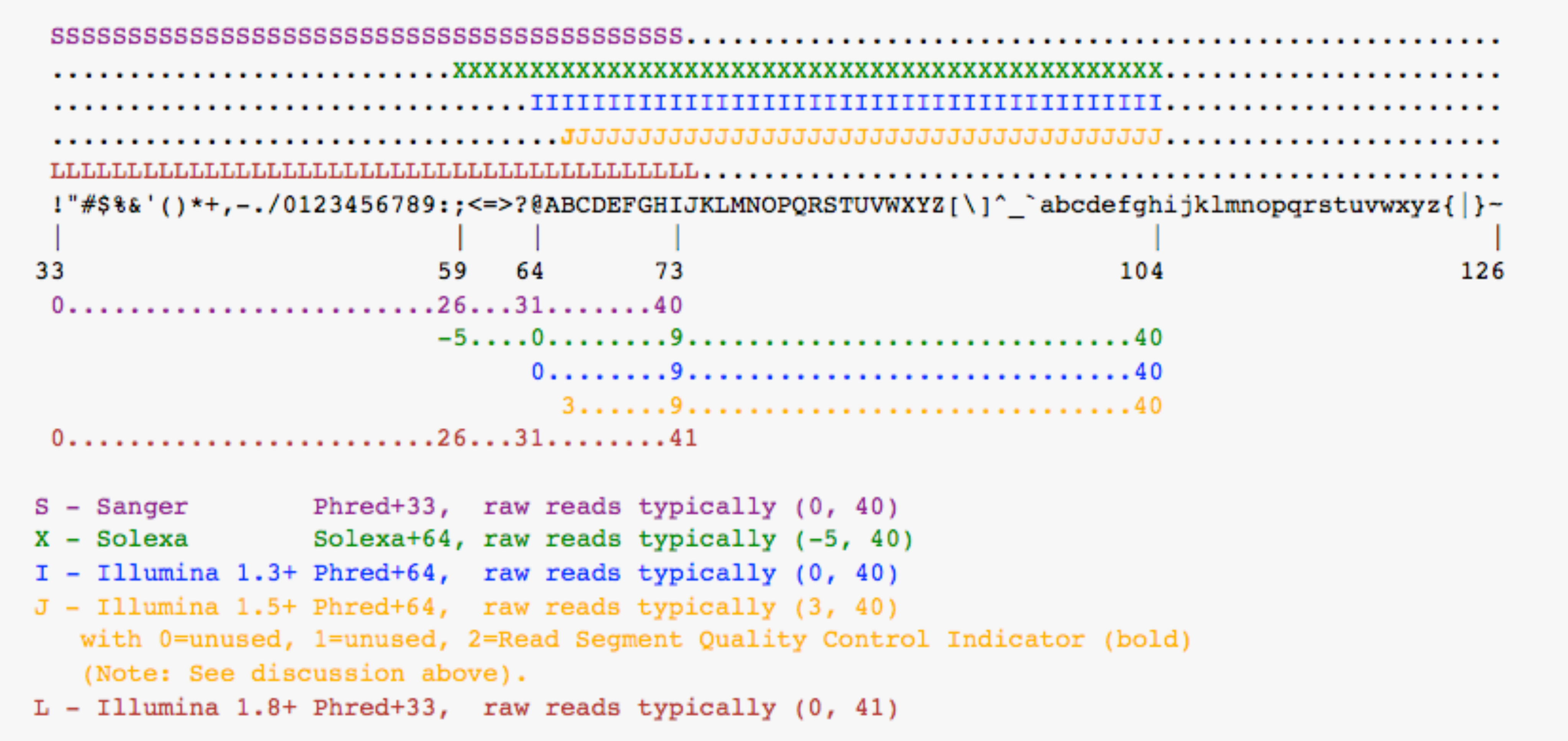
One base call is one character

Corresponding quality is one character too

http://en.wikipedia.org/wiki/FASTQ_format

However, FASTQ is such a cool standard, ...

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0........................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

that one version is not enough!

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 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

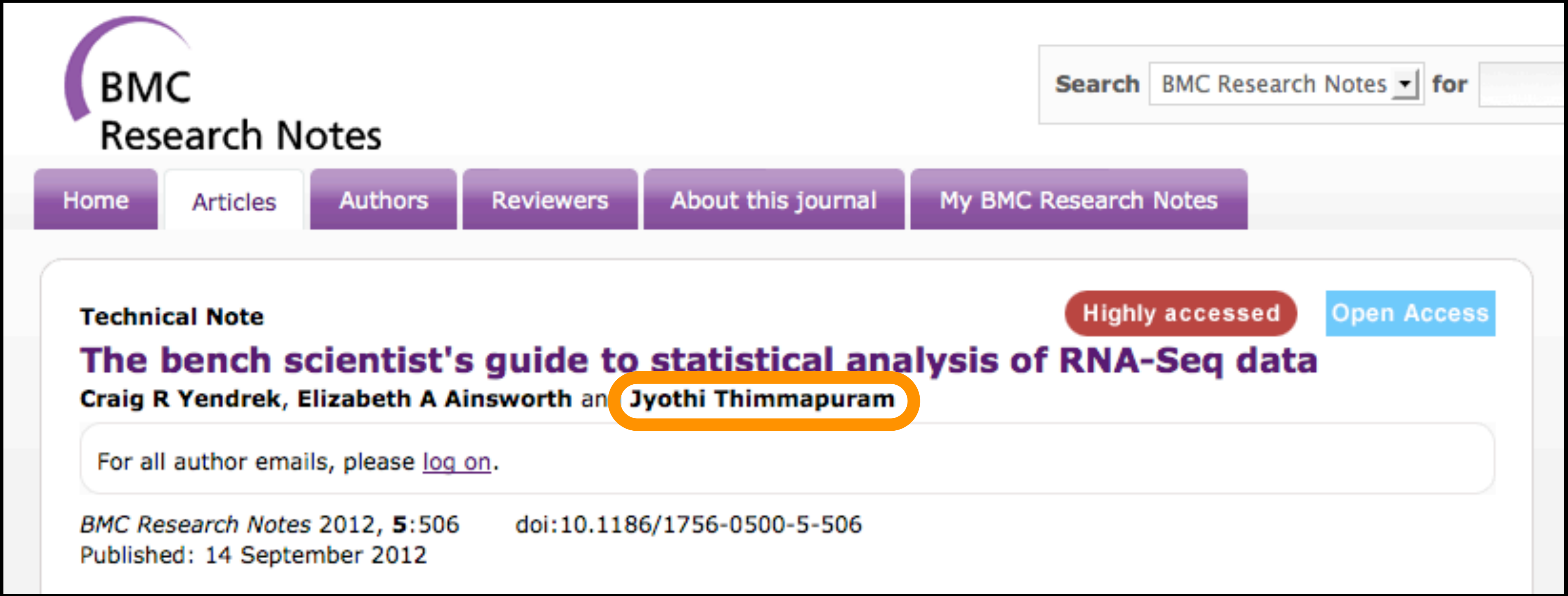
RNA-seq Exercise: A Plan

- 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

Variable Length Reads?

Will that hurt? I dunno, but ...



Technical Note

The bench scientist's guide to statistical analysis of RNA-Seq data

Craig R Yendrek, Elizabeth A Ainsworth an Jyothi Thimmapuram

For all author emails, please log on.

BMC Research Notes 2012, **5**:506 doi:10.1186/1756-0500-5-506
Published: 14 September 2012

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

Welcome, Basic Analysis

Basic analyses into Reusable Workflows

NGS Analysis I: Through Tophat

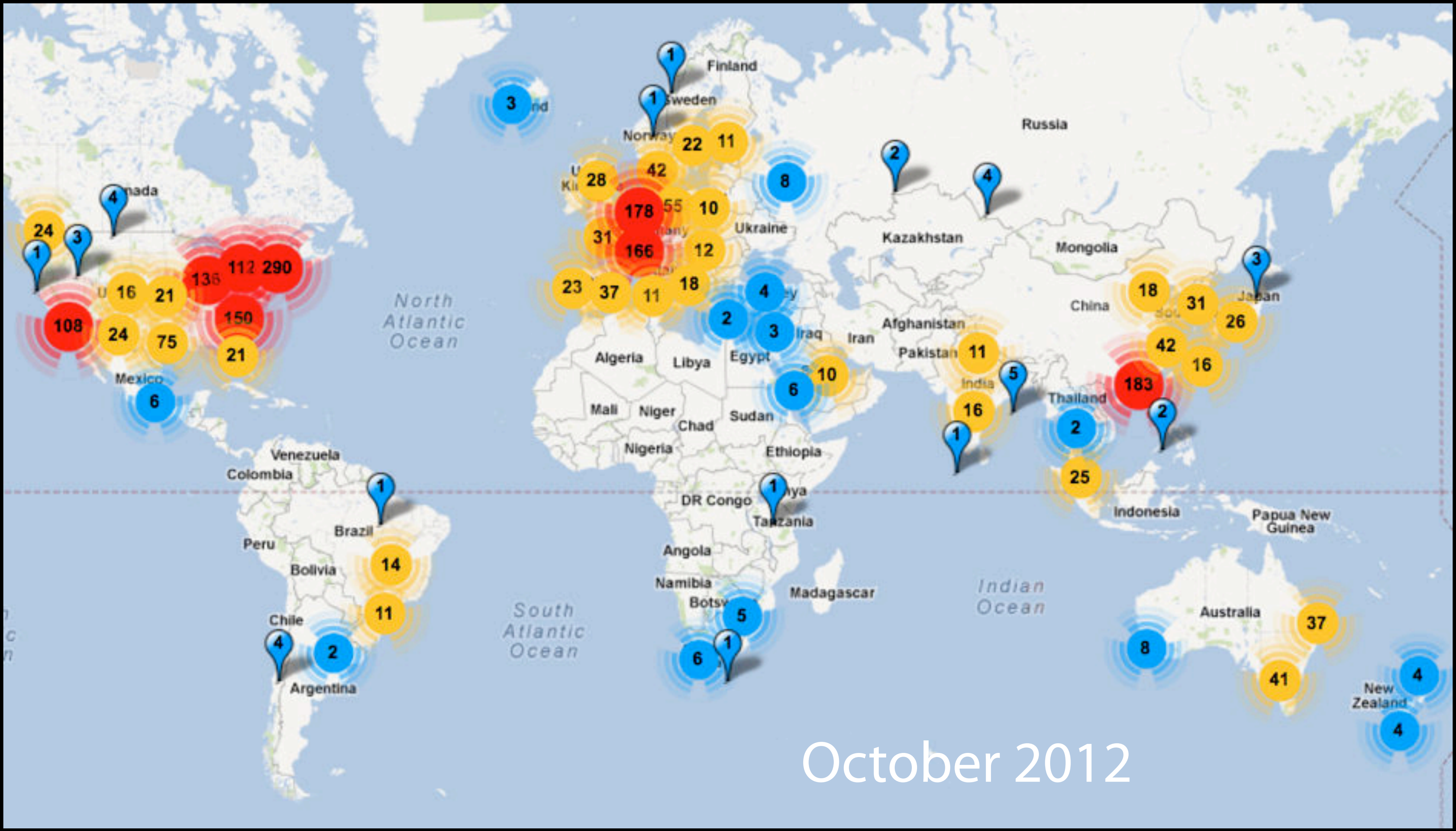
Galaxy Project Overview

NGS Analysis II: Cufflinks, Visualization

Manage, Reuse, and Share your Analyses

Setting up your own Galaxy on the Cloud

The Motivation Slide



October 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

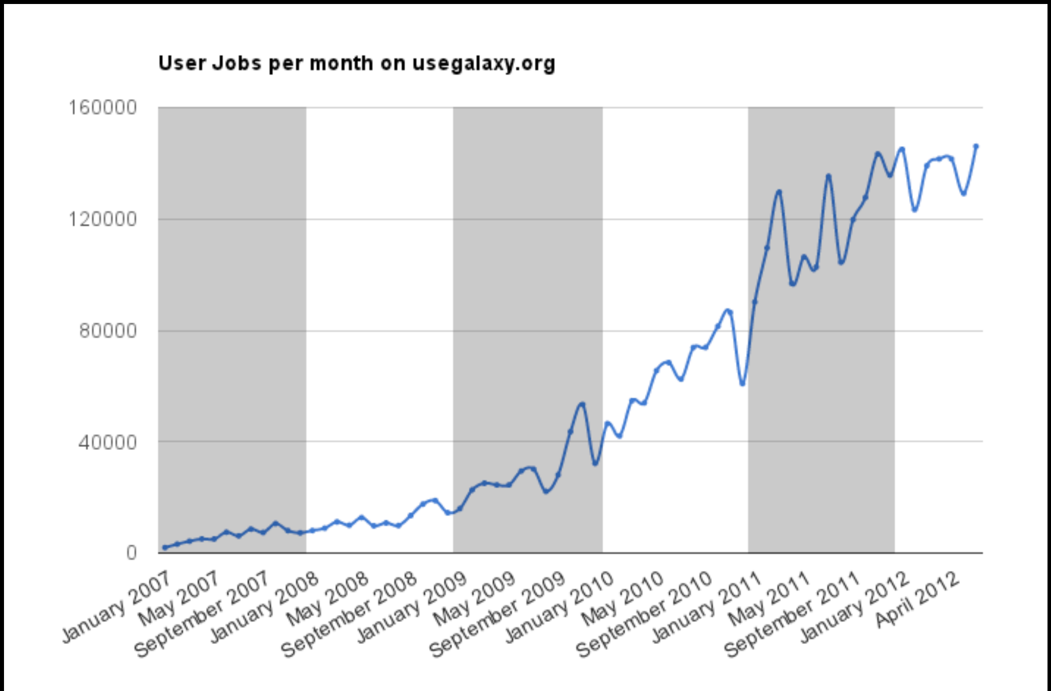
Galaxy is available ...

