GC content correlation

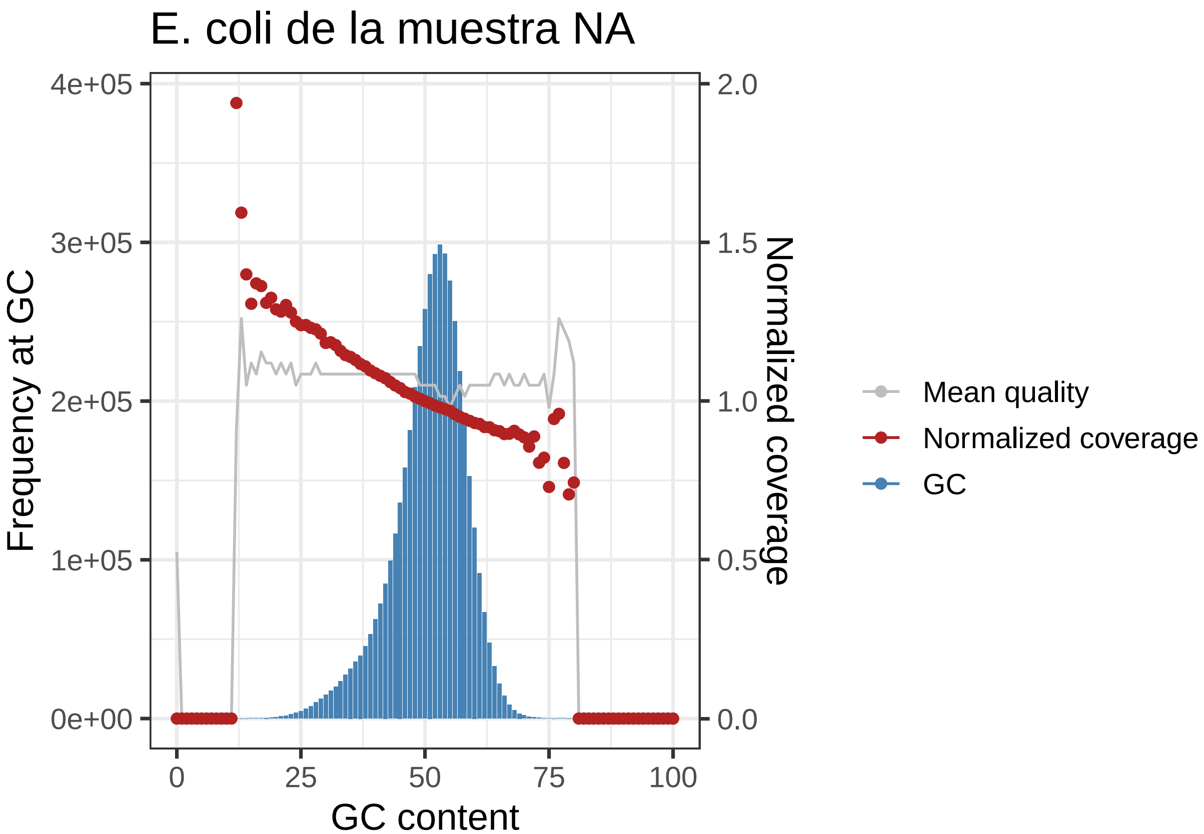
Modesto Redrejo Rodríguez

2022-02-21

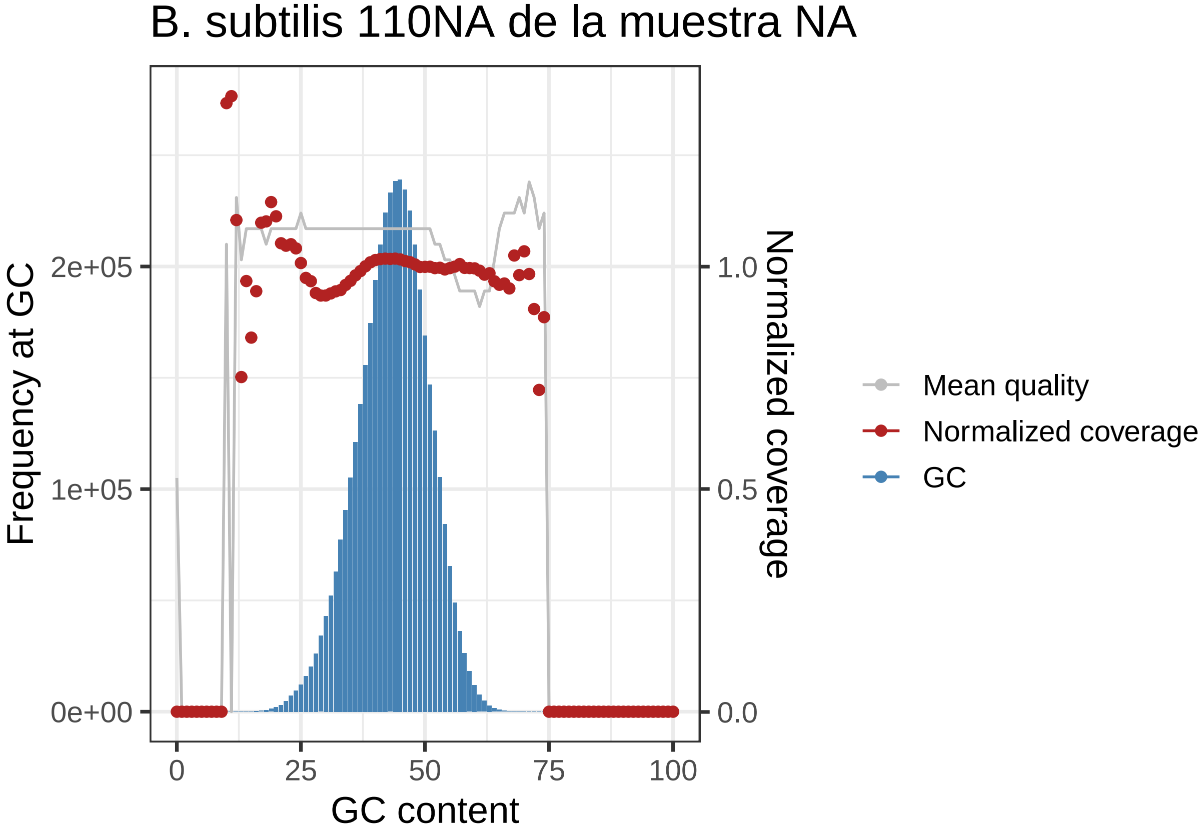
1. **Read files and plot data**

   GC content was analyzed with Picard (see Evernote ELN).

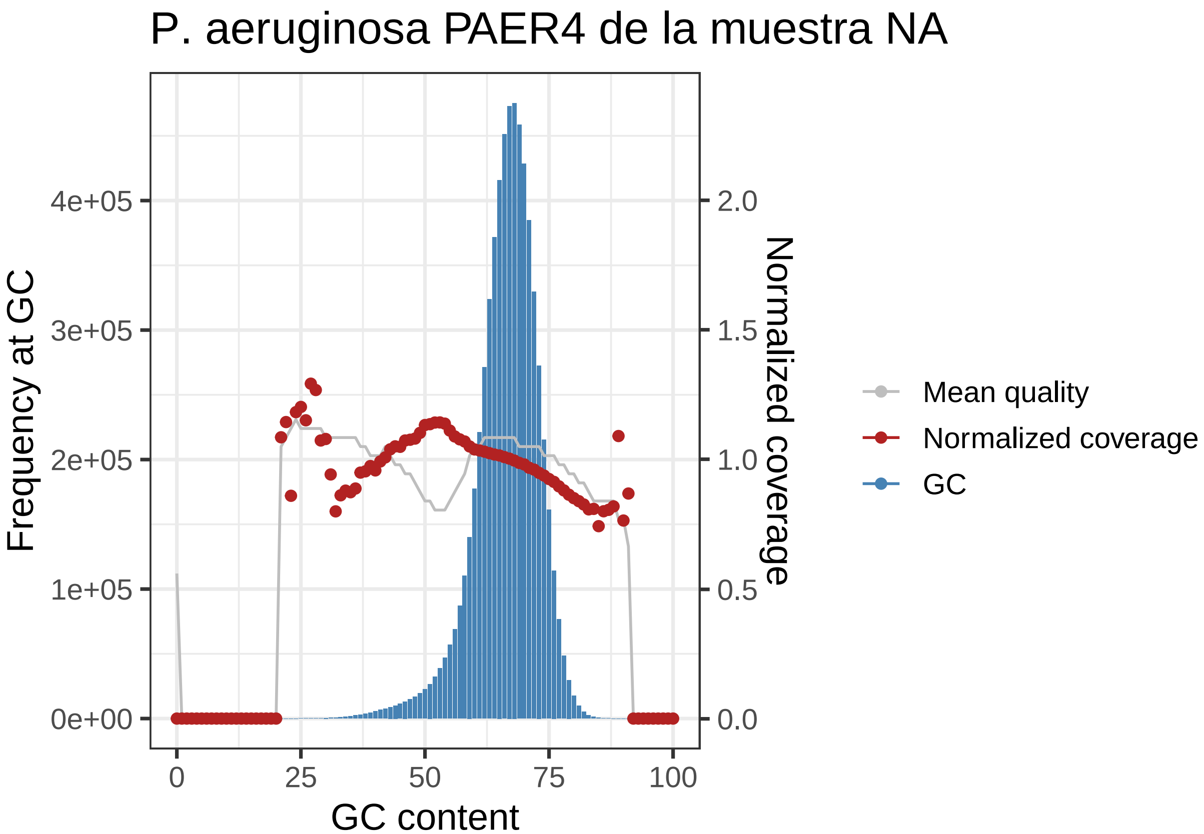
```
#load all the files as a list of dataframes
gc_picard=c("ctrl_gc_bias_coli", "ctrl_gc_bias_subtilis","ctrl_gc_bias_PAE","ctrl_gc_bias_kocuria",
            "ctrl2_gc_bias_coli", "ctrl2_gc_bias_subtilis","ctrl2_gc_bias_PAE","ctrl2_gc_bias_kocuria",
            "N4_gc_bias_coli", "N4_gc_bias_subtilis","N4_gc_bias_PAE","N4_gc_bias_kocuria",
            "D3_gc_bias_coli", "D3_gc_bias_subtilis","D3_gc_bias_PAE","D3_gc_bias_kocuria",
            "C3_gc_bias_coli", "C3_gc_bias_subtilis","C3_gc_bias_PAE","C3_gc_bias_kocuria",
            "N1_gc_bias_coli", "N1_gc_bias_subtilis","N1_gc_bias_PAE","N1_gc_bias_kocuria",
            "F2_gc_bias_coli", "F2_gc_bias_subtilis","F2_gc_bias_PAE","F2_gc_bias_kocuria",
            "N2_gc_bias_coli", "N2_gc_bias_subtilis","N2_gc_bias_PAE","N2_gc_bias_kocuria",
            "B2_gc_bias_coli", "B2_gc_bias_subtilis","B2_gc_bias_PAE","B2_gc_bias_kocuria",
            "N6_gc_bias_coli", "N6_gc_bias_subtilis","N6_gc_bias_PAE","N6_gc_bias_kocuria",
            "A2_gc_bias_coli", "A2_gc_bias_subtilis","A2_gc_bias_PAE","A2_gc_bias_kocuria")
gc <- lapply(gc_picard, function(x) read.csv2(paste(x,".txt",sep=""), skip=6,header=TRUE,sep="\t", colC

#plot all the samples using a loop
library(ggplot2)
library(ggpubr)
cor_matrix <- data.frame(44,3)
colors <- c("Mean quality"="grey","Normalized coverage"="firebrick","GC"="steelblue")
samples <- c("NA","NA2","RepliG","RepliG2","TruePrime","piPolB","piPolB+D","piMDA","piMDA2","piMDA+D","
templates <- c("E. coli","B. subtilis 110NA", "P. aeruginosa PAER4","K. rhizophila")
genomas <- merge(templates, samples ,all=TRUE)
plot_list <- list()
for (i in 1:length(gc)){
  p <- ggplot(data=gc[[i]],aes(x=gc[[i]]$GC,y=gc[[i]]$WINDOWS)) +
    geom_bar(stat="identity",fill="steelblue")+
    geom_line(aes(y=gc[[i]]$MEAN_BASE_QUALITY*7000,color="Mean quality"))+
    geom_point(aes(y=gc[[i]]$NORMALIZED_COVERAGE*200000,color="Normalized coverage"))+
    scale_y_continuous("Frequency at GC", sec.axis=sec_axis(~./200000,name="Normalized coverage"))+
    scale_color_manual(name="",values=colors)+
    xlab("GC content") +
    ggtitle(paste(genomas[i,1],"de la muestra",genomas[i,2])) +
    theme_bw(base_size=14)
  print(p)
}
```

