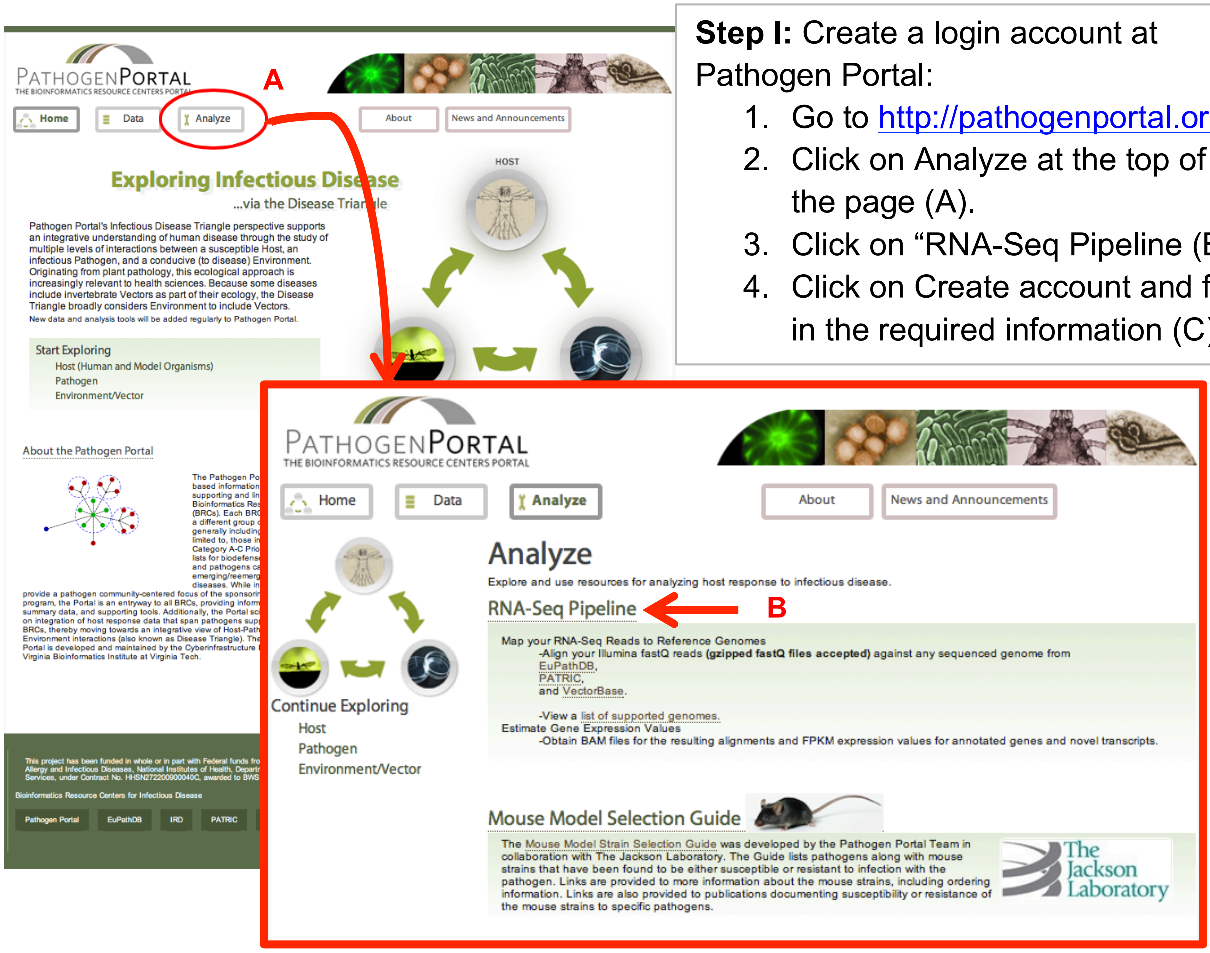
RNA sequence data analysis
(Part 1: using pathogen portal’s RNAseq pipeline)
Exercise 3

Step I: Create a login account at Pathogen Portal:

1. Go to http://pathogenportal.org
2. Click on Analyze at the top of the page (A).
3. Click on “RNA-Seq Pipeline (B).
4. Click on Create account and fill in the required information (C).



