

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>)

Galaxy

Tools

[Get Data](#)

[Send Data](#)

[ENCODE Tools](#)

[Lift-Over](#)

[Text Manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Convert Formats](#)

[Extract Features](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

[Get Genomic Scores](#)

[Operate on Genomic Intervals](#)

[Statistics](#)

[Wavelet Analysis](#)

[Graph/Display Data](#)

[Regional Variation](#)

[Multiple regression](#)

[Multivariate Analysis](#)

[Evolution](#)

[Motif Tools](#)

[Multiple Alignments](#)

[Metagenomic analyses](#)

[FASTA manipulation](#)

[NGS: QC and manipulation](#)

[NGS: Mapping](#)

[NGS: Indel Analysis](#)

[NGS: RNA Analysis](#)

**UCR | Institute for Integrative
Genome Biology**

Welcome to IIGB's Galaxy Server!

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 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 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](#).

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.

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 [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 this service can be sent to support@biocluster.ucr.edu.

Additional Databases and Software Tools

Support requests for including additional reference genomes and software tools on IIGB's Galaxy server can be sent to support@biocluster.ucr.edu

Workshops on Galaxy

Past and future UCR workshop events on using Galaxy are listed [here](#). The user manual from previous workshops can be accessed [here](#).

Enter IIGB's Galaxy Service

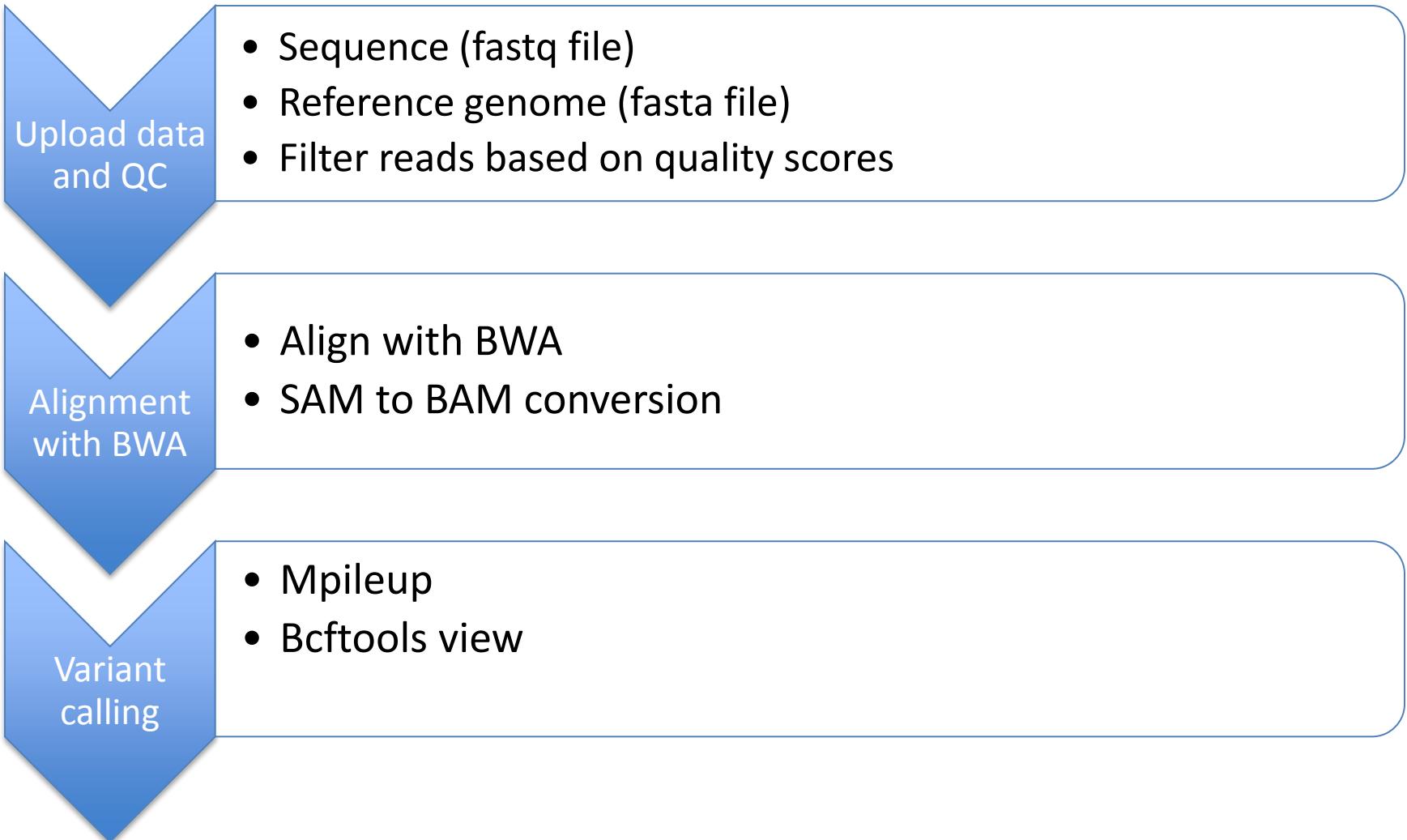
To enter this service, click [here](#).



SNPSeq Analysis dataset

- Data source: SRR038850 sample from experiment published by Kaufman et al (2012, GSE20176)
- FastqFile:
http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/SRR038850.fastq
- TAIR10 Genome:
http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/tair10chr.fasta

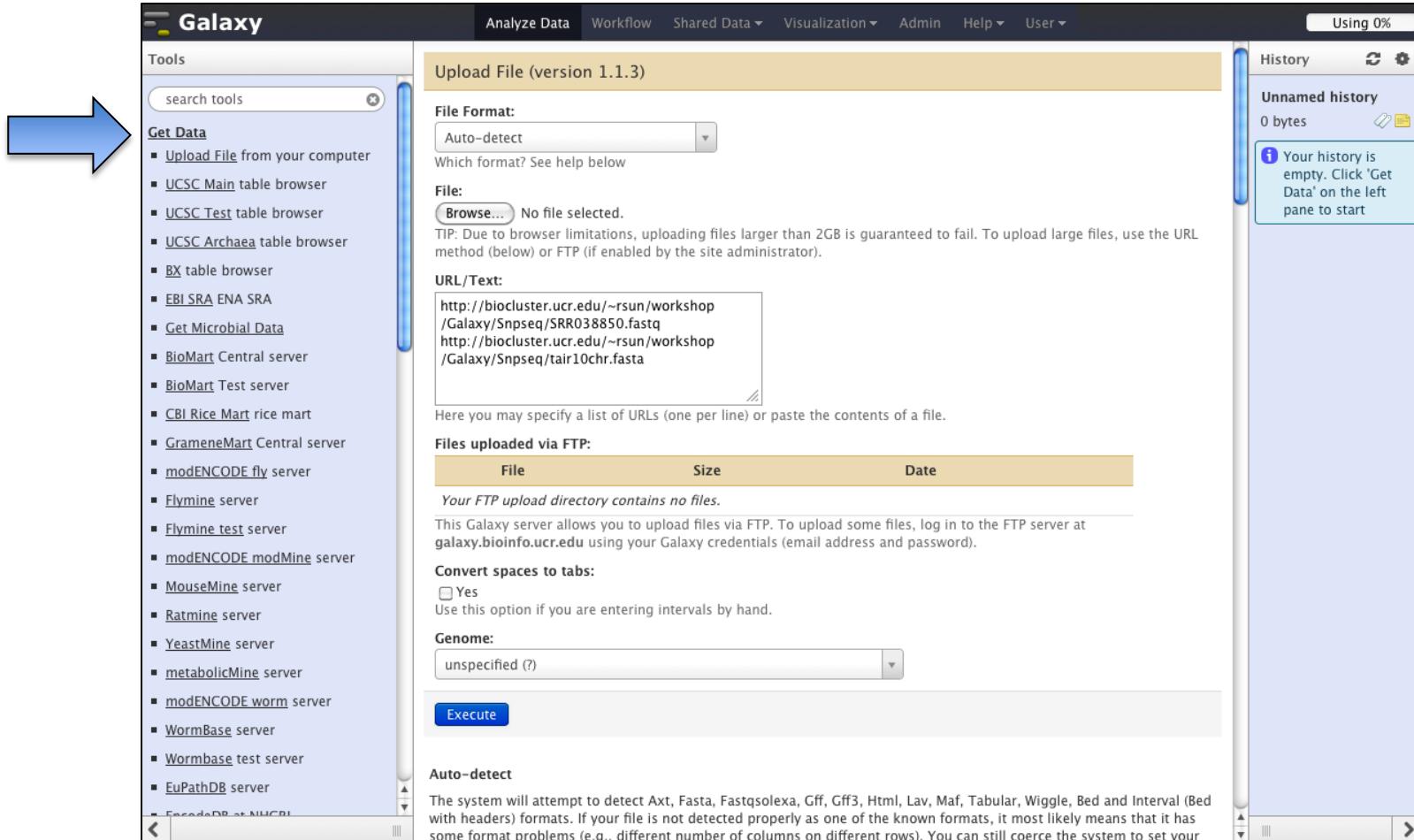
SNP-Seq pipeline



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

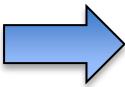


The screenshot shows the Galaxy web interface with the 'Upload File (version 1.1.3)' tool selected. On the left, a sidebar lists various 'Get Data' options, with a blue arrow pointing to the 'Upload File from your computer' link. The main panel shows the 'File Format' set to 'Auto-detect' and the 'URL/Text' field containing the two URLs provided in the text above. The right side of the interface includes a 'History' pane showing an empty history and a message about emptying the history.

http://biocluster.ucr.edu/~rsun/workshop
/Galaxy/Snpseq/SRR038850.fastq
http://biocluster.ucr.edu/~rsun/workshop
/Galaxy/Snpseq/tair10chr.fasta

Convert Fastq file to sanger sequences

- 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.



The screenshot shows the Galaxy web interface with the following details:

- Tools Panel:** On the left, under the "Tools" heading, the "FASTQ Groomer (version 1.0.4)" tool is selected. Other visible tools include "Get Genomic Scores", "Operate on Genomic Intervals", "Statistics", "Wavelet Analysis", "Graph/Display Data", "Regional Variation", "Multiple regression", "Multivariate Analysis", "Evolution", "Motif Tools", "Multiple Alignments", "Metagenomic analyses", "FASTA manipulation", and "NGS: QC and manipulation".
- Tool Configuration:** The "File to groom:" field contains the URL http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/SRR038850.fastq. The "Input FASTQ quality scores type:" dropdown is set to "Sanger".
- Advanced Options:** A "Hide Advanced Options" button is present.
- Execute Button:** A blue "Execute" button is located below the input fields.
- What it does:** This section describes the tool's functions, including conversion options for Sanger and Color Space Sanger formats, handling of quality score ranges, and mapping between Solexa and PHRED scales.
- Quality Score Comparison:** A chart comparing quality scores across various bases (A, T, C, G) and positions (33, 59, 64, 73, 104, 126).
- History Panel:** On the right, the "History" panel shows two entries:
 - Entry 2: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/tair10chr.fasta
 - Entry 1: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/SRR038850.fastq

Features available in history panel

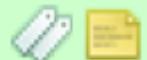
- View, Edit, Delete file
- Size of the file
- Save the file
- Repeat the analysis

1: <http://biocluster.ucr.edu> /
/~nkatiyar/Galaxy_workshop
/Snpseq/SRR038850.fastq

66.8 MB

format: fastq, database: ?

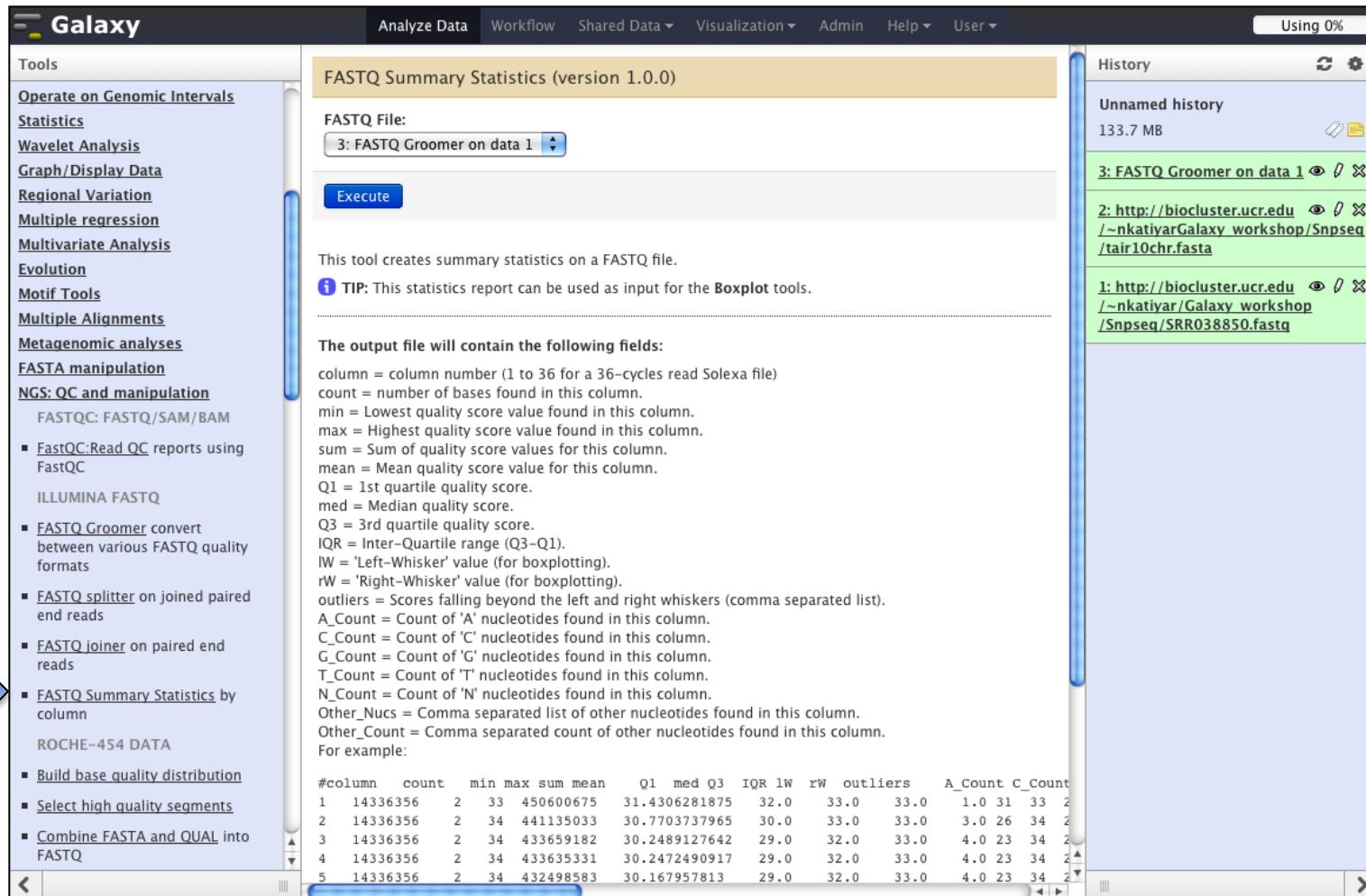
uploaded fastq file



```
@SRR038850.12 HWI-EAS038:3:1:2:1948 le  
CAAGCATTTTTGAATTCGGATTATCCGTTA  
+SRR038850.12 HWI-EAS038:3:1:2:1948 le  
@?@<@?@BBAAB@>===@AB?7=<:6>@A@?:6
```

