Erica Chio, Assigned coursework #6

Due date: April 9th (two weeks), 11pm EST
Submission: PDF document by email to daniel.depledge@nyulangone.org

The objective of this assignment is to align m^{6}A data against the human genome and determine which genes produce m^{6}A-modified mRNAs in test vs. control conditions

exomePeak - http://bioconductor.org/packages/release/bioc/html/exomePeak.html
m6A-viewer - http://dna2.leeds.ac.uk/m6a/

Name	Condition
SRR7992458	iCTRL1
SRR7992461	iCTRL2
SRR7992460	iCTRL3
SRR7992450	iDS1
SRR7992457	iDS2
SRR7992456	iDS3
SRR7992455	mCTRL1
SRR7992454	mCTRL2
SRR7992459	mCTRL3
SRR7992453	mDS1
SRR7992452	mDS2
SRR7992451	mDS3

abbreviation	meaning
i	input RNA
m	m^{6}A enriched RNA
CTRL	control dataset
DS	test dataset
1-3	biological replicates

1. Download data from the SRA, trim, and align (bowtie2) to HG38 genome

```
#!/bin/bash
#SBATCH --job-name=<filename> # Job name
#SBATCH --mail-type=END,FAIL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=Erica.Chio@nyulangone.org # Where to send mail
#SBATCH --ntasks=8 # Run on a single CPU
#SBATCH --mem=16gb # Job memory request
#SBATCH --time=24:00:00 # Time limit hrs:min:sec
#SBATCH --output=<filename>_%j.log # Standard output and error log
```

```
#SBATCH -p cpu_short # Specifies location to submit job
```

```

module purge

#sra
module load sratoolkit/2.9.1
module load fastqc/0.11.7

#trim
module load trimgalore/0.5.0
module load python/cpu/2.7.15-ES

#bowtie
module load bowtie2/2.3.5.1

#samtools
module load samtools/1.3


echo ${SLURM_ARRAY_TASK_ID}

#download sra
fastq-dump SRR79924${SLURM_ARRAY_TASK_ID} --gzip -O /gpfs/scratch/ebc308/AIS/coursework6data/SRR79924${SLURM_ARRAY_TASK_ID} --origfmt

rm -r ~/ncbi

#trim
# !single end reads!
# --fastqc (run FastQC)

trim_galore -o /gpfs/scratch/ebc308/AIS/coursework6data/SRR79924${SLURM_ARRAY_TASK_ID}/ --fastqc /gpfs/scratch/ebc308/AIS/coursework6data/SRR79924${SLURM_ARRAY_TASK_ID}/SRR79924${SLURM_ARRAY_TASK_ID}.fastq.gz

#bowtie
# already a bowtie index in Home_sapiens file.
# -q (reads are in fastq file)
# -x (base name of reference genome. Home_sapies has Bowtie2Index genome)
# --end-to-end  (--very-sensitive to be more stringent - better peaks?)

bowtie2 --threads 8 --end-to-end --very-sensitive -q -x /gpfs/scratch/ebc308/Homo_sapiens/UCSC/hg38/Sequence/Bowtie2Index/genome -U /gpfs/scratch/ebc308/AIS/coursework6data/SRR79924${SLURM_ARRAY_TASK_ID}/SRR79924${SLURM_ARRAY_TASK_ID}_trimmed.fq.gz -S /gpfs/scratch/ebc308/AIS/coursework6data/SRR79924${SLURM_ARRAY_TASK_ID}/SRR79924${SLURM_ARRAY_TASK_ID}.sam

samtools view -S -b /gpfs/scratch/ebc308/AIS/coursework6data/SRR79924${SLURM_ARRAY_TASK_ID}/SRR79924${SLURM_ARRAY_TASK_ID}.sam > /gpfs/scratch/ebc308/AIS/coursework6data/SRR79924${SLURM_ARRAY_TASK_ID}/SRR79924${SLURM_ARRAY_TASK_ID}.bam

#sort bam files
samtools sort /gpfs/scratch/ebc308/AIS/coursework6data/SRR79924${SLURM_ARRAY_TASK_ID}/SRR79924${SLURM_ARRAY_TASK_ID}.bam -o /gpfs/scratch/ebc308/AIS/coursework6data/SRR79924${SLURM_ARRAY_TASK_ID}/SRR79924${SLURM_ARRAY_TASK_ID}_sorted.bam
```

```
sbatch --array=50,51,52,53,54,55,56,57,58,59,60,61 download_trim_bowtie.sh
```