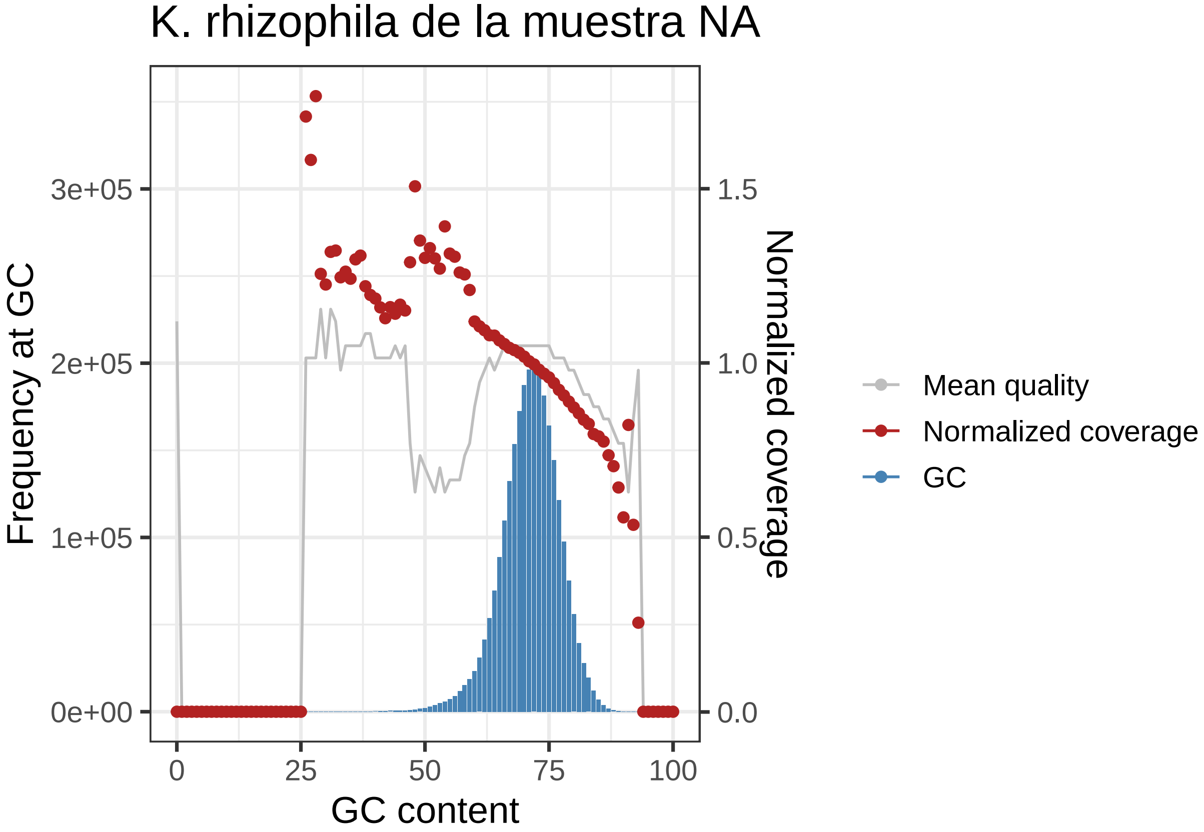


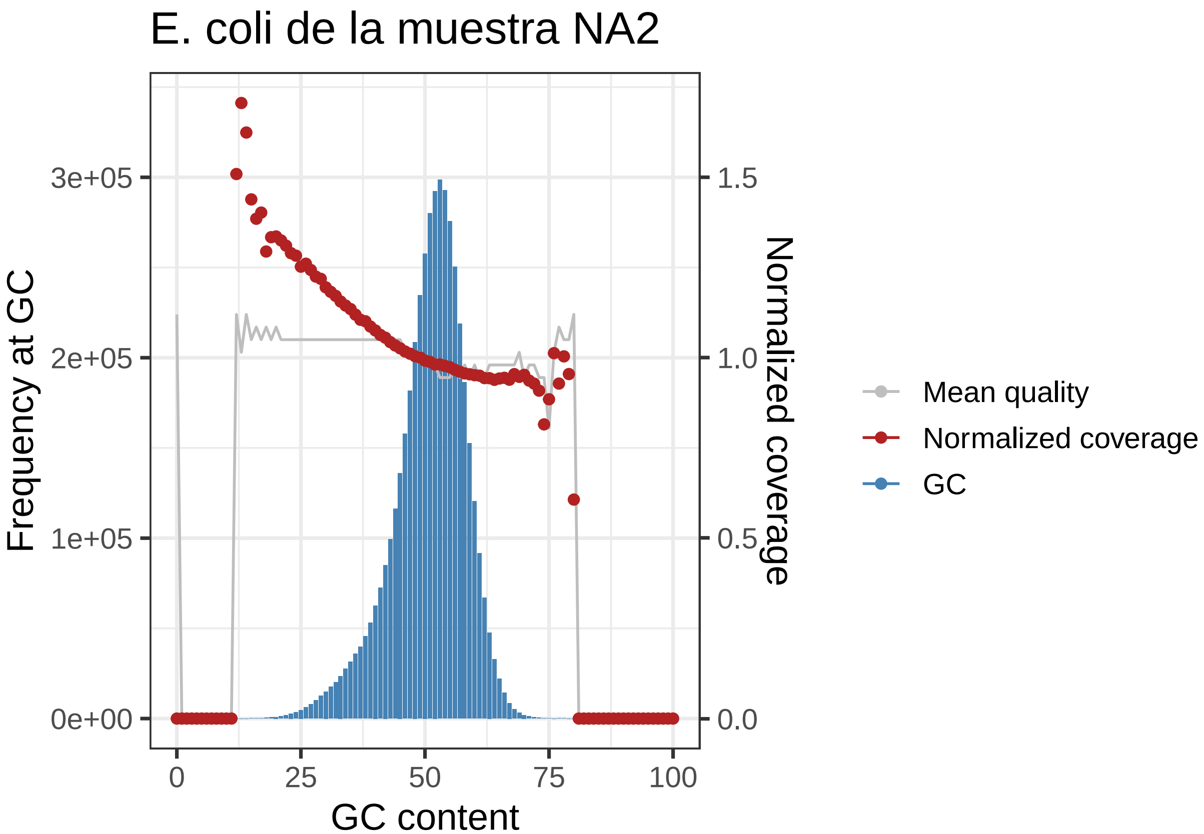
E. coli de la muestra NA

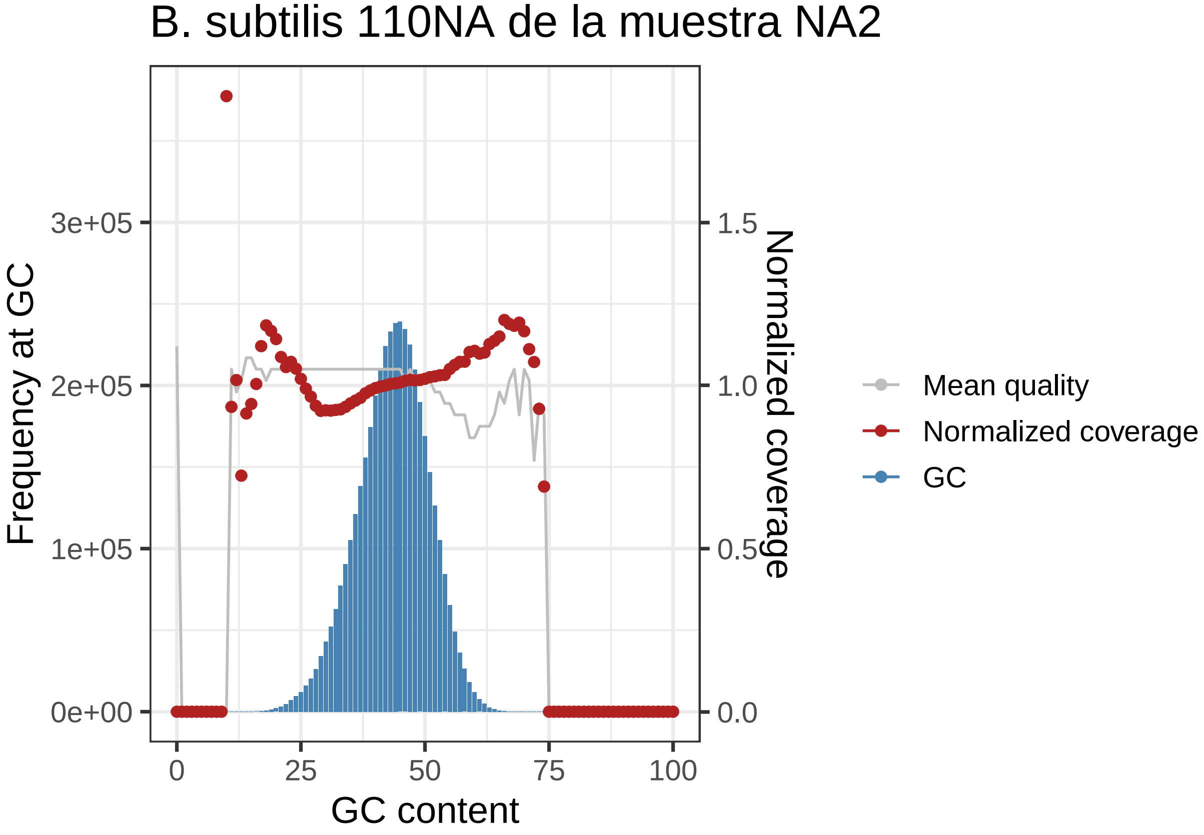


B. subtilis 110NA de la muestra NA

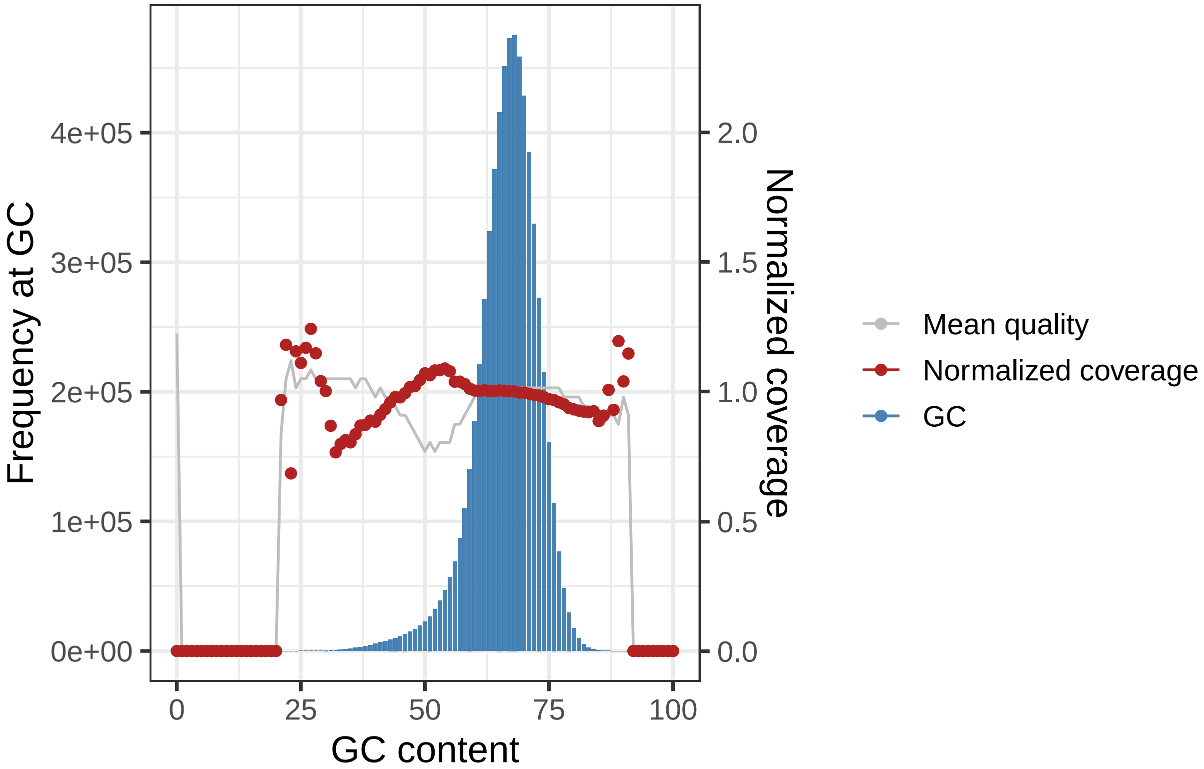

P. aeruginosa PAER4 de la muestra NA
Frequency at GC
0e+00
1e+05
2e+05
3e+05
4e+05
Normalized coverage
0.0
0.5
1.0
1.5
2.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC


K. rhizophila de la muestra NA
Frequency at GC
0e+00
1e+05
2e+05
3e+05
Normalized coverage
0.0
0.5
1.0
1.5
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC

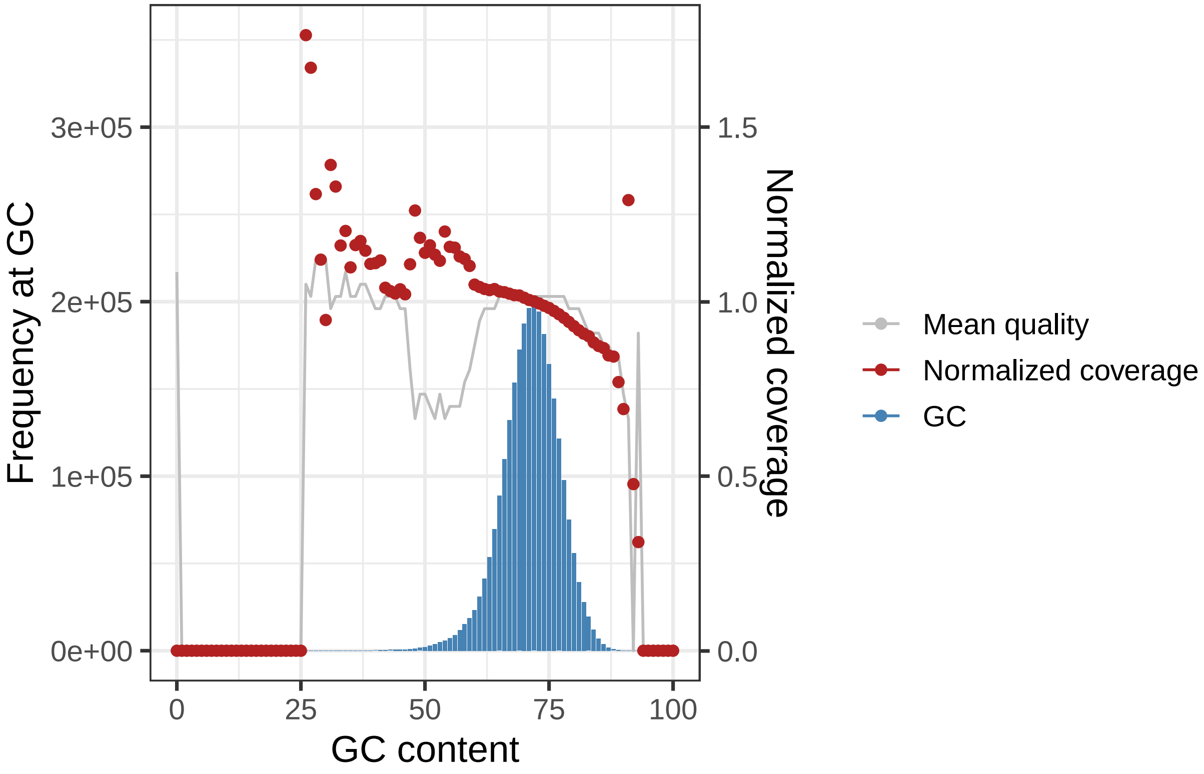

E. coli de la muestra NA2
Frequency at GC
3e+05
2e+05
1e+05
0e+00
Normalized coverage
1.5
1.0
0.5
0.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC


B. subtilis 110NA de la muestra NA2
Frequency at GC
3e+05
2e+05
1e+05
0e+00
Normalized coverage
1.5
1.0
0.5
0.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC

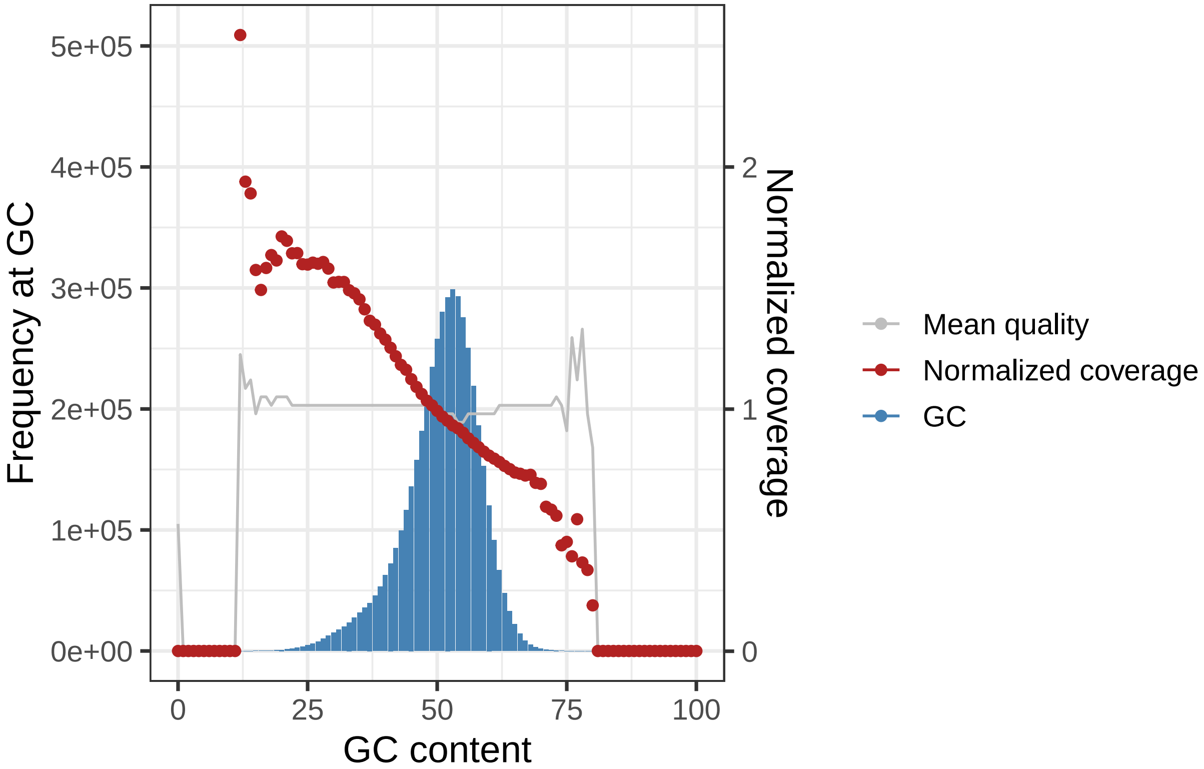
P. aeruginosa PAER4 de la muestra NA2



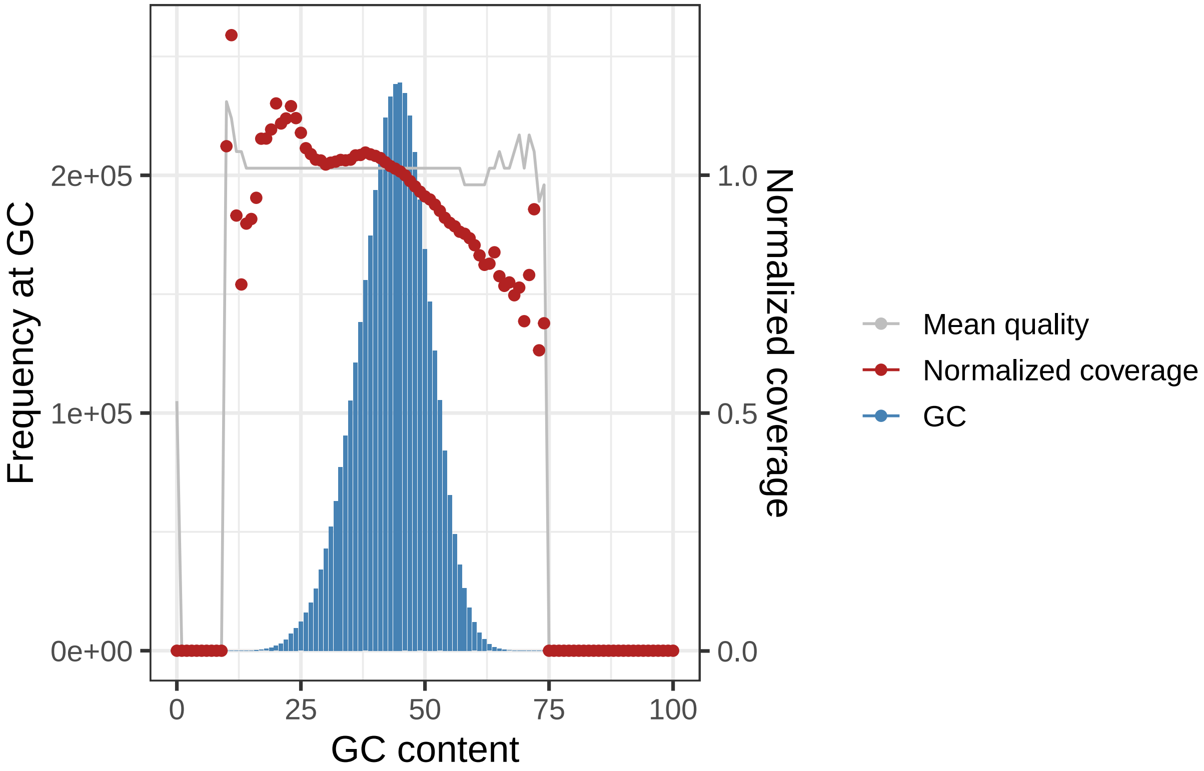
K. rhizophila de la muestra NA2

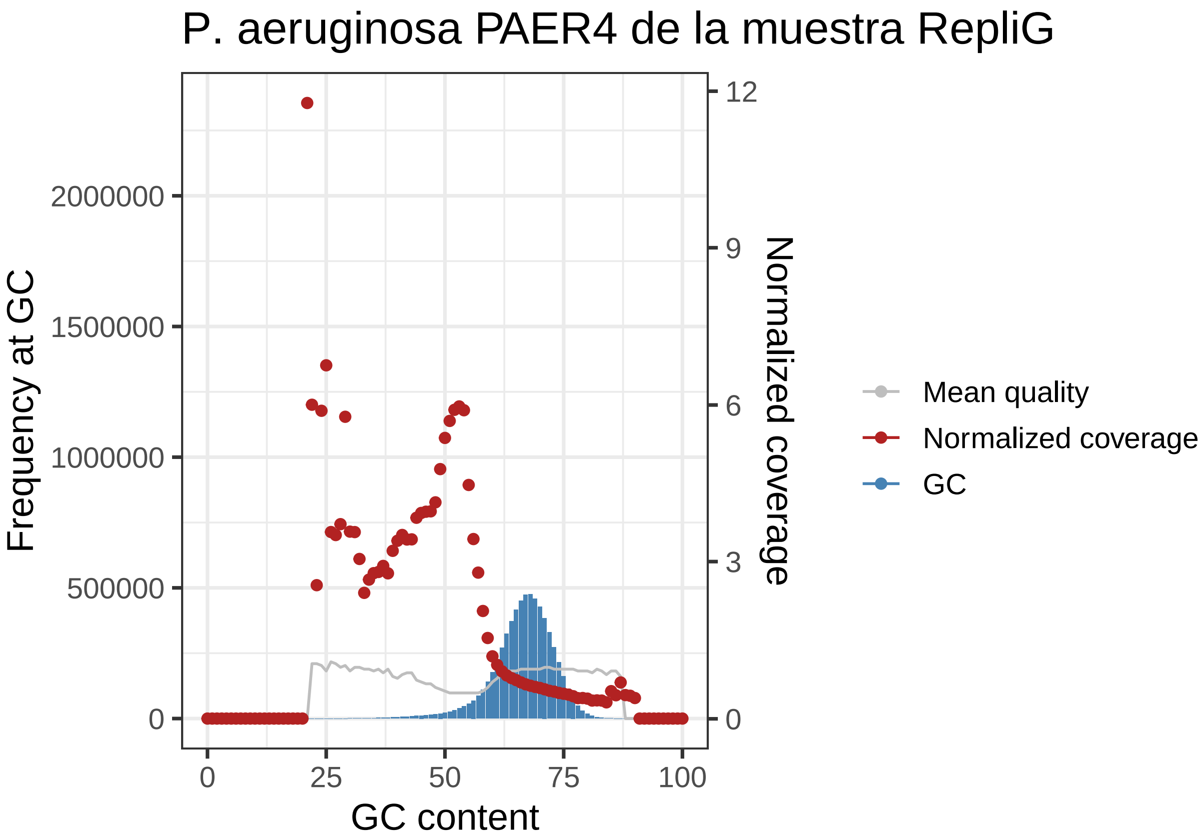


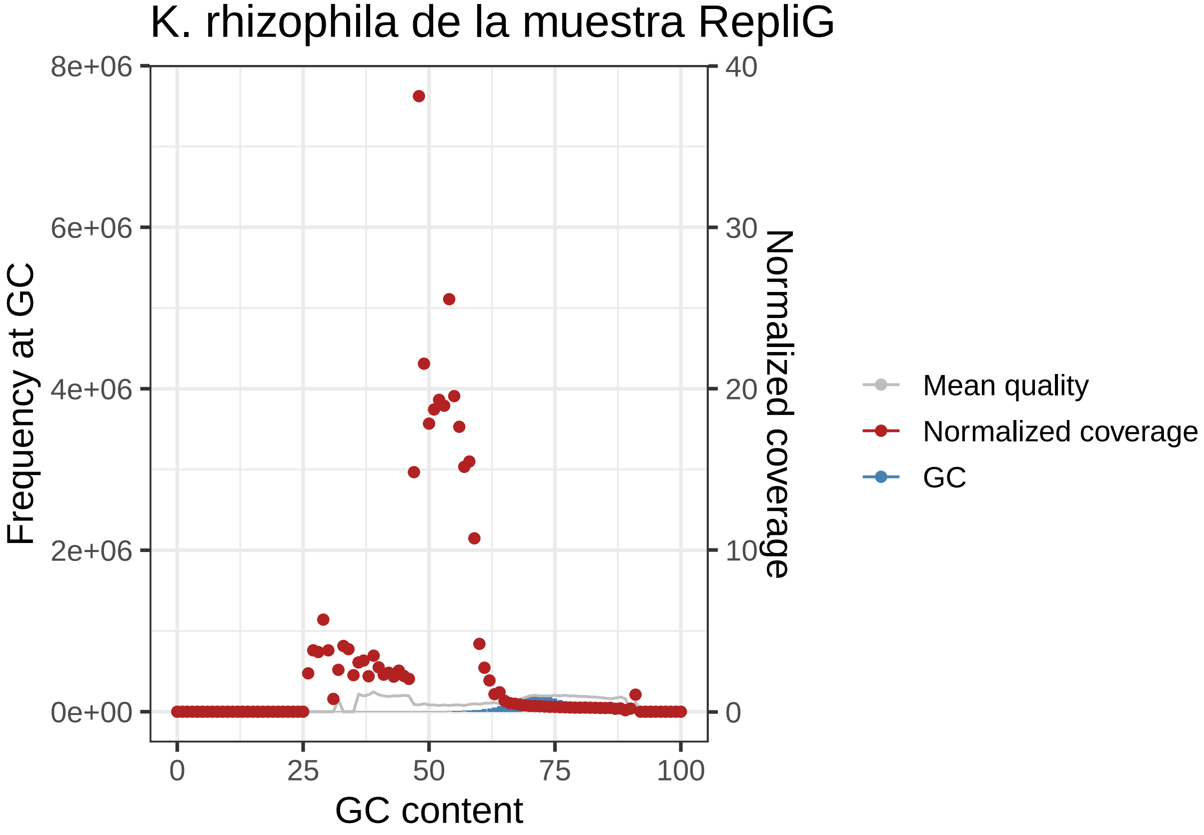
E. coli de la muestra RepliG

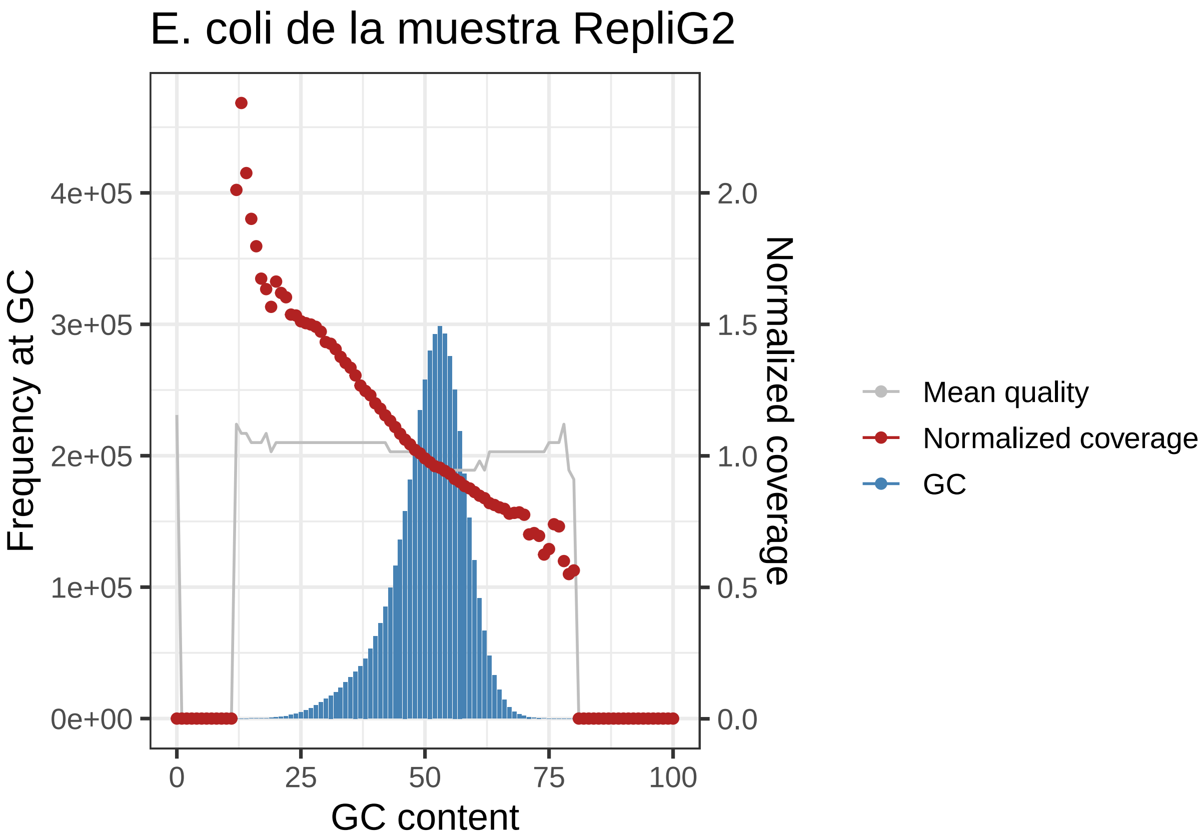


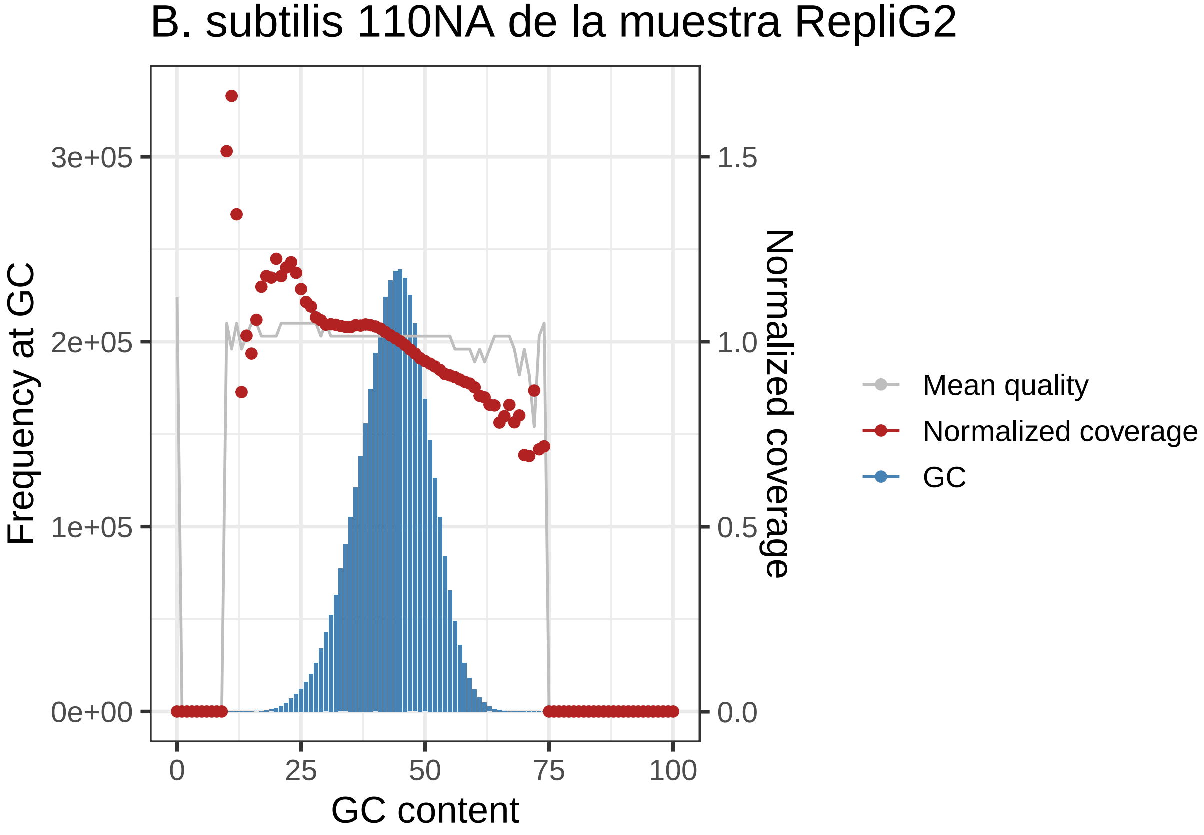
B. subtilis 110NA de la muestra RepliG

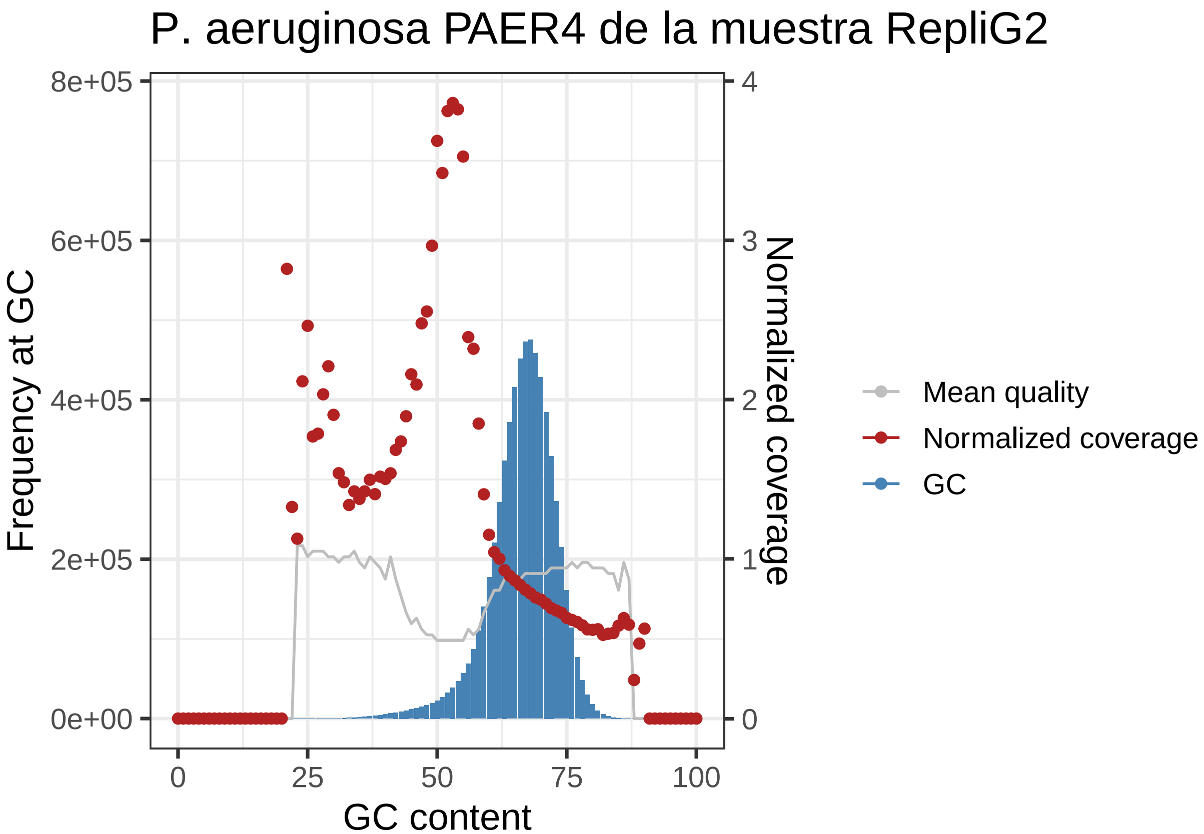


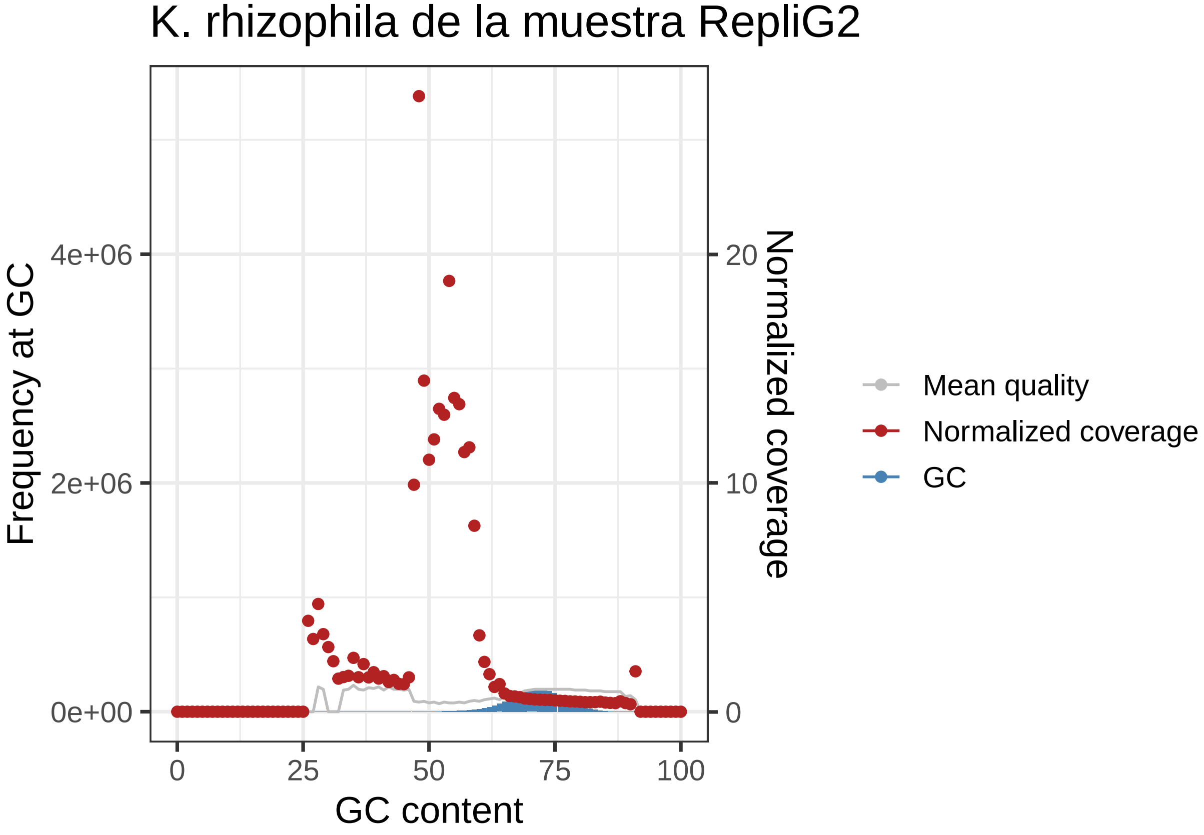

