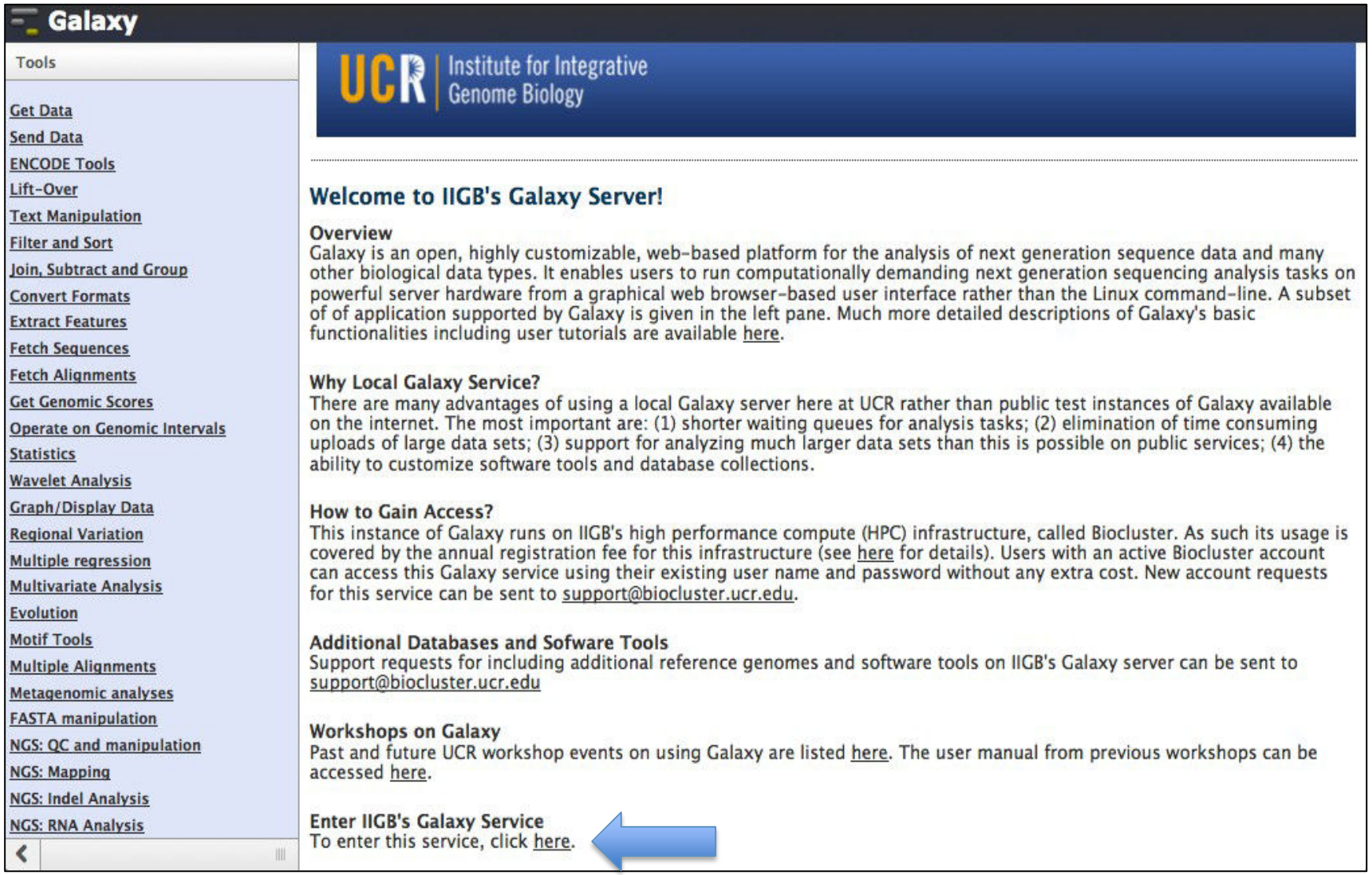
NGS Analysis Using Galaxy

- 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**

UCR Galaxy homepage (https://galaxy.bioinfo.ucr.edu)



SNP-seq Analysis dataset

- Data source: SRR038850 sample from experiment published by Kaufman *et al* (2012, GSE20176)
    - Understand the target molecular mechanisms underlying AP1 function

- Fastq File:

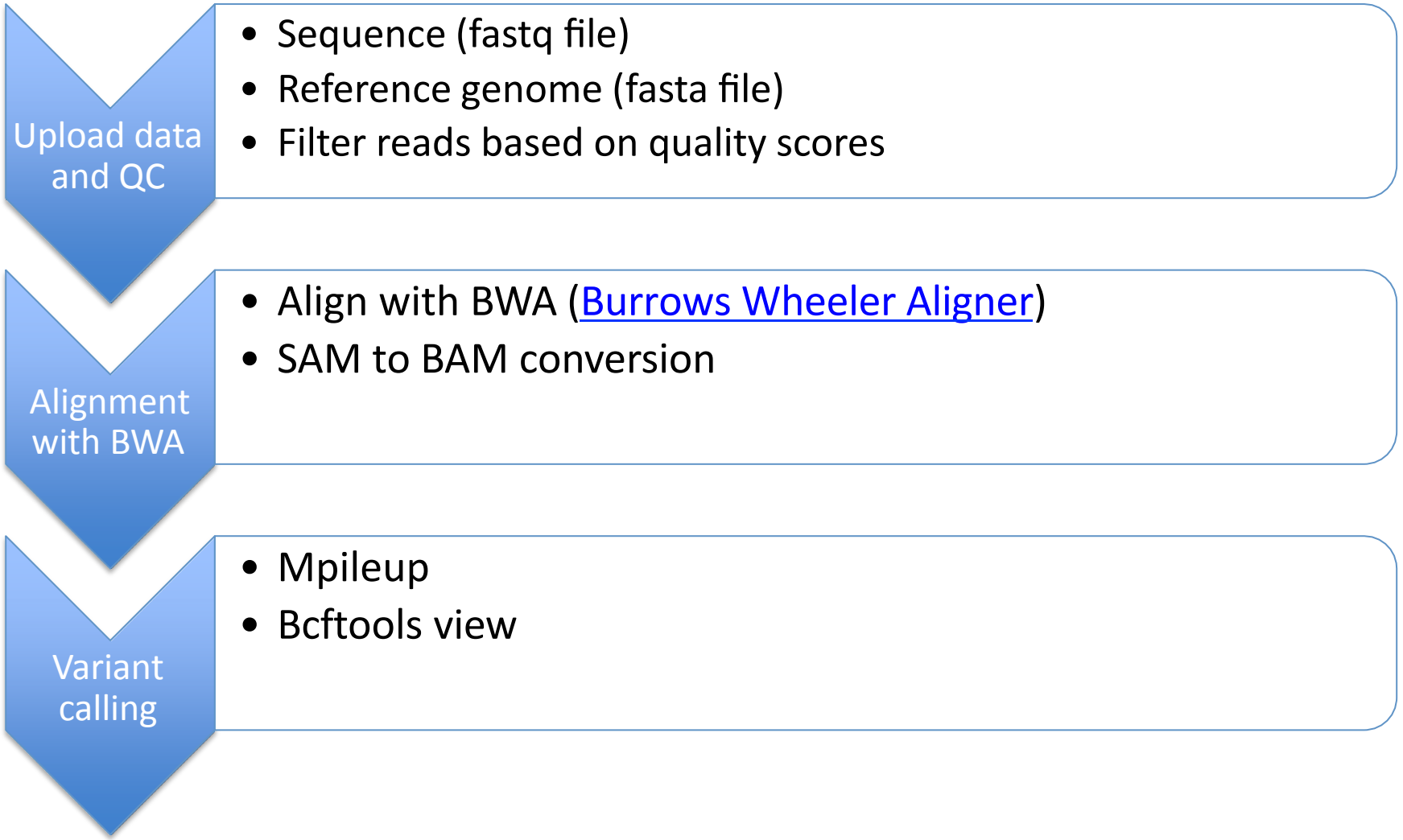
  http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Snpseq/SRR038850.fastq

- TAIR10 Genome:

  http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Snpseq/tair10chr.fasta

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20176

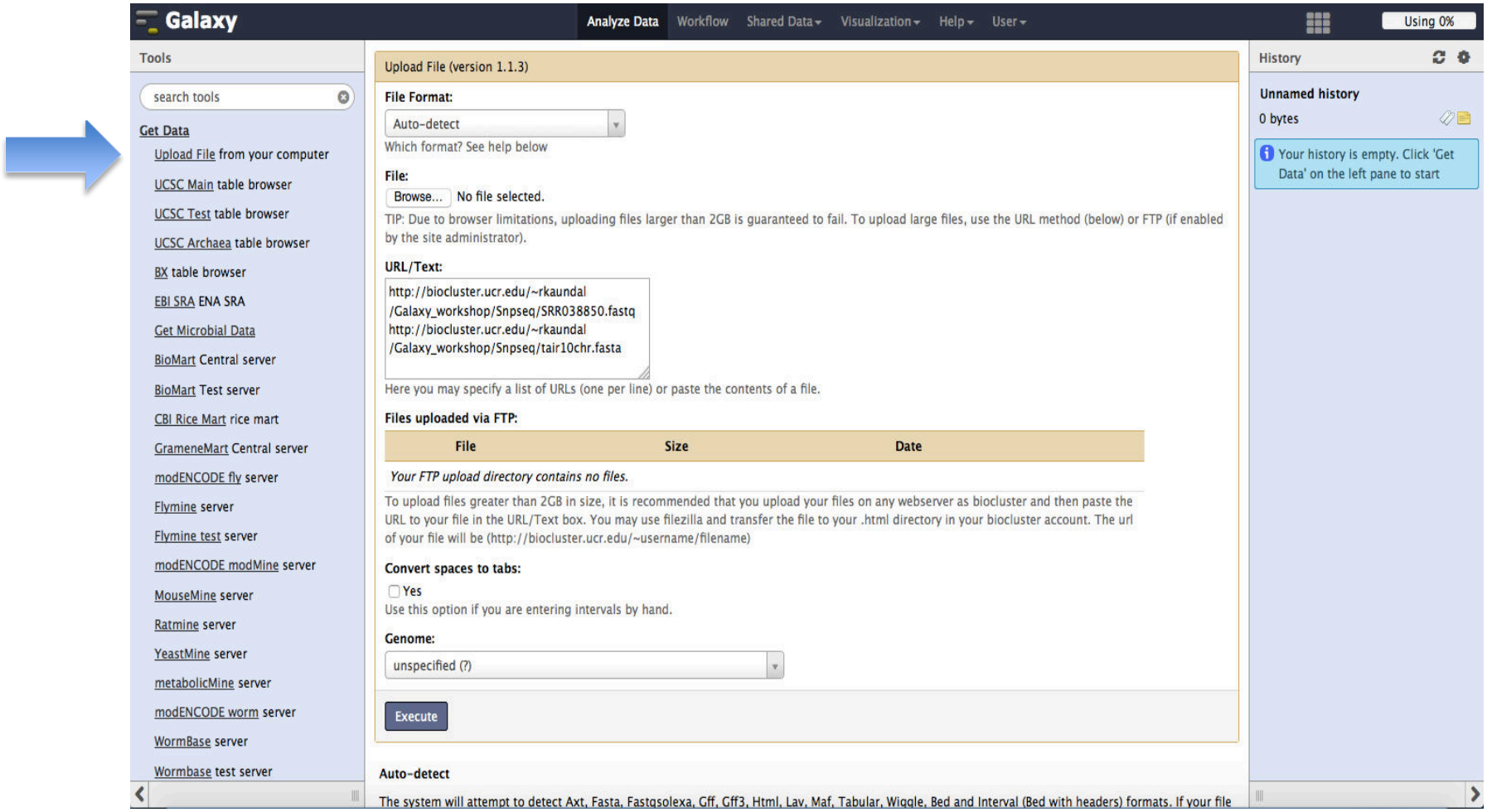
SNP-seq pipeline



Upload data

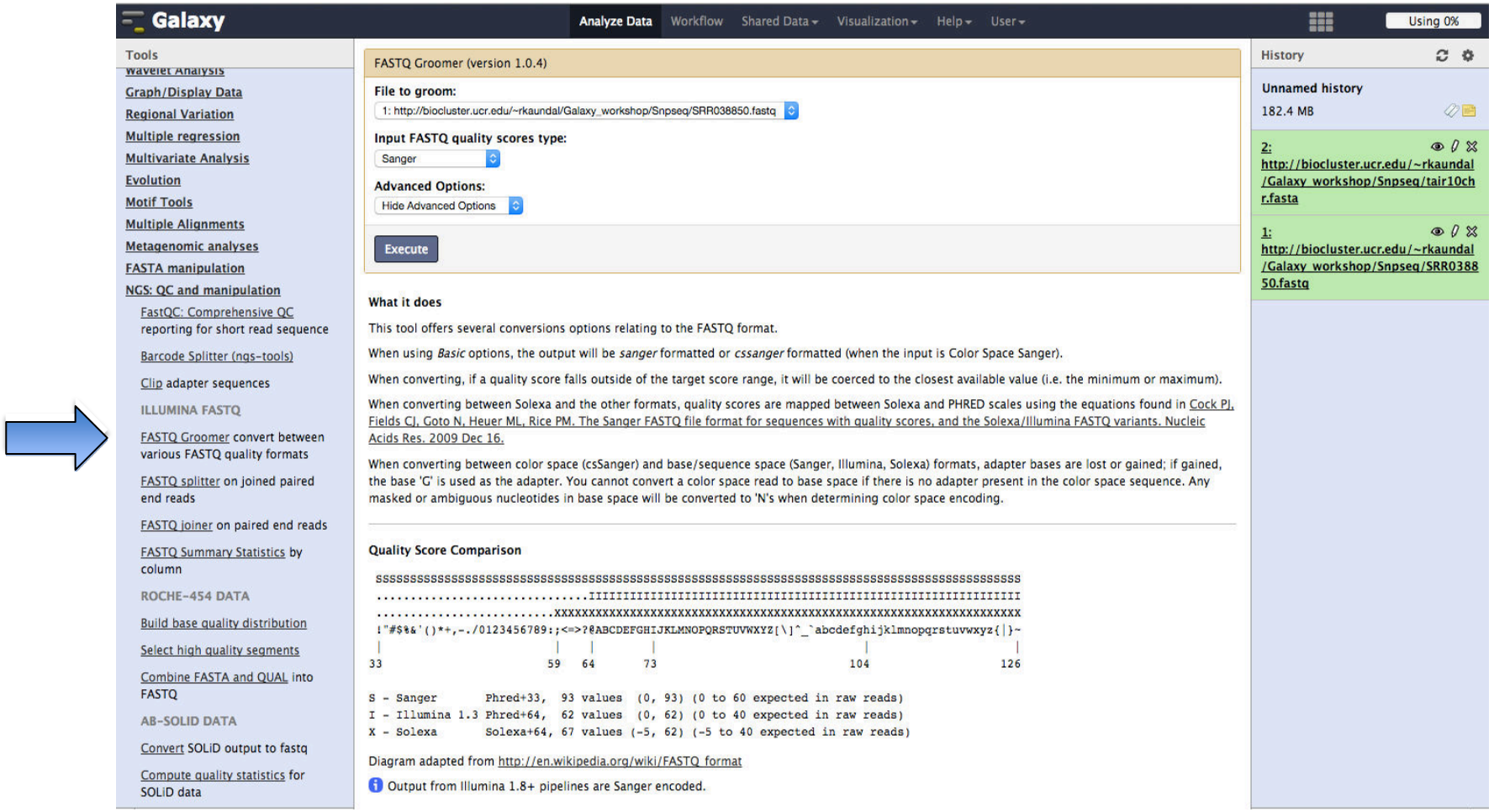
Go to "Get Data", click open "Upload File from your computer". Then specify the following list of URLs in URL/Text box

http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Snpseq/SRR038850.fastq
http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Snpseq/tair10chr.fasta



