VEME 2022
NGS Module
RNA-Seq practical session tutorial
Yun Zhang & Richard Scheuermann, J. Craig Venter Institute, USA

Objectives

Upon completion of this exercise, you will:

- be familiarized with the common RNA sequence processing and analysis methods
- be able to perform RNA sequence processing and analysis using the BV-BRC web services (https://www.bv-brc.org/) and interpret the output results

Background

For this exercise, we will use a dataset from Dr. Scheuermann's lab that explores the transcriptional changes that occur in cell lines infected with the picornavirus EV-D68. In this experiment, SH-SY5Y human neuronal cells were infected with enterovirus EV-D68. Samples were collected at 0, 1, 2, 4, 8 hrs post-infection in duplicate. Changes in transcriptional profiles were evaluated by comparing mock-infected versus D68-infected samples at a given timepoint.

In the practical session, we will limit the analysis to 8 fastq files from the paired-end reads (R1 and R2) from two replicate samples of mock-infected and EV-D68 US-MO strain infected SH-SY5Y cells, both at 8 hrs post infection.

Data files

Input: paired-end bulk RNA sequencing reads
https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq

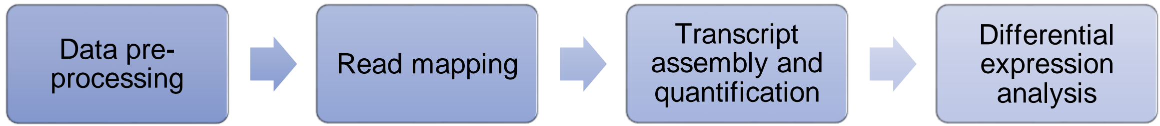
- Mock-infected_8dpi_rep1: D68-Mock_PLT01-A09_S9_R1_001.fastq.gz, D68-Mock_PLT01-A09_S9_R2_001.fastq.gz
- Mock-infected_8dpi_rep2: D68-Mock_PLT01-B09_S21_R1_001.fastq.gz, D68-Mock_PLT01-B09_S21_R2_001.fastq.gz
- MO-infected_8dpi_rep1: D68-MO_PLT01-A11_S11_R1_001.fastq.gz, D68-MO_PLT01-A11_S11_R2_001.fastq.gz
- MO-infected_8dpi_rep2: D68-MO_PLT01-B11_S23_R1_001.fastq.gz, D68-MO_PLT01-B11_S23_R2_001.fastq.gz

Requirements

Output:

- Quantify transcript levels in each sample
- Identify differentially expressed genes between mock and EV-D68 US-MO infected samples at 8 dpi

Workflow



Suggested steps

1. Data pre-processing

1.1 Examine the file format and contents

Data files: https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq

- Mock-infected_8dpi_rep1: D68-Mock_PLT01-A09_S9_R1_001.fastq.gz, D68-Mock_PLT01-A09_S9_R2_001.fastq.gz
- Mock-infected_8dpi_rep2: D68-Mock_PLT01-B09_S21_R1_001.fastq.gz, D68-Mock_PLT01-B09_S21_R2_001.fastq.gz
- MO-infected_8dpi_rep1: D68-MO_PLT01-A11_S11_R1_001.fastq.gz, D68-MO_PLT01-A11_S11_R2_001.fastq.gz
- MO-infected_8dpi_rep2: D68-MO_PLT01-B11_S23_R1_001.fastq.gz, D68-MO_PLT01-B11_S23_R2_001.fastq.gz

1.1.1 Choose an appropriate tool:

Do you want to open the file directly?

1.1.2 Review the results:

1.1.2.1 What is the basic structure of the fastq format? How does it differ from the fasta format?

1.1.2.2 What is the length for the first sequence read in the fastq file?

1.2 Evaluate and improve the quality of the sequence reads

1.2.1 Choose appropriate tools:

1.2.2 Create a workflow:

1.2.3 Review the trimming report:

1.2.4 Review the pre- and post-trimming FastQC reports, and record the results in the table below:

Modules	Pre-QC	Post-QC
Per base sequence quality		
Per tile sequence quality		
Per sequence quality scores		
Per base sequence content		

Per sequence GC content		
Per base N content		
Sequence Length Distribution		
Sequence Duplication Levels		
Overrepresented sequences		
Adapter Content		

2. RNA-Seq analysis

2.1 Choose appropriate tools:

2.2 Create a workflow:

2.3 Review the results:

2.3.1 Examine pipeline log file

2.3.2 Validate the read mapping result by examining the mapping quality stats

2.3.3 Validate the read mapping result by visualizing the read alignment in IGV
How do the coverage plots reflect the intron-exon structures of genes? Is there evidence of alternative splicing?

2.3.4 Examine transcript quantity and differential expression reports

Step by step instructions

1. Data pre-processing

1.1 Examine the file format and contents

Data files: https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq

- Mock-infected_8dpi_rep1: D68-Mock_PLT01-A09_S9_R1_001.fastq.gz, D68-Mock_PLT01-A09_S9_R2_001.fastq.gz
- Mock-infected_8dpi_rep2: D68-Mock_PLT01-B09_S21_R1_001.fastq.gz, D68-Mock_PLT01-B09_S21_R2_001.fastq.gz
- MO-infected_8dpi_rep1: D68-MO_PLT01-A11_S11_R1_001.fastq.gz, D68-MO_PLT01-A11_S11_R2_001.fastq.gz
- MO-infected_8dpi_rep2: D68-MO_PLT01-B11_S23_R1_001.fastq.gz, D68-MO_PLT01-B11_S23_R2_001.fastq.gz

1.1.1 Choose an appropriate tool:

Do you want to open the file directly?

