Alignment Stats

Lucas Michel Todó, Cristina Bancells
Alfred Cortes and Juan R. Gonzalez

Barcelona Global Health Institute (ISGlobal), Campus PRBB

March 6, 2017

Contents

1 Importing Data **2**

2 Plots **2**

2.1 MAPQ 2

2.2 Fragment Length 33

1 Importing Data

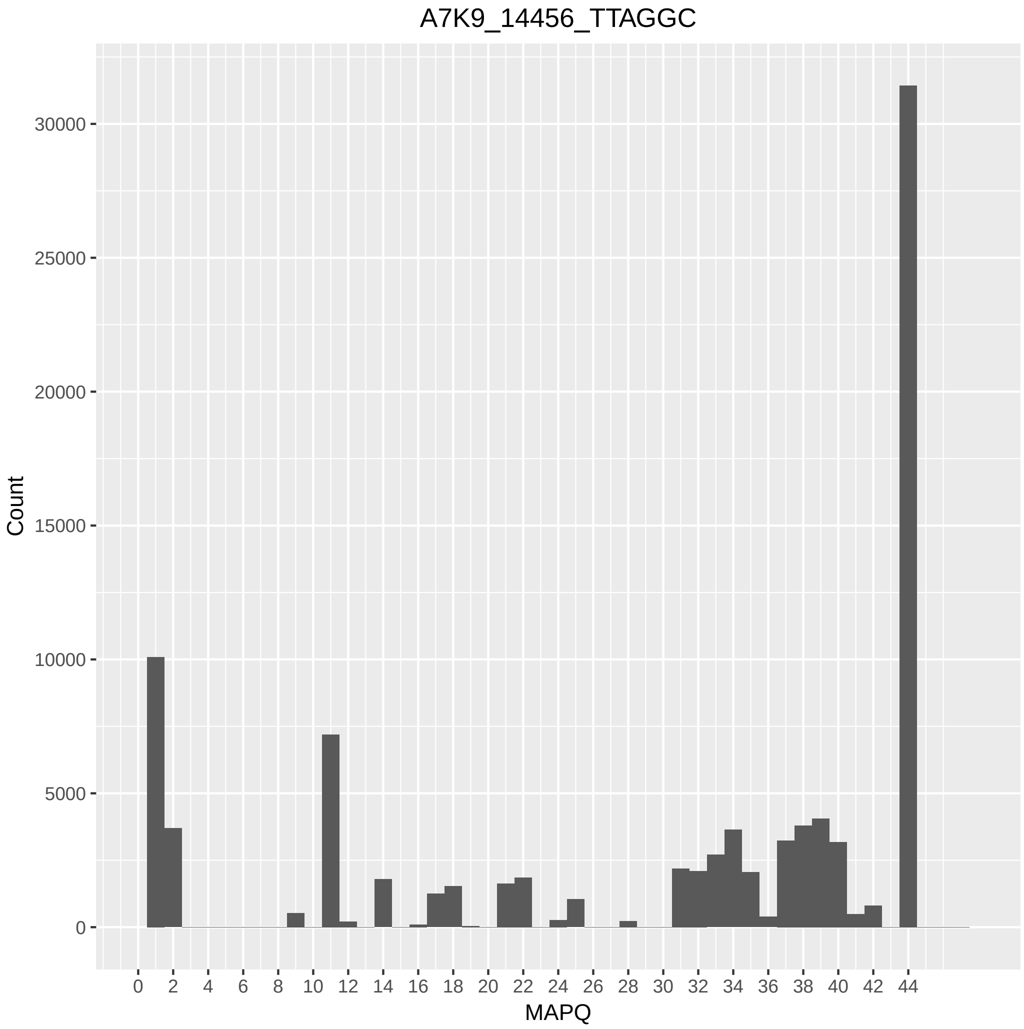
First we import the data for every alignment:

```
####  Fetch files and create iterables ####
file_names <- list.files("/home/lucas/ISGlobal/TestSet/align_tests/inputs/")
file_names <- unique(lapply(file_names, function(x) substr(x,1,17)))

dirs <- list.files("/home/lucas/ISGlobal/TestSet/align_tests/")[substr(list.file
```

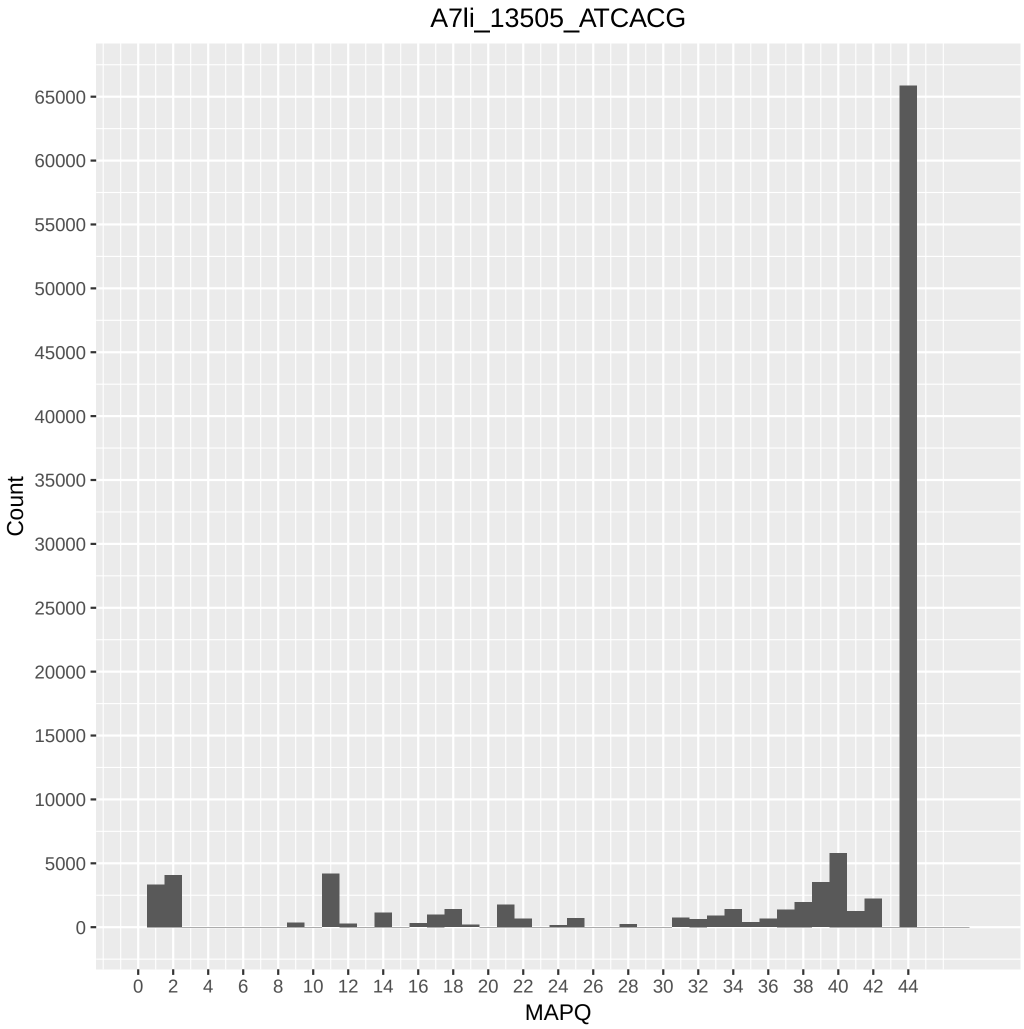
2 Plots

2.1 MAPQ

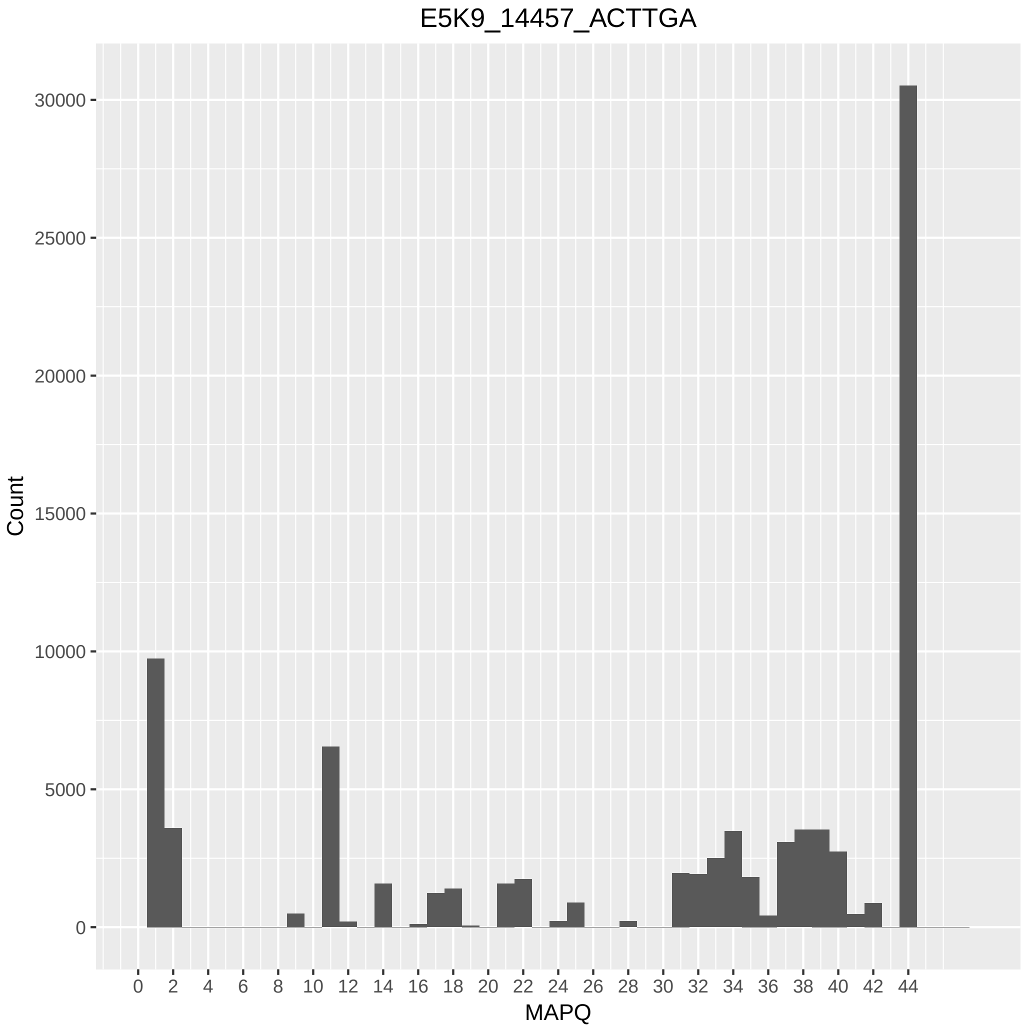
```
for (dir in dirs){
  for (x in file_names){
    mapq <- read.csv2(file = paste0("/home/lucas/ISGlobal/TestSet/align_tests/",
    df <- as.data.frame(as.numeric(mapq))
    colnames(df) <- "MAPQ"
    title <- x
    print(ggplot(df, aes(x = MAPQ)) +
        geom_histogram(binwidth = 1) +
        labs(x = "MAPQ", y = "Count") +
        ggtitle(title) +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_x_continuous(breaks = seq(0, 45, by = 2), limits = c(0,48)) +
        scale_y_continuous(breaks = seq(0,80000, by = 5000)))

    plot.new()
    print(table(df > 5))
  }
}
```