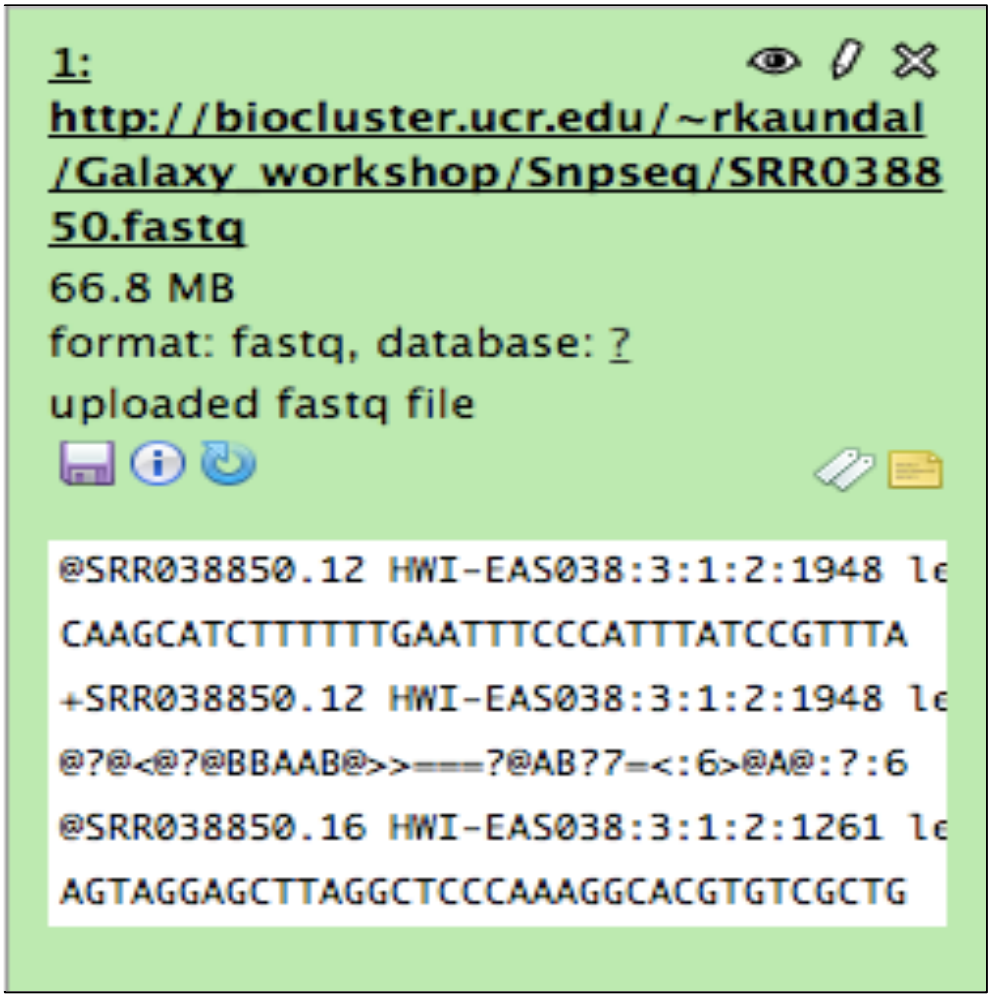
Convert Fastq file to sanger format

- Select NGS: QC and manipulation and Fastq groomer
- The FASTQ Groomer tool is used to verify and convert between the known FASTQ variants.
- After grooming, the user is presented with a valid FASTQ format that is accepted by all downstream analysis tools.



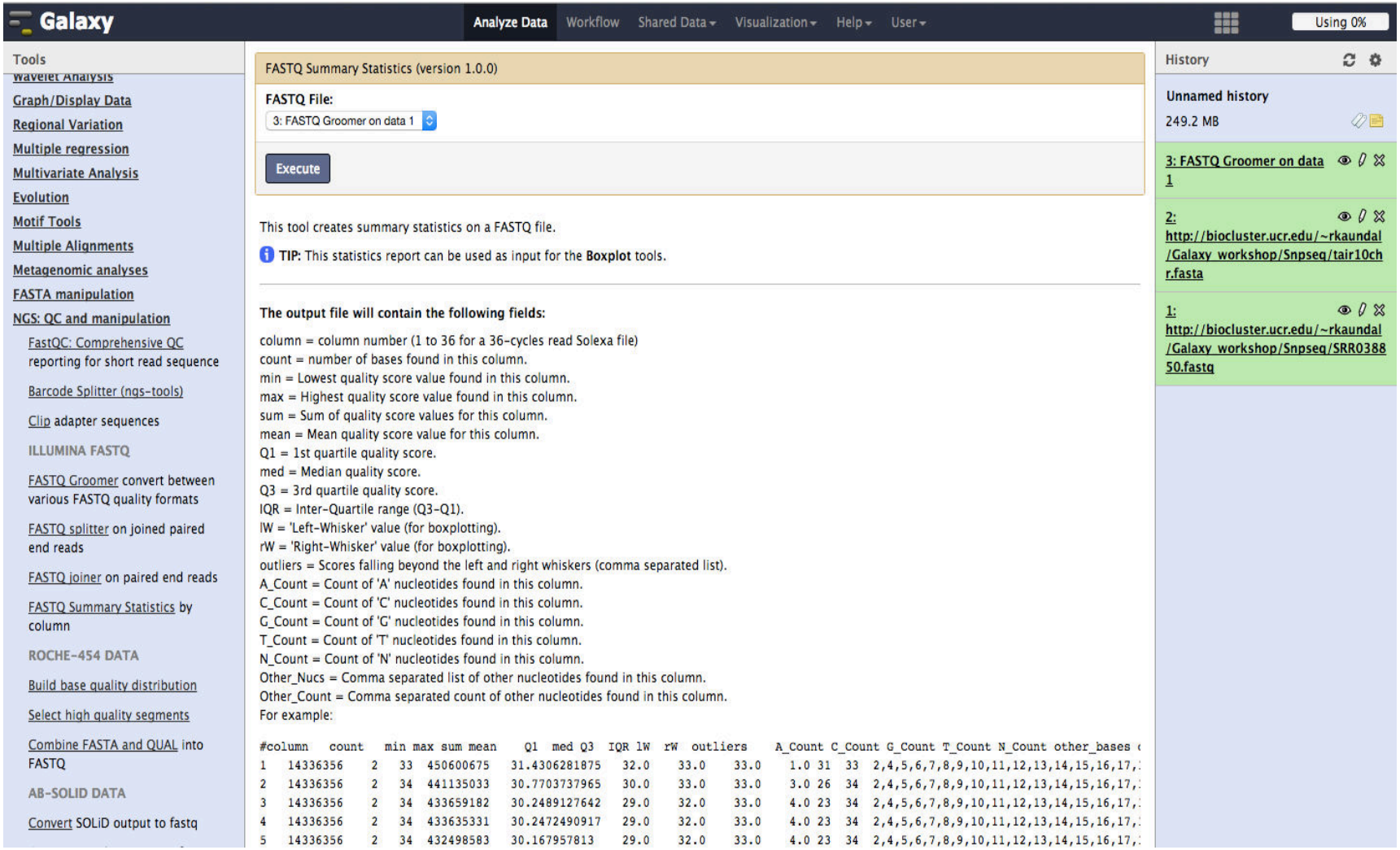
Features available in history panel

- View, Edit, Delete file
- Size of the file
- Save the file
- Repeat the analysis



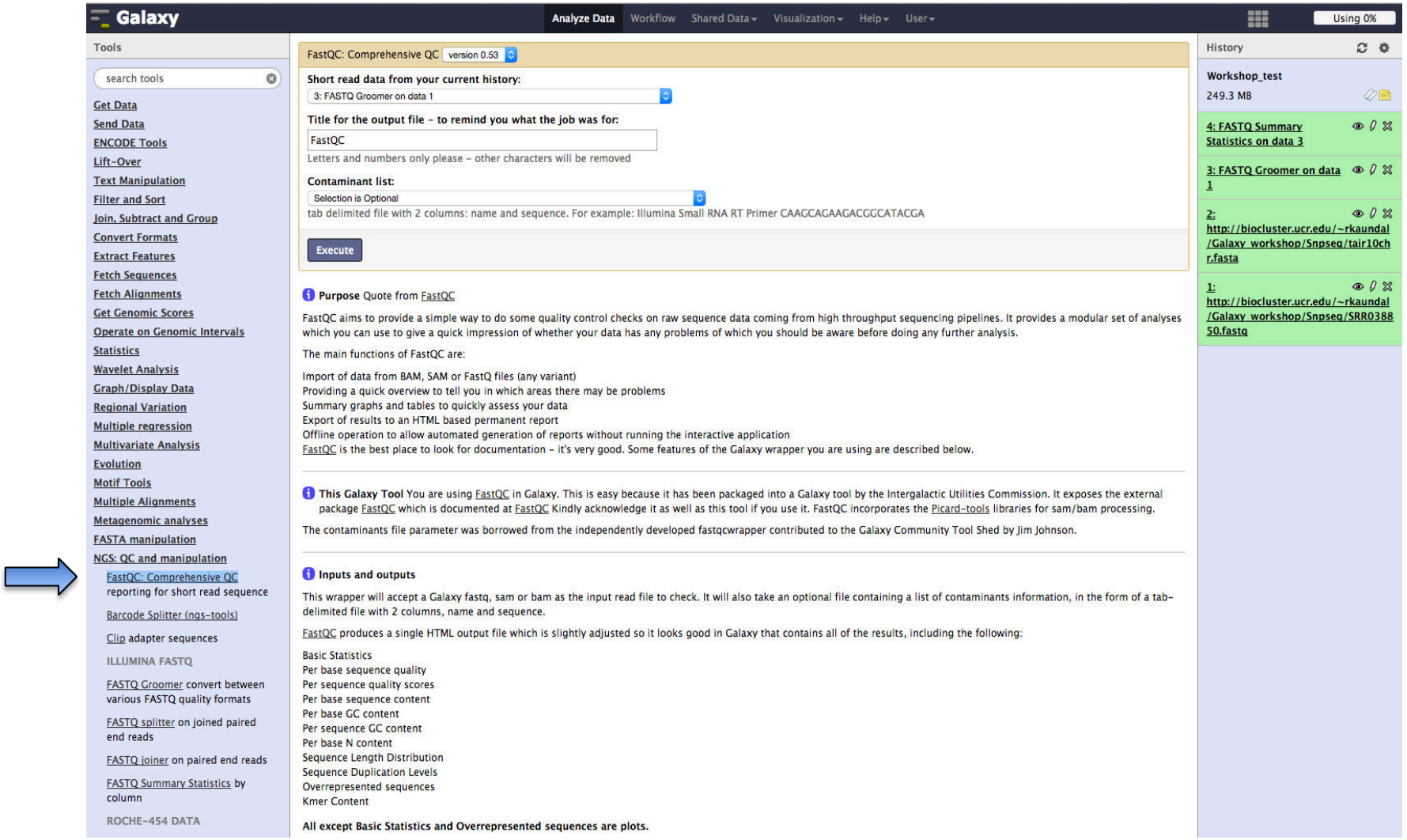
FASTQ Summary Statistics

- To understand the quality properties of the reads, one can run the FASTQ Summary Statistics tool from NGS: QC and manipulation.

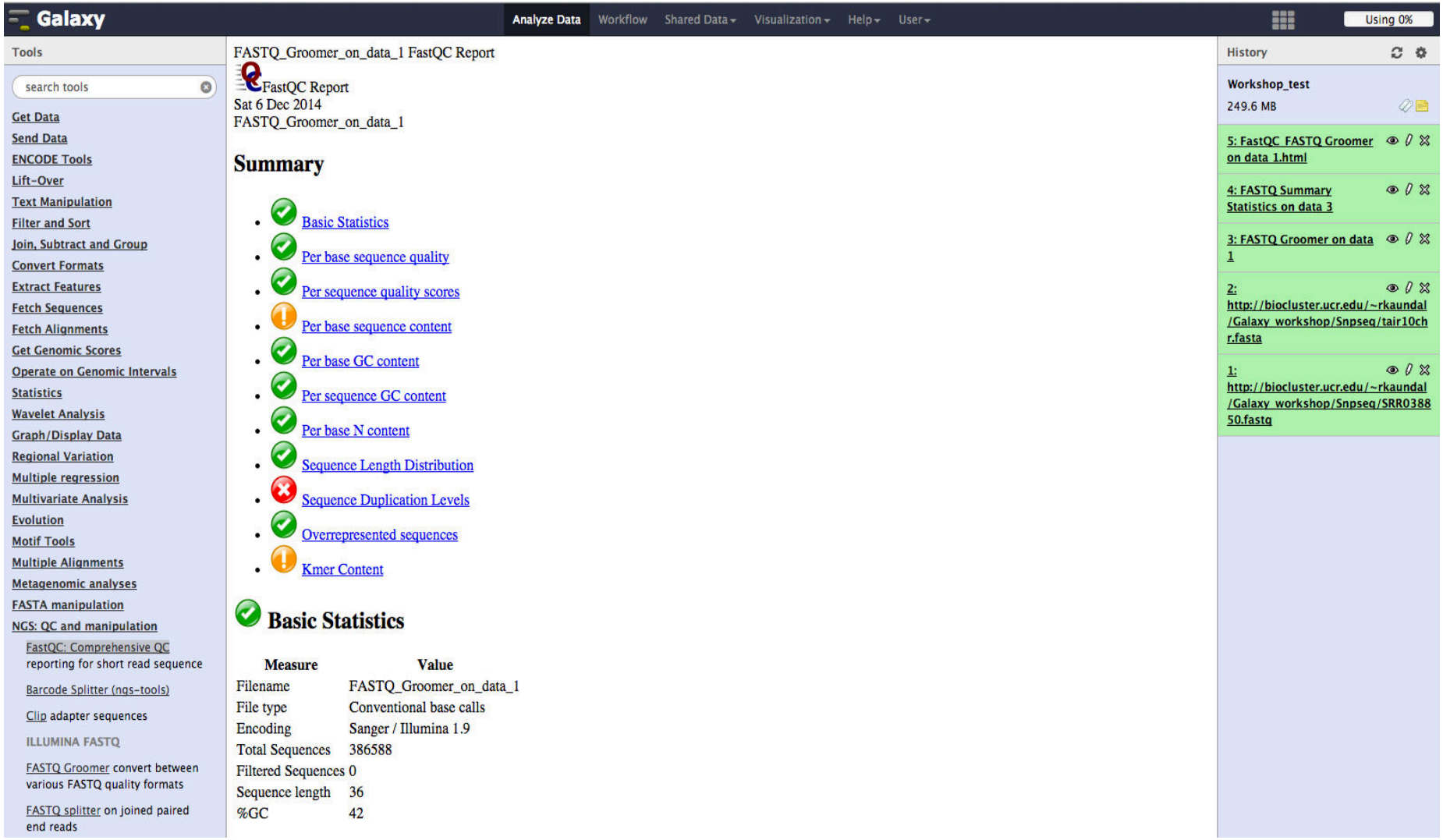


FASTQ Quality control

- To understand the quality properties of the reads, one can also run the FASTQC: Read QC reports from NGS: QC and manipulation.