If the file size is much smaller than the computer RAM size, you can open it in a text editor. Otherwise, use a program to view a few lines of the file.

Preview a large file without loading all of it into memory:

```
$ gunzip -c file.gz | head
$ zmore file.gz
```

1.1.2 Review the results:

1.1.2.1 What is the basic structure of the fastq format? How does it differ from the fasta format?

four lines per sequence:

- Line 1 begins with '@' and is followed by a sequence identifier and an optional description.
- Line 2 is the sequence read.
- Line 3 begins with '+' and is optionally followed by the same sequence identifier.
- Line 4 encodes (Phred+33) the quality values for the sequence in Line 2.

1.1.2.2 What is the length for the first sequence read in the fastq file?

Count the number of characters in line 2 or line 4: 251.

1.2 Evaluate and improve the quality of the sequence reads

1.2.1 Choose appropriate tools:

FastQC: A very popular tool used to provide an overview of basic quality control metrics for raw next generation sequencing data.

Trim Galore: a wrapper tool around cutadapt to apply quality and adapter trimming to FastQ files.

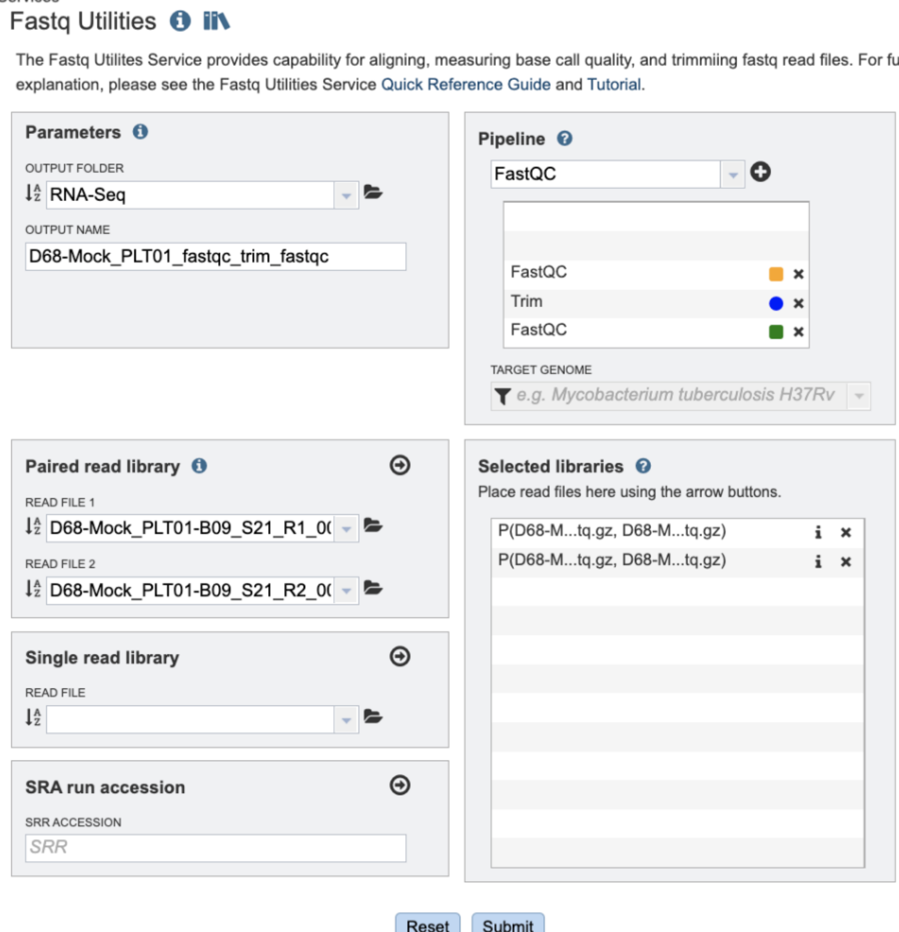
1.2.2 Create a workflow:

BV-BRC > Tools & Services > Fastq utilities > FastQC, Trim, FastQC

tutorial: https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html

input parameters for this exercise:

- Paired read library: read file 1 - D68-Mock_PLT01-A09_S9_R1_001.fastq.gz, read file 2 - D68-Mock_PLT01-A09_S9_R2_001.fastq.gz > click arrow to add the files to the Selected libraries section. Repeat the process for all other read files.
- Pipeline: fastqc > trim > fastqc