A7K9_14456_TTAGGC
Count
30000
25000
20000
15000
10000
5000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

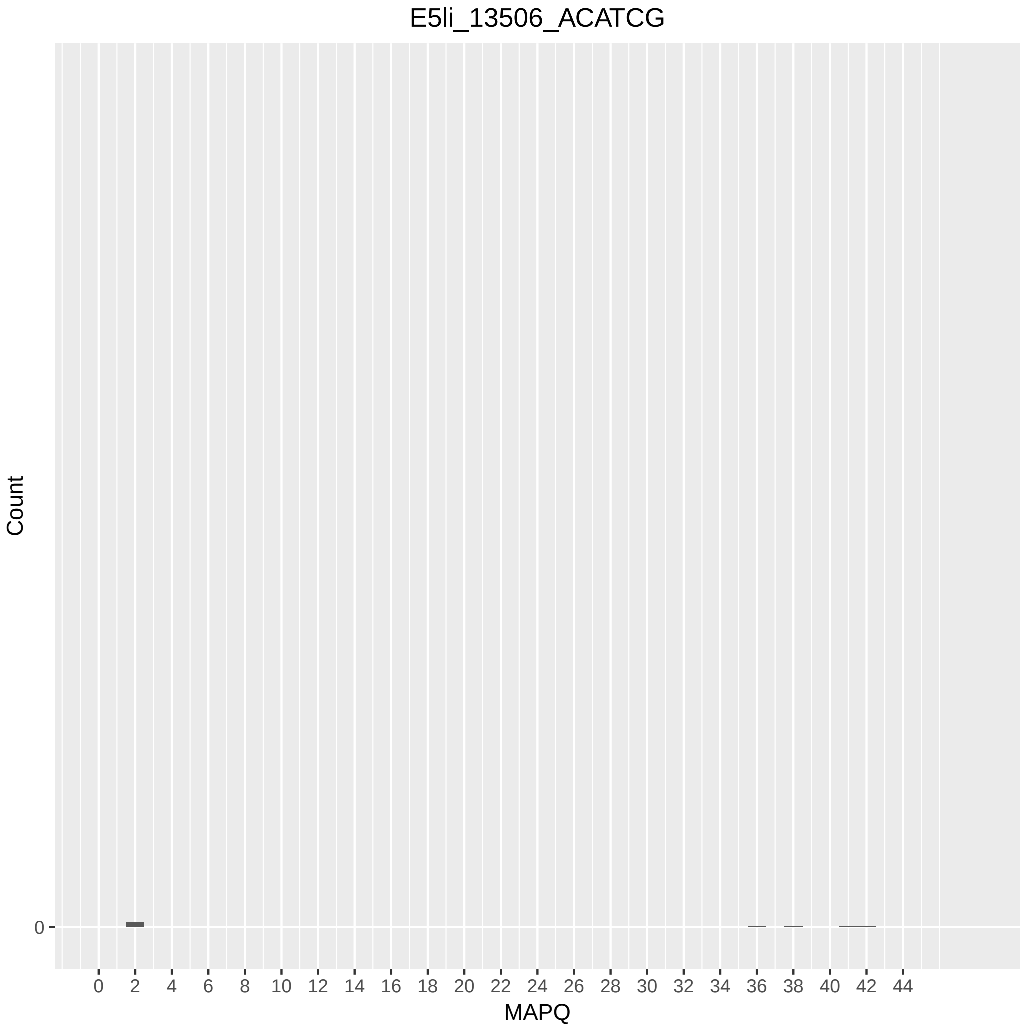
```
##
## FALSE  TRUE
## 32237 77763
```


A7li_13505_ATCACG
Count
65000
60000
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

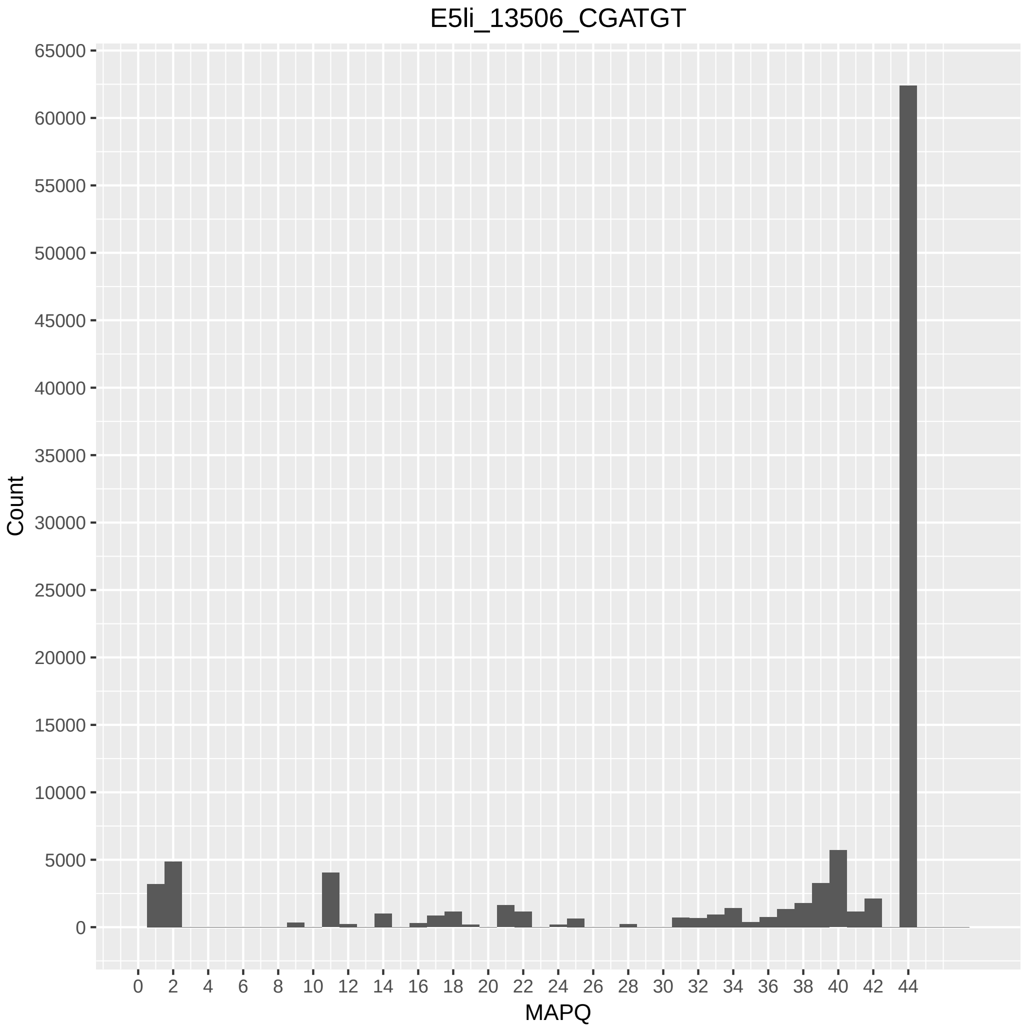
```
##
## FALSE  TRUE
## 10747 99253
```


E5K9_14457_ACTTGA
Count
30000
25000
20000
15000
10000
5000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

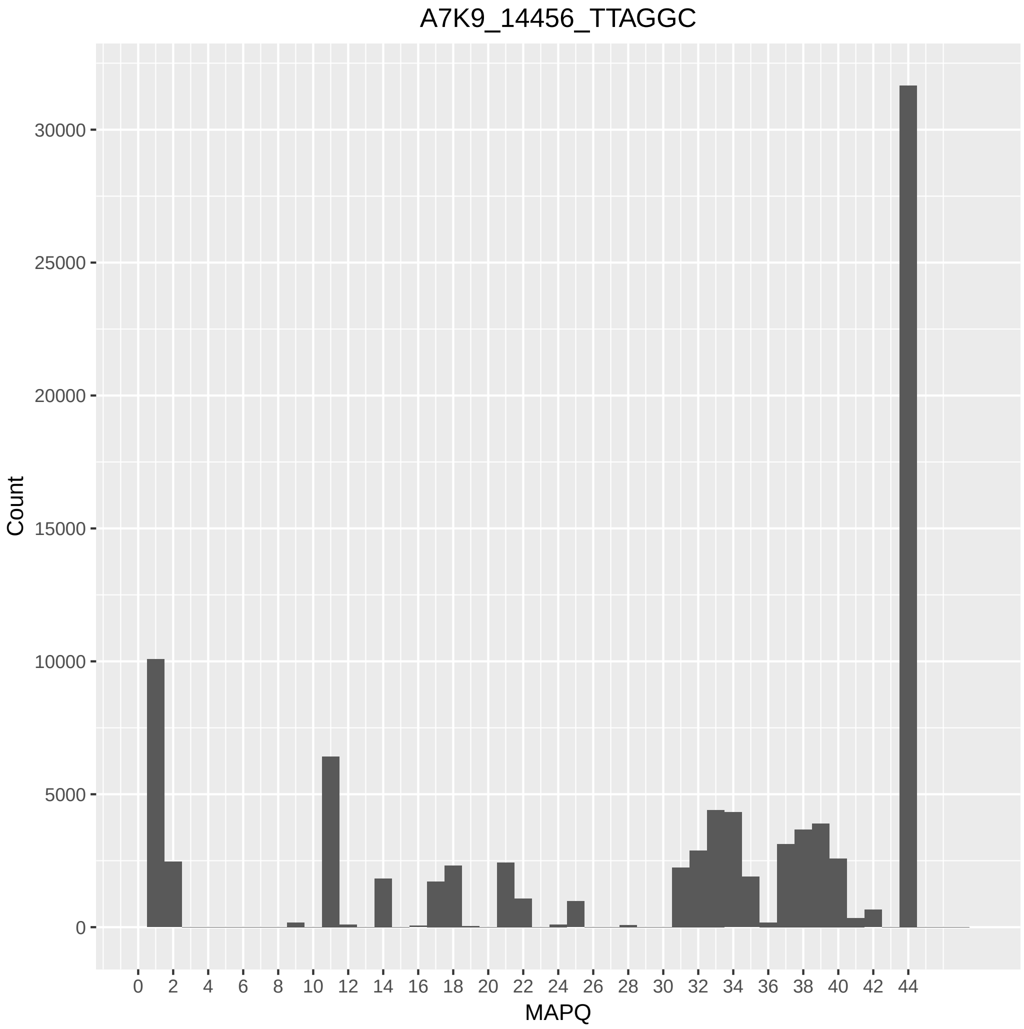
```
##
## FALSE  TRUE
## 36833 73167
```


