
# RNA-Seq Data Analysis using Galaxy

RNASeq Data Analysis Bootcamp Week4

**Alper Kucukural**
**BioCore**
**03/04/2014**

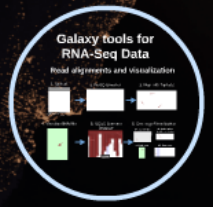


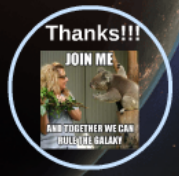

# RNA-Seq Data Analysis using Galaxy

Alper Kucukural

BioCore

03/04/2014

# What is Galaxy?

**Galaxy is**

- an integrated tool management system with a user-friendly (GUI).
- It is designed for running multiple bioinformatics tools.

## Galaxy Basics

- Data upload.
- Grooming the data.
- FastQC Quality Check
- Filtering FastQ files

## Galaxy RNASeq Tools

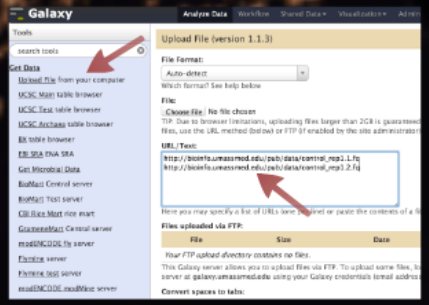
- Read alignments (Tophat2)
- UCSC Genome Browser
- Coverage Calculation
- BedGraph to bigWig

# Galaxy Basics

## Galaxy Overview

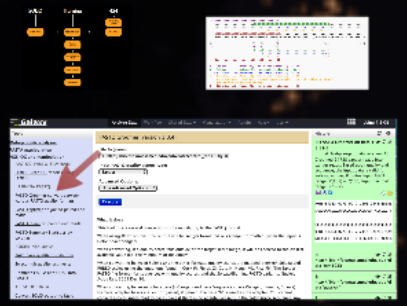
galaxy.umassmed.edu or
usegalaxy.org


## Data Upload



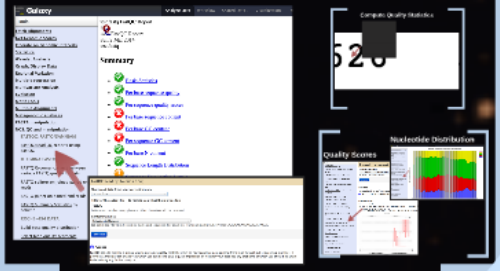

## Grooming the Data

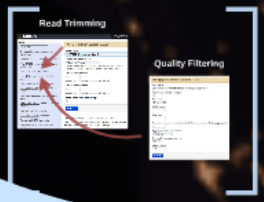
When your read qualities are not formatted for galaxy...

