RNA sequence data analysis in VEuPathDB Galaxy, Part I

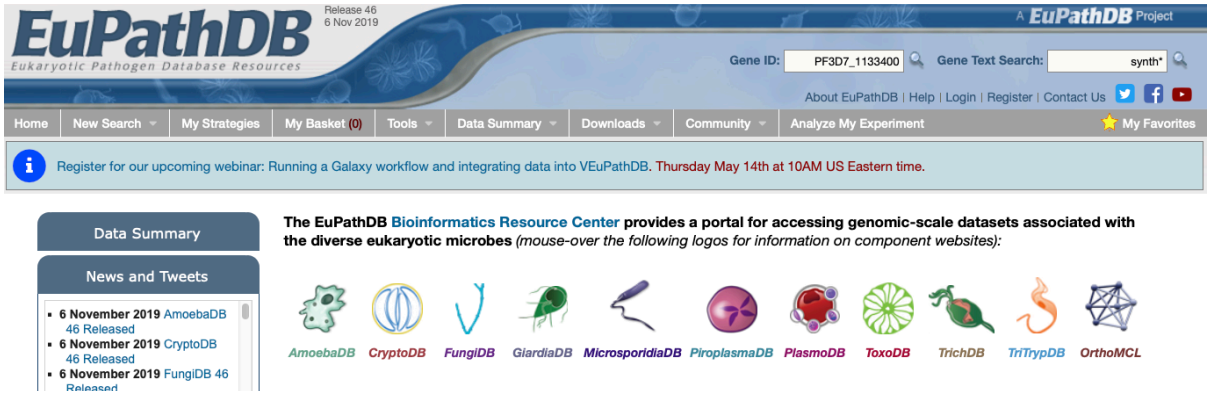
Galaxy is an open, web-based platform for data-intensive biomedical research. Galaxy allows you to perform, reproduce, and share complete analyses without the use of command-line scripting. The EuPathDB project, now known as VEuPathDB, developed its Galaxy instance in collaboration with Globus Genomics. To learn how to use Galaxy, follow this link to access tutorials prepared by the Galaxy Training Network: https://wiki.galaxyproject.org/Learn#Galaxy_101

Learning objectives:

1. Retrieve raw sequence data from the sequence repository EBI using Globus Data Transfer tool;
2. Run an RNA-Seq workflow for paired-end reads.

For this exercise, we will retrieve raw sequence files from a repository, assess the quality of the data, and then run the data through a workflow (or pipeline) that will align the data to a reference genome, calculate expression values and determine differential expression.

You will need to have a VEuPathDB account to use VEuPathDB Galaxy services. If you don't have an account, click on "Register" at the top right corner of the page to set up a free account. The username and password will work on any VEuPathDB site.



One you have an account, follow the steps outlined in the "Setting up your EuPathDB Galaxy account" tutorial to get started.

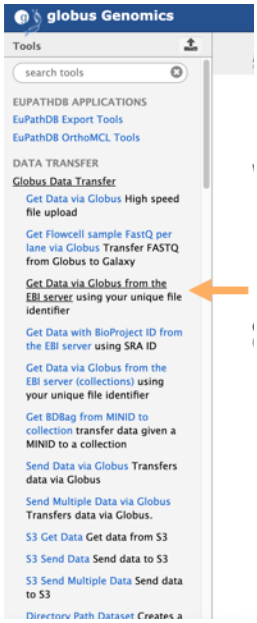
1. Retrieve raw sequence data from the sequence repository EBI using Globus Data Transfer tool;

There are multiple ways to important data into your Galaxy workspace. For this exercise, we will use the '**Get Data via Globus from the EBI: server using your unique file identifier**" tool and enter the sequence repository sample IDs

