RNA-seq Analysis with Galaxy.

Benjamin King, PhD | bking@mdibl.org | Kyle Shank | kshank@mdibl.org |

Contents

Objective

Align a set of RNA-seq reads to the *Mycobacterium smegmatis* MC2 155 genome assembly and perform a basic analysis.

The overall process consists of five discrete steps:

1. Import the reference genome and reference genome annotation.
2. Import the RNA-seq reads.
3. Perform diagnostic analyses of the RNA-seq reads.
4. Align the RNA-seq reads to the reference genome.
5. Generate a count of reads per gene that can be analyzed downstream with `R`.

Step 0: Register for Galaxy

Before proceeding, you must register for a free account on Galaxy. It's fast, and importantly, free. After logging in, you should arrive at a screen that looks like this:

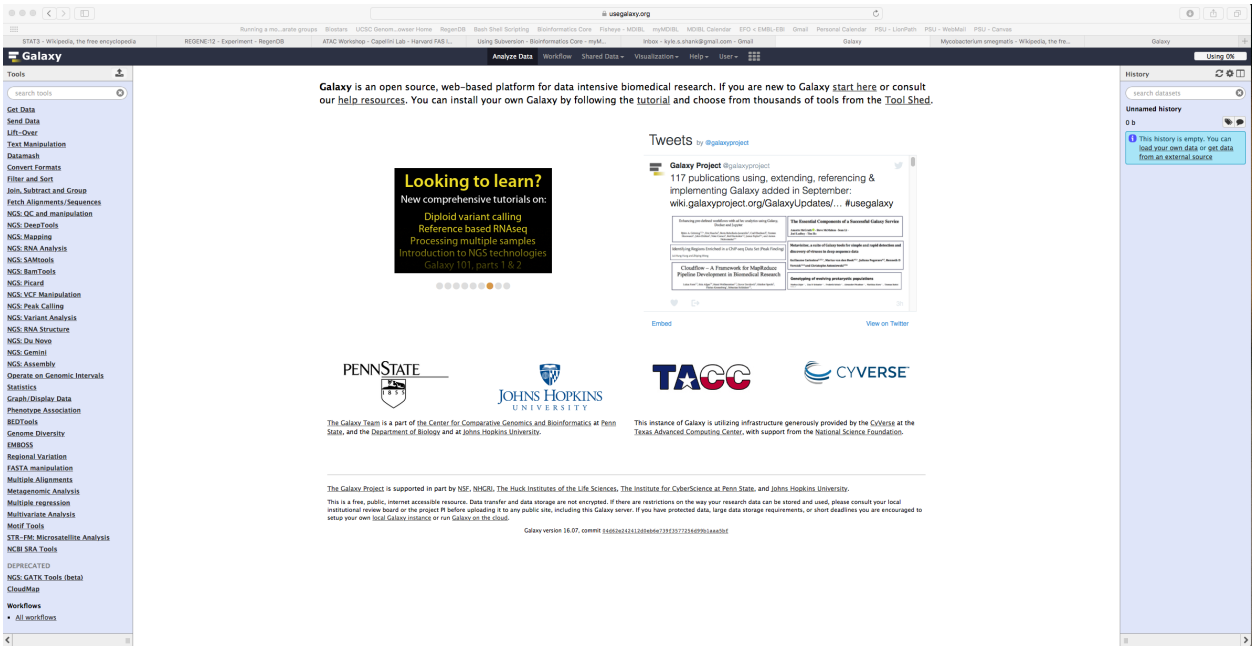


Figure 1:

Step 1: Import your Genome Information

In this step, we're going to download the genome assembly (a FASTA file) and the annotation (a gtf file) onto our local machine from EnsemblBacteria.

Download the Genome Assembly

Click on this link to download the *M. Smegmatis* MC2 155 genome assembly.

Beneath the section labeled "Gene Annotation", click on the **FASTA** link.

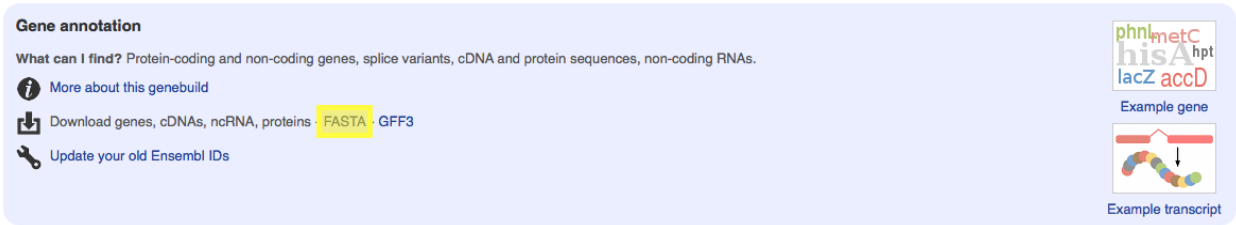


Figure 2: This is the proper place to click to fetch the FASTA file

When prompted, make sure to choose to continue your download as a *Guest*. A directory will then be downloaded. Within it are several sub-directories: `cdna`, `cds`, `dna`, `ncrna`, `pep`. Open `dna`. From here, save the file marked `Mycobacterium_smegmatis_str_mc2_155.ASM1500v1.dna_sm.toplevel.fa.gz` to your working directory on your local machine.