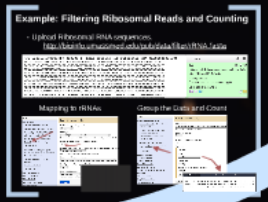


## FASTQC and other QC

Upload this file to your galaxy;
http://bioinfo.umassmed.edu/pub/data/test.fastq

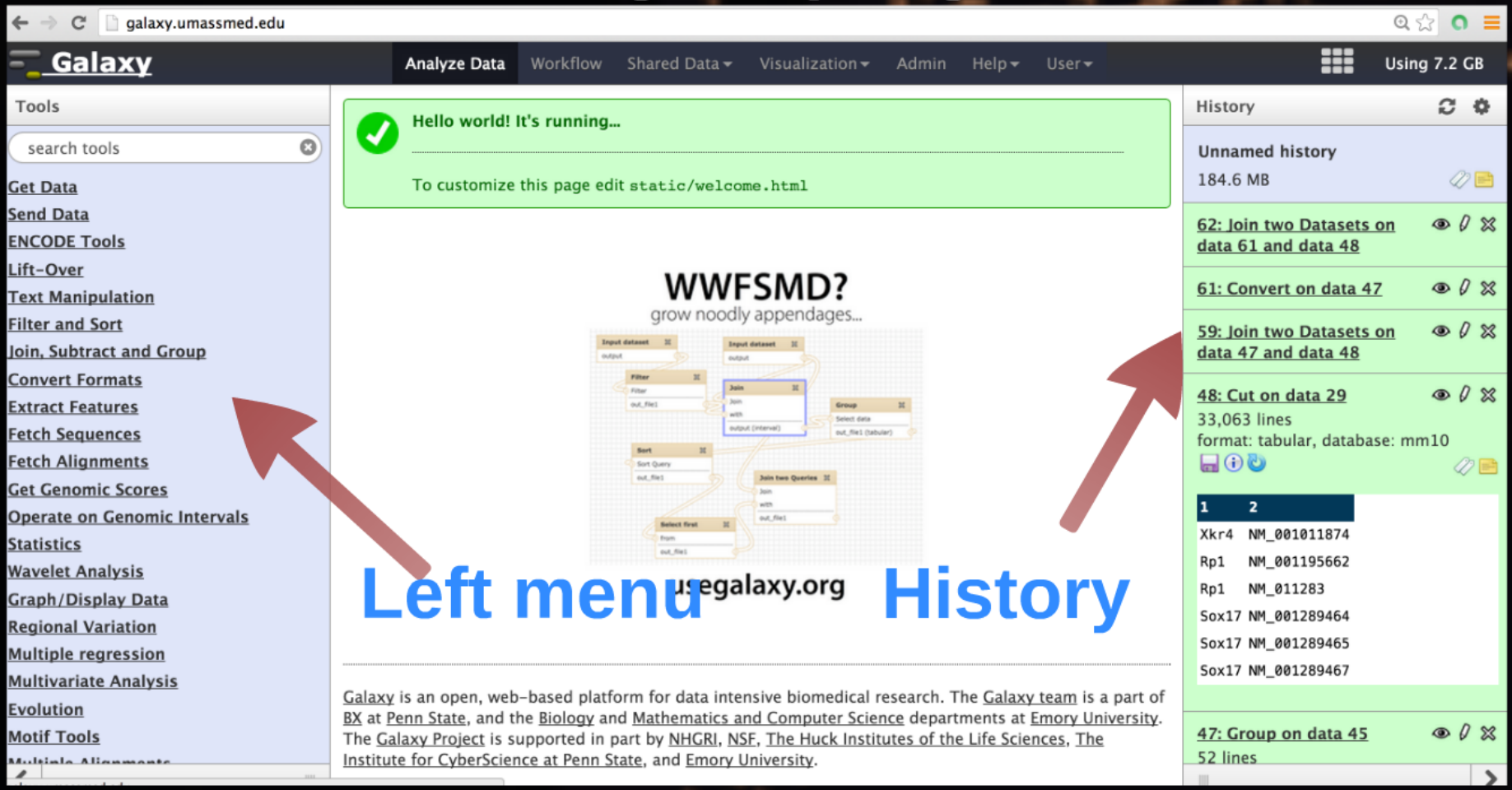


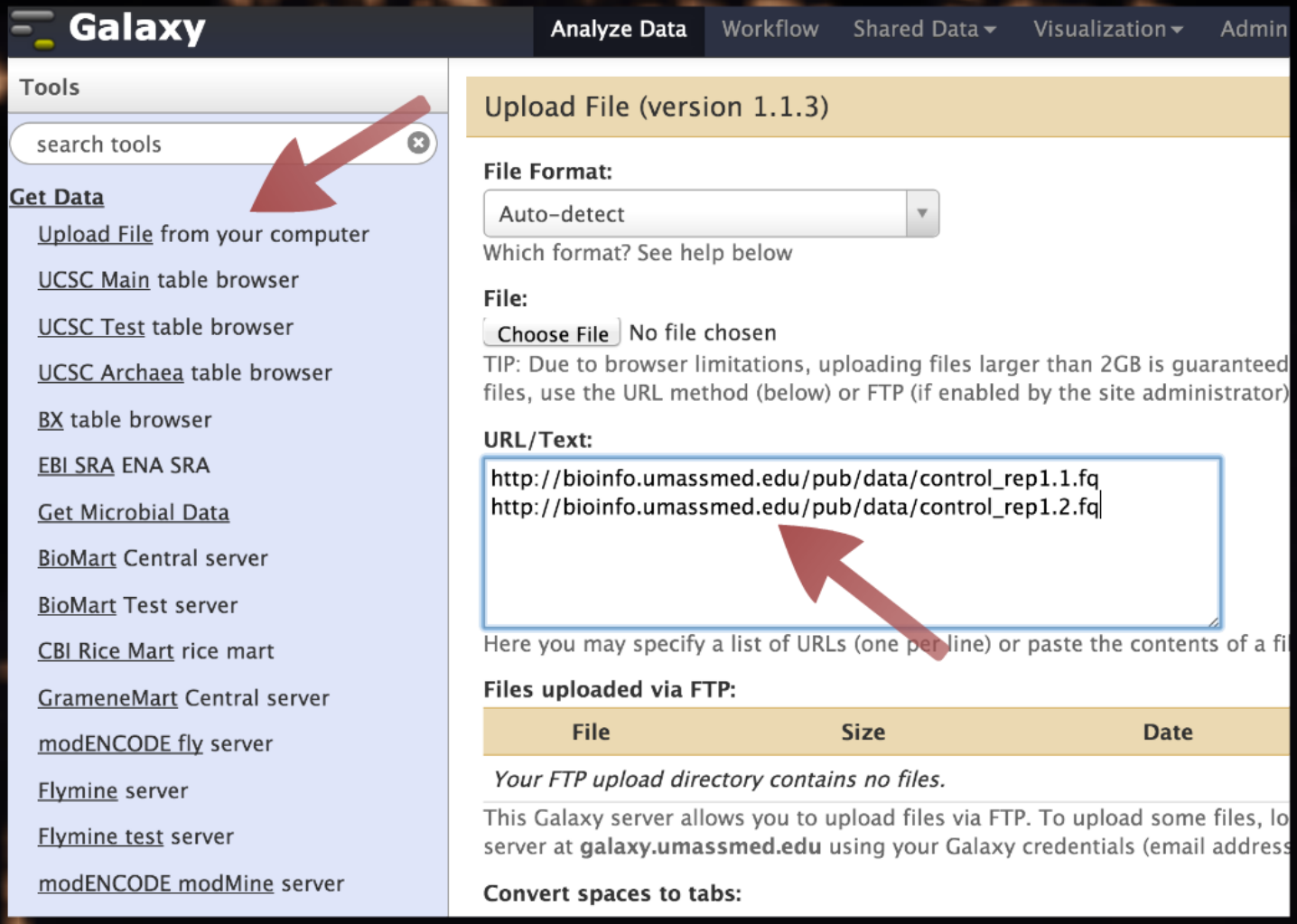



# Galaxy Overview

## galaxy.umassmed.edu or usegalaxy.org

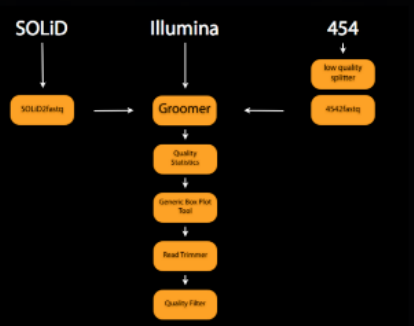


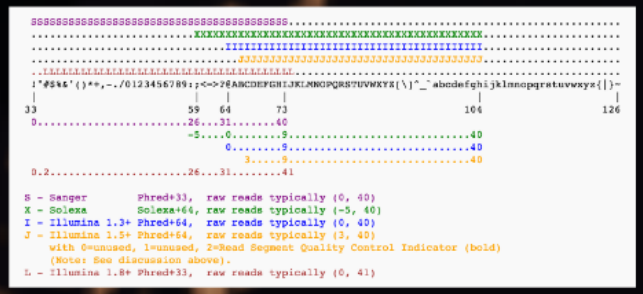
# Data Upload

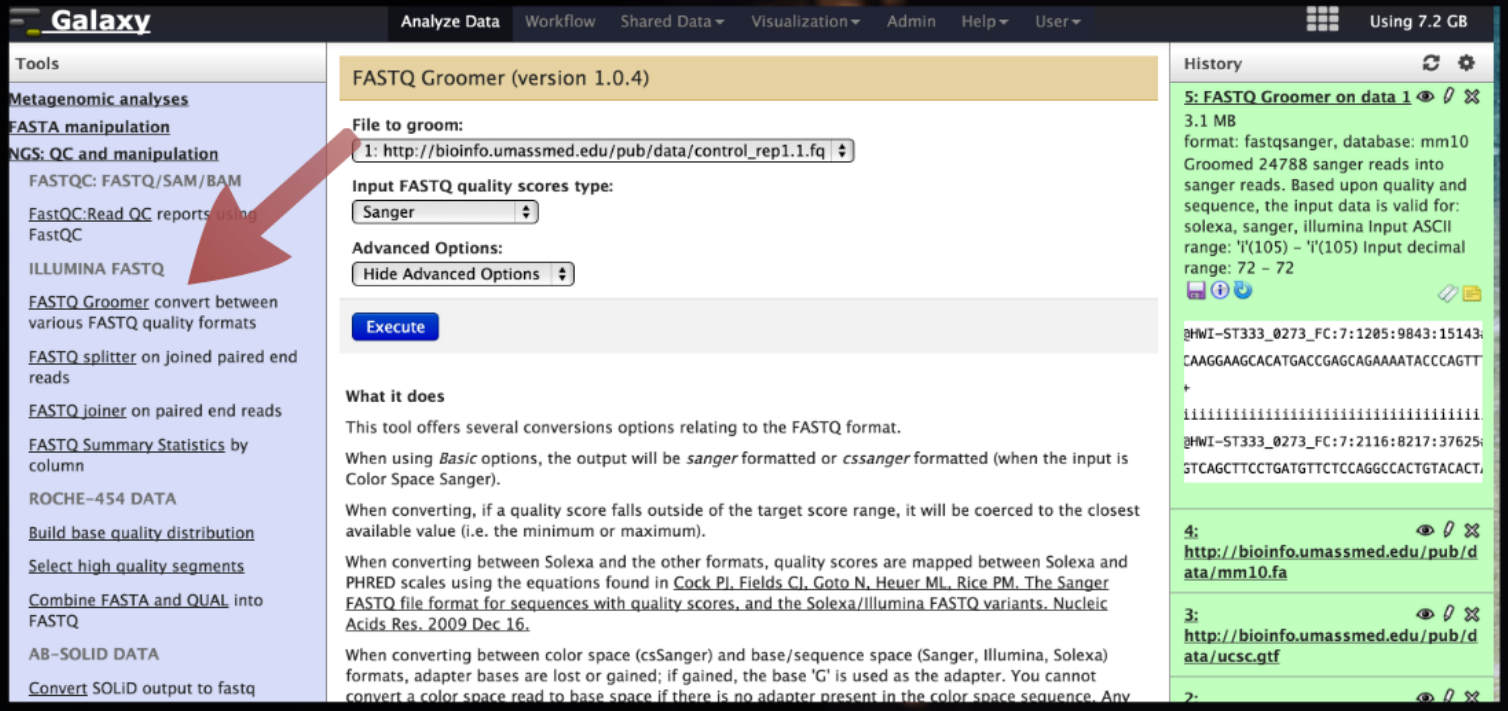

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Tools
search tools
Get Data
Upload File from your computer
UCSC Main table browser
UCSC Test table browser
UCSC Archaea table browser
BX table browser
EBI SRA ENA SRA
Get Microbial Data
BioMart Central server
BioMart Test server
CBI Rice Mart rice mart
GrameneMart Central server
modENCODE fly server
Flymine server
Flymine test server
modENCODE modMine server
Upload File (version 1.1.3)
File Format:
Auto-detect
Which format? See help below
File:
Choose File
No file chosen
TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed
files, use the URL method (below) or FTP (if enabled by the site administrator)
URL/Text:
http://bioinfo.umassmed.edu/pub/data/control_rep1.1.fq
http://bioinfo.umassmed.edu/pub/data/control_rep1.2.fq
Here you may specify a list of URLs (one per line) or paste the contents of a fi
Files uploaded via FTP:
File
Size
Date
Your FTP upload directory contains no files.
This Galaxy server allows you to upload files via FTP. To upload some files, lo
server at galaxy.umassmed.edu using your Galaxy credentials (email address
Convert spaces to tabs:

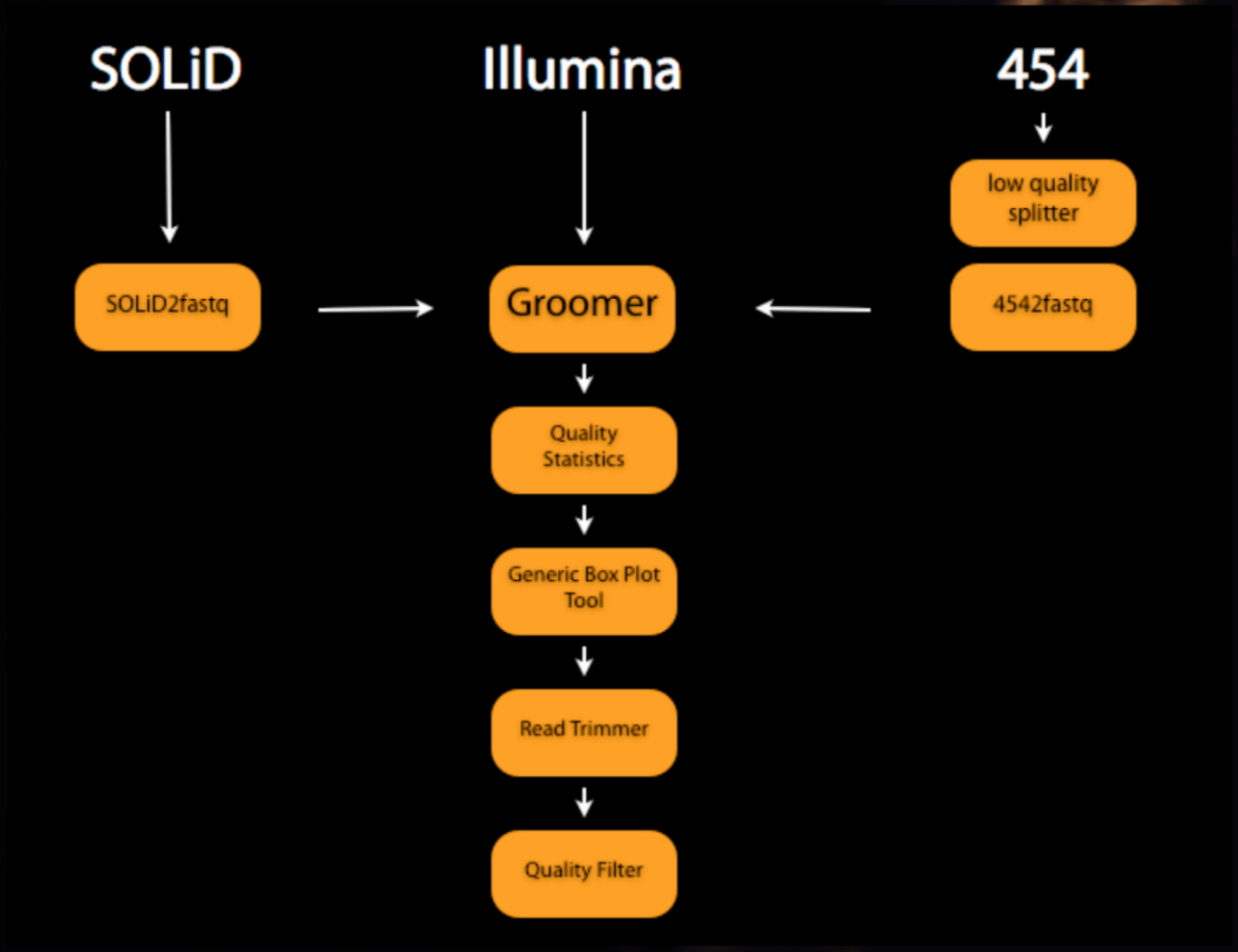
# Grooming the Data

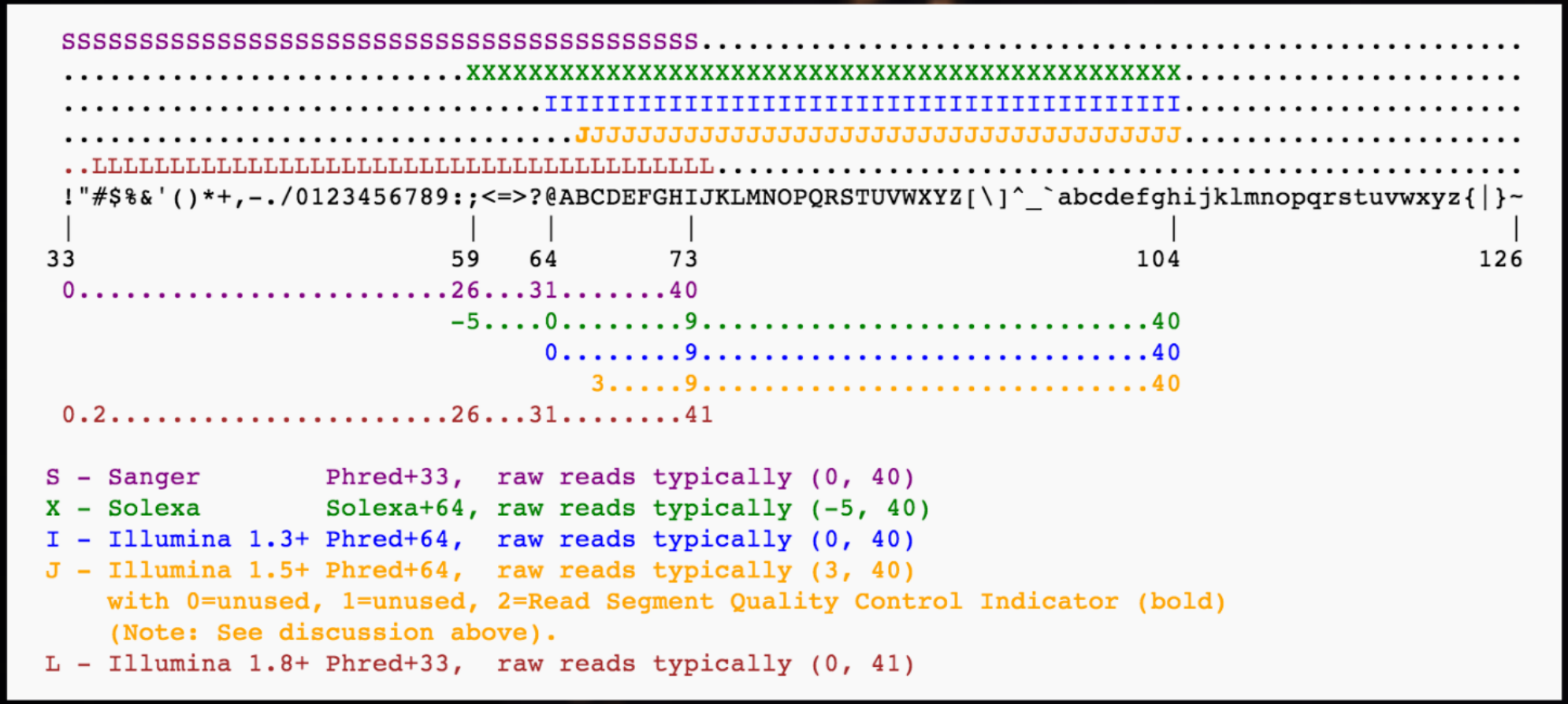
## When your read qualities are not formatted for galaxy...




SOLiD
Illumina
454
low quality splitter
SOLiD2fastq
Groomer
4542fastq
Quality Statistics
Generic Box Plot Tool
Read Trimmer
Quality Filter


SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
!"#$%&'()*+,-./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)


```
  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)
```

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 7.2 GB

**Tools**

- 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

## FASTQ Groomer (version 1.0.4)

**File to groom:** 1: http://bioinfo.umassmed.edu/pub/data/control_rep1.1.fq

**Input FASTQ quality scores type:** Sanger

**Advanced Options:** Hide Advanced Options

Execute

**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

**History**

**5: FASTQ Groomer on data 1**
3.1 MB
format: fastqsanger, database: mm10
Groomed 24788 sanger reads into sanger reads. Based upon quality and sequence, the input data is valid for: solexa, sanger, illumina Input ASCII range: 'i'(105) - 'i'(105) Input decimal range: 72 - 72

```
@HWI-ST333_0273_FC:7:1205:9843:15143
CAAGGAAGCACATGACCGAGCAGAAAATACCCAGTT
+
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
@HWI-ST333_0273_FC:7:2116:8217:37625
GTCAGCTTCCTGATGTTCTCCAGGCCACTGTACACT
```

