

Galaxy for Biologists

California Institute of Technology
March 5, 2013

Dave Clements, Emory University
<http://galaxyproject.org/>



Sternberg Lab



The Galaxy logo, which features a stylized icon of three horizontal bars of varying lengths to the left of the word "Galaxy" in a bold, sans-serif font.

Agenda

- 11:00 **Basic Analysis with Galaxy**
- 12:20 Basic Analysis into Reusable Workflows
- 12:40 Lunch
- 1:20 NGS Quality Control
- 1:40 Galaxy Project Overview
- 2:00 Finish

Goals

1. Introduce **Galaxy**
2. Introduce **bioinformatics concepts and formats**
3. Hands-on experience
 - **Load and integrate data**
 - **Perform bioinformatic analysis with Galaxy**

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

- 11:00 **Basic Analysis with Galaxy**
- 12:20 Basic Analysis into Reusable Workflows
- 12:40 Lunch
- 1:20 NGS Quality Control
- 1:40 Galaxy Project Overview
- 2:00 Finish

Basic Analysis

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

<http://bit.ly/ctorange>

<http://bit.ly/ctwhite>

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

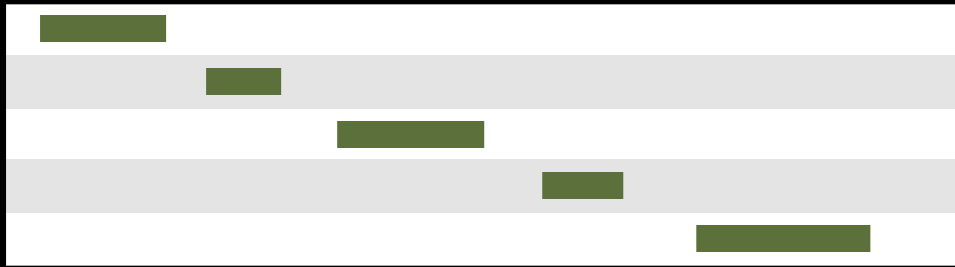
Exons & Repeats: A General Plan

- 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

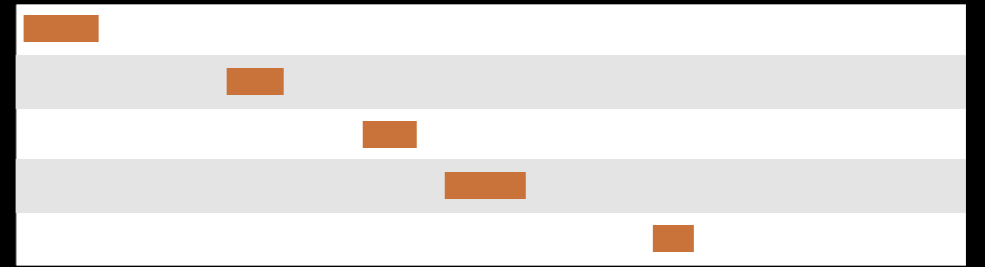
<http://bit.ly/ctorange>

<http://bit.ly/ctwhite>

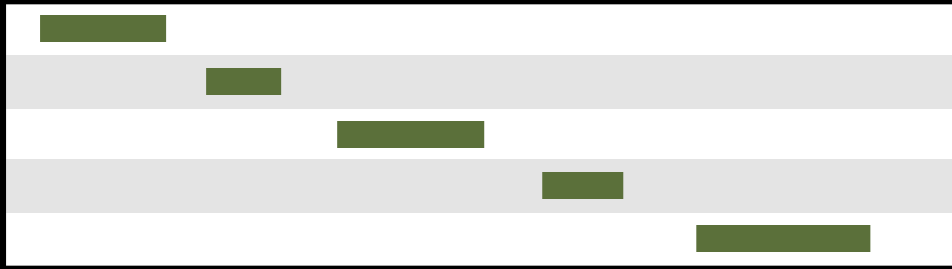
(~ <http://usegalaxy.org/galaxy101>)



