

#### PIPELINE FOR 2015-06-02 ATAC-SEQ DATA ULLI ###

### MARIANA RUIZ VELASCO LEYVA ###

### MAY 28th, 2015 ###

### #PRE-ALIGNMENT QC: FASTQC

(<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

fastqc -o output.dir 1-\*

# Data contains 3 replicates in .txt.gz and .bam formats

**# Trimming of adaptors with trimmomatic, from submaster**

**# ORDER OF OUTPUT FILES MATTERS!!!**

# L001

```
bsub -M 10000 java -jar ./Trimmomatic-0.33/trimmomatic-0.33.jar PE -phred33
-trimlog ./Data_Ulli/trimmed.files/1-CGAGGCTG_S1_L001_R1_001.log
./Data_Ulli/1-CGAGGCTG_S1_L001_R1_001.fastq.gz ./Data_Ulli/1-
CGAGGCTG_S1_L001_R2_001.fastq.gz ./Data_Ulli/trimmed.files/1-
CGAGGCTG_S1_L001_R1_001.trimmed.fq ./Data_Ulli/trimmed.files/1-
CGAGGCTG_S1_L001_R1_001.unpaired.fq ./Data_Ulli/trimmed.files/1-
CGAGGCTG_S1_L001_R2_001.trimmed.fq ./Data_Ulli/trimmed.files/1-
CGAGGCTG_S1_L001_R2_001.unpaired.fq ILLUMINACLIP:./Trimmomatic-
0.33/adapters/NexteraPE-PE.fa:1:30:4 TRAILING:3 MINLEN:20
```

# WORKED!!!

**# Calculated FASTQC for post-trimmed files**

```
fastqc -o /g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/trimmed.files/
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/trimmed.files/1-
CGAGGCTG_S1_L001_R1_001.trimmed.fq
```

```
fastqc -o /g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/trimmed.files/
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/trimmed.files/1-
CGAGGCTG_S1_L001_R2_001.trimmed.fq
```

# WORKED!!!

**# Alignment with Bowtie**

# L001

```
bsub -M 20000 -o
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/bowtie.files/output-bowtie-
L001.txt bowtie -S hg19 -1
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/trimmed.files/1-
CGAGGCTG_S1_L001_R1_001.trimmed.fq -2
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/trimmed.files/1-
CGAGGCTG_S1_L001_R2_001.trimmed.fq
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/bowtie.files/1-
```

```
CGAGGCTG_S1_L0011-CGAGGCTG_S1_L001.aligned.sam  
>/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/bowtie.files/1-  
CGAGGCTG_S1_L0011-CGAGGCTG_S1_L001.aligned.log
```

### **# Removing duplicates and sorting with Samtools**

```
samtools view -bS 1-CGAGGCTG_S1_L001.aligned.sam | samtools sort -o -m  
100000000 - - > 1-CGAGGCTG_S1_L001.sort.bam  
# Done
```

# Creating an index

```
samtools index 1-CGAGGCTG_S1_L001.sort.bam 1-  
CGAGGCTG_S1_L001.sort.bam.bai
```

# Splitting reads

```
samtools view -H 1-CGAGGCTG_S1_L001.sort.bam | grep chr | cut -f2 | sed  
's/SN://g' | grep -v chrM
```

# Also ran for the alignment data the code of P. Stempor, just needed to change to the folder where my other files are

```
Rscript -e "devtools::source_gist('c977a360a702e2b3ad10')"  
# worked for file 1, failed for file 2 and stopped running for the rest
```

# Removing duplicates

```
samtools rmdup 1-CGAGGCTG_S1_L001.sort.bam 1-  
CGAGGCTG_S1_L001.rmdup.bam
```

# Creating index for the duplicates

```
samtools index 1-CGAGGCTG_S1_L001.rmdup.bam 1-  
CGAGGCTG_S1_L001.rmdup.bam.bai
```

