

Mapping RNA sequence data

Part 1: RNA-Rocket RNaseq pipeline

The goal of this exercise is to retrieve an RNA-seq dataset in FASTQ format and run it through an RNA-sequence analysis pipeline. We will be using Pathogen Portal's RNA-Rocket which includes a workflow for mapping RNA-Seq reads to a reference genome, using this mapping to assemble transcripts, mapping transcripts to existing annotations, and determining expression levels. The mapping workflow uses two algorithms, TopHat for aligning reads and Cufflinks for transcript prediction and calculating expression levels. The input required is FASTQ files and the outputs are read alignments (BAM Files), tab delimited assembly and expression files for known genes, isoforms and novel transcripts.

1. Create an account on RNA Rocket

- a. Go to <http://pathogenportal.org>
- b. Click on RNA Rocket
- c. Click on Create an Account and fill in the required information.

The image shows a composite screenshot of the Pathogen Portal website. At the top, the 'PATHOGENPORTAL THE BIOINFORMATICS RESOURCE CENTERS PORTAL' logo is visible. Below the logo is a navigation bar with 'Home', 'Data', 'Analyze', 'About', and 'News and Announcements'. The main content area features 'Explore Infectious Disease' with sub-sections for 'pathogen' and 'Vector'. A red box highlights the 'RNA-Rocket' section, which describes the pipeline: 'Align your Illumina fastQ reads against supported genomes, view supported genomes, and estimate gene expression values using an RNA-Seq Pipeline running on Galaxy.' A red arrow points from this box to the 'RNA-Rocket' workflow diagram in the Galaxy interface. The Galaxy interface shows a workflow with steps: 'READS QUALITY CHECK', 'ALIGNMENT & MAPPING', 'TRANSCRIPT ASSEMBLY', 'MAPING QUALITY CHECK', 'DIFFERENTIAL EXPRESSION ANALYSIS', and 'LOG RATIO PLOTS'. A red circle highlights the 'Login | Create an Account' link in the top right of the Galaxy interface. A red arrow points from this circle to a 'Create account' form. The form includes fields for 'Email address:', 'Password:', 'Confirm password:', and 'Public name:'. Below the form, there is a note: 'Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least four characters in length and contain only lower-case letters, numbers, and the '-' character.' A 'Submit' button is at the bottom of the form.

2. Upload the RNA sequencing reads to your RNA Rocket launch pad. RNA Rocket allows you to directly retrieve FASTQ files of the sequencing reads using SRA accession numbers.

a. **Background:** This exercise will rely on data deposited in the sequence read archive (SRA). The data is based on transcriptomic analysis of three developmental stages of *Plasmodium falciparum*:

1. Salivary gland sporozoites
2. Cultured sporozoites, and
3. Cultured asexual stages.

Each developmental stage was assayed by RNA sequencing (2 replicates per sample). The **study accession number for this data on SRA is SRP033414** and additional information about this experiment may be obtained from GEO:

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52867>

Examining the information available in GEO and under the SRA accession numbers you will notice that this data is paired end. So for each sample there should be two files one for each of the pairs. More information for each sequencing run can be found at:

Salivary gland sporozoites sample 1: <http://www.ncbi.nlm.nih.gov/sra/SRX385640>

Salivary gland sporozoites sample 2: <http://www.ncbi.nlm.nih.gov/sra/SRX385641>

Cultured sporozoites sample 1: <http://www.ncbi.nlm.nih.gov/sra/SRX385642>

Cultured sporozoites sample 2: <http://www.ncbi.nlm.nih.gov/sra/SRX385643>

Asexual stage parasites sample 1: <http://www.ncbi.nlm.nih.gov/sra/SRX385644>

Asexual stage parasites sample 2: <http://www.ncbi.nlm.nih.gov/sra/SRX385645>

The required input file for RNA Rocket's analysis pipeline is a FASTQ file, a text file (similar to FASTA) that includes 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

```

FASTA
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDK
AVQLLREKGLGKAAKKADRLAAEGLVSVKVSDDFITAA
MRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRL
KDPNKPEHKIPQFASRKQLSDAILKEAEEKIKEELKAQ
GKPEKIWDNIIPGKMNSFIADNSQLDSKLTLMGQFYVM
DDKKTVEQVIAEKEKEFGGKIKIVFICFEVGEGLKKT
EDFAAEVAAQL

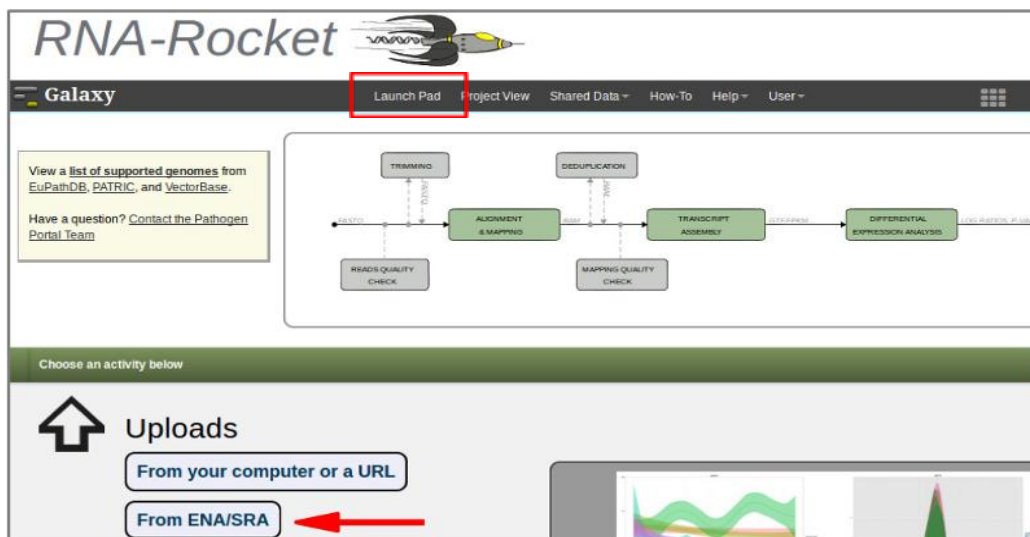
FASTQ
@SRR016080.2 20AKUAAXX:7:1:123:268
TGTAGCATAATGCCGTTTTCTTTGTTCCATTCATC
+
!!&!&4IICIIIIIIII.III3:III3#6IIII1I)
@SRR016080.3 20AKUAAXX:7:1:112:638
TATAGATCTTGGTAACACCCGTTGTATTATTCGCAA
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SRR016080.4 20AKUAAXX:7:1:102:360
TTGCCAGTACAACACCGTTTTGCATCGTTTTTTTTTA
+
IIIIII$IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIID35

```

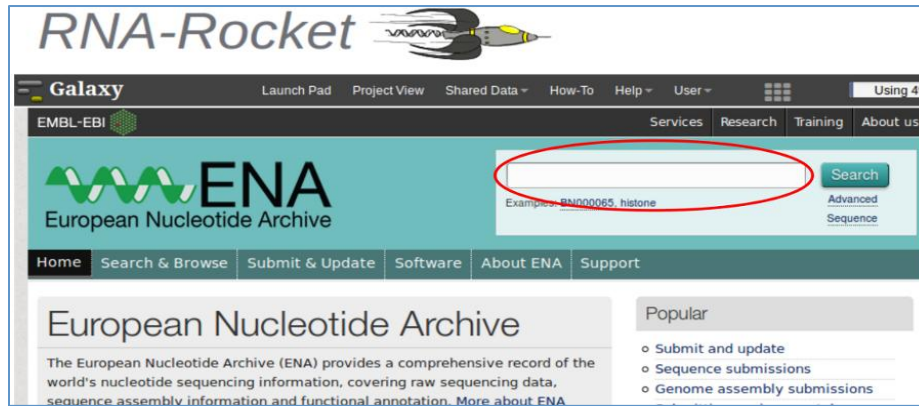
b. Upload data into your Launchpad.