P. aeruginosa PAER4 de la muestra RepliG

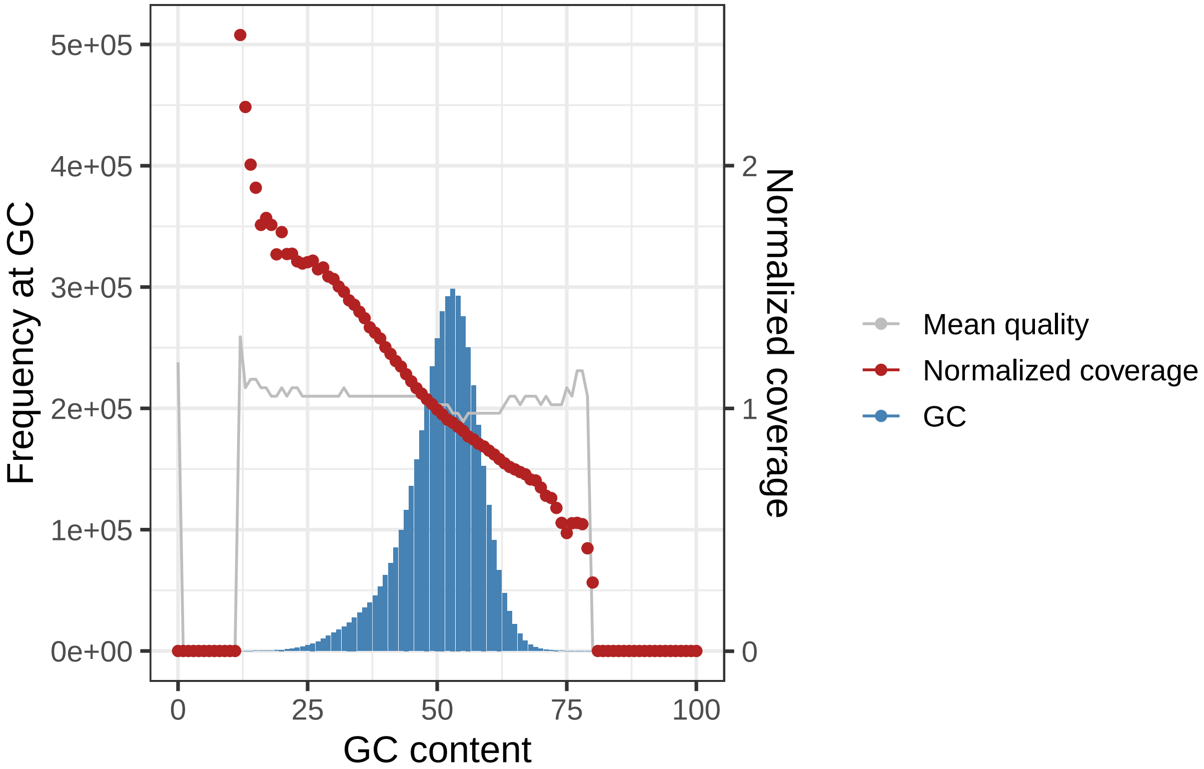

K. rhizophila de la muestra RepliG

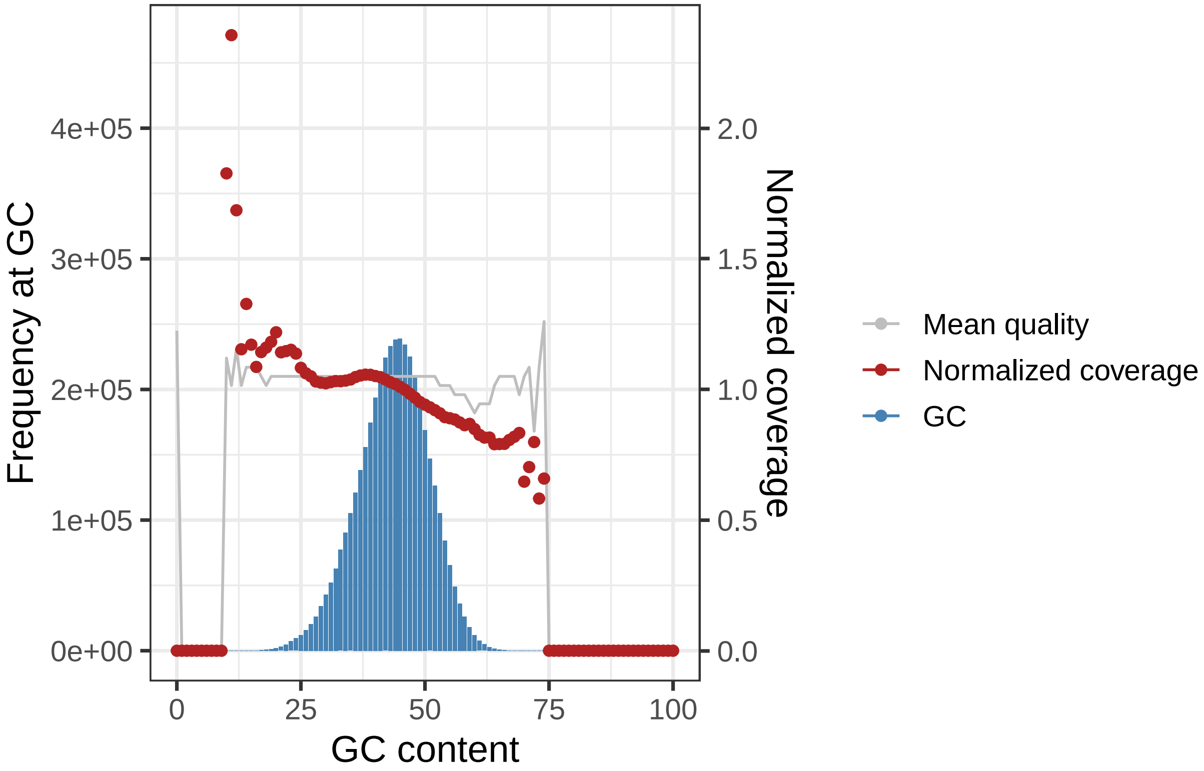

E. coli de la muestra RepliG2

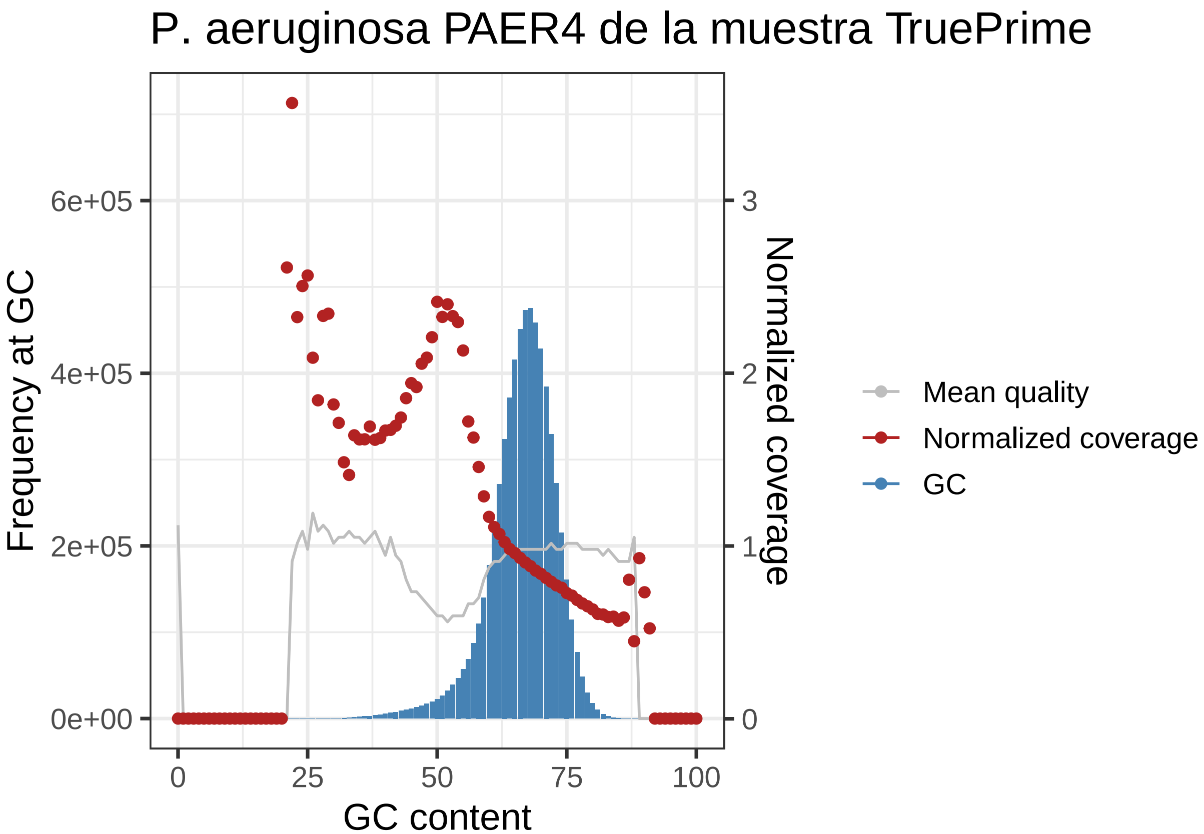

B. subtilis 110NA de la muestra RepliG2

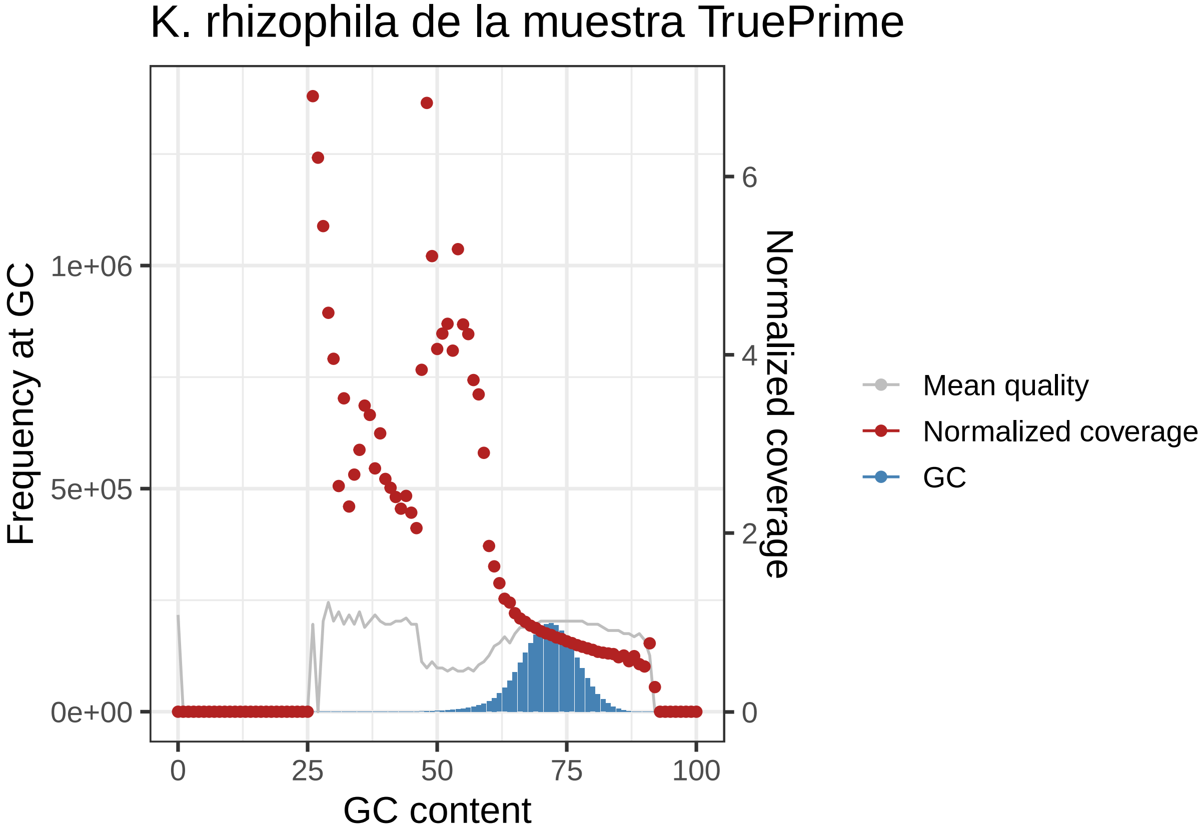

P. aeruginosa PAER4 de la muestra RepliG2
Frequency at GC
8e+05
6e+05
4e+05
2e+05
0e+00
Normalized coverage
4
3
2
1
0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC


K. rhizophila de la muestra RepliG2
Frequency at GC
4e+06
2e+06
0e+00
Normalized coverage
20
10
0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC

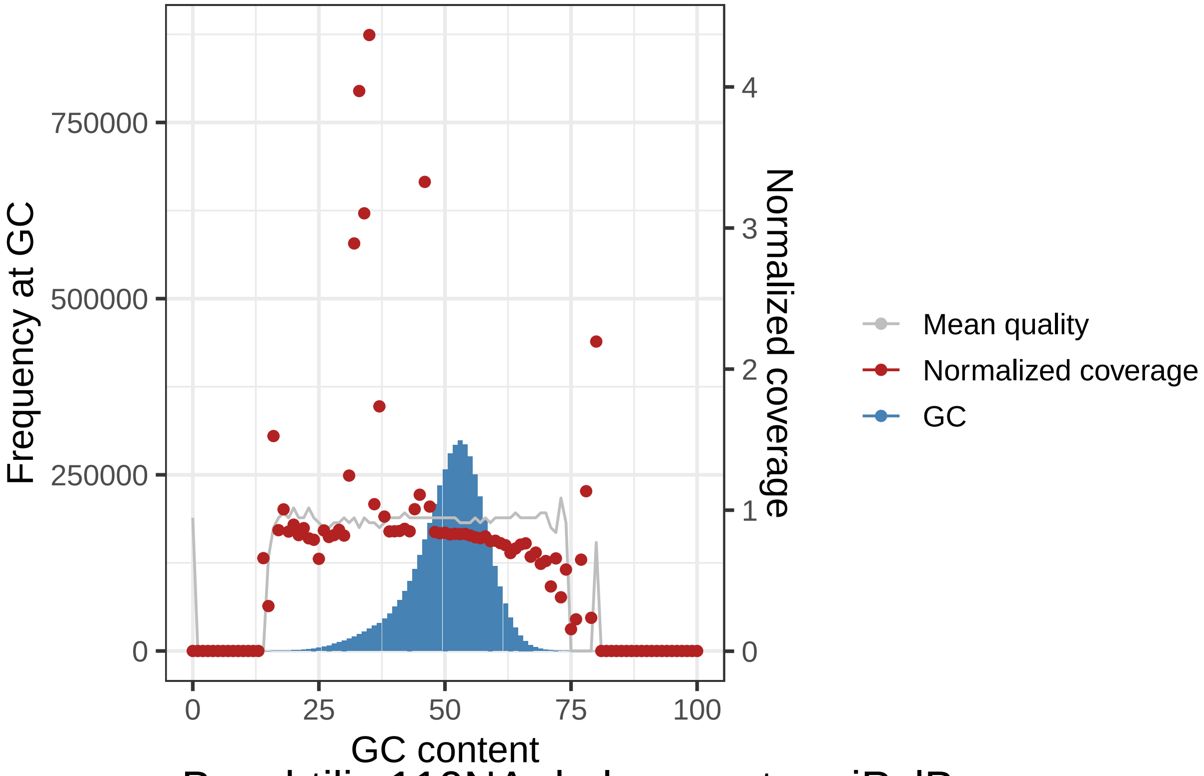

E. coli de la muestra TruePrime


B. subtilis 110NA de la muestra TruePrime

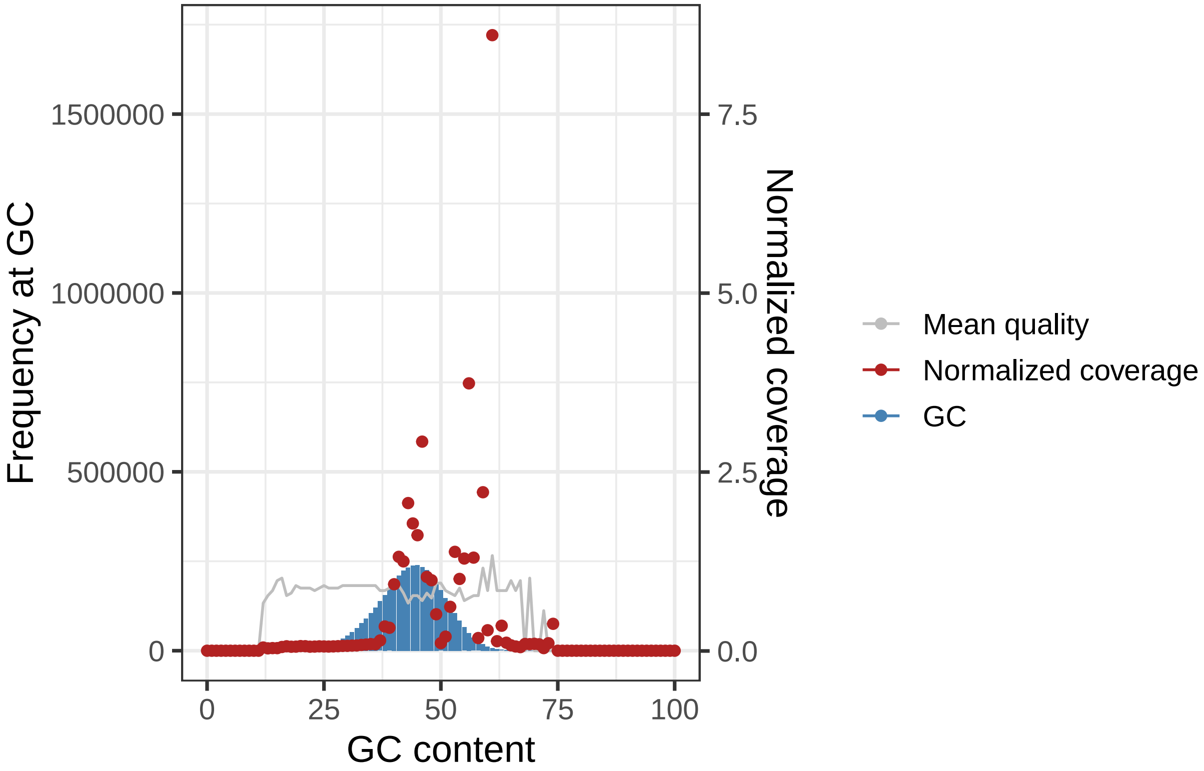

P. aeruginosa PAER4 de la muestra TruePrime
Frequency at GC
6e+05
4e+05
2e+05
0e+00
0
25
50
75
100
GC content
Normalized coverage
3
2
1
0
Mean quality
Normalized coverage
GC


K. rhizophila de la muestra TruePrime
Frequency at GC
1e+06
5e+05
0e+00
0
25
50
75
100
GC content
Normalized coverage
6
4
2
0
Mean quality
Normalized coverage
GC

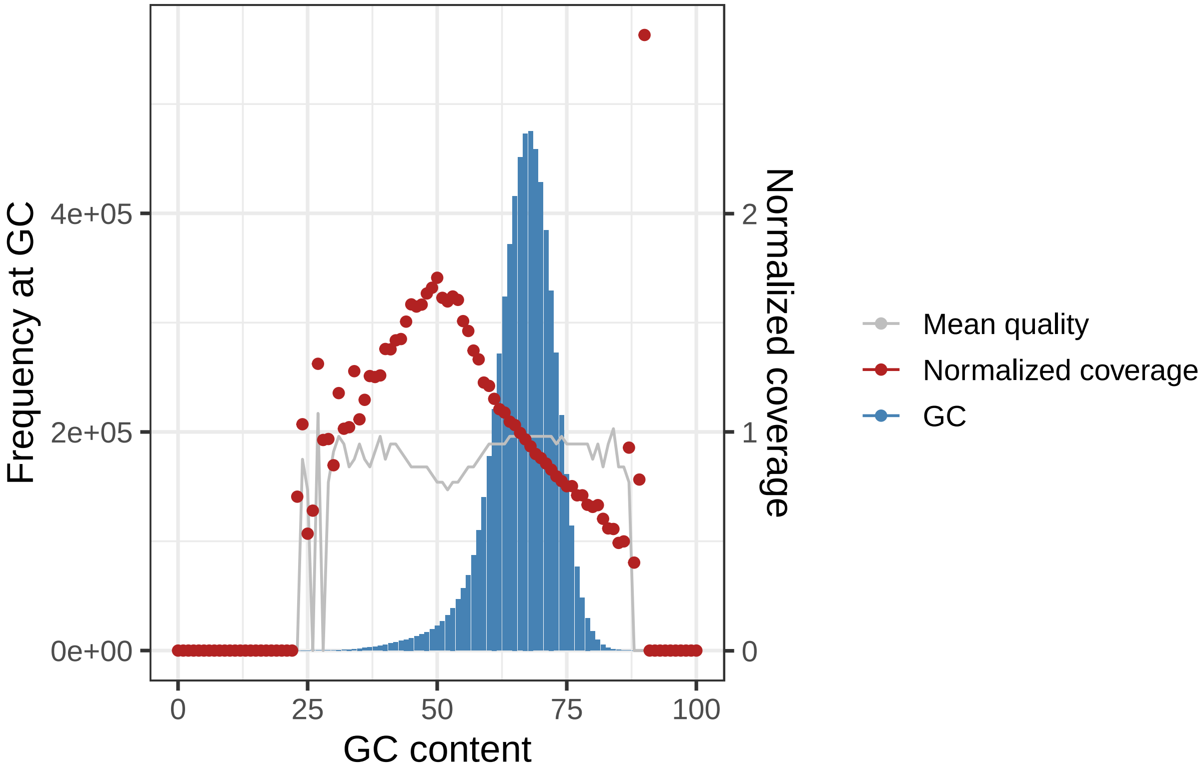
E. coli de la muestra piPolB



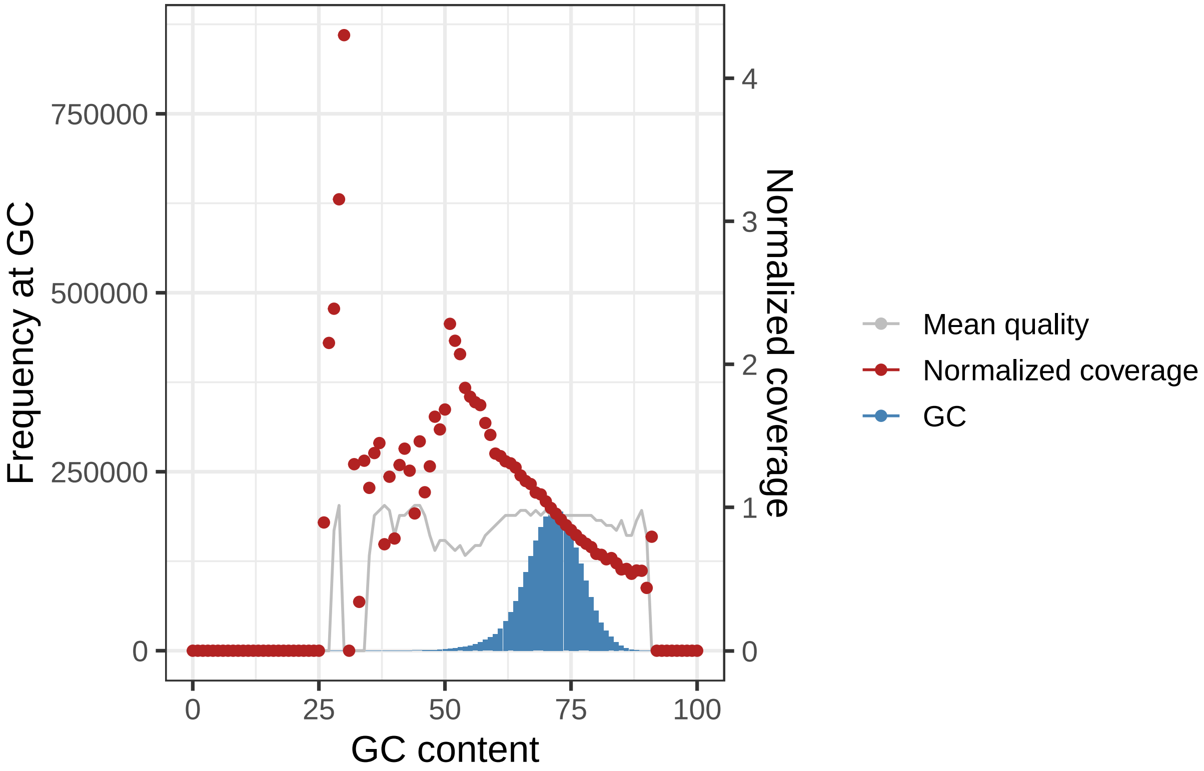
B. subtilis 110NA de la muestra piPolB



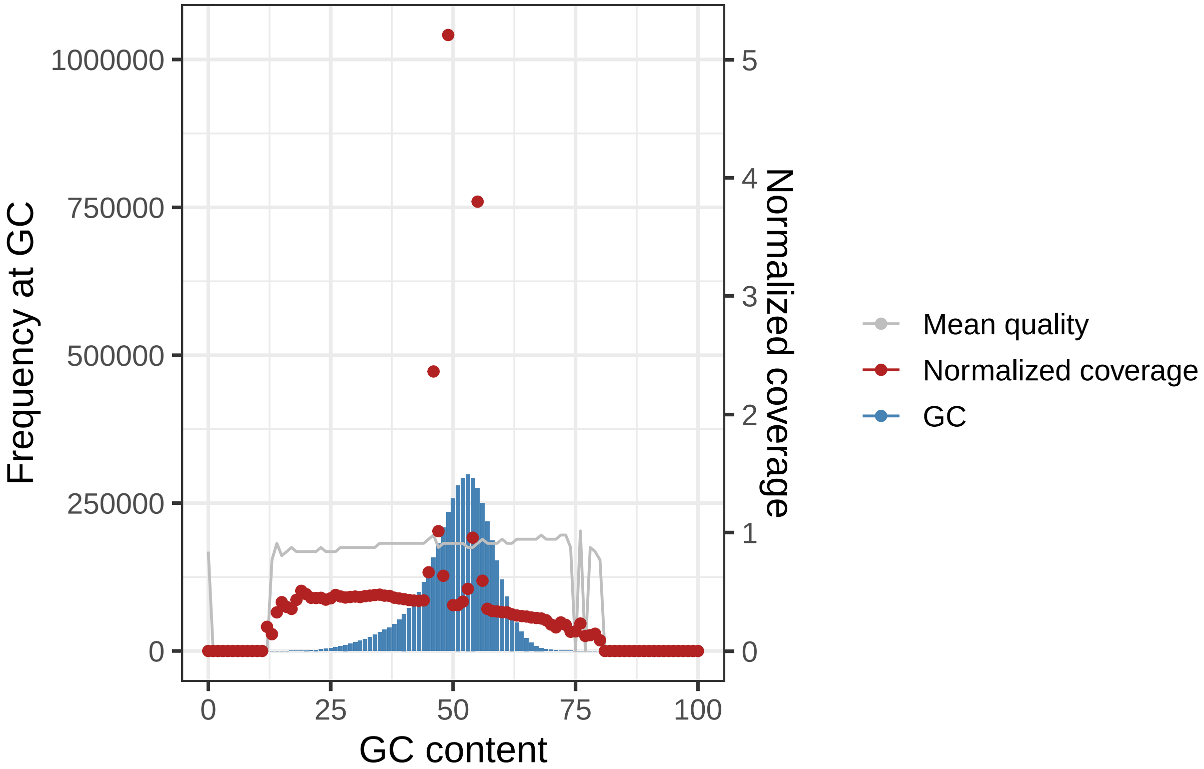
P. aeruginosa PAER4 de la muestra piPolB



