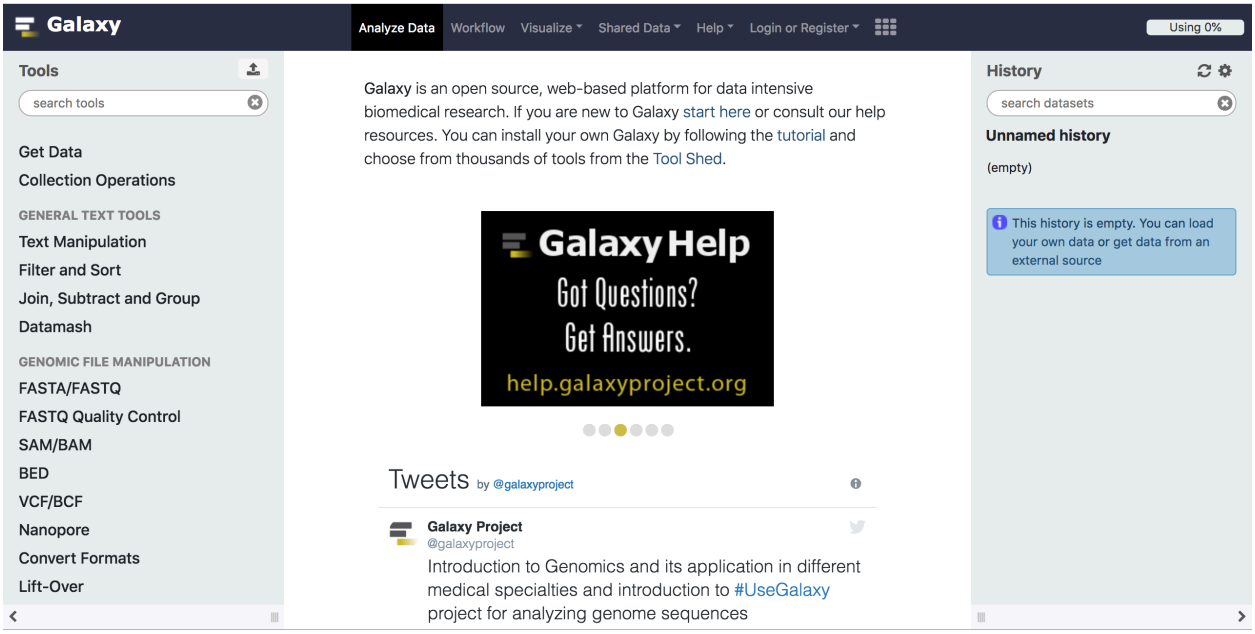
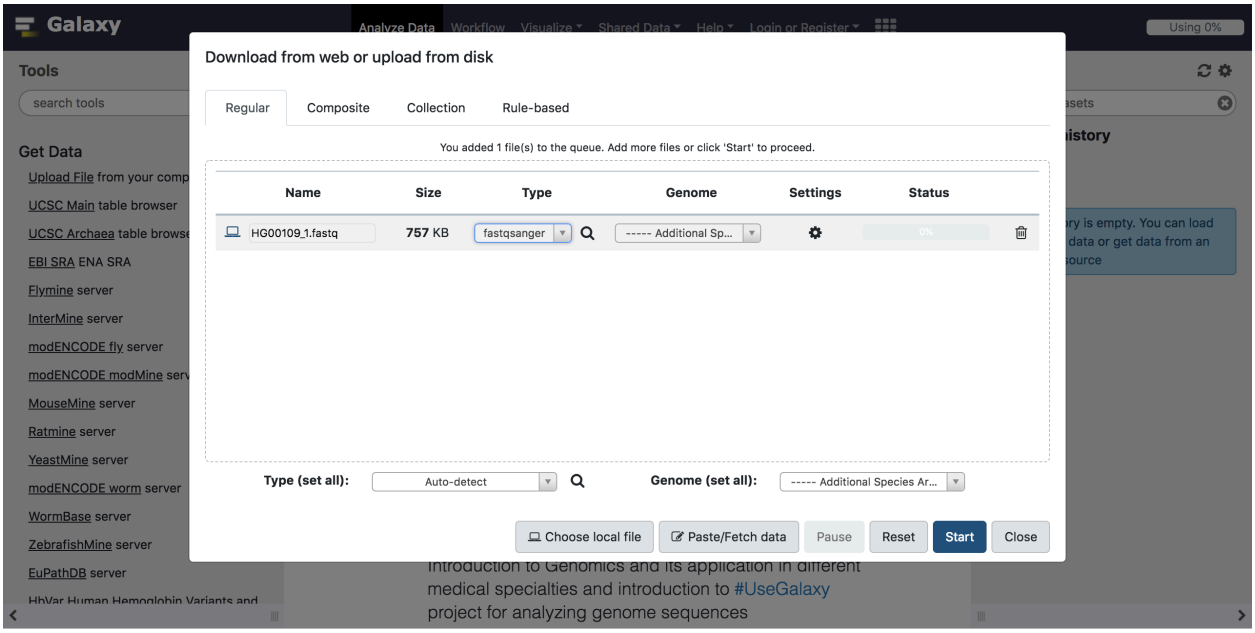
TUTORIAL: How to Use Galaxy for RNA-Seq Analysis

