Introduction to Galaxy

University of North Carolina
Chapel Hill
July 19, 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
11:30  RNA-Seq Example Part I
12:30  Lunch
1:30   RNA-Seq Example Part II
2:20   Galaxy Project Overview
2:40   Break
3:00   RNA-Seq Example Part III
3:40   Sharing, Publishing and Reproducibility
4:00   Setting up Galaxy on the Amazon Cloud
4:30   Done
Introductions

In 40 seconds or less tell us

- your name
- your affiliation(s)
- something about your research
- something about your goals for today
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
   - Demonstrate how to set up a **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
11:30 RNA-Seq Example Part I
12:30 Lunch
1:30  RNA-Seq Example Part II
2:20  Galaxy Project Overview
2:40  Break
3:00  RNA-Seq Example Part III
3:40  Sharing, Publishing and Reproducibility
4:00  Setting up Galaxy on the Amazon Cloud
4:30  Done
Basic Analysis: We have

an assembly of an archaeal organism

gene annotation

TF binding sites

Which genes have most overlapping TFBSs?

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

(~ http://usegalaxy.org/galaxy101)
Exons & TFBSs: A General Plan

• Get some data
  • Sequence, genes/exons, TFBSs
• Mess with it
  • Identify which genes/exons have TFBSs
  • Count TFBSs per exon
  • Visualize, save, download, ... exons with most TFBSs

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

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

TFBS

Overlap pairings

Exon overlap counts
Exon overlap counts
### Exon overlap counts

<table>
<thead>
<tr>
<th>Exon</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

### Exons

<table>
<thead>
<tr>
<th>Exon</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

### Join on exon name

<table>
<thead>
<tr>
<th>Exon</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Visualize results

- New Visualization
- Saved Visualizations
Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:30 Lunch
1:30 RNA-Seq Example Part II
2:20 Galaxy Project Overview
2:40 Break
3:00 RNA-Seq Example Part III
3:40 Sharing, Publishing and Reproducibility
4:00 Setting up Galaxy on the Amazon Cloud
4:30 Done
Agenda

9:00  Welcome
9:20  Basic Analysis with Galaxy
10:40  Break
11:00  Basic Analysis into Reusable Workflows
11:30  RNA-Seq Example Part I
12:30  Lunch
1:30  RNA-Seq Example Part II
2:20  Galaxy Project Overview
2:40  Break
3:00  RNA-Seq Example Part III
3:40  Sharing, Publishing and Reproducibility
4:00  Setting up Galaxy on the Amazon Cloud
4:30  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
The analysis we just finished was about:

- An archaea
- Overlap between exons and TFBs

But, ...

- There is nothing inherently in the analysis about archaea, exons or TFBs.
- 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
Did that work?

**On your own:**
Count # of exons in each TFBS
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
11:30 **RNA-Seq Example Part I**
12:30 Lunch
1:30  RNA-Seq Example Part II
2:20  Galaxy Project Overview
2:40  Break
3:00  RNA-Seq Example Part III
3:40  Sharing, Publishing and Reproducibility
4:00  Setting up Galaxy on the Amazon Cloud
4:30  Done
RNA-seq Exercise

Shared Data → Published Pages

→ RNA-Seq Analysis Exercise
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
- Perform differential gene expression analysis with Cuffdiff
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 → RNA-Seq Example
What is **FASTQ**?

