Single/paired-end RNAseq analysis with Galaxy

Contents:

1. Introduction

RNA sequencing (RNAseq) is high-throughput mRNA/cDNA sequencing. RNAseq is used to discover new genes, splice variants, locate precise transcription product boundaries (junction between exons) and quantify expression genome-wide in a single experiment. RNAseq experiments span a variety of conditions such as: organism parts (e.g., antenna vs. maxillary palps), treatments (e.g., sugar vs. blood meal), development (e.g., embryo vs. larva vs. pupa vs. adult), and insecticide resistance (e.g., susceptible vs. resistant).

Most available RNAseq analysis tools are command-line driven. Galaxy is a web-based platform with a graphical user interface for users without UNIX skills, that can be used for data intensive research such as single and pair end RNAseq reads. To access VectorBase Galaxy, login using your VectorBase username and password.

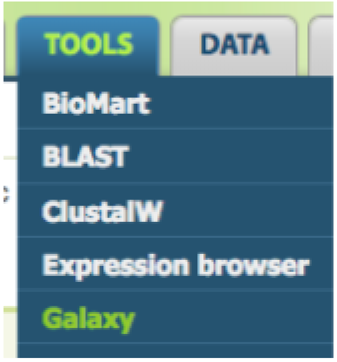


Figure 1. Access *Galaxy from the tools menu or following this link,* https://www.vectorbase.org/galaxy

In Galaxy you can manually select each one of the tools you need for data analyses or use a workflow. The workflows can be viewed and imported from the Shared Data tab and Workflows (Figure 2), on the list of Published Workflows as **RNA-Seq: paired-end FastQ to FPKM** and **RNA-Seq: single end FastQ to FPKM** (Figure 3). In these workflows the RNAseq reads are preprocessed to remove adapters (FastQC) and trim low quality reads (Trimmomatic), alignment is performed (TopHat) to produce BAM files, BAMs are then annotated and reads counted (Cufflinks).

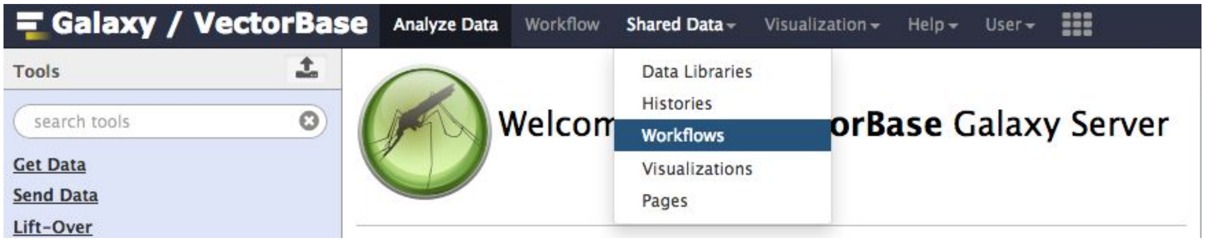
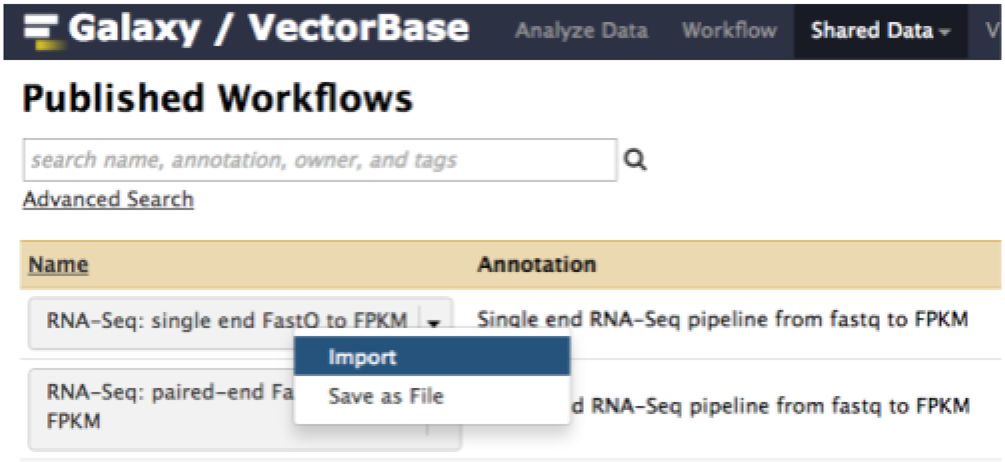