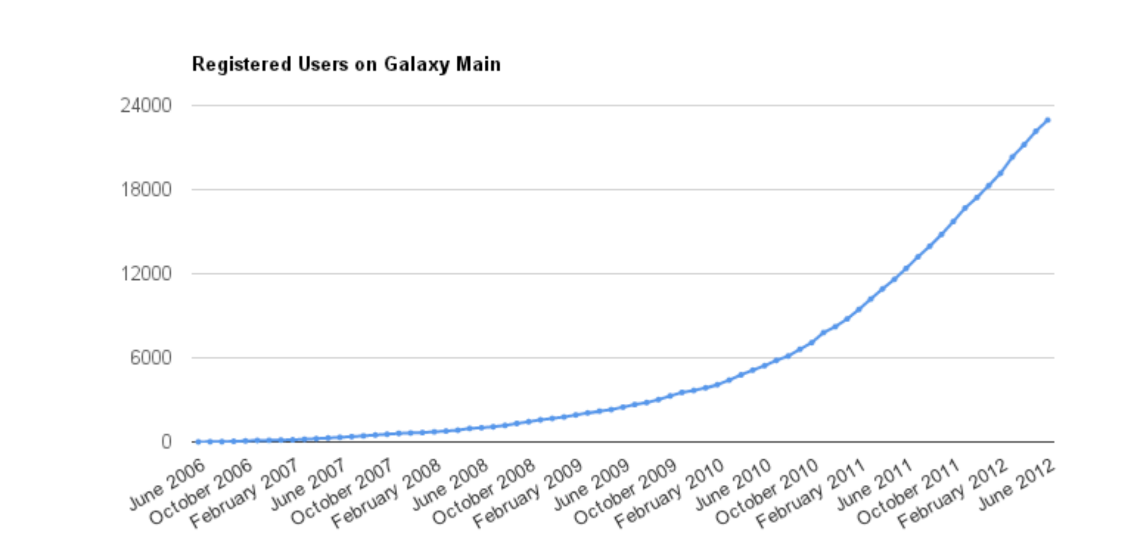
- **As a free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

http://usegalaxy.org

http://usegalaxy.org (a.k.a Main)

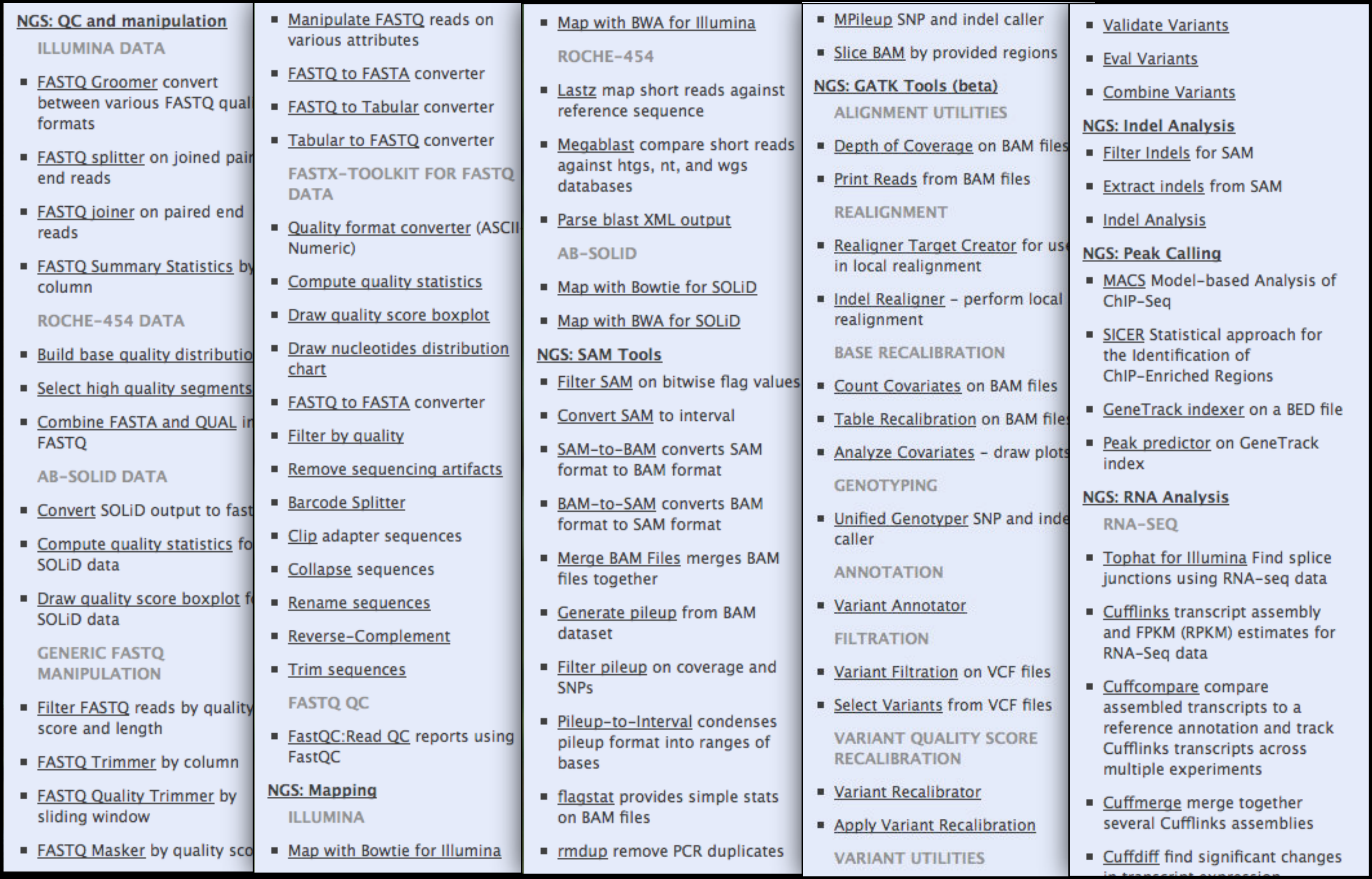
- Public web site
- Anybody can use it
- Persistent
- + 500 users / month
- ~300 TB of user data
- ~140,000 jobs / month
- Hundreds of tools ...





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 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 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

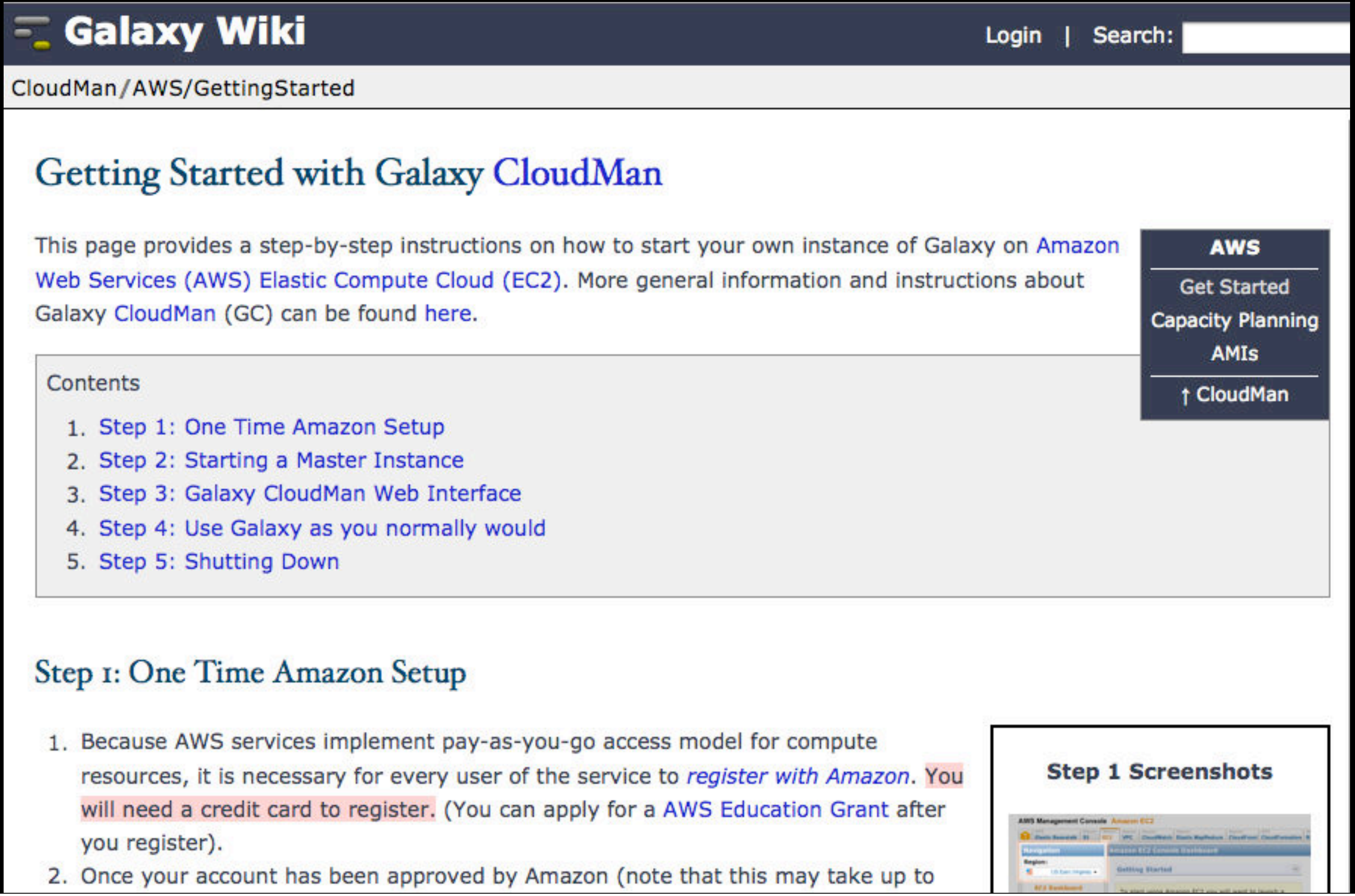
http://usegalaxy.org/cloud

- 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 are using this today.**
- **We will do this today.**

amazon
web services™

http://aws.amazon.com/education

Step by Step Instructions on the Wiki for Amazon



Getting Started with Galaxy CloudMan

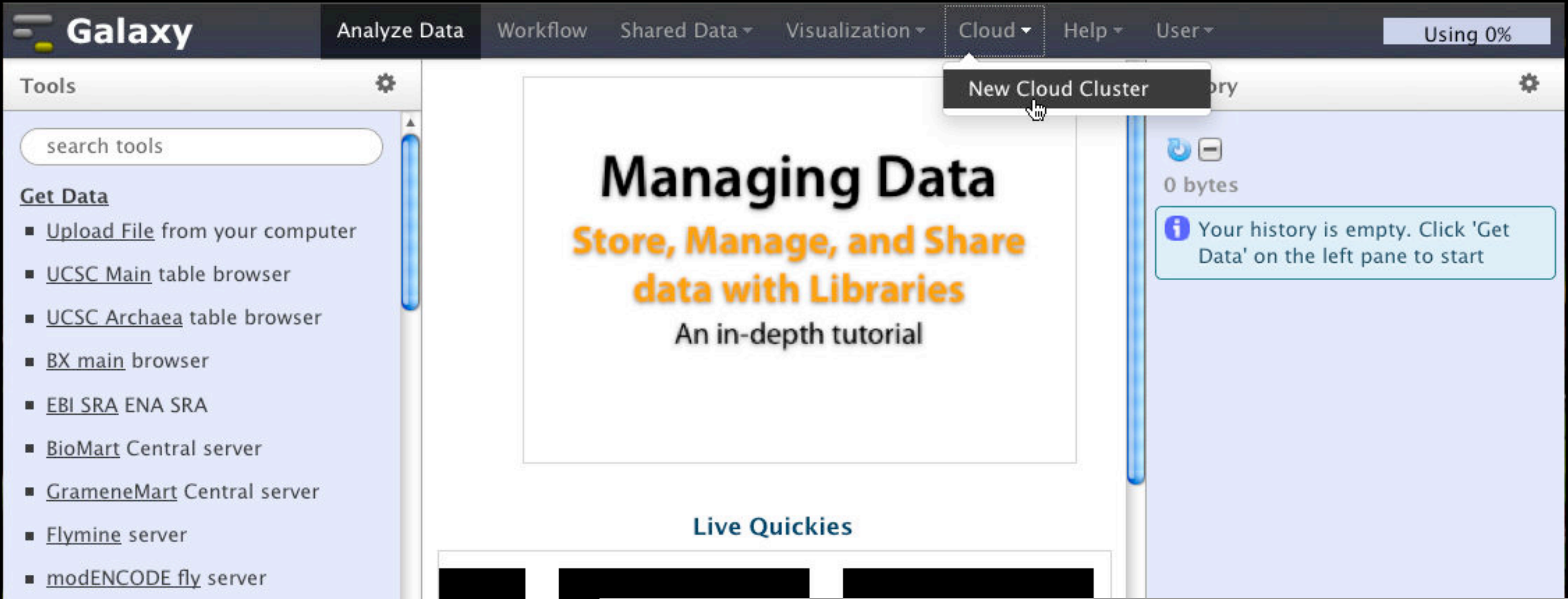
This page provides a step-by-step instructions on how to start your own instance of Galaxy on Amazon Web Services (AWS) Elastic Compute Cloud (EC2). More general information and instructions about Galaxy CloudMan (GC) can be found here.