1.2.3 Review the trimming report: https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq/.D68-MO_PLT01_fastqc_trim_fastqc/D68-MO_PLT01-A11_S11_R1_001.fastq.gz_trimming_report.txt

```
Input filename: /tmp/stage/D68-MO_PLT01-A11_S11_R1_001.fastq.gz
Trimming mode: paired-end
Trim Galore version: 0.6.5dev
Cutadapt version: 2.2
Python version: 3.7.10
Number of cores used for trimming: 8
Quality Phred score cutoff: 20
Quality encoding type selected: ASCII+33
Using Nextera adapter for trimming (count: 122532). Second best hit was smallRNA (count:
1)
Adapter sequence: 'CTGTCTCTTATA' (Nextera Transposase sequence; auto-detected)
Maximum trimming error rate: 0.1 (default)
Minimum required adapter overlap (stringency): 1 bp
Minimum required sequence length for both reads before a sequence pair gets removed: 20
bp
Output file will be GZIP compressed

This is cutadapt 2.2 with Python 3.7.10
Command line parameters: -j 8 -e 0.1 -q 20 -O 1 -a CTGTCTCTTATA /tmp/stage/D68-MO_PLT01-
A11_S11_R1_001.fastq.gz
Processing reads on 8 cores in single-end mode ...
```

```
Finished in 46.03 s (61 us/read; 0.98 M reads/minute).

=== Summary ===
Total reads processed:                 752,296
Reads with adapters:                   368,343 (49.0%)
Reads written (passing filters):       752,296 (100.0%)

Total basepairs processed:   113,596,696 bp
Quality-trimmed:               1,396,683 bp (1.2%)
Total written (filtered):    106,576,682 bp (93.8%)
```

1.2.4 Review the pre- and post-trimming FastQC reports, and record the results in the table below:

Pre-trimming:
https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq/.D68-MO_PLT01_fastqc_trim_fastqc/D68-MO_PLT01-A11_S11_R1_001_fastqc.html

Post-trimming:
https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq/.D68-MO_PLT01_fastqc_trim_fastqc/D68-MO_PLT01-A11_S11_R1_001_ptrim_fastqc.html

Modules	Pre-QC	Post-QC
Per base sequence quality	Steadily drops over the length of reads	Improved
Per tile sequence quality	Okay	Okay
Per sequence quality scores	Okay	Okay
Per base sequence content*	Non-uniform distribution in the beginning of reads	Unchanged
Per sequence GC content	Okay	Okay
Per base N content	Okay	Okay
Sequence Length Distribution	A single peak	A tailed distribution
Sequence Duplication Levels**	A small peak	Unchanged
Overrepresented sequences	None	None
Adapter Content	Nextera	None

* With most RNA-Seq library preparation protocols, there is clear non-uniform distribution of bases for the first 10-15 nucleotides; this is normal and expected depending on the type of library kit used (e.g. TruSeq RNA Library Preparation).
** RNA-seq data often have high-read coverage for highly expressed transcripts.

