RNA sequence data analysis via Galaxy, Part I Uploading data and starting the workflow (Group Exercise)

The goal of this exercise is to use a Galaxy workflow to analyze RNA sequencing data. 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. EuPathDB developed its own Galaxy instance in collaboration with Globus Genomics. Many resources are available to learn how to use Galaxy. The following link has information about additional resources to help you learn how to use Galaxy:

https://wiki.galaxyproject.org/Learn#Galaxy_101

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, calculate expression values and determine differential expression. Part 1, uploading data and starting the workflow will be performed today. The workflows will run overnight and we will view / interpret the results tomorrow in Part 2.

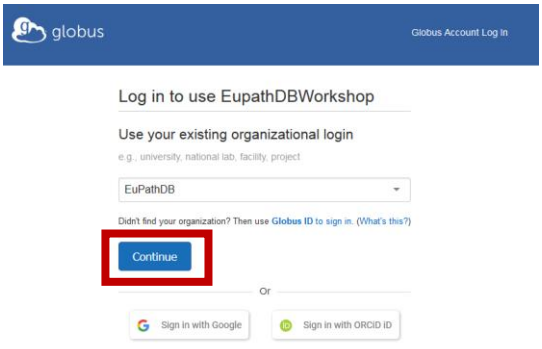
We will be working in groups. Each group will have 4-6 members. One person in the group will run the Galaxy controls on one computer. The other members' roles are to ensure that the correct datasets are used and that the correct workflow parameters are selected.

Section I: Setting up your EuPathDB Galaxy account

Step 1: Access the EuPathDB Galaxy instance at the following URL:

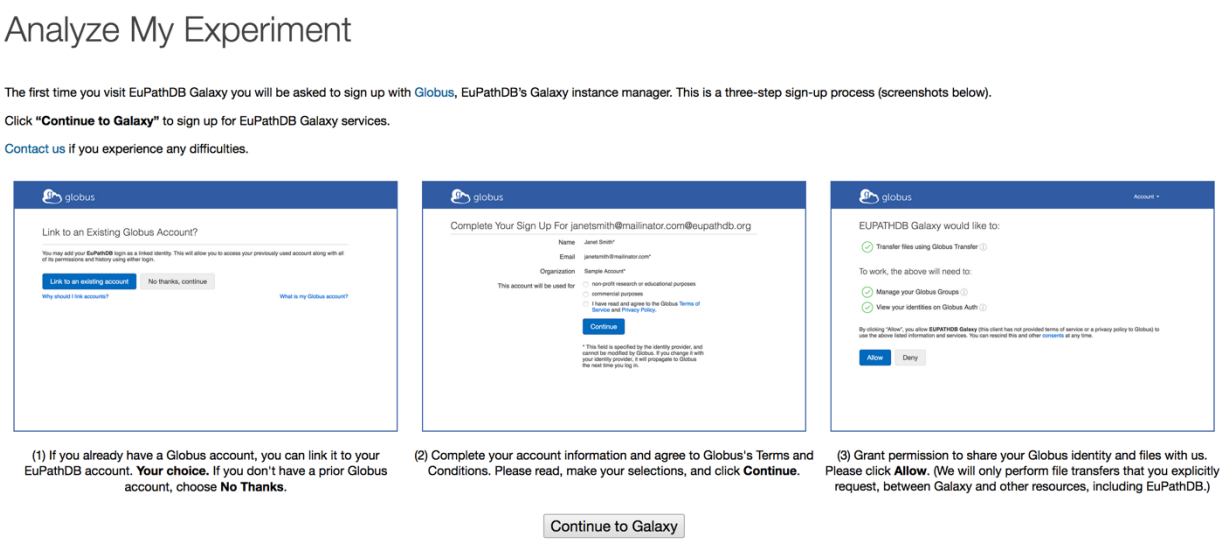
http://eupathdbworkshop.globusgenomics.org/

Step 2: On the next page you will be asked to define your organization. Choose EuPathDB and click Continue.

