Creative Commons

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

Learn how to distribute your work using this licence

Canadian Bioinformatics Workshops

www.bioinformatics.ca

You are free to:

Copy, share, adapt, or re-mix;

Photograph, film, or broadcast;

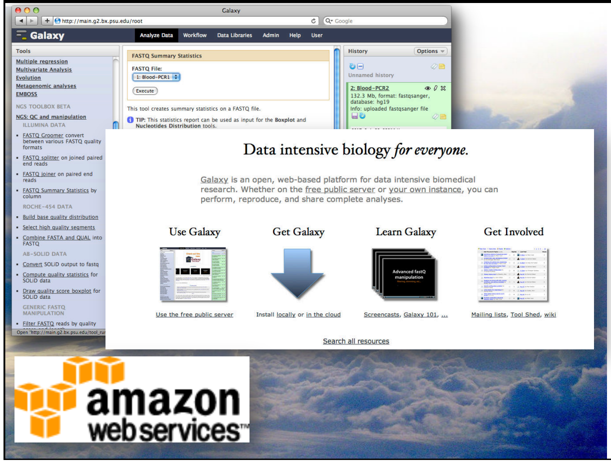
Blog, live-blog, or post video of;

This presentation. Provided that:

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

Module 6 part 1
Galaxy

BF Francis Ouellette
BF Francis Ouellette
Informatics on High Throughput Sequencing
June 9-10, 2014


Galaxy
http://main.g2.bx.psu.edu/root
Google
Galaxy
Analyze Data
Workflow
Data Libraries
Admin
Help
User
Tools
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
ILLUMINA DATA
FASTQ Groomer convert between various FASTQ quality formats
FASTQ splitter on joined paired end reads
FASTQ joiner on paired end reads
FASTQ Summary Statistics by column
ROCHE-454 DATA
Build base quality distribution
Select high quality segments
Combine FASTA and QUAL into FASTQ
AB-SOLID DATA
Convert SOLiD output to fastq
Compute quality statistics for SOLiD data
Draw quality score boxplot for SOLiD data
GENERIC FASTQ MANIPULATION
Filter FASTQ reads by quality
FASTQ Summary Statistics
FASTQ File:
1: Blood-PCR1
Execute
History
Options
Unnamed history
2: Blood-PCR2
Use Galaxy
Get Galaxy
Learn Galaxy
Get Involved
Advanced fastQ manipulation
Use the free public server
Install locally or in the cloud
Screencasts, Galaxy 101, ...
Mailing lists, Tool Shed, wiki
Search all resources
amazon webservices™

132.3 Mb, format: fastqsanger, database: hg19
Info: uploaded fastqsanger file

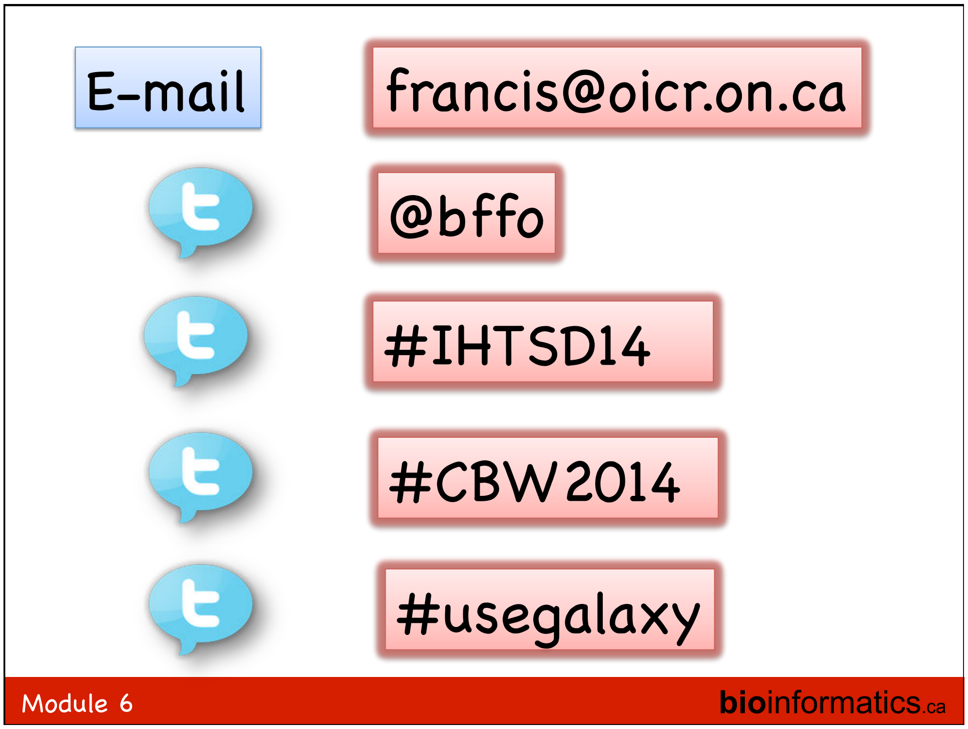
This tool creates summary statistics on a FASTQ file.

TIP: This statistics report can be used as input for the **Boxplot** and **Nucleotides Distribution** tools.

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.

OICR
Ontario Institute for Cancer Research



E-mail francis@oicr.on.ca

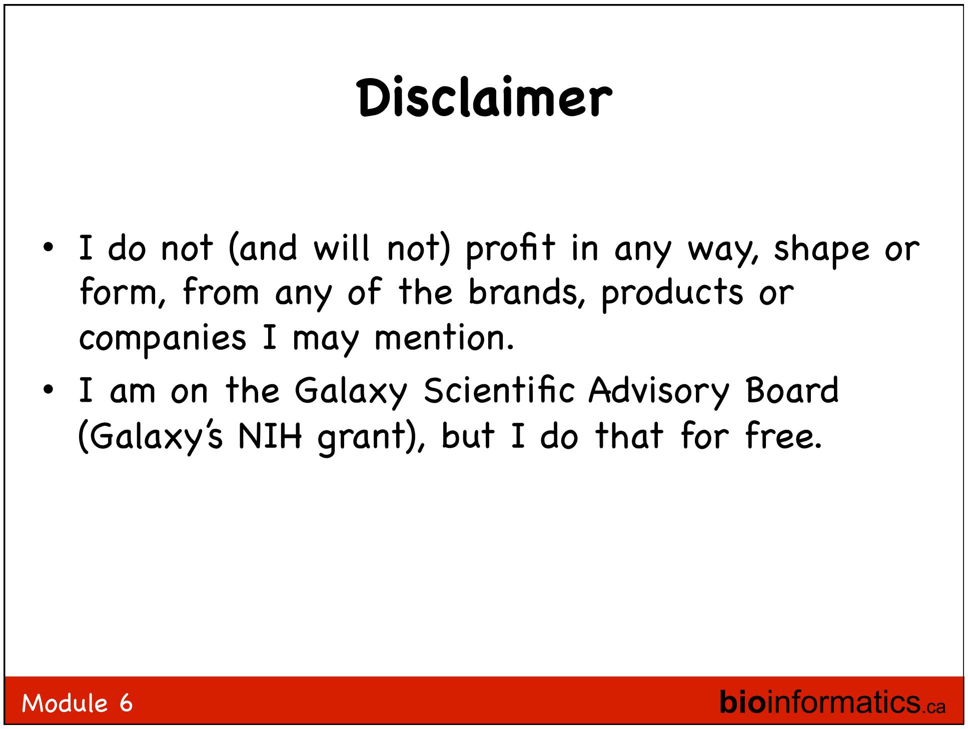
@bffo

#IHTSD14

#CBW2014

#usegalaxy

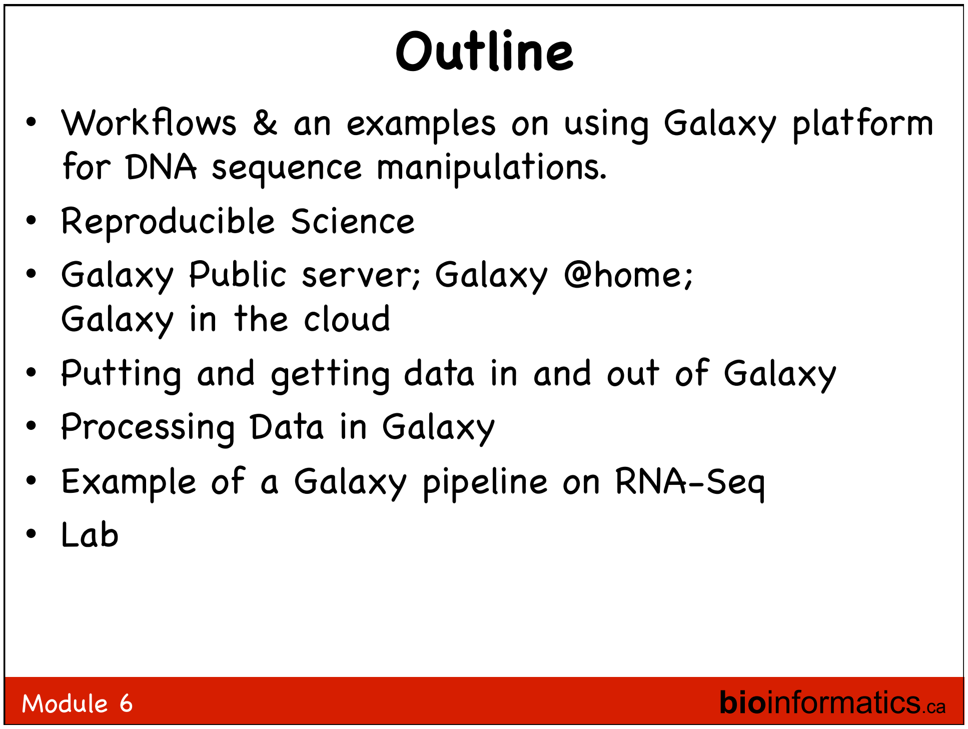
Module 6 bioinformatics.ca



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.

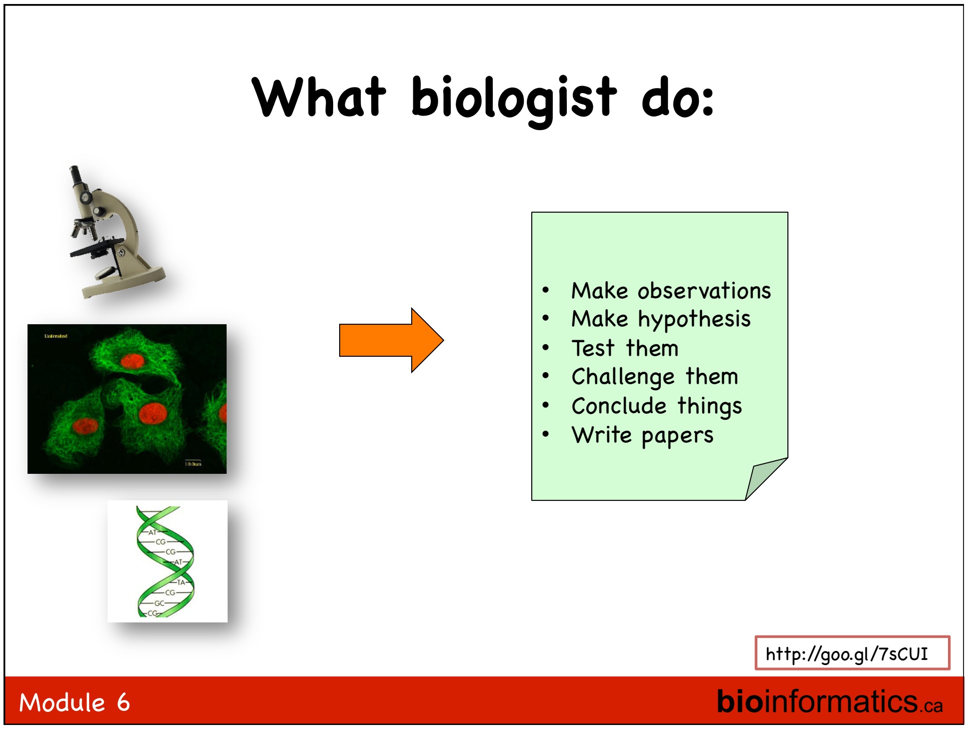
Module 6 bioinformatics.ca



Outline

- 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
- Example of a Galaxy pipeline on RNA-Seq
- Lab

Module 6 bioinformatics.ca

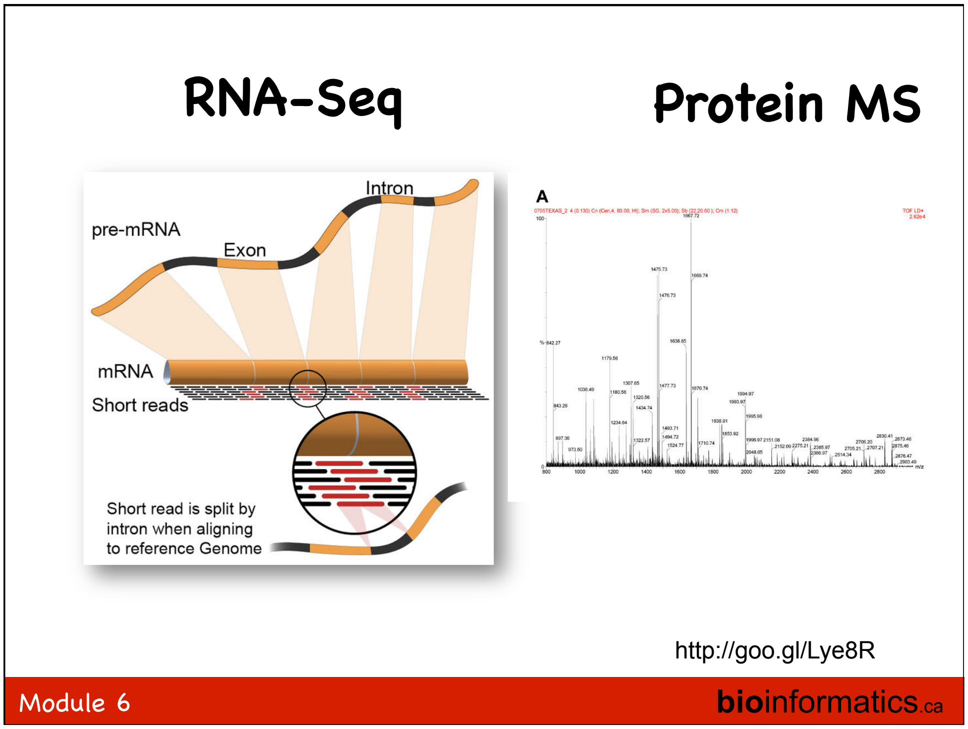


What biologist do:

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

http://goo.gl/7sCUI

Module 6 bioinformatics.ca



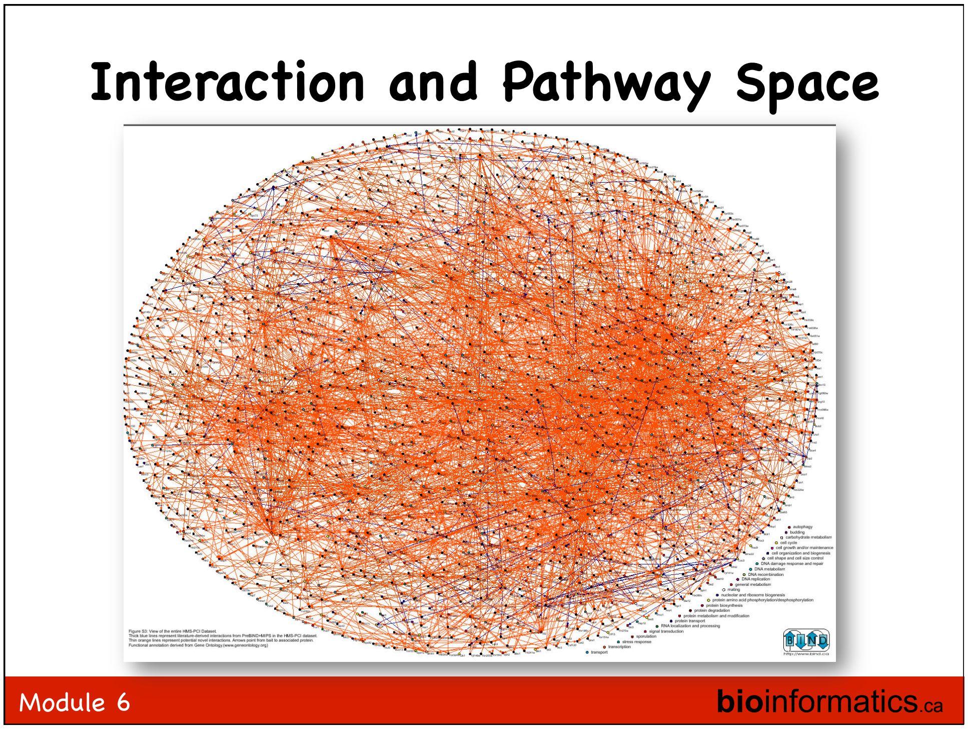
RNA-Seq

Protein MS

http://goo.gl/Lye8R

Module 6

bioinformatics.ca



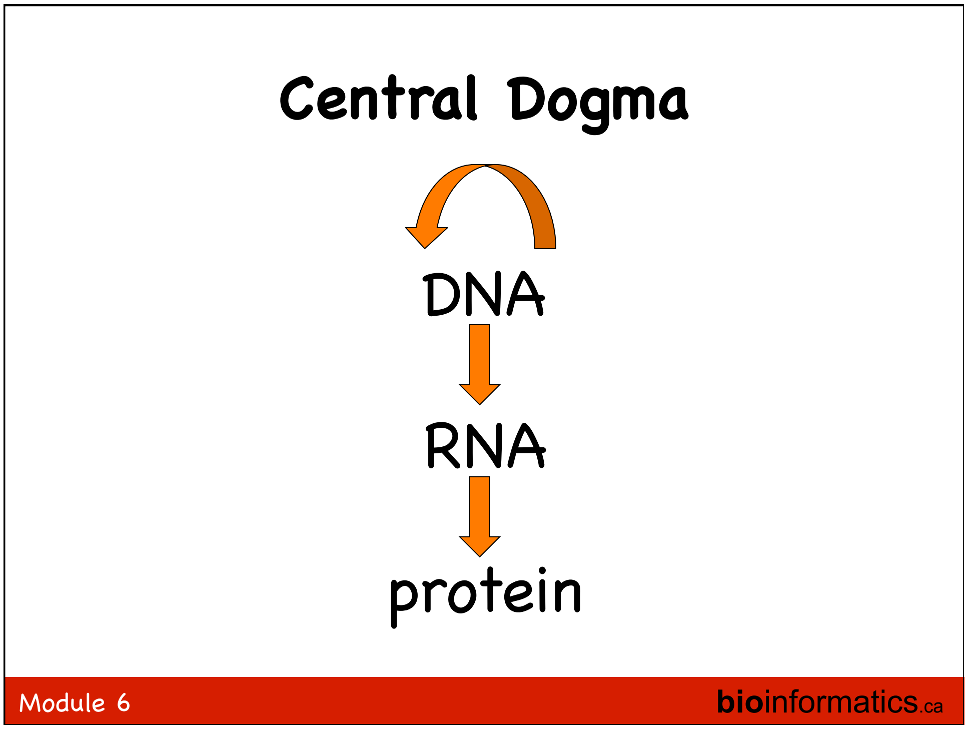
Interaction and Pathway Space

Figure S3: View of the entire HMS-PCI Dataset.
Thick blue lines represent literature-derived interactions from PreBIND+MIPS in the HMS-PCI dataset.
Thin orange lines represent potential novel interactions. Arrows point from bait to associated protein.
Functional annotation derived from Gene Ontology.(www.geneontology.org)

autophagy
budding
carbohydrate metabolism
cell cycle
cell growth and/or maintenance
cell organization and biogenesis
cell shape and cell size control
DNA damage response and repair
DNA metabolism
DNA recombination
DNA replication
general metabolism
mating
nucleolar and ribosome biogenesis
protein amino acid phosphorylation/desphosphorylation
protein biosynthesis
protein degradation
protein metabolism and modification
protein transport
RNA localization and processing
signal transduction
sporulation
stress response
transcription
transport
BIND
http://www.bind.ca