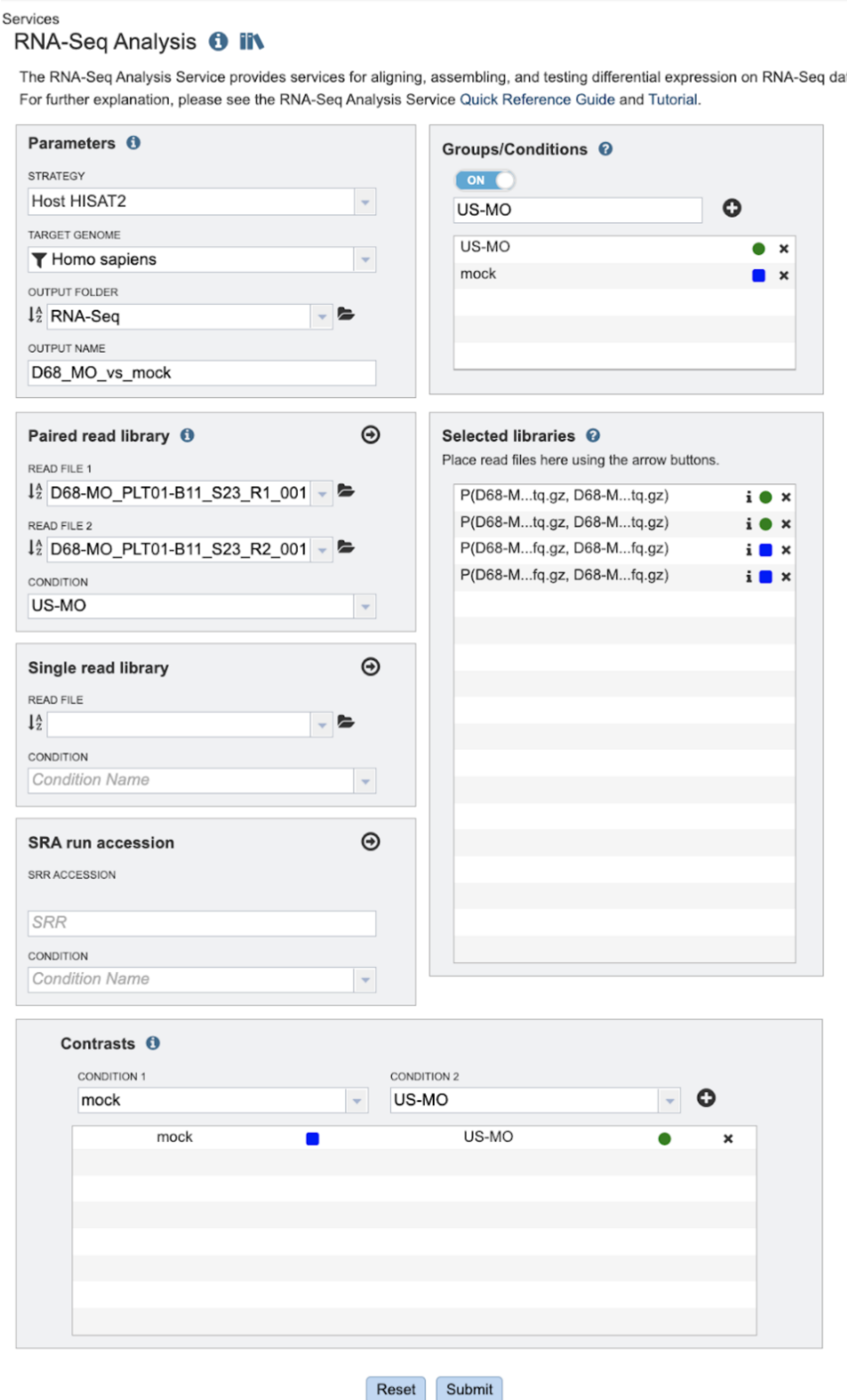
2. RNA-Seq analysis

2.1 Choose appropriate tools:

HISAT2: a read alignment program that supports spliced alignment

HTSeq Count: counts the number of reads mapped to each genomic feature
TPMCalculator: calculates the TPM values for transcripts
DESeq2: differential expression analysis

2.2 Create a workflow:



Data files: sequence reads after quality improvement from step 1.2.
BV-BRC > Tools & Services > RNA-Seq Analysis
tutorial: https://www.bv-brc.org/docs/tutorial/rna_seq/rna_seq.html
input parameters for this exercise:

- strategy: Host HISAT2
- target genome: Homo sapiens[9066.33]
- groups/conditions: mock, US-MO
- paired read library:
  - read file 1: D68-Mock_PLT01-A09_S9_R1_001_ptrim.fq.gz
  - read file 2: D68-Mock_PLT01-A09_S9_R2_001_ptrim.fq.gz
  - condition: mock
  - click arrow to add the files to the Selected libraries section. Repeat the process for all other read files. Make sure to tag mock to mock read files and US-MO to D68 infected read files.
- contrasts: condition 1 - mock, condition 2 - US-MO

2.3 Review the results:
https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq/.D68-MO_vs_mock/9606.33

2.3.1 Pipeline.txt
This log file shows that the RNA-Seq pipeline starts with quality control using fastqc, infers the library preparation type (https://salmon.readthedocs.io/en/latest/library_type.html#fraglibtype) by samples reads using seqtk followed by inference using infer_experiment, then aligns reads to reference genome using Hisat2, assigns reads and quantifies transcript and gene counts using both HTSeq-count and TMPCalculator, and lastly performs differential expression analysis using DEseq2.

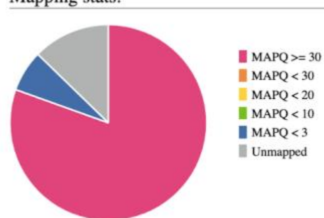
2.3.2 Validate the read mapping result by examining the mapping quality stats: https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq/.D68-MO_vs_mock/9606.33/US-MO/replicate1/D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim.bam.samstat.html

html file: D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim.bam.samstat.html

D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim.bam

1500472 reads, size:140912848 bytes, created 2022-08-12 21:39:27

Mapping stats:



Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences.

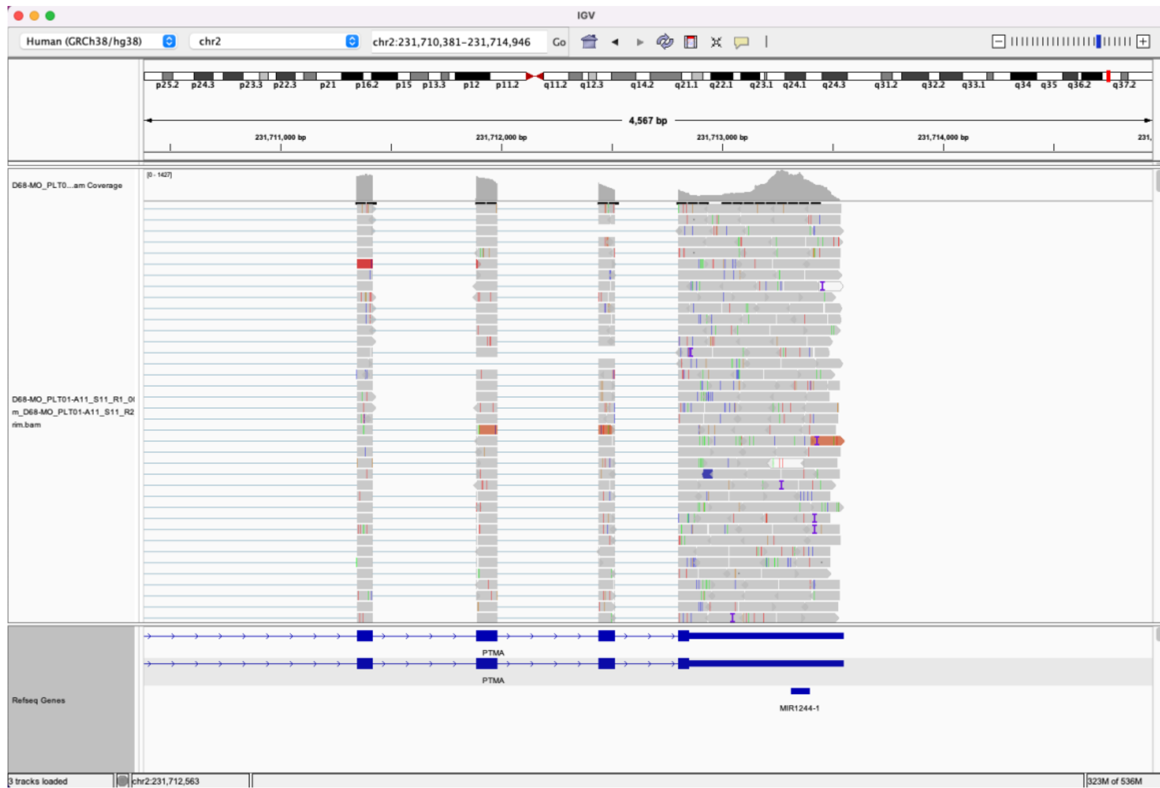
	Number	Percentage
MAPQ >= 30	1207760.0	80.5
MAPQ < 30	0.0	0.0
MAPQ < 20	0.0	0.0
MAPQ < 10	0.0	0.0
MAPQ < 3	102151.0	6.8
Unmapped	190561.0	12.7
Total	1500472.0	100.0

Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences.

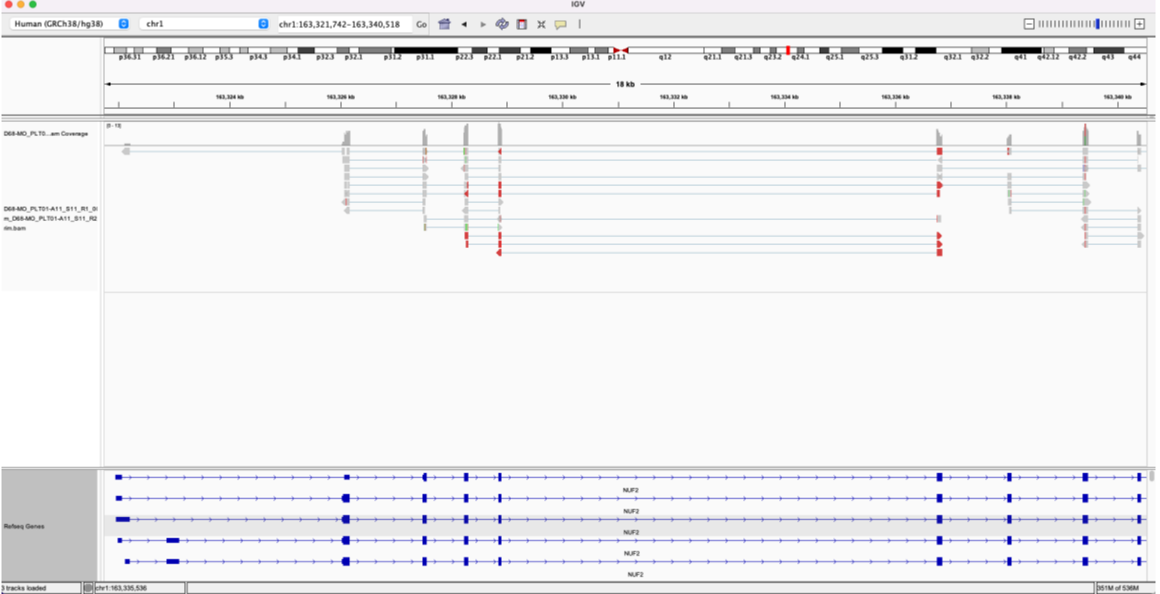
2.3.3 Validate the read mapping result by visualizing the read alignment in IGV

Download the alignment files for the US-MO infected replicate 1 (https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq/.D68-MO_vs_mock/9606.33/US-MO/replicate1: D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim.bam, D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim.bam.bai) to your computer > Open IGV on your computer > Genomes > Select Hosted Genomes > Human (hg38) > File > Load from file: D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim.bam, D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim.bam.bai needs to be in the same directory