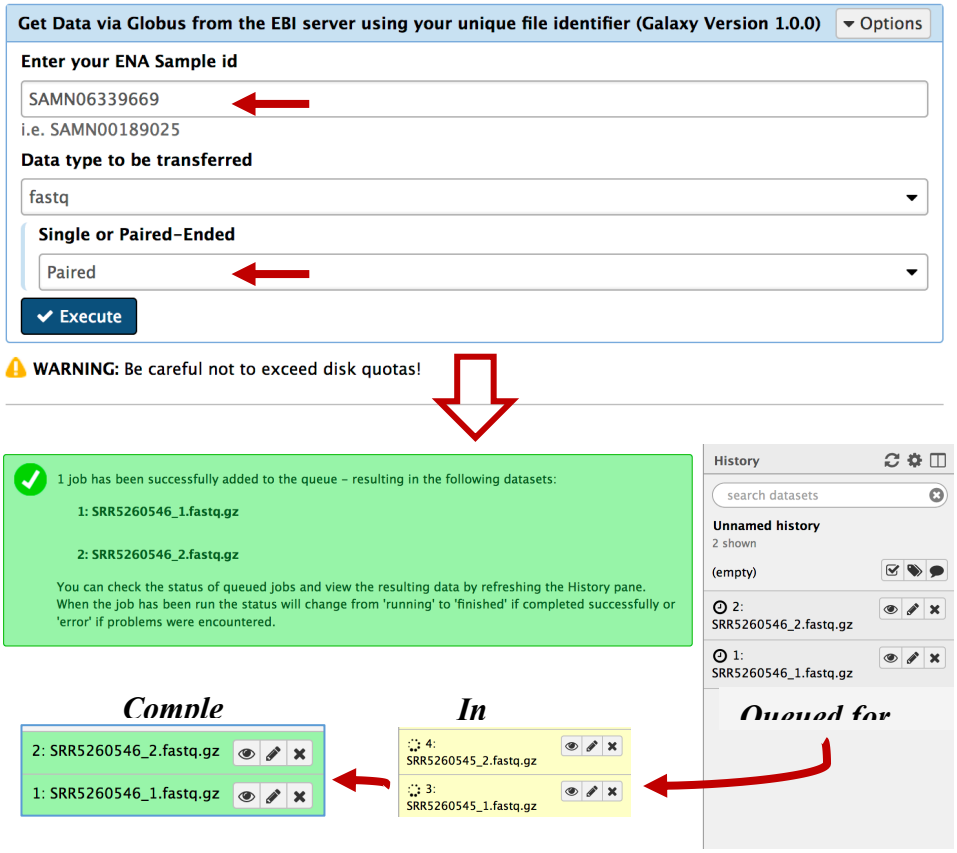
We will examine data from a study called "*Plasmodium berghei* transcriptome for female gametocytes, male gametocytes, and asexual erythrocytic stages"
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5604118/
The data is available in the sequence repositories:
https://www.ebi.ac.uk/ena/data/view/PRJNA374918

Sample Name	Erythrocyte stages (Asexual)	Male gametocytes	Comparison
Sample Accession Numbers	SAMN06339669 SAMN06339670 SAMN06339671	SAMN06339666 SAMN06339667 SAMN06339668	Erythrocyte stages vs. Male gametocytes

Step 1: Click on the "**Globus Data Transfer**" link in the left-hand menu. This will reveal a list of options; click on "**Get Data via Globus from the EBI server**". ***important: do not select the option for transferring a collection.

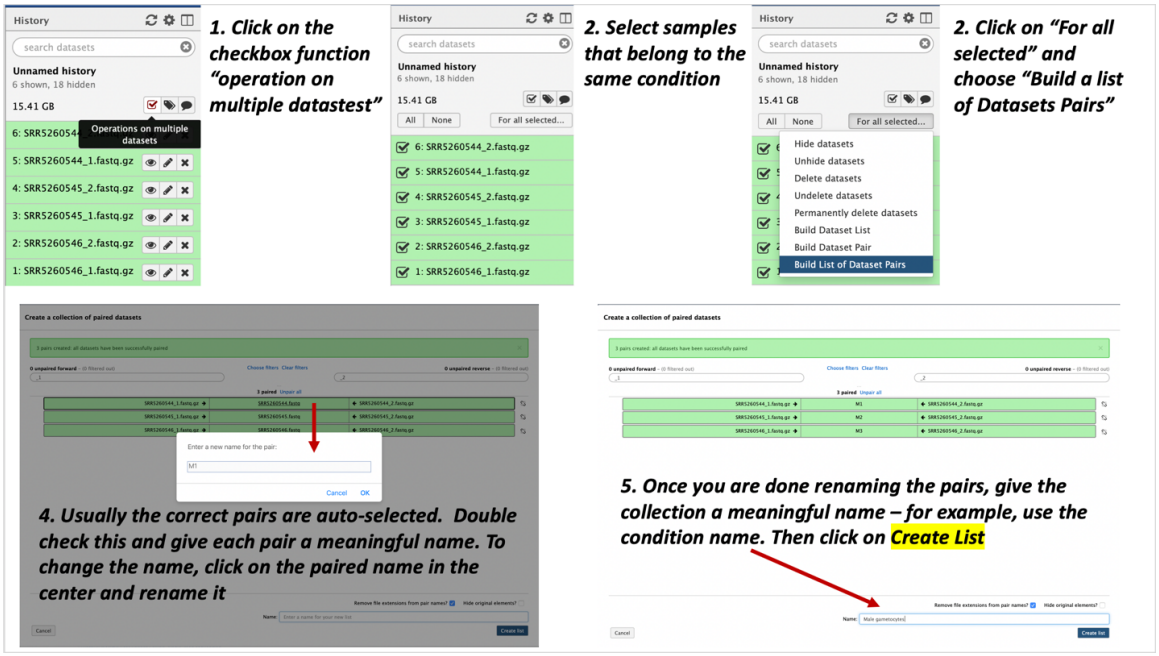


Step 2: In the middle section enter the sample ID and choose whether the run was single or paired end. Click on Execute.



