Accessible, Transparent and Reproducible Analysis with Galaxy

Application of Next Generation Sequencing Technologies for Whole Transcriptome and Genome Analysis

ABRF 2013
Saturday, March 2, 2013
Palm Springs, California, United States

Dave Clements
Emory University

This Workshop

Demonstrate Galaxy with a hands-on walk through of an example RNA-Seq analysis

Introduce Galaxy and Galaxy Project as we go.

Complements talk on Monday:

Galaxy for Core Facilities

(W6) Community Resource Solutions to Analyzing Large Genomic Data Sets

Slides are at

bit.ly/ABRFgxyWS1 wiki.galaxyproject.org/Events

Demonstrate Galaxy with a hands-on walk through of an RNA-Seq analysis

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

http://bit.ly/gxyrnaseq

http://bit.ly/ABRFgxy1
http://bit.ly/ABRFgxy2
http://bit.ly/ABRFgxy3

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

bit.ly/gxyrnaseq bit.ly/ABRFgxy[123]

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 → Data Libraries**

bit.ly/gxyrnaseq bit.ly/ABRFgxy[123]

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!

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ................................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ...........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./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

- 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/gxyrnaseq

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/gxyrnaseq

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/gxyrnaseq

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 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

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.*

http://bit.ly/gxyrnaseq

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

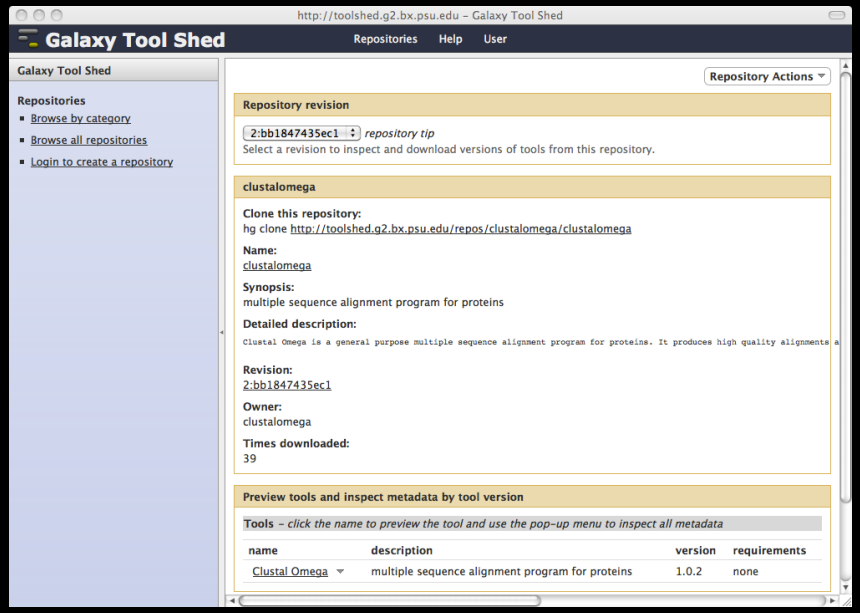
http://getgalaxy.org

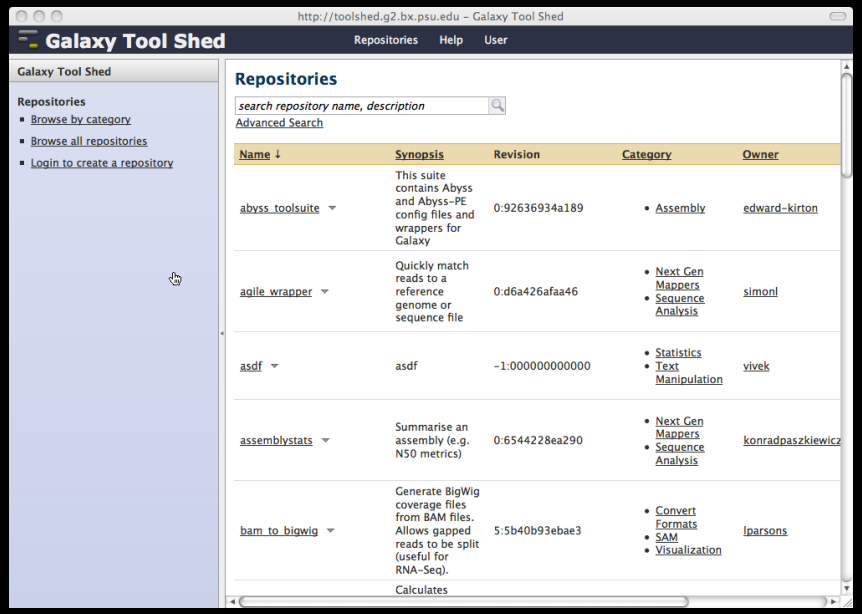
Encourage Local Galaxy Instances

- 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://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

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?

- 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.

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

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

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

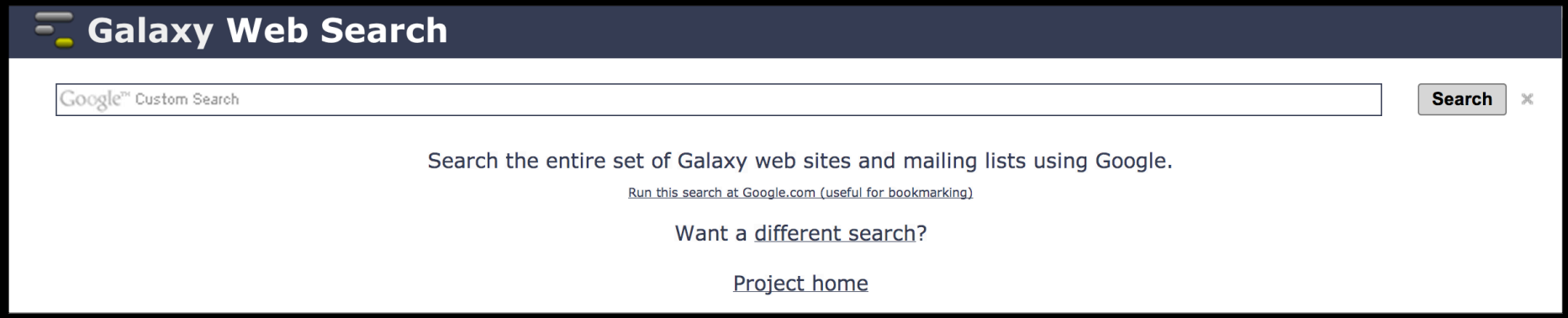
High volume (2900 posts, 2700 members in 2012)

