Introduction to Galaxy

Washington University

St. Louis, Missouri

April 29, 2013

Dave Clements, Emory University

http://galaxyproject.org/

Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

Introductions

In 60 seconds or less tell us

- your name
- your affiliation(s)
- something about your research
- something about what you want to learn

Goals

1. Introduce Galaxy
2. Introduce bioinformatics concepts and formats
3. Hands-on experience
    - Load and integrate data
    - Perform bioinformatic analysis with Galaxy
    - Save, share describe and publish your analyses
    - Visualize your results
    - Set up your own Galaxy server in the cloud

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

Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

Basic Analysis

On human chromosome 22,
which coding exons have the most
repeats in them?

http://cloud1.galaxyproject.org/
http://cloud2.galaxyproject.org/
http://cloud3.galaxyproject.org/

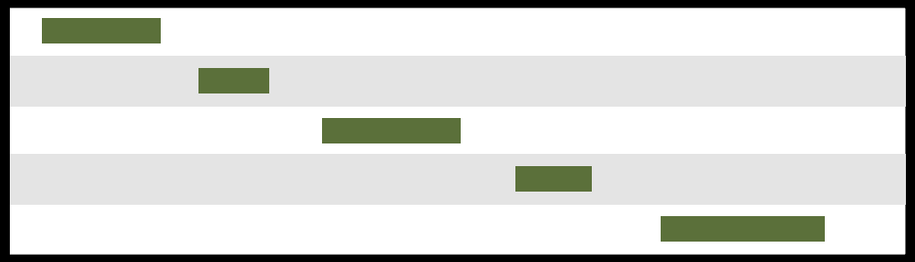
(~ http://usegalaxy.org/galaxy101)

Exons & Repeats: A General Plan

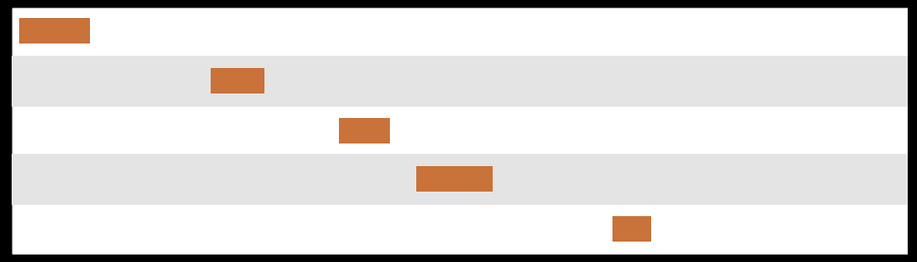
- Get some data
  - Coding exons on chromosome 22
  - Repeats on chromosome 22
- Mess with it
  - Identify which exons have repeats
  - Count repeats per exon
  - Save, download, ... exons with most repeats

http://cloud1.galaxyproject.org/
http://cloud2.galaxyproject.org/
http://cloud3.galaxyproject.org/

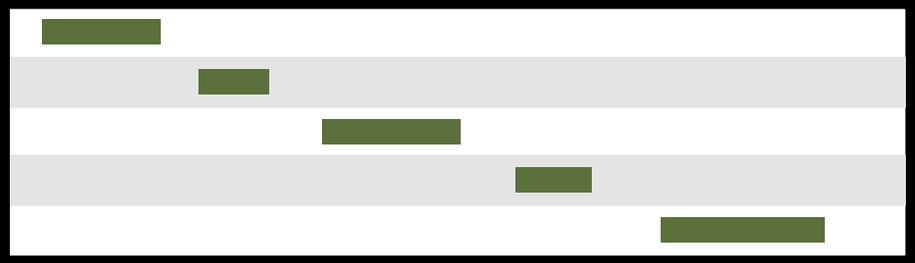
(~ http://usegalaxy.org/galaxy101)