Note that the sample ID resulted in importing two files one for each pair. Repeat this process for each sample you want to import. *If you are working with samples from two conditions and the experiment was done in triplicate and paired end sequenced then you should end up with 12 files; six from each condition.*

Step 3: If you are working with a dataset with biological replicates it is useful to organize the different conditions of your experiment into "Collections". For example, if your experiment included RNAseq from *Plasmodium falciparum* male gametocyte stages (three biological replicates) and erythrocytic stages (three biological replicates), it is useful to organize these into two collections, one that includes all male gametocyte files and the other that includes all the erythrocytic stage files. Using collections also reduces the complexity of the Galaxy workflows. See below:

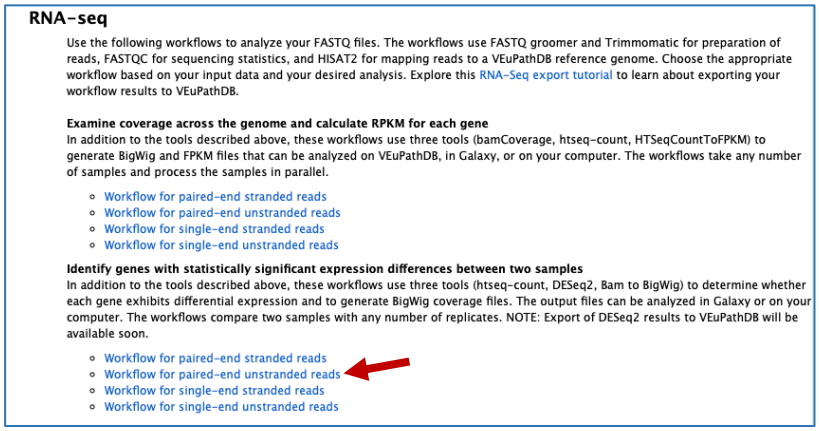


2. Running a workflow in VEuPathDB Galaxy

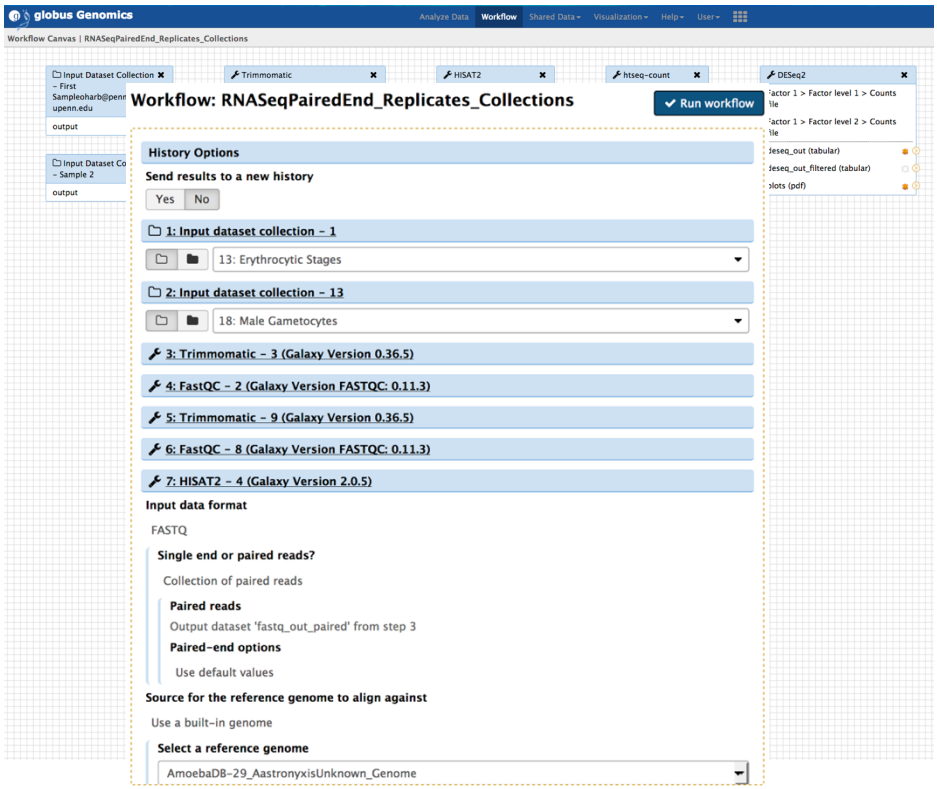
You can create your own workflows in galaxy based on your needs. The tools in the left section can all be added and configured as steps in a workflow that can be run on appropriate datasets. For this exercise we will use a preconfigured workflow that does the following main things:

1. Analyzes the reads in your files and generates FASTQC reports.
2. Trims the reads based on their quality scores and adaptor sequences (Trimmomatic).
3. Aligns the reads to a reference genome using HISAT2 and generates coverage plots.
4. Determines read counts per gene (HTSeq)
5. Determines differential expression of genes between samples (DESeq2).

To use one of the VEuPathDB preconfigured workflows, go to the VEuPathDB Galaxy home page and select the workflow that you would like to run. For this exercise, we will use "**Workflow for paired-end unstranded reads**" – click on this workflow to run it



Configure your workflow - there are multiple steps in the workflow but you do not need to configure all of them. For the purpose of this exercise you will need to configure the following:

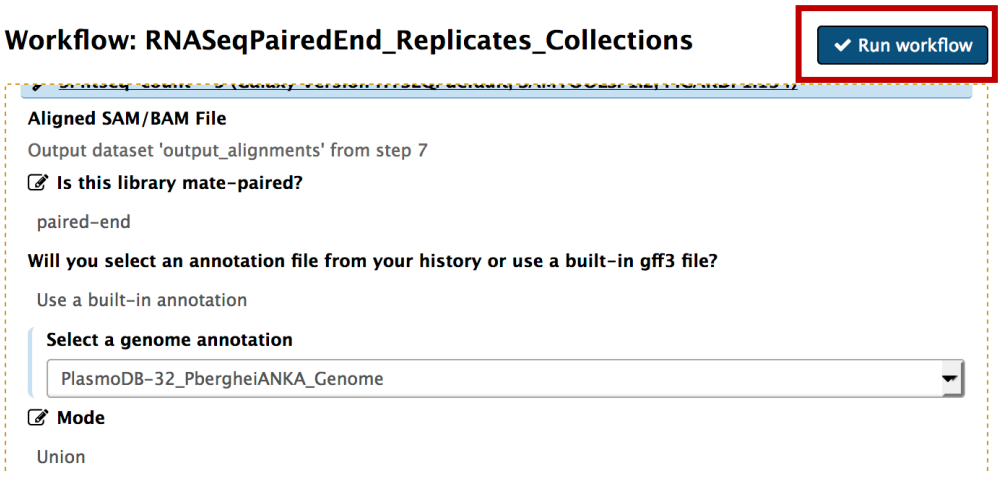


a. Select the input dataset collections. These are the collections of fastq files you just created. Workflow steps 1-2 allow you to select the datasets.

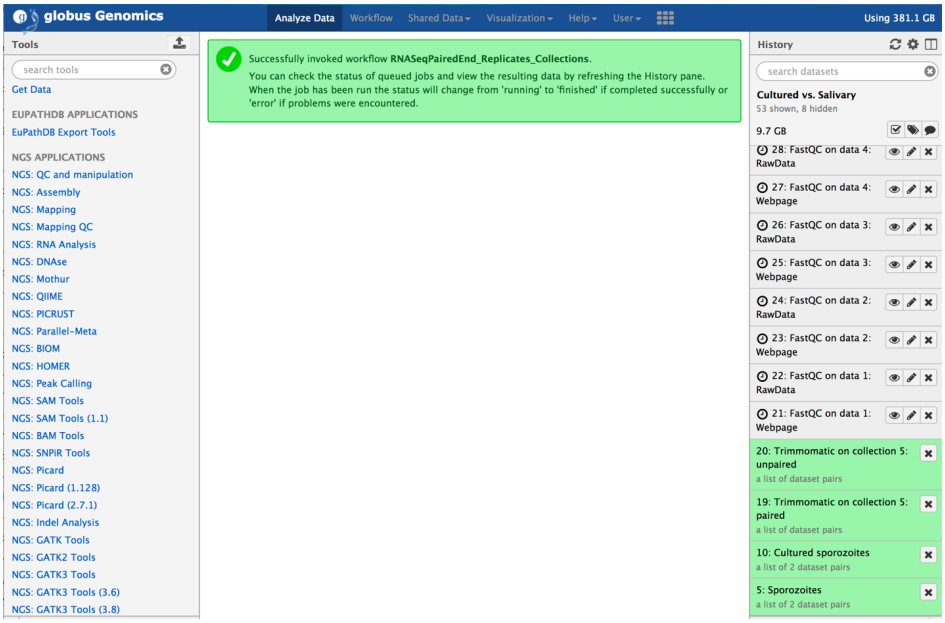
b. Some tools in the workflow require that you select the reference genome to be used. In this workflow both **HISAT2** and **HTSeq** require this (note these tools are in the workflow twice since you have two collections). It is critical that you select the correct genome that matches the experimental organism. For example, if your experiment was performed using *Plasmodium berghei*, the reference genome you select should be *Plasmodium berghei*.