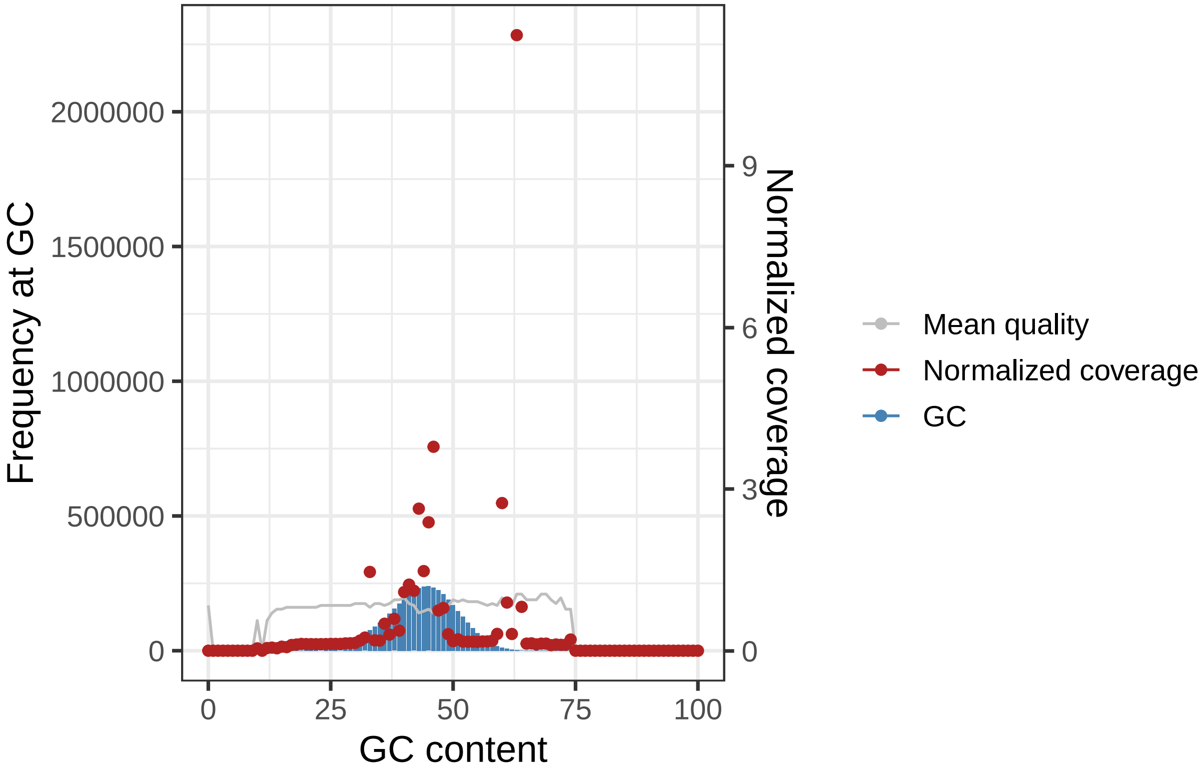
K. rhizophila de la muestra piPolB

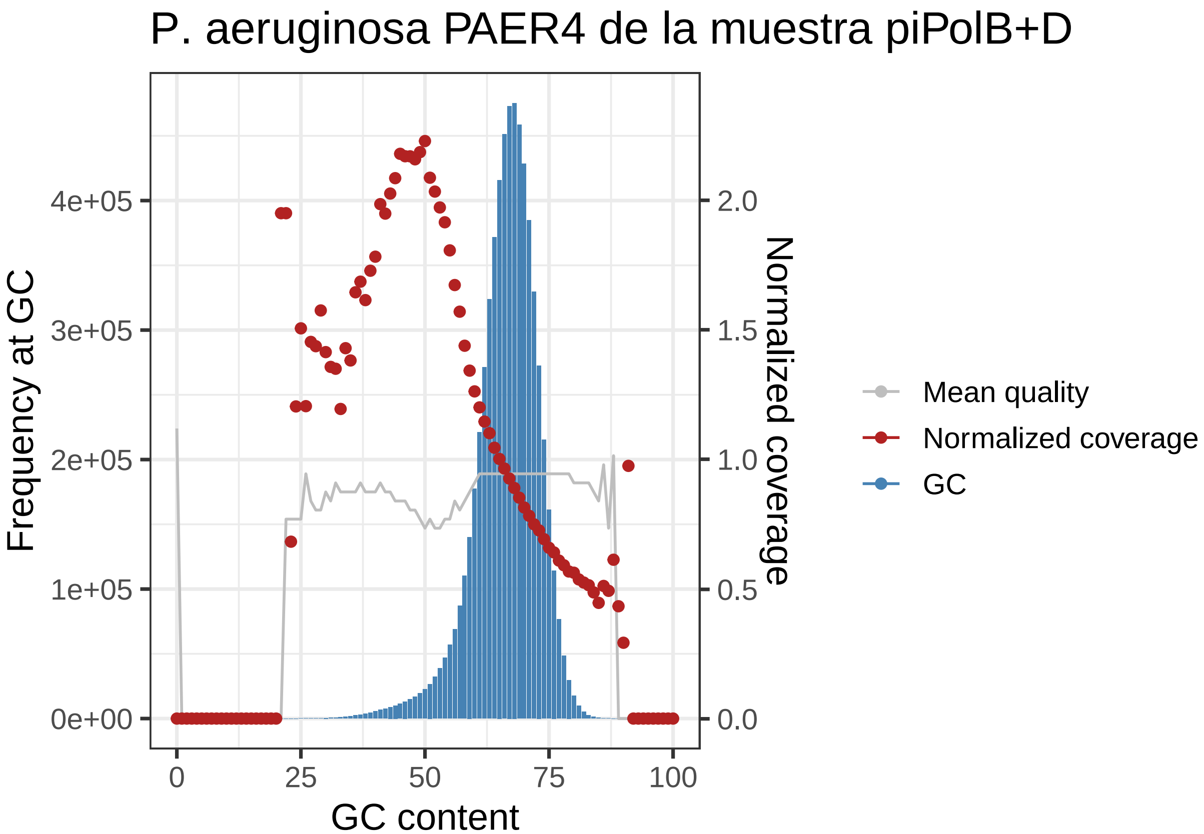


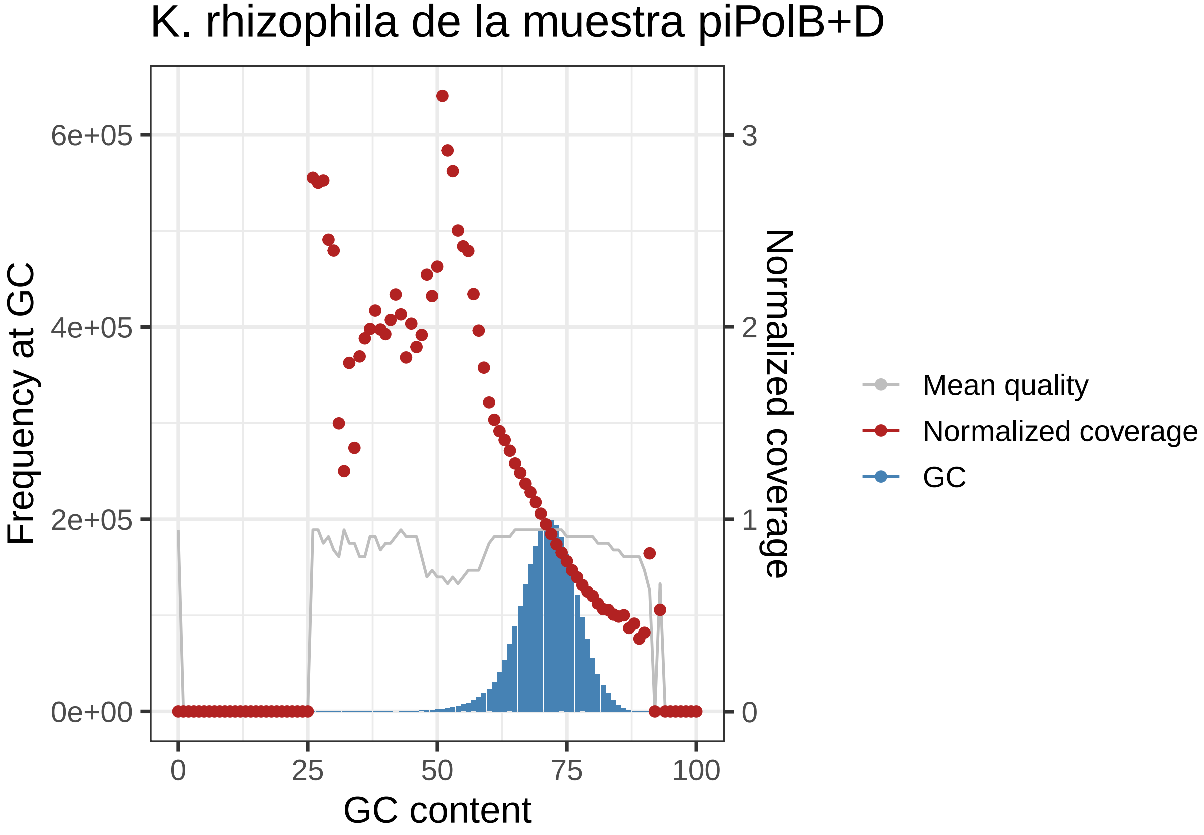
E. coli de la muestra piPolB+D

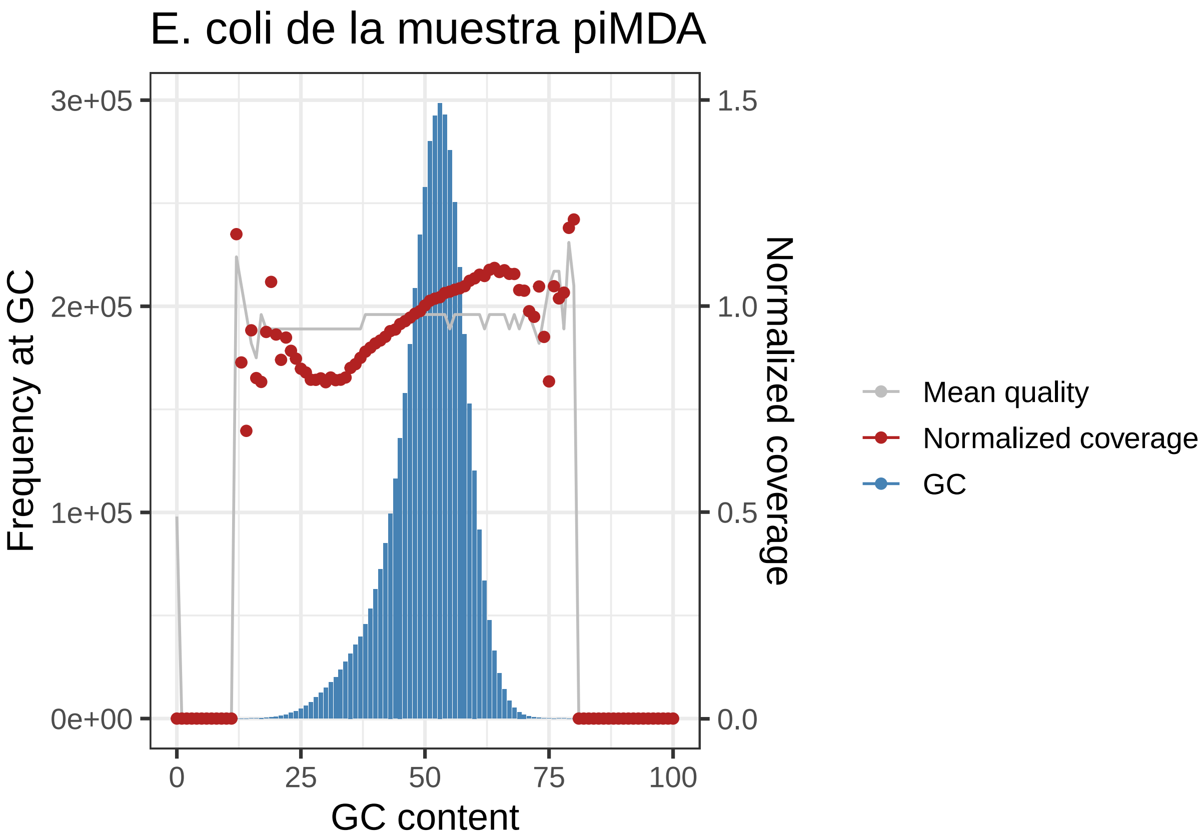


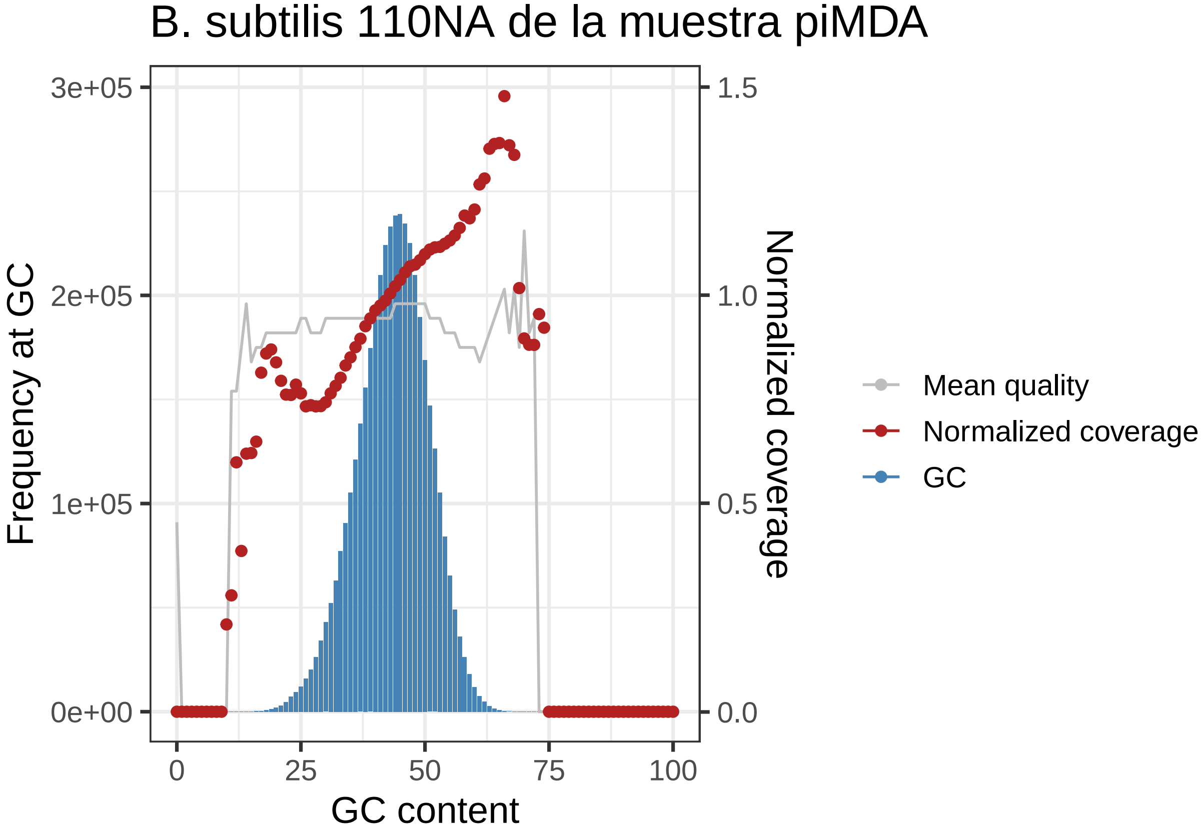
B. subtilis 110NA de la muestra piPolB+D