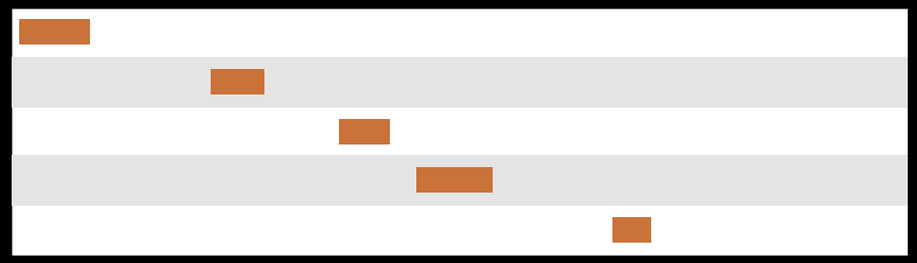
Exons, from UCSC



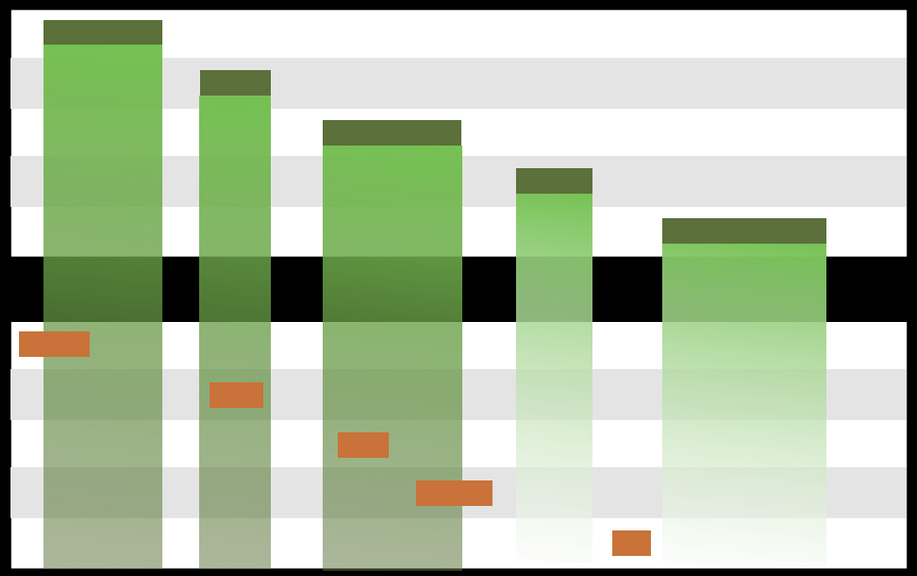
Repeats, from UCSC



Exons, from UCSC



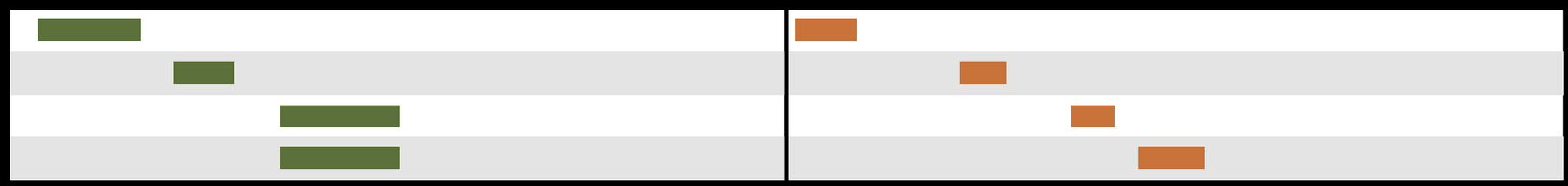
Repeats, from UCSC

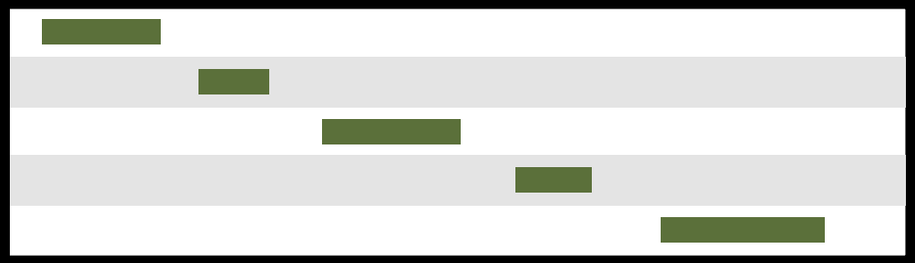


Exons, from UCSC

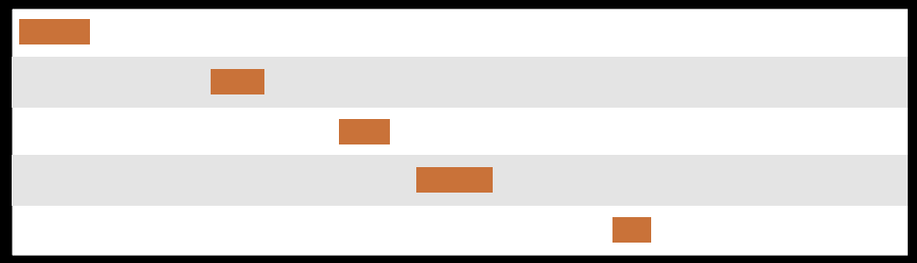
Repeats, from UCSC

Overlap pairings

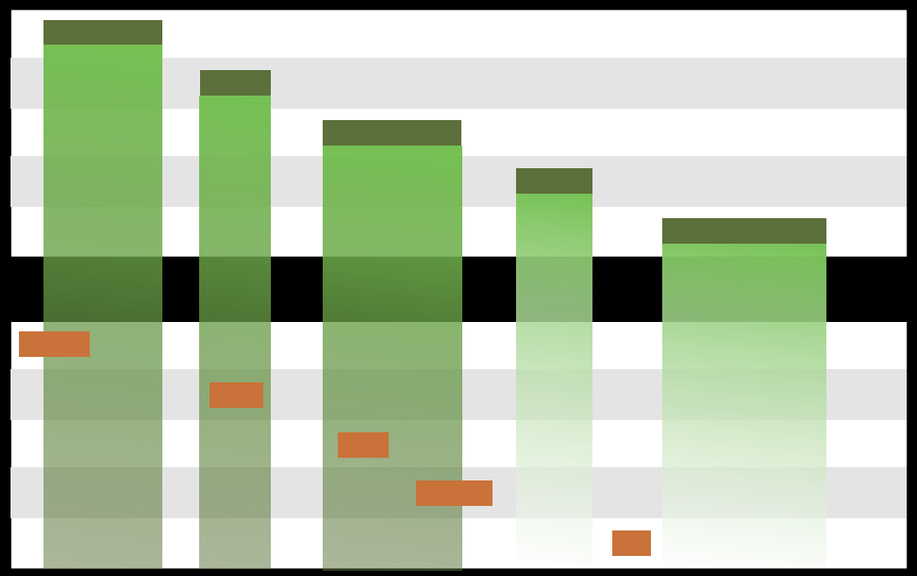




Exons, from UCSC



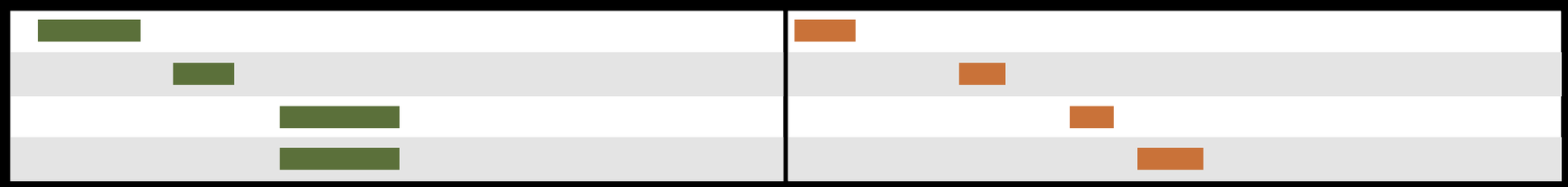
Repeats, from UCSC

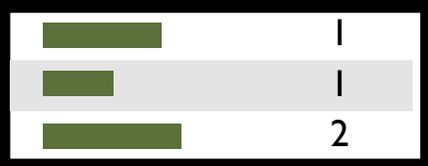


Exons, from UCSC

Repeats, from UCSC

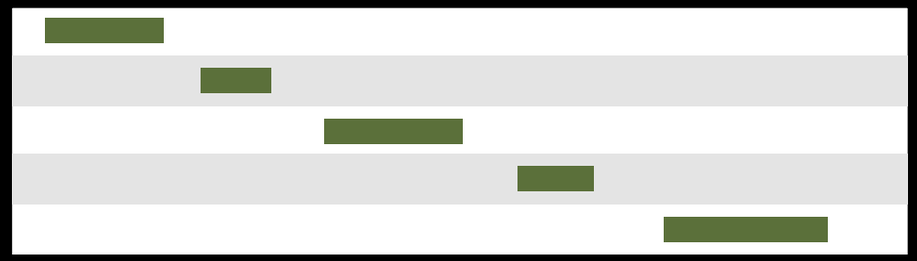
Overlap pairings





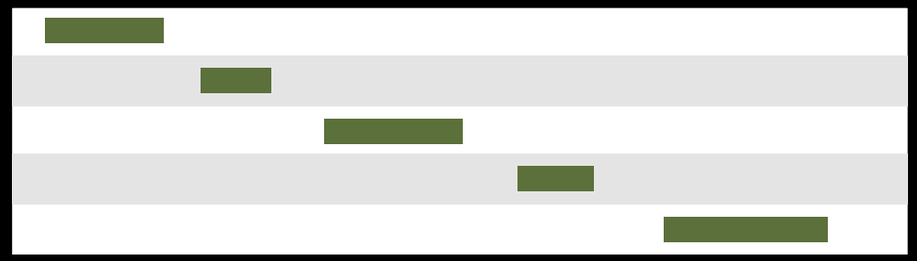
Exon overlap counts

Exon overlap counts



Exons, from UCSC

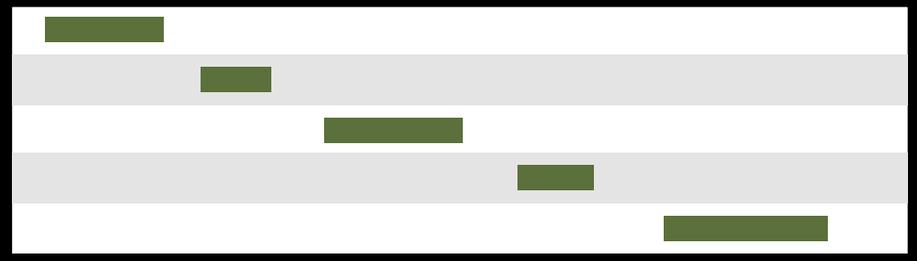
Exon overlap counts



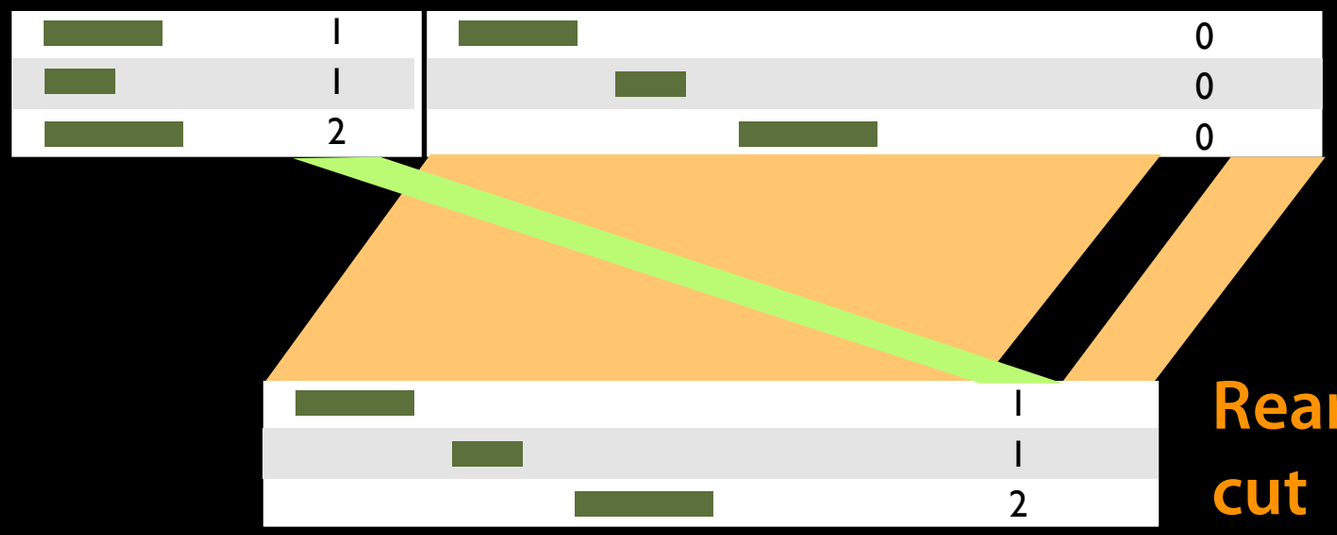
Exons, from UCSC

Join on exon name

Exon overlap counts



Exons, from UCSC



Join on exon name

Rearrange columns w/ cut

Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

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

Exons and Repeats *History* → Reusable *Workflow?*

- The analysis we just finished was about
  - Human chromosome 22
  - Overlap between exons and repeats
- But, ...
  - there is nothing inherently in the analysis about humans, 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.

