






You are free to:

-   Copy, share, adapt, or re-mix;
-    Photograph, film, or broadcast;
-    Blog, live-blog, or post video of;

This presentation. Provided that:

-  You attribute the work to its author and respect the rights and licenses associated with its components.

Slide Concept by Cameron Neylon, who has waived all copyright and related or neighbouring rights. This slide only [ccZero](#). Social Media Icons adapted with permission from originals by Christopher Ross. Original images are available under GPL at: <http://www.thisismyurl.com/free-downloads/15-free-speech-bubble-icons-for-popular-websites>

<http://www.ebi.ac.uk/training/course/embo-practical-course-analysis-high-throughput-sequencing-data-1>

The screenshot shows the EMBL-EBI Training website. At the top, there is a navigation bar with 'Services', 'Research', and 'Training'. Below this is a yellow banner with the word 'Training'. A secondary navigation bar includes 'Training Home', 'Train at EBI', 'Train outside EBI', 'Public events', 'Train Online', 'About Training', and 'Contacts'. The main content area features the title 'EMBO Practical Course on Analysis of High-Throughput Sequencing Data'. Below the title, there is a grey box containing details: Venue (European Bioinformatics Institute, Cambridge, CB10 1SD, United Kingdom), Date (Monday, October 21, 2013 - Saturday, October 26, 2013), Organizers (Gabriella Rustici, EMBL-EBI, UK), Admin support (Holly Foster, EMBL-EBI, UK), Registration Deadline (Friday, August 23, 2013), Acceptance Notification Date (Friday, September 6, 2013), and Participation (Open application with selection). To the right of this text is a purple logo. Below the details is a 'Sponsors' section with the EMBO logo (excellence in life sciences) and the text 'supported by AWS in Education Grant award'.

Day 5 : Galaxy





3




Module 5 part 1 Galaxy

BF Francis Ouellette
EMBO Practical Course on Analysis of High-Throughput Sequencing Data



E-mail	francis@oicr.on.ca
	@bffo
	@EMBOcomm
	@emblebi
	#usegalaxy

Day 5 : Galaxy 

Disclaimer

- I do not (and will not) profit in any way, shape or form, from any of the brands, products or companies I may mention.
- I am on the Galaxy Scientific Advisory Board (Galaxy's NIH grant), but I do that for free.

Outline

- Bioinformatics.ca
- Workflows & an examples on using Galaxy platform for DNA sequence manipulations.
- More about interactions, networks, pathways and visualization.

Day 5 : Galaxy



Galaxy outline

- Bioinformatics.ca
- Workflows & an examples on using Galaxy platform for DNA sequence manipulations.
- Reproducible Science
- Galaxy Public server; Galaxy @home; Galaxy in the cloud
- Putting and getting data in and out of Galaxy
- Processing Data in Galaxy
- Galaxy and #RNASeq
- Lab

Day 5 : Galaxy

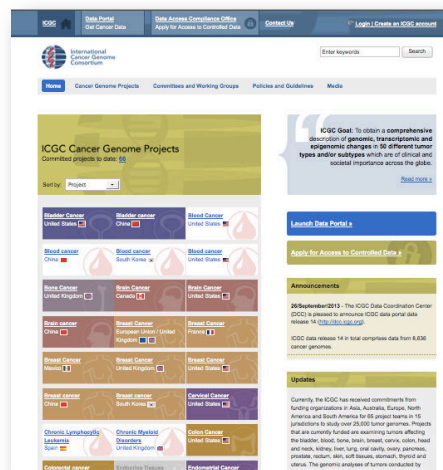




Day 5 : Galaxy



International Cancer Genome Consortium: icgc.org



Day 5 : Galaxy



ICGC data portal: dcc.icgc.org

ICGC Data Portal

Search: BRAF, KRAS G12D, D035108, MU7870, TCGA-06-5858

About Us
The ICGC Data Portal provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.

Data Release 14
September 26th, 2013

Donor Distribution by Primary Site

Cancer projects	41
Cancer primary sites	18
Donors	8,532
Simple somatic mutations	2,184,526
Mutated genes	54,682

Information
Access Raw Data
Methods
Submitter Tools

Tutorial
EXAMPLE QUERIES

1. BRAF missense mutations in colorectal cancer
2. Most frequently mutated genes in stage III malignant lymphoma
3. Brain cancer donors with frameshift mutations and having methylation data available

Tweets

Amber Johns @AmberJay
The new @icgc_dcc data portal is brilliant. Even clinicians can use it! #bravo #pancreaticcancer
13 Retweeted by ICGC DCC

BF Francis Ouellette (de)is
Video for the DCC ICGC data portal tutorial.
vimeo.com/775822668 | dcc.icgc.org | #bioinformatics #cancer #genomics
13 Retweeted by ICGC DCC

Day 5 : Galaxy



http://www.csb.utoronto.ca/

CSB
Cell & Systems Biology
UNIVERSITY OF TORONTO

Department of Cell & Systems Biology
225 Harbord Street
Toronto, Ontario
Canada M5S 3G5
phone: 416-946-3692
fax: 416-978-8532

Featured Hot Paper

Chen, C.Y., Morris, Q., Mitchell, J.A. (2012). Enhancer Identification in Mouse Embryonic Stem Cells Using Integrative Modeling of Chromatin and Genomic Features. BMC Genomics. 13, 152. Abstract

All Publications

Job Opportunities

All positions

Welcome to the Department of Cell & Systems Biology

The molecular biology revolution that dominated the life sciences in the second half of the 20th century has given us an unprecedented ability to explore the behaviour of cells - the fundamental units of life - in terms of molecular processes within and between cells. Researchers in the Department of Cell and Systems Biology, combine many different high-throughput, cell imaging, and physiological methods to characterize and understand cellular and physiological processes in both model (Arabidopsis, Drosophila, Mouse, Zebrafish, Pseudomonads) and non-model organisms.

Latest News

David Hogg won the American Physiological Society 2012 Scholander Award

David Hogg, a PhD student in the Buck Lab, has won the prestigious 2012 Scholander Award from the American Physiological Society's Comparative and Evolutionary Physiology Section. The award recognizes an outstanding research presentation by a young investigator and includes the opportunity to organize a symposium for next year's meeting. Dave's poster entitled: ROS scavenging mimics anoxia by enhancing GABA receptor-mediated electrical suppression in anoxia-tolerant turtle cortex, beat out 29 others.

Day 5 : Galaxy



<http://bioinformatics.ca/>

bioinformatics.ca

Home About Workshops Job Postings Resources

bioinformatics.ca is...
The Canadian Bioinformatics Workshops offer hands-on, intensive workshops to gain bioinformatics skills training in various topic areas relevant to life sciences research today.

New Job Postings **Come to Work HIRING**

Front-End Web Developer and Quality Analyst (Downtown Toronto)
Scientific Systems Administrator (Boston, MA)
Post Doctoral Fellow at the Samuel Lunenfeld Research Institute in the laboratory of Dr. Steven Gallinger (Toronto, ON)
[See More Jobs >](#)

Recent News

May 28, 2012 - 50th Canadian Bioinformatics Workshop being held May 28-June 1, 2012 in Toronto. | [Details](#)

January 20, 2012 - New 3-Day Patent Informatics workshops in collaboration with EBI and EPO occurring

CANADIAN BIOINFORMATICS WORKSHOPS
New Workshop Schedule for 2012

Featured

2-Day Workshop
Exploratory Analysis of Biological Data using R

NEW 5-Day Workshop
Bioinformatics for Cancer Genomics (BiCG)

Upcoming Workshops

Exploratory Analysis of Biological Data using R
September 6 - September 7, 2012
Downtown Toronto, ON
[Apply Now](#) [Award Opportunities](#)

[See More Workshops](#)

Apply now!
[Application Form](#)

Students Postdocs & Fellows:
[Award Opportunities](#)

Meet the instructors:
Faculty

Want to know more?
[Workshop FAQ](#)

Day 5 : Galaxy

bioinformatics.ca

Home About Workshops Job Postings Resources

Home > Workshops > 2013 > Bioinformatics for Cancer Genomics (BiCG) (2013)

WORKSHOPS

Past Workshops

- [+] 2012
- [+] 2011
- [+] 2010
- [+] 2009
- [+] 2008
- [+] 2006
- [+] 2005
- [+] 2004

Workshop Application
Awards
Faculty
Testimonials
Sponsors
FAQ

[User Login](#)

Bioinformatics for Cancer Genomics (BiCG) (2013)

May 27 - May 31, 2013
Downtown Toronto, ON

Closed This workshop has taken place.

Lead Faculty

- John McPherson
- Francis Ouellette
- Anna Lapuk
- Jared Simpson
- Sohrab Shah
- Lincoln Stein
- Paul Boutros
- Obi Griffith

Share this page

Canadian Bioinformatics

Course Objectives

Advertising poster for this workshop is [here](#)

Cancer research has rapidly embraced high throughput technologies into its research, using various microarray, tissue array, and next generation sequencing platforms. The result has been a rapid increase in cancer data output and data types. Now more than ever, having the informatic skills and knowledge of available bioinformatic resources specific to cancer is critical. The CBW has developed a **5-day workshop** covering the key bioinformatics concepts and tools required to analyze cancer genomic data sets. Participants will gain experience in genomic data visualization tools which will be applied throughout the development of the skills required to analyze cancer -omic data for gene expression, genome rearrangement, somatic mutations and copy number variation. The workshop will conclude with analyzing and conducting

Day 5 : Galaxy

Workshops planned for 2014:

<http://bioinformatics.ca/workshops>

1. Exploratory Analysis of Biological Data using R
2. Bioinformatics for Cancer Genomics
3. Informatics for RNA-sequence Analysis
4. Informatics on High Throughput Sequencing Data
5. Pathway and Network Analysis of -omics Data
6. Flow Cytometry Data Analysis using R
7. Microarray Data Analysis
8. Informatics and Statistics for Metabolomics

Day 5 : Galaxy



<http://bioinformatics.ca/workshops/2013>

Access Past Workshops

Canadian Bioinformatics Workshops promotes open access. Past workshop content is available under a Creative Commons License.

Course Material

Canadian Bioinformatics Workshops promotes open access. Past workshop content is available under a Creative Commons License.

Module 1 - Reference-guided Genome Alignment (2013) (Faculty: [Michael Stromberg](#))

[PDF](#) (5MB) | [PPTX](#) (3MB) | [MM](#) (37MB)

Module 2 - Small-variant calling and annotation (2013) (Faculty: [Michael Stromberg](#))

[PDF](#) (14MB) | [PPTX](#) (11MB) | [MM](#) (28MB)

Module 3 - Structural variant calling (includes CNVs) (2013) (Faculty: [Abron Quinlan](#))

[PDF](#) (16MB) | [MM](#) (404MB)

The screenshot shows the website interface for past workshops. The main content area lists several workshops from 2013, including 'Exploratory Analysis of Biological Data using R (2013)', 'Bioinformatics for Cancer Genomics (BCG) (2013)', 'Informatics for RNA-sequence Analysis (2013)', 'Informatics on High Throughput Sequencing Data (2013)', 'Pathway and Network Analysis of -omics Data (2013)', 'Flow Cytometry Data Analysis using R (2013)', 'Microarray Data Analysis (2013)', and 'Informatics and Statistics for Metabolomics (2013)'. Each entry includes the date and an 'Access Content' button. A red arrow points from the 'Access Past Workshops' section of the left sidebar to the 'Access Content' button for the first workshop listed.

Day 5 : Galaxy



E-mail: course_info@bioinformatics.ca

Web: <http://bioinformatics.ca>

Workshop announcement mailing list:

<http://bioinformatics.ca/mailman/listinfo/announce>

Day 5 : Galaxy

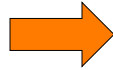
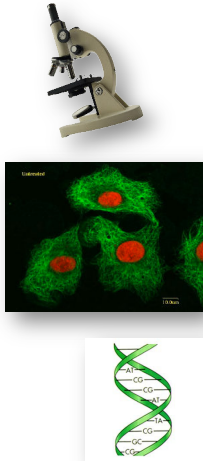


Bioinformatics is about integrating biological themes together with the help of computer tools and biological databases, and gaining new knowledge about the system in study.