E5li_13506_ACATCG
Count
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

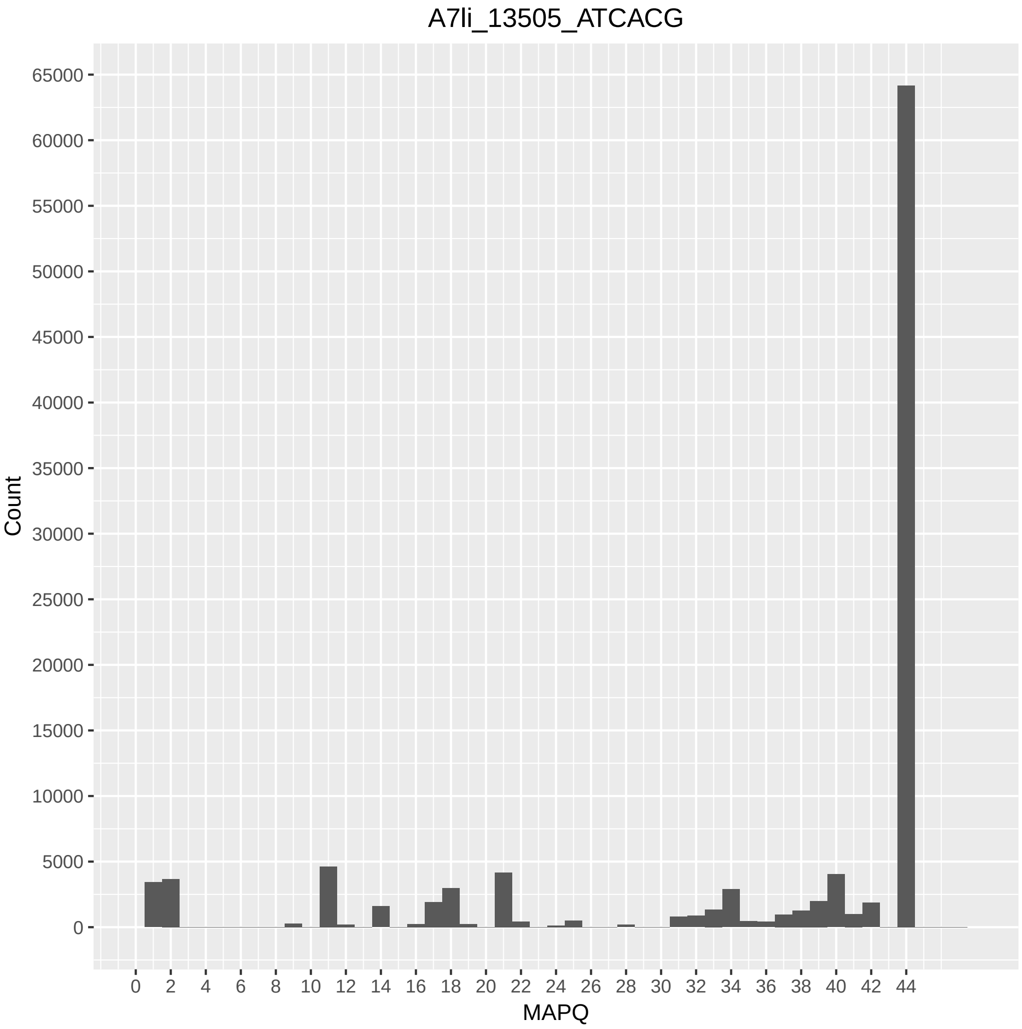
```
##
## FALSE  TRUE
##  2265     5
```


E5li_13506_CGATGT
Count
65000
60000
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

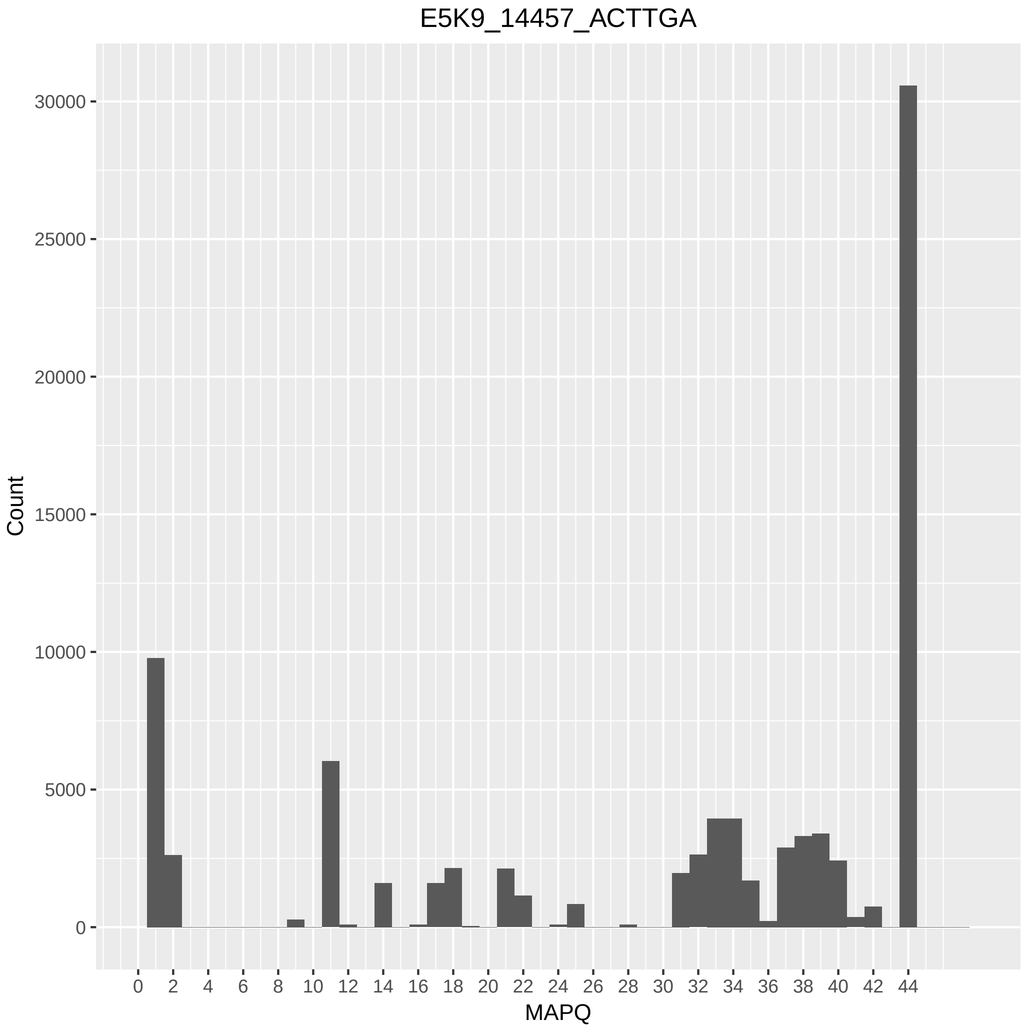
```
##
## FALSE  TRUE
## 15355 94645
```


A7K9_14456_TTAGGC
Count
30000
25000
20000
15000
10000
5000
0
MAPQ
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44

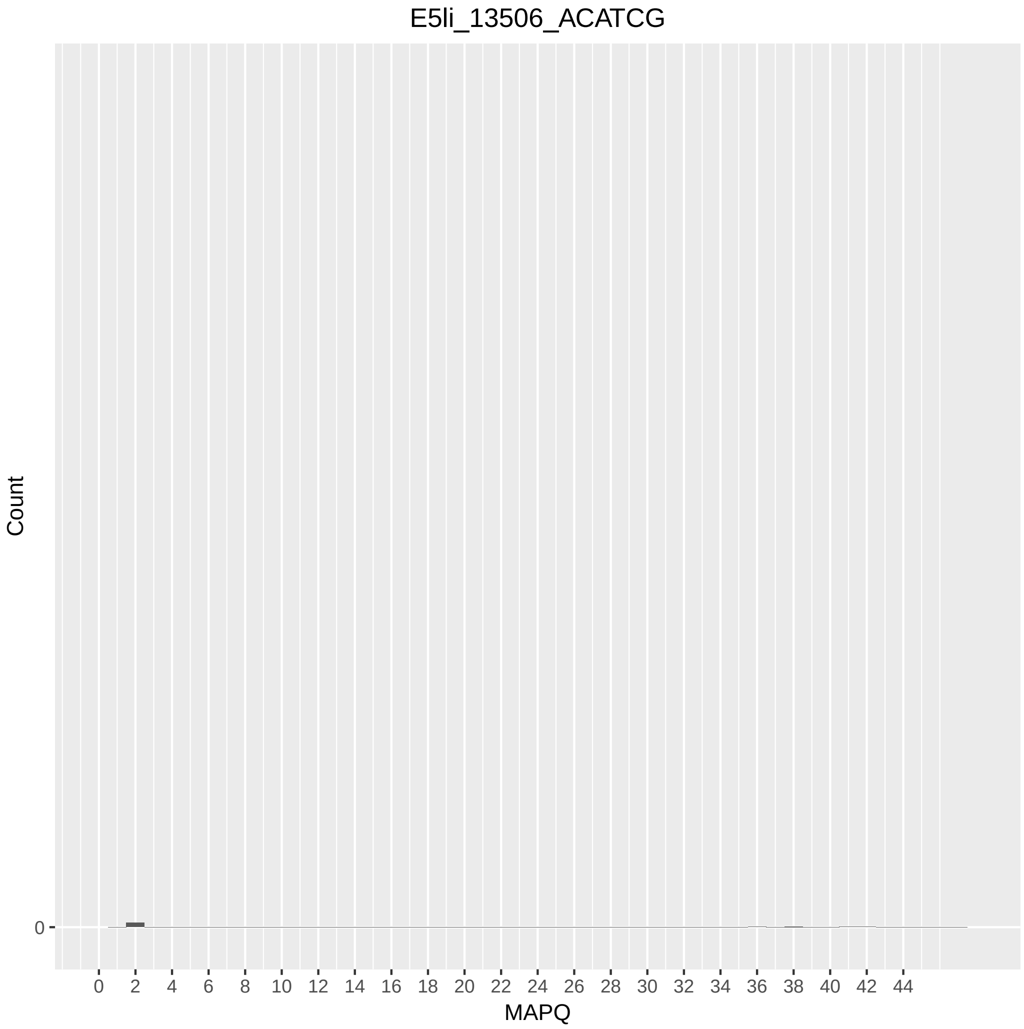
```
##
## FALSE  TRUE
## 30779 79221
```