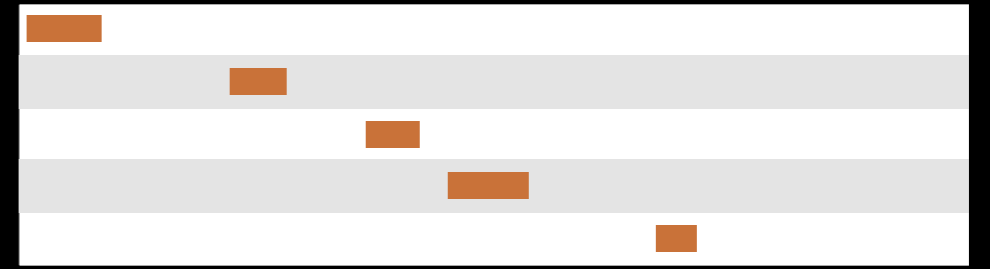
Exons, from UCSC



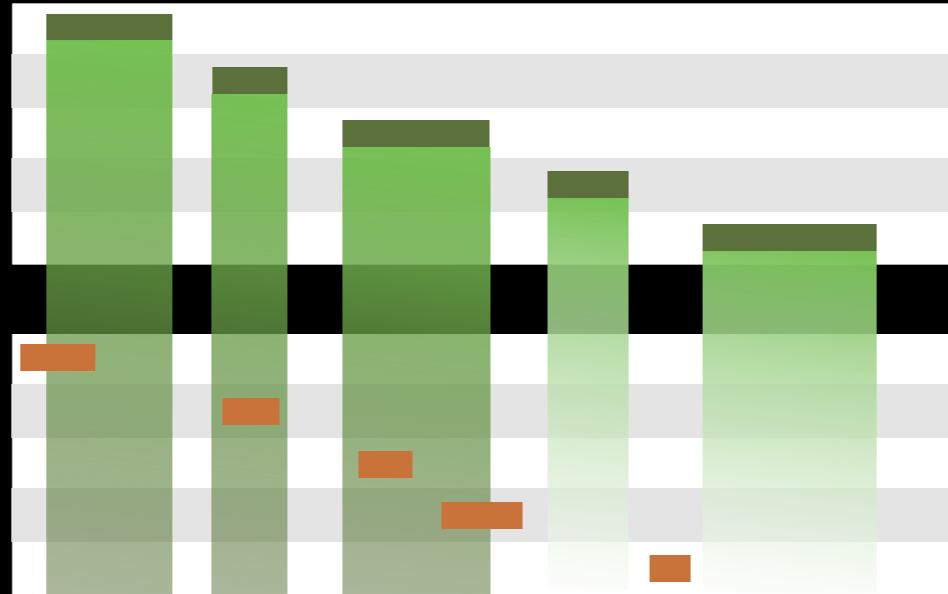
Repeats, from UCSC



Exons, from UCSC



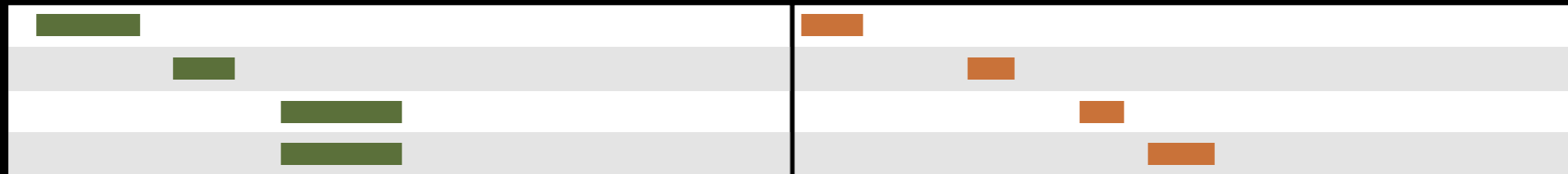
Repeats, from UCSC

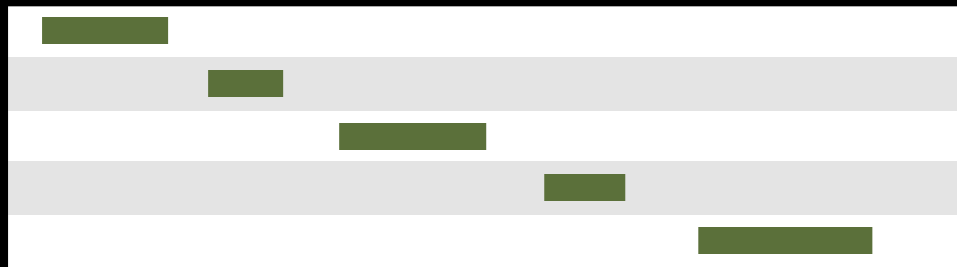


Exons, from UCSC

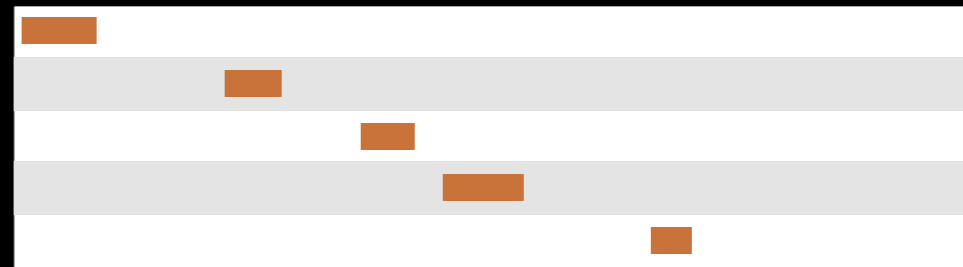
Repeats, from UCSC

Overlap pairings

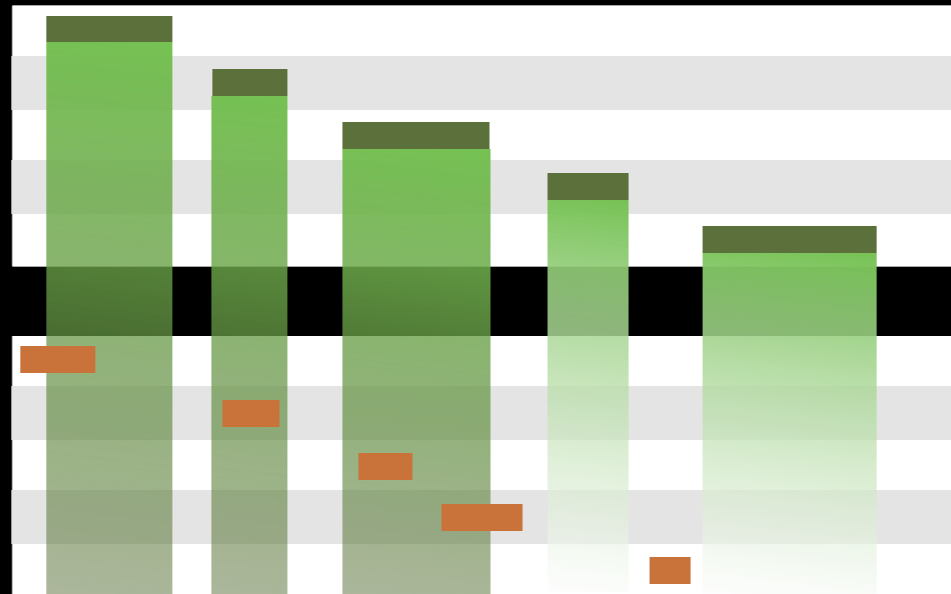




Exons, from UCSC



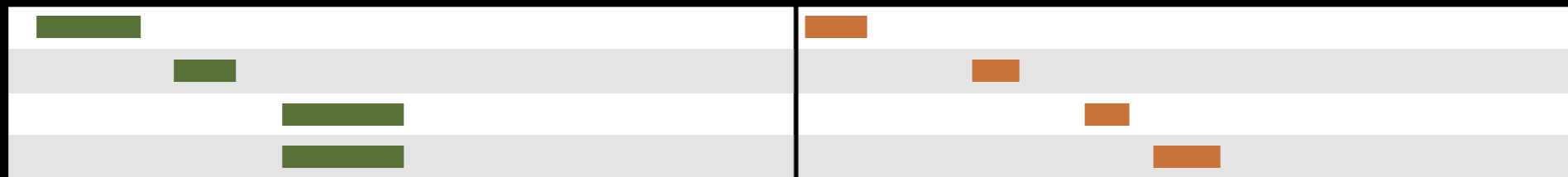
Repeats, from UCSC



Exons, from UCSC

Repeats, from UCSC

Overlap pairings

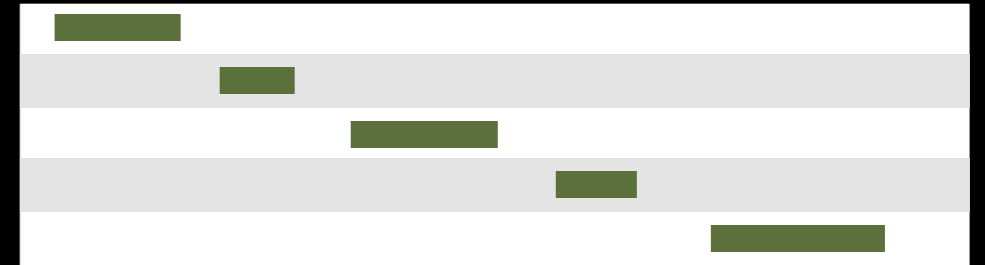


	1
	1
	2

Exon overlap counts



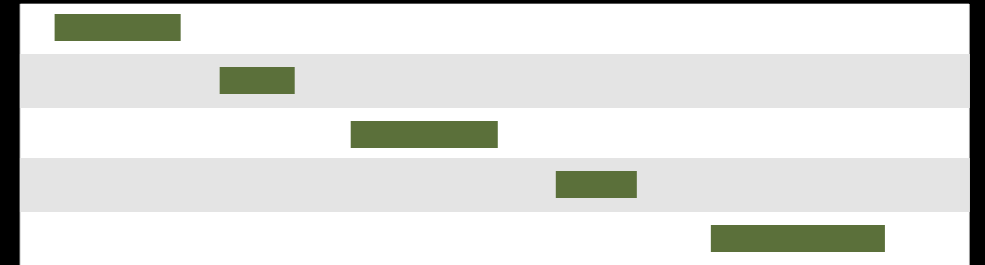
Exon overlap counts