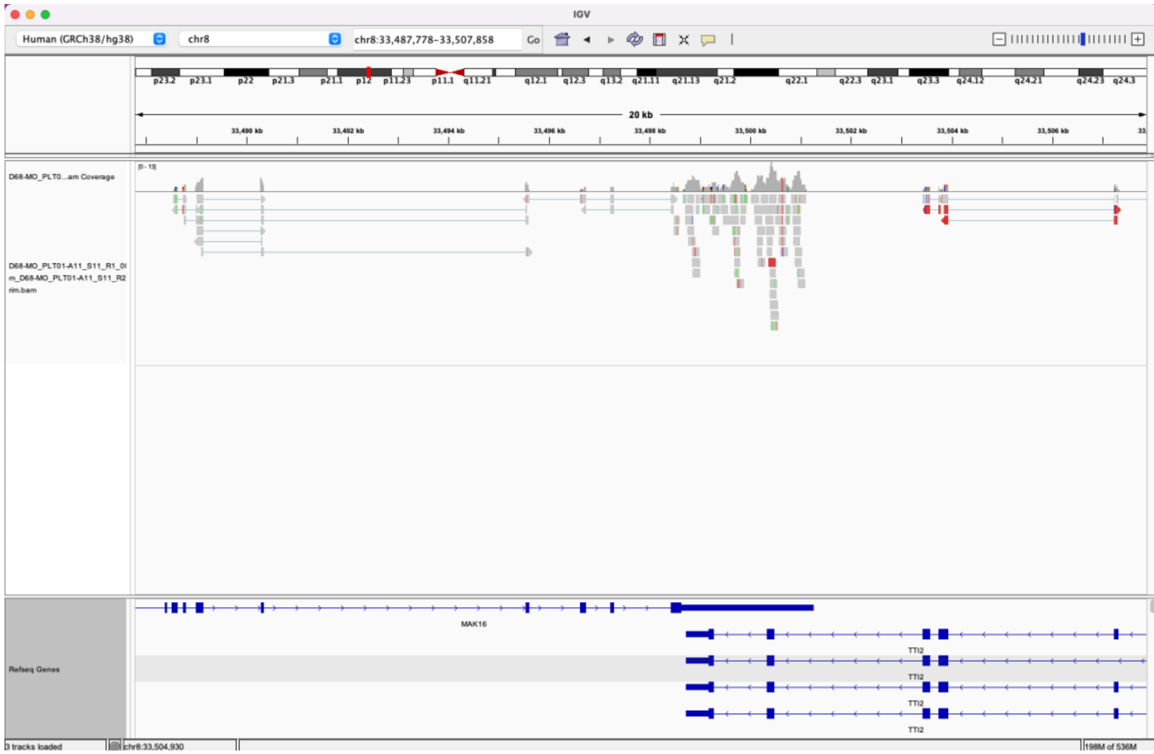
- Since the genome is very large, it's hard to navigate through the alignment by scrolling. You can search gene name by typing in a name in the coordinate text box, e.g., "MIR1244-1", "NUF2", "RGS5", etc.
- Read coverage panel and read alignment panel: You may need to zoom in to see the coverage plot and reads.
- Refseq Genes panel: Right-click the Refseq Genes track name and choose "Expanded" mode.
- Reads panel: clicking in this panel will pop up a window displaying read information
- How do the coverage plots reflect the intron-exon structures of genes? Is there evidence of alternative splicing?



MIR1244-1 and its surround region. MIR1244-1 is a microRNA (miRNA). miRNAs are short (20-24 nt) non-coding RNAs that are involved in post-transcriptional regulation of gene expression in multicellular organisms by affecting both the stability and translation of mRNAs.



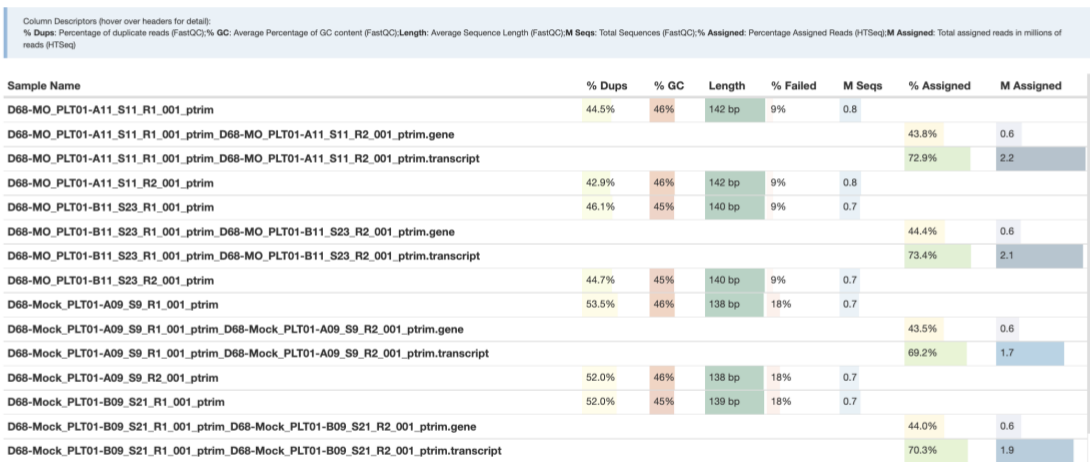
NUF2 and its surrounding region



MAK16 and its surround region

2.3.4 Examine transcript quantity and differential expression report

- View consolidated report 9606.33_report.html
    - General Statistics:



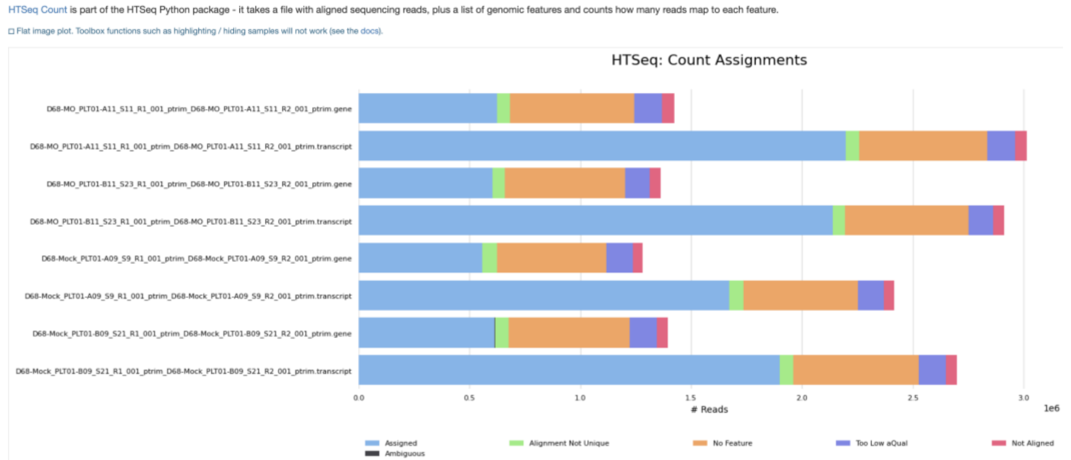
General Statistics

Column Descriptors (hover over headers for detail):
% Dups: Percentage of duplicate reads (FastQC);**% GC**: Average Percentage of GC content (FastQC);**Length**: Average Sequence Length (FastQC);**M Seqs**: Total Sequences (FastQC);**% Assigned**: Percentage Assigned Reads (HTSeq);**M Assigned**: Total assigned reads in millions of reads (HTSeq)