A7li_13505_ATCACG
Count
65000
60000
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44
MAPQ

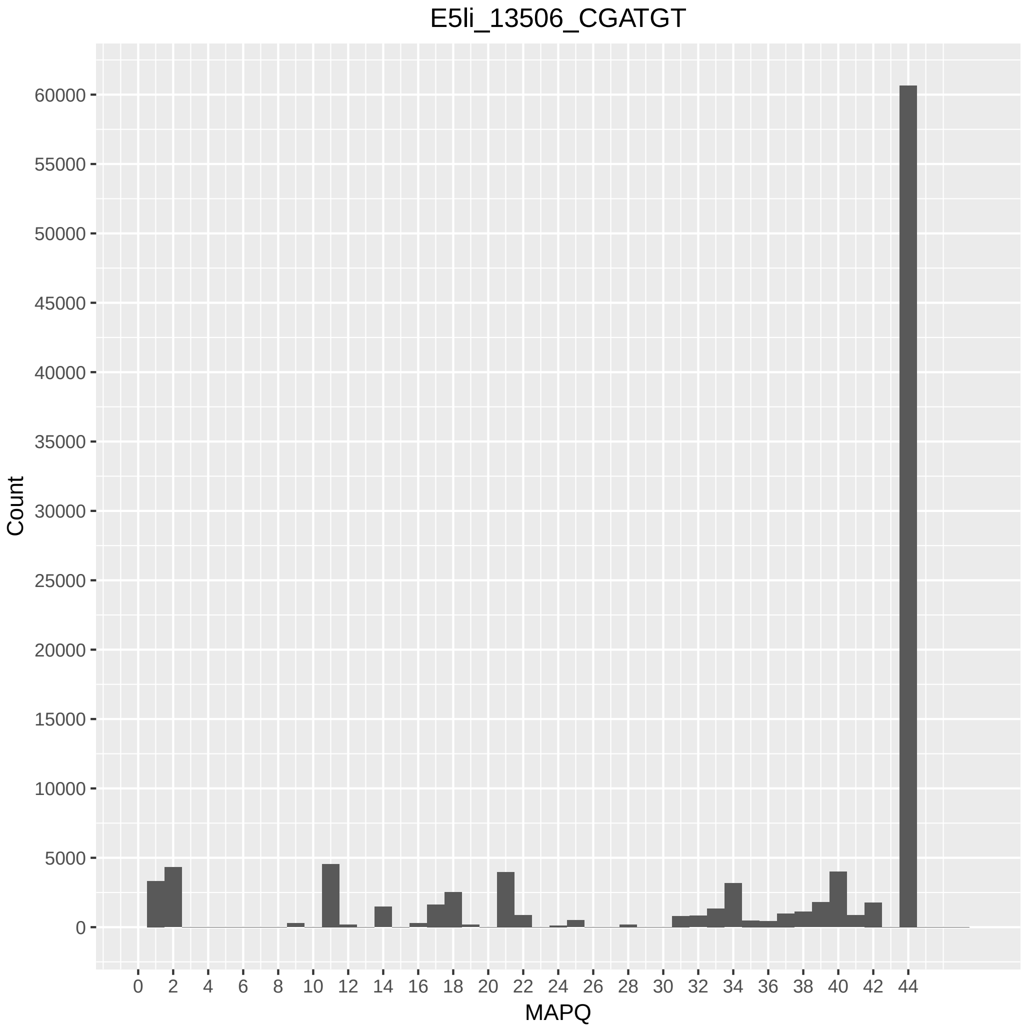
```
##
## FALSE  TRUE
## 10396 99604
```


E5K9_14457_ACTTGA
Count
30000
25000
20000
15000
10000
5000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

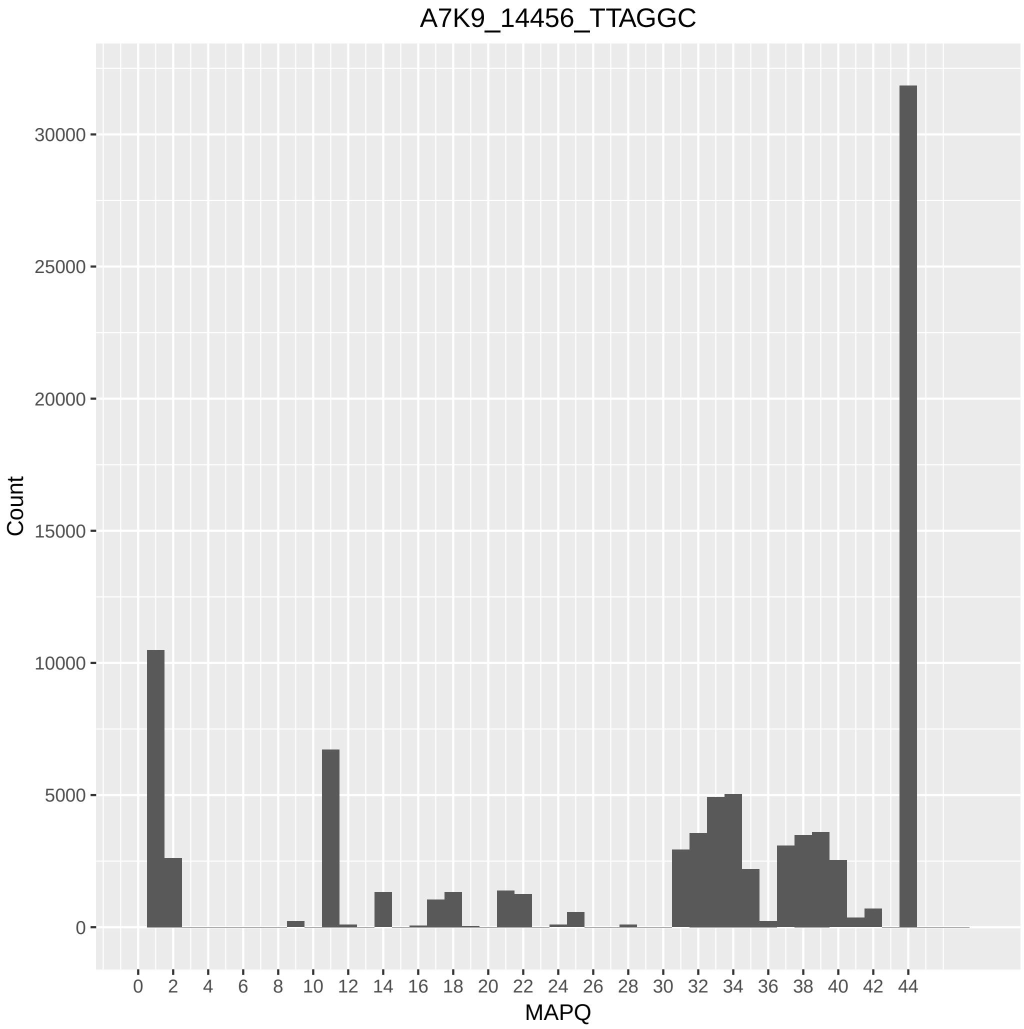
```
##
## FALSE  TRUE
## 35707 74293
```


E5li_13506_ACATCG
Count
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

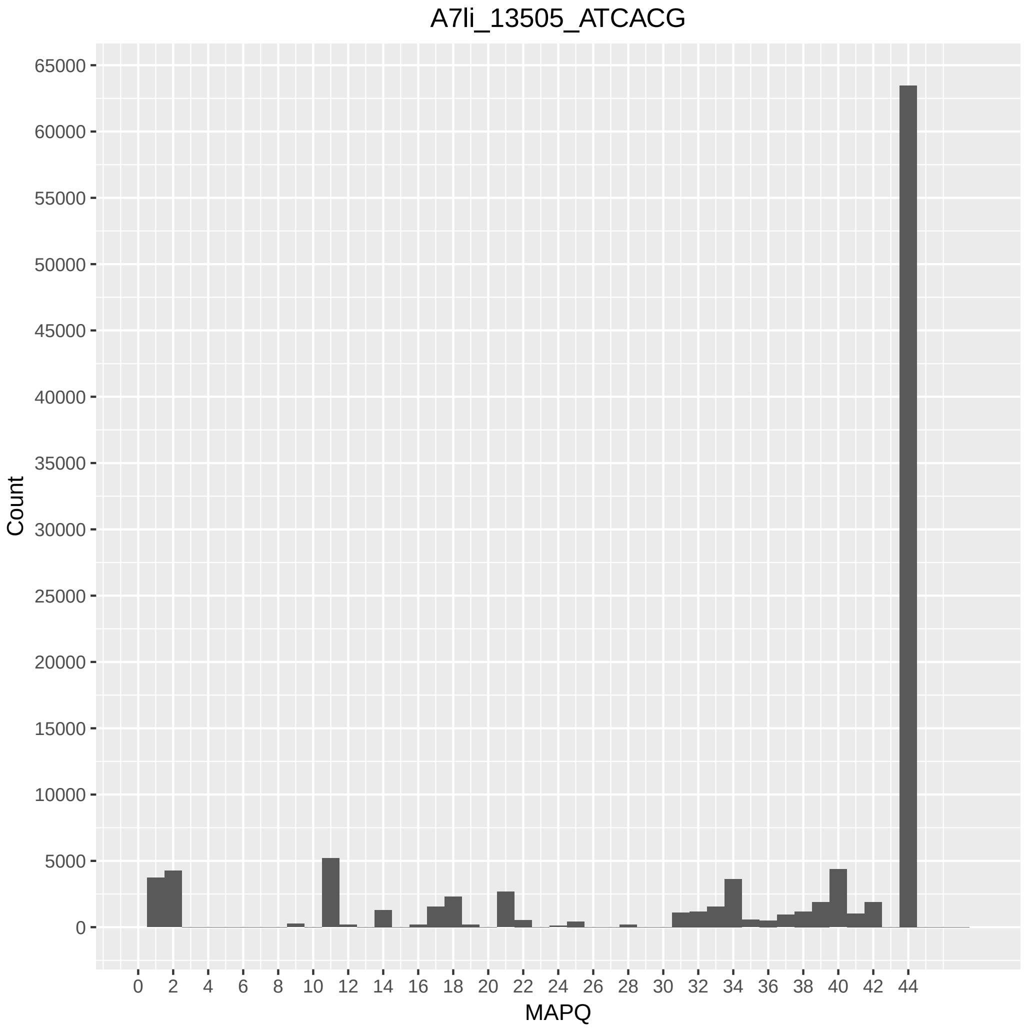
```
##
## FALSE  TRUE
##  2265     5
```


E5li_13506_CGATGT
Count
60000
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