Note: During this exercise you will NOT download any data to your computer. Instead you will be providing information to enable transferring data from ENA/SRA to RNA-Rocket.

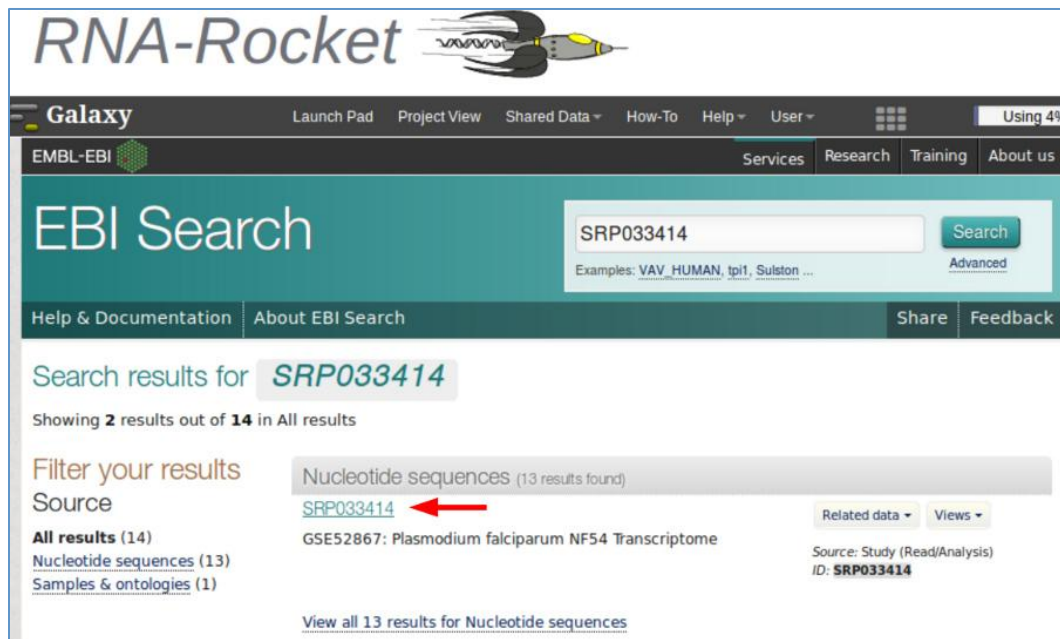
- i. Click on the “Launch Pad” link in the Galaxy menu bar. Then select “From ENA/SRA”.



- ii. On the next page, notice the instructions to use the **global search** on the ENA site. Click on continue.
- iii. Cut and paste the **study accession number (SRP033414)** into the search box (see red circle below). Click on the search icon.



- iii. Depending on RNA-rocket's configuration you may be taken to the EBI search results page where you will need to click on the Study link ID in order to get to the study page. If your page looks like the second screen shot, please proceed to iv.



RNA-Rocket Galaxy

EMBL-EBI Services Research Training About us

ENA
European Nucleotide Archive

Search: Search
Examples: BVD00065, histone
Advanced Sequence

Home Search & Browse Submit & Update About ENA Support

Please subscribe to ena-announce mailing list here: listserv.ebi.ac.uk/mailman/listin... to receive alerts about ENA services.

Study: SRP033414
GSE52867: Plasmodium falciparum NF54 Transcriptome

View: [XML](#) Send Feedback
Download: [XML](#)

Submitting Centre	Study Type	Read Count	Base Count
GEO	Transcriptome Analysis	112,445,306	22,713,951,812

Broker Name
NCBI

Abstract
Summary: Transcriptomic Analysis of Cultured Sporozoites of P. falciparum Overall Design: RNA-seq reads from each of three developmental stages (2 replicates per sample) were mapped to the reference Plasmodium falciparum genome, and gene expression levels were calculated for each sample.