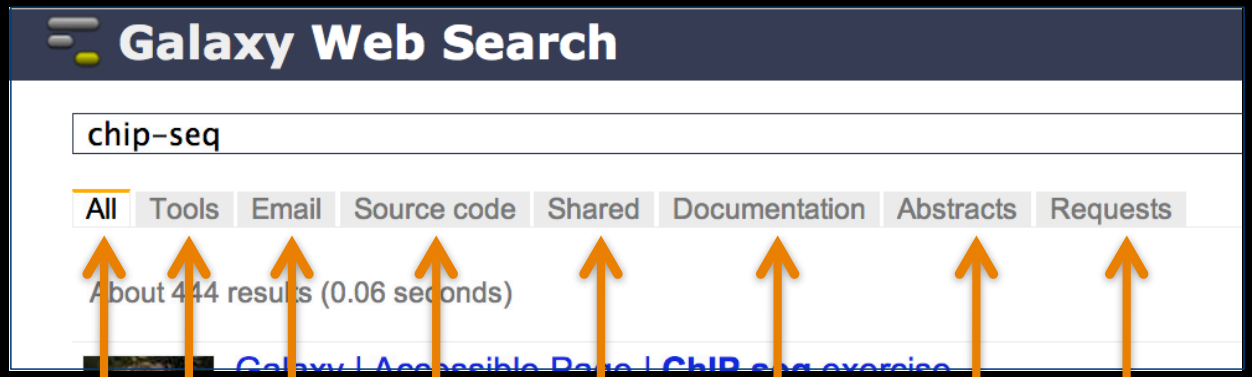
Galaxy-Dev

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

Trello | Help | Notifications | Boards

Galaxy: Development Inbox — Galaxy Project — Public

Inbox

- To add cards, use the http://galaxyproject.org/trello (2 votes, 1)
- Filter and Sort: "Select" tool not dealing with special characters right (1)
- Uploaded fastq file datatype not usable in BWA (1)
- Reference genome request: GATK-ordered hg19 (1)
- Feature request: manually hide datasets (1)

Add a card...

Developer ideas

- Anonymous use of workflows/visualizations (0/2)
- Feature Request: the ability to restart a failed workflow from the point of failure; (6 votes, 2)
- Google Drive / Dropbox / Box / ... integration (1 vote)
- Bug report: always import deleted datasets (2)
- Standalone web application(s) for visualizations
- Enh: Archiving histories (1)
- Modify data library upload completion message (1)
- Display in UI runtime

Add a card...

Bug Reports

- Issues with workflow step hiding not persisting (1 vote, 1)
- Workflow View Broken in Toolshed?
- Unable to run jobs when user job limits are set (1 vote, 4)
- Fix tool tip FASTQ Summary Statistics (1)
- Bug when using data_column
- Velvet wrapper broken when real user jobs are used
- apport.fileutils
- Bug: Running functional tests for migrated or installed tools does not

Add a card...

Issues from Bitbucket

- 5: Option to disable automatic history creation (2 votes, 1)
- 6: Option to require that histories have names (1 vote)
- 8: More flexible output handlers
- 10: Allow overriding parameters when running a workflow (1 vote)
- 20: Suggestion: new tag in tool's XML file - 12/9/08 email from Assaf Gordon
- 21: Real DB key build ontology
- 24: Add ability to password secure tools

Add a card...

Members

CE DB G

Add Members...

Board

Options

Add List

Filter Cards

Activity View all...

Dannon Baker added **API: Library Contents** to Developer ideas and
- sent to the board
- joined

today at 10:39 am

G **g2roboto** on **Feature request: manually hide datasets**

Submitted by @nickstoler

Feb 1 at 4:40 pm

G **g2roboto** added **Feature request: manually hide datasets** to Inbox.

Feb 1 at 4:40 pm

G **g2roboto** on **Reference**

http://bit.ly/gxytrello

http://wiki.galaxyproject.org

Galaxy Wiki

DaveClements Settings Logout | Search: Titles Text

FrontPage

Edit History Actions

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@glaxyproject.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!

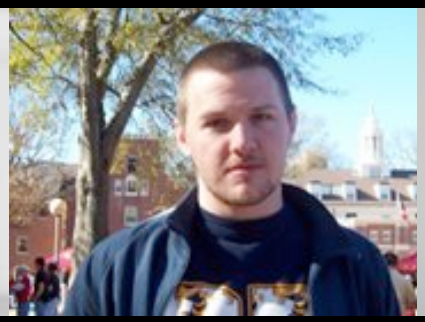
A list of possible topics for the GCC2013 Training Day is now available. Please take a few minutes to review these possibilities and then vote for your favorite three topics.*

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.