Exploring Infectious Disease

...via the Disease Triangle

Pathogen Portal's Infectious Disease Triangle perspective supports an integrative understanding of human disease through the study of multiple levels of interactions between a susceptible Host, an infectious Pathogen, and a conducive (to disease) Environment. Originating from plant pathology, this ecological approach is increasingly relevant to health sciences. Because some diseases include invertebrate Vectors as part of their ecology, the Disease Triangle broadly considers Environment to include Vectors.

New data and analysis tools will be added regularly to Pathogen Portal.

Start Exploring
- Host (Human and Model Organisms)
- Pathogen
- Environment/Vector

About the Pathogen Portal

The Pathogen Po
based information
supporting and lin
Bioinformatics Res
(BRCs). Each BRC
a different group o
generally including
limited to, those in
Category A-C Prio
lists for biodefense
and pathogens ca
emerging/reemerg
diseases. While in
provide a pathogen community-centered focus of the sponsorin
program, the Portal is an entryway to all BRCs, providing inform
summary data, and supporting tools. Additionally, the Portal sc
on integration of host response data that span pathogens sup
BRCs, thereby moving towards an integrative view of Host-Path
Environment interactions (also known as Disease Triangle). The
Portal is developed and maintained by the Cyberinfrastructure
Virginia Bioinformatics Institute at Virginia Tech.

This project has been funded in whole or in part with Federal funds fro
Allergy and Infectious Diseases, National Institutes of Health, Departr
Services, under Contract No. HHSN272200900040C, awarded to BWS

Bioinformatics Resource Centers for Infectious Disease

Pathogen Portal | EuPathDB | IRD | PATRIC

Analyze

Explore and use resources for analyzing host response to infectious disease.

RNA-Seq Pipeline

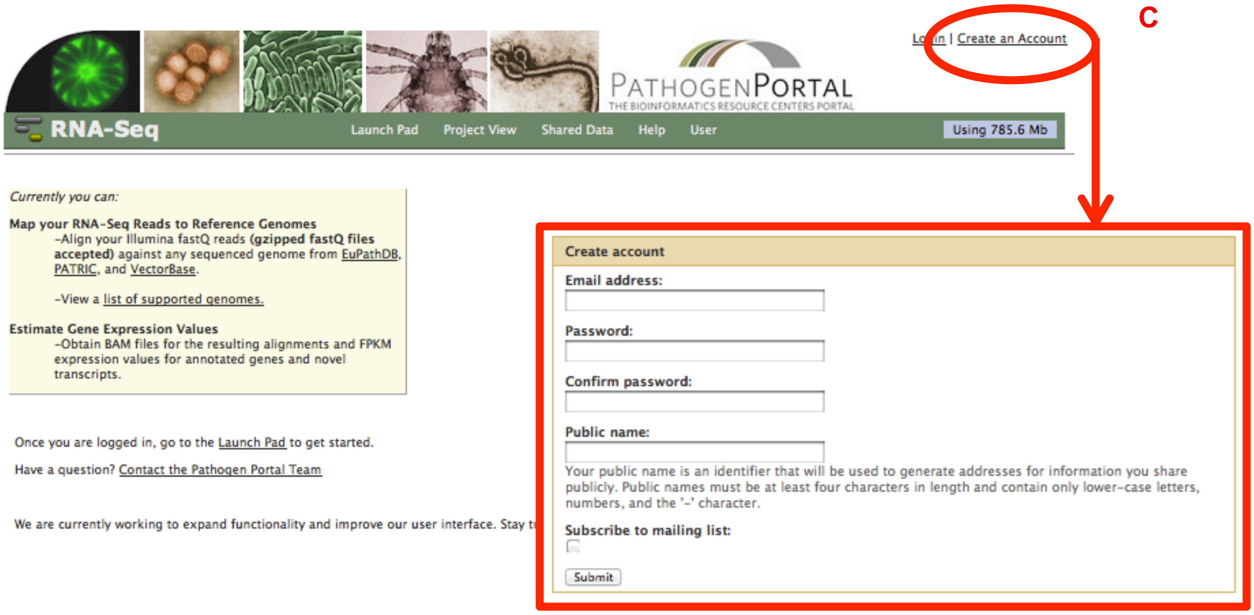
Map your RNA-Seq Reads to Reference Genomes
- -Align your Illumina fastQ reads **(gzipped fastQ files accepted)** against any sequenced genome from EuPathDB, PATRIC, and VectorBase.
- -View a list of supported genomes.