Sample Name	% Dups	% GC	Length	% Failed	M Seqs	% Assigned	M Assigned
D68-MO_PLT01-A11_S11_R1_001_ptrim	44.5%	46%	142 bp	9%	0.8		
D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim.gene						43.8%	0.6
D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim.transcript						72.9%	2.2
D68-MO_PLT01-A11_S11_R2_001_ptrim	42.9%	46%	142 bp	9%	0.8		
D68-MO_PLT01-B11_S23_R1_001_ptrim	46.1%	45%	140 bp	9%	0.7		
D68-MO_PLT01-B11_S23_R1_001_ptrim_D68-MO_PLT01-B11_S23_R2_001_ptrim.gene						44.4%	0.6
D68-MO_PLT01-B11_S23_R1_001_ptrim_D68-MO_PLT01-B11_S23_R2_001_ptrim.transcript						73.4%	2.1
D68-MO_PLT01-B11_S23_R2_001_ptrim	44.7%	45%	140 bp	9%	0.7		
D68-Mock_PLT01-A09_S9_R1_001_ptrim	53.5%	46%	138 bp	18%	0.7		
D68-Mock_PLT01-A09_S9_R1_001_ptrim_D68-Mock_PLT01-A09_S9_R2_001_ptrim.gene						43.5%	0.6
D68-Mock_PLT01-A09_S9_R1_001_ptrim_D68-Mock_PLT01-A09_S9_R2_001_ptrim.transcript						69.2%	1.7
D68-Mock_PLT01-A09_S9_R2_001_ptrim	52.0%	46%	138 bp	18%	0.7		
D68-Mock_PLT01-B09_S21_R1_001_ptrim	52.0%	45%	139 bp	18%	0.7		
D68-Mock_PLT01-B09_S21_R1_001_ptrim_D68-Mock_PLT01-B09_S21_R2_001_ptrim.gene						44.0%	0.6
D68-Mock_PLT01-B09_S21_R1_001_ptrim_D68-Mock_PLT01-B09_S21_R2_001_ptrim.transcript						70.3%	1.9

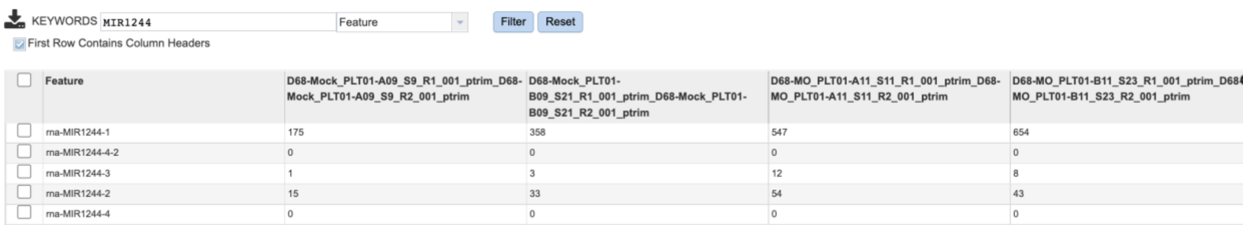
- consolidates statistics from FastQC and HTSeq
- High percentage of duplicate reads reported by FastQC which is as expected

- FastQC:
  - consolidated reports for all samples
- RSeQC:
  - Infer the RNA-Seq library type by sampling reads and then running statistics to infer strandness
- HTSeq-count:



  - statistics on assigned reads

- View transcript quantity reports by HTSeq-count
  raw counts:
  https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq/.D68-MO_vs_mock/9606.33/9606.33.htseq.transcript_counts.tsv
  filter Gene by MIR1244



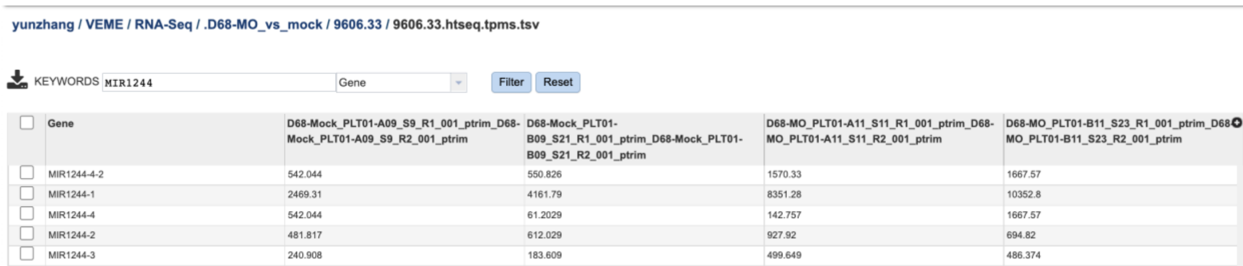
yunzhang / VEME / RNA-Seq / .D68-MO_vs_mock / 9606.33 / 9606.33.htseq.transcript_counts.tsv