2. Use exomePeak to identify genes producing m^{6}A modified transcripts that differ between test and control datasets

```
### STEP ONE - LOAD PACKAGE AND DATA
library("exomePeak")

#gene annotation
gtf <- "/gpfs/scratch/ebc308/Homo_sapiens/UCSC/hg38/Annotation/Genes.gencode/genes.gtf"

#ip_bam - untreated condition
```

```
#SRR7992455   (mCTRL1), SRR7992454 (mCTRL2), SRR7992459   (mCTRL3)
mCTRL1 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992455/SRR7992455_sorted.bam"
mCTRL2 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992454/SRR7992454_sorted.bam"
mCTRL3 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992459/SRR7992459_sorted.bam"

#input_bam - control samples from the untreated condition
# SRR7992458 (iCTRL1), SRR7992461 (iCTRL2), SRR7992460 (iCTRL3)
iCTRL1 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992458/SRR7992458_sorted.bam"
iCTRL2 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992461/SRR7992461_sorted.bam"
iCTRL3 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992460/SRR7992460_sorted.bam"

#treated_ip_bam - treated condition
#SRR7992453 (mDS1), SRR7992452(mDS2), SRR7992451 (mDS3)
mDS1 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992453/SRR7992453_sorted.bam"
mDS2 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992452/SRR7992452_sorted.bam"
mDS3 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992451/SRR7992451_sorted.bam"

#treated_input_bam - control samples from the treated condition
#SRR7992450 (iDS1), SRR7992457 (iDS2), SRR7992456 (iDS3)
iDS1 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992450/SRR7992450_sorted.bam"
iDS2 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992457/SRR7992457_sorted.bam"
iDS3 <- "/gpfs/scratch/ebc308/AIS/coursework6data/SRR7992456/SRR7992456_sorted.bam"

IP_BAM <- c(mCTRL1,mCTRL2,mCTRL3)
INPUT_BAM <- c(iCTRL1,iCTRL2,iCTRL3)
TREATED_IP_BAM <- c(mDS1,mDS2,mDS3)
TREATED_INPUT_BAM <- c(iDS1,iDS2,iDS3)
```

```
### SET WORKING DIRECTORY WHERE YOU WANT YOUR OUTPUT TO BE PLACED
setwd("/gpfs/home/ebc308/AIS_coursework/coursework6/exomePeakResult")

result <- exomepeak(GENE_ANNO_GTF=gtf, IP_BAM=IP_BAM, INPUT_BAM=INPUT_BAM,
TREATED_IP_BAM=TREATED_IP_BAM, TREATED_INPUT_BAM=TREATED_INPUT_BAM)

setwd("/gpfs/home/ebc308/AIS_coursework/coursework6/exomePeakResult/
exomePeak_output")
con_sig_diff_peak <- read.table("con_sig_diff_peak.xls", head = TRUE)
length(unique(con_sig_diff_peak$name)) #167
```

