Transparent, accessible, reproducible analysis with Galaxy

Indiana University
19 October 2012

Dave Clements
Emory University

http://galaxyproject.org/
Acknowledgements

Richard LeDuc
William Barnett
Scott Michaels
Radhika Khetani

National Center for Genome Analysis Support (NCGAS)
The Center for Genomics and Bioinformatics
Indiana University
The Galaxy Team

http://galaxyproject.org/wiki/GalaxyTeam
As science becomes increasingly dependent on computation:

● How best to ensure that analysis are **reproducible**?

● How can methods best be made **accessible** to scientists?

● How to facilitate **transparent** communication of analyses?
A crisis in genomics research: **reproducibility**
Key Reproducibility Problems

- **Datasets**: not all available, difficult to access
- **Tools**: inaccessible, hard to record details
- **Publication**: results, data, methods separate
Microarray Experiment Reproducibility

- 18 Nat. Genetics microarray gene expression experiments
- Less than 50% reproducible

Problems
- missing data (38%)
- missing software, hardware details (50%)
- missing method, processing details (66%)

50 papers citing bwa

31 provide no version and no settings

8 lists versions

4 list settings

7 lists versions and settings

26 do not provide access to data

Nekrutenko & Taylor, "Next-generation sequencing data interpretation: enhancing reproducibility and accessibility" Nature Reviews Genetics 13, 667-672 (September 2012) doi:10.1038/nrg3305
Galaxy: accessible analysis system
Integrating existing tools into a uniform framework

- Defined in terms of an abstract interface (inputs and outputs)
- In practice, mostly command line tools, a declarative XML description of the interface, how to generate a command line
- Designed to be as easy as possible for tool authors, while still allowing rigorous reasoning
Galaxy analysis interface

- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses
- Exact parameters of a step can always be inspected, and easily rerun
Automatically tracks every step of every analysis

<table>
<thead>
<tr>
<th>QNAME</th>
<th>FLAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>HWI-EAS269:3:1:1449:913</td>
<td>99</td>
</tr>
<tr>
<td>HWI-EAS269:3:1:1449:913</td>
<td>147</td>
</tr>
<tr>
<td>HWI-EAS269:3:1:709:832</td>
<td>99</td>
</tr>
<tr>
<td>HWI-EAS269:3:1:709:832</td>
<td>147</td>
</tr>
<tr>
<td>HWI-EAS269:3:1:1422:1087</td>
<td>99</td>
</tr>
<tr>
<td>HWI-EAS269:3:1:1422:1087</td>
<td>147</td>
</tr>
</tbody>
</table>
As well as user-generated metadata and annotation...
Galaxy workflow system

- **Workflows** can be constructed from scratch or extracted from existing analysis histories.
- Facilitate reuse, as well as providing precise reproducibility of a complex analysis.
Tranparency: Sharing and publishing

- All analysis components (datasets, histories, workflows) can be shared among Galaxy users and published.
- Annotation and Galaxy Pages allow analyses to be augmented with textual content and provided in the form of an integrated.
Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

Sergei Kosakovskiy Pond1,2,6,9, Samir Wadhawan3,6,7, Francesca Chiaramonte4, Guruprasad Ananda1,3, Wen-Yu Chung1,3,8, James Taylor1,5,9, Anton Nekrutenko1,3,9 and The Galaxy Team1

This Article
Published in Advance October 3, 2009, doi: 10.1101/gr.894506.109
Copyright © 2009 by Cold Spring Harbor Laboratory Press

OPEN ACCESS ARTICLE

Comparison of Galaxy vs. MEGAN pipeline.

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.]
Give it a spin: usegalaxy.org/galaxy101

Galaxy 101, is a hands-on exercise that demonstrates many Galaxy basics.

Galaxy 101 includes histories, datasets, and workflows, and is itself a Galaxy Page.
Galaxy is available ...

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

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

- Free public web site
- Anybody can use it
- Persistent
- 24,000 registered users
- 300+ TB of user data
- 140,000+ jobs / month
- Hundreds of tools

For example, the first 5 pages of NGS tools
But, it’s a big world

Main has lots of tools, storage, processor, users, ...

- But **not all tools** - there are thousands and adding new tools is not taken lightly

- But **not infinite storage and processors** - Main now has job limits and storage quotas

A centralized solution cannot scale to meet data analysis demands of the whole world
Galaxy is available ...

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

- As open source software that makes integrating your own tools and data and customizing for your own site simple

http://getgalaxy.org
Local Galaxy Instances

- Galaxy is designed for local installation and customization
- Easily integrate new tools
- Easy to deploy and manage on nearly any (unix) system
- Run jobs on existing compute clusters
- Requires an existing computational resource on which to be deployed

http://getgalaxy.org
Encourage Local Galaxy Instances

- Support *increasingly decentralized model* and *improve access to existing resources*
- Focus on building *infrastructure to enable the community to integrate and share* tools, workflows, and best practices
Galaxy Tool Shed

- Allow sites to share “suites” containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies

- Integration with Galaxy instances to automate tool installation and updates

[toolshed.g2.bx.psu.edu]
Public Galaxy Servers

http://galaxyproject.org/wiki/PublicGalaxyServers

Interested in:

- ChIP-chip and ChIP-seq?
  - ✓ Cistrome
- Statistical Analysis?
  - ✓ Genomic Hyperbrowser
- Sequence and tiling arrays?
  - ✓ Oqtans
- Text Mining?
  - ✓ DBCLS Galaxy
- Reasoning with ontologies?
  - ✓ GO Galaxy
- Internally symmetric protein structures?
  - ✓ SymD
Local Galaxy Instances

- Galaxy is designed for local installation and customization
- Easily integrate new tools
- Easy to deploy and manage on nearly any (unix) system
- Run jobs on existing compute clusters
- Requires an existing computational resource on which to be deployed

http://getgalaxy.org
Got your own cluster?

