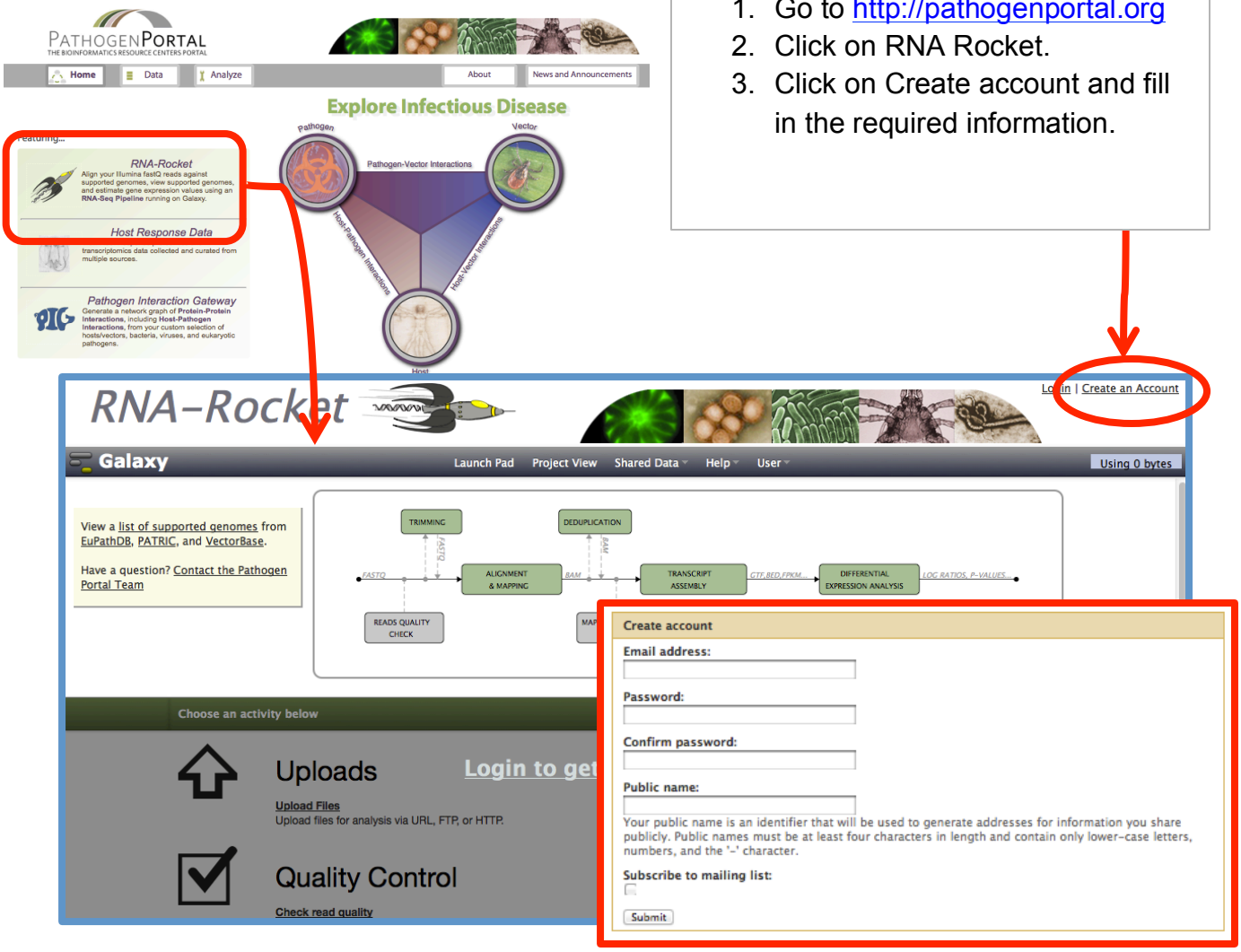
Mapping RNA sequence data
(Part 1: using pathogen portal's RNAseq pipeline)
Exercise 4

The goal of this exercise is to retrieve an RNA-seq dataset in FASTQ format and run it through an RNA-sequence analysis pipeline.

Step I: Create a login account at Pathogen Portal:

1. Go to http://pathogenportal.org
2. Click on RNA Rocket.
3. Click on Create account and fill in the required information.



Step II: Getting data into your launch pad.

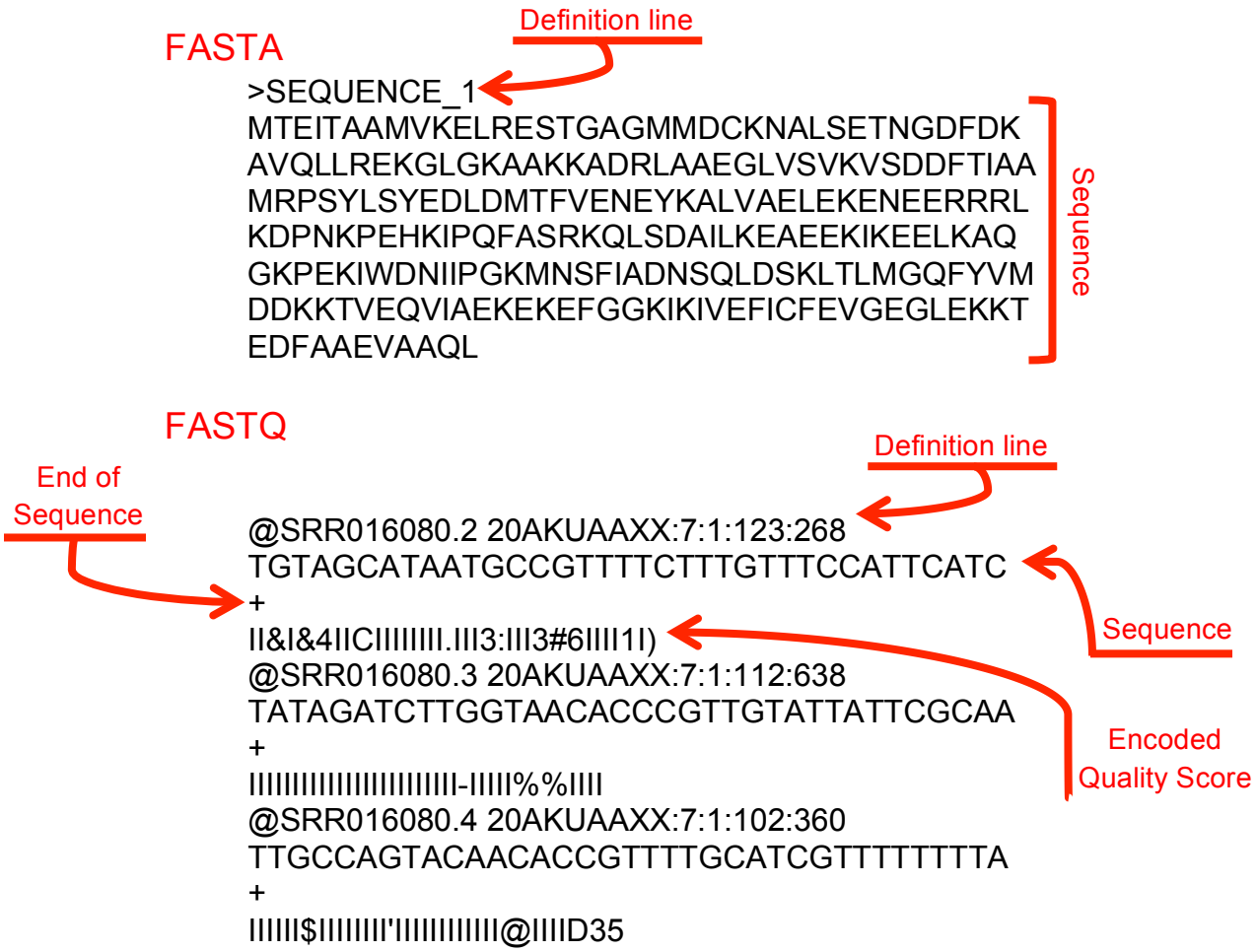
The following exercise is based on data generated from the recent study:
Grisdale *et al.* Transcriptome analysis of the parasite *Encephalitozoon cuniculi*: an in-depth examination of pre-mRNA splicing in a reduced eukaryote. BMC Genomics 2013, 14:207

http://www.biomedcentral.com/1471-2164/14/207

In the paper the authors indicate that the data has been deposited to the sequence read archive (SRA) and a study accession number is provided: SRP017112. You can access this record here:

http://www.ncbi.nlm.nih.gov/sra/SRP017112

The required input format is something called a FASTQ file, which is similar to a FASTA file. These are simple text files that include sequence and additional information about 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. The file that we will be using for this exercise originated from the DNA Data Bank of Japan (DDBJ), which is a mirror of NCBI and EBI.

Here is the record at DDBJ:

http://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP017112

The FastQ files for each time point are available here:

ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/

The 24hr time data are in the folder called: SRX247417
The 48hr time data are in the folder called: SRX229331
The 72hr time data are in the folder called: SRX247418

We will be uploading data directly from the DDBJ FTP site. Each samples is paired end (ie. two files per sample). Also, they indicate that two runs were done for each sample. We are only going to worry about one of the runs for each time point. For the next part of the this exercise feel free to navigate in the FTP site to the desired time point folder or simply use the links provided below:

Group 1 (24hr time point):

Upstream:

ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX247417/SRR769604_1.fastq.bz2

Downstream:

ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX247417/SRR769604_2.fastq.bz2

Group 2 (48hr time point):

Upstream:

ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_1.fastq.bz2

Downstream:

ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_2.fastq.bz2

Group 2 (72hr time point):

Upstream:

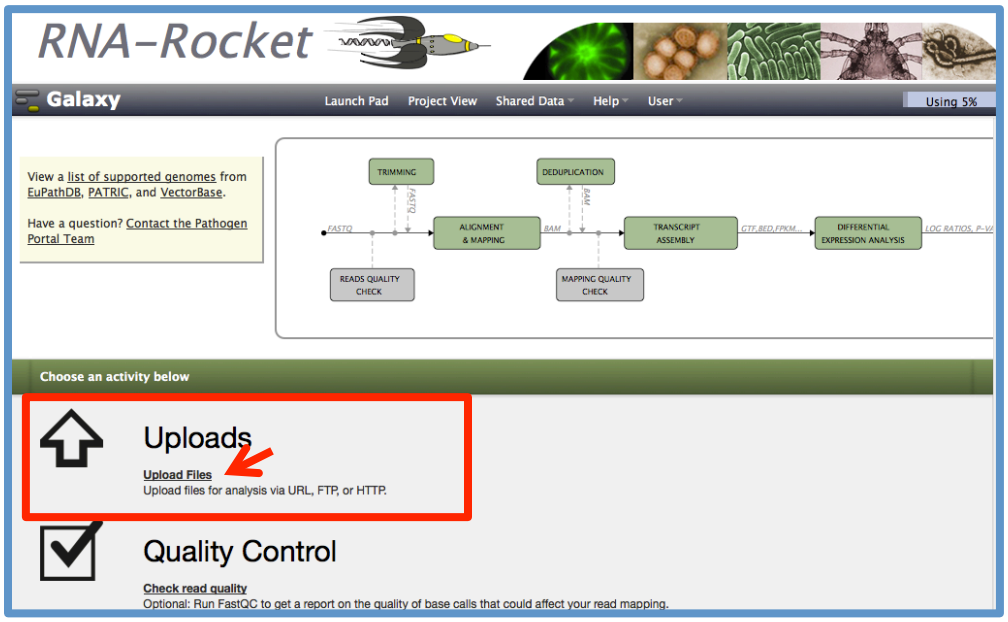
ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX247418/SRR769608_1.fastq.bz2

Downstream:

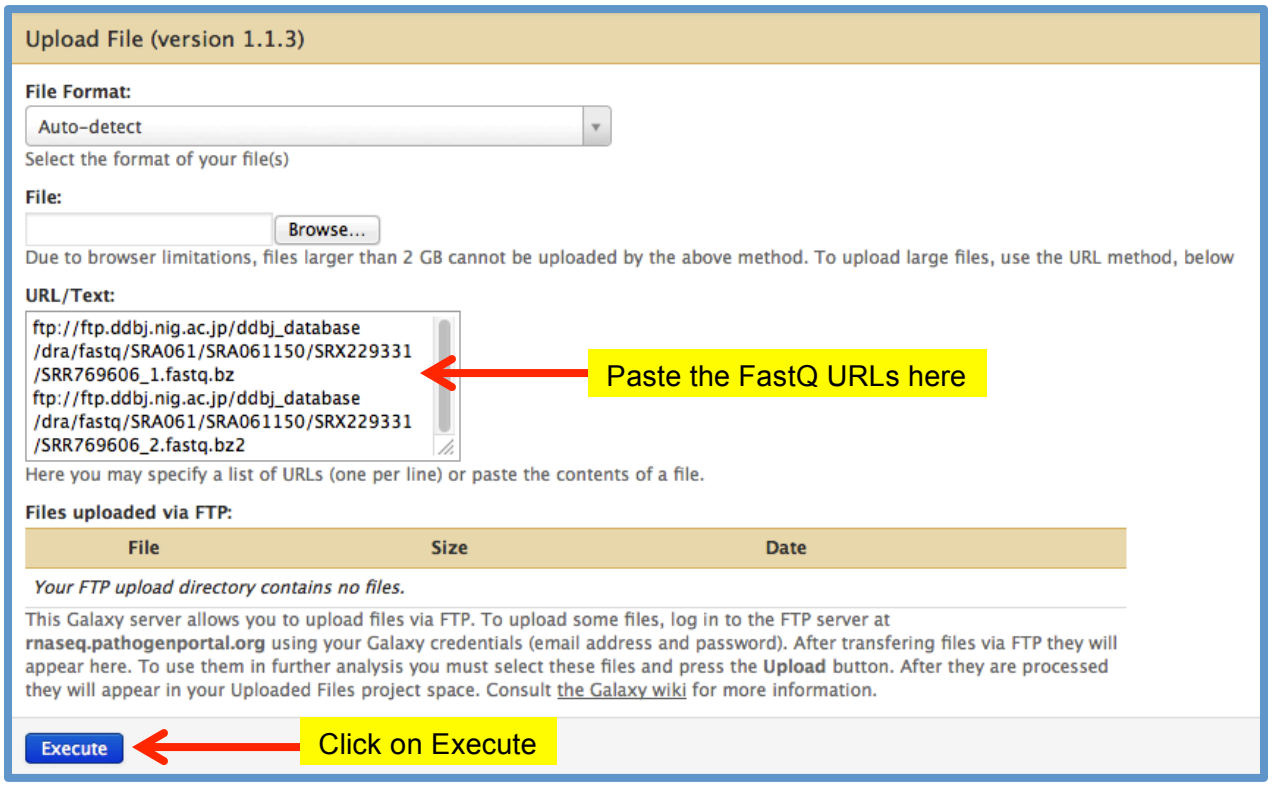
ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX247418/SRR769608_2.fastq.bz2

Here are the steps you take to start uploading data into your Launchpad:

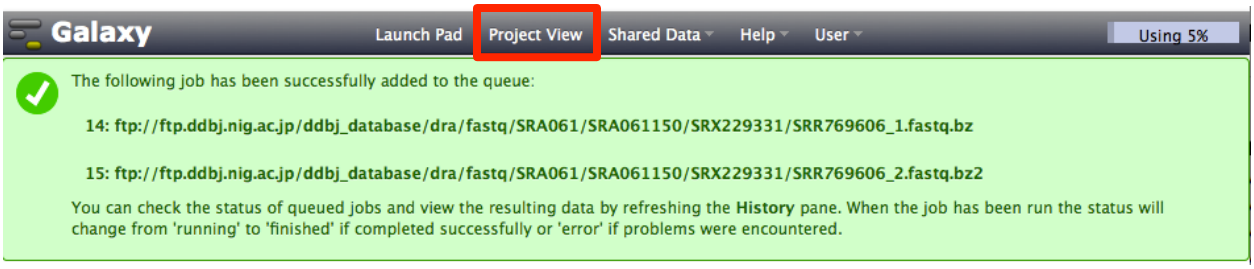
1. Click on the “Upload Files” link



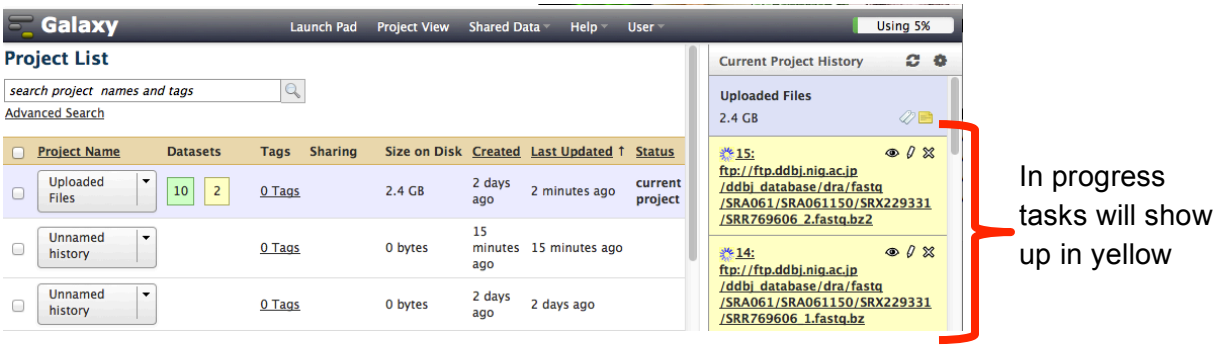
2. On the next page, copy and paste both files for your time point in the “URL/Text” window then click on the “Execute” button.

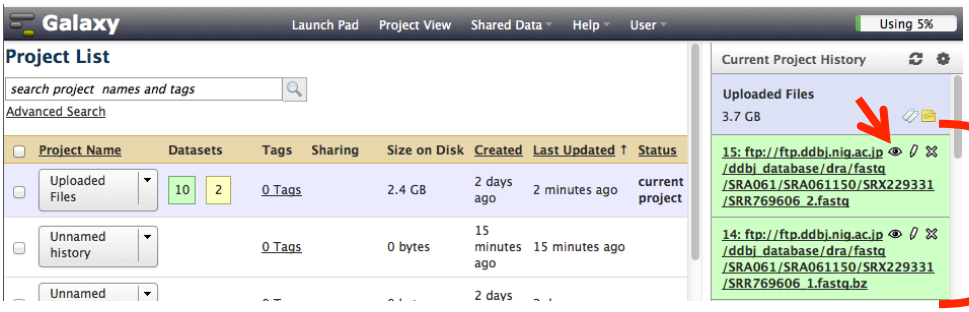


You should now see a window that looks like this:



To view the progress of your upload, click on “Project View” (red square in image above).





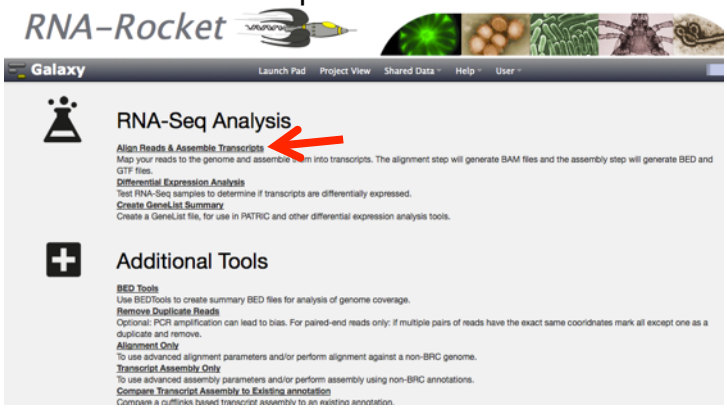
Completed tasks will show up in green

You can inspect the contents of completed tasks (like uploaded files) by clicking on the eye icon next to the name of the file (arrow in above image). Inspecting a FASTQ file should look like this:

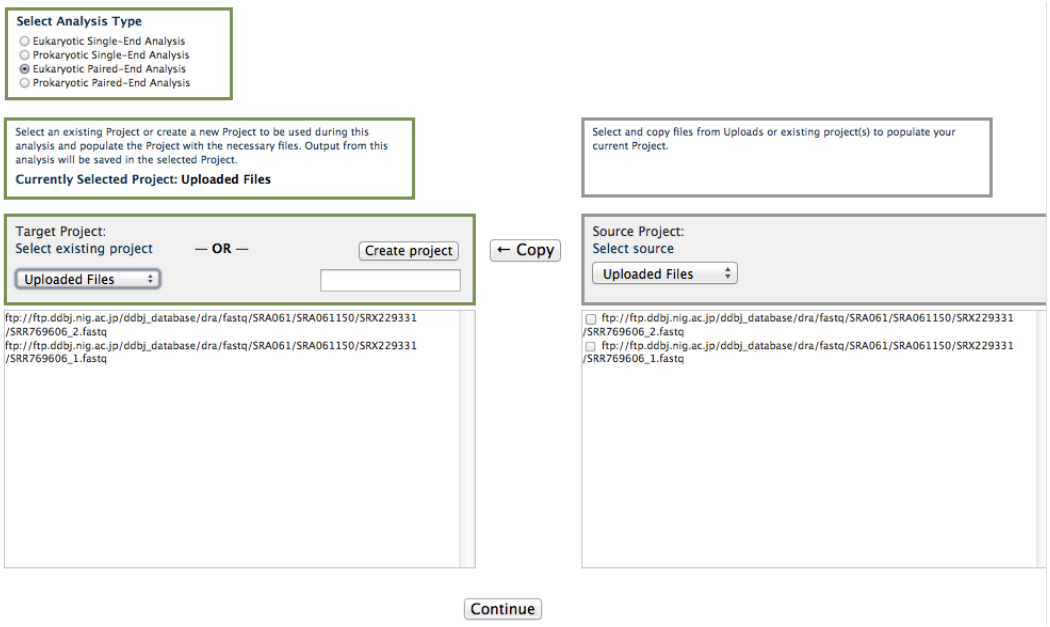
3. Once the RNA-sequence FASTQ file has been uploaded you can start the RNA-seq pipeline. Pathogen portal uses two algorithms for mapping (TopHat) and transcript prediction and expression value calculation (Cufflinks). Note that there are many algorithms and methods for RNA-seq mapping and analysis each with its advantages and disadvantages. You are encouraged to learn more about the algorithm you are using.

    - TopHat: http://tophat.cbcb.umd.edu/
    - Cufflinks: http://cufflinks.cbcb.umd.edu/index.html

- To start the pipeline click on the “Launch Pad” link (red square in above image). On the next page, scroll down to the “RNA-Seq Analysis” section and click on “Align Reads & Assemble Transcripts”.



- On the next page, scroll down and choose the type of analysis (in this case we are analyzing a paired end eukaryotic sample).
- Next select the target project from the drop down menu. You should only have one or two projects one of which will contain both FASTQ files you uploaded (probably called “Uploaded Files”). Once you select the correct project you should see the two FASTQ files contained within it. Next click on continue.