Download the Genome Annotation

Genome annotation can be one of the more difficult problems tackled in bioinformatics, mostly due to the plethora of file formats and transformation tools that are available. To simplify the task in this particular study, we have provided a suitable `GTF` file, availble here.

Import into Galaxy

From the Galaxy homepage, select **Get Data**.

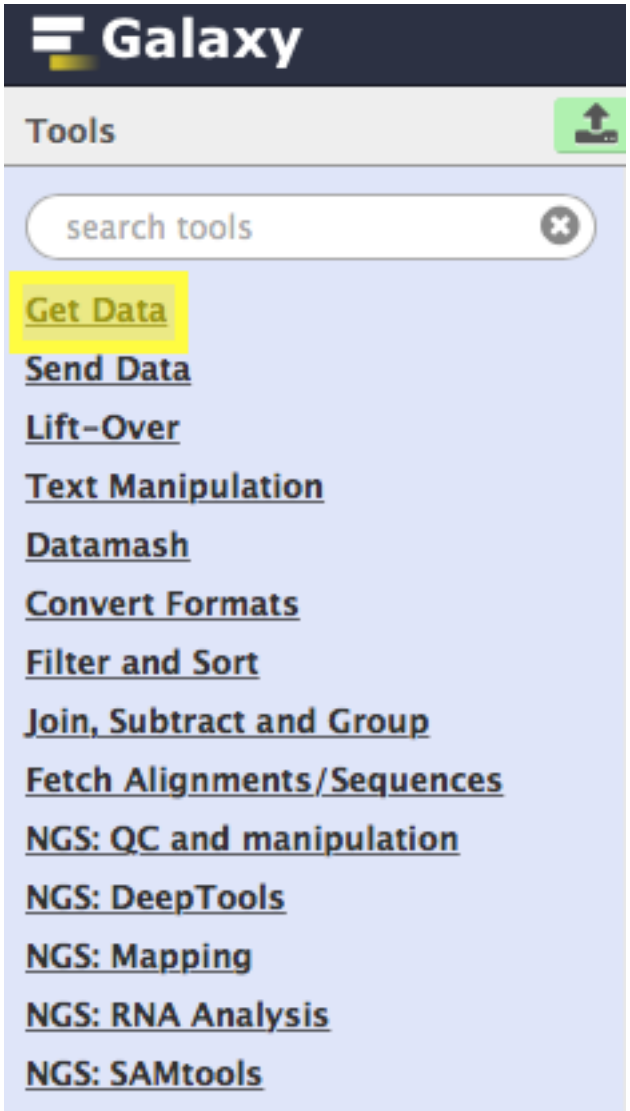


Figure 3:

Next, click **Upload File from your computer**

Drag and drop both files that you wish to upload. Under `Type`, make sure to change the set the Geneome Assembly file to `fasta` and the Genome Annotation file to `gtf`. Click **Start**.