FASTQ Summary Statistics

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

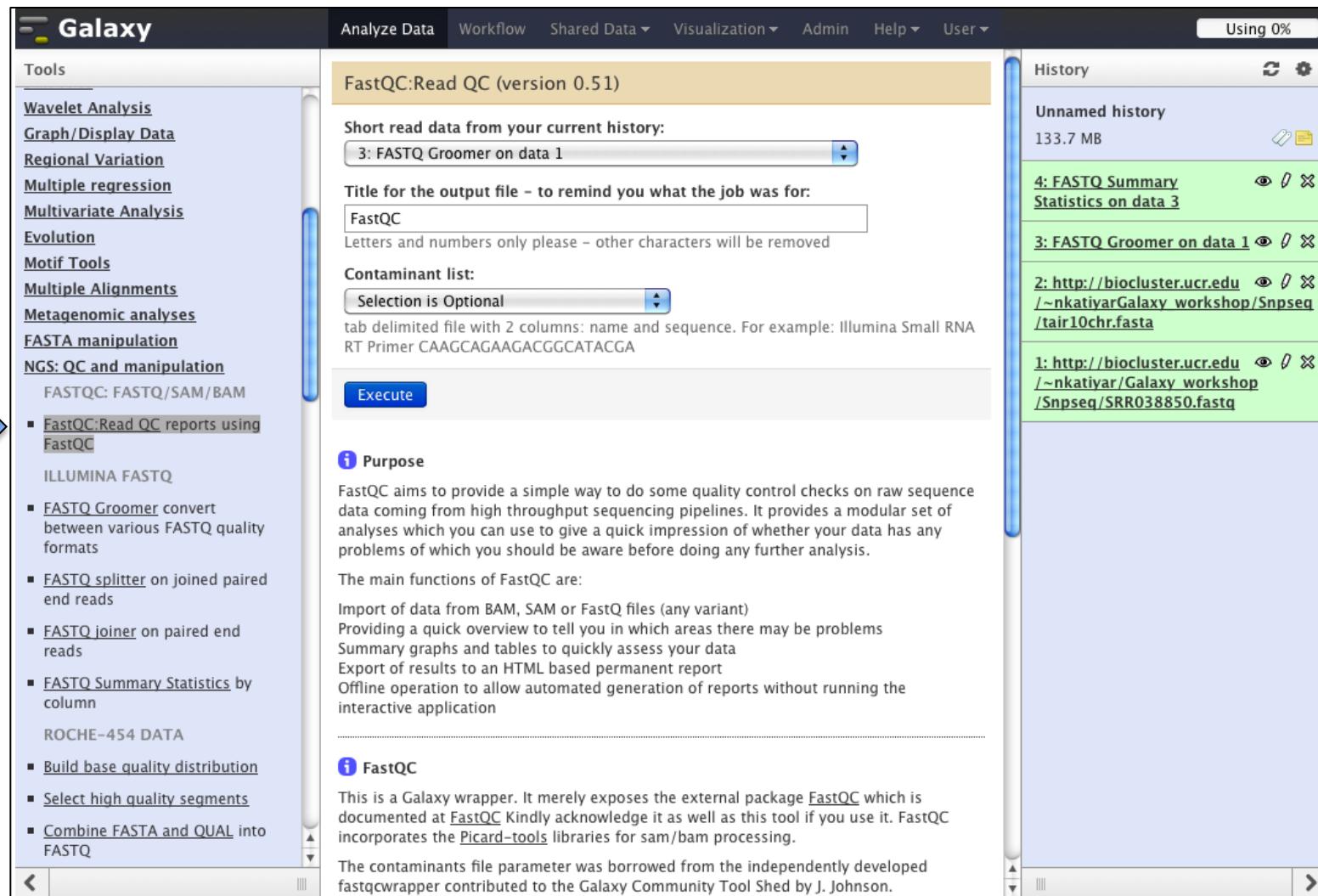


The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.
- Tools Sidebar (Left):**
 - Operate on Genomic Intervals
 - Statistics
 - Wavelet Analysis
 - Graph/Display Data
 - Regional Variation
 - Multiple regression
 - Multivariate Analysis
 - Evolution
 - Motif Tools
 - Multiple Alignments
 - Metagenomic analyses
 - FASTA manipulation
 - NGS: QC and manipulation**
 - FASTQC: FASTQ/SAM/BAM
 - FastQC: Read QC reports using FastQC
 - ILLUMINA FASTQ
 - FASTQ Groomer convert between various FASTQ quality formats
 - FASTQ splitter on joined paired end reads
 - FASTQ joiner on paired end reads
 - FASTQ Summary Statistics by column
 - ROCHE-454 DATA
 - Build base quality distribution
 - Select high quality segments
 - Combine FASTA and QUAL into FASTQ
- Tool Panel (Center):** FASTQ Summary Statistics (version 1.0.0).
 - FASTQ File: 3: FASTQ Groomer on data 1
 - Execute
 - This tool creates summary statistics on a FASTQ file.
 - TIP:** This statistics report can be used as input for the Boxplot tools.
 - The output file will contain the following fields:**
 - column = column number (1 to 36 for a 36-cycles read Solexa file)
 - count = number of bases found in this column.
 - min = Lowest quality score value found in this column.
 - max = Highest quality score value found in this column.
 - sum = Sum of quality score values for this column.
 - mean = Mean quality score value for this column.
 - Q1 = 1st quartile quality score.
 - med = Median quality score.
 - Q3 = 3rd quartile quality score.
 - IQR = Inter-Quartile range (Q3-Q1).
 - lW = 'Left-Whisker' value (for boxplotting).
 - rW = 'Right-Whisker' value (for boxplotting).
 - outliers = Scores falling beyond the left and right whiskers (comma separated list).
 - A_Count = Count of 'A' nucleotides found in this column.
 - C_Count = Count of 'C' nucleotides found in this column.
 - G_Count = Count of 'G' nucleotides found in this column.
 - T_Count = Count of 'T' nucleotides found in this column.
 - N_Count = Count of 'N' nucleotides found in this column.
 - Other_Nucs = Comma separated list of other nucleotides found in this column.
 - Other_Count = Comma separated count of other nucleotides found in this column.
- History Panel (Right):** History, Unnamed history, 133.7 MB.
 - 3: FASTQ Groomer on data 1
 - 2: http://biocluster.ucr.edu/~nkatiyar/Galaxy/workshop/Snpseq/tair10chr.fasta
 - 1: http://biocluster.ucr.edu/~nkatiyar/Galaxy/workshop/Snpseq/SRR038850.fastq

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.



The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User, Using 0%.
- Left Sidebar (Tools):**
 - Wavelet Analysis
 - Graph/Display Data
 - Regional Variation
 - Multiple regression
 - Multivariate Analysis
 - Evolution
 - Motif Tools
 - Multiple Alignments
 - Metagenomic analyses
 - FASTA manipulation
 - NGS: QC and manipulation**
 - FASTQC: FASTQ/SAM/BAM
 - FastQC: Read QC reports using FastQC** (highlighted with a blue arrow)
 - ILLUMINA FASTQ
 - FASTQ Groomer convert between various FASTQ quality formats
 - FASTQ splitter on joined paired end reads
 - FASTQ joiner on paired end reads
 - FASTQ Summary Statistics by column
 - ROCHE-454 DATA
 - Build base quality distribution
 - Select high quality segments
 - Combine FASTA and QUAL into FASTQ
- Central Panel:** FastQC:Read QC (version 0.51).
 - Short read data from your current history:
3: FASTQ Groomer on data 1
 - Title for the output file – to remind you what the job was for:
FastQC
 - Letters and numbers only please – other characters will be removed
 - Contaminant list:
Selection is Optional
 - tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAACGGCATACGA
 - Execute** button
- Right Panel (History):**
 - History
 - Unnamed history
133.7 MB
 - 4: FASTQ Summary Statistics on data 3
 - 3: FASTQ Groomer on data 1
 - 2: http://biocluster.ucr.edu/~nkatiyarGalaxy workshop/Snpseq/tair10chr.fasta
 - 1: http://biocluster.ucr.edu/~nkatiyar/Galaxy workshop/Snpseq/SRR038850.fastq

Quality control output

The screenshot shows the Galaxy web interface with a "FastQC Report" tool running on a dataset named "FASTQ_Groomer_on_data_1".

Tools Panel:

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

Report Content:

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

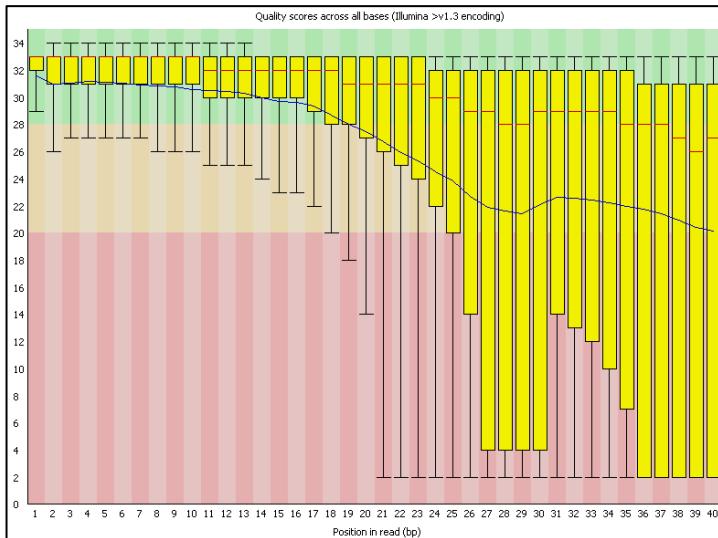
Measure	Value
Filename	FASTQ_Groomer_on_data_1
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	386588
Filtered Sequences	0
Sequence length	36
%GC	42

History:

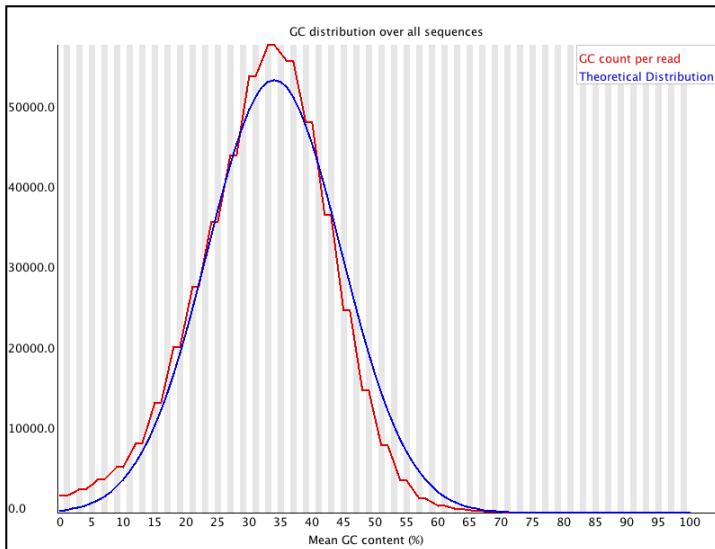
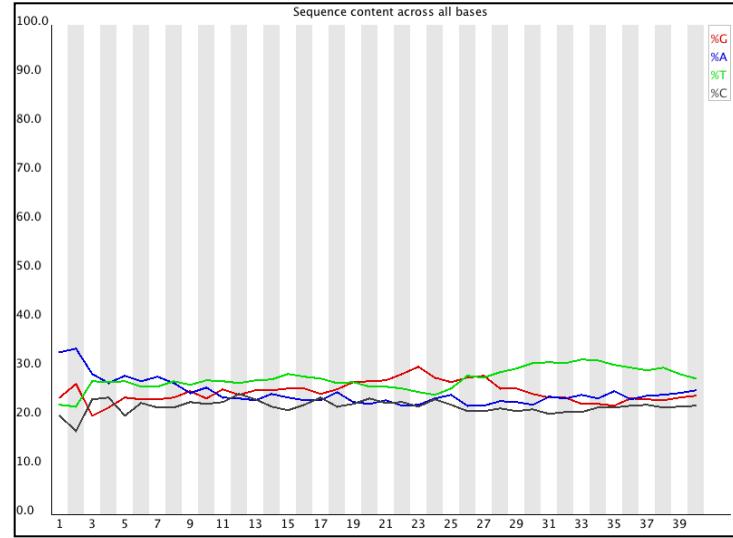
- 15: bcftools view on data 13
- 14: MPileup on data 7 and data 10
- 13: MPileup on data 7 and data 10
- 10: SAM-to-BAM on data 7 and data 9: converted BAM
- 9: Map with BWA for Illumina on data 8 and data 7: mapped reads
- 8: Filter FASTQ on data 3
- 7: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/tair10chr.fasta
- 5: FastQC FASTQ Groomer on data 1.html
- 4: FASTQ Summary Statistics on data 3
- 3: FASTQ Groomer on data 1
- 1: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/SRR038850.fasta