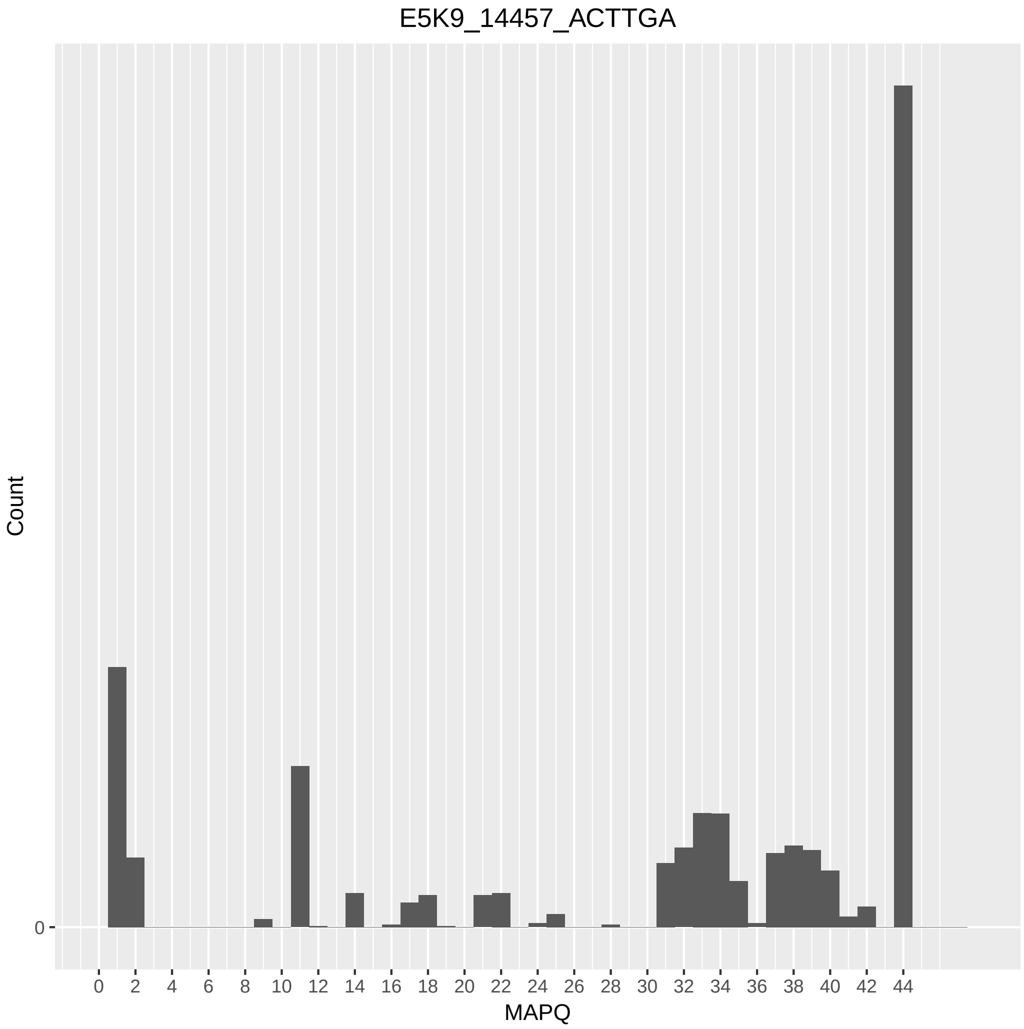
```
##
## FALSE  TRUE
## 14955 95045
```


A7K9_14456_TTAGGC
Count
30000
25000
20000
15000
10000
5000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

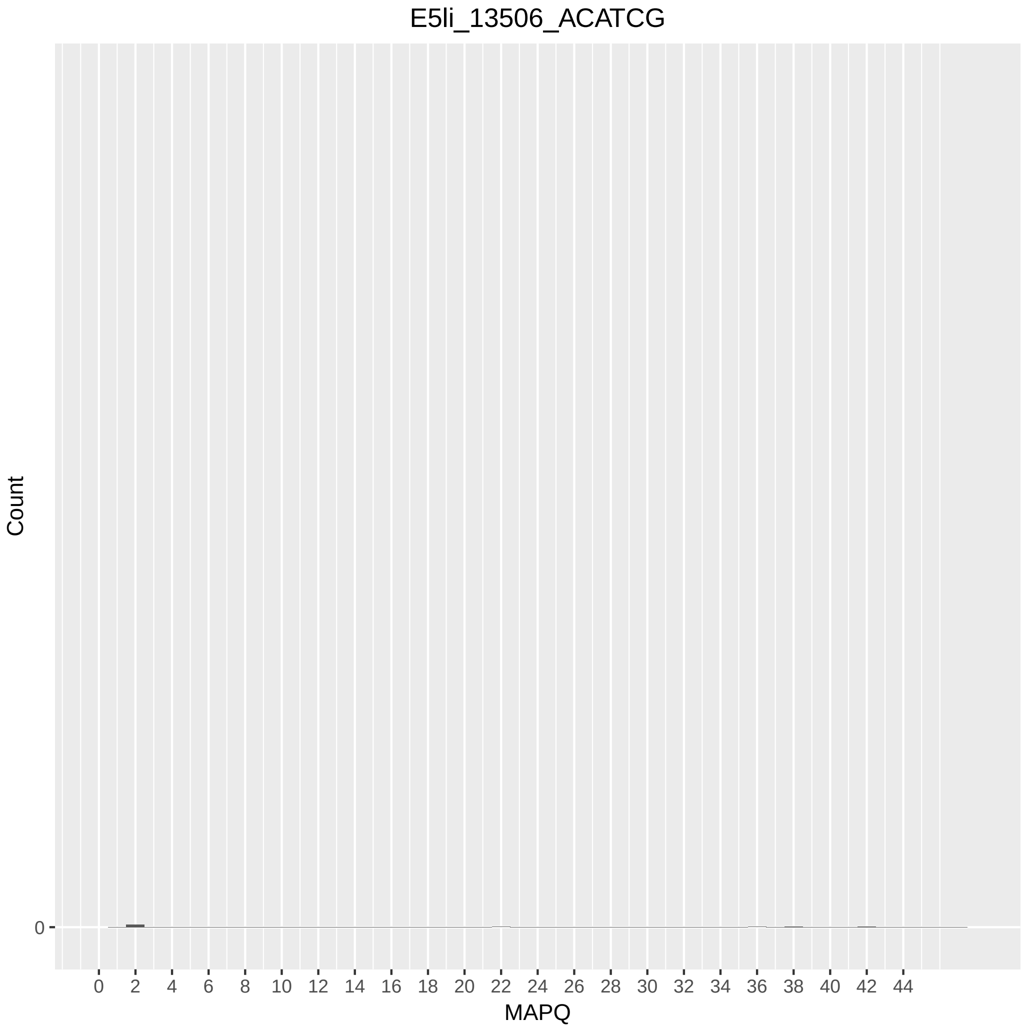
```
##
## FALSE  TRUE
## 31263 78737
```


A7li_13505_ATCACG
Count
65000
60000
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

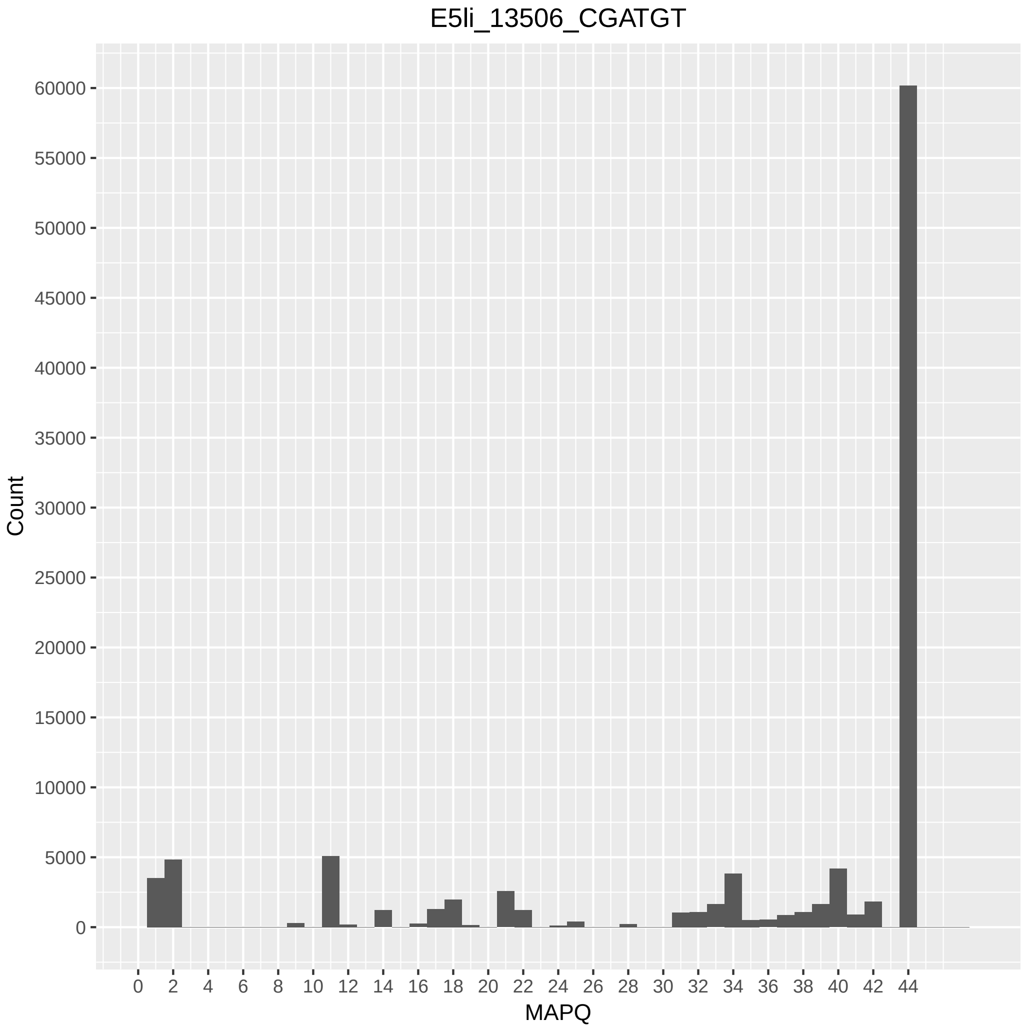
```
##
## FALSE  TRUE
## 11487 98513
```


E5K9_14457_ACTTGA
Count
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

```
##
## FALSE  TRUE
##  4521  9328
```