Module 6

bioinformatics.ca



Central Dogma

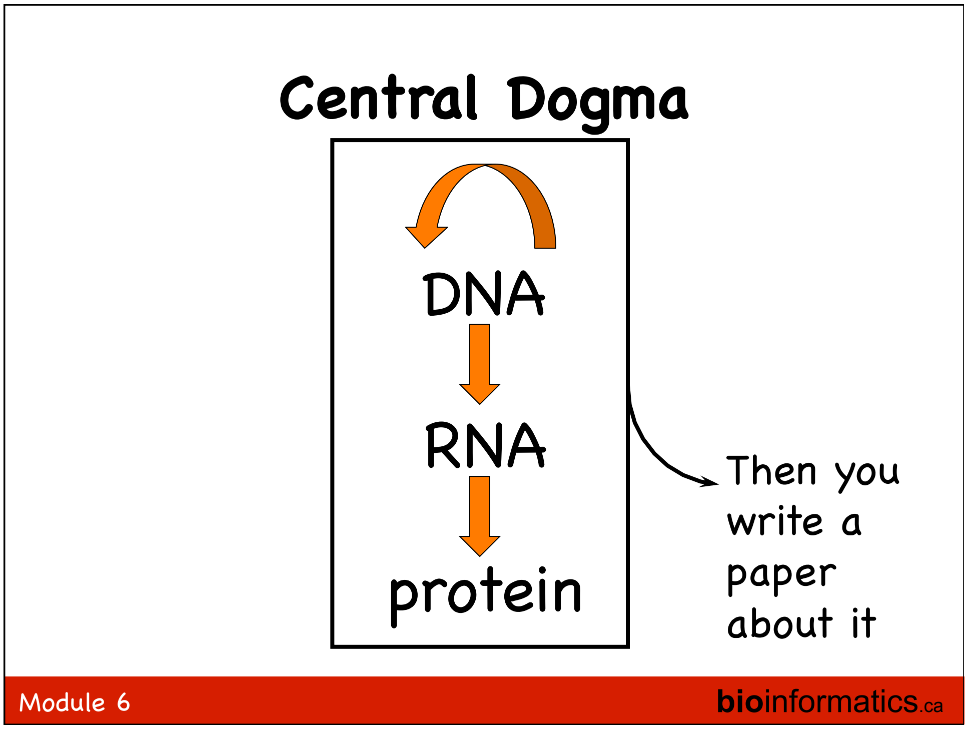
DNA

RNA

protein

Module 6

bioinformatics.ca



Central Dogma

DNA

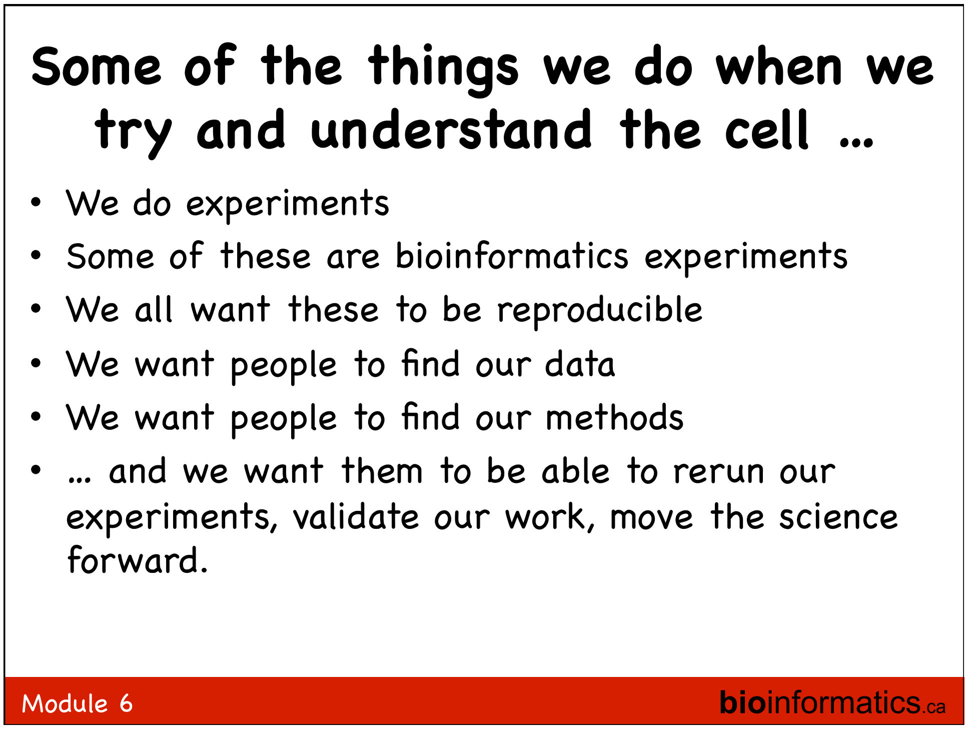
RNA

protein

Then you write a paper about it

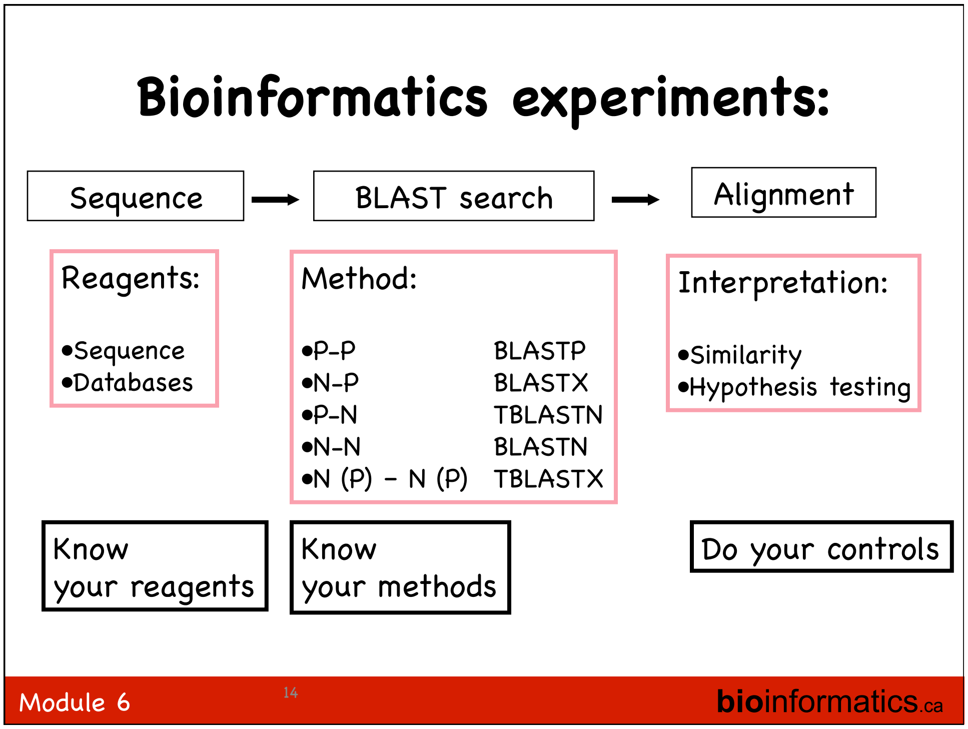
Module 6

bioinformatics.ca



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.



Bioinformatics experiments:

Sequence → BLAST search → Alignment

Reagents:

- •Sequence
- •Databases

Method:

•P-P	BLASTP
•N-P	BLASTX
•P-N	TBLASTN
•N-N	BLASTN
•N (P) – N (P)	TBLASTX

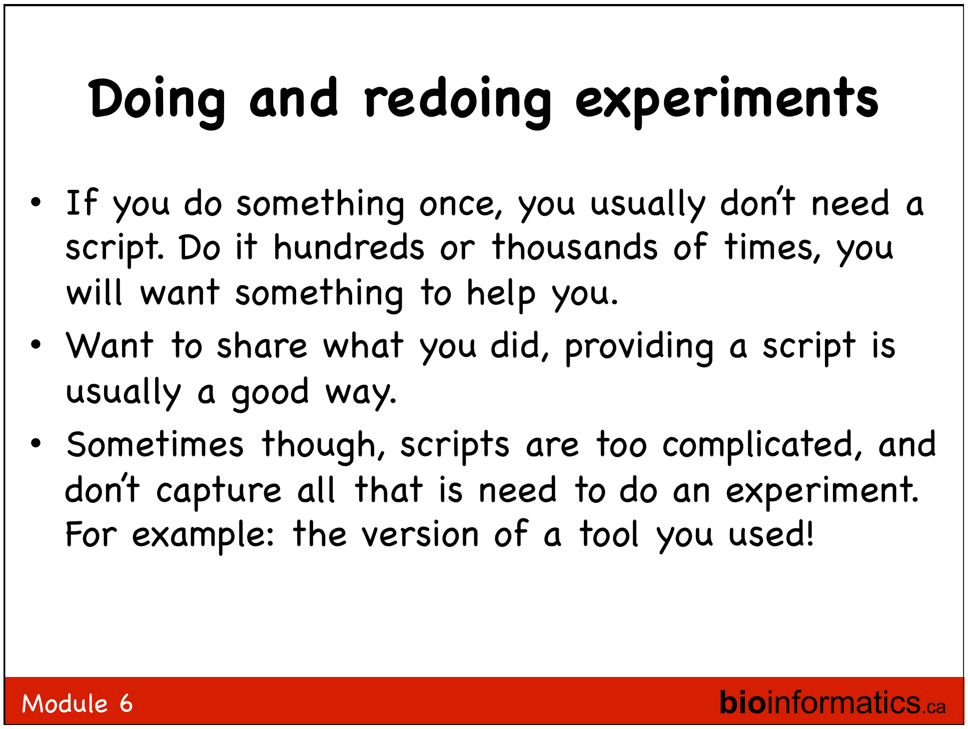
Interpretation:

- •Similarity
- •Hypothesis testing

Know your reagents

Know your methods

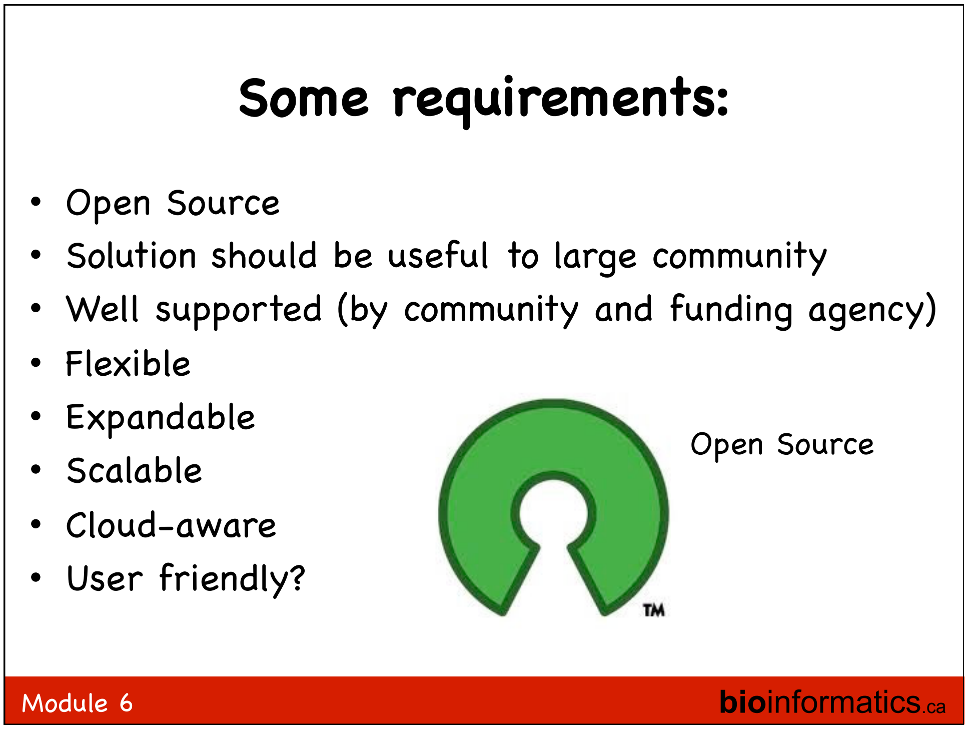
Do your controls



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!

Module 6 bioinformatics.ca



Some requirements:

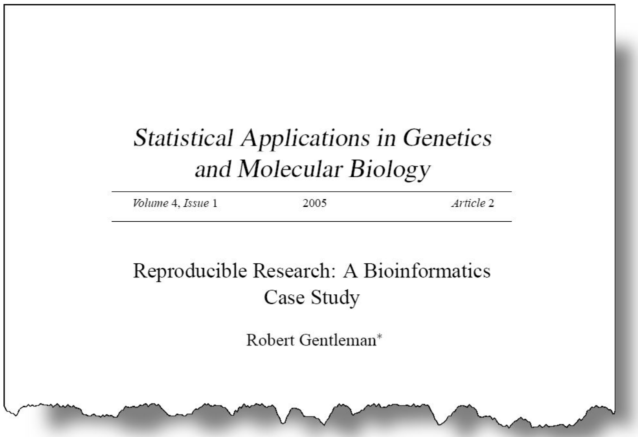
- Open Source
- Solution should be useful to large community
- Well supported (by community and funding agency)
- Flexible
- Expandable
- Scalable
- Cloud-aware
- User friendly?

Open Source

Module 6 bioinformatics.ca

Some solutions

1. R and bioconductor (#rstat)



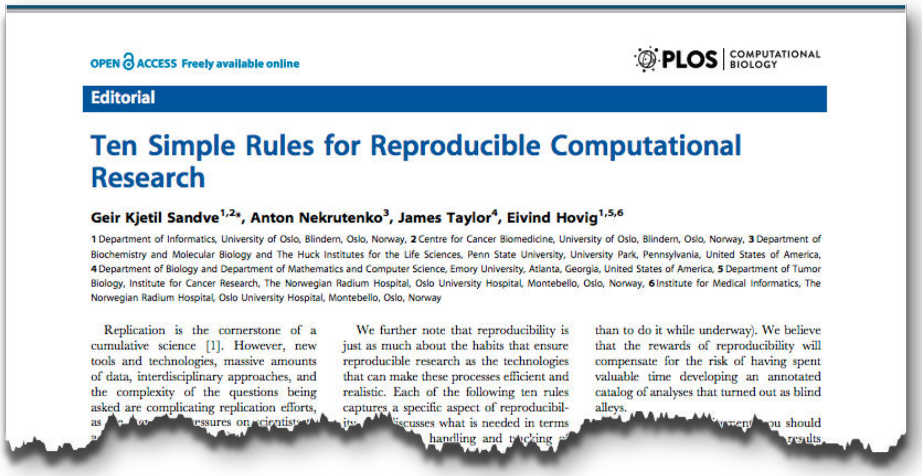
Statistical Applications in Genetics and Molecular Biology

Volume 4, Issue 1 2005 *Article 2*

Reproducible Research: A Bioinformatics Case Study

Robert Gentleman*

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



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve[1,2]*, Anton Nekrutenko[3], James Taylor[4], 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 [illegible] essures on scientists

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 [illegible] iscusses what is needed in terms handling and tracking

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.

ou should results

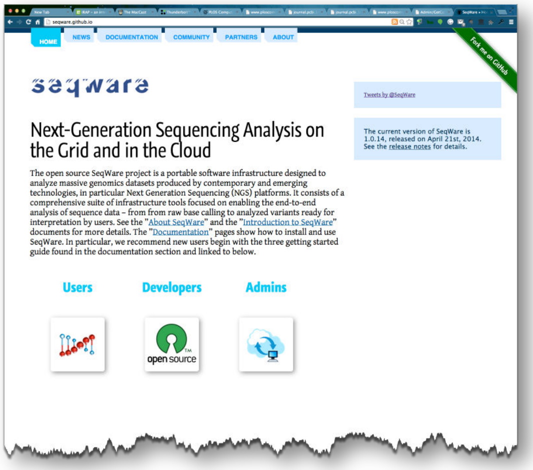
http://goo.gl/j8kCgd

Ten Simple Rules for Reproducible Computational Research

Rule 1: For Every Result, Keep Track of How It Was Produced
Rule 2: Avoid Manual Data Manipulation Steps
Rule 3: Archive the Exact Versions of All External Programs Used
Rule 4: Version Control All Custom Scripts
Rule 5: Record All Intermediate Results, When Possible in Standardized Formats
Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds
Rule 7: Always Store Raw Data behind Plots
Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
Rule 9: Connect Textual Statements to Underlying Results
Rule 10: Provide Public Access to Scripts, Runs, and Results

Some solutions (2)

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



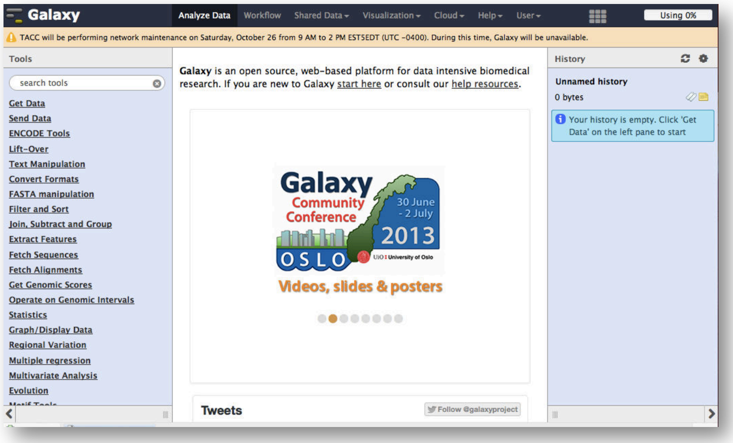
Next-Generation Sequencing Analysis on the Grid and in the Cloud

The open source SeqWare project is a portable software infrastructure designed to analyze massive genomics datasets produced by contemporary and emerging technologies, in particular Next Generation Sequencing (NGS) platforms. It consists of a comprehensive suite of infrastructure tools focused on enabling the end-to-end analysis of sequence data – from from raw base calling to analyzed variants ready for interpretation by users. See the "About SeqWare" and the "Introduction to SeqWare" documents for more details. The "Documentation" pages show how to install and use SeqWare. In particular, we recommend new users begin with the three getting started guide found in the documentation section and linked to below.