Quality control reports

Per base sequence quality



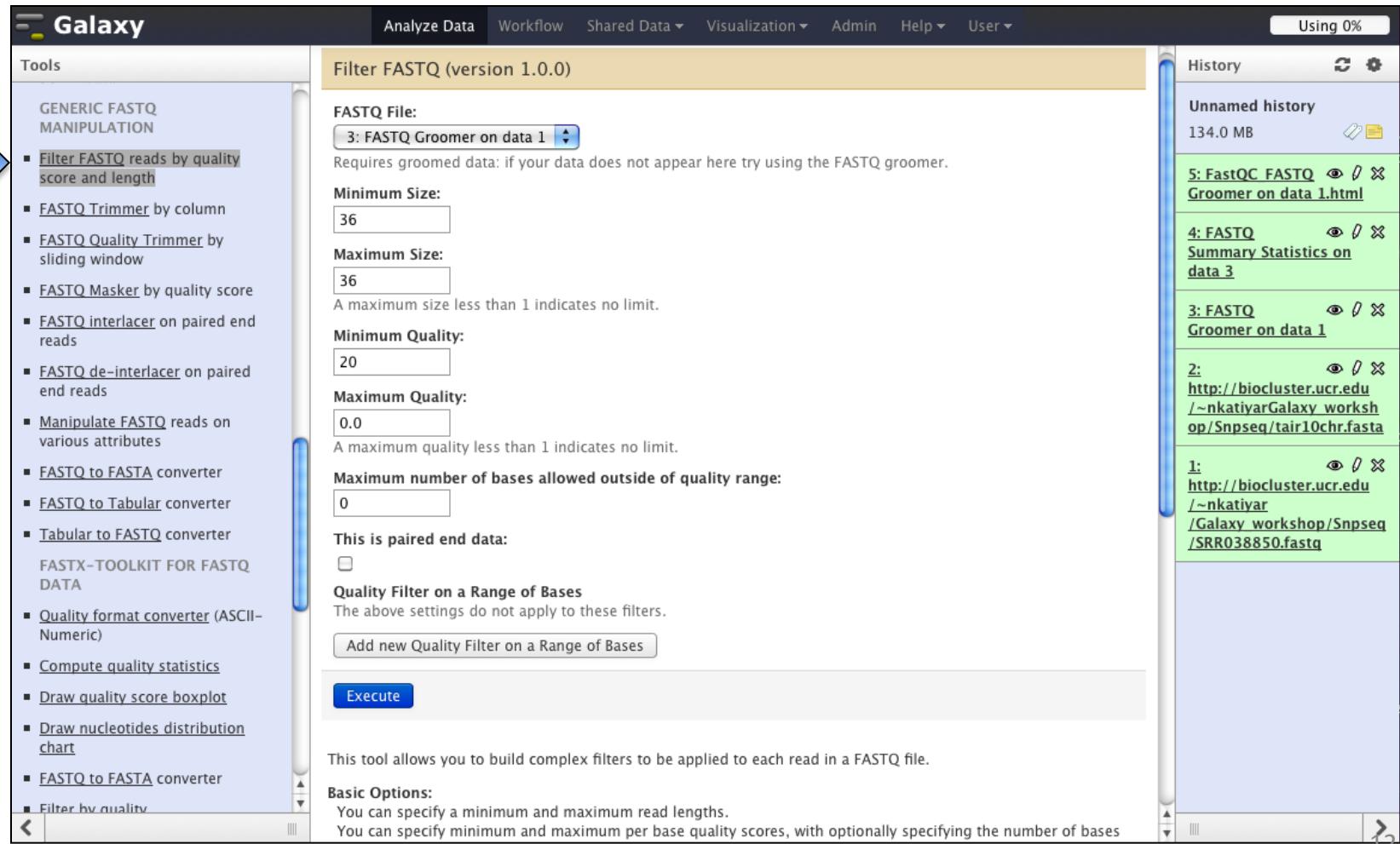
Per base sequence content



Per base 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

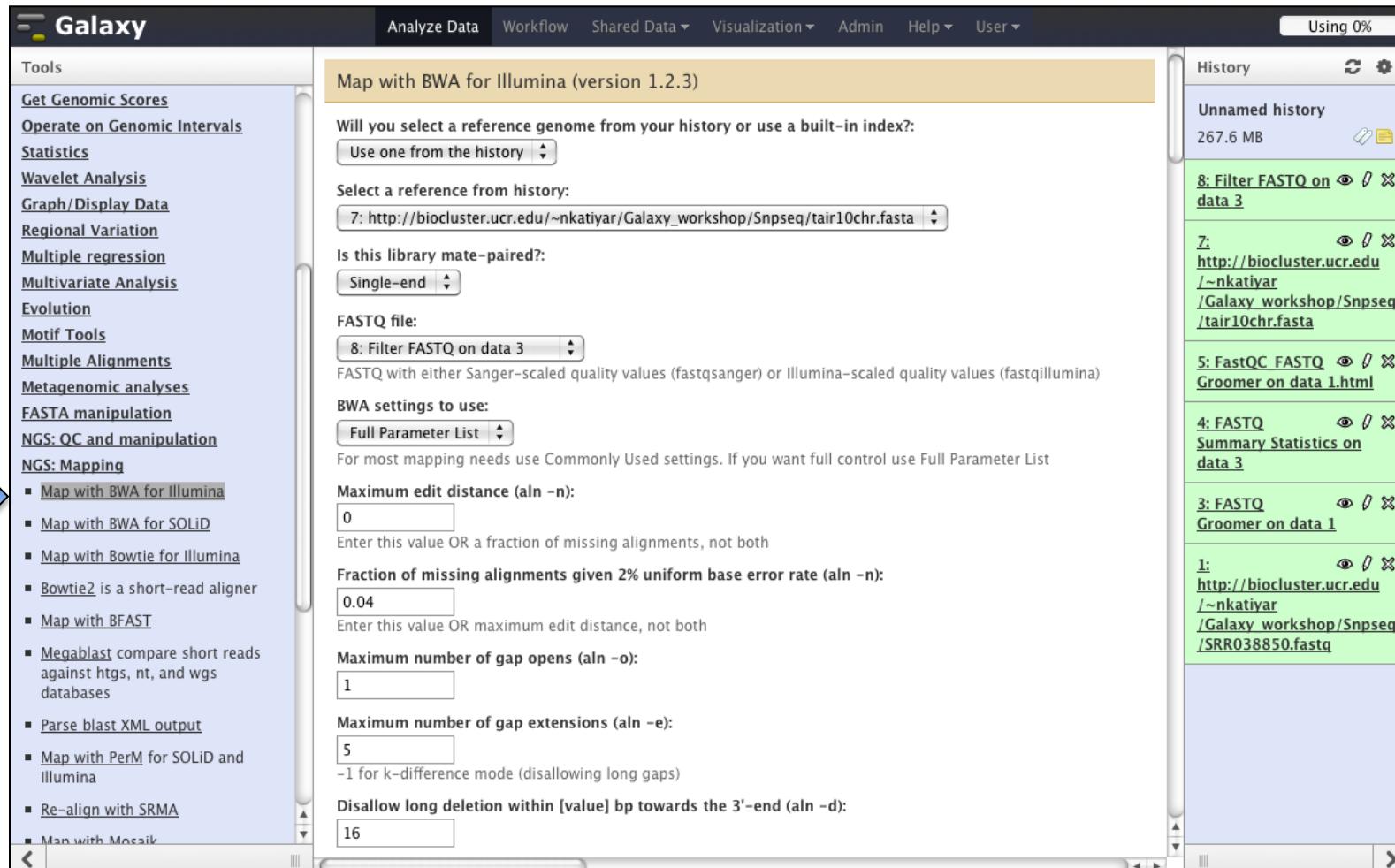


The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User, Using 0%
- Tools Sidebar:** GENERIC FASTQ MANIPULATION
 - Selected:** Filter FASTQ reads by quality score and length
 - FASTQ Trimmer by column
 - FASTQ Quality Trimmer by sliding window
 - FASTQ Masker by quality score
 - FASTQ interlacer on paired end reads
 - FASTQ de-interlacer on paired end reads
 - Manipulate FASTQ reads on various attributes
 - FASTQ to FASTA converter
 - FASTQ to Tabular converter
 - Tabular to FASTQ converter
- Tool Panel:** Filter FASTQ (version 1.0.0)
 - FASTQ File:** 3: FASTQ Groomer on data 1
 - Requires groomed data: if your data does not appear here try using the FASTQ groomer.
 - Minimum Size:** 36
 - Maximum Size:** 36
 - A maximum size less than 1 indicates no limit.
 - Minimum Quality:** 20
 - Maximum Quality:** 0
 - A maximum quality less than 1 indicates no limit.
 - Maximum number of bases allowed outside of quality range:** 0
 - This is paired end data:**
 - Quality Filter on a Range of Bases**
The above settings do not apply to these filters.
Add new Quality Filter on a Range of Bases
 - Execute**
- History Panel:** History, Unnamed history (134.0 MB), 5: FastQC FASTQ Groomer on data 1.html, 4: FASTQ Summary Statistics on data 3, 3: FASTQ Groomer on data 1, 2: http://biocluster.ucr.edu/~nkatiyarGalaxy/workshop/Snpseq/tair10chr.fasta, 1: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/SRR038850.fastq

Alignment with BWA

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

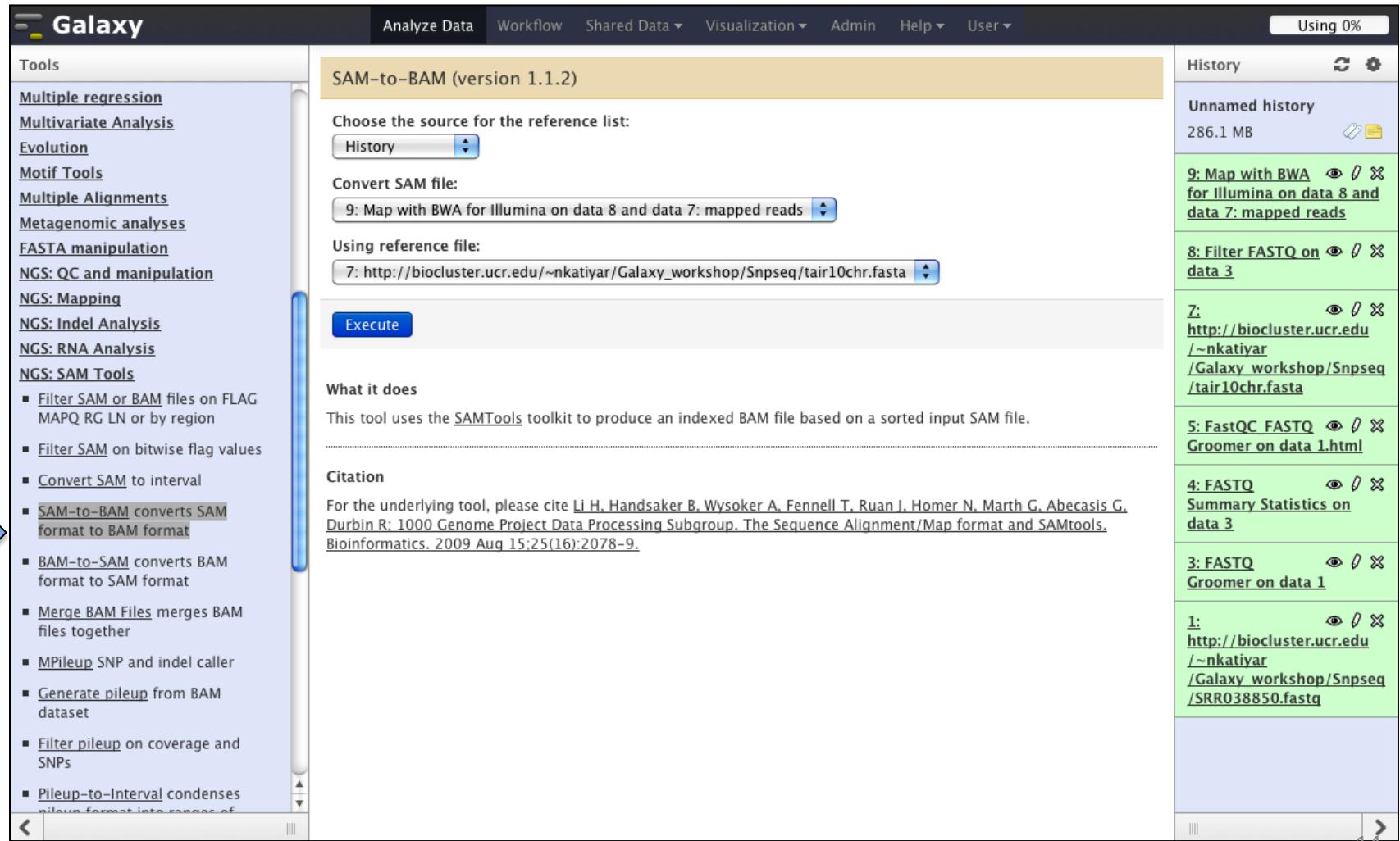


The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User, Using 0%
- Left Sidebar (Tools):**
 - Get Genomic Scores
 - Operate on Genomic Intervals
 - Statistics
 - Wavelet Analysis
 - Graph/Display Data
 - Regional Variation
 - Multiple regression
 - Multivariate Analysis
 - Evolution
 - Motif Tools
 - Multiple Alignments
 - Metagenomic analyses
 - FASTA manipulation
 - NGS: QC and manipulation
 - NGS: Mapping** (selected)
 - Map with BWA for Illumina
 - Map with BWA for SOLiD
 - Map with Bowtie for Illumina
 - Bowtie2 is a short-read aligner
 - Map with BFAST
 - Megablast compare short reads against htgs, nt, and wgs databases
 - Parse blast XML output
 - Map with PerM for SOLiD and Illumina
 - Re-align with SRMA
 - Map with Mosaik
- Central Panel (Map with BWA for Illumina):**
 - Will you select a reference genome from your history or use a built-in index?: Use one from the history
 - Select a reference from history:
7: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/tair10chr.fasta
 - Is this library mate-paired?: Single-end
 - FASTQ file:
8: Filter FASTQ on data 3
FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)
 - BWA settings to use:
Full Parameter List
For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List
 - Maximum edit distance (aln -n): 0
Enter this value OR a fraction of missing alignments, not both
 - Fraction of missing alignments given 2% uniform base error rate (aln -n): 0.04
Enter this value OR maximum edit distance, not both
 - Maximum number of gap opens (aln -o): 1
 - Maximum number of gap extensions (aln -e): 5
-1 for k-difference mode (disallowing long gaps)
 - Disallow long deletion within [value] bp towards the 3'-end (aln -d): 16
- Right Panel (History):**
 - History
 - Unnamed history (267.6 MB)
 - 8: Filter FASTQ on data 3
 - 7: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/tair10chr.fasta
 - 5: FastQC FASTQ Groomer on data 1.html
 - 4: FASTQ Summary Statistics on data 3
 - 3: FASTQ Groomer on data 1
 - 1: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/SRR038850.fastq

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".

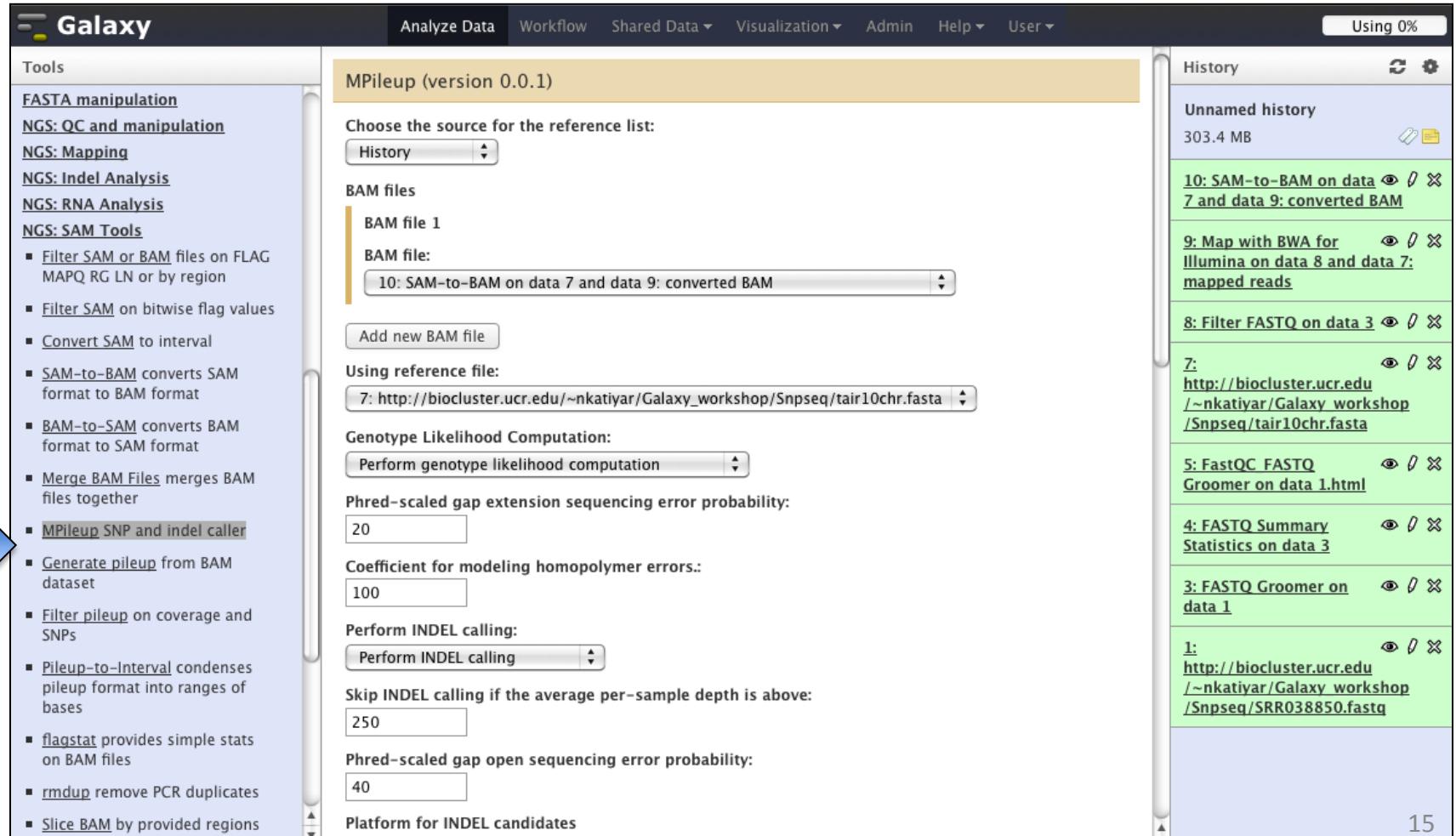


The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.
- Tools Sidebar:** Multiple regression, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments, Metagenomic analyses, FASTA manipulation, NGS: QC and manipulation, NGS: Mapping, NGS: Indel Analysis, NGS: RNA Analysis, NGS: SAM Tools (selected).
 - A blue arrow points to the "Execute" button.
- SAM-to-BAM (version 1.1.2) Panel:**
 - Choose the source for the reference list: History (dropdown).
 - Convert SAM file: 9: Map with BWA for Illumina on data 8 and data 7: mapped reads (dropdown).
 - Using reference file: 7: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/tair10chr.fasta (dropdown).
 - Execute button.
- History Panel:** Unnamed history (286.1 MB), containing:
 - 9: Map with BWA for Illumina on data 8 and data 7: mapped reads
 - 8: Filter FASTQ on data 3
 - 7: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/tair10chr.fasta
 - 5: FastQC FASTQ Groomer on data 1.html
 - 4: FASTQ Summary Statistics on data 3
 - 3: FASTQ Groomer on data 1
 - 1: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/SRR038850.fastq

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".