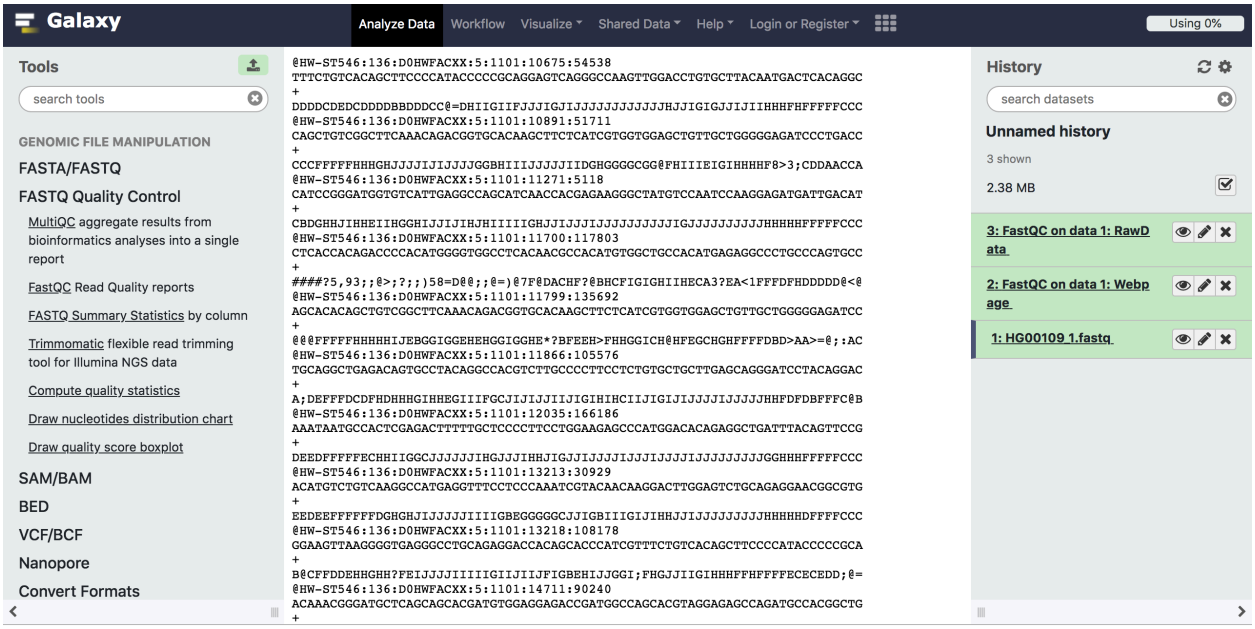
Here is a website (https://bioboot.github.io/bggn213_W19/class-material/lab-13-bggn213.pdf) presenting an alternative method for analyzing RNA-Seq data; Galaxy is a web-based interface (clicking and selecting), allowing one to avoid worrying about code on Terminal. The information is summarized below:



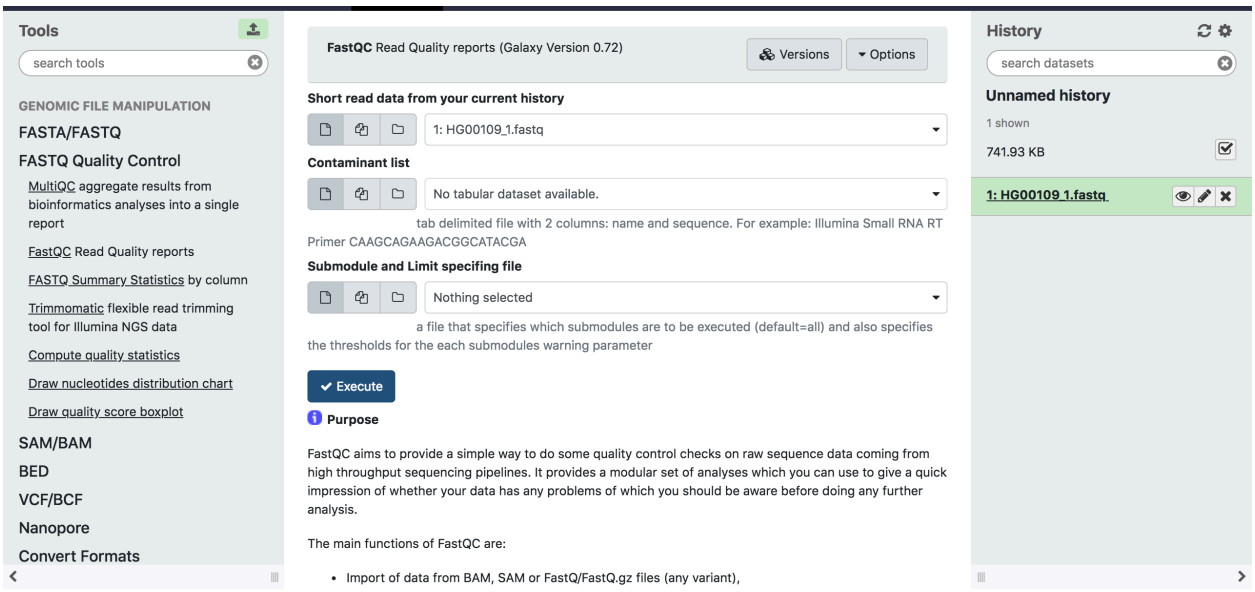
The flow chart (summarized in each step below, before the colon), is still the same as shown in the RNA-Seq section of the course-

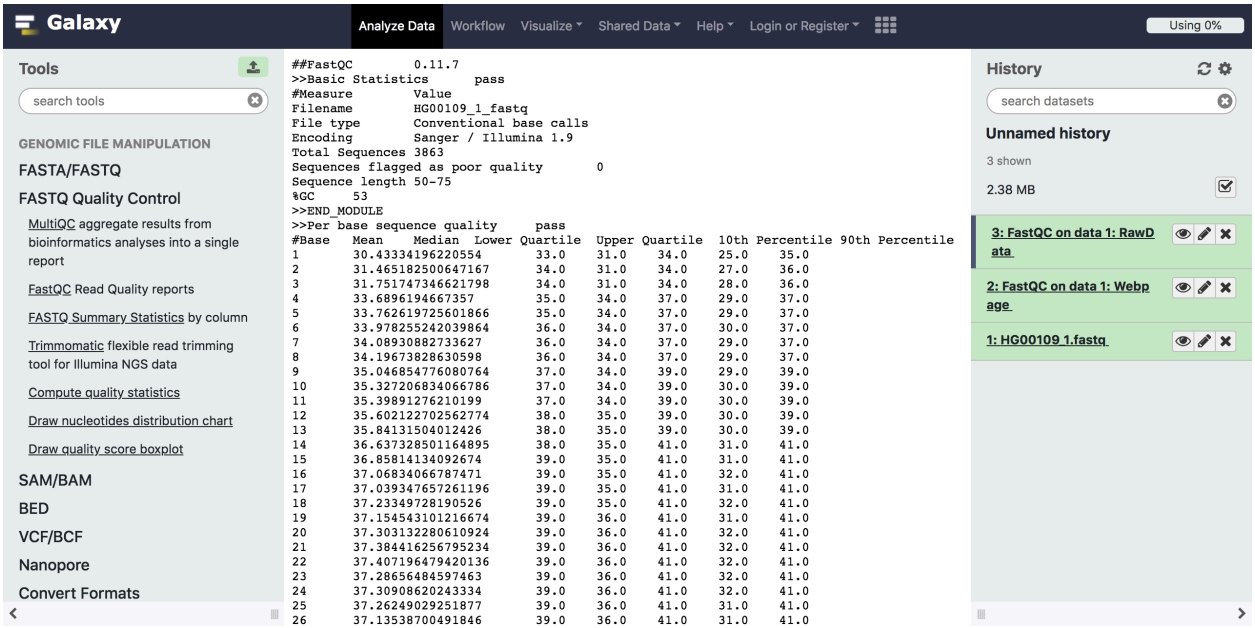
1. Upload files (typically fastq): In Galaxy, upload fastq files with fastqsanger as the type (used for Tophat, see below when aligning to genome), by looking at the left-hand panel (**TOOLS > Get Data > Upload File**); to confirm upload, look at the right-hand panel for a green box with the fastq file available and ready to be analyzed



