Advanced Topics in Bioinformatics: Mapping and predicting ChIP-seq BSs with BWA and Samtools

Juanma Medina

September 12, 2017

```
library("ggplot2")
```

ALIGNMENT

```
# Align the reads and produce a .sai file of all potential candidates
bwa aln -t 10 ../../advTopics17/hg19_bwa ../../advTopics17/ENCFF000QMZ10.fastq > run1.sai

# Transform the aligning file from a single-end input.sai file to a .sam file, where
# repetitive hits are randomly chosen
bwa samse ../../advTopics17/hg19_bwa run1.sai ../../advTopics17/ENCFF000QMZ10.fastq > run1.sam

# SAM to BAM conversion
samtools view -Sb run1.sam > run1.bam
```

Exercise 1: Percentages of mapped reads

```
samtools flagstat run1.bam
```

2,323,367 mapped reads / 2,366,598 total records = 98.17% mapped reads

```
samtools view -F 20 -c run1.bam
```

1,160,488 mapped reads to the FORWARD strand
1,160,488 / 2,323,367 = 49.95%

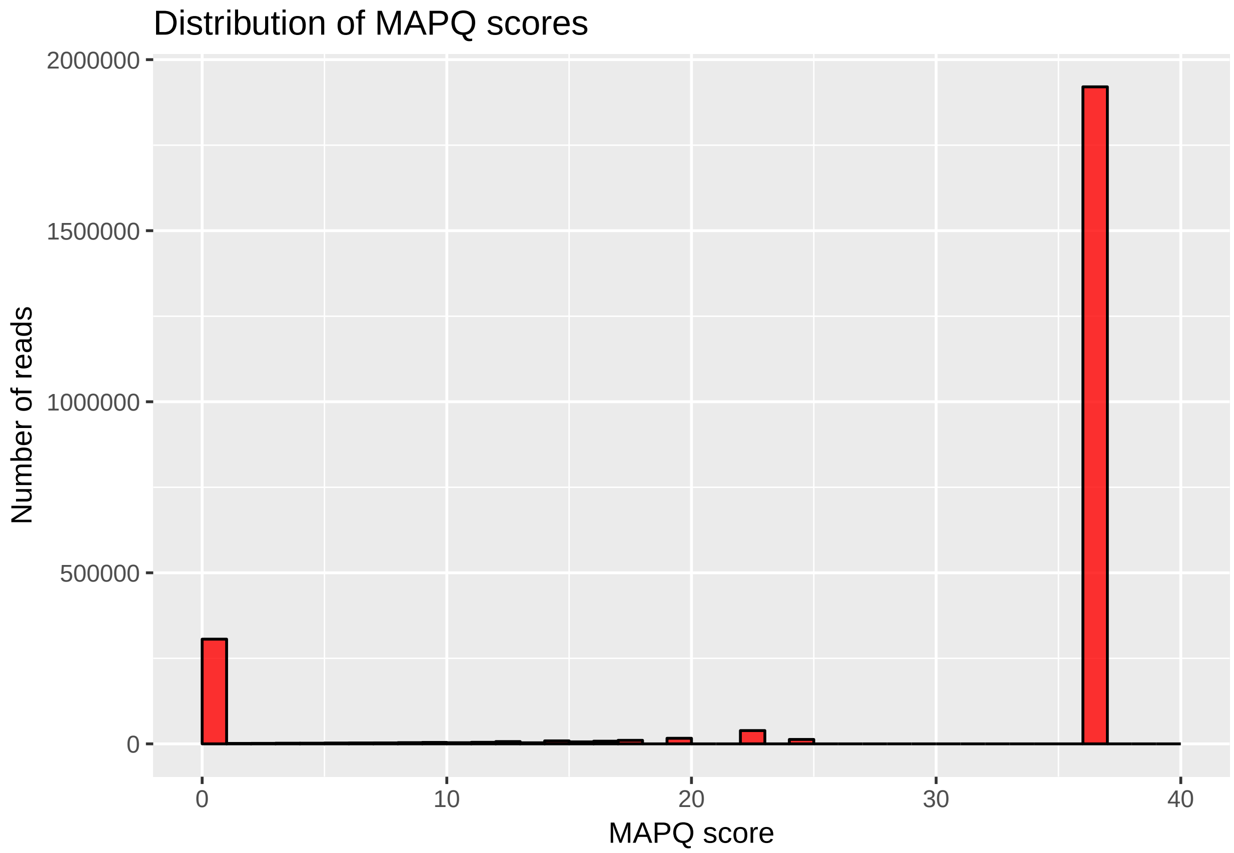
```
samtools view -f 0x10 -c run1.bam
```