# POST-ALIGNMENT Fastqc Report

```
fastqc -o fastqc_reports/ 1-CGAGGCTG_S1_L001.rmdup.bam
```

# Try improving alignment

```
bsub -M 20000 -o  
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/bowtie.files/output-bowtie-  
L001_2.txt bowtie -S -X2000 -m1 hg19 -1  
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/trimmed.files/1-  
CGAGGCTG_S1_L001_R1_001.trimmed.fq -2
```

```
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/trimmed.files/1-
CGAGGCTG_S1_L001_R2_001.trimmed.fq
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/bowtie.files/1-
CGAGGCTG_S1_L0011-CGAGGCTG_S1_L001.2.aligned.sam
>/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/bowtie.files/1-
CGAGGCTG_S1_L0011-CGAGGCTG_S1_L001.2.aligned.log
```

### **# Additional processing (steps not included in pipeline yet!)**

```
fastqc -o fastqc.reports/ *.aligned.sam
```

```
samtools view -b -h 1-CGAGGCTG_S1_L001.rmdup.bam chr1 chr2 chr3 chr4 chr5
chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18
chr19 chr20 chr21 chr22 chrX chrY > 1-CGAGGCTG_S1_L001.nucl.rmdup.bam
# we might also remove chrY
```

```
samtools index 1-CGAGGCTG_S1_L001.nucl.rmdup.bam
# For general statistics: number of reads per chromosome, we can use: samtools
idxstats file.nucl.rmdup.bam
```

```
# Shifting base pairs
```

```
sh /g/scb/zaugg/carnold/scripts/shiftPosBAM.sh
# didn't run but should be a small bug
```

```
samtools view -H 1-CGAGGCTG_S1_L001.nucl.rmdup.bam > 1-
CGAGGCTG_S1_L001.nucl.rmdup.bam.header
```

```
samtools view 1-CGAGGCTG_S1_L001.nucl.rmdup.bam | awk 'BEGIN {OFS = "\t"}
; {if ($9 < 0) $4=$4-5; else $4=$4+4; print $0}' > 1-
CGAGGCTG_S1_L001.nucl.rmdup.bam.shifted.sam
```

```
cat 1-CGAGGCTG_S1_L001.nucl.rmdup.bam.header 1-
CGAGGCTG_S1_L001.nucl.rmdup.bam.shifted.sam | samtools view -bS - > 1-
CGAGGCTG_S1_L001.nucl.rmdup.bam.shifted.bam
# the important final file is the .shifted.bam for calling with macs2
```

```
#for f in
/g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/samtools.files/*.nucl.rmdup.bam
```

```
#do
#     echo "Processing $f"
```

```
    #1. Save the header so it does not get lost
    #samtools view -H $inputFile >$inputFile.header
```

```
    #2. Shift positions according to strand and save as new SAM
    #samtools view $inputFile | awk 'BEGIN {OFS = "\t"} ; {if ($9 < 0) $4=$4-5;
else $4=$4+4; print $0}' > $inputFile.shifted.sam
```

```
    #3. Concatenate header again and converse to BAM
```

```
#cat $inputFile.header $inputFile.shifted.sam | samtools view -bS - >
$inputFile.shifted.bam
```

```
#done
```

**# In Perl pipeline they calculate insert size distribution by parsing through the samtools flags 99 and 163... we can check if we define insert size in this way or other.**

### **# PEAK CALLING!**

```
macs2 callpeak -t /g/scb/zaugg/zaugg_shared/ATACSeq/Data_Ulli/samtools.files/1-
CGAGGCTG_S1_L001.nucl.rmdup.bam.shifted.bam -f BAMPE -g hs -n
Ulli.L001.peak.calling --outdir ../macs2.files/ --nomodel 2>
../macs2.files/macs2_out.log
# function 'callpeak' of macs2 with arguments{-t input file -f format -g Genome
effective size = 'hs' for human --nomodel ATACseq specific}
```

**# Additionally, we can add in between steps for removing intermediate files :)**

**# Next steps in R**