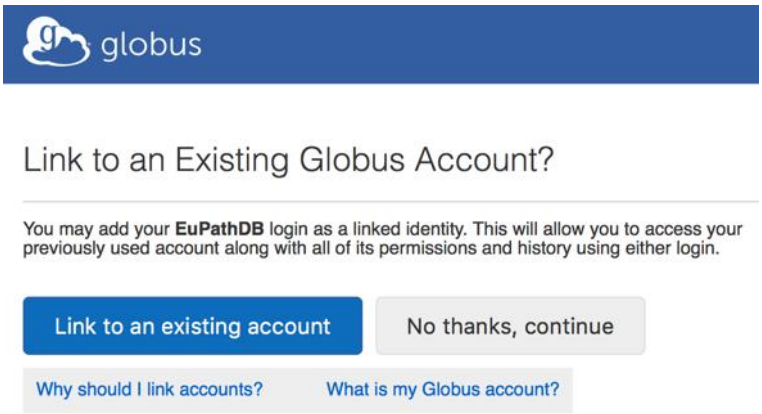


Step 3: Log in to EuPathDB (if you are not logged in already).

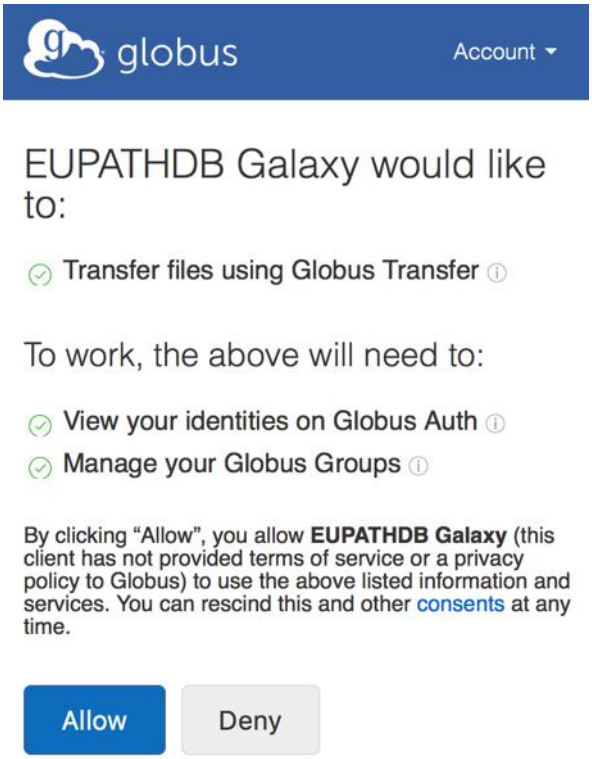
Step 4: Next, sign up for the EuPathDB Galaxy instance.



Step 5: Click on "Continue to Galaxy" and follow the instructions.
Step 6: Click on "No thanks, continue"



Step 7: Click on “Allow”

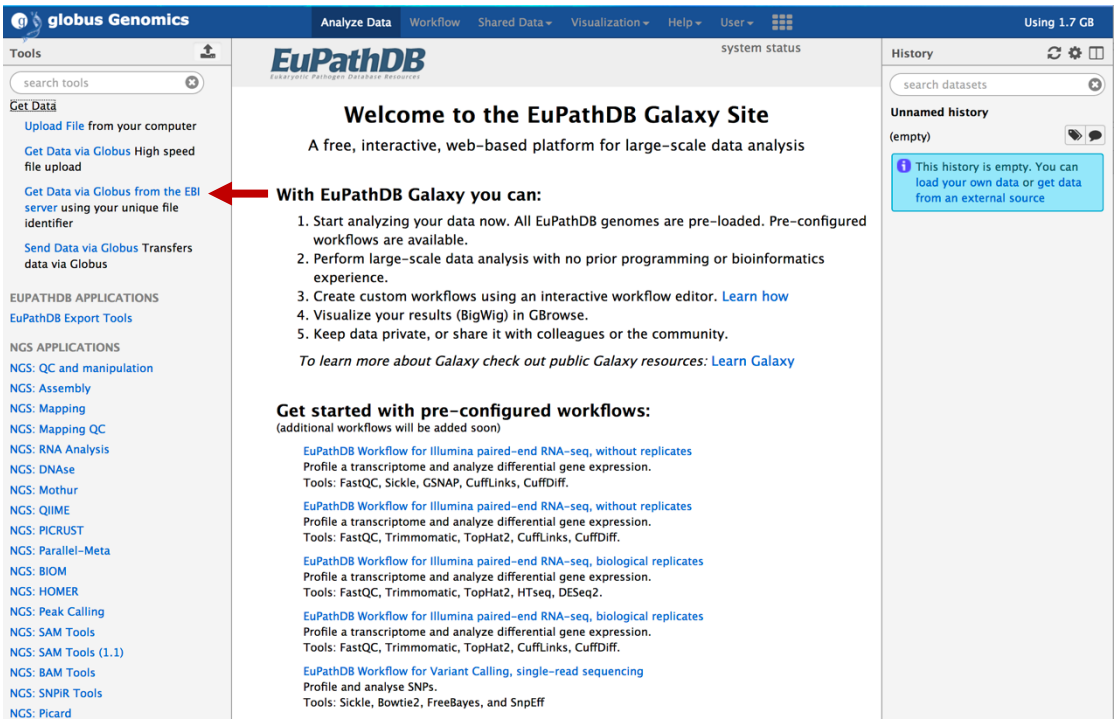


Step 8: Congratulations, you are in!