- iv. Click on the link for File 1 in the column called “Fastq files (galaxy)” for the sample assigned to your group, then click on the back button on your browser and click on the link for File 2 from the same sample. This will begin the file transfer to RNA-Rocket. You may need to scroll down to see the Read Files tab which contains the Fastq files (galaxy) column that you need. You will need to get 2 files, one for each file generated by the paired end sequencing.

Galaxy

Launch Pad Project View Shared Data How-To Help User

NCBI Abstract
Summary: Transcriptomic Analysis of Cultured Sporozoites of P. falciparum Overall Design: RNA-seq reads from each of three developmental stages (2 replicates per sample) were mapped to the reference Plasmodium falciparum genome, and gene expression levels were calculated for each sample.

Navigation Read Files Attributes

Download files

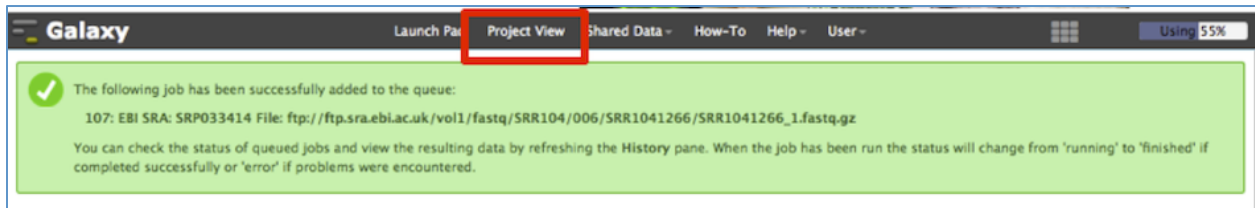
View: [TEXT](#) Download: [TEXT](#)

Select columns

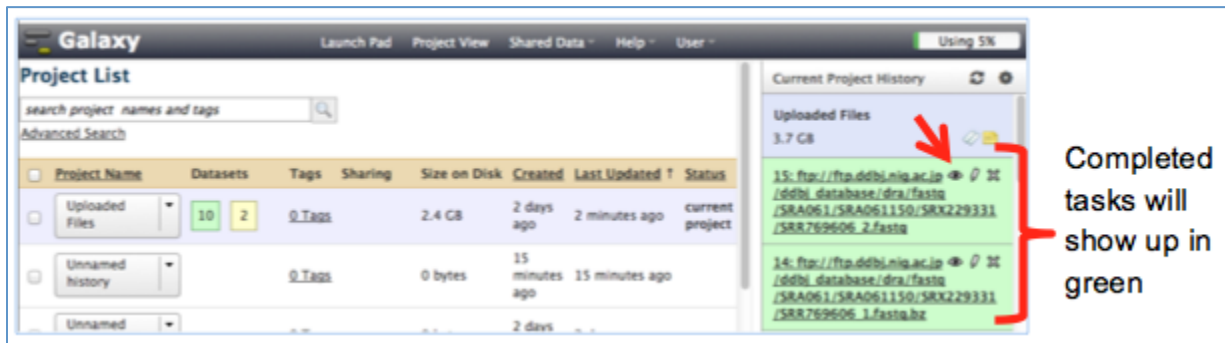
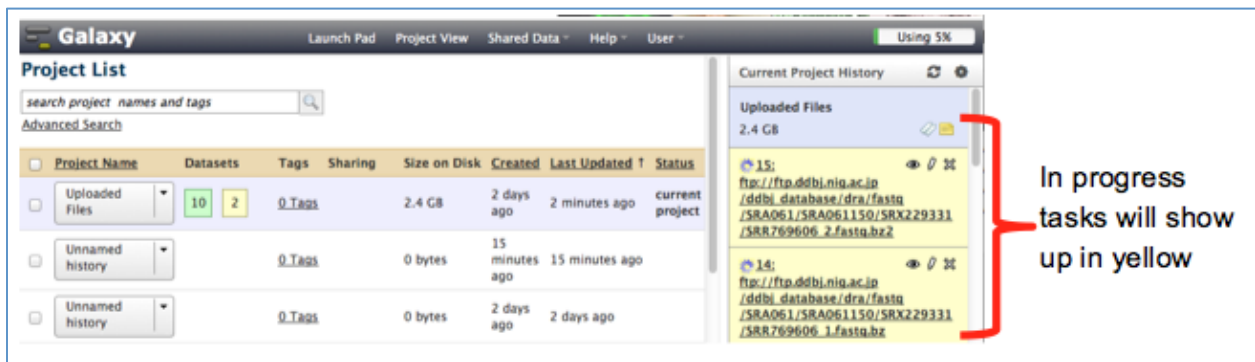
Showing results 1 - 6 of 6 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)
SRP033414	SRP033414	SAMN02428726	SRS509745	SRX385640	SRR1041266	Plasmodium falciparum NF54	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
SRP033414	SRP033414	SAMN02428729	SRS509746	SRX385641	SRR1041267	Plasmodium falciparum NF54	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
SRP033414	SRP033414	SAMN02428728	SRS509747	SRX385642	SRR1041268	Plasmodium falciparum NF54	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
SRP033414	SRP033414	SAMN02428727	SRS509748	SRX385643	SRR1041269	Plasmodium falciparum NF54	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
SRP033414	SRP033414	SAMN02428730	SRS509749	SRX385644	SRR1041270	Plasmodium falciparum NF54	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
SRP033414	SRP033414	SAMN02428734	SRS509750	SRX385645	SRR1041271	Plasmodium falciparum NF54	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2

You should now see a window that looks similar to this:



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



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:

c. Configure and initiate the RNA sequence analysis pipeline.

- i. **Background:** 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.

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

- ii. **Navigate to the workflow.** Click on the “Launch Pad” link in the upper menu bar. On the next page, scroll down to the “RNA-Seq Analysis” section and click on “Map Reads & Assemble Transcripts”.

- iii. **Select Analysis Type.** On the next page, scroll down and choose Eukaryotic Paired-End Analysis under Select Analysis Type. We are analyzing a paired end eukaryotic sample.
- iv. **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.

Select Analysis Type

Eukaryotic Single-End Analysis
 Prokaryotic Single-End Analysis
 Eukaryotic Paired-End Analysis
 Prokaryotic Paired-End Analysis

Select an existing Project or create a new Project to be used during this analysis and populate the Project with the necessary files. Output from this analysis will be saved in the selected Project.

Currently Selected Project: **Uploaded Files**

Target Project: Select existing project — OR — Create project

Uploaded Files

Source Project: Select source

Uploaded Files

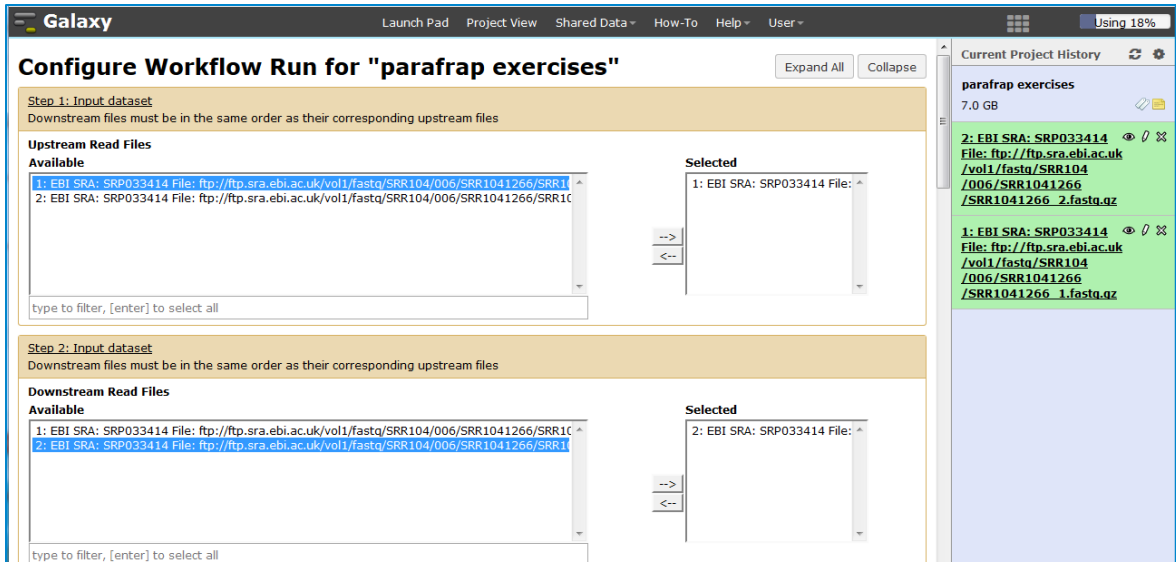
ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_2.fastq
 ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_1.fastq

Continue

- v. **Configure the pipeline.** The pipeline consists of 7 steps.

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

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



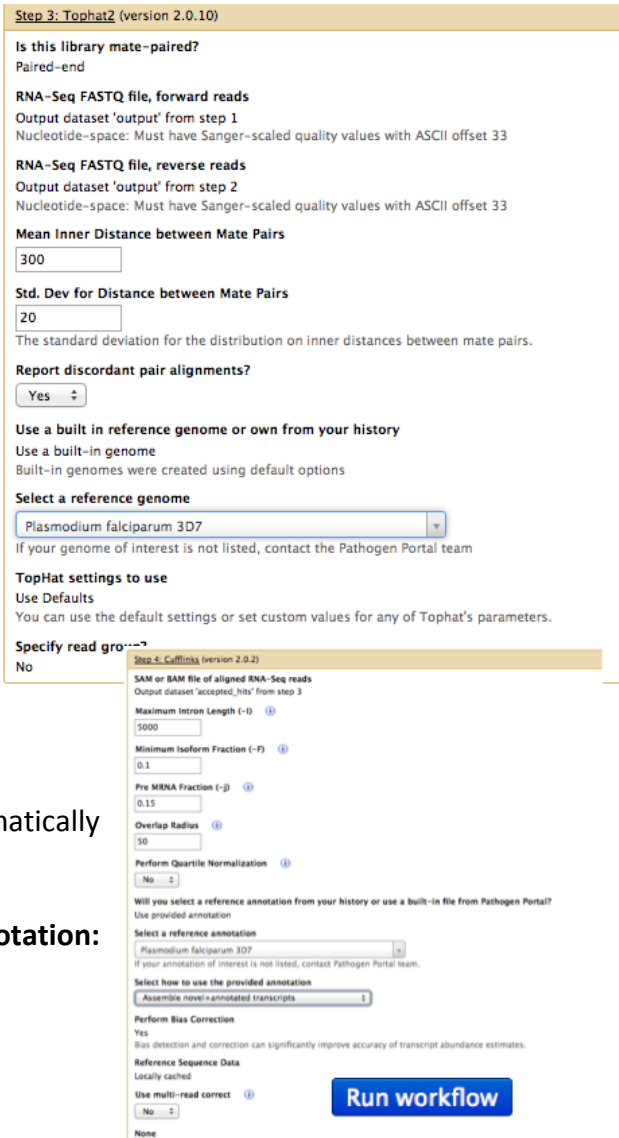
Step3: TopHat2 – Under Select a reference genome choose *Plasmodium falciparum* 3D7. There are a number of options that may be modified, however, for the purposes of this exercise the default parameters may be used.

Step4: Cufflinks –

Set the **Maximum Intron Length (-l): 5000**.

The reference annotation should be automatically selected: *Plasmodium falciparum* 3D7

Select how to use the provided annotation:
Assemble Novel + annotated transcripts.



Once again there are a number of options to modify but we only need to change the maximum Intron Length.

