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 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.

INA-RO	ocket 🚎	≥⊳ ⊿	× 80 0	100 A	Oharb@pcbi.upenn.edu Log
Galaxy		Launch Pad Project View	Shared Data - How-To Hel	lp≁ User≁	
EMBL-EBI 🌒				Services Research Training Abou	us SRX129648
NAA.FI					
	European Nucle	eotide Archive			
NA Home Search & Brow	vse Submit & Update About EN	VA Contact			
 Please subscribe 	e to ena-announce mailing list	t here:listserver.ebi.ac.uk/mailman/	listin to receive alerts about E	ENA services.	
Text search Advanced s	search Sequence search				
Enter or paste text or ENA a	accession number:	Uplo	ad file of accessions:		
Enter or paste text or ENA a	accession number:	Uplo Search Choos	ad file of accessions: re File no file selected	Search	
Enter or paste text or ENA a	accession number:	Uplo Search Choos	ad file of accessions: ae File no file selected	Search	
Enter or paste text or ENA a	accession number:	Search Choos	ad file of accessions: ^{the} File) no file selected	(Search)	
Enter or paste text or ENA a	accession number:	Uplo Search Choose Research	ad file of accessions: le File no file selected Training	Search	About us
Enter or paste text or ENA a	accession number:	Upio Search Choos Research	ad file of accessions: le File no file selected	Search	About us
Enter or paste text or ENA a	Services By topic	Upio Search Choose Research Overview	ad file of accessions: <u>e File</u> no file selected Training Overview Overview	Search Industry Overview	About us Overview
Enter or paste text or ENA a	Services By topic By name (A-Z) By name (A-Z)	Upio Search Choose Research Overview Publications	ad file of accessions: <u>a File</u> no file selected Training Overview Train at EBI Train at EBI	Search Industry Overview Members Area	About us Overview Leadership
Enter or paste text or ENA a MBL-EBI	Services By topic By name (A-2) Help & Support	Research Overview Publications Research groups Research groups	ad file of accessions: <u>e File</u> no file selected Training Overview Train at EBI Train outside EBI	Search Industry Overview Members Area Workshops	About us Overview Leadership Funding of
Inter or paste text or ENA a MBL-EBI	Services By topic By name (A-2) Help & Support	Upio Search Choos Research Overview Publications Research groups Postdocs & PhDs	ad file of accessions: e file no file selected Training Overview Train at EBI Train outide EBI Train outide EBI Train outide EBI	Search Industry Overview Members Area Workshops SME Forum Coded Industry engagements	About us Overview Leadership Funding Background Collaporation
Enter or paste text or ENA a MBL-EBI ews contact us intranet	Services By topic By name (A-2) Help & Support	Research Overview Publications Research groups Postdocs & PhDs	ad file of accessions: <u>e File</u> no file selected Training Verview Train et Ell Train oxitide EBI Train oxitide EBI Train oxitide	Search Industry Overview Members Area Workshops SME Forum Contact Industry programme	About us Overview Leadership Funding Background Categoration
Enter or paste text or ENA a MBL-EBI Intervs rochures iontact us ntranet	Services By topic By name (A-Z) Help & Support	Verview Publications Research Overview Publications Research groups Postdocs & PhDs	ad file of accessions: a file no file selected Training Overview Train at EBI Train outside EBI Train outside EBI Train outside EBI Train outside EBI Train outside EBI	Search Industry Overview Members Area Workshops SME Forum Contact Industry programme	About us Overview Leadership Funding Background Collaboration Jobs Bacele & conung
Enter or paste text or ENA a EMBL-EBI Irochuras Irochuras Intranet	Services By topic By name (A-2) Help & Support	Vplo Search Choos Postilications Research groups Postdocs & PhDs	ad file of accessions: e_fileno file selected Training Verview Train at EBI Train outside EBI Train outside EBI Train outside CBI Contact organisers	Search Industry Overview Members Area Workshops SME Forum Contact Industry programme	About us Overview Leadership Funding Background Collaboration Jobs People & groups News
Enter or paste text or ENA a EMBL-EBI News Nochures Ontact us ntranet	Services By topic By name (A-2) Help & Support	Choose Research Overview Publications Research groups Postdocs & PhDs	ad file of accessions: a file in o file selected Training Overview Train at EBI Train outside EBI Train outside EBI Train online Contact organisers	Search Industry Overview Members Area Workshops SME Forum Contact Industry programme	About us Overview Leadership Funding Background Collaboration Jobs People & groups News Events
Enter or paste text or ENA a EMBL-EBI () leves invectors instact us intranet	Services By topic By name (A-2) Help & Support	Vplo Search Choos Pesearch Overview Publications Research groups Postdocs & PhDs	ad file of accessions: e File no file selected Training Overview Train outside EBI Train outside EBI Train outside CBI Contact organisers	Search Industry Overview Members Area Workshops SMF Forum Contact Industry programme	About us Overview Leadership Funding Background Collaboration Jobs People & groups News Events Visit us
Enter or pasto text or ENA a 2008L-EBI News rochures ontact us ntranet	Services By topic By name (A-2) Help & Support	Choose Research Overview Publications Research groups Postdocs & PhDs	ad file of accessions: a file in o file selected Training Overview Train at EBI Train outSide EBI Train online Contact organisers	Search Industry Overview Members Area Werkshops SME Forum Contact Industry programme	About us Overview Leadership Funding Background Collaboration Jobs People & groups News Events Visit us Contact us
Enter or paste text or ENA a IMBL-EBI () leves involutes instact us intranet	Services By topic By mane (A-Z) Help & Support	Verview Pesearch Overview Postdocs & PhDs	ad file of accessions: e File no file selected Training Train overview Train at EBI Train online Contact organisers	Search Industry Overview Members Area Workshops SMF Forum Contact Industry programme	About us Overview Leadership Funding Background Collaboration Jobs People & groups News Events Visit us Contact us
Enter or paste text or ENA a BMBL-EBI Vews Srochures Jontact us Intranet MBL-EBI, Wellcome Trust (Services By topic By name (A-Z) Help & Support	Idgeshire, CB10 15D, UK +44 (0)12:	ad file of accessions: <u>ar file</u> no file selected Training Overview Train at EBI Train outside EBI Tra	Search Industry Overview Members Area Workshops SME Forum Contact Industry programme	About us Overview Leadership Funding Background Collaboration Jobs People & groups News Events Events Visit us Contact us