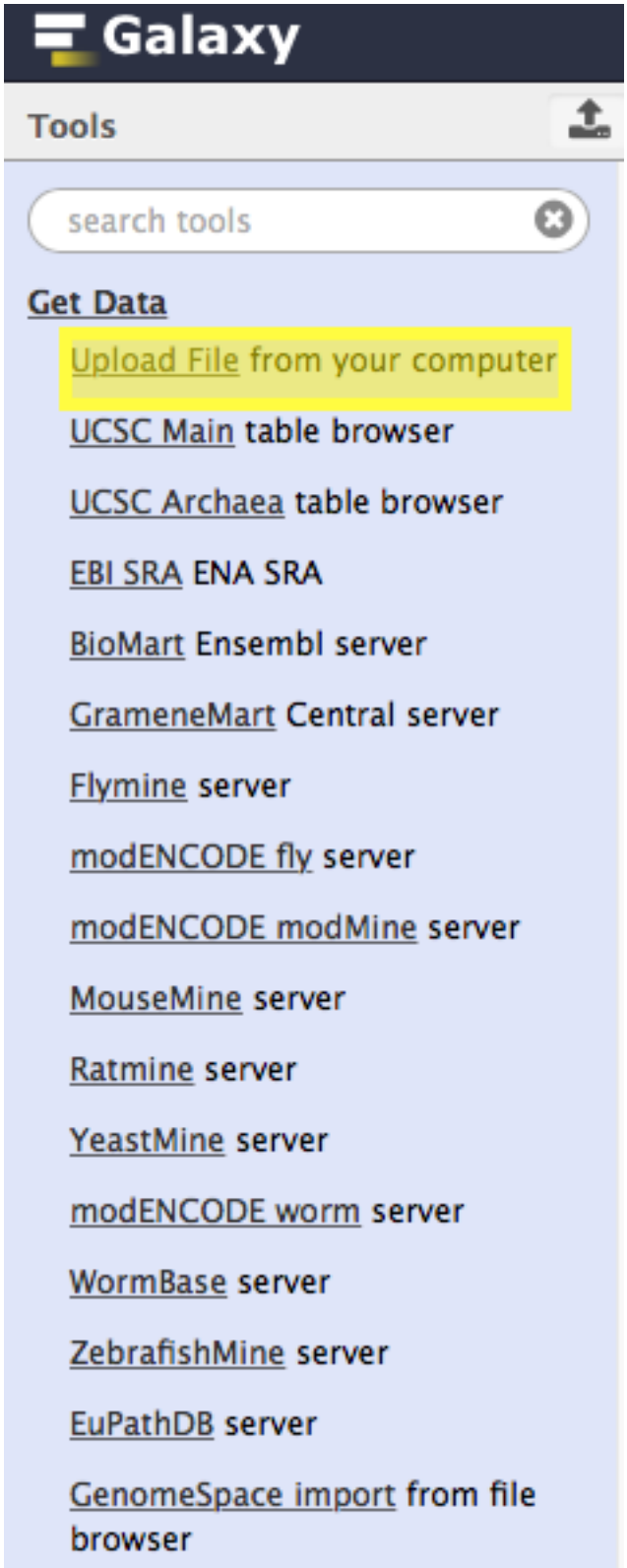

Galaxy
Tools
search tools
Get Data
Upload File from your computer
UCSC Main table browser
UCSC Archaea table browser
EBI SRA ENA SRA
BioMart Ensembl server
GrameneMart Central server
Flymine server
modENCODE fly server
modENCODE modMine server
MouseMine server
Ratmine server
YeastMine server
modENCODE worm server
WormBase server
ZebrafishMine server
EuPathDB server
GenomeSpace import from file browser

Figure 4:

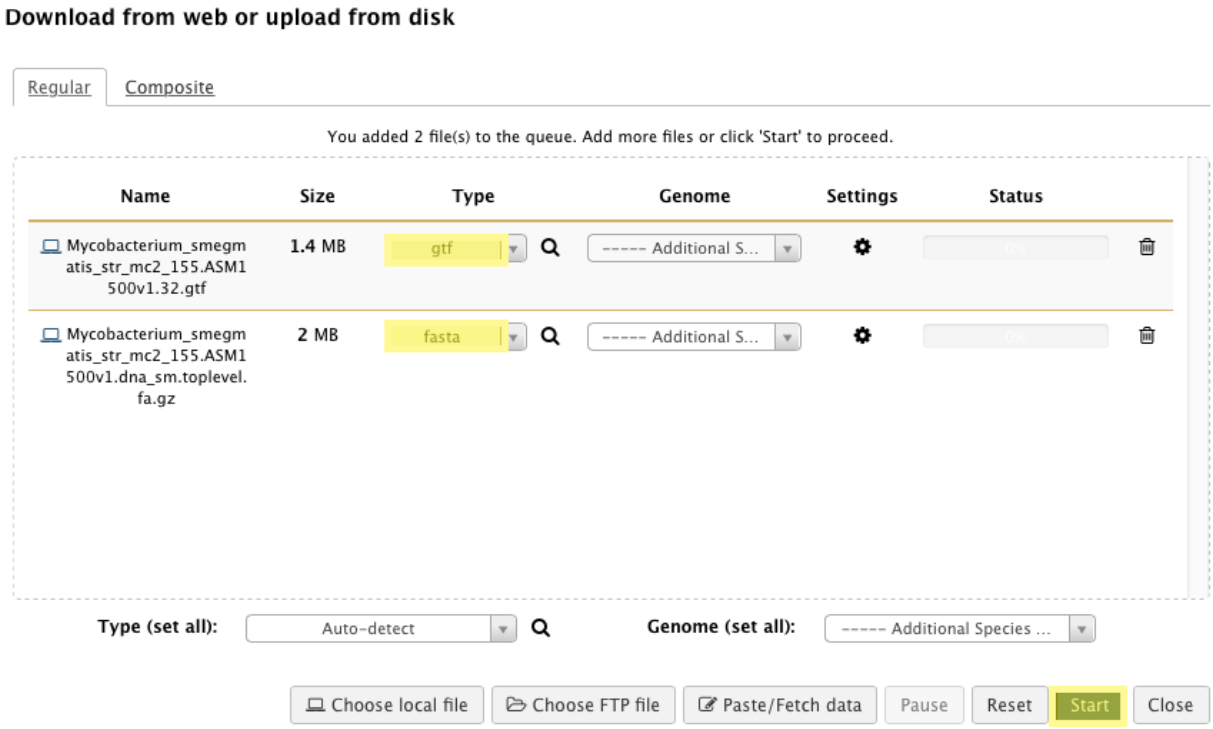


Figure 5:

Click **close**. You can now see in your history bar (on the right) that you've successfully uploaded both files.

Step 2: Import RNA-seq Reads.

The reads for the Giles RNA-Seq study were initially deposited in the NCBI Short Read Archive. The European Nucleotide Archive (ENA) has mirrored those data and has made it easy to upload the FASTQ files into Galaxy.

Find the RNA-seq Reads.

Navigate to the ENA and enter the Gene Expression Omnibus accession for this study, GSE43434. Click **Search**.

On the results page, select the first link (`SRP017906`) to look at the study.