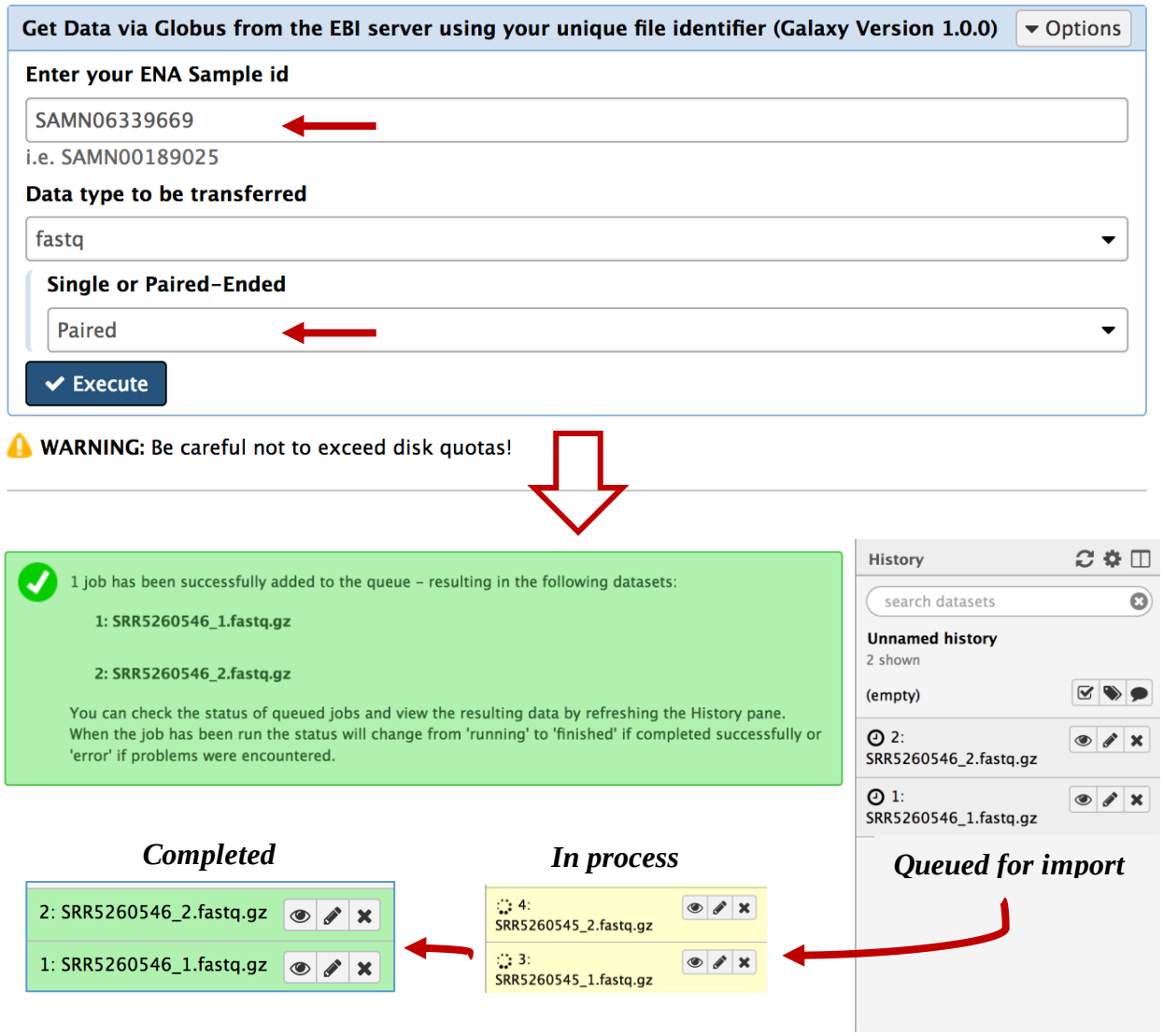
Section II: Importing data to Galaxy

There are multiple ways to important data into your Galaxy workspace. For this exercise, we will use the ‘Download from web or upload from disk’ tool and enter the direct data repository links listed below under ‘Group Assignments’. Remember one person in your group will be starting the workflow. Although all group members can sign up for an account for later use, please only one person should start a workflow today because we don’t want to overload the servers. The samples below were all generated by paired end sequencing, hence there are two files for each sample. The files are fastq files that are compressed (that is why they end in .gz = gzip).