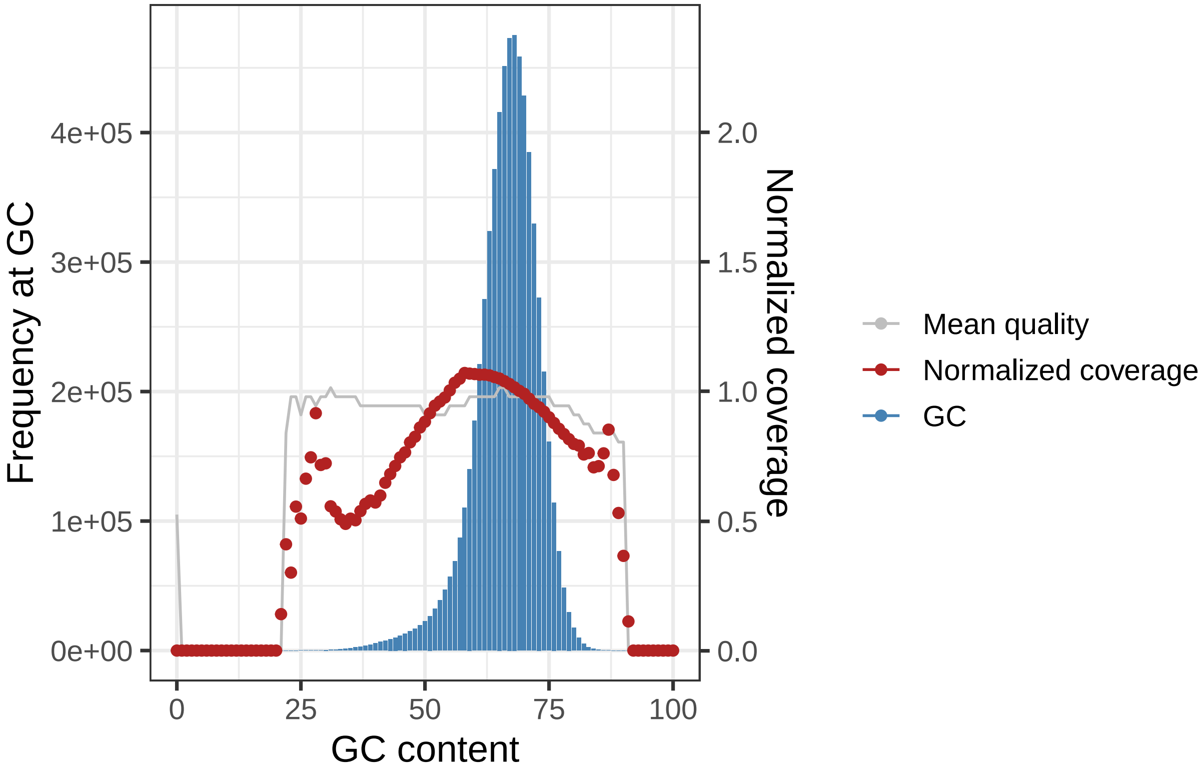

P. aeruginosa PAER4 de la muestra piPolB+D


K. rhizophila de la muestra piPolB+D

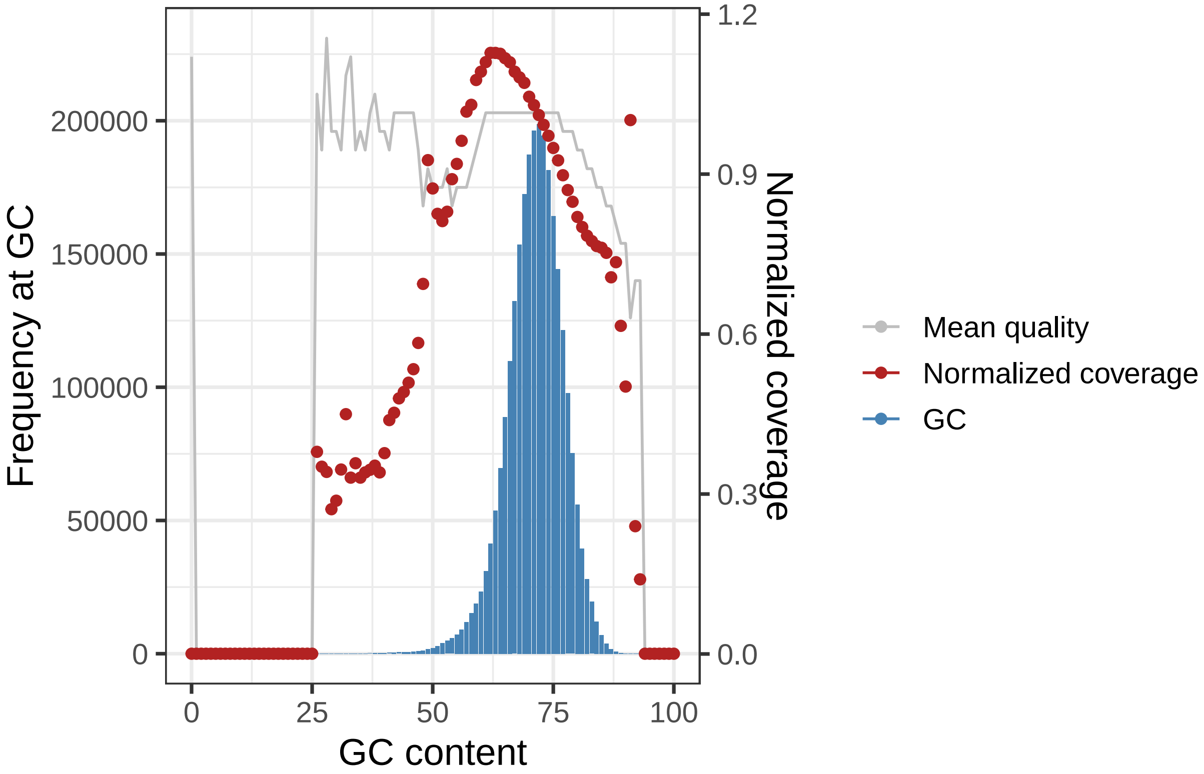

E. coli de la muestra piMDA

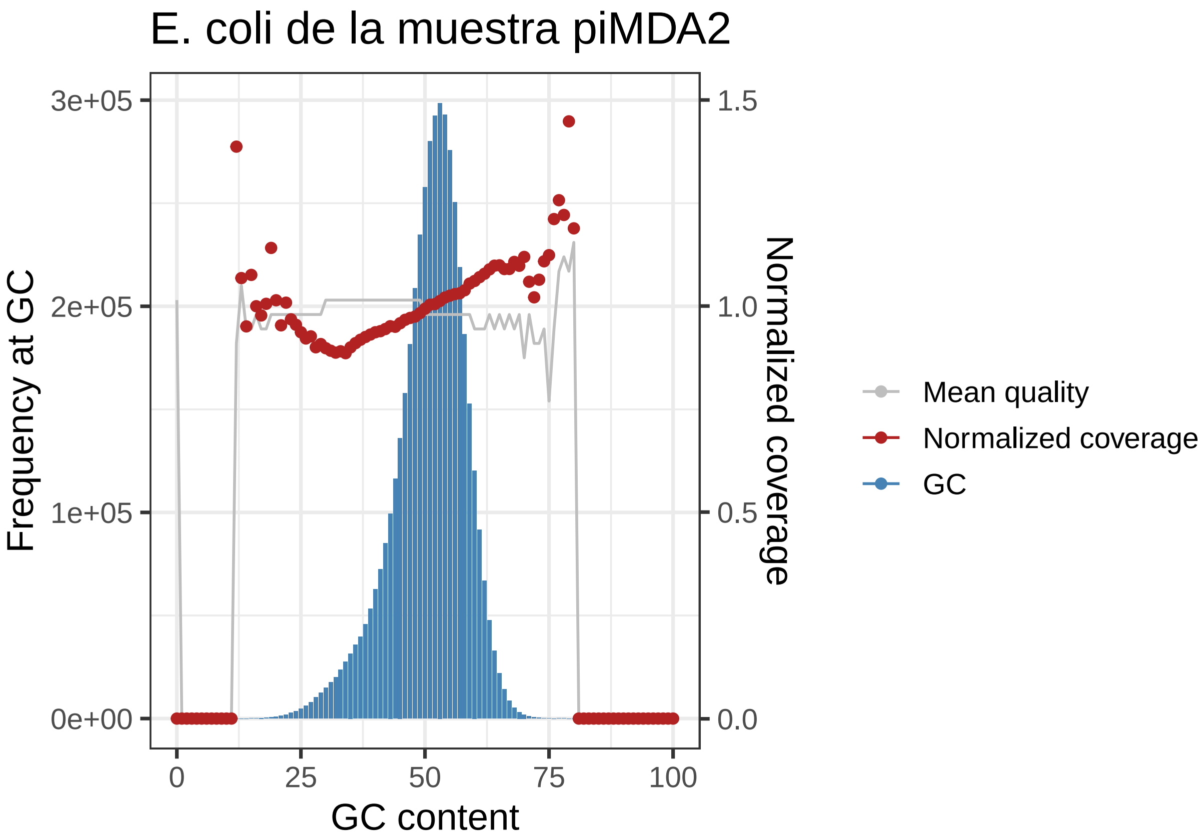

B. subtilis 110NA de la muestra piMDA

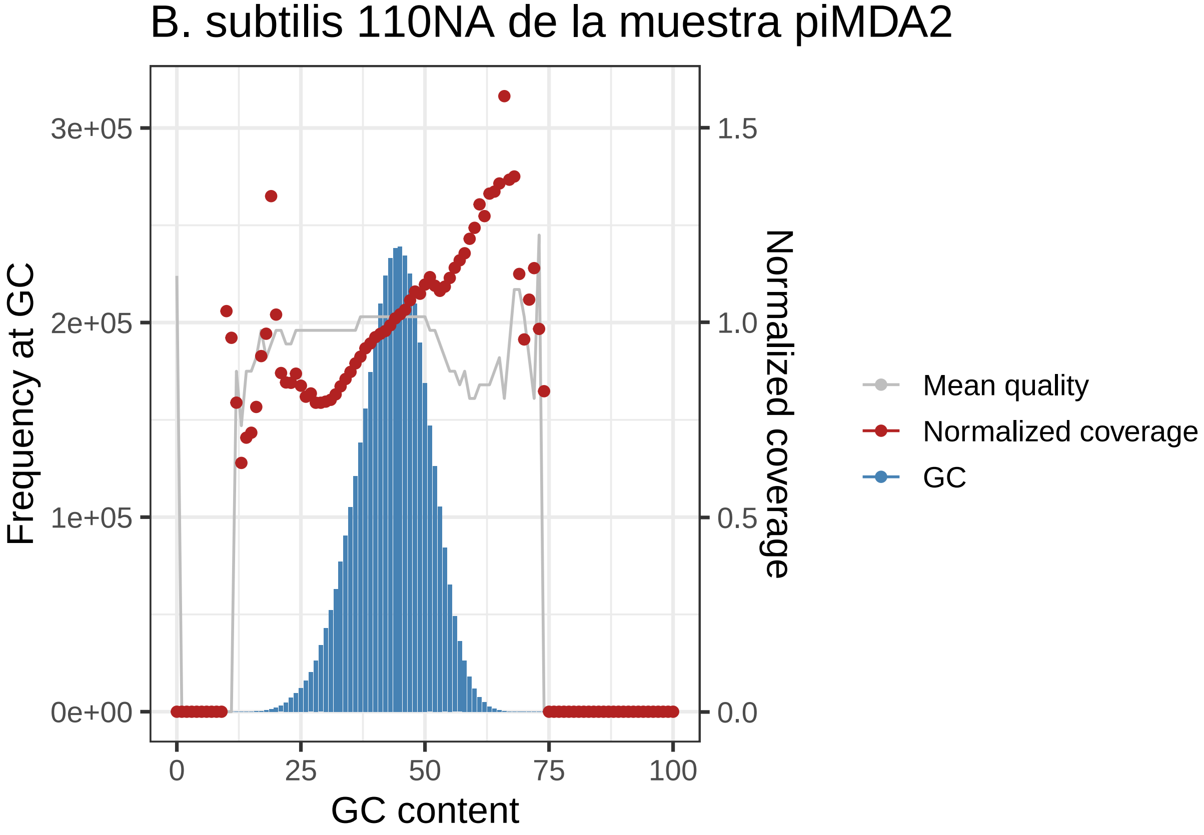
P. aeruginosa PAER4 de la muestra piMDA

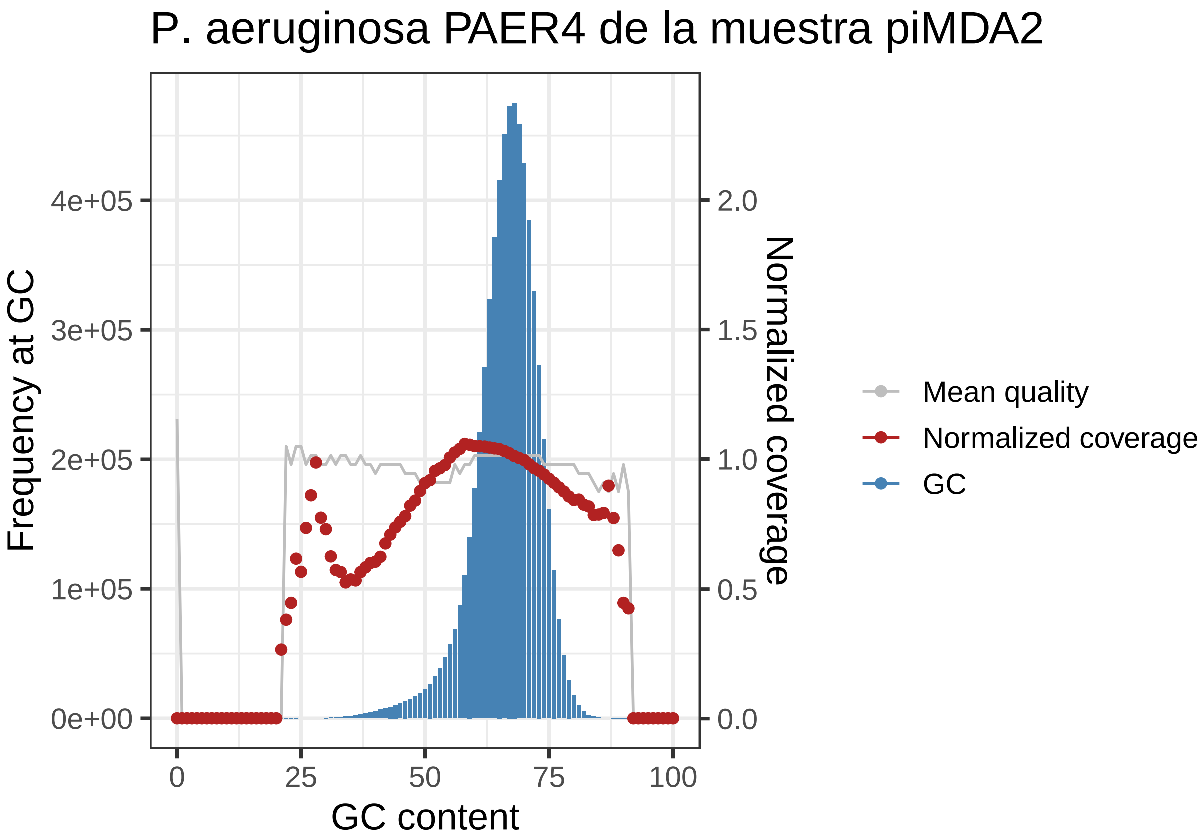


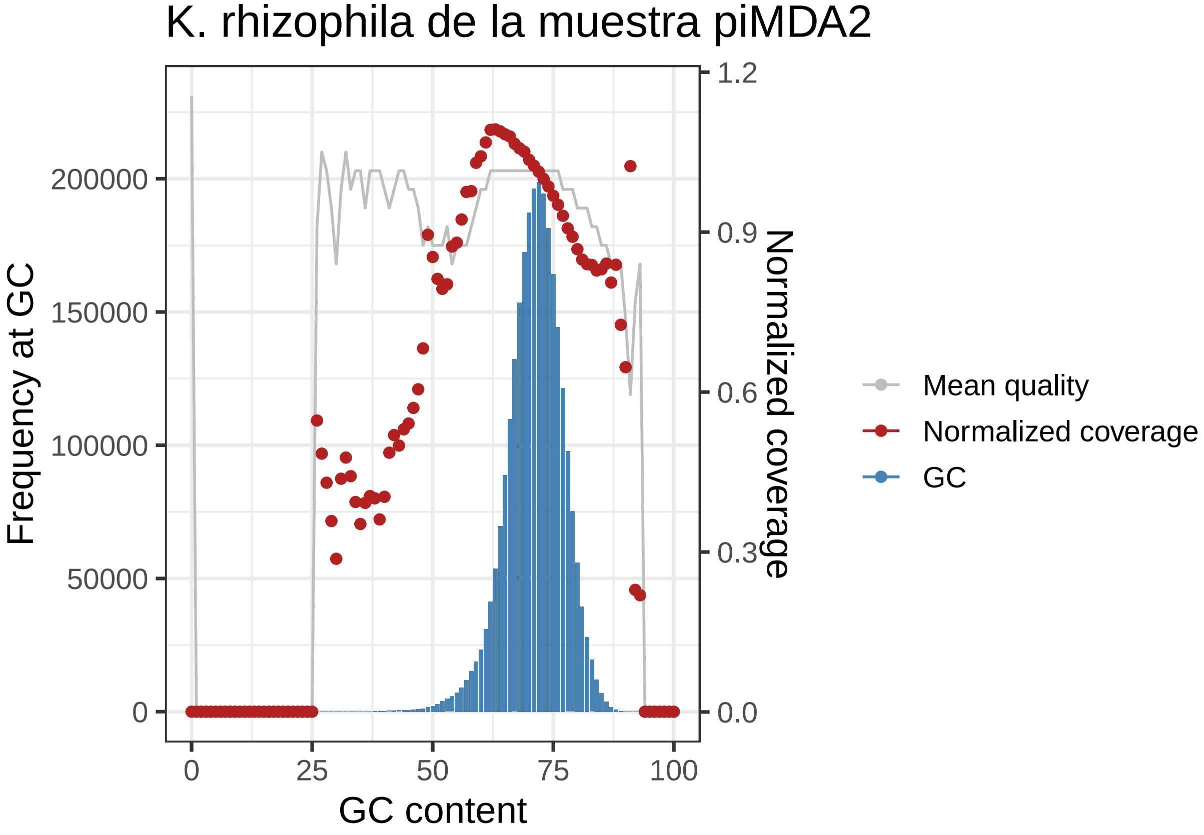
K. rhizophila de la muestra piMDA

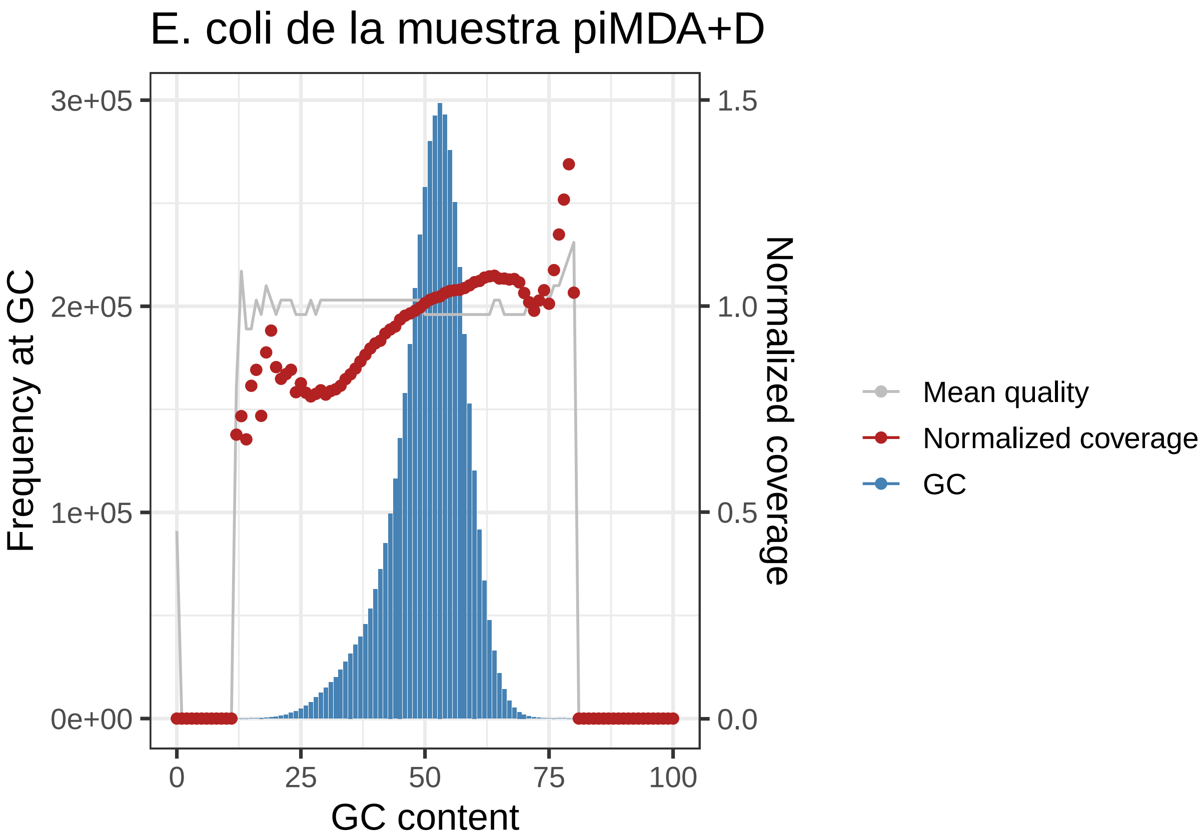


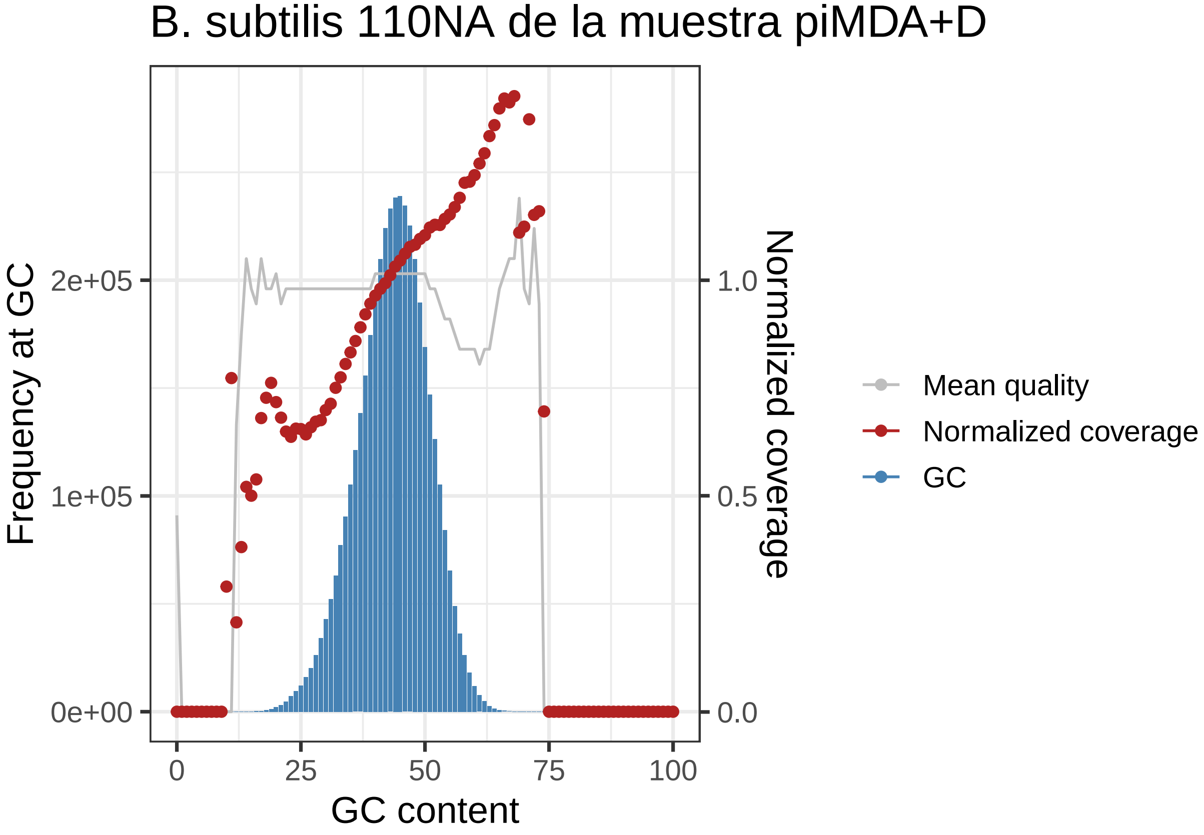

E. coli de la muestra piMDA2
Frequency at GC
3e+05
2e+05
1e+05
0e+00
Normalized coverage
1.5
1.0
0.5