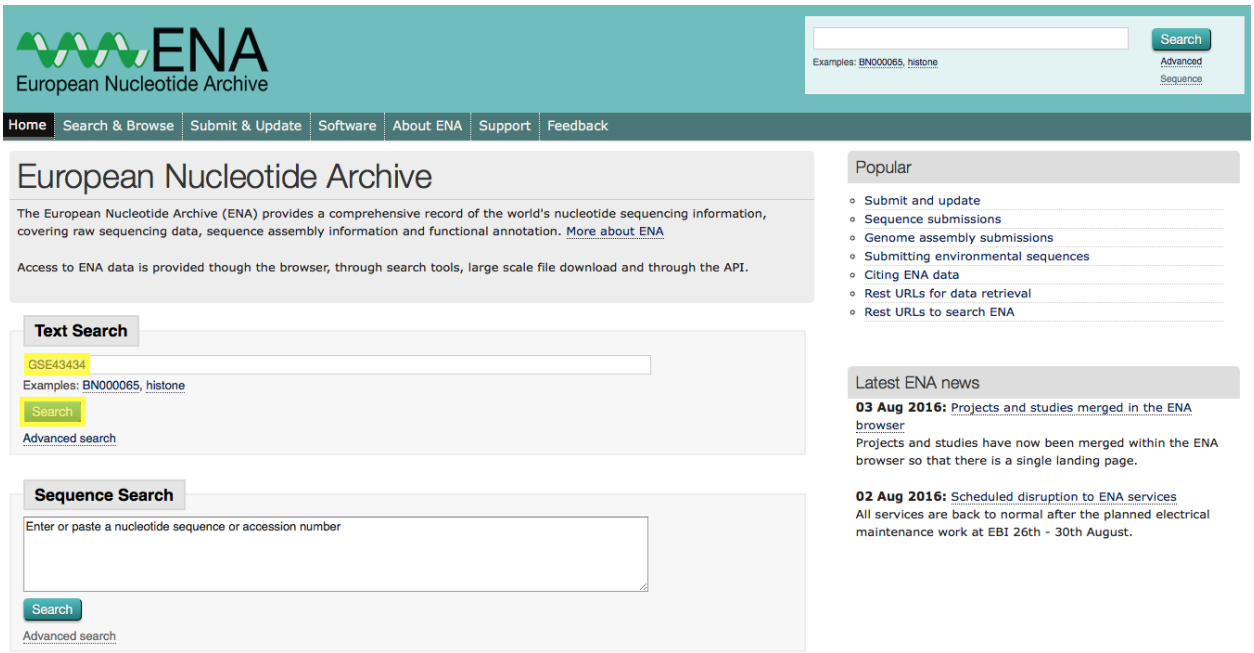


Figure 6:

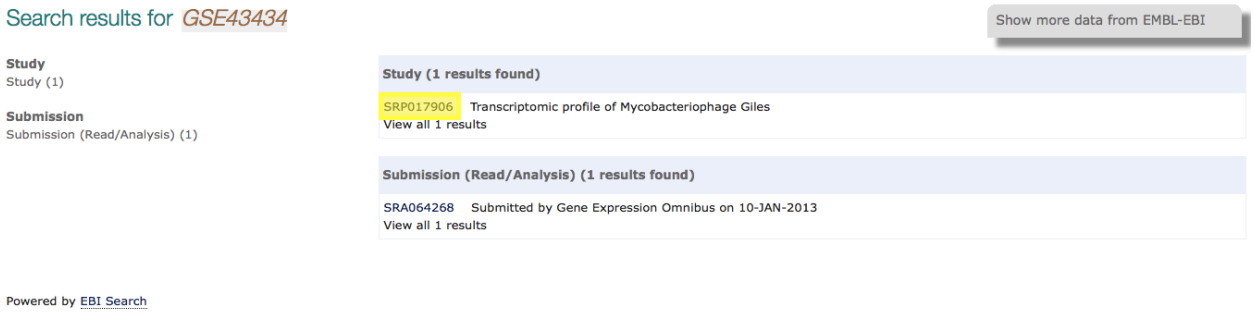
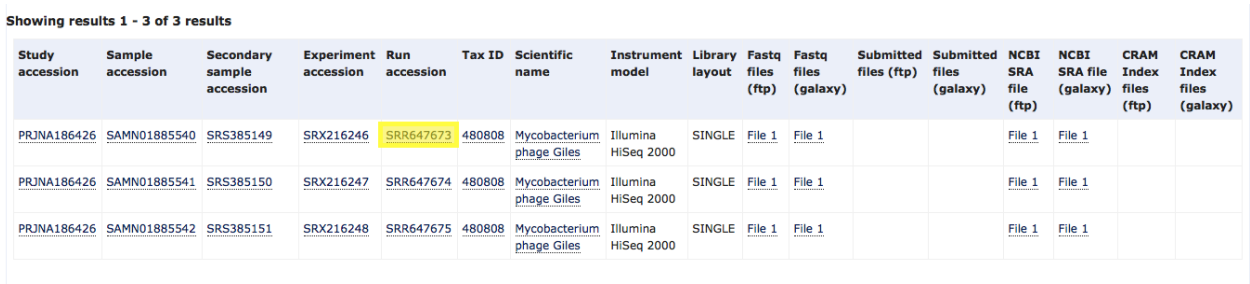


Figure 7:



Showing results 1 - 3 of 3 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)	Submitted files (ftp)	Submitted files (galaxy)	NCBI SRA file (ftp)	NCBI SRA file (galaxy)	CRAM Index files (ftp)	CRAM Index files (galaxy)
PRJNA186426	SAMN01885540	SRS385149	SRX216246	SRR647673	480808	Mycobacterium phage Giles	Illumina HiSeq 2000	SINGLE	File 1	File 1			File 1	File 1		
PRJNA186426	SAMN01885541	SRS385150	SRX216247	SRR647674	480808	Mycobacterium phage Giles	Illumina HiSeq 2000	SINGLE	File 1	File 1			File 1	File 1		
PRJNA186426	SAMN01885542	SRS385151	SRX216248	SRR647675	480808	Mycobacterium phage Giles	Illumina HiSeq 2000	SINGLE	File 1	File 1			File 1	File 1		

Figure 8:

Import into Galaxy

On the results page, make note of the `SSR-` values in the *Run accession* column. Each of these files is an individual run through the sequencing machine. Copy the first string you see (`SSR647673`).