Step 1: Click on the "Get data" link in the left-hand menu. This will reveal a list of options; click on "**Get Data via Globus from the EBI server**"



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

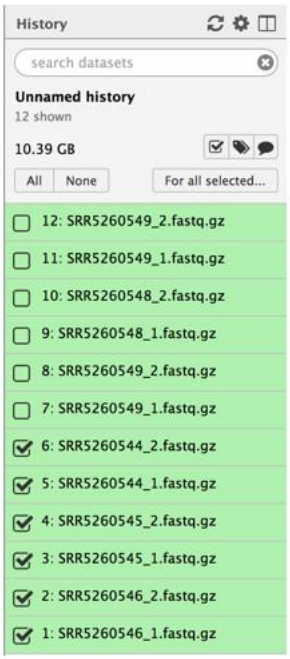


Note that the sample ID resulted in importing two files one for each pair. Repeat this process for each sample you want to import. *If you are working with samples from two conditions and the experiment was done in triplicate and paired end sequenced then you should end up with 12 files; six from each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

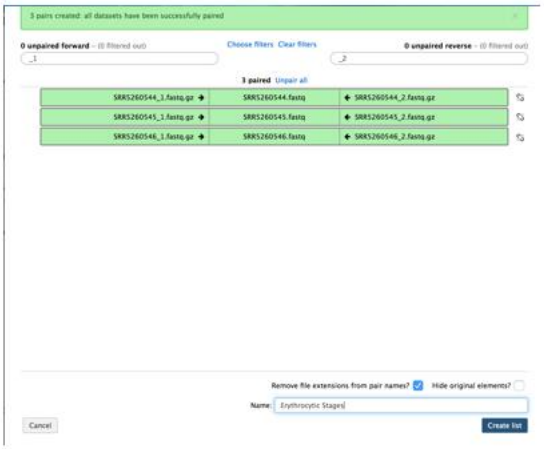
1. Click on the checkbox function "Operation on multiple datasets"

2. Select files that belong to the same condition

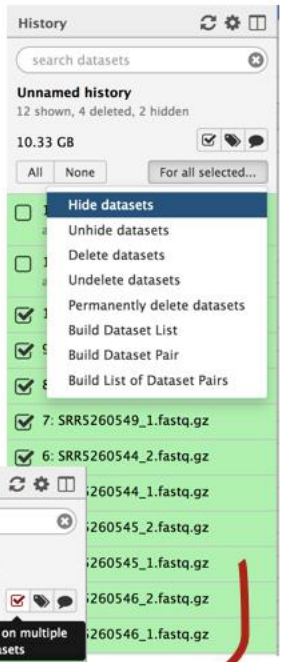


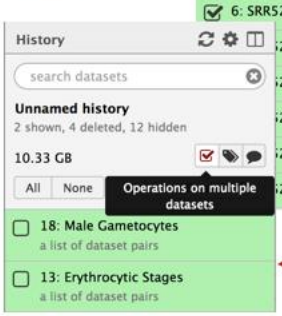
3. Click on for all selected and choose "Build List of Dataset Pairs"

4. Usually the corrected pairs are matched. Double check this and give this collection a meaningful name



5. You can clean up your history by hiding the files that you just put into a collection





Group assignments:

Groups 1, 2 & 3 will be examining 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

Samples:

Erythrocyte stages (Asexual):
SAMN06339669
SAMN06339670
SAMN06339671

Male gametocytes:
SAMN06339666
SAMN06339667
SAMN06339668

Female gametocytes:
SAMN06339663
SAMN06339664
SAMN06339665

Group 1:

Plasmodium berghei male gametocytes vs. erythrocytic stages

Group 2:

Plasmodium berghei female gametocytes vs. erythrocytic stages

Group 3:

Plasmodium berghei male gametocytes vs. female gametocytes

Groups 4, 5 & 6 will be examining data from a study called "*Plasmodium falciparum* NF54 Transcriptome" which examines RNAseq from 3 stages: erythrocytic, salivary gland and cultured sporozoite stages. This study is unpublished but data is accessible in the sequence repositories: https://www.ebi.ac.uk/ena/data/view/PRJNA230379