KEYWORDS MIR1244 Feature Filter Reset

First Row Contains Column Headers

Feature	D68-Mock_PLT01-A09_S9_R1_001_ptrim_D68-Mock_PLT01-A09_S9_R2_001_ptrim	D68-Mock_PLT01-B09_S21_R1_001_ptrim_D68-Mock_PLT01-B09_S21_R2_001_ptrim	D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim	D68-MO_PLT01-B11_S23_R1_001_ptrim_D68-MO_PLT01-B11_S23_R2_001_ptrim
rna-MIR1244-1	175	358	547	654
rna-MIR1244-4-2	0	0	0	0
rna-MIR1244-3	1	3	12	8
rna-MIR1244-2	15	33	54	43
rna-MIR1244-4	0	0	0	0

normalized counts:
https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq/.D68-MO_vs_mock/9606.33/9606.33.htseq.tpms.tsv
filter feature by MIR1244



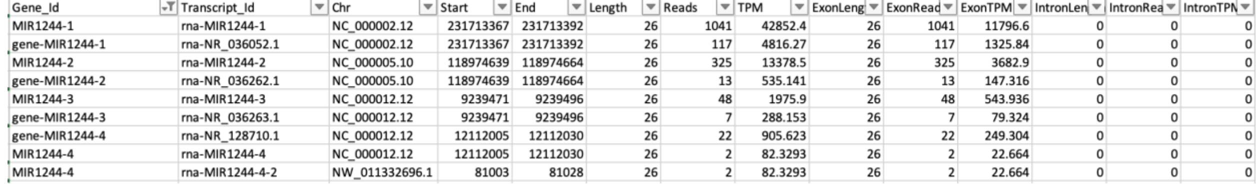
yunzhang / VEME / RNA-Seq / .D68-MO_vs_mock / 9606.33 / 9606.33.htseq.tpms.tsv

KEYWORDS MIR1244 Gene Filter Reset

Gene	D68-Mock_PLT01-A09_S9_R1_001_ptrim_D68-Mock_PLT01-A09_S9_R2_001_ptrim	D68-Mock_PLT01-B09_S21_R1_001_ptrim_D68-Mock_PLT01-B09_S21_R2_001_ptrim	D68-MO_PLT01-A11_S11_R1_001_ptrim_D68-MO_PLT01-A11_S11_R2_001_ptrim	D68-MO_PLT01-B11_S23_R1_001_ptrim_D68-MO_PLT01-B11_S23_R2_001_ptrim
MIR1244-4-2	542.044	550.826	1570.33	1667.57
MIR1244-1	2469.31	4161.79	8351.28	10352.8
MIR1244-4	542.044	61.2029	142.757	1667.57
MIR1244-2	481.817	612.029	927.92	694.82
MIR1244-3	240.908	183.609	499.649	486.374

- View transcript quantity reports by TPMCalculator
  https://www.bv-brc.org/workspace/yunzhang@patricbrc.org/VEME/RNA-Seq/.D68-MO_vs_mock/9606.33/TPMCalculator_Output
  Download all files to your computer.

transcripts.out



Gene_Id	Transcript_Id	Chr	Start	End	Length	Reads	TPM	ExonLeng	ExonRead	ExonTPM	IntronLen	IntronRea	IntronTPN
MIR1244-1	rna-MIR1244-1	NC_000002.12	231713367	231713392	26	1041	42852.4	26	1041	11796.6	0	0	0
gene-MIR1244-1	rna-NR_036052.1	NC_000002.12	231713367	231713392	26	117	4816.27	26	117	1325.84	0	0	0
MIR1244-2	rna-MIR1244-2	NC_000005.10	118974639	118974664	26	325	13378.5	26	325	3682.9	0	0	0
gene-MIR1244-2	rna-NR_036262.1	NC_000005.10	118974639	118974664	26	13	535.141	26	13	147.316	0	0	0
MIR1244-3	rna-MIR1244-3	NC_000012.12	9239471	9239496	26	48	1975.9	26	48	543.936	0	0	0
gene-MIR1244-3	rna-NR_036263.1	NC_000012.12	9239471	9239496	26	7	288.153	26	7	79.324	0	0	0
gene-MIR1244-4	rna-NR_128710.1	NC_000012.12	12112005	12112030	26	22	905.623	26	22	249.304	0	0	0
MIR1244-4	rna-MIR1244-4	NC_000012.12	12112005	12112030	26	2	82.3293	26	2	22.664	0	0	0
MIR1244-4	rna-MIR1244-4-2	NW_011332696.1	81003	81028	26	2	82.3293	26	2	22.664	0	0	0

transcripts.ent

Gene_Id	Transcript_Id	Chr	Type	Type_Nur	start	end	Length	Reads	TPM
MIR1244-1	rna-MIR1244-1	NC_000002.12	exon	1	231713367	231713392	26	1041	814.386
gene-MIR1244-1	rna-NR_036052.1	NC_000002.12	exon	1	231713367	231713392	26	117	91.5305
MIR1244-2	rna-MIR1244-2	NC_000005.10	exon	1	118974639	118974664	26	325	254.251
gene-MIR1244-2	rna-NR_036262.1	NC_000005.10	exon	1	118974639	118974664	26	13	10.1701