Contents

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

Instant CloudMan



Tools

Get Data

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

Managing Data

Store, Manage, and Share data with Libraries

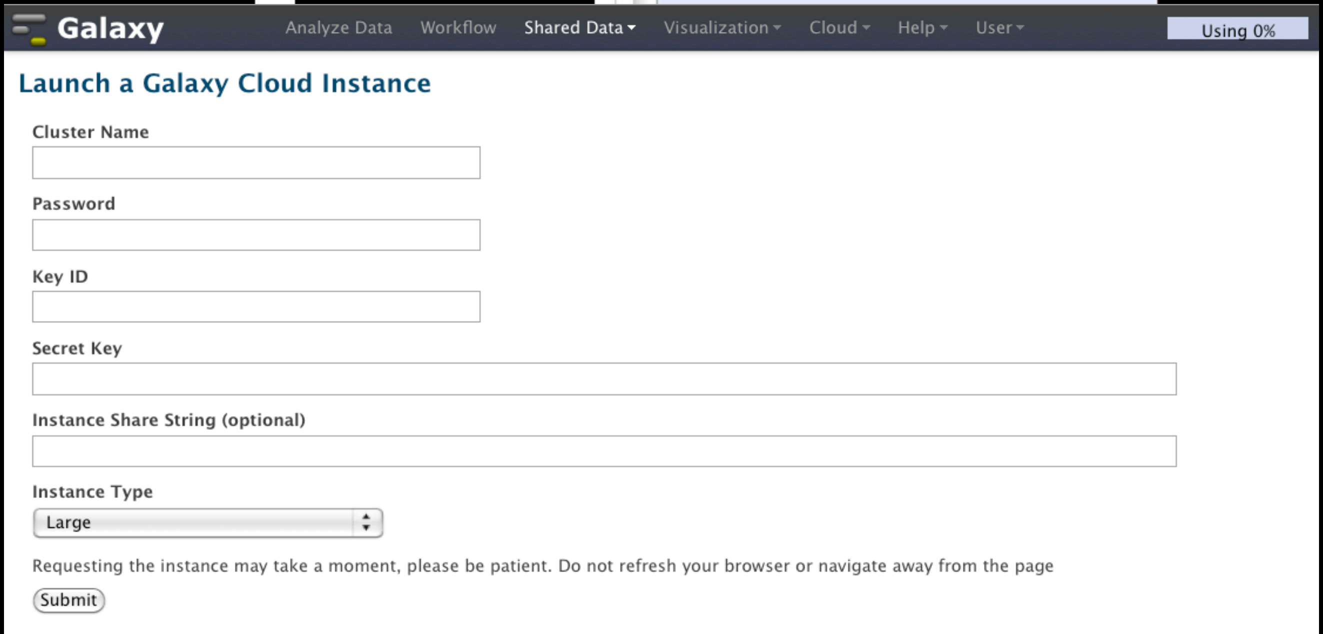
An in-depth tutorial

Live Quickies

0 bytes

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

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



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

Galaxy Community

Tool Shed

Mailing Lists (very active)

Screencasts

Events Calendar, News Feed

Community Wiki

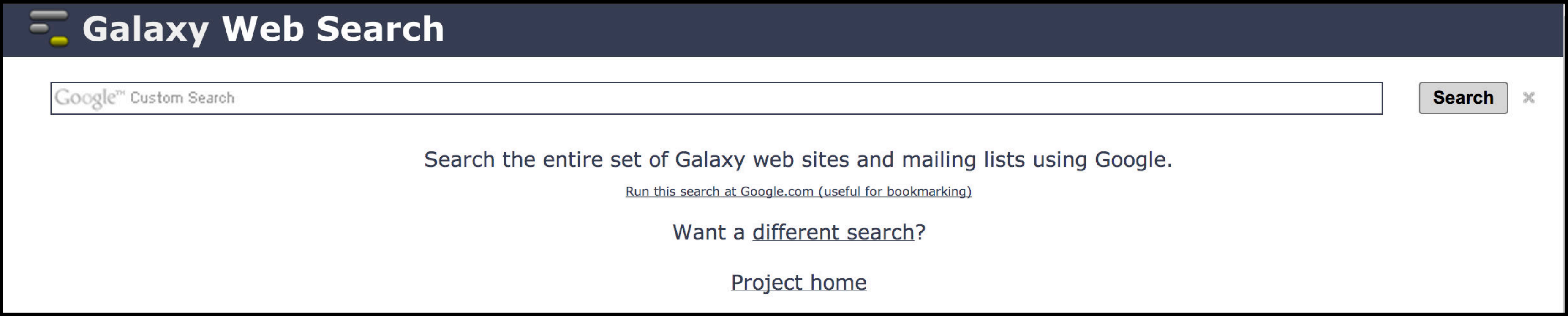
Local Public Installs

CiteULike group, Mendeley mirror

Annual Community Meting

http://galaxyproject.org/wiki

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

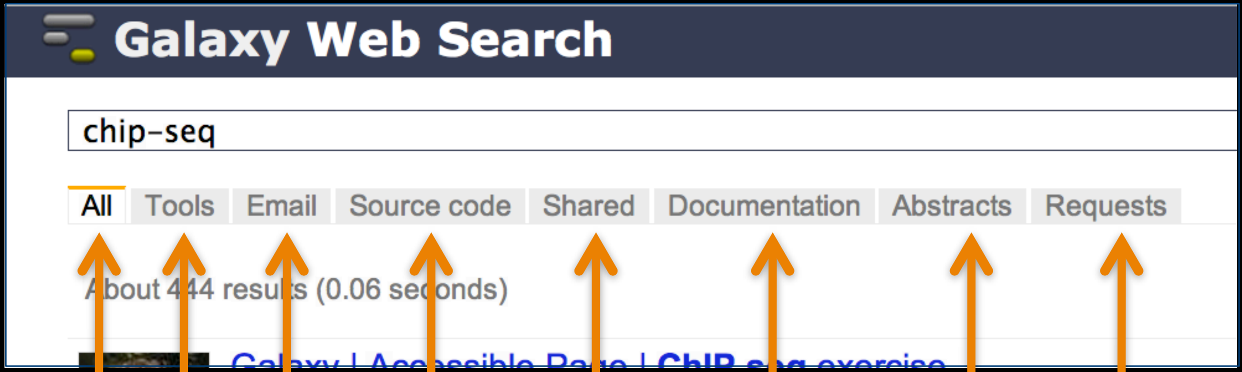


Search the entire set of Galaxy web sites and mailing lists using Google.

Run this search at Google.com (useful for bookmarking)

Want a different search?

Project home



About 444 results (0.06 seconds)

Galaxy | Accessible Page | ChIP-seq exercise

Find

Everything on …

Tools for …

Email about …

Source code for …

Published Histories, Pages, Workflows, about …

Documentation on …

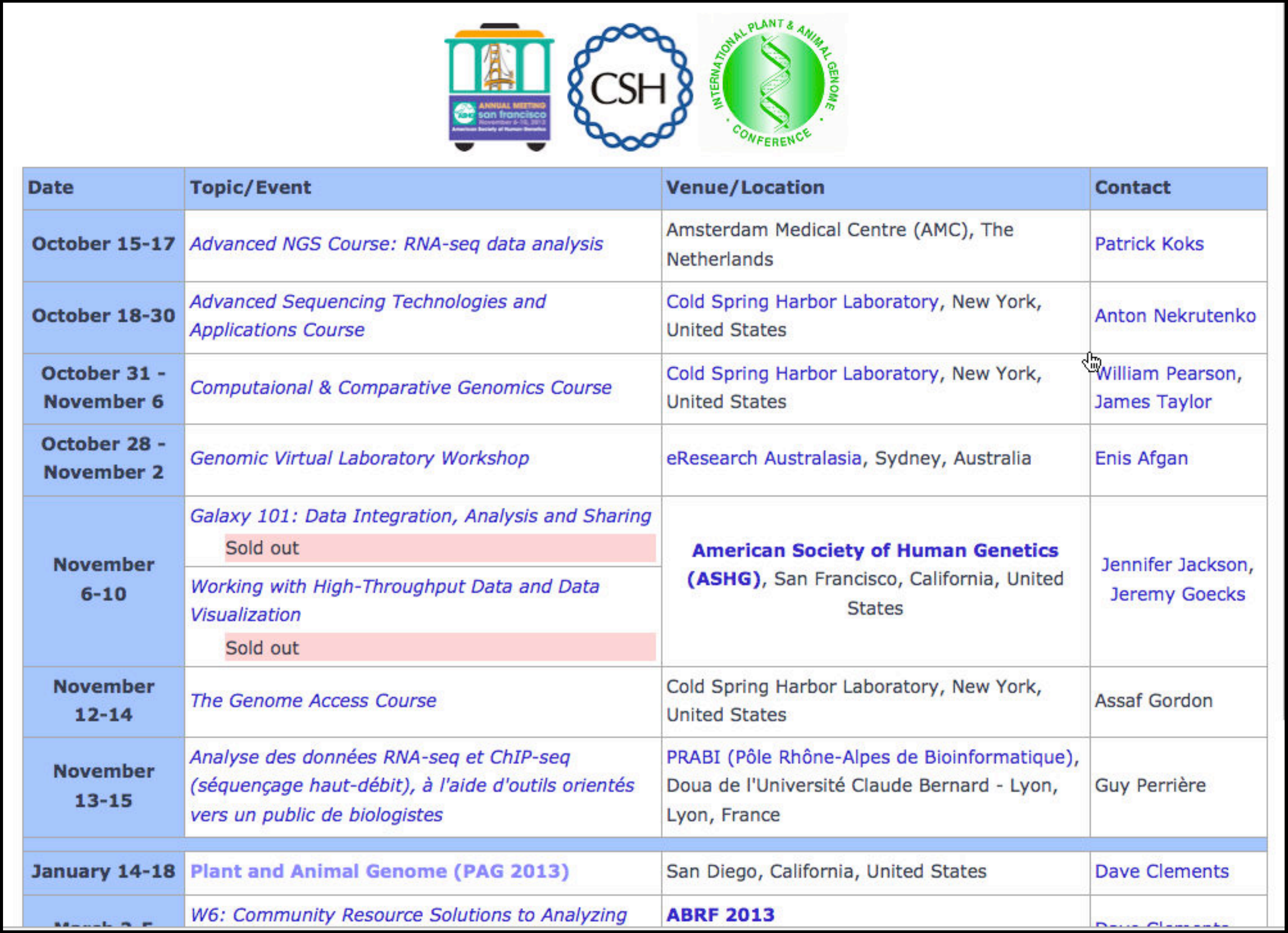
Papers using Galaxy for …

Related feature requests

Galaxy
Community
Conference
30 June
- 2 July
2013
OSLO
UiO : University of Oslo

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 *Working with High-Throughput Data and Data Visualization* Sold out	**American Society of Human Genetics (ASHG)**, San Francisco, California, United States	Jennifer Jackson, Jeremy Goecks
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
March 2-5	*W6: Community Resource Solutions to Analyzing*	**ABRF 2013**	Dave Clements