E5li_13506_ACATCG
Count
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

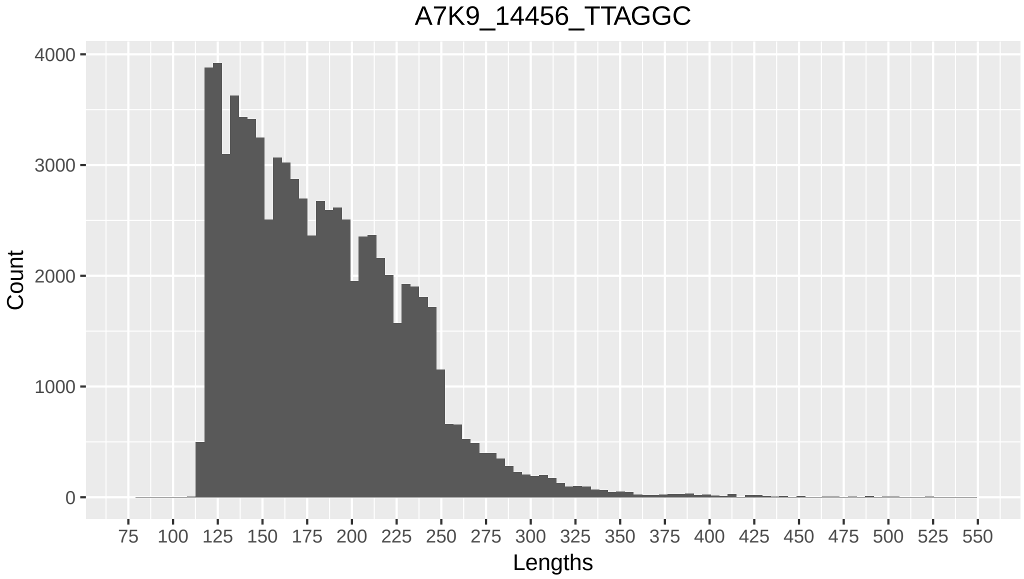
```
##
## FALSE  TRUE
##  2264     6
```


E5li_13506_CGATGT
Count
60000
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
MAPQ

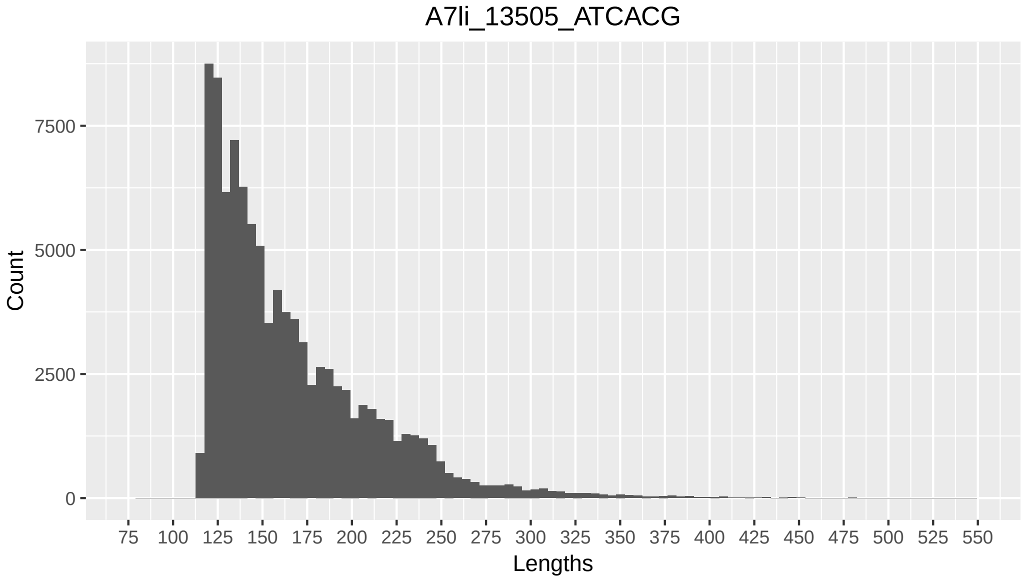
```
##
## FALSE  TRUE
## 15614 94386
```

2.2 Fragment Length

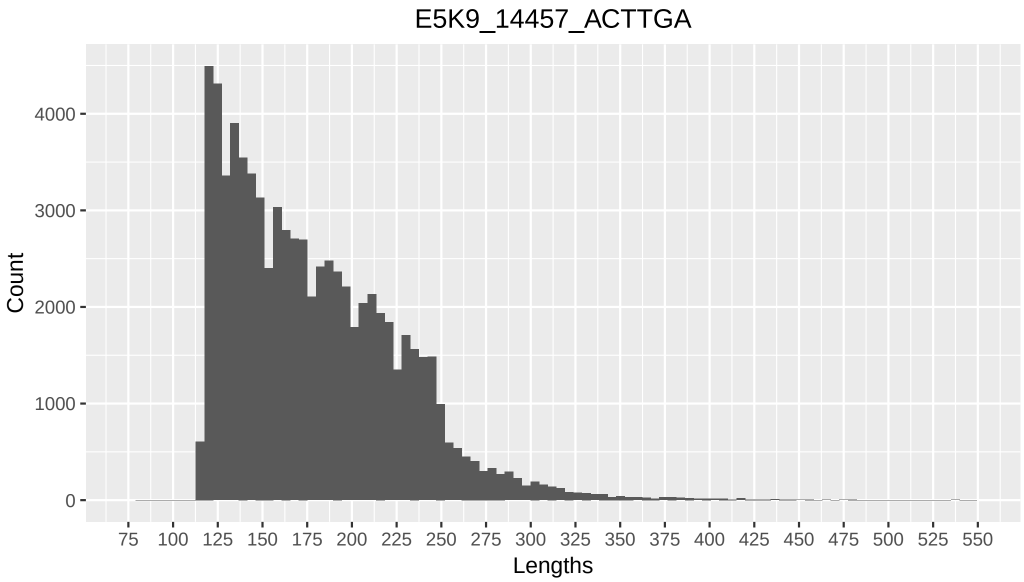
```
for (dir in dirs){
  for (x in file_names){
    lens <- read.csv2(file = paste0("/home/lucas/ISGlobal/TestSet/align_tests/",
    abs_lens <- abs(as.numeric(lens))
    for (element in abs_lens){
      if (element > 3000){abs_lens[element] <- 0}
    }
    df <- as.data.frame(abs_lens[abs_lens != 0])
    colnames(df) <- "len"
    title <- x
    print(ggplot(df, aes(x = len)) +
        geom_histogram(bins = 100) +
        labs(x = "Lengths", y = "Count") +
        ggtitle(title) +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_x_continuous(breaks = seq(75, 550, by = 25), limits = c(75,550)))
    plot.new()
  }
}

## Warning:  Removed 947157 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



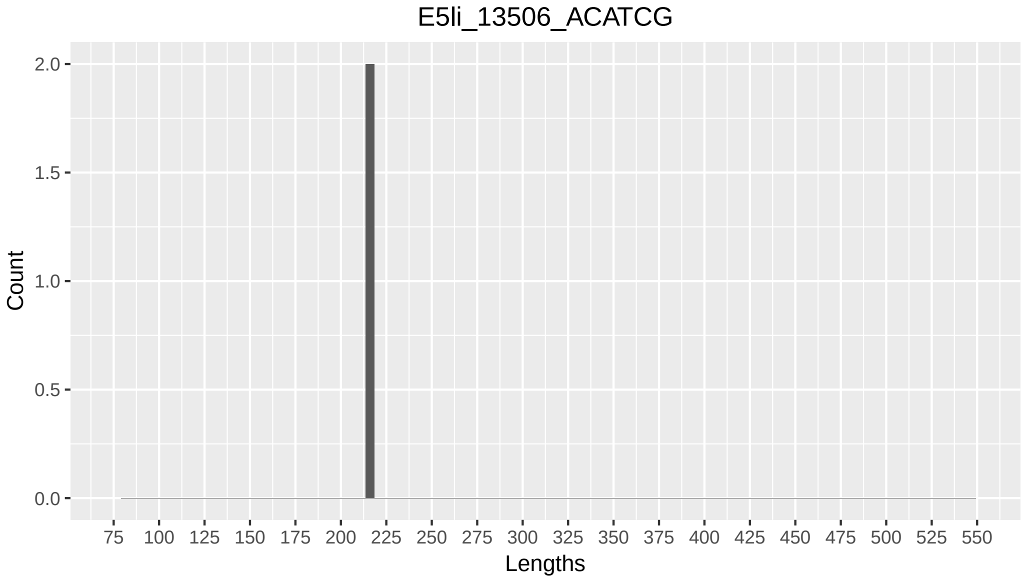
```
## Warning:  Removed 26 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



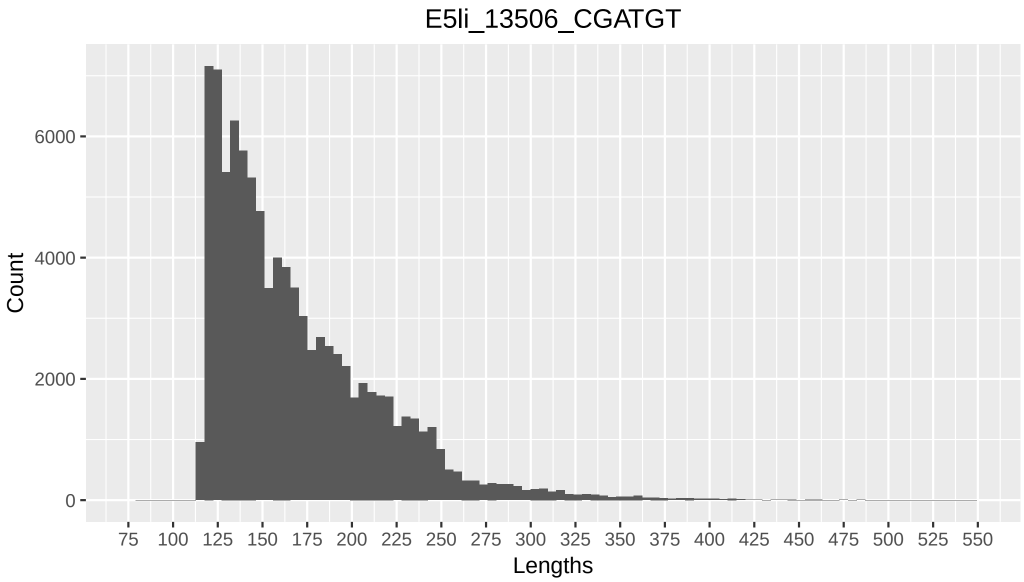
```
## Warning:  Removed 20 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



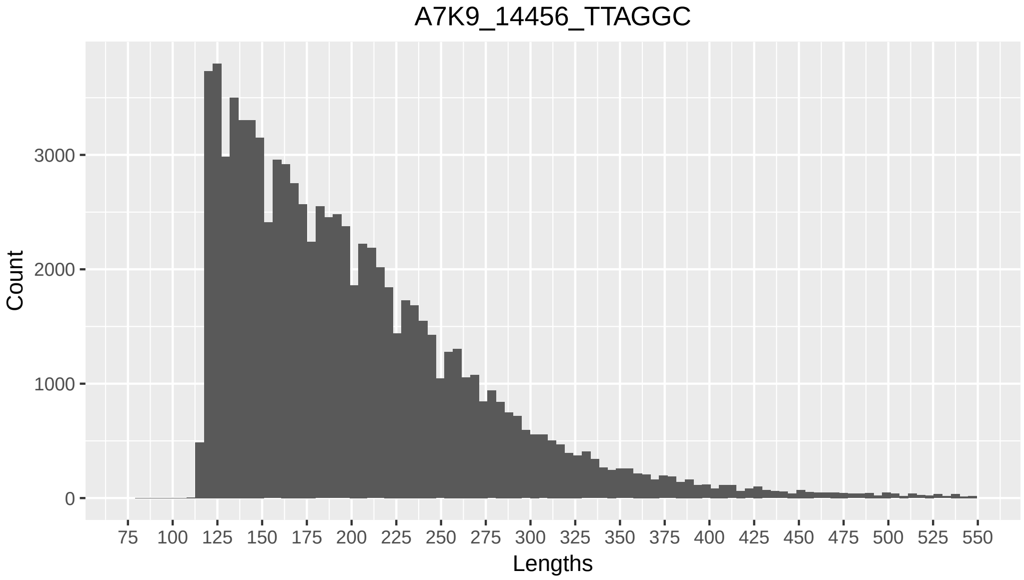
```
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