Estimate Gene Expression Values
- -Obtain BAM files for the resulting alignments and FPKM expression values for annotated genes and novel transcripts.

Continue Exploring
- Host
- Pathogen
- Environment/Vector

Mouse Model Selection Guide

The Mouse Model Strain Selection Guide was developed by the Pathogen Portal Team in collaboration with The Jackson Laboratory. The Guide lists pathogens along with mouse strains that have been found to be either susceptible or resistant to infection with the pathogen. Links are provided to more information about the mouse strains, including ordering information. Links are also provided to publications documenting susceptibility or resistance of the mouse strains to specific pathogens.



Currently you can:

Map your RNA-Seq Reads to Reference Genomes
- -Align your Illumina fastQ reads **(gzipped fastQ files accepted)** against any sequenced genome from EuPathDB, PATRIC, and VectorBase.
- -View a list of supported genomes.

Estimate Gene Expression Values
- -Obtain BAM files for the resulting alignments and FPKM expression values for annotated genes and novel transcripts.

Once you are logged in, go to the Launch Pad to get started.

Have a question? Contact the Pathogen Portal Team

We are currently working to expand functionality and improve our user interface. Stay t

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:
□

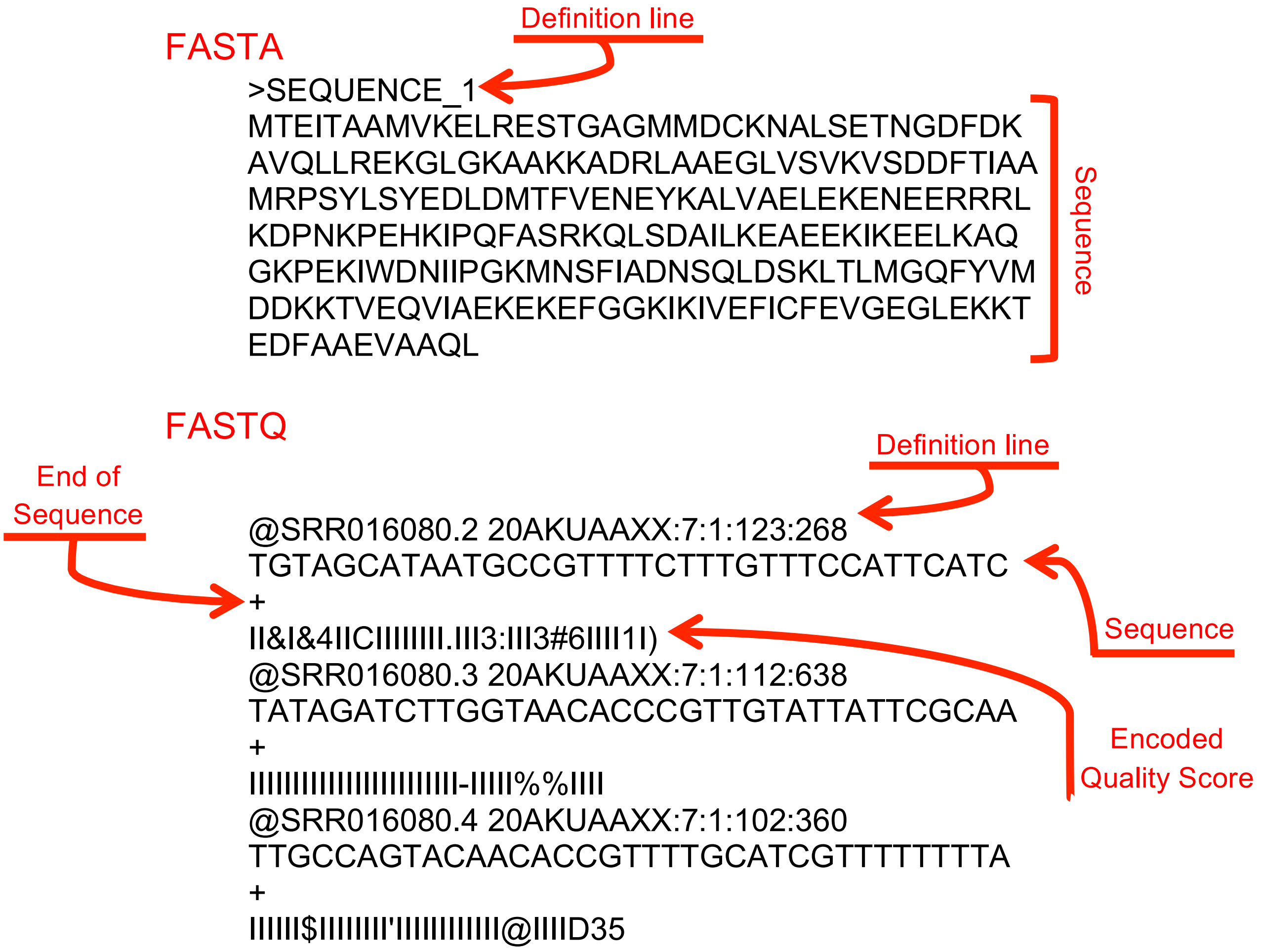
Submit

Step II: Getting data into your launch pad.

For this exercise we will be working with a data set generated from Illumina sequencing of a *Babesia bovis* cDNA library (http://www.ncbi.nlm.nih.gov/pubmed/18987005).

You can read more about the actual sample files here:
http://www.ncbi.nlm.nih.gov/sra/SRX004534

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
TTGCCAGTACAACACCGTTTTGCATCGTTTTTTTA
+
IIIII$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 NCBI:

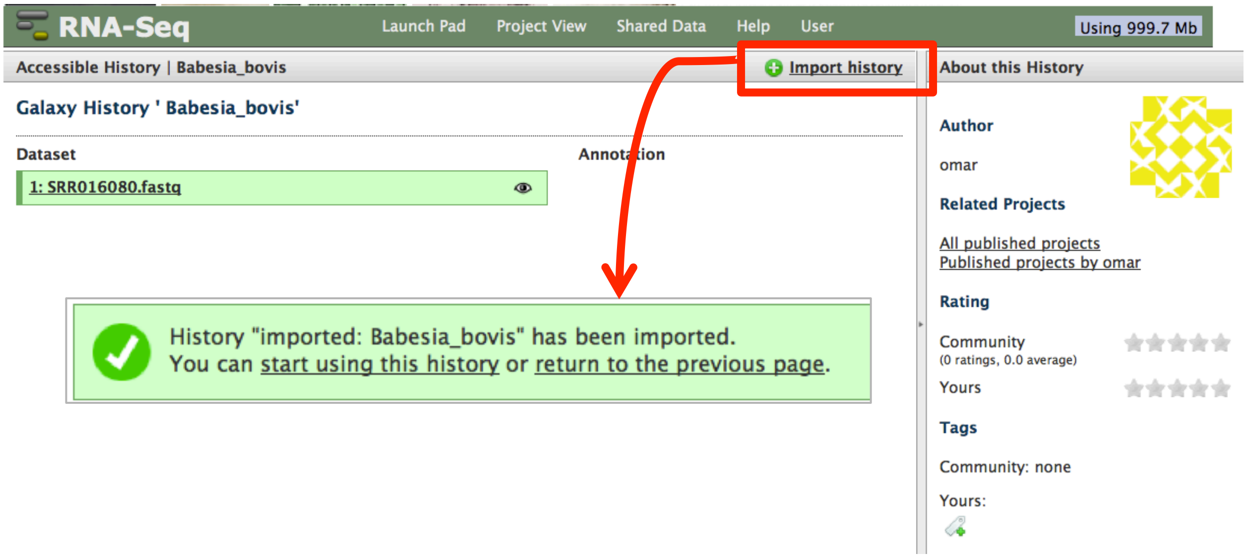
http://www.ncbi.nlm.nih.gov/sra/SRX004534