Return to Galaxy. From the toolbar, select **NCBI SRA Tools**. Then select **Extract reads in FASTQ/A format from NCBI SRA**.

On this page, paste your copied run accession code (`SSR647673`) into the appropriate blank field. Then click execute.

This will add the file (as a pending job) to your history on the right. Not that this process can take quite a bit of time to complete. Repeat the above for the other three `SSR-` files (`SSR647674`,`SSR647675`). Note that these files will be in the `fastqsanger` format.

Step 3: Perform QC

Quality control is an important step in bioinformatic (and all general data analysis) pipelines. We will be using `FastQC`. FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

---

Running FastQC

From the Galaxy toolbar, click **NGS: QC and manipulation**. Then click **FastQC**.

Select one of your `FASTQ` files from the history to read in. Note that these files may have different referneces in your own history (in the example, the 3 `FASTQ` files imported from ENA are called 9: Extract Reads, 10: Extract Reads, and 11: Extra Reads). Then click **Execute**

Examine FastQC Output

You will see two new additions to your history bar: a FastQC "RawData" file, and a FastQC "Webpage". You can download this file to your desktop and examine the FastQC Output. The main file of interest is the `html` file. Upon opening, you should see something similar to this:


Galaxy
Tools
search tools
Get Data
Send Data
Lift-Over
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
NGS: RNA Analysis
NGS: SAMtools
NGS: BamTools
NGS: Picard
NGS: VCF Manipulation
NGS: Peak Calling
NGS: Variant Analysis
NGS: RNA Structure
NGS: Du Novo
NGS: Gemini
NGS: Assembly
Operate on Genomic Intervals
Statistics
Graph/Display Data
Phenotype Association
BEDTools
Genome Diversity
EMBOSS
Regional Variation
FASTA manipulation
Multiple Alignments
Metagenomic Analysis
Multiple regression
Multivariate Analysis
Motif Tools
STR-FM: Microsatellite Analysis
NCBI SRA Tools
Generate pileup format from NCBI sra.
Extract reads in SAM or BAM format from NCBI SRA.
Extract reads in FASTQ/A format from NCBI SRA.

Figure 9:

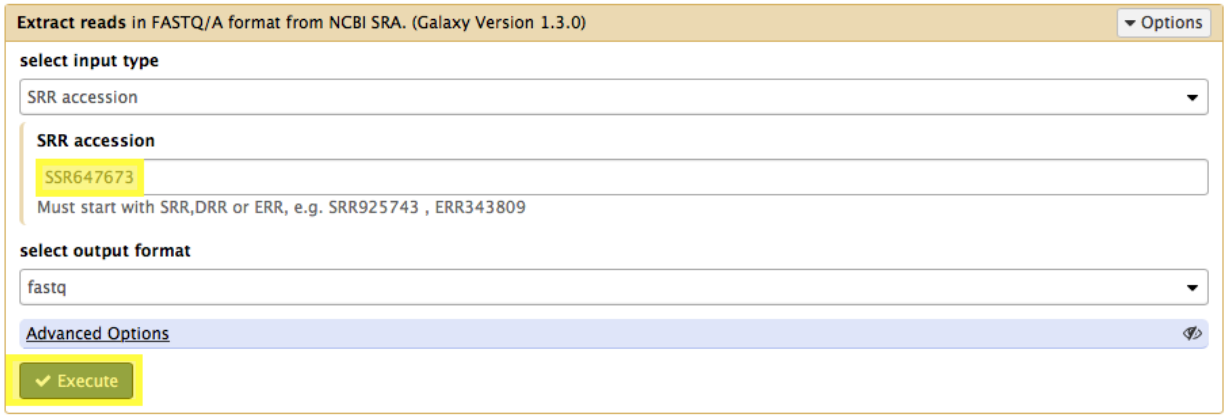


Figure 10:

Step 4: Map to the Genome

Mapping refers to the process of aligning short reads to a reference sequence, whether the reference is a complete genome, transcriptome, or de novo assembly. There are numerous programs that have been developed to map reads to a reference sequence that vary in their algorithms and therefore speed. The program that we utilize in this pipeline is called `bowtie`. More information available here

---

Running Bowtie

Click **NGS: Mapping**. Then click **Map with Bowtie for Illumina**.

In the first blank area ("*Will you select a reference genome...?*"), select **Use one from the history**. Then, choose your reference genome (`Mycobacterium_smegmatis_str_mc2_155.ASM1500v1.dna_sm.toplevel.fa.gz`). Leave the rest of the settings as they are - but make note of the `FASTQ` file that you are performing the mapping on, as you'll need to repeat this step for each of the three *FASTQ* files that you've loaded into Galaxy. When you're ready, click **Execute**.

Repeat this step for the remaining two `FASTQ` files. Note that your output files will be `SAM` files. For more information on `SAM/BAM` files, click here

---

Step 5: Generate Counts per Read

To perform differential analysis, it's necessary to be able to calculate the number of reads mapping to each feature. Here, we think of a feature as an interval (i.e., a range of positions) on a chromosome or a union of such intervals. In the case of RNA-Seq, the features are typically genes, where each gene is considered here as the union of all its exons. One may also consider each exon as a feature, e.g., in order to check for alternative splicing.