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

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

- **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) |
```

RNA-seq Exercise: A Plan

Look at quality Options 1 & 2:

1. NGS QC and Manipulation → Compute Quality Statistics
   NGS QC and Manipulation → 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,
   Statistics in text and graphic formats
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

• 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
  - Can have different thresholds for different regions of the reads.
  - Keeps original read length.
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
Options are not mutually exclusive
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
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://biostars.org/)
  - [http://seqanswers.com/](http://seqanswers.com/)
  - [http://galaxyproject.org/search](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
  • Tophat looks for best place(s) to map reads, and best places to insert introns
  • *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq mapping here.*
Agenda

9:00  Welcome
9:20  Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:30 Lunch
1:30  RNA-Seq Example Part II
2:20  Galaxy Project Overview
2:40  Break
3:00  RNA-Seq Example Part III
3:40  Sharing, Publishing and Reproducibility
4:00  Setting up Galaxy on the Amazon Cloud
4:30  Done
Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:30 Lunch
  1:30 RNA-Seq Example Part II
2:20 Galaxy Project Overview
2:40 Break
3:00 RNA-Seq Example Part III
3:40 Sharing, Publishing and Reproducibility
4:00 Setting up Galaxy on the Amazon Cloud
4:30 Done
RNA-seq Exercise: A Plan

- ... 

- Map the reads to the human reference using Tophat

- Run Cufflinks on Tophat output to assemble reads into transcripts
  
  - Tophat does not make any predictions about how the reads it mapped, assemble together 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
Trackster: Galaxy’s embedded track browser
Create a visualization in Galaxy

or
Vizualization inside Galaxy

- Leverage visualization to *evaluate* and *refine* analyses
- Make the *analyze-visualize-refine* loop seamless and *fast*
- Enable *experimenting* with tools and their parameter *space*
- Support *custom genome browsers*
Agenda

9:00  Welcome
9:20  Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:30 Lunch
1:30  RNA-Seq Example Part II
2:20  Galaxy Project Overview
2:40  Break
3:00  RNA-Seq Example Part III
3:40  Sharing, Publishing and Reproducibility
4:00  Setting up Galaxy on the Amazon Cloud
4:30  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


interested in:

- ChIP-chip and ChIP-sequencing?
  - Cistrome, Nebula
- 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.
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 we will demonstrate how to do this later today

  http://aws.amazon.com/education
Galaxy is available ...

• As a free (for everyone) web service
• As open source software
• On the Cloud

• With Commercial Support
  
  A ready-to-use appliance (BioTeam)
  Cloud-based solutions (Appistry, ABgenomica, AIS)
  Consulting & Customization (Arctix, Deena Bioinformatics)
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
Galaxy Resources and Community: Mailing Lists
http://wiki.galaxyproject.org/MailingLists

**Galaxy-Announce**
Project announcements, low volume, moderated
Low volume (42 posts in 2012, 2100+ members)

**Galaxy-User**
Questions about using Galaxy and usegalaxy.org
High volume (2900 posts in 2012, 2700+ members)

**Galaxy-Dev**
Questions about developing for and deploying Galaxy
High volume (4500 posts in 2012, 900+ members)
Unified Search: http://galaxyproject.org/search
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
- Galaxy Appliance

**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 add it here or send it to Outreach@galaxyproject.org.

**Upcoming Events**

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic/Event</th>
<th>Venue/Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>July 18-23</td>
<td>Introduction to Galaxy Workshop, National Institute of Environmental Health Sciences (NIEHS)</td>
<td>2013 Research Triangle Park, North Carolina, United States</td>
</tr>
<tr>
<td></td>
<td>Introduction to Galaxy Workshop</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of North Carolina, Chapel Hill</td>
<td></td>
</tr>
<tr>
<td>July 19-23</td>
<td>ISMB/ECBB, BOSC and MS SIG 2013</td>
<td>Berlin, Germany</td>
</tr>
<tr>
<td></td>
<td>Talks, posters and workshops. Lots of them!</td>
<td></td>
</tr>
<tr>
<td>July 21-25</td>
<td>Experiences in building a Next-Generation Sequencing Analysis</td>
<td>XSEDE13, San Diego, CA, United States</td>
</tr>
<tr>
<td></td>
<td>Service using Galaxy, Globus Online, and Amazon Web Services</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Supporting Genomics and other Biological Research</td>
<td></td>
</tr>
<tr>
<td>September 28 - October 1</td>
<td>Galaxy Workshop</td>
<td>The Genomic Bioinformatics Workshop, Sydney, Australia</td>
</tr>
<tr>
<td>October 1-3</td>
<td>Beyond the Genome 2013</td>
<td>Beyond the Genome 2013, San Francisco, California, United States</td>
</tr>
<tr>
<td>October 7-8</td>
<td>TTO</td>
<td>NGS &amp; Bioinformatics Suite, Europe</td>
</tr>
<tr>
<td>October 9-11</td>
<td>Galaxy Training Days</td>
<td>GenoToul bioinformatics facilities, IRISA, Toulouse Azeville, France</td>
</tr>
<tr>
<td>October 22-26</td>
<td>High Throughput Data Analysis and Visualization with Galaxy</td>
<td>ASHG 2013, Boston, Massachusetts, United States</td>
</tr>
<tr>
<td>November 6-12</td>
<td>Computational and Comparative Genomics Course</td>
<td>Cold Spring Harbor Laboratory, York, United States</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. Older news items are available in the Galaxy News Archive.

See also:
- Galaxy News Briefs