Here is the record at DDBJ:

http://trace.ddbj.nig.ac.jp/DRASearch/run?acc=SRR016080

- To save download and upload time the file is available as a shared project in PathogenPortal. Go to the following link, then click on "Import History".

http://rnaseq.pathogenportal.org/u/omar/h/babesiabovis



Galaxy History ' Babesia_bovis'

Dataset	Annotation
1: SRR016080.fastq	

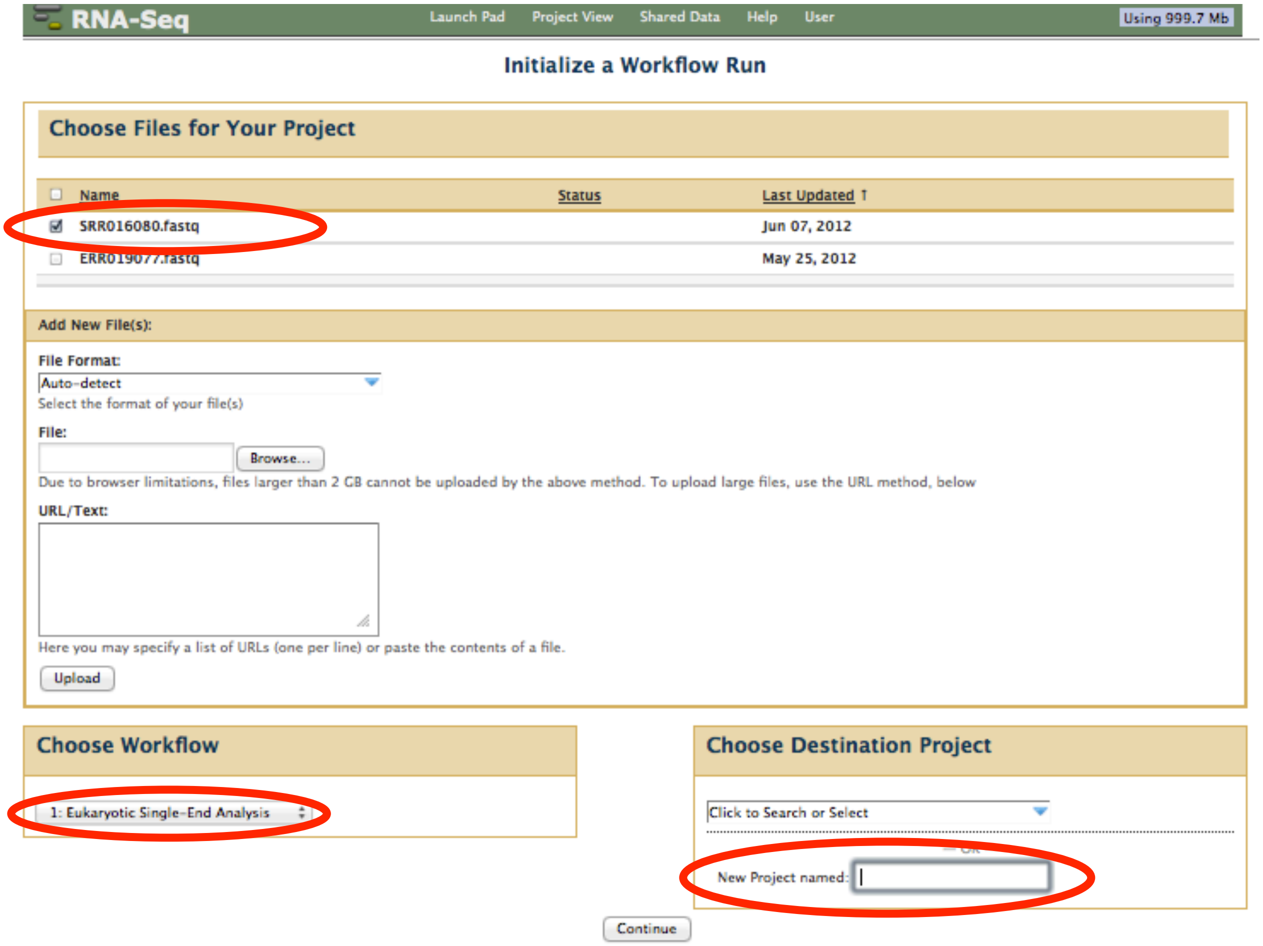
History "imported: Babesia_bovis" has been imported.
You can start using this history or return to the previous page.