To perform this task, we will use the `htseq-count` program.


Galaxy
Tools
search tools
Get Data
Send Data
Lift-Over
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
NGS: QC and manipulation
Trim Galore! adaptive quality and adapter trimmer
FastQC Read Quality reports
FASTQ joiner on paired end reads
updated
Pear Paired-End read merger
new

Figure 11:

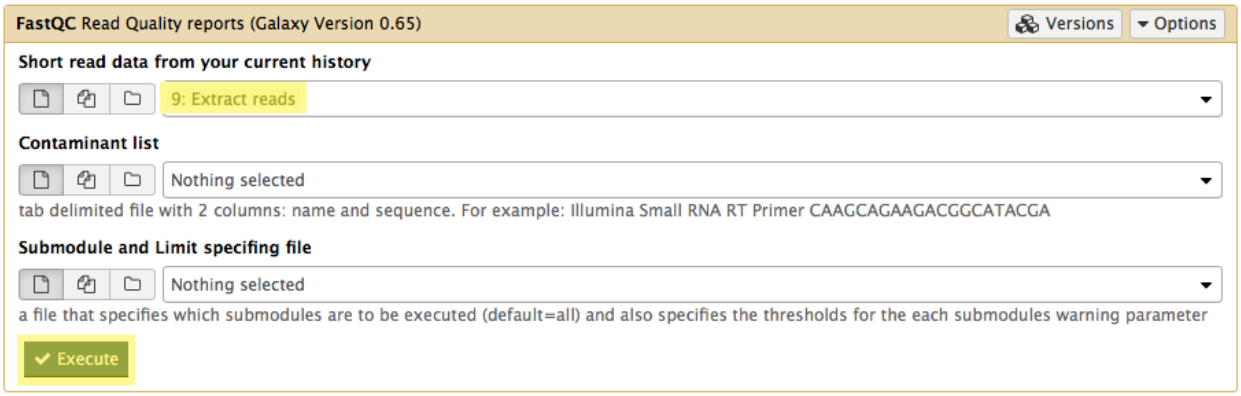
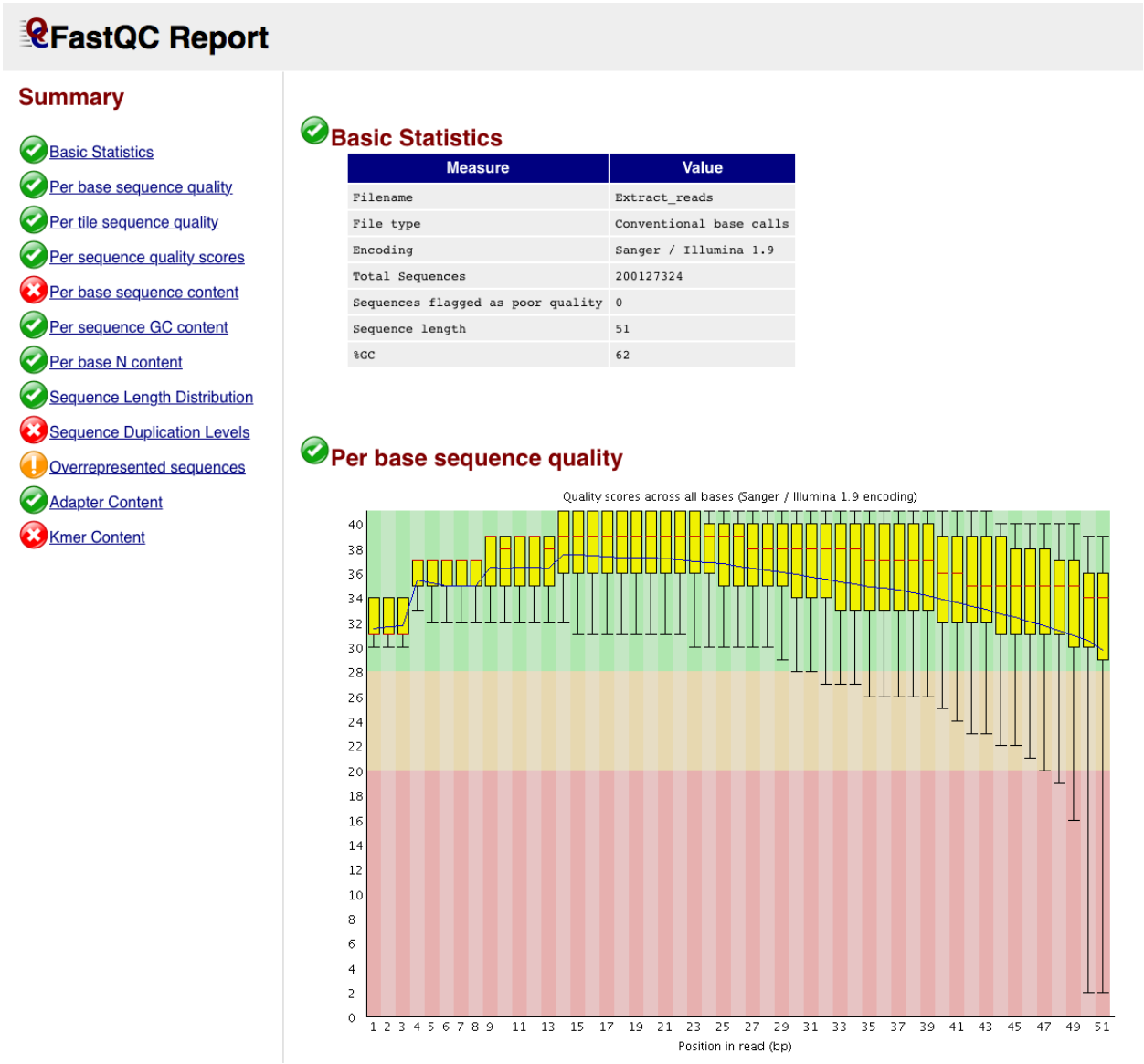