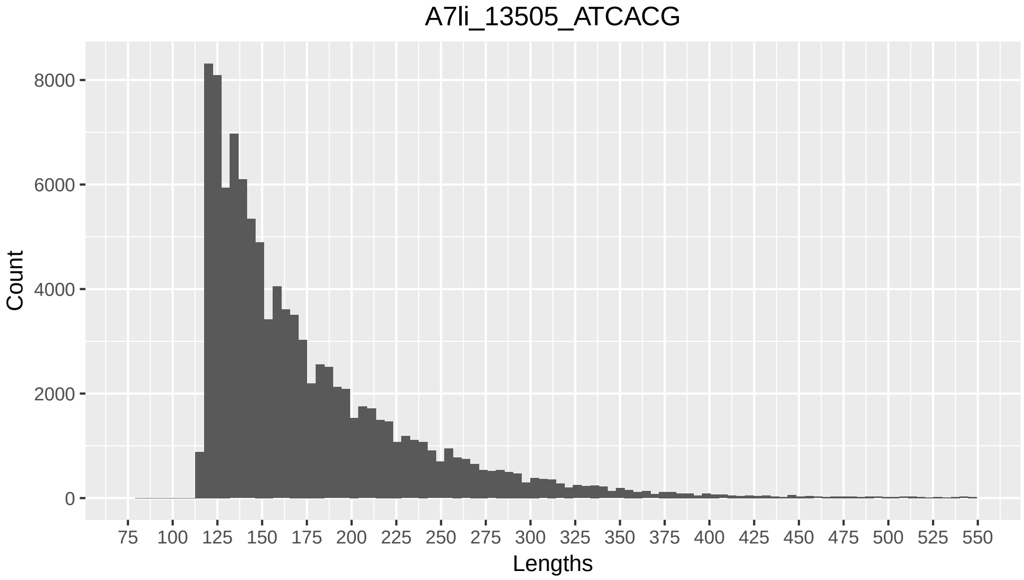
```
## Warning: Removed 1242026 rows containing non-finite values (stat_bin).
## Warning: Removed 1 rows containing missing values (geom_bar).
```



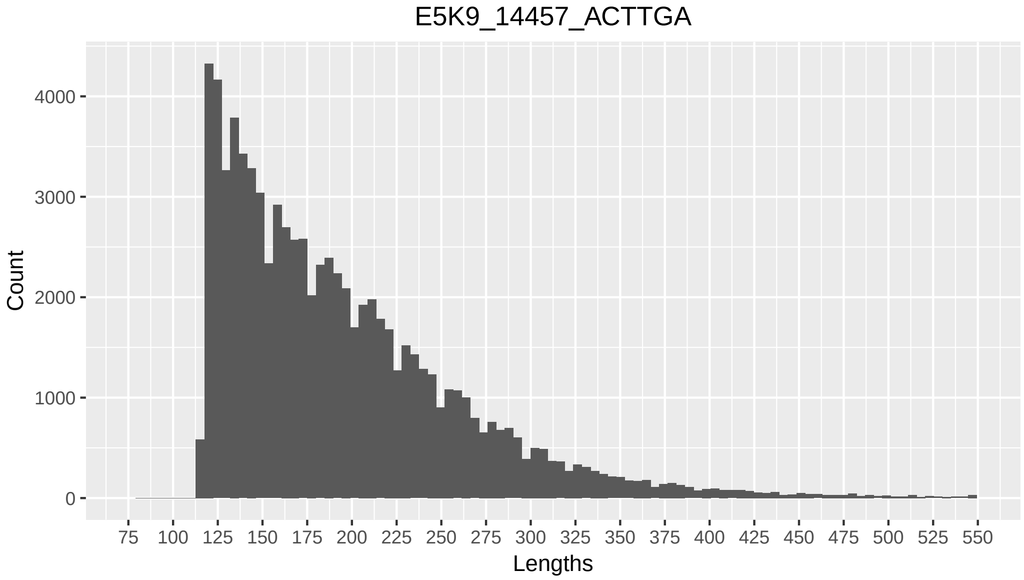
```
## Warning:  Removed 948769 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



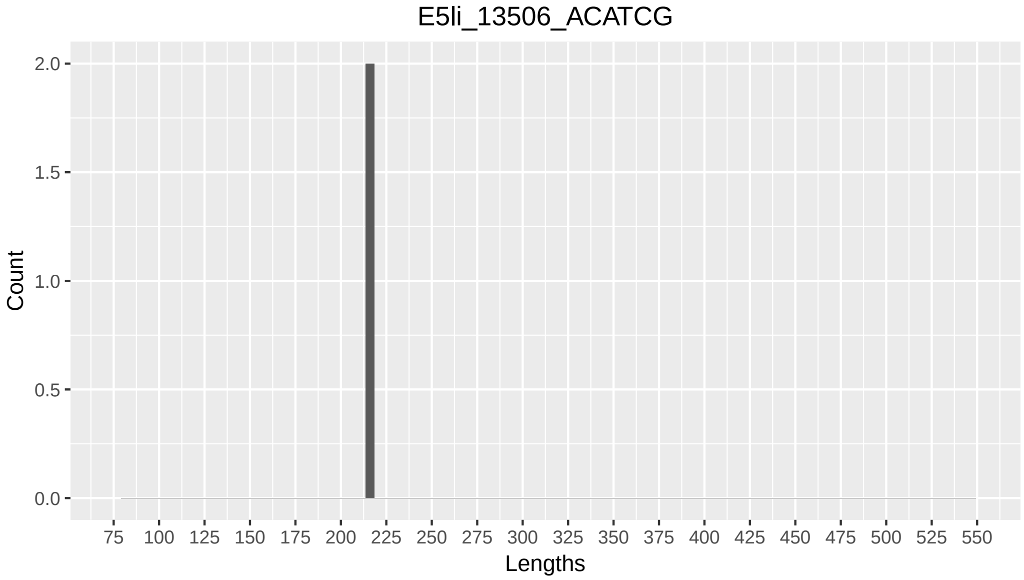
```
## Warning:  Removed 2966 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



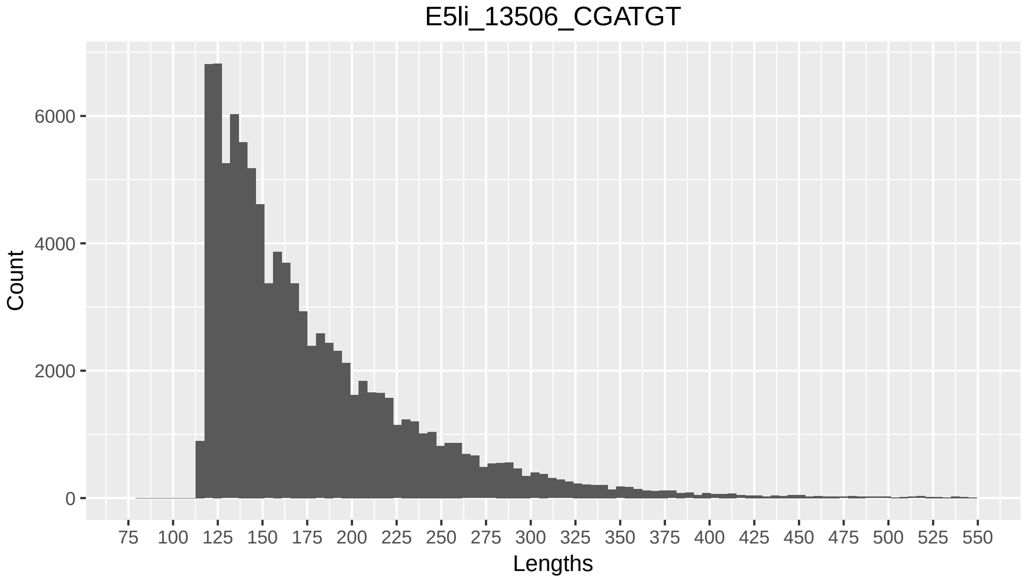
```
## Warning:  Removed 1542 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



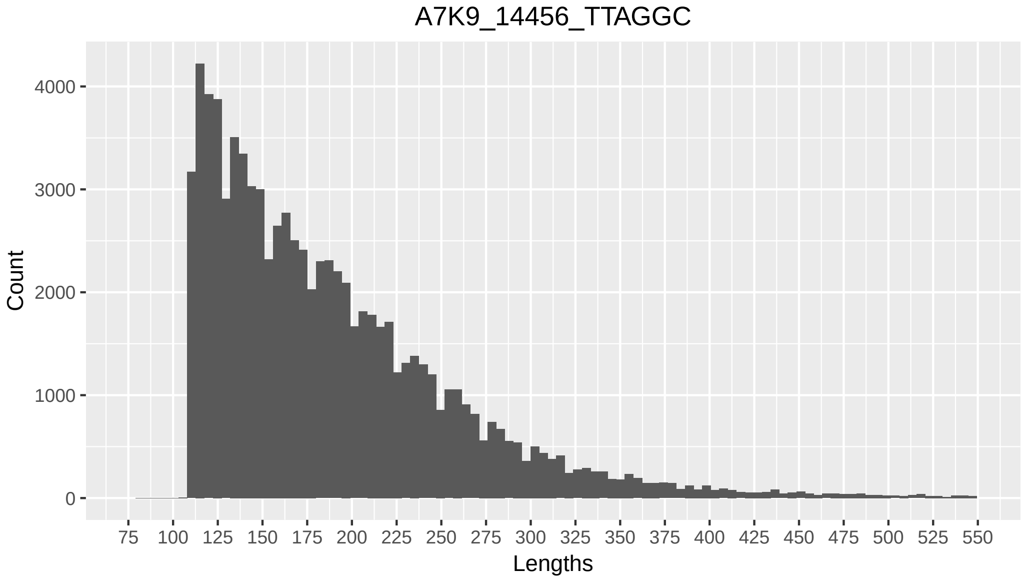
```
## Warning:  Removed 2 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