Samples:

Erythrocytic stages (Asexual):
SAMN02428730
SAMN02428734

Salivary gland sporozoites:
SAMN02428726
SAMN02428729

Cultured sporozoites:
SAMN02428728
SAMN02428727

Group 4:

Plasmodium falciparum salivary sporozoites vs. erythrocytic stages

Group 5:

Plasmodium falciparum cultured sporozoites vs. erythrocytic stages

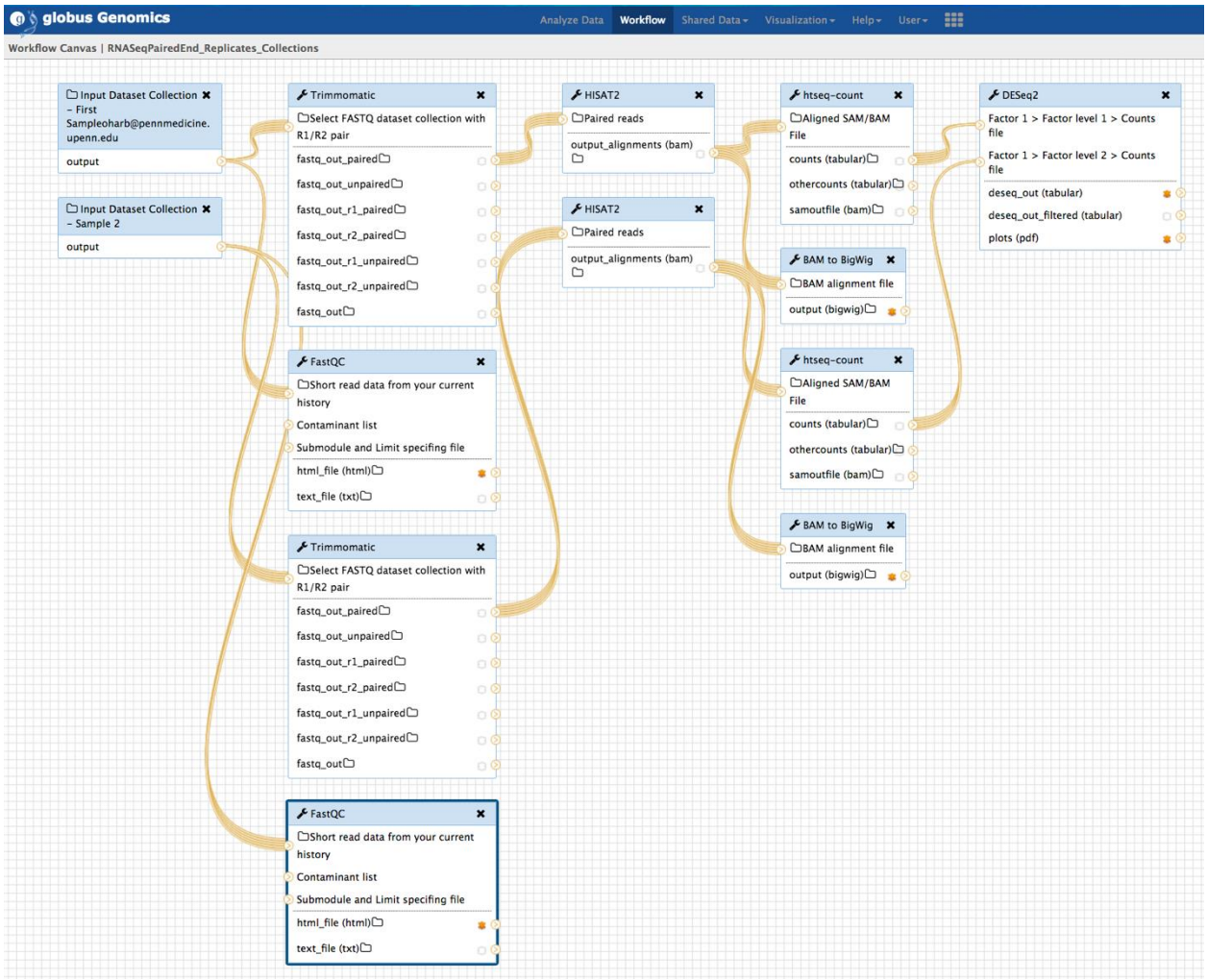
Group 6:

Plasmodium falciparum salivary sporozoites vs. cultured sporozoites

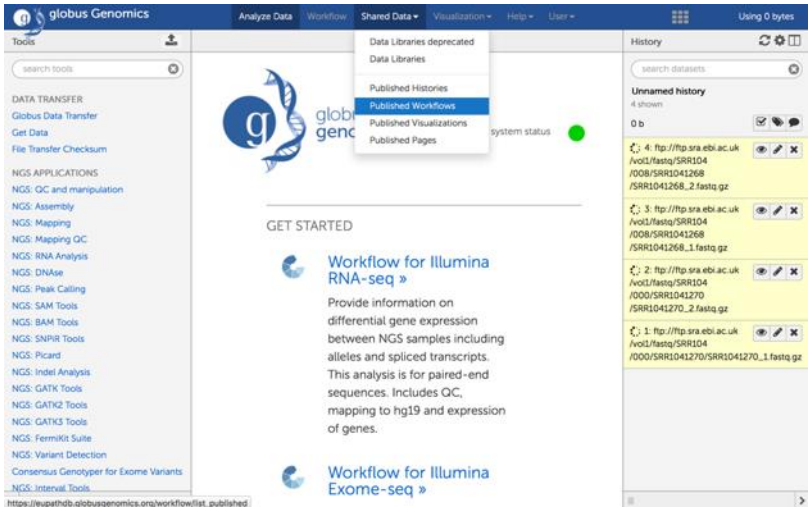
Section II: Running a workflow in Galaxy

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:

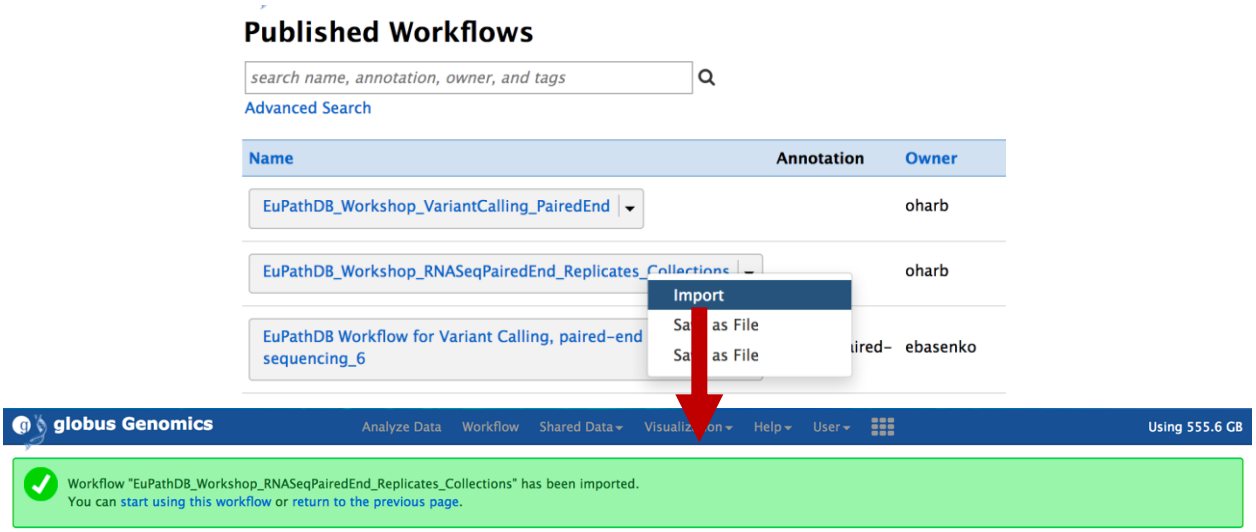
1. Analyzes the reads in your files and generates FASTQC reports.
2. Trims the reads based on their quality scores and adaptor sequences (Trimmomatic).
3. Aligns the reads to a reference genome using HISAT2 and generates coverage plots.
4. Determines read counts per gene (HTSeq)
5. Determines differential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