Figure 12:



Measure	Value
Filename	Extract_reads
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	200127324
Sequences flagged as poor quality	0
Sequence length	51
%GC	62

Figure 13:

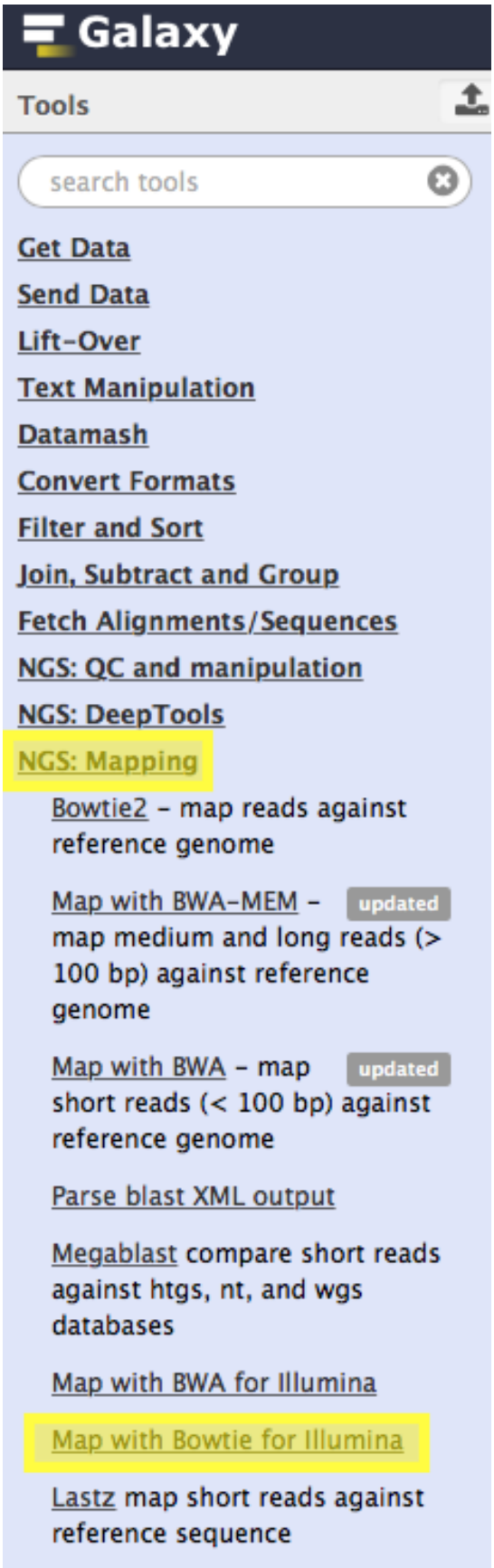

Galaxy
Tools
search tools
Get Data
Send Data
Lift-Over
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
Bowtie2 - map reads against reference genome
Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome
updated
Map with BWA - map short reads (< 100 bp) against reference genome
updated
Parse blast XML output
Megablast compare short reads against htgs, nt, and wgs databases
Map with BWA for Illumina
Map with Bowtie for Illumina
Lastz map short reads against reference sequence

Figure 14:

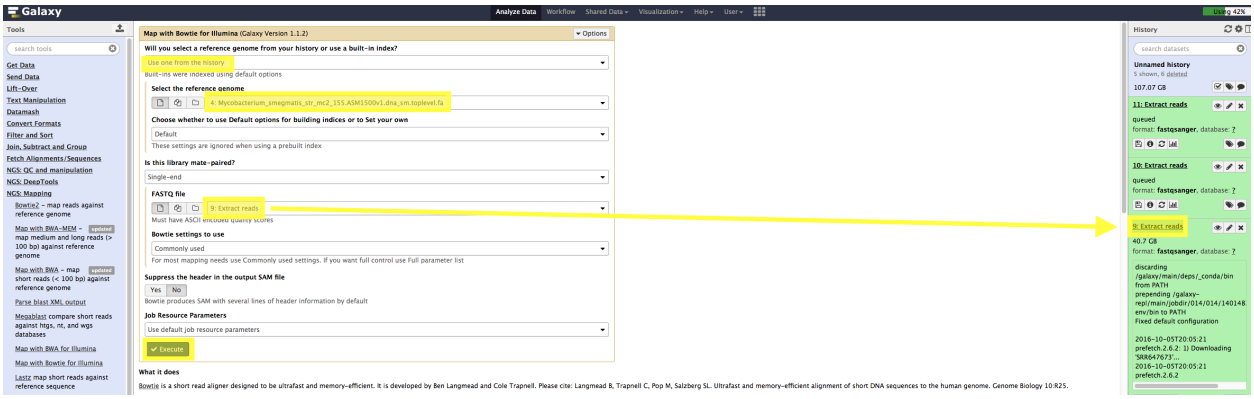


Figure 15:

Running htseq-count

From the toolbar, click **NGS: RNA Analysis**. Then click **htseq-count**.