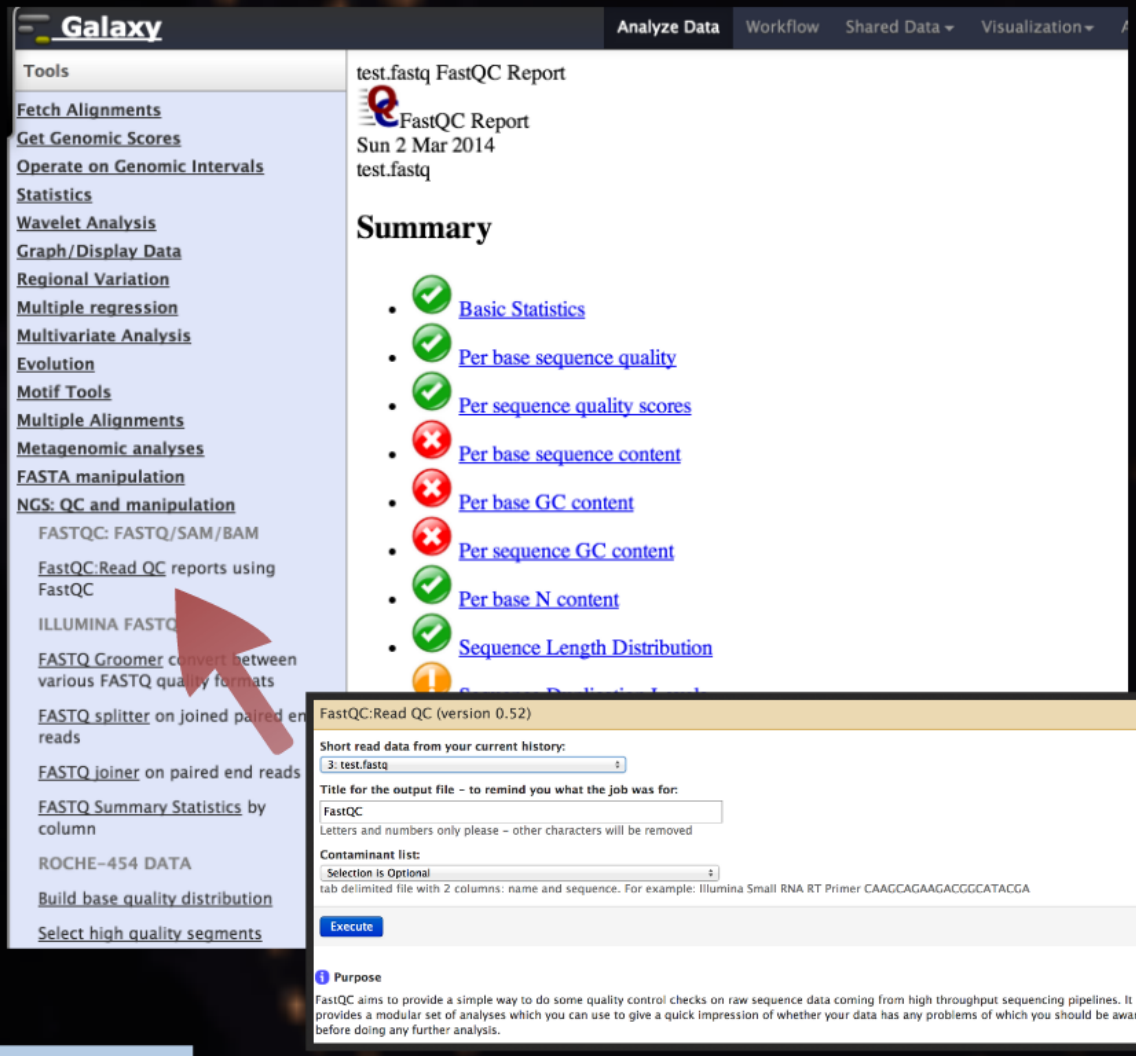
**4:** http://bioinfo.umassmed.edu/pub/data/mm10.fa

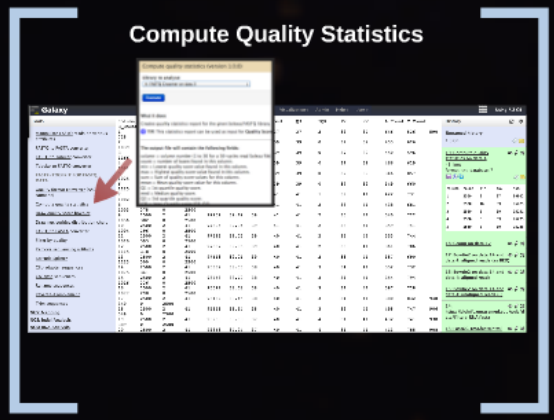
**3:** http://bioinfo.umassmed.edu/pub/data/ucsc.gtf

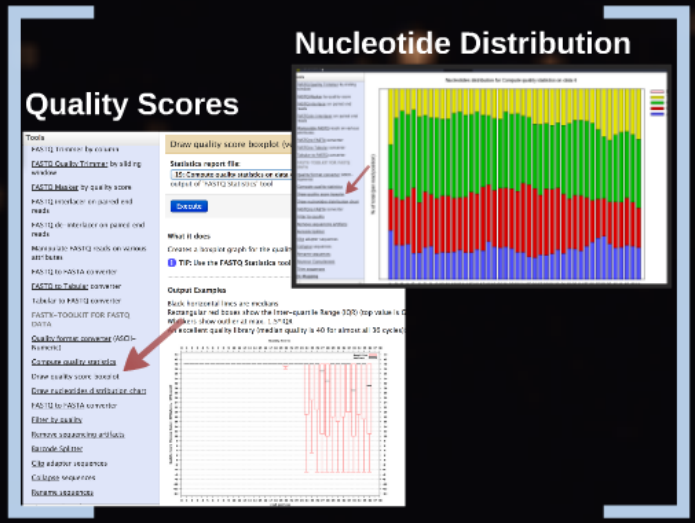
**2:**

# FASTQC and other QC

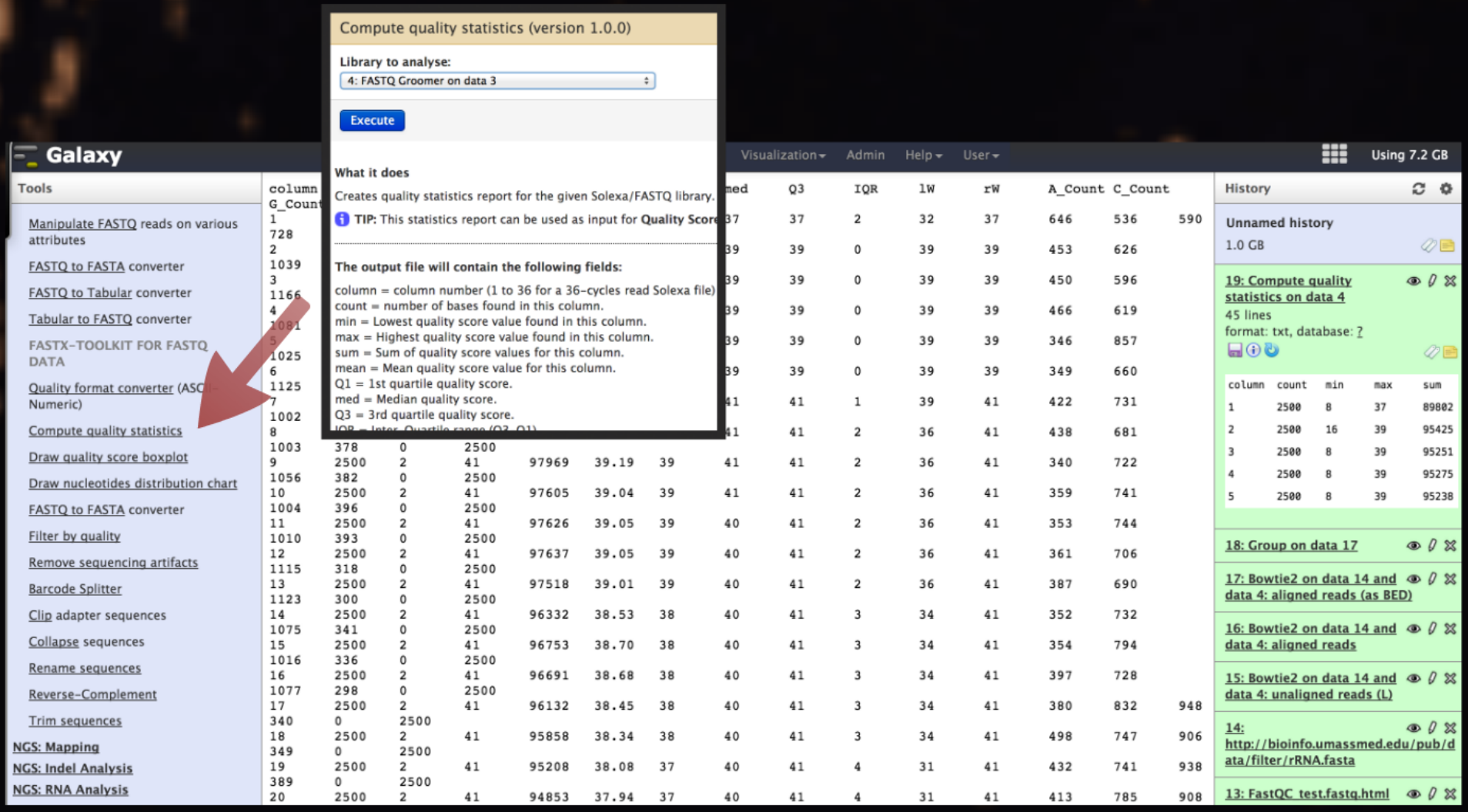
Upload this file to your galaxy;
http://bioinfo.umassmed.edu/pub/data/test.fastq



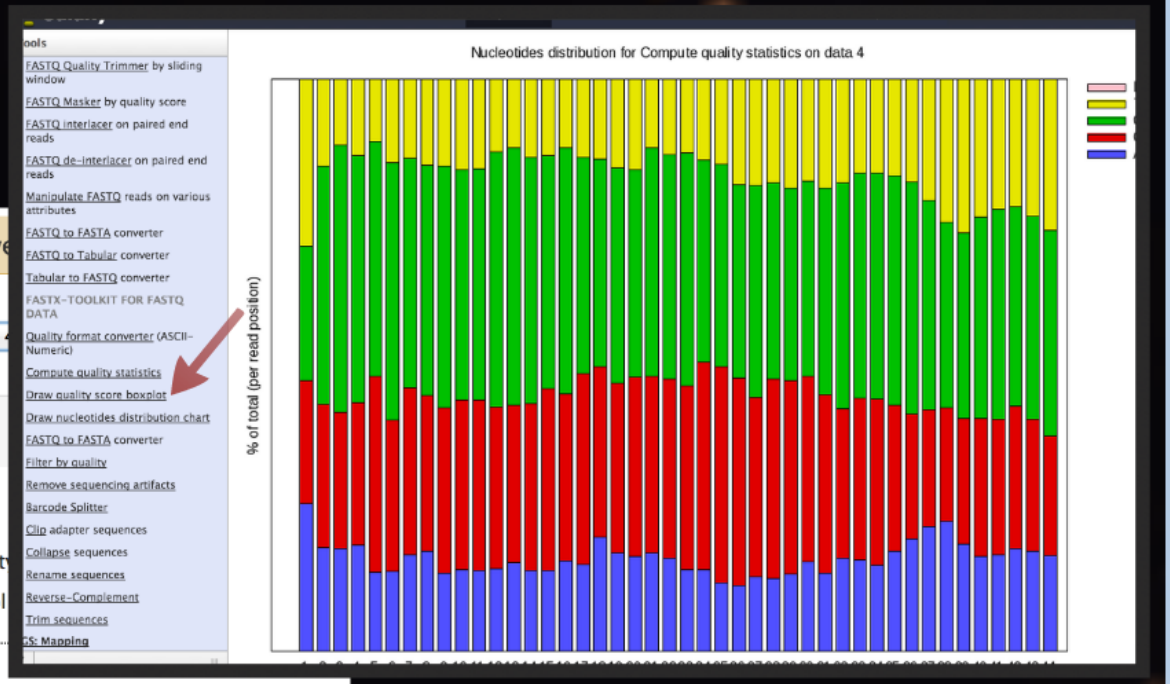



# Compute Quality Statistics



# Nucleotide Distribution




# Quality Scores

Tools

- 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
- FASTX-TOOLKIT FOR FASTQ DATA
- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Rename sequences

