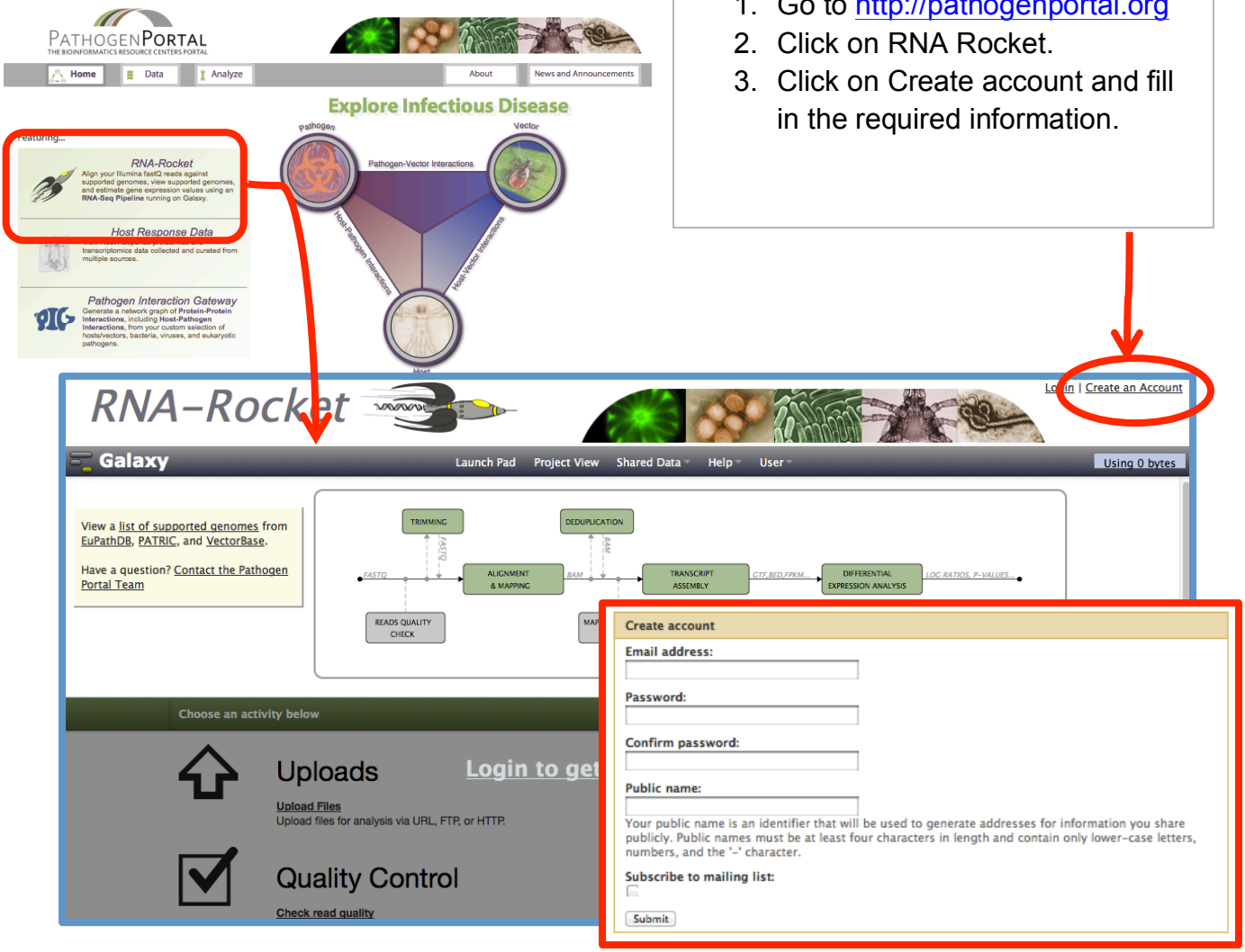
Mapping RNA sequence data
(Part 1: using pathogen portal's RNAseq pipeline)
Exercise 6