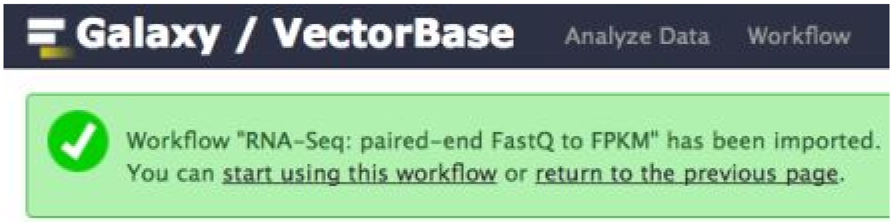
Figure 2. *VectorBase Galaxy home page*

Click on the arrow from the workflow(s) of interest and select import (Figure 3). A green message will confirm that the process was successful (Figure 4). The workflows can be imported to your history and ran.



Name	Annotation
RNA-Seq: single end FastQ to FPKM	Single end RNA-Seq pipeline from fastq to FPKM
RNA-Seq: paired-end Fa FPKM	d RNA-Seq pipeline from fastq to FPKM

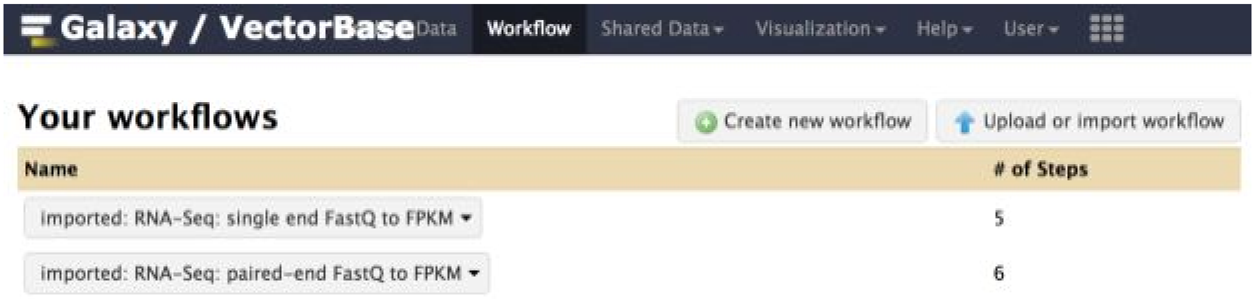
Figure 3. *Published workflows list*



Workflow "RNA-Seq: paired-end FastQ to FPKM" has been imported.
You can start using this workflow or return to the previous page.

Figure 4. *Galaxy green messages inform about successful processes*

Workflows can be visualised from the tab with the same name (Figure 6). From this screen you can also create a new workflow.



Your workflows

Create new workflow | Upload or import workflow

Name	# of Steps
imported: RNA-Seq: single end FastQ to FPKM ▾	5
imported: RNA-Seq: paired-end FastQ to FPKM ▾	6

Figure 6. *Your workflows*

Select 'Create a new workflow'. Using the workflow editor you can chain the tools of interest (Figure 7). Select the programs, match their inputs, outputs and set necessary parameters. A workflow allows to perform, reproduce and share complete analyses.