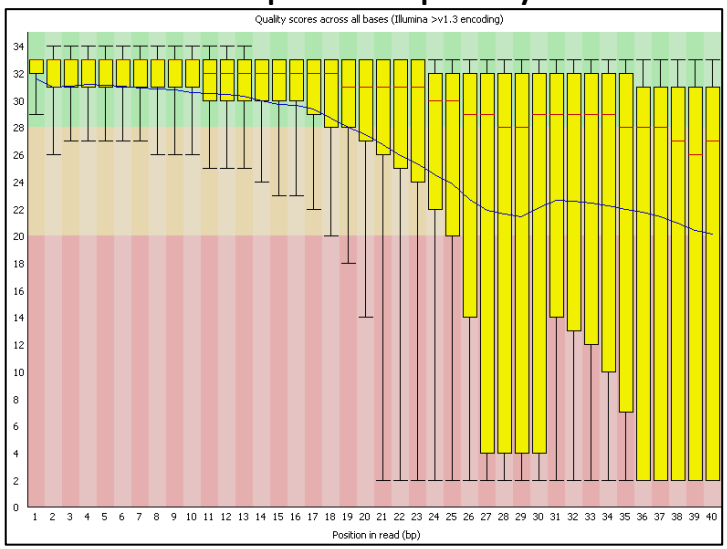


Quality control output

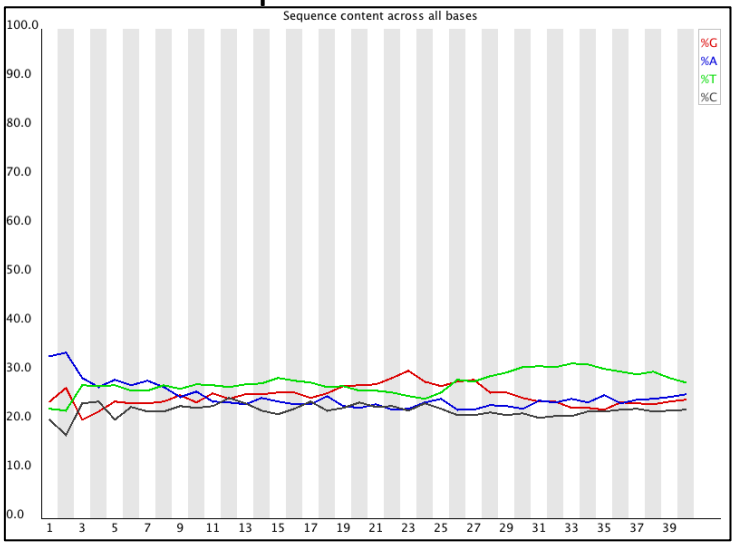


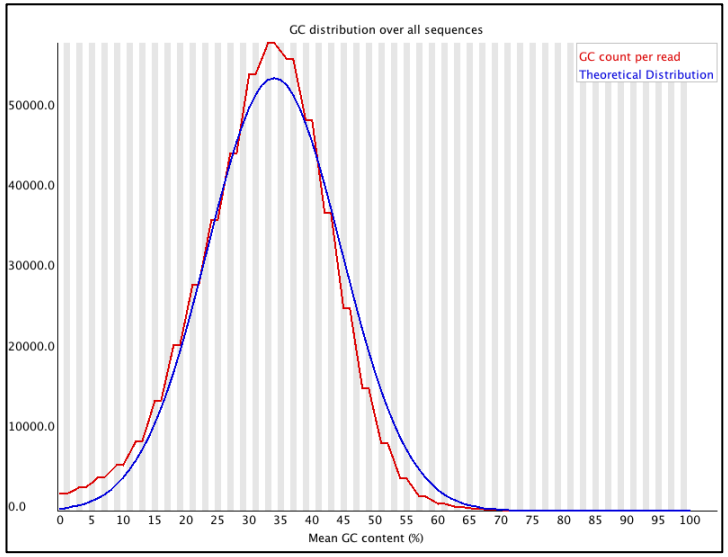
Quality control reports

Per base sequence quality



Per base sequence content

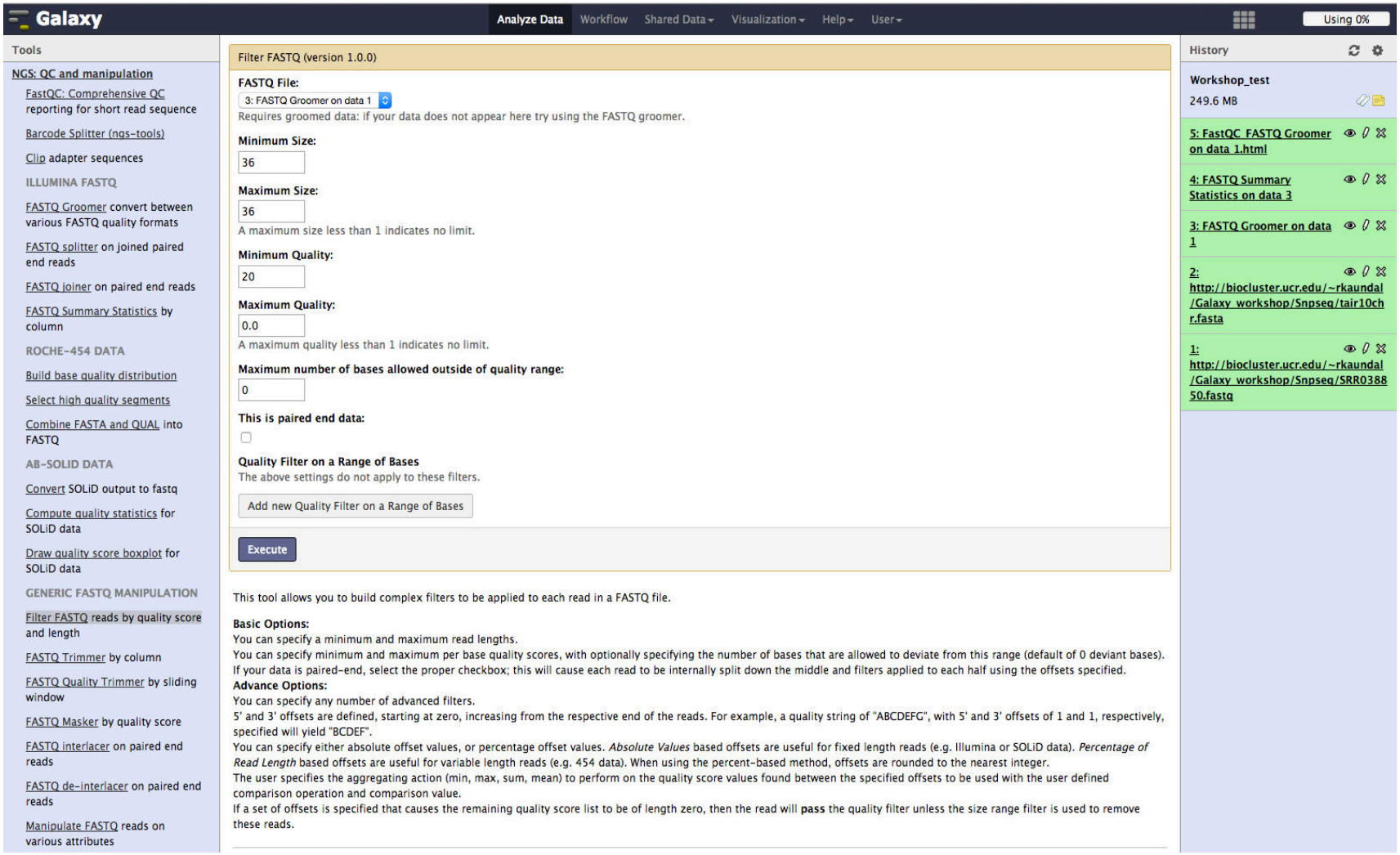




Per sequence GC content

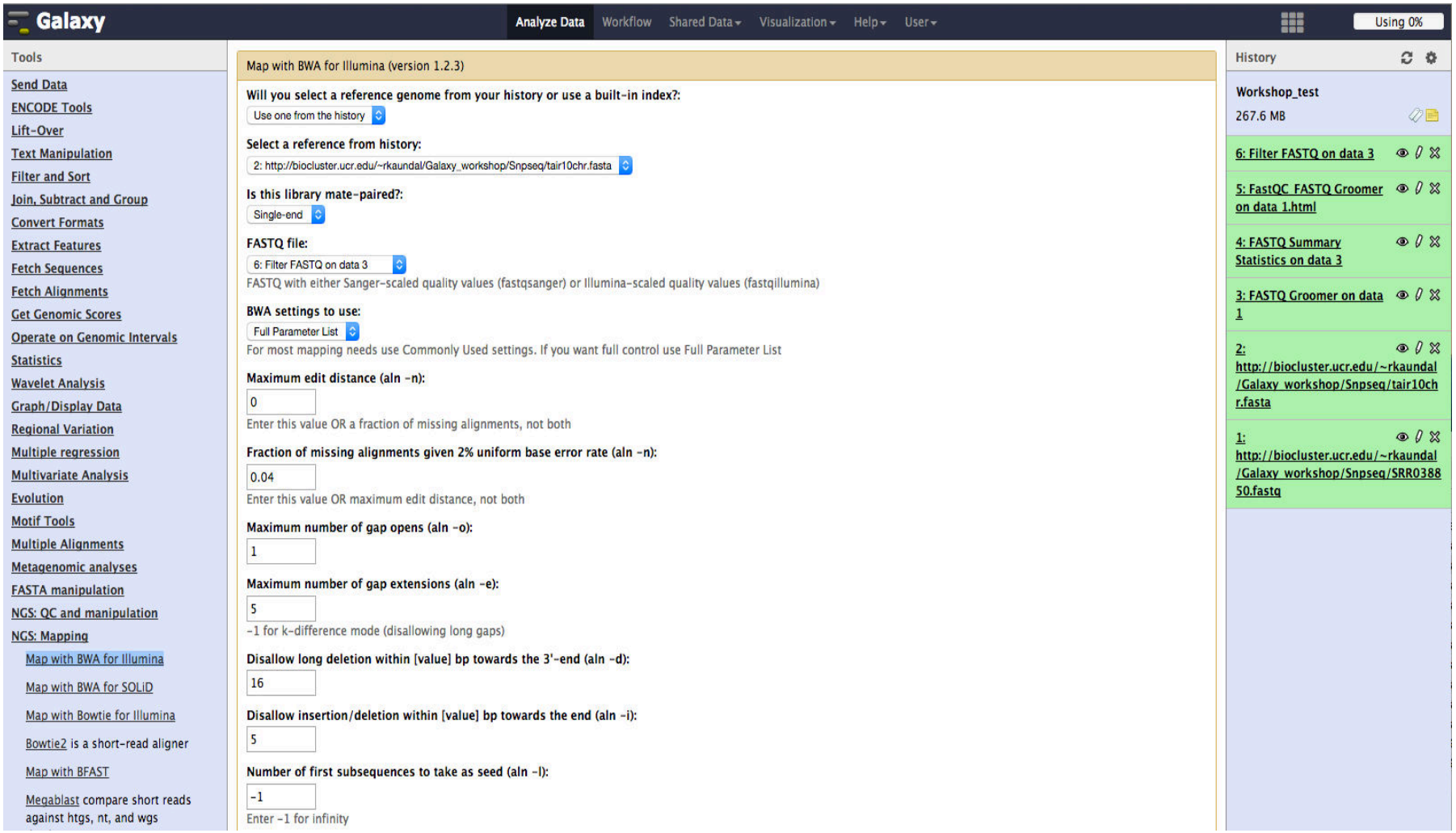
Quality filter

- This tool filters reads based on quality scores. NGS: QC and manipulation -> Generic FASTQ manipulation->Filter FASTQ reads by quality score and length



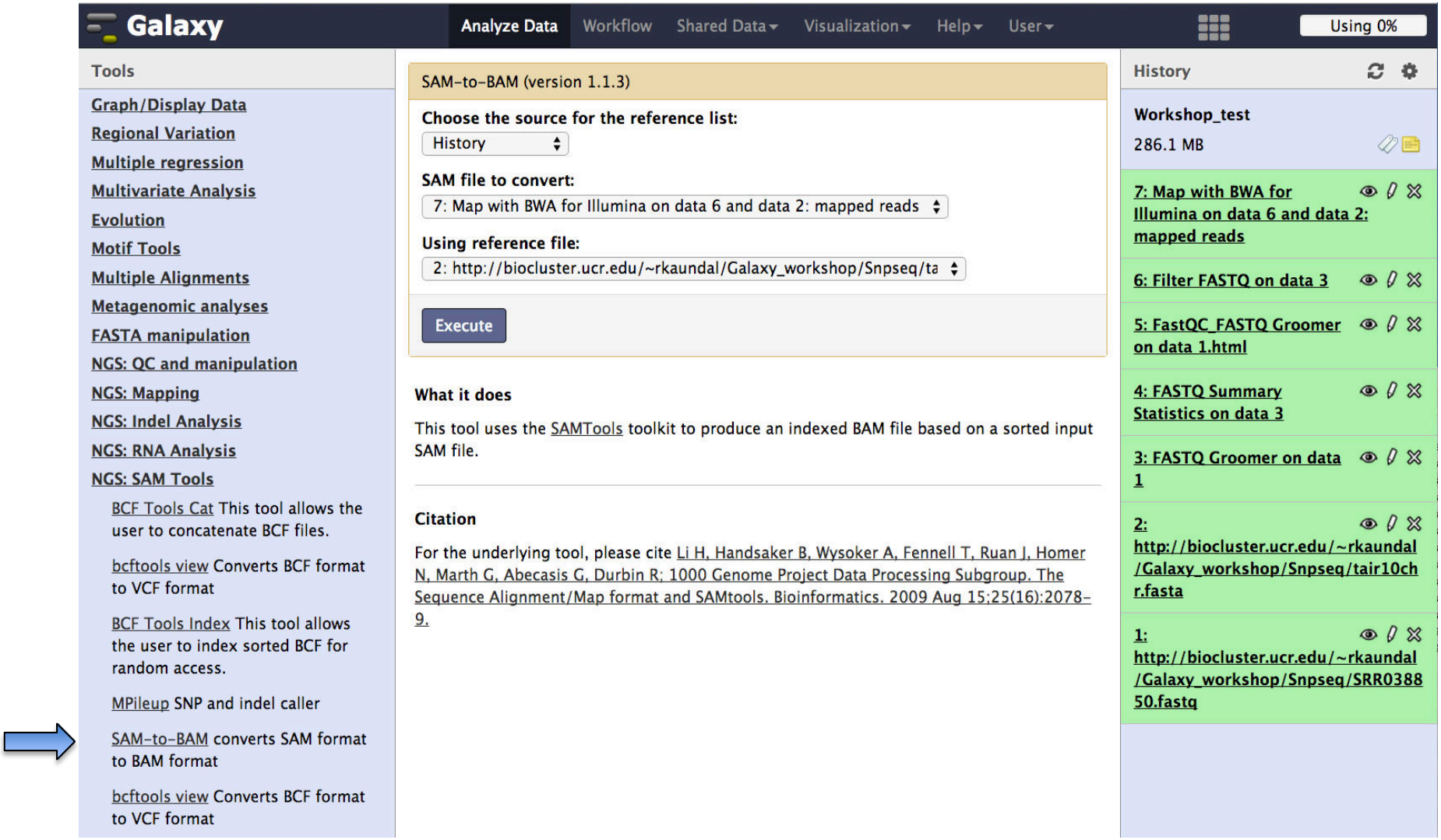
Alignment with BWA

- BWA is a fast and accurate short read aligner that allows mismatches and indels (Link)
- Go to "NGS: Mapping" and click on "Map with BWA". (Settings)



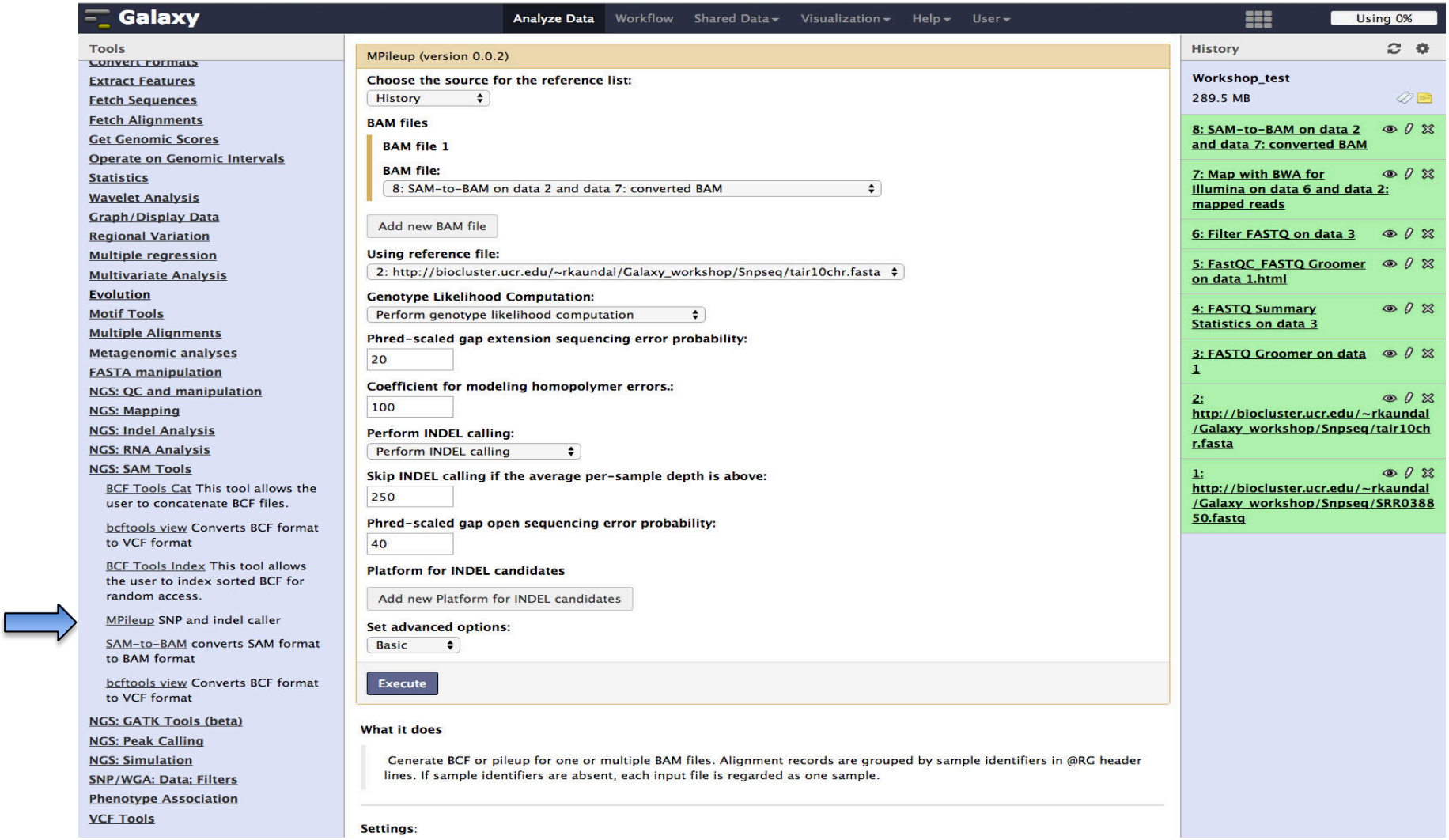
SAM to BAM format conversion

- Produce an indexed BAM file based on a sorted input SAM file.
- Go to "NGS: SAM Tools", then click open "SAM-to-BAM".



