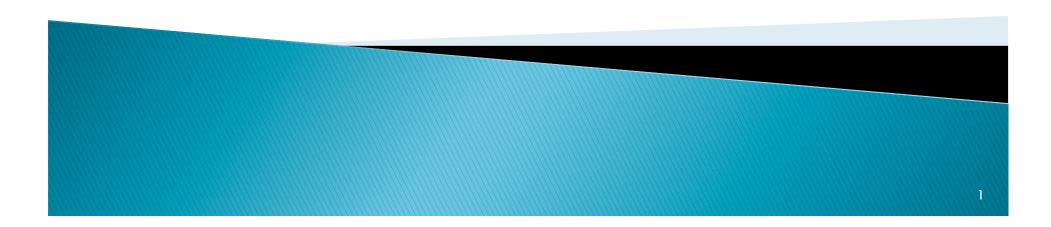


# NGS Analysis Using Galaxy



## Outline

- What is Galaxy
- Galaxy for Bioinformaticians
- Galaxy for Experimental Biologists
- Using Galaxy for NGS Analysis
- NGS Data Visualization and Exploration Using IGV



## Outline

#### What is Galaxy

- Galaxy for Bioinformaticians
- Galaxy for Experimental Biologists
- Using Galaxy for NGS Analysis
- NGS Data Visualization and Exploration Using IGV

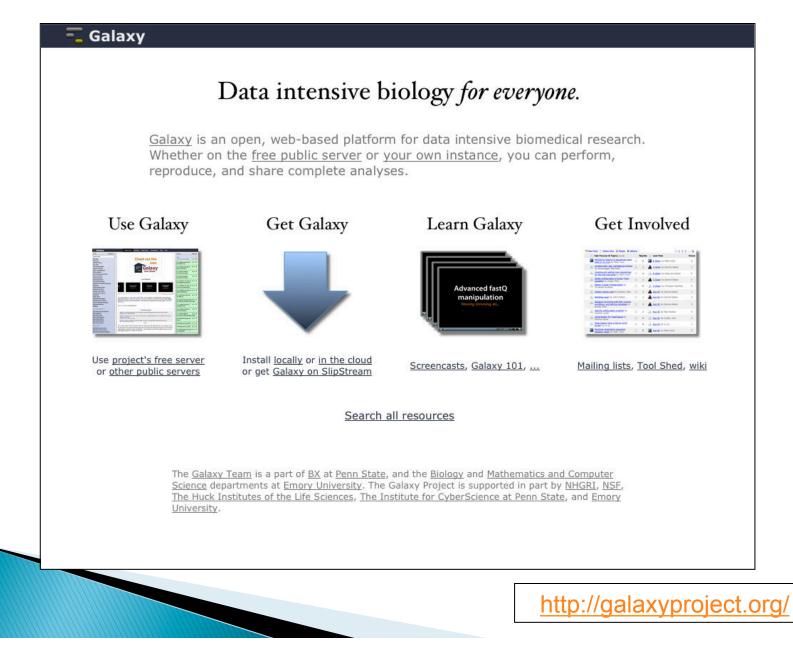


Galaxy, a web-based genome analysis platform

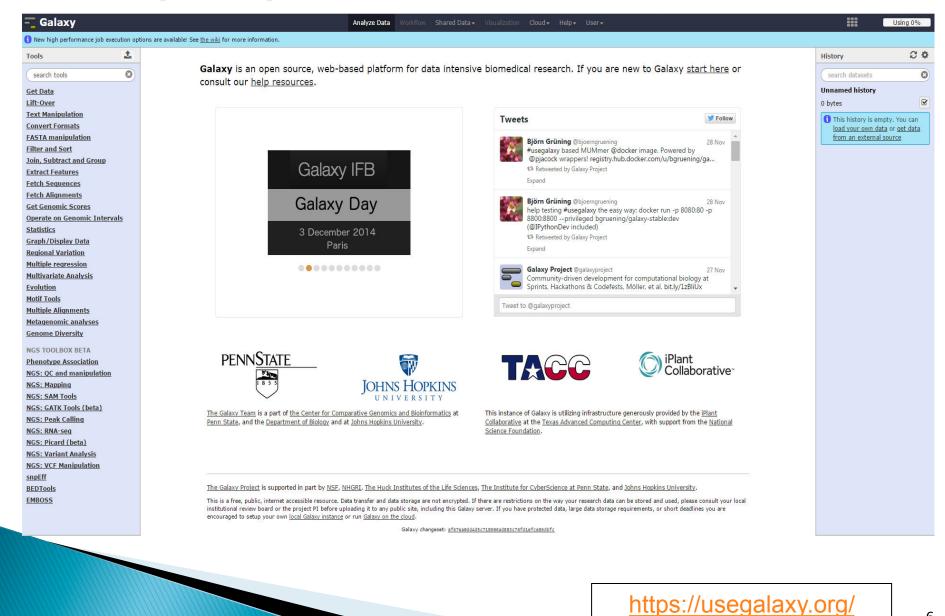
- Galaxy is an open-source framework for integrating various computational tools and databases into a cohesive workspace.
- A web-based service we provide, integrating many popular tools and resources for comparative genomics.
- A completely self-contained application for building your own Galaxy style sites.



### **Galaxy Project Interface**



## Galaxy analysis web interface



## Outline

- What is Galaxy
- Galaxy for Bioinformaticians
- Galaxy for Experimental Biologists
- Using Galaxy for NGS Analysis
- NGS Data Visualization and Exploration Using IGV



Galaxy: the instant web-based tool and data resource integration platform

- Open Source downloadable package that can be deployed in individual labs
- Modularized
  - Add new tools
  - Integrate new data sources
  - Easy to plug in your own components
- Straightforward to run your own private galaxy server



## Galaxy instance

#### https://galaxy.bioinfo.ucr.edu/root

- Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -	Using 0%	
Tools		🚺 History 📿 🕇	0
search tools	UCR Institute for Integrative Genome Biology	Unnamed history 0 bytes	A
<u>Get Data</u>			5
Send Data		Your history is empty. Click 'Get	
ENCODE Tools		Data' on the left pane to start	
<u>Lift-Over</u>	Welcome to IIGB's Galaxy Server!		
Text Manipulation	Overview		
Filter and Sort	Galaxy is an open, highly customizable, web-based platform for the analysis of next generation sequence data and many other biological data types. It enables users to run computationally		
Join, Subtract and Group	demanding next generation sequencing analysis tasks on powerful server hardware from a		
Convert Formats	graphical web browser-based user interface rather than the Linux command-line. A subset of of		
Extract Features Fetch Sequences	application supported by Galaxy is given in the left pane. Much more detailed descriptions of Galaxy's basic functionalities including user tutorials are available here.		
Fetch Alignments	Galaxy's basic functionalities including user tutorials are available <u>inere</u> .		
Get Genomic Scores	Why Local Galaxy Service?		
Operate on Genomic Intervals	There are many advantages of using a local Galaxy server here at UCR rather than public test		
Statistics	instances of Galaxy available on the internet. The most important are: (1) shorter waiting queues for analysis tasks; (2) elimination of time consuming uploads of large data sets; (3) support for		
Wavelet Analysis	analyzing much larger data sets than this is possible on public services; (4) the ability to		
Graph/Display Data	customize software tools and database collections.		
Regional Variation			
Multiple regression	How to Gain Access? This instance of Galaxy runs on IIGB's high performance compute (HPC) infrastructure, called		
Multivariate Analysis	Biocluster. As such its usage is covered by the annual registration fee for this infrastructure (see		
Evolution	here for details). Users with an active Biocluster account can access this Galaxy service using		
Motif Tools	their existing user name and password without any extra cost. New account requests for this service can be sent to support@biocluster.ucr.edu.		
Multiple Alignments	service can be sent to <u>supportionioeuster.uer.eou</u> .		
Metagenomic analyses	Additional Databases and Sofware Tools		
FASTA manipulation	Support requests for including additional reference genomes and software tools on IIGB's Galaxy		
NGS: QC and manipulation	server can be sent to <u>support@biocluster.ucr.edu</u>		
NGS: Mapping	Workshops on Galaxy		
NGS: Indel Analysis	Past and future UCR workshop events on using Galaxy are listed <u>here</u> . The user manual from		
NGS: RNA Analysis	previous workshops can be accessed <u>here</u> .		
NGS: SAM Tools			
NGS: GATK Tools (beta)	Enter IIGB's Galaxy Service To enter this service, click here.		
NGS: Peak Calling	ro enter this service, enter <u>inere</u> .		
NGS: Simulation			
SNP/WGA: Data; Filters		<b>•</b>	
SNP/WGA: QC; LD; Plots		¥.	

## Outline

- What is Galaxy
- Galaxy for Bioinformaticians
- Galaxy for Experimental Biologists
- Using Galaxy for NGS Analysis
- NGS Data Visualization and Exploration Using IGV



#### Galaxy - the one stop shop for Genome Analysis

- Analyze
  - Retrieve data directly from popular data resources or upload your own.
  - Interactively manipulate genomic data with a comprehensive and expanding best-practices toolset.
  - Galaxy is designed to work with many different datatypes. (Link)

#### Visualize

- Trackster is Galaxy's visualization and visual analysis environment.
- See more details (Link)
- Publish and Share
  - Results and step-by-step analysis record (Data Libraries and Histories)
  - Customizable pipelines (Workflows)
  - Complete protocols (Pages)



## **Tools and Data Sources**

#### Data Sources

- Upload file from your computer
- UCSC table browser
- BioMart, modENCODE, GrameneMart, WormBase servers.....

#### Tool Suites

- Text manipulation
- Join, Subtract and Group
- Format converters
- NGS
- Graph plotting
- Motif tools



#### Data Libraries

> Datasets are accessible from within Galaxy or for download.

💳 Galaxy	Analyze Data Workflow Shared Data - Visualization - Cloud - Help - User -
Data Libraries	
search dataset name, info, message,	dbkey
Advanced Search	
Data library name ↓	Data library description
1000 Genomes	Data from the 1000 Genomes Project FTP site
AC-exome	
<u>Bushman</u>	Data for Nature Letter "Complete Khoisan and Bantu genomes from southern Africa"
ChIP-Seq Mouse Example	Data used in examples that demonstrate analysis of ChIP-Seq data
<u>Chobi</u>	
<u>CloudMap</u>	Contains userguide, reference files, and configuration files for the Cloudmap WGS analysis pipeline
Codon Usage Frequencies	
<u>Coleman</u>	IonPGM
Denisovan sequences	Files from 'A high-coverage genome sequence from an archaic Denisovan Individual" Meyer et al. Science 2012 and basic processed data.
Erythroid Epigenetic Landscape	Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration
Evolutionary Trajectories in a Phage	Experimental evolution (Illumina)
GATK	
GCAT	Consortium
Genome Diversity	Nucleotide polymorphisms for several threatened species
guru 1000GP	
<u>He-2010</u>	
<u>Heteroplasmy</u>	Data for Genome Biology 2011 manuscript
iGenomes	Selected files from Illumina iGenomes collection
Illumina BodyMap 2.0	RNA-seq data for the Illumina BodyMap 2.0 project
<u>Illumina iDEA Datasets</u> (sub-sampled)	Sub-samapled versions of datasets used for the Illumina iDEA challenge
Irish whole genome	Irish whole genome sequence and analysis

## Workflows

- Workflows specify the steps in a process.
- Workflows are analysis that are meant to be run, each time with different user-provided datasets.

Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -
Tools	Running workflow "Workflow constructed from history [Expand All Collapse] 'SNPseq_Analysis_Galaxy_workshop2013'''
( search tools O	SNFSEQ_Analysis_Galaxy_workshop2015
<u>Get Data</u>	Step 1: Input dataset
Send Data	Input Dataset 🗇
ENCODE Tools	
Lift-Over Text Manipulation	
Filter and Sort	type to filter
Join, Subtract and Group	
Convert Formats	Step 2: Input dataset
Extract Features	
Fetch Sequences	Input Dataset
Fetch Alignments	7: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/tair10chr.fas
Get Genomic Scores	type to filter
Operate on Genomic Intervals	
<u>Statistics</u>	Step 3: FASTQ Groomer (version 1.0.4)
Wavelet Analysis	
Graph/Display Data	Step 4: FASTQ Summary Statistics (version 1.0.0)
Regional Variation	
Multiple regression	Step 5: FastQC:Read QC (version 0.51)
Multivariate Analysis Evolution	
Motif Tools	Short read data from your current history
Multiple Alignments	Output dataset 'output_file' from step 3
Metagenomic analyses	Title for the output file – to remind you what the job was for FastOC
FASTA manipulation	
NGS: QC and manipulation	Contaminant list
NGS: Mapping	Selection is Optional
NGS: Indel Analysis	
NGS: RNA Analysis	Step 6: Filter FASTQ (version 1.0.0)
NGS: SAM Tools	
NGS: GATK Tools (beta)	Step 7: Map with BWA for Illumina (version 1.2.3)
NGS: Peak Calling	
NGS: Simulation	Step 8: SAM-to-BAM (version 1.1.2)
SNP/WGA: Data; Filters SNP/WGA: QC; LD; Plots	
SNP/WGA: Statistical Models	Step 9: MPileup (version 0.0.1)
Phenotype Association	
VCF Tools	Step 10: bcftools view (version 0.0.1)
Workflows	Send results to a new history
All workflows	
	Run workflow

## Pages

Pages are documentation within the Galaxy that explain the steps and reasoning in a particular history or workflow.

- Galaxy	Analyze Data Workflow Shared Data - Visualization Cloud - Help -	User <del>~</del>			Using 0%
Published Pages search title, annotation, owner, and tags Advanced Search					
Title	Annotation	<u>Owner</u>	Community Rating	Community Tags	Last Updated
Galaxy RNA-seg Analysis Exercise	An exercise that illustrates how to use Galaxy for RNA-seq analyses.	jeremy	****	(ma-seq) (tutorial) (ma) (sl	Nov 18, 2014
honeybee	page associated with Fuller et al. (2014)	webb	****		Sep 28, 2014
Cancer Analyses	Page describes data and workflows in Goecks et al.'s 2014 paper on integrated cancer genomics with Galaxy.	jeremy	*****		Aug 21, 2014
Raisins	NGS Quality Control Using Galaxy: Training Day, GCC 2014	usinggalaxy4	****		Jul 02, 2014
NGS Analysis 2013	Supplemental Tutorials "MiMB: Analysis of Next-Generation Sequencing Data Using Galaxy"	galaxyproject	***		Jun 18, 2014
Extract Workflow	Tutorial: Extract Workflow from a History	galaxyproject	*****		May 13, 2014
Tutorial (Yodosha, 2014)		kawaji	***		May 08, 2014
Galaxy 101 NGS: Introduction to Polymorphism Detection via Variant Analysis	Galaxy 101 NGS Tutorial: Heteroplasmy: Mother-Child mtDNA Variant Polymorphism Detection Step-by- Step	galaxyproject	*****	ngs variant tutorial galaxy101 video	Apr 30, 2014
Controlling for Contamination in Resequencing	Biotechniques 56(3) companion page	aun1	****		Mar 18, 2014
CloudMap	A Cloud-based Pipeline for Analysis of Mutant Genome Sequences	gm2123	***		Feb 08, 2014
Divergent functions of hematopoietic transcription factors in lineage priming and differentiation during erythro-megakaryopoiesis	paper companion site	csm165	****		Jan 15, 2014
FinalProject.Fridland	Elucidation of Conserved and Nonconserved Residues in MspA and Similar Portal Proteins	sfridland	****		Dec 15, 2013
Self-Binding GPCR Seach		bnorgeot	****		Dec 14, 2013
Screencast videos for usegalaxy.org	UseGalaxy.org Screencast Videos at Vimeo	galaxyproject	****		Dec 03, 2013
Using Galaxy 2012	Supplemental information for "Using Galaxy to Perform Large-Scale Interactive Data Analysis" paper in Current Protocols in Bioinformatics, unit 10.5	galaxyproject	****	chip-seq snp maf tutorial interval	Dec 03, 2013
<u>OianYuBioinformatics</u>	BioinformaticsProblemSet	clark	****		Nov 27, 2013
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	*****		Oct 23, 2013
Interactive RNA-seg 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
SNP classification	SNP classification workflow and history for Mutation Detection 2013	Belinda	****		Apr 19, 2013
pipeline	data sets and workflows for the paper "High-throughput analysis of large and possibly unassembled genomes"	webb	*****		Mar 18, 2013
ave-ave	data sets and workflows for the paper "Aye-aye population genomic analyses highlight an important center of endemism in northern Madagascar" (PNAS)	webb	*****		Mar 17, 2013

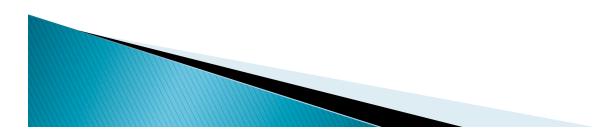
#### History

- Histories are all steps in the process and the used setting.
- Histories can be imported into your session and rerun as it is or modified.

Galaxy	Analyze Data Workflow Shared Data - Visualization - Admin Help - User -	Using 0%
Tools       search tools       Get Data       Send Data       ENCODE Tools       Lift-Over       Text Manipulation	Import and start using history   refresh   show deleted   collapse all         RNASeq_Analysis_Galaxy_workshop_2013         1: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064154.fastq         215.4 MB         format: fastq, database: ?         Info:         uploaded fastq file	History LISTS RNA: Saved Histories op.2 Histories Shared with Me 1.4 C CURRENT HISTORY <u>39: C</u> Create New <u>data</u> Copy History <u>FPKN</u> Copy Datasets
Filter and Sort Join, Subtract and Group Convert Formats Extract Features Fetch Sequences Fetch Alignments Get Genomic Scores Operate on Genomic Intervals Statistics	<pre>@SRR064154.208 HWUSI-EAS627_1:8:1:2:1681 length=38 ANGANNNGGACTTTGAAAAGAGAGTCAAAGAGGCTTG + ?108!!!3C?BCBB<bcbb?bbacabbbbbb@cabab @srr064154.222="" ccgcnnnccactatcacggcagcgctcccacaagc<="" hwusi-eas627_1:8:1:3:1820="" length="38" pre=""></bcbb?bbacabbbbbb@cabab></pre>	38: C       Share or Publish         data       Extract Workflow         diffe       Dataset Security         37: C       Resume Paused Jobs         data       Collapse Expanded Datasets         track       Include Deleted Datasets         36: C       Include Hidden Datasets         data       Unhide Hidden Datasets
Wavelet Analysis Graph/Display Data Regional Variation Multiple regression Multivariate Analysis Evolution Motif Tools Multiple Alignments	3: http://biocluster.ucr.edu/~nkatiyar/Galaxy workshop/Rnaseg/SRR064167.fastq         4: http://biocluster.ucr.edu/~nkatiyar/Galaxy workshop/Rnaseg/tair10chr.fasta         5: http://biocluster.ucr.edu/~nkatiyar/Galaxy workshop/Rnaseg/TAIR10.GTF         6: http://biocluster.ucr.edu/~nkatiyar/Galaxy workshop/Rnaseg/SRR064155.fastq         7: http://biocluster.ucr.edu/~nkatiyar/Galaxy workshop/Rnaseg/SRR064166.fastq	Delete Hidden Datasets <u>35:</u> C Purge Deleted Datasets <u>data</u> Show Structure <u>Export to File</u> <u>34:</u> C Delete <u>data</u> <u>diffe</u> Delete Permanently <u>OTHER ACTIONS</u> <u>33:</u> C Import from File <u>data</u>
Metagenomic analyses FASTA manipulation NGS: QC and manipulation NGS: Mapping NGS: Indel Analysis	8: FASTQ Groomer on data 1     Image: Comparison of the second seco	tracking       32: Cuffdiff on data 19,       data 15, and others: CDS FPKM       differential expression testing

### User Account

- An account is not required to access the Galaxy public Main or Test instances,
- But if used, the data quota is increased and full functionality across sessions opens up, such as naming, saving, sharing, and publishing Galaxy objects (Histories, Workflows, Datasets, Pages).



## Outline

- What is Galaxy
- Galaxy for Bioinformaticians
- Galaxy for Experimental Biologists
- Using Galaxy for NGS Analysis
- NGS Data Visualization and Exploration Using IGV



#### NGS Data

- Raw: Sequencing Reads (FASTQ)
- Derived
  - Alignments against reference genome
    - SAM / BAM
    - VCF / BCF
  - Annotations
    - GFF / GTF
    - BED



### FASTQ Format

- A FASTQ file normally uses four lines per sequence.
- Line 1 begins with a '@' character and is followed by a sequence identifier.
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character, is optionally followed by the same sequence identifier.
- Line 4 encodes the Phred quality values for the sequence in line 2, each value represents the error probability of a given base call.

```
@SRR064154.208 HWUSI-EAS627_1:8:1:2:1681 length=38
ANGANNNGGACTTTGAAAAGAGAGTCAAAGAGTGCTTG
+
?!08!!!3C?BCBB<BCBB?BBACABBBBBBBB@CABAB</pre>
```



## FASTQ Quality Score (Link)

- Quality score represents the error probability of a given basecall.
- In a FASTQ file, quality scores are often represented using the ASCII alphabet.
- For example, a Phred score of 40 can be represented as the ASCII char "I" (40+33= ASCII #73), and an Illumina score of 40 as "h" (40+64=ASCII #104).
- The range of scores will depend on the technology and the base caller used, but will typically be up to 40.

SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS	SSSSSSSSSSSSSS								
·····									
	<b>J</b> JJJJJJJJJJJJJJ	ԵԵԵԵԵԵԵԵԵԵԵԵԵԵԵԵԵԵԵԵ							
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL	LLLLLLLLLLL								
!"#\$%&'()*+,/0123456789:;	<=>?@ABCDEFGHIJKLMN	DPQRSTUVWXYZ[\]^_`abcdefghijl	klmnopqrstuvwxyz{ }~						
1	I I	I	L.						
33 59	64 73	104	126						
0									
-5									
	09								
	39								
0.2	31								
S - Sanger Phred+33,									
	raw reads typicall								
I - Illumina 1.3+ Phred+64,	raw reads typicall;	y (0, 40)							
J - Illumina 1.5+ Phred+64,	raw reads typicall	y (3, 40)							
with 0=unused, 1=unused,	2=Read Segment Qua	lity Control Indicator (bold)	)						
(Note: See discussion ab	ove).								
L - Illumina 1.8+ Phred+33,	raw reads typically	y (O, 41)							

## SAM Format

- SAM stands for Sequence Alignment/Map format.
- For more details:

http://samtools.sourceforge.net/SAM1.pdf

- Consists of header and alignment section
- 11 mandatory fields

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User									
Tools	QNAME FLAG RNAME POS	MAPQCIGAR MRNMMPOSISIZ	ESEQ	QUAL	OPT	History C 4			
search tools	@SQ SN:Chr1 LN:30427671 @SQ SN:Chr2 LN:19698289				20139	SNPseq_Analysis_Galaxy_worksho			
<u>Get Data</u> Send Data	@SQ SN:Chr3 LN:23459830 @SQ SN:Chr4 LN:18585056 @SQ SN:Chr5 LN:26975502					308.4 MB 🖉 🖻			
ENCODE Tools Lift-Over Text Manipulation	@SQ SN:ChrC LN:154478 @SQ SN:ChrM LN:366924					15: bcftools view on data			
Filter and Sort Join, Subtract and Group	@PG ID:bwa PN:bwa VN:0.5.9-r16 SRR038850.12 0 ChrC 2632		CAAGCATCTTTTTTGAATTTCCCATTTATCCGTTTA	@?@<@?@BBAAB@>>===?@AB?7=<:6>@A@:?:6	XT:A:U NM::0 X0:i:1 X1:i:0 XM:i:0 X0:i:0 XG:i:0 MD:Z:36	14: MPileup on data 7 and (1) (log)			
Convert Formats Extract Features	SRR038850.74 0 ChrC 14843 SRR038850.507 16 ChrC 295 SRR038850.576 0 ChrC 9427	1 37 36M * 0 0	TTTTTGGAATAGATTCCATTTTGAGAGAGTTGAAAA AAAAATCTTGGATTCAAAATTGATTTTTTTTAATA ATTATGCCTTGAAGAGGACTCGAACCTCCACGCTCT	BBBBAABAAAABBBBBBBBA?A>B>B@=B=>AAB @@<9<@BA@BB@BBBBBB@>A> <abab@?5?:baa< ?BCCBCCCCBCCBBABBBBB&gt;CBBC@CBBBACBCC</abab@?5?:baa< 	XT:A:R NM:E1 X0:E2 X1:E0 XM:E1 X0:E0 XG:E0 MD:Z:1G XT:A:U NM:E0 X0:E1 X1:E0 XM:E0 XO:E0 XG:E0 MD:Z:36 XT:A:R NM:E0 X0:E2 X1:E0 XM:E0 XO:E0 XG:E0 MD:Z:36	13: MPileup on data 7 and @ 0 🗱 data 10			
Fetch Sequences Fetch Alignments	SRR038850.877 0 Chr3 1353905 SRR038850.889 16 Chr2 896	9 0 36M * 0 0	CTCTCATATCTCCCTCGAATAAAGCTAAATTCTTTG TGTGGCAGCCAAGCGTTCATAGCGACGTTGCTTTTT	BCCCCCCBCCBCBBBCBCCCCBCBBBBBABCBCBB B<@==;7?6?@6?;@BB @@A<B=AACBBAB?CBA</th <th>XT:A:R NM:1:0 X0:1:2 X1:1:0 XM:1:0 X0:1:0 XG:1:0 MD:2:36 XT:A:R NM:1:0 X0:1:2 X1:1:0 XM:1:0 X0:1:0 XG:1:0 MD:2:36 XT:A:R NM:1:0 X0:1:2 X1:1:0 XM:1:0 X0:1:0 XG:1:0 MD:2:36</th> <th>10: SAM-to-BAM on data @ 12 X 7 and data 9: converted BAM</th>	XT:A:R NM:1:0 X0:1:2 X1:1:0 XM:1:0 X0:1:0 XG:1:0 MD:2:36 XT:A:R NM:1:0 X0:1:2 X1:1:0 XM:1:0 X0:1:0 XG:1:0 MD:2:36 XT:A:R NM:1:0 X0:1:2 X1:1:0 XM:1:0 X0:1:0 XG:1:0 MD:2:36	10: SAM-to-BAM on data @ 12 X 7 and data 9: converted BAM			
Get Genomic Scores Operate on Genomic Intervals	SRR038850.1208         16         Chr2         524           SRR038850.1343         0         ChrC         1416	0 23 36M * 0 0	ATGGGGATAGATCATTGCAATTGTTGGTCTTCAACG ACAACGACTAATTCATCGGCTAATATATTTCCGAAA	ABAB@BBBBABBBBBBBB@BCCBBBBBB@@ABBB?C@ BAC?CCCB@BCCCCCBB7@BABCAC@CCB=@CBBCC	XT:A:R NM:i:0 X0:i:2 X1:i:0 XM:i:0 X0:i:0 XG:i:0 MD:Z:36 XT:A:U NM:i:0 X0:i:1 X1:i:1 XM:i:0 X0:i:0 XG:i:0 MD:Z:36	9: Map with BWA for $\textcircled{P}$ 22 Illumina on data 8 and data 7:			
<u>Statistics</u> <u>Wavelet Analysis</u>	SRR038850.1467 0 ChrC 6426 SRR038850.1539 0 ChrC 5206 SRR038850.1560 16 ChrC 3544	3 37 36M * 0 0	TATCGAATACTGGTAATAATATCAGCAAAAGAACGT TAACTCAGTGGTTAGAGTATTGCTTTCATACGGCAG GCCCTATGAGTTAATACGATCACTATGTAGAGAAAG	?5@BCBB@B@B==;CBBBCBCBBBB@>??@ABCBBB 9CCBCCCC@BA@BCBC?5BBB>@CB<=CA@BCBCBA AB=BBBABBBCCBBC?CBCCB=BCCCBBCCCCCCB	XT:A:U NM::0 X0::1 X1::0 XM::0 X0::0 XG::0 MD:2:36 XT:A:U NM::0 X0::1 X1::0 XM::0 X0::0 XG::0 MD:2:36 XT:A:U NM::0 X0::1 X1::0 XM::0 X0::0 XG::0 MD:2:36	mapped reads ~100,000 lines, 8 comments format: sam, database: <u>?</u>			
Graph/Display Data Regional Variation Multiple regression	SRR038850.1597 0 ChrM 3446 SRR038850.1646 0 ChrC 2248	1 37 36M * 0 0	TTTCTCTCGAACTAACATATCATCCACCATCATCG AATCGCCTTTTTTTTATTTGGGAGGATTGAATACA	BBBC@B@CB=BBBBCABAB@@B@BBBABA@@ABA?? B?BCAABBCBCBBBB=BAB;5>@B@??A>AB>:<@=	XT-AC0 NM:10 X0:11 X1:10 XM:10 X0:10 X0:10 MD:2:30 XT-A:U NM:10 X0:11 X1:10 XM:10 X0:10 XG:10 MD:2:36 XT-A:U NM:10 X0:11 X1:10 XM:10 X0:10 XG:10 MD:2:36	BWA Version: 0.5.9-r16 BWA run on single-end data			
Multivariate Analysis Evolution	SRR038850.1762         0         ChrC         2905           SRR038850.1788         16         ChrM         32005		CCAATAAAAAAAAAAGTTCTTTATGATTCTTTTTCC CCAGTAAAGGTCTAATTCTTAGTTTTTCTATTTTAT	BA@<@@@BBBABBA@99AABBAA@=>?BABBBAB@A 9?AAC===B=A>CAAB@BBB>ABBBA8: <b:b:a<b< td=""><td>XT:A:U NM:E0 X0:E1 X1:E0 XM:E0 X0:E0 XG:E0 MD:Z:36 XT:A:U NM:E0 X0:E1 X1:E0 XM:E0 X0:E0 XG:E0 MD:Z:36</td><td>1.QNAME 2.FLAG 3.RNAME 4.POS 5.MAPQ</td></b:b:a<b<>	XT:A:U NM:E0 X0:E1 X1:E0 XM:E0 X0:E0 XG:E0 MD:Z:36 XT:A:U NM:E0 X0:E1 X1:E0 XM:E0 X0:E0 XG:E0 MD:Z:36	1.QNAME 2.FLAG 3.RNAME 4.POS 5.MAPQ			
Motif Tools Multiple Alignments	SRR038850.2044 0 ChrC 10959 SRR038850.2048 16 ChrC 35900 SRR038850.2075 16 ChrC 44900	0 37 36M * 0 0	ATTAATTCTCGCTGGCCGCGCTCCTATAGGGATCATG AGTCTTCTTGGTGGGTATCCTTAATTCTCTTATCTC	ABCCBCCCCCCBBBCCCCCBCCCCCBABBBCCCC AA=BBBBC??BBB;>CBB@>ABBBBBABB@A>BB>A	XT:A:U NM::0 X0::1 X1::0 XM::0 X0::0 XG::0 MD:2:36 XT:A:U NM::0 X0::1 X1::0 XM::0 X0::0 XG::0 MD:2:36 XT:A:U NM::0 X0::1 X1::0 XM::0 X0::0 XG::0 MD:2:36	@SQ SN:Chr1 LN:30427671 @SQ SN:Chr2 LN:19698289			
Metagenomic analyses FASTA manipulation	SRR038850.2075 16 Chrc 44900 SRR038850.2238 0 Chr3 1420298: SRR038850.2329 16 ChrC 6800	3 0 36M * 0 0	TCTTTCCGTACTTTCAACAAATTCACCAATCTTACG GCTCTTCCTATCATTGTGAAGCAGAATTCACCAAGT TAAAACTTACTTTATTGATCATTACATAGAATTCAA	9AABB> =AA => B@@BAB@@ACBAB7B>BB9BBBBCB AAB?CBBCCCCCCCBC => AB>BC@@BBBBBBBBBB A@@B@ABC:ABAAB?BB@BBBBB? = ABCBCBAACBB	XT:A:U NM:E0 X0:E1 X1:E0 XM:E0 X0:E0 XG:E0 MD:Z:36 XT:A:R NM:E0 X0:E2 X1:E0 XM:E0 X0:E0 XG:E0 MD:Z:36 XT:A:U NM:E0 X0:E1 X1:E0 XM:E0 X0:E0 XG:E0 MD:Z:36	<pre>@SQ SN:Chr3 LN:23459830 @SQ SN:Chr4 LN:18585056 @SQ SN:Chr5 LN:26975502</pre>			
NGS: QC and manipulation NGS: Mapping NGS: Indel Analysis	SRR038850.2456         16         ChrC         2453           SRR038850.2498         16         Chr2         744		ACGATCTTCTCAGTGTCAGTATAAAGGATTTTTCCC AAATGGATGGCGCCTTAAGCGCGCGCACCTATACCCGG	;BA?A@ABA?@BBBBBBBBBBBBBBBBBBBBBBBBBBBBB	XT:A:U NM:i:0 X0:i:1 X1:i:0 XM:i:0 X0:i:0 XG:i:0 MD:Z:36 XT:A:R NM:i:0 X0:i:2 X1:i:0 XM:i:0 X0:i:0 XG:i:0 MD:Z:36	esq sh:chrc LN:154478			
NGS: INDEL ANALYSIS NGS: RNA Analysis NGS: SAM Tools	SRR038850.2688 16 Chr3 14198274 SRR038850.2914 16 Chr2 1600	6 0 36M * 0 0	AGCGTATATTTAAGTTGTTGCAGTTAAAAAGCTCGT AGTTTAGGATGTCAAGTTTGCATCAAATATGCCCAC	@BBBBBBBBBBBBBBBCCBB@BBBCBCBBBB9A>B A <bbaa@@bb@b??@@ba=bbbbbbbbcbbbbb<b AABBAA@@BB@BB??@@BA=BBBBBBBBCBBBBBS</bbaa@@bb@b??@@ba=bbbbbbbbcbbbbb<b 	XT:A:R NM:I:0 X0:I:2 X1:E0 XM:I:0 X0:I:0 XG:I:0 MD:Z:36 XT:A:R NM:I:0 X0:I:2 X1:E0 XM:I:0 X0:I:0 XG:I:0 MD:Z:36	8: Filter FASTQ on data 3 @ 0 %			
NGS: SAM TOOIS	SRR038850.3049 16 Chr3 1419923	0 0 36M * 0 0	TCATTGCAATTGTTGGTCTTCAACGAGGAATTCCTA	AABBBABBBBBBBBBBA>@BAB@BCBBBCBBBCAC@	XT:A:R NM:i:0 X0:i:2 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:36	o. ritter rASIQ on data 5 @ V &			