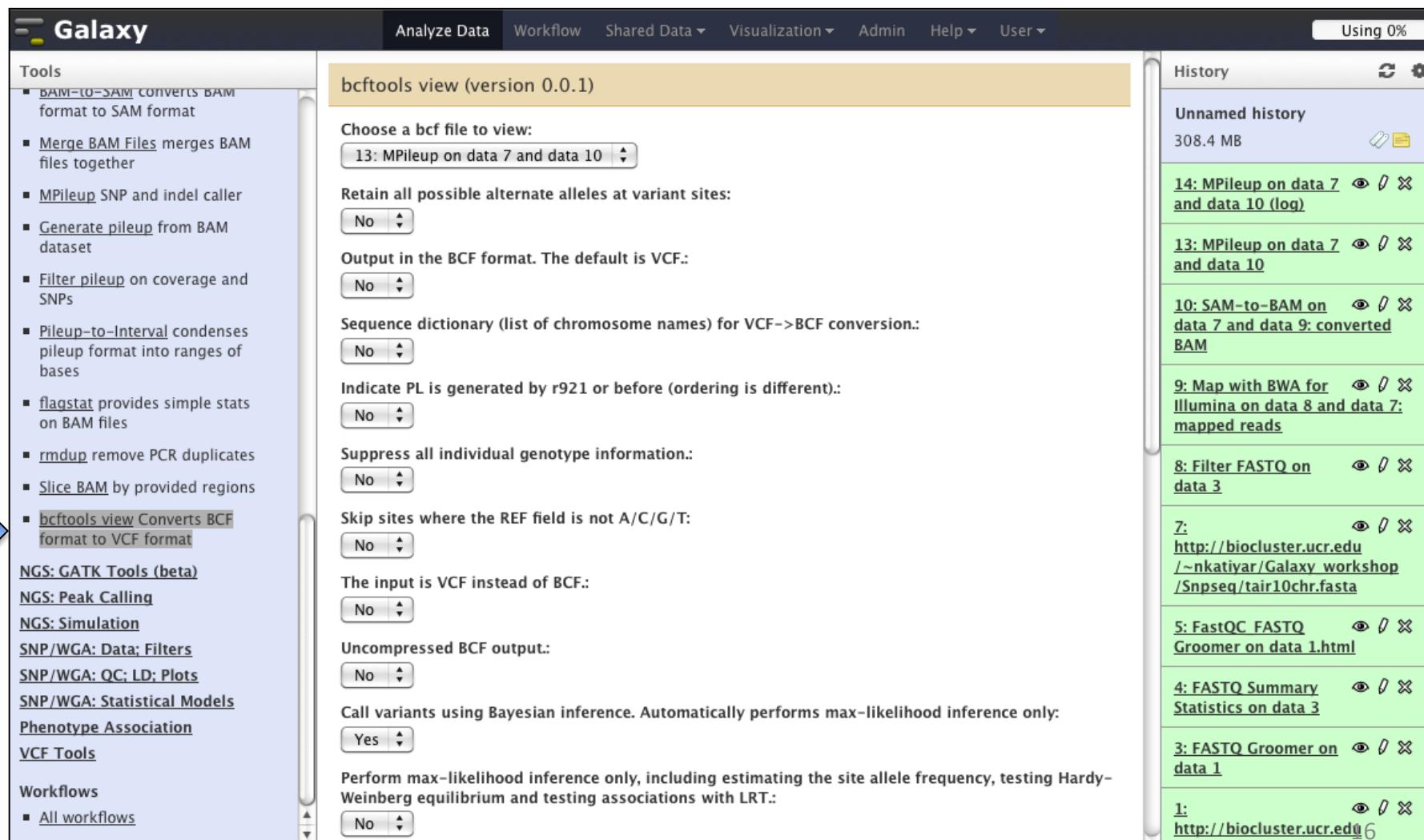


The screenshot shows the Galaxy web interface with the following details:

- Left Sidebar (Tools):** Lists various NGS tools, with the "MPileup SNP and indel caller" option highlighted by a blue arrow.
- Middle Panel (MPileup tool configuration):**
 - Choose the source for the reference list:** Set to "History".
 - BAM files:**
 - BAM file 1:** Set to "10: SAM-to-BAM on data 7 and data 9: converted BAM".
 - Add new BAM file:** Input field is empty.
 - Using reference file:** Set to "7: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/tair10chr.fasta".
 - Genotype Likelihood Computation:** Set to "Perform genotype likelihood computation".
 - Phred-scaled gap extension sequencing error probability:** Set to "20".
 - Coefficient for modeling homopolymer errors.:** Set to "100".
 - Perform INDEL calling:** Set to "Perform INDEL calling".
 - Skip INDEL calling if the average per-sample depth is above:** Set to "250".
 - Phred-scaled gap open sequencing error probability:** Set to "40".
 - Platform for INDEL candidates:** Input field is empty.
- Right Panel (History):** Shows a list of completed workflows:
 - 10: SAM-to-BAM on data 7 and data 9: converted BAM
 - 9: Map with BWA for Illumina on data 8 and data 7: mapped reads
 - 8: Filter FASTQ on data 3
 - 7: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/tair10chr.fasta
 - 5: FastQC FASTQ Groomer on data 1.html
 - 4: FASTQ Summary Statistics on data 3
 - 3: FASTQ Groomer on data 1
 - 1: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/SRR038850.fastq

Bcftools view

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



The screenshot shows the Galaxy web interface with the 'bcftools view (version 0.0.1)' tool selected. The interface includes a top navigation bar with 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The 'History' panel on the right lists several completed workflows, including '13: MPileup on data 7 and data 10' and '14: MPileup on data 7 and data 10 (log)'. The main panel displays configuration options for the 'bcftools view' tool, such as selecting a BCF file, output format (VCF), and various filtering and processing parameters. The 'Tools' sidebar on the left lists other available tools like BAM-to-SAM, Merge BAM Files, and bcftools view.

bcftools view (version 0.0.1)

Choose a bcf file to view:
13: MPileup on data 7 and data 10

Retain all possible alternate alleles at variant sites:
No

Output in the BCF format. The default is VCF:
No

Sequence dictionary (list of chromosome names) for VCF->BCF conversion:
No

Indicate PL is generated by r921 or before (ordering is different):
No

Suppress all individual genotype information:
No

Skip sites where the REF field is not A/C/G/T:
No

The input is VCF instead of BCF:
No

Uncompressed BCF output:
No

Call variants using Bayesian inference. Automatically performs max-likelihood inference only:
Yes

Perform max-likelihood inference only, including estimating the site allele frequency, testing Hardy-Weinberg equilibrium and testing associations with LRT:
No

History

- Unnamed history
308.4 MB
- 14: MPileup on data 7 and data 10 (log)
- 13: MPileup on data 7 and data 10
- 10: SAM-to-BAM on data 7 and data 9: converted BAM
- 9: Map with BWA for Illumina on data 8 and data 7: mapped reads
- 8: Filter FASTQ on data 3
- 7: http://biocluster.ucr.edu/~nkatiyar/Galaxy%20workshop/Snpseq/tair10chr.fasta
- 5: FastQC FASTQ Groomer on data 1.html
- 4: FASTQ Summary Statistics on data 3
- 3: FASTQ Groomer on data 1
- 1: http://biocluster.ucr.edu%206

Rename history

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](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Convert Formats](#)

[Extract Features](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

[Get Genomic Scores](#)

[Operate on Genomic Intervals](#)

[Statistics](#)

[Wavelet Analysis](#)

[Graph/Display Data](#)

[Regional Variation](#)

[Multiple regression](#)

[Multivariate Analysis](#)

[Evolution](#)

[Motif Tools](#)

[Multiple Alignments](#)

[Metagenomic analyses](#)

UCR | Institute for Integrative Genome Biology

Welcome to IIGB's Galaxy Server!

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 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 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](#).

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.

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 [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 this service can be sent to support@biocluster.ucr.edu.

History

Unnamed history 23.0 MB Click to rename history

2: Sort on data 1

1: UCSC Main on Human: knownGene (genome)

Extract workflow from history

The screenshot shows the Galaxy web interface. On the left is a sidebar with various analysis tools. The main content area features the UCR Institute for Integrative Genome Biology logo and a welcome message. A context menu is open on the right, listing options for managing histories, including "Extract Workflow".

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

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Wavelet Analysis

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

javascript:void(0);

UCR | Institute for Integrative Genome Biology

Welcome to IIGB's Galaxy Server!

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 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 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](#).

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.

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 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 this service can be sent to support@biocluster.ucr.edu.

Histo...

HISTORY LISTS

- Untitled
- 23.0
- 2: Some known**

CURRENT HISTORY

- Create New
- Copy History
- Copy Datasets
- Share or Publish
- Extract Workflow**