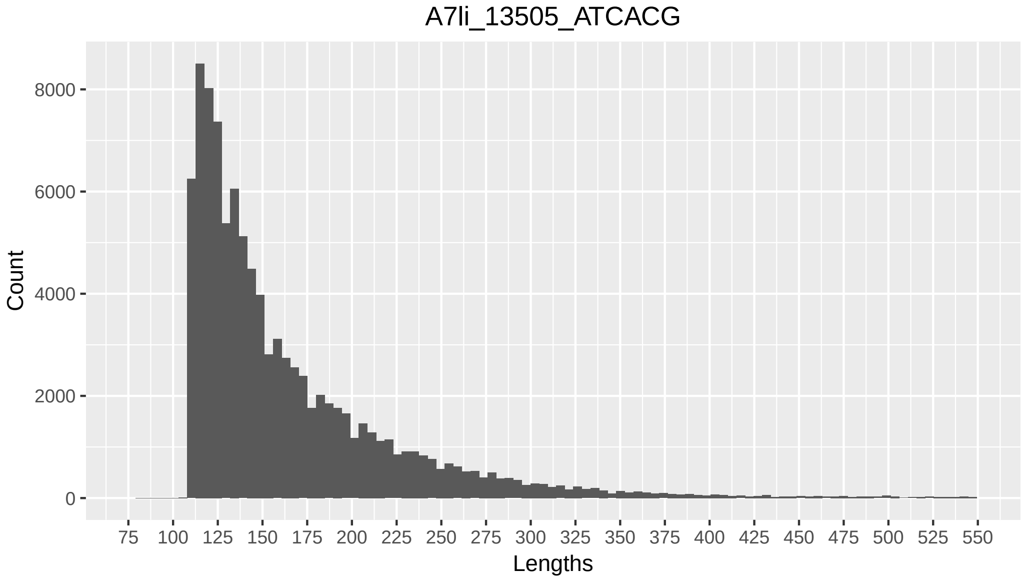
```
## Warning:  Removed 1244840 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



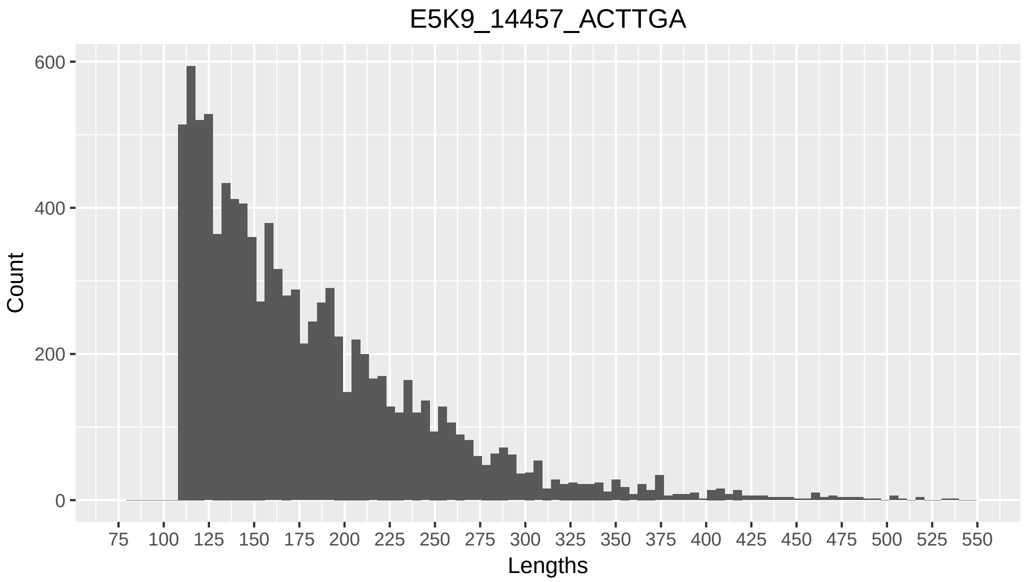
```
## Warning:  Removed 2216571 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



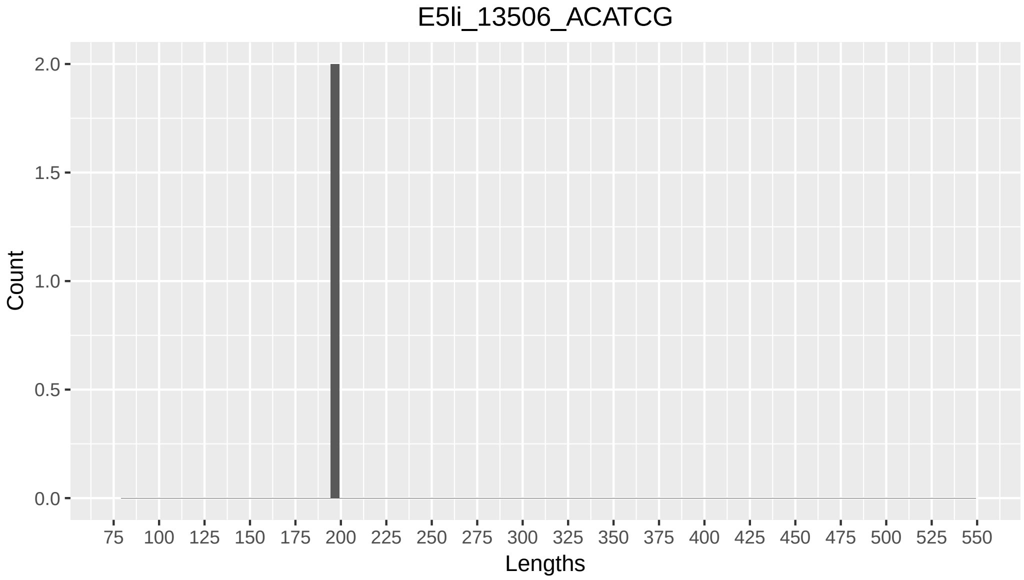
```
## Warning:  Removed 3892 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



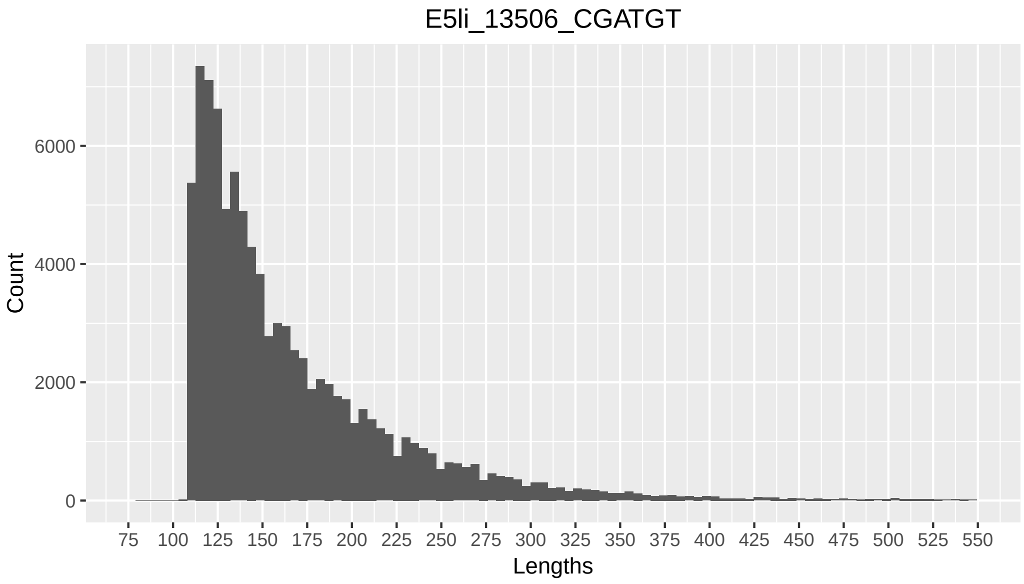
```
## Warning:  Removed 268 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



```
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



```
## Warning:  Removed 1888585 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
```



```
mapq <- read.csv2(file = "/home/lucas/ISGlobal/TestSet/align_tests/params_1/A7K9
df <- as.data.frame(as.numeric(mapq))
colnames(df) <- "MAPQ"
title <- "A7K9_14456_TTAGGC"
print(ggplot(df, aes(x = MAPQ)) +
        geom_histogram(binwidth = 1) +
        labs(x = "MAPQ", y = "Count") +
        ggtitle(title) +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_x_continuous(breaks = seq(0, 45, by = 2), limits = c(0,48)) +
        scale_y_continuous(breaks = seq(0,80000, by = 5000)))
```

And some text after.

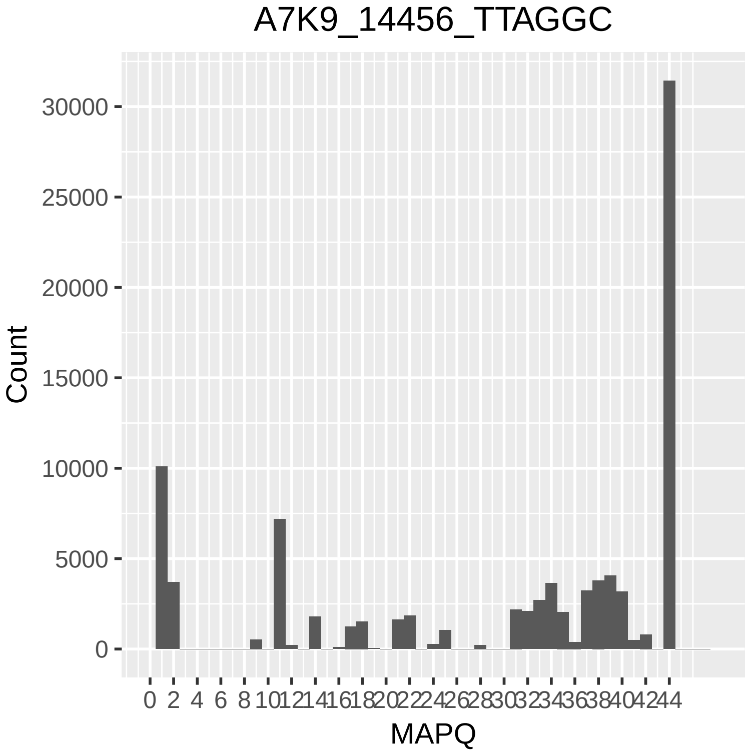


Figure 1: Tralala