## GFF and GTF format

General Feature Format (GFF) (Link)

```
browser position chr22:1000000-10025000
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
```

- Gene Transfer format (GTF) (Link)
  - > The list attribute must begin with 2 mandatory attributes.

23

Gene\_id\_value, transcript\_id\_value

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



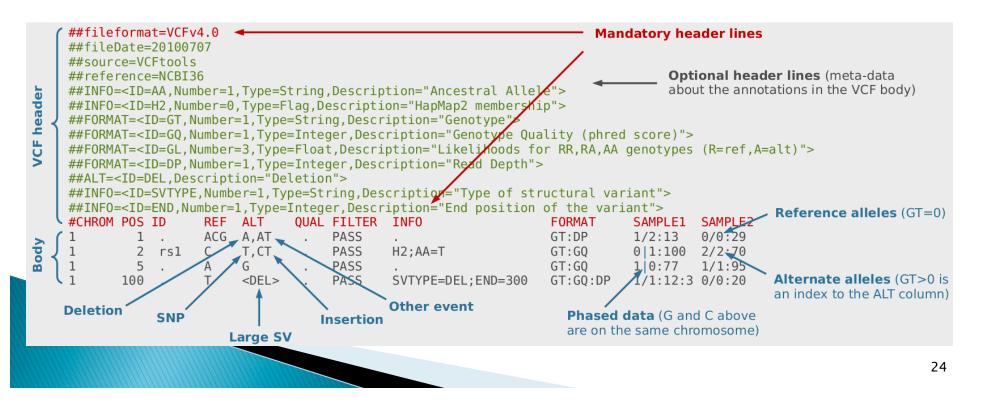
## BED format (Browser Extensible Data) (Link)

> Flexible way to define the data lines in the annotation track.

```
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
```

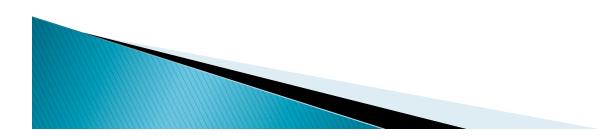
### **BCF / VCF format**

VCF: Variant Calling Format (Link)
 BCF: Binary version of VCF



## Available NGS Analysis Toolsets

- Prepare, Quality Check and Manipulate FASTQ reads
- Mapping
- SAMtools
- SNP and INDEL analysis
- RNA-seq analysis
- Peak calling / ChIP-seq
- Many more.....

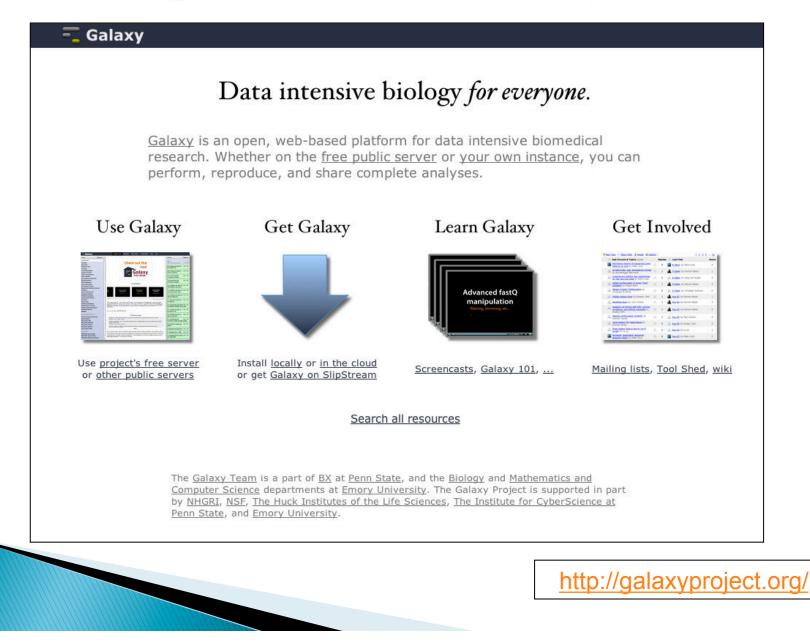


## NGS Analysis Using Galaxy

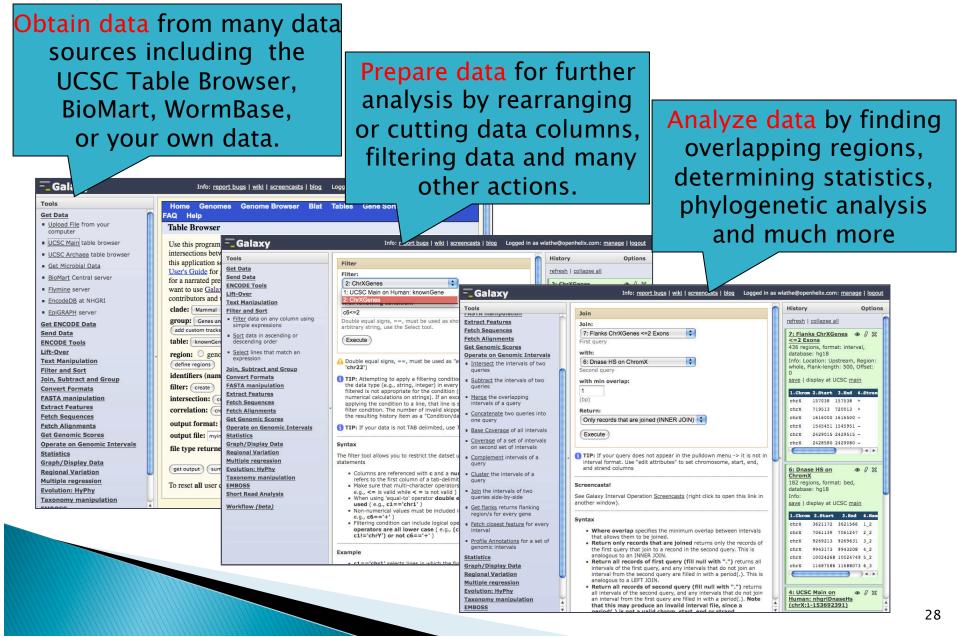
- Galaxy overview and Interface
- Getting Data in Galaxy
- Analyzing Data in Galaxy
  - Quality Control
  - Mapping Data
- History and workflow
- Sequences and Alignment Format
- Galaxy Exercises

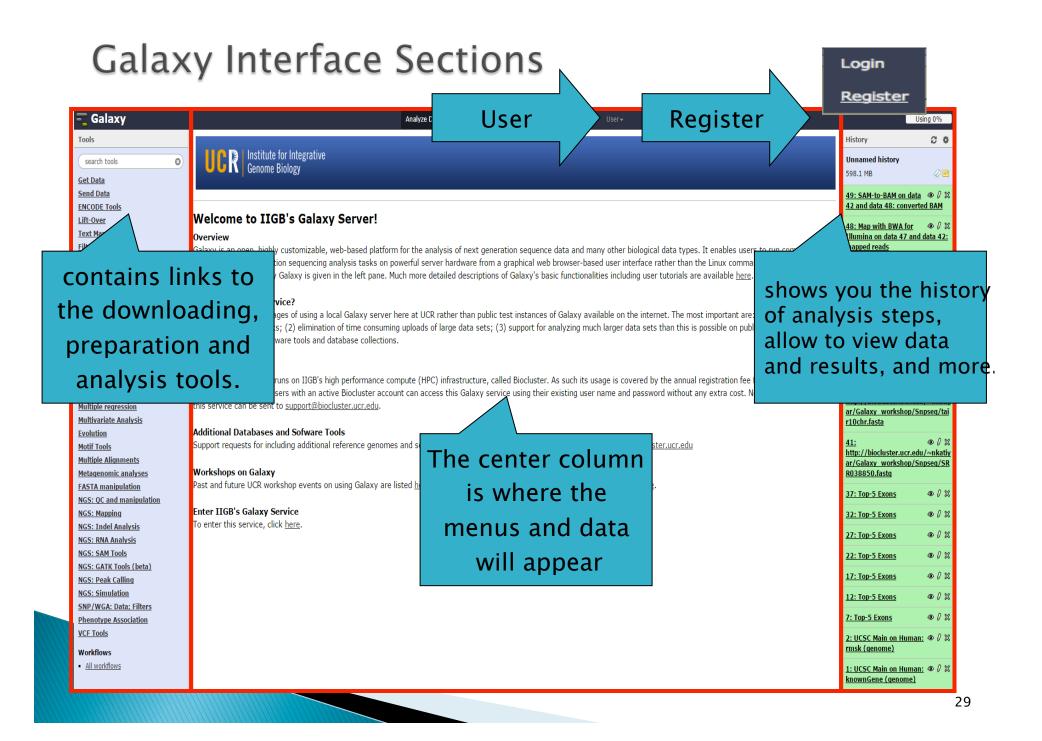


## Getting started with Galaxy



# Galaxy Conceptual Framework





## NGS Analysis Using Galaxy

- Sequences and Alignment Format
- Galaxy overview and Interface
- Getting Data in Galaxy
- Analyzing Data in Galaxy
  - Quality Control
  - Mapping Data
- History and workflow
- Galaxy Exercises