The current version of SeqWare is 1.0.14, released on April 21st, 2014. See the release notes for details.

Some solutions (3)

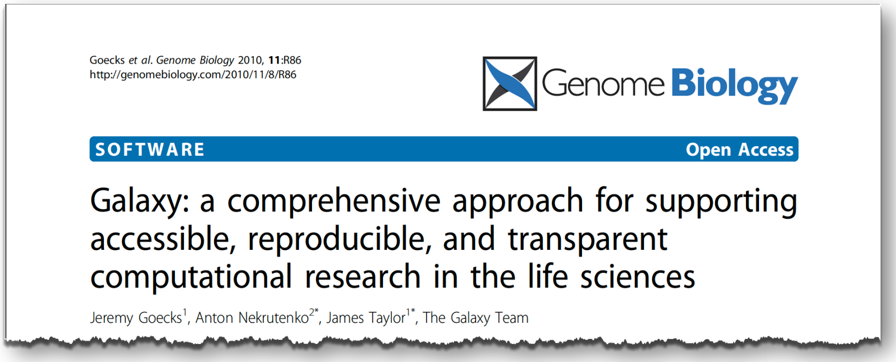
- Galaxy



TACC will be performing network maintenance on Saturday, October 26 from 9 AM to 2 PM EST5EDT (UTC -0400). During this time, Galaxy will be unavailable.

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

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



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

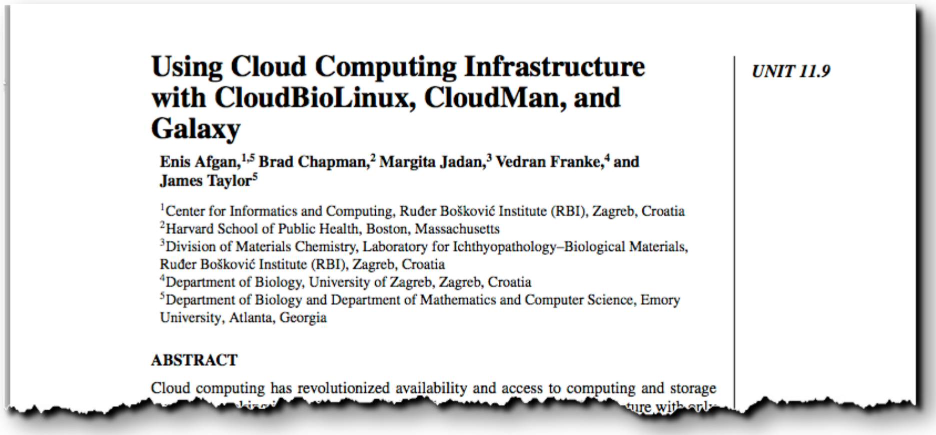
Genome Biology

SOFTWARE Open Access

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

Jeremy Goecks[1], Anton Nekrutenko[2*], James Taylor[1*], The Galaxy Team

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



Using Cloud Computing Infrastructure with CloudBioLinux, CloudMan, and Galaxy

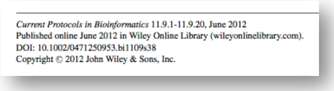
UNIT 11.9

Enis Afgan,[1,5] Brad Chapman,[2] Margita Jadan,[3] Vedran Franke,[4] and James Taylor[5]

[1]Center for Informatics and Computing, Ruđer Bošković Institute (RBI), Zagreb, Croatia
[2]Harvard School of Public Health, Boston, Massachusetts
[3]Division of Materials Chemistry, Laboratory for Ichthyopathology–Biological Materials, Ruđer Bošković Institute (RBI), Zagreb, Croatia
[4]Department of Biology, University of Zagreb, Zagreb, Croatia
[5]Department 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

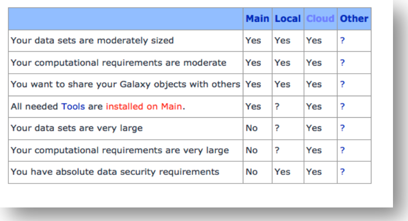


Current Protocols in Bioinformatics 11.9.1-11.9.20, June 2012
Published online June 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471250953.bi1109s38

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

Which Galaxy?

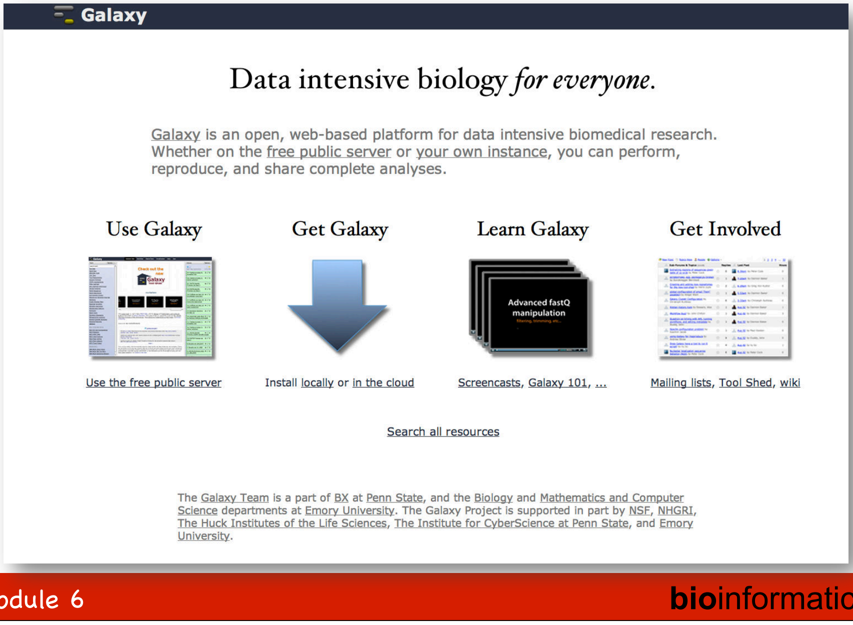
- galaxyproject.org: Galaxy home page
- usegalaxy.org: main Galaxy public server
- getgalaxy.org: source for installing local Galaxy
- usegalaxy.org/cloud: use galaxy in the cloud
- http://goo.gl/mlyOC : Other public Galaxy servers



	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

http://galaxyproject.org/



Data intensive biology *for everyone.*

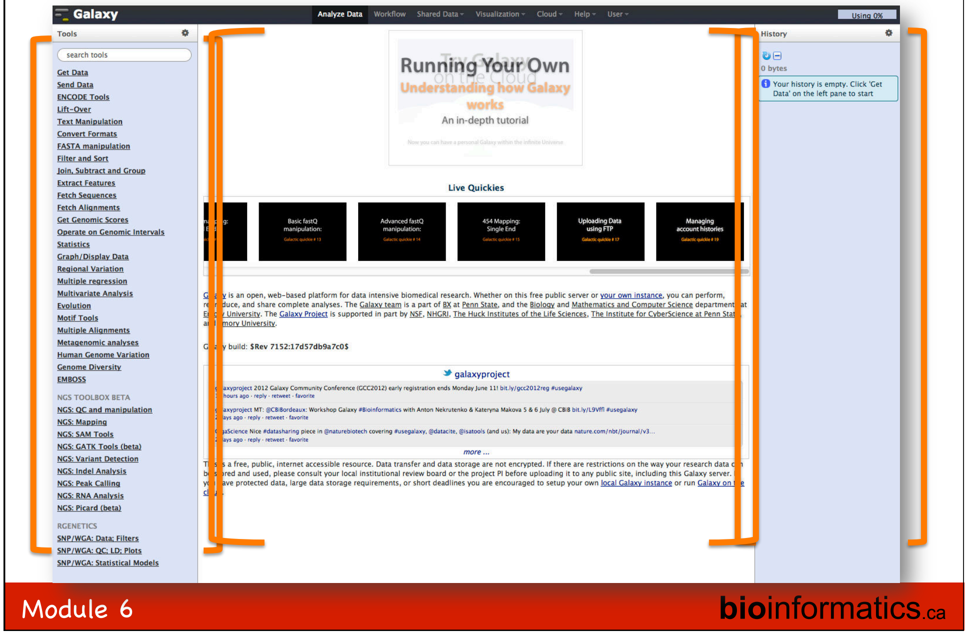
Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy	Get Galaxy	Learn Galaxy	Get Involved
Use the free public server	Install locally or in the cloud	Screencasts, Galaxy 101, ...	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, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

http://usegalaxy.org



Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
Tools
search tools
History
0 bytes
Running Your Own
Understanding how Galaxy works
An in-depth tutorial
Now you can have a personal Galaxy within the infinite Universe
Live Quickies
Basic fastQ manipulation: Galactic quickie # 13
Advanced fastQ manipulation: Galactic quickie # 14
454 Mapping: Single End Galactic quickie # 15
Uploading Data using FTP Galactic quickie # 17
Managing account histories Galactic quickie # 19
galaxyproject
more ...

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
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
Human Genome Variation
Genome Diversity
EMBOSS

NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Variant Detection
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
NGS: Picard (beta)

RGENETICS
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models

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

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science department at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Galaxy build: **$Rev 7152:17d57db9a7c0$**

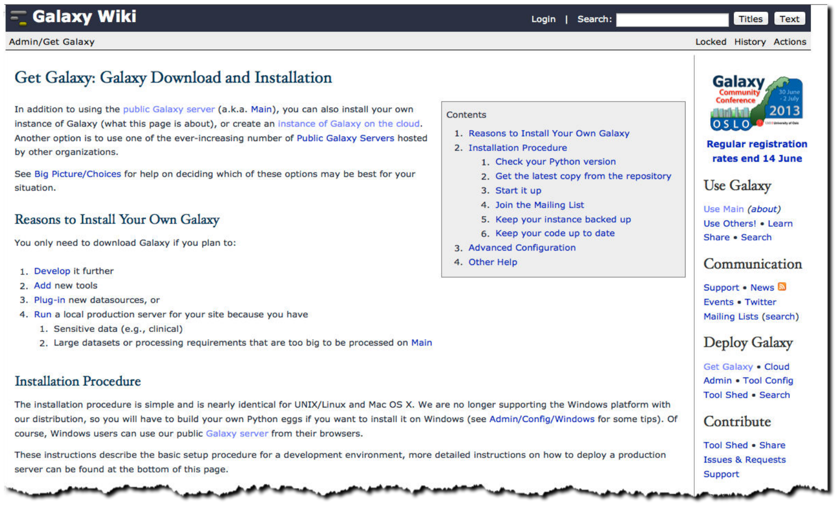
axyproject 2012 Galaxy Community Conference (GCC2012) early registration ends Monday June 11! bit.ly/gcc2012reg #usegalaxy
hours ago · reply · retweet · favorite

axyproject MT: @CBiBordeaux: Workshop Galaxy #Bioinformatics with Anton Nekrutenko & Kateryna Makova 5 & 6 July @ CBiB bit.ly/L9Vffl #usegalaxy
lays ago · reply · retweet · favorite

jaScience Nice #datasharing piece in @naturebiotech covering #usegalaxy, @datacite, @isatools (and us): My data are your data nature.com/nbt/journal/v3...
lays ago · reply · retweet · favorite

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own local Galaxy instance or run Galaxy on the cloud.

getgalaxy.org



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.

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

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.

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

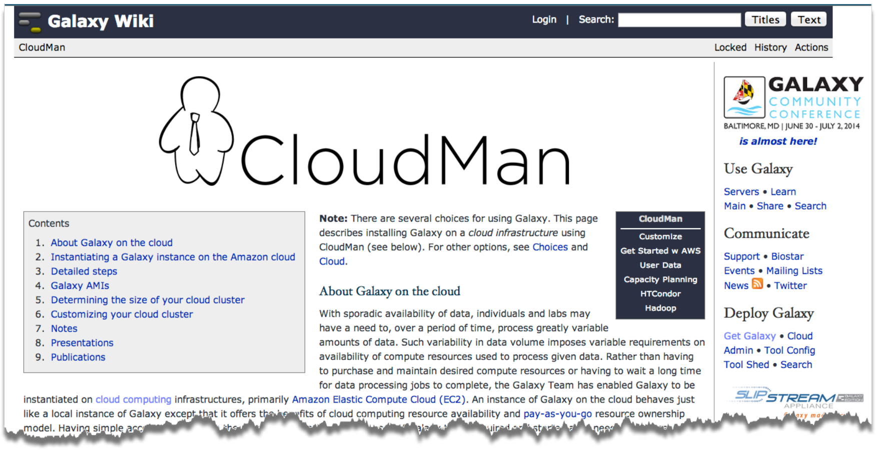
Get Galaxy • Cloud
Admin • Tool Config
Tool Shed • Search

Contribute

Tool Shed • Share
Issues & Requests
Support

Module 6 **bio**informatics.ca

usegalaxy.org/cloud



Contents

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

Note: There are several choices for using Galaxy. This page describes installing Galaxy on a *cloud infrastructure* using CloudMan (see below). For other options, see Choices and Cloud.

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 just 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 acc

GALAXY COMMUNITY CONFERENCE
BALTIMORE, MD | JUNE 30 - JULY 2, 2014

is almost here!

Use Galaxy

Servers • Learn
Main • Share • Search

Communicate

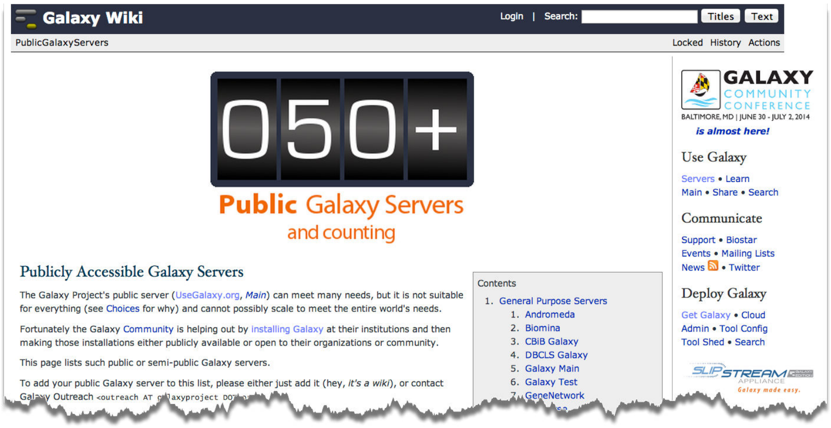
Support • Biostar
Events • Mailing Lists
News • Twitter

Deploy Galaxy

Get Galaxy • Cloud
Admin • Tool Config
Tool Shed • Search

Module 6 **bio**informatics.ca

http://wiki.galaxyproject.org/PublicGalaxyServers



Publicly Accessible Galaxy Servers

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

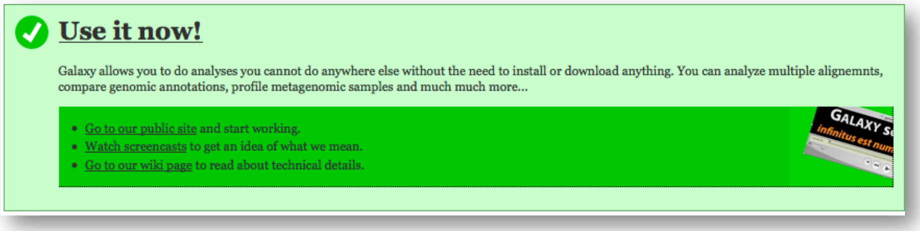
Fortunately the Galaxy Community is helping out by installing Galaxy at their institutions and then making those installations either publicly available or open to their organizations or community.