Draw quality score boxplot (ve

**Statistics report file:**

19: Compute quality statistics on data 4

output of 'FASTQ Statistics' tool

Execute

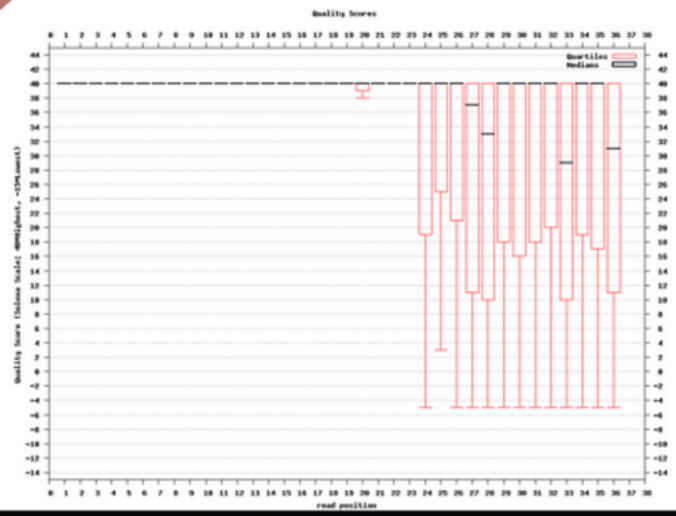
**What it does**

Creates a boxplot graph for the qualit

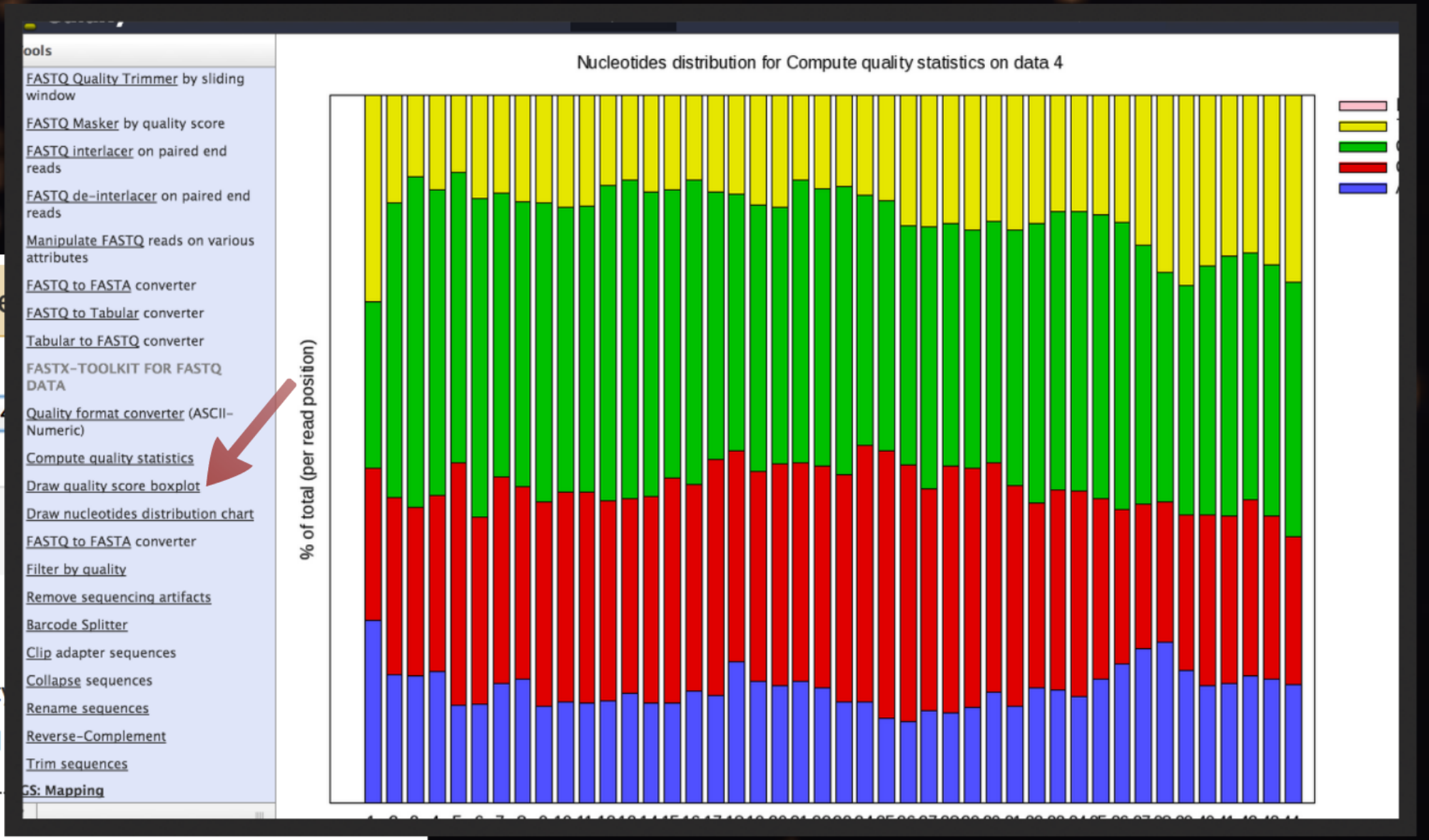
**TIP:** Use the **FASTQ Statistics** tool

**Output Examples**

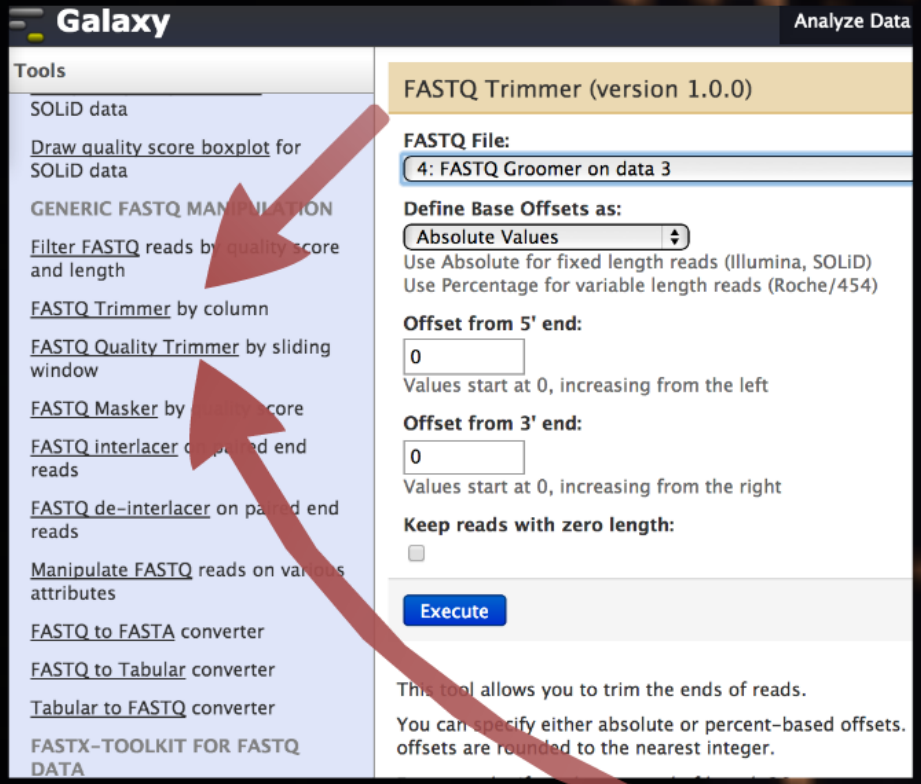
Black horizontal lines are medians
Rectangular red boxes show the Inter-quartile Range (IQR) (top value is Q
Whiskers show outlier at max. 1.5*IQR
An excellent quality library (median quality is 40 for almost all 36 cycles):



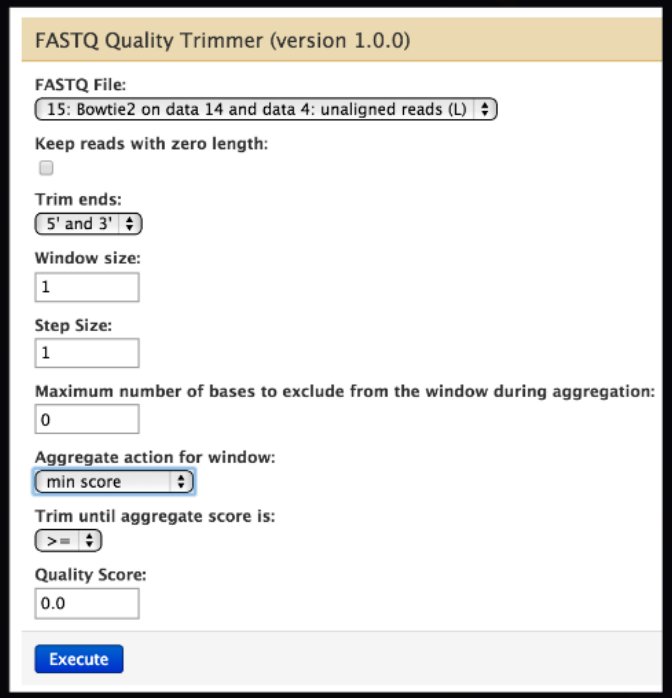
# Nucleotide Distribution



# Read Trimming




# Quality Filtering



# Example: Filtering Ribosomal Reads and Counting

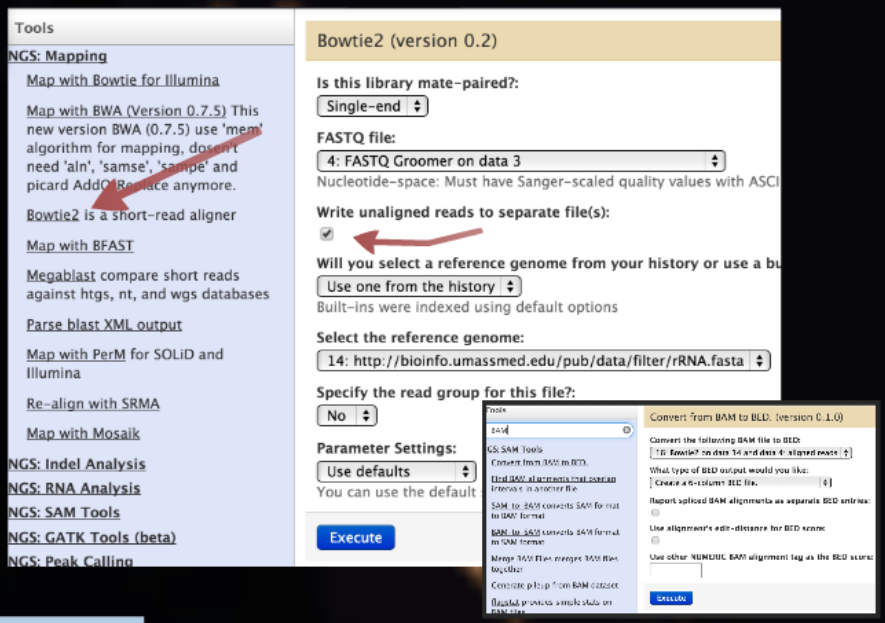
- Upload Ribosomal RNA sequences.
  http://bioinfo.umassmed.edu/pub/data/filter/rRNA.fasta

```
>gi|124517659|ref|NR_003286.1|_Homo_sapiens_18S_ribosomal_RNA_(LOC100008588)
TACCTGGTTGATCCTGCCAGTAGCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTAAGTACGCACG
GCCGGTACAGTGAAACTGCGAATGGCTCATTAAATCAGTTATGGTTCCTTTGGTCGCTCGCTCCTCTCCT
ACTTGGATAACTGTGGTAATTCTAGAGCTAATACATGCCGACGGGCGCTGACCCCCTTCGCGGGGGGGAT
GCGTGCATTTATCAGATCAAAACCAACCCGGTCAGCCCCTCTCCGGCCCCGGCCGGGGGGCGGGCGCCGG
CGGCTTTGGTGACTCTAGATAACCTCGGGCCGATCGCACGCCCCCCGTGGCGGCGACGACCCATTCGAAC
GTCTGCCCTATCAACTTTCGATGGTAGTCGCCGTGCCTACCATGGTGACCACGGGTGACGGGGAATCAGG
GTTCGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTAC
CCACTCCCGACCCGGGGAGGTAGTGACGAAAAATAACAATACAGGACTCTTTCGAGGCCCTGTAATTGGA
ATGAGTCCACTTTAAATCCTTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAAT
TCCAGCTCCAATAGCGTATATTAAAGTTGCTGCAGTTAAAAAGCTCGTAGTTGGATCTTGGGAGCGGGCG
```

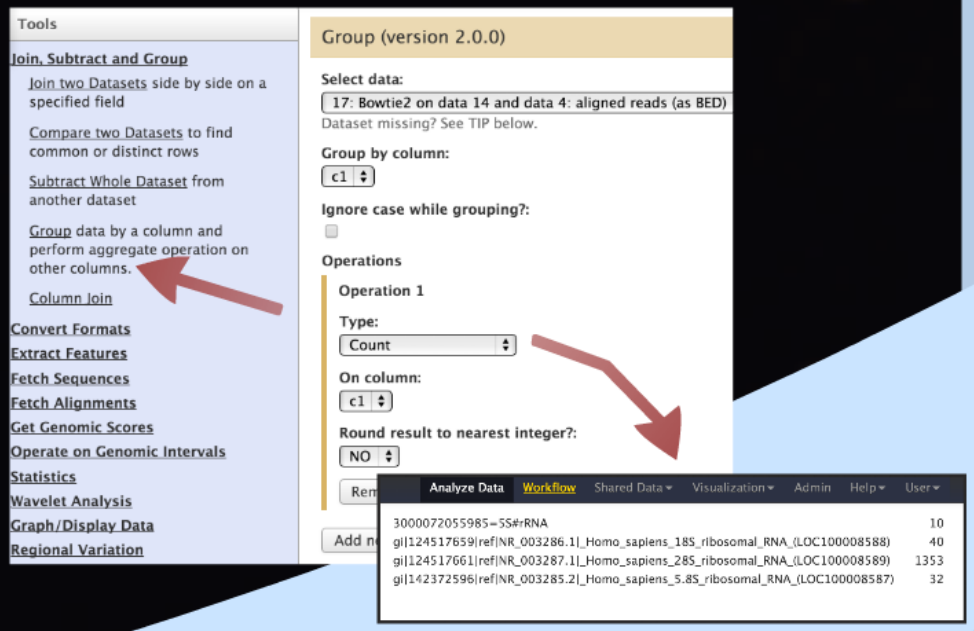
History

14: http://bioinfo.umassmed.edu/pub/data/filter/rRNA.fasta
4 sequences
format: fasta, database: hg19
uploaded fasta file