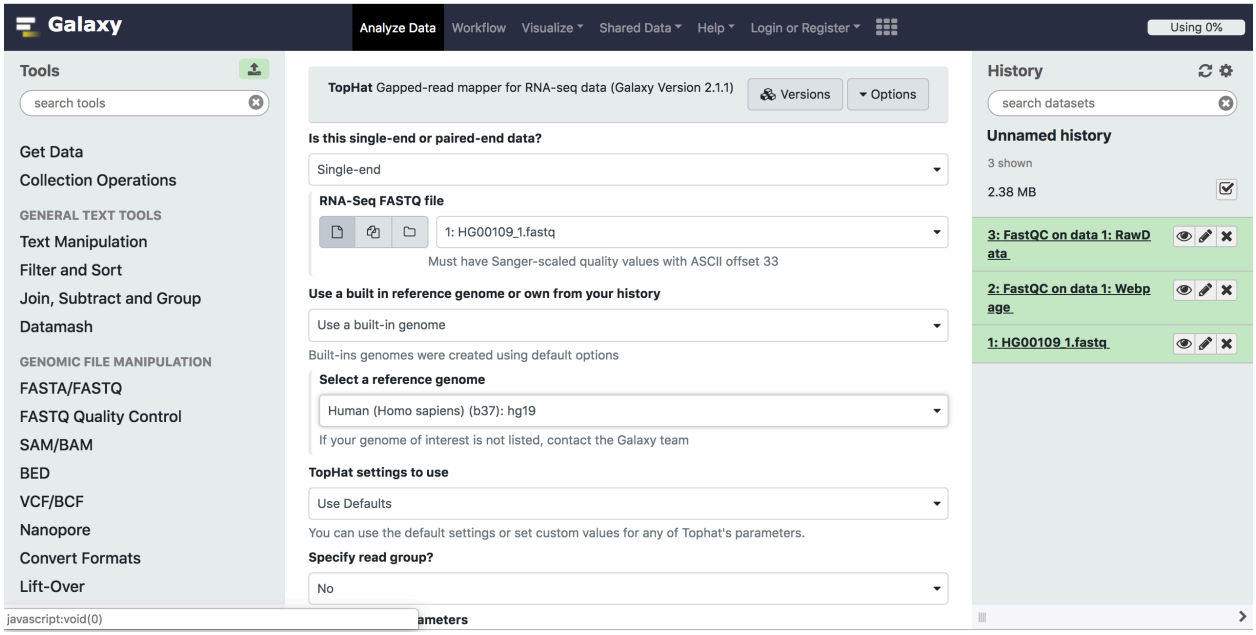


2. Check quality (fastqc): In Galaxy, select FastQC on the uploaded fastq files (**NGS: QC and manipulation > FastQC Read Quality reports**)- the output, a quality report to indicate if any positions have anomalous data (e.g., Phred), can be found on the right-hand panel