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:
Franzén *et al*. Transcriptome profiling of Giardia intestinalis using strand-specific RNA-seq. PLoS Comput Biol. 2013;9(3)

http://www.ncbi.nlm.nih.gov/pubmed/23555231

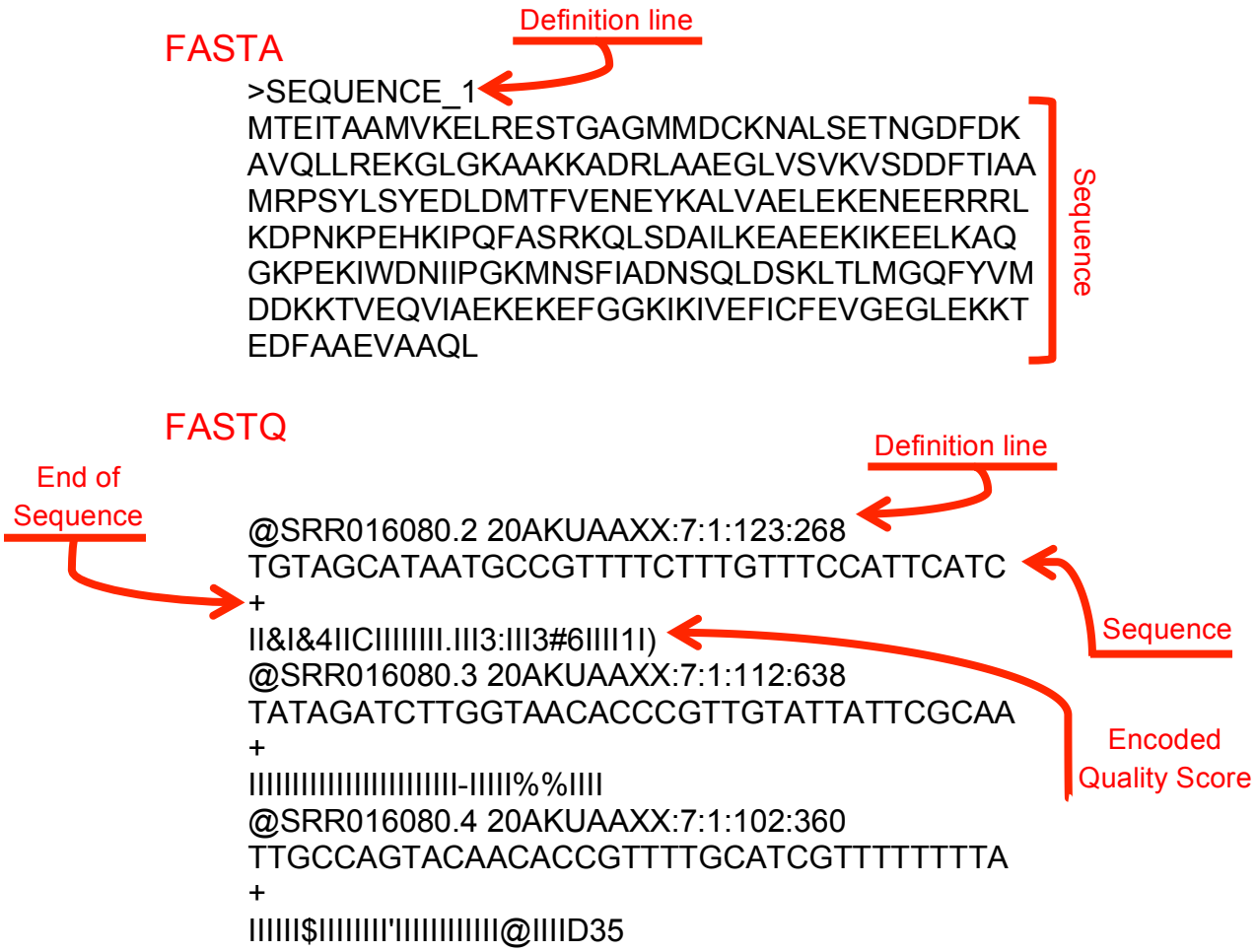
The paper examines transcription in Giardia assemblages A (WB), B (GS) and E (P15). In the paper the authors indicate that the data has been deposited to the sequence read archive (SRA) and they provide a link to GEO:

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM895812

Examining the information available in GEO and under the SRA accession numbers you will notice that this data is paired end and strand specific. So for each assemblage there should be two files for the forward strand (one for each pair) and two files for the reverse strand.

Assemblage A (WB): http://www.ncbi.nlm.nih.gov/sra/SRX129645
Assemblage A (AS175): http://www.ncbi.nlm.nih.gov/sra/SRX129648
Assemblage B (GS): http://ncbi.nlm.nih.gov/sra/SRX129649
Assemblage E (P15): http://www.ncbi.nlm.nih.gov/sra/SRX129646

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.
- 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
- RNArocket allows you to use SRA accession numbers and directly retrieve FASTQ files.

