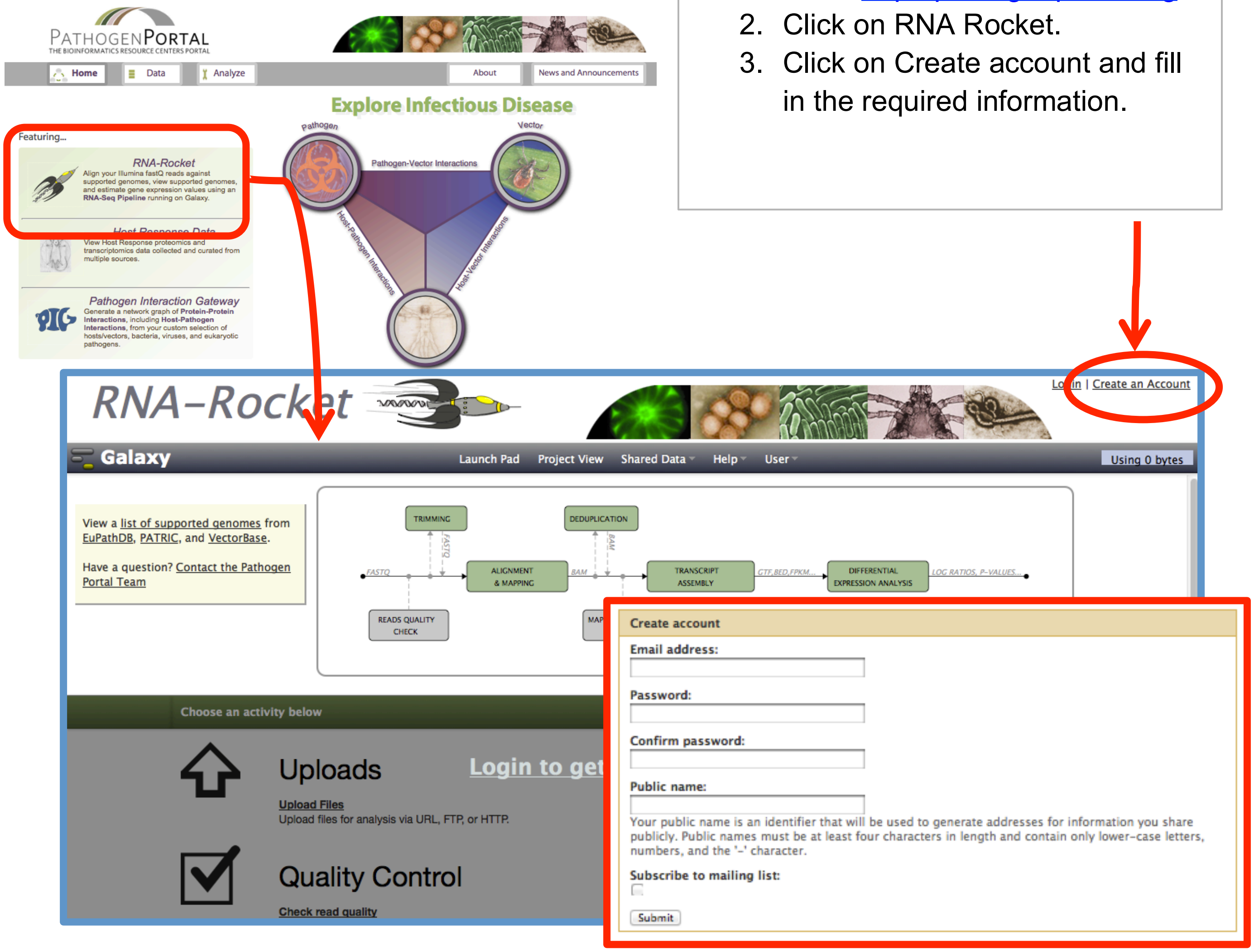
Mapping RNA sequence data
(Part 1: using pathogen portal’s RNAseq pipeline)
Exercise 3

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.



RNA-Rocket: Align your Illumina fastQ reads against supported genomes, view supported genomes, and estimate gene expression values using an RNA-Seq Pipeline running on Galaxy.

Host Response Data: View Host Response proteomics and transcriptomics data collected and curated from multiple sources.

Pathogen Interaction Gateway: Generate a network graph of Protein-Protein Interactions, including Host-Pathogen Interactions, from your custom selection of hosts/vectors, bacteria, viruses, and eukaryotic pathogens.

View a list of supported genomes from EuPathDB, PATRIC, and VectorBase.

Have a question? Contact the Pathogen Portal Team

Upload Files
Upload files for analysis via URL, FTP, or HTTP.

Create account

Email address:

Password:

Confirm password:

Public name:

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.

Subscribe to mailing list:
□

Submit

Step II: Getting data into your launch pad.

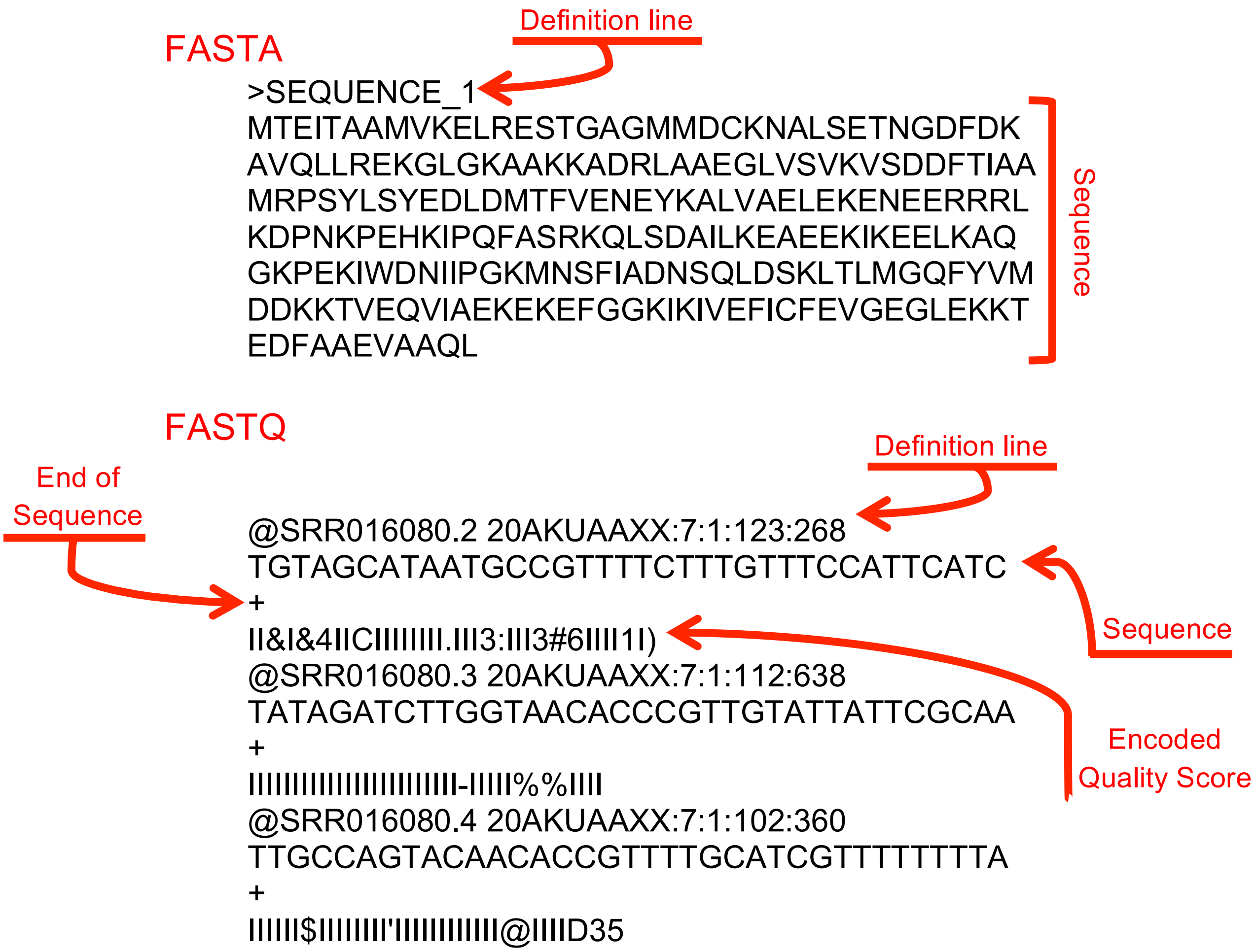
This exercise will rely on data deposited in the sequence read archive (SRA). The data is based on transcriptomic analysis of three developmental stages of *Plasmodium falciparum*: 1. Cultured asexual stages, 2. Cultured sporozoites, and 3. Salivary gland Sporozoites. Two replicates of each developmental stage were sample were generated. 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.

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 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.).



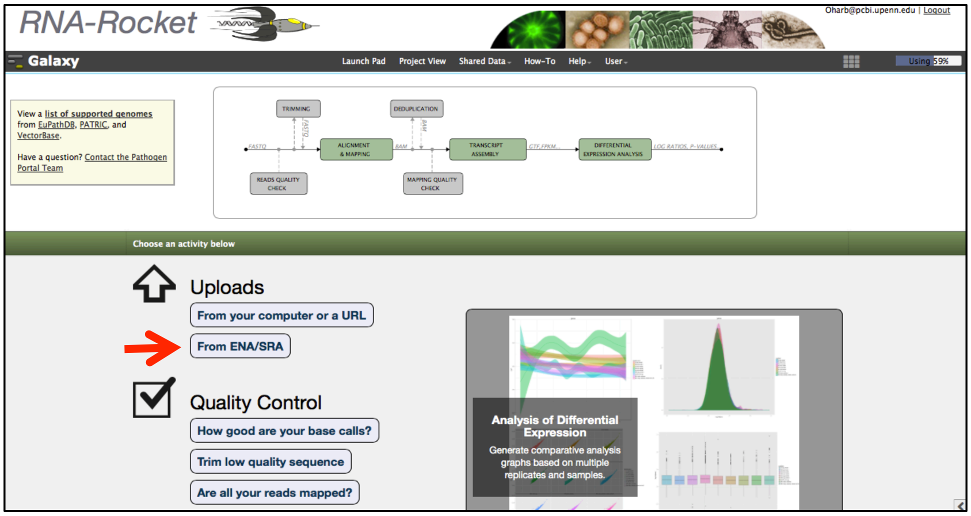
```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDK
AVQLLREKGLGKAAKKADRLAAEGLVSVKVSDDFTIAA
MRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRL
KDPNKPEHKIPQFASRKQLSDAILKEAEEKIKEELKAQ
GKPEKIWDNIIPGKMNSFIADNSQLDSKLTLMGQFYVM
DDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKT
EDFAAEVAAQL
```

```
@SRR016080.2 20AKUAAXX:7:1:123:268
TGTAGCATAATGCCGTTTTCTTTGTTTCCATTCATC
+
II&I&4IICIIIIIIII.III3:III3#6IIII1I)
@SRR016080.3 20AKUAAXX:7:1:112:638
TATAGATCTTGGTAACACCCGTTGTATTATTCGCAA
+
IIIIIIIIIIIIIIIIIIIIIIII-IIIII%%IIII
@SRR016080.4 20AKUAAXX:7:1:102:360
TTGCCAGTACAACACCGTTTTGCATCGTTTTTTTTA
+
IIIIII$IIIIIIII'IIIIIIIIIIII@IIIID35
```

- 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.

- Here are the steps you take to start uploading 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 SRA to RNA-Rocket.

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.



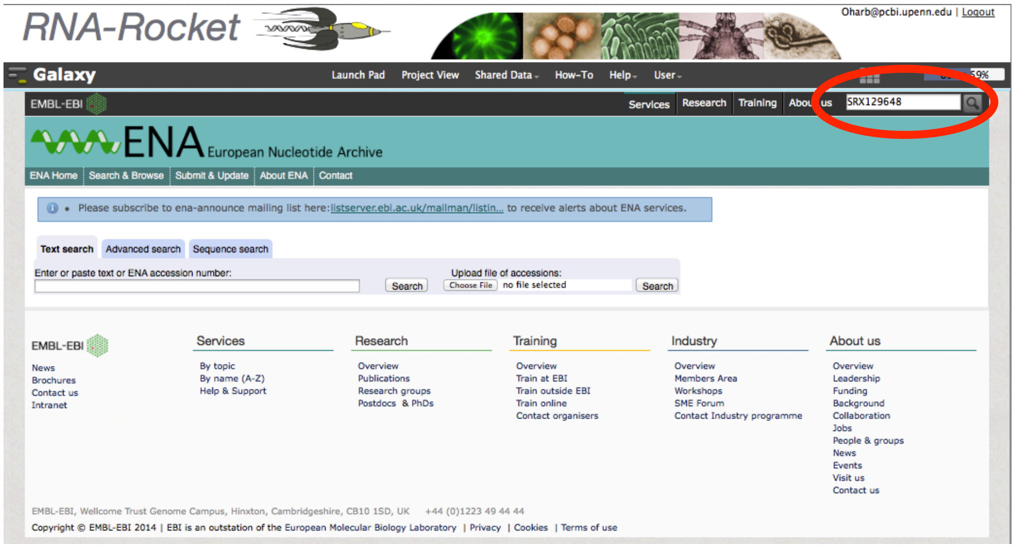
View a list of supported genomes from EuPathDB, PATRIC, and VectorBase.

Have a question? Contact the Pathogen Portal Team

Analysis of Differential Expression

Generate comparative analysis graphs based on multiple replicates and samples.

3. Cut and paste the study accession number (SRP033414) into the global search box (see red circle below). Click on the search icon.



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

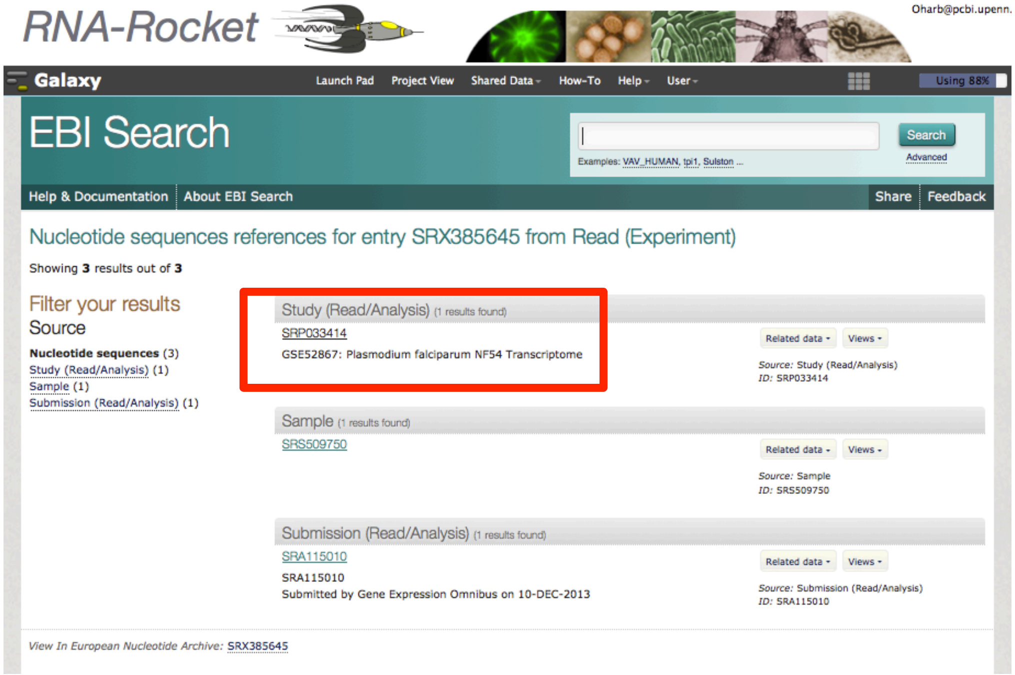
Enter or paste text or ENA accession number:

Upload file of accessions:

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK +44 (0)1223 49 44 44

Copyright © EMBL-EBI 2014 | EBI is an outstation of the European Molecular Biology Laboratory | Privacy | Cookies | Terms of use

4. Click on the Study link obtained.



Nucleotide sequences references for entry SRX385645 from Read (Experiment)

Showing **3** results out of **3**

Filter your results

Source

Nucleotide sequences (3)
Study (Read/Analysis) (1)
Sample (1)
Submission (Read/Analysis) (1)

Study (Read/Analysis) (1 results found)

SRP033414

GSE52867: Plasmodium falciparum NF54 Transcriptome

Source: Study (Read/Analysis)
ID: SRP033414

Sample (1 results found)

SRS509750

Source: Sample
ID: SRS509750

Submission (Read/Analysis) (1 results found)

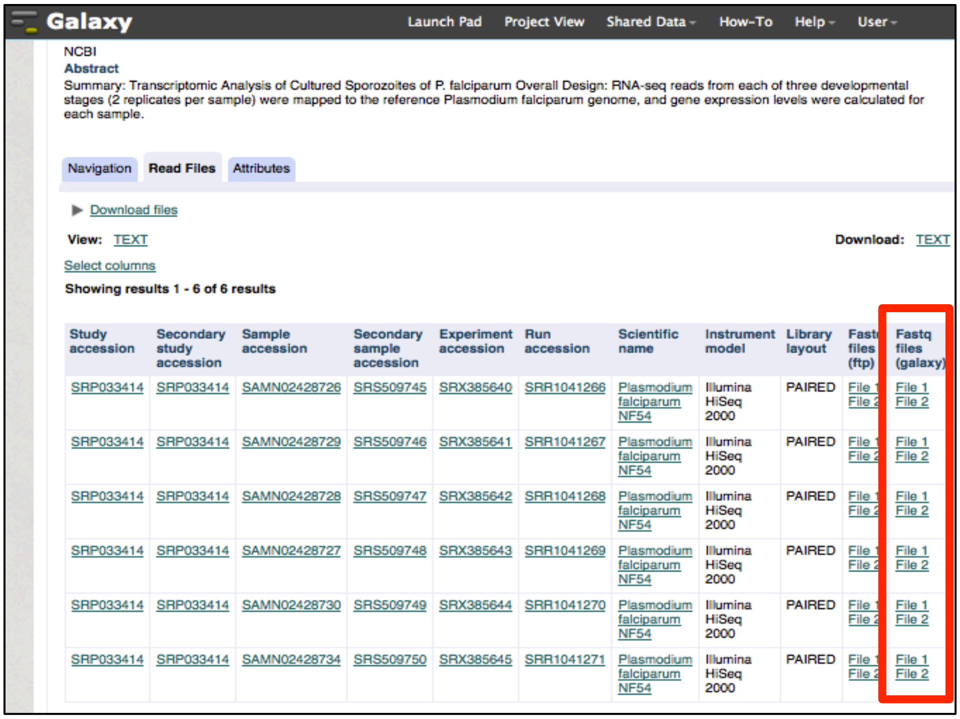
SRA115010

SRA115010
Submitted by Gene Expression Omnibus on 10-DEC-2013

Source: Submission (Read/Analysis)
ID: SRA115010

View In European Nucleotide Archive: SRX385645

5. To transfer files to RNA-Rocket, click on the File 1 or File 2 in the column called “Fastq files (galaxy)”. Remember, you have to get 2 files, one for each pair. Click on the link for File 1 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.



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.