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

Contents

1. General Purpose Servers
   1. Andromeda
   2. Biomina
   3. CBiB Galaxy
   4. DBCLS Galaxy
   5. Galaxy Main
   6. Galaxy Test
   7. GeneNetwork

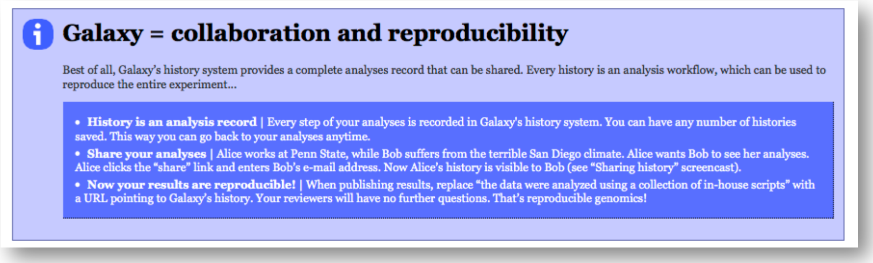


Use it now!

Galaxy allows you to do analyses you cannot do anywhere else without the need to install or download anything. You can analyze multiple alignemnts, 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.

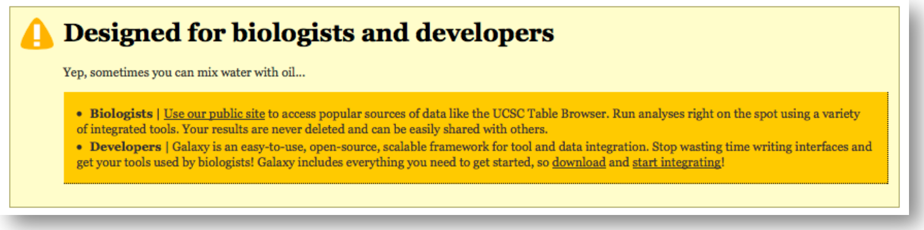


Galaxy = collaboration and reproducibility

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 easy to work with collaborators down the hall, or across the globe.

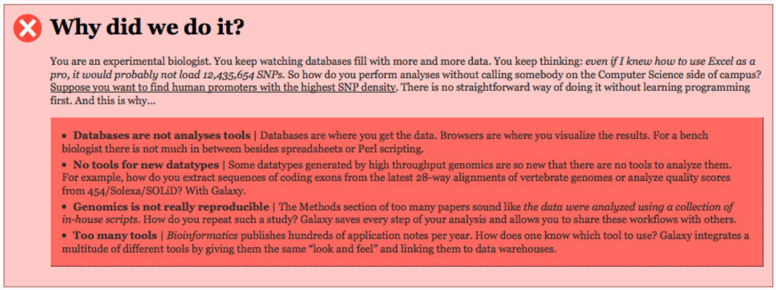


Designed for biologists and developers

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.



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...

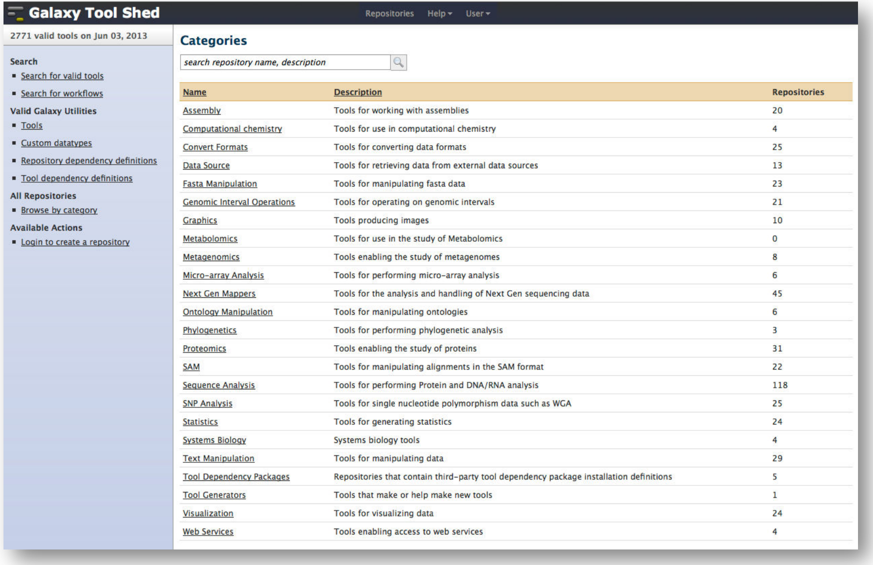
- **Databases are not analyses tools** | Databases are where you get the data. Browsers are where you visualize the results. For a bench biologist there is not much in between besides spreadsheets or Perl scripting.
- **No tools for new datatypes** | Some datatypes generated by high throughput genomics 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/Solexa/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

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!

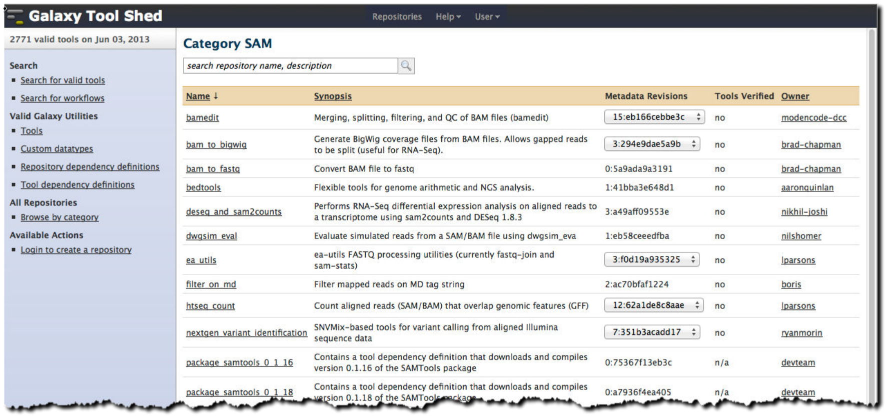
Galaxy Toolshed: http://toolshed.g2.bx.psu.edu/



Categories

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

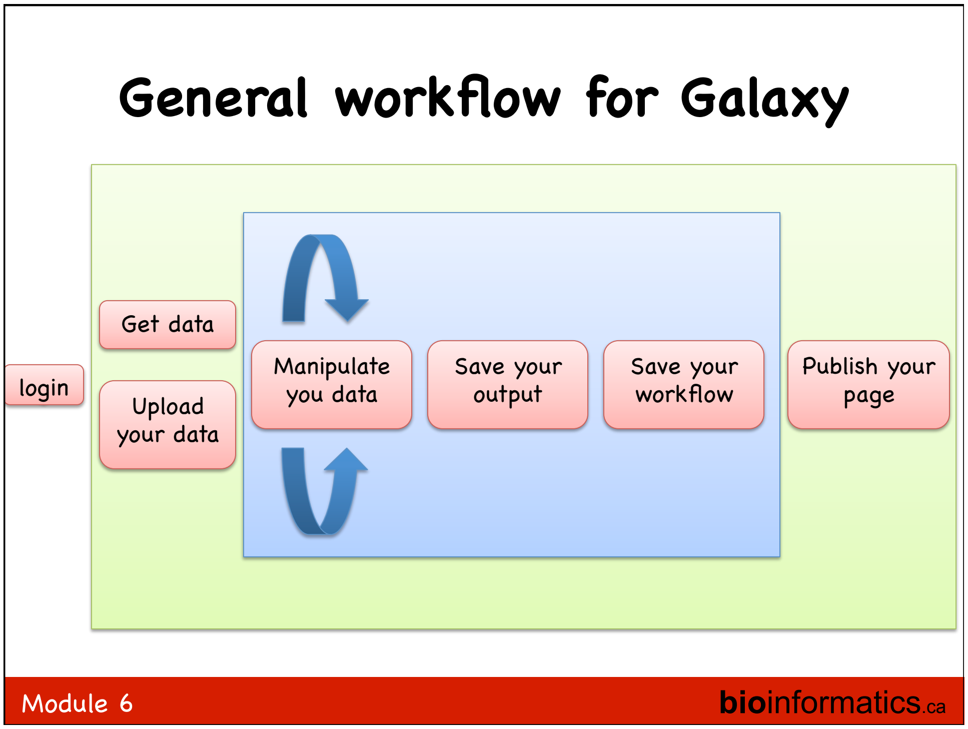
Galaxy Toolshed: SAM



Category SAM

Name ↓	Synopsis	Metadata Revisions	Tools Verified	Owner
bamedit	Merging, splitting, filtering, and QC of BAM files (bamedit)	15:eb166cebbe3c	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_eva	1:eb58ceeedfba	no	nilshomer
ea_utils	ea-utils FASTQ processing utilities (currently fastq-join and sam-stats)	3:f0d19a935325	no	lparsons
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	lparsons
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:a7936f4ea405	n/a	devteam

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



General workflow for Galaxy

login

Get data

Upload your data

Manipulate you data

Save your output

Save your workflow

Publish your page

Module 6

bioinformatics.ca

Time for sponsor announcement!

PRINCESS MARGARET GENOMICS CENTRE

http://www.pmgenomics.ca/

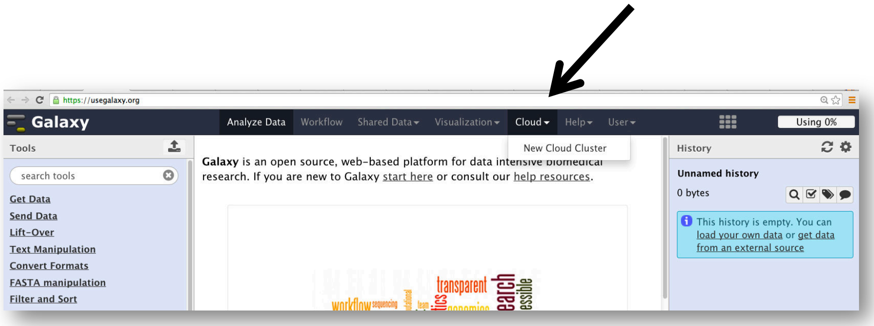
Zhibin Lu

Module 6

bioinformatics.ca

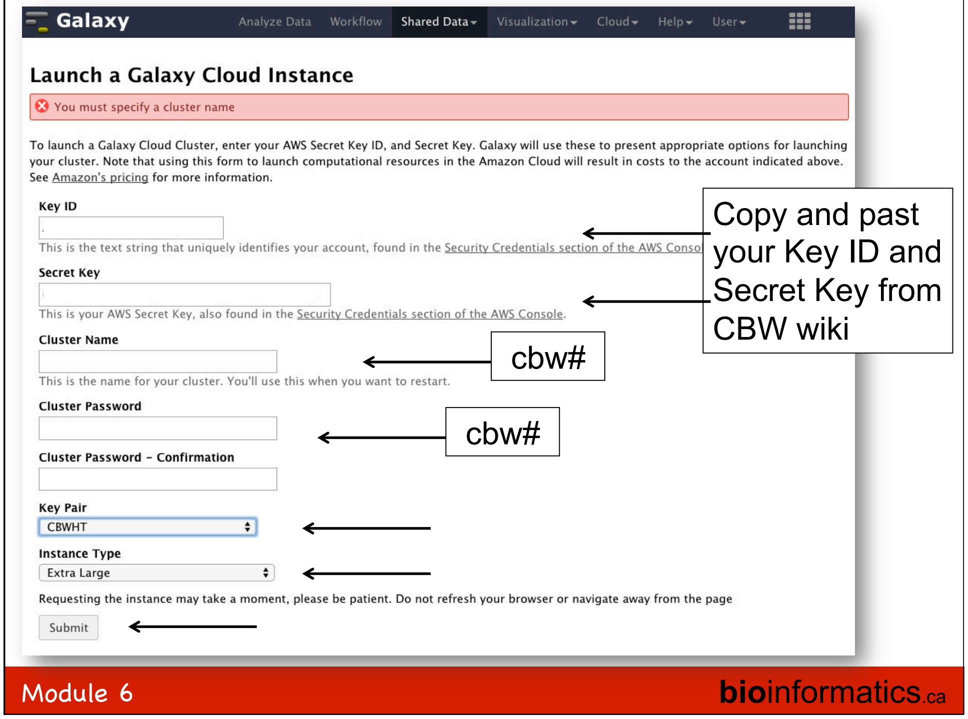
Log onto Galaxy

- Go to https://usegalaxy.org
- Under the tab Cloud click on New Cloud Cluster



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

This history is empty. You can load your own data or get data from an external source



Launch a Galaxy Cloud Instance

You must specify a cluster name

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 your cluster. Note that using this form to launch computational resources in the Amazon Cloud will result in costs to the account indicated above. See Amazon's pricing for more information.

Key ID

This is the text string that uniquely identifies your account, found in the Security Credentials section of the AWS Conso

Secret Key

This is your AWS Secret Key, also found in the Security Credentials section of the AWS Console.

Cluster Name

This is the name for your cluster. You'll use this when you want to restart.

Cluster Password

Cluster Password - Confirmation

Key Pair: CBWHT

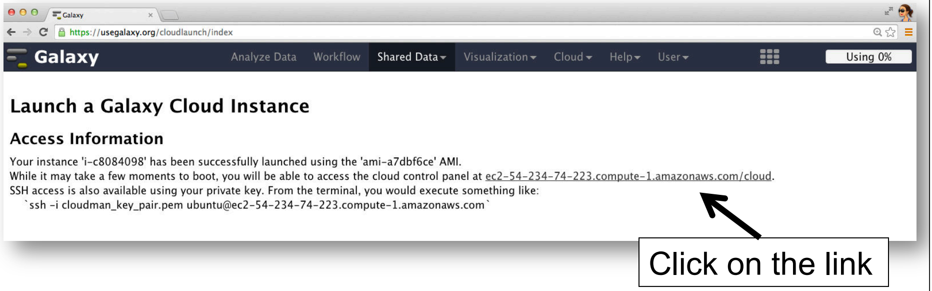
Instance Type: Extra Large

Requesting the instance may take a moment, please be patient. Do not refresh your browser or navigate away from the page

Copy and past your Key ID and Secret Key from CBW wiki

cbw#

cbw#



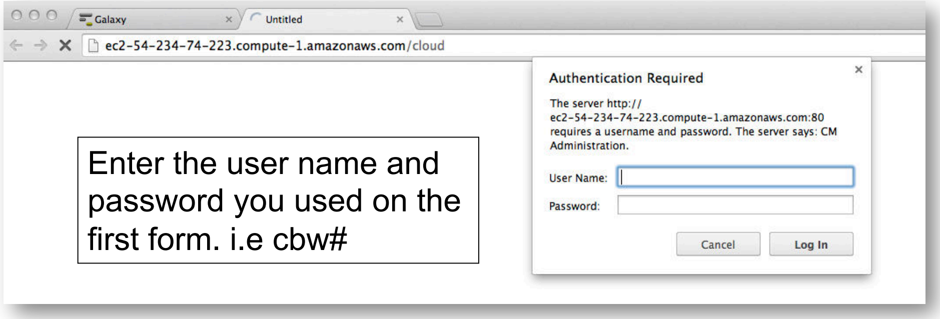
Launch a Galaxy Cloud Instance

Access Information

Your instance 'i-c8084098' has been successfully launched using the 'ami-a7dbf6ce' AMI.
While it may take a few moments to boot, you will be able to access the cloud control panel at ec2-54-234-74-223.compute-1.amazonaws.com/cloud.
SSH access is also available using your private key. From the terminal, you would execute something like:

```
    `ssh -i cloudman_key_pair.pem ubuntu@ec2-54-234-74-223.compute-1.amazonaws.com`
```

Click on the link



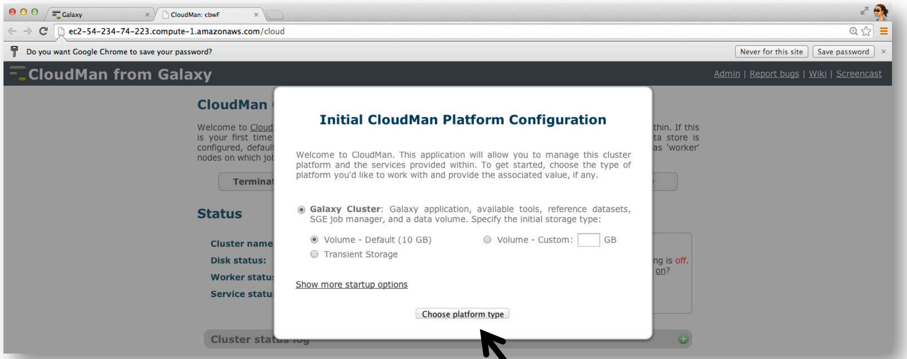
Authentication Required

The server http://
ec2-54-234-74-223.compute-1.amazonaws.com:80
requires a username and password. The server says: CM Administration.

User Name:

Password:

Enter the user name and password you used on the first form. i.e cbw#



ec2-54-234-74-223.compute-1.amazonaws.com/cloud

Do you want Google Chrome to save your password?

Initial CloudMan Platform Configuration

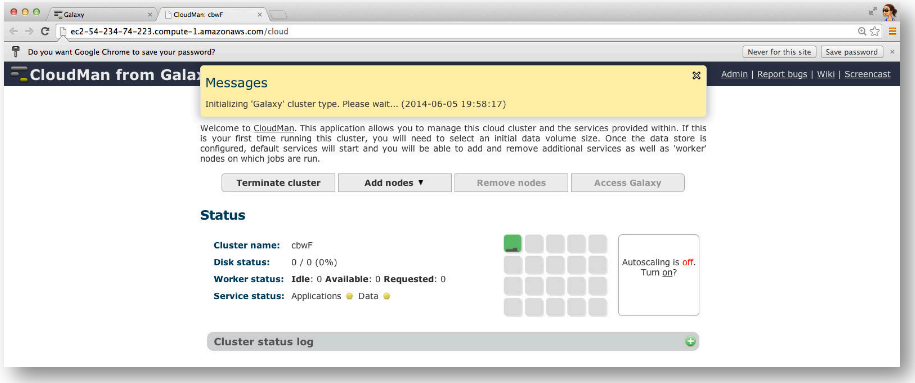
Welcome to CloudMan. This application will allow you to manage this cluster platform and the services provided within. To get started, choose the type of platform you'd like to work with and provide the associated value, if any.

● **Galaxy Cluster**: Galaxy application, available tools, reference datasets, SGE job manager, and a data volume. Specify the initial storage type:

● Volume - Default (10 GB) ○ Volume - Custom: GB

○ Transient Storage

Keep default
Press in Choose platform type



ec2-54-234-74-223.compute-1.amazonaws.com/cloud

Do you want Google Chrome to save your password?