Datasets	Genes Modified
All Three Datasets	167

I first ran exomePeak for all three datasets together to get how many genes were consistently modified (con_sig_diff file to get the consistently significantly differentiated peaks) in all the datasets. I then ran exomePeak for each treated dataset individually to get the number of genes modified for that specific dataset. I picked the con_sig_diff_peak.xls of each dataset because it shows the peaks that consistently show up in every dataset - which indicate the highest confidence.*

```
#DS1
### SET WORKING DIRECTORY WHERE YOU WANT YOUR OUTPUT TO BE PLACED
setwd("/gpfs/home/ebc308/AIS_coursework/coursework6/DS1")
result1 <- exomepeak(GENE_ANNO_GTF=gtf, IP_BAM=IP_BAM, INPUT_BAM=INPUT_BAM,
TREATED_IP_BAM=c(mDS1), TREATED_INPUT_BAM=c(iDS1))

#DS2
### SET WORKING DIRECTORY WHERE YOU WANT YOUR OUTPUT TO BE PLACED
setwd("/gpfs/home/ebc308/AIS_coursework/coursework6_exomePeak/DS2")
result2 <- exomepeak(GENE_ANNO_GTF=gtf, IP_BAM=IP_BAM, INPUT_BAM=INPUT_BAM,
TREATED_IP_BAM=c(mDS2), TREATED_INPUT_BAM=c(iDS2))

#DS3
### SET WORKING DIRECTORY WHERE YOU WANT YOUR OUTPUT TO BE PLACED
setwd("/gpfs/home/ebc308/AIS_coursework/coursework6_exomePeak/DS3")
result3 <- exomepeak(GENE_ANNO_GTF=gtf, IP_BAM=IP_BAM, INPUT_BAM=INPUT_BAM,
TREATED_IP_BAM=c(mDS3), TREATED_INPUT_BAM=c(iDS3))
```

- How many genes are modified in each individual dataset?

```

# set directory
# open con_sig_diff to get consistently signficantly differentiated peaks
# get all of the modified gene names from each dataset
# unique to ensure no overlap of gene names

setwd("/gpfs/home/ebc308/AIS_coursework/coursework6/DS1/exomePeak_output")
ds1_con_sig_diff_peak <- read.table("con_sig_diff_peak.xls", head = TRUE)
length(unique(ds1_con_sig_diff_peak$name)) #171

setwd("/gpfs/home/ebc308/AIS_coursework/coursework6/DS2/exomePeak_output")
ds2_con_sig_diff_peak <- read.table("con_sig_diff_peak.xls", head = TRUE)
length(unique(ds2_con_sig_diff_peak$name)) #155

setwd("/gpfs/home/ebc308/AIS_coursework/coursework6/DS3/exomePeak_output")
ds3_con_sig_diff_peak <- read.table("con_sig_diff_peak.xls", head = TRUE)
length(unique(ds3_con_sig_diff_peak$name)) #468
```

Dataset	Genes Modified
DS1	171
DS2	155
DS3	468

- How many genes are present in at least two test but none of the control datasets?

To get how many genes are present in at least two datasets, I merged all the unique names of genes of each con_sig_diff and then counted how many showed up at least two times.

```

# get all of the modified gene names from each dataset
# unique to ensure no overlap of gene names
ds1_names <- unique(ds1_con_sig_diff_peak$name)
ds2_names <- unique(ds2_con_sig_diff_peak$name)
ds3_names <- unique(ds3_con_sig_diff_peak$name)
```

```
# merge all the names of genes modified from all datasets together
total <- c(as.vector(ds1_names),as.vector(ds2_names), as.vector(ds3_names))

# table() creates a table with the frequency of each element and element name
mergedTogether <- DataFrame(table(total))
nrow(subset(mergedTogether, Freq >= 2 )) #153
```

Datasets	Genes Modified
At Least Two Datasets	153

3. Use m^{6}A viewer to show the m^{6}A peak structure in IFNB1 (screenshot is fine)

- Note: peaks will only be visible in the test dataset as IFNB1 is not detected in the control datasets

