

RNAseq Mapping 1a

This addendum will allow you to take advantage of the EuPathDB RNAseq export tool which provides a mechanism to query your RNAseq results (FPKM values) using EuPathDB search tools.

To generate FPKM values you need a tool called Cufflinks and read alignment files – BAM files. Our workflow from yesterday generated BAM alignment files from a tool called HISAT2.

Follow these steps to generate FPKM values:

1. Find the tool called Cufflinks by typing the word cufflinks in the tool search box on the left-hand side.
2. Click on the tool to access its parameters.
3. Modify the cufflinks parameters
 - Change the input file to collection and select one of the HISAT2 collections
 - Change the Use Reference Annotation from “No” to “use reference annotation”
 - Select the appropriate reference genome from the drop down list
 - Click on execute.

Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data (Galaxy Version CUFFLINKS: 2.1.1) Versions Options

SAM or BAM file of aligned RNA-Seq reads
78: HISAT2 on collection 55 **a**

Dataset collection this is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Max Intron Length
300000

Min Isoform Fraction
0.1

Pre MRNA Fraction
0.15

Perform quartile normalization
No
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation **b**
Use reference annotation

Will you select an annotation file from your history or use a built-in gff3 file?
Use a built-in annotation

Select a genome annotation **c**
AmoebaDB-29_AastronyxisUnknown_Genome

Perform Bias Correction
No
Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Use multi-read correct
No
Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.

d Execute

Tools

cufflinks

NGS: RNA Analysis

CUFFLINKS PACKAGE

Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data

Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments

Cuffmerge merge together several Cufflinks assemblies

Cuffdiff find significant changes in transcript expression, splicing, and promoter use

CUFFLINKS2 PACKAGE

Cuffquant Precompute gene expression levels

Cuffnorm Create normalized expression levels

StringTie transcript assembly and quantification

FILTERING

Filter Combined Transcripts using tracking file

Ballgown Flexible, isoform-level differential expression analysis

VISUALIZATION

cummeRbund R package designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output

NGS: HOMER

findPeaks performs all of the peak calling and transcript identification analysis

After Cufflinks is done running, the next step is to take the FPKM output files from the collection outputs and put them into a single collection. Notice that cufflinks generates three types of FPKM files (or collections in this case): 1. Gene expression 2. Transcript expression 3. Assembled transcripts. We will only worry about the gene expression files for this section.

- Since we have collections of output files we will need to show all hidden files so we can generate the single collection. To do this, click on the word hidden in the upper right-hand side of the screen
- This will expose all hidden files.
- Click on the check box to perform an operation on multiple datasets (arrow in above image)
- Find all files containing the words “gene expression” and select all the cufflinks files (**NOT** the collections)
- Build a dataset list by clicking on the “for all selected” button and select “Build dataset list”
- Rename each of the datasets in the list and give this collection a meaningful name.