http://galaxyproject.org/wiki/Events

Galaxy URLs to Remember

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

http://usegalaxy.org/galaxy101

Agenda

Welcome, Basic Analysis

Basic analyses into Reusable Workflows

NGS Analysis I: Through Tophat

Galaxy Project Overview

NGS Analysis II: Cufflinks, Visualization

Manage, Reuse, and Share your Analyses

Setting up your own Galaxy on the Cloud

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

Visualize

Send data results to **external** genome browsers

Trackster: Galaxy’s genome browser

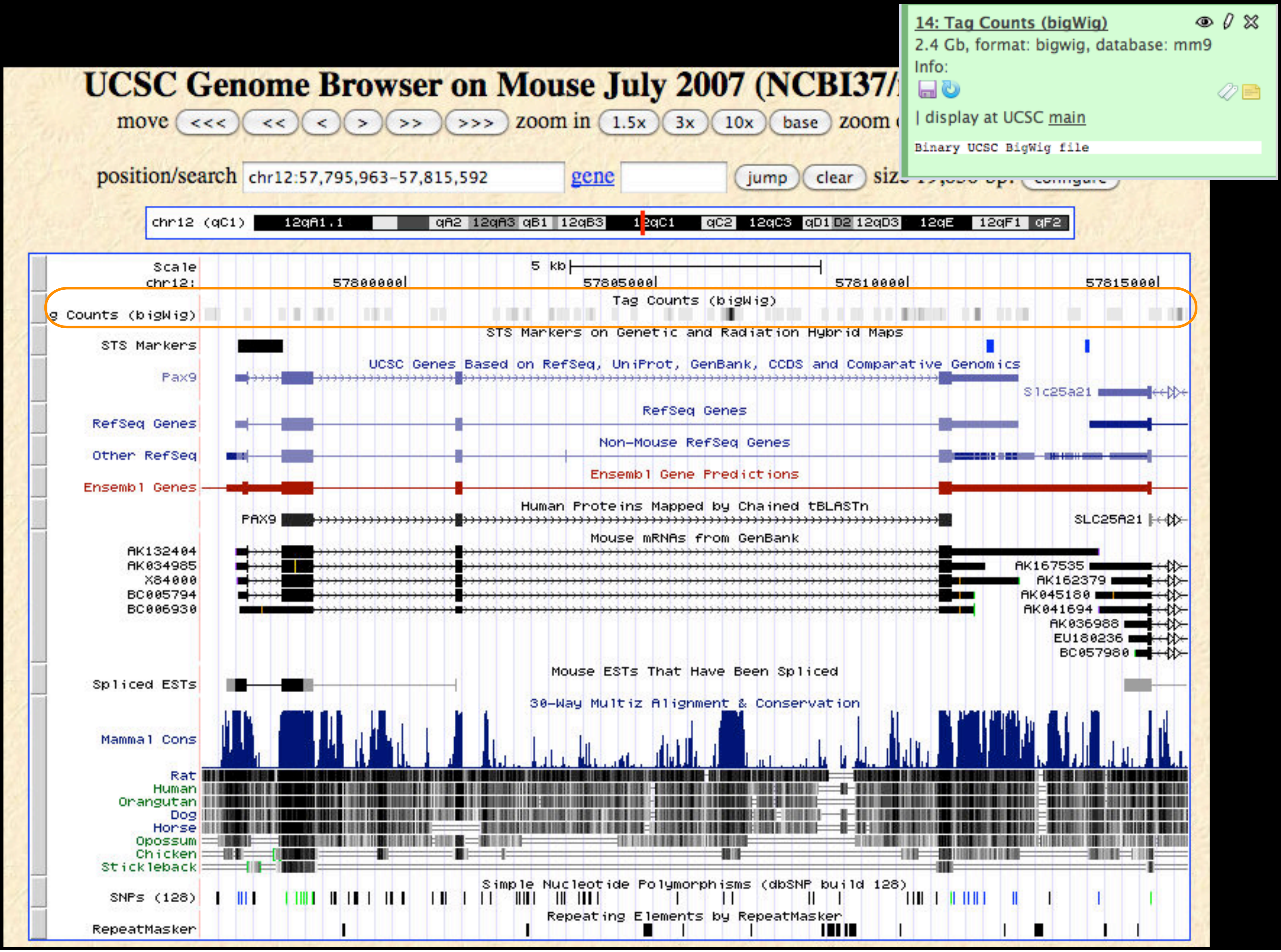
External Genome Browsers

UCSC

Ensembl

GBrowse

IGV



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

14: Tag Counts (bigWig)

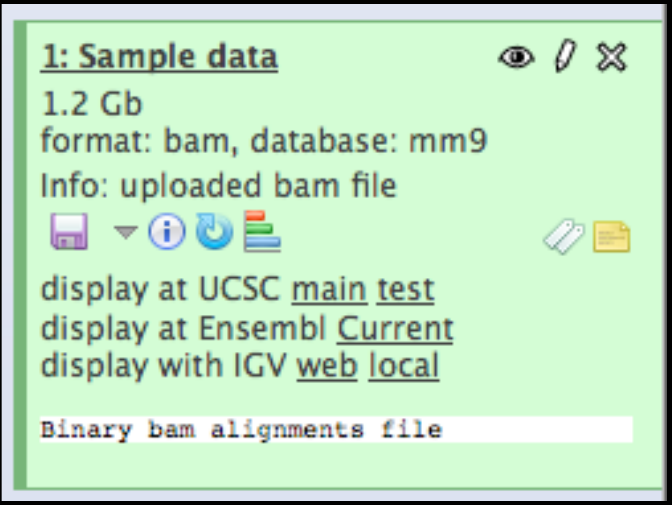
2.4 Gb, format: bigwig, database: mm9

Info:

| display at UCSC main

```
Binary UCSC BigWig file
```

Integrative Genomics Viewer (IGV)

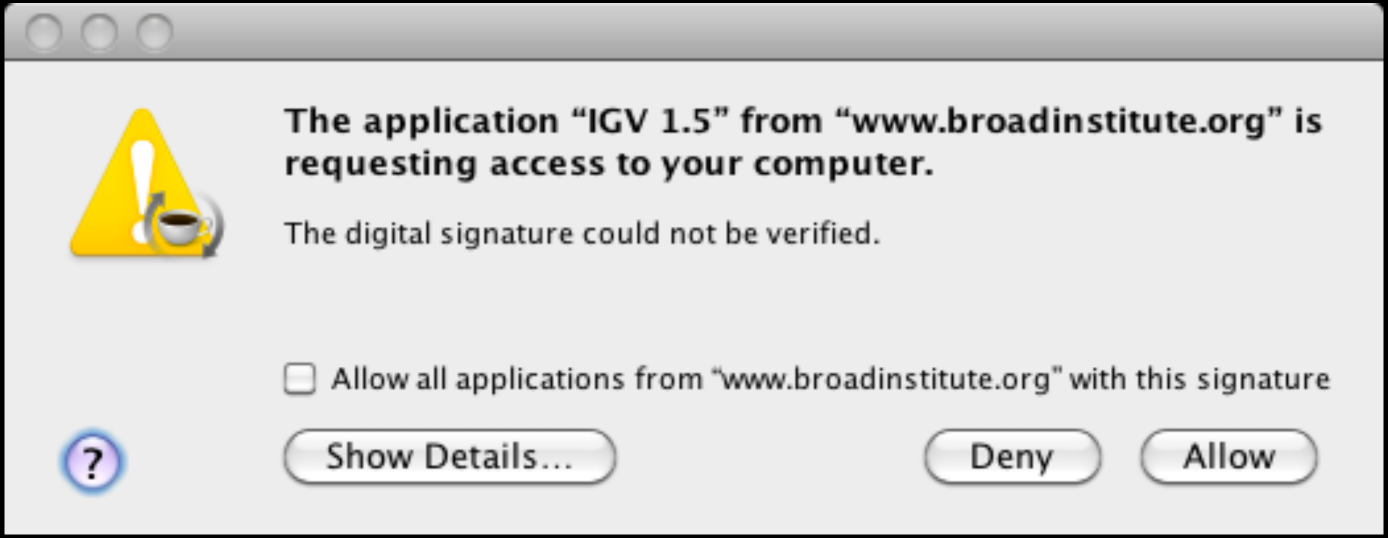


1: Sample data

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

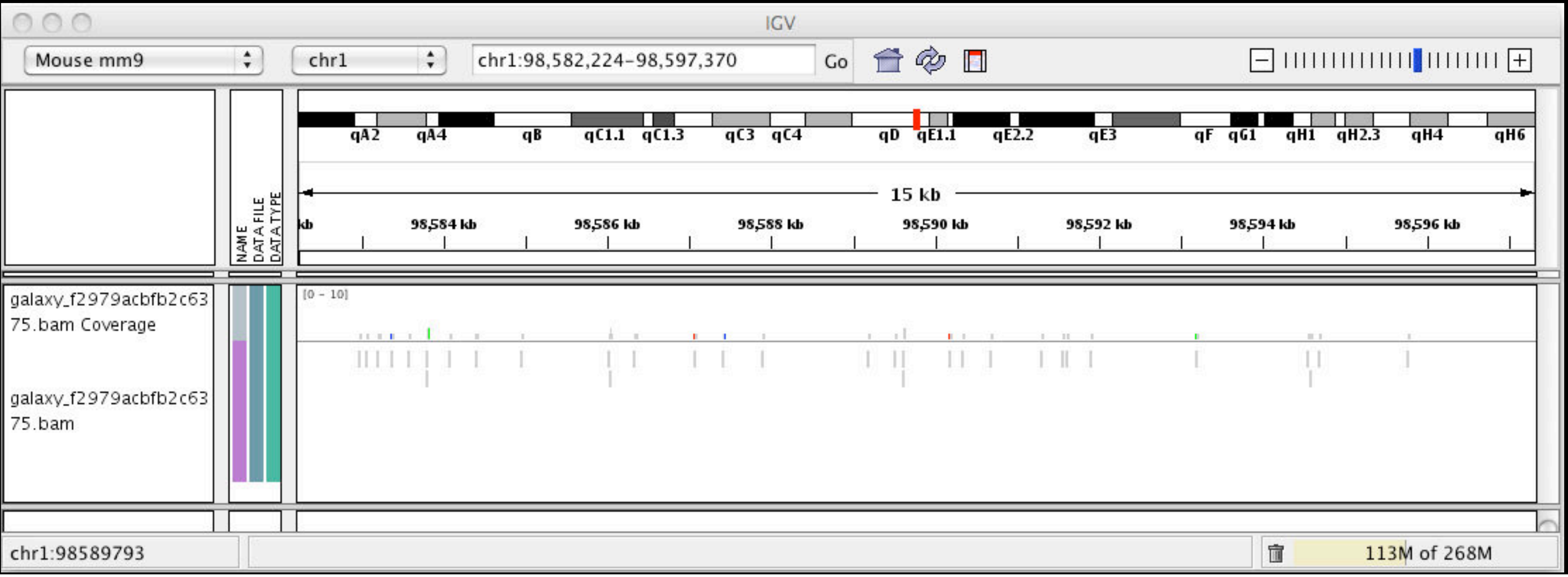
Binary bam alignments file



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



IGV

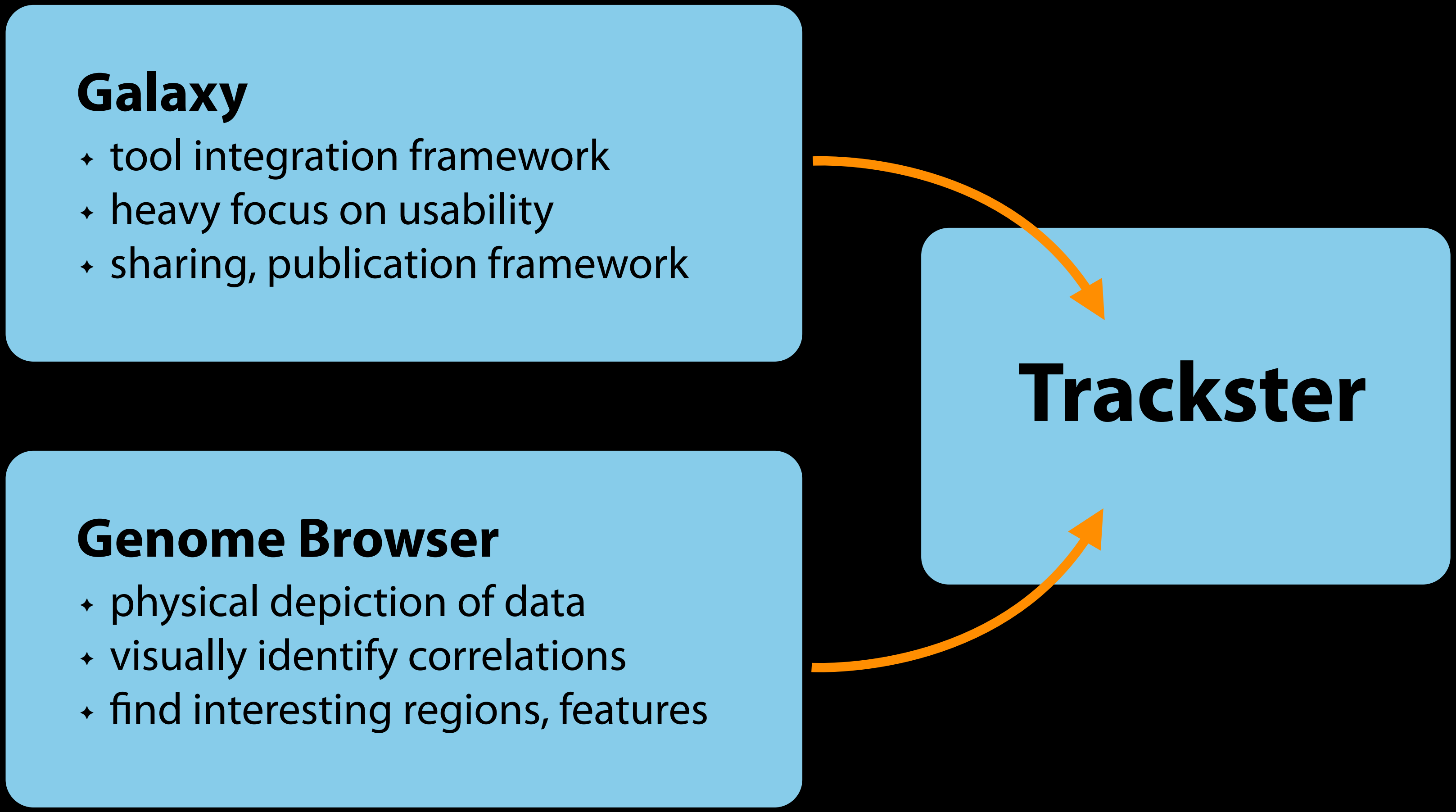
Mouse mm9 | chr1 | chr1:98,582,224-98,597,370 | Go