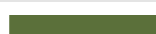
Exons, from UCSC

	1
	1
	2

Exon overlap counts



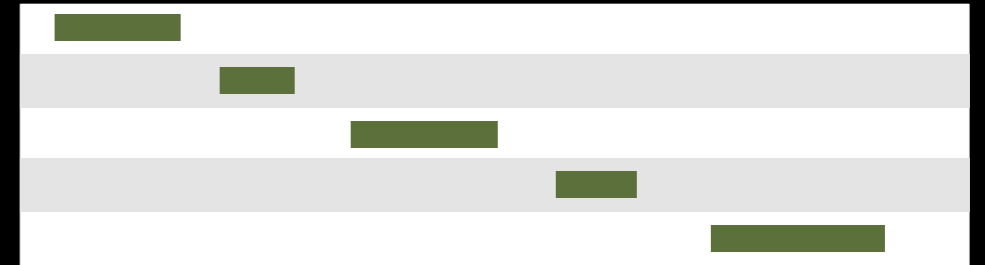
Exons, from UCSC

	1		0
	1		0
	2		0





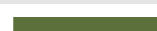
Join on exon name

	1
	1
	2




Exon overlap counts



Exons, from UCSC

	1		0
	1		0
	2		0

Join on exon name

	1
	1
	2

Rearrange columns w/
cut

Agenda

- 11:00 **Basic Analysis with Galaxy**
- 12:20 Basic Analysis into Reusable Workflows
- 12:40 Lunch
- 1:20 NGS Quality Control
- 1:40 Galaxy Project Overview
- 2:00 Finish

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

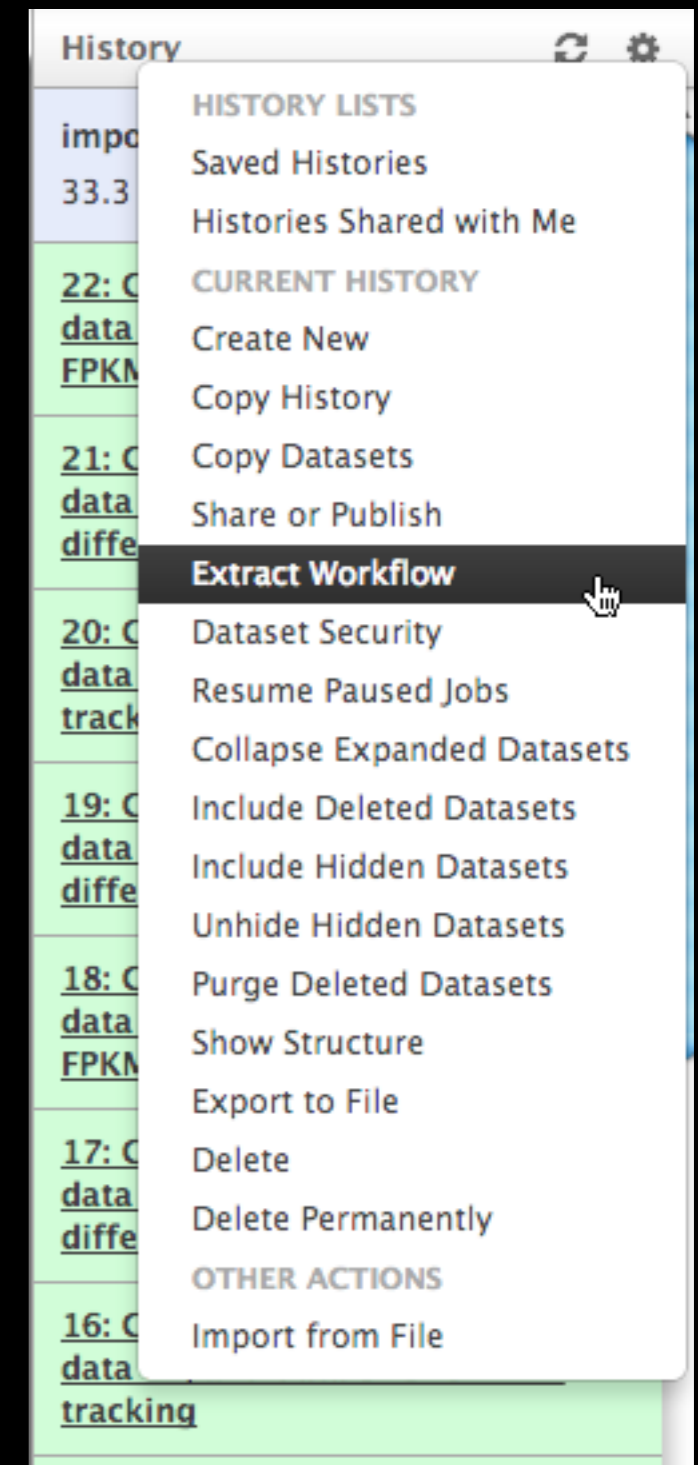
On your own:

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

11:00 **Basic Analysis with Galaxy**

12:20 Basic Analysis into Reusable
Workflows

12:40 Lunch

1:20 NGS Quality Control

1:40 Galaxy Project Overview

2:00 Finish



Agenda

11:00 **Basic Analysis with Galaxy**

12:20 Basic Analysis into Reusable
Workflows

12:40 Lunch

1:20 NGS Quality Control

1:40 Galaxy Project Overview

2:00 Finish

NGS Quality Control

in the context of an RNA-seq Exercise

Start this exercise

<http://bit.ly/gxyrnaseq>

(<http://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise>)

You can continue on your own through the end of today

<http://bit.ly/ctorange>

<http://bit.ly/ctwhite>

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
- Visualize it

<http://bit.ly/gxyrnaseq>

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

<http://bit.ly/gxyRNASEX>

RNA-seq Exercise: A Plan

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

<http://bit.ly/gxyRNASEX>

RNA-seq Exercise: A Plan

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

<http://bit.ly/gxyRNASEX>

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality: Option 3
 - NGS QC and Manipulation → **Fastqc**
 - Gives you a lot **more information but little control** over how it is calculated or presented.

<http://bit.ly/gxyRNASEX>

RNA-seq Exercise: Trim as we see fit.

“For the love of all that is holy, please trim your reads!”

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

- Option 1
 - **NGS QC and Manipulation** → **FASTQ Trimmer by column**
 - Trim same number of columns from every record
 - Can specify different trim for 5' and 3' ends

<http://bit.ly/gxyrnaseq>

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/gxyrnaseq>

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/gxyrnaseq>

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!

<http://bit.ly/gxyrnaseq>

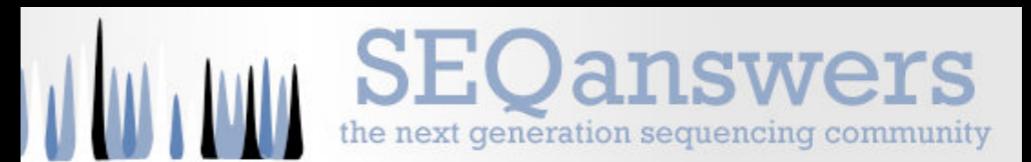
Trim? *As we see fit?*

- 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>



<http://bit.ly/gxyrnaseq>

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 this is where we suspend this exercise.

<http://bit.ly/gxyrnaseq>

Agenda

11:00 **Basic Analysis with Galaxy**

12:20 Basic Analysis into Reusable
Workflows

12:40 Lunch

1:20 NGS Quality Control

1:40 Galaxy Project Overview

2:00 Finish

What is Galaxy?

- An open, web-based platform for **accessible, reproducible,** and **transparent** computational biomedical research.
- Galaxy is available as:
 - **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
 - **Free cloud images** that can be deployed by informatics novices

<http://galaxyproject.org>

A free for everyone web service: usegalaxy.org

The screenshot displays the Galaxy web interface. At the top, the navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. A 'Using 3%' indicator is visible in the top right corner.

The left sidebar contains a 'Tools' section with a search bar and a list of tool categories: 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, and Operate on Genomic Intervals. The 'Operate on Genomic Intervals' section includes a list of actions: Intersect, Subtract, Merge, Concatenate, and Base Coverage.

The main workspace shows the 'Join (version 1.0.0)' tool configuration. The 'Join:' dropdown is set to '1: Exons' (First dataset), and the 'with:' dropdown is set to '2: Repeats' (Second dataset). The 'with min overlap:' field is set to '1' (bp). The 'Return:' dropdown is set to 'Only records that are joined (INNER JOIN)'. An 'Execute' button is located below the configuration fields.

A tip message states: 'TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.'

Below the tip, there are sections for 'Screencasts!' and 'Syntax'. The 'Syntax' section explains that 'Where overlap' specifies the minimum overlap between intervals and that 'Return only records that are joined' returns only the records of the intersection.

The right sidebar shows the 'History' panel, which lists previous jobs. The current job, '13: Join on data 2 and data 1', is highlighted in green. It shows 267 regions in interval format for the hg19 database. Below the job details is a table with the following data:

1. Chrom	2. Start	3. End	4. Name
chr22	17586742	17586844	NM_014331
chr22	17600280	17602017	NM_031891
chr22	17600280	17602017	NM_031891
chr22	17600280	17602017	NM_031891
chr22	17600280	17602017	NM_031891
chr22	17600280	17602017	NM_031891
chr22	17600280	17602017	NM_031891

Other jobs listed in the history include '12: Cluster on data 2', '11: Complement on data 2', '10: Coverage on data 2 and data 1', and '9: Base Coverage on data 2'.

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

Open Source Software: getgalaxy.org

- Galaxy is designed for local installation and customization
 - Easily integrate new tools
 - Requires a computational resource on which to be deployed

