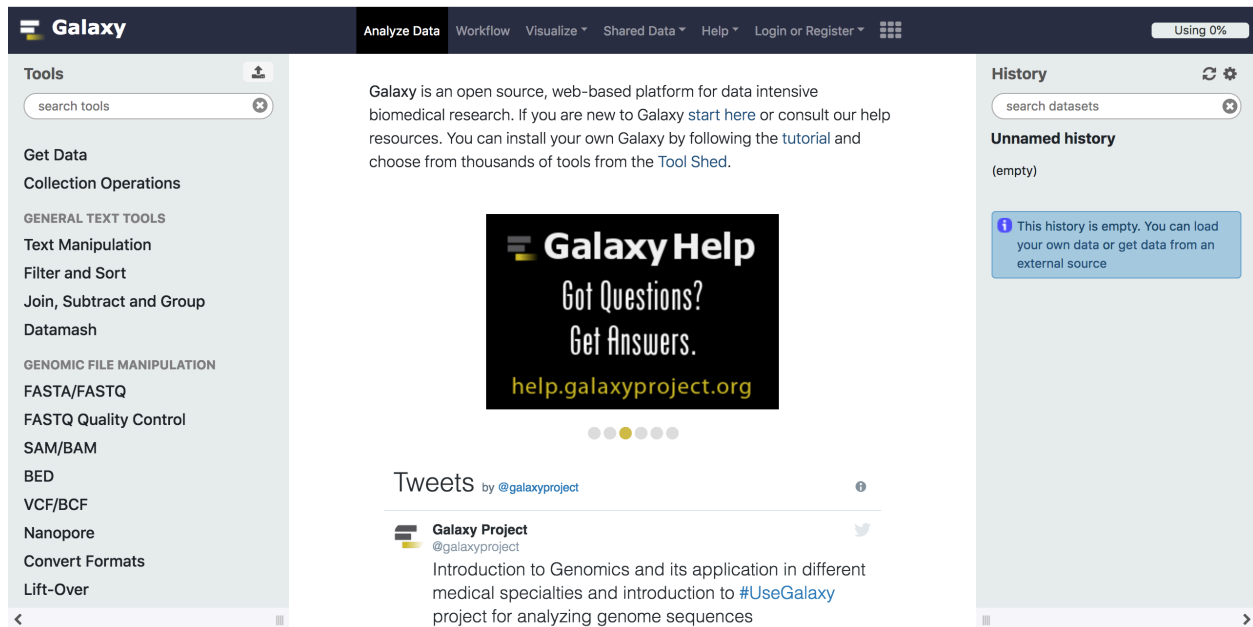


TUTORIAL: How to Use Galaxy for RNA-Seq Analysis

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

Tools

search tools

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

MultiQC aggregate results from bioinformatics analyses into a single report

FastQC Read Quality reports

FASTQ Summary Statistics by column

Trimmomatic flexible read trimming tool for Illumina NGS data

Compute quality statistics

Draw nucleotides distribution chart

Draw quality score boxplot

SAM/BAM

BED

VCF/BCF

Nanopore

Convert Formats

FastQC Read Quality reports (Galaxy Version 0.72)

VersionsOptions

Short read data from your current history

1: HG00109_1.fastq

Contaminant list

No tabular dataset available.

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifying file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Execute

Purpose

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/FastQ.gz files (any variant),

