

atac-seq pipeline

Amal Thomas

August 6, 2016

A computational pipeline for processing and analyzing atac-seq data

Contents

1	Overview	2
2	Pre-processing	2
2.1	Quality checks	2
2.1.1	Input files	2
2.1.2	Quality check	2
2.1.3	Remove adapters	2
2.2	Mapping reads	2
2.2.1	Alignment	2
2.3	Quality controls	3
2.3.1	Filter unwanted reads	3
2.3.2	Remove duplicate reads	3
2.3.3	Remove low quality reads	3
2.4	Peak Calling	3
2.4.1	Shift coordinates	3
2.4.2	MACS2 peak calling	4
2.5	Visualization	4
2.5.1	bigBed & bigWig file	4
3	Post-processing	5
3.1	Plot insert size	5
3.2	Replicate correlation	5
3.3	Differential analysis	6

1 Overview

In our pipeline, we provide a handy work flow to process the atac-seq data and identify those regions with differential chromatin profile. The pipeline has mainly two parts:

- Pre-processing part: Here we process the input raw data and prepare it for the downstream analysis. The main steps include quality checks of the raw data, read alignment to the appropriate reference genome and subsequent filtering, peak calling and creating UCSC Track Hub.
- Post-processing part: Here we try to identify those chromosomal regions that show significant difference in the chromatin accessibility.

Users are assumed to have prior experience with UNIX/Linux environment and common bioinformatics tools.

2 Pre-processing

2.1 Quality checks

2.1.1 Input files

The input to our pipeline is raw next generation sequencing data in the FASTQ format. The raw files from Sequence Read Archive (SRA), have to be decompressed and properly split to generate the right FASTQ files.

2.1.2 Quality check

FastQC provides an easy way to analyse the quality of raw sequencing data. This will give us a hint whether our data has any experimental error. One can employ the tool by the following command:

```
fastqc <read1.fastq> <read2.fastq> -o <outputdir>
```

2.1.3 Remove adapters

Sequencing adapters can be easily removed using trimgalore program. Typical command for paired end reads is:

```
trim_galore --paired <read1.fastq> <read2.fastq> \
--output_dir <output directory>
```

Optional arguments:

To remove low quality reads use `--quality` parameter:

```
--quality <cutoff>
```

2.2 Mapping reads

2.2.1 Alignment

The trimmed read are mapped to respective genome using bowtie2. Since the minimum distance between two Tn5 binding sites are about 38bp, only fragment size greater than this is kept.

```
bowtie2 -X 2000 --fr --no-discordant --no-mixed \
--minins 38 -x <bowtie-index> -1 <read1_trimmed.fastq> \
-2 <read2_trimmed.fastq> -S <read_mapped.sam>
```

Create sorted bam file: The output of mapping is in sam format, which is human readable. The downside of sam format is the huge file size. We can use samtools to convert samfile into smaller and compressed format called bam.

```
samtools view -bS read_mapped.sam \  
|samtools sort - read_mapped.sorted.bam
```

2.3 Quality controls

2.3.1 Filter unwanted reads

Here we keep only those reads that is mapped to standard chromosomes. Reads mapped to mitochondrial DNA are also filtered out.

```
samtools view -h read_mapped.sorted.bam \  
|awk 'substr($0,1,1) == "@" || (length($3)<=5 && \  
$3!="chrM" && $3 != "*"){{print $0}}' \  
|samtools view -bS - > read_mapped.sorted.chr.bam
```

2.3.2 Remove duplicate reads

PCR duplicated reads are removed using Picard Tools MarkDuplicates

```
java -Xmx2g -jar MarkDuplicates.jar \  
INPUT = read_mapped.sorted.chr.bam \  
METRICS_FILE=read_mapped.markdup.metrics" \  
OUTPUT=read_mapped.sorted.chr.nodup.bam" \  
REMOVE_DUPLICATES=true \  
ASSUME_SORTED=true
```

2.3.3 Remove low quality reads

Reads with Phred quality scores less than 30 are filtered out using the command:

```
samtools view -b -h -q 30 read_mapped.sorted.chr.nodup.bam \  
> read_mapped.sorted.chr.nodup.filt.bam
```

The output bam file can be indexed using:

```
samtools index read_mapped.sorted.chr.nodup.filt.bam
```

2.4 Peak Calling

2.4.1 Shift coordinates

Inorder to represent the center of Tn5 binding site we shift the reads aligning to the postive strand by +4 bp, and those to the negative strand by 5 bp. We can do this easily by converting the bam file to bed format and then shift the coordinates.

```
bamToBed -i read_mapped.sorted.chr.nodup.filt.bam > temp.bed
```

```
samtools view read_mapped.sorted.chr.nodup.filt.bam \  
| awk 'BEGIN{OFS="\t"}{print $1,$9}' > insert.temp
```

```
paste temp.bed insert.temp | awk 'BEGIN{OFS="\t"}{split($4,name,"/"); \  
if(name[1]== $7 && $6 == "+" ) {print $1,$2+4,$3,$4,$8,$6} \  
else if(name[1]== $7 && $6 == "-" ) {print $1,$2,$3-5,$4,$8,$6} \  
'
```

```
> read_mapped.sorted.chr.nodup.filt.shift.bed

rm temp.bed .insert.temp
```