Author: omar

Community (0 ratings, 0.0 average)

- Once the RNA-sequence FASTQ file has been imported into your history 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.
    - Select the file you just imported.
    - Choose the workflow – in the case were running a “Eukaryotic Single End Analysis”.
    - Choose a destination project. You can give this a name in the “New Project Named” window.
    - Click on Continue.



Initialize a Workflow Run

Choose Files for Your Project

□ Name	Status	Last Updated ↑
☑ SRR016080.fastq		Jun 07, 2012
□ ERR019077.fastq		May 25, 2012

Add New File(s):

File Format: Auto-detect
Select the format of your file(s)

File:

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:

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

Choose Workflow

1: Eukaryotic Single-End Analysis

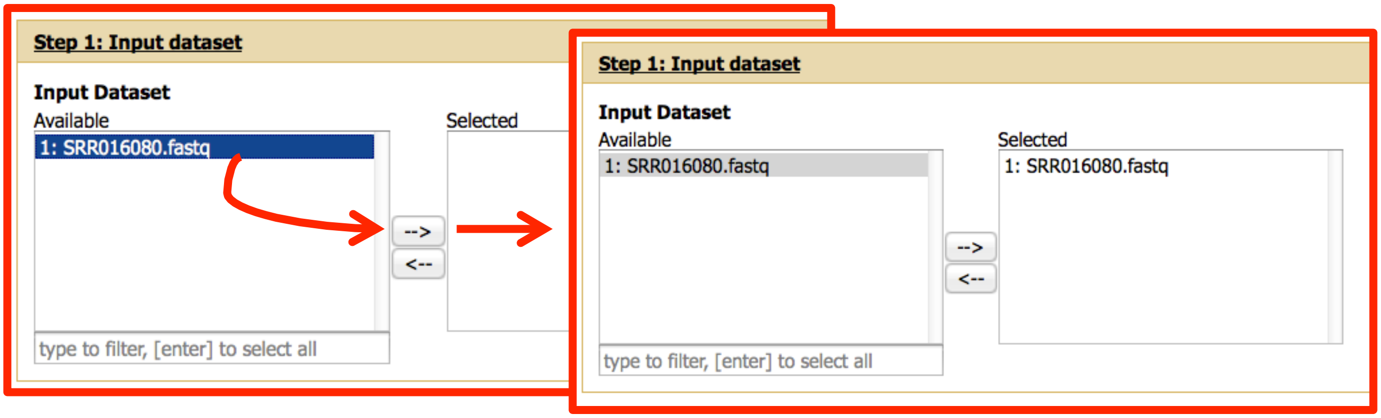
Choose Destination Project

Click to Search or Select

New Project named:

- The next page allows you to configure the pipeline and set both TopHat and Cufflinks parameters:
    - Step1: Select input dataset.
    - Step2: Configure TopHat.
    - Step3: Configure Cufflinks.