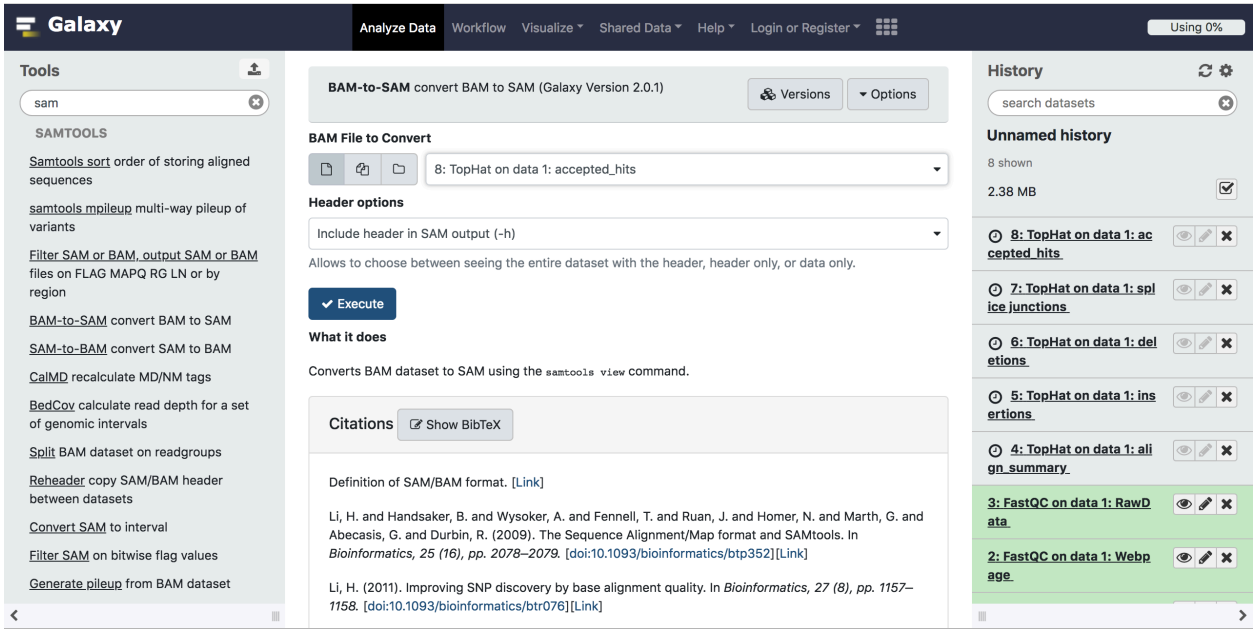


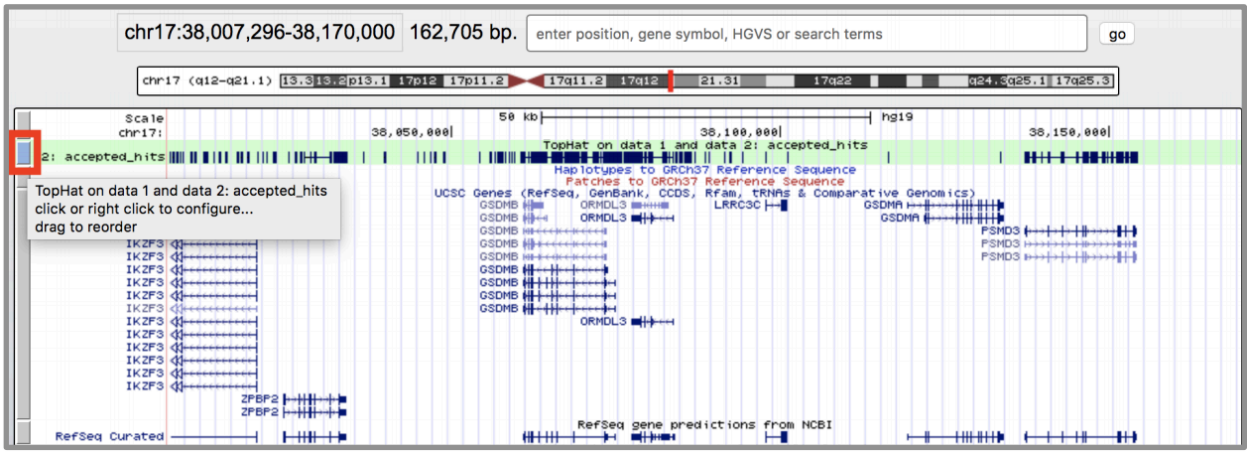


3. Align to genome (used STAR): In Galaxy, Tophat (**NGS: RNA Analysis > Tophat**) is used to map reads onto the genome (accounts for splice junctions)- select fastq files, single- or paired-end, mean inner distance between mate pairs, and reference genome (e.g., hg19)- this will give five outputs- accepted hits, insertions, deletions, splice junctions, and alignment history in BAM format

