Alignment Stats

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

Barcelona Global Health Institute (ISGlobal), Campus PRBB

March 13, 2017

Contents

1 Importing Data

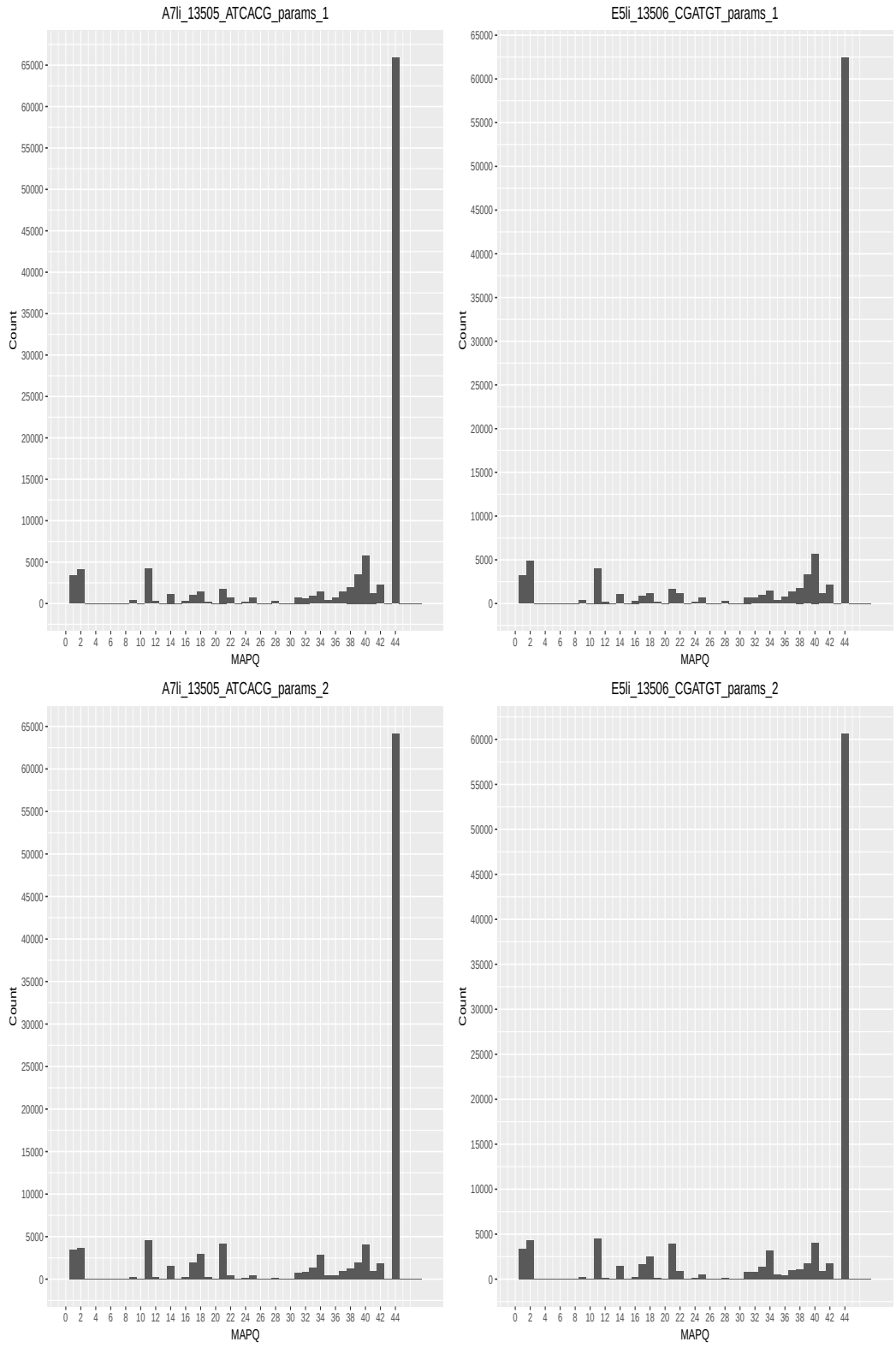
First we import the data for every alignment:

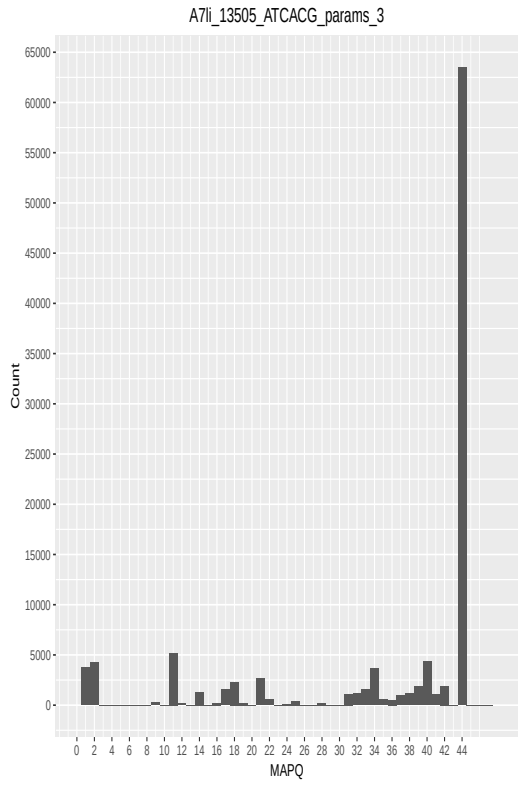
```
####  Fetch files and create iterables ####
file_names <- list.files("/home/lucas/ISGlobal/TestSet/align_tests2/inputs/")
file_names <- unique(lapply(file_names, function(x) substr(x,1,17)))

dirs <- list.files("/home/lucas/ISGlobal/TestSet/align_tests2/")[substr(list.fil
```

