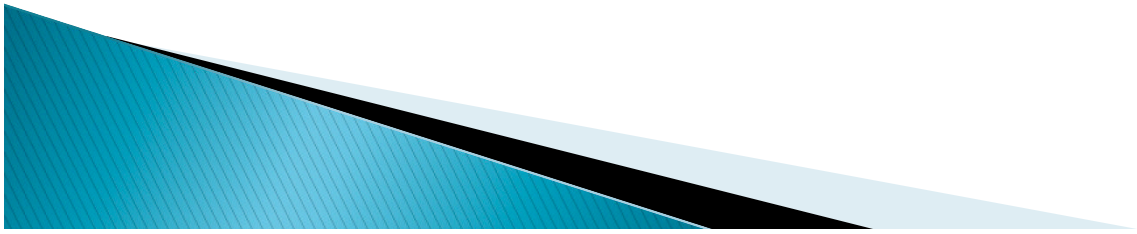
NGS Analysis Using Galaxy

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

Data intensive biology *for everyone.*

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

Use Galaxy

Use project's free server or other public servers

Get Galaxy

Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy

Screencasts, Galaxy 101, ...

Get Involved

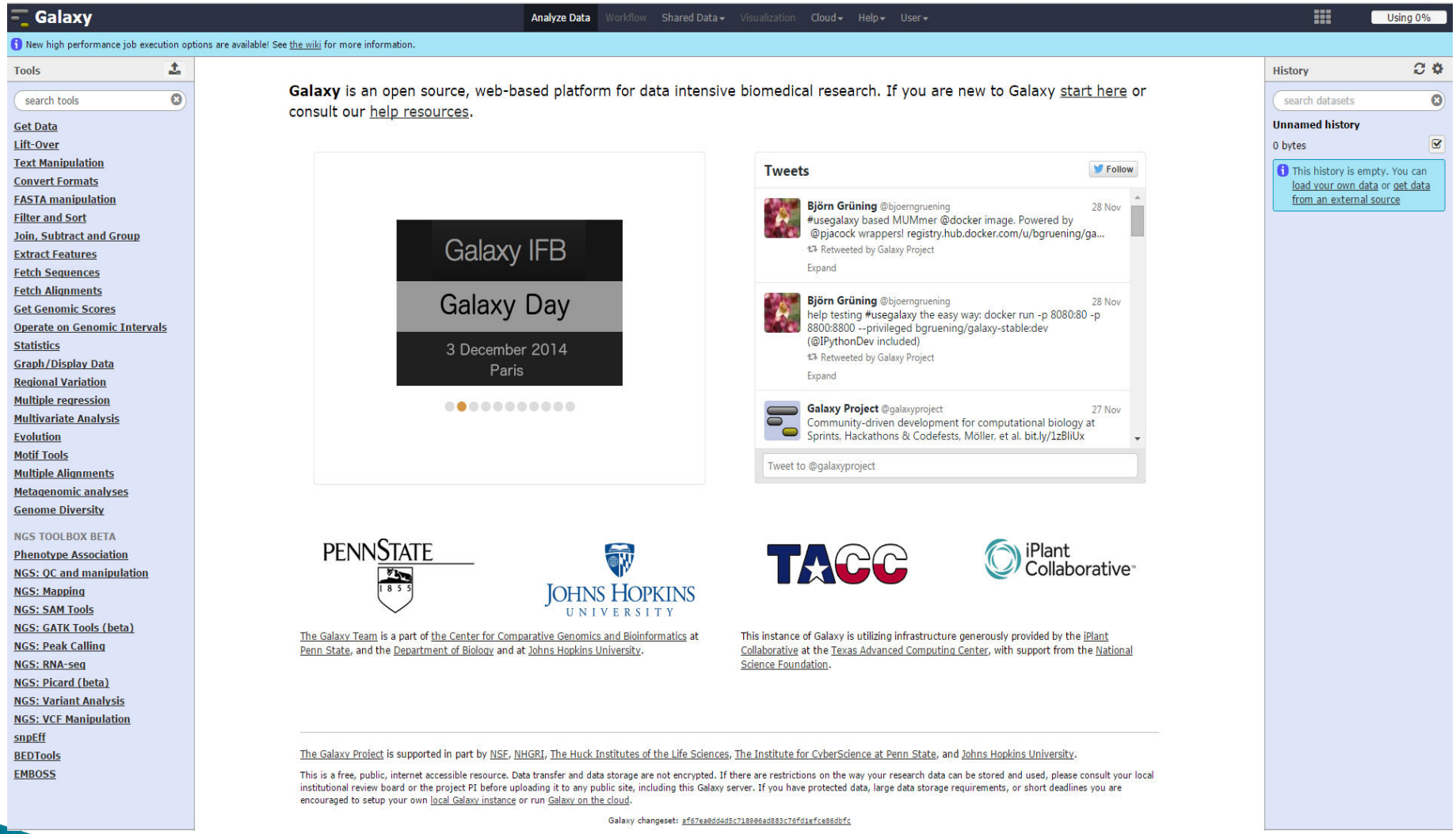
Mailing lists, Tool Shed, wiki

Search all resources

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

http://galaxyproject.org/

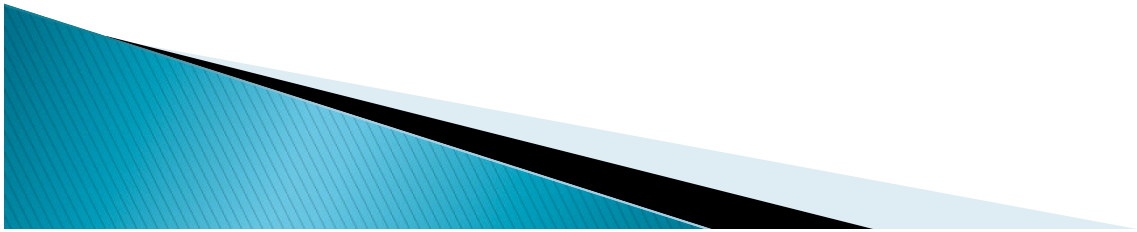
Galaxy analysis web interface



https://usegalaxy.org/

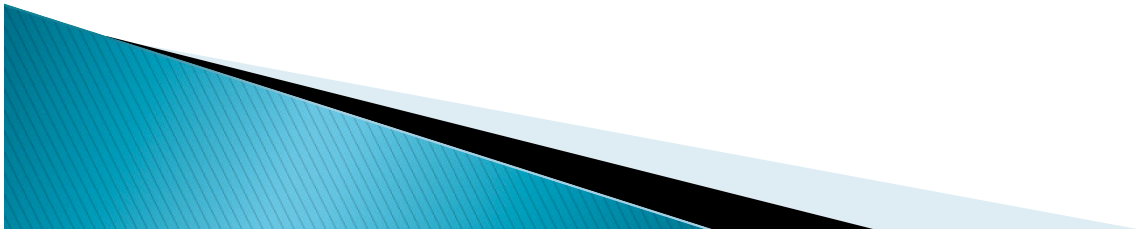
Outline

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



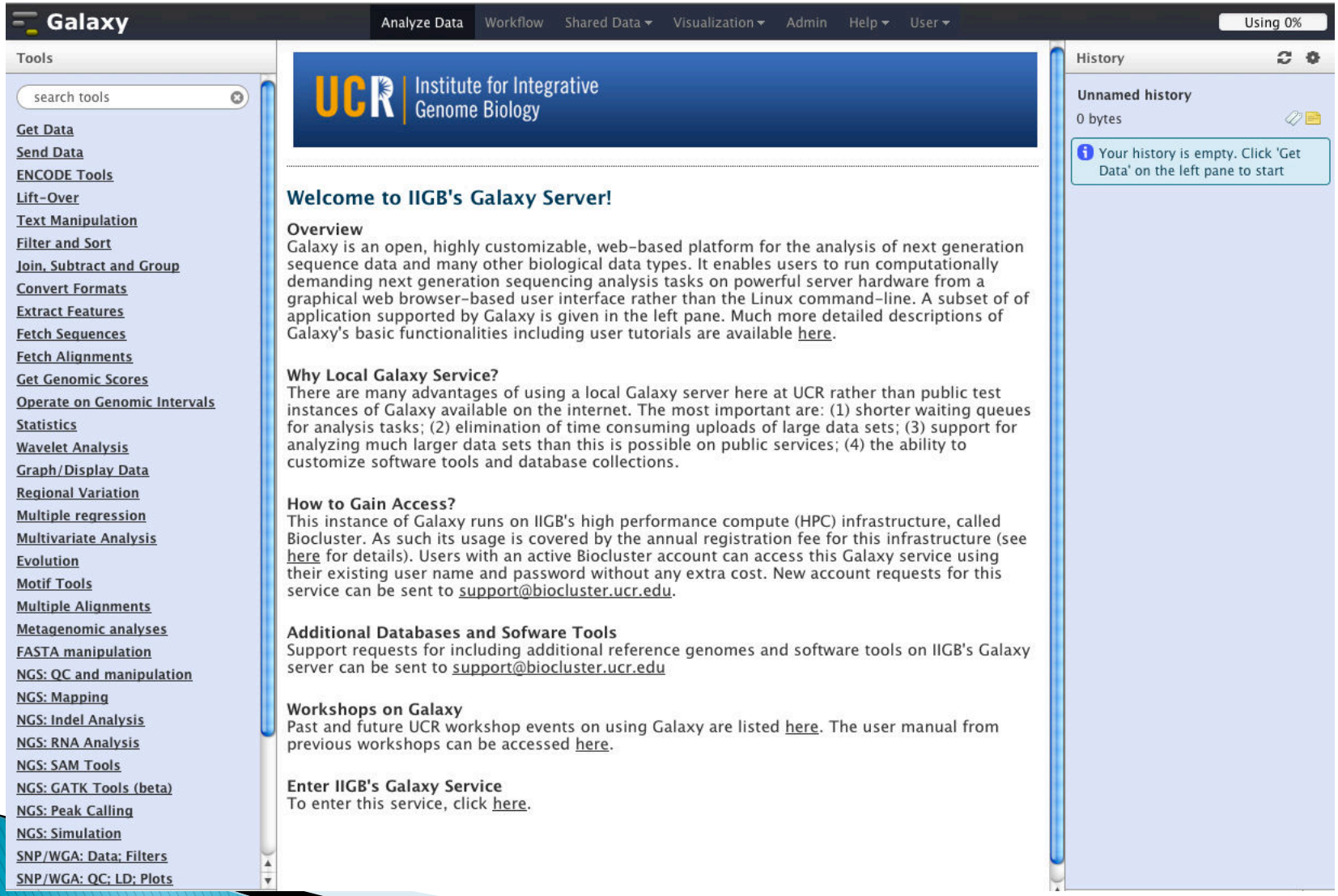
Galaxy: the instant web-based tool and data resource integration platform

- 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



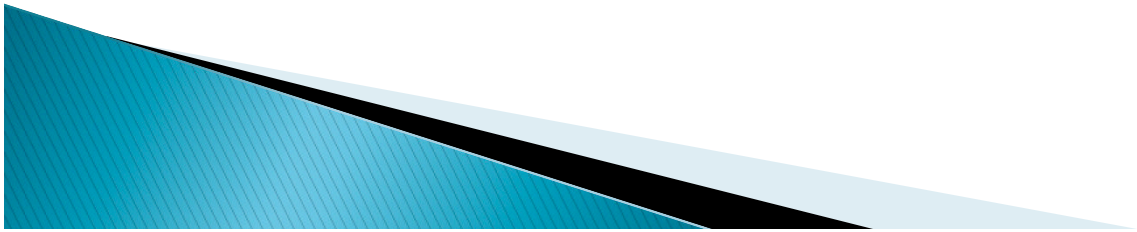
Galaxy instance

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



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

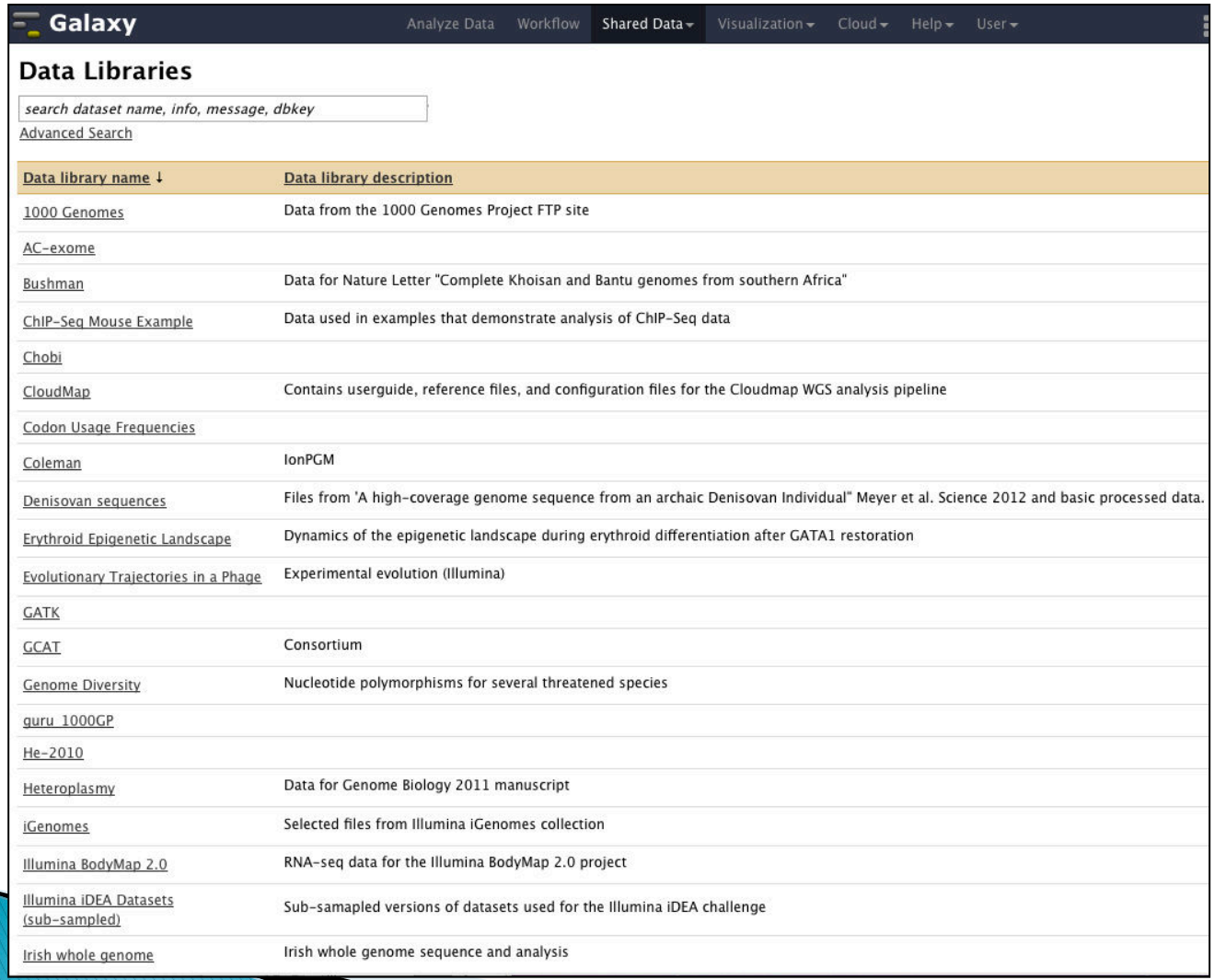
- Analyze
  - Retrieve data directly from popular data resources or upload your own.
  - Interactively manipulate genomic data with a comprehensive and expanding best-practices toolset.
  - Galaxy is designed to work with many different datatypes. (Link)
- Visualize
  - Trackster is Galaxy's visualization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

