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 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 indepth 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.

Upload File (version 1.1.3)			
File Format: Auto-detect Select the format of your file(s) File: Browse Due to browser limitations, files large URL/Text: ftp://ftp.ddbj.nig.ac.jp/ddbj_databa /dra/fastq/SRA061/SRA061150/SRX /SRR769606_1.fastq.bz ftp://ftp.ddbj.nig.ac.jp/ddbj_databa	e er than 2 GB cannot be up se (229331	Ioaded by the above method. To upload large files, use the URL metho Paste the FastQ URLs here	od, below
/dra/fastq/SRA061/SRA061150/SRX /SRR769606_2.fastq.bz2	229331		
Here you may specify a list of URLs (o	one per line) or paste the o	contents of a file.	
Files uploaded via FTP: File	Size	Date	
Your FTP upload directory contains r	no files.		
This Galaxy server allows you to uplo rnaseq.pathogenportal.org using yo appear here. To use them in further a they will appear in your Uploaded File	ad files via FTP. To uploa ur Galaxy credentials (em inalysis you must select tl is project space. Consult	d some files, log in to the FTP server at ail address and password). After transfering files via FTP they will hese files and press the Upload button. After they are processed <u>the Galaxy wiki</u> for more information.	
Execute CI	lick on Execute		

You should now see a window that looks like this:

Galaxy	Launch Pad	Project View	Shared Data -	Help –	User →	Using 5%
The following job has been successful	ly added to the	e queue:				
, 14: ftp://ftp.ddbj.nig.ac.jp/ddbj_da	atabase/dra/f	astq/SRA061/S	RA061150/SRX2	29331/S	RR769606_1.fastq.bz	
15: ftp://ftp.ddbj.nig.ac.jp/ddbj_da	atabase/dra/f	astq/SRA061/S	RA061150/SRX2	29331/S	RR769606_2.fastq.bz2	
You can check the status of queued jo change from 'running' to 'finished' if c	bs and view th ompleted succ	e resulting data essfully or 'error	by refreshing the r' if problems wer	History e encount	pane. When the job has been run the st ered.	atus will

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

💳 Galaxy	Launch Pad Project View	Shared Data - Help - User -	Using 5%	
Project List			Current Project History 🛛 🗢	
search project names and tags	Ð,		Uploaded Files 2.4 GB	
Datasets Tags	Sharing Size on Disk	Created Last Updated ↑ Status	<u>**15:</u> ● ℓ ×	
Uploaded VIII 2 0 Tac	<u>gs</u> 2.4 GB	2 days 2 minutes ago current project	ftp://ftp.ddbj.nig.ac.jp /ddbj_database/dra/fastq /SRA061/SRA06145/SRA229331	In progress
Unnamed history	gs0 bytes	15 minutes 15 minutes ago ago	<u>/3RK/09000 2.fastq.bzz</u> <u>≫14:</u> ● Ø % ftp://ftp.ddbj.nig.ac.jp	up in yellow
Unnamed history	gs 0 bytes	2 days 2 days ago 2	/ddbj_database/dra/fastq /SRA061/SRA061150/SRX229331 /SRR769606_1.fastq.bz	

5	Galaxy	Launch Pad	Project View	Shared D	ata - Help -	User -	Using 5%
Pro	oject List						Current Project History
sea Adva	rch project names and tags anced Search	0					Uploaded Files 3.7 GB
	Project Name Datasets	Tags Sharing	Size on Disk	Created	Last Updated	<u>Status</u>	15: ftp://ftp.ddbj.nig.ac.jp ● ℓ 🗴 Completed
	Uploaded Tiles	<u>0 Tags</u>	2.4 GB	2 days ago	2 minutes ago	current project	/sRA061/sRA061150/58X229331 /sRR769606 2.fastq /sRR769606 2.fastq
	Unnamed + history	<u>0 Tags</u>	0 bytes	15 minutes ago	15 minutes ago		14: ftp://ftp.ddbj.nig.ac.jp ● ℓ % /ddbj.database/dra/fastq /SRA061/SRA061150/SRX229331
_	Unnamed -	^ - -		2 days	~ .		<u>/SRR769606_1.fastq.bz</u>

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:

写 Galaxy	Launch Pad F oject Viev	w Shared Data - H	elp 👻 User 👻		Using 5%
This dataset is large and only the first megaby Show all Save	e is shown below.			0	Current Project History
					Uploaded Files
<pre>@SRR769606.1 HWI-ST765:7:1101:1527:2028 lengt</pre>	n=101				3.7 GB 🖉 🖻
ATTGGATTGGAGTTTTCGAAGATTGGAGTGGCCTCGAGCCTCAC +SRR769606.1 HHI-ST75517111011527:2028 Lengt a_cceeex ^a aJQlbae ^a eye ^a dXJQXHp_ggf_eHOOU ^a BBBB §SRR769606.2 HHI-ST75517111011533:2056 Lengt CCACCTTGGACAACAGGGACACAGAGAGACATTCATCGACCTGATGT +SRP769606.2 HHI-ST75517110113312056 Lengt	ACACAGGAAAGAAGTATTCGAAGGC h=101 BBBBBBBBBBBBBBBBBBBBBBBBBBBBB n=101 TGTGTGCCTGCCTCCCTGTTAGTTATC h=101	GTATATGGACATTTCGAG BBBBBBBBBBBBBBBBBBBB GTTCCGGTCTTCTTCAGG	STACAAGCTCGA BBBBBBBBBBBB CAATCATCAATT		15: ftp://ftp.ddbj.nig.ac.jp
<pre>rakrosucz.m.s.Job/filibiai/filibia</pre>	lhghghhhiiiiiiiiiiiiigggggg =101 3GCGGTACATCAAGGAACACATGTAT h=101 gfddRXZ^^`bccbcab`abcb]bdd	eeeeeccaccccccdcb wggaacgggaatgcaatg L_bcc^_a_accc``bb`	ccccbccccccd AGCCTGTGGAAG _bcb][GY^bbc		14: ftp://ftp.ddbj.nig.ac.jp ● ℓ ⊗ /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: <u>http://tophat.cbcb.umd.edu/</u>
 - o Cufflinks: <u>http://cufflinks.cbcb.umd.edu/index.html</u>
- 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.

Select Analysis Type © Eukaryotic Single-End Analysis © Prokaryotic Single-End Analysis © Eukaryotic Brierd-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	← Сору	Source Project: Select source Uploaded Files
ftp://ftp.ddbj.mlp.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331 //SR769606_12astq ftp://ftp.ddbj.mlp.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331 /SR769606_1.fastq		☐ tp://ftp.ddbj.nig.ac.jp/dbj_database/dra/fastq/SRA061/SRA061150/SRX229331 /SRR79606_12.fastq htp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331 /SRR769606_1.fastq
ſ	Continue	

- 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.

<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 *Encephalitozoon cuniculi* EC2

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 а reference annotation: Encephalitozoon cuniculi EC2 Select how to use the provided annotation: Assemble Novel + annotated transcripts.

Click on the Run Workflow button.

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 guality 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 Ŧ 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?

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
Use multi-read correct () Run workflow
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.