Step 1: Select input dataset and click on the arrow to move it from the “Available” window to the “Selected” window:

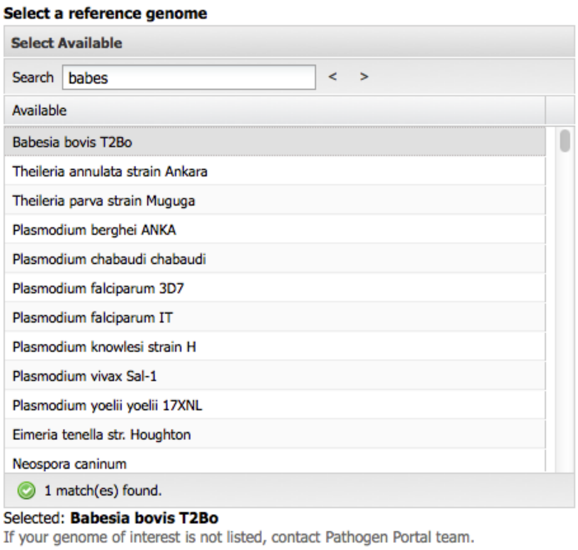


Available: 1: SRR016080.fastq

Selected: 1: SRR016080.fastq

Step 2: Configure TopHat:

- Select the reference organism – in this case *Babesia bovis*. You can start typing the name of the organism in the search window. This will automatically search for the closest match and select it for you.
- We will leave most of the TopHat paramaters at the default values.
- Change the “Maximum intron length” field to 2000.



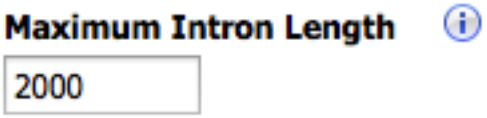
Search: babes

Babesia bovis T2Bo
Theileria annulata strain Ankara
Theileria parva strain Muguga
Plasmodium berghei ANKA
Plasmodium chabaudi chabaudi
Plasmodium falciparum 3D7
Plasmodium falciparum IT
Plasmodium knowlesi strain H
Plasmodium vivax Sal-1
Plasmodium yoelii yoelii 17XNL
Eimeria tenella str. Houghton
Neospora caninum

1 match(es) found.

Selected: **Babesia bovis T2Bo**

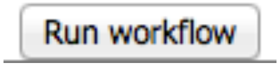
If your genome of interest is not listed, contact Pathogen Portal team.



