Erica Chio, Assigned coursework #2

Due date: February 20th (two weeks), 11pm EST

Submission:PDF document by email to daniel.depledge@nyulangone.org

1. Download dataset(s) from SRA associated with the following accessions

- SRR1523657
- SRX747060
- SRX4146457

note that run accessions (SRR) detail a single file while experiment accession (SRX) may detail multiple files

```
#!/bin/bash
#SBATCH --job-name=<NAME> # 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=1 # Run on a single CPU
#SBATCH --mem=8gb # Job memory request
#SBATCH --time=08:00:00 # Time limit hrs:min:sec
#SBATCH --output=<NAME>%j.log # Standard output and error log
#SBATCH -p cpu_short # Specifies location to submit job
```

```
module load sratoolkit/2.9.1

fastq-dump --split-files $1 --gzip -O /gpfs/scratch/ebc308/AIS/coursework2data
--origfmt

rm -r ~/ncbi
```

```
sbatch downloadSRA.sh SRR1523657
sbatch downloadSRA.sh SRX747060
sbatch downloadSRA.sh SRR7240634

sbatch downloadSRA.sh SRR7240635
```

2. What sequencing methodologies were employed? Are these single- or paired-end datasets?

SRR1523657

Homo sapiens; RNA-Seq

Instrument:Illumina HiSeq 2500

Strategy:RNA-Seq

Source:TRANSCRIPTOMIC

Selection:cDNA

Layout:PAIRED

SRX747060 (SRR1634756)

Homo sapiens; miRNA-Seq

*Instrument:*Illumina HiSeq 2000/

Strategy:*miRNA-Seq*

Source:*TRANSCRIPTOMIC*

Selection:*size fractionation*

Layout:/PAIRED

SRX4146457 (SRR7240634 & SRR7240635)

Homo sapiens; RNA-Seq

*Instrument:*Illumina HiSeq 2000/

Strategy:*RNA-Seq*

Source:*TRANSCRIPTOMIC*

Selection:*cDNA*

Layout:/PAIRED

3. Examine datasets with FASTQC

```
module purge


module load fastqc/0.11.7
```

```
fastqc -o /gpfs/scratch/ebc308/AIS/coursework2data/fastqData/$1 /gpfs/scratch/ebc308/AIS/coursework2data/$2 /gpfs/scratch/ebc308/AIS/coursework2data/$3
```

```
sbatch fastqData.sh SRR1523657 SRR1523657_1.fastq.gz SRR1523657_2.fastq.gz
```

```
sbatch fastqData.sh SRX747060 SRX747060_1.fastq.gz SRX747060_2.fastq.gz
```

```
sbatch fastqData.sh SRR7240634 SRR7240634_1.fastq.gz SRR7240634_2.fastq.gz
sbatch fastqData.sh SRR7240635 SRR7240635_1.fastq.gz SRR7240635_2.fastq.gz
```

Do datasets need adapter and/or quality trimming?

SRR1523657
Needs Adapter Trimming

SRX747060 (SRR1634756)
Needs Adapter Trimming

SRX4146457 (SRR7240634 & SRR7240635)
Does Not Need Adapter Trimming

4. Trim data with Trim Galore AND Trimmomatic

Trim Galore

```
module purge


module load python/cpu/2.7.15-ES
module load trimgalore/0.5.0
module load fastqc/0.11.7
```

```
trim_galore --fastqc -q 30 --paired --fastqc_args "--output_dir /gpfs/scratch/
ebc308/AIS/coursework2data/trimgalore" /gpfs/scratch/ebc308/AIS/
coursework2data/$1 /gpfs/scratch/ebc308/AIS/coursework2data/$2 --gzip
```

```
sbatch trimgalore.sh SRR1523657_1.fastq.gz SRR1523657_2.fastq.gz

sbatch trimgalore.sh SRX747060_1.fastq.gz SRX747060_2.fastq.gz

sbatch trimgalore.sh SRR7240634_1.fastq.gz SRR7240634_2.fastq.gz

sbatch trimgalore.sh SRR7240635_1.fastq.gz SRR7240635_2.fastq.gz
```

note: saves output file in home directory, moving to scratch directory & running there