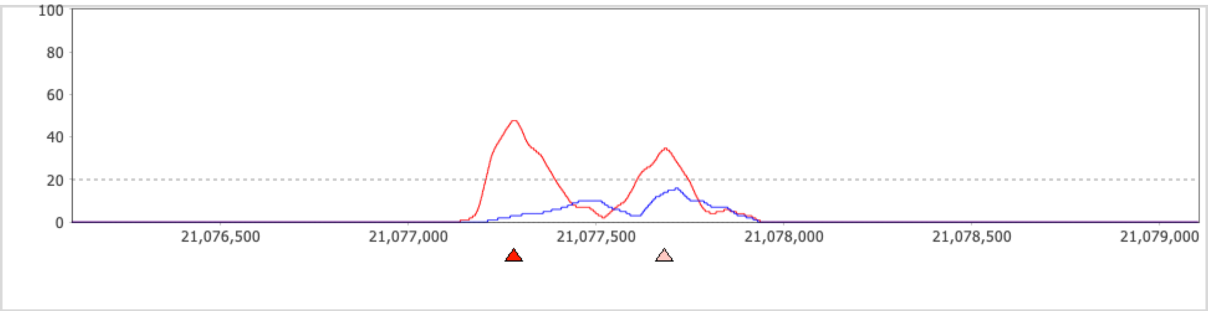
m^{6}A viewer requires files to be sorted and indexed:

```
module purge
module load samtools/1.3

# copy all bam files to new folder
cp /gpfs/scratch/ebc308/AIS/coursework6data/SRR79924${SLURM_ARRAY_TASK_ID}/SRR79924${SLURM_ARRAY_TASK_ID}_sorted.bam /gpfs/scratch/ebc308/AIS/coursework6data/indexed

# index all files in the new folder
samtools index /gpfs/scratch/ebc308/AIS/coursework6data/indexed/SRR79924${SLURM_ARRAY_TASK_ID}_sorted.bam
```

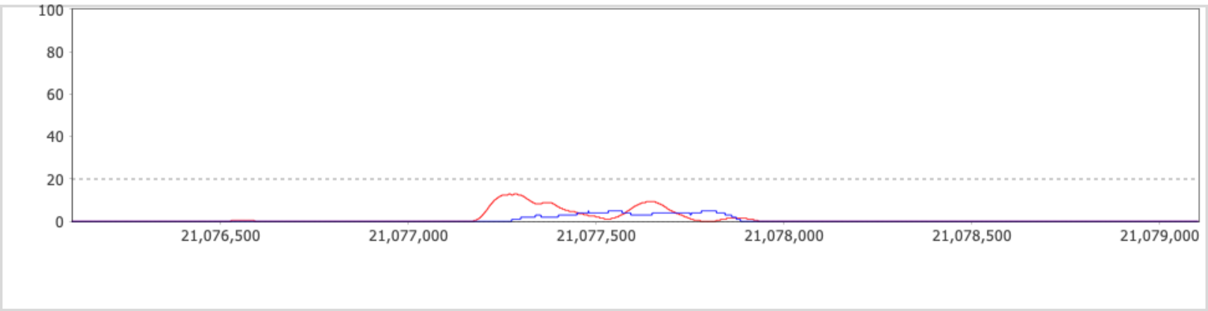
all .bai files must be in the same folder as .bam folders for m^{6}A viewer to read indexed BAM files

Location: Chromsome 9, 21076104-21079104

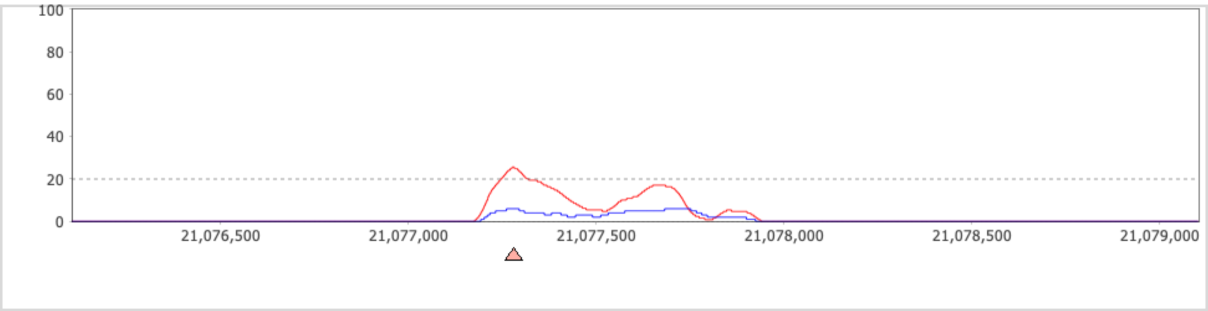
mDS1 vs iDS1



mDS2 vs iDS2



mDS3 vs iDS3



Provide full coding, justification for parameters chosen, pertinent figures and tables (with legends) -all in a single pdf file

#appliedsequencinginformatics