Dataset Security

Resume Paused Jobs

Collapse Expanded Datasets

Include Deleted Datasets

Include Hidden Datasets

Unhide Hidden Datasets

Delete Hidden Datasets

Purge Deleted Datasets

Show Structure

Export to File

Delete

Delete Permanently

OTHER ACTIONS

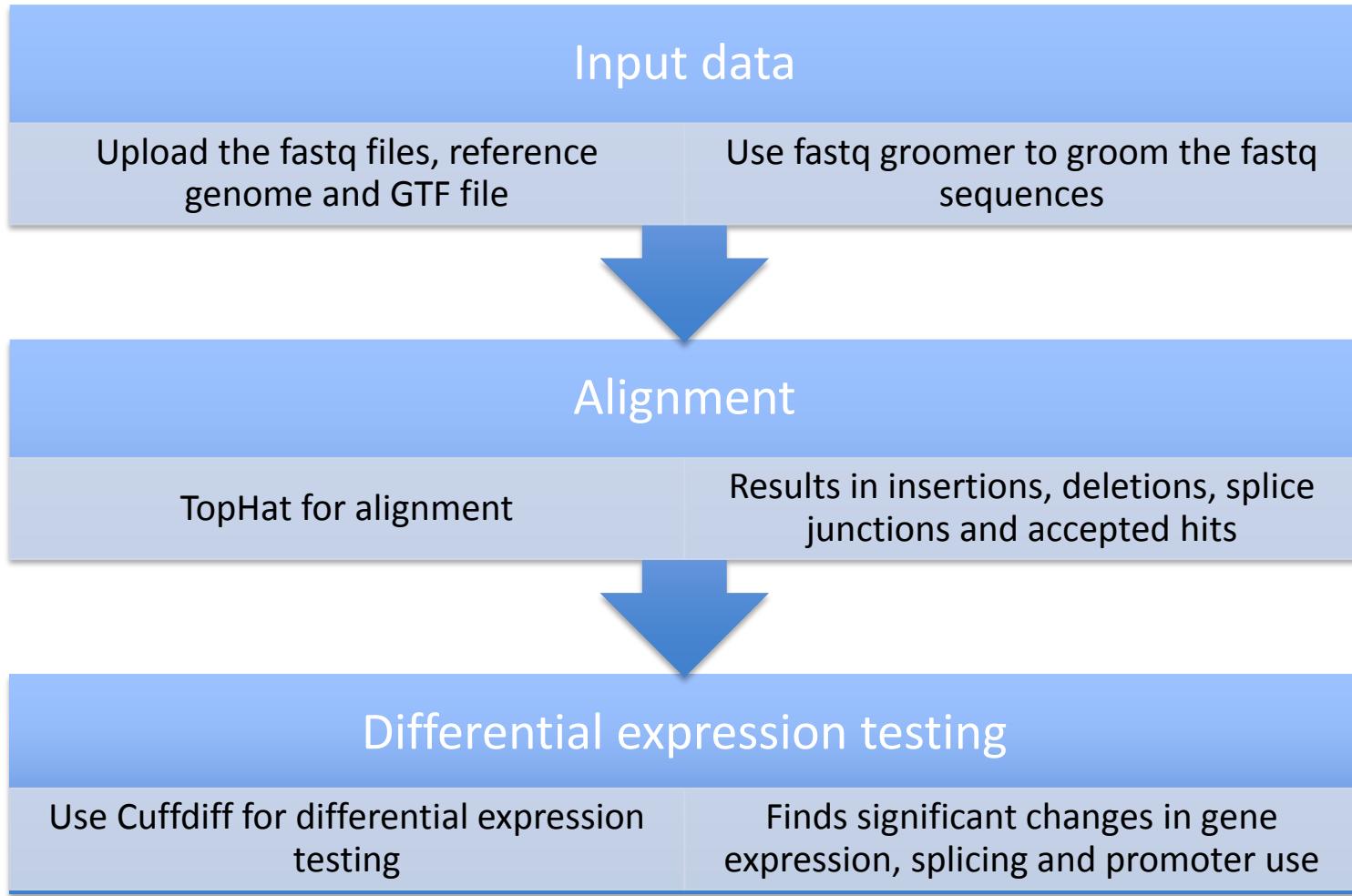
Import from File

Exercise 2: RNASeq Analysis

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

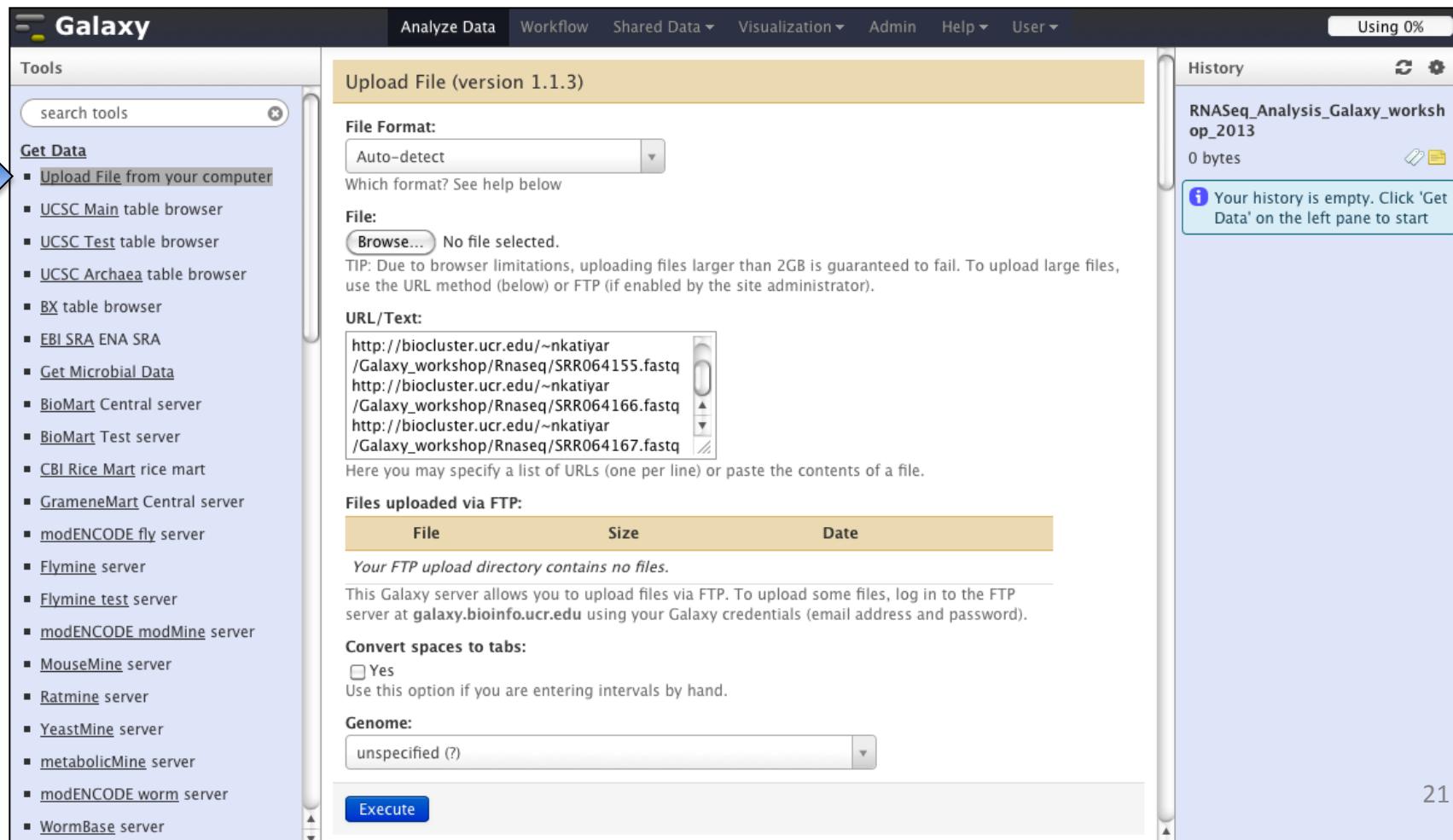
Samples	Factors	Fastq
AP3_f14	AP3	http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064154.fastq
AP3_f14	AP3	http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064155.fastq
T1_f14	TRL	http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064166.fastq
T1_f14	TRL	http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064167.fa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

RNASeq Analysis workflow



Upload Data

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

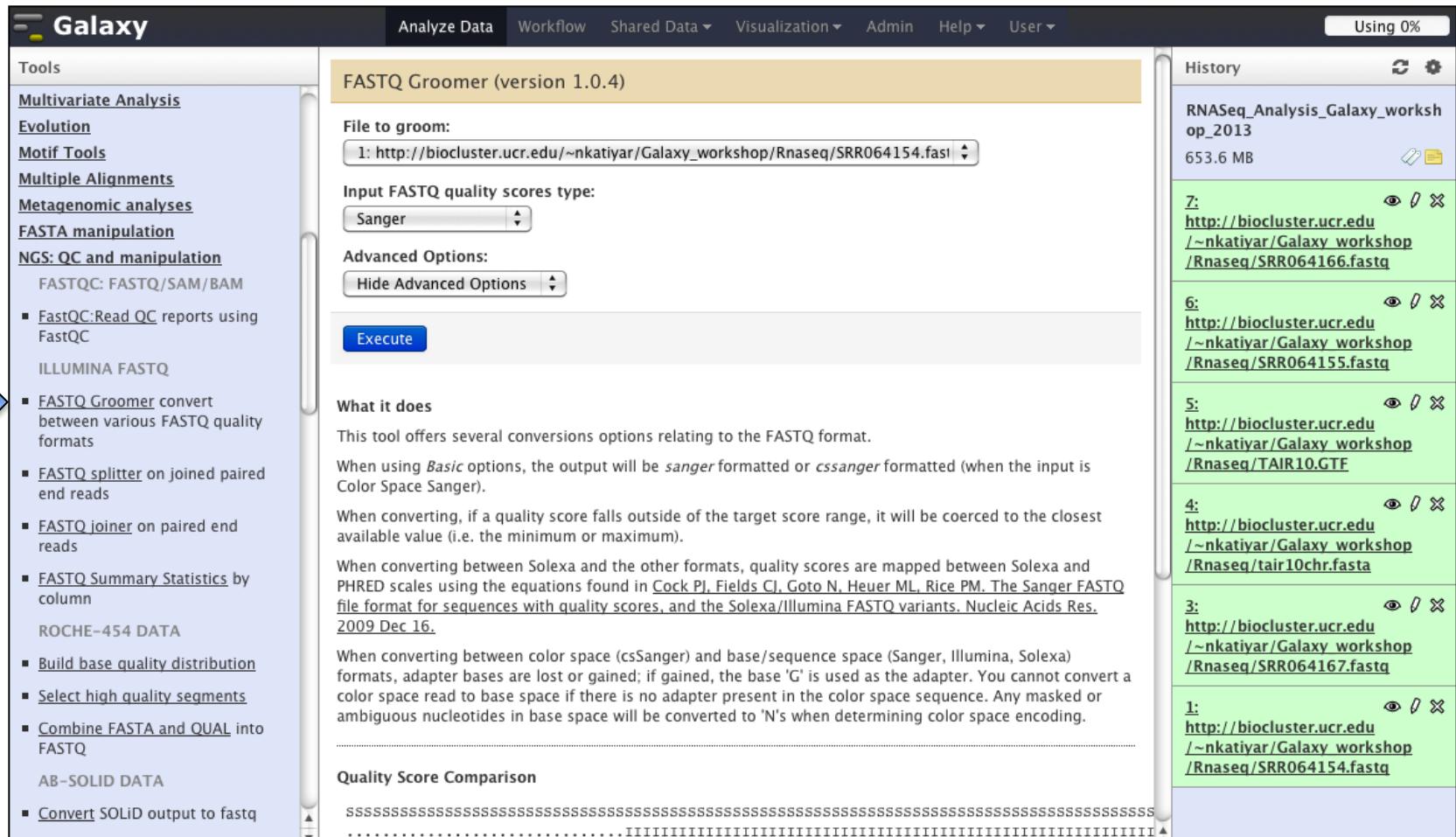


The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.
- Left Sidebar (Tools):** search tools, Get Data (highlighted by a blue arrow), UCSC Main table browser, UCSC Test table browser, UCSC Archaea table browser, BX table browser, EBI SRA ENA SRA, Get Microbial Data, BioMart Central server, BioMart Test server, CBI Rice Mart rice mart, GrameneMart Central server, modENCODE fly server, Flymine server, Flymine test server, modENCODE modMine server, MouseMine server, Ratmine server, YeastMine server, metabolicMine server, modENCODE worm server, WormBase server.
- Middle Panel (Upload File):**
 - File Format:** Auto-detect.
 - File:** No file selected. TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).
 - URL/Text:** A text area containing four URLs:
`http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064155.fastq`
`http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064166.fastq`
`http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064167.fastq`
 - Files uploaded via FTP:** Your FTP upload directory contains no files.
 - Convert spaces to tabs:** Yes (checkbox checked). Use this option if you are entering intervals by hand.
 - Genome:** unspecified (?)
- Right Panel (History):** RNASeq_Analysis_Galaxy_workshop_2013, 0 bytes, Your history is empty. Click 'Get Data' on the left pane to start.

Fastq Groomer

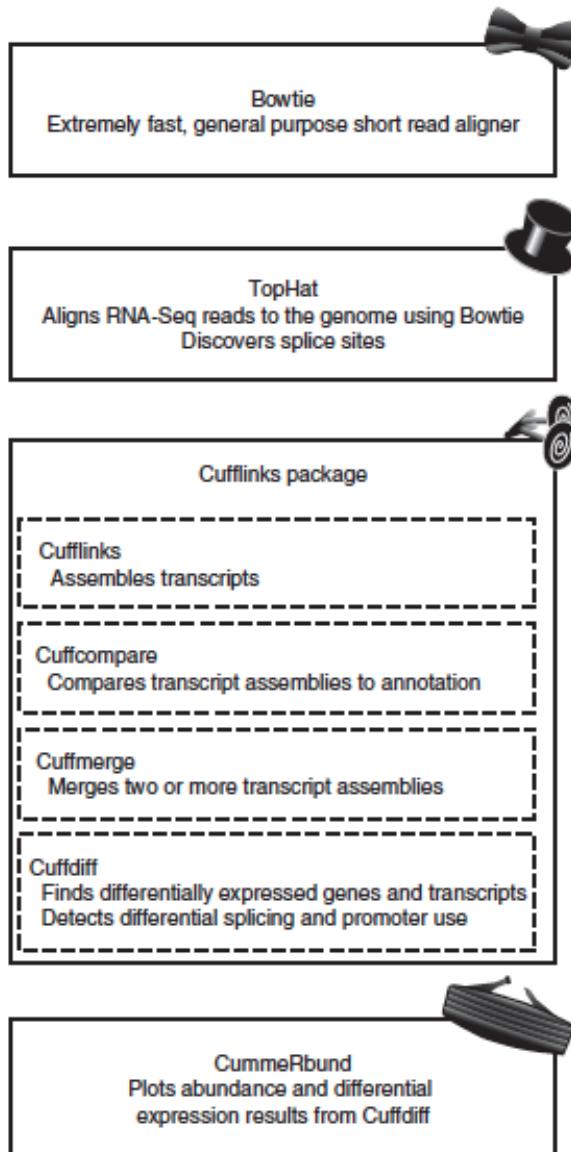
- Select NGS: QC and manipulation and Fastq groomer
- Run Fastq Groomer for all the 4 fastq sequences.



The screenshot shows the Galaxy web interface with the following details:

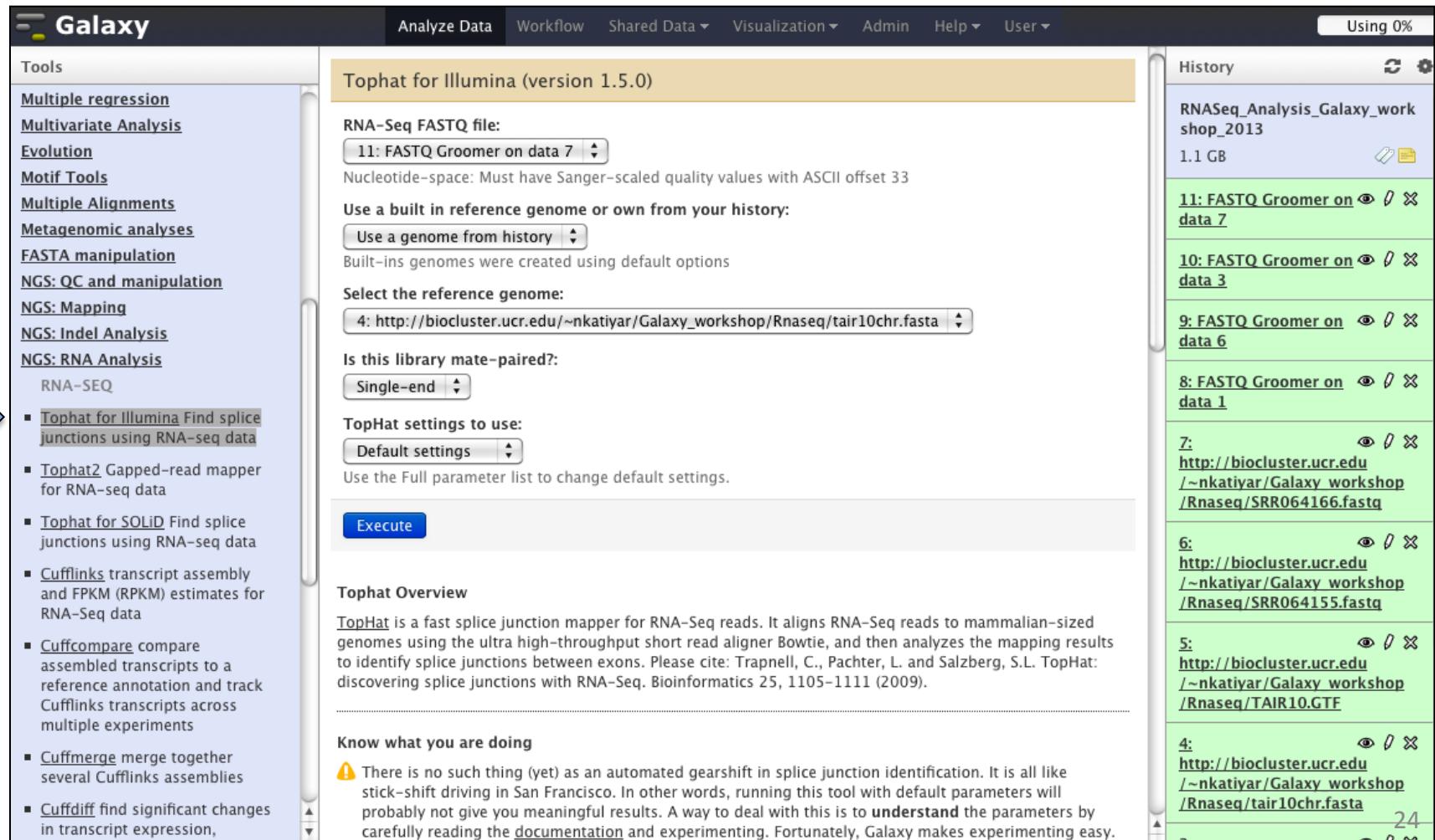
- Header:** Galaxy, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User, Using 0%.
- Tools Sidebar (Left):**
 - Multivariate Analysis
 - Evolution
 - Motif Tools
 - Multiple Alignments
 - Metagenomic analyses
 - FASTA manipulation
 - NGS: QC and manipulation**
 - FASTQC: FASTQ/SAM/BAM
 - FastQC:Read QC reports using FastQC
 - ILLUMINA FASTQ
 - FASTQ Groomer convert between various FASTQ quality formats
 - FASTQ splitter on joined paired end reads
 - FASTQ joiner on paired end reads
 - FASTQ Summary Statistics by column
 - ROCHE-454 DATA
 - Build base quality distribution
 - Select high quality segments
 - Combine FASTA and QUAL into FASTQ
 - AB-SOLID DATA
 - Convert SOLiD output to fastq
- Tool Panel (Center):** FASTQ Groomer (version 1.0.4).
 - File to groom: 1: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064154.fastq
 - Input FASTQ quality scores type: Sanger
 - Advanced Options: Hide Advanced Options
 - Execute button
- Description Panel (Bottom Left):**
 - 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 Solexa and PHRED scales using the equations found in [Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2009 Dec 16.](#)
 - When converting between color space (csSanger) and base/sequence space (Sanger, Illumina, Solexa) formats, adapter bases are lost or gained; if gained, the base 'G' is used as the adapter. You cannot convert a color space read to base space if there is no adapter present in the color space sequence. Any masked or ambiguous nucleotides in base space will be converted to 'N's when determining color space encoding.
- History Panel (Right):** Lists recent analysis workflows.
 - RNASeq_Analysis_Galaxy_workshop_2013 (653.6 MB)
 - 7: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064166.fastq
 - 6: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064155.fastq
 - 5: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/TAIR10.GTF
 - 4: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/tair10chr.fasta
 - 3: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064167.fastq
 - 1: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064154.fastq