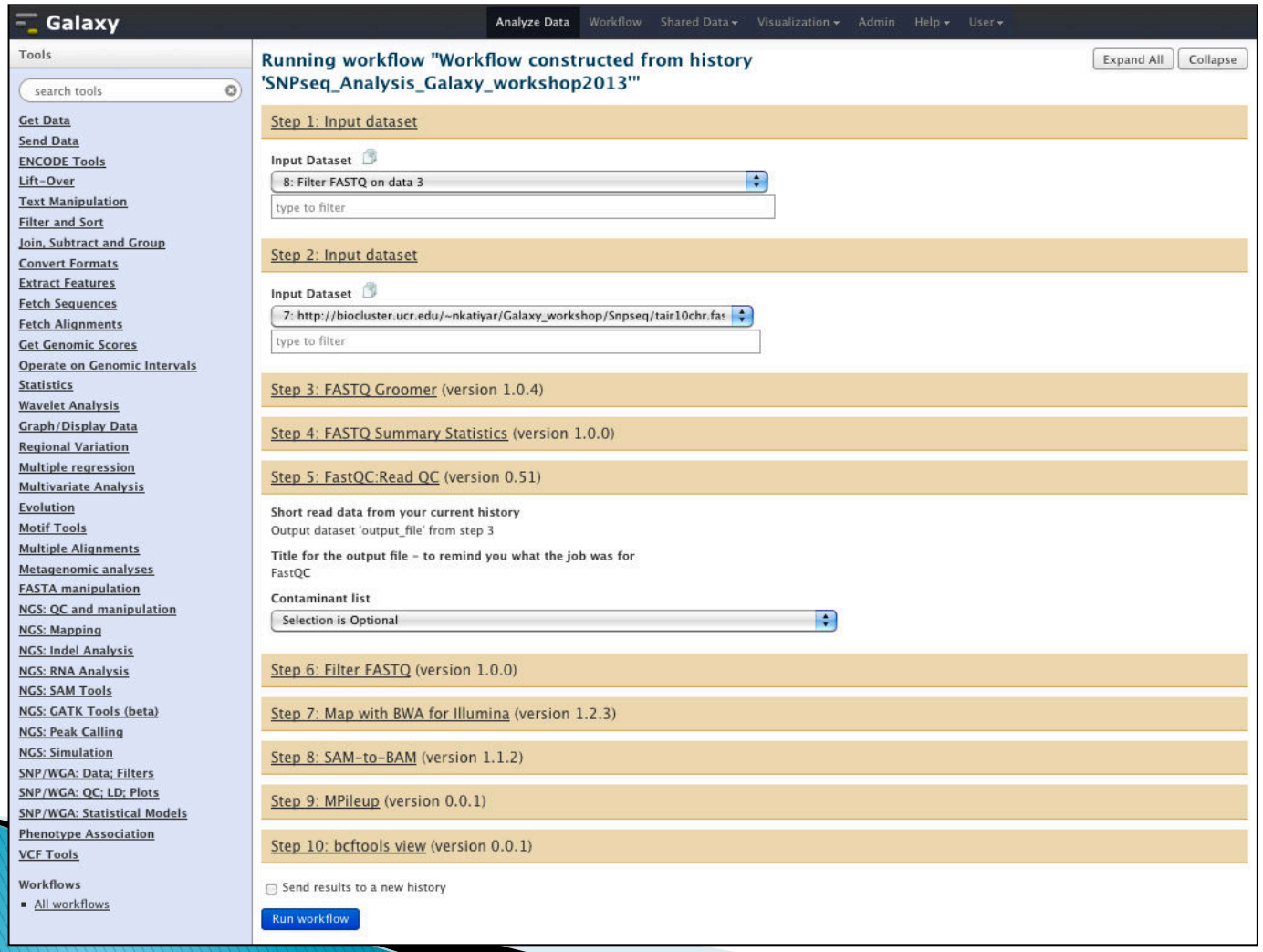
Data Libraries

- Datasets are accessible from within Galaxy or for download.



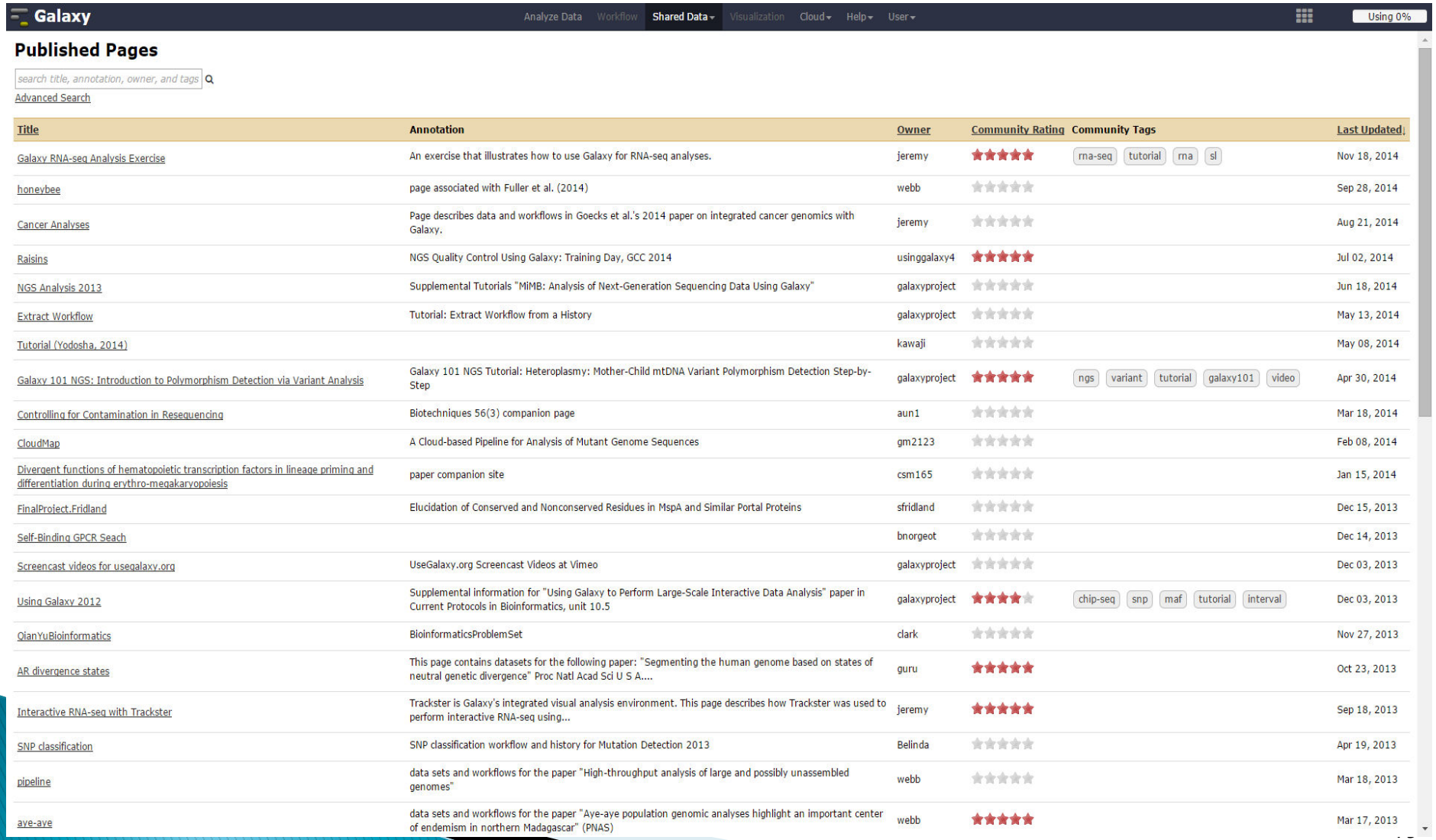
Workflows

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



Pages

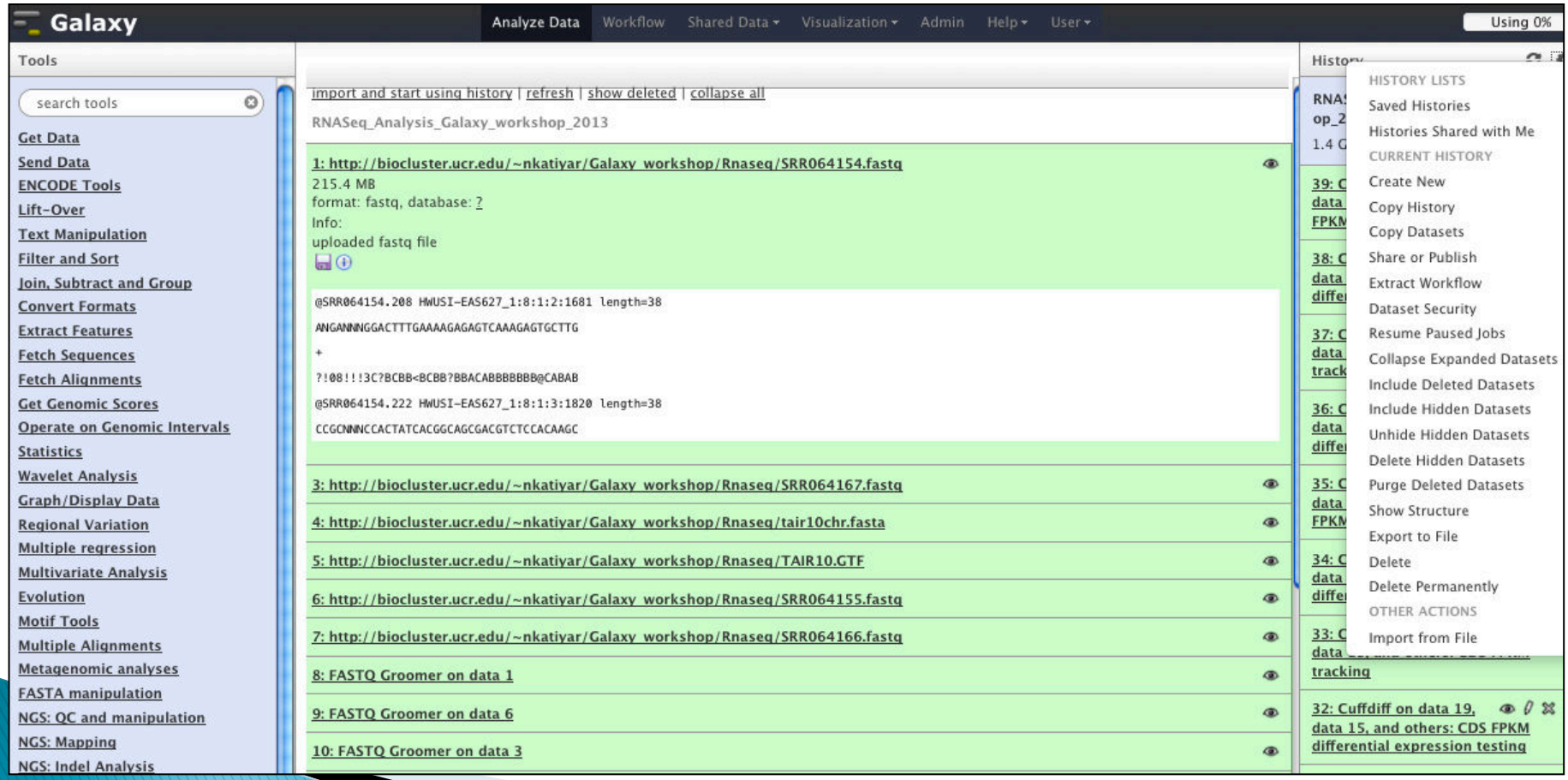
- Pages are documentation within the Galaxy that explain the steps and reasoning in a particular history or workflow.



Title	Annotation	Owner	Community Rating	Community Tags	Last Updated
Galaxy RNA-seq Analysis Exercise	An exercise that illustrates how to use Galaxy for RNA-seq analyses.	jeremy	★★★★★	rna-seq, tutorial, rna, 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
QianYuBioinformatics	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-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
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
aye-aye	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.



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).

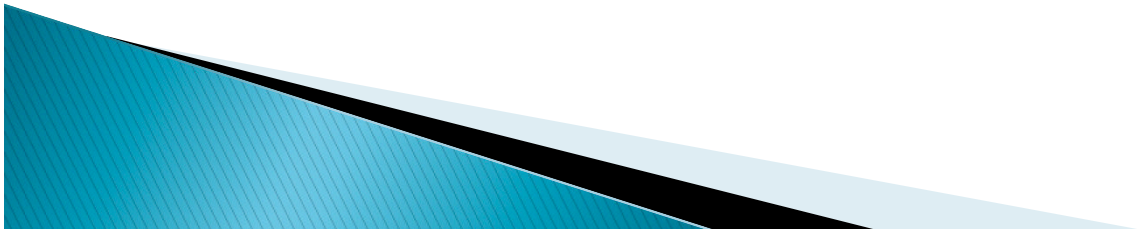
Outline

- 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
```

FASTQ Quality Score (Link)

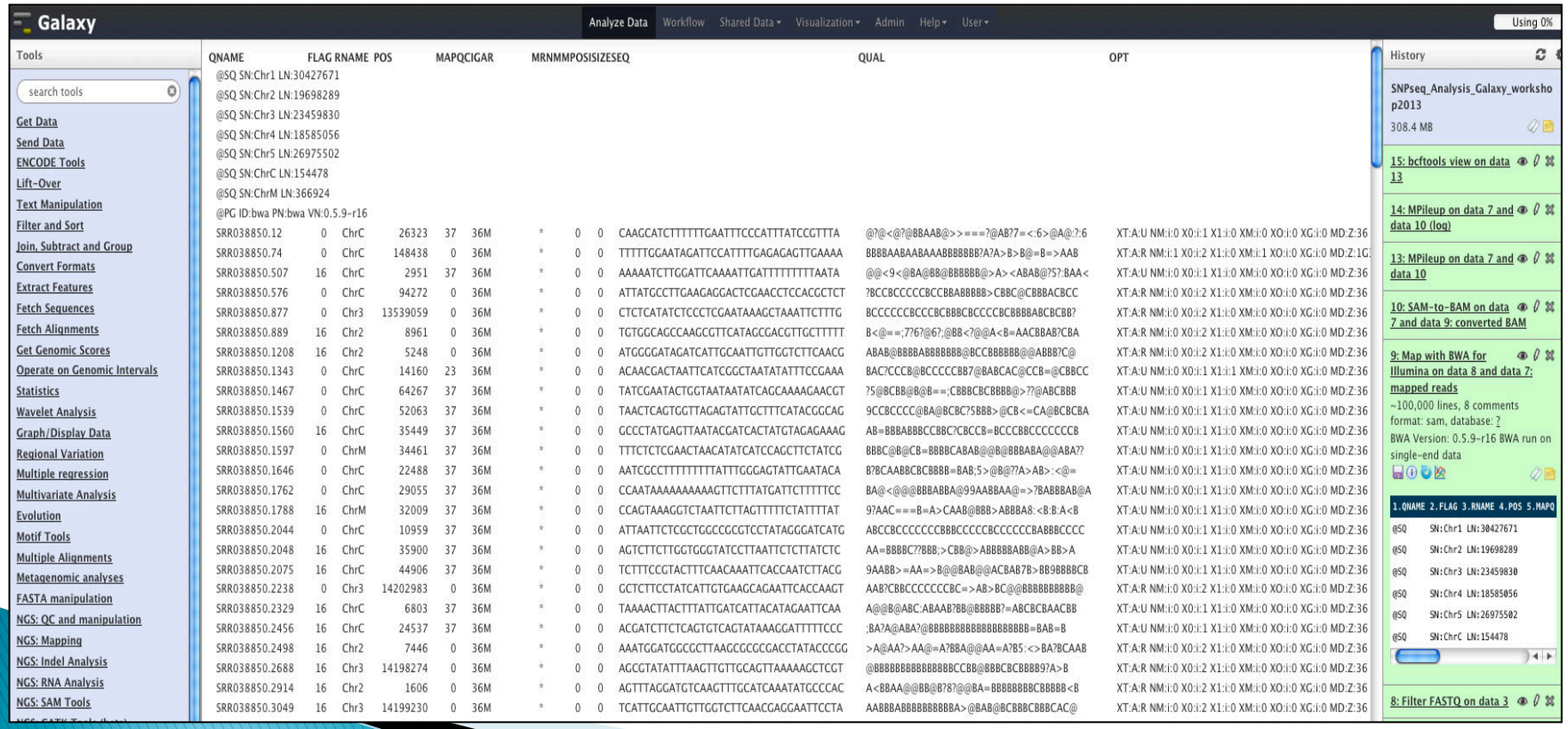
- Quality score represents the error probability of a given basecall.
- In a FASTQ file, quality scores are often represented using the ASCII alphabet.
- For example, a Phred score of 40 can be represented as the ASCII char “I” (40+33= ASCII #73), and an Illumina score of 40 as “h” (40+64=ASCII #104).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

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0.2......................26...31.........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 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



GFF and GTF format

- General Feature Format (GFF) (Link)

```
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
```

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

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

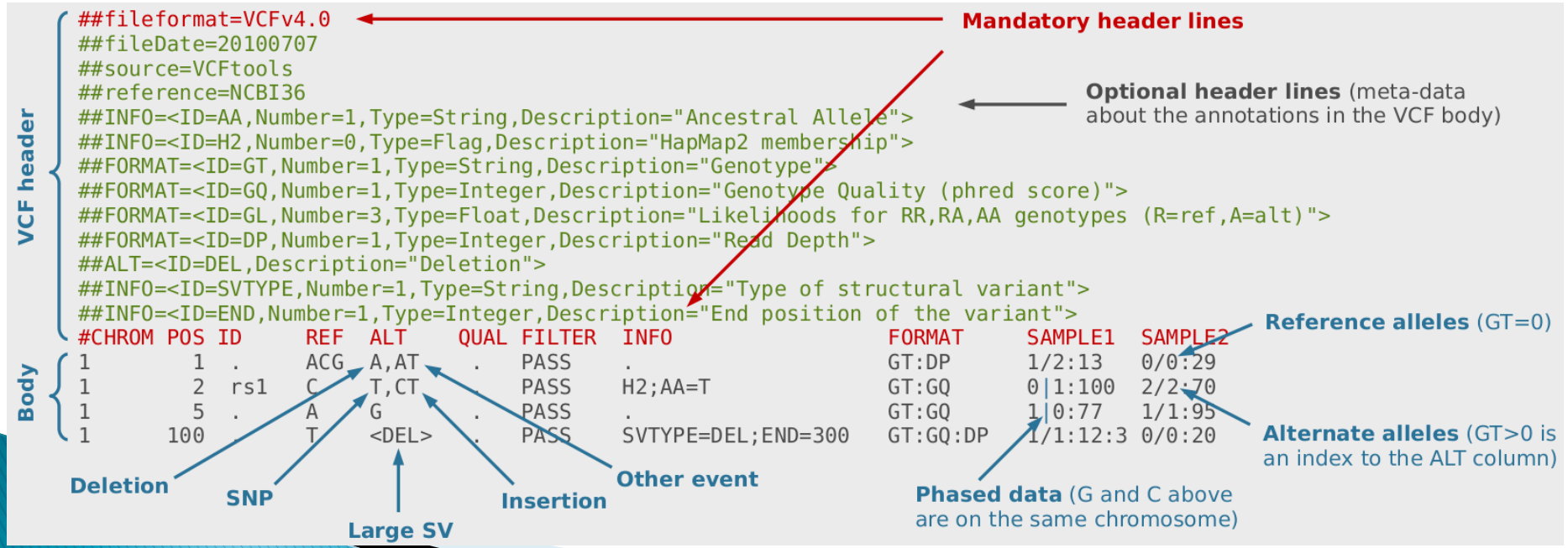
BED format (Browser Extensible Data) (Link)

- 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

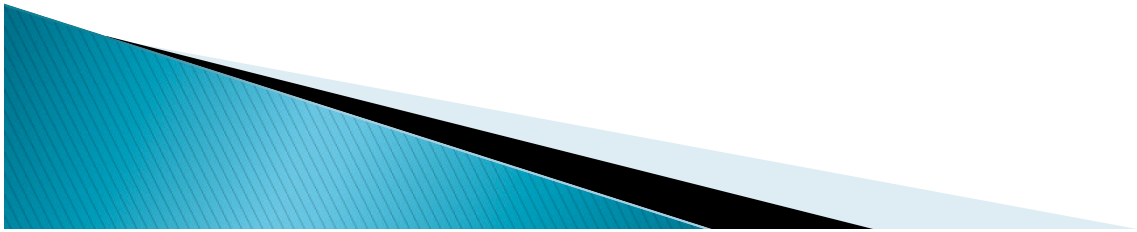


Available NGS Analysis Toolsets

- 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

Data intensive biology *for everyone.*

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

Use Galaxy

Use project's free server or other public servers

Get Galaxy

Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy

Screencasts, Galaxy 101, ...

Get Involved

Mailing lists, Tool Shed, wiki

Search all resources

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

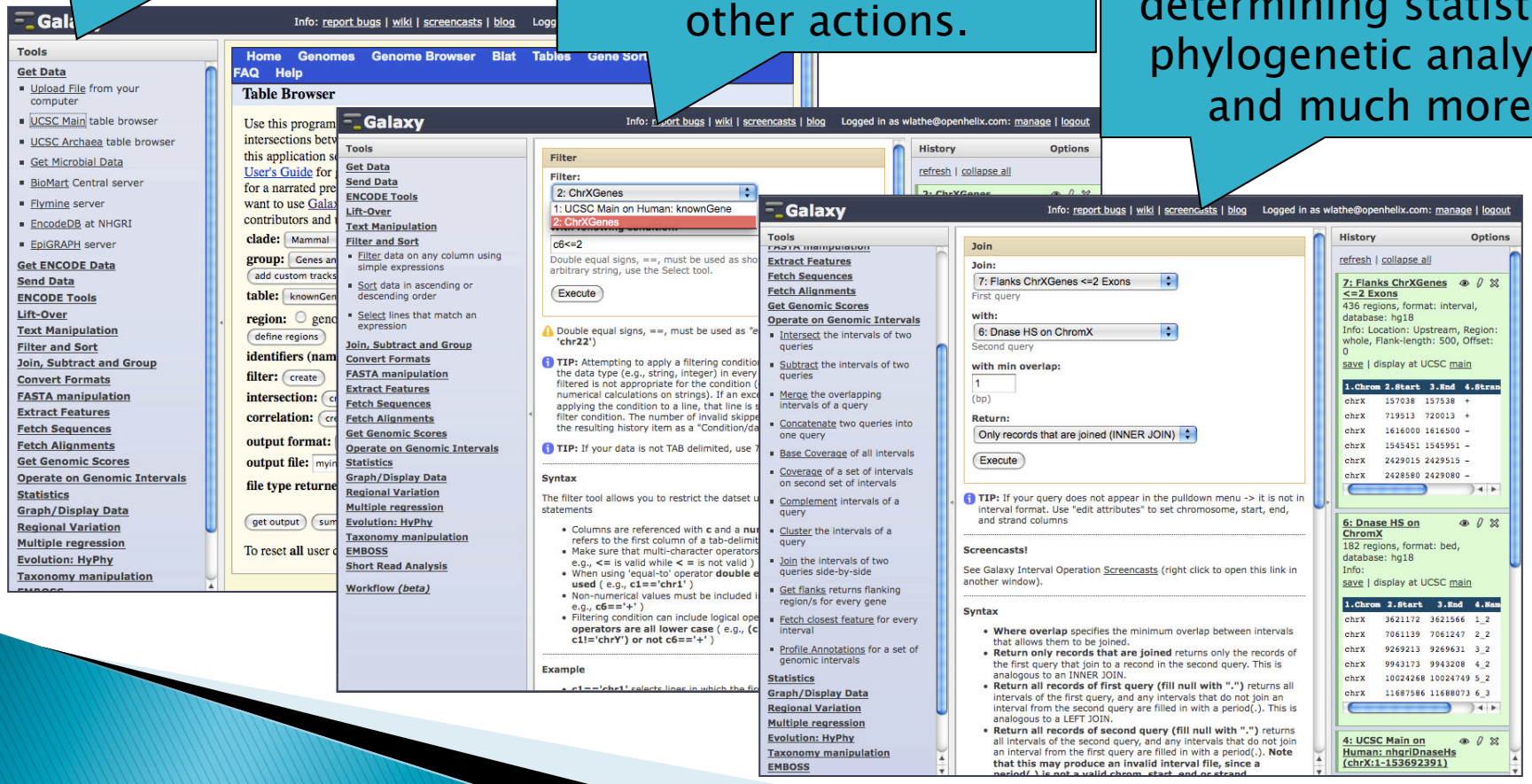
http://galaxyproject.org/

Galaxy Conceptual Framework

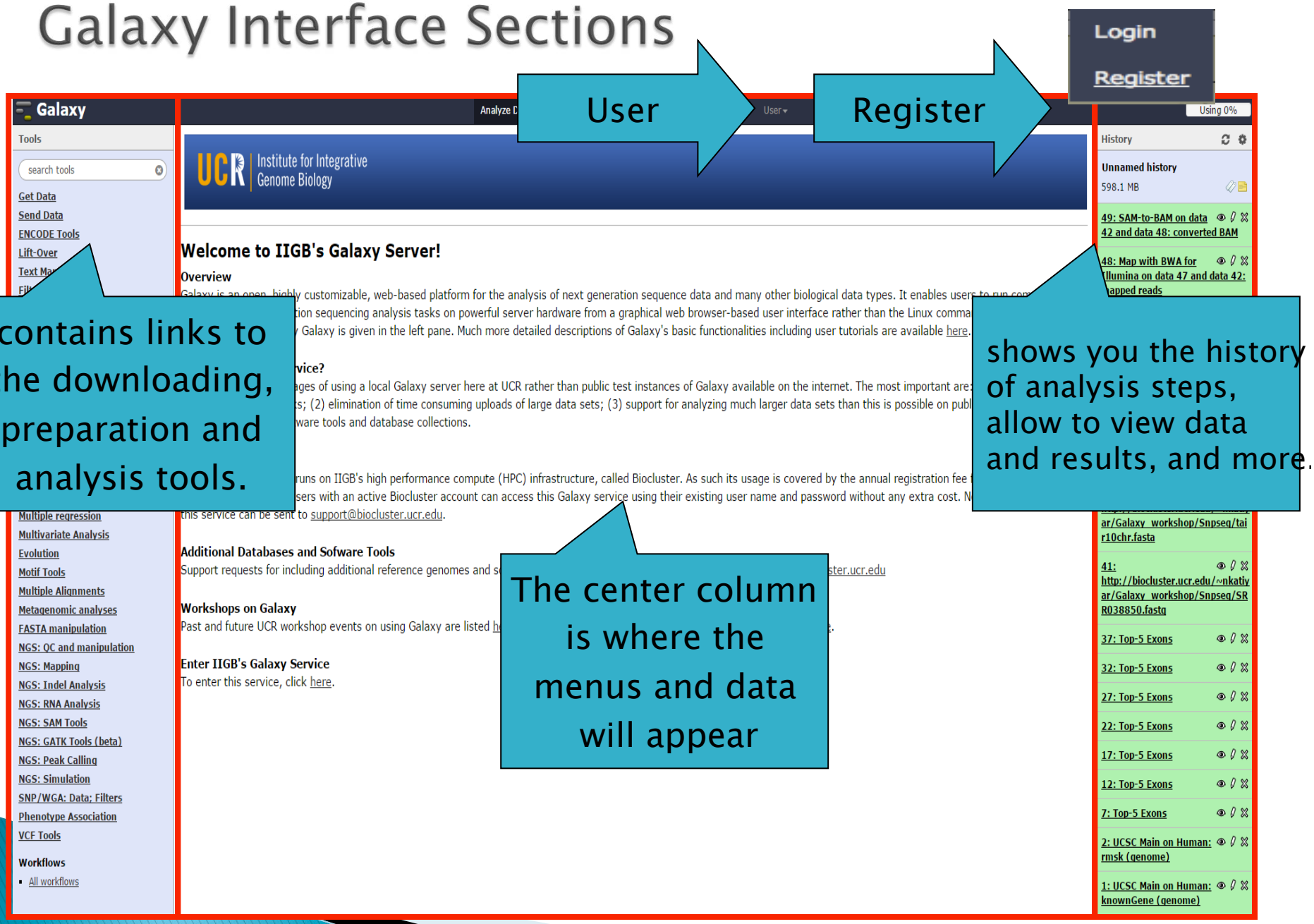
Obtain data from many data sources including the UCSC Table Browser, BioMart, WormBase, or your own data.

Prepare data for further analysis by rearranging or cutting data columns, filtering data and many other actions.

Analyze data by finding overlapping regions, determining statistics, phylogenetic analysis and much more



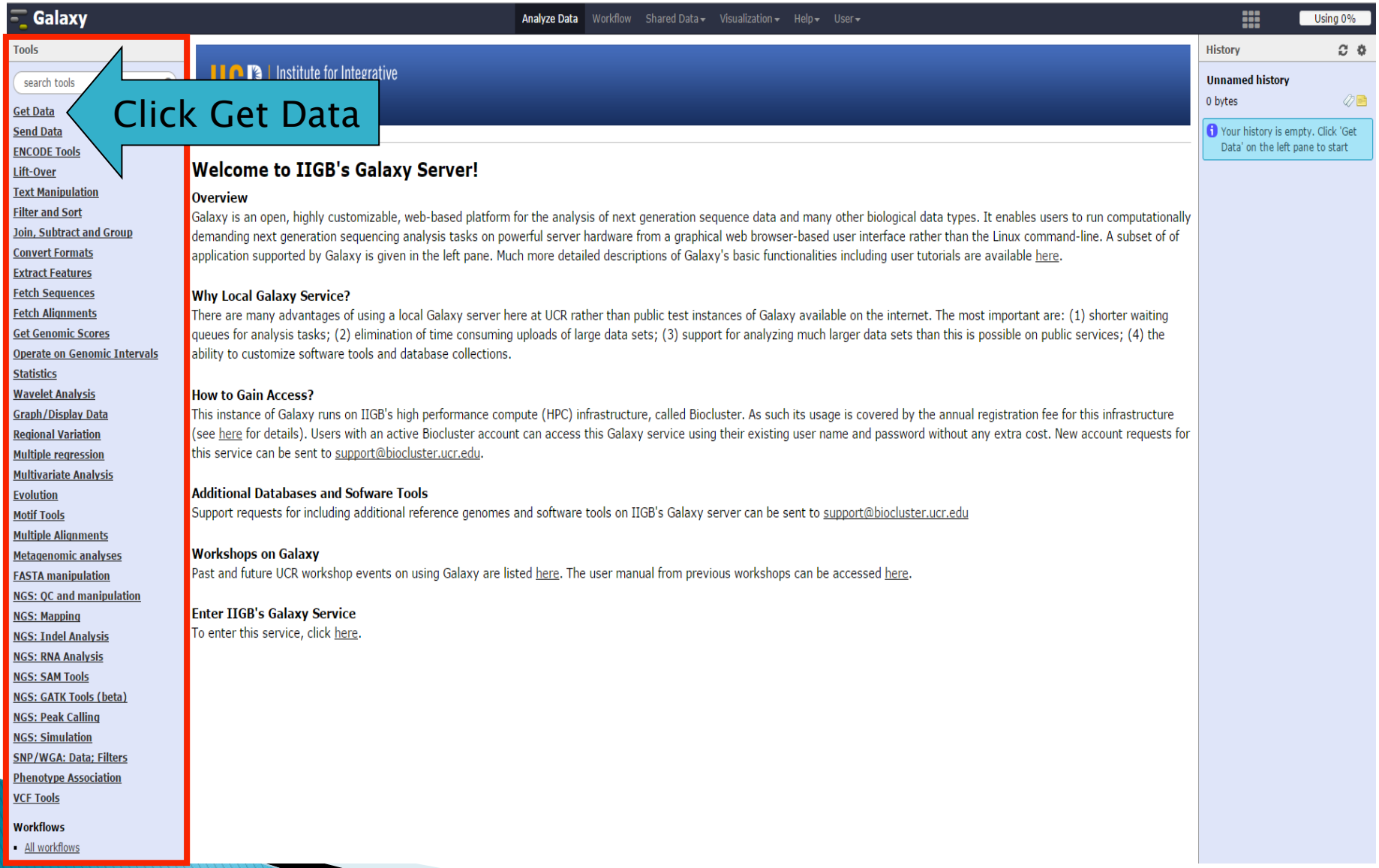
Galaxy Interface Sections



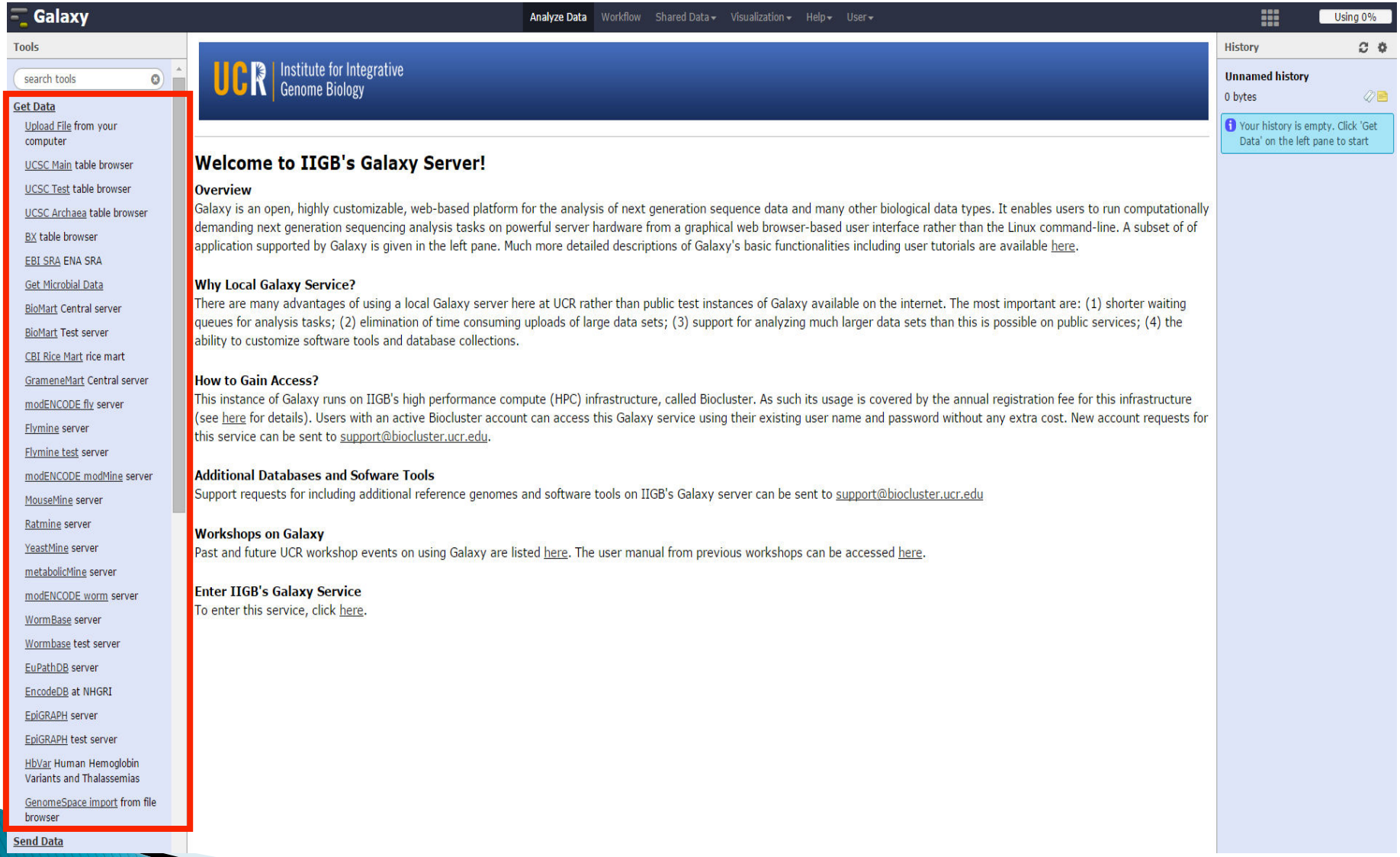
NGS Analysis Using Galaxy

- 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

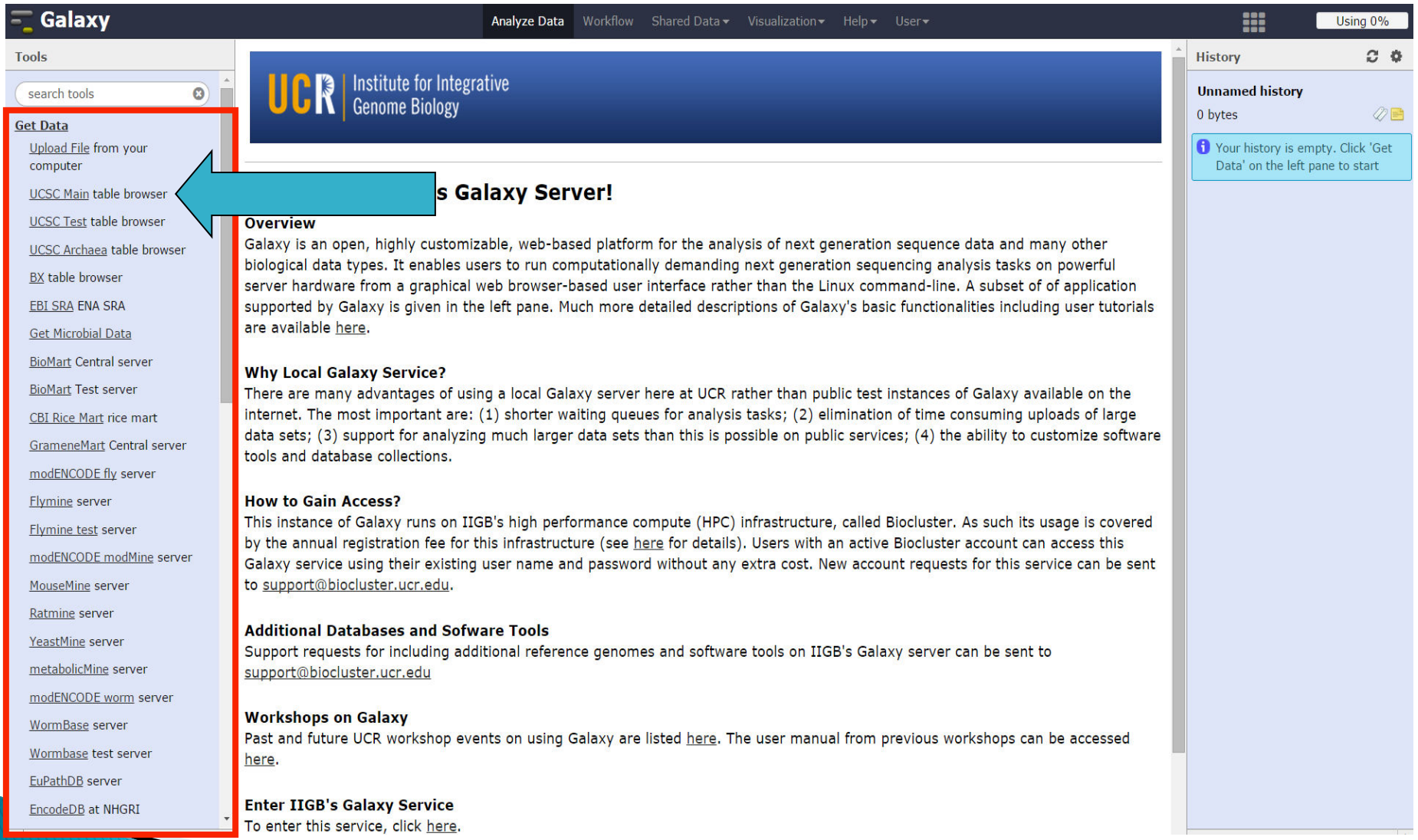
Getting Data



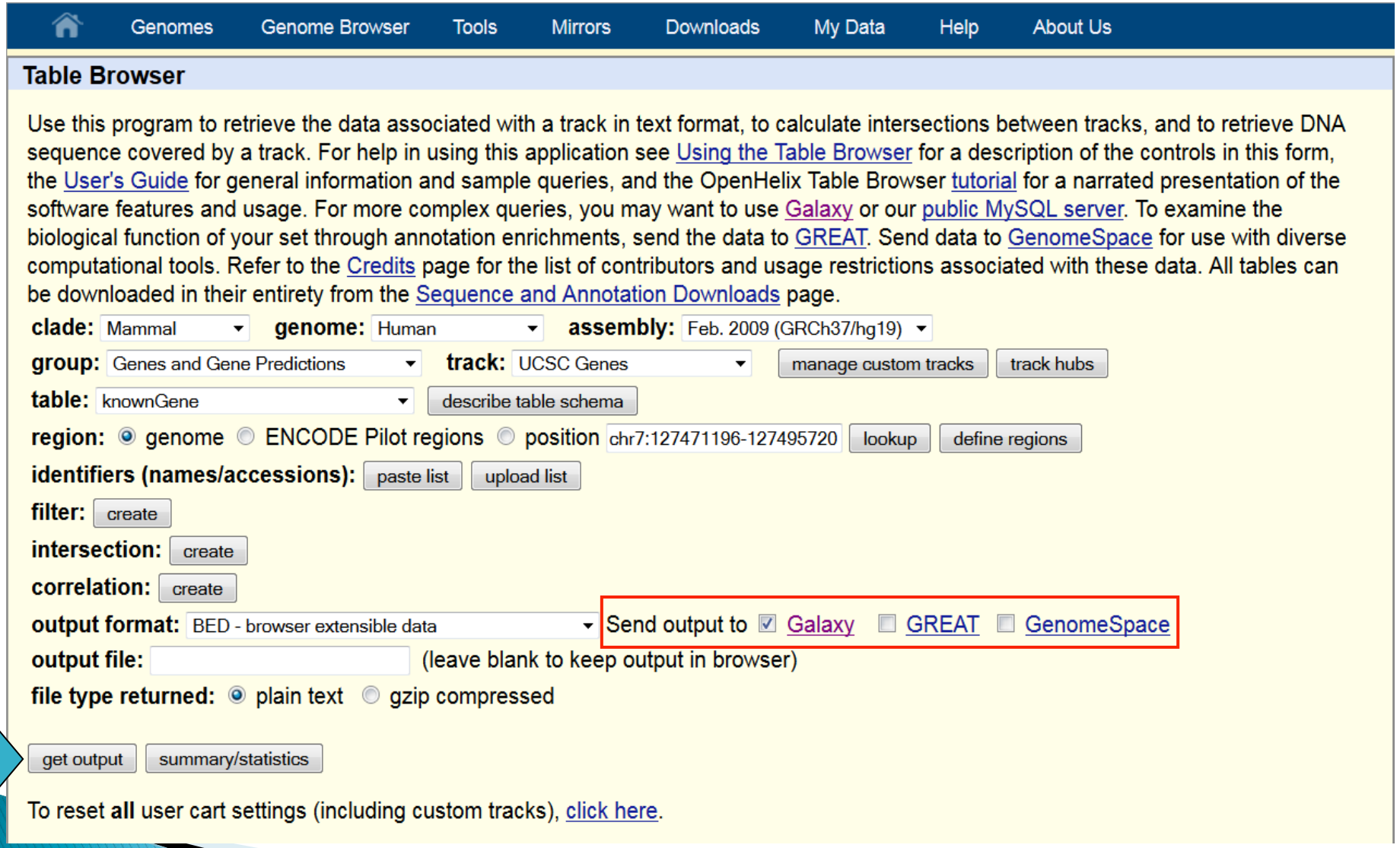
Getting Data



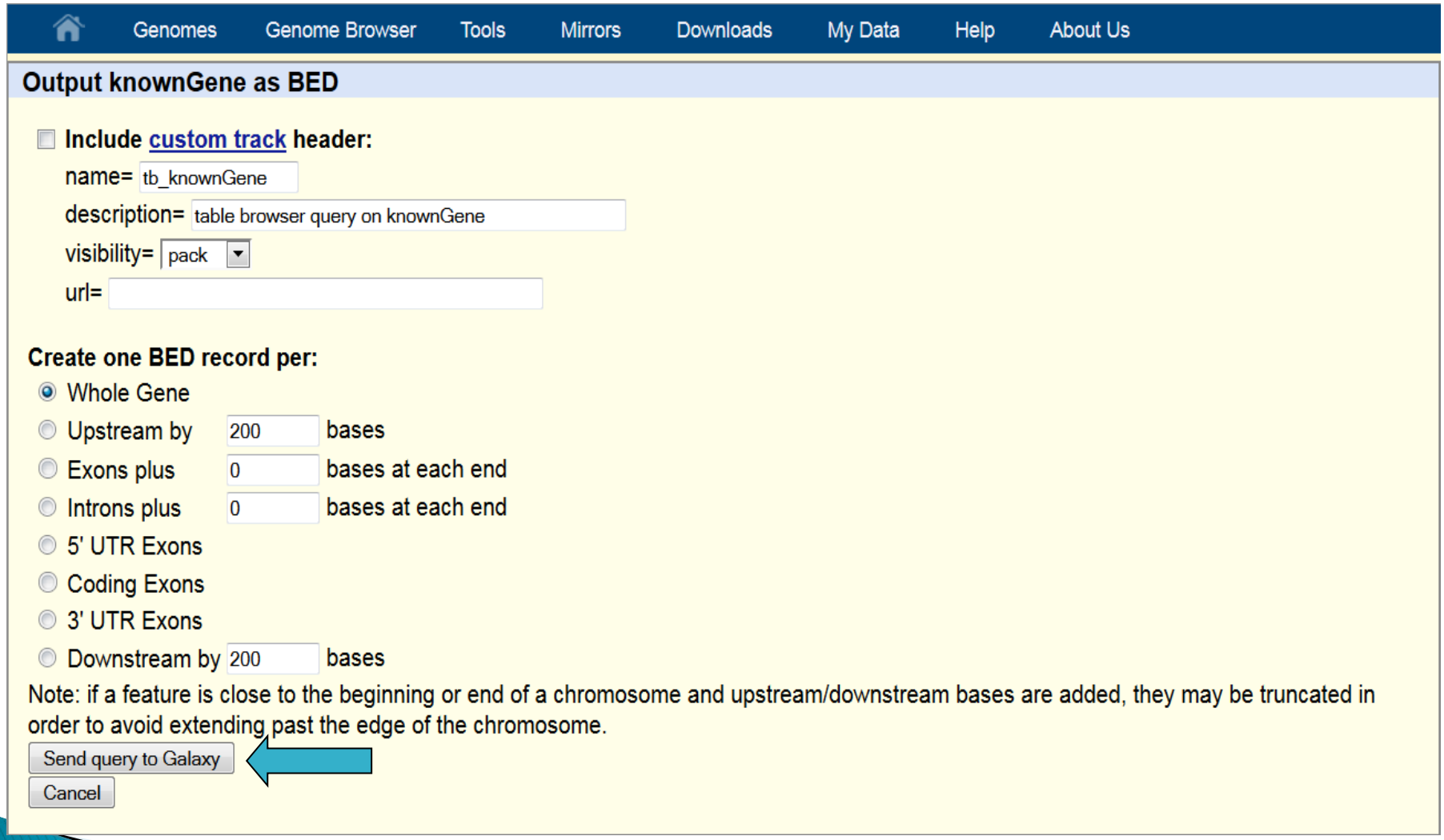
Import data from UCSC genome browser



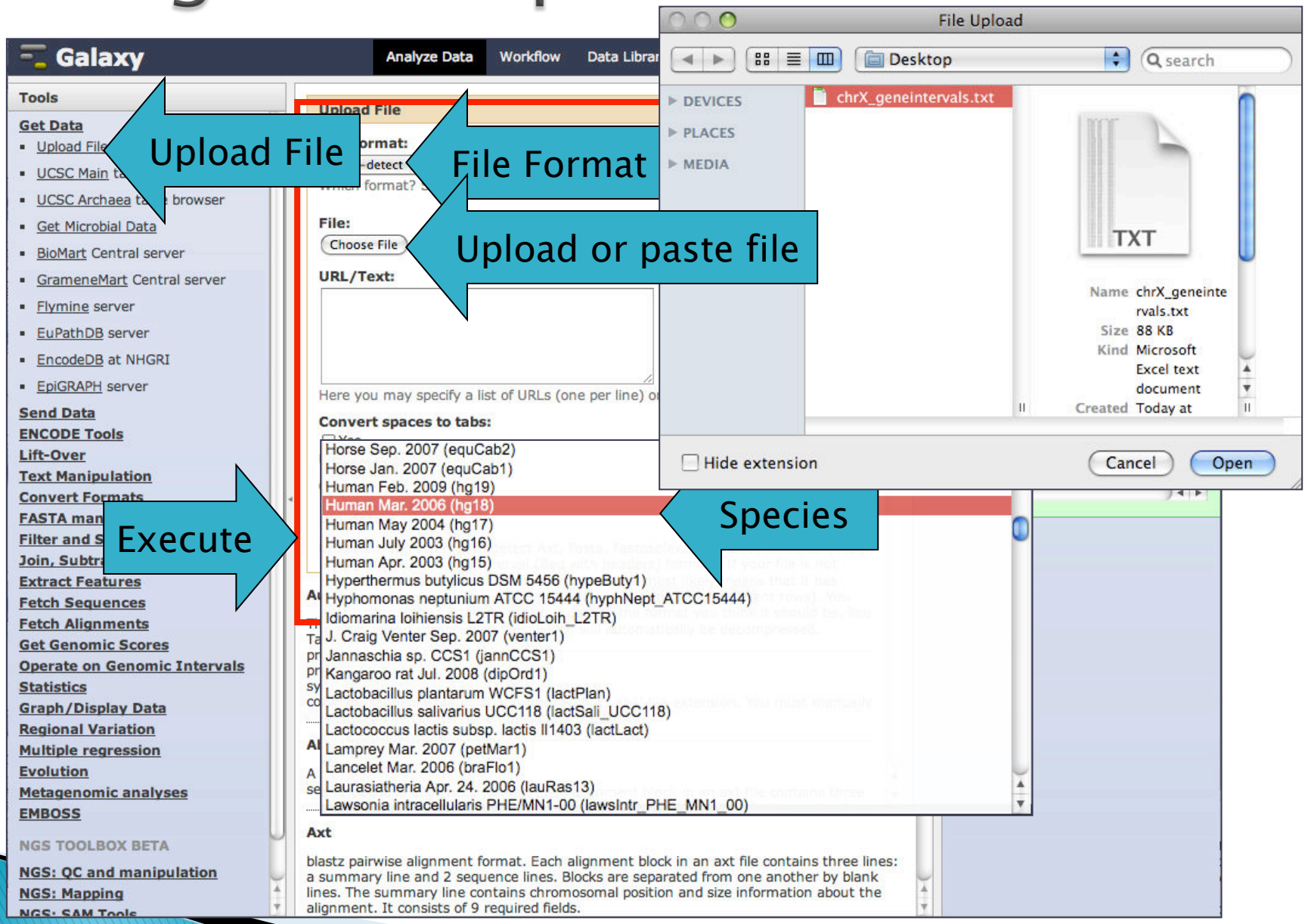
Import data from UCSC genome browser



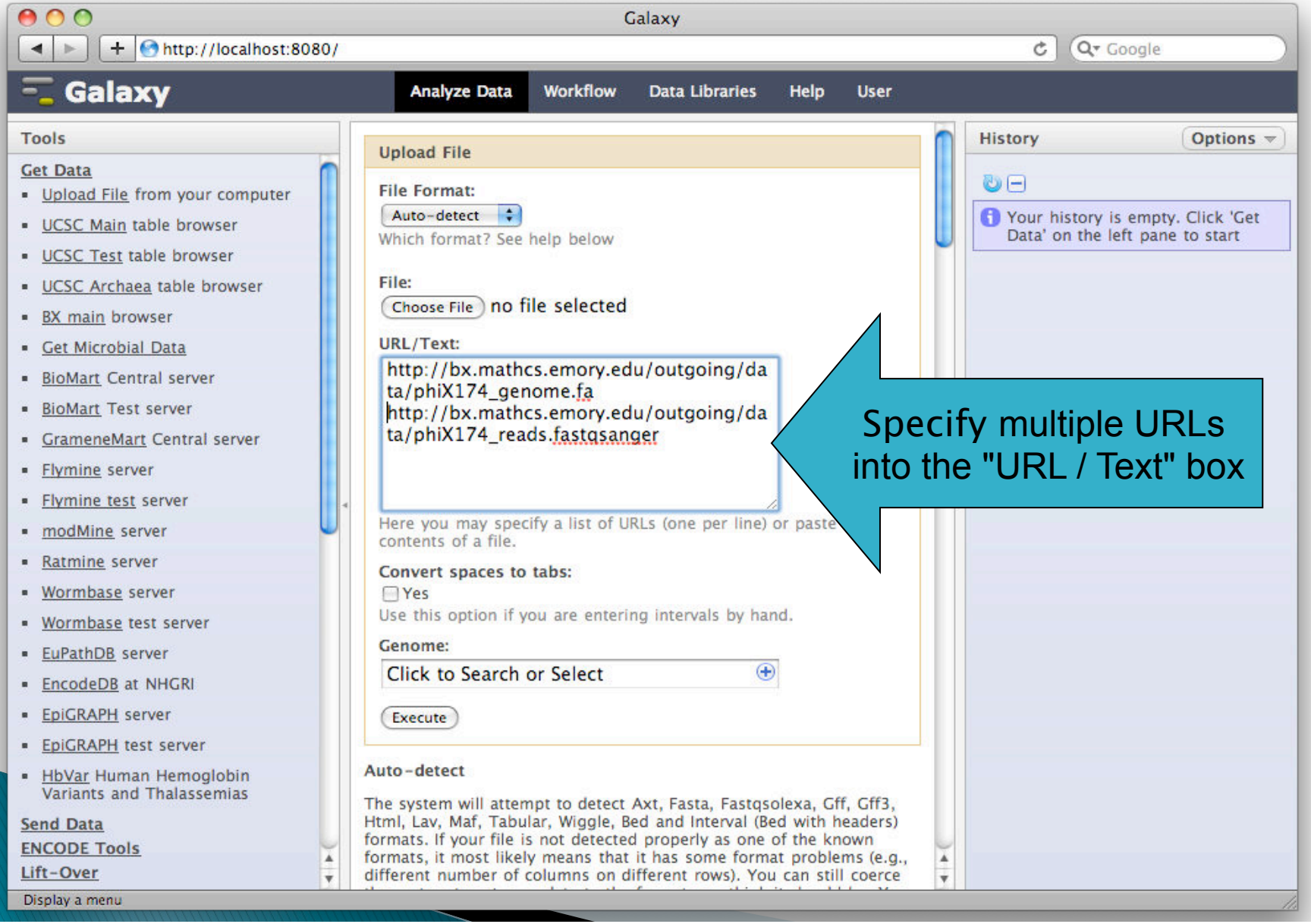
Send query to Galaxy from UCSC genome browser



Getting Data: Upload File



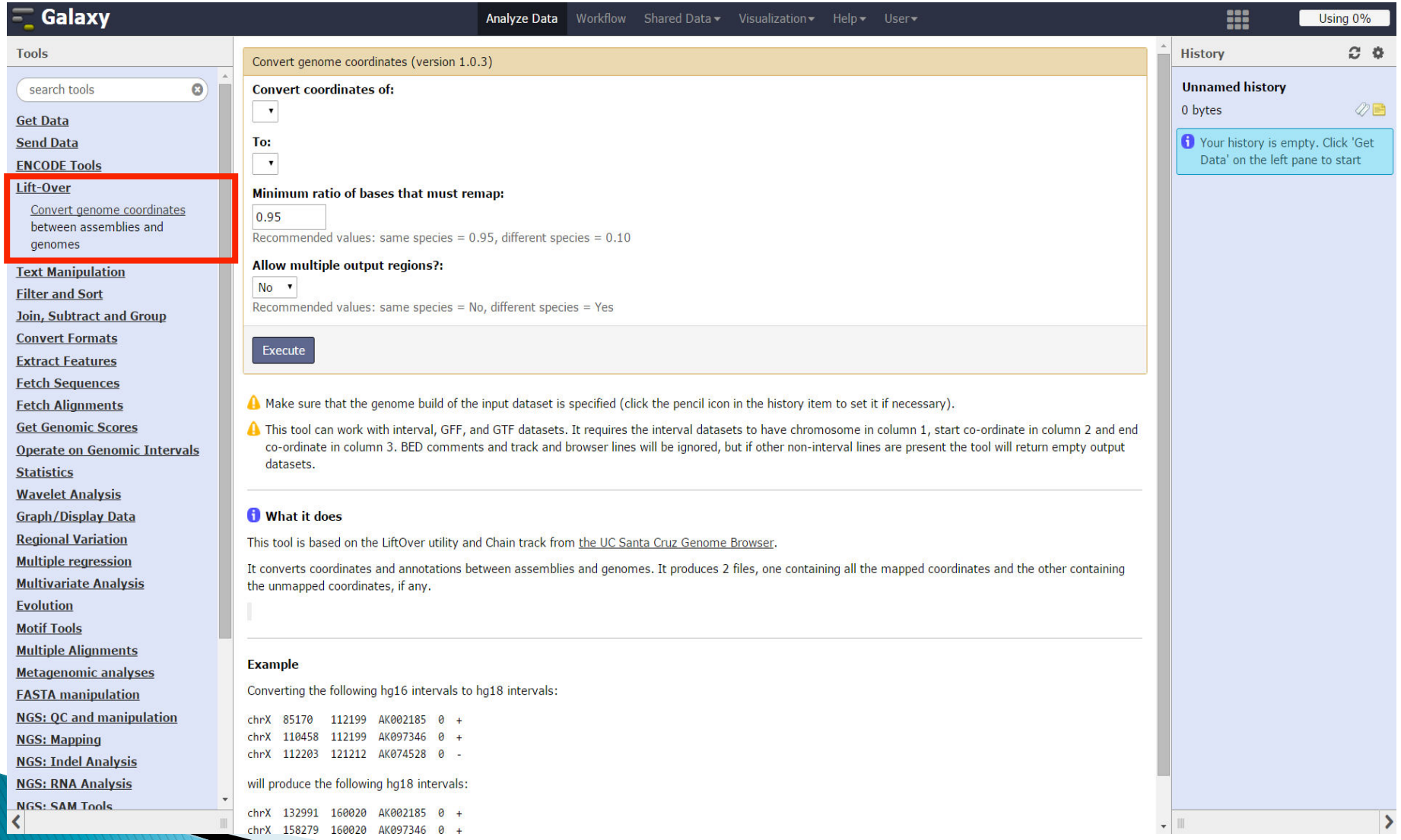
Getting Data: Upload File



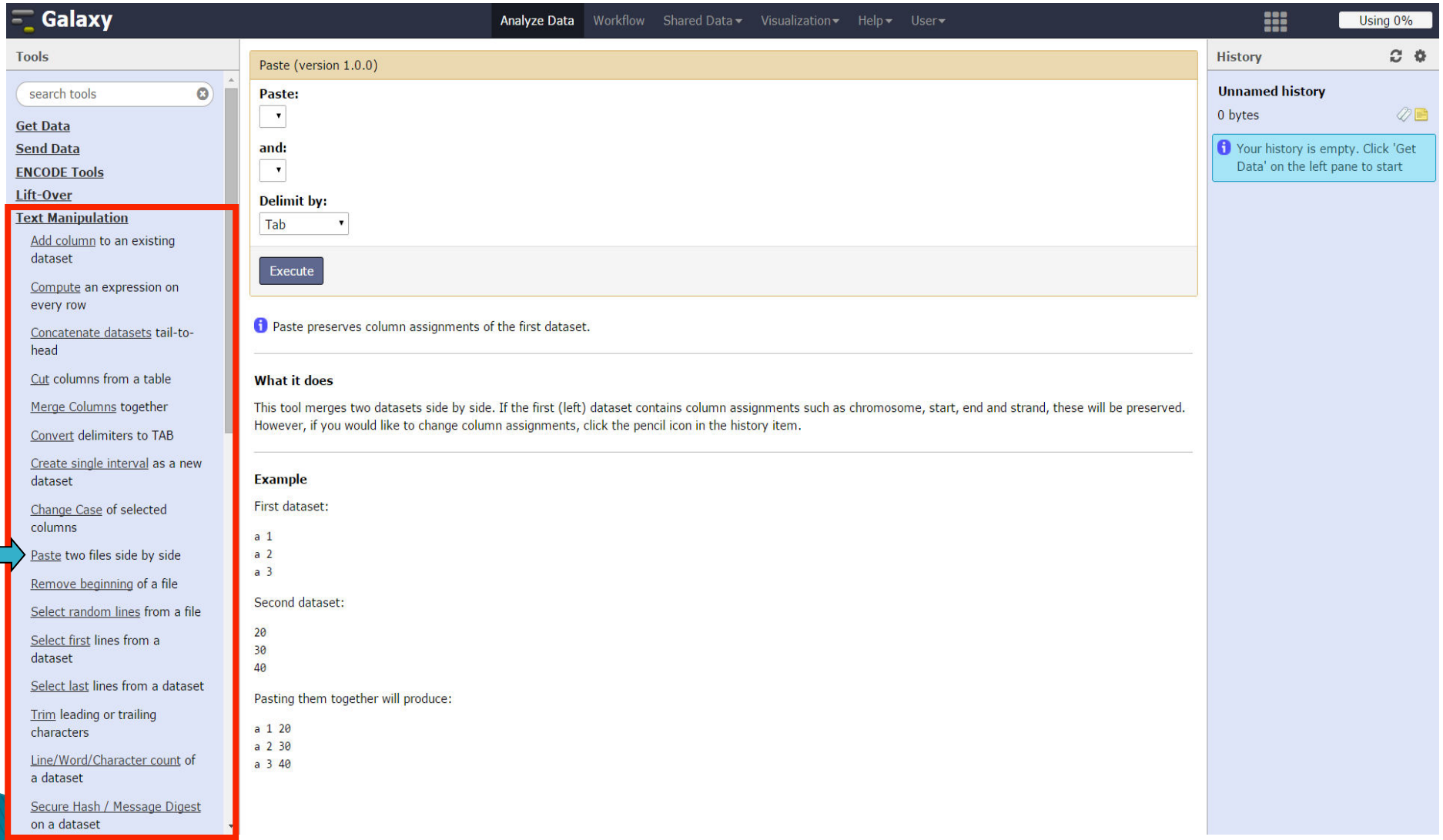
NGS Analysis Using Galaxy

- 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

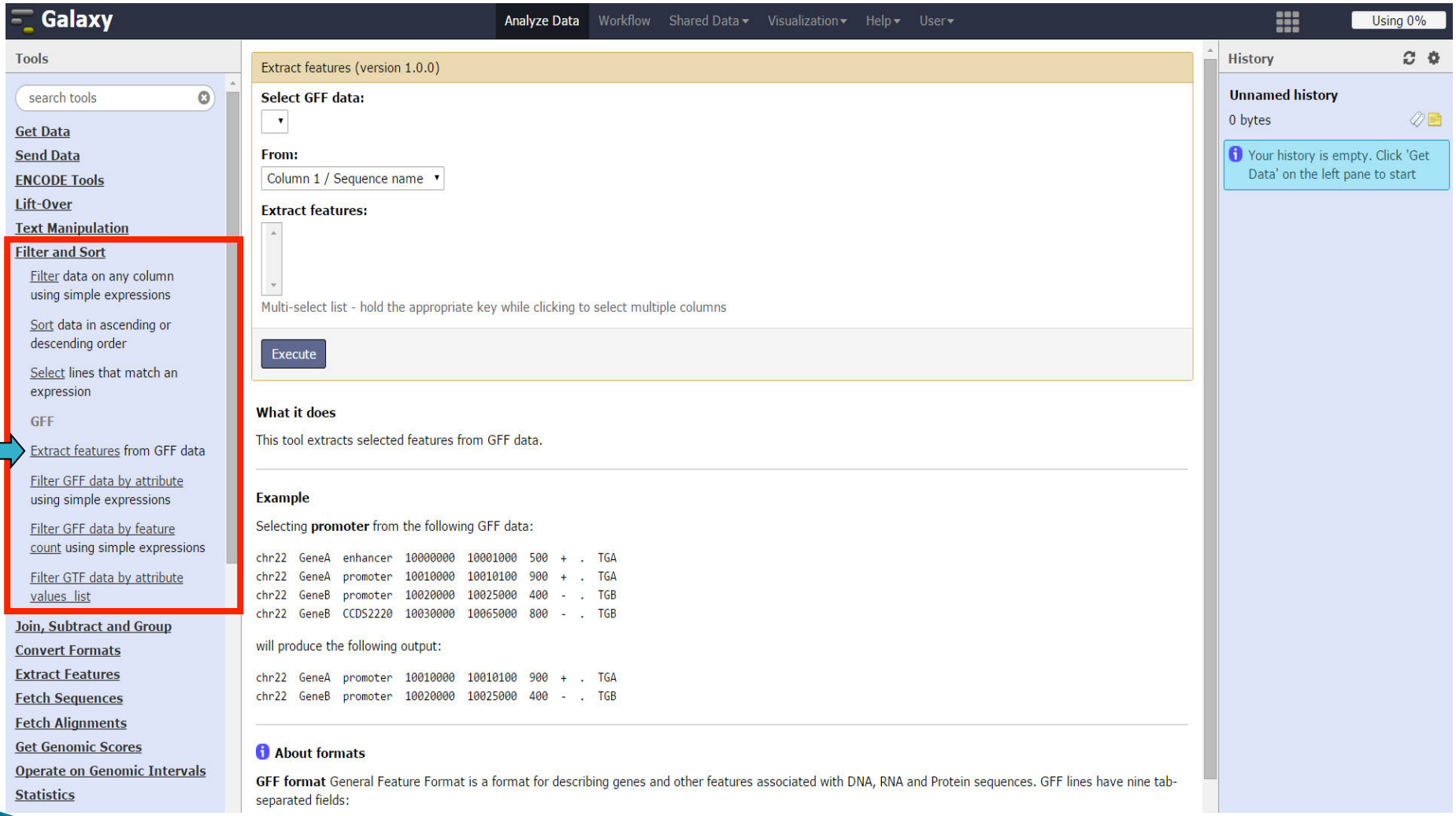
Lift-Over: convert genome coordinates



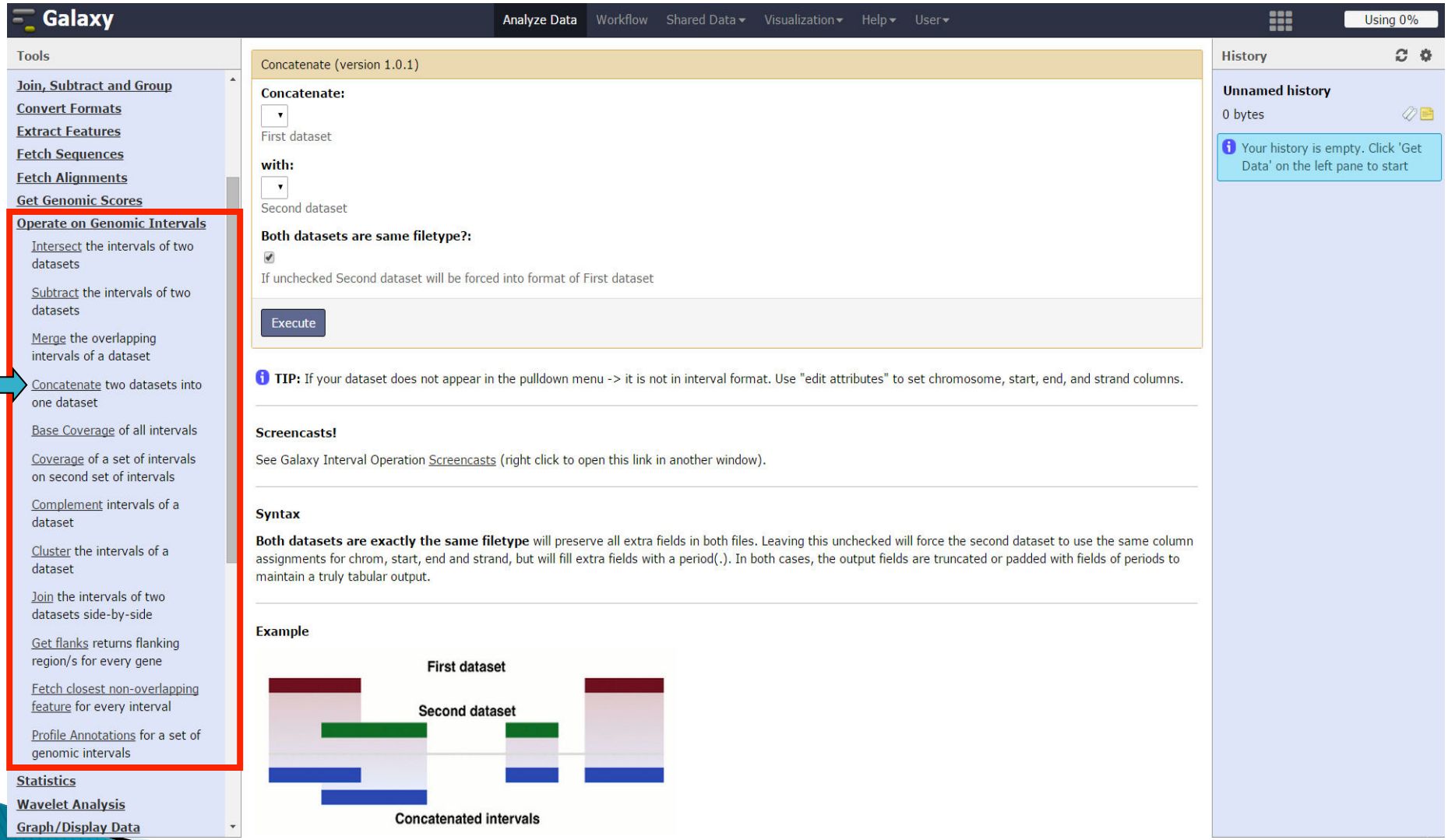
Text manipulation tools



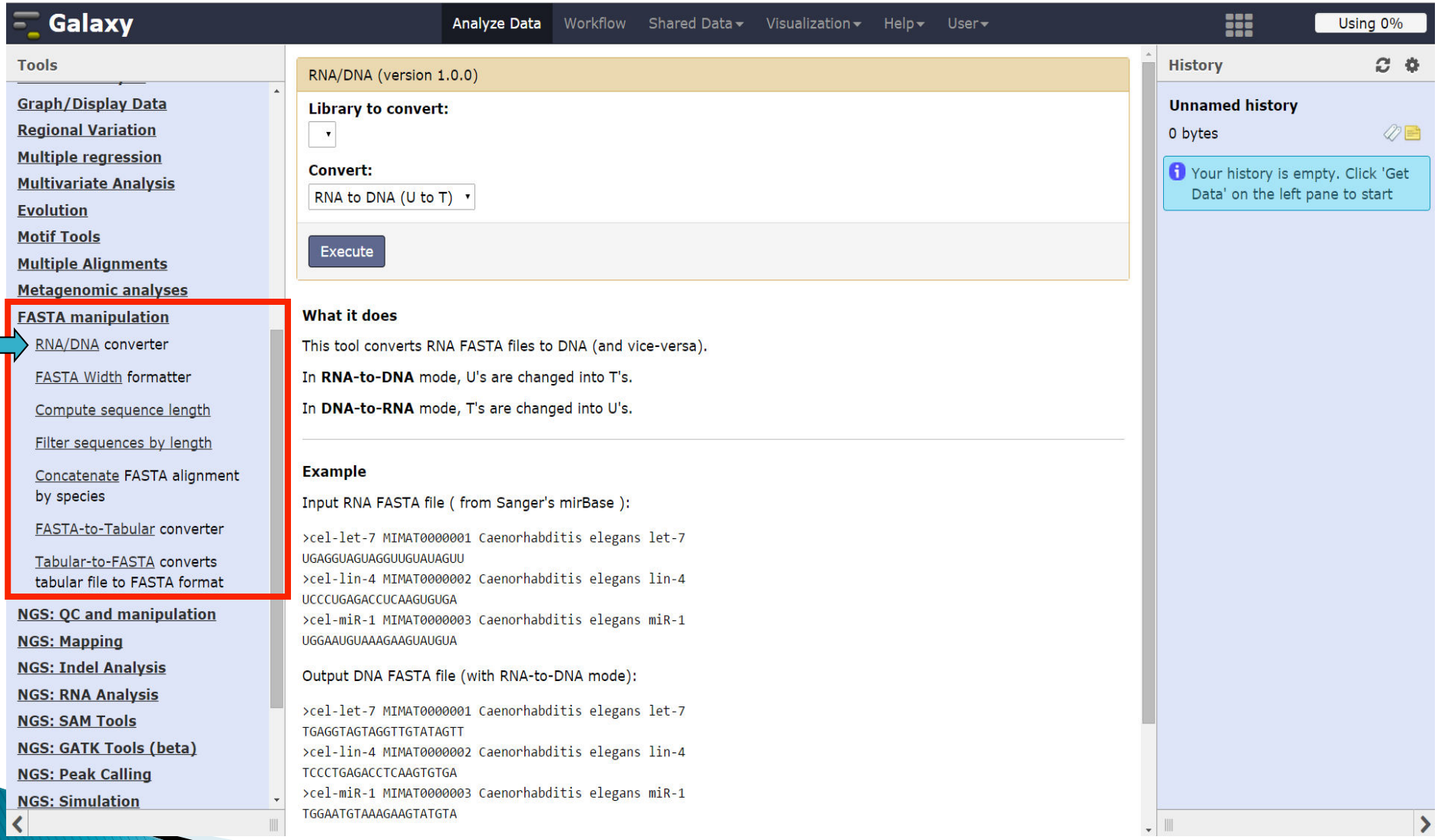
Filter and Sort on Galaxy



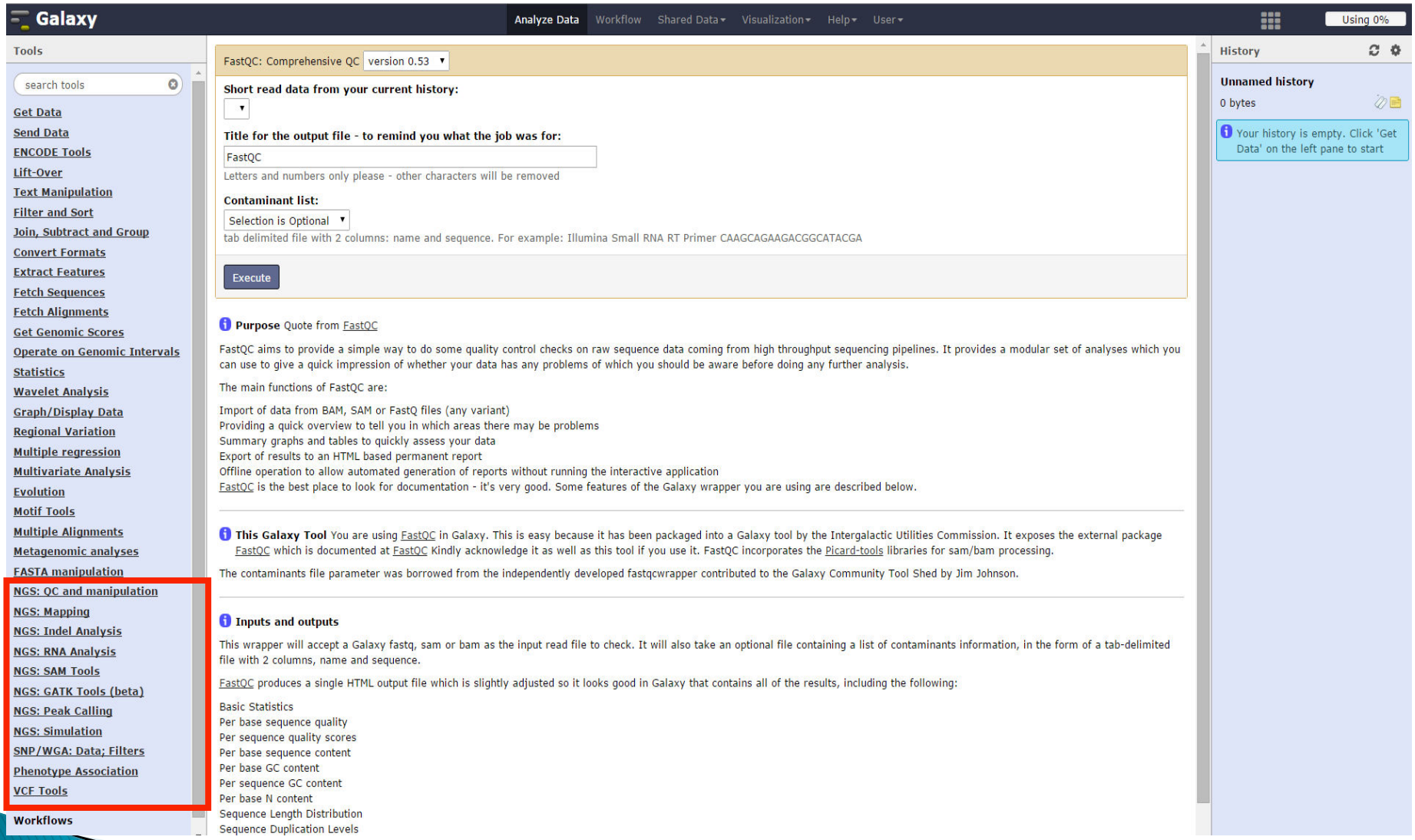
Operate on Genomic Intervals



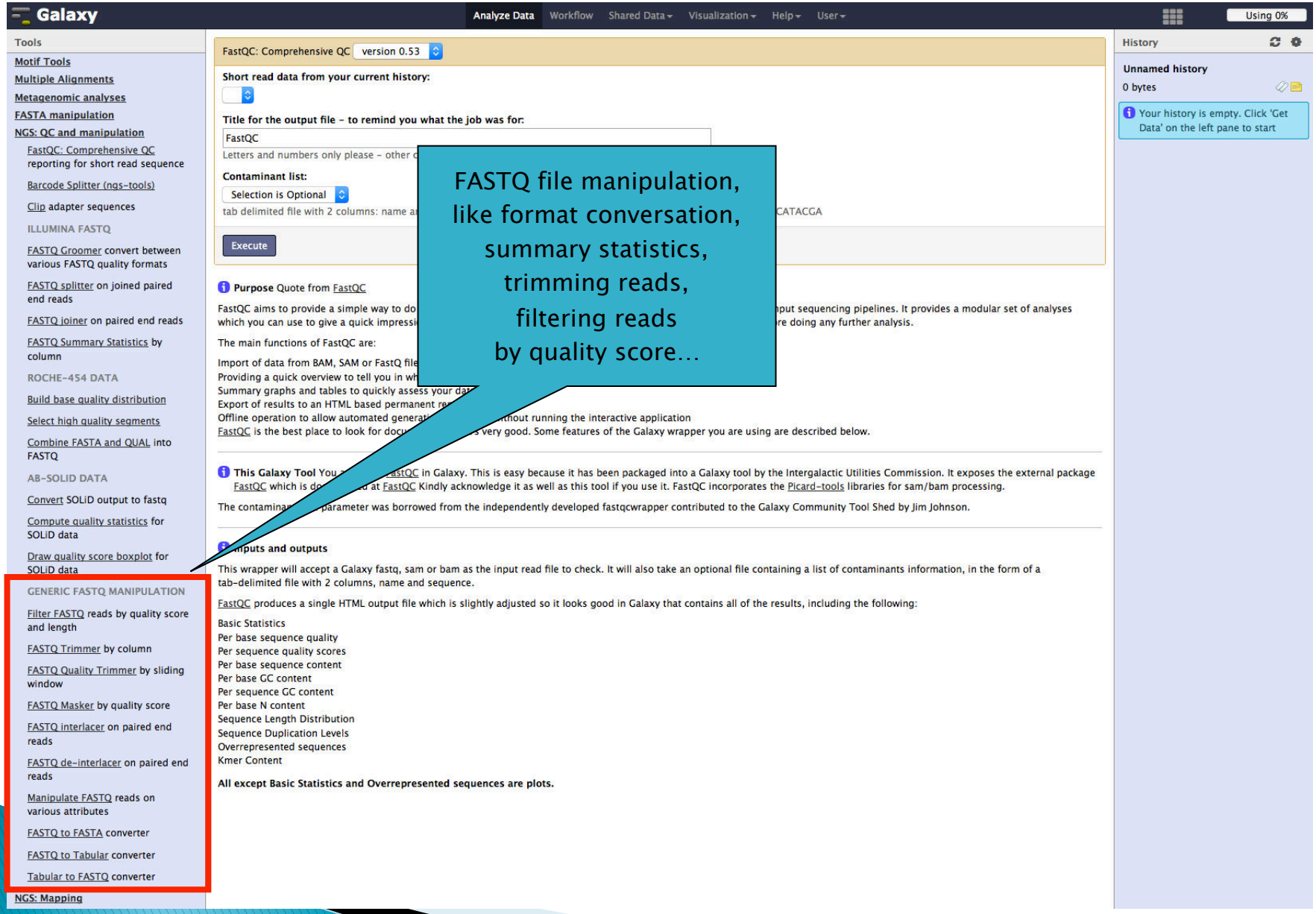
FASTA manipulation



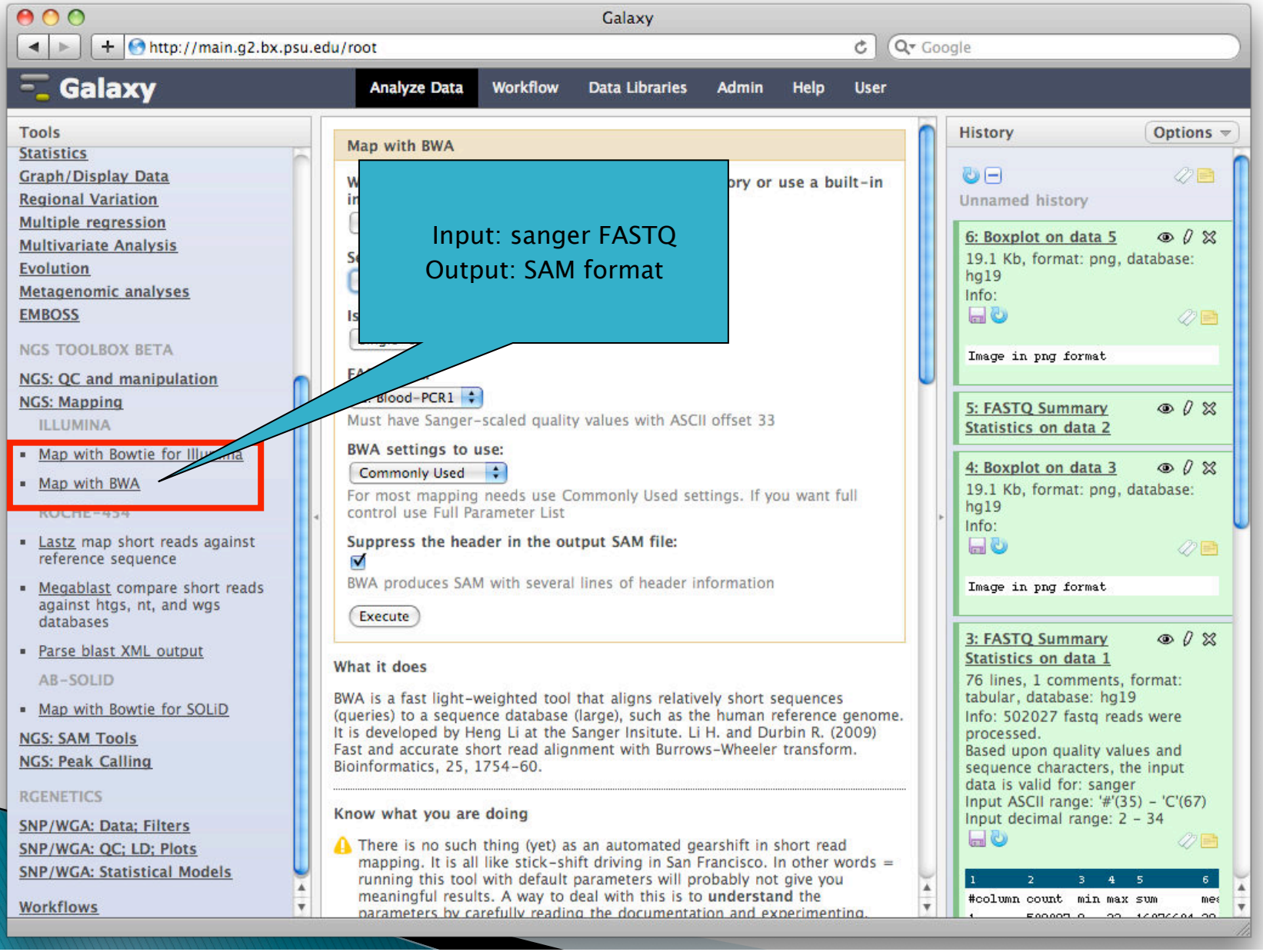
Analyzing Data: Next Generation Sequencing



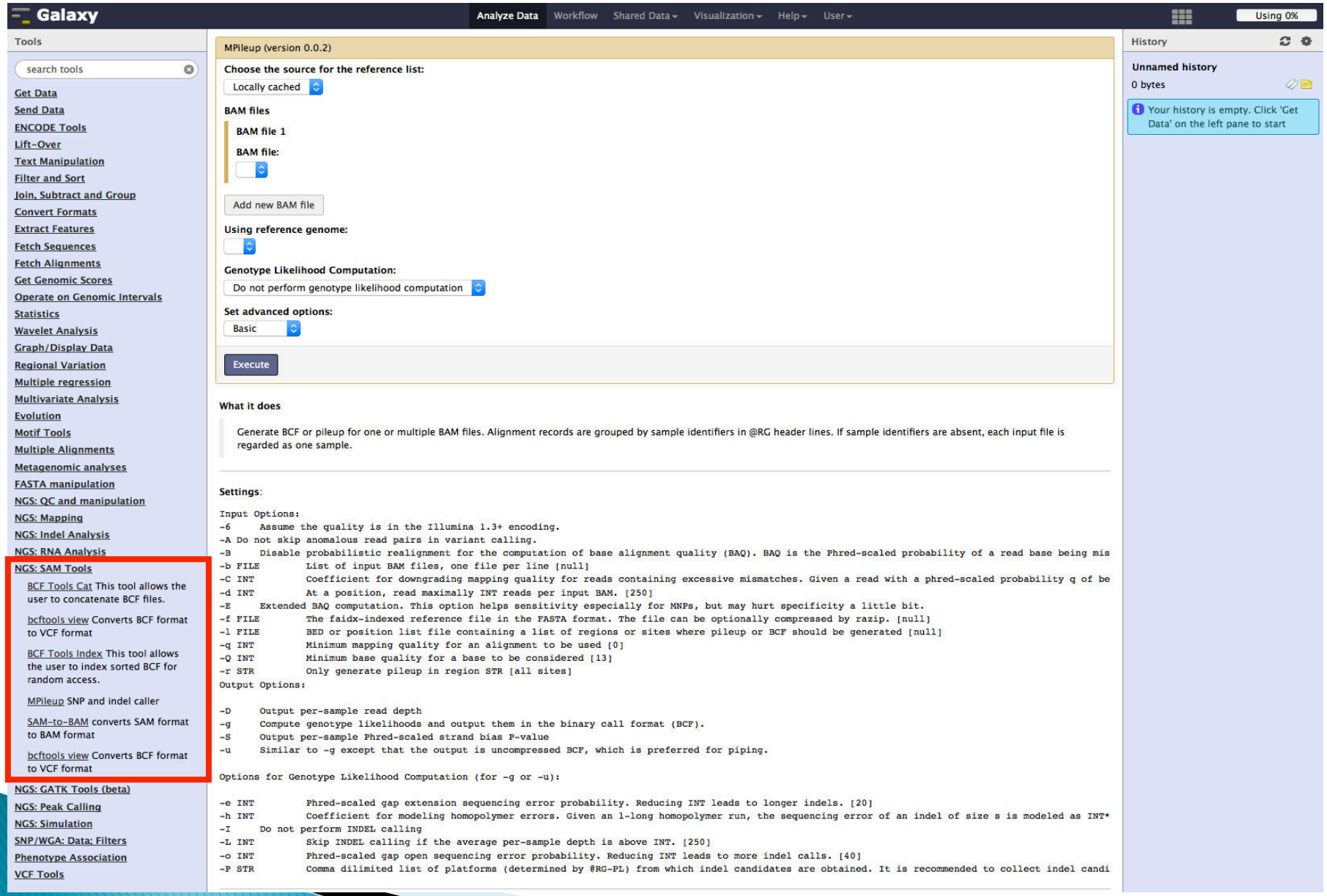
Analyzing Data: Next Generation Sequencing



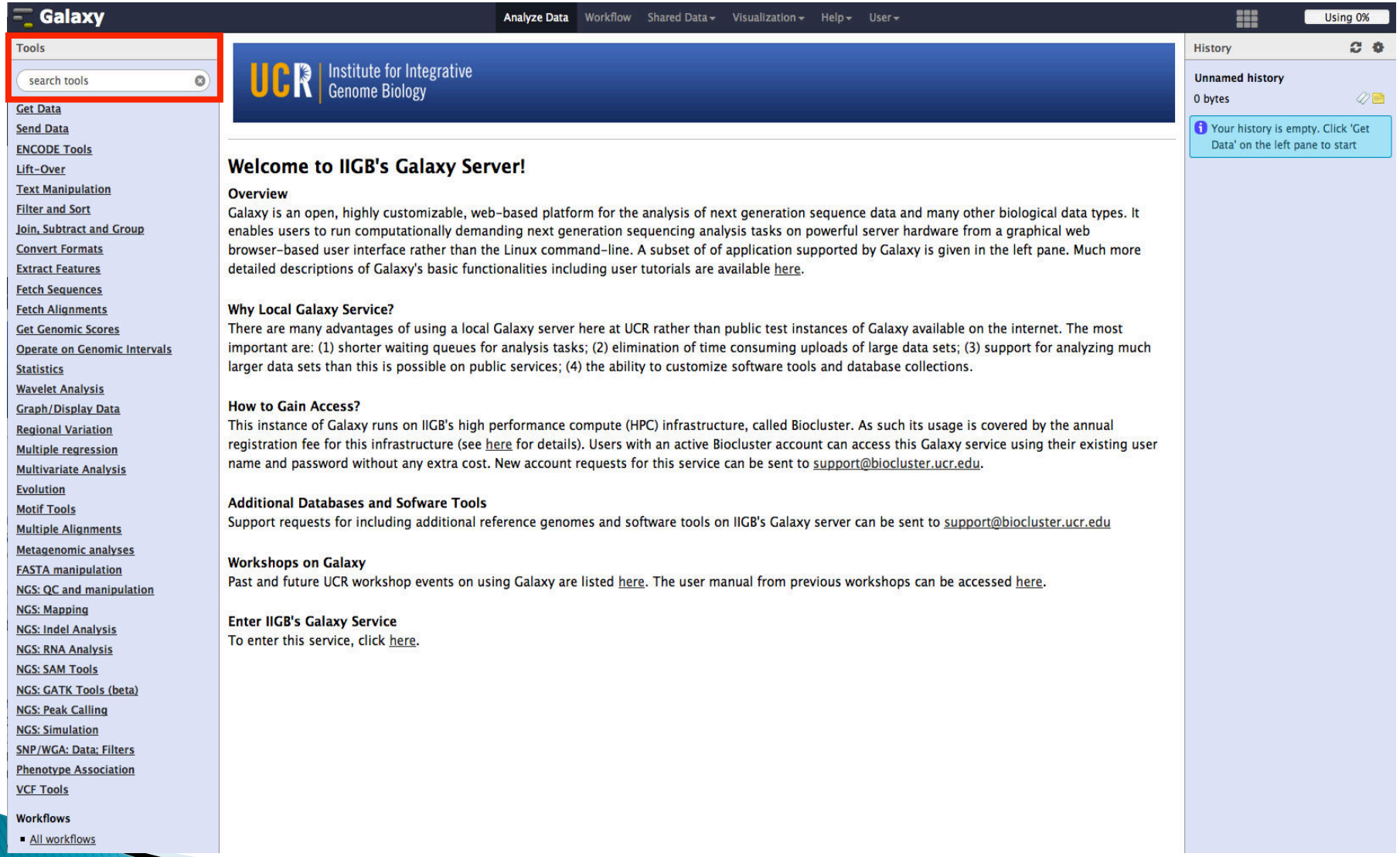
Analyzing Data: Next Generation Sequencing



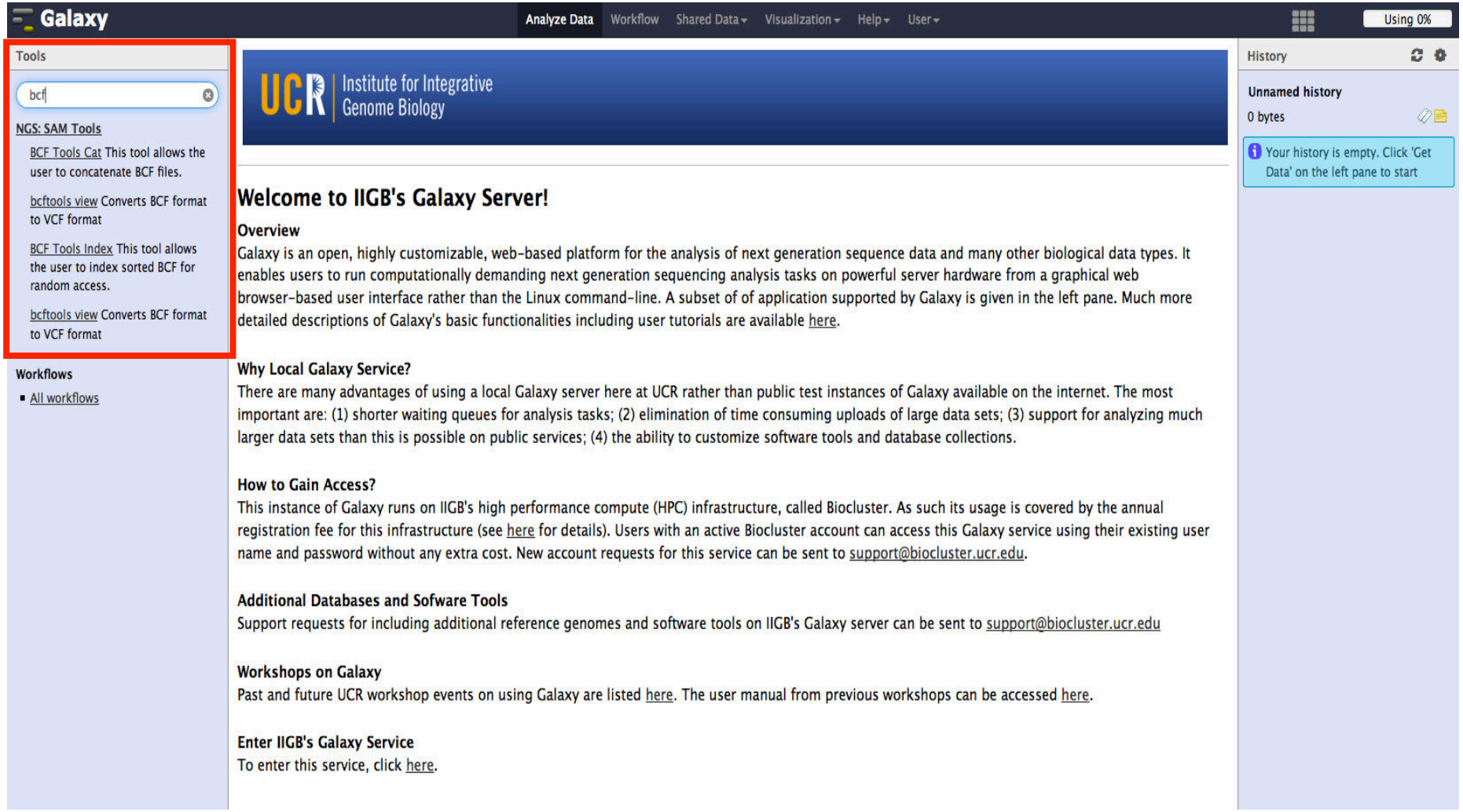
Analyzing Data: Next Generation Sequencing



Analyzing Data: Next Generation Sequencing



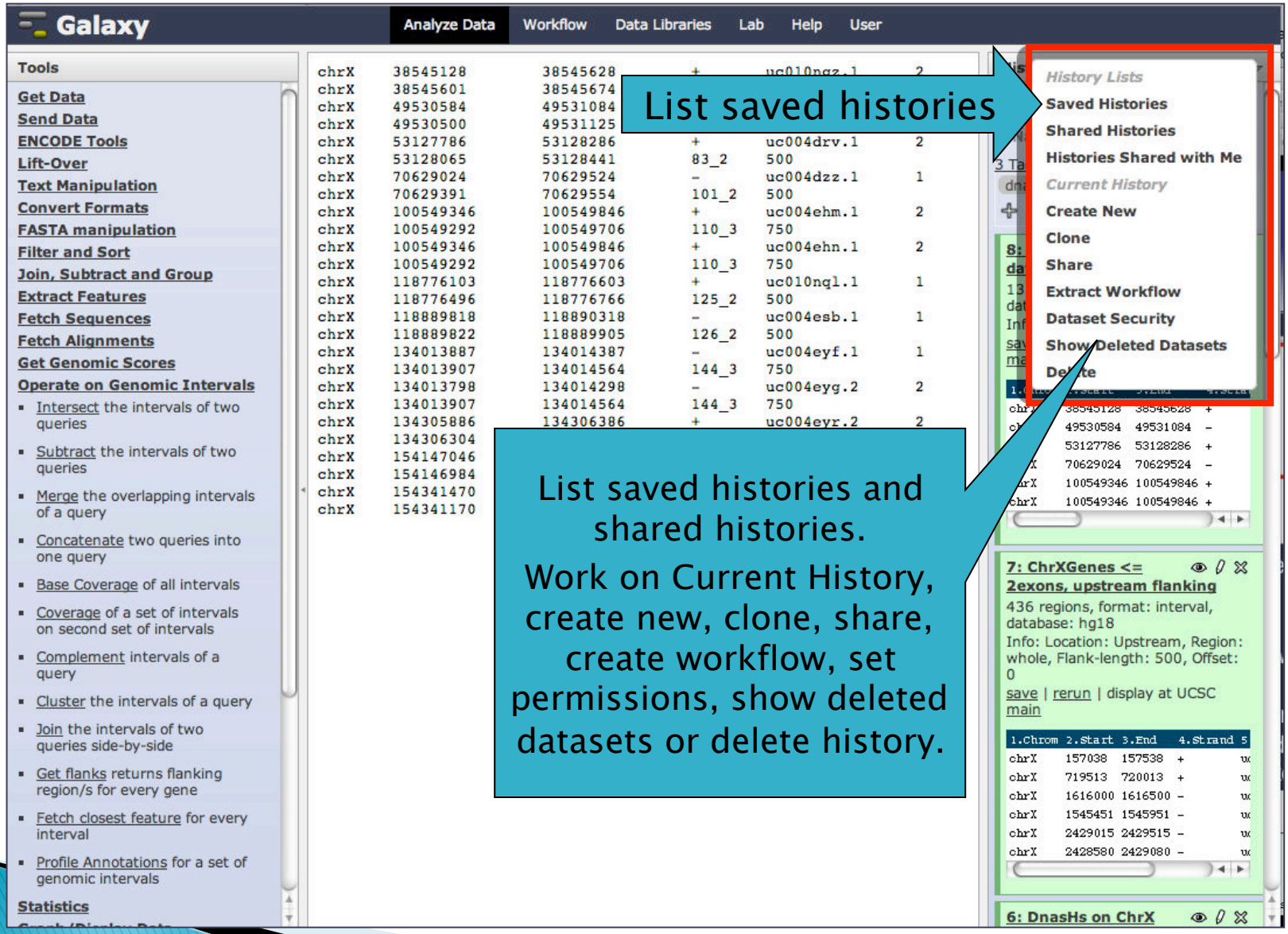
Analyzing Data: Next Generation Sequencing



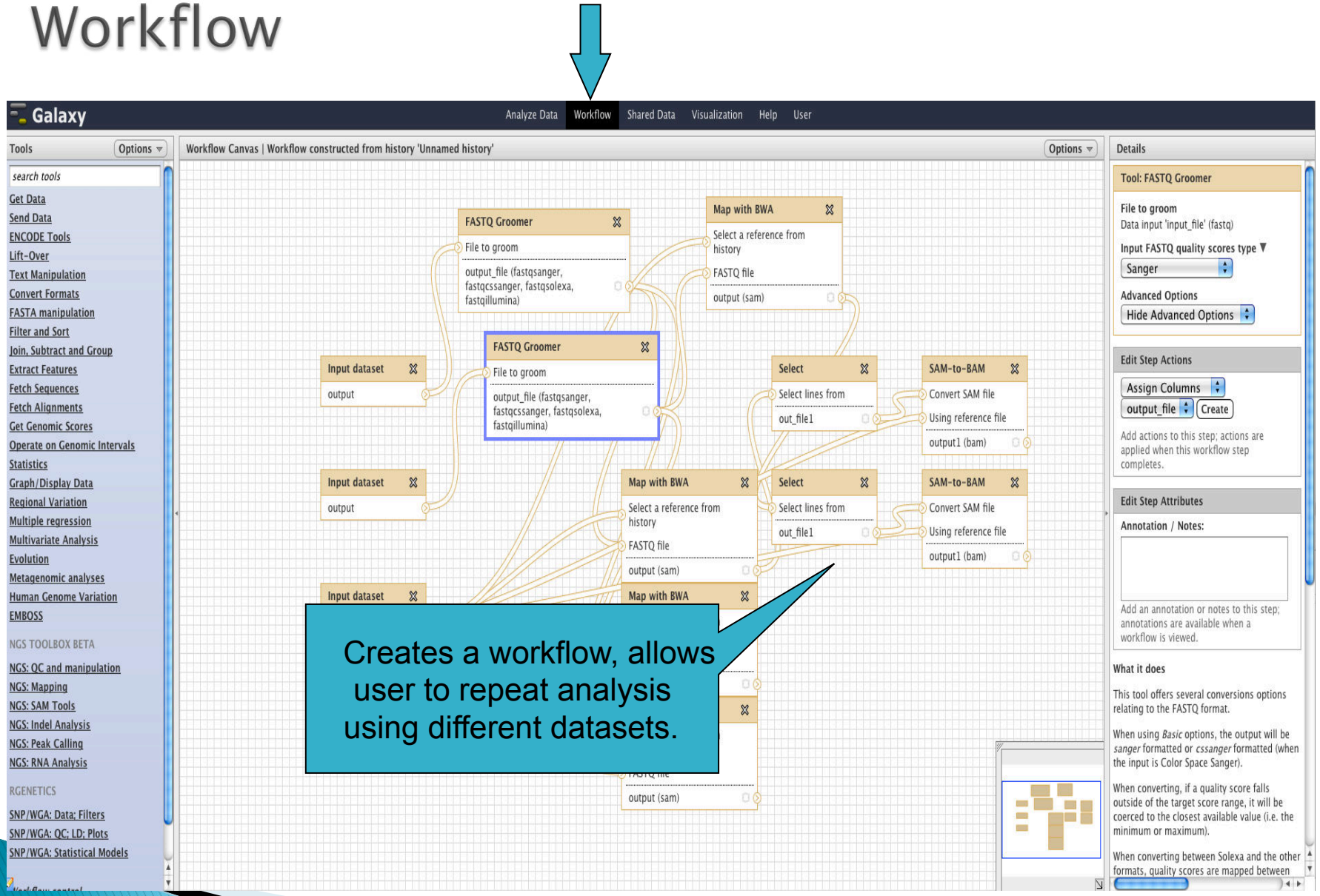
NGS Analysis Using Galaxy

- 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


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
Options
search tools
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
Metagenomic analyses
Human Genome Variation
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
RGENETICS
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
Workflow Canvas | Workflow constructed from history 'Unnamed history'
Options
FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
Map with BWA
Select a reference from history
FASTQ file
output (sam)
Input dataset
output
Select
Select lines from
out_file1
SAM-to-BAM
Convert SAM file
Using reference file
output1 (bam)
Creates a workflow, allows user to repeat analysis using different datasets.
Details
Tool: FASTQ Groomer
File to groom
Data input 'input_file' (fastq)
Input FASTQ quality scores type
Sanger
Advanced Options
Hide Advanced Options
Edit Step Actions
Assign Columns
output_file
Create
Add actions to this step; actions are applied when this workflow step completes.
Edit Step Attributes
Annotation / Notes:
Add an annotation or notes to this step; annotations are available when a workflow is viewed.
What it does
This tool offers several conversions options relating to the FASTQ format.
When using Basic options, the output will be sanger formatted or cssanger formatted (when the input is Color Space Sanger).
When converting, if a quality score falls outside of the target score range, it will be coerced to the closest available value (i.e. the minimum or maximum).
When converting between Solexa and the other formats, quality scores are mapped between

What's next?

- Galaxy exercises
  - SNP-Seq
  - RNA-Seq

- Visualization
  - IGV (Integrative Genomics Viewer)
  - http://www.broadinstitute.org/igv/