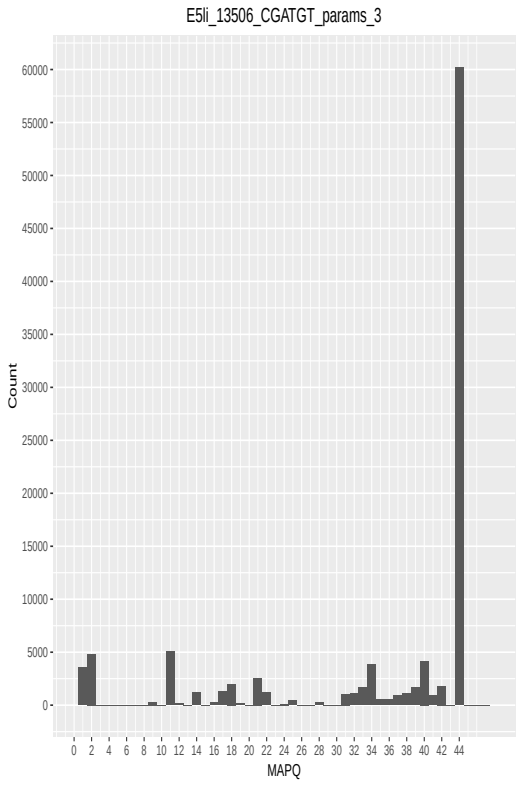
2 Plots

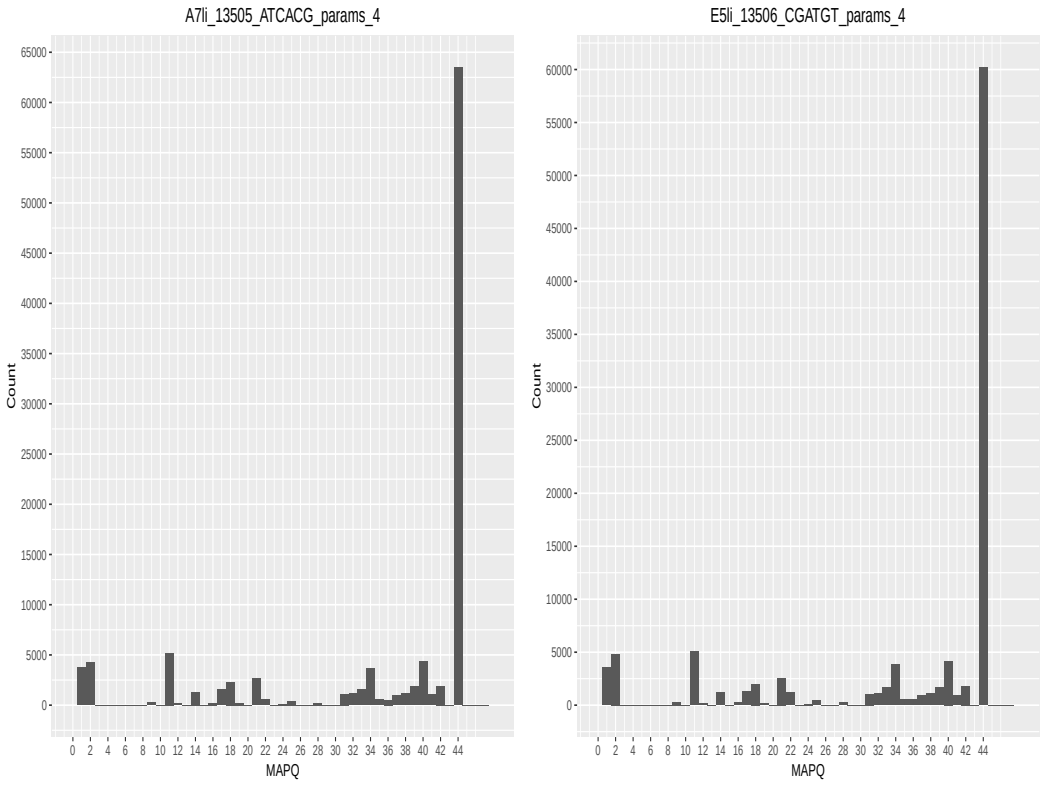
2.1 MAPQ

```
par(mfrow = c(1, 2))
for (dir in dirs){
  for (x in file_names){
    mapq <- read.csv2(file = paste0("/home/lucas/ISGlobal/TestSet/align_tests2/"
    df <- as.data.frame(as.numeric(mapq))
    colnames(df) <- "MAPQ"
    title <- paste0(x,"_",dir)
    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()
  }
}
```


A7li_13505_ATCACG_params_1
E5li_13506_CGATGT_params_1
A7li_13505_ATCACG_params_2
E5li_13506_CGATGT_params_2
Count
MAPQ