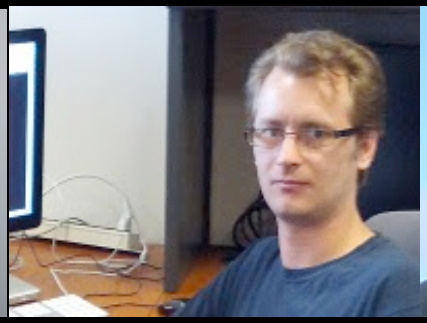
Registration & abstract submission are now open

http://galaxyproject.org/GCC2013


Enis Afgan


Dannon Baker


Dan Blankenberg


Dave Bouvier


Dave Clements


Nate Coraor


Carl Eberhard


Jeremy Goecks


Sam Guerler


Jen Jackson


Greg von Kuster


Ross Lazarus


Anton Nekrutenko


James Taylor

You

The Galaxy Team

http://wiki.galaxyproject.org/GalaxyTeam

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/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

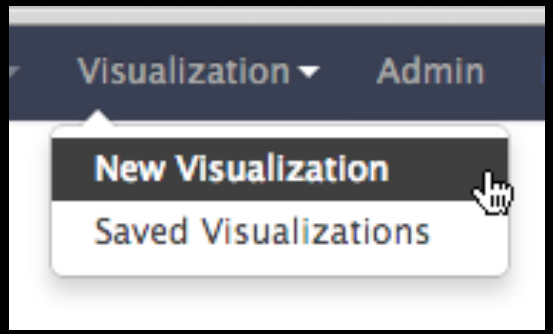
Traditional browser strengths:

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

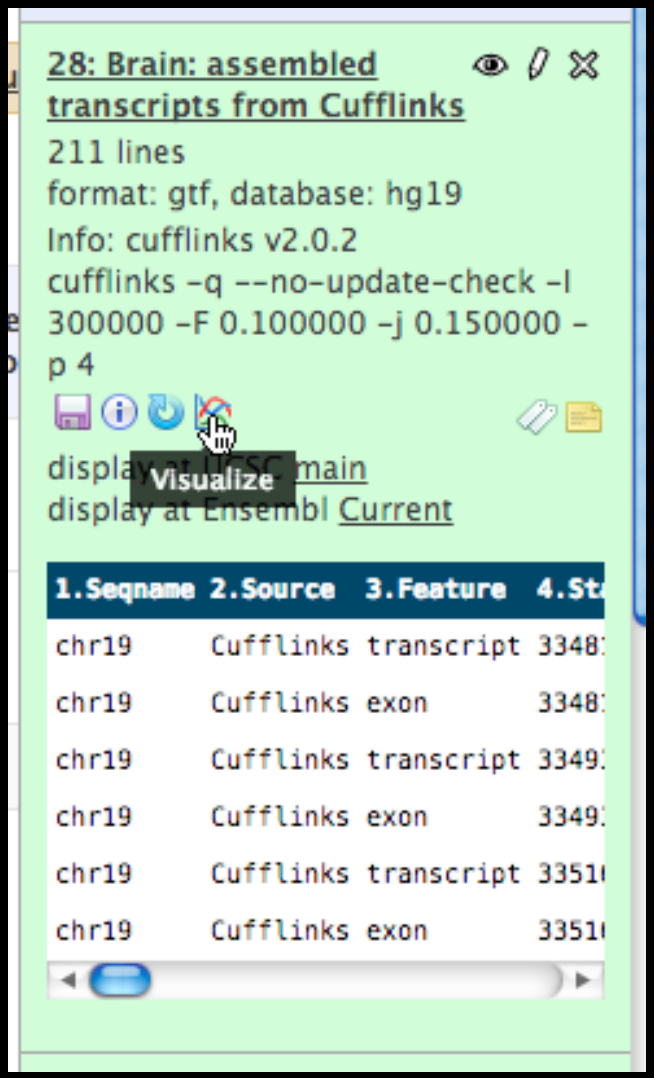
But, *wouldn't it be nice to*

- 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?

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?