- Galaxy Updates
- Galaxy on Twitter
- Events
- Learn
- Support
- About the Galaxy Project

**News Items**

**New CloudMan Release**

We just released an update to Galaxy CloudMan. CloudMan offers an easy way to get a personal and completely functional instance of Galaxy in the cloud in just a few minutes, without any manual configuration.

IMPORTANT - please read

Any new cluster will automatically start using this version of CloudMan. Existing clusters will be given an option to do an automatic update once the main interface page is refreshed. Note that this upgrade is a major version upgrade and thus the migration is rather complicated. The migration process has been automated but will take a little while to complete. If you have made customizations to your cluster in terms of adding file systems, upgrading the database, or similar, we do not recommend you perform the upgrade. Note that this upgrade comes with (and requires) a new AMI (ami-11806cf8), which will automatically be used when starting an instance via CloudLaunch.

This update brings a large number of updates and new features, the most prominent ones being:
- Unification of galaxytools and galaxydata file systems into a single galaxy file system. This change makes it possible to utilize the Galaxy Tool Shed when installing tools into Galaxy.
- Added initial support for Hadoop-like workloads.
- Added additional support for cluster federation via HTCondor.
- Added a new file system service for instance's transient storage, allowing it to be used across the cluster over NFS.
- Added a service for the Galaxy Reports webapp.
- Added optional Loggly based off-site logging support.
- Added tags to all resources utilized by CloudMan.

For more details on the new features, see the the CHANGELOG and for even more details see, all 291 commit messages from 7 contributors.

Enjoy and please let us know what you think.

Enis Akgun

**SlipStream Appliance: Galaxy Edition**

**Posted to the Galaxy News on 2013-07-08**
Galaxy is hiring post-docs and software engineers

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
11:30 RNA-Seq Example Part I
12:30 Lunch
1:30 RNA-Seq Example Part II
2:20 Galaxy Project Overview
2:40 Break
3:00 RNA-Seq Example Part III
3:40 Sharing, Publishing and Reproducibility
4:00 Setting up Galaxy on the Amazon Cloud
4:30 Done
Agenda

9:00 Welcome
9:20 Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:30 Lunch
1:30 RNA-Seq Example Part II
2:20 Galaxy Project Overview
2:40 Break
3:00 RNA-Seq Example Part III
3:40 Sharing, Publishing and Reproducibility
4:00 Setting up Galaxy on the Amazon Cloud
4:30 Done
RNA-Seq Example: Part III

- Run Cufflinks on Tophat output to assemble reads into transcripts
- Run Cuffdiff on Tophat output to find significant differences in expression.

- Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq differential expression analysis here.

Cuffdiff

- Which Transcript definitions to use?
  - IGenomes
  - Adrenal or Brain from Cufflinks
  - Run Cuffmerge on Adrenal & Brain Cufflinks files
- Depends on what you care about.
  - I care about a timely workshop, so I’ll use IGenomes.
Cuffdiff

- Produces 11 output files, all explained in doc
- We’ll focus on gene/transcript differential expression testing files (also care about gene/transcript FPKM files)
- Column 7 ("status") can be FAIL, NOTEST, LOWDATA or OK

  - Filter and Sort → Filter
    - c7 == ‘OK’ or C7 == ‘LOWDATA’
- Column 14 ("significant") can be yes or no
  - c14 == ‘yes’
Agenda

9:00  Welcome
9:20  Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:30 Lunch
1:30  RNA-Seq Example Part II
2:20  Galaxy Project Overview
2:40  Break
3:00  RNA-Seq Example Part III
3:40  Sharing, Publishing and Reproducibility
4:00  Setting up Galaxy on the Amazon Cloud
4:30  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

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 Kosakovs\-ky 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\(^1\)
Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovskiy Pond¹,²,⁶,³, Samir Wadhawan³,⁶,⁷, Francesca Chiaromonte⁴, Guruprasad Ananda¹,³, Wen-Yu Chung¹,³,⁸, James Taylor¹,⁵,⁹, Anton Nekrutenko¹,³,⁹ and The Galaxy Team¹

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, SAMIR WADHAWAN, FRANCESCA CHIAROMONTE, GURUPRASAD ANANDA, WEN-YU CHUNG, JAMES TAYLOR, ANTON NEKRUTENKO and THE GALAXY TEAM

Correspondence should be addressed to SKP, IT, 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 there the analyses are rendered through Galaxy via the above workflow and uploaded:

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
11:30 RNA-Seq Example Part I
12:30 Lunch
1:30  RNA-Seq Example Part II
2:20  Galaxy Project Overview
2:40  Break
3:00  RNA-Seq Example Part III
3:40  Sharing, Publishing and Reproducibility
4:00  Setting up Galaxy on the Amazon Cloud
4:30  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 were using this today.

http://aws.amazon.com/education
Could do this step by step, but ...


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 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
Instant CloudMan
http://usegalaxy.org/cloudlaunch
Agenda

9:00  Welcome
9:20  Basic Analysis with Galaxy
10:40 Break
11:00 Basic Analysis into Reusable Workflows
11:30 RNA-Seq Example Part I
12:30 Lunch
1:30  RNA-Seq Example Part II
2:20  Galaxy Project Overview
2:40  Break
3:00  RNA-Seq Example Part III
3:40  Sharing, Publishing and Reproducibility
4:00  Setting up Galaxy on the Amazon Cloud
4:30  Done, almost
Instant Feedback

Acknowledgements

Barrie Hayes  
Hemant Kelkar  
Erin Foster  
Erin Morris  
Julia Shaw-Kokot  
John Wysor  
You

Tom Randall  
Trudy Mackay  
The Galaxy Team

The University of North Carolina  
Health Sciences Library  
AWS Education Grant  
NIH  NSF  Huck Institute  
Penn State University  Emory University

Thanks


Dave Clements
Galaxy Project
Emory University
clements@galaxyproject.org