- Move tool execution to other systems
- Galaxy works with DRMAA compliant cluster job schedulers (which is most of them).
- Galaxy is just another client to your scheduler.
Galaxy is available ...

- As a free (for everyone) web service integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- As open source software that makes integrating your own tools and data and customizing for your own site simple
- On the Cloud

http://usegalaxy.org/cloud
Galaxy CloudMan
http://usegalaxy.org/cloud

• Start with a fully configured and populated (tools and data) Galaxy instance.
• Allows you to scale up and down your compute assets as needed.
• Someone else manages the data center.
• We’ll use Amazon for the Galaxy for Biologists workshop later today.

http://aws.amazon.com/education
Step by Step Instructions on the Wiki for Amazon

Galaxy Wiki

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.

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 an AWS Education Grant after you register).
2. Once your account has been approved by Amazon (note that this may take up to
Launch a CloudMan instance directly from Main, and transfer your current history.
Galaxy Community & Resources

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

http://galaxyproject.org/wiki
Mailing Lists

Galaxy-User messages / month

Galaxy-Dev messages / month

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

**Find**
- Everything on ...
- Tools for ...
- Email about ...
- Source code for ...
- Published Histories, Pages, Workflows, about ...
- Related feature requests
- Papers using Galaxy for ...
- Documentation on ...
## Other Upcoming Galaxy Events

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic/Event</th>
<th>Venue/Location</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>October 15-17</td>
<td>Advanced NGS Course: RNA-seq data analysis</td>
<td>Amsterdam Medical Centre (AMC), The Netherlands</td>
<td>Patrick Koks</td>
</tr>
<tr>
<td>October 18-30</td>
<td>Advanced Sequencing Technologies and Applications Course</td>
<td>Cold Spring Harbor Laboratory, New York, United States</td>
<td>Anton Nekrutenko</td>
</tr>
<tr>
<td>October 31 - November 6</td>
<td>Computational &amp; Comparative Genomics Course</td>
<td>Cold Spring Harbor Laboratory, New York, United States</td>
<td>William Pearson, James Taylor</td>
</tr>
<tr>
<td>October 28 - November 2</td>
<td>Genomic Virtual Laboratory Workshop</td>
<td>eResearch Australasia, Sydney, Australia</td>
<td>Enis Afgan</td>
</tr>
<tr>
<td>November 6-10</td>
<td>Galaxy 101: Data Integration, Analysis and Sharing</td>
<td>American Society of Human Genetics (ASHG), San Francisco, California, United States</td>
<td>Jennifer Jackson, Jeremy Goecks</td>
</tr>
<tr>
<td></td>
<td>Working with High-Throughput Data and Data Visualization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>November 12-14</td>
<td>The Genome Access Course</td>
<td>Cold Spring Harbor Laboratory, New York, United States</td>
<td>Assaf Gordon</td>
</tr>
<tr>
<td>November 13-15</td>
<td>Analyse des données RNA-seq et ChIP-seq (séquençage haut-débit), à l’aide d’outils orientés vers un public de biologistes</td>
<td>PRABI (Pôle Rhône-Alpes de Bioinformatique), Doua de l’Université Claude Bernard - Lyon, Lyon, France</td>
<td>Guy Perrière</td>
</tr>
<tr>
<td>January 14-18</td>
<td>Plant and Animal Genome (PAG 2013)</td>
<td>San Diego, California, United States</td>
<td>Dave Clements</td>
</tr>
<tr>
<td>March 2-5</td>
<td>W6: Community Resource Solutions to Analyzing</td>
<td>ABRF 2013</td>
<td>Dave Clements</td>
</tr>
</tbody>
</table>

[http://galaxyproject.org/wiki/Events](http://galaxyproject.org/wiki/Events)
Visualization

Send data results to **external** genome browsers:
- UCSC, Ensembl, GBrowse, IGV

**Trackster**: Galaxy’s genome browser
Trackster

View your data from within Galaxy
- No data transfers to external site
- Use it locally, even without internet access

Supports common filetypes
- BAM, BED, GFF/GTF, WIG

Unique features
- custom genomes
- highly interactive
But really, why *another* genome browser

From static browsing to **visual analysis**

**Visual feedback and experimentation** needed for complex tools with many parameters

**Leverage Galaxy strengths:** a very sound model for abstracting interfaces to analysis tools and already integrates an enormous number
Dynamic Filtering
Integrating Tools and Visualization
Exploring Parameter Space with Trackster
Galaxy URLs to Remember

http://galaxyproject.org
http://usegalaxy.org
http://getgalaxy.org
http://usegalaxy.org/galaxy101

and

https://galaxy.indiana.edu/
Galaxy at Indiana University

- Backend
  - Runs on UITS Supercomputers
  - Supported by the National Center for Genome Analysis Support

- Requests for new bioinformatic tools addressed within 48 hours

- Use your IU user name and password
Thank you.