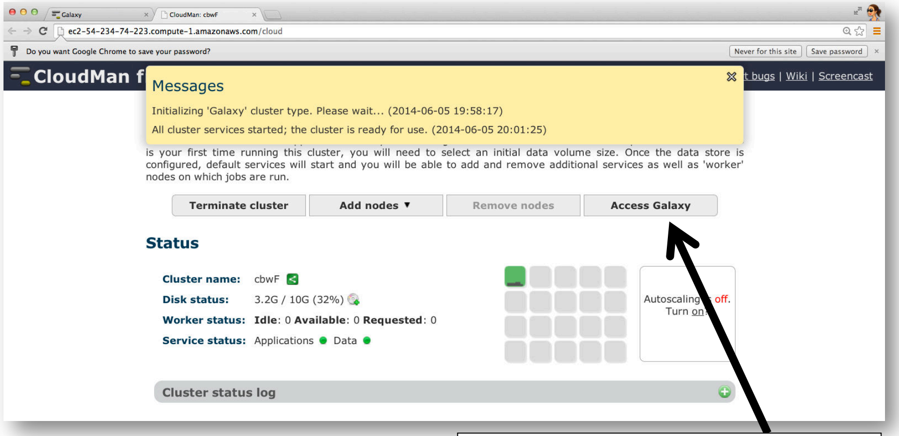
Messages

Initializing 'Galaxy' cluster type. Please wait... (2014-06-05 19:58:17)

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 'worker' nodes on which jobs are run.

Status

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



Messages

Initializing 'Galaxy' cluster type. Please wait... (2014-06-05 19:58:17)

All cluster services started; the cluster is ready for use. (2014-06-05 20:01:25)

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 'worker' nodes on which jobs are run.

Status

Cluster name: cbwF

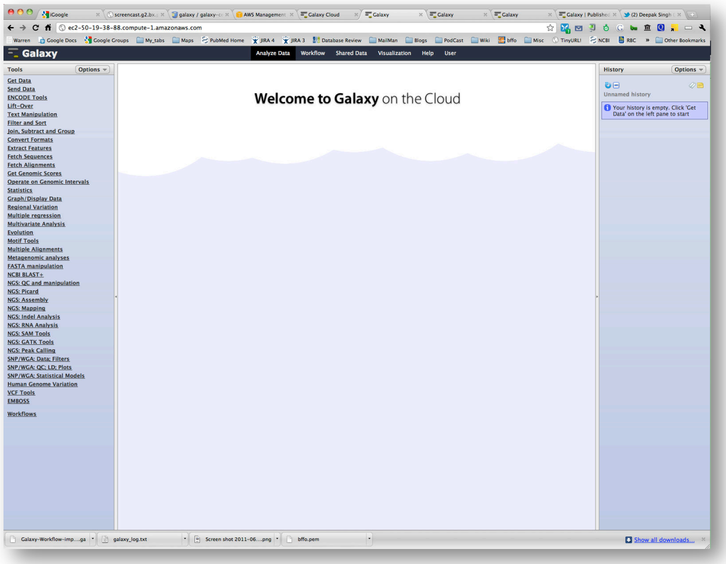
Disk status: 3.2G / 10G (32%)

Worker status: **Idle**: 0 **Available**: 0 **Requested**: 0

Service status: Applications ● Data ●

Autoscaling is off.
Turn on

Press on Access 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
NCBI BLAST+
NGS: QC and manipulation
NGS: Picard
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
EMBOSS

Workflows

Welcome to Galaxy on the Cloud

Unnamed history

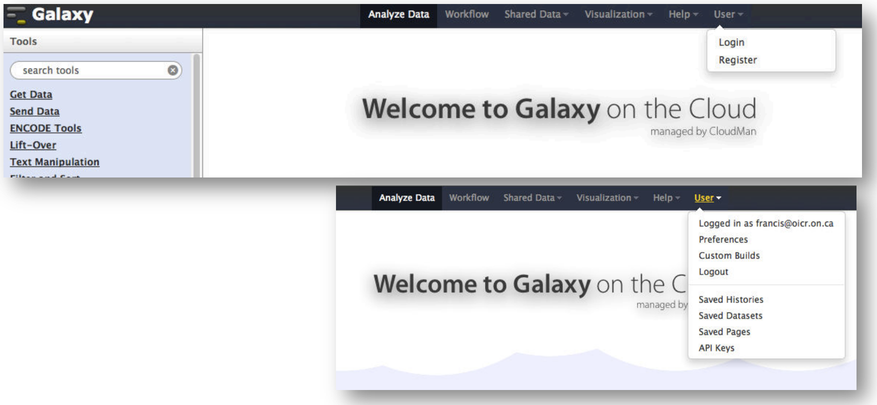
Your history is empty. Click 'Get Data' on the left pane to start

What is next?

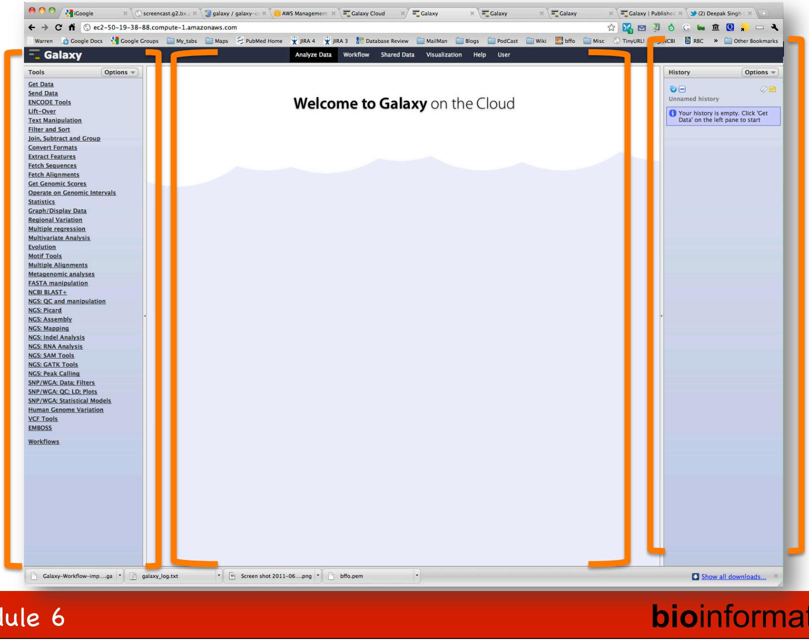
- I'm going to tell you about getting data in and out of Galaxy
- Doing operations in Galaxy
- Understanding the user interface.
- Linking multiple steps into "pipelines"
- Do an RNASeq mapping experiment
- Sharing pipeline with colleagues, and making them public.
- How to learn more …

1st thing to do before we start:

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



Galaxy in the cloud



Welcome to Galaxy on the Cloud

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

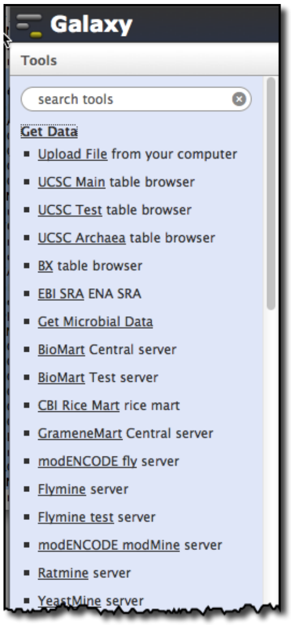
- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple 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

Galaxy cloud

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

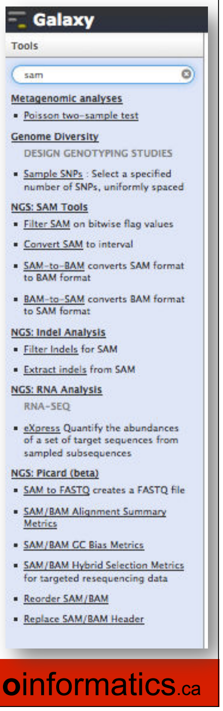
usegalaxy.org

- > 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



- … 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 using the search tool.
- E.g. look for "sam"

usegalaxy.org



Metagenomic analyses
- Poisson two-sample test

Genome Diversity

DESIGN GENOTYPING STUDIES
- Sample SNPs : Select a specified number of SNPs, uniformly spaced

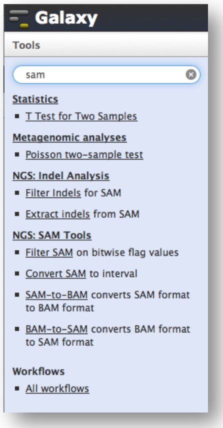
NGS: SAM Tools
- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format

NGS: Indel Analysis
- Filter Indels for SAM
- Extract indels from SAM

NGS: RNA Analysis

RNA-SEQ
- eXpress Quantify the abundances of a set of target sequences from sampled subsequences

NGS: Picard (beta)
- SAM to FASTQ creates a FASTQ file
- SAM/BAM Alignment Summary Metrics
- SAM/BAM GC Bias Metrics
- SAM/BAM Hybrid Selection Metrics for targeted resequencing data
- Reorder SAM/BAM
- Replace SAM/BAM Header



Statistics
- T Test for Two Samples

Metagenomic analyses
- Poisson two-sample test

NGS: Indel Analysis
- Filter Indels for SAM
- Extract indels from SAM

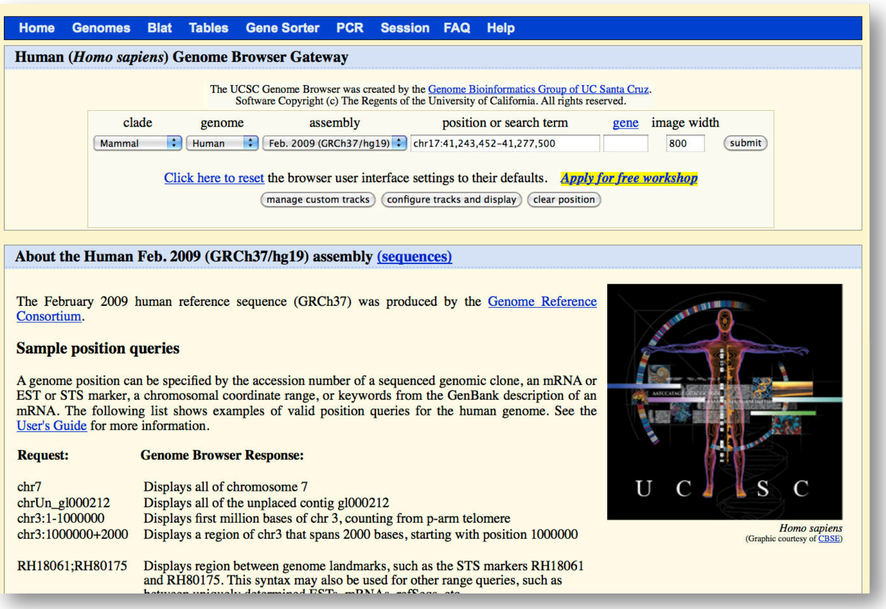
NGS: SAM Tools
- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format

Workflows
- All workflows

Galaxy cloud

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!



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

Click here to reset the browser user interface settings to their defaults. *Apply for free workshop*

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_gl000212	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

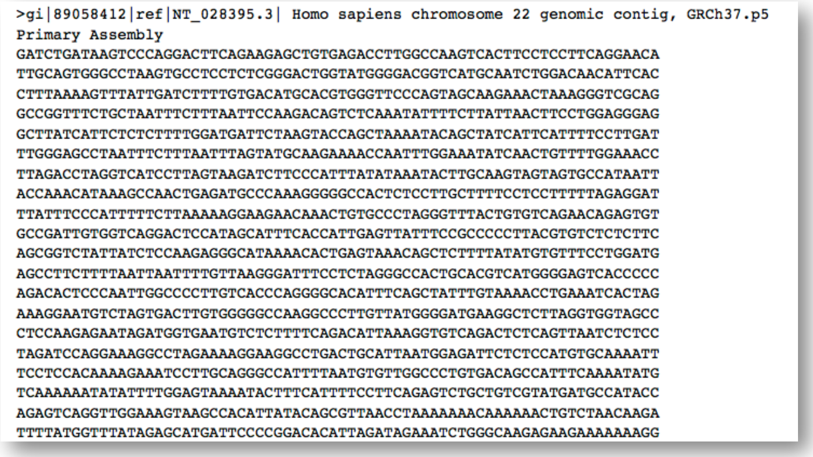
Homo sapiens
(Graphic courtesy of CBSE)

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)

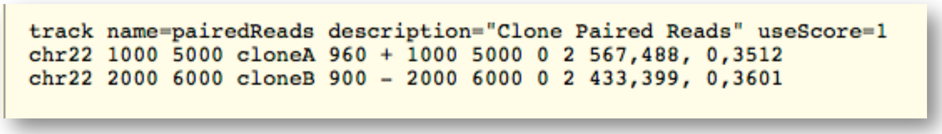
Examples of Data Formats for UCSC:

- Sequence (FASTA):

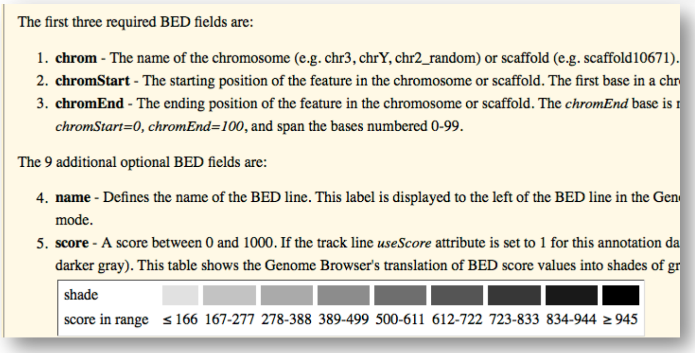


```
>gi|89058412|ref|NT_028395.3| Homo sapiens chromosome 22 genomic contig, GRCh37.p5
Primary Assembly
GATCTGATAAGTCCCAGGACTTCAGAAGAGCTGTGAGACCTTGGCCAAGTCACTTCCTCCTTCAGGAACA
TTGCAGTGGGCCTAAGTGCCTCCTCTCGGGACTGGTATGGGGACGGTCATGCAATCTGGACAACATTCAC
CTTTAAAAGTTTATTGATCTTTTGTGACATGCACGTGGGTTCCCAGTAGCAAGAAACTAAAGGGTCGCAG
GCCGGTTTCTGCTAATTTCTTTAATTCCAAGACAGTCTCAAATATTTTCTTATTAACTTCCTGGAGGGAG
GCTTATCATTCTCTCTTTTGGATGATTCTAAGTACCAGCTAAAATACAGCTATCATTCATTTTCCTTGAT
TTGGGAGCCTAATTTCTTTAATTTAGTATGCAAGAAAACCAATTTGGAAATATCAACTGTTTTGGAAACC
TTAGACCTAGGTCATCCTTAGTAAGATCTTCCCATTTATATAAATACTTGCAAGTAGTAGTGCCATAATT
ACCAAACATAAAGCCAACTGAGATGCCCAAAGGGGGCCACTCTCCTTGCTTTTCCTCCTTTTTAGAGGAT
TTATTTCCCATTTTTCTTAAAAAGGAAGAACAAACTGTGCCCTAGGGTTTACTGTGTCAGAACAGAGTGT
GCCGATTGTGGTCAGGACTCCATAGCATTTCACCATTGAGTTATTTCCGCCCCCTTACGTGTCTCTCTTC
AGCGGTCTATTATCTCCAAGAGGGCATAAAACACTGAGTAAACAGCTCTTTTATATGTGTTTCCTGGATG
AGCCTTCTTTTAATTAATTTTGTTAAGGGATTTCCTCTAGGGCCACTGCACGTCATGGGGAGTCACCCCC
AGACACTCCCAATTGGCCCCTTGTCACCCAGGGGCACATTTCAGCTATTTGTAAAACCTGAAATCACTAG
AAAGGAATGTCTAGTGACTTGTGGGGGCCAAGGCCCTTGTTATGGGGATGAAGGCTCTTAGGTGGTAGCC
CTCCAAGAGAATAGATGGTGAATGTCTCTTTTCAGACATTAAAGGTGTCAGACTCTCAGTTAATCTCTCC
TAGATCCAGGAAAGGCCTAGAAAAGGAAGGCCTGACTGCATTAATGGAGATTCTCTCCATGTGCAAAATT
TCCTCCACAAAAGAAATCCTTGCAGGGCCATTTTAATGTGTTGGCCCTGTGACAGCCATTTCAAAATATG
TCAAAAAATATATTTTGGAGTAAAATACTTTCATTTTCCTTCAGAGTCTGCTGTCGTATGATGCCATACC
AGAGTCAGGTTGGAAAGTAAGCCACATTATACAGCGTTAACCTAAAAAAACAAAAAACTGTCTAACAAGA
TTTTATGGTTTATAGAGCATGATTCCCCGGACACATTAGATAGAAATCTGGGCAAGAGAAGAAAAAAAGG
```

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 chr
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is r *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

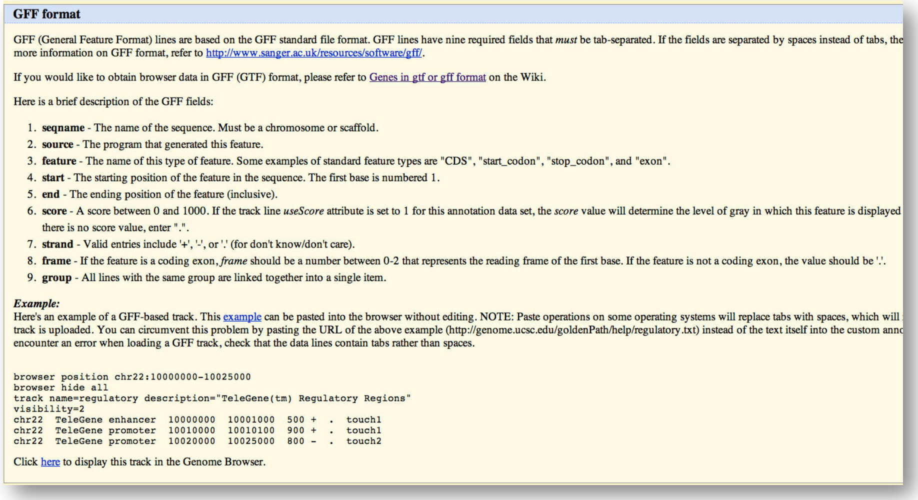
The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Gen mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation da darker gray). This table shows the Genome Browser's translation of BED score values into shades of gr

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

http://goo.gl/agfWu

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
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 annc
encounter an error when loading a GFF track, check that the data lines contain tabs rather than spaces.