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

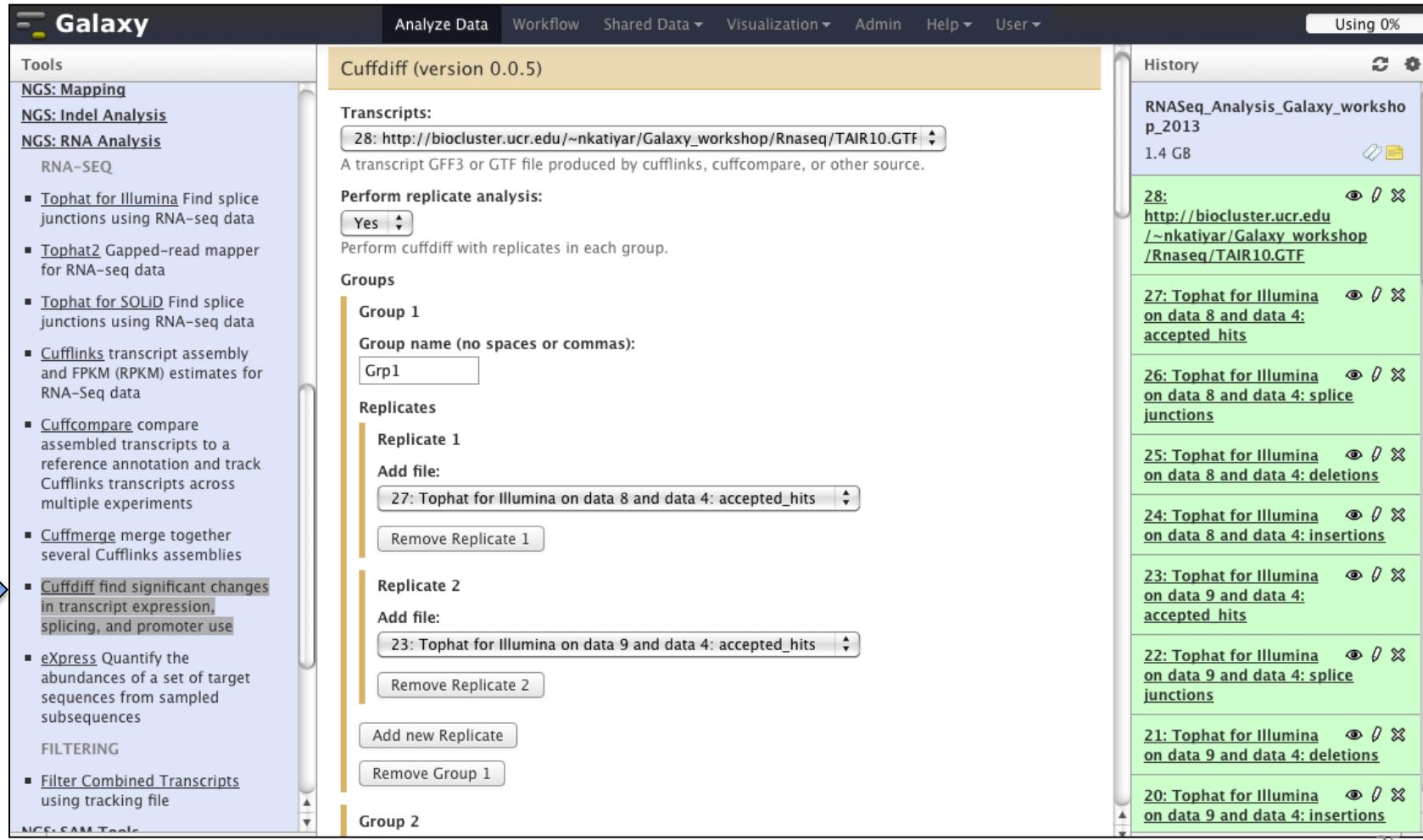


The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User, Using 0%.
- Tools Sidebar (Left):** Multiple regression, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments, Metagenomic analyses, FASTA manipulation, NGS: QC and manipulation, NGS: Mapping, NGS: Indel Analysis, NGS: RNA Analysis, RNA-SEQ:
 - Tophat for Illumina Find splice junctions using RNA-seq data
 - Tophat2 Gapped-read mapper for RNA-seq data
 - Tophat for SOLiD Find splice junctions using RNA-seq data
 - Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
 - Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
 - Cuffmerge merge together several Cufflinks assemblies
 - Cuffdiff find significant changes in transcript expression
- Main Content Area:**
 - Tophat for Illumina (version 1.5.0)**
 - RNA-Seq FASTQ file:** 11: FASTQ Groomer on data 7
 - Nucleotide-space:** Must have Sanger-scaled quality values with ASCII offset 33
 - Use a built in reference genome or own from your history:** Use a genome from history
 - Built-ins genomes were created using default options**
 - Select the reference genome:** 4: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/tair10chr.fasta
 - Is this library mate-paired?** Single-end
 - TopHat settings to use:** Default settings
 - Execute** button
- Bottom Information:**
 - Tophat Overview:** Tophat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Please cite: Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111 (2009).
 - Know what you are doing:** A warning message: There is no such thing (yet) as an automated gearshift in splice junction identification. It is all like stick-shift driving in San Francisco. In other words, running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to understand the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.
- History Panel (Right):** Lists previous workflow steps:
 - 11: FASTQ Groomer on data 7
 - 10: FASTQ Groomer on data 3
 - 9: FASTQ Groomer on data 6
 - 8: FASTQ Groomer on data 1
 - 7: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064166.fasta
 - 6: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064155.fasta
 - 5: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/TAIR10.GTF
 - 4: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/tair10chr.fasta

Find Significant Changes

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



The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy, Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.
- Left Sidebar (Tools):**
 - NGS: Mapping**
 - NGS: Indel Analysis**
 - NGS: RNA Analysis**
 - RNA-SEQ**
 - Tophat for Illumina** Find splice junctions using RNA-seq data
 - Tophat2** Gapped-read mapper for RNA-seq data
 - Tophat for SOLiD** Find splice junctions using RNA-seq data
 - Cufflinks** transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
 - Cuffcompare** compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
 - Cuffmerge** merge together several Cufflinks assemblies
 - Cuffdiff** find significant changes in transcript expression, splicing, and promoter use
 - eXpress** Quantify the abundances of a set of target sequences from sampled subsequences
 - FILTERING**
 - Filter Combined Transcripts** using tracking file
- Main Panel (Cuffdiff version 0.0.5):**
 - Transcripts:** 28: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/TAIR10.GTF
 - A transcript GFF3 or GTF file produced by cufflinks, cuffcompare, or other source.
 - Perform replicate analysis:** Yes
 - Perform cuffdiff with replicates in each group.
 - Groups:**
 - Group 1**
 - Group name (no spaces or commas):** Grp1
 - Replicates**
 - Replicate 1**
 - Add file:** 27: Tophat for Illumina on data 8 and data 4: accepted_hits
 - Remove Replicate 1**
 - Replicate 2**
 - Add file:** 23: Tophat for Illumina on data 9 and data 4: accepted_hits
 - Remove Replicate 2**
 - Add new Replicate**
 - Remove Group 1**
 - Group 2**
- History:**
 - 28: http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/TAIR10.GTF
 - 27: Tophat for Illumina on data 8 and data 4: accepted hits
 - 26: Tophat for Illumina on data 8 and data 4: splice junctions
 - 25: Tophat for Illumina on data 8 and data 4: deletions
 - 24: Tophat for Illumina on data 8 and data 4: insertions
 - 23: Tophat for Illumina on data 9 and data 4: accepted hits
 - 22: Tophat for Illumina on data 9 and data 4: splice junctions
 - 21: Tophat for Illumina on data 9 and data 4: deletions
 - 20: Tophat for Illumina on data 9 and data 4: insertions

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

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](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Convert Formats](#)

[Extract Features](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

[Get Genomic Scores](#)

[Operate on Genomic Intervals](#)

[Statistics](#)

[Wavelet Analysis](#)

[Graph/Display Data](#)

[Regional Variation](#)

[Multiple regression](#)

[Multivariate Analysis](#)

[Evolution](#)

[Motif Tools](#)

[Multiple Alignments](#)

[Metagenomic analyses](#)

[FASTA manipulation](#)

[NGS: QC and manipulation](#)

[NCS: Mapping](#)

test_id	gene_id	gene	locus	sample_1	sample_2	status
AT1G01010.1	AT1G01010	ANAC001	Chr1:3630-5899	Grp1	Grp2	OK
AT1G01020.1	AT1G01020	ARV1	Chr1:5927-8737	Grp1	Grp2	NOTEST
AT1G01020.2	AT1G01020	ARV1	Chr1:5927-8737	Grp1	Grp2	OK
AT1G01030.1	AT1G01030	NGA3	Chr1:11648-13714	Grp1	Grp2	NOTEST
AT1G01040.1	AT1G01040	DCL1	Chr1:23145-33153	Grp1	Grp2	OK
AT1G01040.2	AT1G01040	DCL1	Chr1:23145-33153	Grp1	Grp2	OK
AT1G01046.1	AT1G01046	MIR838A	Chr1:23145-33153	Grp1	Grp2	OK
AT1G01050.1	AT1G01050	AtPPa1	Chr1:23145-33153	Grp1	Grp2	OK
AT1G01060.1	AT1G01060	LHY	Chr1:33378-37871	Grp1	Grp2	OK
AT1G01060.2	AT1G01060	LHY	Chr1:33378-37871	Grp1	Grp2	NOTEST
AT1G01060.3	AT1G01060	LHY	Chr1:33378-37871	Grp1	Grp2	OK
AT1G01060.4	AT1G01060	LHY	Chr1:33378-37871	Grp1	Grp2	OK
AT1G01060.5	AT1G01060	LHY	Chr1:33378-37871	Grp1	Grp2	NOTEST
AT1G01070.1	AT1G01070	AT1G01070	Chr1:38751-40944	Grp1	Grp2	OK
AT1G01070.2	AT1G01070	AT1G01070	Chr1:38751-40944	Grp1	Grp2	NOTEST
AT1G01073.1	AT1G01073	AT1G01073	Chr1:44676-44787	Grp1	Grp2	NOTEST
AT1G01080.1	AT1G01080	AT1G01080	Chr1:45295-47019	Grp1	Grp2	OK
AT1G01080.2	AT1G01080	AT1G01080	Chr1:45295-47019	Grp1	Grp2	NOTEST
AT1G01090.1	AT1G01090	PDH-E1 ALPHA	Chr1:47484-49286	Grp1	Grp2	OK
AT1G01100.1	AT1G01100	AT1G01100	Chr1:50074-51199	Grp1	Grp2	OK
AT1G01100.2	AT1G01100	AT1G01100	Chr1:50074-51199	Grp1	Grp2	OK
AT1G01100.3	AT1G01100	AT1G01100	Chr1:50074-51199	Grp1	Grp2	OK
AT1G01100.4	AT1G01100	AT1G01100	Chr1:50074-51199	Grp1	Grp2	OK
AT1G01110.1	AT1G01110	IQD18	Chr1:52238-54692	Grp1	Grp2	OK
AT1G01110.2	AT1G01110	IQD18	Chr1:52238-54692	Grp1	Grp2	OK
AT1G01115.1	AT1G01115	AT1G01115	Chr1:56623-56740	Grp1	Grp2	NOTEST
AT1G01120.1	AT1G01120	KCS1	Chr1:57268-59167	Grp1	Grp2	OK
AT1G01130.1	AT1G01130	AT1G01130	Chr1:61904-63811	Grp1	Grp2	OK
AT1G01140.1	AT1G01140	CIPK9	Chr1:64165-67625	Grp1	Grp2	OK
AT1G01140.2	AT1G01140	CIPK9	Chr1:64165-67625	Grp1	Grp2	NOTEST
AT1G01140.3	AT1G01140	CIPK9	Chr1:64165-67625	Grp1	Grp2	NOTEST

History

RNASeq_Analysis_Galaxy_workspace_2013
1.4 GB

39: Cuffdiff on data 19, data 15, and others: transcript FPKM tracking

38: Cuffdiff on data 19, data 15, and others: transcript differential expression testing
41,622 lines
format: tabular, database: ?
cuffdiff v2.1.1 (4046M) cuffdiff
--no-update-check -q --library-norm-method geometric
--dispersion-method pooled -p 4 -c 10 --FDR 0.050000 --labels "Grp1","Grp2" /home/galaxy/galaxy-dist/database/files/000/dataset_297.dat /home/galaxy/galaxy-dist/databases

1 2 3 4

test_id	gene_id	gene	locus
AT1G01010.1	AT1G01010	ANAC001	Chr1:3630-5899
AT1G01020.1	AT1G01020	ARV1	Chr1:5927-8737
AT1G01020.2	AT1G01020	ARV1	Chr1:5927-8737
AT1G01030.1	AT1G01030	NGA3	Chr1:11648-13714
AT1G01040.1	AT1G01040	DCL1	Chr1:23145-33153
AT1G01040.2	AT1G01040	DCL1	Chr1:23145-33153
AT1G01046.1	AT1G01046	MIR838A	Chr1:23145-33153
AT1G01050.1	AT1G01050	AtPPa1	Chr1:23145-33153
AT1G01060.1	AT1G01060	LHY	Chr1:33378-37871
AT1G01060.2	AT1G01060	LHY	Chr1:33378-37871
AT1G01060.3	AT1G01060	LHY	Chr1:33378-37871
AT1G01060.4	AT1G01060	LHY	Chr1:33378-37871
AT1G01060.5	AT1G01060	LHY	Chr1:33378-37871
AT1G01070.1	AT1G01070	AT1G01070	Chr1:38751-40944
AT1G01070.2	AT1G01070	AT1G01070	Chr1:38751-40944
AT1G01073.1	AT1G01073	AT1G01073	Chr1:44676-44787
AT1G01080.1	AT1G01080	AT1G01080	Chr1:45295-47019
AT1G01080.2	AT1G01080	AT1G01080	Chr1:45295-47019
AT1G01090.1	AT1G01090	PDH-E1 ALPHA	Chr1:47484-49286
AT1G01100.1	AT1G01100	AT1G01100	Chr1:50074-51199
AT1G01100.2	AT1G01100	AT1G01100	Chr1:50074-51199
AT1G01100.3	AT1G01100	AT1G01100	Chr1:50074-51199
AT1G01100.4	AT1G01100	AT1G01100	Chr1:50074-51199
AT1G01110.1	AT1G01110	IQD18	Chr1:52238-54692
AT1G01110.2	AT1G01110	IQD18	Chr1:52238-54692
AT1G01115.1	AT1G01115	AT1G01115	Chr1:56623-56740
AT1G01120.1	AT1G01120	KCS1	Chr1:57268-59167
AT1G01130.1	AT1G01130	AT1G01130	Chr1:61904-63811
AT1G01140.1	AT1G01140	CIPK9	Chr1:64165-67625
AT1G01140.2	AT1G01140	CIPK9	Chr1:64165-67625
AT1G01140.3	AT1G01140	CIPK9	Chr1:64165-67625

Output File from Galaxy

- SNPseq
 - 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/~nkatiyar/Galaxy_workshop/Snpseq/
- RNASeq
 - Save four Bam files and four .bai files
 - Already saved them at
 - http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/

Downloading bam index file (.bai)

Galaxy

Analyze Data Workflow Shared Data ▾ Visualization ▾ Admin Help ▾ User ▾ Using 0%

History

15: Tophat for Illumina on data 11 and data 4: accepted hits
11.3 MB format: bam, database: ? TopHat v1.3.2 Settings: Output files: "/tmp/tmpm3lcD8 /dataset_273.*.ebwt" Line rate: 6 (line is 64 bytes) Lines per side: 1 (side is 64 bytes) Offset rate: 5 (one in 32) FTable chars: 10 Strings: unpacked Max bucket size: default Max buck