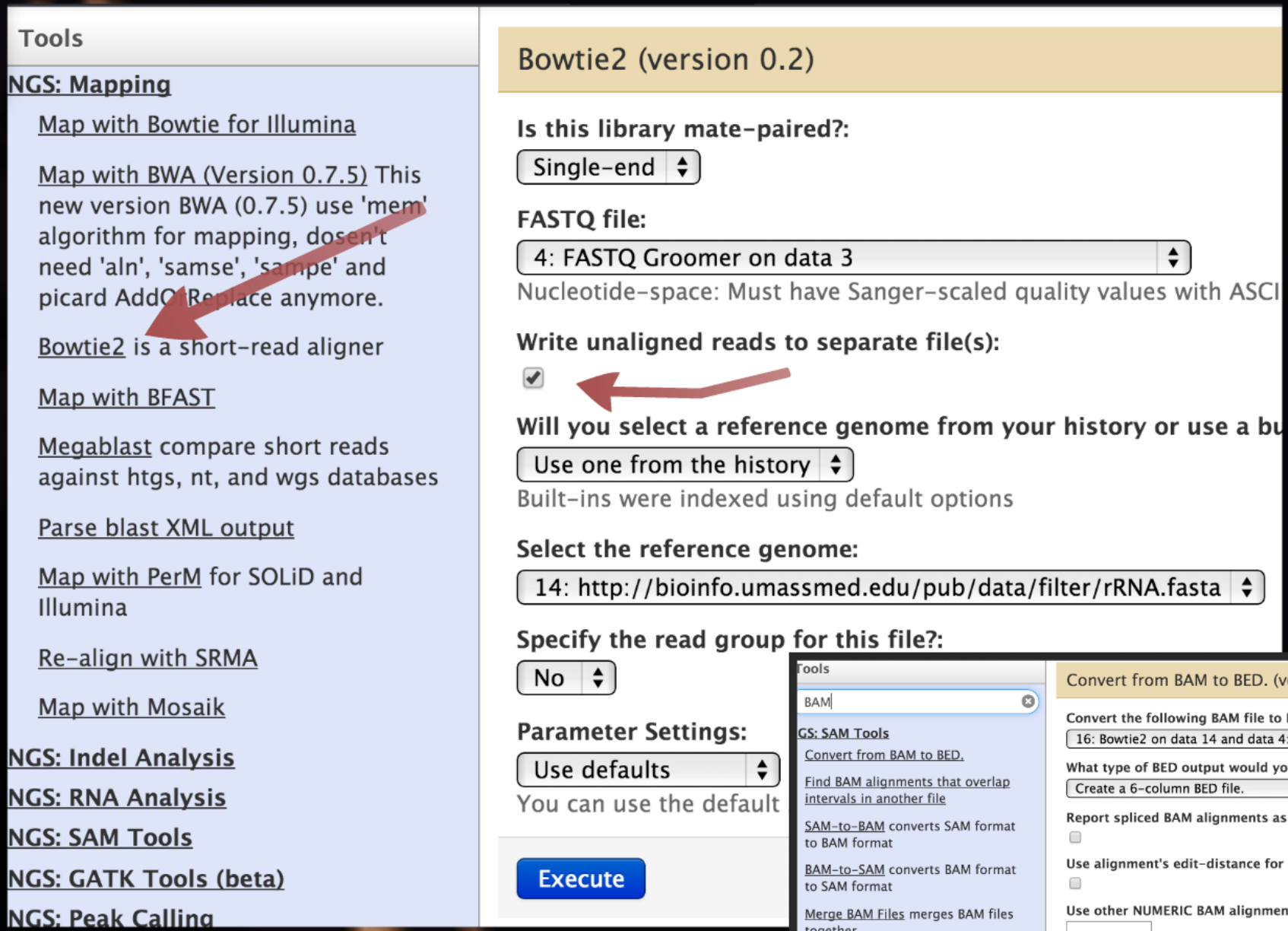
## Mapping to rRNAs




## Group the Data and Count



# Mapping to rRNAs



ssmed.edu/pub/data/filter/rRNA.fasta

# for this file?:

Tools

BAM

**GS: SAM Tools**

- Convert from BAM to BED.
- Find BAM alignments that overlap intervals in another file
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- flagstat provides simple stats on BAM files

## Convert from BAM to BED. (version 0.1.0)

**Convert the following BAM file to BED:**

16: Bowtie2 on data 14 and data 4: aligned reads

**What type of BED output would you like:**

Create a 6-column BED file.

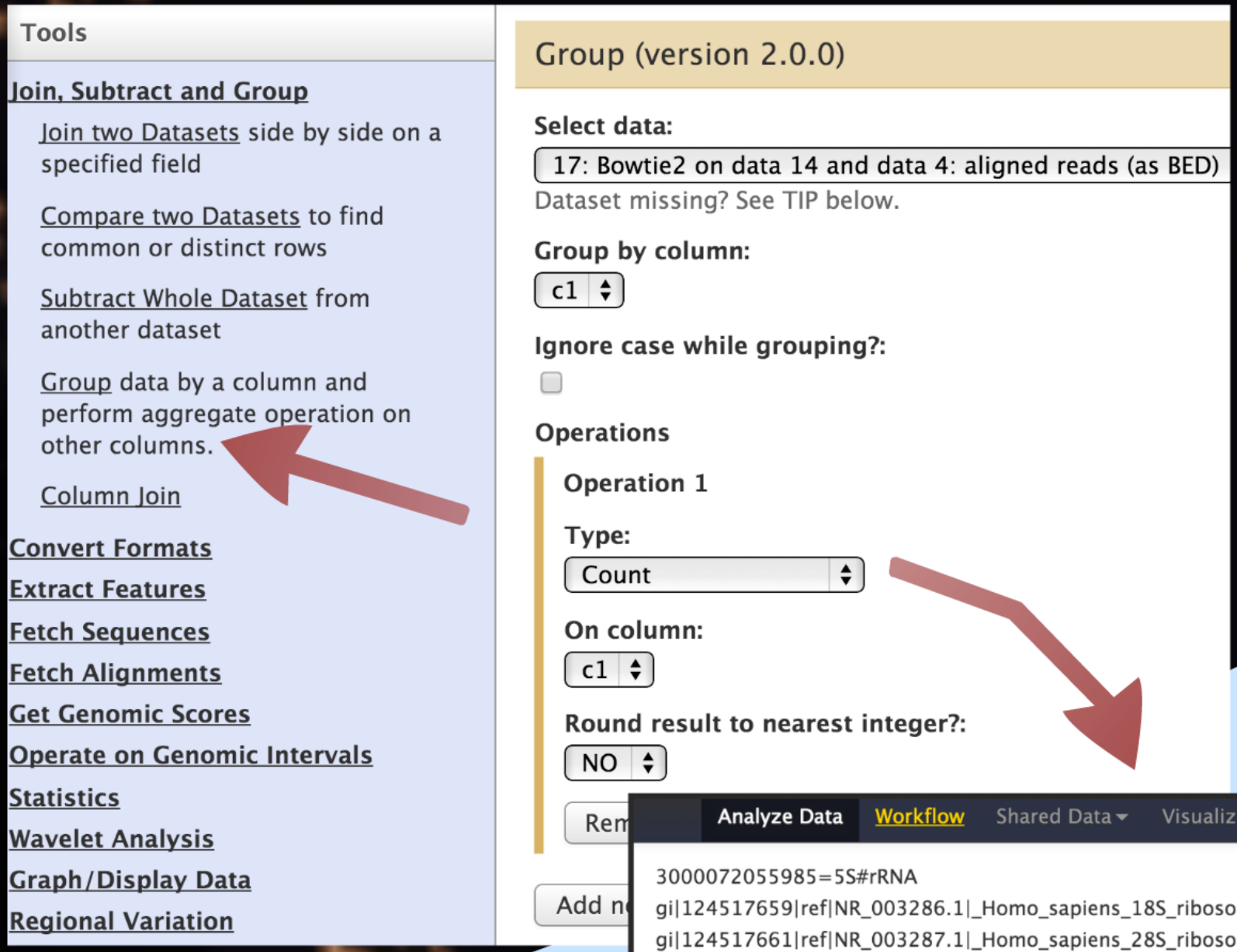
**Report spliced BAM alignments as separate BED entries:**

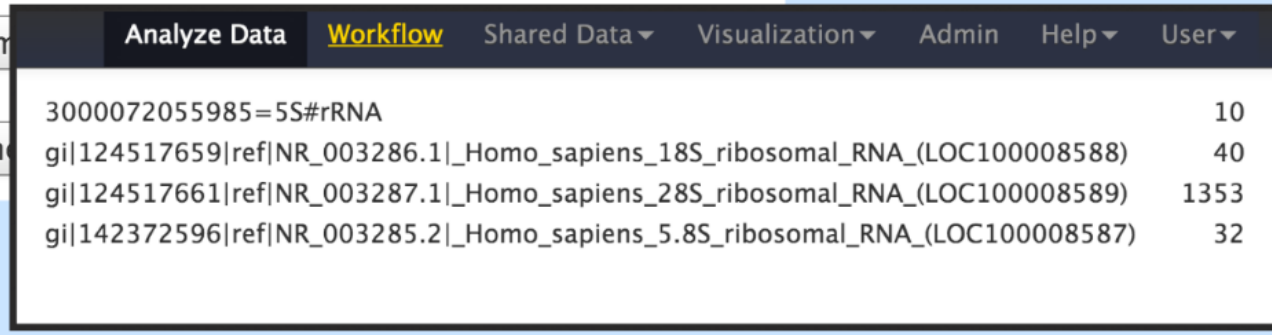
**Use alignment's edit-distance for BED score:**

**Use other NUMERIC BAM alignment tag as the BED score:**

Execute

# Group the Data and Count



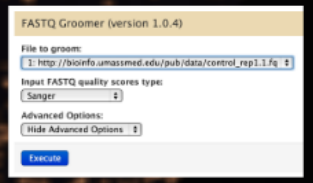
# Homework:

## Galaxy tools for RNA-Seq Data

### Read alignments and visualization

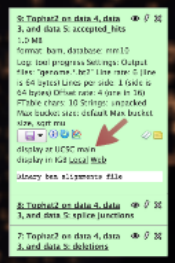
1. Upload

2. FastQ Groomer

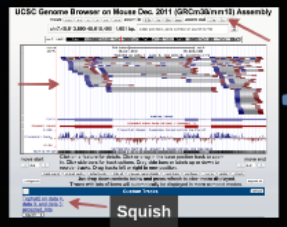



3. Map with Tophat2

4. Visualize BAM file



5. UCSC Genome Browser

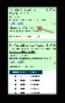



6. Coverage Visualization

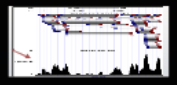
6a. Coverage

6b. bedGraph to bigWig conversion

6c.Visualization

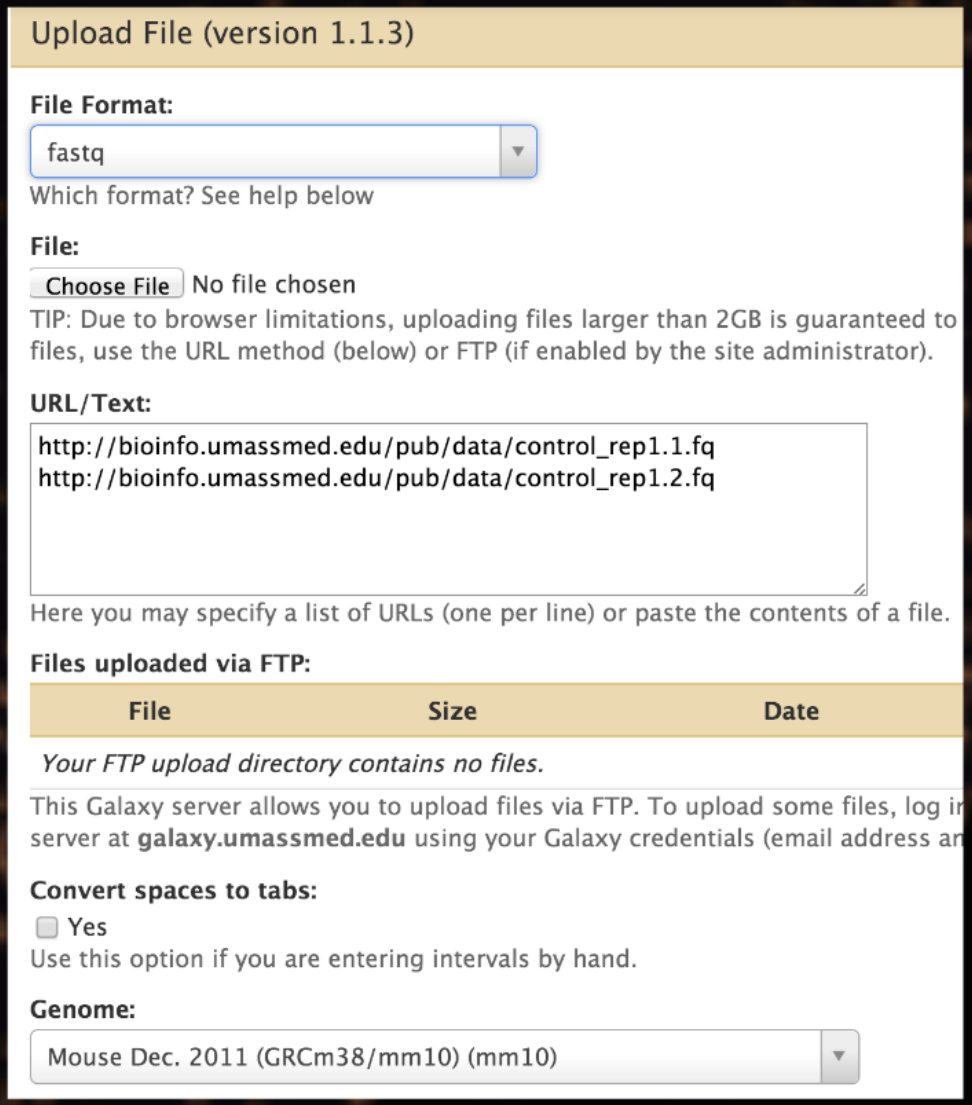


6d. Browser



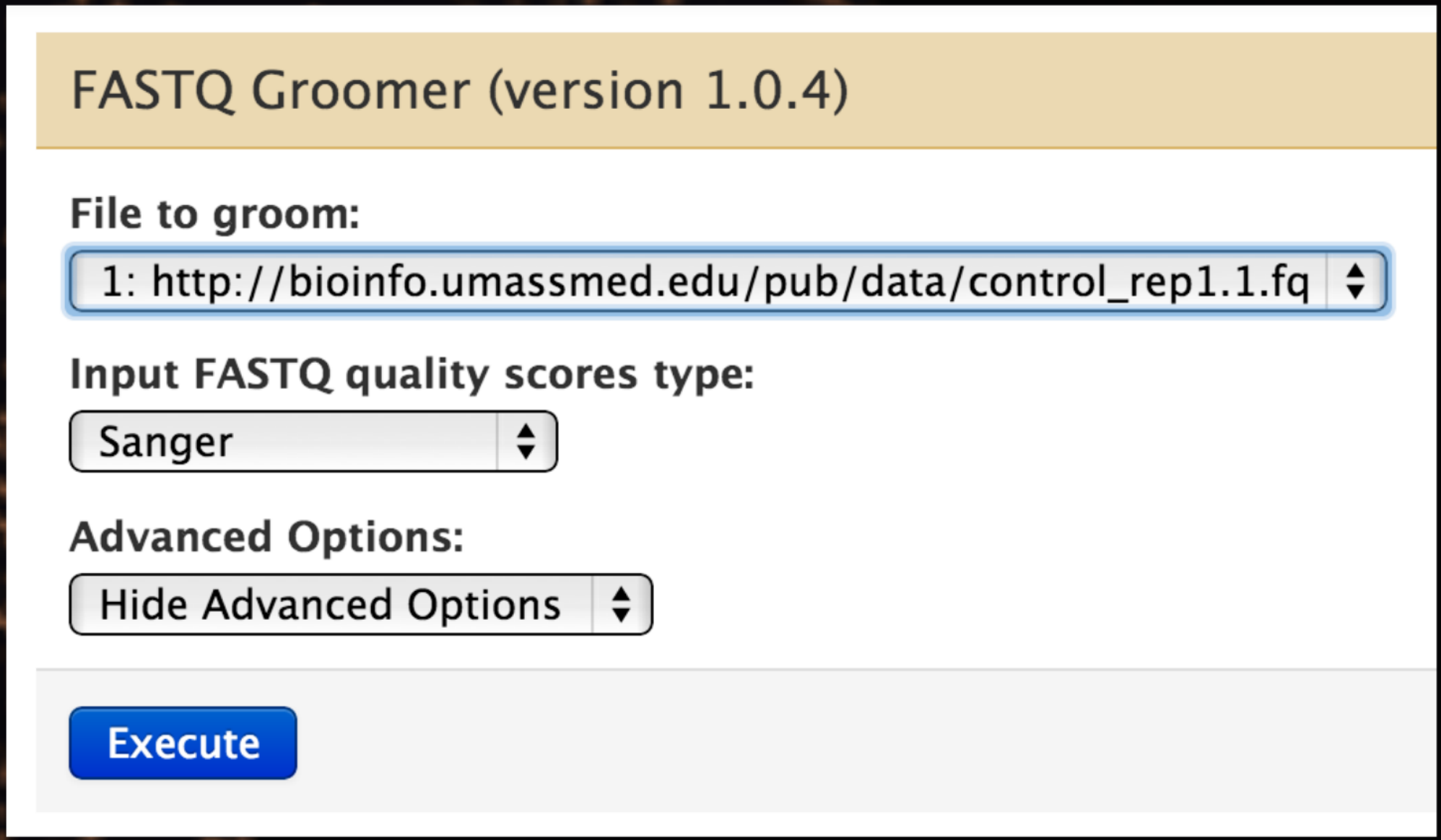
# 1. Upload

## 1a. Upload fastq files:

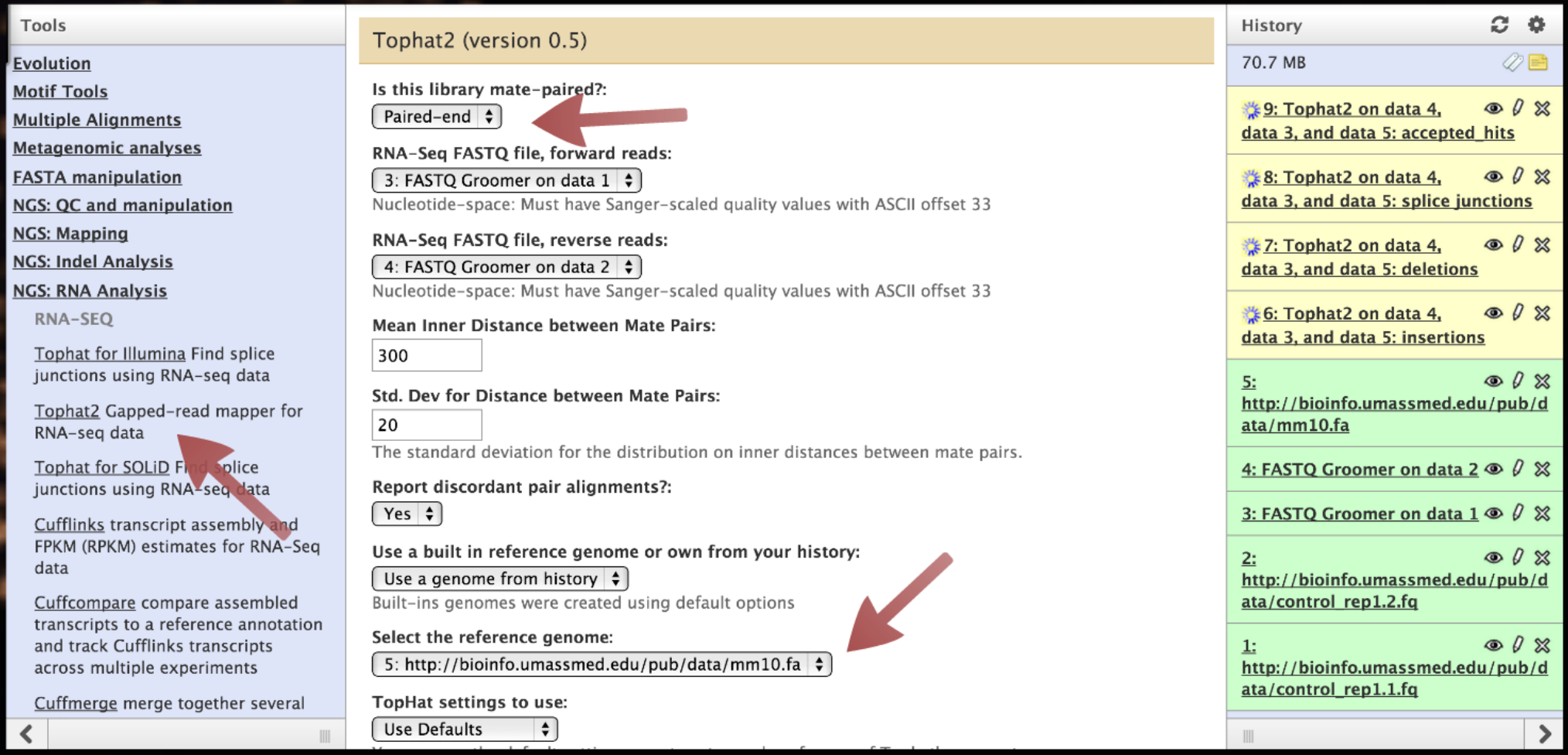