Day 5 : Galaxy



What biologist do:



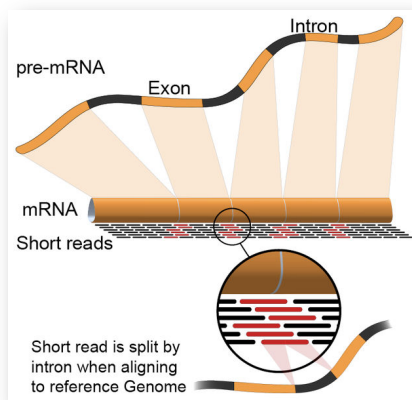
- Make observations
- Make hypothesis
- Test them
- Challenge them
- Conclude things
- Write papers

<http://goo.gl/7sCUI>

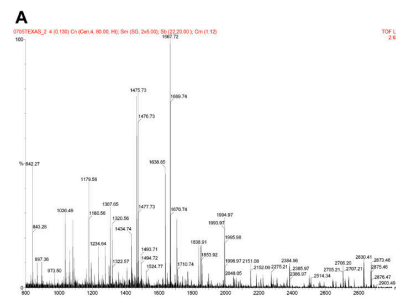
Day 5 : Galaxy



RNA-Seq



Protein MS

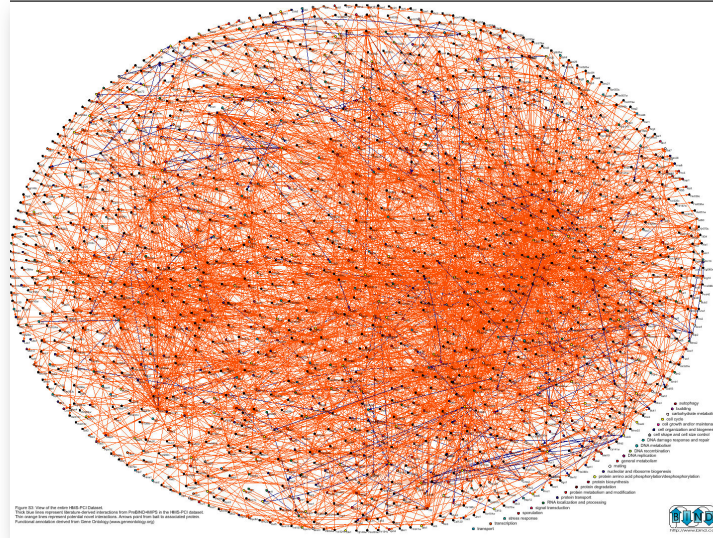


<http://goo.gl/Lye8R>

Day 5 : Galaxy



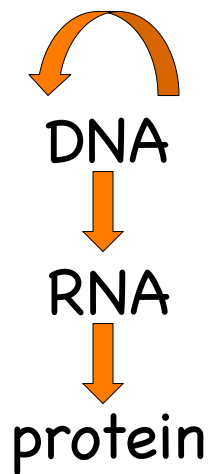
Interaction and Pathway Space



Day 5 : Galaxy

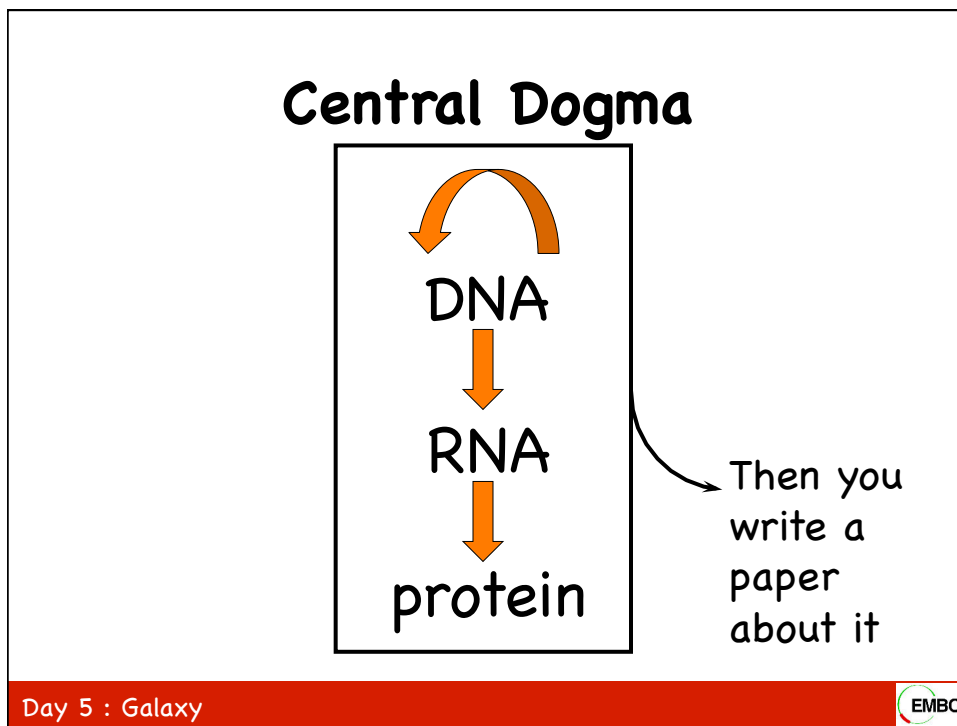


Central Dogma



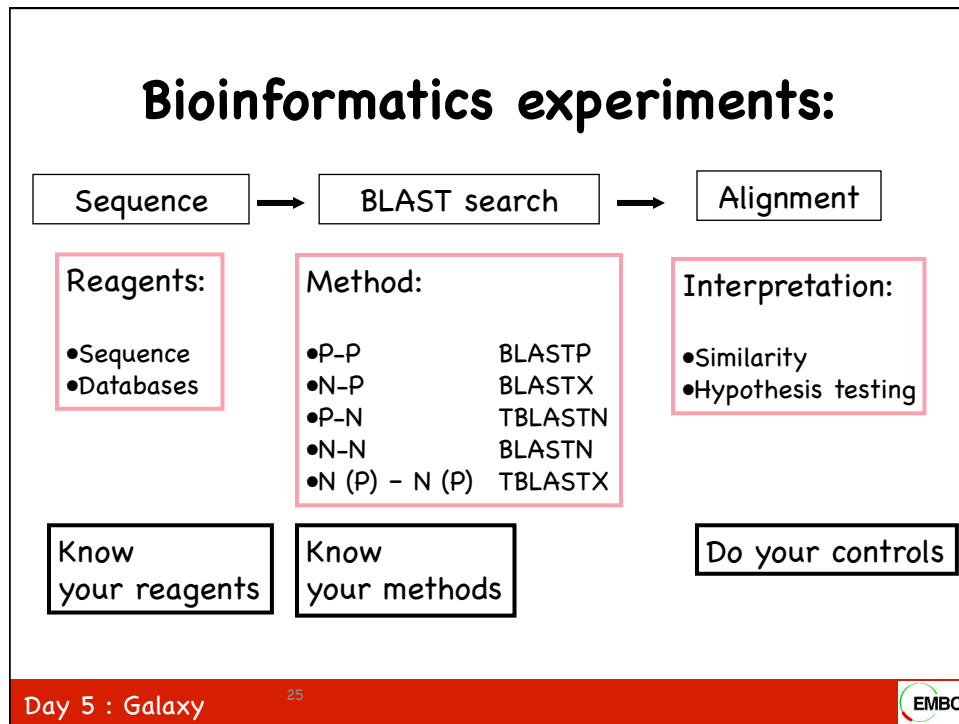
Day 5 : Galaxy





Some of the things we do when we try and understand the cell ...

- We do experiments
- Some of these are bioinformatics experiments
- We all want these to be reproducible
- We want people to find our data
- We want people to find our methods
- ... and we want them to be able to rerun our experiments, validate our work, move the science forward.



Doing and redoing experiments

- If you do something once, you usually don't need a script. Do it hundreds or thousands of times, you will want something to help you.
- Want to share what you did, providing a script is usually a good way.
- Sometimes though, scripts are too complicated, and don't capture all that is need to do an experiment. For example: the version of a tool you used!

Some requirements:

- Open Source
- Solution should be useful to large community
- Well supported (by community and funding agency)
- Flexible
- Expandable
- Scalable
- Cloud-aware
- User friendly?



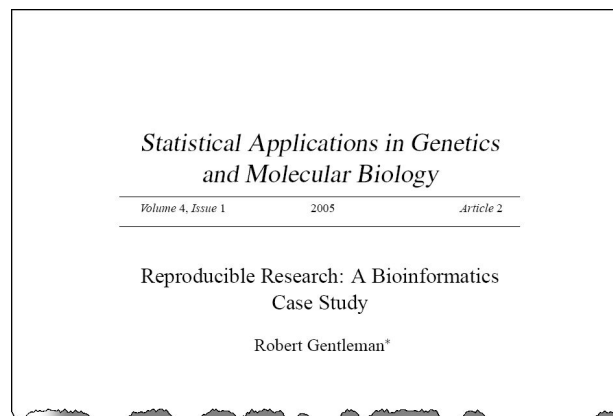
Open Source

Day 5 : Galaxy



Some solutions

1. R and bioconductor (#rstat)



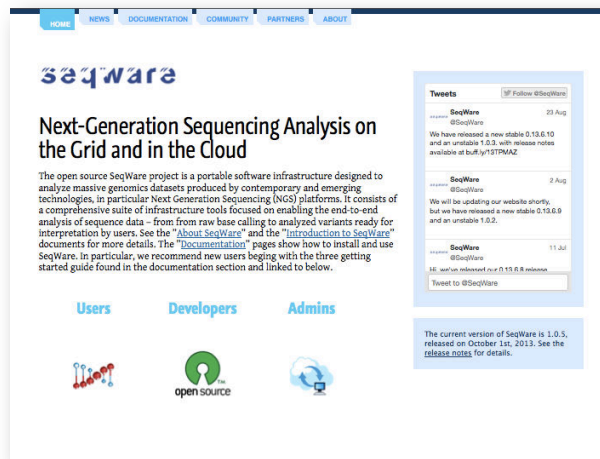
<http://www.ncbi.nlm.nih.gov/pubmed/16646837>

Day 5 : Galaxy



Some solutions (2)

- SeqWare : <http://seqware.github.io/>



Day 5 : Galaxy



Some workshop thoughts:

- Since 1999 we've been experimenting with servers of one kind or another.
- At 1st we had sponsorship from Sun Microsystems, and had a 4 CPU E450.
- A few of years we had a Linux laptop in the back of the room: a web server and sharing some data.
- In 2007-8 we were doing courses in same building as OICR, and has the systems group carve us 40 cores from the cluster.
- In 2010 We experimented with my MacPro in my office serving up Virtual Machines for students (it didn't work)

Day 5 : Galaxy



2011



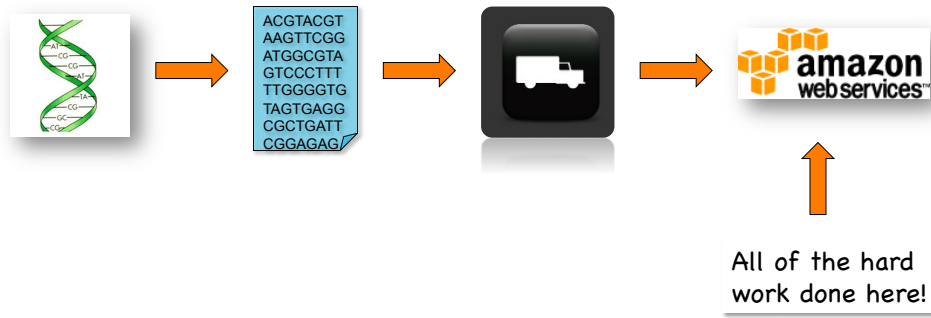
<http://aws.amazon.com/education/>




Day 5 : Galaxy


Genomic companies already there!

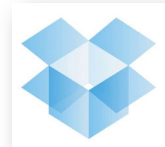
- Complete Genomics' pipeline:



Day 5 : Galaxy


Most people already there!

- Google docs
- Dropbox
- Netflix
- Twitter



Day 5 : Galaxy



Amazon Web Services (AWS)

- Infinite storage (scalable): S3 (simple storage service)
- Compute per hour: EC2 (elastic cloud computing)
- Ready when you are High Performance Computing
- Multiple football fields of HPC throughout the world
- HPC are expanded at one contained at a time:



<http://goo.gl/7PVAt>



Day 5 : Galaxy



Some solutions (3)

- Galaxy



Day 5 : Galaxy



Goecks *et al.* *Genome Biology* 2010, 11:R86
<http://genomebiology.com/2010/11/8/R86>



SOFTWARE

Open Access

Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences

Jeremy Goecks¹, Anton Nekrutenko^{2*}, James Taylor^{1*}, The Galaxy Team

<http://genomebiology.com/2010/11/8/R86>

Day 5 : Galaxy



Using Cloud Computing Infrastructure with CloudBioLinux, CloudMan, and Galaxy

Enis Afgan,^{1,5} Brad Chapman,² Margita Jadan,³ Vedran Franke,⁴ and James Taylor⁵

¹Center for Informatics and Computing, Ruder Bošković Institute (RBI), Zagreb, Croatia
²Harvard School of Public Health, Boston, Massachusetts
³Division of Materials Chemistry, Laboratory for Ichthyopathology–Biological Materials, Ruder Bošković Institute (RBI), Zagreb, Croatia
⁴Department of Biology, University of Zagreb, Zagreb, Croatia
⁵Department of Biology and Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia

ABSTRACT
Cloud computing has revolutionized availability and access to computing and storage

UNIT 11.9

Current Protocols in Bioinformatics 11.9.1-11.9.20, June 2012
Published online June 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471250953.bi1109s38
Copyright © 2012 John Wiley & Sons, Inc.

<http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1109s38/pdf>

Day 5 : Galaxy

Which Galaxy?


- galaxyproject.org: Galaxy home page
- usegalaxy.org: main Galaxy public server
- getgalaxy.org: source for installing local Galaxy
- usegalaxy.org/cloud: use galaxy in the cloud
- <http://goo.gl/mlyOC> : Other public Galaxy servers

	Main	Local	Cloud	Other
Your data sets are moderately sized	Yes	Yes	Yes	?
Your computational requirements are moderate	Yes	Yes	Yes	?
You want to share your Galaxy objects with others	Yes	Yes	Yes	?
All needed Tools are installed on Main.	Yes	?	Yes	?
Your data sets are very large	No	?	Yes	?
Your computational requirements are very large	No	?	Yes	?
You have absolute data security requirements	No	Yes	Yes	?

<http://goo.gl/x3DXm>

Day 5 : Galaxy


http://galaxyproject.org/



Data intensive biology *for everyone.*


Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the [free public server](#) or [your own instance](#), you can perform, reproduce, and share complete analyses.

Use Galaxy




Use the [free public server](#)

Get Galaxy




Install locally or in the [cloud](#)

Learn Galaxy



Screencasts, Galaxy 101, ...


Get Involved



Mailing lists, Tool Shed, wiki

[Search all resources](#)

The Galaxy Team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NH&GRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Day 5 : Galaxy


http://usegalaxy.org



The screenshot shows the Galaxy web interface with a search bar, a list of tools on the left, and a central area with a tutorial banner and live quickies. Orange brackets highlight the search bar and the tool list.

Day 5 : Galaxy


getgalaxy.org

Galaxy Wiki Login | Search: Titles Text

Admin/Get Galaxy Locked History Actions

Get Galaxy: Galaxy Download and Installation

In addition to using the [public Galaxy server](#) (a.k.a. Main), you can also install your own instance of Galaxy (what this page is about), or create an [instance of Galaxy on the cloud](#). Another option is to use one of the ever-increasing number of [Public Galaxy Servers](#) hosted by other organizations.

See [Big Picture/Choices](#) for help on deciding which of these options may be best for your situation.

Reasons to Install Your Own Galaxy

You only need to download Galaxy if you plan to:

1. Develop it further
2. Add new tools
3. Plug-in new datasources, or
4. Run a local production server for your site because you have
 1. Sensitive data (e.g., clinical)
 2. Large datasets or processing requirements that are too big to be processed on Main


Installation Procedure

The installation procedure is simple and is nearly identical for UNIX/Linux and Mac OS X. We are no longer supporting the Windows platform with our distribution, so you will have to build your own Python eggs if you want to install it on Windows (see [Admin/Config/Windows](#) for some tips). Of course, Windows users can use our [public Galaxy server](#) from their browsers.

These instructions describe the basic setup procedure for a development environment, more detailed instructions on how to deploy a production server can be found at the bottom of this page.

Contents

1. Reasons to Install Your Own Galaxy
2. Installation Procedure
 1. Check your Python version
 2. Get the latest copy from the repository
 3. Start it up
 4. Join the Mailing List
 5. Keep your instance backed up
 6. Keep your code up to date
3. Advanced Configuration
4. Other Help



Regular registration rates end 14 June

Use Galaxy
[Use Main \(about\)](#)
[Use Others!](#) • [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](#)

Day 5 : Galaxy



usegalaxy.org/cloud

Galaxy Wiki Login | Search: Titles Text

CloudMan Locked History Actions

Contents

1. About Galaxy on the cloud
2. Instantiating a Galaxy instance on the Amazon cloud
3. Screencast
4. Detailed steps
5. Galaxy AMIs
6. Determining the size of your cloud cluster
7. Customizing your cloud cluster
8. Notes
9. Presentations
10. Publications

Note: There are several choices for using Galaxy. This page describes installing Galaxy on a *cloud infrastructure* (see below). For other options, see [Big Picture/Choices](#).

About Galaxy on the cloud


With sporadic availability of data, individuals and labs may have a need to, over a period of time, process greatly variable amounts of data. Such variability in data volume imposes variable requirements on availability of compute resources used to process given data. Rather than having to purchase and maintain desired compute resources or having to wait a long time for data processing jobs to complete, the Galaxy Team has enabled Galaxy to be instantiated on cloud computing infrastructures, primarily Amazon Elastic Compute Cloud (EC2). An instance of Galaxy on the cloud behaves like a local instance of Galaxy except that it offers the benefits of cloud computing resource availability and pay-as-you-go resource ownership model. Having simple access to Galaxy on the cloud enables as many instances of Galaxy to be acquired and started as is needed to process given data. Once the need subsides, those instances can be released as simply as they were acquired. With such a paradigm, one pays only for the resources they need and use while all the other concerns and costs are eliminated. To see how much using Amazon cloud might cost, you can use the [AWS cost calculator](#). When calculating the total cost, in addition to the EC2 instance, you will have EBS volumes associated with your cluster. There are a total of three EBS volumes associated with each Galaxy cluster: your data volume (size is decided by you when setting up the cluster, say 100GB to begin with), tools volume (10GB), and indices volume (700GB). (Note, the indices volume can be greatly reduced if you don't need all the genome data).

Slides with talks about Galaxy Cloud (including screenshots) can be found [below](#).

Instantiating a Galaxy instance on the Amazon cloud

Important note: preferentially use the [CloudLaunch](https://main-g2.bx.psu.edu/cloudlaunch) interface at <https://main-g2.bx.psu.edu/cloudlaunch>.

For the purposes of executing Galaxy on the cloud, we have packaged Galaxy and Galaxy-required tools as a virtual machine (VM) image that resides with Amazon (referred to as an AMI - Amazon Machine Image). This VM acts as a complete unit that can be easily instantiated offering the same functionality as any other instance of Galaxy.



Regular registration rates end 14 June

Use Galaxy
[Use Main \(about\)](#)
[Use Others!](#) • [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](#)

Day 5 : Galaxy



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

Publicly Accessible Galaxy Servers

Galaxy's public server (UseGalaxy.org, Main) can meet many needs, but it is not suitable for everything (see Big Picture/Choices for why) and cannot possibly scale to meet the entire world's needs.

Fortunately the Galaxy Community is helping out by installing Galaxy at their institutions and then making those installations either publicly available or open to their organizations or community.

This page lists such public or semi-public Galaxy servers.

To add your public Galaxy server to this list, please either just add it (hey, it's a wiki), or contact Galaxy Outreach <outreach AT galaxyproject DOT org>.

Andromeda

- **Link:** [Andromeda](#)
- **Domain/Purpose:** This is a fully populated Galaxy instance.
- **Comments:** Andromeda is hosted at the SURFsara High Performance Computing (HPC) cloud. The installation is supported by Enis Afgan (CloudPlan project) and Matthias de Hollander (NIOO).
- **Quotas:** Registered users: 10GB; Anonymous users: 10MB
- **Sponsor(s):** Netherlands Bioinformatics Centre (NBIC) and BIG Grid SURFsara

galaxy

- **Link:** [galaxy](#)
- **Domain/Purpose:** Hosts the BALL (Biochemical Algorithms Library) Project tools, i.e. computer aided drug design and molecular modelling based on protein and ligand structure data.

Contents

1. Andromeda
2. ballaxy
3. CBIB Galaxy
4. Cidrome Analysis Pipeline
5. DBCLS Galaxy
6. Galaxy Main
7. Galaxy Test
8. GeneNetwork
9. GenBoree
10. Genomic Hyperbrowser
11. Gene Ontology (GO)
12. Gigalaxy
13. GOWIS-viz
14. Huttenhower Lab
15. IBDSite
16. INRA-URGI
17. MSTARX
18. Nebula
19. NELLY
20. Netherlands Metabolomics
21. cDISE
22. OPPL Galaxy
23. Optans
24. P-Galaxy
25. PopGenIE
26. Regulatory Genomics
27. RepeatExplorer
28. RNA-Rocket @ Pathogen Portal
29. Stem Cell Discovery Engine
30. South Green
31. SymD
32. Wageningen University
33. Yeoman

Day 5 : Galaxy



✓ Use it now!

Galaxy allows you to do analyses you cannot do anywhere else without the need to install or download anything. You can analyze multiple alignments, compare genomic annotations, profile metagenomic samples and much much more...

- [Go to our public site](#) and start working.
- [Watch screencasts](#) to get an idea of what we mean.
- [Go to our wiki page](#) to read about technical details.



- Galaxy integrates input data sources
- Galaxy allows you to use many tools that you don't need to install and maintain.
- Galaxy allows you to maintain workflows, reuse them, and share them.
- Galaxy lets you "publish" experiments.
- Galaxy has fully entered the "next-gen" space.
- Galaxy works in the cloud.

Day 5 : Galaxy



Galaxy = collaboration and reproducibility

Best of all, Galaxy's history system provides a complete analyses record that can be shared. Every history is an analysis workflow, which can be used to reproduce the entire experiment...

- **History is an analysis record** | Every step of your analyses is recorded in Galaxy's history system. You can have any number of histories saved. This way you can go back to your analyses anytime.
- **Share your analyses** | Alice works at Penn State, while Bob suffers from the terrible San Diego climate. Alice wants Bob to see her analyses. Alice clicks the "share" link and enters Bob's e-mail address. Now Alice's history is visible to Bob (see "Sharing history" screencast).
- **Now your results are reproducible!** | When publishing results, replace "the data were analyzed using a collection of in-house scripts" with a URL pointing to Galaxy's history. Your reviewers will have no further questions. That's reproducible genomics!

- Galaxy strongly believes on reproducibility!
- Galaxy is very good at keeping a history of what you did, and allow you do it again when you need to, or allow somebody else to do it again.
- Galaxy makes it very to work with collaborators down the hall, or across the ocean.

Day 5 : Galaxy



Designed for biologists and developers

Yep, sometimes you can mix water with oil...

- **Biologists** | [Use our public site](#) to access popular sources of data like the UCSC Table Browser. Run analyses right on the spot using a variety of integrated tools. Your results are never deleted and can be easily shared with others.
- **Developers** | Galaxy is an easy-to-use, open-source, scalable framework for tool and data integration. Stop wasting time writing interfaces and get your tools used by biologists! Galaxy includes everything you need to get started, so [download](#) and [start integrating!](#)

- Galaxy is designed with biologists in mind, and basically thinks like we do (most of the time!)
- Galaxy has a healthy developer community, and is very present in forums of other Open Source initiatives.

Day 5 : Galaxy



✘ Why did we do it?

You are an experimental biologist. You keep watching databases fill with more and more data. You keep thinking: *even if I knew how to use Excel as a pro, it would probably not load 12,435,654 SNPs*. So how do you perform analyses without calling somebody on the Computer Science side of campus? Suppose you want to find human promoters with the highest SNP density. There is no straightforward way of doing it without learning programming first. And this is why...

- **Databases are not analyses tools** | Databases are where you get the data. Browsers are where you visualize the results. For a bench biologist there is not much in between besides spreadsheets or Perl scripting.
- **No tools for new datatypes** | Some datatypes generated by high throughput genomics are so new that there are no tools to analyze them. For example, how do you extract sequences of coding exons from the latest 28-way alignments of vertebrate genomes or analyze quality scores from 454/Selex/SOLID? With Galaxy.
- **Genomics is not really reproducible** | The Methods section of too many papers sound like *the data were analyzed using a collection of in-house scripts*. How do you repeat such a study? Galaxy saves every step of your analysis and allows you to share these workflows with others.
- **Too many tools** | *Bioinformatics* publishes hundreds of application notes per year. How does one know which tool to use? Galaxy integrates a multitude of different tools by giving them the same "look and feel" and linking them to data warehouses.

- To help biologists deal with tools and data.
- Funding: NIH, NSF, & Penn State University.
- Development: Emory University and Penn State
- <http://wiki.galaxyproject.org/>
- <http://wiki.galaxyproject.org/Learn>

Day 5 : Galaxy



Challenge with multiple sites/model

- Not all galaxy are created the same
- Galaxy team moving to an "empty" shell, and cafeteria model: take only what you need.
- Adding tools and updating tools causes problems sometimes, but Galaxy team is working to make this easier
- The Toolshed is a great solution for this!

Day 5 : Galaxy



Galaxy Toolshed: http://toolshed.g2.bx.psu.edu/

Galaxy Tool Shed
2771 valid tools on Jun 03, 2013

Search
 • Search for valid tools
 • Search for workflows

Valid Galaxy Utilities
 • Tools
 • Custom datatypes
 • Repository dependency definitions
 • Tool dependency definitions

All Repositories
 • Browse by category

Available Actions
 • Login to create a repository

Categories
 search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	20
Computational chemistry	Tools for use in computational chemistry	4
Convert Formats	Tools for converting data formats	25
Data Source	Tools for retrieving data from external data sources	13
Fasta Manipulation	Tools for manipulating fasta data	23
Genomic Interval Operations	Tools for operating on genomic intervals	21
Graphics	Tools producing images	10
Metabolomics	Tools for use in the study of Metabolomics	0
Metagenomics	Tools enabling the study of metagenomes	8
Micro-array Analysis	Tools for performing micro-array analysis	6
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	45
Ontology Manipulation	Tools for manipulating ontologies	6
Phylogenetics	Tools for performing phylogenetic analysis	3
Proteomics	Tools enabling the study of proteins	31
SAM	Tools for manipulating alignments in the SAM format	22
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	118
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	25
Statistics	Tools for generating statistics	24
Systems Biology	Systems biology tools	4
Text Manipulation	Tools for manipulating data	29
Tool Dependency Packages	Repositories that contain third-party tool dependency package installation definitions	5
Tool Generators	Tools that make or help make new tools	1
Visualization	Tools for visualizing data	24
Web Services	Tools enabling access to web services	4

Day 5 : Galaxy



Galaxy Toolshed: SAM

Galaxy Tool Shed
2771 valid tools on Jun 03, 2013

Search
 • Search for valid tools
 • Search for workflows

Valid Galaxy Utilities
 • Tools
 • Custom datatypes
 • Repository dependency definitions
 • Tool dependency definitions

All Repositories
 • Browse by category

Available Actions
 • Login to create a repository

Category SAM
 search repository name, description

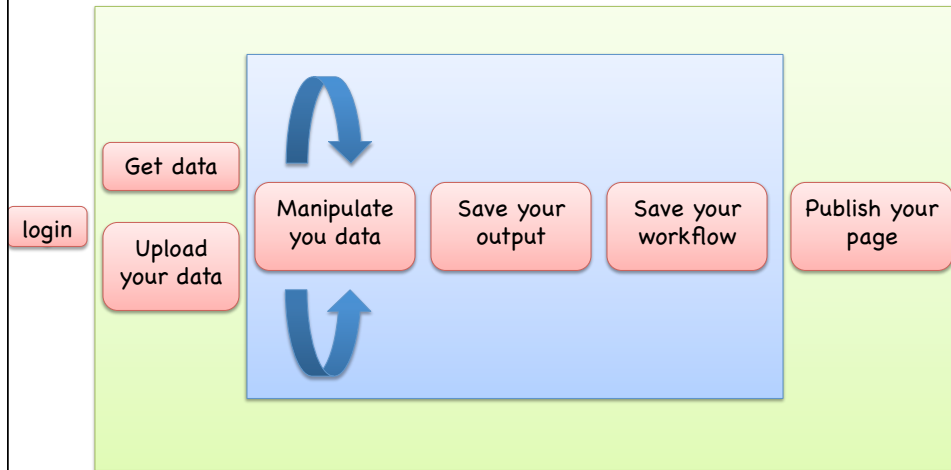
Name	Synopsis	Metadata Revisions	Tools Verified	Owner
bamedit	Merging, splitting, filtering, and QC of BAM files (bamedit)	15:eb166cbb3c	no	modencode-dcc
bam_to_bigwig	Generate BigWig coverage files from BAM files. Allows gapped reads to be split (useful for RNA-Seq).	3:294e9dae5a9b	no	brad-chapman
bam_to_fastq	Convert BAM file to fastq	0:5a9ada9a3191	no	brad-chapman
bedtools	Flexible tools for genome arithmetic and NGS analysis.	1:41bba3e648d1	no	aaronquinlan
deseq_and_sam2counts	Performs RNA-Seq differential expression analysis on aligned reads to a transcriptome using sam2counts and DESeq 1.8.3	3:a49aff09553e	no	nikhil-joshi
dwgsim_eval	Evaluate simulated reads from a SAM/BAM file using dwgsim_eval	1:eb58ceefdb	no	nishomec
ea_utils	ea-utils FASTQ processing utilities (currently fastq-join and sam-stats)	3:f0d19a935325	no	learsons
filter_on_md	Filter mapped reads on MD tag string	2:ac70bfaf1224	no	boris
htseq_count	Count aligned reads (SAM/BAM) that overlap genomic features (GFF)	12:62a1de8c8aae	no	learsons
nextgen_variant_identification	SNVMix-based tools for variant calling from aligned illumina sequence data	7:351b3acadd17	no	ryanmorin
package_samtools_0.1.16	Contains a tool dependency definition that downloads and compiles version 0.1.16 of the SAMTools package	0:75367f13eb3c	n/a	devteam
package_samtools_0.1.18	Contains a tool dependency definition that downloads and compiles version 0.1.18 of the SAMTools package	0:a7936f4e405	n/a	devteam

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

Day 5 : Galaxy



General workflow for Galaxy



Day 5 : Galaxy



Time for sponsor announcement!