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.

								5	Send Fe	ad: <u>XML</u> edback 🖂				
Submitting Cen GEO	tre	Platform	Mod	tel nina Genome A	Analyzer	Read Count		Base Cou	nt					
ibrary Layout		Ibrary Strategy	Libr TR/	ary Source	IC C	Library Selecti cDNA	on	Library Na GSM8958 GLAS175	ame 15: P33 BI	NAsea				
Broker Name														
Navigation Re	ad Files A	tributes												
lew: <u>TEXT</u>	5								Downlo		г			
Select columns	: 1 - 2 of 1 res	uits							[-			
ielect columns ihowing 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)	Submitted files (ftp)	CoL tax ID	Submitted files (galaxy)	CoL scientifi name
Select columns Showing results Study accession PRJNA153531	Secondary study accession	Sample accession SAMN00811621	Secondary sample accession SRS300498	Experiment accession SRX129648	Run accession SRR44517	Scientific name	Instrument model Illumina Genome Analyzer	Library layout PAIRED	Fast files (ftp) File 1 File 2	Fastq files (galaxy) File 1 File 2	Submitted files (ftp)	CoL tax ID	Submitted files (galaxy)	CoL scientifi name

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).

00	Galaxy	_	Launch Pad	Project View	Shared D	ata - Help -	User 👻	Using 5%	
Pr se	arch project names	and tags	Q					Current Project History 2 * Uploaded Files 2.4 GB 🖉 🖻	
	Project Name Uploaded Files	Datasets	Tags Sharing 0 Tags	Size on Disk 2.4 GB	Created 2 days ago	Last Updated ↑ 2 minutes ago	Status current project	The: The:	ow
	Unnamed history	•	<u>0 Tags</u>	0 bytes	15 minutes ago	15 minutes ago		<u> </u>	
	Unnamed history	•	<u>0 Tags</u>	0 bytes	2 days ago	2 days ago		/ddbj_database/dra/fastq /SRA061/SRA061150/SRX229331 /SRR769606_1.fastq.bz	

💳 Galaxy	Launch Pad	Project View	Shared Da	ata - Help -	User 🗸		Using 5%	
Project List search project names and tags Advanced Search	Q					C U 3.	Jploaded Files	Completed
Project Name Datasets Uploaded Image: Constraint of the second s	Tags Sharing 0 Tags	Size on Disk	Created 2 days ago	Last Updated	Current project	1 // //	5: ftp://ftp.ddbj.nig.ac.jp ● Ø ⊗ ddbj database/dra/fastg SRA061/SRA061150/SRX229331 SRR769606 2.fastg	tasks will
Unnamed +	<u>0 Tags</u>	0 bytes	15 minutes ago	15 minutes ago	1		4: ftp://ftp.ddbj.nig.ac.jp	green
Unnamed -			2 days				SRR769606 1.fastq.bz	

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:

S Galaxy	Launch Rad	iour Charod Data -	Holp = Usor =			Using 5%
	Launch Pad Toject V	iew Silareu Data	neip osei			Using 5%
This dataset is large and only the first megab	yte is shown below.				Current Project Hist	tory 🖸 🔅
Show all Save						
					Uploaded Files	
ASER769606.1 HWT=ST765:7:1101:1527:2028 leng	th=101				3.7 GB	47 📑
ATTGGATTGGAGTTTTCGAAGATTGGAGTGGCCTCGAGCCTCAG	CGACACAGGAAAGAAGTATTCGAAG	GCGTATATGGACATTTCGA	GGTACAAGCTCGA			
+SRR769606.1 HWI-ST765:7:1101:1527:2028 leng	th=101				15: ftp://ftp.ddbj.ni	ig.ac.jp 👁 🖉 💥
a cceeeR`eJQ[bae^eYe^dXJQXHP qqf eHOOU^BBBE	BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB	BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB	BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB		/ddbj database/dra	a/fastq
@SRR769606.2 HWI-ST765:7:1101:1533:2056 leng	th=101				/SRA061/SRA0611	50/SRX229331
CCACCTTGACAACAGGAGACACAGAGAACTTCATCGACCTGATG	TTGTGTGCTGCCTCCCTGTTAGTTA	TCGTTCCGGTCTTCTTCAG	GCAATCATCAATT		/SRR769606 2.fast	đ
+SRR769606.2 HWI-ST765:7:1101:1533:2056 leng	th=101			-		
bbbeeeegggfgiiiiihhhhiffhihiiihhhiiiiihfh	iihghghhhiiiiiiiiiiiigggg	ggeeeeeccaccccccdc	bccccbcccccd		14: ftp://ftp.ddbj.ni	g.ac.jp 👁 🖉 💥
@SRR769606.3 HWI-ST765:7:1101:1845:2018 leng	th=101				/ddbj_database/dra	<u>a/fastq</u>
TGATTGAGAGGTATGTCGGCGAGTCTGTGTTTATGCTTGGGATC	CGGCGGTACATCAAGGAACACATGT	ATGGGAACGGGAATGCAAT	GAGCCTGTGGAAG		/SRA061/SRA0611	50/SRX229331
+SRR769606.3 HWI-ST765:7:1101:1845:2018 leng	th=101				/SRR769606_1.fast	<u>q.bz</u>
Z cceR`eeac^gdefhdaf` eghd^caaegfga^aaeg^ae	bgfddRXZ^``bccbcab`abcb1b	dd bcc^ a accc``bb`	bcb1[GY^bbc			

- 6. Once the RNA-sequence FASTQ file has been uploaded you can start the RNAseq 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 "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.

Select an costing import or create a new Yroject to be used during this analysis and populate the Project: Select an costing import of create a new Yroject to be used during this analysis will be saved in the selected Project. Currently Selected Project: Uploaded Files Import of create a new Yroject is the project is select and copy hilds from this Select and copy hilds or existing project(s) to populate your current Project. Select existing project OR Create project Uploaded Files Import of create a new Yroject is project is pro	Select Analysis Type © Eukayotic Single-End Analysis © Prokaryotic Single-End Analysis © Eukaryotic Faterd-End Analysis © Prokaryotic Patred-End Analysis	
Target Project: Select existing project OR Create project Select source Select source Uploaded Files Image: project Image: project Uploaded Files Image: project Uploaded Files Image: project Select source Select source Select source </td <td>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</td> <td>Select and copy files from Uploads or existing project(s) to populate your current Project.</td>	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.
ftp://ftp.ddbj.mg.ac.jp/ddbj_database/dra/fastq/SRA0611/SRA061150/SRX229331 /SRR769606_2.fastq [SRR769606_1.fastq [SRR769606_1.fastq (SRR769606_1.fastq (SRR769606_1.fastq	Target Project: Select existing project - OR - Create project Uploaded Files :	← Copy Source Project: Select source Uploaded Files ♀
	hp://thu/ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331 SRR76506_1stq tp://thu/ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331 SRR769606_1.fastq	☐ thc://thc.ddbj.njg.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061150/SRX229331 /SRR206060_2.fastq ☐ thc://thc.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.

Step 1: Input dataset Downstream files must be in the same order as their corresponding upstream files	9.4 GB
Upstream Read Files L: EBI SRA: SRX129648 File: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR445/SRR445171/SRR445171_1.fastq.gz type to filter	2: EBI SRA: SRX129648 File: ⊕ 0 🕸 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SR R445/SRR445171/SRR445171 2.fast g.gz
<u>Step 2: Input dataset</u> Downstream files must be in the same order as their corresponding upstream files	1: EBI SRA: SRX129648 File:
Downstream Read Files Image: Comparison of the start of	<u>q.qz</u>

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

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 guality 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 Giardia Assemblage A isolate WB 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 3: Tophat2 (version 2.0.10)

No

Step 4: Cufflinks (version 2.0.2) SAM or BAM file of aligned RNA-Seg reads Output dataset 'accepted_hits' from step 3 Maximum Intron Length (-I) 500 Minimum Isoform Fraction (-F) 0.1 Pre MRNA Fraction (-j) 间 0.15 Overlap Radius 间 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 Giardia Assemblage A isolate WB 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 (1) Run workflow No 🗘 None

Click on the Run Workflow button.

transcripts.

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.