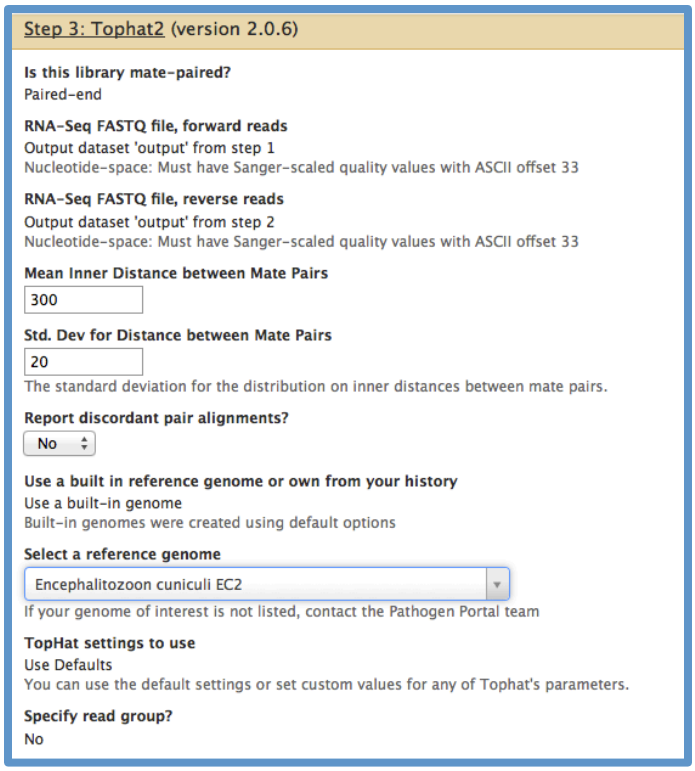


- The next page allows you to configure the pipeline:

  Step1: Select the upstream read file (ends in _1) and click on the arrow to move it to the “Selected” window.

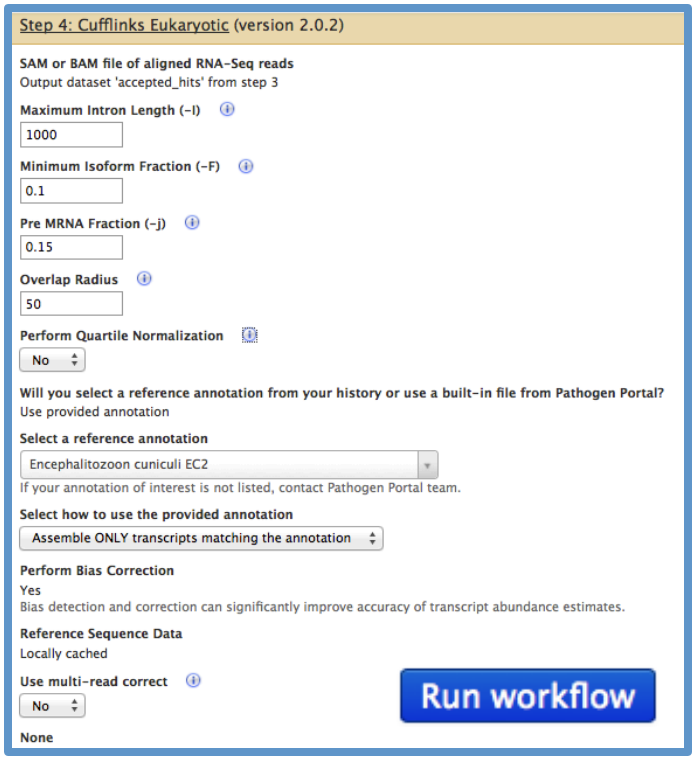
  Step2: Select the downstream read file (ends in _2) and click on the arrow to move it to the “Selected” window.

Step3: Configure TopHat – there are a number of options that may be modified, however, for the purposes of this exercise the default parameters may be used. The only required change is the reference genome -- select *Encephalitozoon cuniculi* EC2



Step4: Configure Cufflinks – once again there are a number of options to modify. For the purposes of this exercise change the following:
Maximum Intron Length (-I): 1000
Select a reference annotation: *Encephalitozoon cuniculi* EC2
Select how to use the provided annotation: Assemble Novel + annotated transcripts.

Click on the Run Workflow button.



After you start the workflow you should get a confirmation window that indicates all the steps that have been added to the queue. The progress of your workflow can be viewed to the right. Completed tasks are in green, running tasks are in yellow and tasks waiting in the queue are in grey.