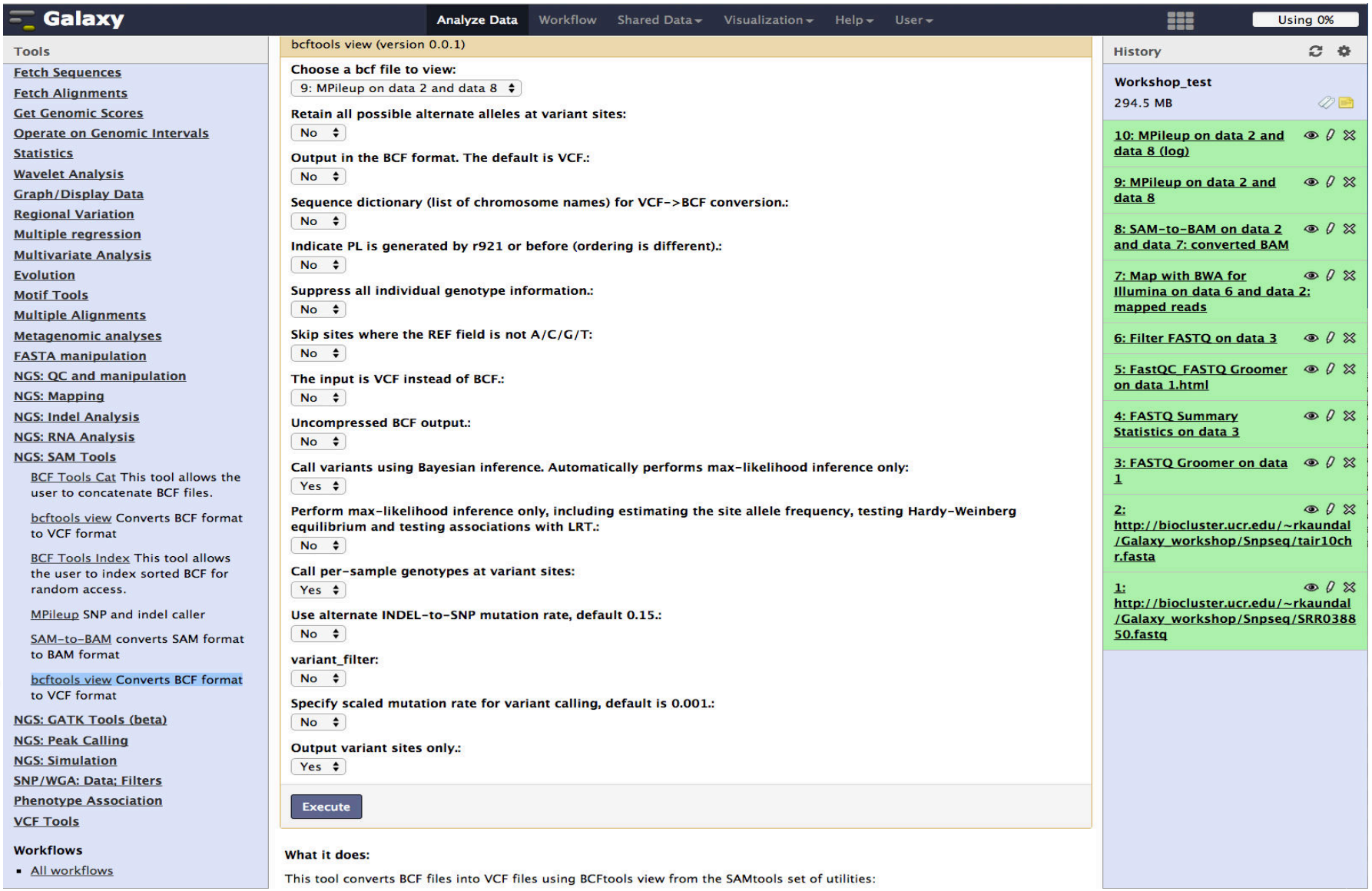
Variants calling with Mpileup

- SNP and INDEL caller to Generate BCF (Binary Variant Format) for one or multiple BAM files
- Go to "NGS: SAM Tools", then click open "MPileup SNP and indel caller".

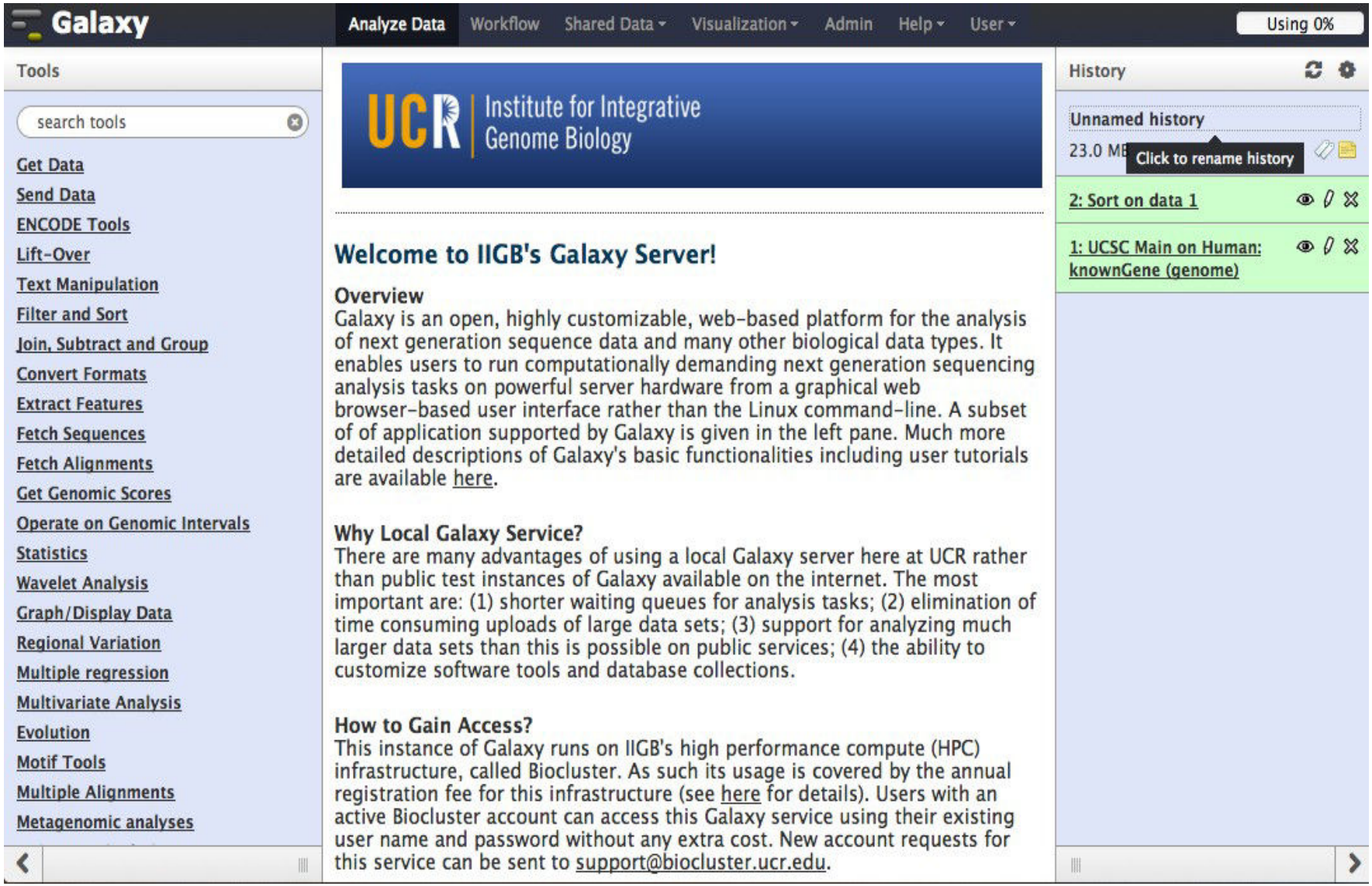


Bcftools view

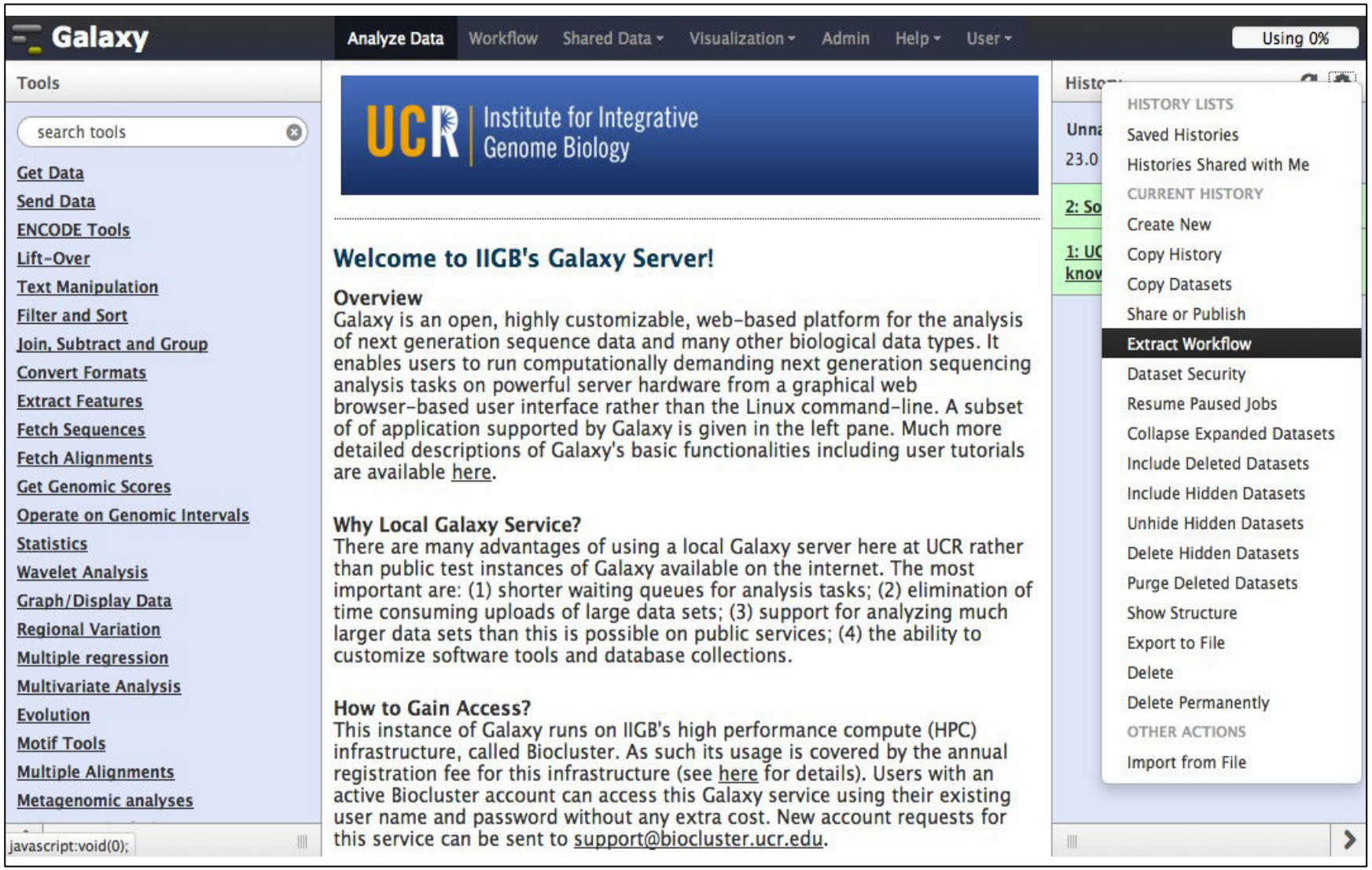
- Converts BCF format to VCF format.
- Go to "NGS: SAM Tools", then click open "bcftools view"



Rename history



Extract workflow from history

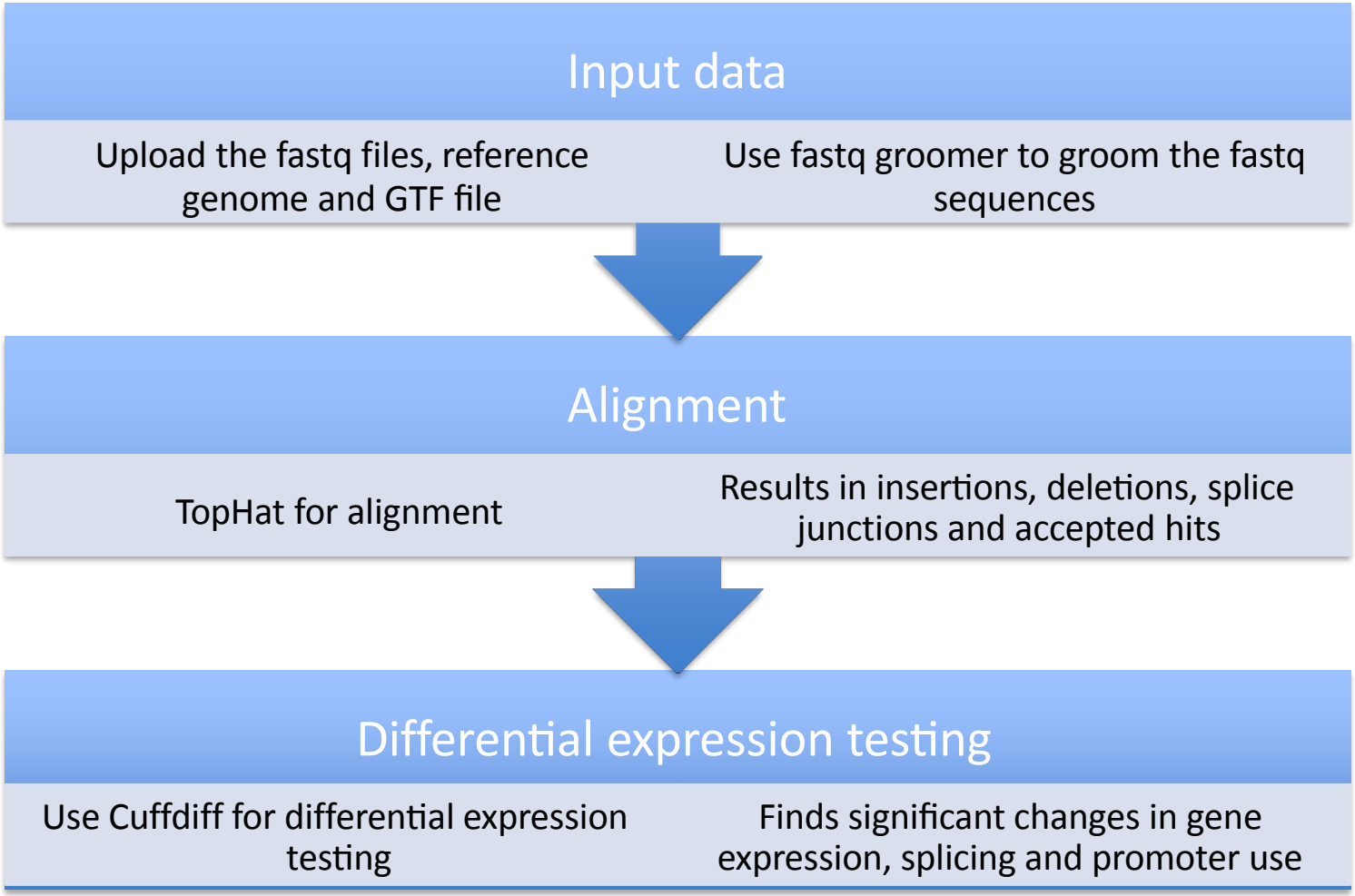


Exercise 2: RNA-seq Analysis

- Data source: RNA-seq experiment SRA023501
- Four Samples:

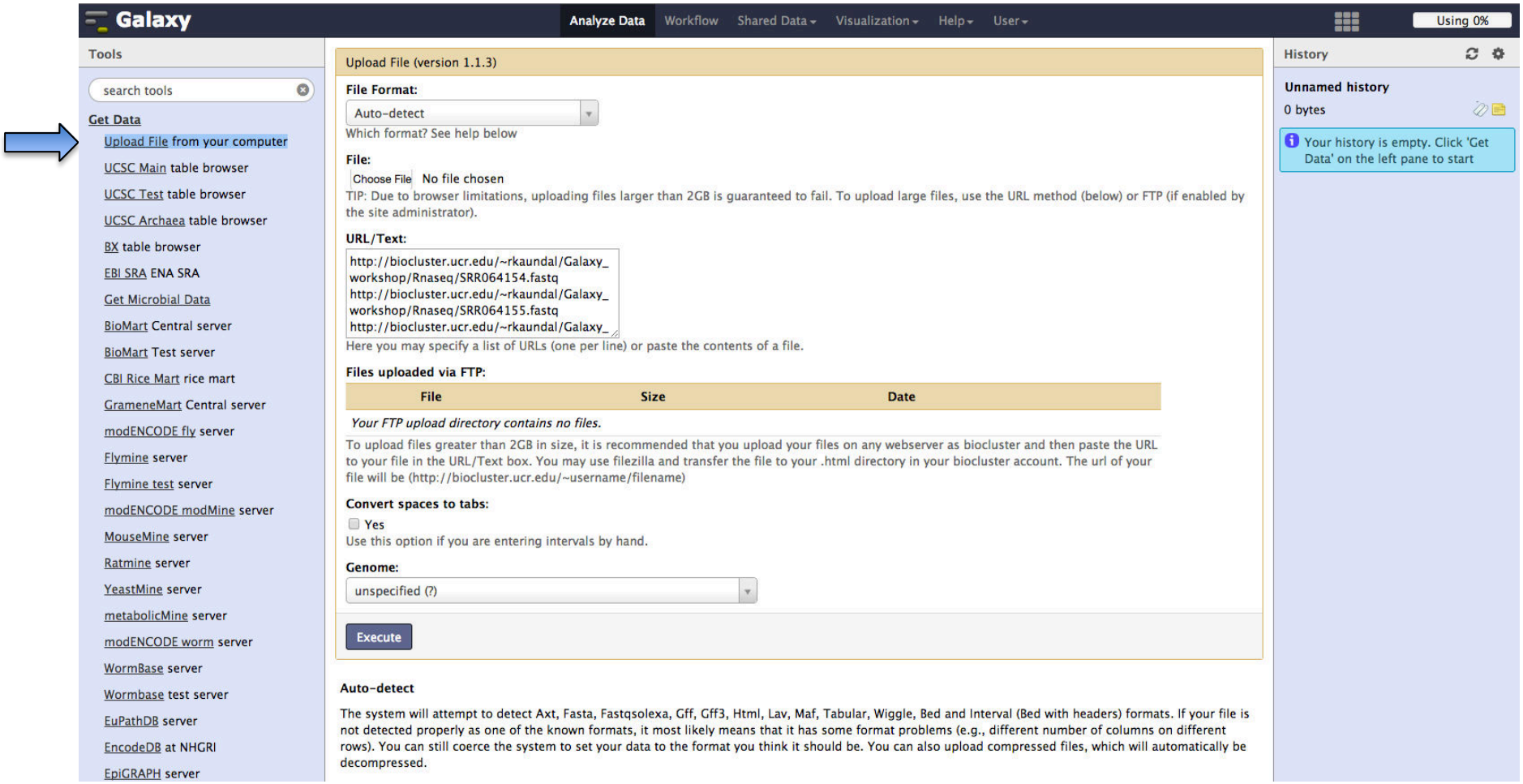
Samples	Factors	Fastq
AP3_f14	AP3	http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/SRR064154.fastq
AP3_f14	AP3	http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/SRR064155.fastq
T1_f14	TRL	http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/SRR064166.fastq
T1_f14	TRL	http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/SRR064167.fastq

RNA-seq Analysis workflow



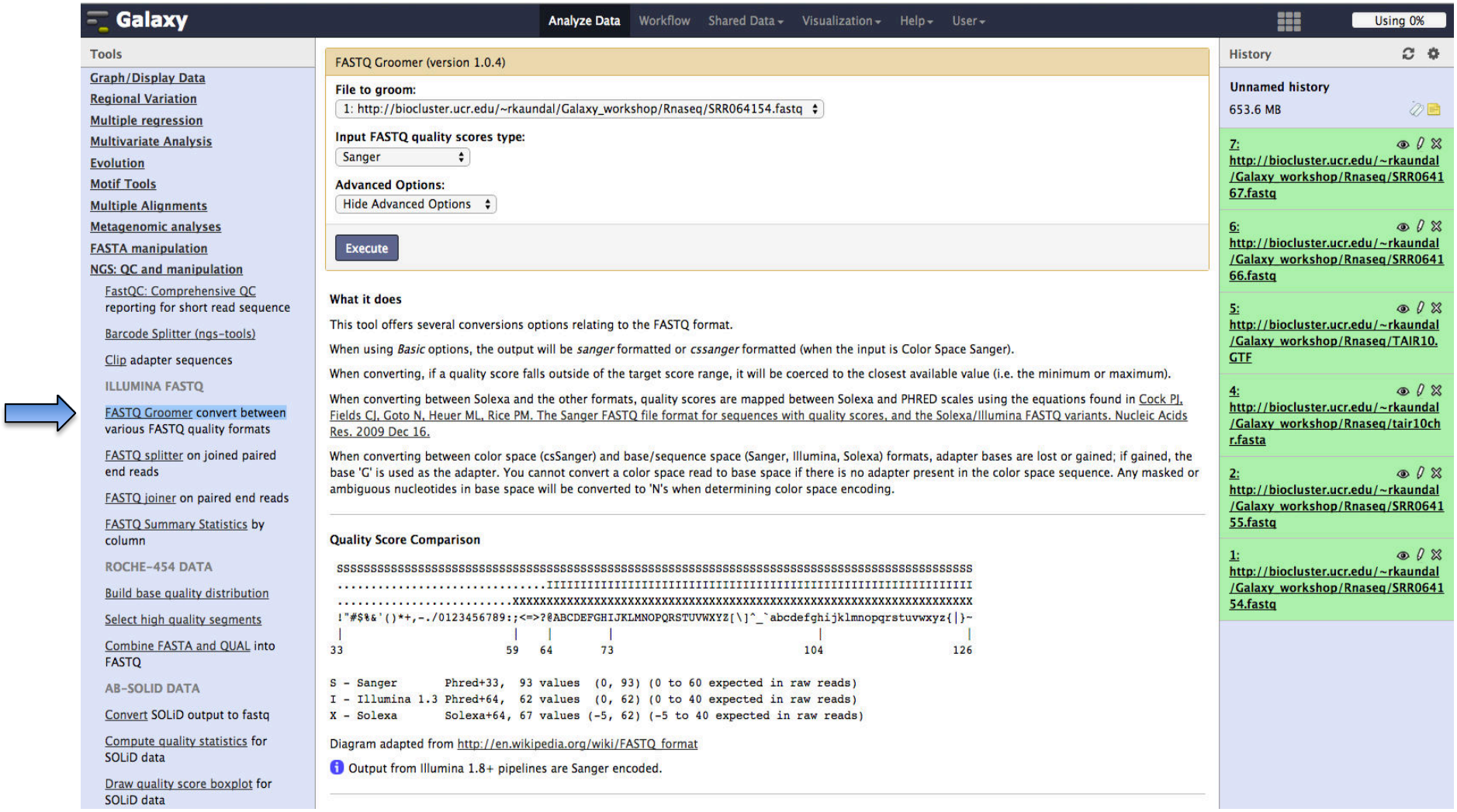
Upload Data

- Upload four fastq files with URL
- Upload tair10chr.fa with URL (http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/tair10chr.fasta)
- Upload TAIR10.GTF with URL, specify the format "gtf" (http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/TAIR10.GTF)

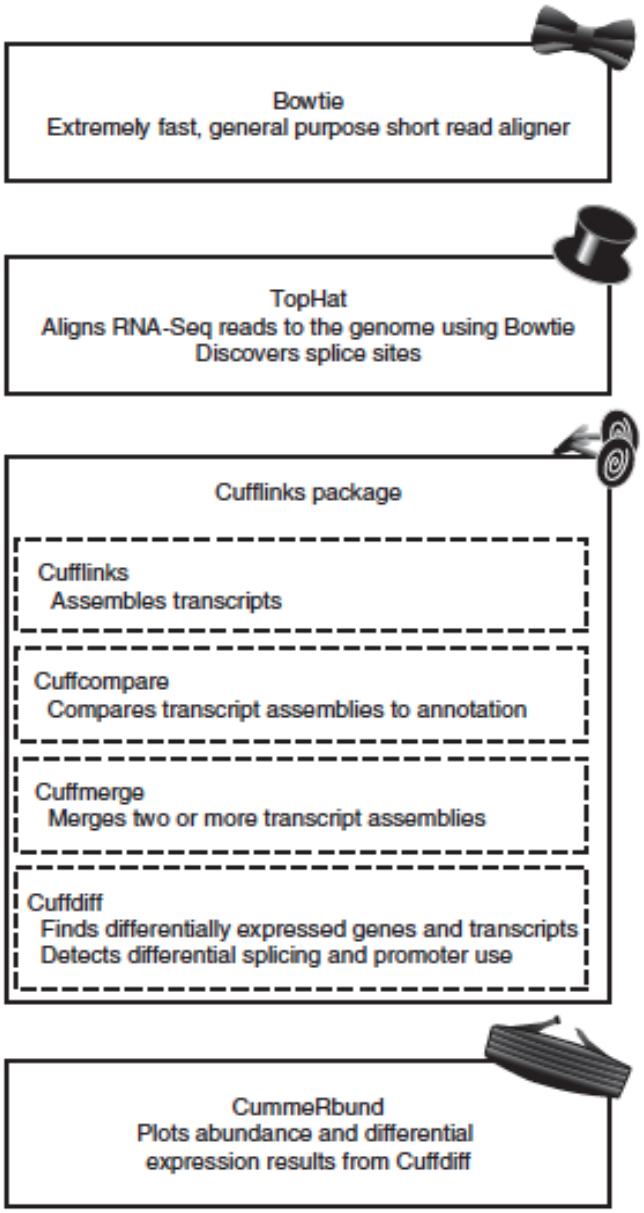


Fastq Groomer

- Select NGS: QC and manipulation and Fastq groomer
- Run Fastq Groomer for all the 4 fastq sequences.