▶ Download files

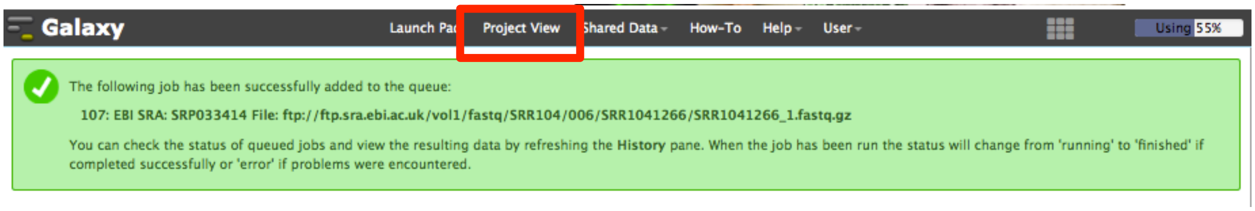
View: TEXT **Download:** TEXT

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	Fast 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:

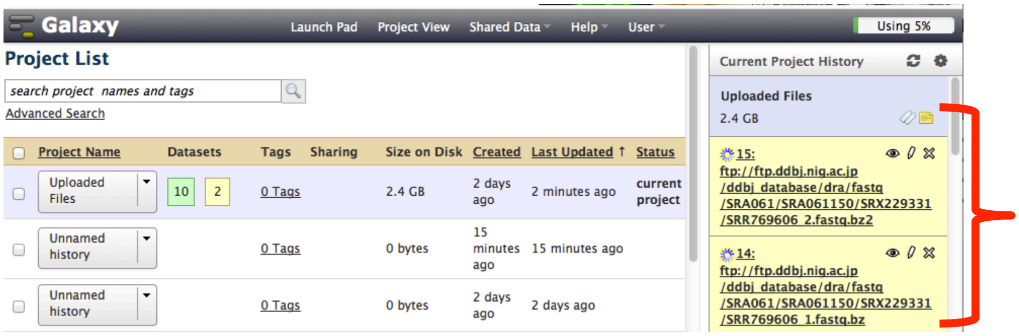


The following job has been successfully added to the queue:

107: EBI SRA: SRP033414 File: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/006/SRR1041266/SRR1041266_1.fastq.gz

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

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



Project List

	Project Name	Datasets		Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
□	Uploaded Files	10	2	0 Tags		2.4 GB	2 days ago	2 minutes ago	current project
□	Unnamed history			0 Tags		0 bytes	15 minutes ago	15 minutes ago	
□	Unnamed history			0 Tags		0 bytes	2 days ago	2 days ago	

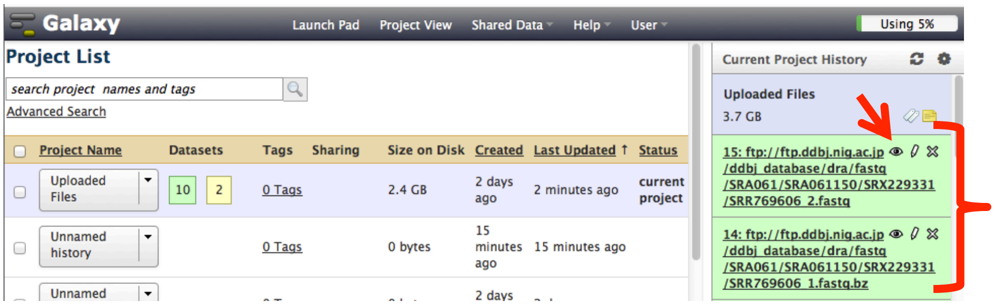
Current Project History

Uploaded Files
2.4 GB

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

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

In progress tasks will show up in yellow



Project List

	Project Name	Datasets		Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
□	Uploaded Files	10	2	0 Tags		2.4 GB	2 days ago	2 minutes ago	current project
□	Unnamed history			0 Tags		0 bytes	15 minutes ago	15 minutes ago	
□	Unnamed						2 days		

Current Project History

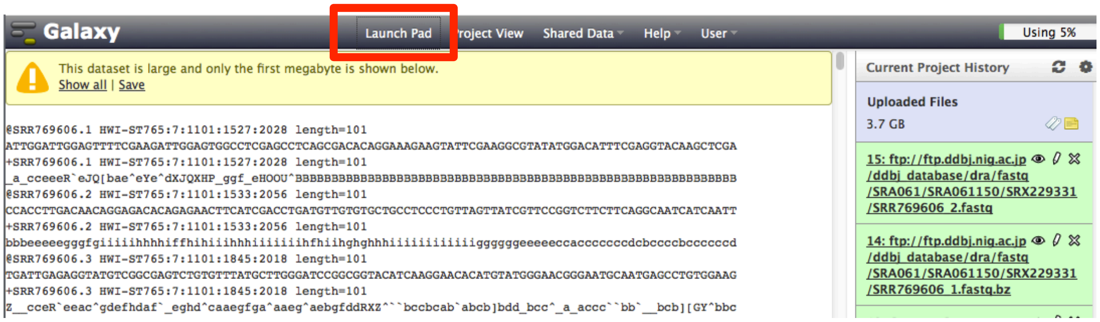
Uploaded Files
3.7 GB

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

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

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:



This dataset is large and only the first megabyte is shown below.
Show all | Save

```
@SRR769606.1 HWI-ST765:7:1101:1527:2028 length=101
ATTGGATTGGAGTTTTCGAAGATTGGAGTGGCCTCGAGCCTCAGCGACACAGGAAAGAAGTATTCGAAGGCGTATATGGACATTTCGAGGTACAAGCTCGA
+SRR769606.1 HWI-ST765:7:1101:1527:2028 length=101
_a_cceeeR`eJQ[bae^eYe^dXJQXHP_ggf_eHOOU^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@SRR769606.2 HWI-ST765:7:1101:1533:2056 length=101
CCACCTTGACAACAGGAGACACAGAGAACTTCATCGACCTGATGTTGTGTGCTGCCTCCCTGTTAGTTATCGTTCCGGTCTTCTTCAGGCAATCATCAATT
+SRR769606.2 HWI-ST765:7:1101:1533:2056 length=101
bbbeeeeegggfgiiiiihhhhiffhihiiihhhiiiiiiihfhiihghghhhiiiiiiiiiiiigggggeeeeeccaccccccdcbccccbcccccccd
@SRR769606.3 HWI-ST765:7:1101:1845:2018 length=101
TGATTGAGAGGTATGTCGGCGAGTCTGTGTTTATGCTTGGGATCCGGCGGTACATCAAGGAACACATGTATGGGAACGGGAATGCAATGAGCCTGTGGAAG
+SRR769606.3 HWI-ST765:7:1101:1845:2018 length=101
Z__cceR`eeac^gdefhdaf`_eghd^caaegfga^aaeg^aebgfddRXZ^``bccbcab`abcb]bdd_bcc^_a_accc``bb`__bcb][GY^bbc
```

Current Project History

Uploaded Files
3.7 GB

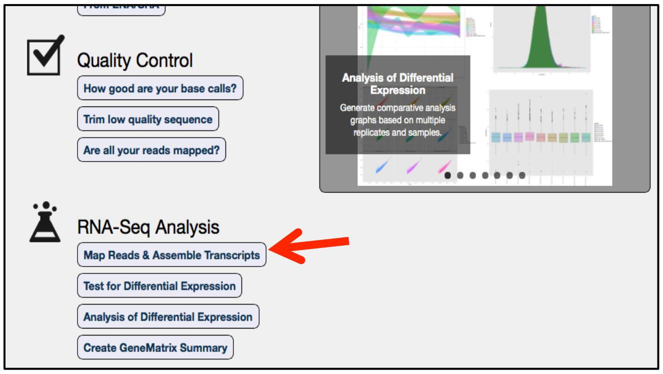
15: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_2.fastq

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

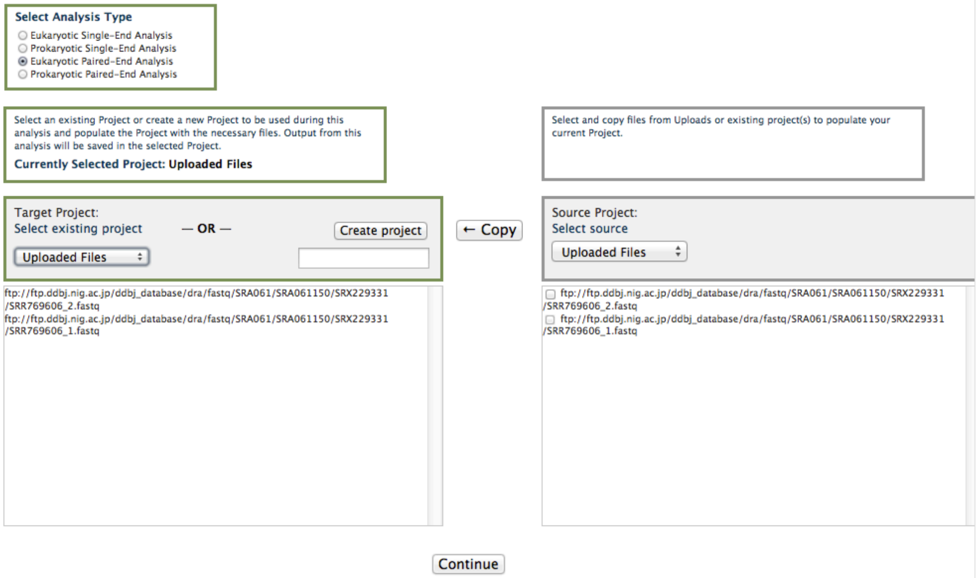
6. Once the RNA-sequence FASTQ file has been uploaded you can start the RNA-seq pipeline. Pathogen portal uses two algorithms for mapping (TopHat) and transcript prediction and expression value calculation (Cufflinks). Note that there are many algorithms and methods for RNA-seq mapping and analysis each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


Quality Control
How good are your base calls?
Trim low quality sequence
Are all your reads mapped?
Analysis of Differential Expression
Generate comparative analysis graphs based on multiple replicates and samples.
RNA-Seq Analysis
Map Reads & Assemble Transcripts
Test for Differential Expression
Analysis of Differential Expression
Create GeneMatrix Summary

- 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.



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

Select and copy files from Uploads or existing project(s) to populate your current Project.

Target Project:
Select existing project — OR —
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
```

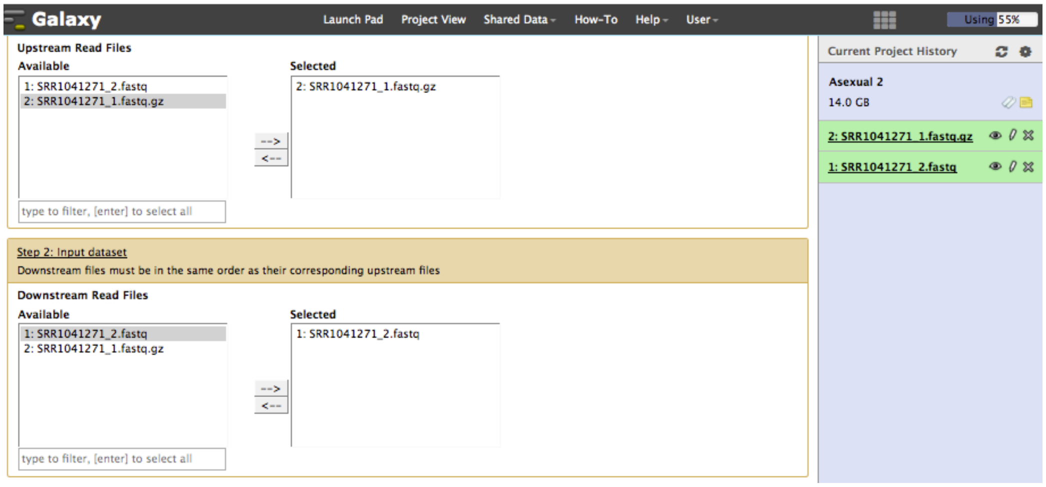
Source Project:
Select source
Uploaded Files

```
□ 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
```

- The next page allows you to configure the pipeline:

<u>Step1:</u> Select the upstream read file (ends in _1) and click on the arrow to move it to the "Selected" window.

<u>Step2:</u> Select the downstream read file (ends in _2) and click on the arrow to move it to the "Selected" window.



Upstream Read Files

Available
1: SRR1041271_2.fastq
2: SRR1041271_1.fastq.gz

Selected
2: SRR1041271_1.fastq.gz

type to filter, [enter] to select all

Step 2: Input dataset

Downstream files must be in the same order as their corresponding upstream files

Downstream Read Files

Available
1: SRR1041271_2.fastq
2: SRR1041271_1.fastq.gz

Selected
1: SRR1041271_2.fastq

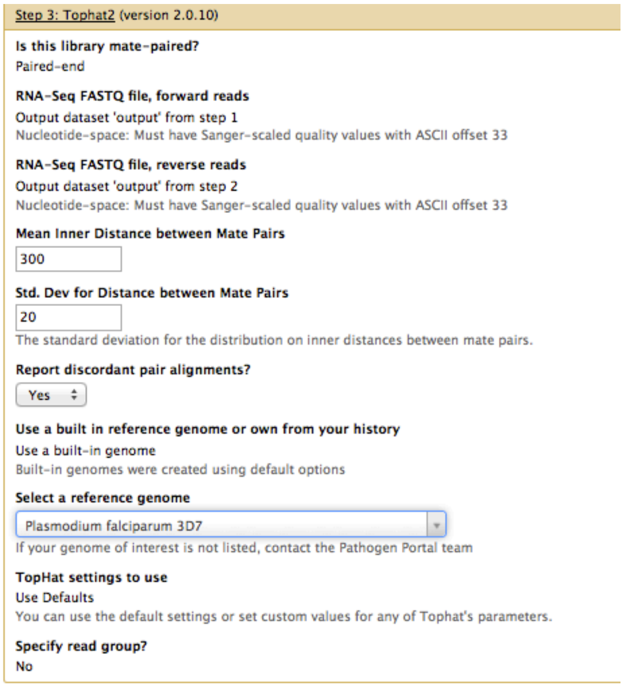
type to filter, [enter] to select all

Current Project History

Asexual 2
14.0 GB

2: SRR1041271_1.fastq.gz
1: SRR1041271_2.fastq

<u>Step3:</u> 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 *Plasmodium falciparum* 3D7



Step 3: Tophat2 (version 2.0.10)

Is this library mate-paired?
Paired-end

RNA-Seq FASTQ file, forward reads
Output dataset 'output' from step 1
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

RNA-Seq FASTQ file, reverse reads
Output dataset 'output' from step 2
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs
300

Std. Dev for Distance between Mate Pairs
20
The standard deviation for the distribution on inner distances between mate pairs.

Report discordant pair alignments?
Yes

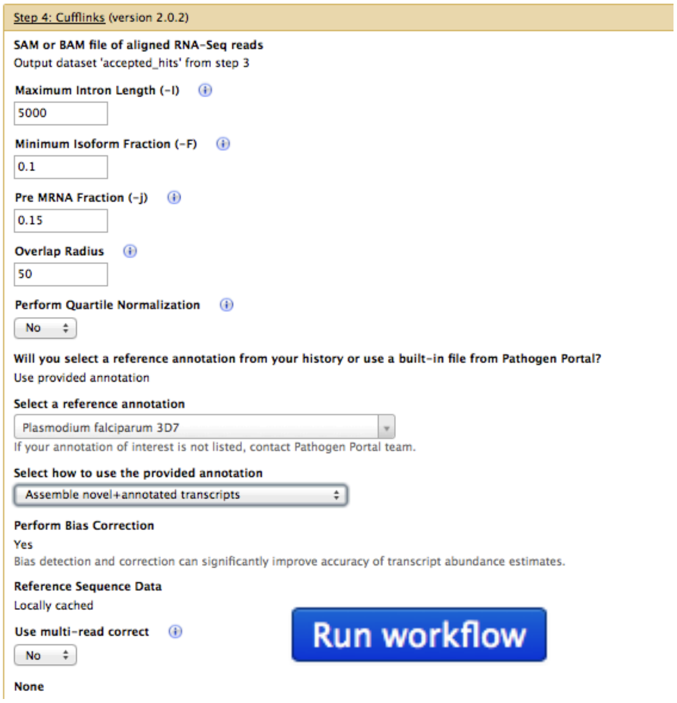
Use a built in reference genome or own from your history
Use a built-in genome
Built-in genomes were created using default options

Select a reference genome
Plasmodium falciparum 3D7
If your genome of interest is not listed, contact the Pathogen Portal team

TopHat settings to use
Use Defaults
You can use the default settings or set custom values for any of Tophat's parameters.

Specify read group?
No

<u>Step4:</u> 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): 5000
The reference annotation should be automatically selected: *Plasmodium falciparum* 3D7
Select how to use the provided annotation: Assemble Novel + annotated transcripts.



Step 4: Cufflinks (version 2.0.2)

SAM or BAM file of aligned RNA-Seq reads
Output dataset 'accepted_hits' from step 3

Maximum Intron Length (-I)
5000

Minimum Isoform Fraction (-F)
0.1

Pre MRNA Fraction (-j)
0.15

Overlap Radius
50

Perform Quartile Normalization
No

Will you select a reference annotation from your history or use a built-in file from Pathogen Portal?
Use provided annotation

Select a reference annotation
Plasmodium falciparum 3D7
If your annotation of interest is not listed, contact Pathogen Portal team.

Select how to use the provided annotation
Assemble novel+annotated transcripts

Perform Bias Correction
Yes
Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

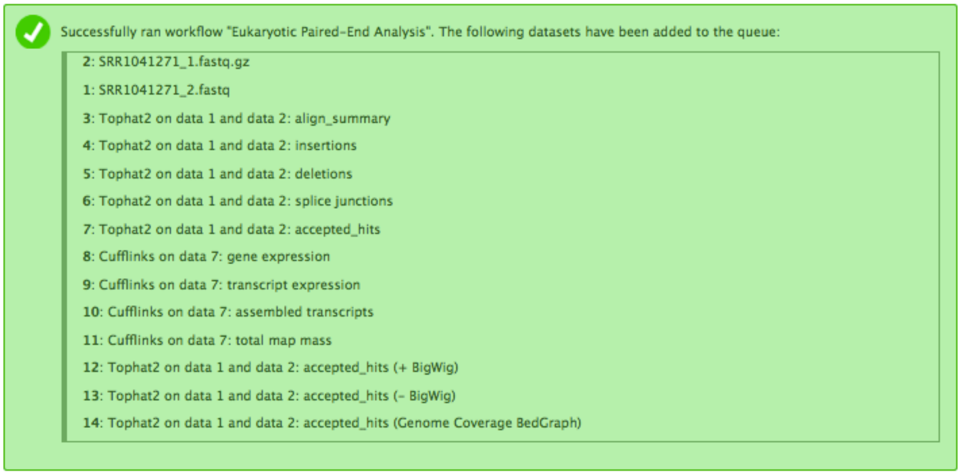
Reference Sequence Data
Locally cached

Use multi-read correct
No

None

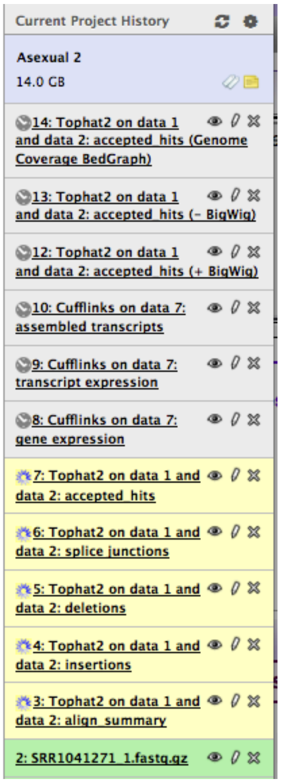
Click on the Run Workflow button.

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.



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