History

search datasets

Unnamed history

1 shown

741.93 KB

1: HG00109_1.fastq

Galaxy

Analyze DataWorkflowVisualizeShared DataHelpLogin or RegisterUsing 0%

Tools

search tools

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

MultiQC aggregate results from bioinformatics analyses into a single report

FastQC Read Quality reports

FASTQ Summary Statistics by column

Trimmomatic flexible read trimming tool for Illumina NGS data

Compute quality statistics

Draw nucleotides distribution chart

Draw quality score boxplot

SAM/BAM

BED

VCF/BCF

Nanopore

Convert Formats

```
##FastQC 0.11.7
>>Basic Statistics pass
#Measure Value
Filename HG00109_1.fastq
File type Conventional base calls
Encoding Sanger / illumina 1.9
Total Sequences 3863
Sequences flagged as poor quality 0
Sequence length 50-75
%GC 53
>>END_MODULE
>>Per base sequence quality
#Base Mean Median Lower Quartile Upper Quartile 10th Percentile 90th Percentile
1 30.43334196220554 33.0 31.0 34.0 25.0 35.0
2 31.465182500647167 34.0 31.0 34.0 27.0 36.0
3 31.751747346621798 34.0 31.0 34.0 28.0 36.0
4 33.6896194667357 35.0 34.0 37.0 29.0 37.0
5 33.762619725601866 35.0 34.0 37.0 29.0 37.0
6 33.978255242039864 36.0 34.0 37.0 30.0 37.0
7 34.08930882733627 36.0 34.0 37.0 29.0 37.0
8 34.19673828630598 36.0 34.0 37.0 29.0 37.0
9 35.046854776080764 37.0 34.0 39.0 29.0 39.0
10 35.327206834066786 37.0 34.0 39.0 30.0 39.0
11 35.39891276210199 37.0 34.0 39.0 30.0 39.0
12 35.602122702562774 38.0 35.0 39.0 30.0 39.0
13 35.84131504012426 38.0 35.0 39.0 30.0 39.0
14 36.637328501164895 38.0 35.0 41.0 31.0 41.0
15 36.85814134092674 39.0 35.0 41.0 31.0 41.0
16 37.06834066787471 39.0 35.0 41.0 32.0 41.0
17 37.039347657261196 39.0 35.0 41.0 31.0 41.0
18 37.23349728190526 39.0 35.0 41.0 32.0 41.0
19 37.154543101216674 39.0 36.0 41.0 31.0 41.0
20 37.303132280610924 39.0 36.0 41.0 32.0 41.0
21 37.384416256795234 39.0 36.0 41.0 32.0 41.0
22 37.407196479420136 39.0 36.0 41.0 32.0 41.0
23 37.28656484597463 39.0 36.0 41.0 32.0 41.0
24 37.30908620243334 39.0 36.0 41.0 32.0 41.0
25 37.26249029251877 39.0 36.0 41.0 31.0 41.0
26 37.13538700491846 39.0 36.0 41.0 31.0 41.0
```

History

search datasets

Unnamed history

3 shown

2.38 MB

3: FastQC on data 1: RawData

2: FastQC on data 1: Webpage

1: HG00109_1.fastq

- 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

Galaxy Analyze Data Workflow Visualize Shared Data Help Login or Register Using 0%

Tools search tools

Get Data
Collection Operations

GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Datamash

GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
Convert Formats
Lift-Over

TopHat Gapped-read mapper for RNA-seq data (Galaxy Version 2.1.1) Versions Options

Is this single-end or paired-end data?
Single-end

RNA-Seq FASTQ file
1: HG00109_1.fastq
Must have Sanger-scaled quality values with ASCII offset 33

Use a built in reference genome or own from your history
Use a built-in genome

Built-ins genomes were created using default options
Select a reference genome
Human (Homo sapiens) (b37): hg19
If your genome of interest is not listed, contact the Galaxy team

TopHat settings to use
Use Defaults

You can use the default settings or set custom values for any of Tophat's parameters.

Specify read group?
No

History search datasets

Unnamed history
3 shown
2.38 MB

3: FastQC on data 1: RawData
2: FastQC on data 1: Webpage
1: HG00109_1.fastq

- 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

Galaxy Analyze Data Workflow Visualize Shared Data Help Login or Register Using 0%

Tools sam

SAMTOOLS
Samtools sort order of storing aligned sequences
samtools mpileup multi-way pileup of variants
Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region
BAM-to-SAM convert BAM to SAM
SAM-to-BAM convert SAM to BAM
CalMD recalculate MD/NM tags
BedCov calculate read depth for a set of genomic intervals
Split BAM dataset on readgroups
Reheader copy SAM/BAM header between datasets
Convert SAM to interval
Filter SAM on bitwise flag values
Generate pileup from BAM dataset

BAM-to-SAM convert BAM to SAM (Galaxy Version 2.0.1) Versions Options

BAM File to Convert
8: TopHat on data 1: accepted_hits

Header options
Include header in SAM output (-h)
Allows to choose between seeing the entire dataset with the header, header only, or data only.

Execute

What it does
Converts BAM dataset to SAM using the samtools view command.