Trimmomatic

```
module purge

module load trimmomatic/0.36

java -jar /gpfs/share/apps/trimmomatic/0.36/trimmomatic-0.36.jar PE /gpfs/
scratch/ebc308/AIS/coursework2data/$1 /gpfs/scratch/ebc308/AIS/coursework2data/
$2 $3 $4 $5 $6 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

```
sbatch trimmomatic.sh SRR1523657_1.fastq.gz SRR1523657_2.fastq.gz
SRR1523657_1P.fastq.gz SRR1523657_1U.fastq.gz SRR1523657_2P.fastq.gz
SRR1523657_2U.fastq.gz

sbatch trimmomatic.sh SRX747060_1.fastq.gz SRX747060_2.fastq.gz
SRX747060_1P.fastq.gz SRX747060_1U.fastq.gz SRX747060_2P.fastq.gz
SRX747060_2U.fastq.gz

sbatch trimmomatic.sh SRR7240634_1.fastq.gz SRR7240634_2.fastq.gz
SRR7240634_1P.fastq.gz SRR7240634_1U.fastq.gz SRR7240634_2P.fastq.gz
```

```
SRR7240634_2U.fastq.gz

sbatch trimmomatic.sh SRR7240635_1.fastq.gz SRR7240635_2.fastq.gz
SRR7240635_1P.fastq.gz SRR7240635_1U.fastq.gz SRR7240635_2P.fastq.gz
SRR7240635_2U.fastq.gz
```

note: ran in scratch folder

which performs better, how did you assess this? How did you equalize parameters?

Performing FastQC again on the newly trimmed data

```
module purge

module load fastqc/0.11.7

fastqc -o /gpfs/scratch/ebc308/AIS/coursework2data/$1 /gpfs/scratch/ebc308/AIS/
coursework2data/$2 /gpfs/scratch/ebc308/AIS/coursework2data/$3
```

```
sbatch fastqData2.sh trimmomatic/fastq trimmomatic/SRR1523657_1P.fastq.gz
trimmomatic/SRR1523657_2P.fastq.gz

sbatch fastqData2.sh trimmomatic/fastq trimmomatic/SRX747060_1P.fastq.gz
trimmomatic/SRX747060_2P.fastq.gz

sbatch fastqData2.sh trimmomatic/fastq trimmomatic/SRR7240634_1P.fastq.gz
trimmomatic/SRR7240634_2P.fastq.gz

sbatch fastqData2.sh trimmomatic/fastq trimmomatic/SRR7240635_1P.fastq.gz
trimmomatic/SRR7240635_2P.fastq.gz
```

```
sbatch fastqData2.sh trimgalore/fastq trimgalore/SRR1523657_1_val_1.fq.gz
trimgalore/SRR1523657_2_val_2.fq.gz
```

```
sbatch fastqData2.sh trimgalore/fastq trimgalore/SRX747060_1_val_1.fq.gz
trimgalore/SRX747060_2_val_2.fq.gz

sbatch fastqData2.sh trimgalore/fastq trimgalore/SRR7240634_1_val_1.fq.gz
trimgalore/SRR7240634_2_val_2.fq.gz

sbatch fastqData2.sh trimgalore/fastq trimgalore/SRR7240635_1_val_1.fq.gz
trimgalore/SRR7240635_2_val_2.fq.gz
```

Trimmomatic
Generally, sequence length distribution worsened

Trimgalore
Generally, sequence length distribution also worsened, but adapter content would improve.

Thus, I believe that trim galore would be a better trimmer all around because it was able to improve at least one testing parameter. Parameters were kept generally defaulted to equalize, besides the trimgalore -q to being 30, instead of the default 20. This q flag is the length of low quality ends that are cut off.

Adapter trimming vs base quality trimming - which has the bigger impact in each dataset?

Adapter trimming had a bigger impact on the dataset for each of the datasets. I determined this by looking a the fastq reports and logs to see what percentage of was adapter trims, and what percentage was quality trims.

Please hand in a short write-up answering the above questions and including all coding used

#appliedsequencinginformatics