## Getting Data

-	Galaxy	Analyze Data Workflow Shared Data → Visualization → Help → User →		Using 0%
Tool	s 🖌		History	2 ¢
60	arch tools	Institute for Integrative	Unnamed history	
	/		0 bytes	47 🖻
<u>Get I</u>		k Get Data	Your history is emp	atu Cliak Cat
	ODE Tools		Data' on the left pa	
	Over V	Welcome to IIGB's Galaxy Server!		
	t Manipulation	Overview		
	er and Sort	Galaxy is an open, highly customizable, web-based platform for the analysis of next generation sequence data and many other biological data types. It enables users to run computationally		
Join,	, Subtract and Group	demanding next generation sequencing analysis tasks on powerful server hardware from a graphical web browser-based user interface rather than the Linux command-line. A subset of of		
Conv	vert Formats	application supported by Galaxy is given in the left pane. Much more detailed descriptions of Galaxy's basic functionalities including user tutorials are available here.		
Extra	act Features			
<u>Fetc</u>	h Sequences	Why Local Galaxy Service?		
	<u>h Alignments</u>	There are many advantages of using a local Galaxy server here at UCR rather than public test instances of Galaxy available on the internet. The most important are: (1) shorter waiting		
	<u>Genomic Scores</u>	queues for analysis tasks; (2) elimination of time consuming uploads of large data sets; (3) support for analyzing much larger data sets than this is possible on public services; (4) the		
	rate on Genomic Intervals	ability to customize software tools and database collections.		
	istics			
	<u>relet Analysis</u>	How to Gain Access?		
	<u>ph/Display Data</u>	This instance of Galaxy runs on IIGB's high performance compute (HPC) infrastructure, called Biocluster. As such its usage is covered by the annual registration fee for this infrastructure		
	ional Variation	(see here for details). Users with an active Biocluster account can access this Galaxy service using their existing user name and password without any extra cost. New account requests for		
	tiple regression	this service can be sent to <u>support@biocluster.ucr.edu</u> .		
	<u>tivariate Analysis</u>	Additional Databases and Sofware Tools		
	<u>ution</u> <u>f Tools</u>	Support requests for including additional reference genomes and software tools on IIGB's Galaxy server can be sent to support@biocluster.ucr.edu		
	tiple Alignments	support requests for including additional reference genomes and software tools on read 5 dataxy server can be serve to <u>support existence reading</u>		
	agenomic analyses	Workshops on Galaxy		
	TA manipulation	Past and future UCR workshop events on using Galaxy are listed here. The user manual from previous workshops can be accessed here.		
	: QC and manipulation			
	: Mapping	Enter IIGB's Galaxy Service		
	: Indel Analysis	To enter this service, click <u>here</u> .		
NGS	: RNA Analysis			
NGS	: SAM Tools			
NGS	: GATK Tools (beta)			
	: Peak Calling			
	: Simulation			
	/WGA: Data; Filters			
	notype Association			
VCF	<u>Tools</u>			
Wor	kflows			
• <u>All</u>	l workflows			
				31
				51

## Getting Data

🗧 Galaxy	Analyze Data Workflow Shared Data + Visualization + Help + User +	808 809 809	Using 0%
Tools		History	00
search tools	UCR Institute for Integrative Genome Biology	Unnamed history	
Get Data	UUN Genome Biology	0 bytes	02
Upload File from your		1 Your history is em	pty. Click 'Get
computer		Data' on the left p	ane to start
UCSC Main table browser	Welcome to IIGB's Galaxy Server!		
UCSC Test table browser	Overview		
UCSC Archaea table browser	Galaxy is an open, highly customizable, web-based platform for the analysis of next generation sequence data and many other biological data types. It enables users to run computationally demanding next generation sequencing analysis tasks on powerful server hardware from a graphical web browser-based user interface rather than the Linux command-line. A subset of of		
<u>BX</u> table browser	application supported by Galaxy is given in the left pane. Much more detailed descriptions of Galaxy's basic functionalities including user tutorials are available here.		
EBI SRA ENA SRA			
<u>Get Microbial Data</u>	Why Local Galaxy Service? There are many advantages of using a local Galaxy server here at UCR rather than public test instances of Galaxy available on the internet. The most important are: (1) shorter waiting		
BioMart Central server	queues for analysis tasks; (2) elimination of time consuming uploads of large data sets; (3) support for analyzing much larger data sets than this is possible on public services; (4) the		
<u>BioMart</u> Test server	ability to customize software tools and database collections.		
<u>CBI Rice Mart</u> rice mart			
GrameneMart Central server	How to Gain Access? This instance of Galaxy runs on IIGB's high performance compute (HPC) infrastructure, called Biocluster. As such its usage is covered by the annual registration fee for this infrastructure		
modENCODE fly server	(see <u>here</u> for details). Users with an active Biocluster account can access this Galaxy service using their existing user name and password without any extra cost. New account requests for		
<u>Flymine</u> server	this service can be sent to <u>support@biocluster.ucr.edu</u> .		
Flymine test server	Additional Databases and Sofware Tools		
modENCODE modMine server	Support requests for including additional reference genomes and software tools on IIGB's Galaxy server can be sent to <u>support@biocluster.ucr.edu</u>		
MouseMine server Ratmine server			
YeastMine server	Workshops on Galaxy		
metabolicMine server	Past and future UCR workshop events on using Galaxy are listed here. The user manual from previous workshops can be accessed here.		
modENCODE worm server	Enter IIGB's Galaxy Service		
WormBase server	To enter this service, click <u>here</u> .		
Wormbase test server			
EuPathDB server			
EncodeDB at NHGRI			
EpiGRAPH server			
EpiGRAPH test server			
HbVar Human Hemoglobin Variants and Thalassemias			
<u>GenomeSpace import</u> from file browser			
<u>Send Data</u>			

## Import data from UCSC genome browser

- Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User-		Using 0%
Tools		History	C 0
search tools	UCR Institute for Integrative Genome Biology	Unnamed histo 0 bytes	ry 🖉 🖻
Upload File from your			empty. Click 'Get
computer		Data' on the l	eft pane to start
UCSC Main table browser	s Galaxy Server!		
UCSC Test table browser UCSC Archaea table browser	Overview Galaxy is an open, highly customizable, web-based platform for the analysis of next generation sequence data and many other		
BX table browser	biological data types. It enables users to run computationally demanding next generation sequencing analysis tasks on powerful		
EBI SRA ENA SRA	server hardware from a graphical web browser-based user interface rather than the Linux command-line. A subset of of application supported by Galaxy is given in the left pane. Much more detailed descriptions of Galaxy's basic functionalities including user tutorials		
Get Microbial Data	are available <u>here</u> .		
BioMart Central server			
BioMart Test server	Why Local Galaxy Service? There are many advantages of using a local Galaxy server here at UCR rather than public test instances of Galaxy available on the		
CBI Rice Mart rice mart	internet. The most important are: (1) shorter waiting queues for analysis tasks; (2) elimination of time consuming uploads of large		
GrameneMart Central server	data sets; (3) support for analyzing much larger data sets than this is possible on public services; (4) the ability to customize software tools and database collections.		
modENCODE fly server			
<u>Flymine</u> server	How to Gain Access?		
<u>Flymine test</u> server	This instance of Galaxy runs on IIGB's high performance compute (HPC) infrastructure, called Biocluster. As such its usage is covered by the annual registration fee for this infrastructure (see <u>here</u> for details). Users with an active Biocluster account can access this		
modENCODE modMine server	Galaxy service using their existing user name and password without any extra cost. New account requests for this service can be sent		
MouseMine server	to <u>support@biocluster.ucr.edu</u> .		
<u>Ratmine</u> server	Additional Databases and Sofware Tools		
<u>YeastMine</u> server	Support requests for including additional reference genomes and software tools on IIGB's Galaxy server can be sent to		
<u>metabolicMine</u> server	support@biocluster.ucr.edu		
<u>modENCODE worm</u> server WormBase server	Workshops on Galaxy		
	Past and future UCR workshop events on using Galaxy are listed here. The user manual from previous workshops can be accessed		
<u>Wormbase</u> test server EuPathDB server	<u>here</u> .		
EncodeDB at NHGRI	Enter IIGB's Galaxy Service		
LICOUCDD ACHINON	To enter this service, click <u>here</u> .		

## Import data from UCSC genome browser

Â	Genomes	Genome Browser	Tools	Mirrors	Downloads	My Data	Help	About Us	
Table B	rowser								
sequenc the User software biologica computa be down clade: 1 group: [ table: k region: identifie filter: 1 intersec correlat output f file type	ie covered by a is Guide for generational tools. R al function of y ational tools. R aloaded in their Mammal Genes and Generation Genes and Generation Genes and Generation Generation: create tion: create format: BED - file: e returned: @ ut summary/s	e Predictions	using this and nd sample of mplex quer otation enrice age for the equence ar track: UC describe tab gions o polist upload	pplication s queries, ar ies, you m chments, s list of cont assemi CSC Genes le schema osition chr list Ser to keep ou	see Using the Tand the OpenHeli ay want to use ( send the data to tributors and usa ion Downloads p bly: Feb. 2009 (G v 7:127471196-1274§	able Browser f         x Table Browser f         x Table Browser f         Galaxy or our         GREAT. Send         age restriction         bage restriction         bage.         iRCh37/hg19)         manage custom         95720         lookup         Galaxy       G	for a desc ser <u>tutoria</u> <u>public My</u> d data to <u>g</u> is associa tracks	ription of the con I for a narrated pr <u>SQL server</u> . To e <u>GenomeSpace</u> fo	trols in this form, resentation of the xamine the or use with diverse
									- A
									34

## Send query to Galaxy from UCSC genome browser

Â	Genomes	Geno	me Browser	Tools	Mirrors	Downloads	My Data	Help	About Us
Output	knownGe	ne as BE	D						
	ude <u>custon</u>		ader:						
	1e= tb_known			_		1			
			query on known	Gene					
	oility=   pack	•			_				
url=									
Create	one BED re	cord per							
	ole Gene	cora per	•						
	stream by	200	bases						
	ons plus	0	bases at ea	ch end					
	ons plus	0	bases at ea						
	TR Exons	U	bases at ca	on end					
	ling Exons								
	vnstream by	200	bases						
				or end of	a chromo	some and upst	ream/downstrea	m bases	are added, they may be truncated in
			the edge of t					in bases	are added, they may be transated in
	uery to Galaxy		Ŭ						
Cancel									