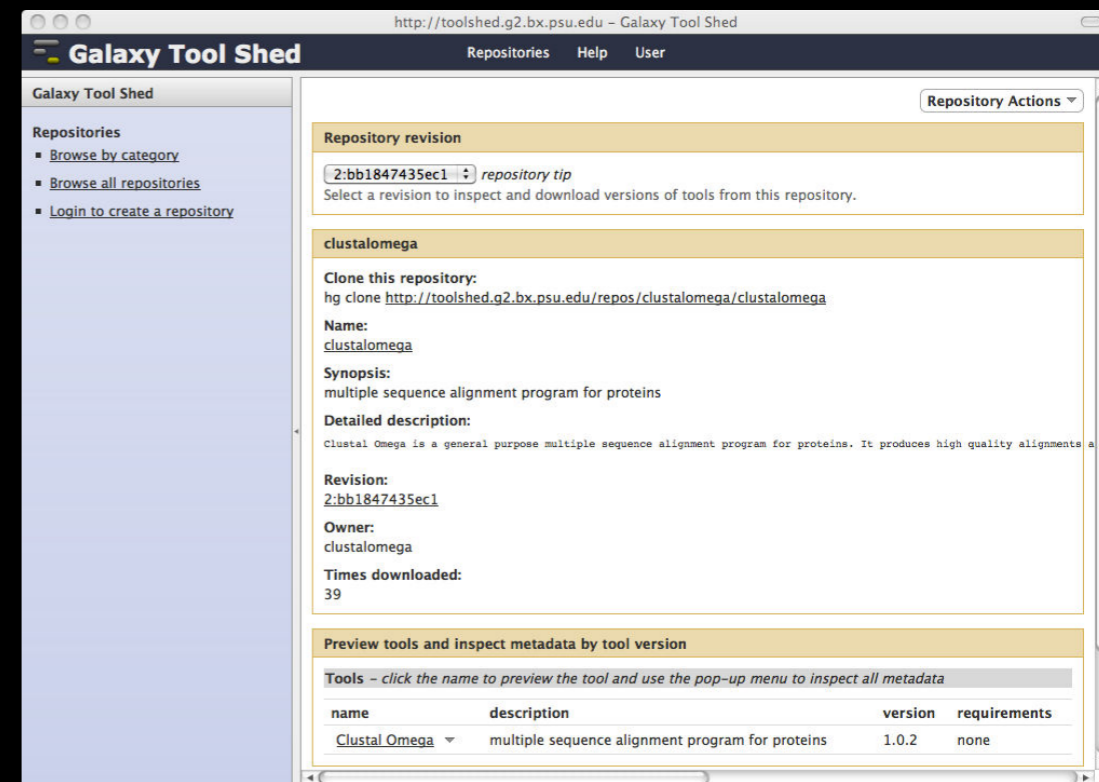
<http://getgalaxy.org>

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**

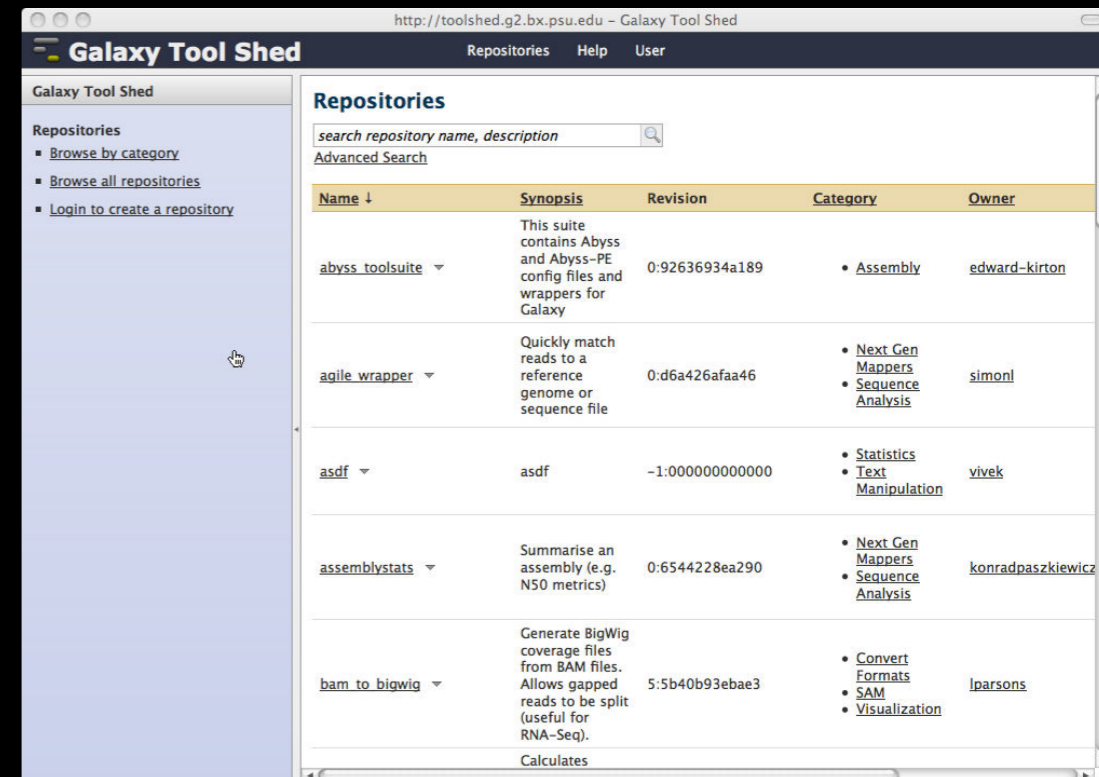
Galaxy Tool Shed

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



The screenshot shows the Galaxy Tool Shed interface for a specific repository revision. The browser address bar indicates the URL is <http://toolshed.g2.bx.psu.edu>. The page title is "Galaxy Tool Shed". The main content area displays the "Repository revision" section with a dropdown menu showing the revision ID "2:bb1847435ec1". Below this, the repository name "clustalomega" is shown. The "Clone this repository" section provides the hg clone command: `hg clone http://toolshed.g2.bx.psu.edu/repos/clustalomega/clustalomega`. The "Name" is "clustalomega", the "Synopsis" is "multiple sequence alignment program for proteins", and the "Detailed description" states: "Clustal Omega is a general purpose multiple sequence alignment program for proteins. It produces high quality alignments". Other metadata includes "Revision: 2:bb1847435ec1", "Owner: clustalomega", and "Times downloaded: 39". A table below shows the tool version details:

name	description	version	requirements
Clustal Omega	multiple sequence alignment program for proteins	1.0.2	none



The screenshot shows the Galaxy Tool Shed interface displaying a list of repositories. The browser address bar indicates the URL is <http://toolshed.g2.bx.psu.edu>. The page title is "Galaxy Tool Shed". The main content area displays the "Repositories" section with a search bar and a table of repository details:

Name ↓	Synopsis	Revision	Category	Owner
abyss_toolsuite	This suite contains Abyss and Abyss-PE config files and wrappers for Galaxy	0:92636934a189	• Assembly	edward-kirton
agile_wrapper	Quickly match reads to a reference genome or sequence file	0:d6a426faa46	• Next Gen Mappers • Sequence Analysis	simonl
asdf	asdf	-1:000000000000	• Statistics • Text Manipulation	vivek
assemblystats	Summarise an assembly (e.g. N50 metrics)	0:6544228ea290	• Next Gen Mappers • Sequence Analysis	konradpaszkiewicz
bam_to_bigwig	Generate BigWig coverage files from BAM files. Allows gapped reads to be split (useful for RNA-Seq). Calculates	5:5b40b93ebae3	• Convert Formats • SAM • Visualization	lparsons

Encourage **Public Galaxy** Instances

<http://wiki.galaxyproject.org/PublicGalaxyServers>

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?