Acknowledgements

Nalini Raghavachari
David Needleman
Jim Vincent

The Galaxy Team
especially
Dannon Baker

ABRF

AWS Education Grant

NIH NSF Huck Institute
Penn State University Emory University

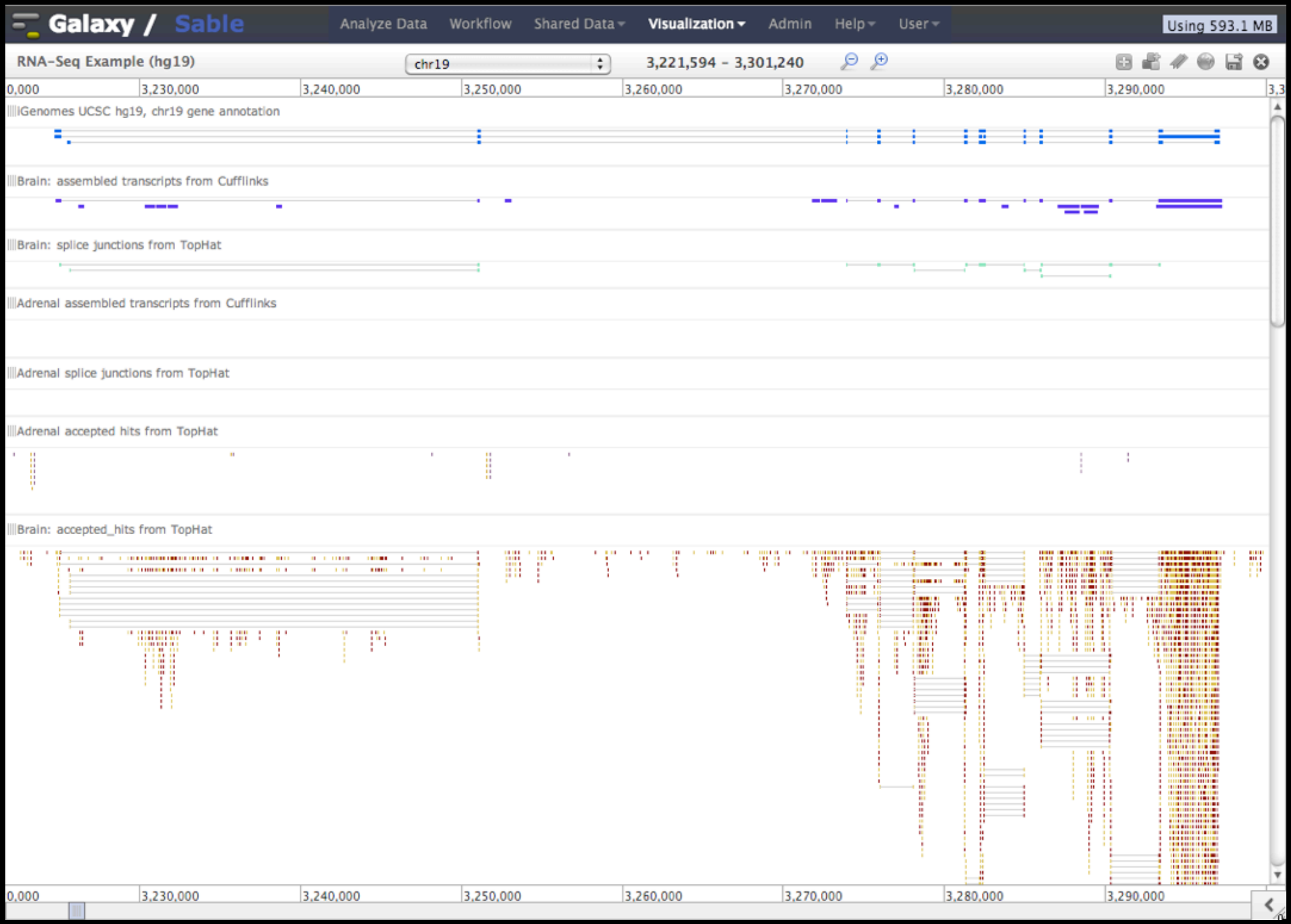
Thanks

Dave Clements

Galaxy Project
Emory University

clements@galaxyproject.org

Trackster: Galaxy's embedded track browser



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

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 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

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

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 T

OPEN ACCESS ARTICLE

This Article

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

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.]

http://usegalaxy.org/u/aun1/p/windshield-splatter