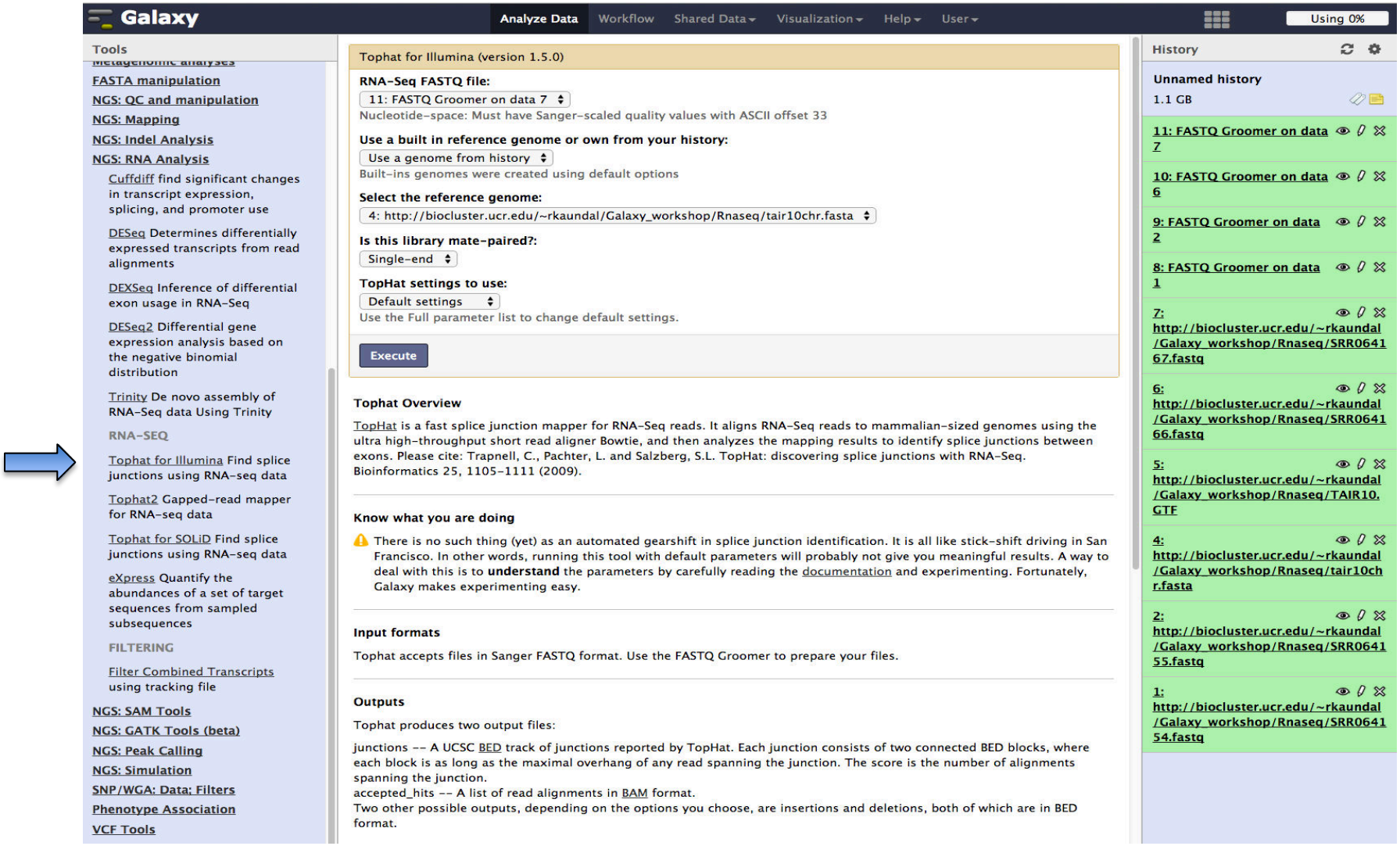


RNA-seq Analysis



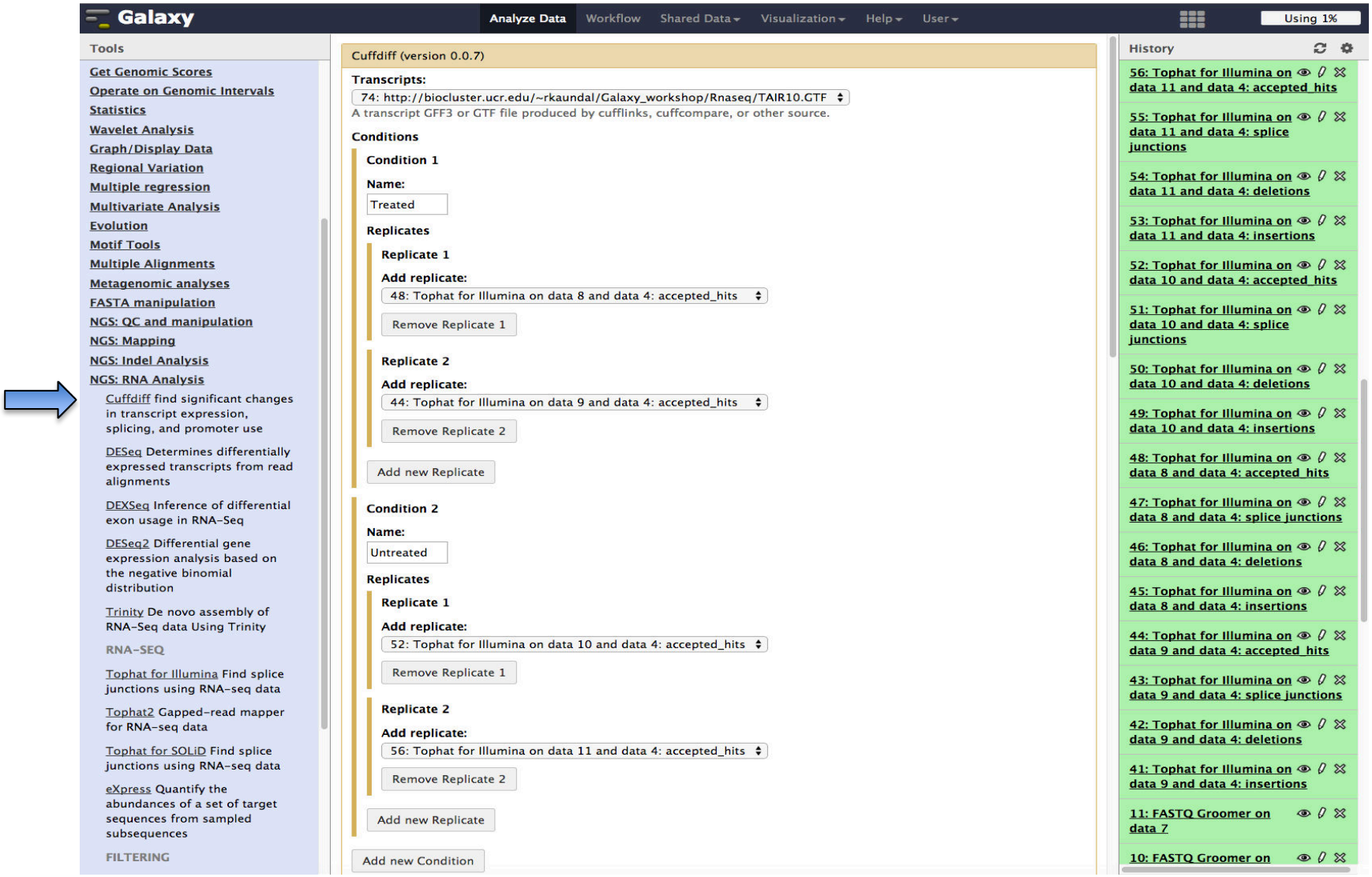
Alignment with TopHat

- TopHat is a fast splice junction mapper for RNA-Seq reads, it can identify splice junctions between exons.
- Go to "NGS RNA analysis", click open "Tophat for illumina"
- Similarly repeat this process for all the 4 fastq groomed sequences



Find Significant Changes

- Cuffdiff find significant changes in transcript expression.
- Go to "NGS RNA analysis", click open "Cuffdiff"



Cuffdiff Output

- TSS... files report on Transcription Start Sites
- splicing... report on splicing
- CDS... track coding region expression
- transcript... track transcripts
- gene... rolls up the transcripts into their genes
  - gene/transcript FPKM tracking: gives information about the gene/ transcript (length, nearest ref id, TSS, etc) and the confidence intervals for FPKM for each condition.
  - gene/transcript differential expression testing: gives the expression change between groups, a status of whether there was enough data for that value to be accurate (OK is good, FAIL and NOTEST are bad. LOWDATA is somewhere in between). Finally, it gives a p-value.
  - see more details… Link

Find Significant Changes

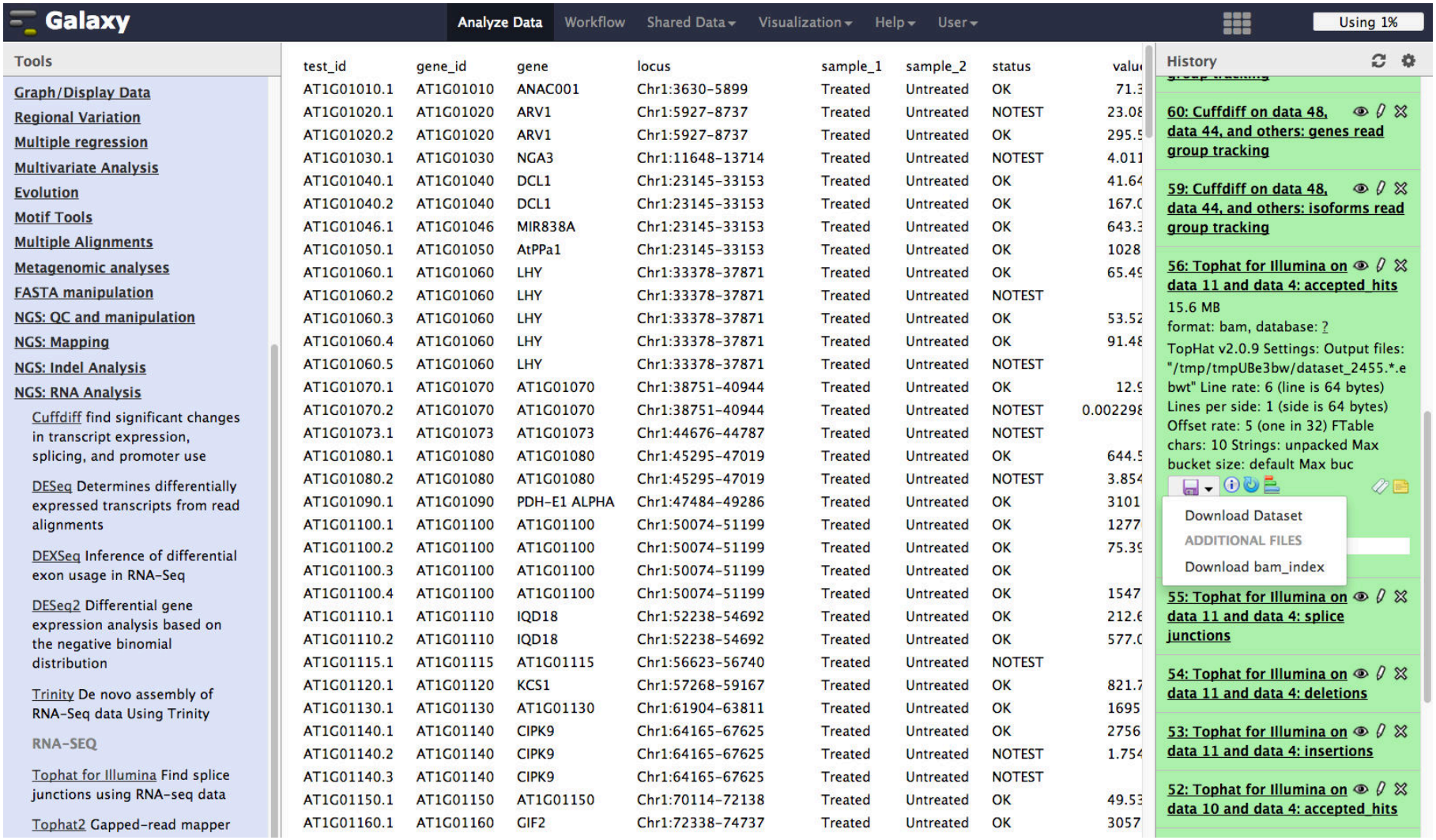
- Cuffdiff find significant changes in transcript expression

Output Files from Galaxy

- SNP-seq
  - Save Bam file BWA generated
  - Save .bai file (index of BAM) BWA generated
  - Save vcf file Samtools mpileup generated
  - Already saved them at:
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Snpseq/

- RNA-seq
  - Save four Bam files and four .bai files
  - Already saved them at:
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/

Downloading bam index file (.bai)

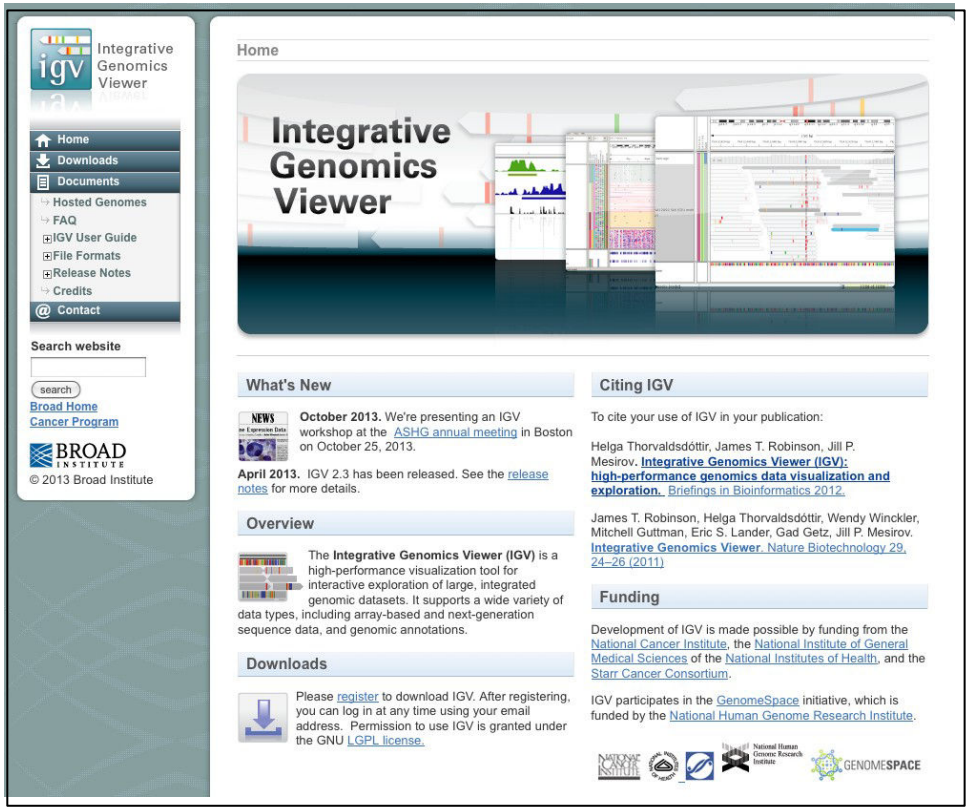


Outline

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

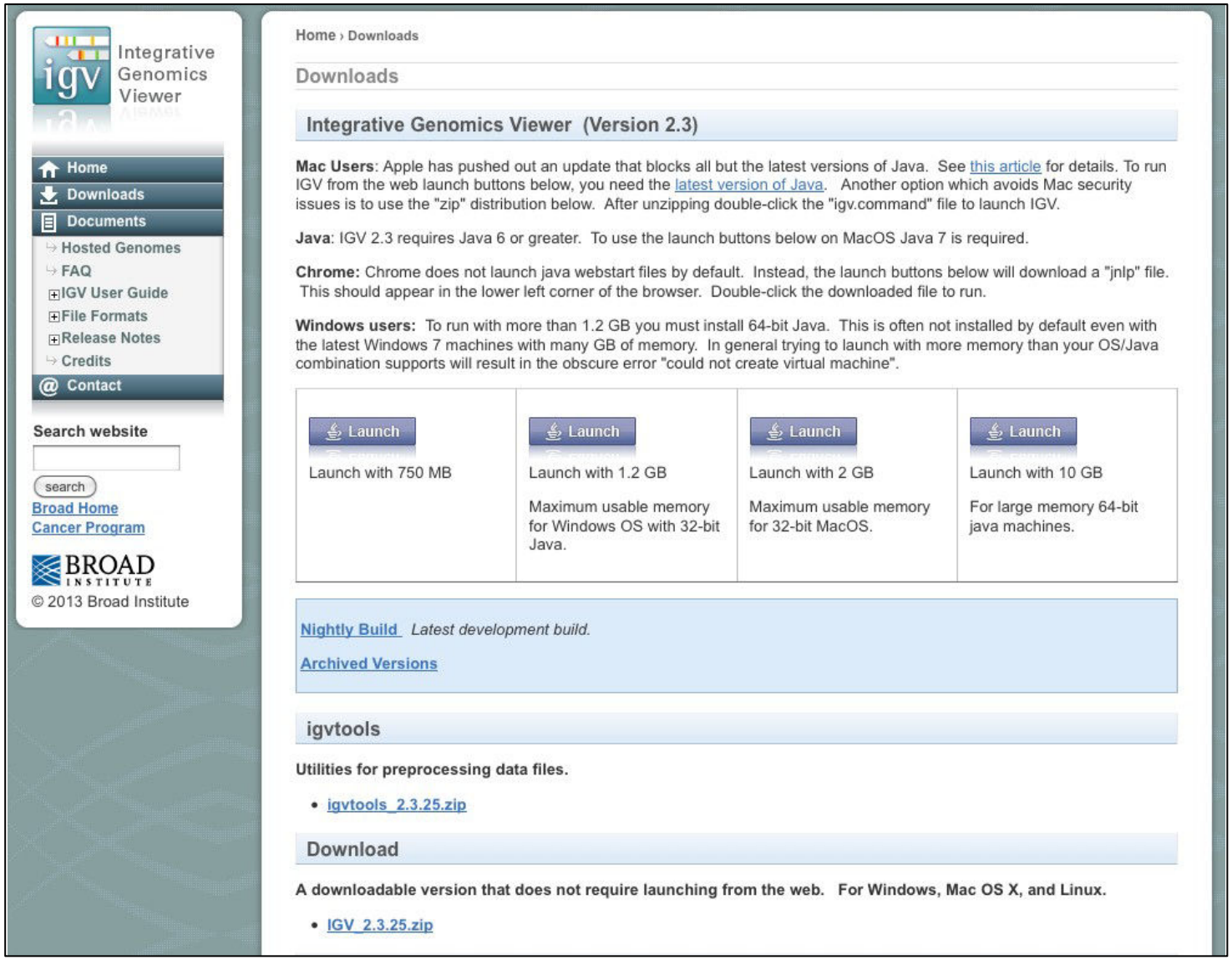
Why IGV

- IGV is an integrated visualization tool of large data types
- View large dataset easily
- Faster navigation on browsing
- Run it locally on your computer
- Easy to use interface