c. Another very important parameter to check in the htseq-count step is the Feature type. The default is usually set to exon. Make sure you change this to **gene**. To change this to gene, click on the edit icon, the type the word "gene". This is case sensitive so be careful about this.

d. Once you are sure everything is configured correctly, click on "Run Workflow" at the top.



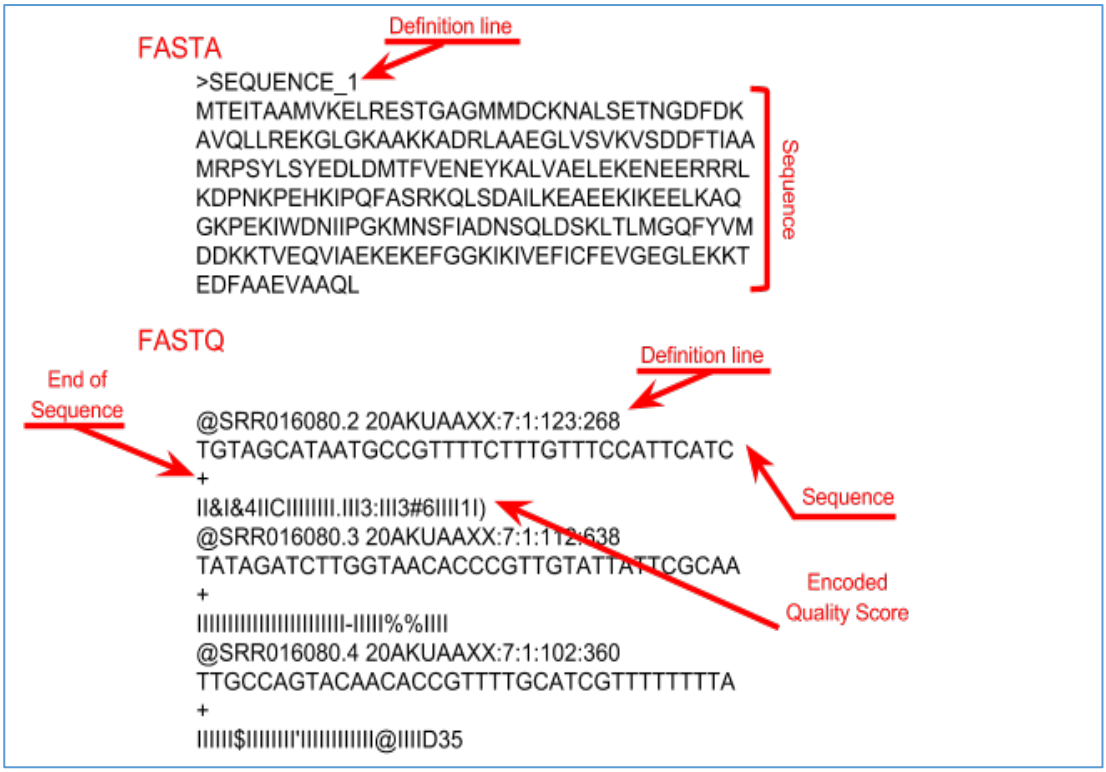
The steps will start running in the history section on the right. Grey means they are waiting to start. Yellow means they are running. Green means they have completed. Red means there was an error in the step.



Appendix:

FASTQ files are text files (similar to FASTA) that include sequence quality information and details in addition to the sequence (ie. name, quality scores, sequencing machine ID, lane number etc.). FASTQ files are large and as a result not all sequencing repositories will store this format. However, tools are available to convert, for example, NCBI's SRA format to FASTQ. Sequence data is housed in three repositories that are synchronized on a regular basis.

- The sequence read archive at GenBank
- The European Nucleotide Archive at EMBL
- The DNA data bank of Japan



Additional resources (tool manuals):

Trimmomatic
FastQC
HISAT2
HTseq
DEseq2