Download Dataset ADDITIONAL FILES Download bam_index

14: Tophat for Illumina on data 11 and data 4: splice junctions

13: Tophat for Illumina on data 11 and data 4: deletions

12: Tophat for Illumina on data 11 and data 4: insertions

11: FASTQ Groomer on data 7

10: FASTQ Groomer on data 3

9: FASTQ Groomer on data 6

8: FASTQ Groomer on data 1

7: http://biocluster.ucr.edu

test_id	gene_id	gene	locus	sample_1	sample_2
AT1G01010.1	AT1G01010	ANAC001	Chr1:3630-5899	Grp1	Grp2
AT1G01020.1	AT1G01020	ARV1	Chr1:5927-8737	Grp1	Grp2
AT1G01020.2	AT1G01020	ARV1	Chr1:5927-8737	Grp1	Grp2
AT1G01030.1	AT1G01030	NGA3	Chr1:11648-13714	Grp1	Grp2
AT1G01040.1	AT1G01040	DCL1	Chr1:23145-33153	Grp1	Grp2
AT1G01040.2	AT1G01040	DCL1	Chr1:23145-33153	Grp1	Grp2
AT1G01046.1	AT1G01046	MIR838A	Chr1:23145-33153	Grp1	Grp2
AT1G01050.1	AT1G01050	AtPPa1	Chr1:23145-33153	Grp1	Grp2
AT1G01060.1	AT1G01060	LHY	Chr1:33378-37871	Grp1	Grp2
AT1G01060.2	AT1G01060	LHY	Chr1:33378-37871	Grp1	Grp2
AT1G01060.3	AT1G01060	LHY	Chr1:33378-37871	Grp1	Grp2
AT1G01060.4	AT1G01060	LHY	Chr1:33378-37871	Grp1	Grp2
AT1G01060.5	AT1G01060	LHY	Chr1:33378-37871	Grp1	Grp2
AT1G01070.1	AT1G01070	AT1G01070	Chr1:38751-40944	Grp1	Grp2
AT1G01070.2	AT1G01070	AT1G01070	Chr1:38751-40944	Grp1	Grp2
AT1G01073.1	AT1G01073	AT1G01073	Chr1:44676-44787	Grp1	Grp2
AT1G01080.1	AT1G01080	AT1G01080	Chr1:45295-47019	Grp1	Grp2
AT1G01080.2	AT1G01080	AT1G01080	Chr1:45295-47019	Grp1	Grp2
AT1G01090.1	AT1G01090	PDH-E1 ALPHA	Chr1:47484-49286	Grp1	Grp2
AT1G01100.1	AT1G01100	AT1G01100	Chr1:50074-51199	Grp1	Grp2
AT1G01100.2	AT1G01100	AT1G01100	Chr1:50074-51199	Grp1	Grp2
AT1G01100.3	AT1G01100	AT1G01100	Chr1:50074-51199	Grp1	Grp2
AT1G01100.4	AT1G01100	AT1G01100	Chr1:50074-51199	Grp1	Grp2
AT1G01110.1	AT1G01110	IQD18	Chr1:52238-54692	Grp1	Grp2
AT1G01110.2	AT1G01110	IQD18	Chr1:52238-54692	Grp1	Grp2
AT1G01115.1	AT1G01115	AT1G01115	Chr1:56623-56740	Grp1	Grp2
AT1G01120.1	AT1G01120	KCS1	Chr1:57268-59167	Grp1	Grp2
AT1G01130.1	AT1G01130	AT1G01130	Chr1:61904-63811	Grp1	Grp2
AT1G01140.1	AT1G01140	CIPK9	Chr1:64165-67625	Grp1	Grp2
AT1G01140.2	AT1G01140	CIPK9	Chr1:64165-67625	Grp1	Grp2
AT1G01140.3	AT1G01140	CIPK9	Chr1:64165-67625	Grp1	Grp2
AT1G01150.1	AT1G01150	AT1G01150	Chr1:70114-72138	Grp1	Grp2
AT1G01160.1	AT1G01160	GIF2	Chr1:72338-74737	Grp1	Grp2
AT1G01160.2	AT1G01160	GIF2	Chr1:72338-74737	Grp1	Grp2
AT1G01170.1	AT1G01170	AT1G01170	Chr1:72338-74737	Grp1	Grp2
AT1G01170.2	AT1G01170	AT1G01170	Chr1:72338-74737	Grp1	Grp2

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

The image shows the official website for the Integrative Genomics Viewer (IGV) on the left and a screenshot of the IGV software interface on the right.

Website (Left):

- Header:** "Integrative Genomics Viewer" with a logo.
- Navigation Bar:** Home, Downloads, Documents, Hosted Genomes, FAQ, IGV User Guide, File Formats, Release Notes, Credits, Contact.
- Search:** Search website input field with a search button.
- Broad Institute:** Broad Home, Cancer Program links, and a logo.
- Footer:** © 2013 Broad Institute.

Software Interface (Right):

- Title:** "Home" and "Integrative Genomics Viewer".
- Content:** A large screenshot of the IGV interface showing multiple tracks of genomic data, including tracks for expression levels, genomic annotations, and sequence variants.
- Section Headers:** What's New, Citing IGV, Overview, Downloads.
- What's New:** October 2013: We're presenting an IGV workshop at the [ASHG annual meeting](#) in Boston on October 25, 2013. April 2013: IGV 2.3 has been released. See the [release notes](#) for more details.
- Citing IGV:** To cite your use of IGV in your publication: Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Briefings in Bioinformatics* 2012; 24–26 (2011).
- Overview:** The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.
- Downloads:** Please [register](#) to download IGV. After registering, you can log in at any time using your email address. Permission to use IGV is granted under the [GNU LGPL license](#).
- Logos:** National Cancer Institute, National Human Genome Research Institute, Broad Institute, GENOME SPACE.

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

IGV download

 Integrative Genomics Viewer ASHG

[Home](#) [Downloads](#) [Documents](#)

- [Hosted Genomes](#)
- [FAQ](#)
- [IGV User Guide](#)
- [File Formats](#)
- [Release Notes](#)
- [Credits](#)

[@ Contact](#)

Search website

[Broad Home](#) [Cancer Program](#)

 © 2013 Broad Institute

Home > Downloads

Downloads

Integrative Genomics Viewer (Version 2.3)

Mac Users: Apple has pushed out an update that blocks all but the latest versions of Java. See [this article](#) for details. To run IGV from the web launch buttons below, you need the [latest version of Java](#). Another option which avoids Mac security issues is to use the "zip" distribution below. After unzipping double-click the "igv.command" file to launch IGV.

Java: IGV 2.3 requires Java 6 or greater. To use the launch buttons below on MacOS Java 7 is required.

Chrome: Chrome does not launch java webstart files by default. Instead, the launch buttons below will download a "jnlp" file. This should appear in the lower left corner of the browser. Double-click the downloaded file to run.

Windows users: To run with more than 1.2 GB you must install 64-bit Java. This is often not installed by default even with the latest Windows 7 machines with many GB of memory. In general trying to launch with more memory than your OS/Java combination supports will result in the obscure error "could not create virtual machine".

 Launch Launch with 750 MB	 Launch Launch with 1.2 GB Maximum usable memory for Windows OS with 32-bit Java.	 Launch Launch with 2 GB Maximum usable memory for 32-bit MacOS.	 Launch Launch with 10 GB For large memory 64-bit java machines.
--	--	---	---

[Nightly Build](#) Latest development build.

[Archived Versions](#)

igvtools

Utilities for preprocessing data files.

- [igvtools_2.3.25.zip](#)

Download

A downloadable version that does not require launching from the web. For Windows, Mac OS X, and Linux.

- [IGV_2.3.25.zip](#)

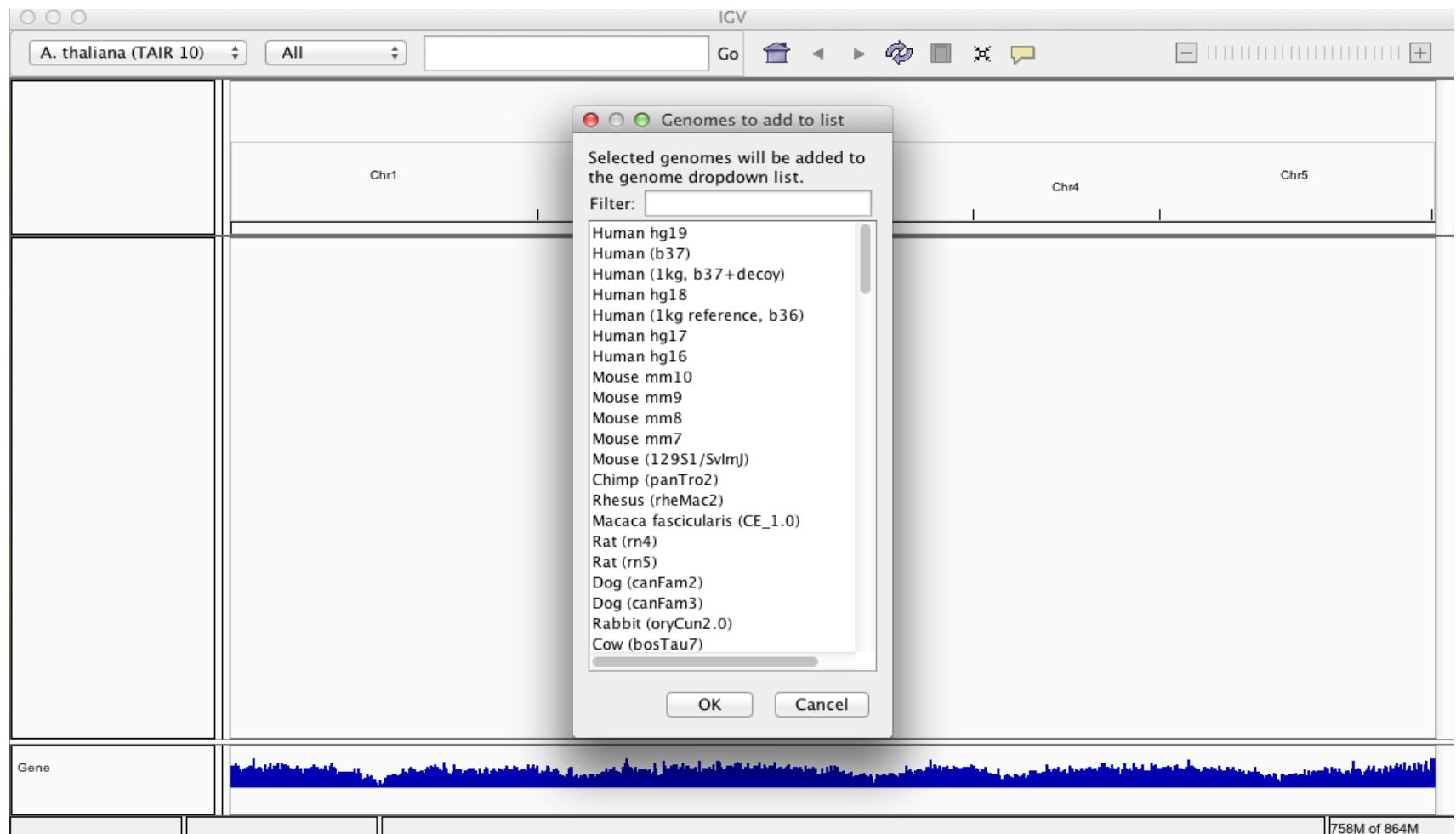
IGV interface



Load data

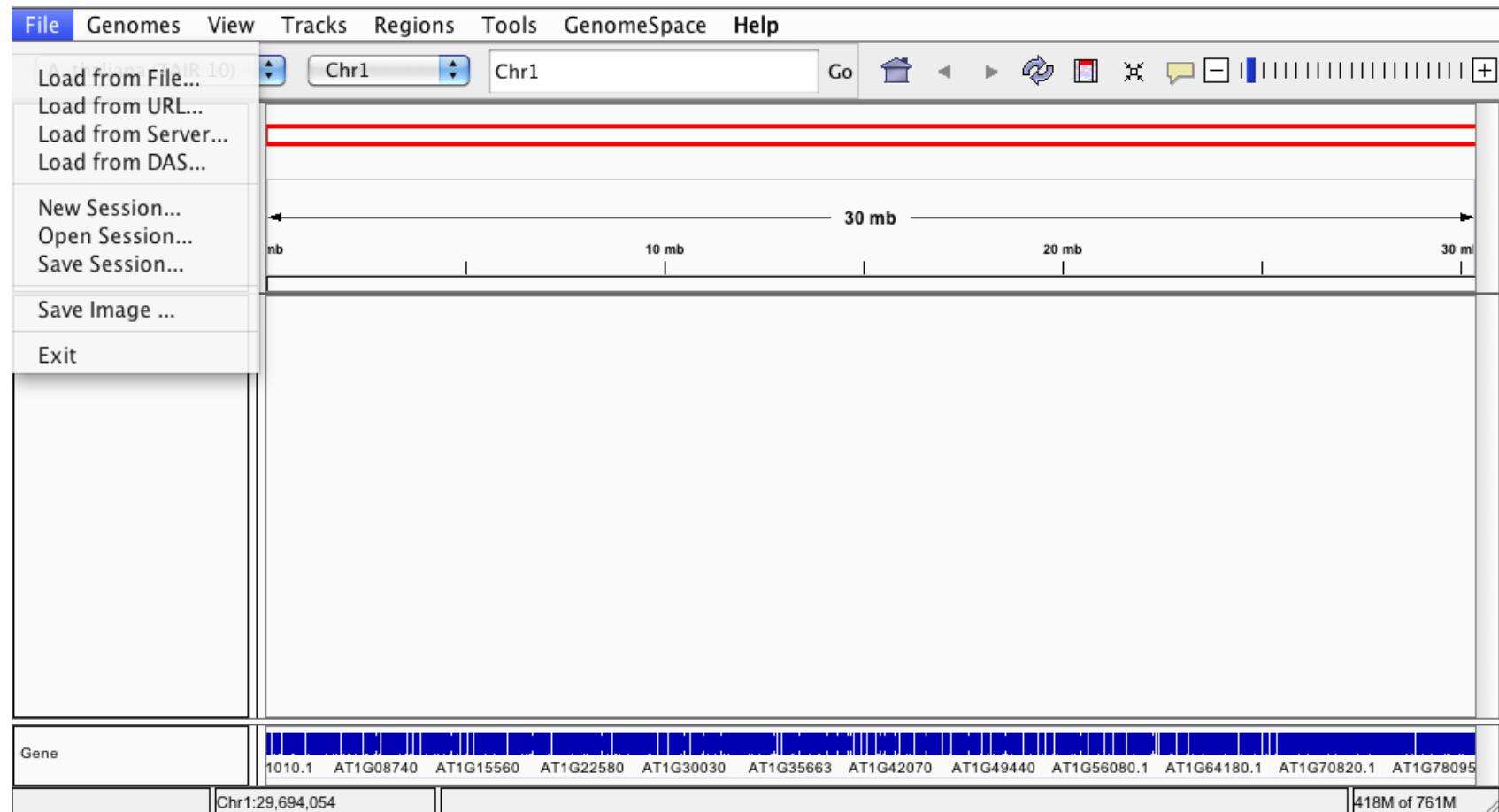
Select genome: Click the genome drop-down list in the toolbar and select the genome