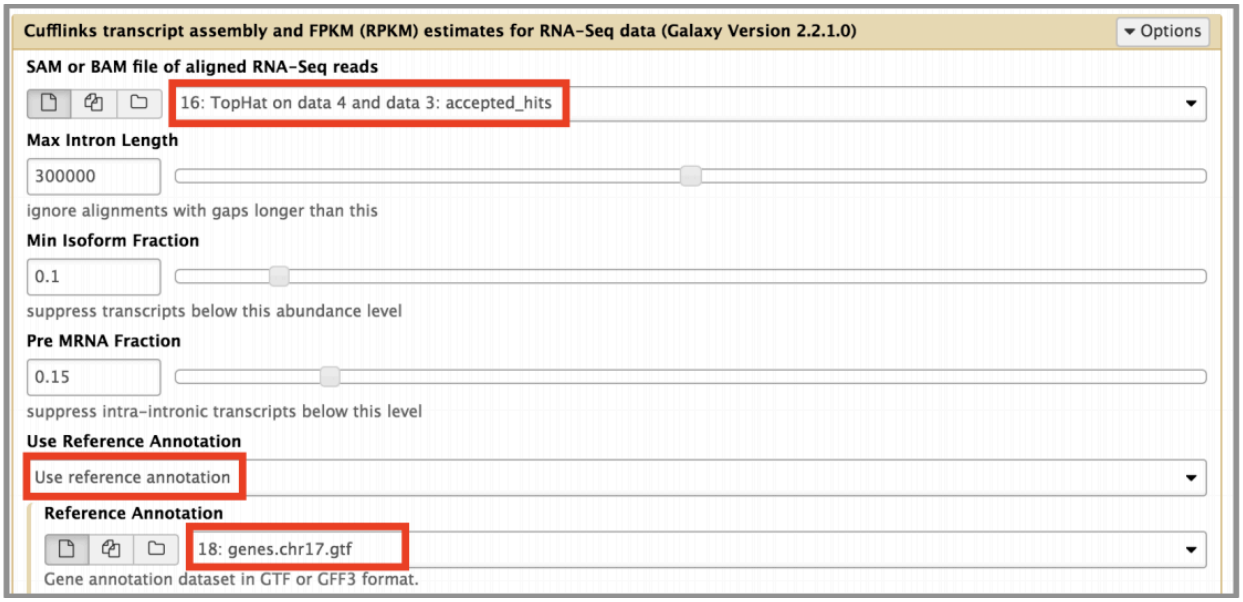


4. View BAM files (accepted hits) by converting to SAM files (**NGS: SAMtools > BAM-to-SAM**)- you can inspect the data results by clicking on "display at UCSC main" for the accepted hits file