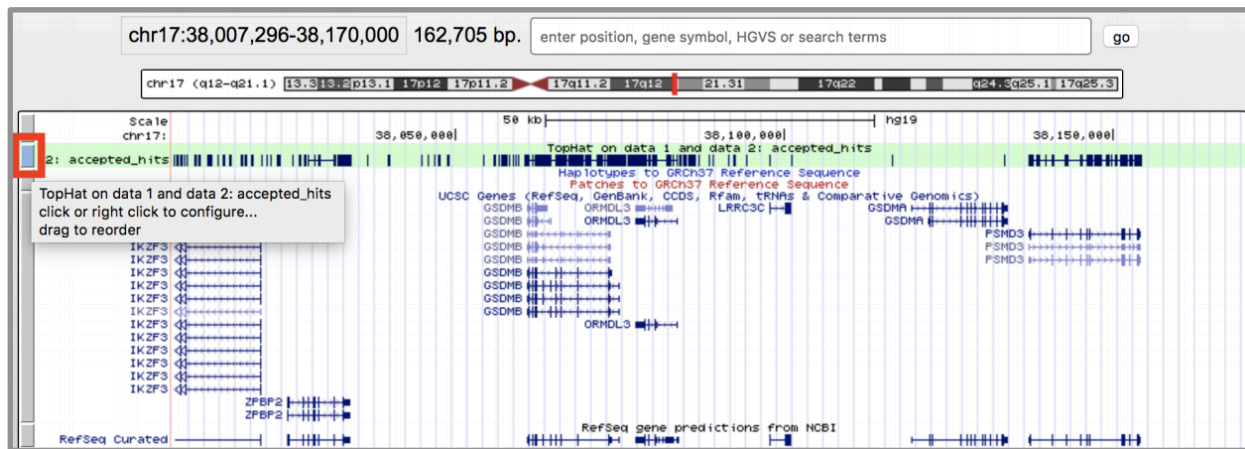
Citations Show BibTeX

Definition of SAM/BAM format. [Link]
Li, H. and Handsaker, B. and Wysoker, A. and Fennell, T. and Ruan, J. and Homer, N. and Marth, G. and Abecasis, G. and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. In *Bioinformatics*, 25 (16), pp. 2078–2079. [doi:10.1093/bioinformatics/btp352][Link]
Li, H. (2011). Improving SNP discovery by base alignment quality. In *Bioinformatics*, 27 (8), pp. 1157–1158. [doi:10.1093/bioinformatics/btr076][Link]

History search datasets

Unnamed history
8 shown
2.38 MB

8: TopHat on data 1: accepted_hits
7: TopHat on data 1: splice junctions
6: TopHat on data 1: deletions
5: TopHat on data 1: insertions
4: TopHat on data 1: alignment summary
3: FastQC on data 1: RawData
2: FastQC on data 1: Webpage



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**)

Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data (Galaxy Version 2.2.1.0) Options

SAM or BAM file of aligned RNA-Seq reads

16: TopHat on data 4 and data 3: accepted_hits ▼

Max Intron Length

ignore alignments with gaps longer than this

Min Isoform Fraction

suppress transcripts below this abundance level

Pre MRNA Fraction

suppress intra-intronic transcripts below this level

Use Reference Annotation

Use reference annotation ▼

Reference Annotation

18: genes.chr17.gtf ▼

Gene annotation dataset in GTF or GFF3 format.

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

The screenshot shows the Galaxy web interface with the **htseq-count** tool selected. The tool description is "Count aligned reads in a BAM file that overlap features in a GFF file (Galaxy Version 0.9.1)". The configuration fields are as follows:

- Aligned SAM/BAM File:** 9: BAM-to-SAM on data 8: converted SAM
- GFF File:** 12: genes.chr17.gtf
- Mode:** Union (Note: Mode to handle reads overlapping more than one feature. (--mode))
- Stranded:** Yes (Note: Specify whether the data is from a strand-specific assay. **Be sure to choose the correct value** (see help for more information). (--stranded))
- Minimum alignment quality:** 10 (Note: Skip all reads with alignment quality lower than the given minimum value. (--minqual))
- Feature type:** exon (Note: Feature type (3rd column in GFF file) to be used. All features of other types are ignored. The default, suitable for Ensembl GTF files, is exon. (--type))

The right sidebar shows the **History** panel with a search bar and a list of datasets. The current dataset is "12: genes.chr17.gtf".

- 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

The screenshot shows the Galaxy web interface with the **DESeq2** tool selected. The tool description is "Determines differentially expressed features from count tables (Galaxy Version 2.11.40.2)". The configuration fields are as follows:

- Factor:** 1: Factor
 - Specify a factor name, e.g. effects_drug_x or cancer_markers:** FactorName (Note: Only letters, numbers and underscores will be retained in this field)
- Factor level:** 1: Factor level
 - Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control':** FactorLevel (Note: Only letters, numbers and underscores will be retained in this field)
- Counts file(s):**
 - 14: htseq-count on data 12 and data 9 (no feature)
 - 13: htseq-count on data 12 and data 9
 - 12: genes.chr17.gtf
 - 9: BAM-to-SAM on data 8: converted SAM
 - 7: TopHat on data 1: splice junctions
 - 6: TopHat on data 1: deletions
 - 5: TopHat on data 1: insertions

The right sidebar shows the **History** panel with a search bar and a list of datasets. The current dataset is "12: genes.chr17.gtf".

- 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).
```