galaxy_f2979acbfb2c6375.bam Coverage

galaxy_f2979acbfb2c6375.bam

chr1:98589793

113M of 268M



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

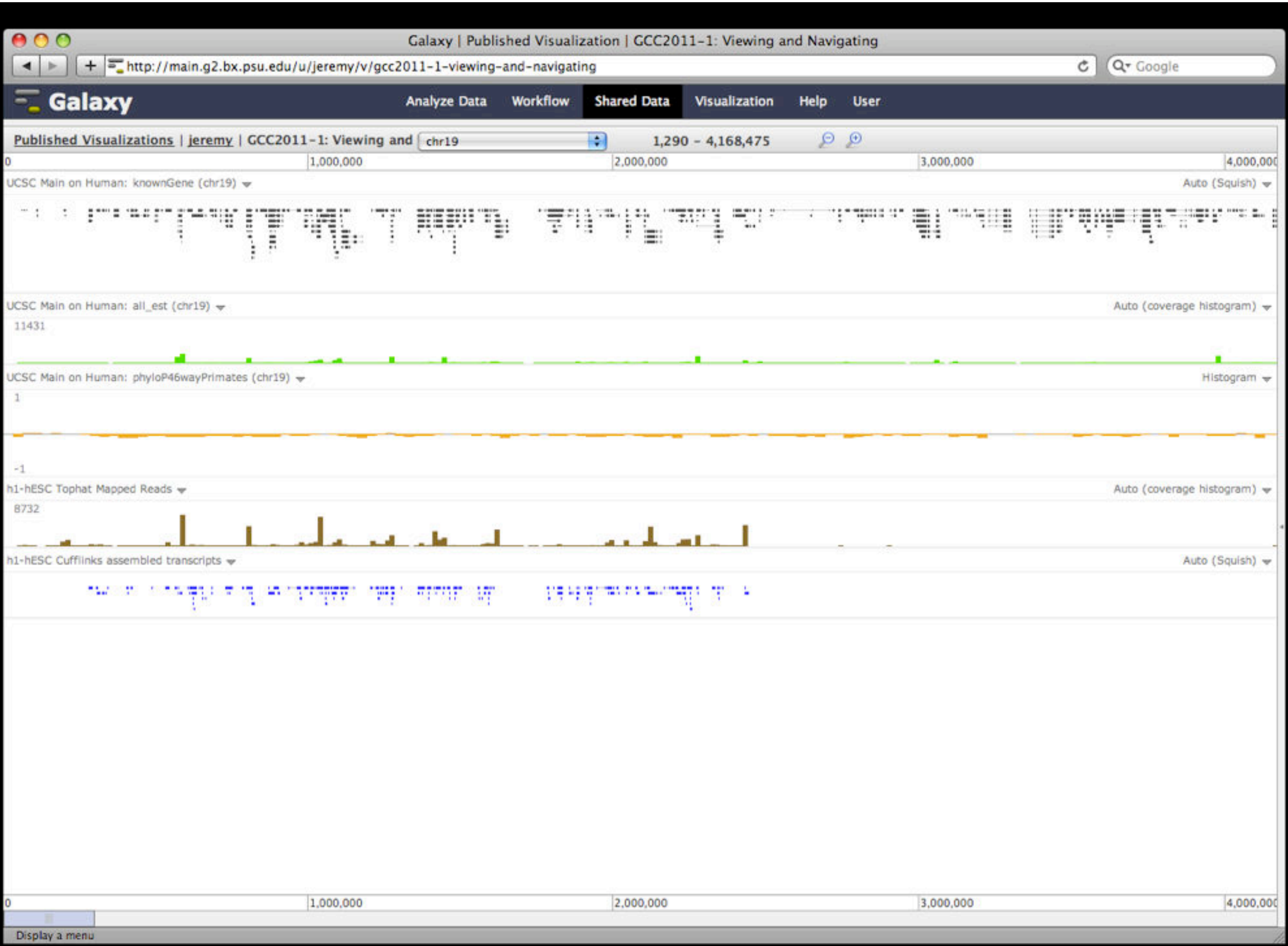
- No data transfers to external site
- Use it locally, even without internet access