<http://www.pmgenomics.ca/>



Zhibin Lu

Day 5 : Galaxy



Setup for this workshop done ahead of time: <https://usegalaxy.org/cloud>

Day 5 : Galaxy EMBO

<http://usegalaxy.org/cloudlaunch>

Key ID

This is the text string that uniquely identifies your account, found in the [Security Credentials](#) section of the [AWS Console](#).

Secret Key

This is your AWS Secret Key, also found in the [Security Credentials](#) section of the [AWS Console](#).

Key ID: AKIAIQMAL3HP52I3VQZA
 Secret Key: e/LyD7rYh6vXLkIzOIimMURpvQ2iycUYRSXayBnc

Request Instances Wizard

Provide the details for your instance(s). You may also decide whether you want to launch your instances as "on-demand" or "spot" instances.

Number of Instances: 1 Instance Type: Micro (t1.micro, 613 MB)

Type	CPU Units	CPU Cores	Memory
Micro (t1.micro) ★ Free tier eligible	Up to 2 ECUs	1 Core	613 MB
Small (m1.small)	1 ECU	1 Core	1.7 GB
High-CPU Medium (c1.medium)	5 ECUs	2 Cores	1.7 GB
Medium (m1.medium)	2 ECUs	1 Core	3.7 GB
Large (m1.large)	4 ECUs	2 Cores	7.5 GB
Extra Large (m1.xlarge)	8 ECUs	4 Cores	15 GB
High-Memory Extra Large (m2.xlarge)	6.5 ECUs	2 Cores	17.1 GB
High-Memory Double Extra Large (m2.2xlarge)	13 ECUs	4 Cores	34.2 GB
High-Memory Quadruple Extra Large (m2.4xlarge)	26 ECUs	8 Cores	68.4 GB
High-CPU Extra Large (c1.xlarge)	20 ECUs	8 Cores	7 GB

Day 5 : Galaxy

Launch a Galaxy Cloud Instance

To launch a Galaxy Cloud Cluster, enter your AWS Secret Key ID, and Secret Key. Galaxy will use these to present appropriate options for launching an Amazon Cloud cluster. See [Amazon's pricing](#) for more information.

Key ID: AKIAIQMAL3HP5213VQZA

Secret Key: eLjV07Yn6vXkizQlimMUrpnQ2yctYR8kayfnc

Cluster Name: eb100

Cluster Password: *****

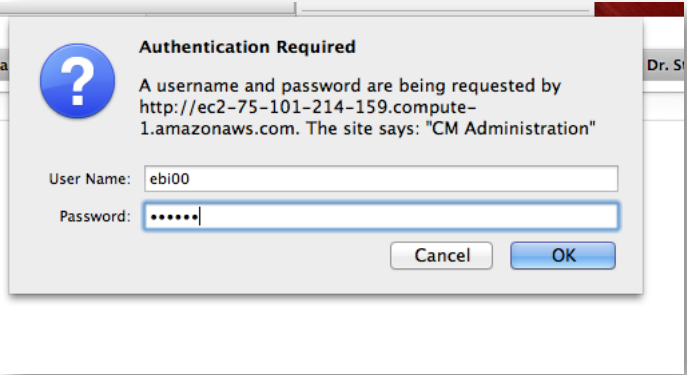
Cluster Password - Confirmation: *****

Key Pair: cloudman_keypair

Instance Type: Extra Large

- Key ID
- Secret Key
- **ebiXX**
- **ebiXX**
- Retype password
- Cloudman keypair
- **Extra Large**

Day 5 : Galaxy




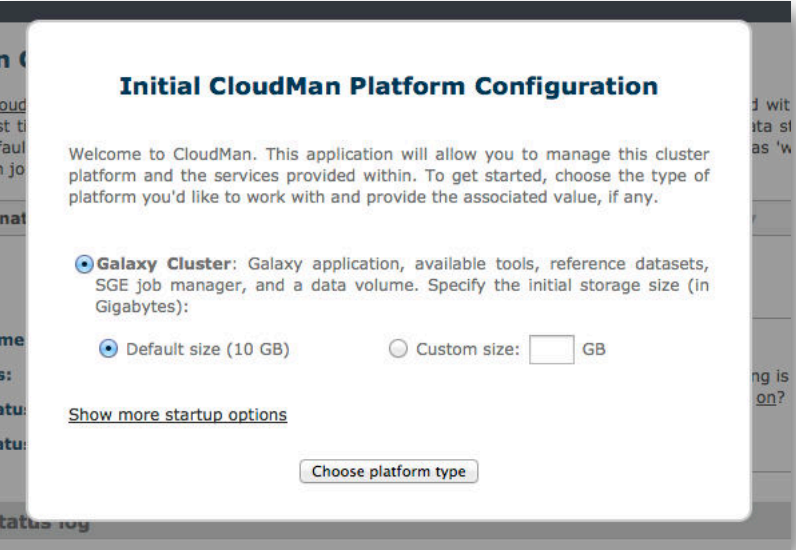
Authentication Required

A username and password are being requested by <http://ec2-75-101-214-159.compute-1.amazonaws.com>. The site says: "CM Administration"

User Name:

Password:

Day 5 : Galaxy 




Initial CloudMan Platform Configuration

Welcome to CloudMan. This application will allow you to manage this cluster platform and the services provided within. To get started, choose the type of platform you'd like to work with and provide the associated value, if any.

- Galaxy Cluster:** Galaxy application, available tools, reference datasets, SGE job manager, and a data volume. Specify the initial storage size (in Gigabytes):
 - Default size (10 GB)
 - Custom size: GB

[Show more startup options](#)

Day 5 : Galaxy 

Messages
Initializing 'Galaxy' cluster type. Please wait... (2013-10-19 17:14:53)

Welcome to **CloudMan**. This application allows you to manage this cloud cluster and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as add nodes on which jobs are run.

Buttons: Terminate cluster, Add nodes, Remove nodes, Access Galaxy

Status

Cluster name: eb100
Disk status: 0 / 0 (0%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications Data

Autoscaling is off. Turn on?

Cluster status log

Messages
Initializing 'Galaxy' cluster type. Please wait... (2013-10-19 17:14:53)

All cluster services started; the cluster is ready for use. (2013-10-19 17:17:18)

this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as add nodes on which jobs are run.

Buttons: Terminate cluster, Add nodes, Remove nodes, Access Galaxy

Status

Cluster name: eb100
Disk status: 2.9G / 10G (29%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications Data

Autoscaling is off. Turn on?

Cluster status log

Day 5 : Galaxy

Galaxy

Tools

- Get Data
- Send Data
- ENCODE Tools
- Life-Cycle
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution

Welcome to Galaxy on the Cloud
managed by CloudMan

History

Unnamed history
0 bytes

Your history is empty. Click 'Get Data' on the left pane to start.

Day 5 : Galaxy

Analyze Data Workflow Shared Data

Create account

Email address:
francis@oicr.on.ca

Password:
•••••

Confirm password:
•••••

Public name:
ebi00

generate addresses for intl characters in length and contain only lower-case letters, numbers, and the '-' c'

Submit

Day 5 : Galaxy

Welcome to Galaxy on the Cloud

Tools: Get Data, Send Data, INSDIC Tools, Life, Draw, Task Manipulation, Filter, and Sort, Jvarkit, Submit and Coma, Convert Formats, Enrich Statistics, Enrich Sequences, Enrich Alignments, Gen. Genomic Scores, Overview, Get Genomic Intervals, Statistics, Graph Display Data, Repeat Selection, Multiple representation, Multivariate Analysis, Evaluation, Model Tools, Multiple Alignments, Measurement, analysis, FASTA manipulation, MCL, BLAST, MCL, GC, and manipulation, MCL, Pfam, MCL, Assembly, MCL, Motifs, MCL, InDel Analysis, MCL, RNA Analysis, MCL, SAM Tools, MCL, CAS, Tools, MCL, PAX, Calling, SNP, WGS, Data, Filters, SNP, WGS, GC, GC, Frag, SNP, WGS, Statistical Models, Human Genome, Variation, VCF, Tools, DRACO, Workflows

History: Unsaved history, Your history is empty. Click 'Get Data' on the left pane to start.

Day 5 : Galaxy

- Software on AMI:
 - java openjdk-7-jre, 1.7.0_21
 - R 3.0.1 (include bioconductor, cummeRbund, ggplot2, edgeR, HMMcopy),
 - Bedtools 2.17.0
 - Hyra 0.5.3
 - JointSNVMix 0.8.0-b2,
 - PennCNV
 - Apolloh, MATLAB runtime library v7.7,
 - HMMcopy 0.1.1,
 - MutationSeq,
 - bowtie2 2.1.0,
 - tophat2 2.0.8b
 - cufflinks2 2.1.1
 - picard 1.9.1
 - bwa 0.74,
 - samtools 0.1.19,
 - GATK 2.5-2,
 - vcftools 0.1.10, tabix 0.2.5,
 - ANNOVAR,
- All software is located in /usr/local directory

Day 5 : Galaxy



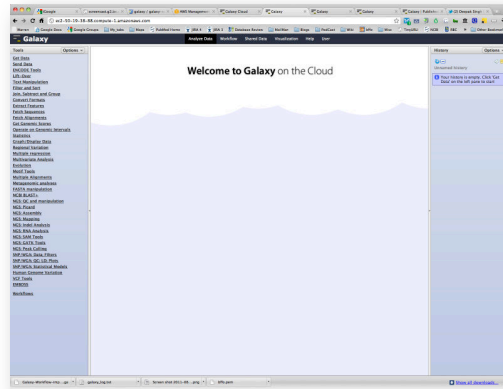
- We will load publically available data sets that reference GRCh37 (aka hg19)
- <http://aws.amazon.com/publicdatasets/>
 - Free available S3 buckets are 1000 Genomes, GenBank, Ensembl and others.

Day 5 : Galaxy



At this point ...

- If all is good, you have started your instance of Galaxy on AWS, and should look some thing like:



Day 5 : Galaxy



What is next?

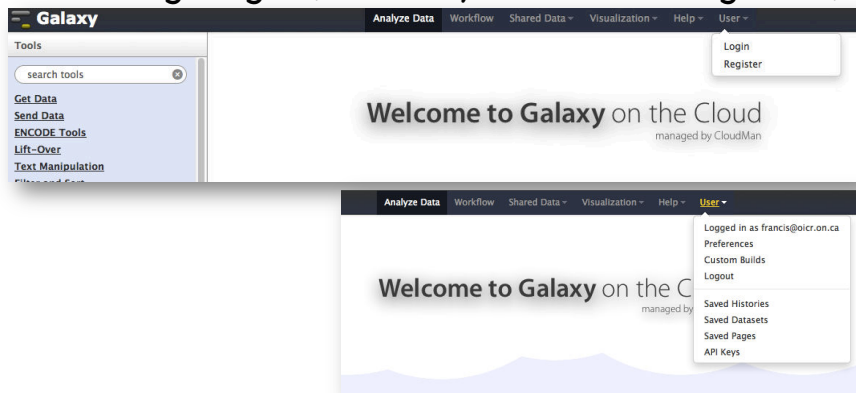
- I'm going to tell you about getting data in and out of Galaxy
- Doing operations in Galaxy
- Understanding the user interface.
- Linking multiple steps into "pipelines"
- Do an RNASeq mapping experiment
- Sharing pipeline with colleagues, and making them public.
- How to learn more ...