2.4.2 MACS2 peak calling

Now we are ready to perform peak calling. We will use peak caller called MACS2. The shifted bed file is provided as the input for the program.

```
macs2 callpeak --gsize hs \
    -f BED \
    --treatment read_mapped.sorted.chr.nodup.filt.shift.bed \
    --name read_mapped \
    --keep-dup all \
    --call-summits \
    --shift -100 --extsize 200 \
    --nomodel --nolambda \
    --verbose 3
```

The `--shift` parameter will shift the reads in 5' to 3' direction and `--extsize` parameter will extend the 3' end to make 200 bp reads. This will ensure that center of each reads correspond to the original Tn5 cutting sites. For *e.g.*

Original reads:

chr1	500	550	read1	.	+
chr1	700	750	read2	.	-

Applying `--shift -100`: negative 100 will shift in 3' to 5' direction

chr1	400	450	read1	.	+
chr1	800	850	read2	.	-

Applying `--extsize 200`:

chr1	400	600	read1	.	+
chr1	650	850	read2	.	-

Remove blacklisted regions: We have to remove some problematic regions (microsatellites, centromeres, telomeres etc.) from the identified peaks. Encode blacklisted regions can be downloaded from this [link](#).

```
intersectBed -v -a read_mapped_peaks.narrowPeak \
-b hg19_consensusBlacklist.bed > read_mapped_peaks.narrowPeak.blacklistcleared
```

2.5 Visualization

2.5.1 bigBed & bigWig file

bigBed file: `bedToBigBed` can be downloaded from the ucsc [site](#). The following command creates bigBed file of the identified peaks:

```
awk '{OFS="\t"; print $1, $2, $3}' \
read_mapped_peaks.narrowPeak.blacklistcleared | sort -k 1,1 -k 2,2n \
> temp.track
```

```
bedToBigBed temp.track hg19.sizes.txt read.bb
rm temp.track
```

Here hg19.sizes is a file with sizes of the chromosomes.

bigWig file: First we create a bed file with read size of one at Tn5 cut site.

```
cat read_mapped.filt.sorted.chr.nodup.shift.bed | \
awk 'BEGIN{OFS="\t"}{if($6 == "-") $2=$3-1; \
print $1, $2, $2+1, $4, $5, $6}' \
| sort -k 1,1 -k2,2n > read_mapped.filt.sorted.chr.nodup.shift.center.bed
```

Then we extend reads to both sides of the cut site center by 15bp. This region represents the accessible sites for Tn5 binding.

```
bedtools slop -i read_mapped.filt.sorted.chr.nodup.shift.center.bed \
-g hg19.sizes -b 15 > read_mapped.15bpshifted.bed
```

Find the librarysize and normalising factor using:

```
librarySize=$(samtools view -c -F 4 read_mapped.sorted.chr.nodup.filt.bam)
expr="1000000 / $librarySize"
scaling_factor=$(echo $expr | bc -l)
```

Find the number of reads falling into Tn5 accessible regions using:

```
bedtools genomecov -i read_mapped.15bpshifted.bed \
-g hg19.sizes.txt -bg -scale $scaling_factor > read.bedgraph
```

Finally the bedGraphToBigWig available [ucsc site](#) can be used to create bigWig file from the generated bigwig file.

```
bedGraphToBigWig read.bedgraph hg19.sizes.txt read.bw
```

3 Post-processing

3.1 Plot insert size

We will use the `plotinsertsize.R` script to generate the insert size distribution. The script requires an input file with insert size of aligned reads. Column 9 of the BAM file contains the insertsize length. We will extract it from the BAM file using the command:

```
samtools view -f66 read_mapped.sorted.chr.nodup.filt.bam \
|cut -f 9|sed 's/^-//' > insertsize.txt
```

Once the insertsize file is generated, then run the Rscript:

```
Rscript plotinsertsize.R insertsize.txt
```

3.2 Replicate correlation

Genome-wide replicate correlation: Most of the experiments are performed to have one or more replicates. Having more replicates ensures the reproducibility of the results. It is a good practise to check the concordance of the replicates before proceeding to further analysis. The `genomewide-replicate-correlation.R` finds the correlation between samples by counting the number of reads in 1000 bp non-overlapping window. The `correlation.xml` file contains the information of the input files. To run the script:

```
Rscript genomewide-replicate-correlation.R correlation.xml
```

3.3 Differential analysis

Here we use sliding window approach mentioned in the csaw paper. Using a sliding window instead of apriori called peaks makes type I error easier to control. Window-based approach has been shown to outperform some existing peak-based methods when the relevant parameters are set intelligently. The script `differential-atac.R` follow the Bioconductor workflow created by Aaron Lun for differential chip-seq analysis. The code was adapted for ATAC-seq data. The input files and path to the blacklist BED file should be mentioned in the XML file.

```
Rscript differential_atac.R differential.xml
```

3.4 Motif Discovery