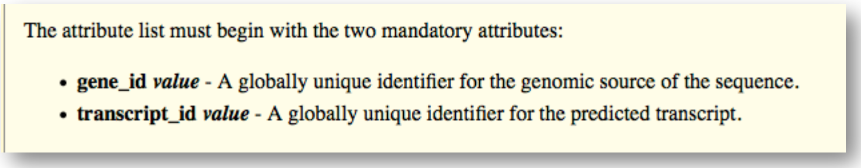
```
browser position chr22:10000000-10025000
browser hide all
track name=regulatory description="TeleGene(tm) Regulatory Regions"
visibility=2
chr22  TeleGene enhancer  10000000  10001000  500 +  .  touch1
chr22  TeleGene promoter  10010000  10010100  900 +  .  touch1
chr22  TeleGene promoter  10020000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

Click here to display this track in the Genome Browser.

http://goo.gl/agfWu

Gene Transfer Format (GTF)

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

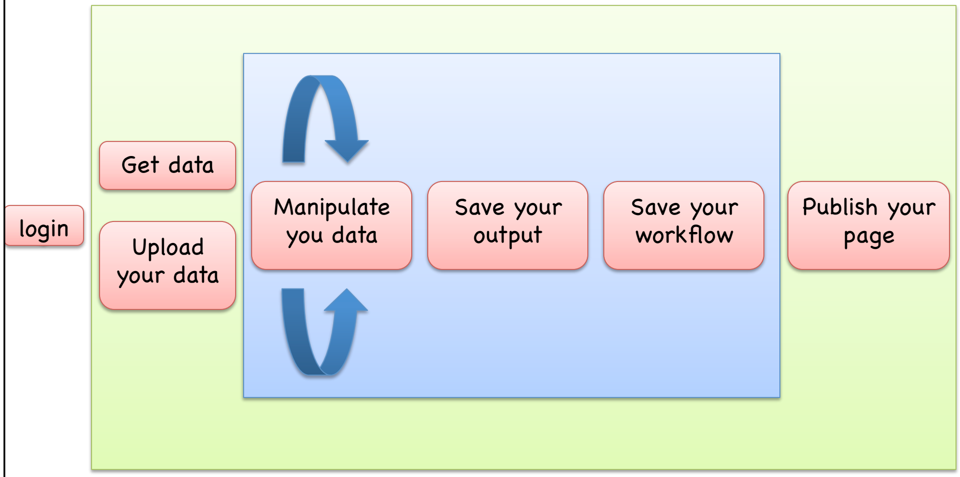


The attribute list must begin with the two mandatory attributes:

- **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.

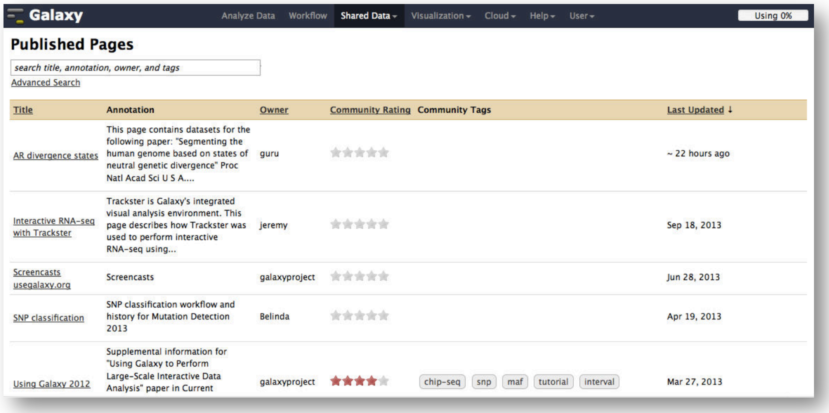
gene_id "Em:U62317.C22.6.mRNA"; transcript_id "Em:U62317.C22.6.mRNA"; exon_number 1

General workflow for Galaxy



Pages in Galaxy

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

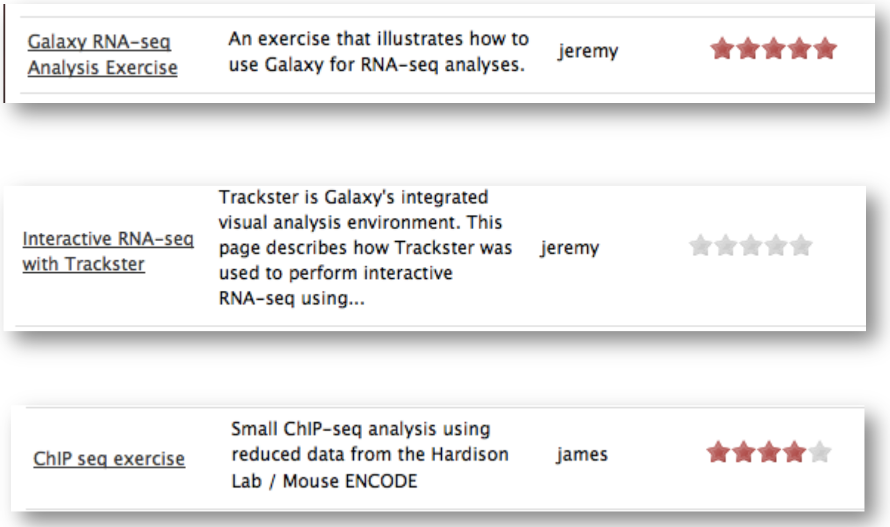


Published Pages

search title, annotation, owner, and tags

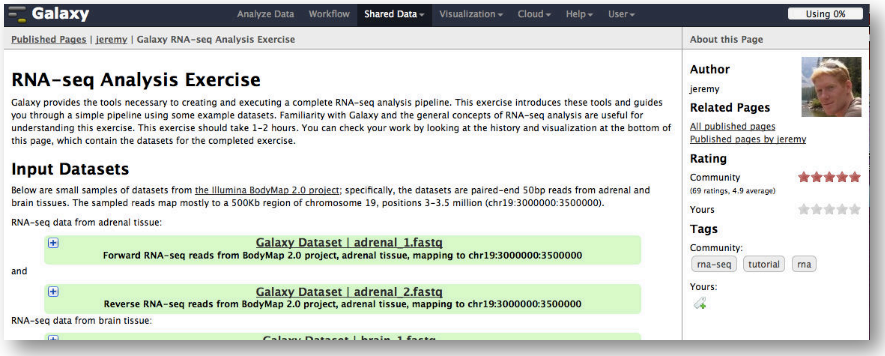
Advanced Search

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



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	★★★★☆

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



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:

RNA-Seq Analysis Exercise

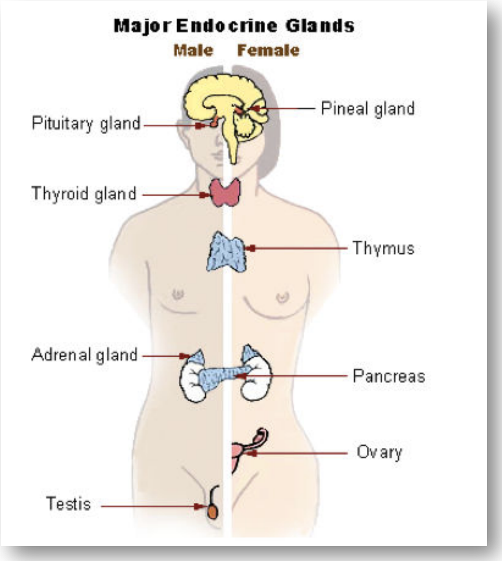
- Human BodyMap 2.0 data from Illumina.
- **paired-end** 50bp reads from **adrenal** and **brain** tissues. The sampled reads map mostly to a 500Kb region of chromosome 19, positions 3-3.5 million (chr19:3000000-3500000).

Human BodyMap 2.0 data from Illumina

Posted on May 24, 2011 by Thibaut (Genebuild) — 33 Comments ↓

I'd like to introduce you an exciting new data set that we've introduced in Ensembl release 62: RNASeq 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, testes, thyroid, and white blood

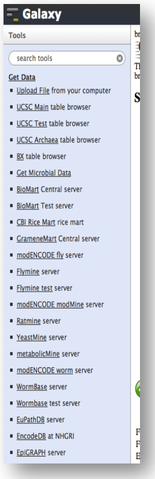
712 people like Ensembl.

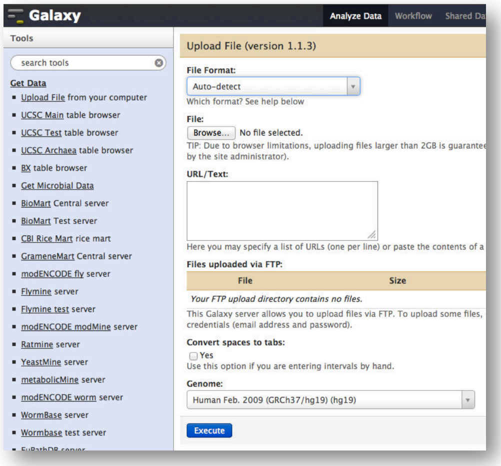


http://en.wikipedia.org/wiki/Adrenal_gland

Getting data

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





Upload File (version 1.1.3)

File Format: Auto-detect

Which format? See help below

File: No file selected.

TIP: Due to browser limitations, uploading files larger than 2GB is guarantee by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a

Files uploaded via FTP:

File	Size
Your FTP upload directory contains no files.	

This Galaxy server allows you to upload files via FTP. To upload some files, credentials (email address and password).

Convert spaces to tabs:

□ Yes

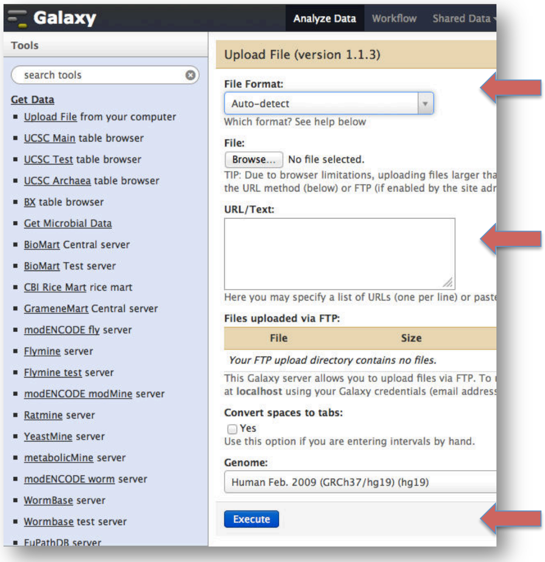
Use this option if you are entering intervals by hand.

Genome: Human Feb. 2009 (GRCh37/hg19) (hg19)

Get 4 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

Load file(s) to Galaxy



Get Data
- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX table browser
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server
- WormBase server
- Wormbase test server
- EuPathDB server

Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:
Browse... No file selected.
TIP: Due to browser limitations, uploading files larger tha
the URL method (below) or FTP (if enabled by the site adr

URL/Text:

Here you may specify a list of URLs (one per line) or paste

Files uploaded via FTP:

File	Size
Your FTP upload directory contains no files.	

This Galaxy server allows you to upload files via FTP. To
at **localhost** using your Galaxy credentials (email address

Convert spaces to tabs:
□ Yes
Use this option if you are entering intervals by hand.

Genome:
Human Feb. 2009 (GRCh37/hg19) (hg19)

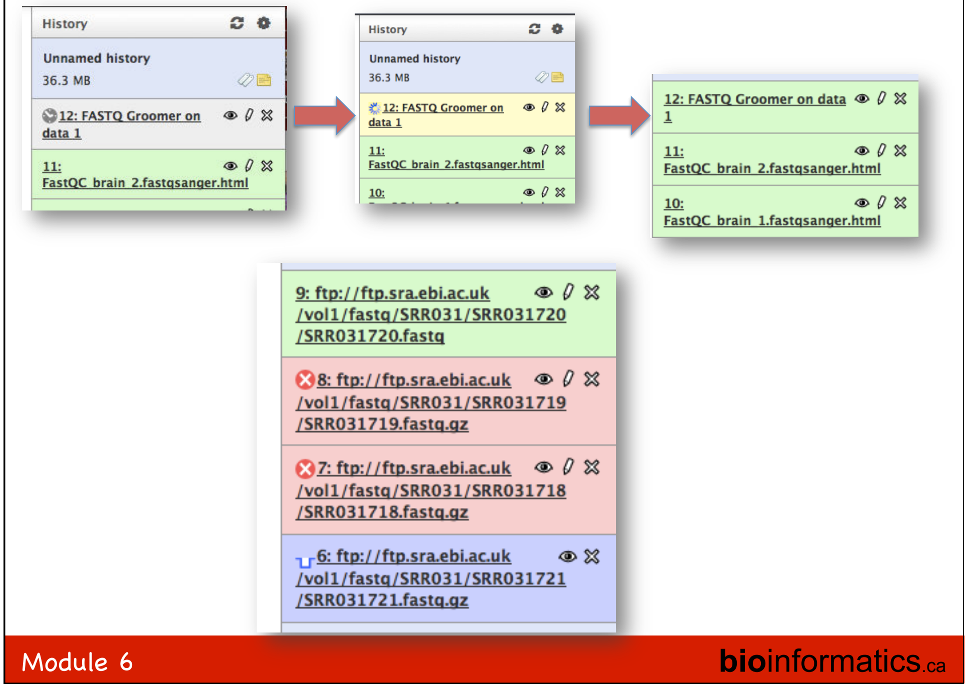
Execute

autodetect

Paste URL
from previous
page

Execute

Module 6 bioinformatics.ca



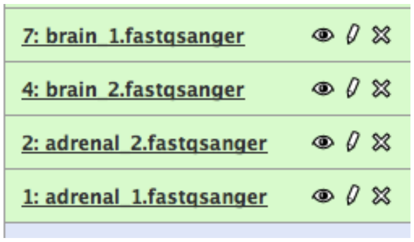
History
Unnamed history
36.3 MB
12: FASTQ Groomer on data 1
11: FastQC_brain_2.fastqsanger.html

History
Unnamed history
36.3 MB
12: FASTQ Groomer on data 1
11: FastQC_brain_2.fastqsanger.html
10:

12: FASTQ Groomer on data 1
11: FastQC_brain_2.fastqsanger.html
10: FastQC_brain_1.fastqsanger.html

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

Module 6 bioinformatics.ca



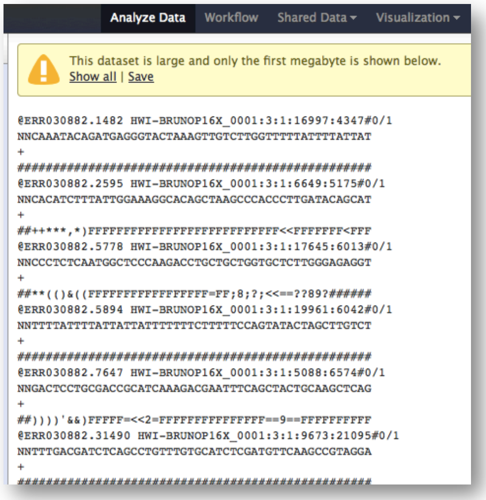
"Poke the eye"

"Edit attribute"

"Delete"

"Numbers may vary with usage"

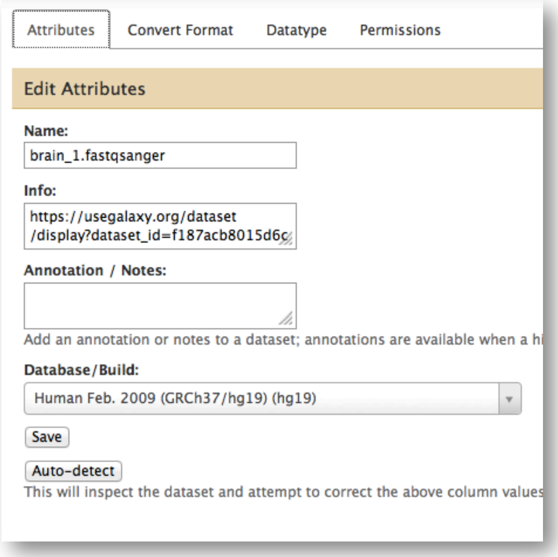
"poke the eye"



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
NNCAAATACAGATGAGGGTACTAAAGTTGTCTTGGTTTTTATTTTATTAT
+
##################################################
@ERR030882.2595 HWI-BRUNOP16X_0001:3:1:6649:5175#0/1
NNCACATCTTTATTGGAAAGGCACAGCTAAGCCCACCCTTGATACAGCAT
+
##++***,*)FFFFFFFFFFFFFFFFFFFFFFFFFFF<<FFFFFFF<FFF
@ERR030882.5778 HWI-BRUNOP16X_0001:3:1:17645:6013#0/1
NNCCCTCTCAATGGCTCCCAAGACCTGCTGCTGGTGCTCTTGGGAGAGGT
+
##**(()&((FFFFFFFFFFFFFFFFF=FF;8;?;<<==??89?######
@ERR030882.5894 HWI-BRUNOP16X_0001:3:1:19961:6042#0/1
NNTTTTATTTTATTATTATTTTTTTCTTTTTCCAGTATACTAGCTTGTCT
+
##################################################
@ERR030882.7647 HWI-BRUNOP16X_0001:3:1:5088:6574#0/1
NNGACTCCTGCGACCGCATCAAAGACGAATTTCAGCTACTGCAAGCTCAG
+
##))))'&&)FFFFF=<<2=FFFFFFFFFFFFFFF==9==FFFFFFFFFF
@ERR030882.31490 HWI-BRUNOP16X_0001:3:1:9673:21095#0/1
NNTTTGACGATCTCAGCCTGTTTGTGCATCTCGATGTTCAAGCCGTAGGA
+
##################################################
```