**1b. Upload genome File:**
http://bioinfo.umassmed.edu/pub/data/mm10.fa => fasta

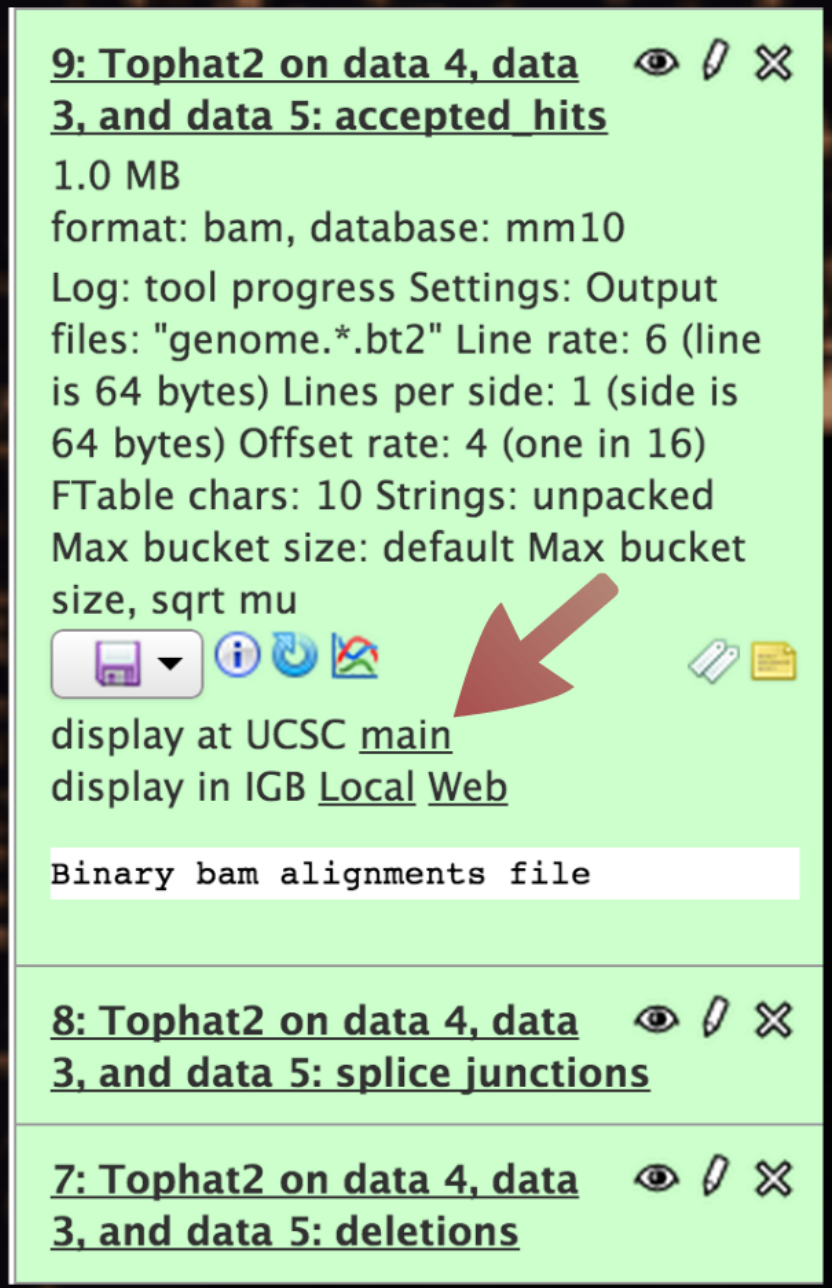
# 2. FastQ Groomer



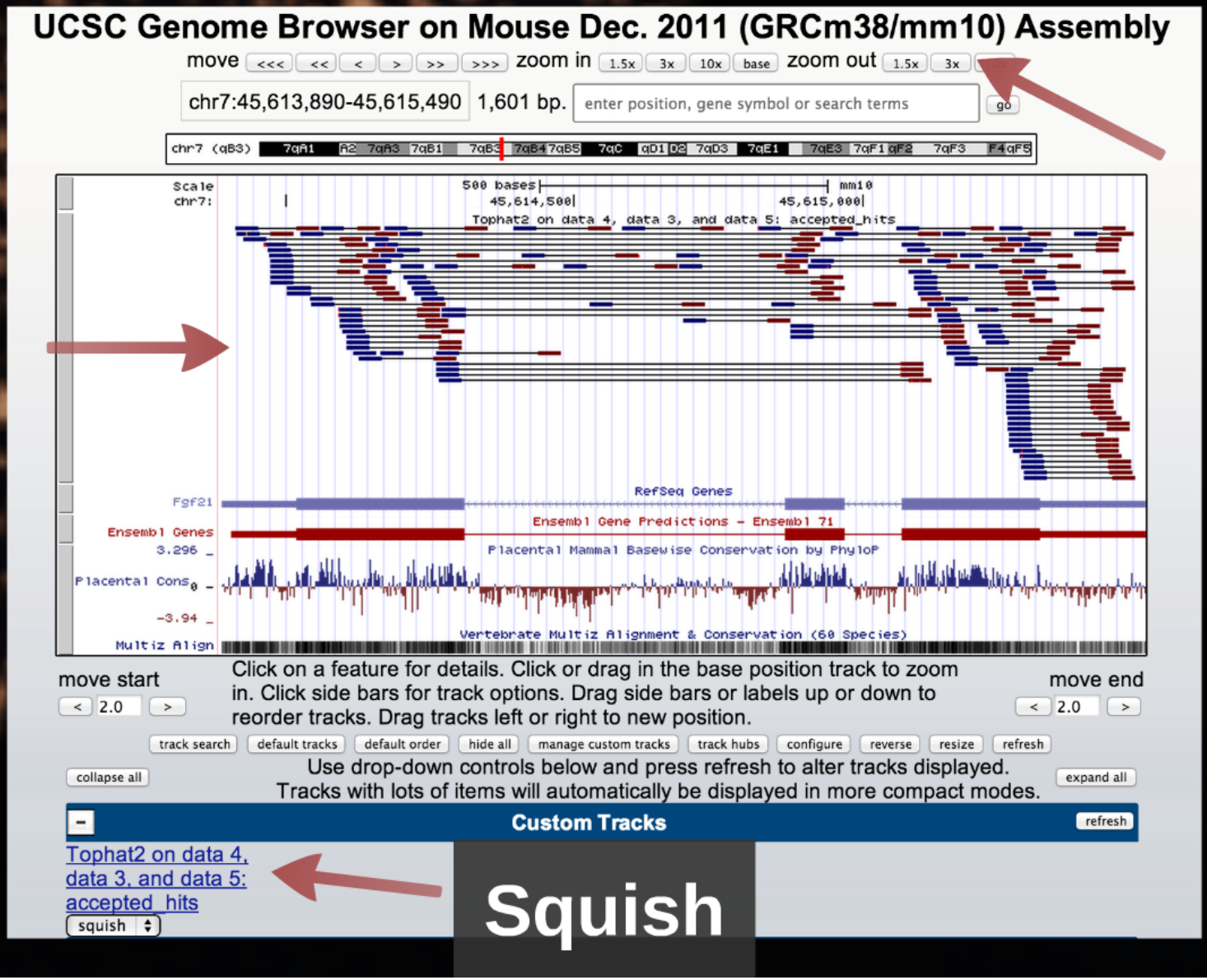
# 3. Map with Tophat2



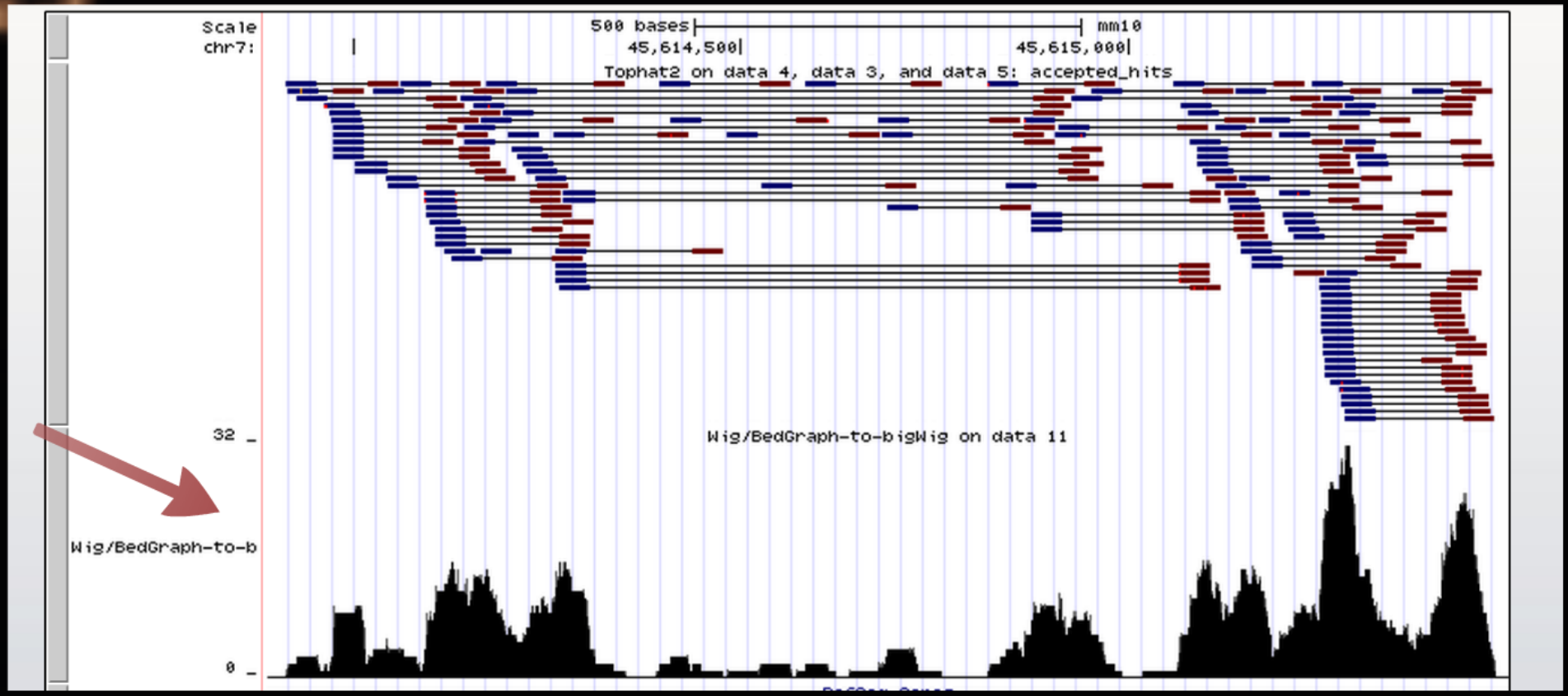
# 4. Visualize BAM file



# 5. UCSC Genome Browser

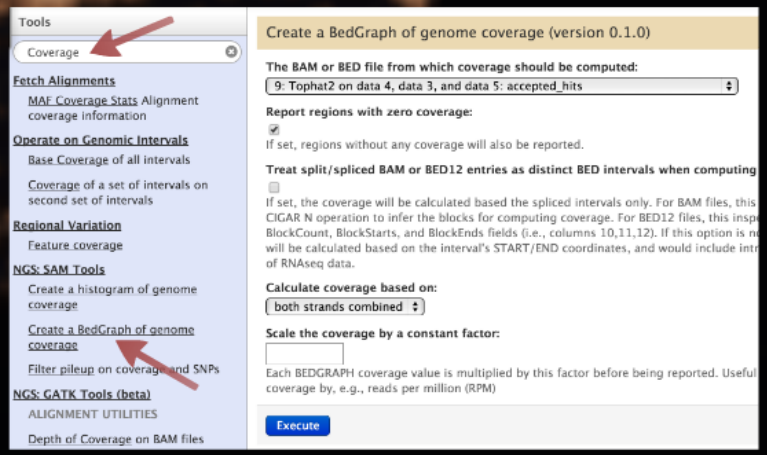


# 6d. Browser

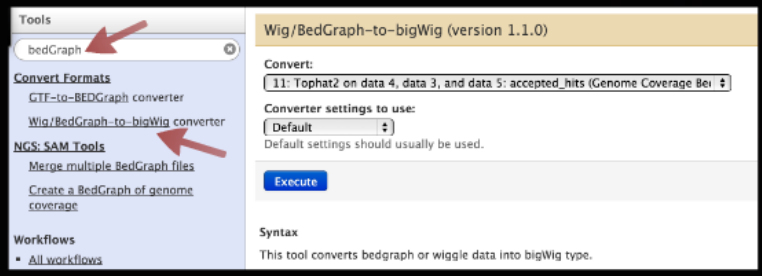

Scale
chr7:
500 bases
mm10
45,614,500
45,615,000
Tophat2 on data 4, data 3, and data 5: accepted_hits
32
Wig/BedGraph-to-bigWig on data 11
Wig/BedGraph-to-b
0