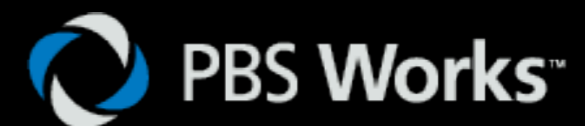
✓ Andromeda

Requires a computational resource on which to be deployed

Control **where** tool execution happens

Galaxy **works with any DRMAA** compliant cluster job scheduler (which is most of them).

Galaxy is **just another client** to your scheduler.



Galaxy is available *on the cloud*

- Using this today
- 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



<http://usegalaxy.org/cloud>

<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 Meeting

<http://wiki.galaxyproject.org>

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>

Galaxy Web Search

Google™ Custom Search x

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](#)

Galaxy Web Search

chip-seq

All Tools Email Source code Shared Documentation Abstracts Requests

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


Community can create, vote and comment on issues

The screenshot shows a Trello board for 'Galaxy: Development Inbox'. The board is organized into four columns: 'Inbox', 'Developer ideas', 'Bug Reports', and 'Issues from Bitbucket'. Each column contains several cards representing different issues or feature requests. The cards include titles, descriptions, and interactive elements like votes and comments. For example, in the 'Developer ideas' column, there is a card titled 'Anonymous use of workflows/visualizations' with 0/2 votes. In the 'Bug Reports' column, there is a card titled 'Workflow View Broken in Toolshed?' with 1 vote. In the 'Issues from Bitbucket' column, there is a card titled '5: Option to disable automatic history creation' with 2 votes. The right side of the board shows a 'Members' section with a grid of member avatars, an 'Add Members...' button, a 'Board' section with 'Options', 'Add List', and 'Filter Cards' buttons, and an 'Activity' section showing recent actions by members like 'Dannon Baker' and 'g2roboto'.

<http://bit.ly/gxytrello>

Galaxy Wiki DaveClements Settings Logout | Search: Titles Text

FrontPage Edit History Actions




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](#).



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](#)




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!](#)




Topic voting now open!



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

News

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, Canada

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](#)

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.

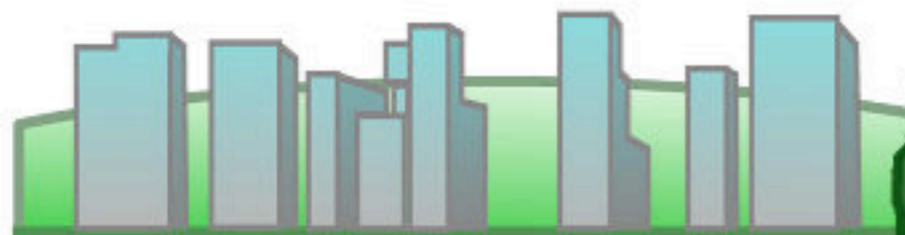
News Items

[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](#)