Day 5 : Galaxy



1st thing to do before we start:

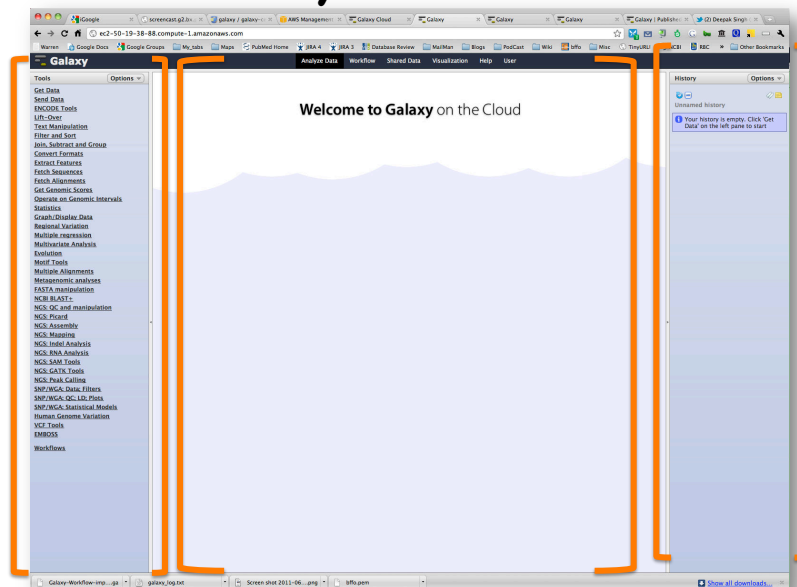
- This is important, irrespective of which cloud you are using: Login (1st time you need to "register")



Day 5 : Galaxy



Galaxy in the cloud



Day 5 : Galaxy



- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Assembly
- NGS: Mapping
- NGS: Indel Analysis
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: GATK Tools
- NGS: Peak Calling
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- Human Genome Variation
- VCF Tools

Day 5 : Galaxy



Galaxy cloud

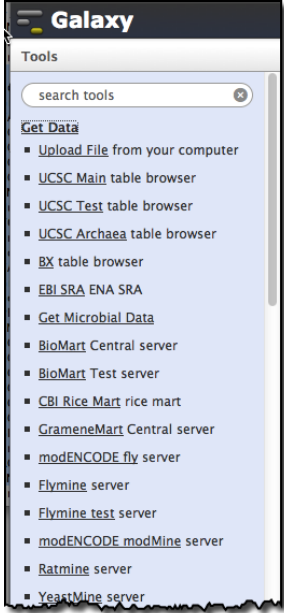
- < NGS: Assembly
- < NGS: GATK Tools
- < SNP/WGA: Statistical Models
- < Human Genome Variation
- < VCF Tools

usegalaxy.org

- > Genome Diversity
- > Phenotype Association
- > EMBOSS
- > NGS Toolbox Beta
- > NGS: GATK Tools (beta)
- > NGS: Variant Detection
- > NGS: Picard (beta)
- > BEDTools
- > snpEff
- > RGENETICS
- > SNP/WGA: Statistical Models


Day 5 : Galaxy



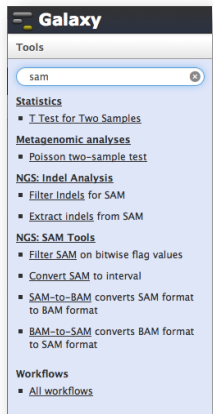
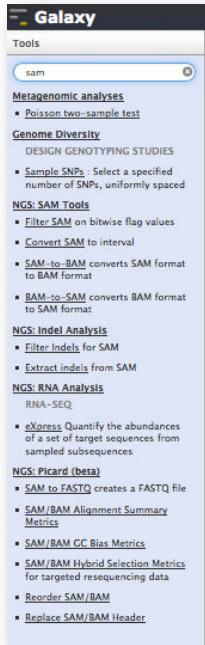


- ... and each item, when you click on it expands to lots more choices!
- What I find most useful when I know the name of the tool I'm looking for is to simply use the search tool.
- E.g. look for "sam"

Day 5 : Galaxy




usegalaxy.org

Galaxy cloud

Day 5 : Galaxy



UCSC Genome Browser: source of data for Galaxy

- Browse many Eukaryotic genomes (yeast to human)
- Most annotations are there
- Important evolutionary and variation data representation.
- Very flexible and configurable views
- Graphical and table views (Galaxy uses this)
- Upload your data into custom tracks and share with colleagues
- Client/server application with it's issues, but a great app!

Day 5 : Galaxy



<http://genome.ucsc.edu/>

The screenshot shows the UCSC Genome Browser homepage. The navigation menu on the left includes: Genomes, Blat, Tables, Gene Sorter, PCR, VistGene, Proteome, Session, FAQ, Help. The main content area has a header 'UCSC Genome Bioinformatics' and a navigation bar. Below the header, there is a 'Welcome to the UCSC Genome Browser website' message. A 'SCHEDULED MAINTENANCE' notice is visible, stating that the browser will be down on Sunday, August 1st, from 2:00 to 3:00 p.m. PDT. The 'News' section features a recent article: '23 July 2010 - BigBed/BigWig Paper Published' by Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. The article discusses the release of new data formats for BigBed and BigWig files.

Day 5 : Galaxy



Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#).
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade: Mammal | genome: Human | assembly: Feb. 2009 (GRCh37/hg19) | position or search term: chr17:41,243,452-41,277,500 | gene: | image width: 800 | submit

[Click here to reset](#) the browser user interface settings to their defaults. [Apply for free workshop](#)

[manage custom tracks](#) [configure tracks and display](#) [clear position](#)


About the Human Feb. 2009 (GRCh37/hg19) assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the [Genome Reference Consortium](#).


Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the [User's Guide](#) for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gI000212	Displays all of the unplaced contig gl000212
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061;RH80175	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, or STSs, etc.



Homo sapiens
(Graphic courtesy of CH3D)

Day 5 : Galaxy 

UCSC Genes

KRAS (uc001rgr.2) at chr12:25386769-25403863 - Homo sapiens Ras family small GTP binding protein K-Ras2 (KRAS) mRNA, complete cds.
 KRAS (uc001rqp.1) at chr12:25386769-25403863 - c-K-ras2 protein isoform b precursor
 KRAS (uc001rqp.1) at chr12:25386769-25403863 - c-K-ras2 protein isoform a precursor
 RASGEF1A (uc001jap.1) at chr19:43689984-43762367 - RasGEF domain family, member 1A
 RAP1GDS1 (uc003lhr.3) at chr4:99182527-99365010 - RAP1, GTP-GDP dissociation stimulator 1 isoform
 RAP1GDS1 (uc003lhr.3) at chr4:99182527-99365010 - RAP1, GTP-GDP dissociation stimulator 1 isoform
 RASBP2 (uc002wif.2) at chr20:4760670-4804291 - Ras association domain family 2
 RASBP2 (uc002wif.2) at chr20:4760670-4797869 - Ras association domain family 2
 RASBP4 (uc001jbo.2) at chr10:45455219-45490170 - Ras association domain family 4
 SRF2 (uc002uol.2) at chr21:182756472-182795462 - sperm specific antigen 2 isoform 1
 RIN1 (uc001ohh.1) at chr11:64499942-65104003 - ras inhibitor RIN1
 RASGRP2 (uc009yvpv.2) at chr11:64494384-64512928 - Ras guanyl releasing protein 2
 RASGRP2 (uc009yvpv.2) at chr11:64494384-64512329 - Ras guanyl releasing protein 2
 RASGRP2 (uc009yvpv.2) at chr11:64494384-64511630 - Ras guanyl releasing protein 2
 RASBP3 (uc001had.2) at chr11:206680879-206762615 - Ras association (RalGDS/AF-6) domain family 5
 MAPKAP1 (uc004bpv.2) at chr9:128199674-128469513 - mitogen-activated protein kinase associated

RefSeq Genes

KRAS at chr12:25386769-25403863 - (NM_033360) GTPase Kras isoform a precursor
 KRAS at chr12:25386769-25403863 - (NM_004985) GTPase Kras isoform b precursor


Non-Human RefSeq Genes

kras at chr11:15251234-15259784 - (NM_001080033) v-Ki-ras2 Kirsten rat sarcoma viral oncogene
 kras-b at chr11:15251234-15259784 - (NM_001080033) v-Ki-ras2 Kirsten rat sarcoma viral oncogene
 kras-d at chr11:15251234-15259784 - (NM_001080033) v-Ki-ras2 Kirsten rat sarcoma viral oncogene
 KRAS at chr12:25386769-25403863 - (NM_001029981) GTPase Kras precursor
 KRAS at chr12:25386769-25403863 - (NM_001029981) GTPase Kras
 Kras at chr12:25386769-25403863 - (NM_001029981) GTPase Kras
 KRAS at chr12:25386769-25403863 - (NM_001100001) GTPase Kras
 Kras-a at chr12:25386769-25403863 - (NM_001100001) GTPase Kras precursor
 spps at chr12:25386769-25403863 - (NM_001095100) sarcospan (Kras oncogene-associated gene)

Human Aligned mRNA Search Results

BC062299 - Homo sapiens sarcospan (Kras oncogene-associated gene), mRNA (cDNA clone MGC:71179 IMAGE:6376868), complete cds.
 AF193917 - Homo sapiens Ras family small GTP binding protein K-Ras2 (KRAS) mRNA, complete cds.
 BC013572 - Homo sapiens v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog, mRNA (cDNA clone MGC:8977 IMAGE:3878884), complete cds.
 M54968 - Homo sapiens K-ras oncogene protein (KRAS) mRNA, complete cds.
 DQ975649 - Synthetic construct clone IMAGE:100002879; FL164775.01X; RFPD0839B1015B0 v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
 DQ975649 - Synthetic construct clone IMAGE:100002879; FL164775.01X; RFPD0839B1015B0 v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
 AK292510 - Homo sapiens cDNA FLJ76682 complete cds, highly similar to Homo sapiens v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
 AK314353 - Homo sapiens cDNA, FLJ95117, highly similar to Homo sapiens sarcospan (Kras oncogene-associated gene) (SPFN), mRNA.

Non-Human Aligned mRNA Search Results

Day 5 : Galaxy 

Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr12:25,386,769-25,403,863 gene jump clear size 17,095 bp. configure [Apply for free workshop](#)

Scale chr12: 25395000 5 kb 25395000 25400000

UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics

RefSeq Genes

Human mRNAs From GenBank

Spliced ESTs

RepeatMasker

move start Click on a feature for details. Click or drag in the base position track to zoom in. move end

Day 5 : Galaxy

Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr12:25,386,769-25,403,863 gene jump clear size 17,095 bp. configure [Apply for free workshop](#)

UCSC Genes

Human mRNAs

Spliced ESTs

RepeatMasker

Phenotype and Disease Associations

Genes and Gene Prediction Tracks

mRNA and EST Tracks

Expression

Regulation

Comparative Genomics

Variation and Repeat

Day 5 : Galaxy

Other Examples of Data Format outputs from UCSC:

- Tab-separated
- Sequence (FASTA)
- Browser Extensible Data format (BED)
- General Feature Format (GFF)
- Gene Transfer Format (GTF)

Day 5 : Galaxy



Examples of Data Formats for UCSC:

- Sequence (FASTA):

```
>gi|89058412|ref|NT_028395.3| Homo sapiens chromosome 22 genomic contig, GRCh37.p5
Primary Assembly
GATCTGATAAGTCCAGGACTTCAGAAGAGCTGTGAGACCTTGGCCAAAGTCACTTCCTCCTCAGGAACA
TTGCAGTGGGCTAAGTGCCTCCTCGGACTGGTATGGGGACGGTCATGCAATCTGGACAACATTCAC
CTTTAAAAGTTTATTTGATCTTTGTGACATGCACGTGGGTTCCAGTAGCAAGAACTAAGGGTCCGAC
GCCGGTTTCGTAAATTTCTTAATTCCAAGACAGTCTCAAAATTTTCTTATTAACCTCCTGGAGGGAG
GCTTATCATTCTCTCTTTTGGATGATTC TAAGTACCAGTAAAATACAGCTATCATTCTTTCTTTGAT
TTGGGAGCCTAATTTCTTTAATTTAGTATGCAAGAAAACCAATTTGGAAATATCAACTGTTTGGAAACC
TTAGACCTAGGTCATCCTTAGTAAGATCTTCCCATTTATATAAACTACTGCAAGTAGTAGTGCATAAAT
ACCAAACATAAAGCCAACCTGAGATGCCAAAGGGGGCCACTCTCCTTGTCTTTTCTCCTTTTTAGAGGAT
TTATTTCCCATTTTCTTAAAAAGGAAGAACAACTGTGCCCTAGGGTTACTGTGTCAGAACAGAGTGT
GCCGATTTGGTTCAGGACTCCATAGCATTTCACCATTGAGTTATTTCCGCCCTTACGTGTCTCTCTTC
AGCGGTCTATTATCTCCAAGAGGGCATAAAACACTGAGTAAACAGCTCTTTTATATGTCTTCTGGATG
AGCCTCTTTTAAATTAATTTTGTAAAGGATTTCTCTAGGGCCACTGCACGTATGGGGAGTCACCCCC
AGACACTCCCAATTTGGCCCTTGTCAACCAGGGGCACATTTTCAGCTATTTGTAAAACCTGAAATCACTAG
AAAGGAATGCTAGTACTTGTGGGGCCAAAGCCCTTGTATGGGGATGAAGCTCTTAGTGGTAGCC
CTCCAAGAGAATAGATGGTGAATGTCTCTTTTCAGACATTAAGGTGTCAGACTCTCAGTTAATCTCTCC
TAGATCCAGGAAAGCCTAGAAAAGGAAGCCCTGACTGCATTAATGGAGATTTCTCCATGTGCAAAAT
TCTCCCAAAAAGAAATCCTGTGAGGGCCATTTAATGTGTGGCCCTGTGACAGCCATTTCAAATATG
TCAAAAATATATTTTGGAGTAAAATACTTTCATTTTCTTCCAGAGTCTGTGTCGTATGATGCCATACC
AGAGTCAGGTTGGAAAGTAAGCCACATTTATACAGGCTTAACCTAAAAAACAAAAACTGTCTAACAGA
TTTTATGGTTTATAGACATGATTTCCCGGACACATTAGATAGAAATCTGGGCAAGAGAAGAAAAAAGG
```

Day 5 : Galaxy



Browser Extensible Data format (BED)

```
track name=pairedReads description="Clone Paired Reads" useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 1.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is numbered 1. The *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the score value will determine the level of gray in which this feature is displayed (darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray.

shade	score in range
	≤ 166
	167-277
	278-388
	389-499
	500-611
	612-722
	723-833
	834-944
	≥ 945

<http://goo.gl/agfWu>

Day 5 : Galaxy



General Feature Format (GFF)

GFF format

GFF (General Feature Format) lines are based on the GFF standard file format. GFF lines have nine required fields that *must* be tab-separated. If the fields are separated by spaces instead of tabs, the more information on GFF format, refer to <http://www.sanger.ac.uk/resources/software/gff/>.

If you would like to obtain browser data in GFF (GTF) format, please refer to [Genes in gtf or gff format](#) on the Wiki.

Here is a brief description of the GFF fields:

1. **seqname** - The name of the sequence. Must be a chromosome or scaffold.
2. **source** - The program that generated this feature.
3. **feature** - The name of this type of feature. Some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
4. **start** - The starting position of the feature in the sequence. The first base is numbered 1.
5. **end** - The ending position of the feature (inclusive).
6. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the score value will determine the level of gray in which this feature is displayed (darker gray). If there is no score value, enter ".".
7. **strand** - Valid entries include '+', '-', or '.' (for don't know/don't care).
8. **frame** - If the feature is a coding exon, *frame* should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
9. **group** - All lines with the same group are linked together into a single item.

Example:

Here's an example of a GFF-based track. This [example](#) can be pasted into the browser without editing. NOTE: Paste operations on some operating systems will replace tabs with spaces, which will track is uploaded. You can circumvent this problem by pasting the URL of the above example (<http://genome.ucsc.edu/goldenPath/help/regulatory.txt>) instead of the text itself into the custom annotation track. If you encounter an error when loading a GFF track, check that the data lines contain tabs rather than spaces.