Supports common filetypes

- BAM, BED, GFF/GTF, WIG

Unique features

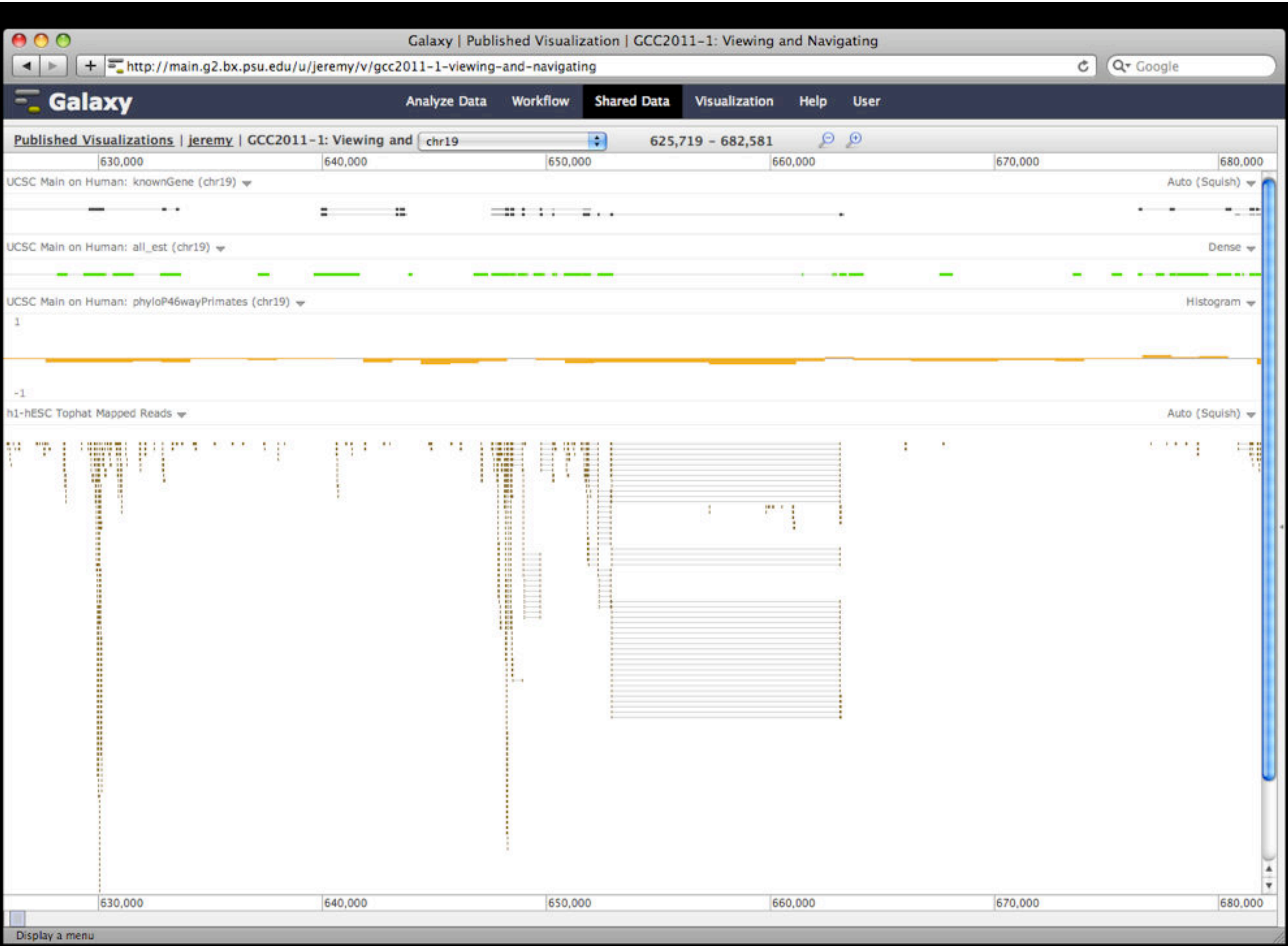
- custom genomes
- highly interactive

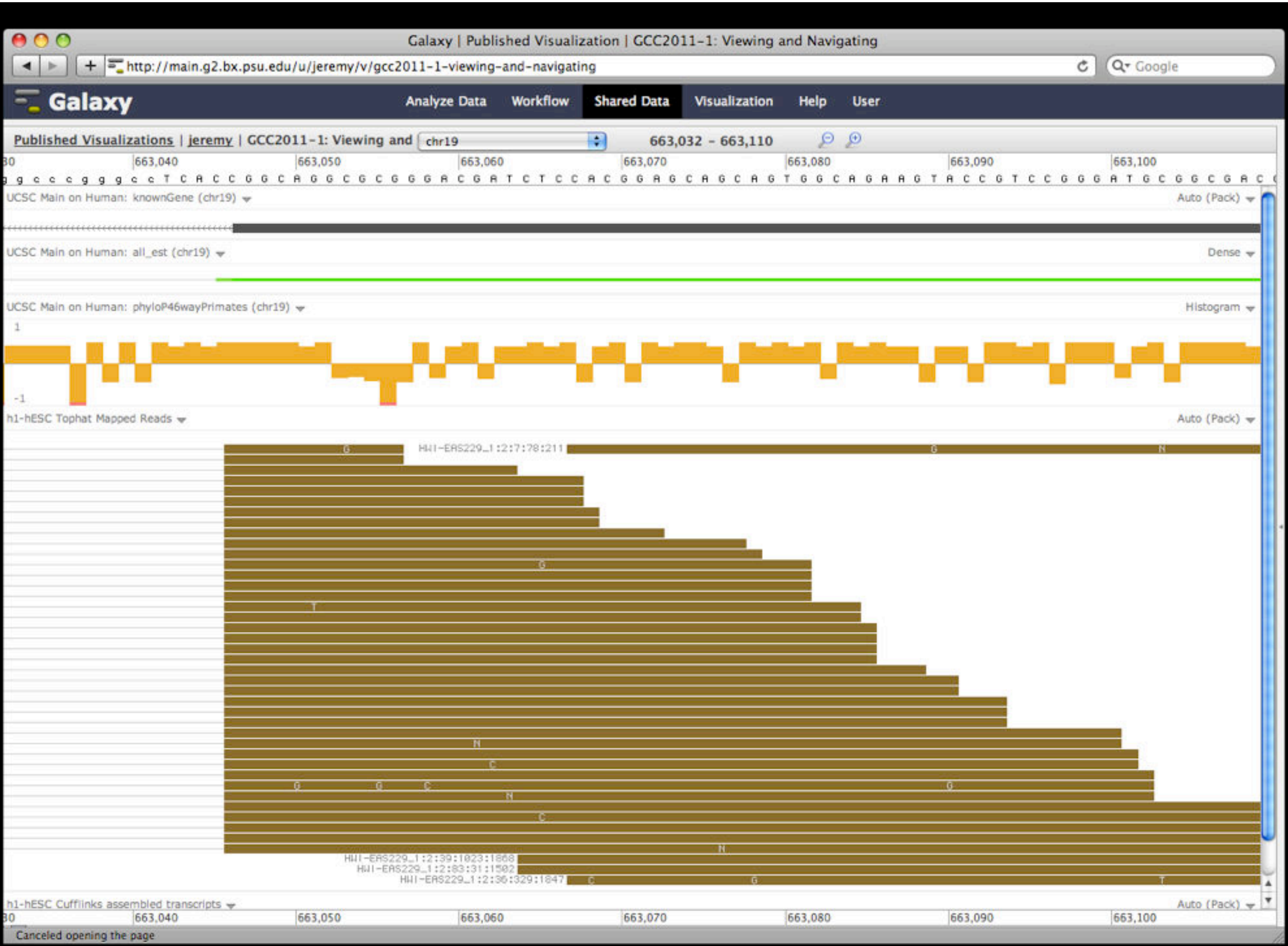


Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating

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

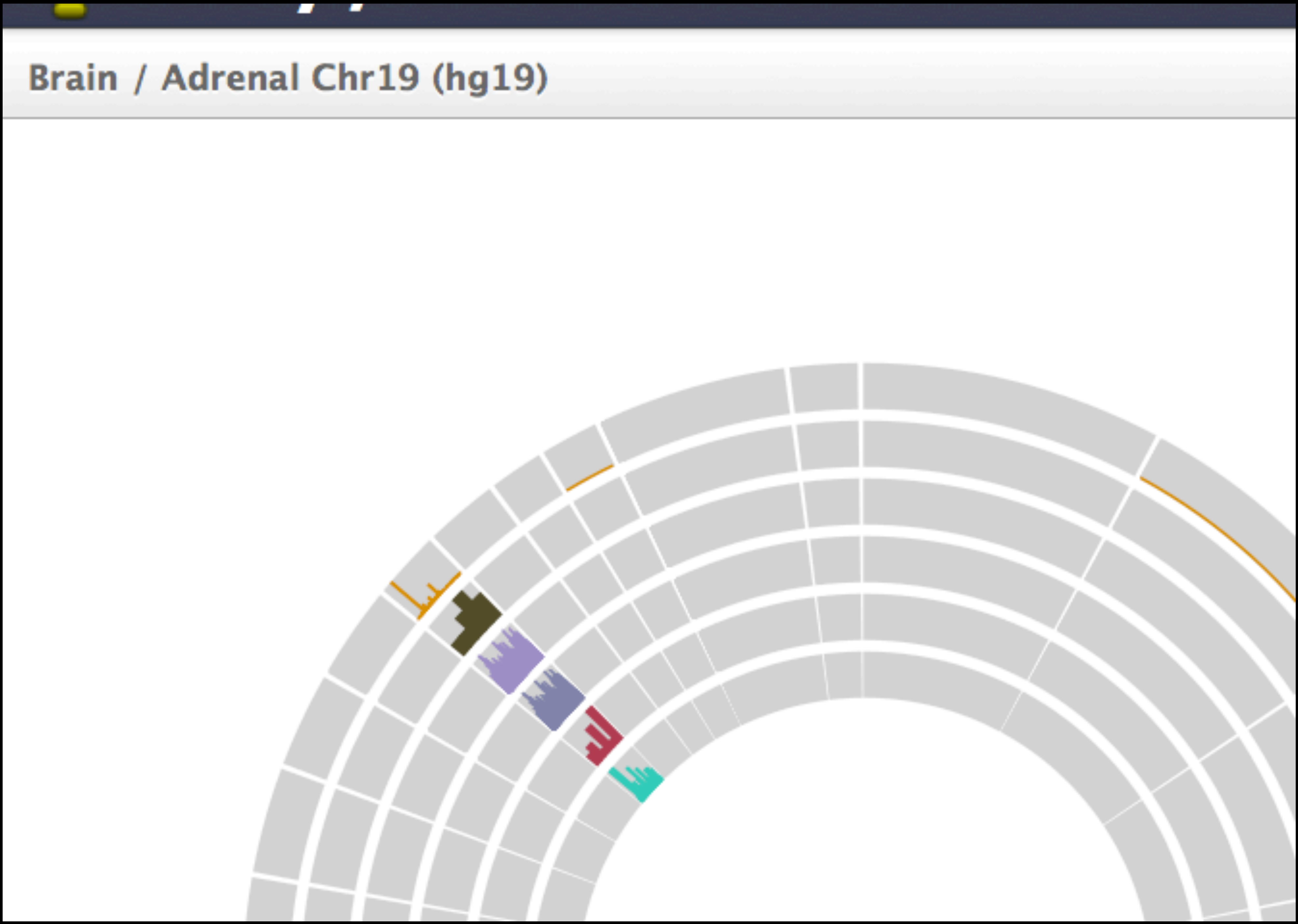
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-1: Viewing and
chr19
1,290 - 4,168,475
0
1,000,000
2,000,000
3,000,000
4,000,000
UCSC Main on Human: knownGene (chr19)
Auto (Squish)
UCSC Main on Human: all_est (chr19)
Auto (coverage histogram)
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
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-1: Viewing and
chr19
663,032 - 663,110
30
663,040
663,050
663,060
663,070
663,080
663,090
663,100
g g c c c g g g c c T C A C C G G C A G G C G C G G G A C G A T C T C C A C G G A G C A G C A G T G G C A G A A G T A C C G T C C G G G A T G C G G C G A C
UCSC Main on Human: knownGene (chr19)
Auto (Pack)
UCSC Main on Human: all_est (chr19)
Dense
UCSC Main on Human: phyloP46wayPrimates (chr19)
Histogram
1
-1
h1-hESC Tophat Mapped Reads
Auto (Pack)
HWI-EAS229_1:2:7:78:211
HWI-EAS229_1:2:39:1023:1868
HWI-EAS229_1:2:83:31:1502
HWI-EAS229_1:2:36:329:1847
h1-hESC Cufflinks assembled transcripts
Auto (Pack)
30
663,040
663,050
663,060
663,070
663,080
663,090
663,100

Canceled opening the page



Brain / Adrenal Chr19 (hg19)

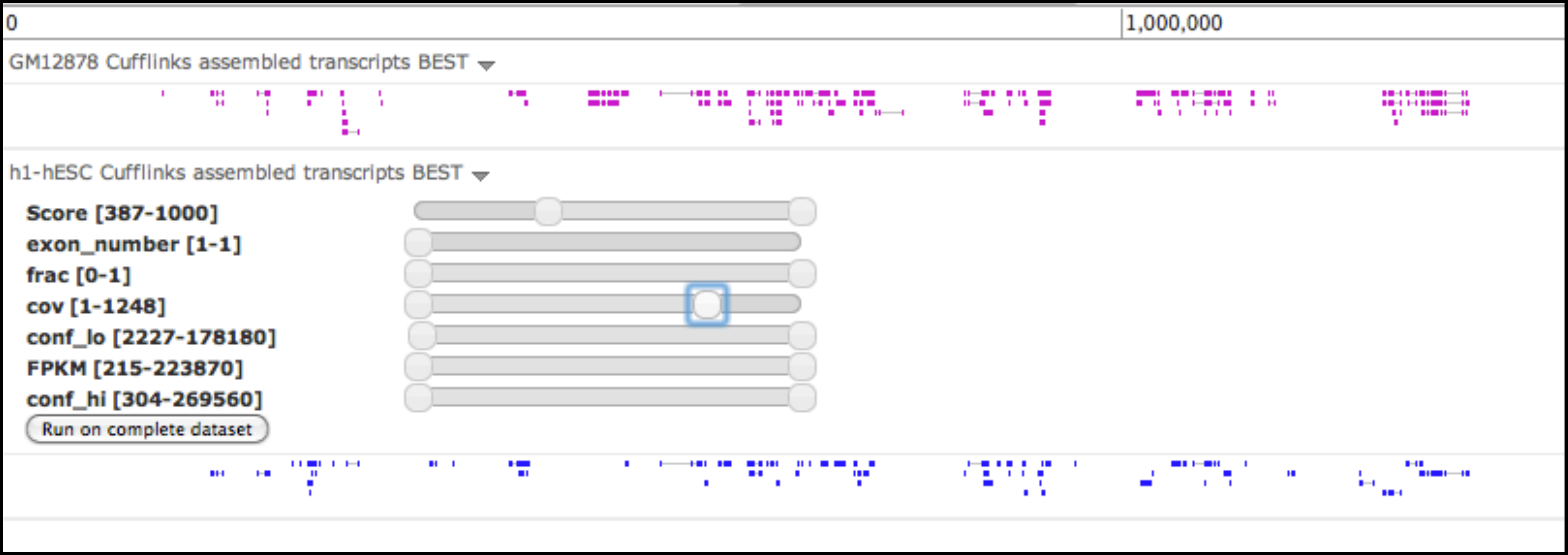
But really, why *another* genome browser

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

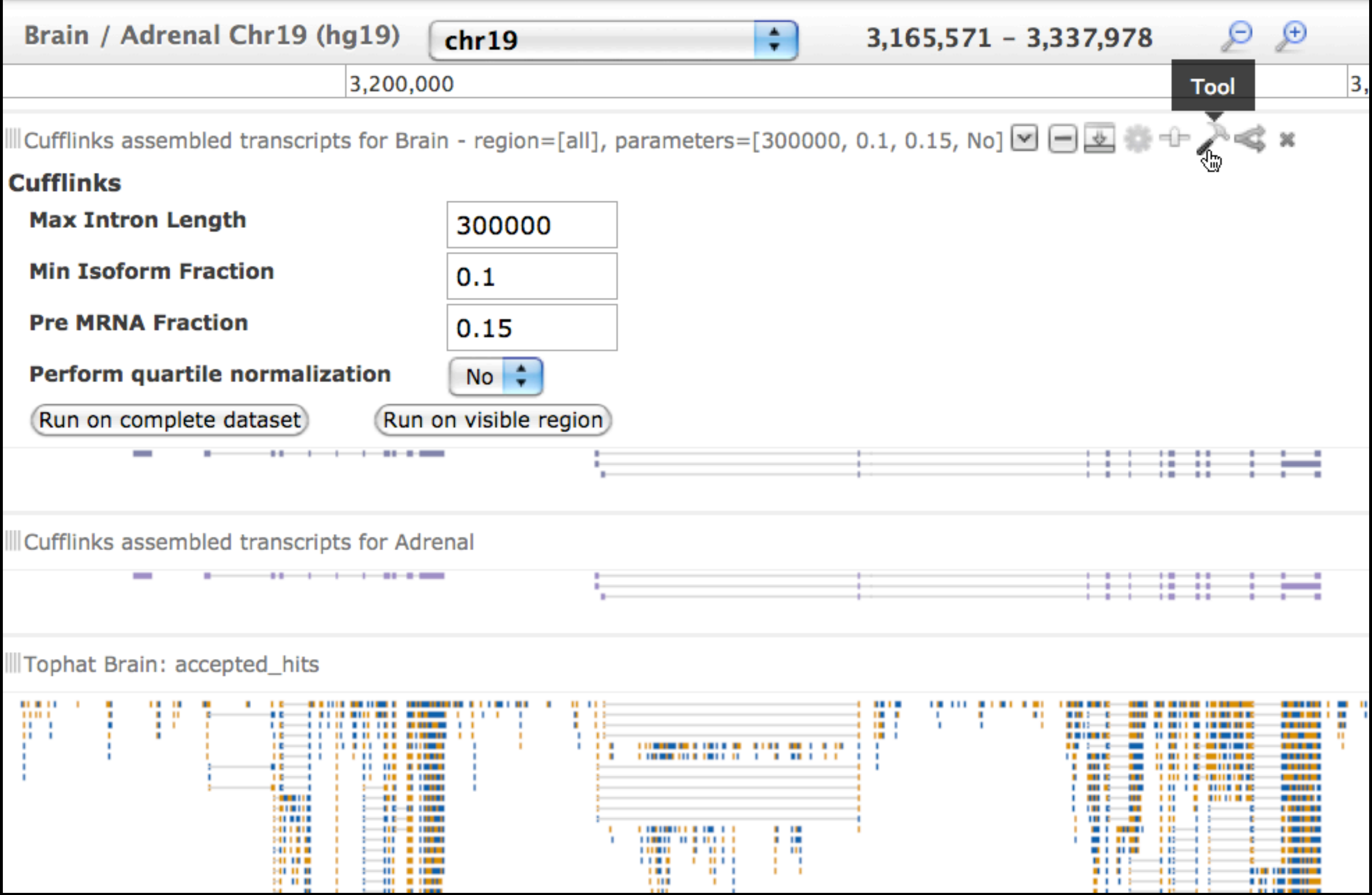
Dynamic Filtering



0
1,000,000
GM12878 Cufflinks assembled transcripts BEST
h1-hESC Cufflinks assembled transcripts BEST

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



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

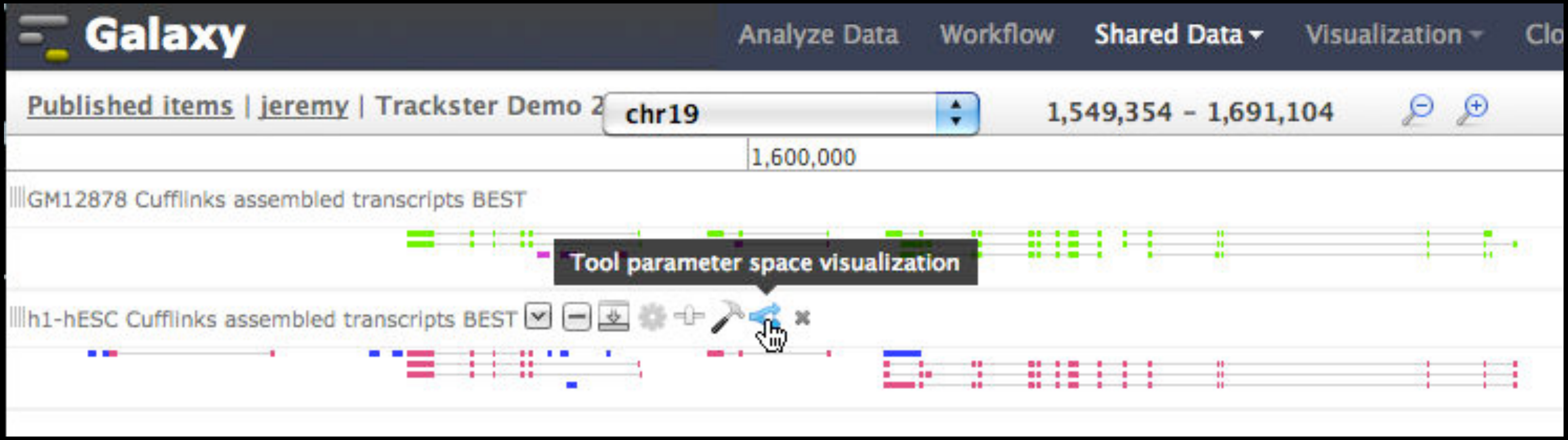
Cufflinks assembled transcripts for Brain - region=[all], parameters=[300000, 0.1, 0.15, No]