"Edit attributes"



Edit Attributes

Name: brain_1.fastqsanger

Info: https://usegalaxy.org/dataset/display?dataset_id=f187acb8015d6c

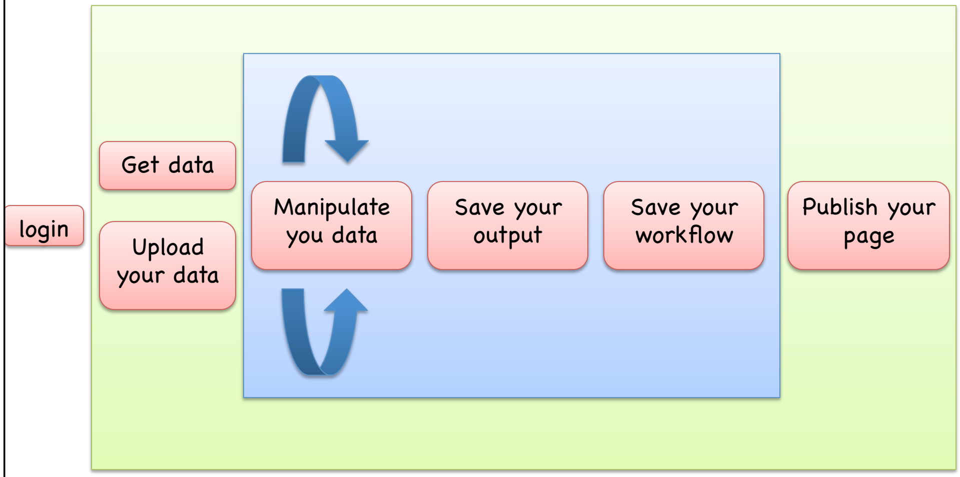
Annotation / Notes:

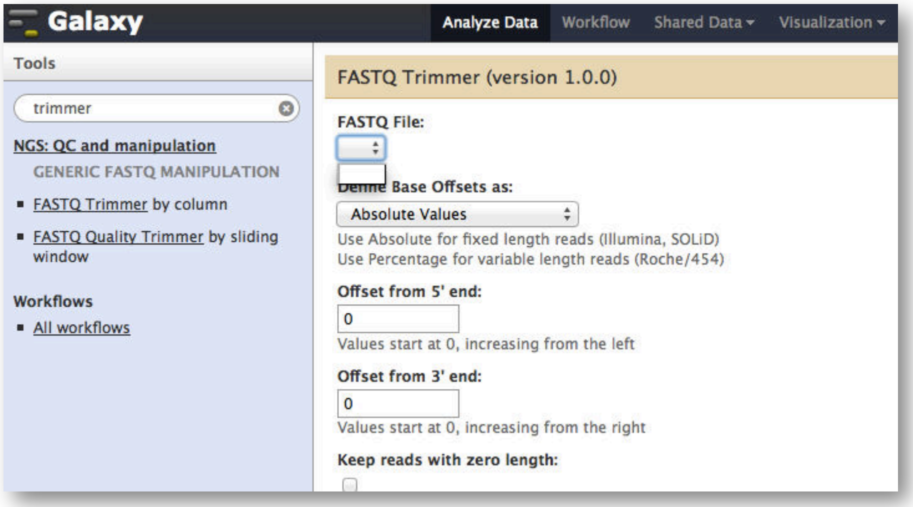
Add an annotation or notes to a dataset; annotations are available when a hi

Database/Build: Human Feb. 2009 (GRCh37/hg19) (hg19)

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

General workflow for Galaxy





FASTQ Trimmer (version 1.0.0)

FASTQ File:

Define Base Offsets as: Absolute Values

Use Absolute for fixed length reads (Illumina, SOLiD)
Use Percentage for variable length reads (Roche/454)

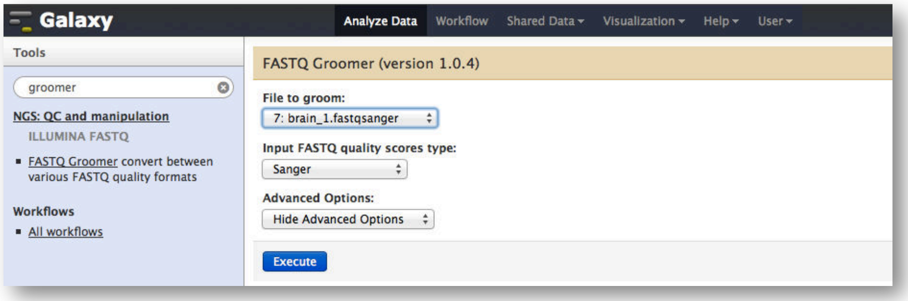
Offset from 5' end: 0

Values start at 0, increasing from the left

Offset from 3' end: 0

Values start at 0, increasing from the right

Keep reads with zero length: □



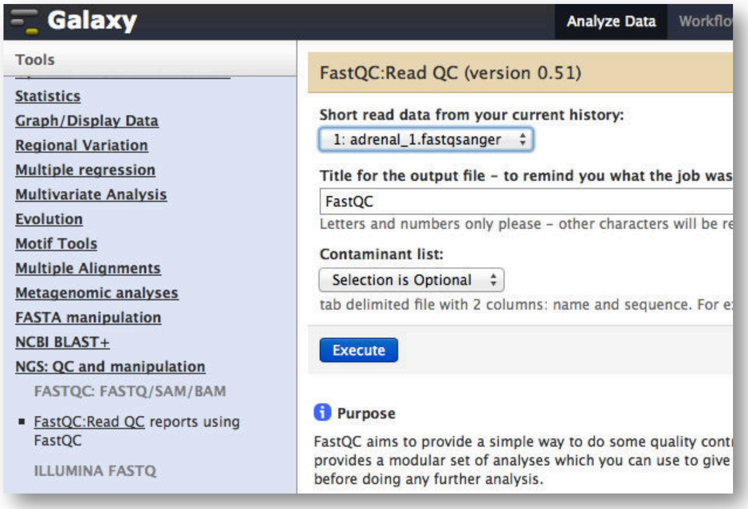
FASTQ Groomer (version 1.0.4)

File to groom: 7: brain_1.fastqsanger

Input FASTQ quality scores type: Sanger

Advanced Options: Hide Advanced Options

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



FastQC:Read QC (version 0.51)

Short read data from your current history:
1: adrenal_1.fastqsanger

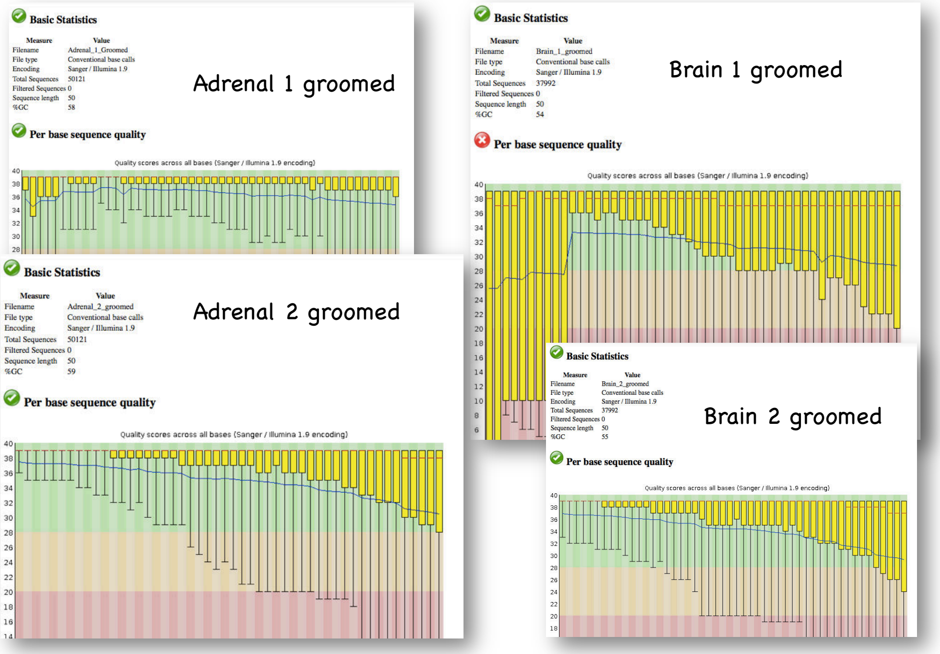
Title for the output file - to remind you what the job was
FastQC
Letters and numbers only please - other characters will be re

Contaminant list:
Selection is Optional
tab delimited file with 2 columns: name and sequence. For e

Purpose

FastQC aims to provide a simple way to do some quality cont
provides a modular set of analyses which you can use to give
before doing any further analysis.

Module 6 bioinformatics.ca



Adrenal 1 groomed

Basic Statistics

Measure	Value
Filename	Adrenal_1_Groomed
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	50121
Filtered Sequences	0
Sequence length	50
%GC	58

Per base sequence quality

Adrenal 2 groomed

Basic Statistics

Measure	Value
Filename	Adrenal_2_groomed
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	50121
Filtered Sequences	0
Sequence length	50
%GC	59

Per base sequence quality

Brain 1 groomed

Basic Statistics

Measure	Value
Filename	Brain_1_groomed
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	37992
Filtered Sequences	0
Sequence length	50
%GC	54

Per base sequence quality

Brain 2 groomed

Basic Statistics

Measure	Value
Filename	Brain_2_groomed
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	37992
Filtered Sequences	0
Sequence length	50
%GC	55

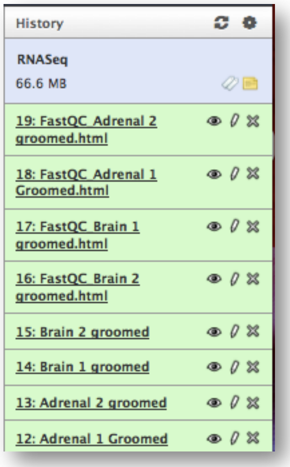
Per base sequence quality

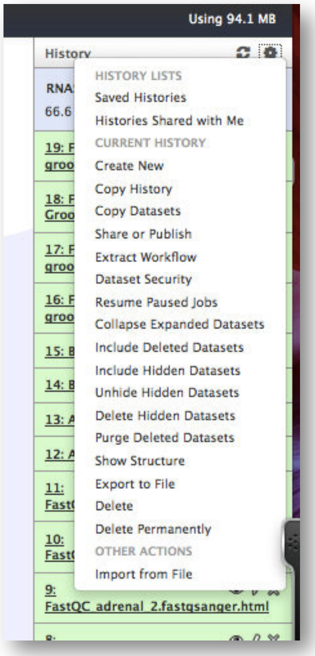
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6

Module 6 bioinformatics.ca

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

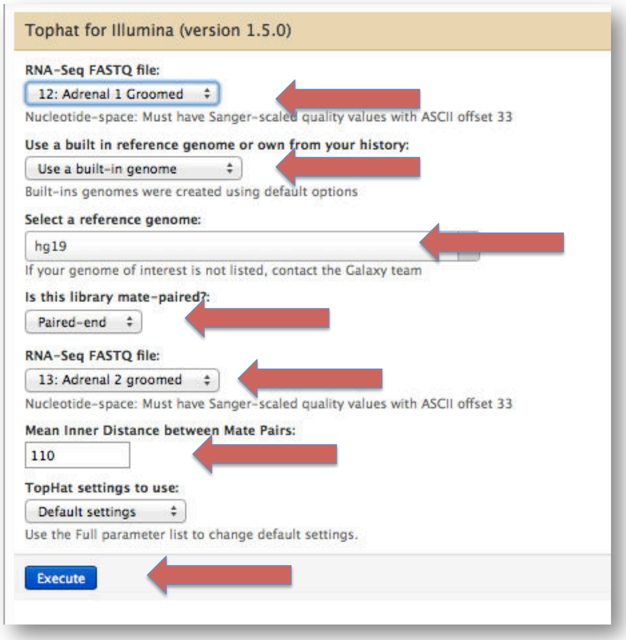

History
RNASeq
66.6 MB
19: FastQC Adrenal 2 groomed.html
18: FastQC Adrenal 1 Groomed.html
17: FastQC Brain 1 groomed.html
16: FastQC Brain 2 groomed.html
15: Brain 2 groomed
14: Brain 1 groomed
13: Adrenal 2 groomed
12: Adrenal 1 Groomed



"Numbers may vary with usage"

[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: 12: Adrenal 1 Groomed

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

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

Built-ins genomes were created using default options

Select a reference genome: hg19

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

Is this library mate-paired?: Paired-end

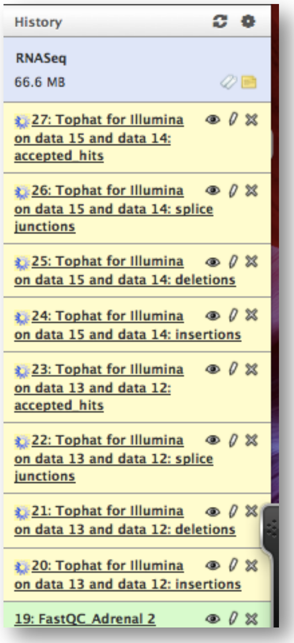
RNA-Seq FASTQ file: 13: Adrenal 2 groomed

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

Mean Inner Distance between Mate Pairs: 110

TopHat settings to use: Default settings

Use the Full parameter list to change default settings.



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

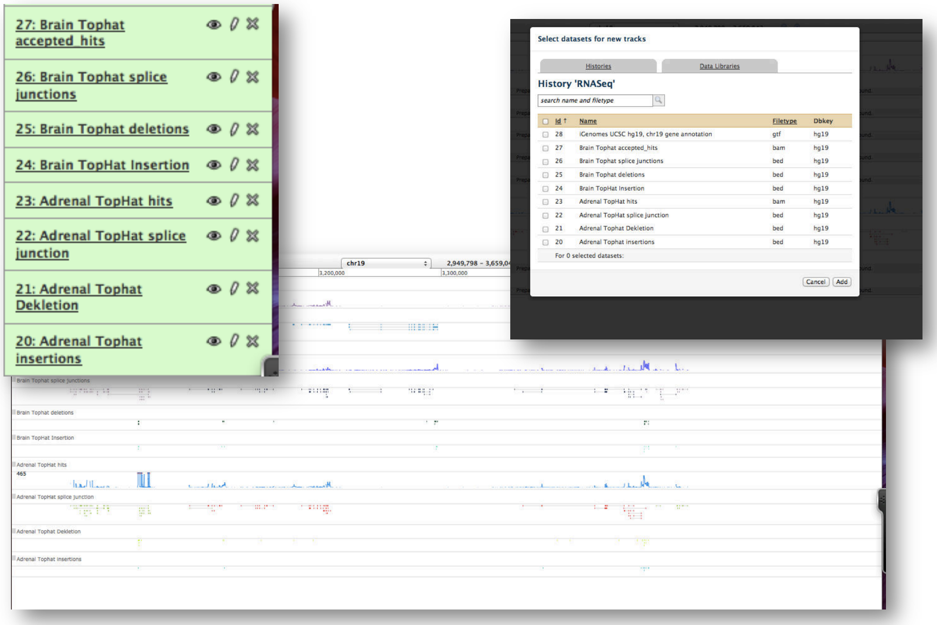
22: Tophat for Illumina on data 13 and data 12: splice junctions

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

This takes about ~ 30 minutes …



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 Dekletion

20: Adrenal Tophat insertions

Select datasets for new tracks

Histories | Data Libraries

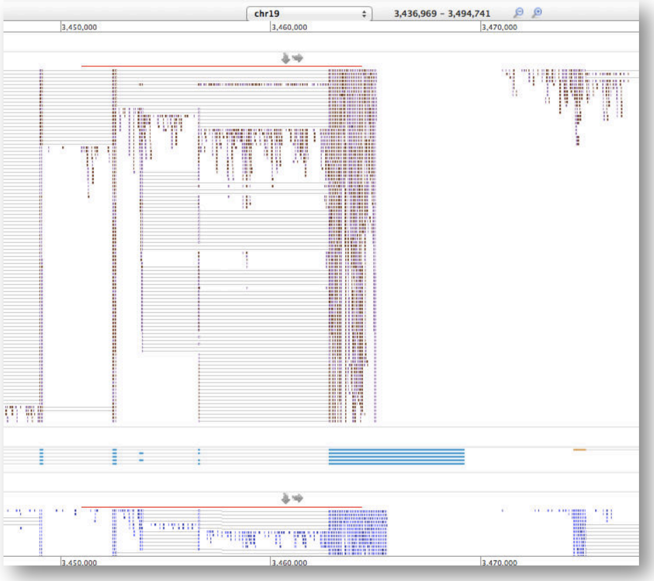
History 'RNASeq'

search name and filetype

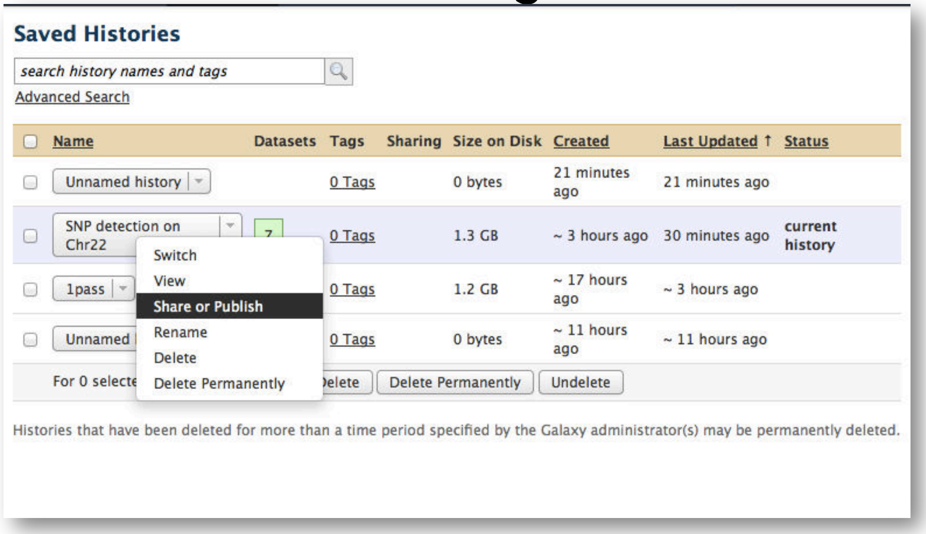
□ Id ↑	Name	Filetype	Dbkey
□ 28	iGenomes UCSC hg19, chr19 gene annotation	gtf	hg19
□ 27	Brain Tophat accepted_hits	bam	hg19
□ 26	Brain Tophat splice junctions	bed	hg19
□ 25	Brain Tophat deletions	bed	hg19
□ 24	Brain TopHat Insertion	bed	hg19
□ 23	Adrenal TopHat hits	bam	hg19
□ 22	Adrenal TopHat splice junction	bed	hg19
□ 21	Adrenal Tophat Dekletion	bed	hg19
□ 20	Adrenal Tophat insertions	bed	hg19