In the first input ("Aligned SAM/BAM file"), select one of the `SAM` files generated from `bowtie`. Then, make sure your `gtf` file is in the second input ("GFF File"). Leave the rest of the parameters as they are. Click **Execute**.

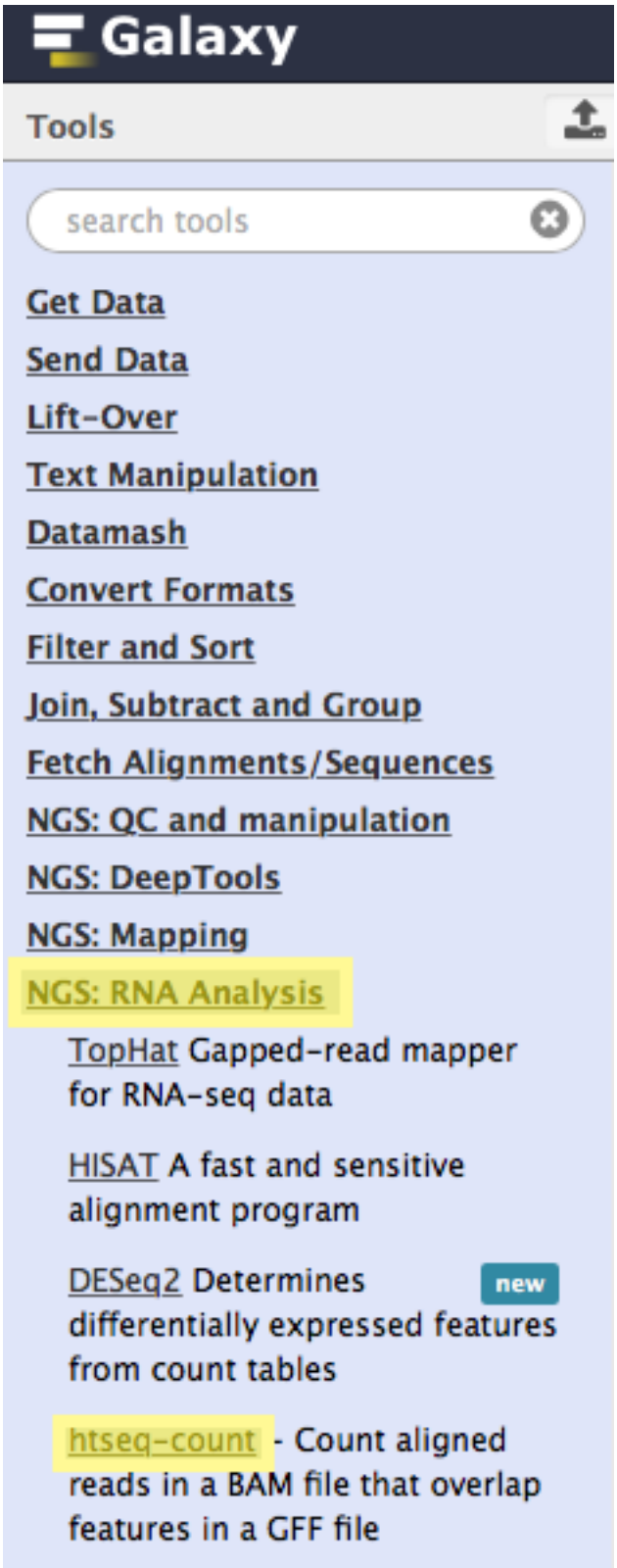

Galaxy
Tools
search tools
Get Data
Send Data
Lift-Over
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
NGS: RNA Analysis
TopHat Gapped-read mapper for RNA-seq data
HISAT A fast and sensitive alignment program
DESeq2 Determines differentially expressed features from count tables
new
htseq-count - Count aligned reads in a BAM file that overlap features in a GFF file

Figure 16:

htseq-count – Count aligned reads in a BAM file that overlap features in a GFF file (Galaxy Version 0.6.1galaxy1) Options

Aligned SAM/BAM File

14: Map with Bowtie for Illumina on data 11 and data 4: mapped reads

GFF File

19: Mycobacterium_smegmatis_str_mc2_155.ASM1500v1.32.gtf

Mode

Union

(--mode)

Stranded

Yes

(--stranded)

Minimum alignment quality

10

Skip all reads with alignment quality lower than the given minimum value. (-minaqual)

Feature type

exon

Feature type (3rd column in GFF file) to be used. All features of other types are ignored. The default, suitable for RNA-Seq and Ensembl GTF files, is exon. (--type)

ID Attribute

gene_id

GFF attribute to be used as feature ID. Several GFF lines with the same feature ID will be considered as parts of the same feature. The feature ID is used to identity the counts in the output table. All features of the specified type MUST have a value for this attribute. The default, suitable for RNA-Seq and Ensembl GTF files, is gene_id.

Additional BAM Output

Yes No

Write out all SAM alignment records into an output BAM file, annotating each line with its assignment to a feature or a special counter (as an optional field with tag 'XF').

Force sorting of SAM/BAM file by NAME

Yes No

This option can be used for for paired-end data that has many unmapped mates. Use this if you get the warning about paired end data missing or not being properly sorted.

Execute

Figure 17: