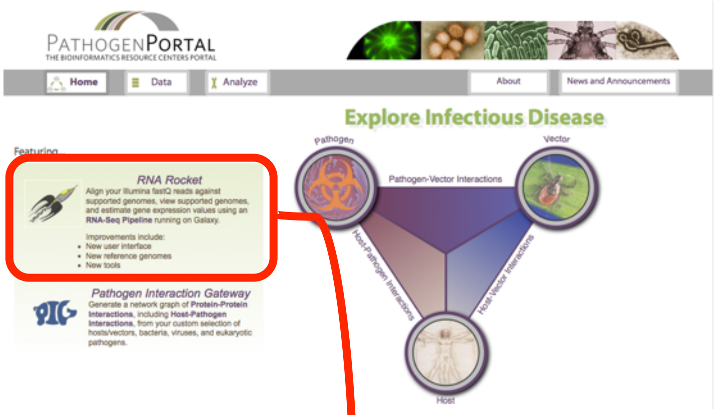
RNA sequence data analysis
(Part 1: using pathogen portal’s RNAseq pipeline)
Exercise 3

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


PATHOGENPORTAL
THE BIOINFORMATICS RESOURCE CENTERS PORTAL
Home
Data
Analyze
About
News and Announcements
Explore Infectious Disease
Pathogen
Vector
Host
Pathogen-Vector Interactions
Host-Pathogen Interactions
Host-Vector Interactions
Featuring...

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.

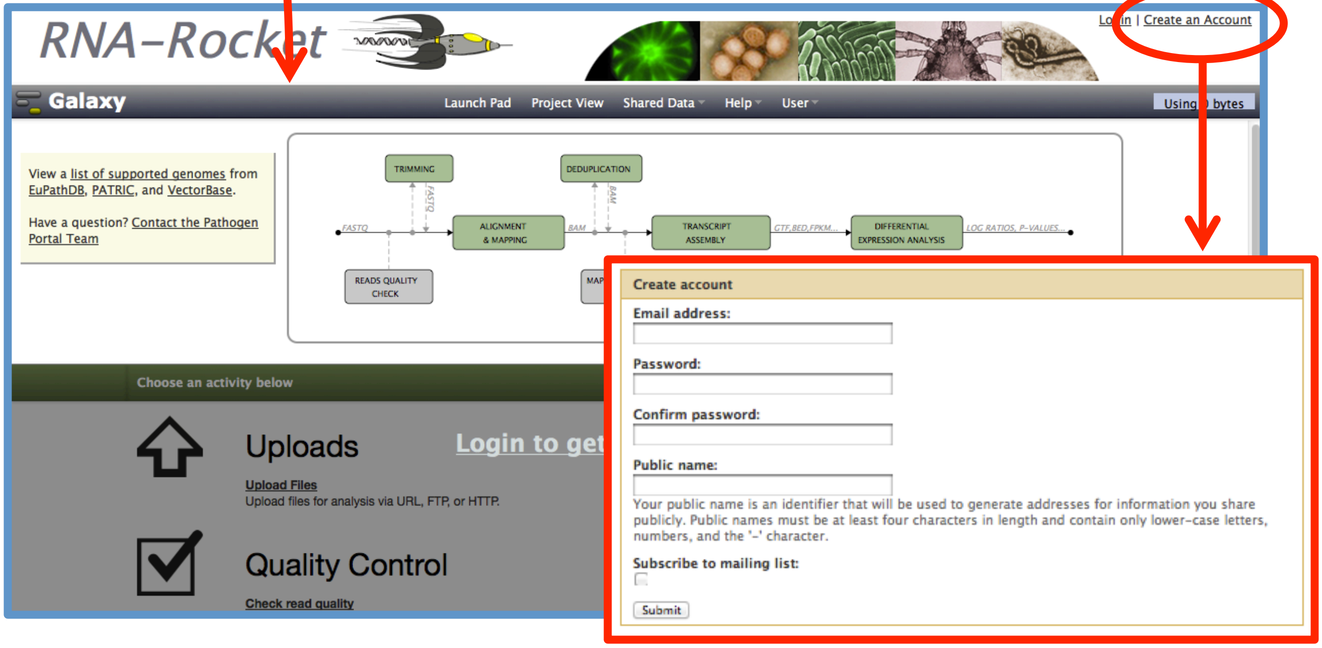
Improvements include:
- New user interface
- New reference genomes
- New tools

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.

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.



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

Have a question? Contact the Pathogen Portal Team

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.

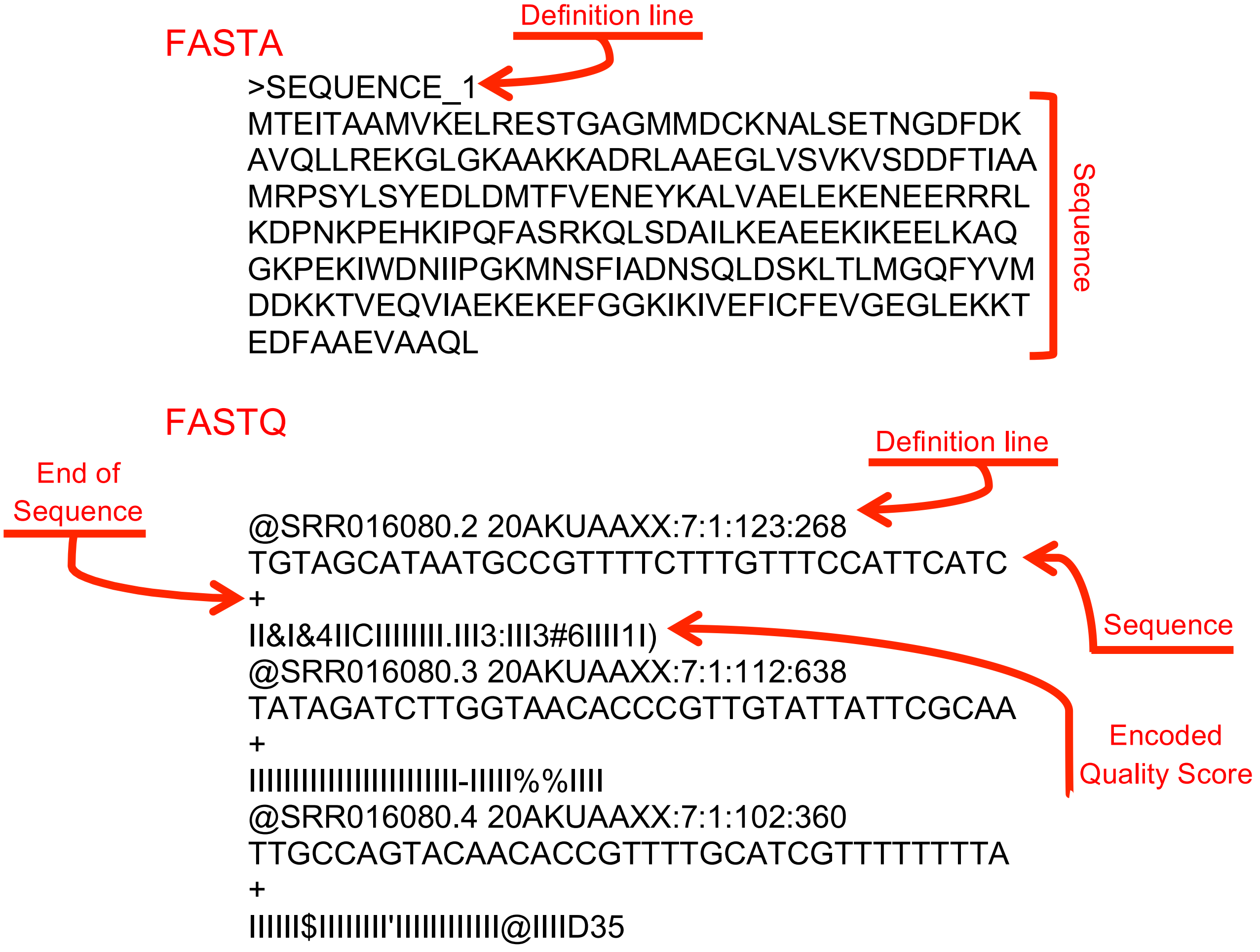
The following exercise is based on data generated from the recent study:
Grisdale *et al.* Transcriptome analysis of the parasite *Encephalitozoon cuniculi*: an in-depth examination of pre-mRNA splicing in a reduced eukaryote. BMC Genomics 2013, 14:207

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.).



FASTA

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

FASTQ

```
@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. 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:

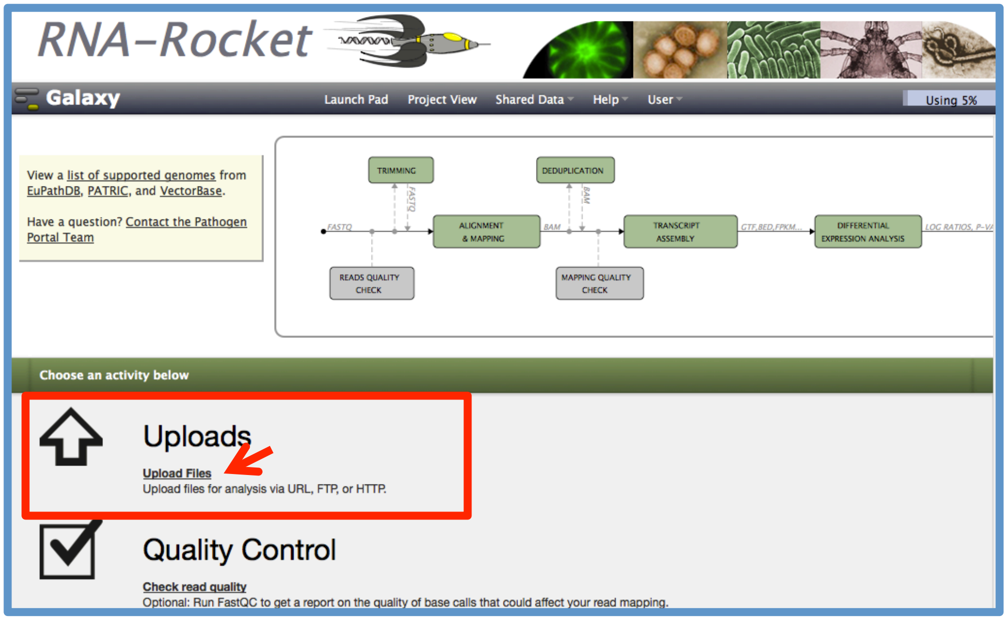
ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX247418/SRR769608_1.fastq.bz2

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



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

Have a question? Contact the Pathogen Portal Team

Uploads

Upload Files

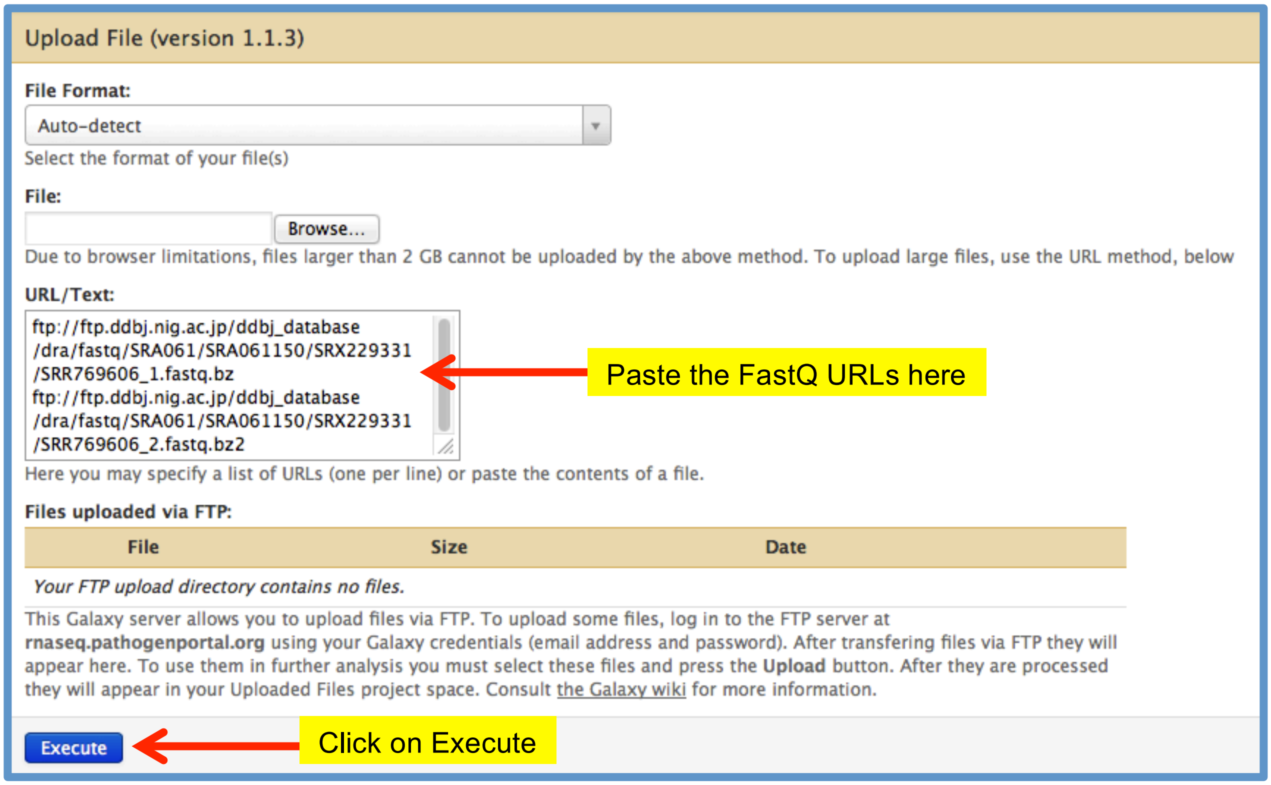
Upload files for analysis via URL, FTP, or HTTP.

Quality Control

Check read quality

Optional: Run FastQC to get a report on the quality of base calls that could affect your read mapping.

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.



Upload File (version 1.1.3)

File Format:

Auto-detect

Select the format of your file(s)

File:

Browse...

Due to browser limitations, files larger than 2 GB cannot be uploaded by the above method. To upload large files, use the URL method, below

URL/Text:

```
ftp://ftp.ddbj.nig.ac.jp/ddbj_database
/dra/fastq/SRA061/SRA061150/SRX229331
/SRR769606_1.fastq.bz
ftp://ftp.ddbj.nig.ac.jp/ddbj_database
/dra/fastq/SRA061/SRA061150/SRX229331
/SRR769606_2.fastq.bz2
```

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
Your FTP upload directory contains no files.		

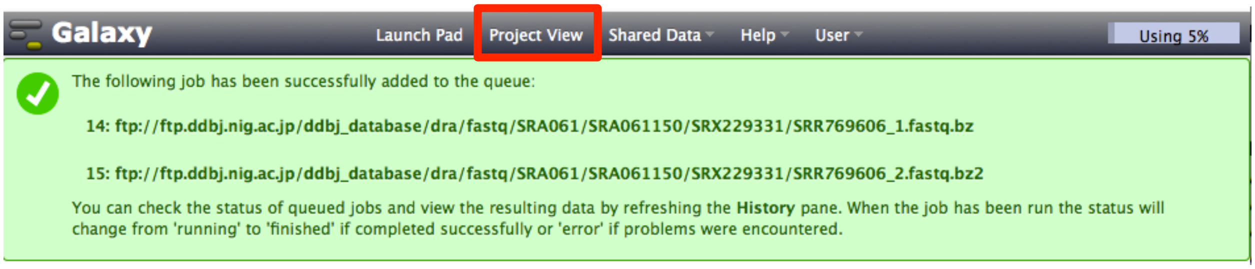
This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at **rnaseq.pathogenportal.org** using your Galaxy credentials (email address and password). After transfering files via FTP they will appear here. To use them in further analysis you must select these files and press the **Upload** button. After they are processed they will appear in your Uploaded Files project space. Consult the Galaxy wiki for more information.

Execute

Paste the FastQ URLs here

Click on Execute

You should now see a window that looks like this:



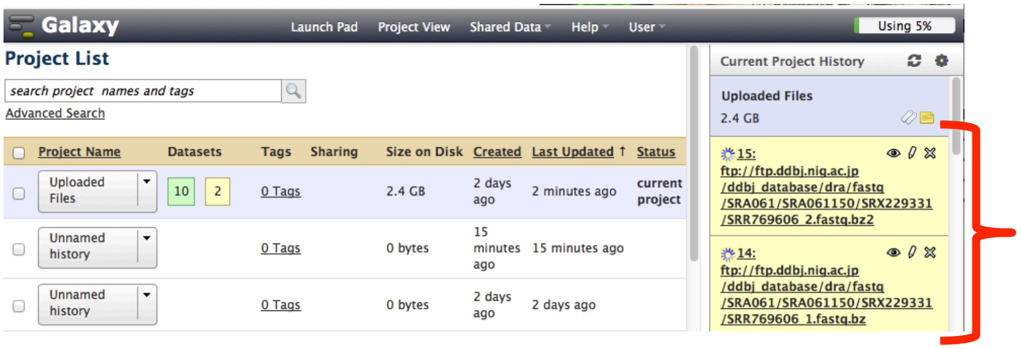
The following job has been successfully added to the queue:

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

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

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

search project names and tags