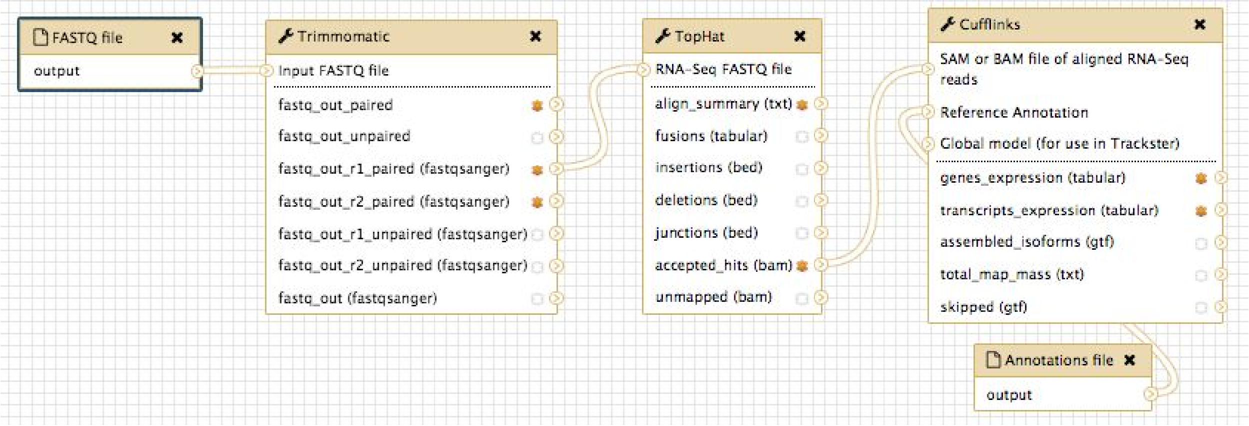


Figure 7. RNAseq single-end *Workflow Canvas*

2. Quality Control

FastQC[1] is not included in the workflow, however, it is recommended to run your FASTQ files through the Galaxy VB FastQC (provides an overview of read quality, GC content, sequence over-representation etc.), prior to running the workflow as this will give better understanding of the quality of your reads before trimming. The reads are also processed through **Trimmomatic** to trim off bad quality reads and adapters before alignment.

[1] This 11:33 min video demonstrates the use of FastQC, https://youtu.be/bz93ReOv87Y

3. Alignment

Once your FASTQ files are QCed, they're mapped to the reference genome using **TopHat** with default settings (tweak the parameters to suit your requirements). The workflows are set to "Use a genome from history", the following references are available in VB Galaxy (*A. aegypti* liverpool (AaegL3), *A. albimanus* (AalbS1), *A. albopictus* (AaloF1), *G. fuscipes* (Gfusl1), *R. microplus* (ADMZ02)). If you're aligning against a different genome, change the TopHat options to "Use a genome from history", and then upload a genome to your history and run. Genomes can be downloaded to your history from VB Galaxy (Shared Data > Data Libraries), alternatively, download your genome from VectorBase and upload to history, https://www.vectorbase.org/downloads.

The alignment produces a SAM file converted to BAM format which is put through Cufflinks to produce the normalized read counts. TopHat also produces an alignment statistics summary file, given the extension (.summary) in your workflow, this gives the total reads counts in your FASTQ files, reads aligned and percentages of aligned reads.

4. Normalization & Read Counts

The number of reads overlapping transcripts/genes are counted and normalized for exons and library size in **Cufflinks**, the output of Cufflinks are a transcript and gene-based FPKM files. With the FPKM file, you can do your differential expression and further analysis.

5. Workflow Overview

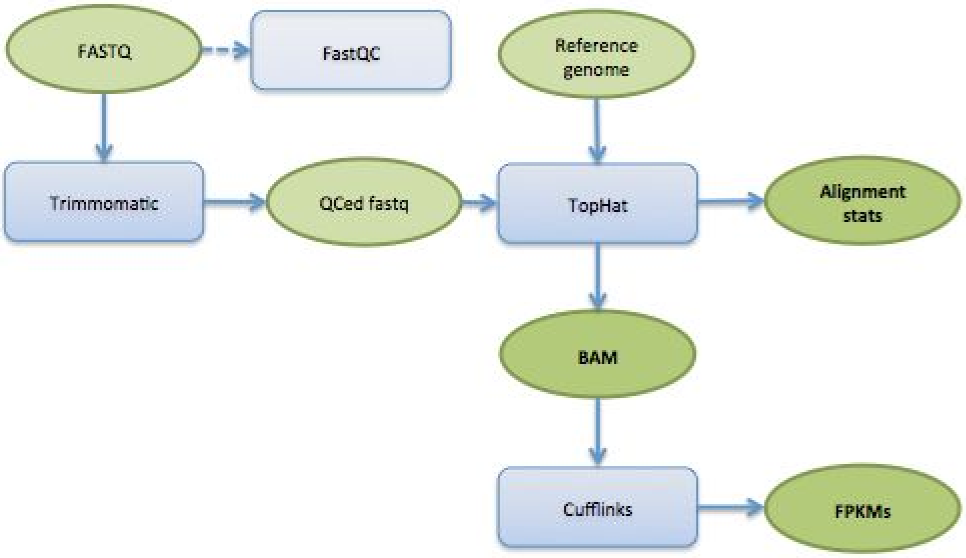


Figure 5. *Overview of the workflow with data files in green and software in blue. FastQC is recommended prior to running the workflow.*