http://www.broadinstitute.org/igv/

IGV download

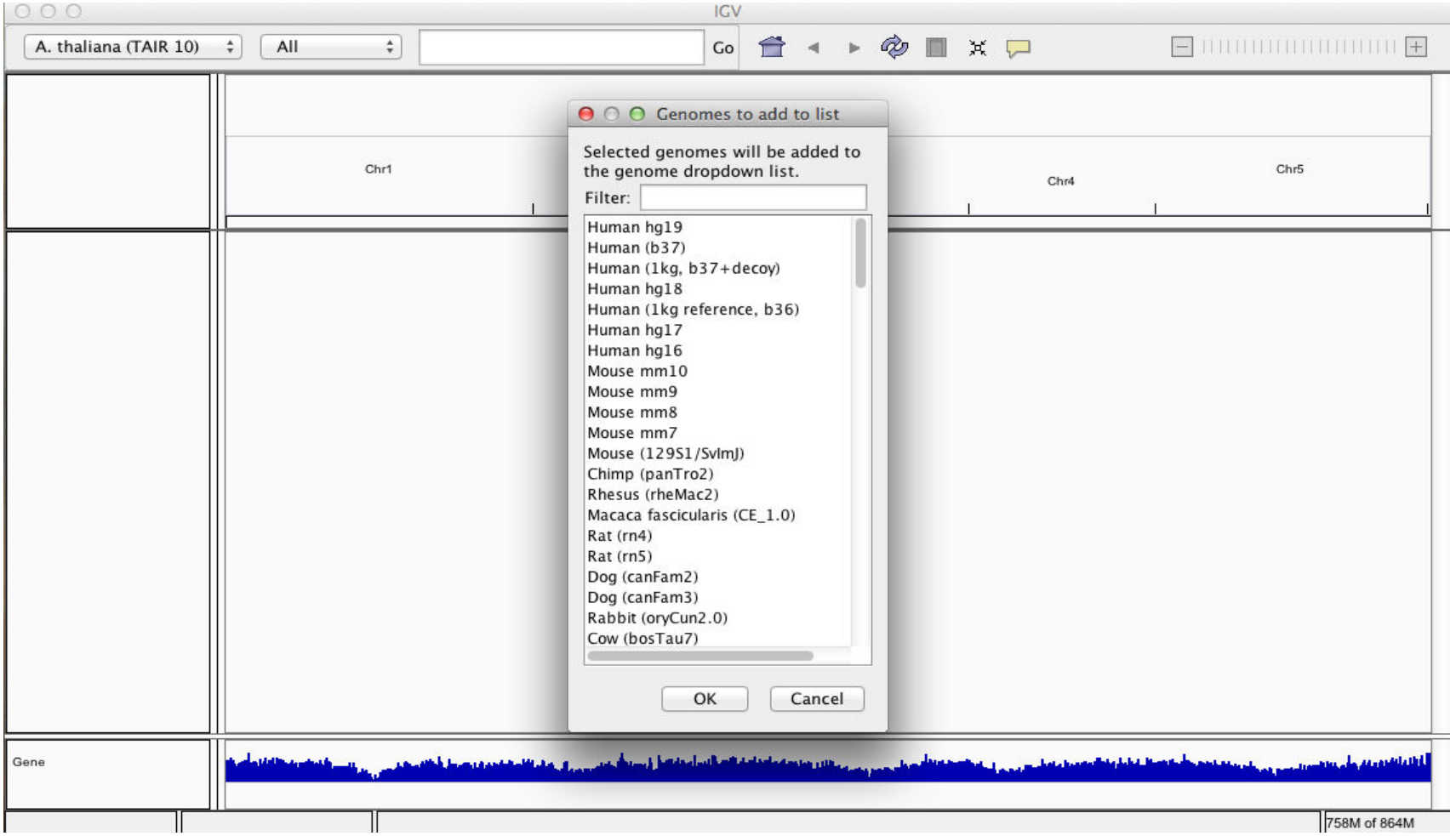


IGV interface

Load data

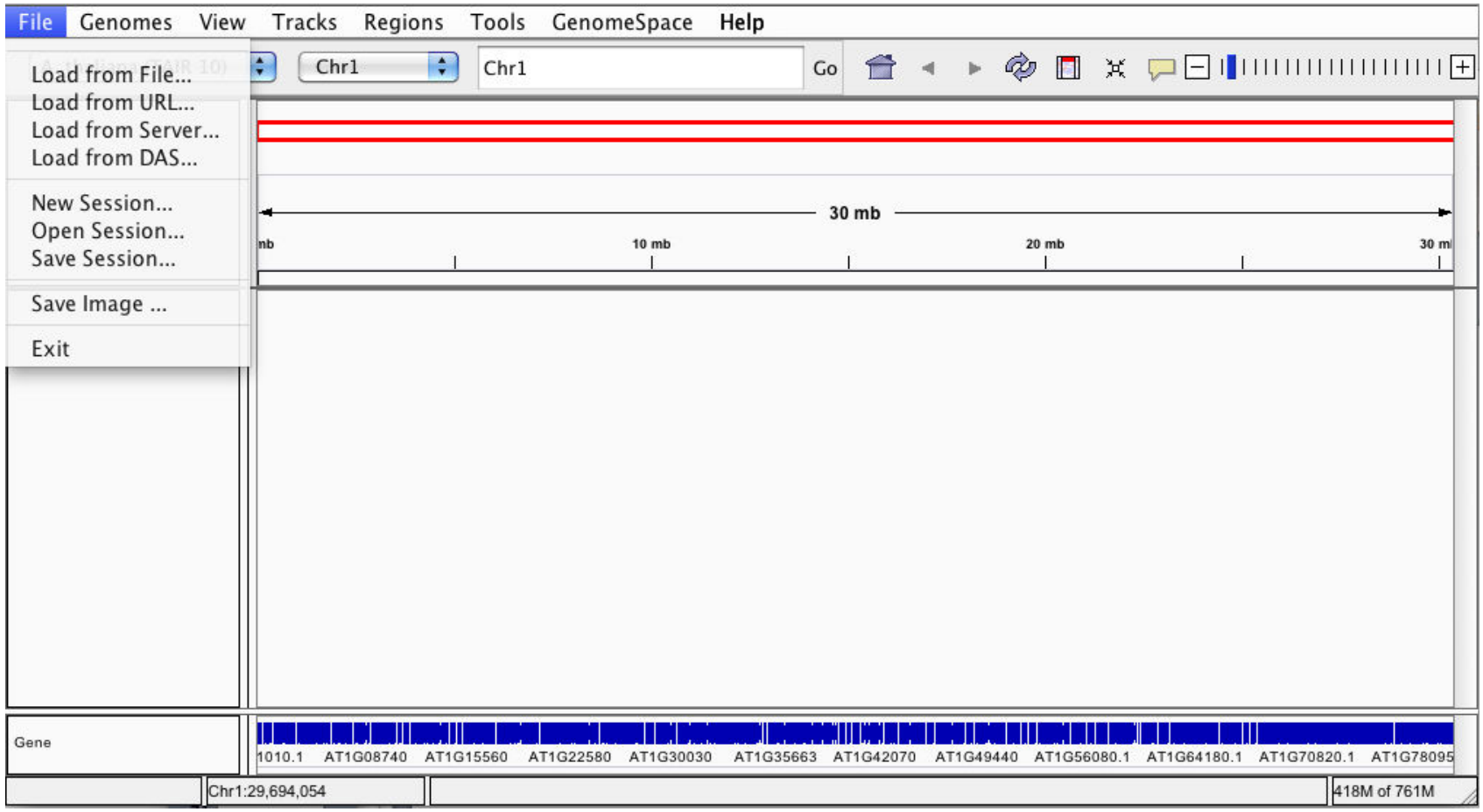
Select genome: Click the genome drop-down list in the toolbar and select the genome

Select chromosome



Load data files

- Load from URL, file, server



Toolbar

- Genome drop-down box: loads a genome
- Chromosome drop-down box: zooms to a chromosome
- Search box: Displays the chromosome location being shown. To scroll to a different location, enter the gene name, locus or track name and click Go.
- Whole genome view: Zooms to whole genome view.
- Define a region: Defines a region of interest on the chromosome.
- Zoom slider: Zooms in and out on a chromosome.

Change Display Options

- IGV offers several display options for tracks
- Zoom in and Zoom out
- Modify Track Height
- Sort the Tracks
- Filter the Tracks
- Group the Tracks
- Sort Tracks based on Region of Interest

Variants Visualization in IGV

- Load TAIR10 genome to IGV
- Load BAM file “SRR038850.bam” to IGV with “Load from URL”
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Snpseq/SAM-to-BAM_BAM.bam
- Load VCF file “var.raw.vcf” to IGV with “Load from URL”
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Snpseq/var.raw.vcf

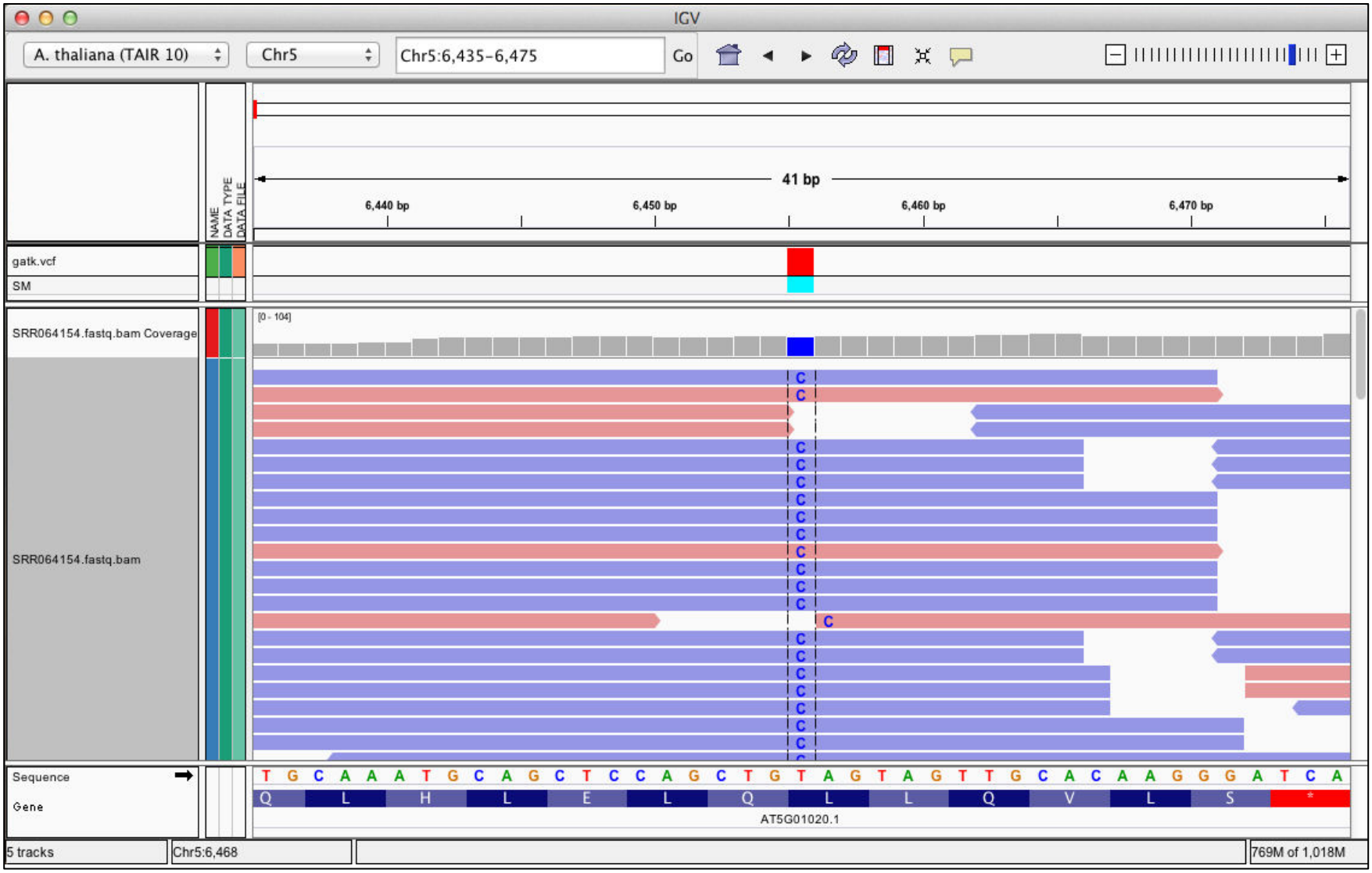
Zoom In Screen

Zoom in to : Chr5:57,073-57,142

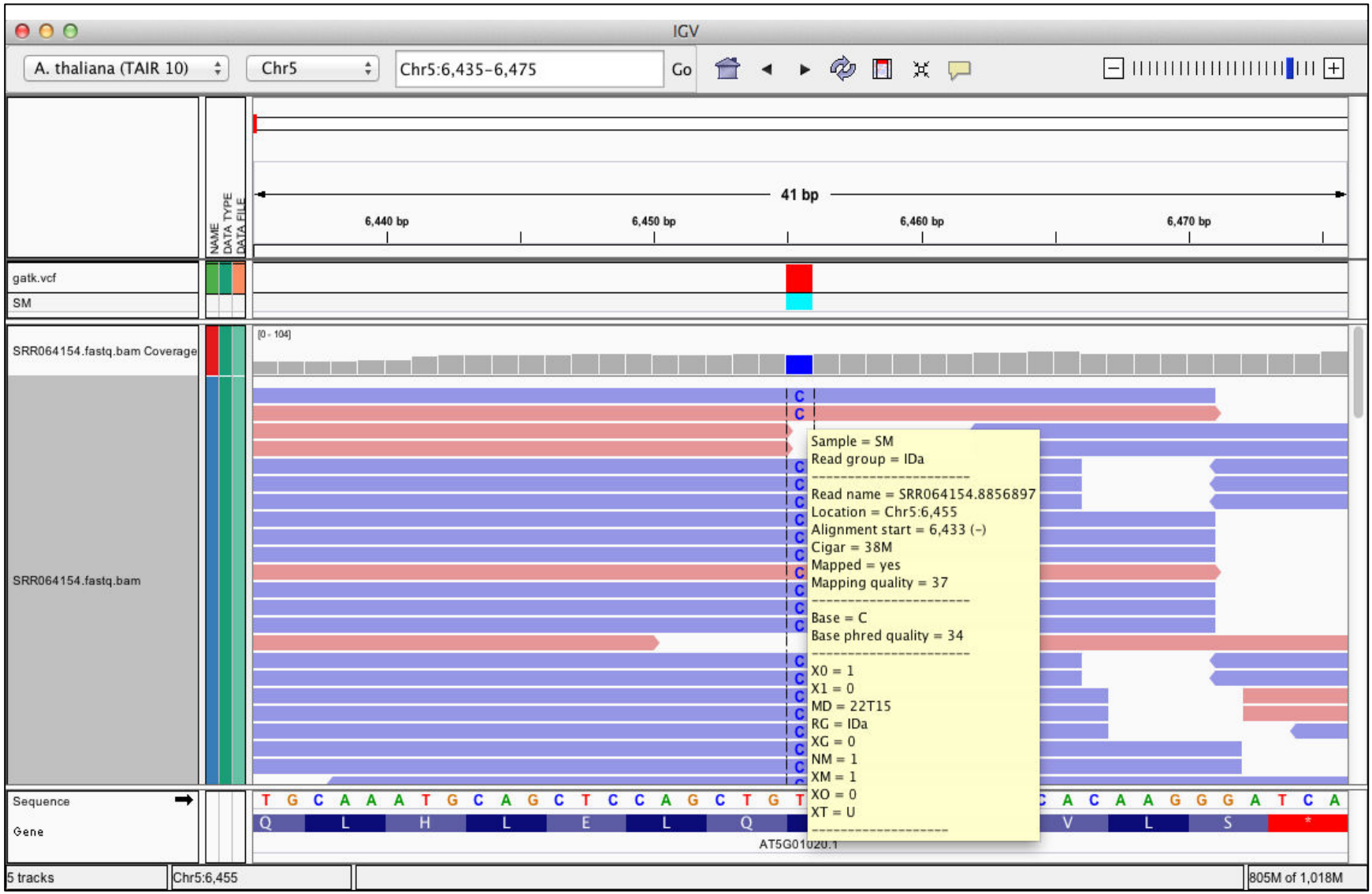
Variants Visualization in IGV

- Load TAIR10 genome to IGV
- Load BAM file “SRR064154.bam” to IGV with “Load from URL”
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Snpseq/SRR064154.fastq.bam
- Load VCF file “var.raw.vcf” to IGV with “Load from URL”
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Snpseq/gatk.vcf