Advanced Search

☐	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

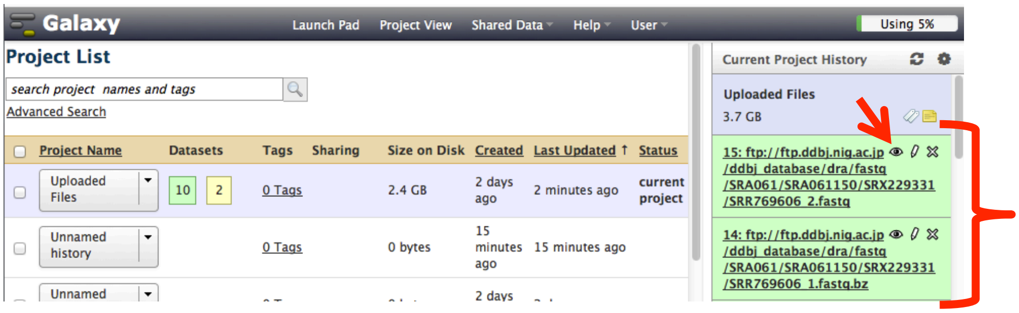
Uploaded Files

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

search project names and tags

Advanced Search

□	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

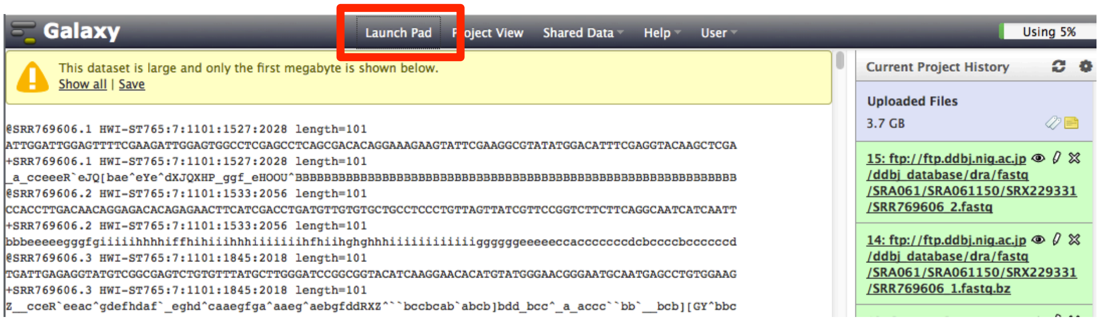
Uploaded Files
3.7 GB

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^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@SRR769606.2 HWI-ST765:7:1101:1533:2056 length=101
CCACCTTGACAACAGGAGACACAGAGAACTTCATCGACCTGATGTTGTGTGCTGCCTCCCTGTTAGTTATCGTTCCGGTCTTCTTCAGGCAATCATCAATT
+SRR769606.2 HWI-ST765:7:1101:1533:2056 length=101
bbbeeeeegggfgiiiiihhhhiffhihiiihhhiiiiiiihfhiihghghhhiiiiiiiiiiiigggggeeeeeccaccccccdcbccccbcccccd
@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

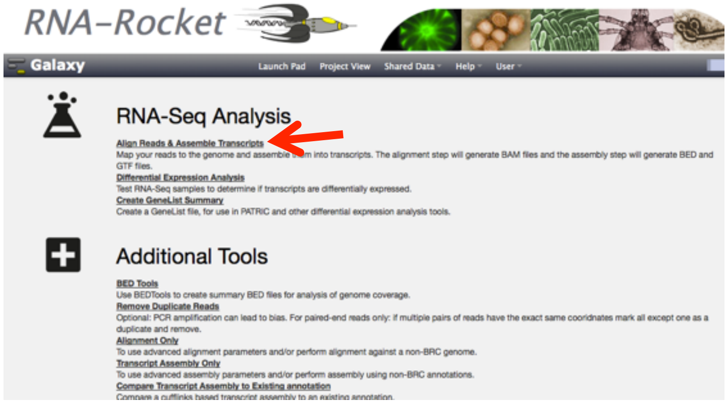
15: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_2.fastq

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

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".



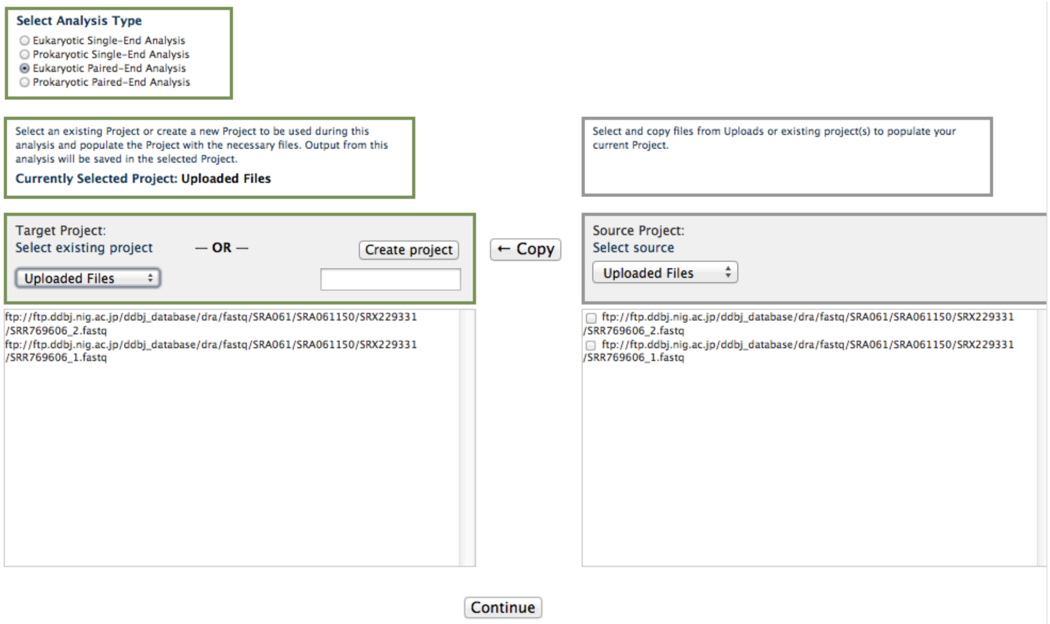
RNA-Seq Analysis

Align Reads & Assemble Transcripts