For 0 selected datasets:

Cancel | Add



sharing



Saved Histories

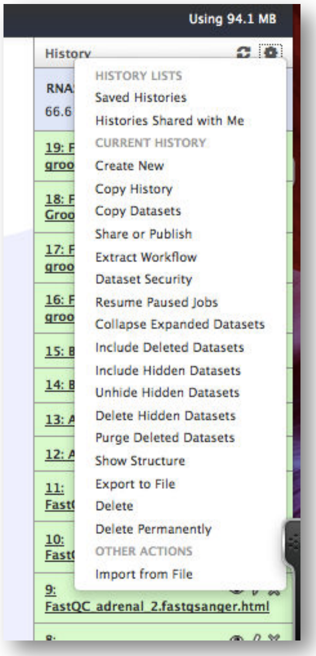
search history names and tags

Advanced Search

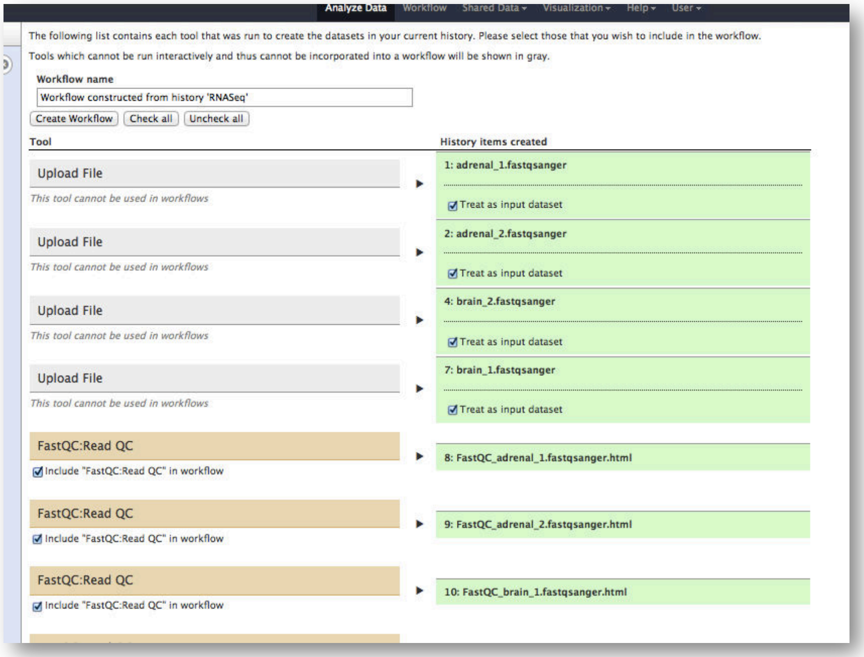
□	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
□	Unnamed history		0 Tags		0 bytes	21 minutes ago	21 minutes ago	
□	SNP detection on Chr22	7	0 Tags		1.3 GB	~ 3 hours ago	30 minutes ago	current history
□	1pass		0 Tags		1.2 GB	~ 17 hours ago	~ 3 hours ago	
□	Unnamed		0 Tags		0 bytes	~ 11 hours ago	~ 11 hours ago	

For 0 selecte elete | Delete Permanently | Undelete

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



- Share history with colleagues
- Extract workflow



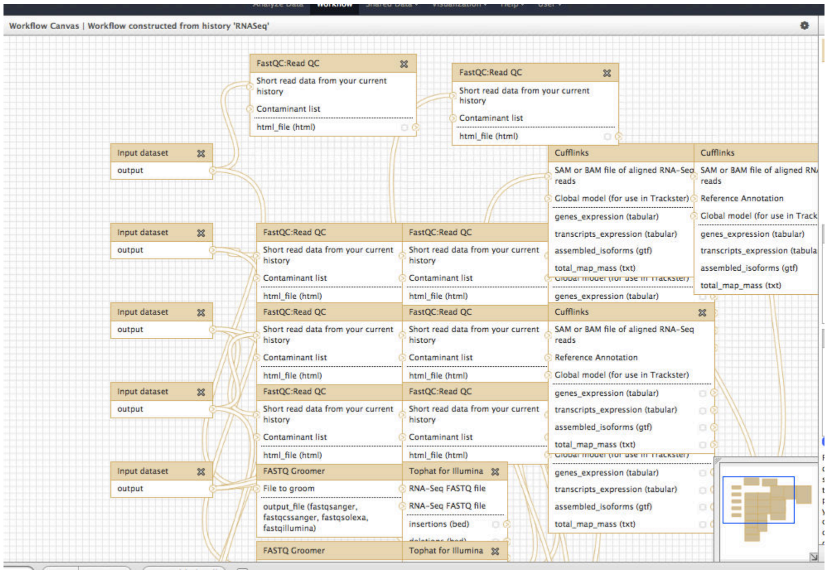
The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

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'

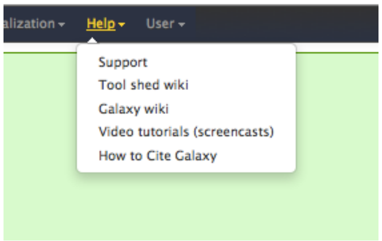
Create Workflow | Check all | Uncheck all

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

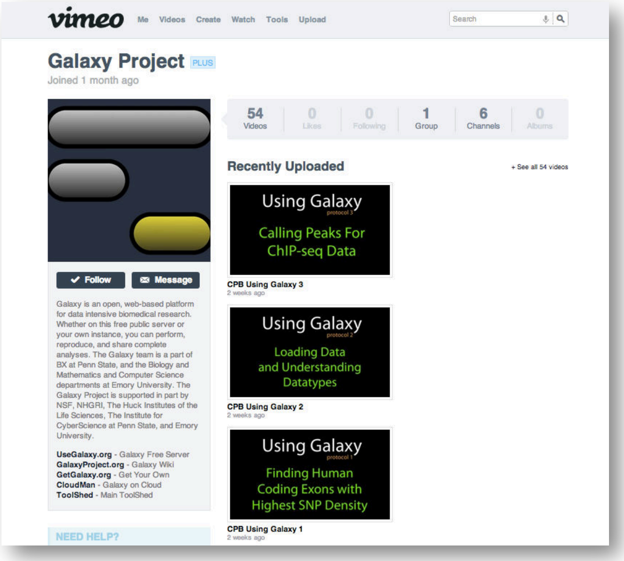

Workflow Canvas | Workflow constructed from history 'RNASeq'
Input dataset
output
FastQC:Read QC
Short read data from your current history
Contaminant list
html_file (html)
FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
Tophat for Illumina
RNA-Seq FASTQ file
insertions (bed)
Cufflinks
SAM or BAM file of aligned RNA-Seq reads
Reference Annotation
Global model (for use in Trackster)
genes_expression (tabular)
transcripts_expression (tabular)
assembled_isoforms (gtf)
total_map_mass (txt)

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

- https://vimeo.com/galaxyproject



https://vimeo.com/galaxyproject



Galaxy Project PLUS

Joined 1 month ago

54	0	0	1	6	0
Videos	Likes	Following	Group	Channels	Albums

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. 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, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

UseGalaxy.org - Galaxy Free Server
GalaxyProject.org - Galaxy Wiki
GetGalaxy.org - Get Your Own
CloudMan - Galaxy on Cloud
ToolShed - Main ToolShed

Recently Uploaded

+ See all 54 videos

Using Galaxy protocol 3 Calling Peaks For ChIP-seq Data
CPB Using Galaxy 3
2 weeks ago

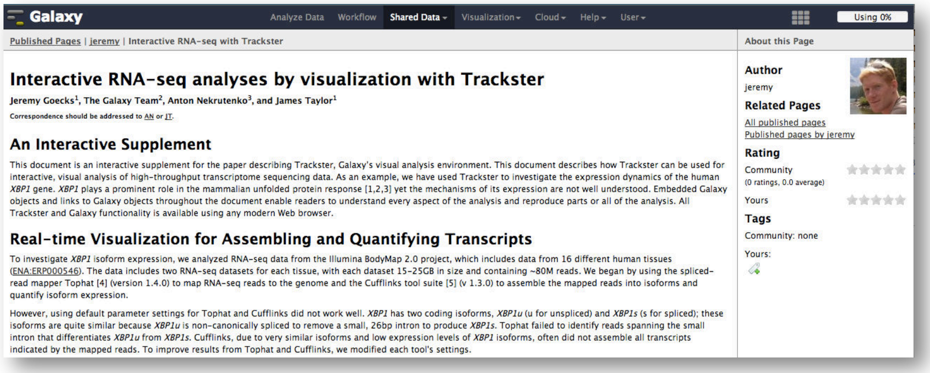
Using Galaxy protocol 2 Loading Data and Understanding Datatypes
CPB Using Galaxy 2
2 weeks ago

Using Galaxy protocol 1 Finding Human Coding Exons with Highest SNP Density
CPB Using Galaxy 1
2 weeks ago

Using Galaxy
protocol 3

Calling Peaks For ChIP-seq Data

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



Interactive RNA-seq analyses by visualization with Trackster

Jeremy Goecks[1], The Galaxy Team[2], Anton Nekrutenko[3], and James Taylor[1]

Correspondence should be addressed to AN or JT.

An Interactive Supplement

This document is an interactive supplement for the paper describing Trackster, Galaxy's visual analysis environment. This document describes how Trackster can be used for interactive, visual analysis of high-throughput transcriptome sequencing data. As an example, we have used Trackster to investigate the expression dynamics of the human *XBP1* gene. *XBP1* plays a prominent role in the mammalian unfolded protein response [1,2,3] yet the mechanisms of its expression are not well understood. Embedded Galaxy objects and links to Galaxy objects throughout the document enable readers to understand every aspect of the analysis and reproduce parts or all of the analysis. All Trackster and Galaxy functionality is available using any modern Web browser.

Real-time Visualization for Assembling and Quantifying Transcripts

To investigate *XBP1* isoform expression, we analyzed RNA-seq data from the Illumina BodyMap 2.0 project, which includes data from 16 different human tissues (ENA:ERP000546). The data includes two RNA-seq datasets for each tissue, with each dataset 15-25GB in size and containing ~80M reads. We began by using the spliced-read mapper Tophat [4] (version 1.4.0) to map RNA-seq reads to the genome and the Cufflinks tool suite [5] (v 1.3.0) to assemble the mapped reads into isoforms and quantify isoform expression.

However, using default parameter settings for Tophat and Cufflinks did not work well. *XBP1* has two coding isoforms, *XBP1u* (u for unspliced) and *XBP1s* (s for spliced); these isoforms are quite similar because *XBP1u* is non-canonically spliced to remove a small, 26bp intron to produce *XBP1s*. Tophat failed to identify reads spanning the small intron that differentiates *XBP1u* from *XBP1s*. Cufflinks, due to very similar isoforms and low expression levels of *XBP1* isoforms, often did not assemble all transcripts indicated by the mapped reads. To improve results from Tophat and Cufflinks, we modified each tool's settings.

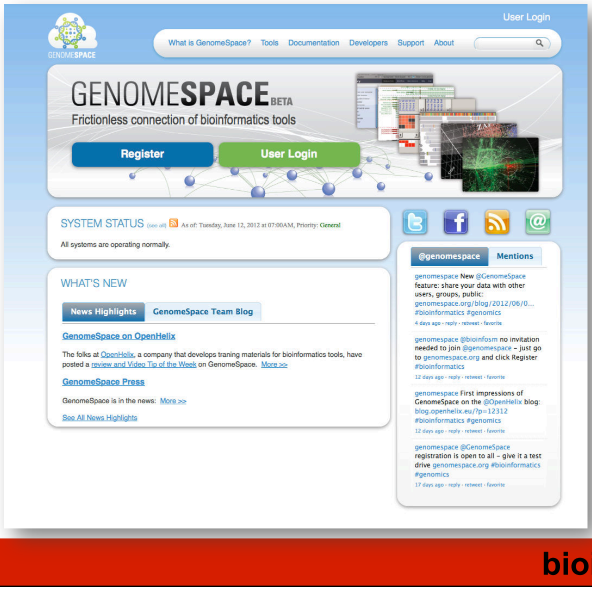
Author
jeremy

Related Pages
All published pages
Published pages by jeremy

Rating
Community (0 ratings, 0.0 average) ☆☆☆☆☆
Yours ☆☆☆☆☆

Tags
Community: none
Yours:

http://genomespace.org/



Frictionless connection of bioinformatics tools

SYSTEM STATUS (see all) As of: Tuesday, June 12, 2012 at 07:00AM, Priority: General

All systems are operating normally.

WHAT'S NEW

GenomeSpace on OpenHelix

The folks at OpenHelix, a company that develops traning materials for bioinformatics tools, have posted a review and Video Tip of the Week on GenomeSpace. More >>

GenomeSpace Press

GenomeSpace is in the news: More >>

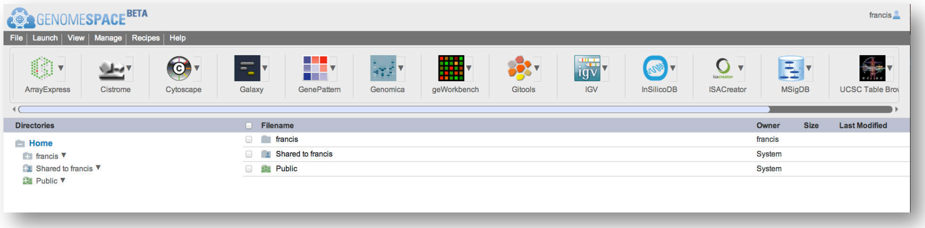
See All News Highlights

genomespace New @GenomeSpace feature: share your data with other users, groups, public: genomespace.org/blog/2012/06/0... #bioinformatics #genomics
4 days ago · reply · retweet · favorite

genomespace @bioinfosm no invitation needed to join @genomespace – just go to genomespace.org and click Register #bioinformatics
12 days ago · reply · retweet · favorite

genomespace First impressions of GenomeSpace on the @OpenHelix blog: blog.openhelix.eu/?p=12312 #bioinformatics #genomics
12 days ago · reply · retweet · favorite

genomespace @GenomeSpace registration is open to all – give it a test drive genomespace.org #bioinformatics #genomics
17 days ago · reply · retweet · favorite



□ Filename	Owner	Size	Last Modified
□ francis	francis		
□ Shared to francis	System		
□ Public	System		

ArrayExpress:	http://www.ebi.ac.uk/arrayexpress/
Cystrome:	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/

Useful Resources

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

Galaxy Project
@galaxyproject Penn State | Emory
Galaxy = bringing developers and biologists together. Reproducible science is our goal.
http://usegalaxy.org

- BioStaR
  - biostars.org
  - Twitter: @biostarquestion

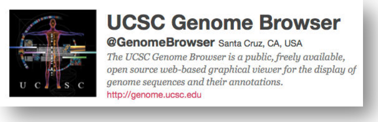
★ BioStar

Useful Resources

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

OpenHelix Staff
@OpenHelix US
Training on Bioinformatics and Genomics Resources
www.openhelix.com
http://blog.openhelix.com/

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



UCSC Genome Browser
@GenomeBrowser Santa Cruz, CA, USA
The UCSC Genome Browser is a public, freely available, open source web-based graphical viewer for the display of genome sequences and their annotations.
http://genome.ucsc.edu

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

SEQanswers
the next generation sequencing community

Module 6 bioinformatics.ca

Papers of interest:

- Robert Gentleman, 2005, Reproducible research: a bioinformatics case Source, 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

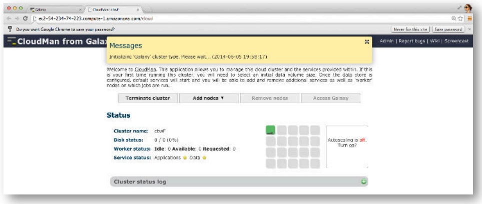
  http://www.ncbi.nlm.nih.gov/pubmed/?term=20738864
- Afgan E, Chapman B, Jadan M, Franke V, Taylor J. (2012) Using cloud computing infrastructure with CloudBioLinux, CloudMan, and Galaxy. Curr Protoc Bioinformatics. Chapter 11:Unit11.9. doi: 10.1002/0471250953.bi1109s38.

  http://www.ncbi.nlm.nih.gov/pubmed/22700313
- Goecks J1, Eberhard C, Too T; Galaxy Team, Nekrutenko A, Taylor J. (2013) Web-based visual analysis for high-throughput genomics.
  BMC Genomics. 2013 Jun 13;14:397
  http://www.ncbi.nlm.nih.gov/pubmed/23758618

Module 6 bioinformatics.ca

Before Coffee Break

- Go to page 20 (or there about) and do: **Log onto Galaxy**
- Login info will be on wiki (at that time)
- Once you have this image, you can go on break:



Messages

Initializing 'Galaxy' cluster type. Please wait... (2014-06-05 19:58:17)

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 'worker' nodes on which jobs are run.

Status

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

Autoscaling is off. Turn on?

After Break we will be doing lab

- Want to acknowledge Florence Cavalli and Zhibin Lu for great work they have done to help me with the cloud, some of the slides and with the accuracy of the slides.
- That said, all errors, mistakes, old URLs etc are my fault, entirely!

@bffo

We are on a Coffee Break & Networking Session

- For those of you not here, watching video, maybe you want to register for workshop?
- More details at http://bioinformatics.ca