Group 1 (Assemblage A (WB)):

Accession number: SRX129645

Group 2 (Assemblage A (AS175)):

Accession number: SRX129648

Group 3 (Assemblage B (GS)):

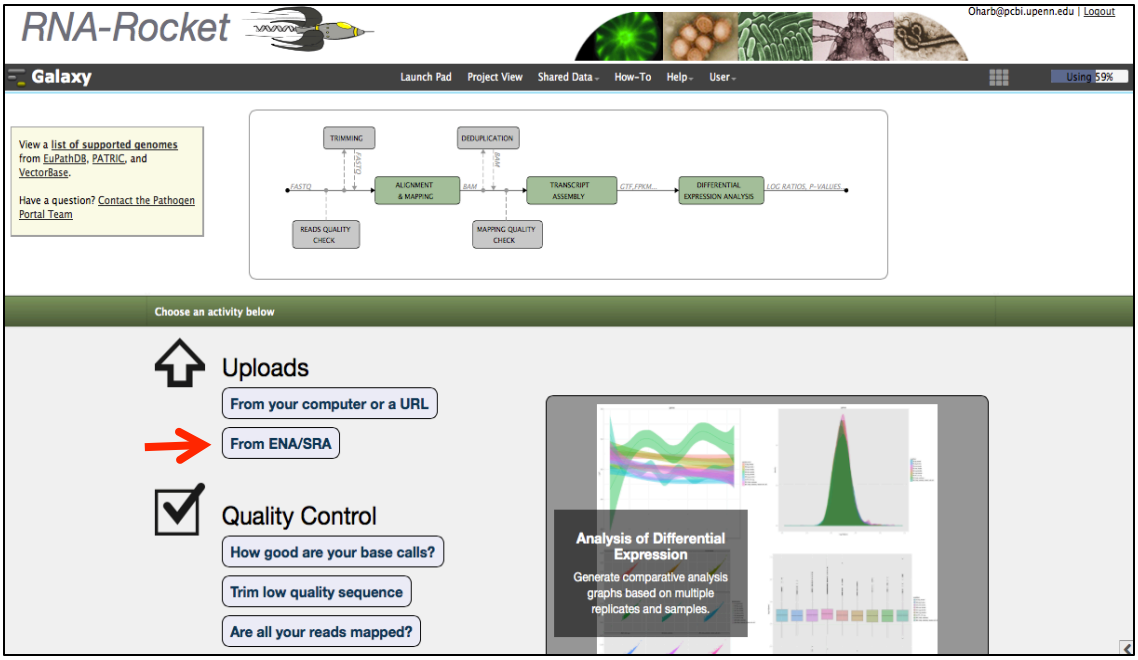
Accession number: SRX129649

Group 4 (Assemblage E (P15)):