1,162,879 mapped reads to the REVERSE strand
1,162,879 / 2,323,367 = 50.05%

Exercise 2: MAPQ scores distribution

```
samtools view run1.bam | cut -f 5 > scores.txt
```

```
# Data reading
mapq <- read.table("scores.txt", header = F)
names(mapq) <- "score"
```



Exercise 3: Binding sites identification

```
samtools sort -o run1.sorted run1.bam
```

Method 1: Samtools "depth" function

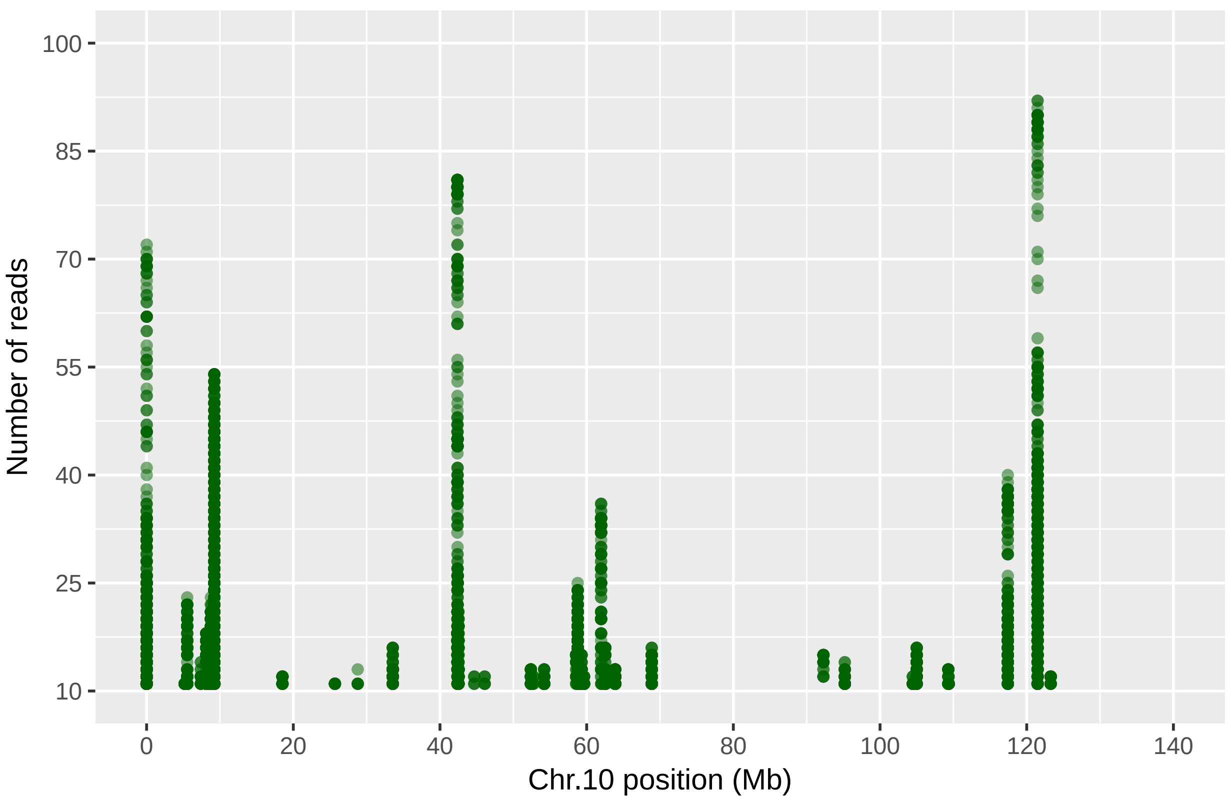
```
samtools depth -l 50 -Q 37 -q 30  run1.sorted | awk '$3 > 10' > bs.depth.txt
```