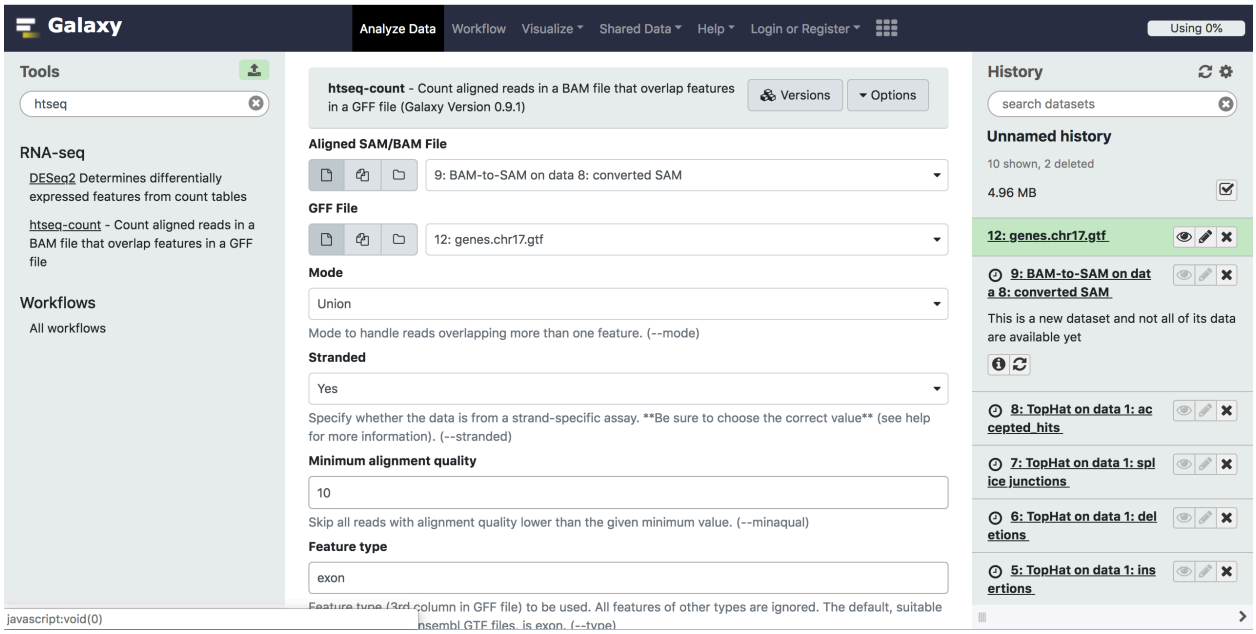




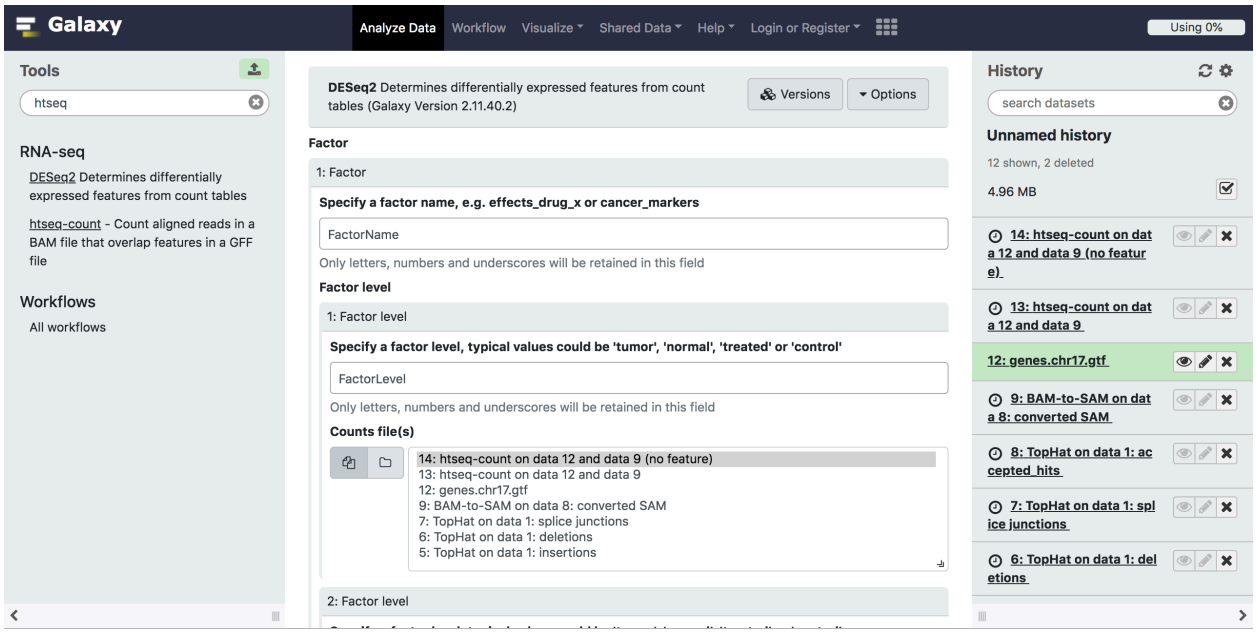
5. Sort and Index (used samtools): In Galaxy, to calculate gene expression, click on Cufflinks (**NGS: RNA Analysis, > Cufflinks**)- load the accepted hits file (BAM or SAM) from Tophat and the reference genome annotation (recall, this is the .gtf file; upload via **Upload File**)



6. Count reads mapped to genes (used featureCounts): In Galaxy, use **htseq-count,** imputing the accepted hits file and GFF file; additionally, select appropriate mode, strandedness, minimum alignment quality, feature type, ID attribute, as was done in class with choosing the appropriate flags



7. Differential expression (DESeq2): In Galaxy, **DESeq2** is also used for differential expression analysis- input the count data and select the appropriate conditions (factor, factor level, etc.), generating a tabular file and reporting



8. Interesting Biology- such as volcano plots to show proportion of genes that are both significantly regulated and have a high fold change (can be plotted in R/RStudio)

```
ggplot(as.data.frame(res), aes(log2FoldChange, -log10(pvalue), col=sig)) +
    geom_point() +
    ggtitle("Volcano plot")
```

```
## Warning: Removed 13578 rows containing missing values (geom_point).
```