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

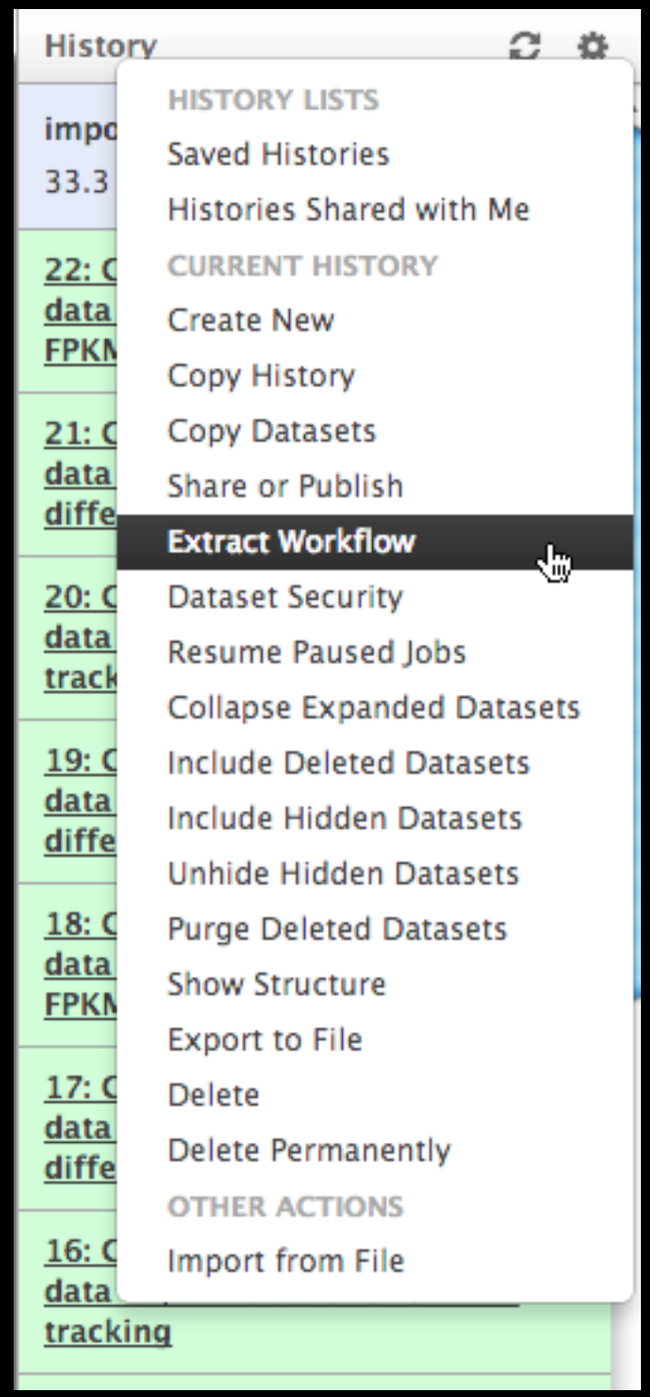
Did that work?

On your own:

Count # of exons in each repeat

Did that work? *Why not?*

Edit workflow: doc assumptions



Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

RNA-seq Exercise

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

http://bit.ly/GxyRNASeqEx

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality
- Trim as we see fit.
- Map the reads to the human reference using Tophat
- Run Cufflinks on Tophat output to assemble reads into transcripts
- Run Cuffdiff on Tophat output to find significant differences

http://bit.ly/GxyRNASeqEx

RNA-seq Exercise: A Plan, Except ...

- At WashU, these steps have likely already been done:
  - Trim as we see fit.
  - Map the reads to the human reference using Tophat
- So, we won't run them today, but we will explain them.

GTAC: http://bit.ly/WashURNA
http://bit.ly/GxyRNASeqEx

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
  - All datasets are FASTQ and from the Body Map 2.0 project
- Shared Data → Published Histories → RNA-Seq with QC through Tophat

http://bit.ly/GxyRNASeqEx

What is FASTQ?

- Specifies sequence (FASTA) and quality scores (PHRED)
- Text format, 4 lines per entry

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

- FASTQ is such a cool standard, there are 3 (or 5) of them!

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger       Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
 I - Illumina 1.3 Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
 X - Solexa       Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

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

RNA-seq Exercise: A Plan

Look at quality Options 1 & 2:

1. NGS QC and Manipulation → Compute Quality Statistics and then Draw quality score boxplot

   No control over how it is calculated or presented.

2. NGS QC and Manipulation → FastQ Summary Statistics, Graph / Display Data → Boxplot of quality statistics

   Lots of control over what the box plot looks like, but no additional information

http://bit.ly/GxyRNASeqEx

RNA-seq Exercise: A Plan

- Look at quality: Option 3
  - NGS QC and Manipulation → **FastQC**
  - Gives you a lot a lot more information but little control over how it is calculated or presented.
- This is what we have done

http://bit.ly/GxyRNASeqEx

RNA-seq Exercise: A Plan

- Look at quality
- Trim as we see fit: Option 1
- **NGS QC and Manipulation → FASTQ Trimmer by column**
- Trim same number of columns from every record
- Can specify different trim for 5' and 3' ends

RNA-seq Exercise: A Plan

"***For the love of all that is holy, please trim your reads!***"

Chris Mason, ABRF NGS Study Report, March 4, 2013

- Look at quality
- Trim as we see fit: Option 1
- **NGS QC and Manipulation → FASTQ Trimmer by column**
- Trim same number of columns from every record
- Can specify different trim for 5' and 3' ends

RNA-seq Exercise: A Plan

- 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/GxyRNASeqEx

RNA-seq Exercise: A Plan

- 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/GxyRNASeqEx

Trim? *As we see fit?*

- Introduced 3 options
  - One preserves original read length, two don't
  - One preserves number of reads, two don't
  - Two keep/make every read the same length, one does not
  - One preserves pairings, two don't
  - Options are not mutually exclusive!

Trim? *As we see fit?*

- Choice depends on downstream tools
- Find out assumptions & requirements for downstream tools and make appropriate choice(s) now.
- How to do that?
  - http://biostars.org/
  - http://seqanswers.com/
  - http://galaxyproject.org/search

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.*
- And thanks to GTAC, we won't run it today.

http://bit.ly/GxyRNASeqEx

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.*

http://bit.ly/11Qvnwh: Lance Parsons' RNA-Seq Tutorial

http://bit.ly/Y7yKzv: UC Davis (Nik Joshi?) RNA-Seq Tutorial

http://bit.ly/YfI0E0: UAB RNA-Seq Tutorial @ GCC2012

http://bit.ly/GxyRNASeqEx

Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

What is Galaxy?

- **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

  http://usegalaxy.org

  However, ***a centralized solution cannot scale to meet the analysis needs of the entire world.***

Galaxy is available ...

- As a free (for everyone) web service

  http://usegalaxy.org

- As open source software

  http://getgalaxy.org

As Open Source Software: 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 a computational resource on which to be deployed

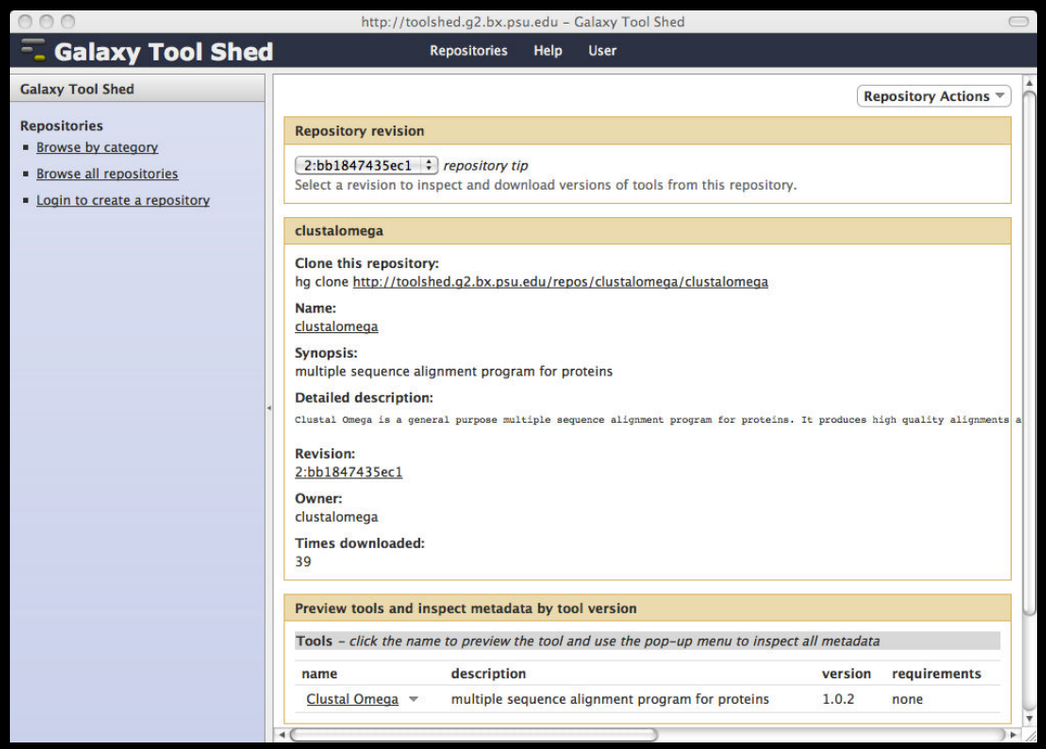
http://getgalaxy.org

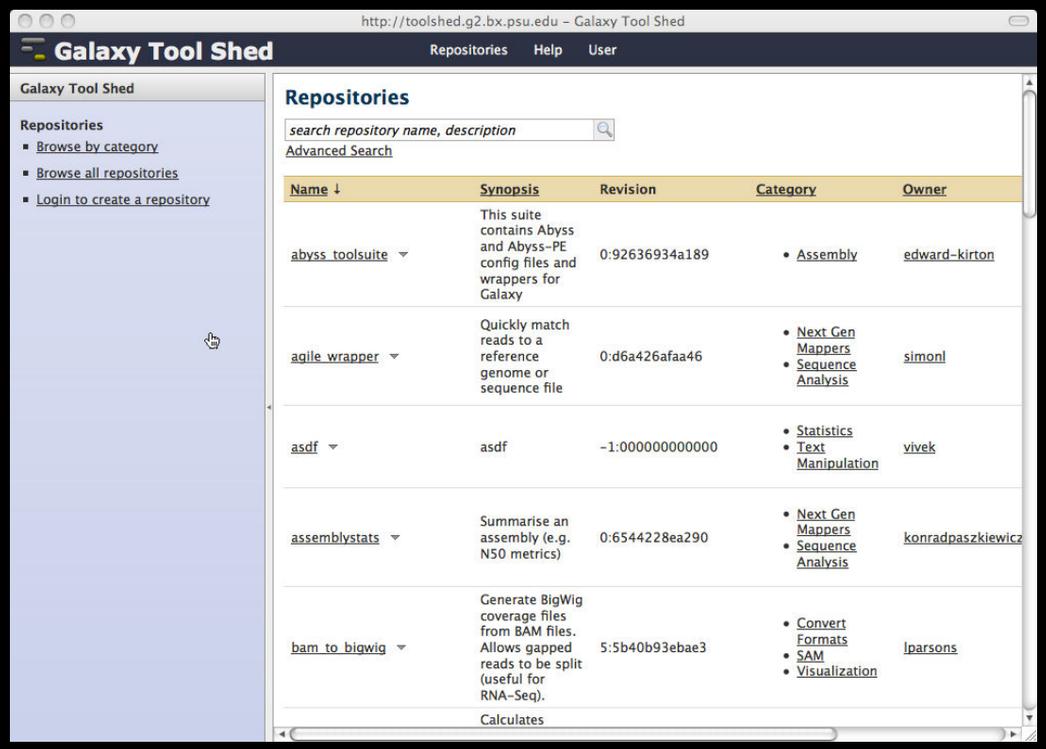
Encourage Local Galaxy Instances

- Encourage and support 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

http://toolshed.g2.bx.psu.edu





Encourage Public Galaxy Instances

http://bit.ly/gxyServers

Interested in:

ChIP-chip and ChIP-seq?
- ✓ Cistrome

Statistical Analysis?
- ✓ Genomic Hyperbrowser

Protein synthesis?
- ✓ GWIPS-viz

de novo assembly?
- ✓ CBIIT Galaxy

Reasoning with ontologies?
- ✓ OPPL Galaxy

Repeats!
- ✓ RepeatExplorer

Everything?
- ✓ Andromeda

Plus many more

As Open Source Software: 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 a **computational resource** on which to be deployed

http://getgalaxy.org

Got your own cluster?

- 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

  http://usegalaxy.org

- As open source software

  http://getgalaxy.org

- *On the Cloud*

  http://usegalaxy.org/cloud

  We are using this right now, and you will set up your own instance today

  http://aws.amazon.com/education

Galaxy Resources and Community

Mailing Lists (very active)
Unified Search
Issues Board
Events Calendar, News Feed
Community Wiki
GalaxyAdmins
Screencasts
Tool Shed
Public Installs
CiteULike group, Mendeley mirror
Annual Community Meting

http://wiki.galaxyproject.org

Galaxy Resources and Community: Mailing Lists

http://wiki.galaxyproject.org/MailingLists

Galaxy-Announce

Project announcements, low volume, moderated

Low volume (42 posts, 1600 members in 2012)

Galaxy-User

Questions about using Galaxy and usegalaxy.org

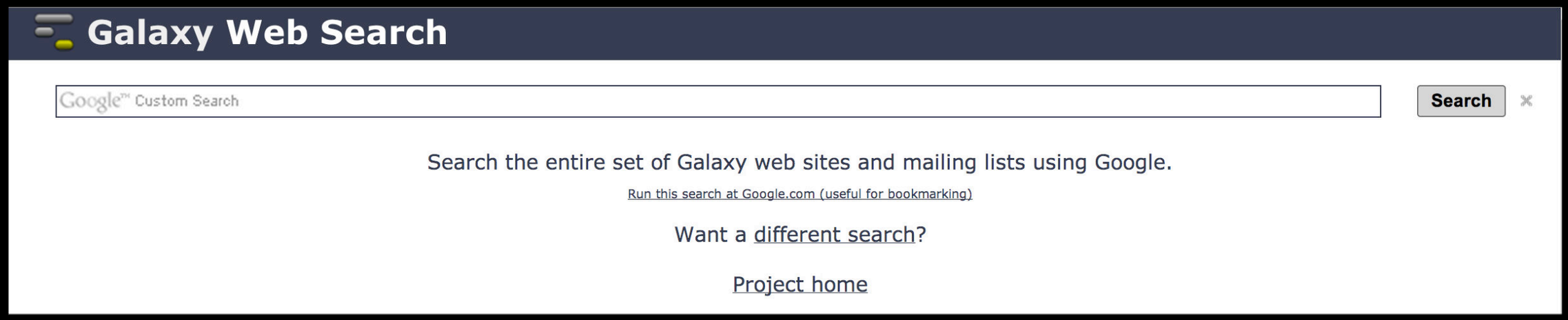
High volume (2900 posts, 2700 members in 2012)

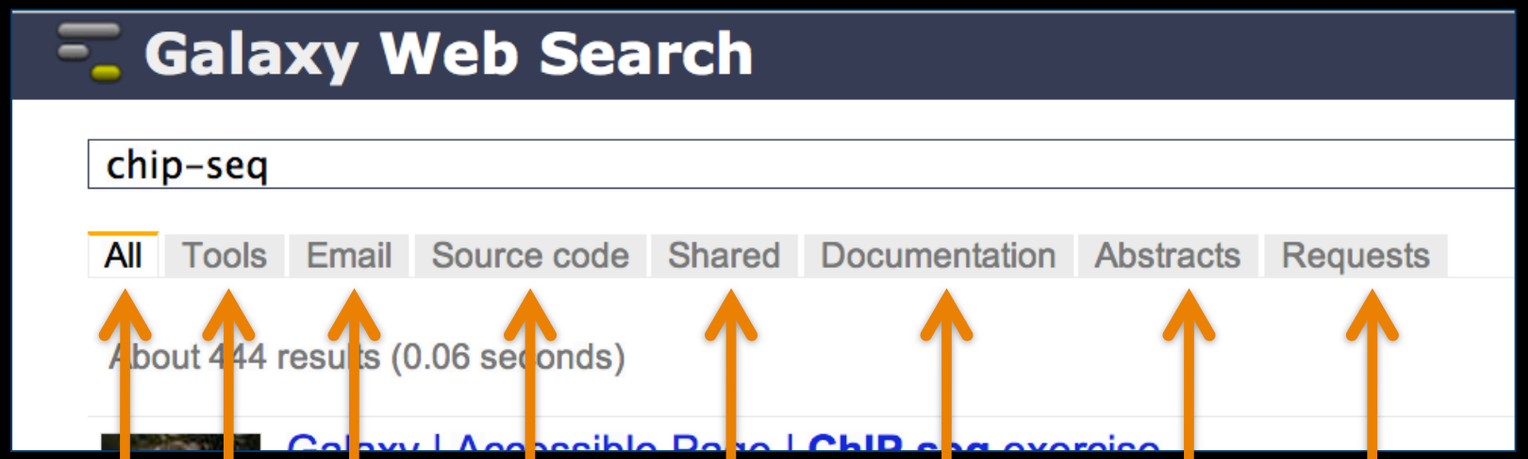
Galaxy-Dev

Questions about developing for and deploying Galaxy

High volume (4500 posts, 850 members in 2012)

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





Find

Everything on …

Tools for …

Email about …

Source code for …

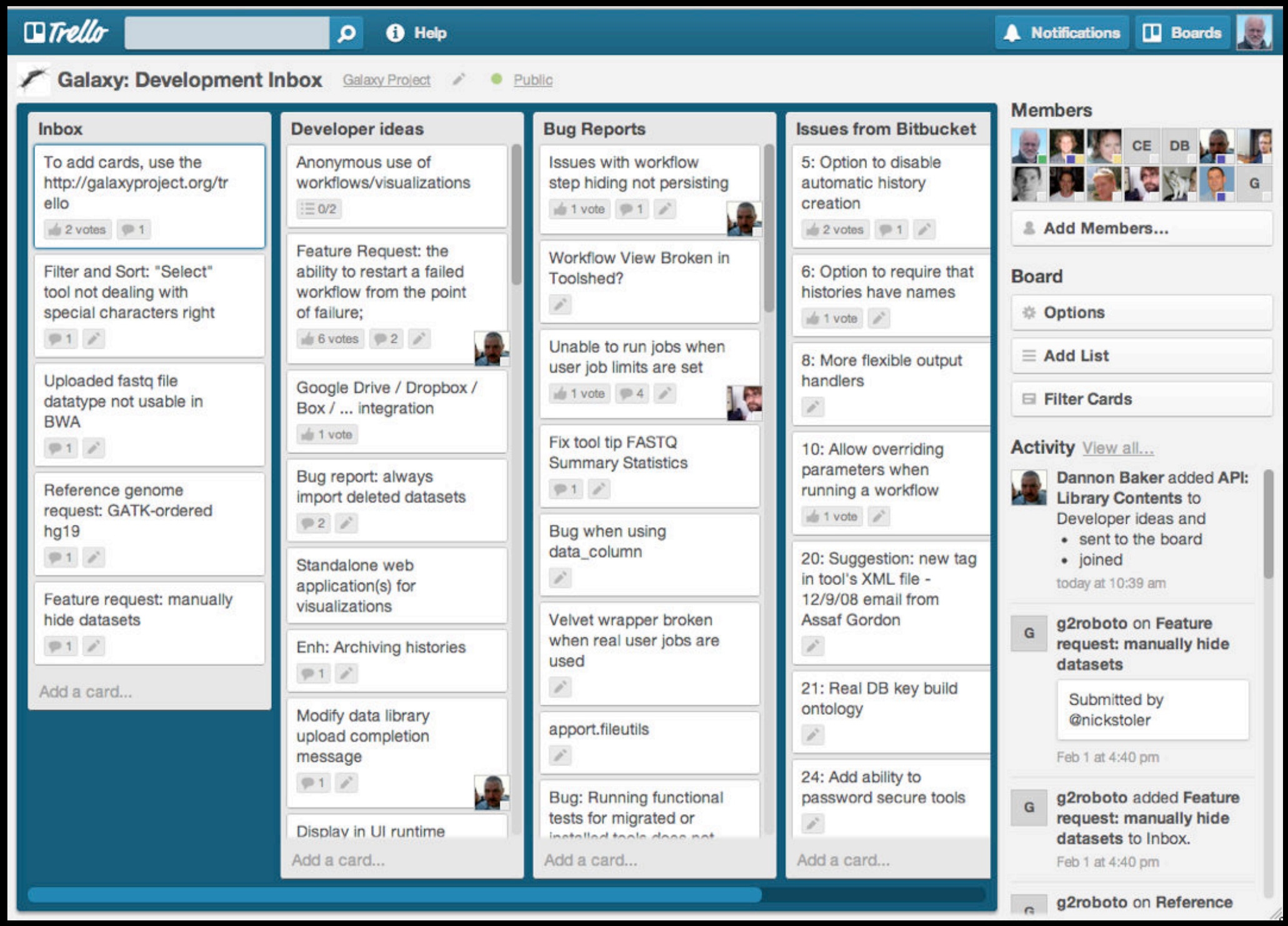
Published Histories, Pages, Workflows, about …

Documentation on …

Papers using Galaxy for …

Related feature requests

Community can create, vote and comment on issues



http://bit.ly/gxyissues

http://wiki.galaxyproject.org

Galaxy

Galaxy is an open, web-based platform for *accessible, reproducible*, and *transparent* computational biomedical research.

- **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible:** Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent:** Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

This is the Galaxy Community Wiki. It describes all things Galaxy.

Use Galaxy

Galaxy's public service web site makes analysis tools, genomic data, tutorial demonstrations, persistent workspaces, and publication services available to any scientist. Extensive user documentation (applicable to any public or local Galaxy instance) is available on this wiki and elsewhere.

usegalaxy.org

Deploy Galaxy

Galaxy is open source for all organizations. Local Galaxy servers can be set up by downloading and customizing the Galaxy application.

- Admin
- Cloud

getgalaxy.org

Community & Project

Galaxy has a large and active user community and many ways to Get Involved.

- Community
- News
- Events
- Support
- Galaxy Project

Contribute

- **Users:** Share your histories, workflows, visualizations, data libraries, and Galaxy Pages, enabling others to use and learn from them.
- **Deployers and Developers:** Contribute tool definitions to the Galaxy Tool Shed (making it easy for others to use those tools on their installations), and code to the core release.
- **Everyone: Get Involved!**

GCC2013 Training Day

Topic voting now open!

Galaxy web search

Use Galaxy

Project Server *(Use it!)*
Other Servers • Learn
Share • Search

Communication

Support • News
Events • Twitter
Mailing Lists (search)

Deploy Galaxy

Get Galaxy • Cloud
Admin • Tool Config
Tool Shed • Search

Contribute

Tool Shed • Share
Issues & Requests
Support

Galaxy Project

Home • About
Community
Big Picture

Events

Galaxy Event Horizon

Events with Galaxy-related content are listed here.

Also see the Galaxy Events Google Calendar for a listing of events and deadlines that are relevant to the Galaxy Community. This is also available as an RSS feed.

If you know of any event that should be added to this page and/or to the Galaxy Event Calendar, please add it here or send it to outreach@galaxyproject.org .

Upcoming Events

Date	Topic/Event	Venue/Location
February 4	*Introduction to Galaxy Boot Camp*	UC Davis Bioinformatics Core Davis, California, United States
March 2-5	*Accessible, Transparent and Reproducible Analysis With Galaxy*, part of SW1: Application of NGS Platforms for Whole Transcriptome and Genome Analysis *Galaxy for Core Facilities*, part of "W6: Community Resource Solutions to Analyzing Large Genomic Data Sets"	**ABRF 2013** Palm Springs, California, United States
March 26-28	*RNA Technologies and Analysis Workshop*	DOE JGI User Meeting
April 5-6	*2013 GMOD Meeting*	Cambridge, United Kingdom, immediately prior to Biocuration 2013
April 7-10	*GO Galaxy Workshop*	Biocuration 2013, Cambridge, United Kingdom
April 9-11	Workshop: *Integrated Research Data Management for Next Gen Sequencing Analysis Using Galaxy and Globus Online Software-as-a-Service* Talk: *Integrated Research Data management and Analysis in NGS using Globus Online, Galaxy and Amazon Web Services*	BioIT World, Boston, Massachusetts, United States
May 14-16	Tutorial: *Exploring and Enabling Biomedical Data Analysis with Galaxy*	Great Lakes Bioinformatics Conference (GLBIO) 2013, Pittsburgh, Pennsylvania, United States
May 21 May 29	*Initiation à l'utilisation de Galaxy* Les deux ateliers sont maintenant complets	Cycle "Bioinformatique par la pratique" 2013, INRA Jouy-en-Josas, France
May 22 May 30	Analyse de données issues de séquenceurs nouvelle génération sous Galaxy Les deux ateliers sont maintenant complets	
June 6-7	*Informatics on High Throughput Sequencing Data Workshop*	Toronto, Ontario, Cananda

News

News

Announcements of interest to the Galaxy Community. These can include items from the Galaxy Team or the Galaxy community and can address anything that is of wide interest to the community.

The Galaxy News is also available as an RSS feed.

See Add a News Item below for how to get an item on this page, and the RSS feed. Older news items are available in the Galaxy News Archive.

See also

- Distribution News Briefs
- Galaxy Updates
- Galaxy on Twitter
- Events
- Learn
- Support
- About the Galaxy Project

News Items

February 2013 Galaxy Update
GCC2013 Training Day Topics: Vote!
Galaxy Project Openings
Jan 11, 2013 Distribution & News Brief
January 2013 GalaxyAdmins
January 2013 Galaxy Update
Dec 20, 2012 Distribution & News Brief
Galaxy Internships @ EMBL
Nominate GCC2013 Training Topics
Dec 3, 2012 Distribution & News Brief
December 2012 Galaxy Update
Nov 14, 2012 Distribution & News Brief
NGS Analysis by Viz. with Trackster
November 2012 GalaxyAdmins

News Archive

News Items

February 2013 Galaxy Update

The February 2013 Galaxy Update is now available.

Highlights:

- Three new public Galaxy servers
- New papers
- Open Positions at five different institutions
- GCC2013 Training Day Topic voting, Registration, and Sponsorships
- January GalaxyAdmins Web Meetup slides and screencast
- Other Upcoming Events and Deadlines
- Galaxy Distributions
- Tool Shed Contributions
- Other News

If you have anything you would like to see in the March *Galaxy Update*, please let us know.

Dave Clements and the Galaxy Team

Posted to the Galaxy News on 2013-02-01

GCC2013 Training Day Topics: Vote!

A list of possible topics for the GCC2013 Training Day is now available. Please take a few minutes to review these possibilities and then vote for your favorite three topics.*

Your votes will determine not only the topics that are offered, but also which topics should be offered more than once, assigned to which rooms, and which ones should not be scheduled at the same time. Your vote matters.

galaxyproject.org/GCC2013

Galaxy Community Conference

STARTING @ $125

(GIGA)n SCIENCE

GCC2013 Training Day

galaxyproject.org/GCC2013

Galaxy Community Conference

STARTING @ $125

Poster abstracts due May 3

(GIGA)n SCIENCE

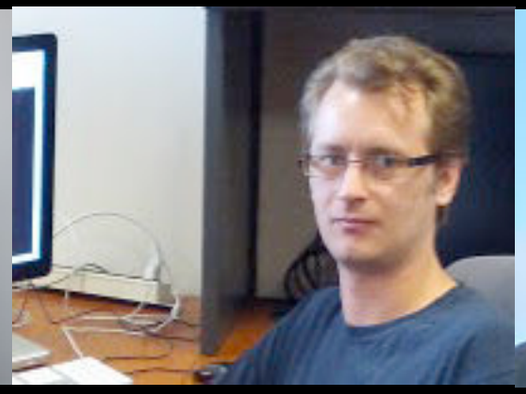
GCC2013 Training Day

The Galaxy Team

Enis Afgan

Dannon Baker

Dan Blankenberg



Dave Bouvier

Dave Clements

Nate Coraor

Carl Eberhard

Dorine Francheteau

Jeremy Goecks

Sam Guerler

Jen Jackson

Greg von Kuster

Ross Lazarus

Anton Nekrutenko

James Taylor

http://wiki.galaxyproject.org/GalaxyTeam

Galaxy is hiring post-docs and software engineers
at both Emory and Penn State.

Please help.

http://wiki.galaxyproject.org/GalaxyIsHiring

Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

RNA-seq Exercise: A Plan

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

http://bit.ly/GxyRNASeqEx

RNA-seq Exercise: A Different *(better?)* Plan

- We've now walked through the analysis
- Suggests our level of iteration / experimentation is the whole workflow.
- Use Visualization as we go, to guide analysis as we go.

http://bit.ly/GxyRNASeqEx

Visualizing Genomics

Supported external browsers

- UCSC
- Ensembl
- GBrowse
- IGB
- IGV

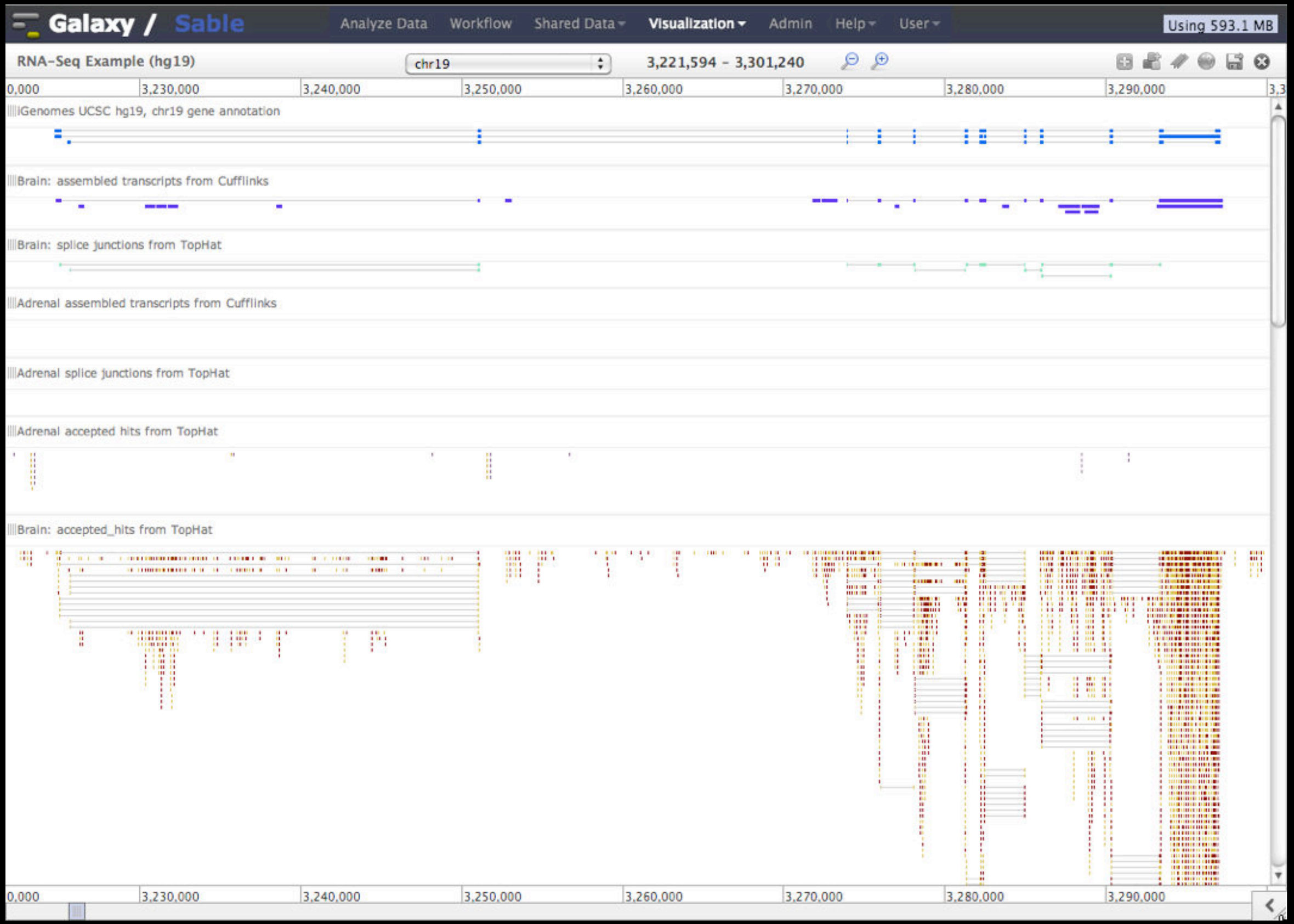
Traditional browser strengths:

- Showing what is nearby
- what else is happening here
- highlighting correlations
- integrating many datasets

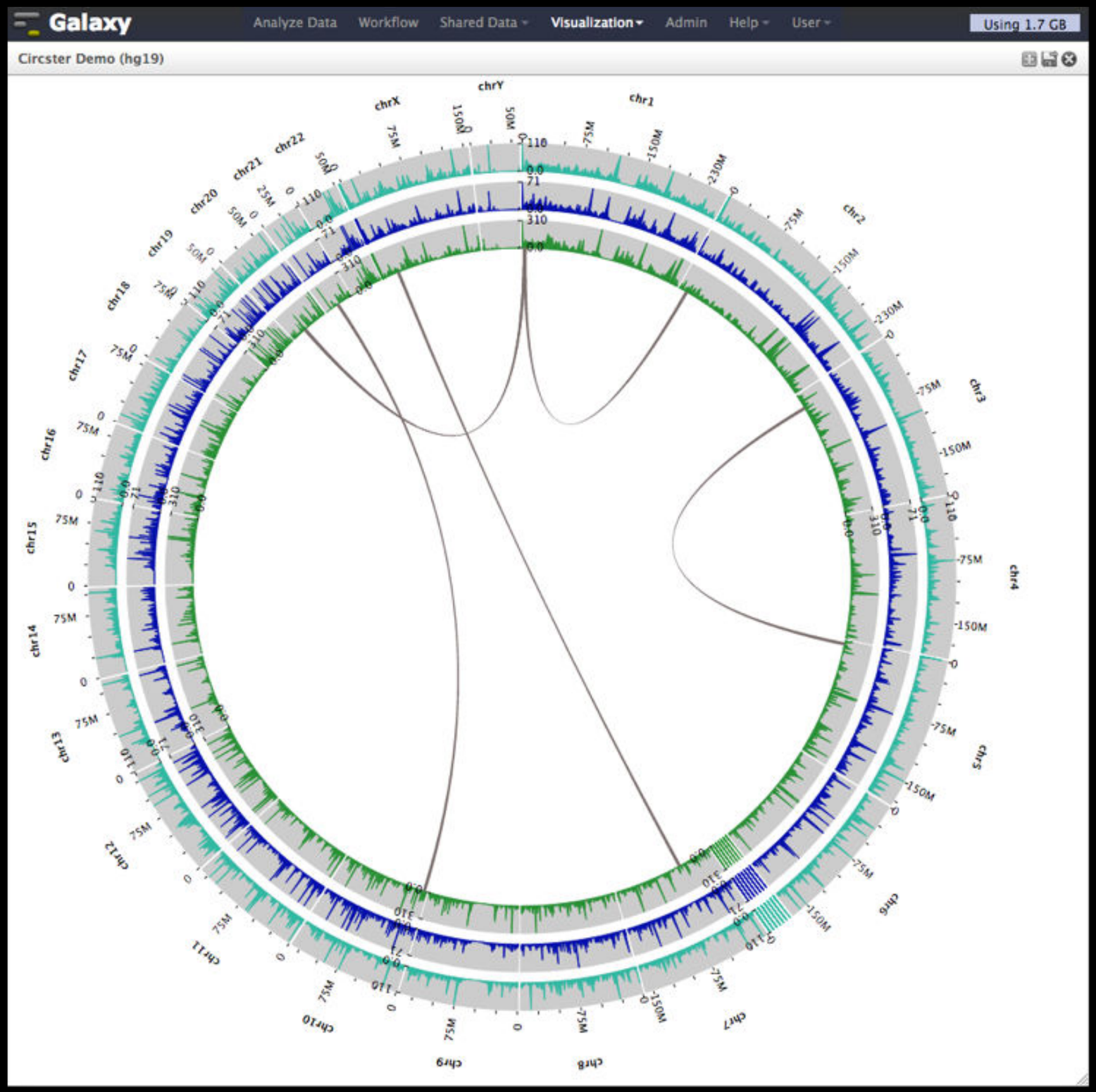
But, *wouldn't it be nice to*

- Use visualization to evaluate and refine analyses?
- Expose some basic analyses in visualization to make it more informative?
- Make that analyze-visualize-refine loop seamless and fast? That is, integrate the two?
- Use visualization to learn tools and explore their parameter space?
- Not be tied to a predefined reference genome?

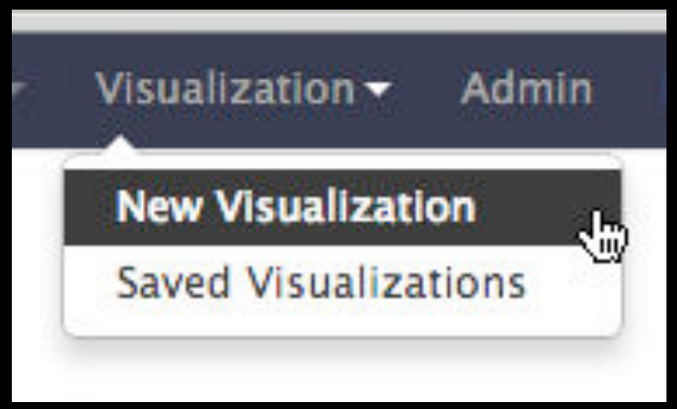
Trackster: Galaxy's embedded track browser



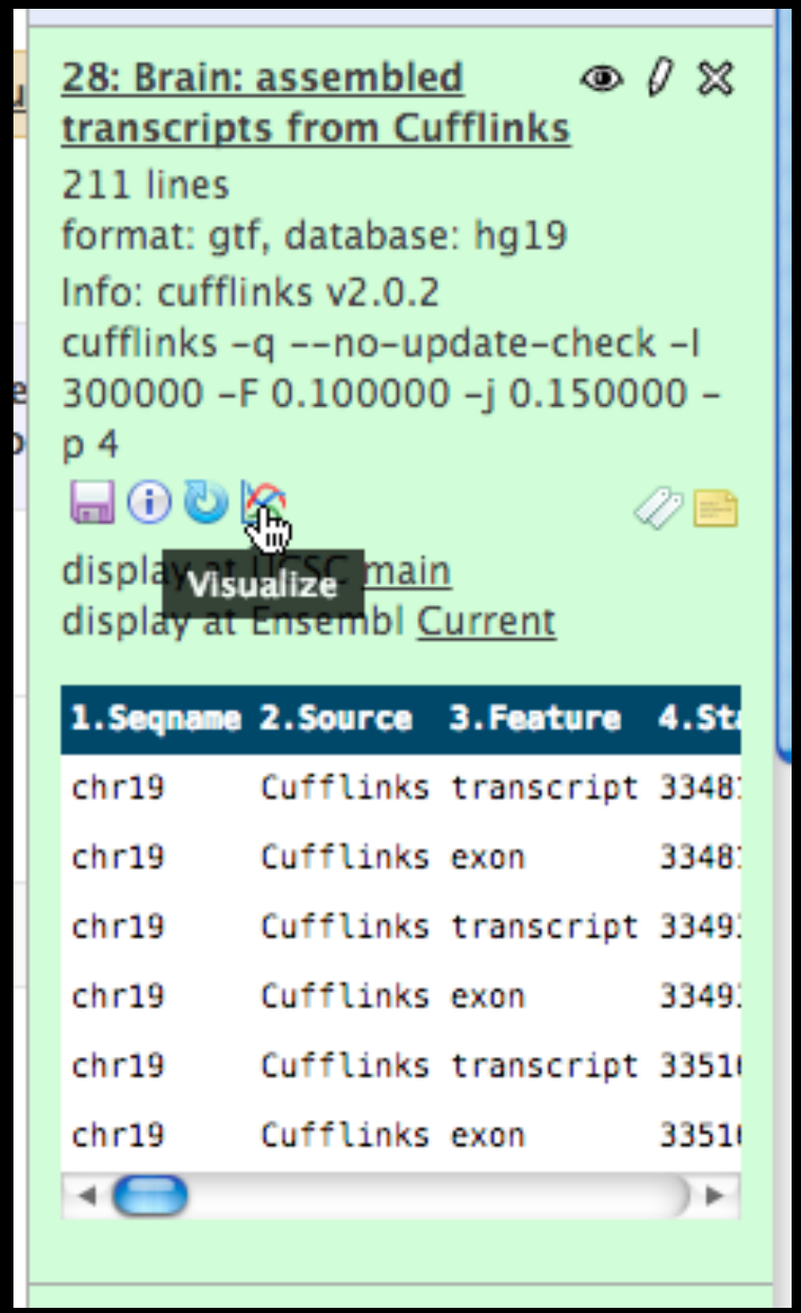
Circster


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 1.7 GB
Circster Demo (hg19)
chrY
chrX
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22

Create a visualization in Galaxy



or



Isn't it nice to

- To do all those things we talked about?
  - Use visualization to evaluate and refine analyses?
  - Expose some basic analyses in visualization to make it more informative?
  - Make that analyze-visualize-refine loop seamless and fast? That is, integrate the two?
  - Use visualization to learn tools and explore their parameter space?
  - Not be tied to a predefined reference genome?

Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

More Galaxy Terminology

Share:

Make something available to someone else

Publish:

Make something available to everyone

Galaxy Page:

Analysis documentation within Galaxy; easy to embed any Galaxy object

Sharing & Publishing enables Reproducibility

Reproducibility: Everybody talks about it, but ...

Galaxy aims to push the goal of reproducibility from the bench to the bioinformatics realm

All analysis in Galaxy is recorded without any extra effort from the user.

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

Sharing & Publishing enables Reproducibility

CSH PRESS GENOME RESEARCH

EXPRESSION ANALYSIS illumina Apply today for the Cancer GWAS Grant.

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

Copyright © 2009 by Cold Spring Harbor Laboratory Press

- » Abstract *Free*
- Full Text (PDF) *Free*

Current Issue

October 2010, 20 (10)

Sharing & Publishing enables Reproducibility

CSH PRESS GENOME RESEARCH

EXPRESSION ANALYSIS illumina Apply today for the Cancer GWAS Grant.

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

Copyright © 2009 by Cold Spring Harbor Laboratory Press

» Abstract *Free*
Full Text (PDF) *Free*

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.]

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

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) – a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Galaxy History | Galaxy vs MEGAN
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 manuscript and **Figure 3A**):

Galaxy History | metagenomic analysis

This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

Galaxy Workflow | metagenomic analysis
Generic workflow for performing a metagenomic analysis on NGS data.

Accessing the Data

Windshield Splatter datasets analyzed in this manuscript can be accessed through this Galaxy Library. From

About this Page

Author

aun1

Related Pages

All published pages
Published pages by aun1

Rating

Community
(6 ratings, 5.0 average)

Tags

Community:
paper galaxy megan

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

Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:20 RNA-Seq Example Part I
12:20 Lunch (on your own)
1:35 Galaxy Project Overview
1:55 RNA-Seq Example Part II
2:45 Break
3:05 Sharing, Publishing and Reproducibility
3:25 Setting up your own Galaxy Cluster on AWS
5:00 Done

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.**

- **You will set up an instance now**

http://aws.amazon.com/education

Could do this step by step, but ...

http://bit.ly/GXYAWSGetStarted

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

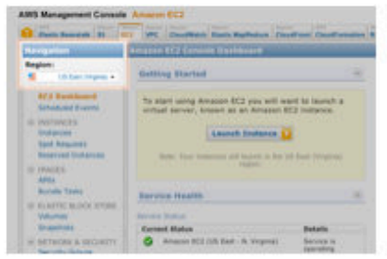
Contents

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 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 supported in at this time (see screenshot 1.2).
3. Click **Network & Security → Key Pairs** or **My Resources → *n* Key Pairs** (see screenshot 1.3 - if it does not look like this, then try using the Chrome browser) and then click **Create Key Pair**. Enter a memorable name for the key pair, e.g., `GalaxyCloud` and click **Create**.
4. *Save your private key!* The previous step creates the key pair and downloads a copy to your machine with the name `MemorableName.pem`. Save this file and protect it like you would your password. The key pair can be used to access started instances from