Cufflinks

Max Intron Length	300000
Min Isoform Fraction	0.1
Pre MRNA Fraction	0.15
Perform quartile normalization	No

Run on complete dataset
Run on visible region
Cufflinks assembled transcripts for Adrenal
Tophat Brain: accepted_hits

Exploring Parameter Space with Trackster



Published items | jeremy | Trackster Demo 2

chr19

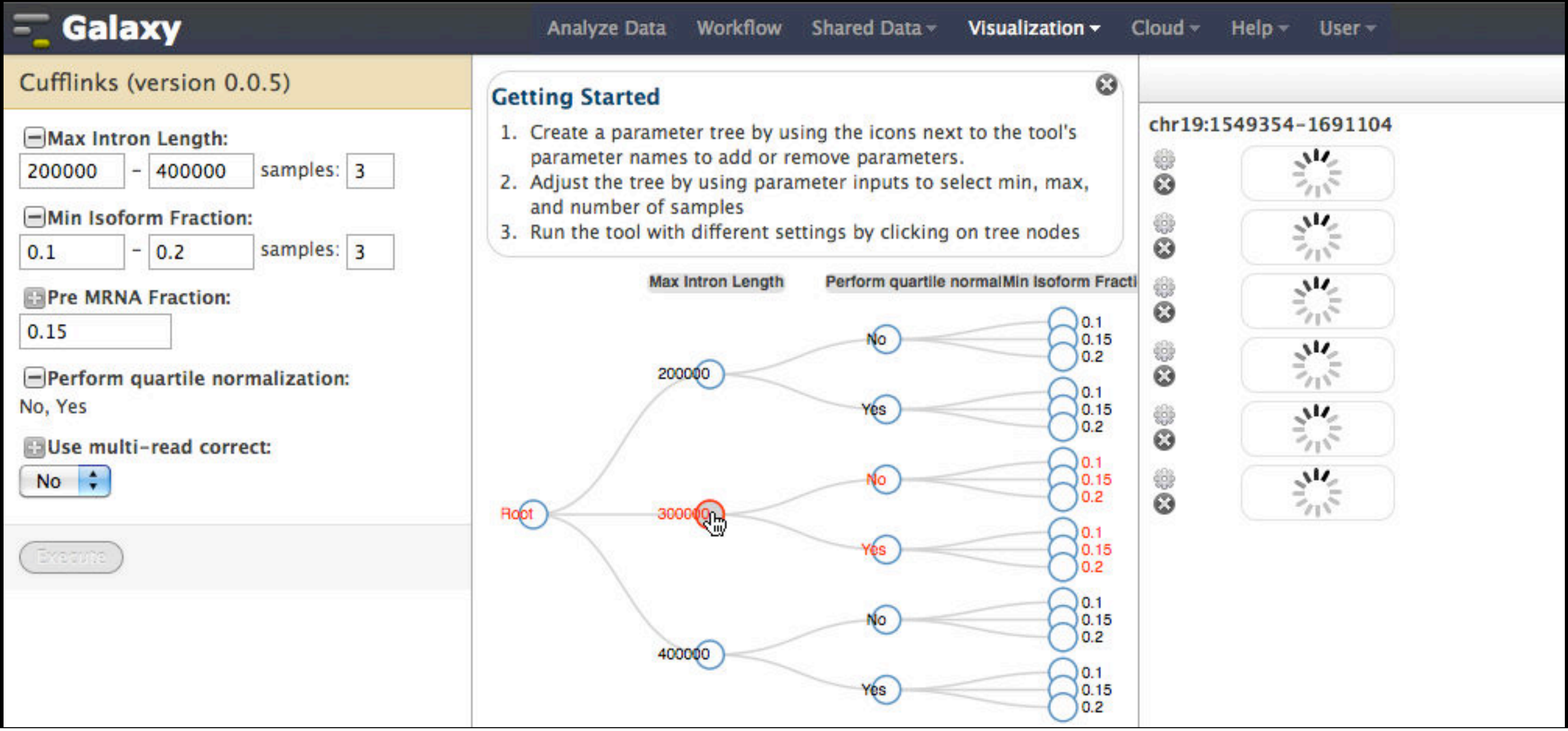
1,549,354 - 1,691,104

1,600,000

GM12878 Cufflinks assembled transcripts BEST

Tool parameter space visualization

h1-hESC Cufflinks assembled transcripts BEST



Cufflinks (version 0.0.5)

Max Intron Length: 200000 - 400000 samples: 3

Min Isoform Fraction: 0.1 - 0.2 samples: 3

Pre MRNA Fraction: 0.15

Perform quartile normalization: No, Yes

Use multi-read correct: No

Getting Started

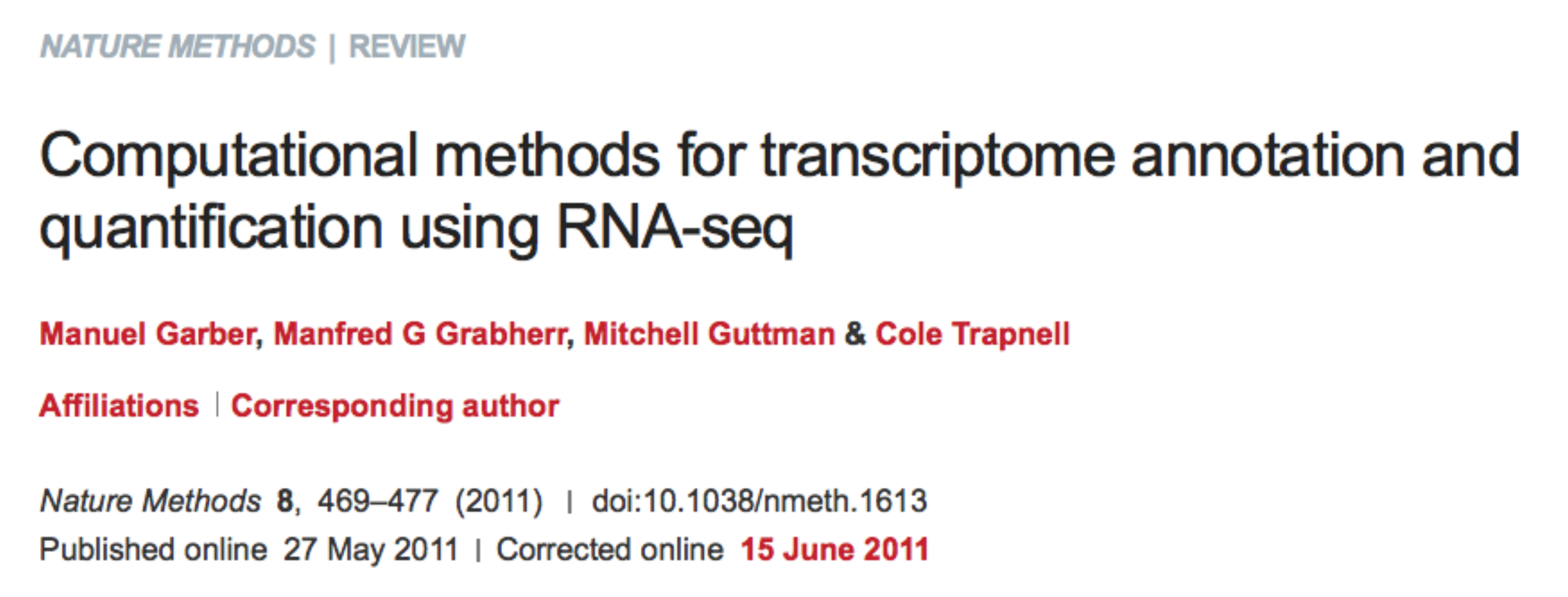
1. Create a parameter tree by using the icons next to the tool's parameter names to add or remove parameters.
2. Adjust the tree by using parameter inputs to select min, max, and number of samples
3. Run the tool with different settings by clicking on tree nodes

chr19:1549354-1691104

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 @

Two RNA-seq Papers



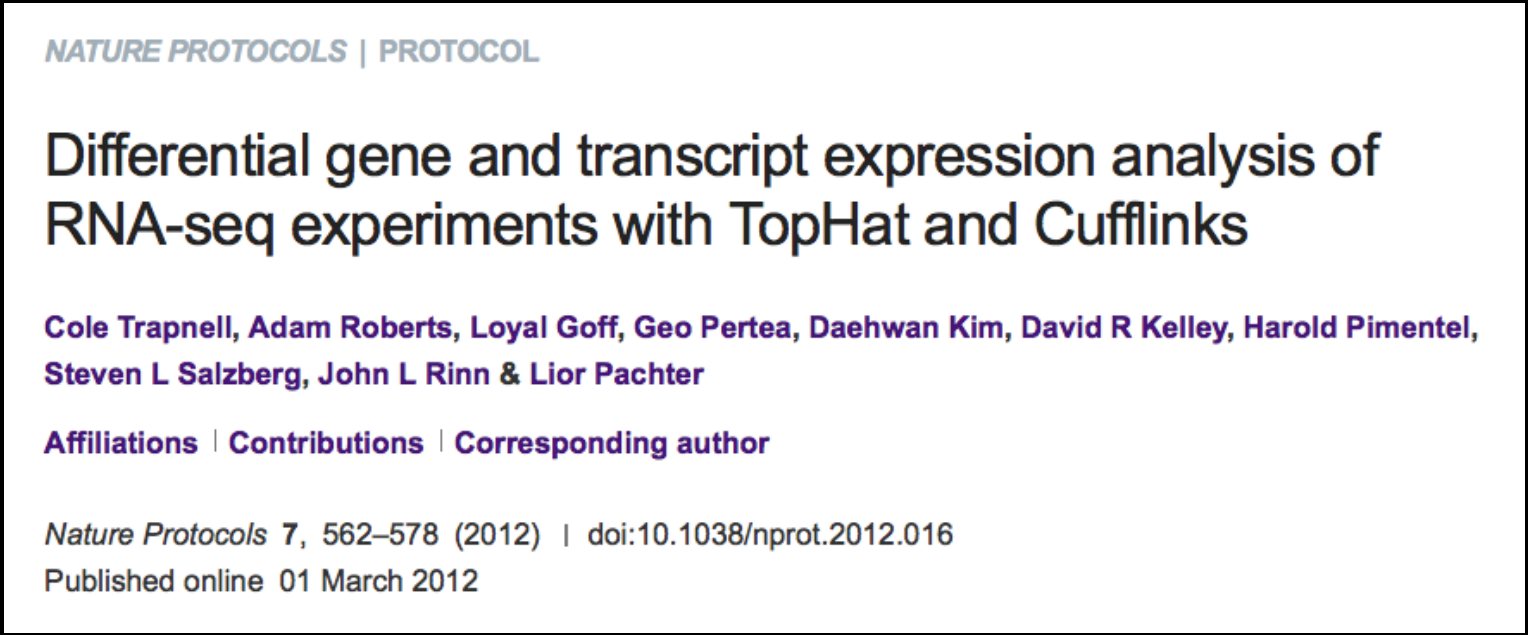
NATURE METHODS | REVIEW

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

Welcome, Basic Analysis

Basic analyses into Reusable Workflows

NGS Analysis I: Through Tophat

Galaxy Project Overview

NGS Analysis II: Cufflinks, Visualization

Manage, Reuse, and Share your Analyses

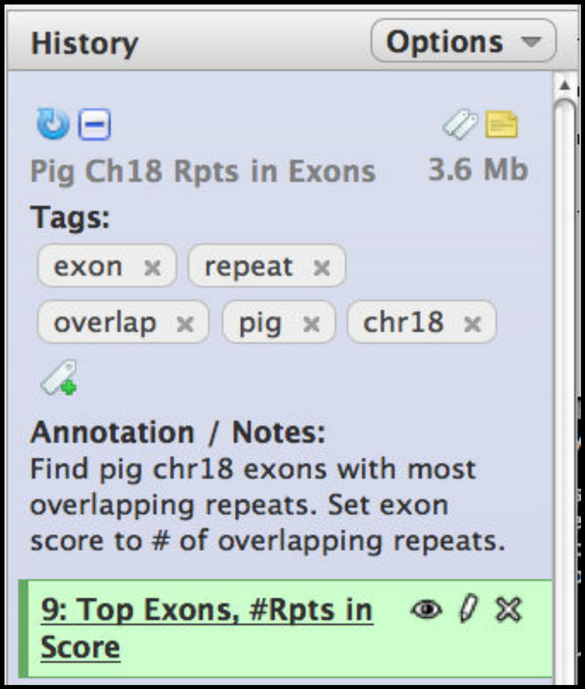
Setting up your own Galaxy on the Cloud

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



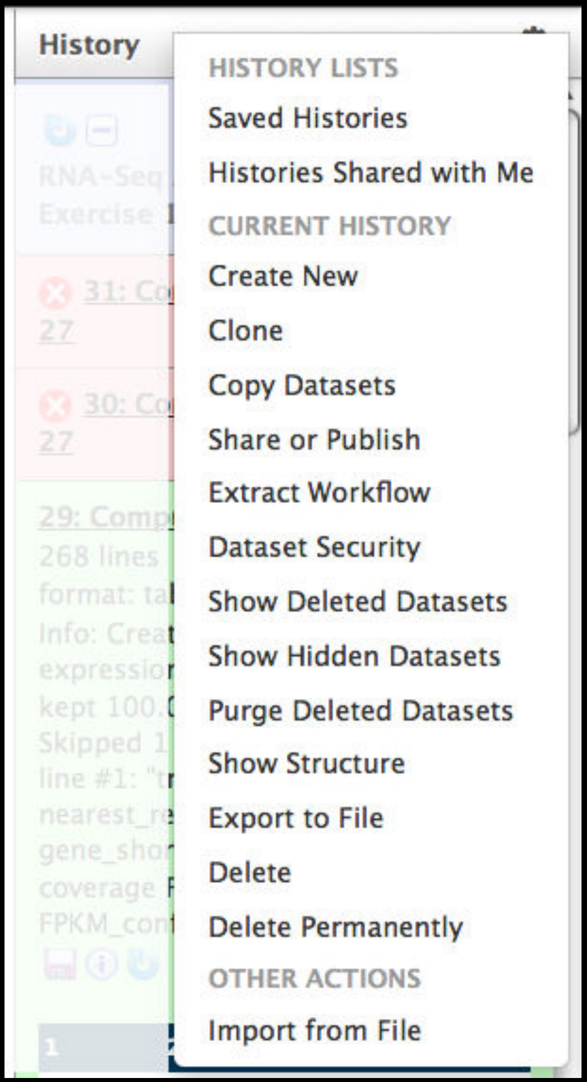
Pig Ch18 Rpts in Exons 3.6 Mb

Tags:

exon × repeat × overlap × pig × chr18 ×

Annotation / Notes:
Find pig chr18 exons with most overlapping repeats. Set exon score to # of overlapping repeats.

9: Top Exons, #Rpts in Score



Some More Galaxy Terminology

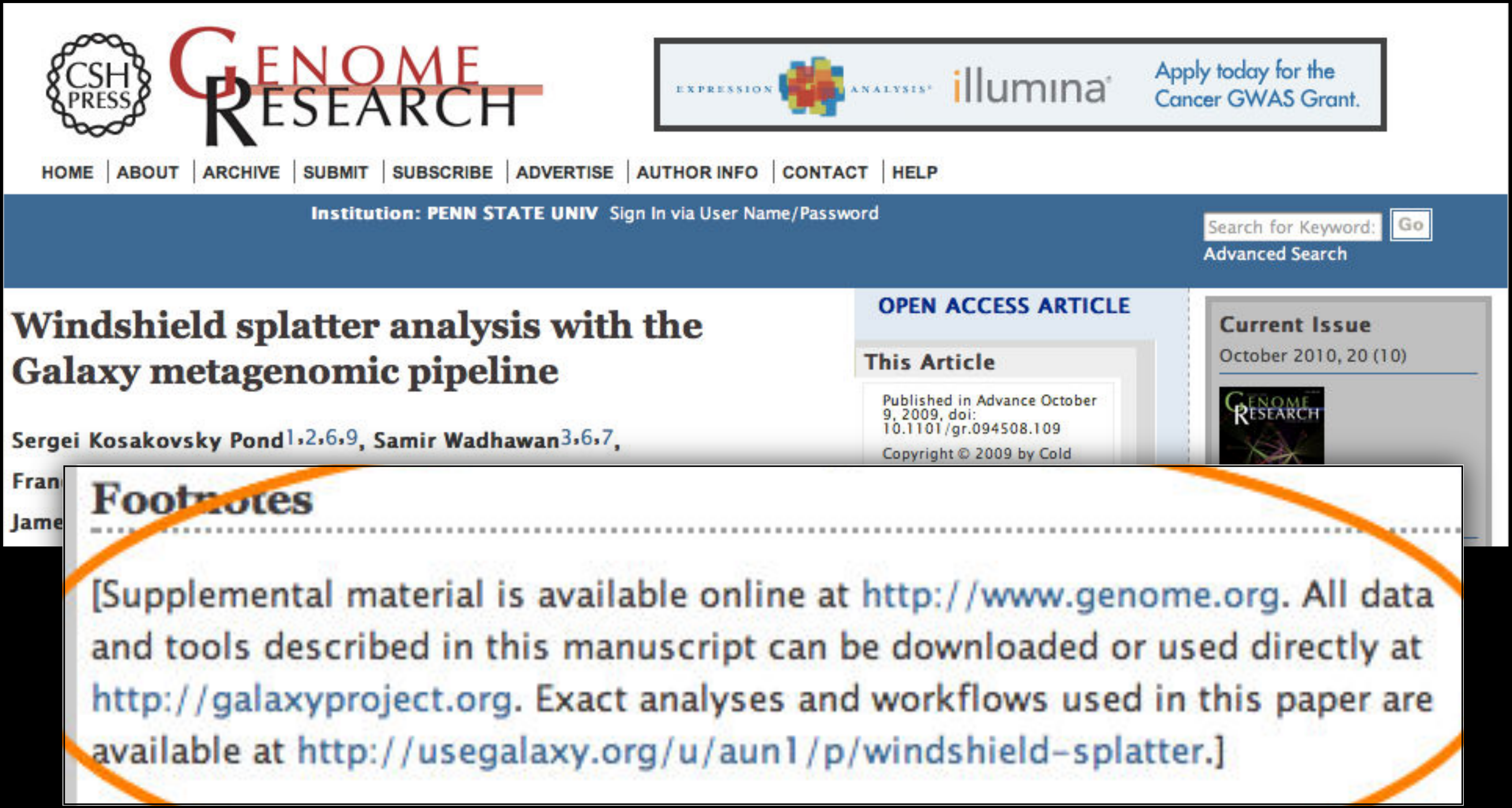
Share:

Make something available to someone else

Publish:

Make something available to everyone and easy to find

Sharing and Publishing Your Work



Institution: PENN STATE UNIV Sign In via User Name/Password

Windshield splatter analysis with the Galaxy metagenomic pipeline

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

Fran

Jame

OPEN ACCESS ARTICLE

This Article

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

Copyright © 2009 by Cold

Current Issue

October 2010, 20 (10)

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

Welcome, Basic Analysis

Basic analyses into Reusable Workflows

NGS Analysis I: Through Tophat

Galaxy Project Overview

NGS Analysis II: Cufflinks, Visualization

Manage, Reuse, and Share your Analyses

Setting up your own Galaxy on the Cloud

Let’s Launch Our Own Galaxy Server

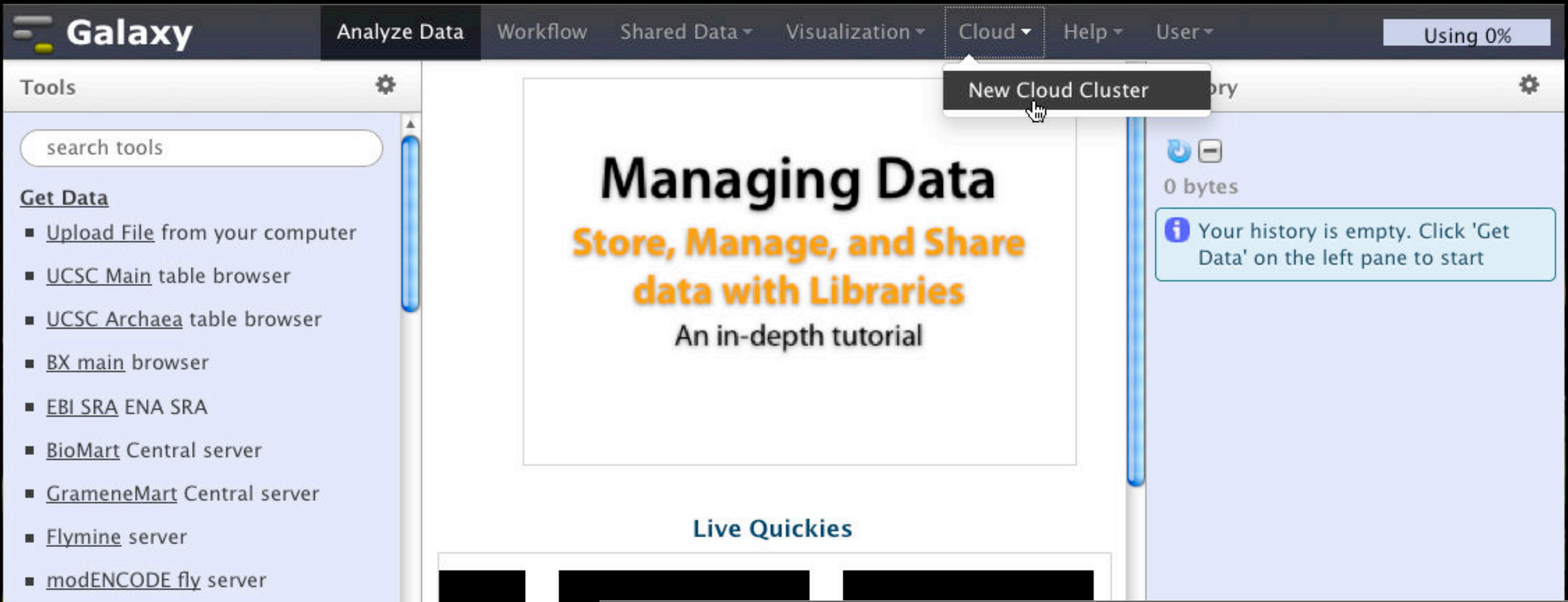
http://bit.ly/gxyawsgetstarted

(http://galaxyproject.org/wiki/CloudMan/AWS/GettingStarted)

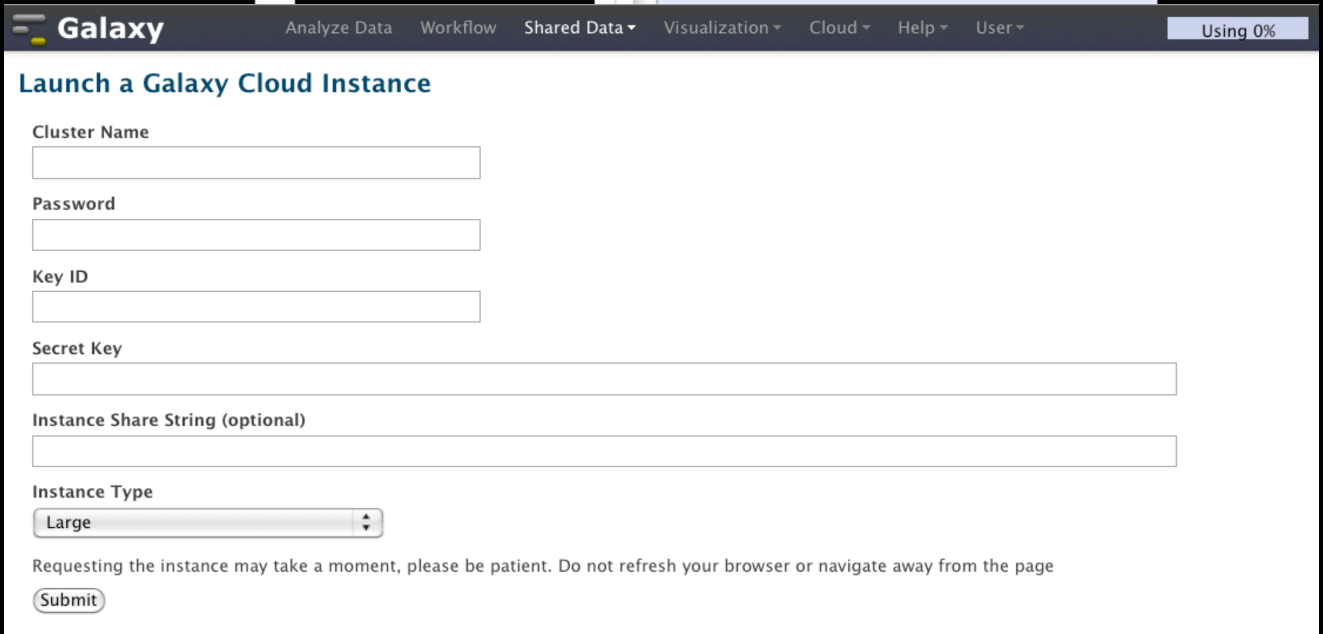
http://bit.ly/PurdueCred

(Access credentials)

Instant CloudMan



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



Launch a Galaxy Cloud Instance

Cluster Name

Password

Key ID

Secret Key

Instance Share String (optional)

Instance Type

Large

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

Submit

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

Workshop Feedback

Please help.

http://bit.ly/gxypurdue

http://bit.ly/gxypurdue

Dave Clements
Emory University

clements@galaxyproject.org
outreach@galaxyproject.org

http://galaxyproject.org/GCC2013

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?

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?*