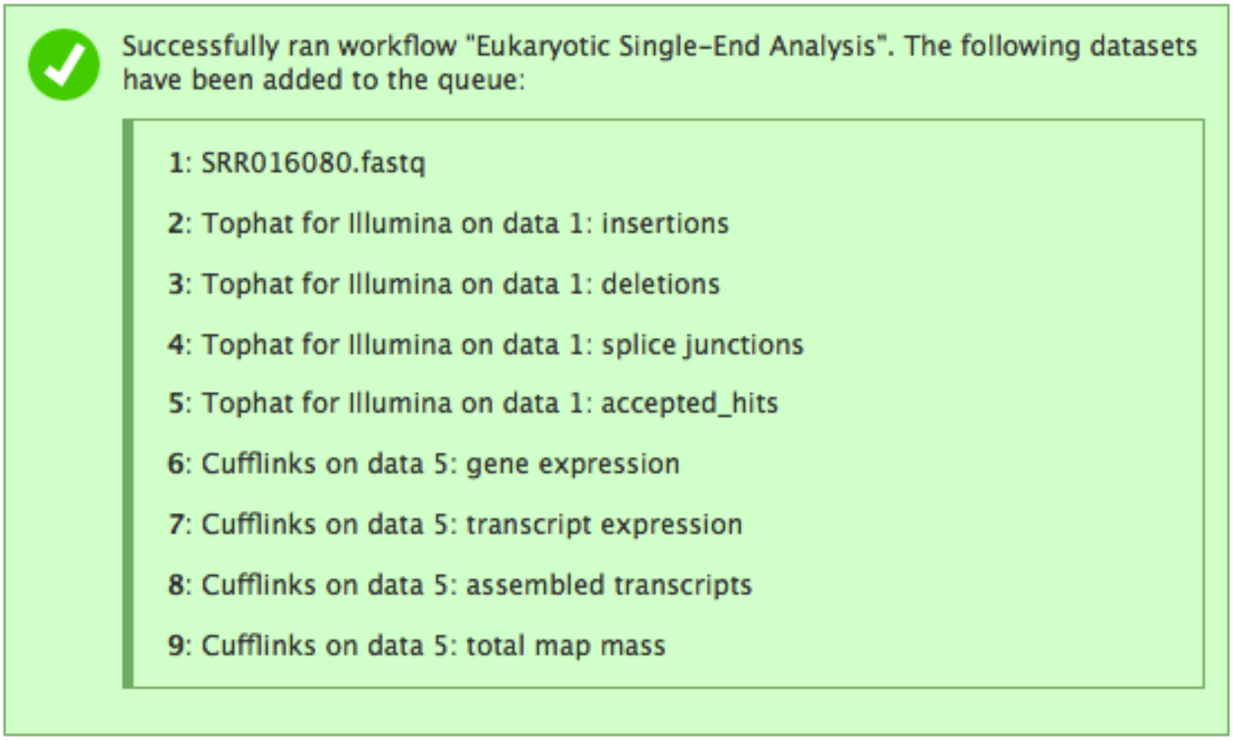
Maximum Intron Length: 2000

Step 3: Configure cufflinks:

- Change the “Maximum intron length” field to 2000.
- Select the reference annotation – in this case *Babesia bovis* (exactly as you did above).

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:



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

1: SRR016080.fastq

2: Tophat for Illumina on data 1: insertions

3: Tophat for Illumina on data 1: deletions

4: Tophat for Illumina on data 1: splice junctions

5: Tophat for Illumina on data 1: accepted_hits

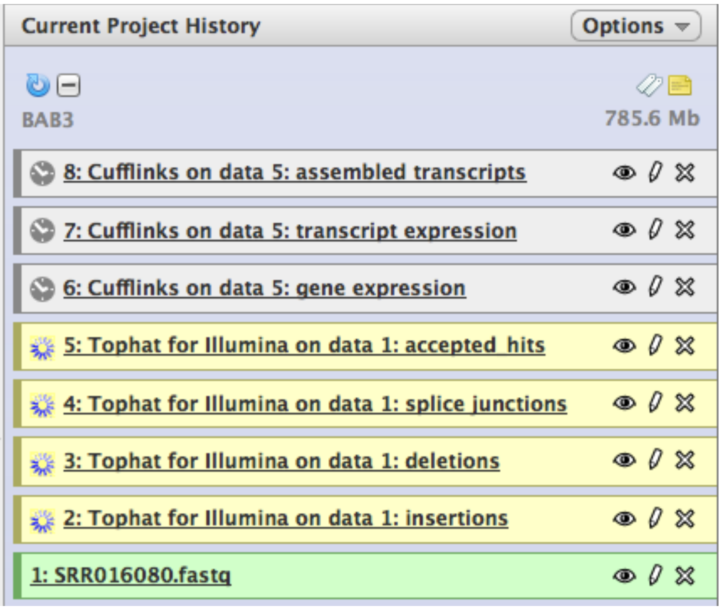
6: Cufflinks on data 5: gene expression

7: Cufflinks on data 5: transcript expression

8: Cufflinks on data 5: assembled transcripts

9: Cufflinks on data 5: total map mass

You can check the progress of your workflow by clicking on the "Project View" link. Completed tasks are in green, running tasks are in yellow and tasks waiting in the queue are in grey:



BAB3 785.6 Mb

8: Cufflinks on data 5: assembled transcripts

7: Cufflinks on data 5: transcript expression

6: Cufflinks on data 5: gene expression

5: Tophat for Illumina on data 1: accepted_hits

4: Tophat for Illumina on data 1: splice junctions

3: Tophat for Illumina on data 1: deletions

2: Tophat for Illumina on data 1: insertions

1: SRR016080.fastq