# 6. Coverage Visualization

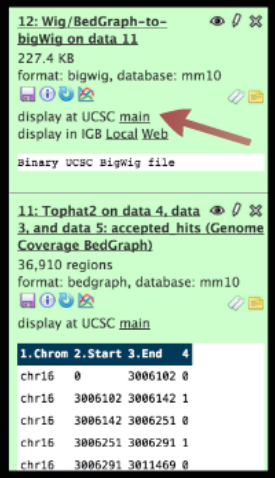
## 6a. Coverage



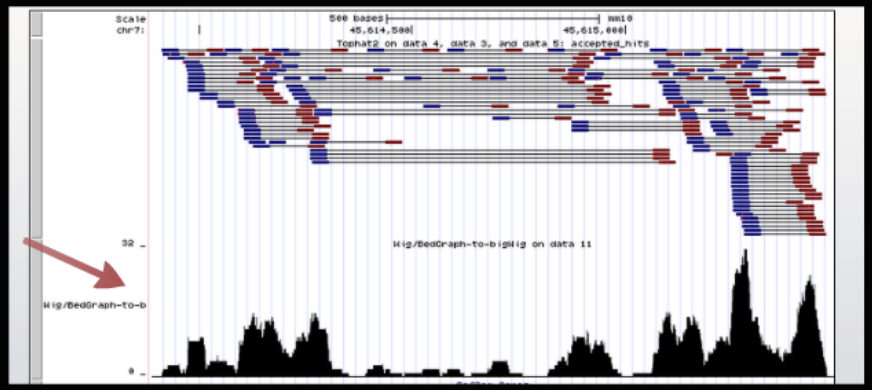

## 6b. bedGraph to bigWig conversion



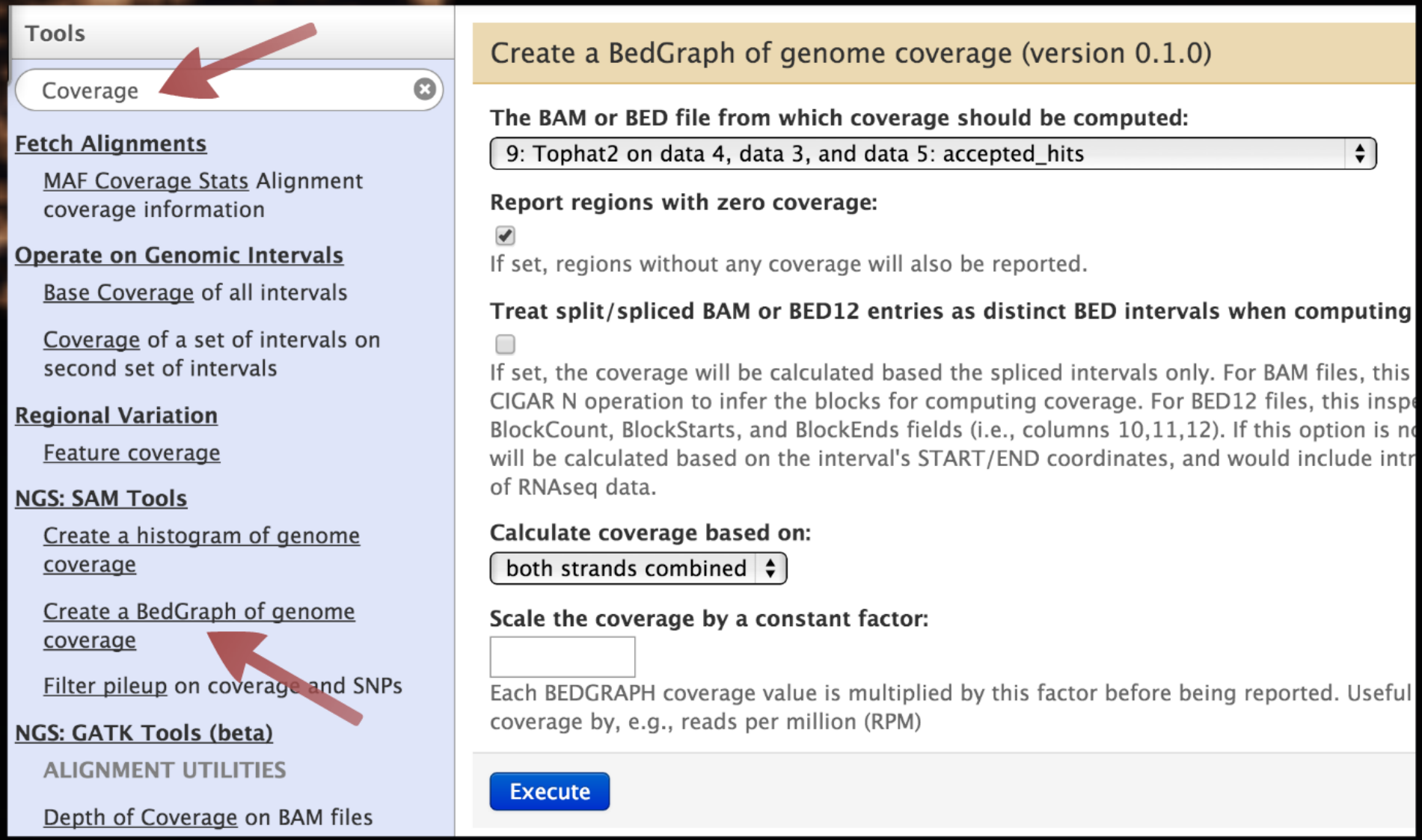

## 6c.Visualization



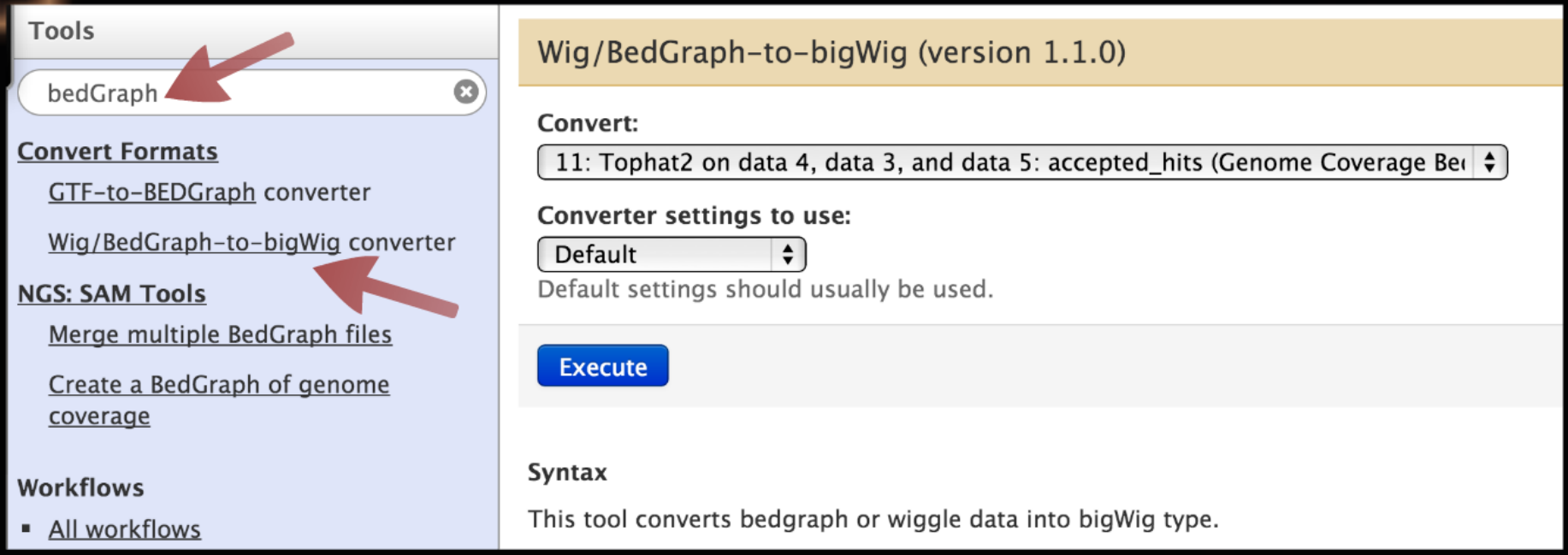

## 6d. Browser



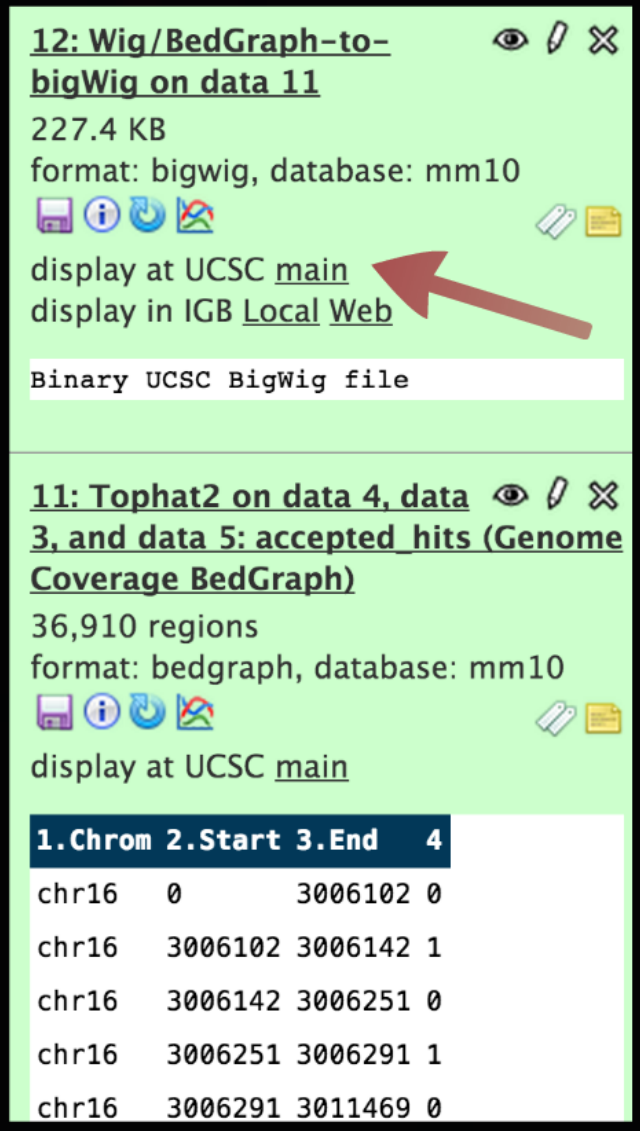
# 6a. Coverage


Tools
Coverage
Fetch Alignments
MAF Coverage Stats Alignment coverage information
Operate on Genomic Intervals
Base Coverage of all intervals
Coverage of a set of intervals on second set of intervals
Regional Variation
Feature coverage
NGS: SAM Tools
Create a histogram of genome coverage
Create a BedGraph of genome coverage
Filter pileup on coverage and SNPs
NGS: GATK Tools (beta)
ALIGNMENT UTILITIES
Depth of Coverage on BAM files
Create a BedGraph of genome coverage (version 0.1.0)
The BAM or BED file from which coverage should be computed:
9: Tophat2 on data 4, data 3, and data 5: accepted_hits
Report regions with zero coverage:
If set, regions without any coverage will also be reported.
Treat split/spliced BAM or BED12 entries as distinct BED intervals when computing
If set, the coverage will be calculated based the spliced intervals only. For BAM files, this
CIGAR N operation to infer the blocks for computing coverage. For BED12 files, this inspe
BlockCount, BlockStarts, and BlockEnds fields (i.e., columns 10,11,12). If this option is n
will be calculated based on the interval's START/END coordinates, and would include intr
of RNAseq data.
Calculate coverage based on:
both strands combined
Scale the coverage by a constant factor:
Each BEDGRAPH coverage value is multiplied by this factor before being reported. Useful
coverage by, e.g., reads per million (RPM)
Execute

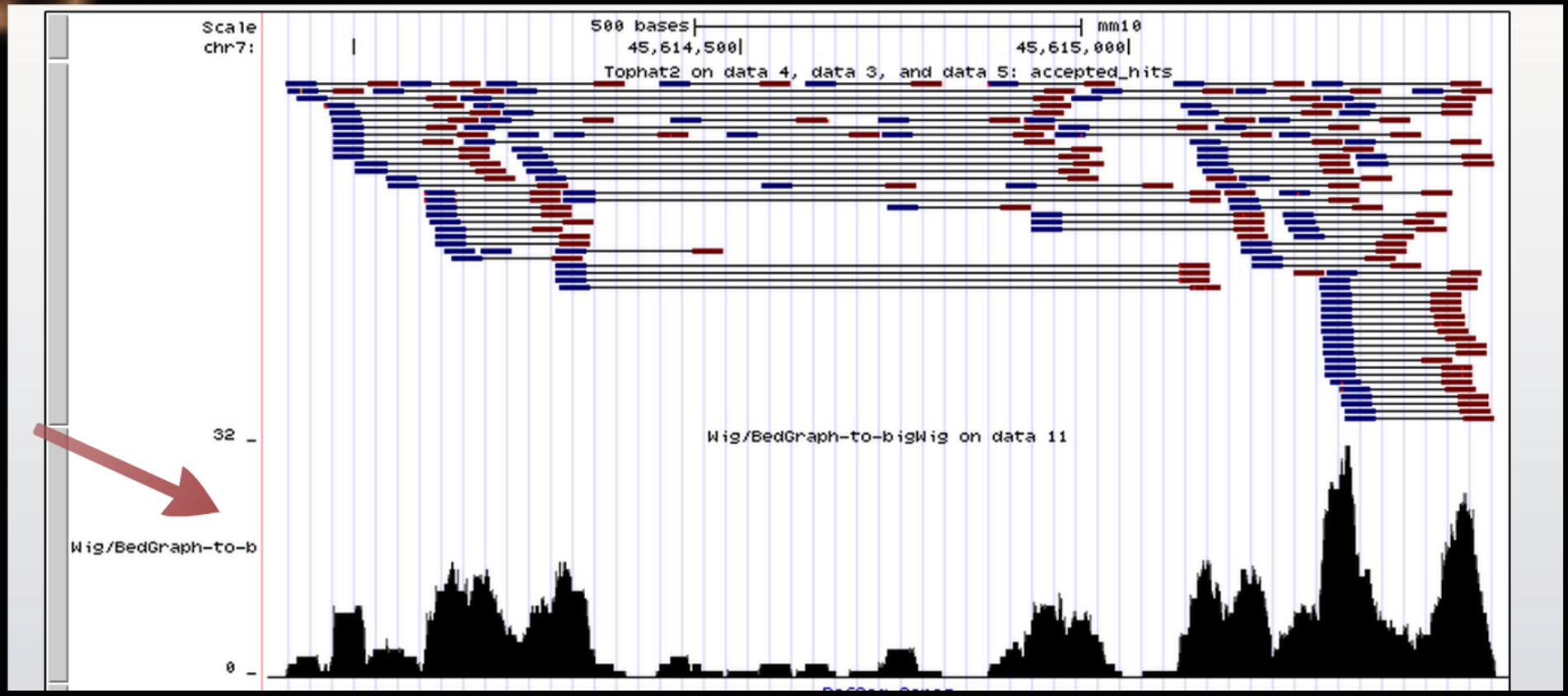
# 6b. bedGraph to bigWig conversion



# 6c.Visualization



# 6d. Browser


Scale
chr7:
500 bases
mm10
45,614,500
45,615,000
Tophat2 on data 4, data 3, and data 5: accepted_hits
32
Wig/BedGraph-to-bigWig on data 11
Wig/BedGraph-to-b
0

# Thanks!!!


JOIN ME
AND TOGETHER WE CAN
RULE THE GALAXY
© culture-nomad.com