```
browser position chr22:10000000-10025000
browser hide all
track name=regulatory description="TeleGene(tm) Regulatory Regions"
visibility=2
chr22 TeleGene enhancer 10000000 10001000 500 + . touch1
chr22 TeleGene promoter 10010000 10010100 900 + . touch1
chr22 TeleGene promoter 10020000 10025000 800 - . touch2
```

Click [here](#) to display this track in the Genome Browser.

<http://goo.gl/agfWu>

Day 5 : Galaxy



Gene Transfer Format (GTF)

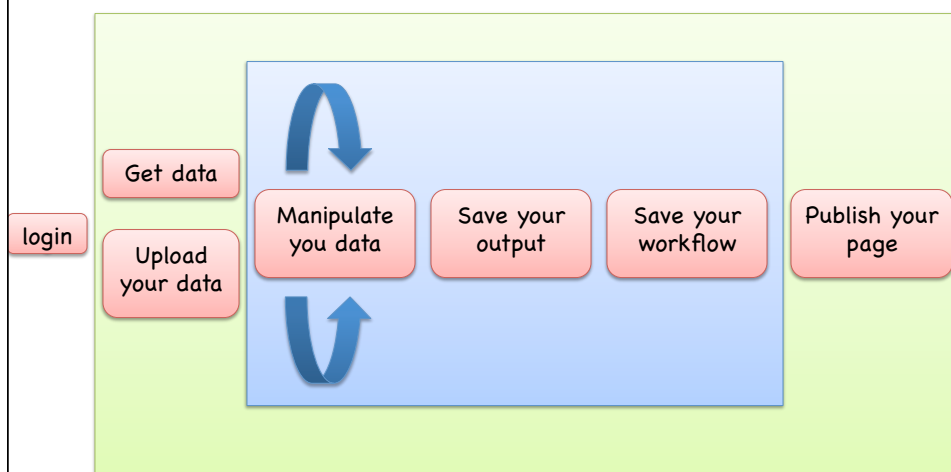
- Like GFF, but specific to exon and CDS features, and has one extra field:

The attribute list must begin with the two mandatory attributes:

- **gene_id value** - A globally unique identifier for the genomic source of the sequence.
- **transcript_id value** - A globally unique identifier for the predicted transcript.

```
gene_id "Em:U62317.C22.6.mRNA"; transcript_id "Em:U62317.C22.6.mRNA"; exon_number 1
```

General workflow for Galaxy



Pages in Galaxy

- https://usegalaxy.org/page/list_published

Title	Annotation	Owner	Community Rating	Community Tags	Last Updated
AR divergence states	This page contains datasets for the following paper: "Segmenting the human genome based on states of neutral genetic divergence" Proc Natl Acad Sci U S A...	guru	★★★★★		~ 22 hours ago
Interactive RNA-seq with Trackster	Trackster is Galaxy's integrated visual analysis environment. This page describes how Trackster was used to perform interactive RNA-seq using...	jeremy	★★★★★		Sep 18, 2013
Screencasts usegalaxy.org	Screencasts	galaxyproject	★★★★★		Jun 28, 2013
SNP classification	SNP classification workflow and history for Mutation Detection 2013	Belinda	★★★★★		Apr 19, 2013
Using Galaxy 2012	Supplemental information for "Using Galaxy to Perform Large-Scale Interactive Data Analysis" paper in Current	galaxyproject	★★★★★	chip-seq snp maf tutorial interval	Mar 27, 2013

Day 5 : Galaxy



[Galaxy RNA-seq Analysis Exercise](#)

An exercise that illustrates how to use Galaxy for RNA-seq analyses.

jeremy



[Interactive RNA-seq with Trackster](#)

Trackster is Galaxy's integrated visual analysis environment. This page describes how Trackster was used to perform interactive RNA-seq using...

jeremy



[ChIP seq exercise](#)

Small ChIP-seq analysis using reduced data from the Hardison Lab / Mouse ENCODE

james



Day 5 : Galaxy



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

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 0%

Published Pages | [jeremy](#) | Galaxy RNA-seq Analysis Exercise

RNA-seq Analysis Exercise

Galaxy provides the tools necessary to creating and executing a complete RNA-seq analysis pipeline. This exercise introduces these tools and guides you through a simple pipeline using some example datasets. Familiarity with Galaxy and the general concepts of RNA-seq analysis are useful for understanding this exercise. This exercise should take 1-2 hours. You can check your work by looking at the history and visualization at the bottom of this page, which contain the datasets for the completed exercise.

Input Datasets

Below are small samples of datasets from [the Illumina BodyMap 2.0 project](#): specifically, the datasets are paired-end 50bp reads from adrenal and brain tissues. The sampled reads map mostly to a 500Kb region of chromosome 19, positions 3-3.5 million (chr19:3000000-3500000).

RNA-seq data from adrenal tissue:

- Galaxy Dataset | adrenal_1.fastq
Forward RNA-seq reads from BodyMap 2.0 project, adrenal tissue, mapping to chr19:3000000-3500000

and

- Galaxy Dataset | adrenal_2.fastq
Reverse RNA-seq reads from BodyMap 2.0 project, adrenal tissue, mapping to chr19:3000000-3500000

RNA-seq data from brain tissue:

- Galaxy Dataset | brain_1.fastq

About this Page

Author
jeremy

Related Pages
[All published pages](#)
[Published pages by jeremy](#)

Rating
Community (69 ratings, 4.9 average)
Yours: ★★★★★

Tags
Community: rna-seq tutorial rna
Yours:

Day 5 : Galaxy



RNA-Seq Analysis Exercise

- Human BodyMap 2.0 data from Illumina.
- **paired-end** 50bp reads from **adrenal** and **brain** tissues. The sampled reads map mostly to a 500Kb region of chromosome 19, positions 3-3.5 million (chr19:3000000-3500000).

Ensembl Blog
News about the Ensembl Project and its genome browser

Home Workspaces Future Plans About Us Contact Us ensembl

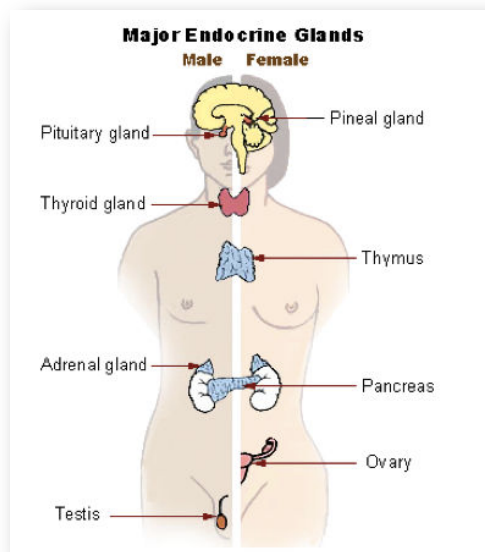
Human BodyMap 2.0 data from Illumina

Posted on May 24, 2011 by Tibault (Ensembl) — 33 Comments

I'd like to introduce you an exciting new data set that we've introduced in Ensembl release 62: RNA-seq data from Illumina's Human BodyMap 2.0 project. The data, generated on HiSeq 2000 instruments in 2010, consist of 16 human tissue types, including adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testis, thyroid, and white blood

Day 5 : Galaxy





http://en.wikipedia.org/wiki/Adrenal_gland

Day 5 : Galaxy



Getting data

- Most of time, you will get from a file on your computer, or from a URL.

Day 5 : Galaxy



Get 5 files

- adrenal_1
https://usegalaxy.org/dataset/display?dataset_id=d44d2a324474d1aa&to_ext=fastqsanger
- adrenal_2
https://usegalaxy.org/dataset/display?dataset_id=d08360a1c0ffdc62&to_ext=fastqsanger
- brain_1
https://usegalaxy.org/dataset/display?dataset_id=f187acb8015d6c7f&to_ext=fastqsanger
- brain_2
https://usegalaxy.org/dataset/display?dataset_id=08c45996966d7ded&to_ext=fastqsanger
- Annotations from this part of Chr 19
https://usegalaxy.org/dataset/display?dataset_id=f09f4a0bd8a19fd6&to_ext=gtf

Day 5 : Galaxy



Load file(s) to Galaxy

The screenshot shows the Galaxy web interface with the 'Upload File (version 1.1.3)' tool selected. The 'File Format' dropdown is set to 'Auto-detect'. The 'URL/Text' field is empty. The 'Execute' button is visible at the bottom. Red arrows point to the 'File Format' dropdown, the 'URL/Text' field, and the 'Execute' button.

autodetect

Paste URL from previous page


Execute

Day 5 : Galaxy



The screenshot shows three Galaxy history panels connected by red arrows, illustrating a workflow. The first panel shows step 12: FASTQ Groomer on data 1 and step 11: FastQC brain 2.fastqsanger.html. The second panel shows step 12: FASTQ Groomer on data 1, step 11: FastQC brain 2.fastqsanger.html, and step 10: FastQC brain 1.fastqsanger.html. The third panel shows step 12: FASTQ Groomer on data 1, step 11: FastQC brain 2.fastqsanger.html, and step 10: FastQC brain 1.fastqsanger.html. Below these panels is a list of data sources:

- 9: ftp://ftp.sra.ebi.ac.uk /vol1/fastq/SRR031/SRR031720/SRR031720.fastq
- 8: ftp://ftp.sra.ebi.ac.uk /vol1/fastq/SRR031/SRR031719/SRR031719.fastq.gz
- 7: ftp://ftp.sra.ebi.ac.uk /vol1/fastq/SRR031/SRR031718/SRR031718.fastq.gz
- 6: ftp://ftp.sra.ebi.ac.uk /vol1/fastq/SRR031/SRR031721/SRR031721.fastq.gz

Day 5 : Galaxy 


The screenshot shows four history items in a list:

- 7: brain 1.fastqsanger
- 4: brain 2.fastqsanger
- 2: adrenal 2.fastqsanger
- 1: adrenal 1.fastqsanger

Annotations to the right of the list explain the icons:

- "Poke the eye" (eye icon)
- "Edit attribute" (pencil icon)
- "Delete" (X icon)

A red arrow points to the first item with the text: "Numbers may vary with usage"

Day 5 : Galaxy 

"poke the eye"

Analyze Data Workflow Shared Data Visualization

Warning: This dataset is large and only the first megabyte is shown below.
[Show all](#) | [Save](#)

```

@ERR030882.1482 HWI-BRUNOP16X_0001:3:1:16997:4347#0/1
NNCAAATACAGATGAGGGTACTAAAGTGTCTTGGTTTTTATTTATTTAT
+
#####
@ERR030882.2595 HWI-BRUNOP16X_0001:3:1:6649:5175#0/1
NNCACATCTTTATTGGAAAGGCACAGCTAAGCCCACCTTGATACAGCMT
+
#####
@ERR030882.5778 HWI-BRUNOP16X_0001:3:1:17645:6013#0/1
NNCCCTCTCAATGGCTCCCAAGACCTGCTGCTGCTCTTGGGAGAGGT
+
#####
@ERR030882.5894 HWI-BRUNOP16X_0001:3:1:19961:6042#0/1
NNTTTATTTATTTATTTATTTTCTTTTCCAGTATACAGCTTGTCT
+
#####
@ERR030882.7647 HWI-BRUNOP16X_0001:3:1:5088:6574#0/1
NNGACTCTCGGACCGCATCAAGACGAATTCAGCTACTGCAAGCTCAG
+
#####
@ERR030882.31490 HWI-BRUNOP16X_0001:3:1:9673:21095#0/1
NNTTTGACGATCAGCCGTGTTGTGCATCGATGTTCAAGCCGTAGGA
+
#####
    