Map your reads to the genome and assemble them into transcripts. The alignment step will generate BAM files and the assembly step will generate BED and GTF files.
Differential Expression Analysis
Test RNA-Seq samples to determine if transcripts are differentially expressed.
Create GeneList Summary
Create a GeneList file, for use in PATRIC and other differential expression analysis tools.

Additional Tools

BED Tools
Use BEDTools to create summary BED files for analysis of genome coverage.
Remove Duplicate Reads
Optional: PCR amplification can lead to bias. For paired-end reads only: if multiple pairs of reads have the exact same cooridnates mark all except one as a duplicate and remove.
Alignment Only
To use advanced alignment parameters and/or perform alignment against a non-BRC genome.
Transcript Assembly Only
To use advanced assembly parameters and/or perform assembly using non-BRC annotations.
Compare Transcript Assembly to Existing annotation
Compare a cufflinks based transcript assembly to an existing annotation.

- 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: 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

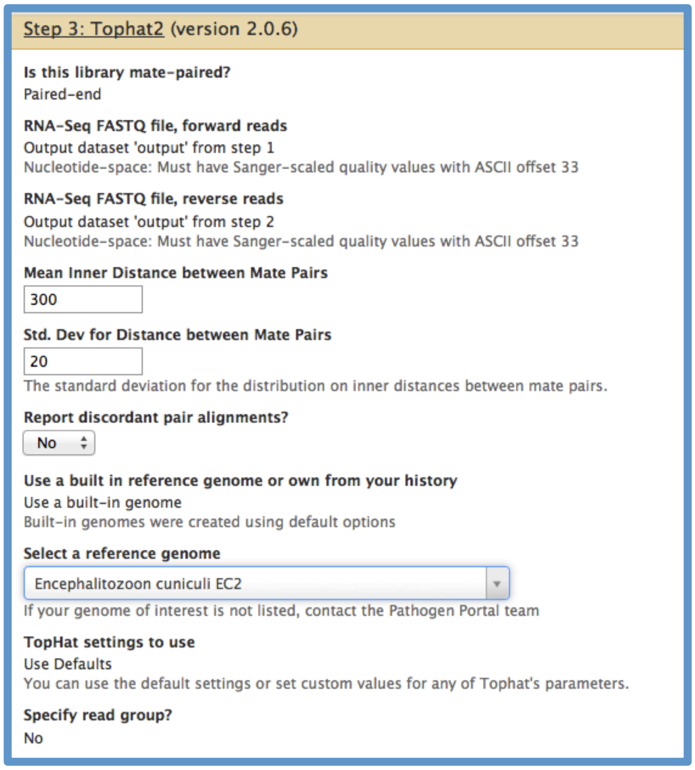
□ ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_2.fastq
□ ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_1.fastq

- The next page allows you to configure the pipeline:

  __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



Step 3: Tophat2 (version 2.0.6)

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?
No

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

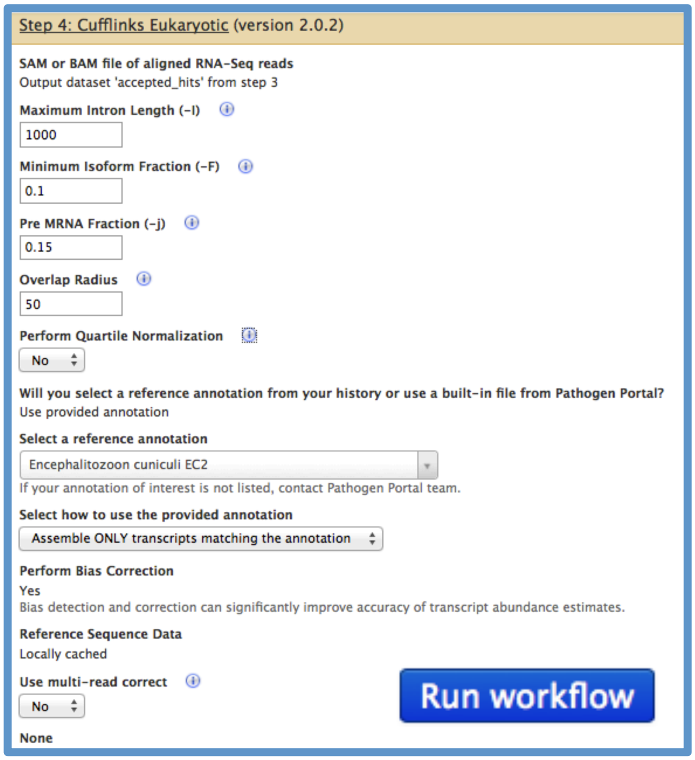
Select a reference genome
Encephalitozoon cuniculi EC2
If your genome of interest is not listed, contact the Pathogen Portal team

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

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.



Step 4: Cufflinks Eukaryotic (version 2.0.2)

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

Maximum Intron Length (-I)
1000

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
Encephalitozoon cuniculi EC2
If your annotation of interest is not listed, contact Pathogen Portal team.

Select how to use the provided annotation
Assemble ONLY transcripts matching the annotation

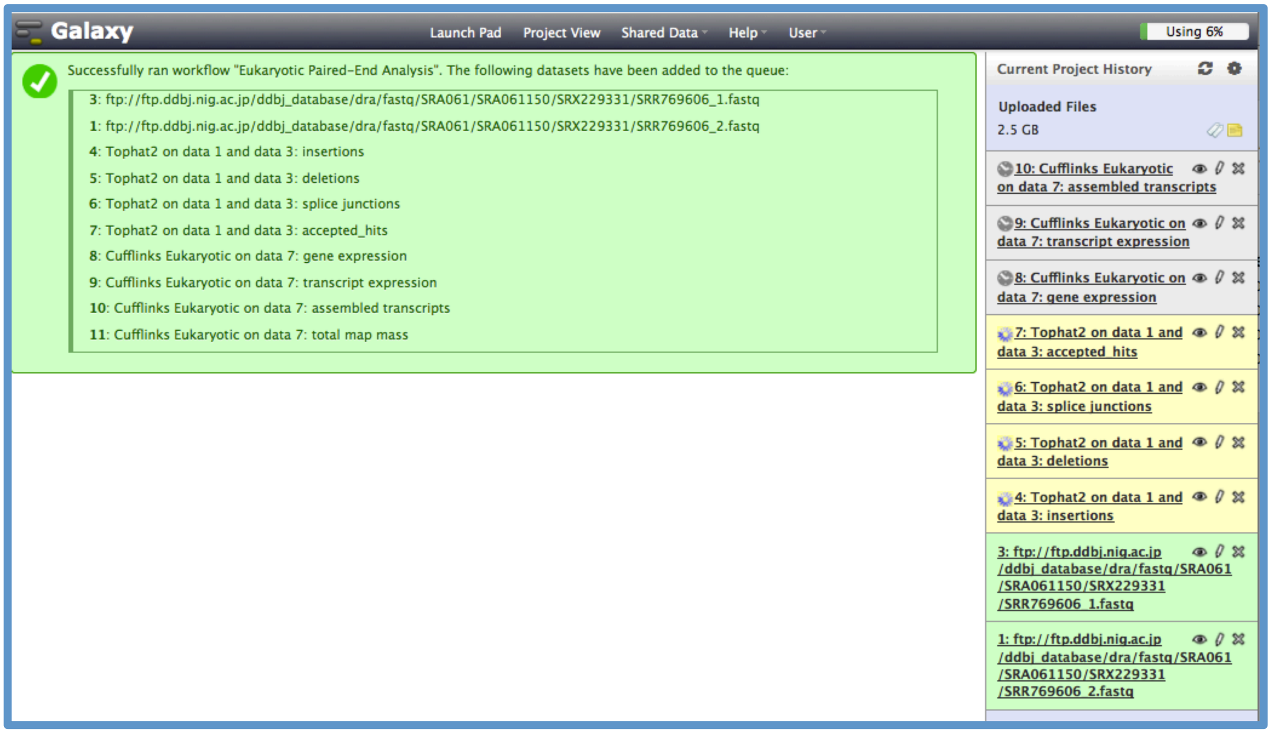
Perform Bias Correction
Yes
Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Reference Sequence Data
Locally cached

Use multi-read correct
No

None

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:

- 3: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_1.fastq
- 1: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_2.fastq
- 4: Tophat2 on data 1 and data 3: insertions
- 5: Tophat2 on data 1 and data 3: deletions
- 6: Tophat2 on data 1 and data 3: splice junctions
- 7: Tophat2 on data 1 and data 3: accepted_hits
- 8: Cufflinks Eukaryotic on data 7: gene expression
- 9: Cufflinks Eukaryotic on data 7: transcript expression
- 10: Cufflinks Eukaryotic on data 7: assembled transcripts
- 11: Cufflinks Eukaryotic on data 7: total map mass

Uploaded Files
2.5 GB

- 10: Cufflinks Eukaryotic on data 7: assembled transcripts
- 9: Cufflinks Eukaryotic on data 7: transcript expression
- 8: Cufflinks Eukaryotic on data 7: gene expression
- 7: Tophat2 on data 1 and data 3: accepted_hits
- 6: Tophat2 on data 1 and data 3: splice junctions
- 5: Tophat2 on data 1 and data 3: deletions
- 4: Tophat2 on data 1 and data 3: insertions
- 3: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_1.fastq
- 1: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331/SRR769606_2.fastq