Method 2: Samtools "mpileup" function

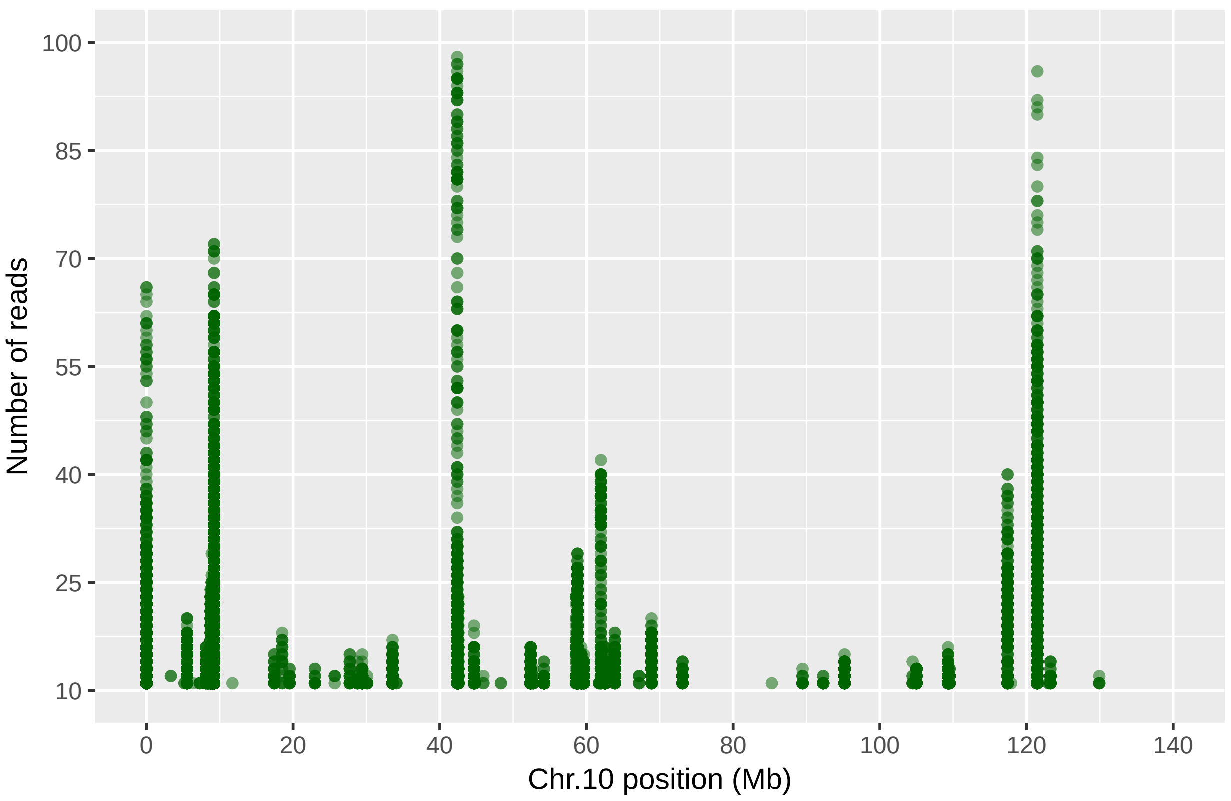
```
samtools mpileup -d 50 -q 37 -Q 40 -s -o pre.mpileup.txt run1.sorted
awk '$4 > 10' pre.mpileup.txt > bs.mpileup.txt
```

```
# Data reading
bs.depth <- read.table("bs.depth.txt", header = F)
bs.mpileup <- read.table("bs.mpileup.txt", header = F)
names(bs.depth) <- c("chr", "locus", "depth")
names(bs.mpileup) <- c("chr", "locus", ".", "depth", ".", "baseQ", "mapQ")
```

ChIP peaks (estimated using depth)



ChIP peaks (estimated using mpileup)



Exercises 4 and 5: Discussion

Establishing an intuitive but arbitrary threshold of 25 reads required to predict a ChIP binding site, there are 7 of them clearly differentiated. In both plots, the reads representated have been filtered by using the Samtools "depth" and "mpileup" useful functions, although it can be seen in the histogram that most of them have a good MAPQ score (37). In the first case (represented in the first plot), they have been strictly required to have a minimum length of 50, a mapping quality of 37 and a base quality of 30. In the second case, they have been required to not exceed a limit of 50 reads per same initial base (in order to avoid potential PCR constructs biasing the results), and again a mapping and base qualities of 37 and 40 respectively. An additional minimum depth of coverage of 10 has been set in both cases to filter out noise.

Both filtered sets output a similar number of reads (6665 and 7467 respectively), and their corresponding distributions look quite symmetrical, which is enough initial assumption to correctly predict the ChIP binding sites located within chromosome 10. Depending on the possible downstream analysis, further criteria should be considered, with a special mention at the potential ChIP hotspots that appear more clearly in the second plot with a slightly -but still decent- lower coverage