```

"Edit attributes"

Attributes Convert Format Datatype Permissions

Edit Attributes

Name:

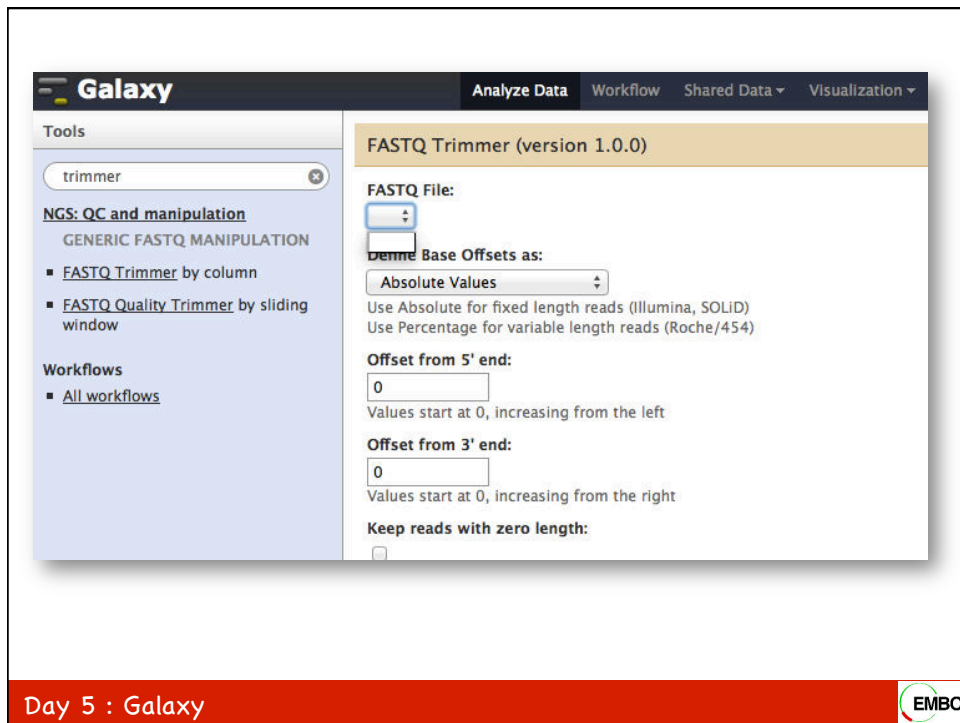
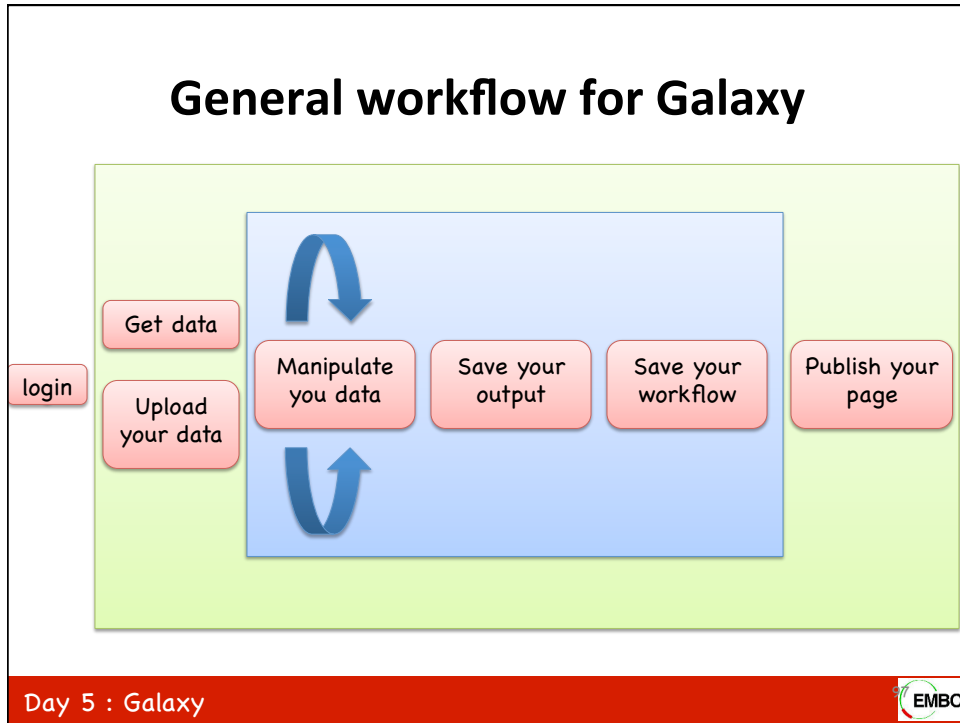
Info:

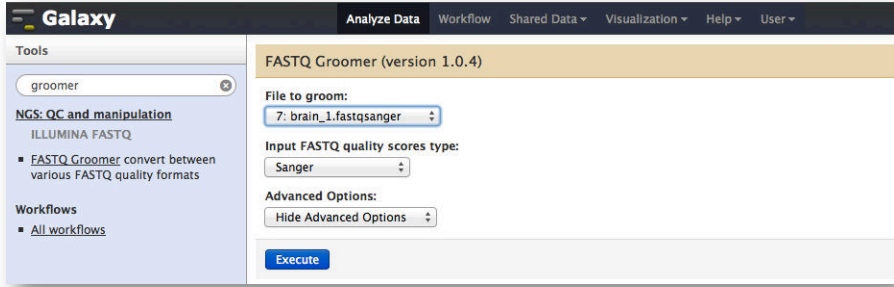
Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available when a h

Database/Build:

This will inspect the dataset and attempt to correct the above column values






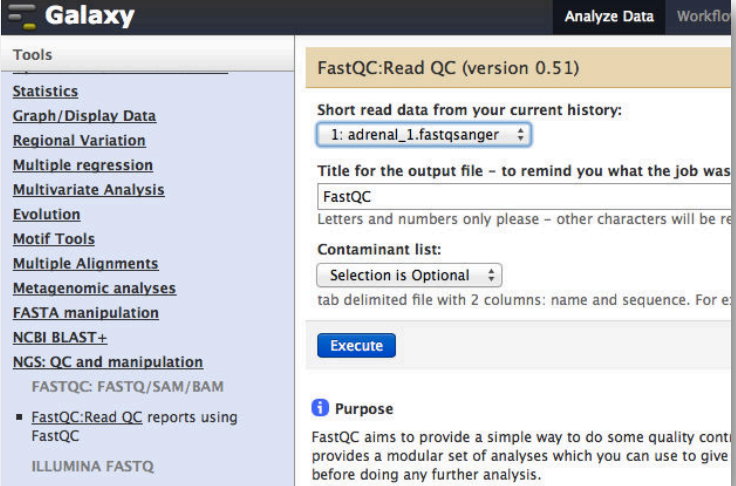
The screenshot shows the Galaxy web interface with the FASTQ Groomer tool (version 1.0.4) selected. The tool configuration includes:

- File to groom:** 7: brain_1.fastqsanger
- Input FASTQ quality scores type:** Sanger
- Advanced Options:** Hide Advanced Options
- Execute** button

The left sidebar shows the 'Tools' menu with 'NGS: QC and manipulation' selected, and 'FASTQ Groomer' listed under 'ILLUMINA FASTQ'.

Day 5 : Galaxy 


Step 1: quality control [NGS: QC and manipulation >] FASTQC tool

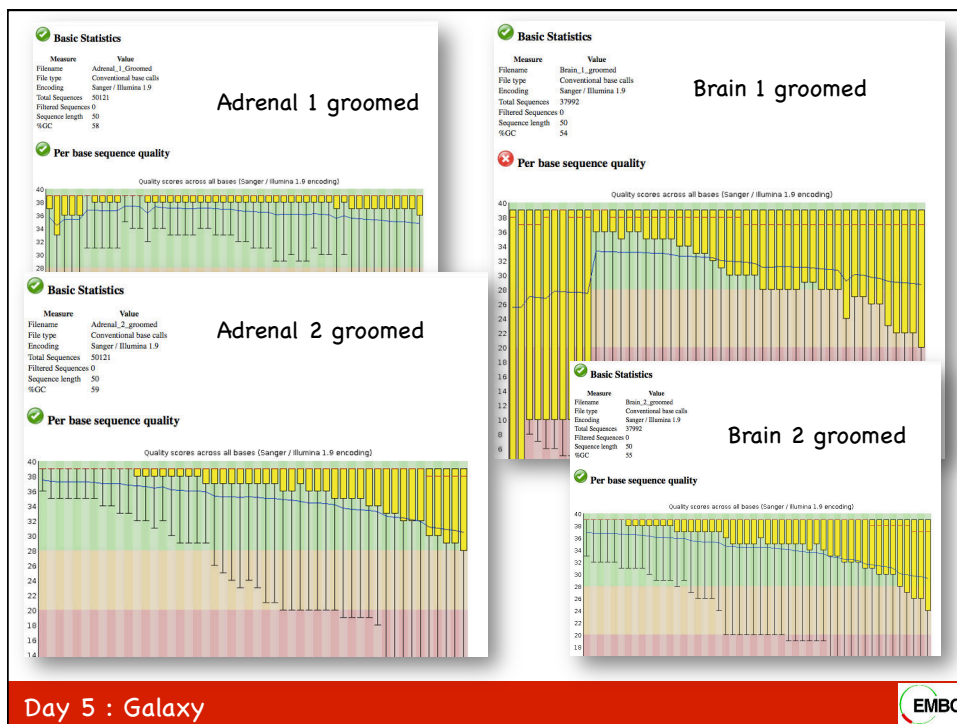


The screenshot shows the Galaxy web interface with the FastQC:Read QC tool (version 0.51) selected. The tool configuration includes:

- Short read data from your current history:** 1: adrenal_1.fastqsanger
- Title for the output file - to remind you what the job was:** FastQC
- Contaminant list:** Selection is Optional
- Execute** button
- Purpose:** FastQC aims to provide a simple way to do some quality control. It provides a modular set of analyses which you can use to give a quick overview of the quality of your data before doing any further analysis.

The left sidebar shows the 'Tools' menu with 'NGS: QC and manipulation' selected, and 'FastQC:Read QC' listed under 'FASTQ: FASTQ/SAM/BAM'.


Day 5 : Galaxy 



Need to remove bad bases in reads?

- Assume a median quality score of below 20 to be unusable.
- Given this criterion, is trimming needed for the datasets?
- If so, which base pairs should be trimmed?
- [NGS: QC and manipulation >] FASTQ Trimmer

↑ "Numbers may vary with usage"

Day 5 : Galaxy 

[NGS: RNA Analysis >] Tophat tool

- Step 1
- Use the [NGS: RNA Analysis >] Tophat tool
- To map RNA-seq reads to the hg19 Canonical Female build.
- Because the reads are paired, you'll need to set mean inner distance between pairs; this is the average distance in base pairs between reads, not the total insert/fragment size.
- Use a mean inner distance of 110 for BodyMap data.

Tophat for Illumina (version 1.5.0)

RNA-Seq FASTQ file:
 ←

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Use a built in reference genome or own from your history:
 ←

Built-ins genomes were created using default options

Select a reference genome:
 ←

If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:
 ←

RNA-Seq FASTQ file:
 ←

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs:
 ←

TopHat settings to use:
 ←

Use the Full parameter list to change default settings.

←

Day 5 : Galaxy

History

RNASeq
66.6 MB

- 27: Tophat for Illumina on data 15 and data 14: accepted hits
- 26: Tophat for Illumina on data 15 and data 14: splice junctions
- 25: Tophat for Illumina on data 15 and data 14: deletions
- 24: Tophat for Illumina on data 15 and data 14: insertions
- 23: Tophat for Illumina on data 13 and data 12: accepted hits
- 22: Tophat for Illumina on data 13 and data 12: splice junctions
- 21: Tophat for Illumina on data 13 and data 12: deletions
- 20: Tophat for Illumina on data 13 and data 12: insertions
- 19: FastQC Adrenal 2

I was hoping this would be lunch time now ...
~ 30 minutes ...

Day 5 : Galaxy

27: Brain TopHat accepted_hits

26: Brain TopHat splice junctions

25: Brain TopHat deletions

24: Brain TopHat Insertion

23: Adrenal TopHat hits

22: Adrenal TopHat splice junction

21: Adrenal TopHat Deletion

20: Adrenal TopHat insertions

Select datasets for new tracks

History: RNaseq

search name and filepath

<input type="checkbox"/>	ID	Name	Filetype	Library
<input type="checkbox"/>	28	Genomes UCSC hg19, chr19 gene annotation	gff	hg19
<input type="checkbox"/>	27	Brain TopHat accepted_hits	bam	hg19
<input type="checkbox"/>	26	Brain TopHat splice junctions	bed	hg19
<input type="checkbox"/>	25	Brain TopHat deletions	bed	hg19
<input type="checkbox"/>	24	Brain TopHat Insertion	bed	hg19
<input type="checkbox"/>	23	Adrenal TopHat hits	bam	hg19
<input type="checkbox"/>	22	Adrenal TopHat splice junction	bed	hg19
<input type="checkbox"/>	21	Adrenal TopHat Deletion	bed	hg19
<input type="checkbox"/>	20	Adrenal TopHat insertions	bed	hg19

For 0 selected datasets: Cancel Add

Day 5 : Galaxy

chr19 3,450,000 3,436,969 - 3,494,741 3,470,000

Day 5 : Galaxy

[NGS: RNA Analysis >] Cufflinks

- Remember to include reference annotations
- Do Adrenal 1st, then Brain

Cufflinks (version 0.0.5)

SAM or BAM file of aligned RNA-Seq reads:
23: Adrenal TopHat hits

Max Intron Length:
300000

Min Isoform Fraction:
0.1

Pre mRNA Fraction:
0.15

Perform quartile normalization:
No

Removes top 25% of genes from FPKM denominator to improve accuracy.

Use Reference Annotation:
Use reference annotation

Reference Annotation:
28: iGenomes UCSC hg19, chr19 gene annotation
Gene annotation dataset in GTF or GFF3 format.

Perform Bias Correction:
No

Bias detection and correction can significantly improve accuracy of FPKM values.

Use multi-read correct:
No

Tells Cufflinks to do an initial estimation procedure to more accurately estimate transcript abundances.

Execute

Day 5 : Galaxy



[NGS: RNA Analysis >] Cuffdiff

[NGS: RNA Analysis >] Cuffcompare

Day 5 : Galaxy



sharing

Saved Histories

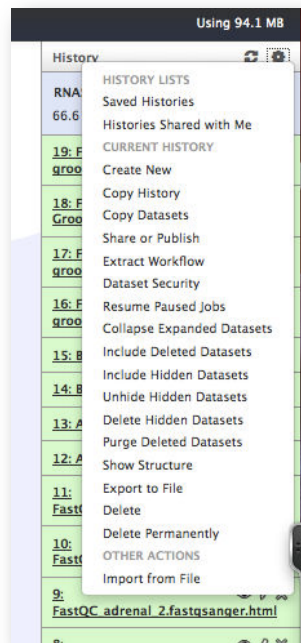
[Advanced Search](#)

<input type="checkbox"/>	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
<input type="checkbox"/>	Unnamed history		0 Tags		0 bytes	21 minutes ago	21 minutes ago	
<input type="checkbox"/>	SNP detection on Chr22	7	0 Tags		1.3 GB	~ 3 hours ago	30 minutes ago	current history
<input type="checkbox"/>	1pass		0 Tags		1.2 GB	~ 17 hours ago	~ 3 hours ago	
<input type="checkbox"/>	Unnamed		0 Tags		0 bytes	~ 11 hours ago	~ 11 hours ago	

For 0 selected items:

Histories that have been deleted for more than a time period specified by the Galaxy administrator(s) may be permanently deleted.

Day 5 : Galaxy



- Share history with neighbor
- Extract workflow

Day 5 : Galaxy




Analyze Data Workflow Shared Data Visualization Help User

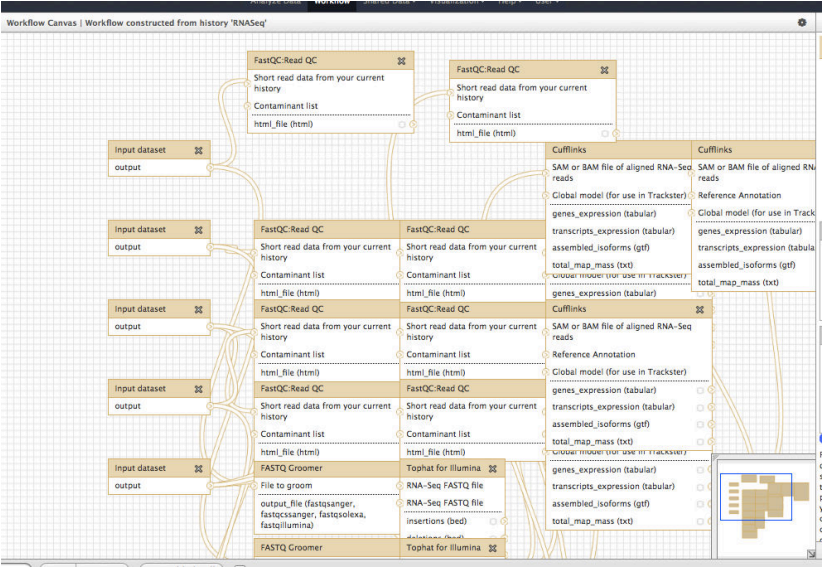
The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow. Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

Workflow name
Workflow constructed from history 'RNASeq'
Create Workflow Check all Uncheck all


Tool	History items created
Upload File <i>This tool cannot be used in workflows</i>	1: adrenal_1.fastqsanger <input checked="" type="checkbox"/> Treat as input dataset
Upload File <i>This tool cannot be used in workflows</i>	2: adrenal_2.fastqsanger <input checked="" type="checkbox"/> Treat as input dataset
Upload File <i>This tool cannot be used in workflows</i>	4: brain_2.fastqsanger <input checked="" type="checkbox"/> Treat as input dataset
Upload File <i>This tool cannot be used in workflows</i>	7: brain_1.fastqsanger <input checked="" type="checkbox"/> Treat as input dataset
FastQC:Read QC <input checked="" type="checkbox"/> Include "FastQC:Read QC" in workflow	8: FastQC_adrenal_1.fastqsanger.html
FastQC:Read QC <input checked="" type="checkbox"/> Include "FastQC:Read QC" in workflow	9: FastQC_adrenal_2.fastqsanger.html
FastQC:Read QC <input checked="" type="checkbox"/> Include "FastQC:Read QC" in workflow	10: FastQC_brain_1.fastqsanger.html

Day 5 : Galaxy 

Workflow Canvas | Workflow constructed from history 'RNASeq'

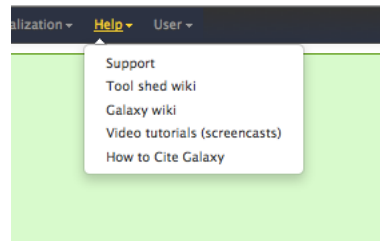


The diagram illustrates a multi-step bioinformatics pipeline. It starts with multiple 'Input dataset' nodes on the left. These feed into several 'FastQC:Read QC' tools, which then connect to 'FASTQ Croomer' and 'Tophat for Illumina' tools. The outputs of these tools are then processed by 'Cufflinks' tools, which generate various output files such as 'genes_expression (tabular)', 'transcripts_expression (tabular)', 'assembled_isoforms (gtf)', and 'total_map_mass (txt)'. The workflow is represented as a network of nodes and connecting lines on a grid background.

Day 5 : Galaxy 

Remember, lots of tutorials, videos, mailing list, twitter etc ...

- <https://vimeo.com/galaxyproject>



Day 5 : Galaxy



<https://vimeo.com/galaxyproject>