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

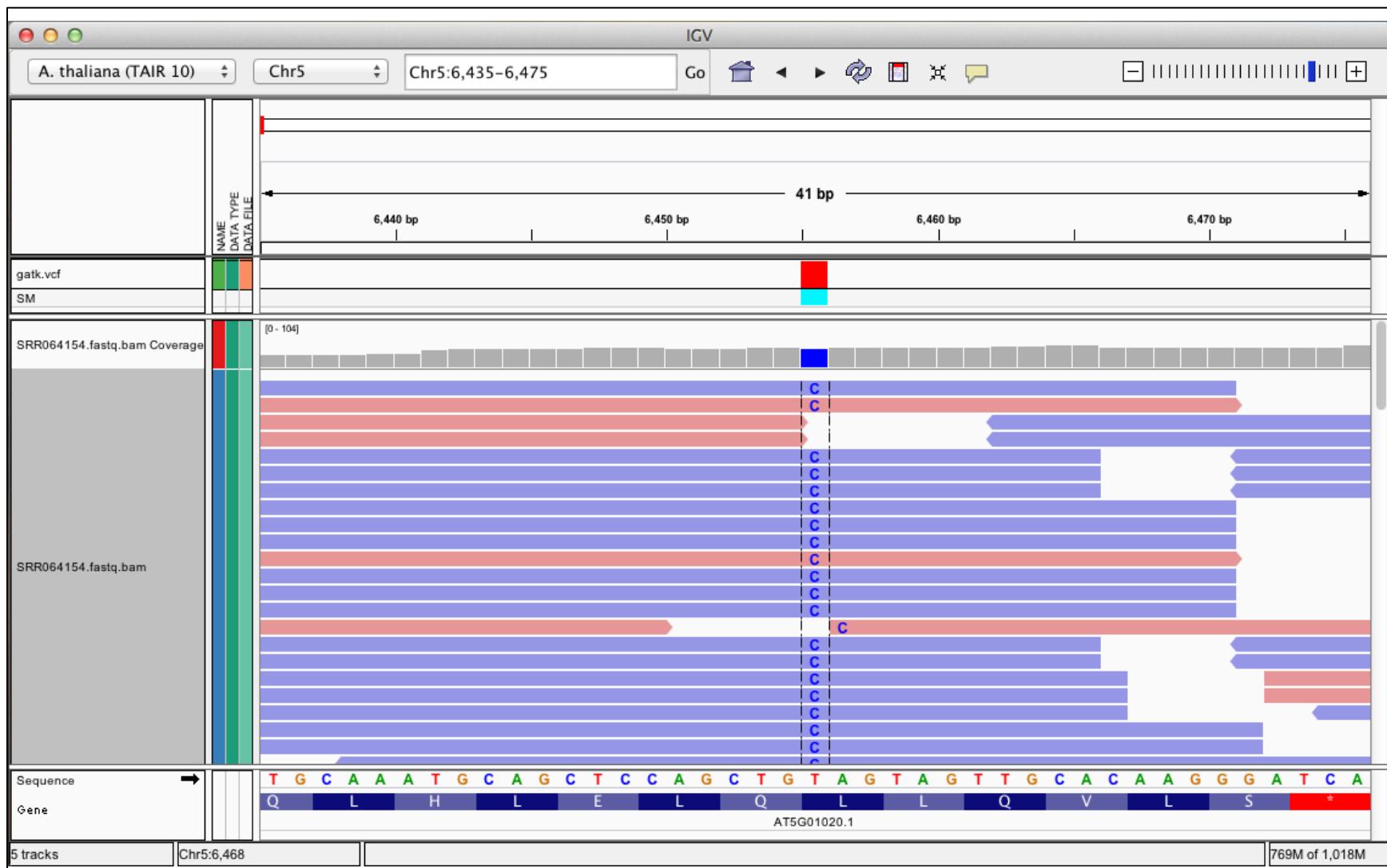
- Load TAIR10 genome to IGV
- Load BAM file “SRR038850.bam” to IGV with “Load from URL”
 - http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/SAM-to-BAM_BAM.bam
- Load VCF file “var.raw.vcf” to IGV with “Load from URL”
 - http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Snpseq/var.raw.vcf

Zoom In Screen

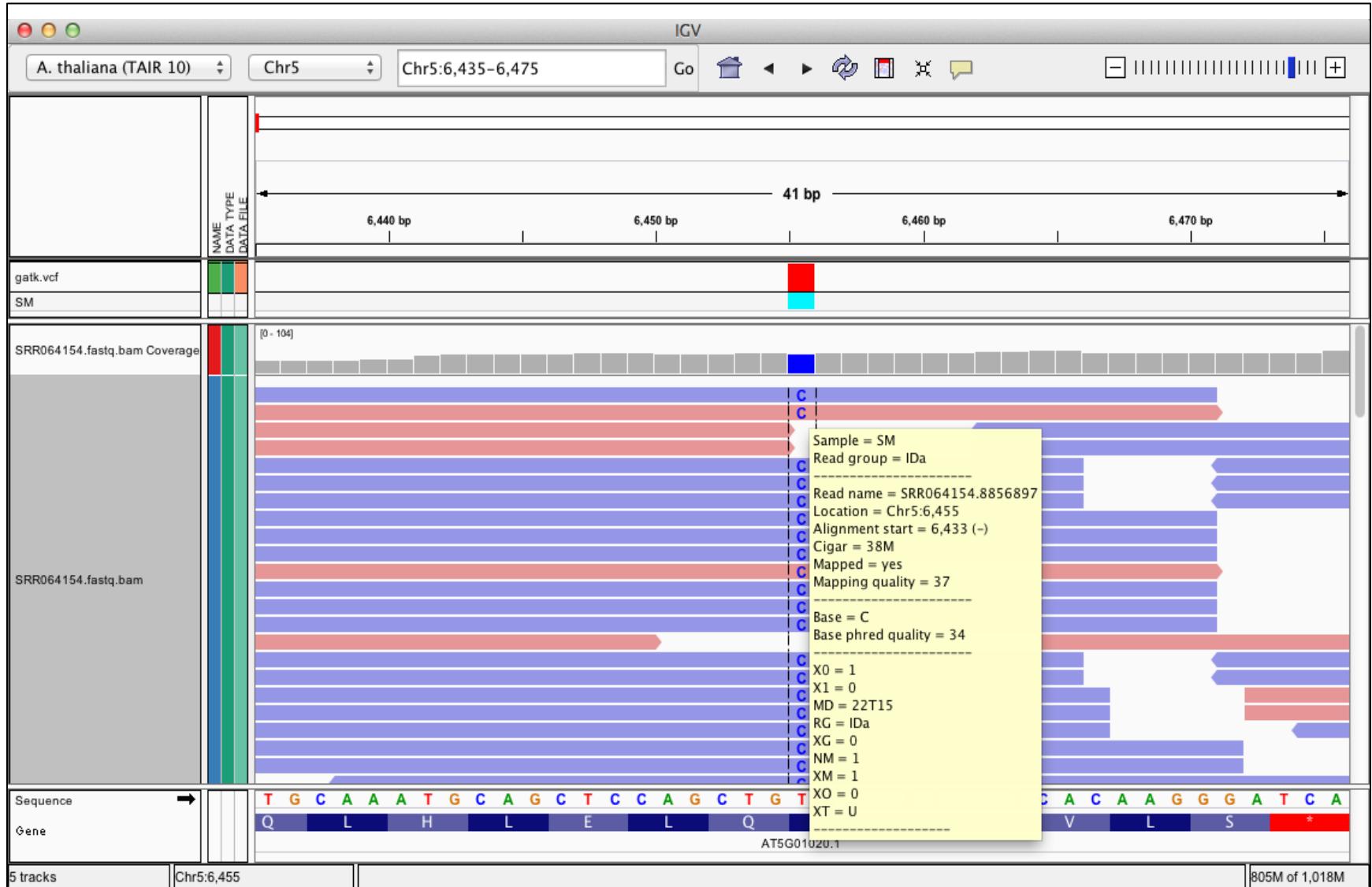
Zoom in to : Chr5:57,073-57,142



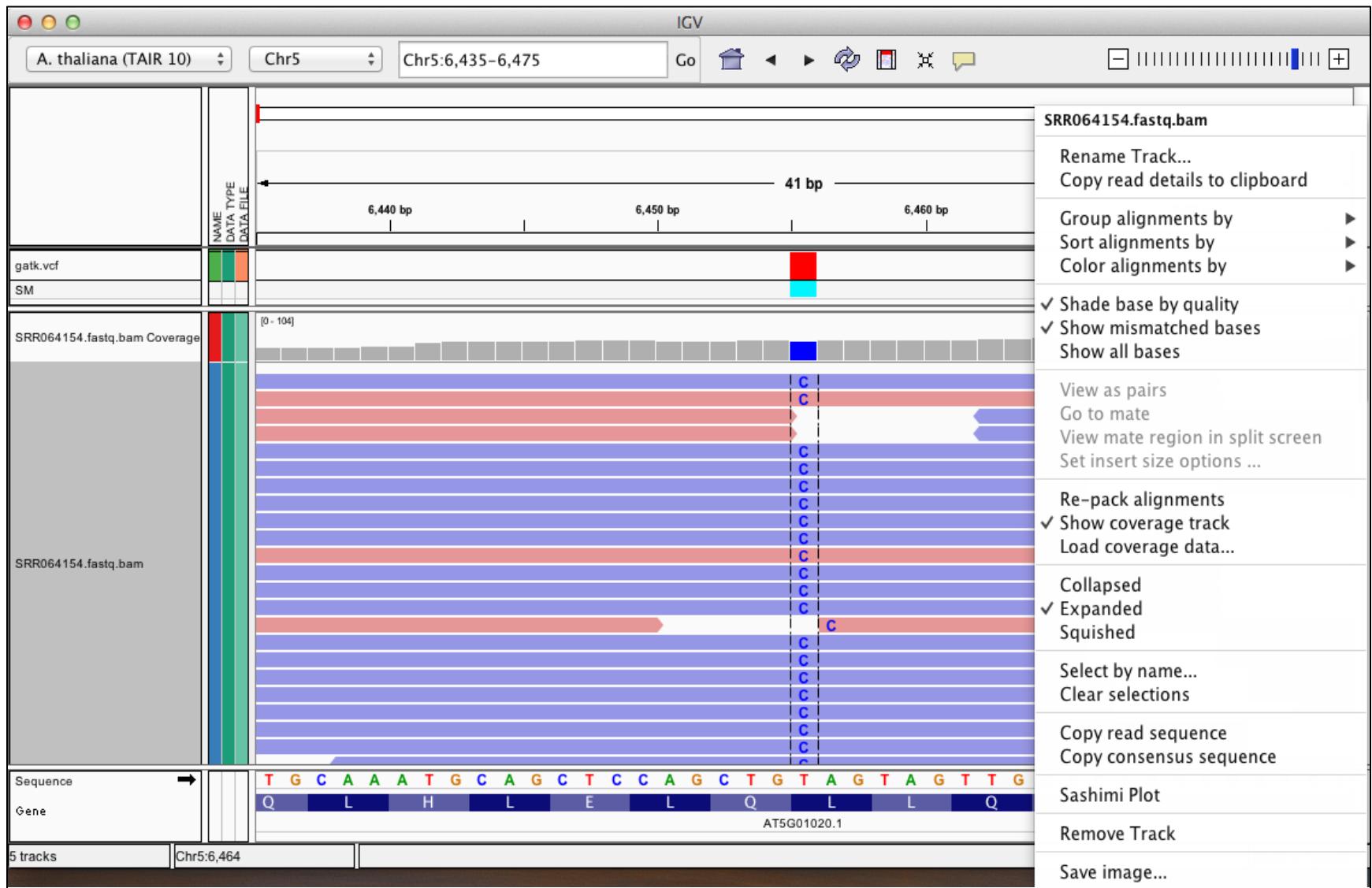
Zoom in position (chr5:6,435-6,475)



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



Right click on track for more options



Show all bases in IGV

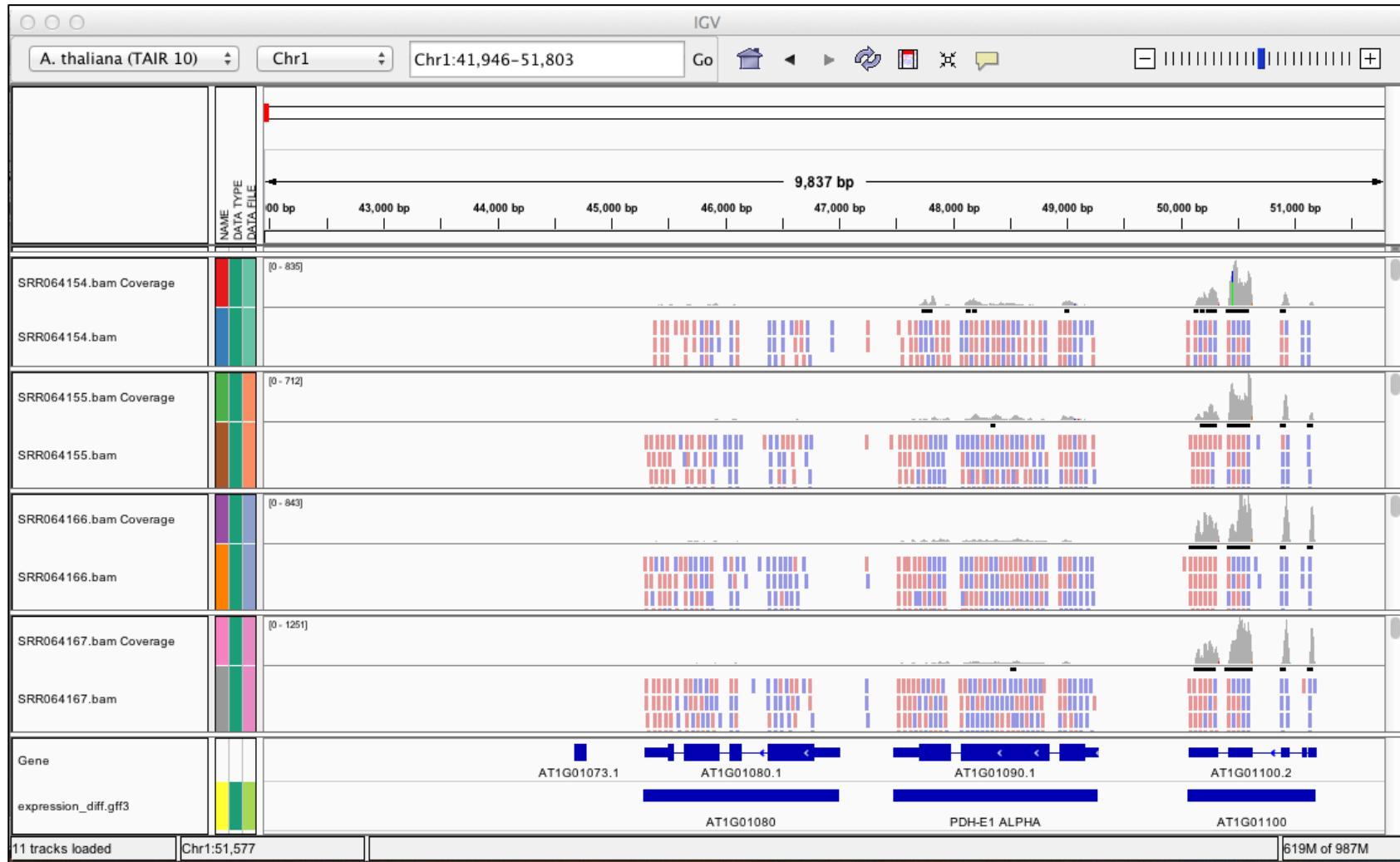


RNAseq Results Visualization

- Load four BAM files to IGV
 - http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064154.bam
 - http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064155.bam
 - http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064166.bam
 - http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/SRR064167.bam
- Load gene differential expression GFF3 file “expression_diff.gff3” to IGV
 - http://biocluster.ucr.edu/~nkatiyar/Galaxy_workshop/Rnaseq/expression_diff.gff3

Exercise2: RNAseq Result Visualization

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



Exercise2: RNAseq Result Visualization

Zoom in Chr1:49,457-51,457