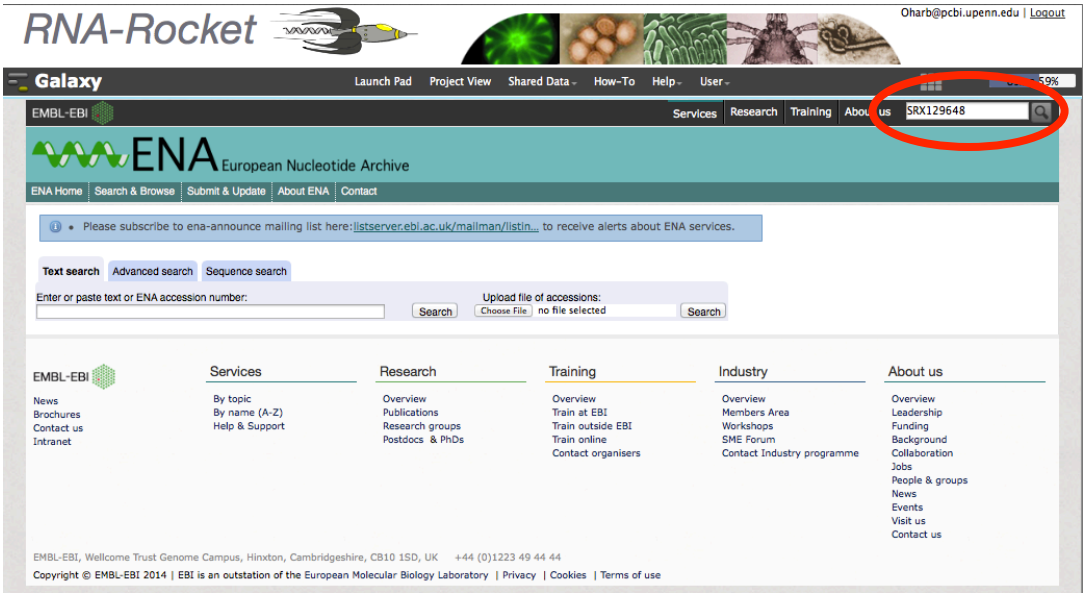
Accession number: SRX129646

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

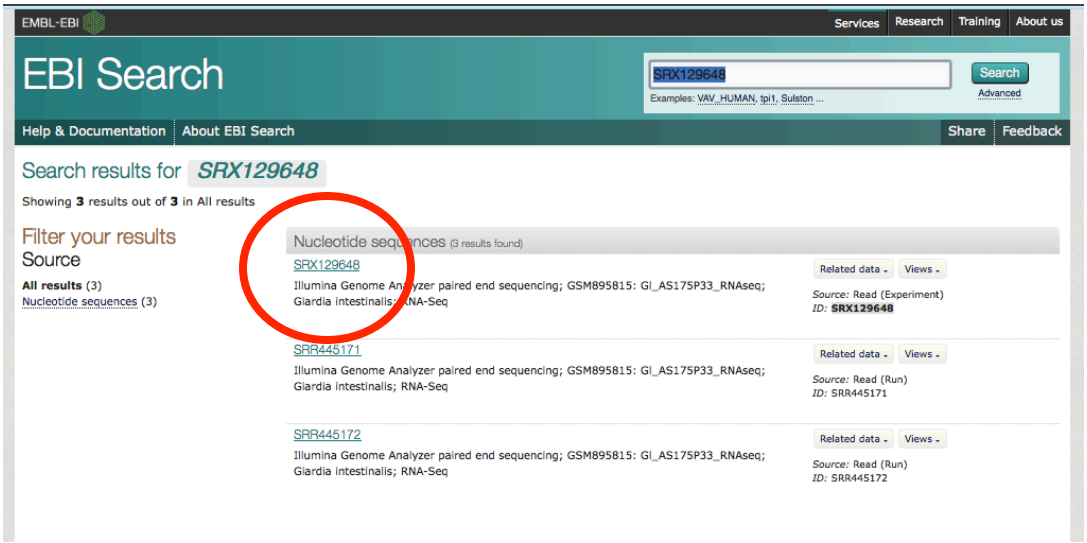
1. Click on the "Upload Files" link



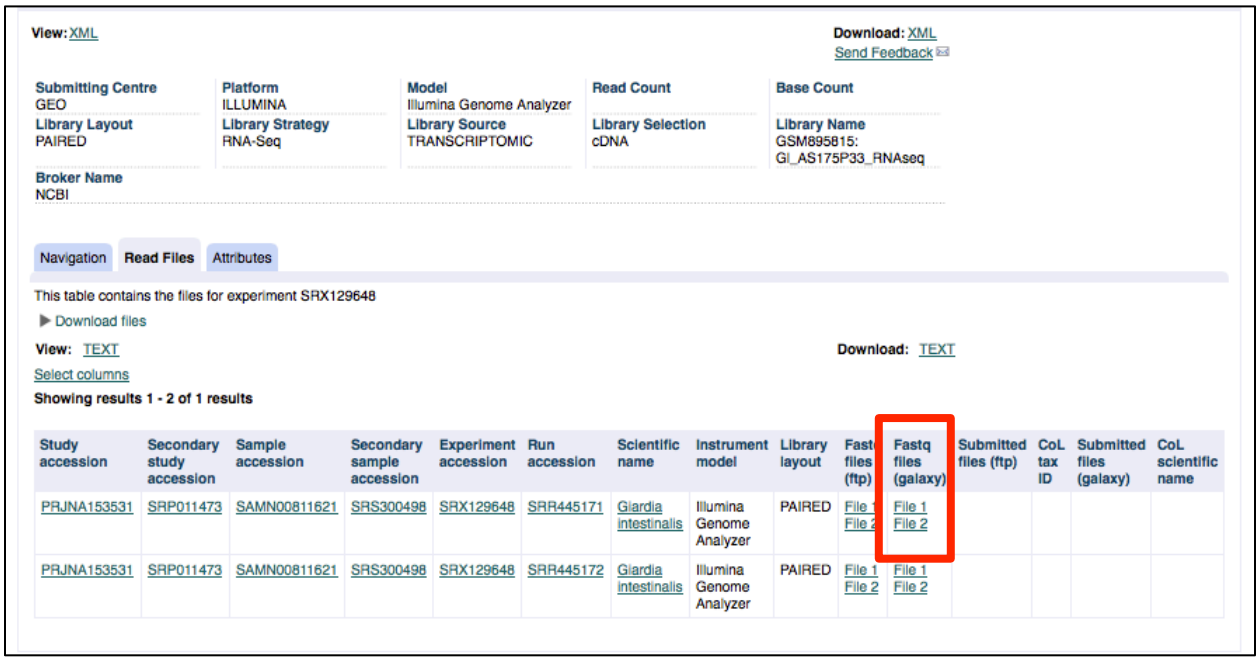
2. On the next page, notice the instructions to use the global search on the ENA site. Next click on continue.
3. Cut and paste your accession number into the global search box. Click on the search icon.