A screenshot of the Galaxy Project channel page on Vimeo. The page header shows the Vimeo logo and navigation options like 'Me', 'Videos', 'Create', 'Watch', 'Tools', and 'Upload'. The channel name 'Galaxy Project' is displayed with a 'PLUS' badge and the text 'Joined 1 month ago'. Below this, statistics are shown: 54 Videos, 0 Likes, 0 Following, 1 Group, 6 Channels, and 0 Albums. A 'Recently Uploaded' section features three video thumbnails, each titled 'Using Galaxy' with a subtitle: 'Calling Peaks For ChIP-seq Data', 'Loading Data and Understanding Datatypes', and 'Finding Human Coding Exons with Highest SNP Density'. A 'NEED HELP?' link is visible at the bottom left of the channel page.

Day 5 : Galaxy



The screenshot shows a Vimeo video player interface. The video title is "Using Galaxy protocol 3 Calling Peaks For ChIP-seq Data". The video player includes a search bar, navigation links (Me, Videos, Create, Watch, Tools, Upload), and social sharing options (LIKE, LATER, SHARE). The video progress bar shows 09:25. The EMBO logo is visible in the bottom right corner of the slide.

Day 5 : Galaxy



<https://usegalaxy.org/u/jeremy/p/interactive-rna-seq-with-trackster>

The screenshot shows a Galaxy web page titled "Interactive RNA-seq analyses by visualization with Trackster". The page includes a navigation bar with links like "Analyze Data", "Workflow", "Shared Data", "Visualization", "Cloud", "Help", and "User". The main content area contains the title, author information (Jeremy Goecks, The Galaxy Team, Anton Nekrutenko, and James Taylor), and a description of the interactive supplement. The right sidebar shows "About this Page" with author details, related pages, and a rating system.

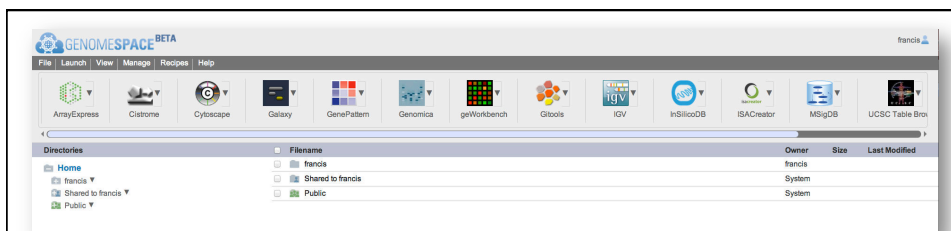
Day 5 : Galaxy



<http://genomespace.org/>



Day 5 : Galaxy



ArrayExpress: <http://www.ebi.ac.uk/arrayexpress/>
 Cistrome: <http://www.cistrome.org>
 Cytoscape: <http://www.cytoscape.org/>
 Galaxy: <http://usegalaxy.org>
 GenePattern: <http://www.broadinstitute.org/cancer/software/genepattern/>
 Genomica: <http://genomica.weizmann.ac.il/>
 geWorkbench: <http://www.geworkbench.org>
 Gitools: <http://www.gitools.org/>
 IGV: <http://www.broadinstitute.org/igv/>
 InSilico DB: <https://insilico.ulb.ac.be/>
 ISACreator: <http://isatab.sourceforge.net/tools.html>
 MSigDB: <http://www.broadinstitute.org/gsea/msigdb/>
 UCSC GB: <http://genome.ucsc.edu/>

Day 5 : Galaxy



Useful Resources

- Galaxy
 - usegalaxy.org and usegalaxy.org/cloud
 - Twitter: @galaxyproject #usegalaxy
 - User's mailing list:
<http://lists.bx.psu.edu/listinfo/galaxy-user>



- BioStaR
 - biostars.org
 - Twitter: @biostarquestion

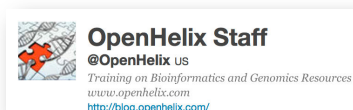


Day 5 : Galaxy

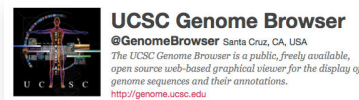


Useful Resources

- OpenHelix
 - <http://www.openhelix.com/>
 - Twitter: @openhelix
 - Blog: <http://blog.openhelix.com/>



- UCSC
 - <http://genome.ucsc.edu/>
 - Twitter: @GenomeBrowser
 - More tutorials: <http://genome.ucsc.edu/training.html>



- SEQanswers
 - Forum for NGS technologies
<http://seqanswers.com/>



Day 5 : Galaxy



Papers that should be of interest

- Robert Gentleman, 2005, Reproducible research: a bioinformatics case, Stat Appl Genet Mol Biol. 2005;4:Article2.
<http://www.ncbi.nlm.nih.gov/pubmed/?term=16646837>
- Goecks J, Nekrutenko A, Taylor J; Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 2010, 11:R86
<http://www.ncbi.nlm.nih.gov/pubmed/?term=20738864>
- Afgan E, Chapman B, Jadan M, Franke V, Taylor J. (2012) Using cloud computing infrastructure with CloudBioLinux, CloudMan, and Galaxy. Curr Protoc Bioinformatics. Chapter 11:Unit11.9. doi: 10.1002/0471250953.bi1109s38.
<http://www.ncbi.nlm.nih.gov/pubmed/22700313>

Center for Informatics and Computing, Ruder Bošković Institut
 Day 5 : Galaxy



From @bffb on twitter:



Day 5 : Galaxy



OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve^{1,2*}, Anton Nekrutenko³, James Taylor⁴, Eivind Hovig^{1,5,6}

1 Department of Informatics, University of Oslo, Blindern, Oslo, Norway, 2 Centre for Cancer Biomedicine, University of Oslo, Blindern, Oslo, Norway, 3 Department of Biochemistry and Molecular Biology and The Huck Institutes for the Life Sciences, Penn State University, University Park, Pennsylvania, United States of America, 4 Department of Biology and Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia, United States of America, 5 Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway, 6 Institute for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway

Replication is the cornerstone of a cumulative science [1]. However, new tools and technologies, massive amounts of data, interdisciplinary approaches, and the complexity of the questions being asked are complicating replication efforts, as are increased pressures on scientists to advance their research [2]. As full replication of studies on independently collected data is often not feasible, there has recently been a call for reproducible research as an attainable minimum standard for assessing the value of scientific claims [3]. This requires that papers in experimental science describe the results and provide a sufficiently clear protocol to allow successful repetition and extension of analyses based on original data [4].

The importance of replication and reproducibility has recently been exemplified through studies showing that scientific papers commonly leave out experimental details essential for reproduction [5], studies showing difficulties with replicating published experimental results [6], an increase in retracted papers [7], and

We further note that reproducibility is just as much about the habits that ensure reproducible research as the technologies that can make these processes efficient and realistic. Each of the following ten rules captures a specific aspect of reproducibility, and discusses what is needed in terms of information handling and tracking of procedures. If you are taking a bare-bones approach to bioinformatics analysis, i.e., running various custom scripts from the command line, you will probably need to handle each rule explicitly. If you are instead performing your analyses through an integrated framework (such as GenePattern [10], Galaxy [11], LONI pipeline [12], or Taverna [13]), the system may already provide full or partial support for most of the rules. What is needed on your part is then merely the knowledge of how to exploit these existing possibilities.

In a pragmatic setting, with publication pressure and deadlines, one may face the need to make a trade-off between the ideals of reproducibility and the need to get the research out while it is still relevant.

than to do it while underway). We believe that the rewards of reproducibility will compensate for the risk of having spent valuable time developing an annotated catalog of analyses that turned out as blind alleys.

As a minimal requirement, you should at least be able to reproduce the results yourself. This would satisfy the most basic requirements of sound research, allowing any substantial future questioning of the research to be met with a precise explanation. Although it may sound like a very weak requirement, even this level of reproducibility will often require a certain level of care in order to be met. There will for a given analysis be an exponential number of possible combinations of software versions, parameter values, pre-processing steps, and so on, meaning that a failure to take notes may make exact reproduction essentially impossible.

With this basic level of reproducibility in place, there is much more that can be wished for. An obvious extension is to go from a level where you can reproduce

Day 5 : Galaxy

EMBO

In the news ...

- As of December 2013, I will be the new PLOS Comp Biology Education Editor, and be sure I will be adding Galaxy tutorials and quick guides to the roster!

In the news ...



- FGED: Functional Genomics Data Society
- fged.org
- Gabry and I are 'officers' of this society



Day 5 : Galaxy



Acknowledgements: the CBW gang



Michelle Brazas



Day 5 : Galaxy



Time for sponsor announcement!



<http://www.pmggenomics.ca/>



Zhibin Lu

Day 5 : Galaxy



Ontario

OICR
Ontario Institute
for Cancer Research

Informatics and Biocomputing at the OICR



If you have time ...

- Import your Dm RNA, and repeat experiments (from Monday) ...
- Do Jeremy's trackster lab:
<https://usegalaxy.org/u/jeremy/p/interactive-rna-seq-with-trackster>