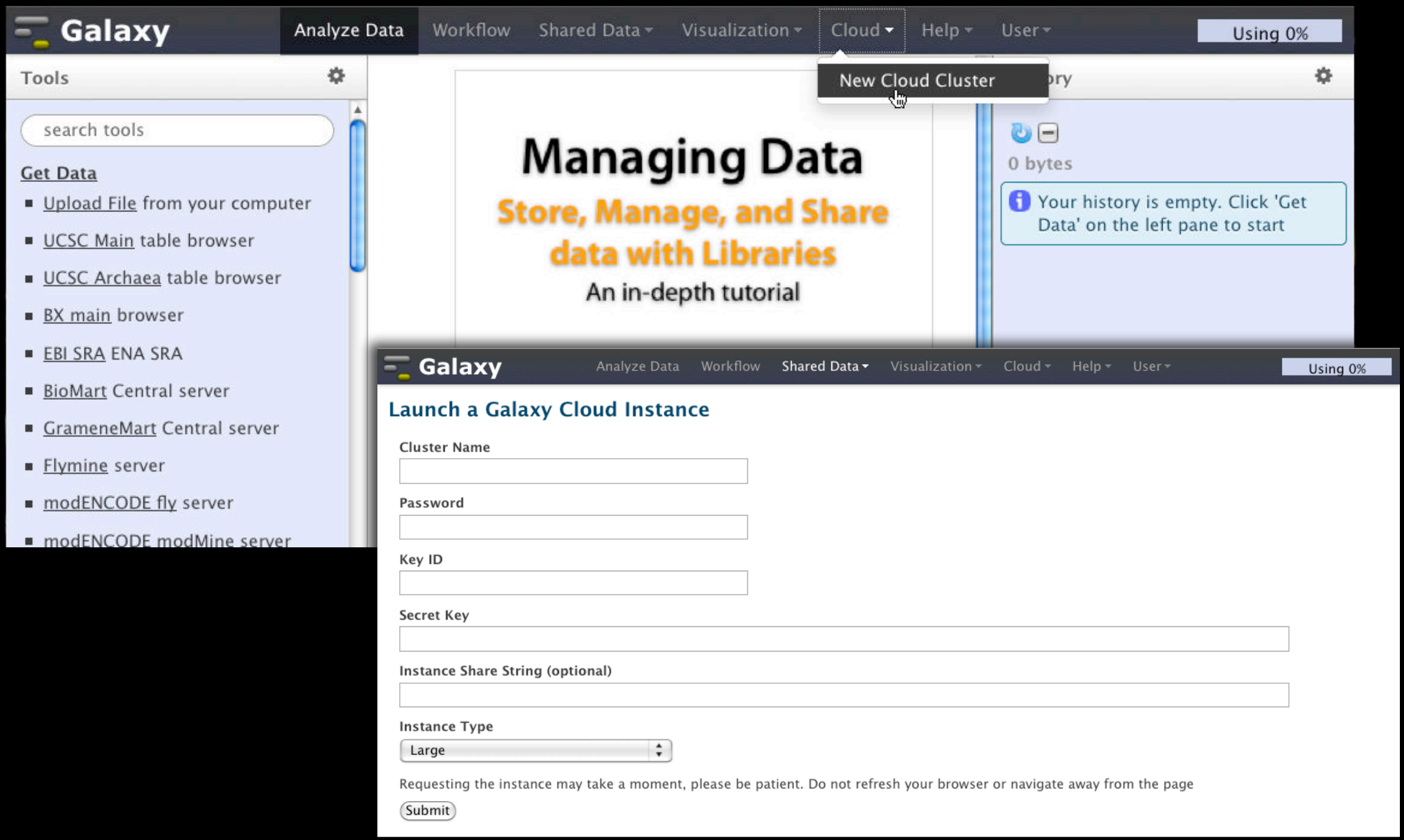
Step 1 Screenshots



1.2. Set region

Instant CloudMan

http://usgalaxy.org/cloudlaunch



Instant CloudMan

AWS Credentials

http://bit.ly/WashUMon

Instant Feedback

http://bit.ly/20130429Gxy

Acknowledgements

Kristi Holmes
Bob Engeszer
Maureen Donlin

The Galaxy Team
You!

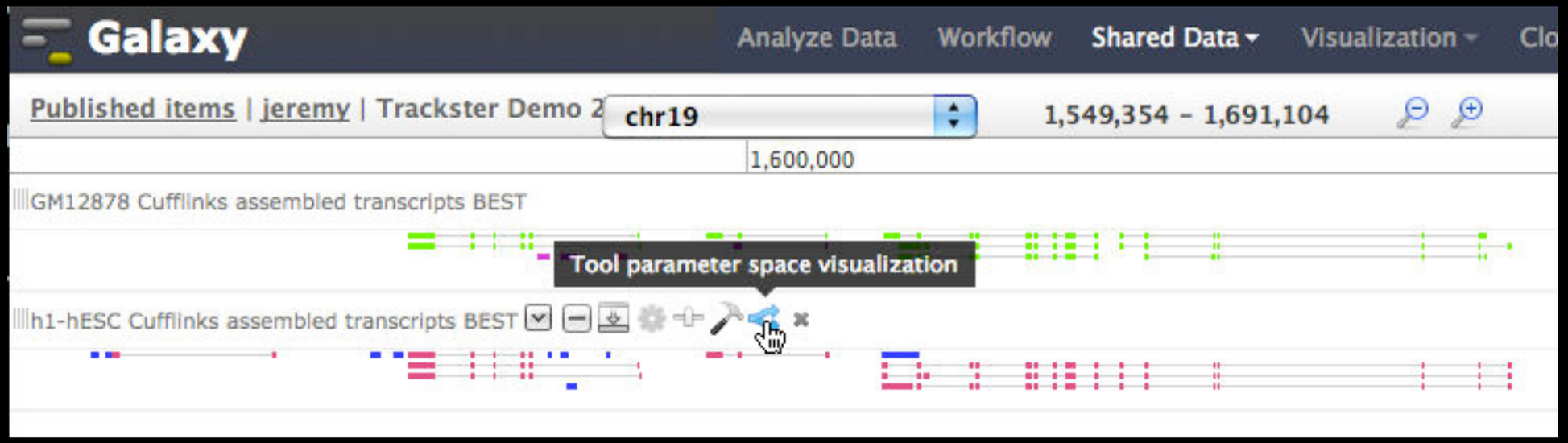
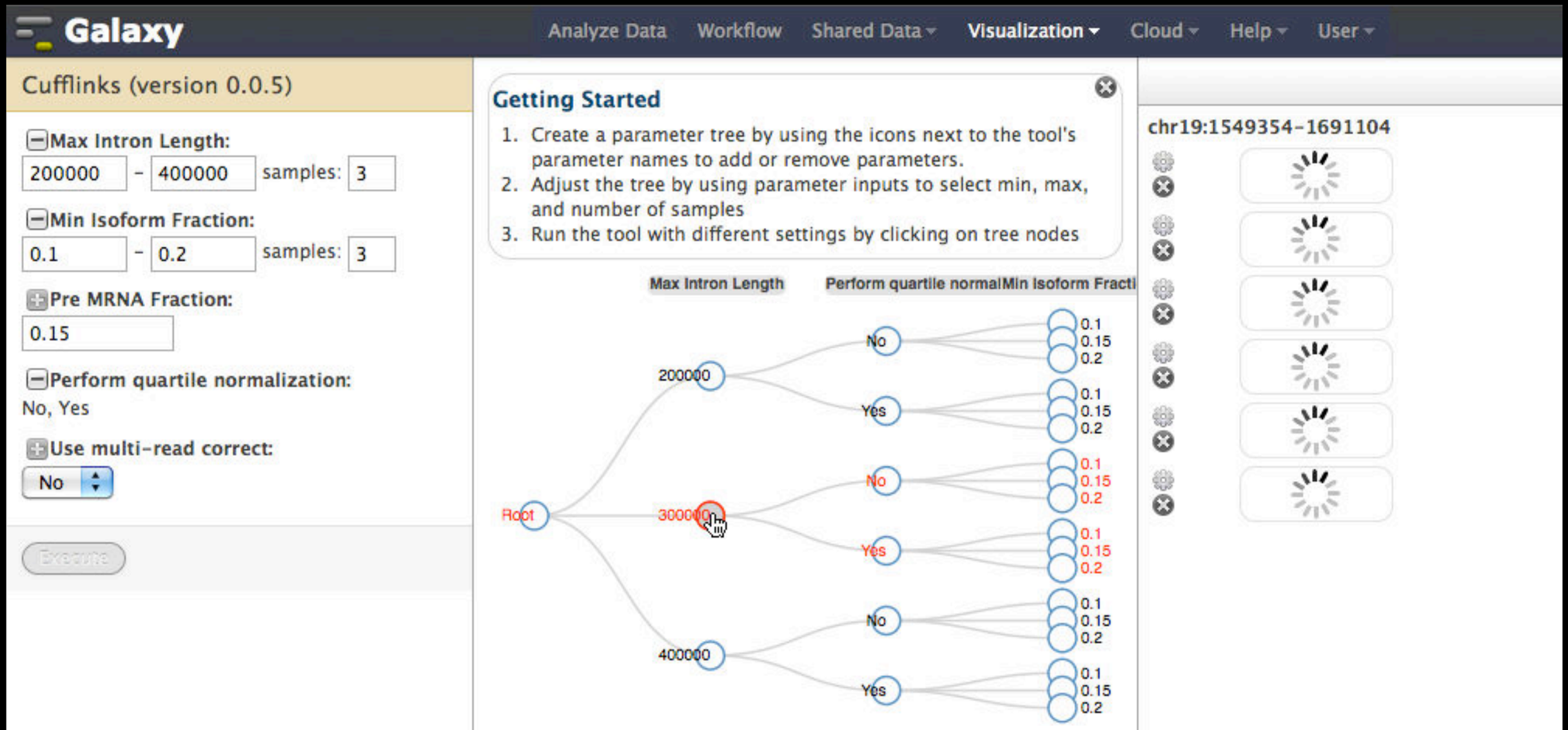
Bernard Becker Medical Library
Washington University in St. Louis

AWS Education Grant

NIH NSF Huck Institute
Penn State University Emory University

http://bit.ly/20130429Gxy

Thanks

http://bit.ly/20130429Gxy

Exploring Parameter Space with Trackster