Galaxy Community Conference



OSLO



UiO : University of Oslo

STARTING
AT

€95

Registration & abstract
submission are now open

<http://galaxyproject.org/GCC2013>

GCC2013
Training
Day





Enis Afgan



Guru Ananda



Dannon Baker



Dan Blankenberg



Dave Bouvier



Dave Clements



Nate Coraor



Carl Eberhard



Jeremy Goecks



Jen Jackson



Greg von Kuster



Ross Lazarus



Anton Nekrutenko



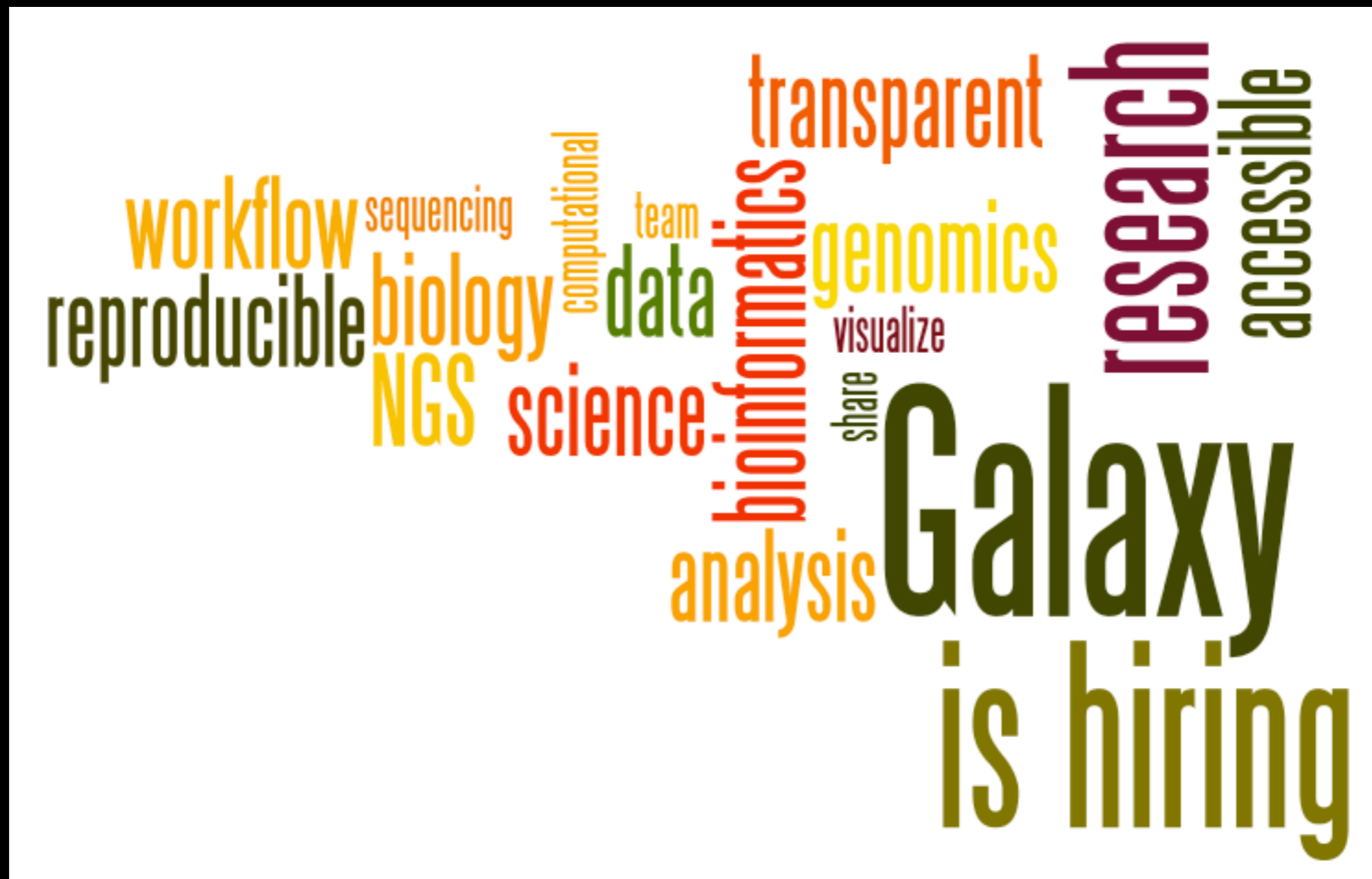
James Taylor



The Galaxy Team

<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>

Acknowledgements

Paul Sternberg

Mary Alvarez

Daniella Raciti

James Vincent

Nalini Raghavachari

You

Caltech
Salzberg Lab

ABRF

AWS Education Grant

NIH NSF Huck Institute

Penn State University Emory University

Feedback Please!

<http://bit.ly/Caltech2013>

Thanks



Dave Clements

**Galaxy Project
Emory University**

clements@galaxyproject.org

<http://bit.ly/Caltech2013>

RNA-seq Exercise: A Plan, *continued*

- 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.*

<http://bit.ly/gxyrnaseq>

RNA-seq Exercise: A Plan, *continued*

- ...
- 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/gxyrnaseq>

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/gxyrnaseq>

Visualizing Genomics

Supported external browsers

- UCSC
- Ensembl
- GBrowse
- IGB
- IGV

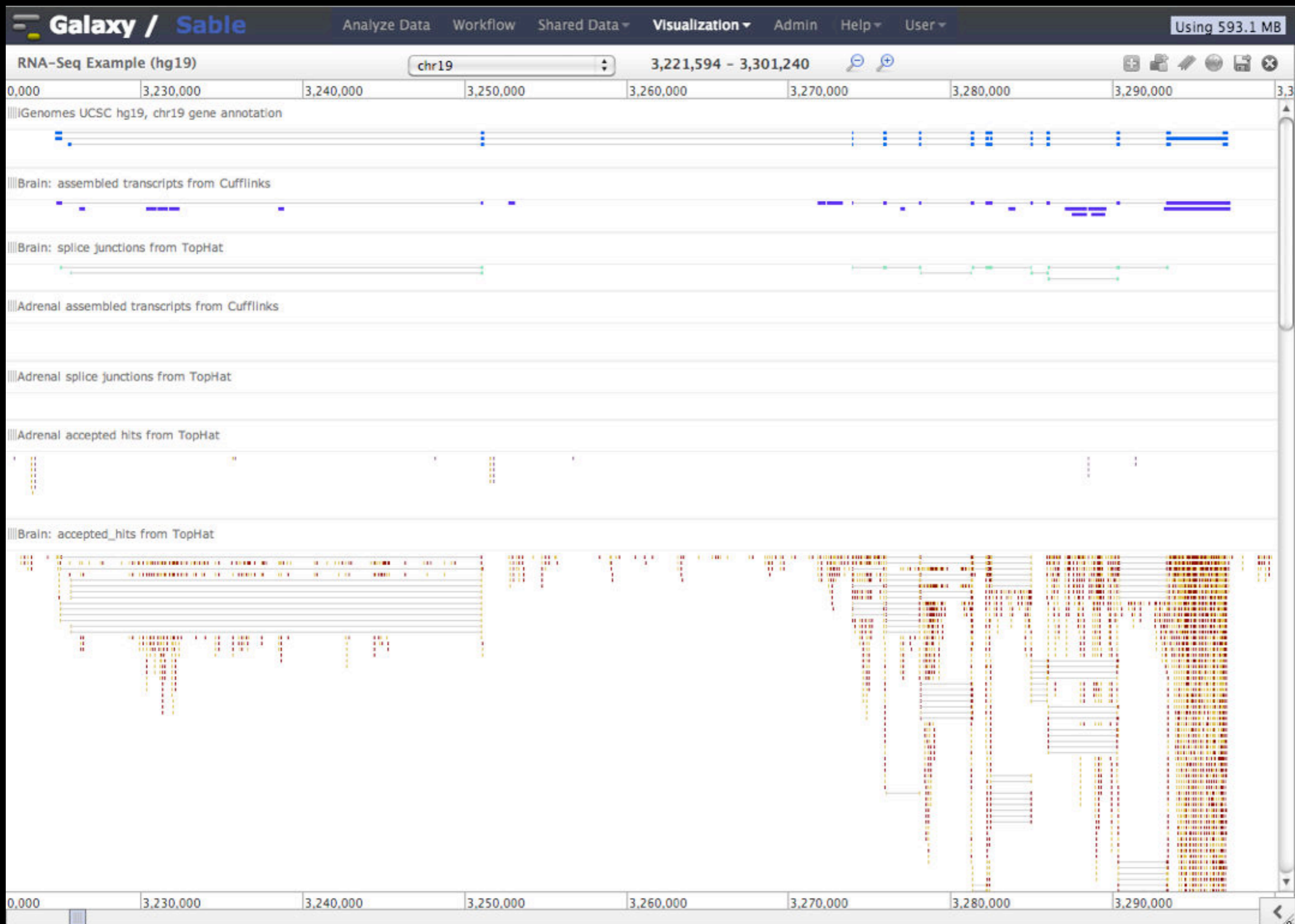
Traditional browser strengths:

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

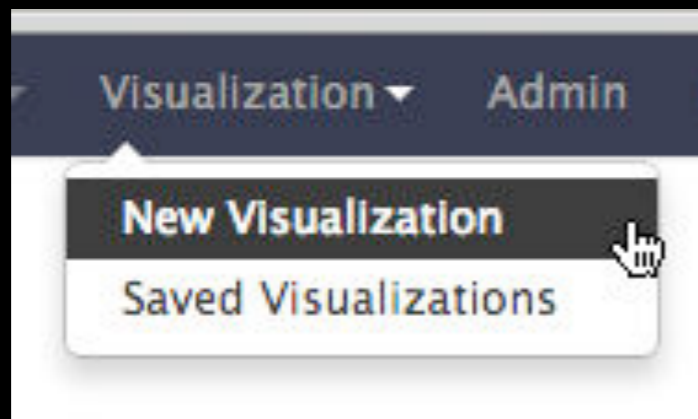
But, *wouldn't it be nice to*

- 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?**

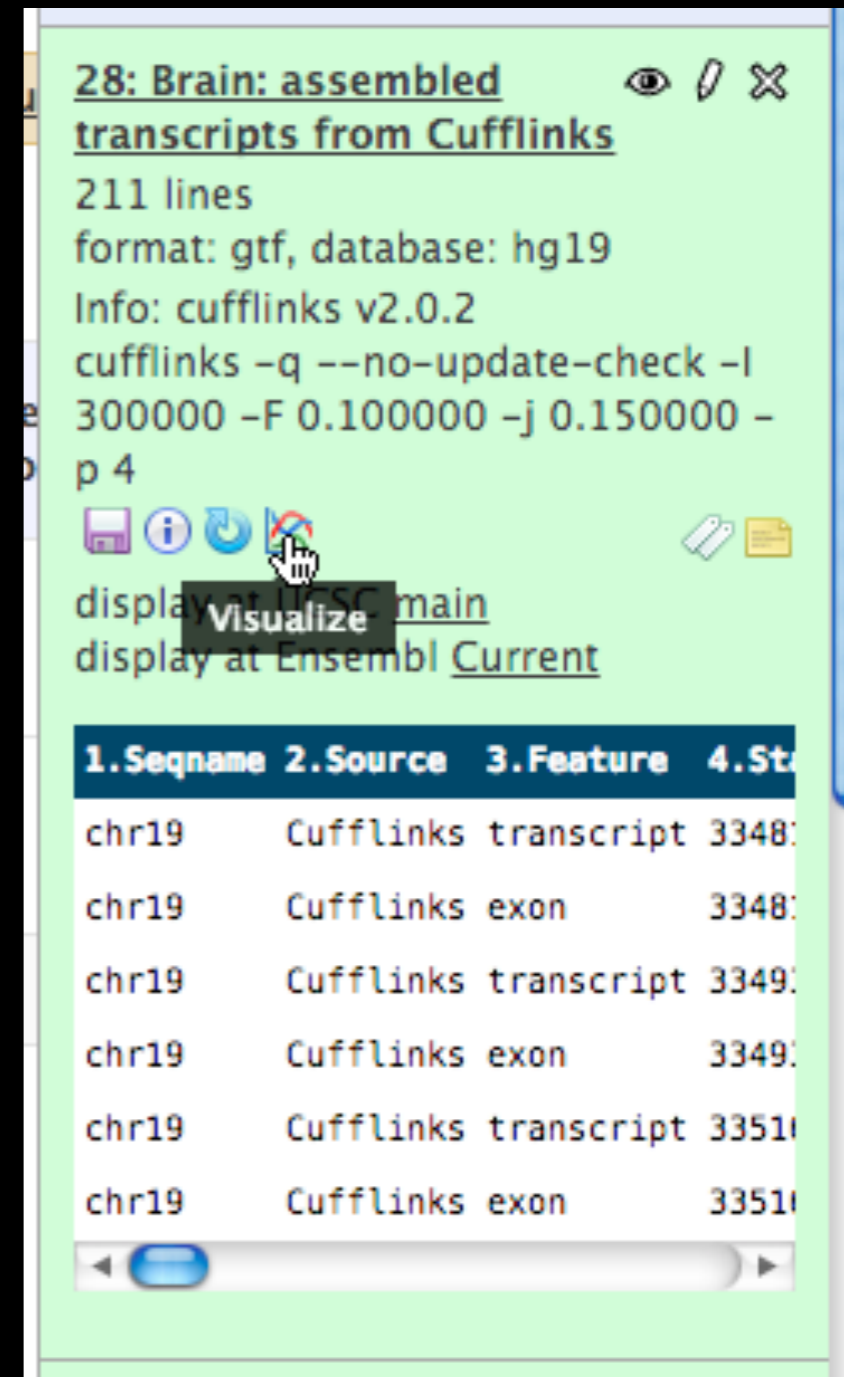
Trackster: Galaxy's embedded track browser



Create a visualization in Galaxy



or



A screenshot of a Galaxy visualization panel. The title is '28: Brain: assembled transcripts from Cufflinks'. Below the title, it says '211 lines', 'format: gtf, database: hg19', and 'Info: cufflinks v2.0.2'. There is a command line: 'cufflinks -q --no-update-check -l 300000 -F 0.100000 -j 0.150000 -p 4'. Below the command line, there are several icons: a folder, an information icon, a refresh icon, a visualize icon (which is highlighted with a mouse cursor), and a tag icon. Below the icons, there are two links: 'display at [HGSC main](#)' and 'display at [Ensembl Current](#)'. At the bottom, there is a table with four columns: '1. Seqname', '2. Source', '3. Feature', and '4. Start'. The table contains several rows of data for chromosome 19, showing Cufflinks transcripts and exons.

1. Seqname	2. Source	3. Feature	4. Start
chr19	Cufflinks	transcript	33480
chr19	Cufflinks	exon	33480
chr19	Cufflinks	transcript	33490
chr19	Cufflinks	exon	33490
chr19	Cufflinks	transcript	33510
chr19	Cufflinks	exon	33510

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?

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

Let's all share...

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**



GENOME
RESEARCH

EXPRESSION



ANALYSIS

illumina

Apply today for the
Cancer GWAS Grant.

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP

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

Search for Keyword:

Go

Advanced Search

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond^{1,2,6,9}, Samir Wadhawan^{3,6,7},

Francesca Chiaromonte⁴, Guruprasad Ananda^{1,3}, Wen-Yu Chung^{1,3,8},

James T

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>.]

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

Current Issue

October 2010, 20 (10)



<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