# Getting Data: Upload File

		File Upload	
💳 Galaxy	Analyze Data Workflow Data Libr	ar 🔺 🕨 🔠 🧰 🛅 Desktop	c search
Tools		DEVICES ChrX_geneintervals.txt	<u>_</u>
Get Data		r benees	Lavor A
		▶ PLACES	
	Eile Format	▶ MEDIA	
UCSC Main ta	format?		
UCSC Archaea ta browser			
Get Microbial Data	File:		
BioMart Central server	Choose File Upload or	nasta fila	TXT
		baste me	
<u>GrameneMart</u> Central server	URL/Text:		Name chrX_geneinte
<ul> <li><u>Flymine</u> server</li> </ul>			rvals.txt
<ul> <li>EuPathDB server</li> </ul>	,		Size 88 KB
EncodeDB at NHGRI			Kind Microsoft
			Excel text
<ul> <li><u>EpiGRAPH</u> server</li> </ul>	Here you may specify a list of URLs (one per line)	o	document
Send Data	Convert spaces to tabs:		Created Today at
ENCODE Tools			
Lift-Over	Horse Sep. 2007 (equCab2)	Hide extension	Cancel Open
Text Manipulation	Horse Jan. 2007 (equCab1) Human Feb. 2009 (hg19)		
Convert Formats	1 Human Mar. 2006 (hg18)		74 1
FASTA man	Human May 2004 (hg17)	< Species	
Filter and S EXECUTE	Human July 2003 (hg16)		
Join, Subtra	Human Apr. 2003 (hg15)	a) formation of your play is not	
Extract Features	Hyperthermus butylicus DSM 5456 (hypeButy1)		
Fetch Sequences	At Hyphomonas neptunium ATCC 15444 (hyphNep	ot_ATCC15444)	
Fetch Alignments	Idiomarina Ioihiensis L2TR (idioLoih_L2TR)	atically be decompressed.	
Get Genomic Scores	Ta J. Craig Venter Sep. 2007 (venter1) pr Jannaschia sp. CCS1 (jannCCS1)		
Operate on Genomic Intervals	pr Kangaroo rat Jul. 2008 (dipOrd1)		
Statistics	sy Lactobacillus plantarum WCFS1 (lactPlan)		
Graph/Display Data	Lactobacillus salivarius UCC118 (lactSali_UCC1	118)	
Regional Variation	Lactococcus lactis subsp. lactis II1403 (lactLact)	)	
Multiple regression	Al Lamprey Mar. 2007 (petMar1)		
Evolution	A Lancelet Mar. 2006 (braFlo1)		
Metagenomic analyses	se Laurasiatheria Apr. 24. 2006 (lauRas13) Lawsonia intracellularis PHE/MN1-00 (lawsIntr_		
EMBOSS			1
NGS TOOLBOX BETA	Axt		
	blastz pairwise alignment format. Each alignment b	lock in an axt file contains three lines:	
NGS: QC and manipulation	a summary line and 2 sequence lines. Blocks are se		
NGS: Mapping	<ul> <li>lines. The summary line contains chromosomal pos alignment. It consists of 9 required fields.</li> </ul>	ition and size information about the	
NGS: SAM Tools	angimente re consists or 5 required fields.		

# Getting Data: Upload File

	Galaxy	
+ Ohttp://localhost:8080	/	C Q Google
- Galaxy	Analyze Data Workflow Data Libraries Help User	
Get Data         • Upload File       from your computer         • UCSC Main       table browser         • UCSC Test       table browser         • UCSC Archaea       table browser         • UCSC Archaea       table browser         • BX main       browser         • Get Microbial Data       BioMart         • BioMart       Test server	Upload File File Format: Auto-detect Which format? See help below File: Choose File no file selected URL/Text: http://bx.mathcs.emory.edu/outgoing/da ta/phiX174_genome.fa http://bx.mathcs.emory.edu/outgoing/da	History Options Opt
<u>GrameneMart</u> Central server <u>Flymine</u> server <u>Flymine test</u> server <u>modMine</u> server <u>Ratmine</u> server <u>Wormbase</u> server <u>Wormbase</u> test server     EuPathDB server	ta/phiX174_reads.fastqsanger SPE	cify multiple URLs the "URL / Text" box
EncodeDB at NHGRI     EpiGRAPH server     EpiGRAPH test server	Click to Search or Select (Execute)	
HbVar Human Hemoglobin Variants and Thalassemias     Send Data     ENCODE Tools     Lift-Over  Display a menu	Auto-detect The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabular, Wiggle, Bed and Interval (Bed with headers) formats. If your file is not detected properly as one of the known formats, it most likely means that it has some format problems (e.g., different number of columns on different rows). You can still coerce	×

NGS Analysis Using Galaxy

- Sequences and Alignment Format
- Galaxy overview and Interface
- Getting Data in Galaxy
- Analyzing Data in Galaxy
  - Lift–Over
  - Text manipulation tools
  - Filter and Sort
  - Operate on Genomic Intervals
  - Quality Control
  - Mapping Data
- History and workflow
  - Galaxy Exercises

## Lift-Over: convert genome coordinates

🗧 Galaxy	Analyze Data Workflow Shared Data → Visualization → Help → User →			Using 0%
Tools	Convert genome coordinates (version 1.0.3)	Histo	ory	C 0
search tools (3)	Convert coordinates of:	Unn	amed history	
Get Data		0 byt	tes	Ø 🖻
Send Data	To:	<b>0</b> Y	'our history is en	npty. Click 'Get
ENCODE Tools			ata' on the left	
Lift-Over	Minimum ratio of bases that must remap:			
Convert genome coordinates	0.95			
between assemblies and	Recommended values: same species = 0.95, different species = 0.10			
genomes	Allow multiple output regions?:			
Text Manipulation				
Filter and Sort	Recommended values: same species = No, different species = Yes			
Join, Subtract and Group				
Convert Formats	Execute			
Extract Features				
Fetch Sequences	A Make sure that the genome build of the input dataset is specified (click the pencil icon in the history item to set it if necessary).			
<u>Fetch Alignments</u> Get Genomic Scores				
Operate on Genomic Intervals	This tool can work with interval, GFF, and GTF datasets. It requires the interval datasets to have chromosome in column 1, start co-ordinate in column 2 and end co-ordinate in column 3. BED comments and track and browser lines will be ignored, but if other non-interval lines are present the tool will return empty output			
Statistics	datasets.			
Wavelet Analysis				
Graph/Display Data	What it does			
Regional Variation				
Multiple regression	This tool is based on the LiftOver utility and Chain track from the UC Santa Cruz Genome Browser.			
Multivariate Analysis	It converts coordinates and annotations between assemblies and genomes. It produces 2 files, one containing all the mapped coordinates and the other containing the unmapped coordinates, if any.			
Evolution	and animapped coordinates, in any.			
Motif Tools				
Multiple Alignments				
Metagenomic analyses	Example			
FASTA manipulation	Converting the following hg16 intervals to hg18 intervals:			
NGS: QC and manipulation	chrX 85170 112199 AK002185 0 +			
NGS: Mapping	chrX 110458 112199 AK097346 0 +			
NGS: Indel Analysis	chrX 112203 121212 AK074528 0 -			
NGS: RNA Analysis	will produce the following hg18 intervals:	_		
NGS: SAM Tools	chrX 132991 160020 AK002185 0 +			
<	chrX 158279 160020 AK097346 0 +	•		>

## Text manipulation tools

🗧 Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User -		Using 0%
Tools	Paste (version 1.0.0)	History	C 0
search tools	Paste:	Unnamed history	
Get Data		0 bytes	47 🖻
Send Data	and:	O Your history is en	
ENCODE Tools		Data' on the left	pane to start
Lift-Over	Delimit by:		
Text Manipulation Add column to an existing	Tab		
dataset			
<u>Compute</u> an expression on every row	Execute		
<u>Concatenate datasets</u> tail-to- head	Paste preserves column assignments of the first dataset.		
<u>Cut</u> columns from a table	What it does		
Merge Columns together	This tool merges two datasets side by side. If the first (left) dataset contains column assignments such as chromosome, start, end and strand, these will be preserved.		
Convert delimiters to TAB	However, if you would like to change column assignments, click the pencil icon in the history item.		
<u>Create single interval</u> as a new dataset	Example		
<u>Change Case</u> of selected columns	First dataset:		
Paste two files side by side	a 2		
Remove beginning of a file	a 3		
Select random lines from a file	Second dataset:		
Select first lines from a	20		
dataset	30 40		
<u>Select last</u> lines from a dataset	Pasting them together will produce:		
Trim leading or trailing	a 1 20		
characters	a 2 30		
Line/Word/Character count of a dataset	a 3 40		
Secure Hash / Message Digest			
on a dataset			

### Filter and Sort on Galaxy

🗧 Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User-		Using 0
Tools	Extract features (version 1.0.0)	History	C
search tools	Select GFF data:	Unnamed history	
Get Data		0 bytes	¢
Send Data	From:	1 Your history is e	mpty. Click '(
NCODE Tools	Column 1 / Sequence name	Data' on the left	pane to star
.ift-Over	Extract features:		
Fext Manipulation			
ilter and Sort			
Filter data on any column	•		
using simple expressions	Multi-select list - hold the appropriate key while clicking to select multiple columns		
Sort data in ascending or		-	
descending order	Execute		
<u>Select</u> lines that match an expression			
	What it does		
GFF	This tool extracts selected features from GFF data.		
Extract features from GFF data	This tool excludes selected reactines from OT data.		
Filter GFF data by attribute			
using simple expressions	Example		
Filter GFF data by feature	Selecting <b>promoter</b> from the following GFF data:		
count using simple expressions	chr22 GeneA enhancer 10000000 10001000 500 + . TGA		
Filter GTF data by attribute	chr22 GeneA promoter 10010000 10010100 900 + . TGA		
<u>values list</u>	chr22 GeneB promoter 10020000 10025000 400 TGB chr22 GeneB CCD52220 10030000 10065000 800 TGB		
loin, Subtract and Group			
Convert Formats	will produce the following output:		
xtract Features	chr22 GeneA promoter 10010000 10010100 900 + . TGA		
etch Sequences	chr22 GeneB promoter 10020000 10025000 400 TGB		
etch Alignments			
Get Genomic Scores	About formats		
Operate on Genomic Intervals	GFF format General Feature Format is a format for describing genes and other features associated with DNA, RNA and Protein sequences. GFF lines have nine tab-		
Statistics	separated fields:		

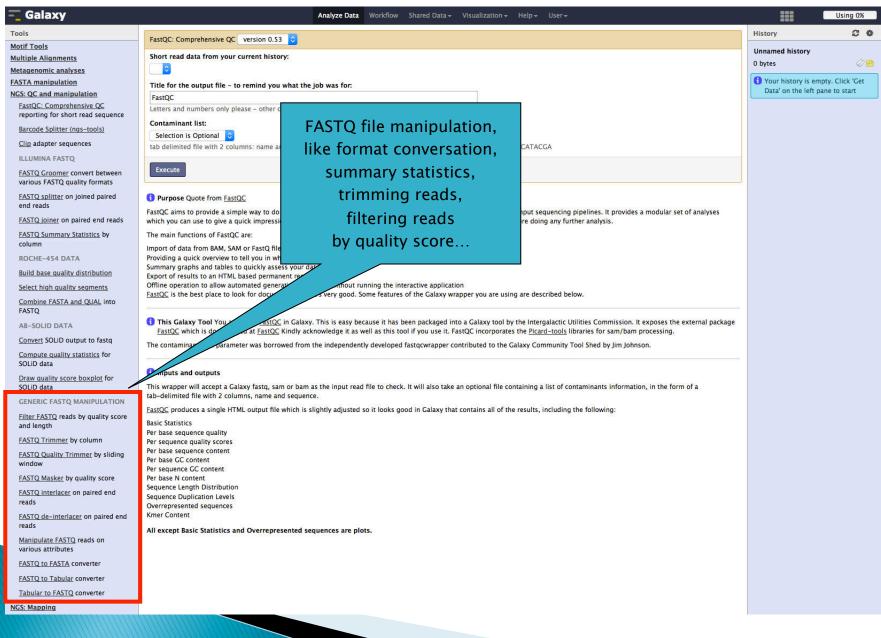
#### **Operate on Genomic Intervals**

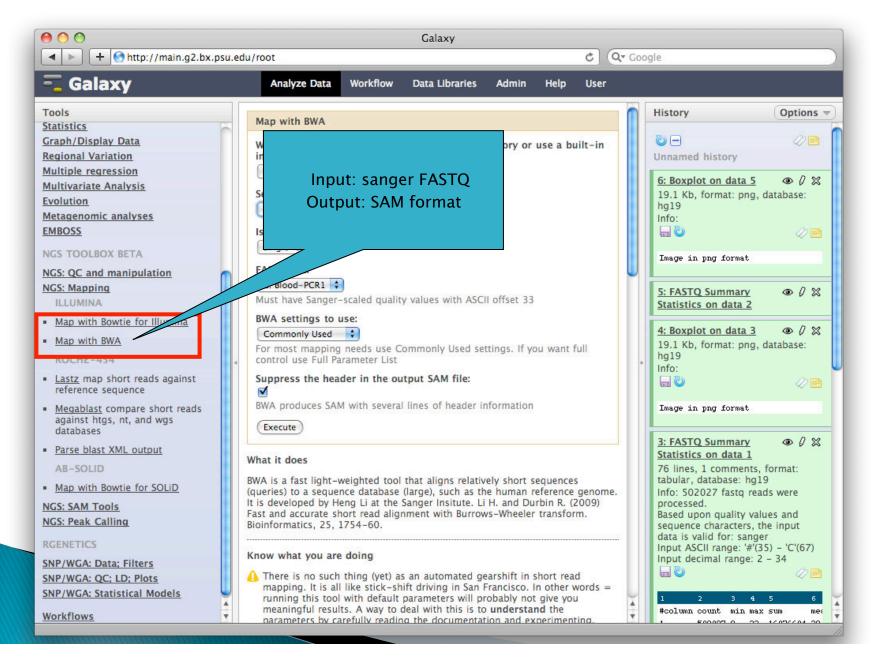
Salaxy Analyze Data Workflow Shared Data - Visualization - Help - User -		Using 0%
Tools Concatenate (version 1.0.1)	History	C 🕈
Join, Subtract and Group	Unnamed history	
Convert Formats	0 bytes	47 🖻
Extract Features First dataset		et Lie i
Fetch Sequences with:	Your history is e Data' on the left	
Fetch Alignments	Data off the felt	e pune to start
Second dataset		
Operate on Genomic Intervals Both datasets are same filetype?:		
Intersect the intervals of two datasets are sume met per the sum sume met per the sum		
If unchecked Second dataset will be forced into format of First dataset		
Subtract the intervals of two datasets		
Execute Execute		
Merge the overlapping intervals of a dataset		
TD: If your detects does not encour in the collideum means to it is not in interval formet. Use "wdit attributes" to get domain and and attributes and and attributes at the state of		
Concatenate two datasets into one dataset		
Base Coverage of all intervals Screencasts!		
Coverage of a set of intervals on second set of intervals		
Complement intervals of a dataset Syntax		
Cluster the intervals of a dataset are exactly the same filetype will preserve all extra fields in both files. Leaving this unchecked will force the second dataset to use the same column assignments for chrom, start, end and strand, but will fill extra fields with a period(.). In both cases, the output fields are truncated or padded with fields of periods to maintain a truly tabular output.		
Join the intervals of two		
datasets side-by-side Example		
Get flanks returns flanking		
region/s for every gene First dataset		
Fetch closest non-overlapping       feature for every interval       Second dataset		
Profile Annotations for a set of		
genomic intervals		
Statistics		
Wavelet Analysis		
Graph/Display Data Concatenated intervals		

## FASTA manipulation

- Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User -	Using 0%
Tools	RNA/DNA (version 1.0.0)	History 📿 🌣
<u>Graph/Display Data</u>	Library to convert:	Unnamed history
Regional Variation		0 bytes 🖉 📑
Multiple regression	Convert:	
<u>Multivariate Analysis</u>		• Your history is empty. Click 'Get Data' on the left pane to start
Evolution	RNA to DNA (U to T)	
Motif Tools		
Multiple Alignments	Execute	
<u>Metagenomic analyses</u>		
FASTA manipulation	What it does	
RNA/DNA converter	This tool converts RNA FASTA files to DNA (and vice-versa).	
FASTA Width formatter	In <b>RNA-to-DNA</b> mode, U's are changed into T's.	
Compute sequence length	In <b>DNA-to-RNA</b> mode, T's are changed into U's.	
Filter sequences by length		
Concatenate FASTA alignment	Example	
by species	Input RNA FASTA file ( from Sanger's mirBase ):	
FASTA-to-Tabular converter	>cel-let-7 MIMAT0000001 Caenorhabditis elegans let-7	
Tabular-to-FASTA converts	UGAGGUAGUAGGUUGUAUAGUU	
tabular file to FASTA format	>cel-lin-4 MIMAT0000002 Caenorhabditis elegans lin-4	
NGS: QC and manipulation	UCCCUGAGACCUCAAGUGUGA >cel-miR-1 MIMAT0000003 Caenorhabditis elegans miR-1	
NGS: Mapping	UGGAAUGUAAAGAAGUAUGUA	
NGS: Indel Analysis	Output DNA FASTA file (with RNA-to-DNA mode):	
NGS: RNA Analysis		
NGS: SAM Tools	>cel-let-7 MIMAT0000001 Caenorhabditis elegans let-7	
NGS: GATK Tools (beta)	TGAGGTAGTAGGTTGTATAGTT >cel-lin-4 MIMAT0000002 Caenorhabditis elegans lin-4	
NGS: Peak Calling	TCCCTGAGACCTCAAGTGTGA	
NGS: Simulation	<pre>&gt;cel-miR-1 MIMAT0000003 Caenorhabditis elegans miR-1</pre>	
	TGGAATGTAAAGAAGTATGTA	• III

🗧 Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User -	Using	0%
Tools	FastQC: Comprehensive QC version 0.53	History	0
search tools	Short read data from your current history:	Unnamed history	
Get Data		0 bytes	0
Send Data	Title for the output file - to remind you what the job was for:	<b>1</b> Your history is empty. Click	k 'Get
ENCODE Tools	FastQC	Data' on the left pane to st	
Lift-Over	Letters and numbers only please - other characters will be removed		
Text Manipulation	Contaminant list:		
Filter and Sort	Selection is Optional		
Join, Subtract and Group	tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA		
Convert Formats			
Extract Features	Execute		
Fetch Sequences			
Fetch Alignments	A Demonstration from Endog		
Get Genomic Scores	<b>1</b> Purpose Quote from <u>FastQC</u>		
Operate on Genomic Intervals	FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.		
Statistics			
<u>Wavelet Analysis</u>	The main functions of FastQC are:		
<u>Graph/Display Data</u>	Import of data from BAM, SAM or FastQ files (any variant) Providing a quick overview to tell you in which areas there may be problems		
Regional Variation	Sumary graphs and tables to guickly assess your data		
Multiple regression	Export of results to an HTML based permanent report		
Multivariate Analysis	Offline operation to allow automated generation of reports without running the interactive application <u>FastQC</u> is the best place to look for documentation - it's very good. Some features of the Galaxy wrapper you are using are described below.		
Evolution	Tastoc is the best place to look for documentation. It's very good, some reatures of the dataxy mapper you are using are described below.		
Motif Tools			
Multiple Alignments Metagenomic analyses	1 This Galaxy Tool You are using FastQC in Galaxy. This is easy because it has been packaged into a Galaxy tool by the Intergalactic Utilities Commission. It exposes the external package FastQC which is documented at FastQC Kindly acknowledge it as well as this tool if you use it. FastQC incorporates the Picard-tools libraries for sam/bam processing.		
FASTA manipulation	The contaminants file parameter was borrowed from the independently developed fastqcwrapper contributed to the Galaxy Community Tool Shed by Jim Johnson.		
NGS: QC and manipulation	The contaminants the parameter was borrowed from the independently developed fastqcwrapper contributed to the Galaxy Community foor shed by Jim Johnson.		
NGS: Mapping	1 Inputs and outputs		
<u>NGS: Indel Analysis</u>			
<u>NGS: RNA Analysis</u>	This wrapper will accept a Galaxy fastq, sam or bam as the input read file to check. It will also take an optional file containing a list of contaminants information, in the form of a tab-delimited file with 2 columns, name and sequence.		
NGS: SAM Tools	FastOC produces a single HTML output file which is slightly adjusted so it looks good in Galaxy that contains all of the results, including the following:		
NGS: GATK Tools (beta)	8 Decision		
NGS: Peak Calling	Basic Statistics Per base sequence quality		
NGS: Simulation	Per sequence quality scores		
SNP/WGA: Data; Filters	Per base sequence content		
Phenotype Association	Per base GC content Per sequence GC content		
VCF Tools	Per base N content		
Workflows	Sequence Length Distribution		
	_ Sequence Duplication Levels		





= Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User -		Using 0%
Tools	MPileup (version 0.0.2)	History	0 0
search tools	Choose the source for the reference list:	Unnamed history	
Get Data	Locally cached	0 bytes	47 🖻
Send Data	BAM files	1 Your history is e	moty Click 'Get
ENCODE Tools	BAM file 1	Data' on the left	
Lift-Over			
Text Manipulation	BAM file:		
Filter and Sort			
Join, Subtract and Group			
Convert Formats	Add new BAM file		
Extract Features	Using reference genome:		
Fetch Sequences			
Fetch Alignments	Genotype Likelihood Computation:		
Get Genomic Scores	Do not perform genotype likelihood computation		
Operate on Genomic Intervals	Set advanced options:		
Statistics	Basic O		
Wavelet Analysis Graph/Display Data			
Regional Variation	Execute		
Multiple regression	caeture		
Multivariate Analysis			
Evolution	What it does		
Motif Tools	Generate BCF or pileup for one or multiple BAM files. Alignment records are grouped by sample identifiers in @RG header lines. If sample identifiers are absent, each input file is		
Multiple Alignments	regarded as one sample.		
Metagenomic analyses			
FASTA manipulation	Setting .		
NGS: QC and manipulation	Settings:		
NGS: Mapping	Input Options: -6 Assume the quality is in the Illumina 1.3+ encoding.		
NGS: Indel Analysis	- A bo not skip anomalous read pairs in variant calling.		
NGS: RNA Analysis	-B Disable probabilistic realignment for the computation of base alignment quality (BAQ). BAQ is the Phred-scaled probability of a read base being mis		
NGS: SAM Tools	-b FILE List of input BAM files, one file per line [null] -C INT Coefficient for downgrading mapping quality for reads containing excessive mismatches. Given a read with a phred-scaled probability q of be		
BCF Tools Cat This tool allows the user to concatenate BCF files.	-d INT At a position, read maximaly INT reads per input BAM. [250]		
	-E Extended BAQ computation. This option helps sensitivity especially for MNPs, but may hurt specificity a little bit.		
bcftools view Converts BCF format to VCF format	-f FILE The faidx-indexed reference file in the FASTA format. The file can be optionally compressed by razip. [null] -1 FILE BED or position list file containing a list of regions or sites where pileup or BCF should be generated [null]		
BCF Tools Index This tool allows	-q INT Minimum mapping quality for an alignment to be used [0]		
the user to index sorted BCF for	-Q INT Minimum base guality for a base to be considered [13] -r STR Only generate pileup in region STR [all sites]		
random access.	-r sik only generate pileup in region sik (all sites) Output Options:		
MPileup SNP and indel caller			
SAM-to-BAM converts SAM format	-D Output per-sample read depth -g Compute genotype likelihoods and output them in the binary call format (BCF).		
to BAM format	-S Output per-sample Phred-scaled strand bias P-value		
bcftools view Converts BCF format	-u Similar to -g except that the output is uncompressed BCF, which is preferred for piping.		
to VCF format	Options for Genotype Likelihood Computation (for -g or -u):		
NGS: GATK Tools (beta)			
NGS: Peak Calling	-e INT Phred-scaled gap extension sequencing error probability. Reducing INT leads to longer indels. [20] -h INT Coefficient for modeling homopolymer errors. Given an 1-long homopolymer run, the sequencing error of an indel of size s is modeled as INT*		
NGS: Simulation	-n INT Coefficient for modeling nomopolymer errors. Given an 1-long nomopolymer run, the sequencing error of an indel of size s is modeled as INT* -I Do not perform INDEL calling		
SNP/WGA: Data; Filters	-L INT Skip INDEL calling if the average per-sample depth is above INT. [250]		
Phenotype Association	-o INT Phred-scaled gap open sequencing error probability. Reducing INT leads to more indel calls. [40] -P STR Comma dilimited list of platforms (determined by @RG-PL) from which indel candidates are obtained. It is recommended to collect indel candi		
VCF Tools	- A DIA COMMUN DELEMENTED DI PLOLIDIMO (DECEMBINED DY END-ED) ILUM WHICH INDER CONDUCTED LI E DECOMMENDED DE CONTECT INDER CANADA		

💳 Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User -		Using 0%
Tools		History	0
	Institute for Integrative	Unnamed history	
search tools	UCR Institute for Integrative Genome Biology	0 bytes	0
<u>Get Data</u>		o bytes	~
Send Data		<ol> <li>Your history is en</li> </ol>	
ENCODE Tools		Data' on the left p	pane to start
Lift-Over	Welcome to IIGB's Galaxy Server!		
Text Manipulation	Overview		
Filter and Sort	Galaxy is an open, highly customizable, web-based platform for the analysis of next generation sequence data and many other biological data types. It		
Join, Subtract and Group	enables users to run computationally demanding next generation sequencing analysis tasks on powerful server hardware from a graphical web		
Convert Formats	browser-based user interface rather than the Linux command-line. A subset of of application supported by Galaxy is given in the left pane. Much more		
Extract Features	detailed descriptions of Galaxy's basic functionalities including user tutorials are available here.		
Fetch Sequences			
Fetch Alignments	Why Local Galaxy Service?		
Get Genomic Scores	There are many advantages of using a local Galaxy server here at UCR rather than public test instances of Galaxy available on the internet. The most		
Operate on Genomic Intervals	important are: (1) shorter waiting queues for analysis tasks; (2) elimination of time consuming uploads of large data sets; (3) support for analyzing much larger data sets than this is possible on public services; (4) the ability to customize software tools and database collections.		
Statistics	larger data sets than this is possible on public services, (4) the ability to customize software tools and database collections.		
Wavelet Analysis	How to Gain Access?		
Graph/Display Data	This instance of Galaxy runs on IIGB's high performance compute (HPC) infrastructure, called Biocluster. As such its usage is covered by the annual		
Regional Variation	registration fee for this infrastructure (see here for details). Users with an active Biocluster account can access this Galaxy service using their existing user		
Multiple regression	name and password without any extra cost. New account requests for this service can be sent to <u>support@biocluster.ucr.edu</u> .		
Multivariate Analysis Evolution			
Motif Tools	Additional Databases and Sofware Tools		
Multiple Alignments	Support requests for including additional reference genomes and software tools on IIGB's Galaxy server can be sent to support@biocluster.ucr.edu		
Metagenomic analyses			
FASTA manipulation	Workshops on Galaxy		
NGS: QC and manipulation	Past and future UCR workshop events on using Galaxy are listed here. The user manual from previous workshops can be accessed here.		
NGS: Mapping			
NGS: Indel Analysis	Enter IIGB's Galaxy Service		
NGS: RNA Analysis	To enter this service, click here.		
NGS: SAM Tools			
NGS: GATK Tools (beta)			
NGS: Peak Calling			
NGS: Simulation			
SNP/WGA: Data; Filters			
Phenotype Association			
VCF Tools			
Workflows			
All workflows			

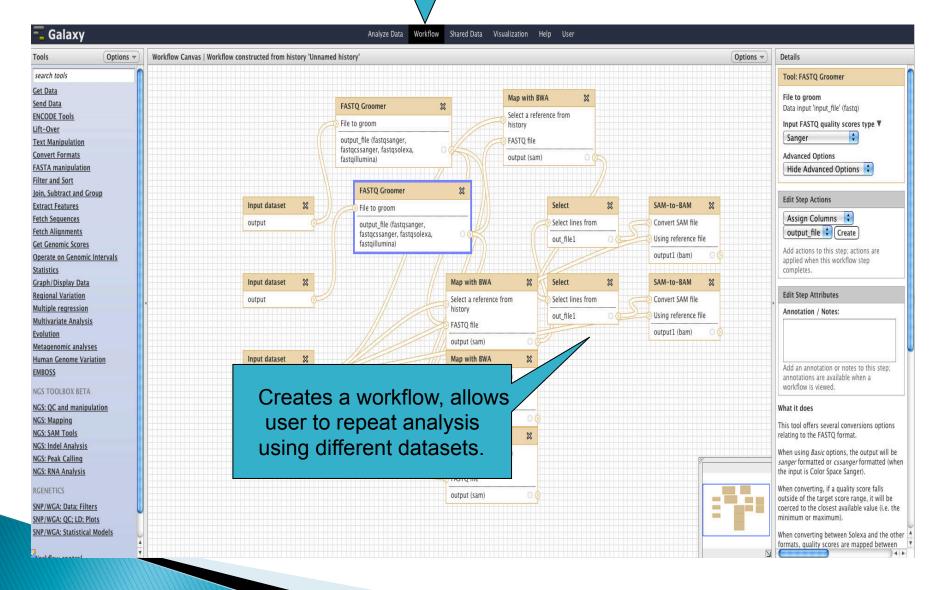
- Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User -		Using 0%
Tools		History	0 0
bcf	UCR Institute for Integrative Genome Biology	Unnamed history	
NGS: SAM Tools		0 bytes	Q =
<u>BCF Tools Cat</u> This tool allows the user to concatenate BCF files.		Your history is emp Data' on the left part	
<u>bcftools view</u> Converts BCF format to VCF format	Welcome to IIGB's Galaxy Server!		
BCF Tools Index This tool allows	Overview Galaxy is an open, highly customizable, web-based platform for the analysis of next generation sequence data and many other biological data types. It		
the user to index sorted BCF for random access.	enables users to run computationally demanding next generation sequencing analysis tasks on powerful server hardware from a graphical web		
bcftools view Converts BCF format	browser-based user interface rather than the Linux command-line. A subset of of application supported by Galaxy is given in the left pane. Much more detailed descriptions of Galaxy's basic functionalities including user tutorials are available <u>here</u> .		
to VCF format			
Vorkflows All workflows	Why Local Galaxy Service? There are many advantages of using a local Galaxy server here at UCR rather than public test instances of Galaxy available on the internet. The most		
	important are: (1) shorter waiting queues for analysis tasks; (2) elimination of time consuming uploads of large data sets; (3) support for analyzing much larger data sets than this is possible on public services; (4) the ability to customize software tools and database collections.		
	Hermite Grin America		
	How to Gain Access? This instance of Galaxy runs on IIGB's high performance compute (HPC) infrastructure, called Biocluster. As such its usage is covered by the annual registration fee for this infrastructure (see <u>here</u> for details). Users with an active Biocluster account can access this Galaxy service using their existing user name and password without any extra cost. New account requests for this service can be sent to <u>support@biocluster.ucr.edu</u> .		
	Additional Databases and Sofware Tools Support requests for including additional reference genomes and software tools on IIGB's Galaxy server can be sent to <u>support@biocluster.ucr.edu</u>		
	Workshops on Galaxy		
	Past and future UCR workshop events on using Galaxy are listed <u>here</u> . The user manual from previous workshops can be accessed <u>here</u> .		
	Enter IIGB's Galaxy Service		
	To enter this service, click <u>here</u> .		

NGS Analysis Using Galaxy

- Sequences and Alignment Format
- Galaxy overview and Interface
- Getting Data in Galaxy
- Analyzing Data in Galaxy
  - Lift-Over
  - Text manipulation tools
  - Filter and Sort
  - Operate on Genomic Intervals
  - Quality Control
  - Mapping Data
- History and workflow
- Galaxy Exercises

## History: History Options

## Workflow



#### What's next?

- Galaxy exercises
  - SNP-Seq
  - RNA-Seq
- Visualization
  - IGV (Integrative Genomics Viewer)
  - <u>http://www.broadinstitute.org/igv/</u>