Step 5: BAM to BigWig – No change needed

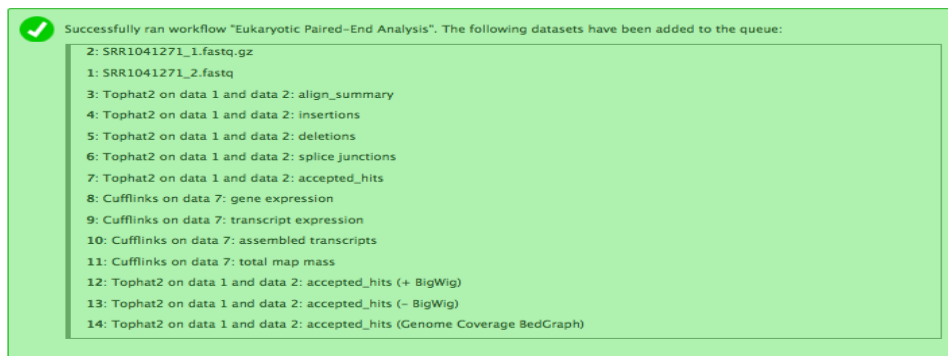
Step 6: BAM to BigWig – No change needed

Step 7: Create a BedGraph of genome coverage – No change needed

Click on the Run Workflow button.

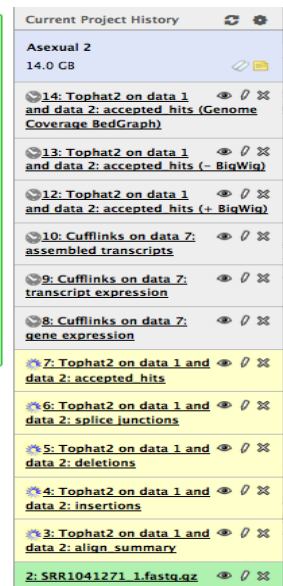
After you start the workflow you should get a confirmation window listing all the steps that have been added to the queue. The progress of your workflow can be viewed to the right. Completed tasks are in green, running tasks are in yellow and tasks waiting in the queue are in grey.

The workflow will run overnight and we will view the results and calculate differential expression in a subsequent exercise.



Successfully ran workflow "Eukaryotic Paired-End Analysis". The following datasets have been added to the queue:

- 2: SRR1041271_1.fastq.gz
- 1: SRR1041271_2.fastq
- 3: Tophat2 on data 1 and data 2: align_summary
- 4: Tophat2 on data 1 and data 2: insertions
- 5: Tophat2 on data 1 and data 2: deletions
- 6: Tophat2 on data 1 and data 2: splice junctions
- 7: Tophat2 on data 1 and data 2: accepted_hits
- 8: Cufflinks on data 7: gene expression
- 9: Cufflinks on data 7: transcript expression
- 10: Cufflinks on data 7: assembled transcripts
- 11: Cufflinks on data 7: total map mass
- 12: Tophat2 on data 1 and data 2: accepted_hits (+ BigWig)
- 13: Tophat2 on data 1 and data 2: accepted_hits (- BigWig)
- 14: Tophat2 on data 1 and data 2: accepted_hits (Genome Coverage BedGraph)



Current Project History

Asexual 2
14.0 GB

- 14: Tophat2 on data 1 and data 2: accepted_hits (Genome Coverage BedGraph)
- 13: Tophat2 on data 1 and data 2: accepted_hits (- BigWig)
- 12: Tophat2 on data 1 and data 2: accepted_hits (+ BigWig)
- 10: Cufflinks on data 7: assembled transcripts
- 9: Cufflinks on data 7: transcript expression
- 8: Cufflinks on data 7: gene expression
- 7: Tophat2 on data 1 and data 2: accepted_hits
- 6: Tophat2 on data 1 and data 2: splice junctions
- 5: Tophat2 on data 1 and data 2: deletions
- 4: Tophat2 on data 1 and data 2: insertions
- 3: Tophat2 on data 1 and data 2: align_summary
- 2: SRR1041271_1.fastq.gz