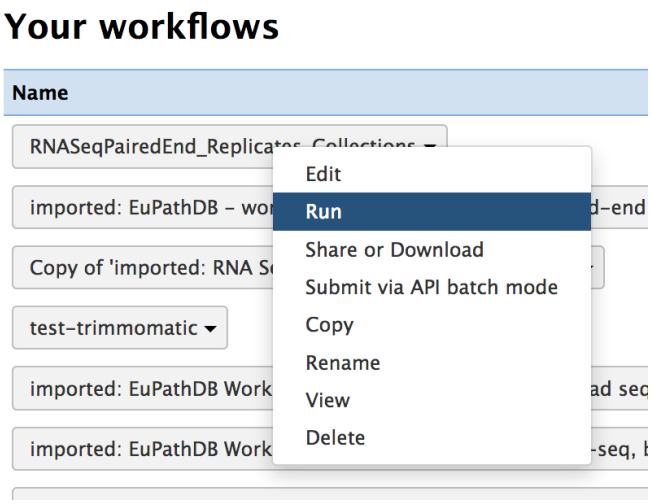
Step 1: Import the workflow called "**EuPathDB_Workshop_RNASeqPairedEnd_Replicates_Collections**" – click on the shared data menu item and select "Published Workflows" from the menu.



Step 2: Click on the arrow next to the appropriate workflow and select import.



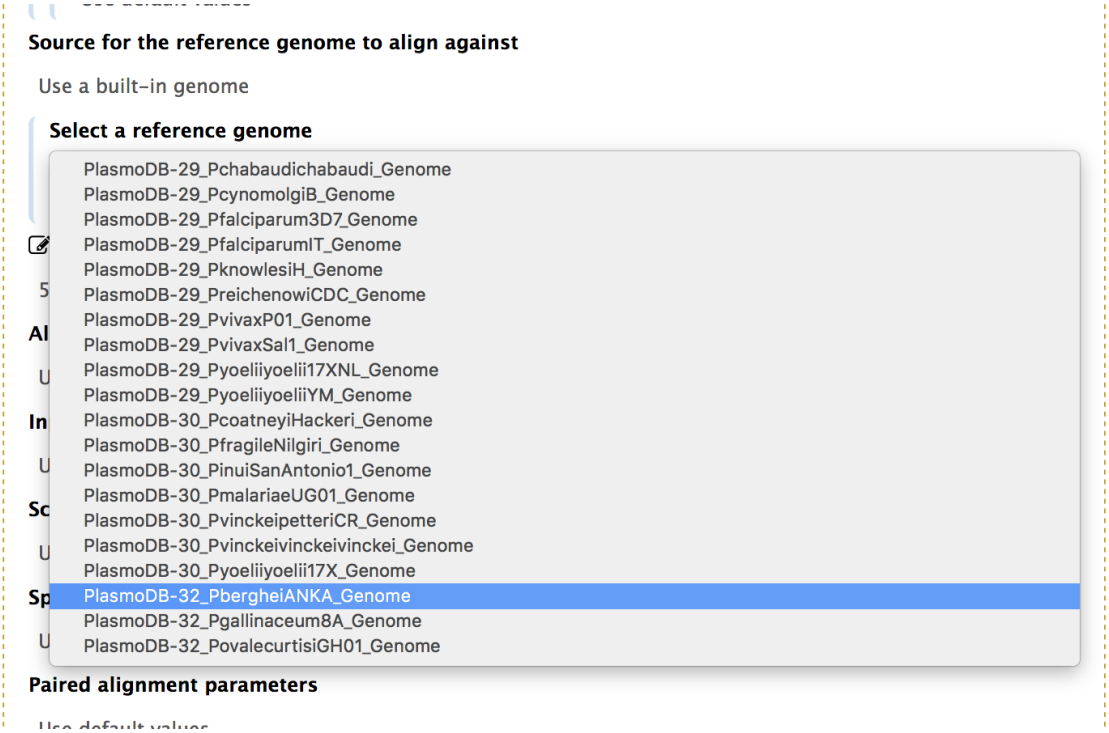
Step 3: Click on "Workflow" in the menu at the top of the page. On the next page click on the arrow next to workflow you just imported and select the "Run" option.



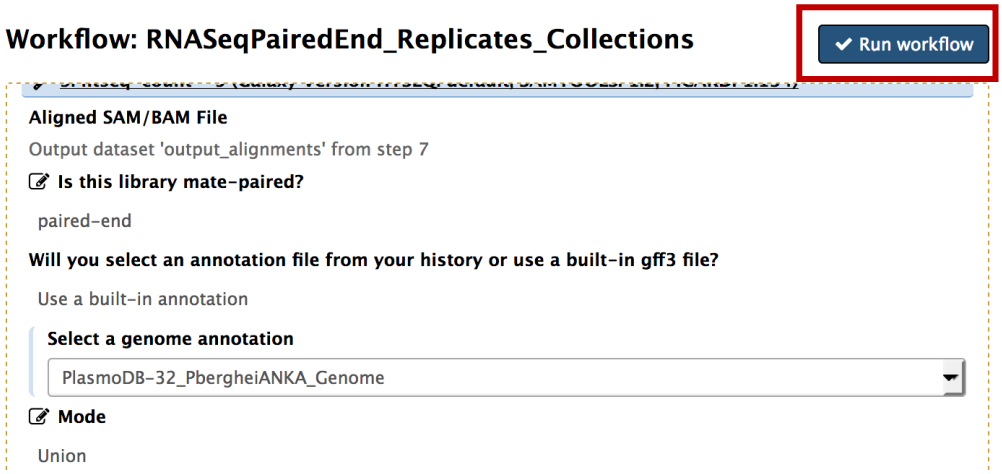
Step 4: Configure your workflow – there are multiple steps in the workflow but you do not need to configure all of them. For the purpose of this exercise you will need to configure the following:

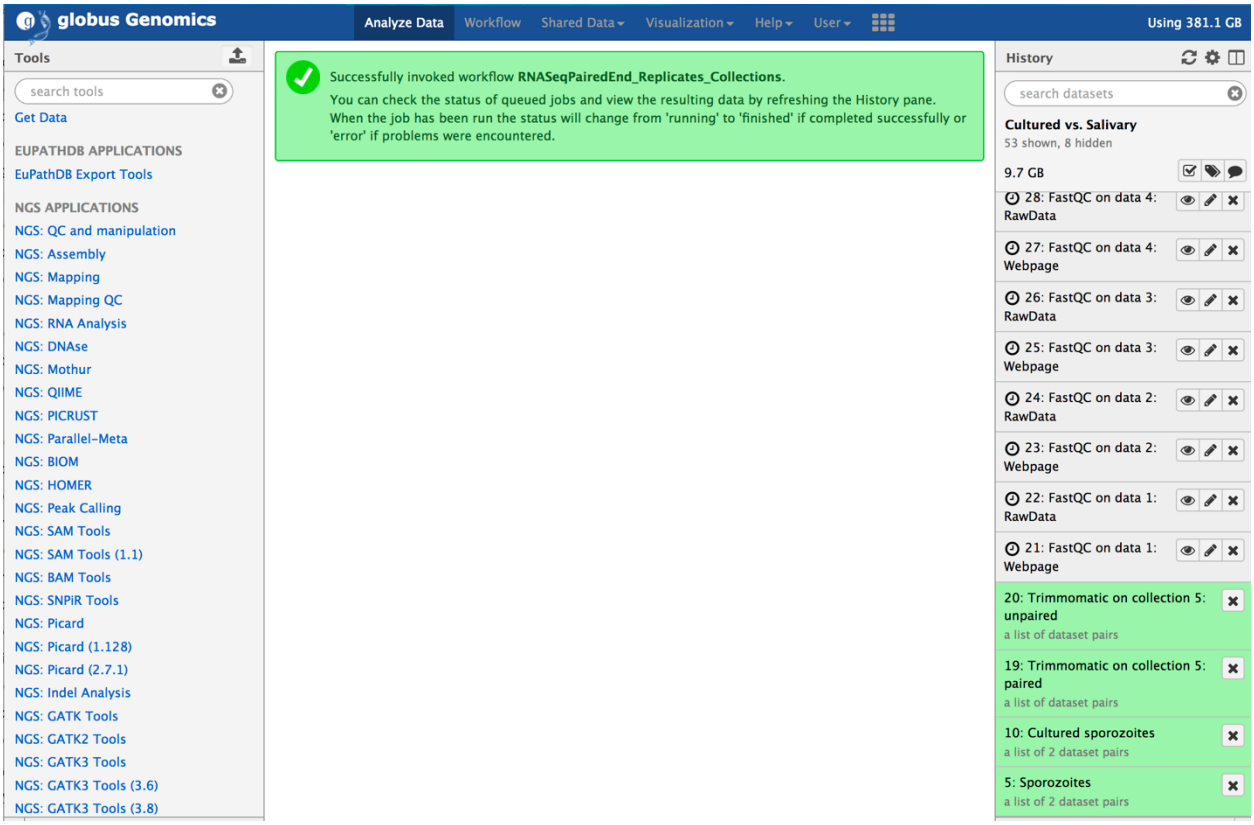
a. Select the input dataset collections. These are the collections of fastq files you just created. Workflow steps 1-2 allow you to select the datasets.

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. So, for example, if your experiment was performed using *Plasmodium berghei*, the reference genome you select should be *Plasmodium berghei*.



c. 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 file 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