- **Input files**: FASTQ, reference genome (fasta), annotation in gtf format.
- **Output files**: align summary, BAM, transcript-based FPKM, gene-based FPKM.
- **Software**: FastQC (Galaxy 0.63), Trimmomatic (Galaxy V 0.32.3), TopHat (Galaxy V 0.9), Cufflinks (Galaxy V 2.2.1.0).
- **FPKM**: fragments per kilobase per million reads (fragments are counted and divided by total length of all exons in gene/transcript), this value is then divided by library size (total number of reads in the sample).
- **BAM**: binary version of the aligned read.

5. Sample data set to test the paired-end workflow

To test the paired-end workflow we have a sample data set. Click on Shared data and Histories (Figure 6). Import to history (Figure 7) and set the workflow to run.

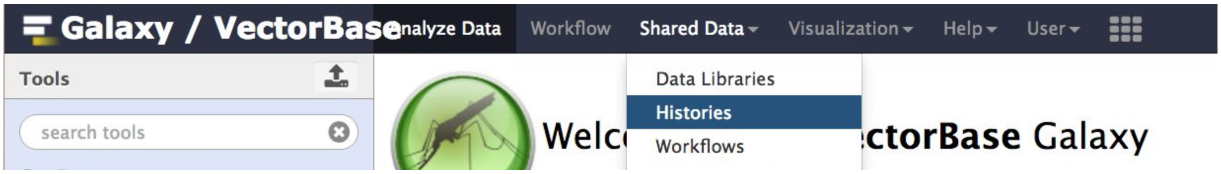
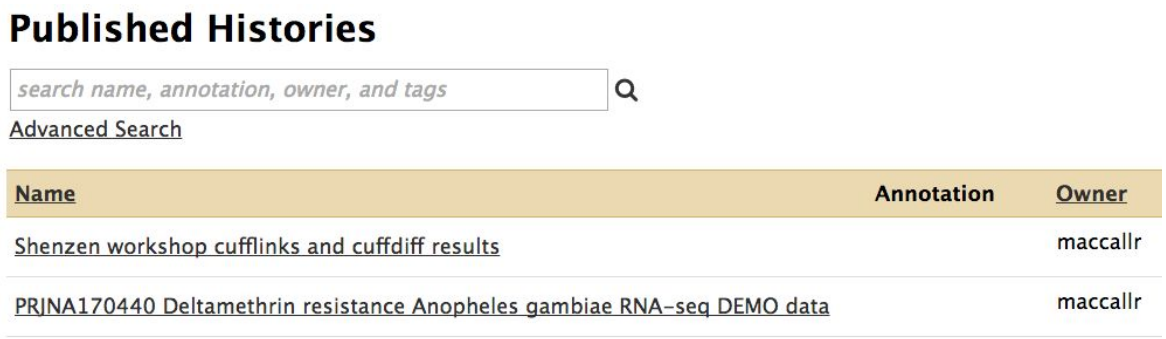


Figure 6. *Access shared data sets*



Published Histories

search name, annotation, owner, and tags

Advanced Search

Name	Annotation	Owner
Shenzen workshop cufflinks and cuffdiff results		maccallr
PRJNA170440 Deltamethrin resistance Anopheles gambiae RNA-seq DEMO data		maccallr

Figure 7. *Click on the second data set (PRJNA170440[2]) and select 'Import history'*

Note: If you need help with any step of this tutorial please contact us at info@vectorbase.org.

[2] Bonizzoni M, Afrane Y, Dunn WA, Atieli FK et al. Comparative transcriptome analyses of deltamethrin-resistant and -susceptible *Anopheles gambiae* mosquitoes from Kenya by RNA-Seq. PLoS One 2012;7(9):e44607. PMID: 22970263