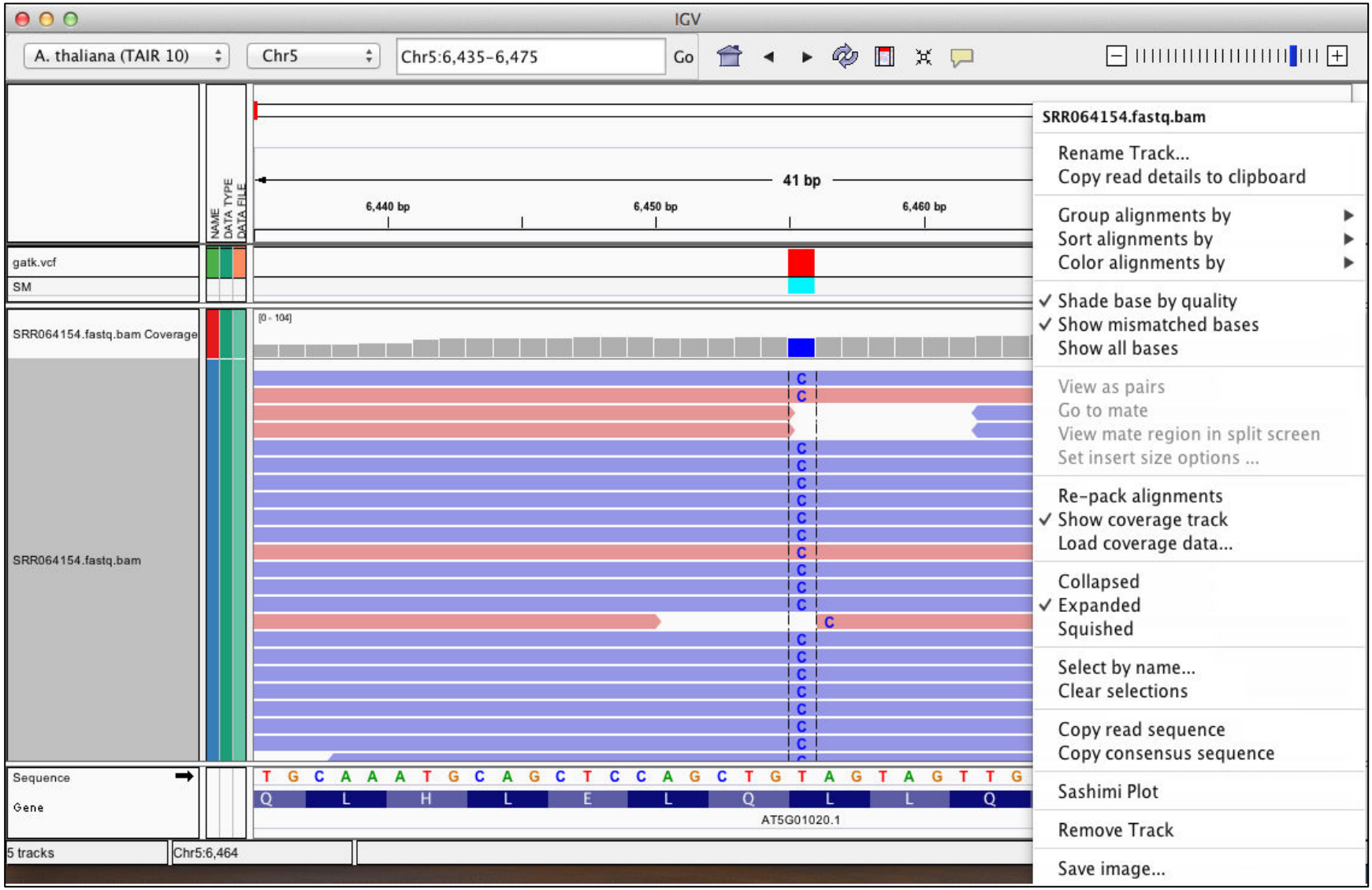
Zoom in position (chr5:6,435-6,475)



Zoom in position (chr5:6,435-6,475)



Right click on track for more options



Show all bases in IGV

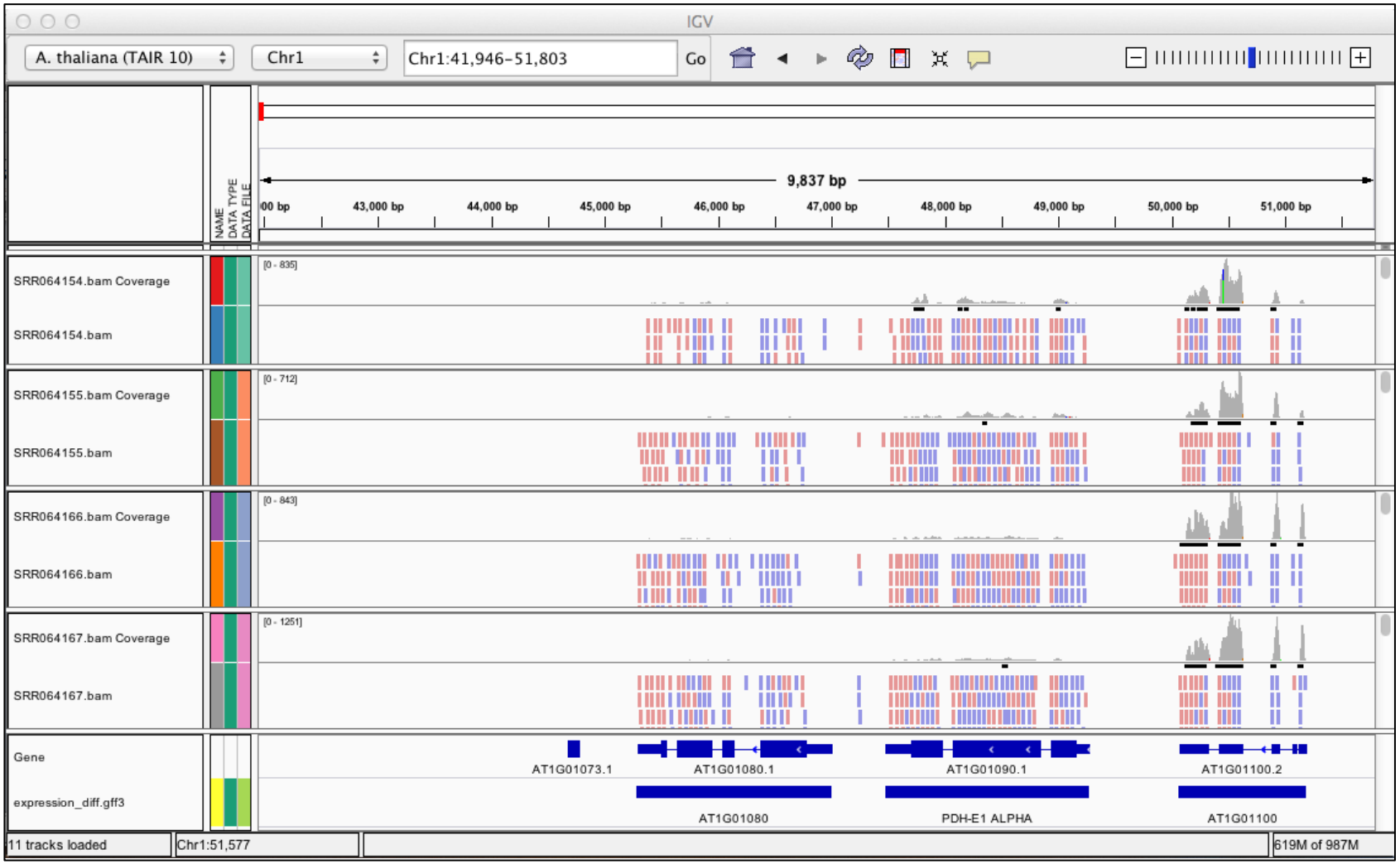
RNA-seq Results Visualization

- Load four BAM files to IGV
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/SRR064154.bam
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/SRR064155.bam
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/SRR064166.bam
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/SRR064167.bam

- Load gene differential expression GFF3 file “expression_diff.gff3” to IGV
  - http://biocluster.ucr.edu/~rkaundal/Galaxy_workshop/Rnaseq/expression_diff.gff3

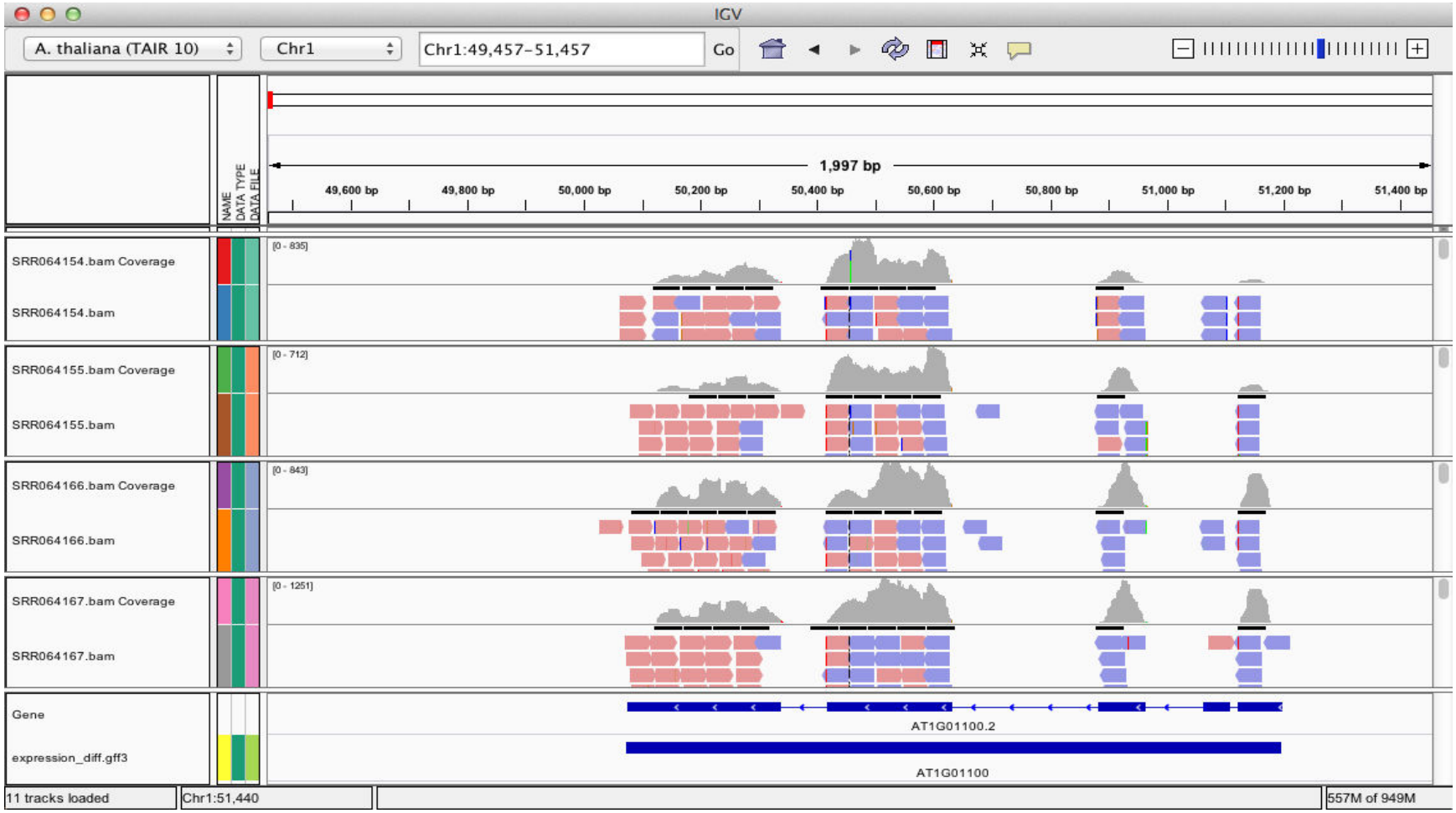
Exercise 2: RNA-seq Result Visualization

Zoom in to Chr1:41,351-51,208

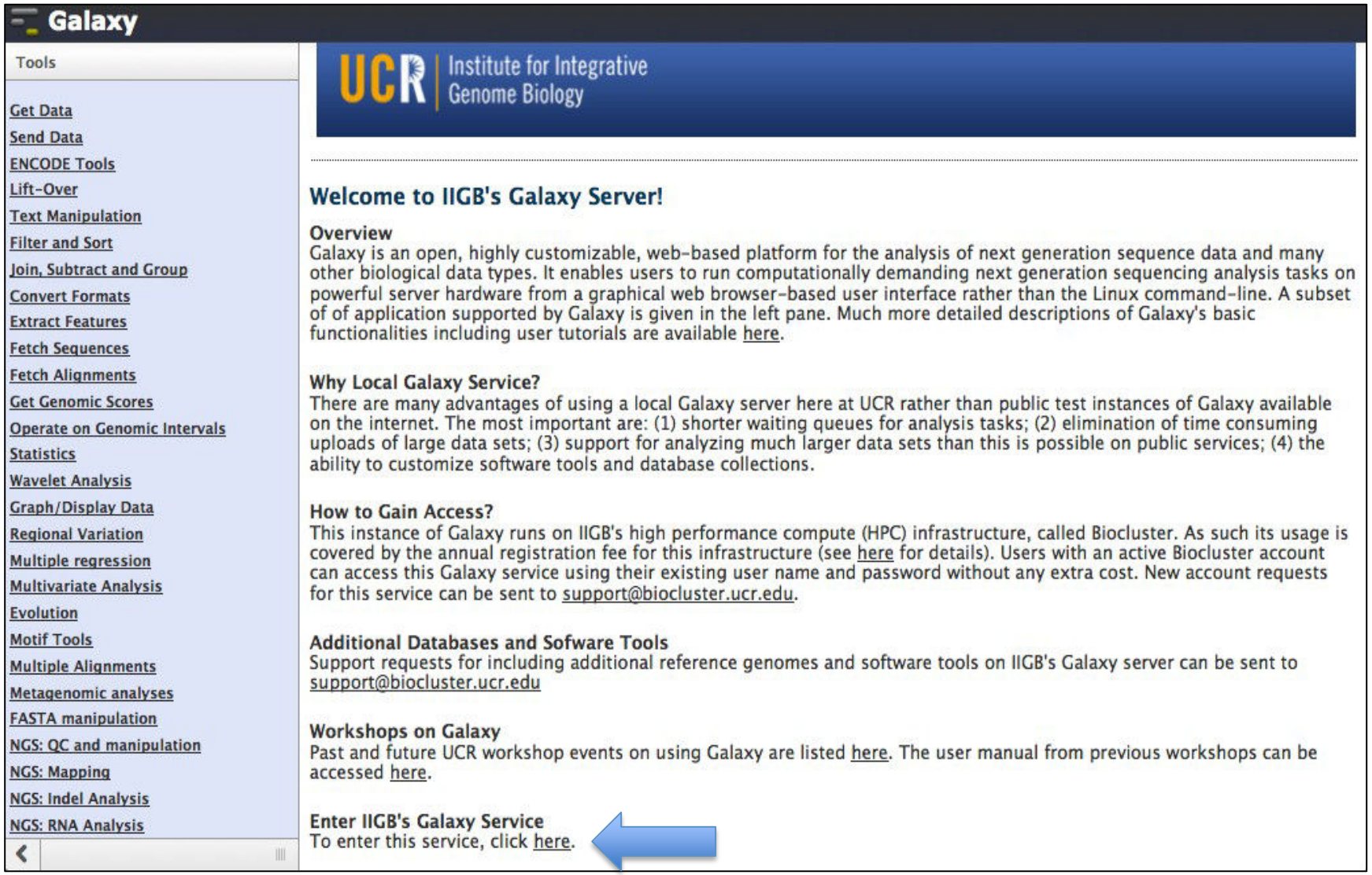


Exercise 2: RNA-seq Result Visualization

Zoom in Chr1:49,457-51,457



UCR Galaxy homepage (https://galaxy.bioinfo.ucr.edu)



Thank You