A7li_13505_ATCACG_params_3
Count
MAPQ

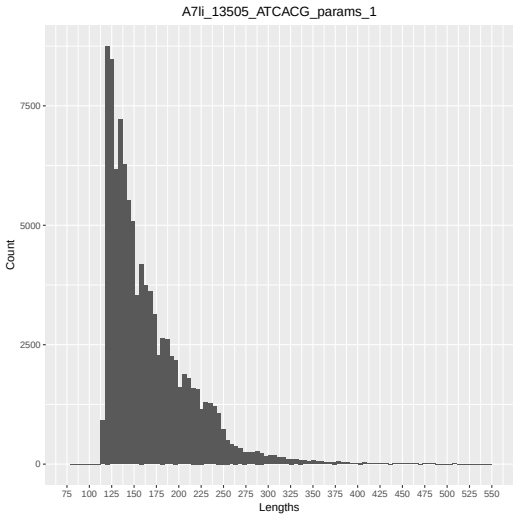

E5li_13506_CGATGT_params_3
Count
MAPQ

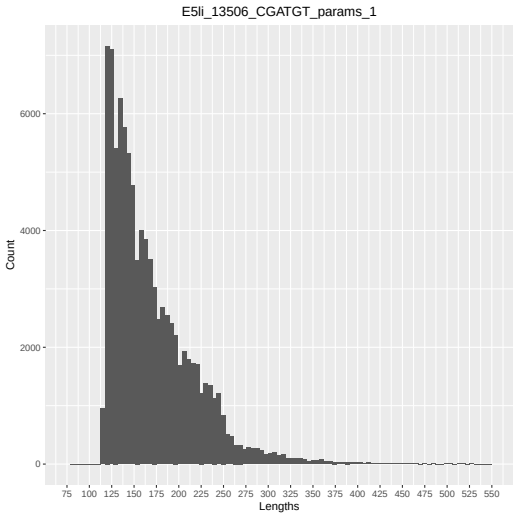

A7li_13505_ATCACG_params_4
E5li_13506_CGATGT_params_4
Count
MAPQ

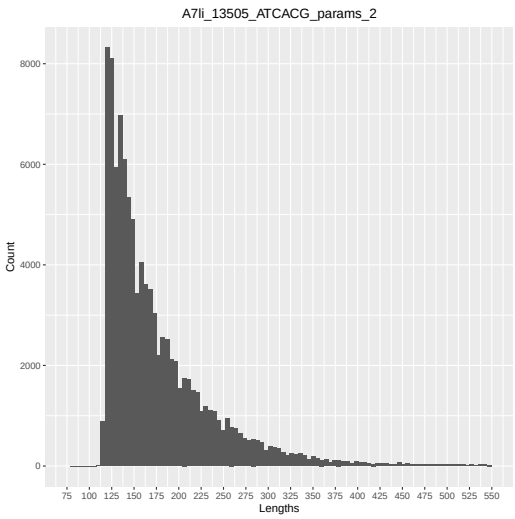
2.2 Fragment Length

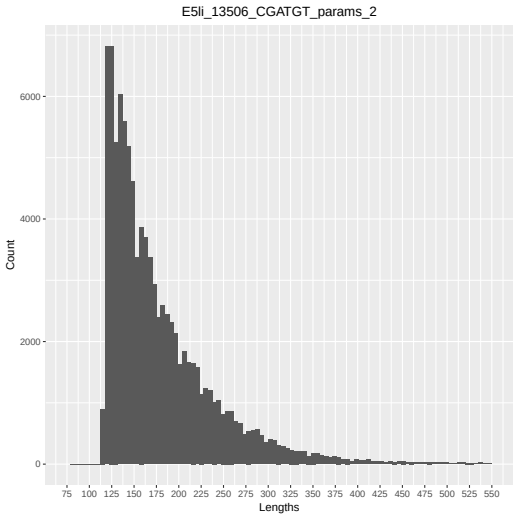
```
for (dir in dirs){
  for (x in file_names){
    lens <- read.csv2(file = paste0("/home/lucas/ISGlobal/TestSet/align_tests2/"
    abs_lens <- abs(as.numeric(lens))
    for (element in abs_lens){
      if (element > 3000){abs_lens[element] <- 0}
    }
    df <- as.data.frame(abs_lens[abs_lens != 0])
    colnames(df) <- "len"
    title <- paste0(x,"_",dir)
    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()
  }
}
```

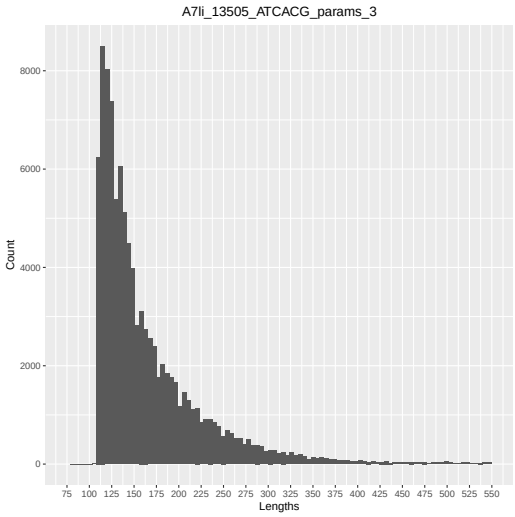
```
## Warning:  Removed 26 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
## Warning:  Removed 1242026 rows containing non-finite values (stat_bin).
## Warning:  Removed 1 rows containing missing values (geom_bar).
## Warning:  Removed 2966 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 3892 rows containing non-finite values (stat_bin).
## 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).
## Warning:  Removed 3892 rows containing non-finite values (stat_bin).
## 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).
```