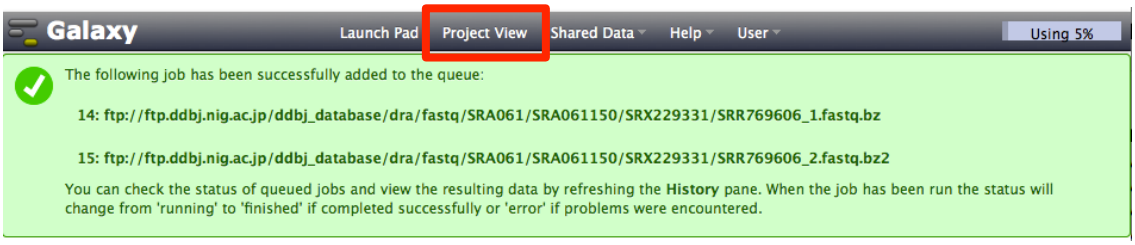
4. Select the record that matches your search accession number, usually the first one.



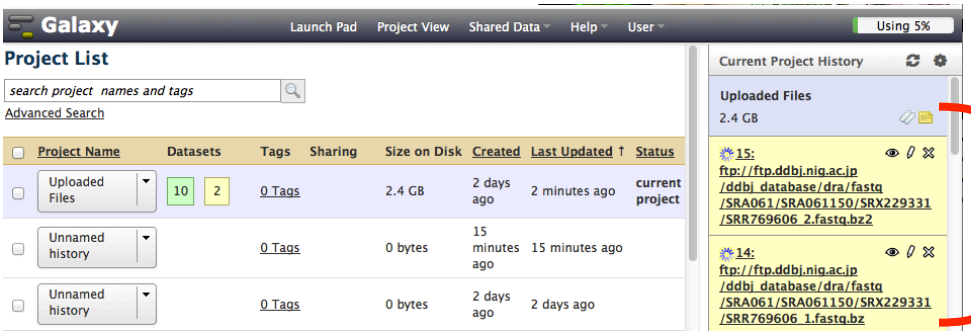
5. On the next page you can select the files to load into RNArocket. We will use the "Fastq files (Galaxy)" since RNArocket is built on Galaxy. Remember, you have to get 4 files. Two (paired) for each strand. Click on the link for File 1 for the first run, then click on the back button on your browser and click on the link for File 2.



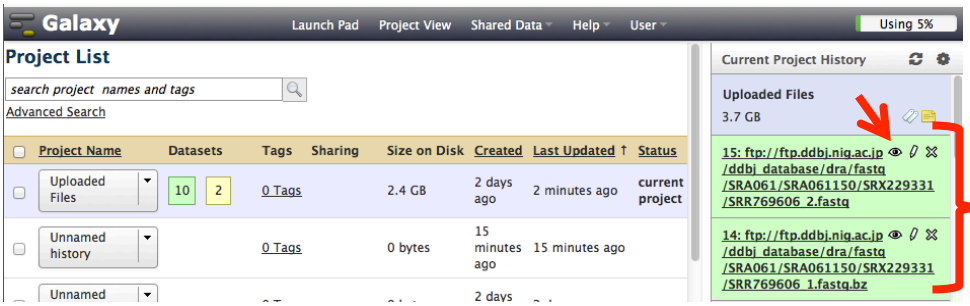
You should now see a window that looks like this:



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



In progress tasks will show up in yellow



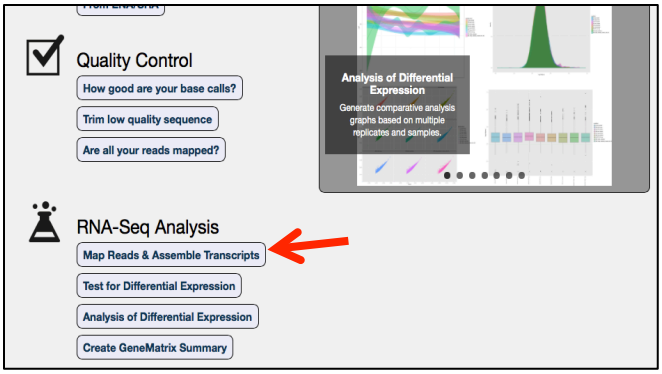
Completed tasks will show up in green

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:

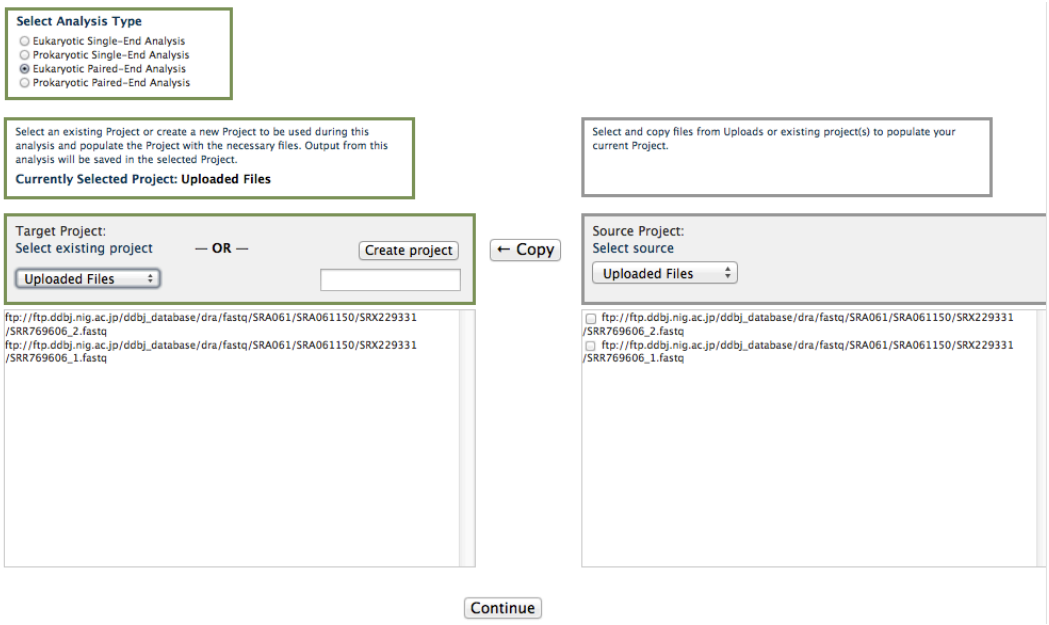
6. Once the RNA-sequence FASTQ file has been uploaded you can start the RNA-seq pipeline. Pathogen portal uses two algorithms for mapping (TopHat) and transcript prediction and expression value calculation (Cufflinks). Note that there are many algorithms and methods for RNA-seq mapping and analysis each 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 "Map Reads & Assemble Transcripts".



