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/ABRFgxyy1


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

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

What is FASTQ?

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

```
@SEQ_ID
GATTTGGGGTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((****++))%%%++)%62 values (0, 62) (0 to 40 expected in raw reads)
```

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

```
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.

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

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.

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

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.

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

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.

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.
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 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](http://galaxyproject.org/search)

**Galaxy Web Search**

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

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

Want a different search?

Project home

**Find**

- Everything on …
- Tools for …
- Email about …
- Source code for …
- Published Histories, Pages, Workflows, about …
- Related feature requests
- Papers using Galaxy for …
- Documentation on …
Community can create, vote and comment on issues

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

**Usegalaxy.org**

**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!
**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 edit it here or send it to outreach@broadinstitute.org.

### Upcoming Events

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic/Event</th>
<th>Venue/Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>February 4</td>
<td>Introduction to Galaxy Boot Camp</td>
<td>UC Davis Biostatistics Core, Davis, California, United States</td>
</tr>
<tr>
<td>March 1- 5</td>
<td>Accessible, Transparent and Reproducible Analysis With Galaxy, part of ABRF 2013</td>
<td>ABRF 2013, Palm Springs, California, United States</td>
</tr>
<tr>
<td>March 26-28</td>
<td>RNA Technologies and Analysis Workshop</td>
<td>DOE JGI User Meeting</td>
</tr>
<tr>
<td>April 5-6</td>
<td>2013 GDOO Meeting</td>
<td>Cambridge, United Kingdom</td>
</tr>
<tr>
<td>April 7-10</td>
<td>GO Galaxy Workshop</td>
<td>Bocca 2013, Cambridge, United Kingdom</td>
</tr>
<tr>
<td>April 9-11</td>
<td>Workshop: Integrated Research Data Management For Next Gen Sequencing Using Galaxy and Cloud Online Software-as-a-Service</td>
<td>BiCIT, Boston, Massachusetts, United States</td>
</tr>
<tr>
<td>May 14-16</td>
<td>Tutorial: Exploring and Enabling Biomedical Data Analysis with Galaxy</td>
<td>Great Lakes Bioinformatics Conference (GLBio) 2013, Pittsburgh, Pennsylvania, United States</td>
</tr>
<tr>
<td>May 21-29</td>
<td>Initiation à l'utilisation de Galaxy</td>
<td>Cycle &quot;Bioinformatique par la pratique&quot; 2013, INRA Jouy-en-Josas, France</td>
</tr>
<tr>
<td>June 6-7</td>
<td>Informatics or High Throughput Sequencing Data Workshop</td>
<td>Toronto, Ontario, Canada</td>
</tr>
</tbody>
</table>

---

**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. Old 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 voting, Registration, and Sponsorships
- January GalaxyAdmins Web Meeting dates and screencast
- Other Upcoming Events and Deadlines
- Galaxy Distributions
- Toolshed 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 @ SML
Nominate GCC2013 Training Topics
Dec 3, 2012 Distribution & News Brief
December 2012 Galaxy Update
Nov 14, 2012 Distribution & News Brief
NGS Analysis using Vis. with Vizier
November 2012 GalaxyAdmins
Registration & abstract submission are now open

http://galaxyproject.org/GCC2013
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.

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

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
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.
Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovskyy Pond1,2,6,9, Samir Wadhawan3,6,7,
Francesca Chiaramonte4, Guruprasad Ananda1,3, Wen-Yu Chung1,3,8,
James Thomas9

[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.]