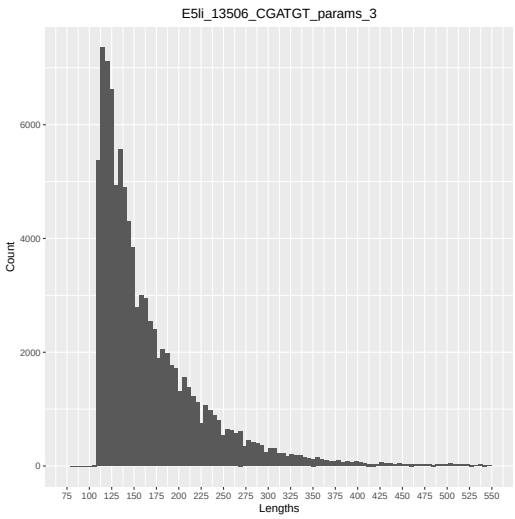

A7li_13505_ATCACG_params_1
Count
Lengths
0
2500
5000
7500
75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550

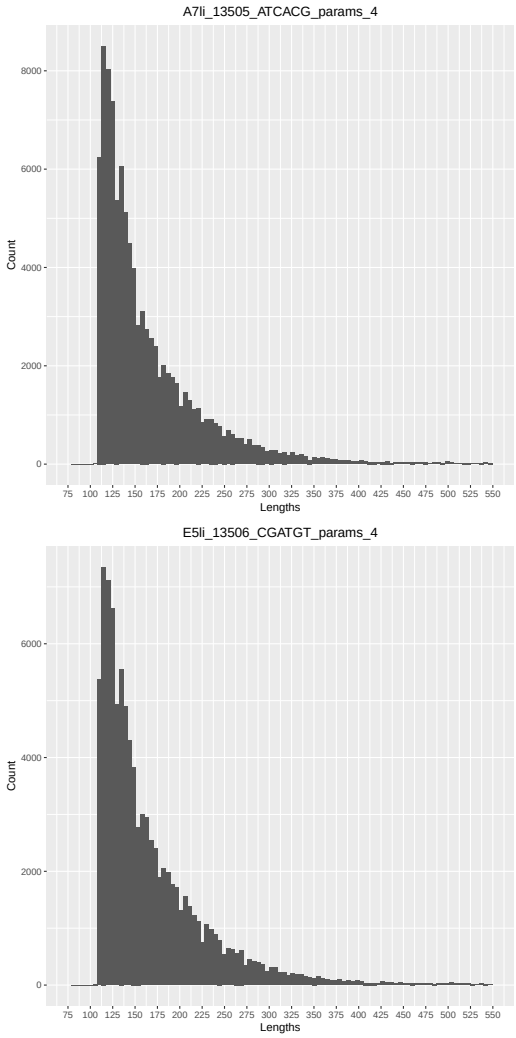

E5li_13506_CGATGT_params_1
Count
Lengths
0
2000
4000
6000
75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550


A7li_13505_ATCACG_params_2
Count
Lengths
0
2000
4000
6000
8000
75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550


E5li_13506_CGATGT_params_2
Count
Lengths
0
2000
4000
6000
75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550


A7li_13505_ATCACG_params_3
Count
Lengths
0
2000
4000
6000
8000
75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550


E5li_13506_CGATGT_params_3
Count
Lengths
0
2000
4000
6000
75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550



And some text after.