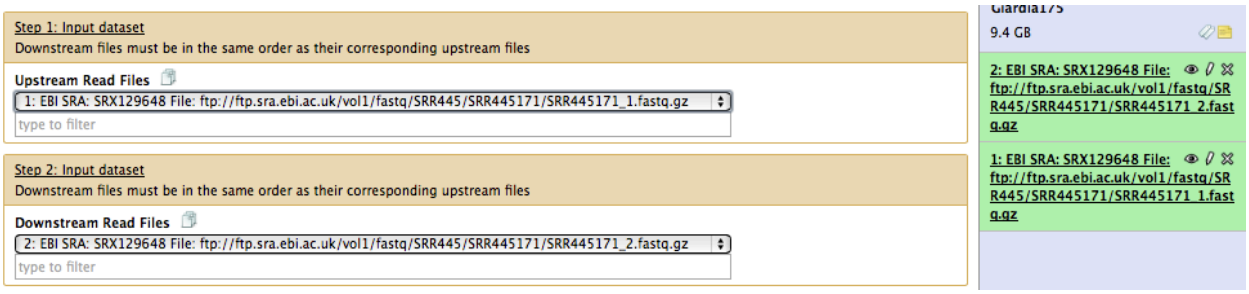
- On the next page, scroll down and choose the type of analysis (in this case we are analyzing a paired end eukaryotic sample).
- Next select the target project from the drop down menu. You should only have one or two projects one of which will contain both FASTQ files you uploaded (probably called “Uploaded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



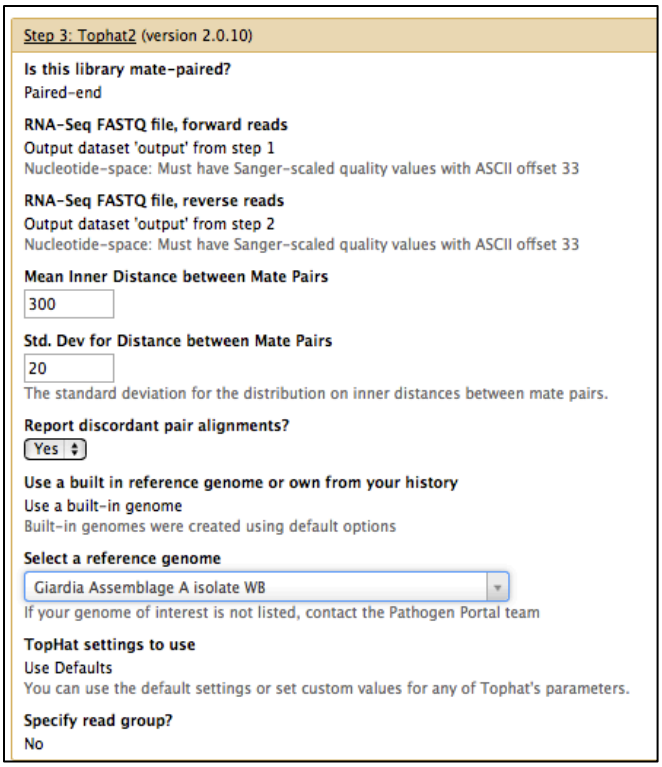
- The next page allows you to configure the pipeline:

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

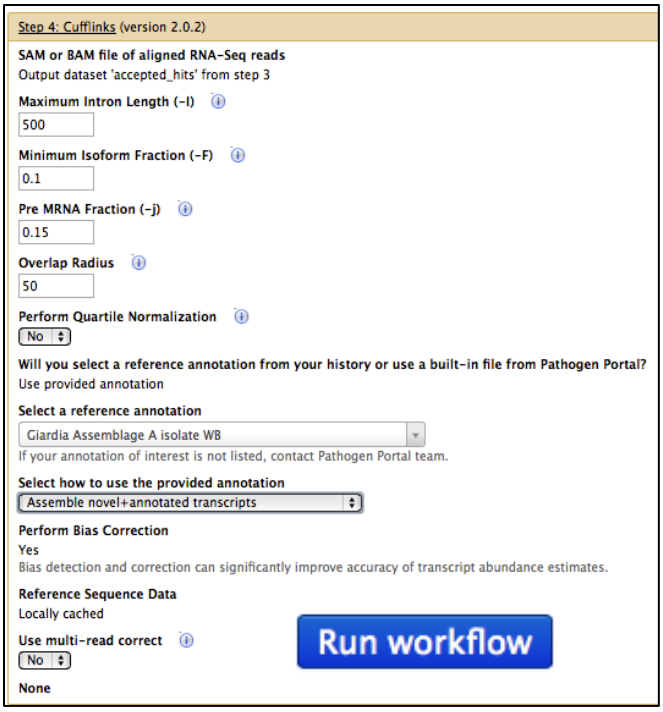
Step2: Select the downstream read file (ends in _2) and click on the arrow to move it to the “Selected” window.



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 *Giardia* Assemblage A isolate WB



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): 500
The reference annotation should be automatically selected: *Giardia* Assemblage A isolate WB
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.