0.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC

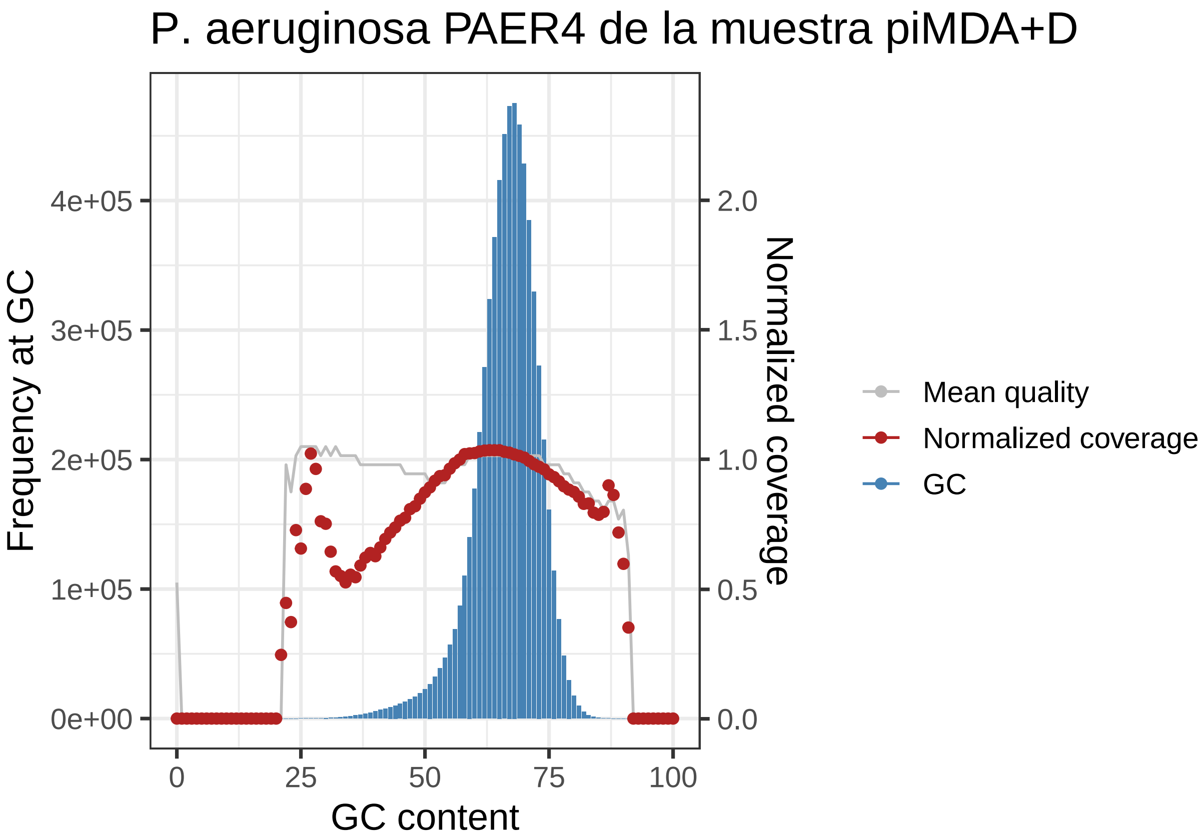

B. subtilis 110NA de la muestra piMDA2
Frequency at GC
3e+05
2e+05
1e+05
0e+00
Normalized coverage
1.5
1.0
0.5
0.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC

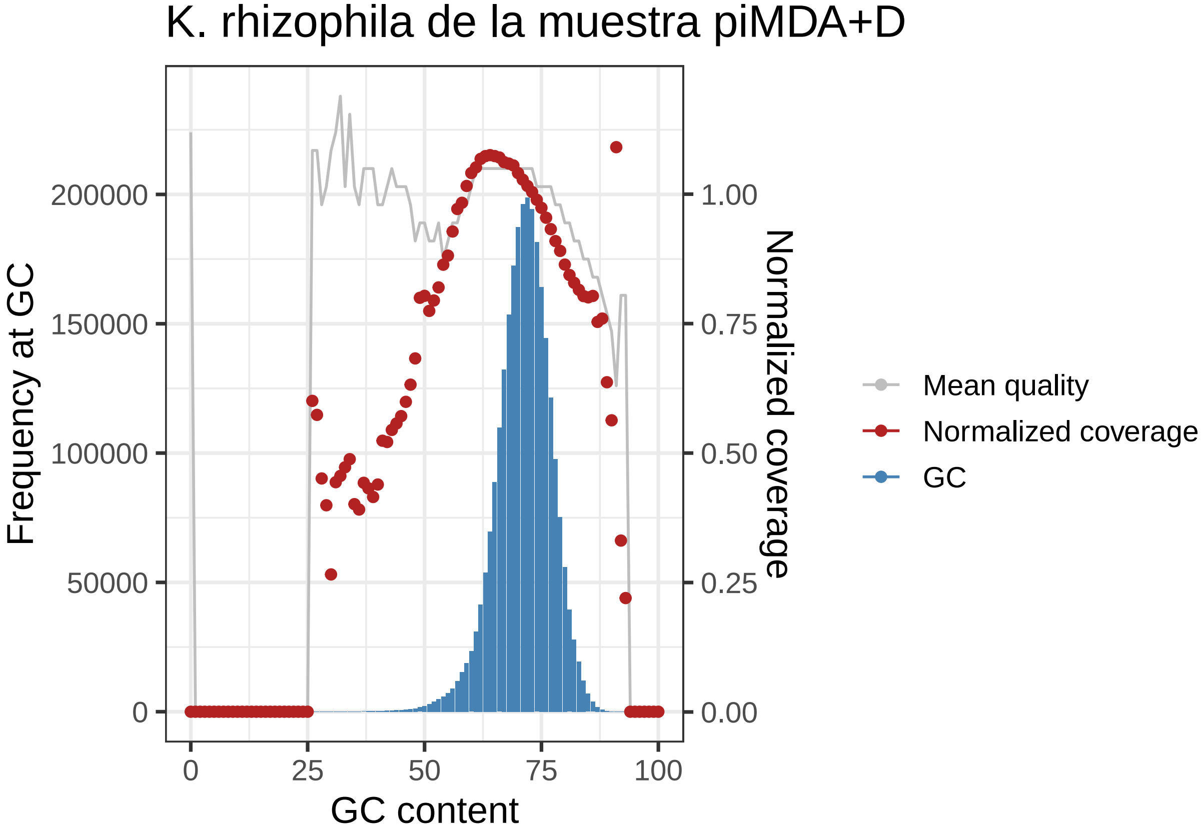

P. aeruginosa PAER4 de la muestra piMDA2
Frequency at GC
4e+05
3e+05
2e+05
1e+05
0e+00
Normalized coverage
2.0
1.5
1.0
0.5
0.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC

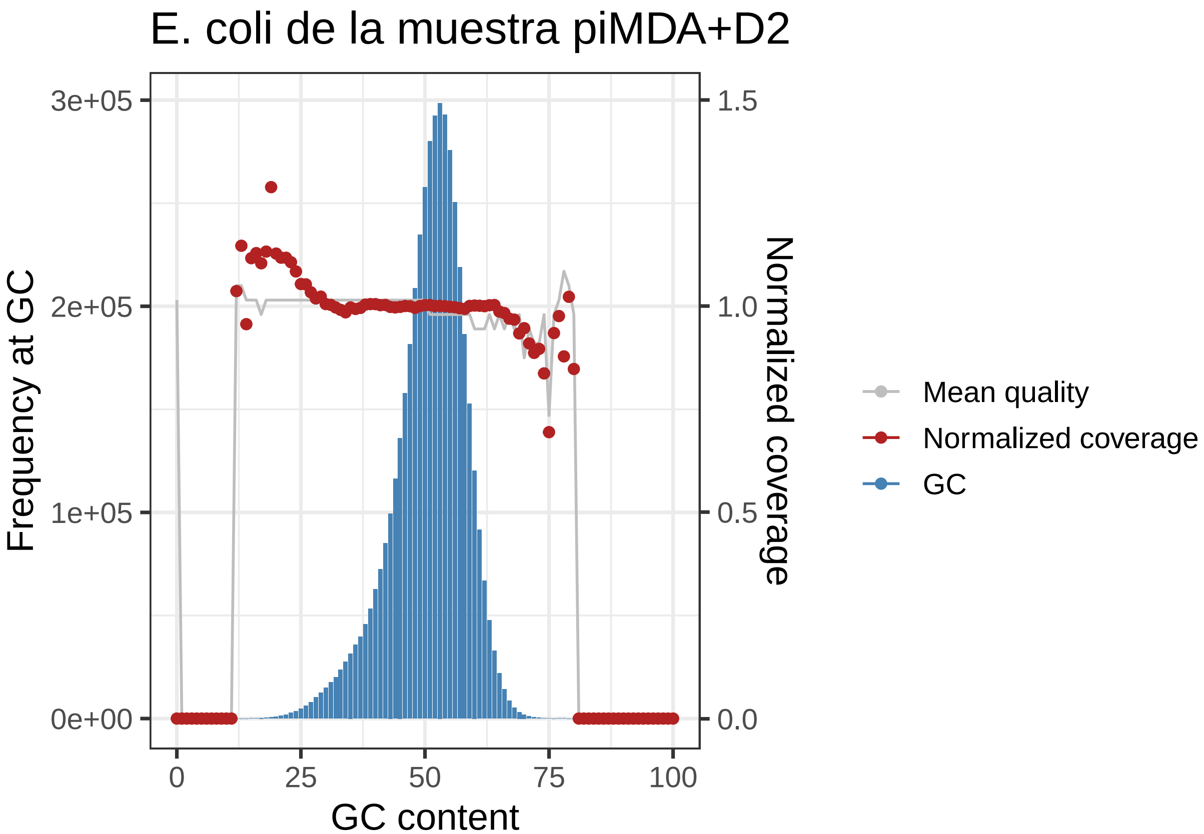

K. rhizophila de la muestra piMDA2
Frequency at GC
200000
150000
100000
50000
0
Normalized coverage
1.2
0.9
0.6
0.3
0.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC

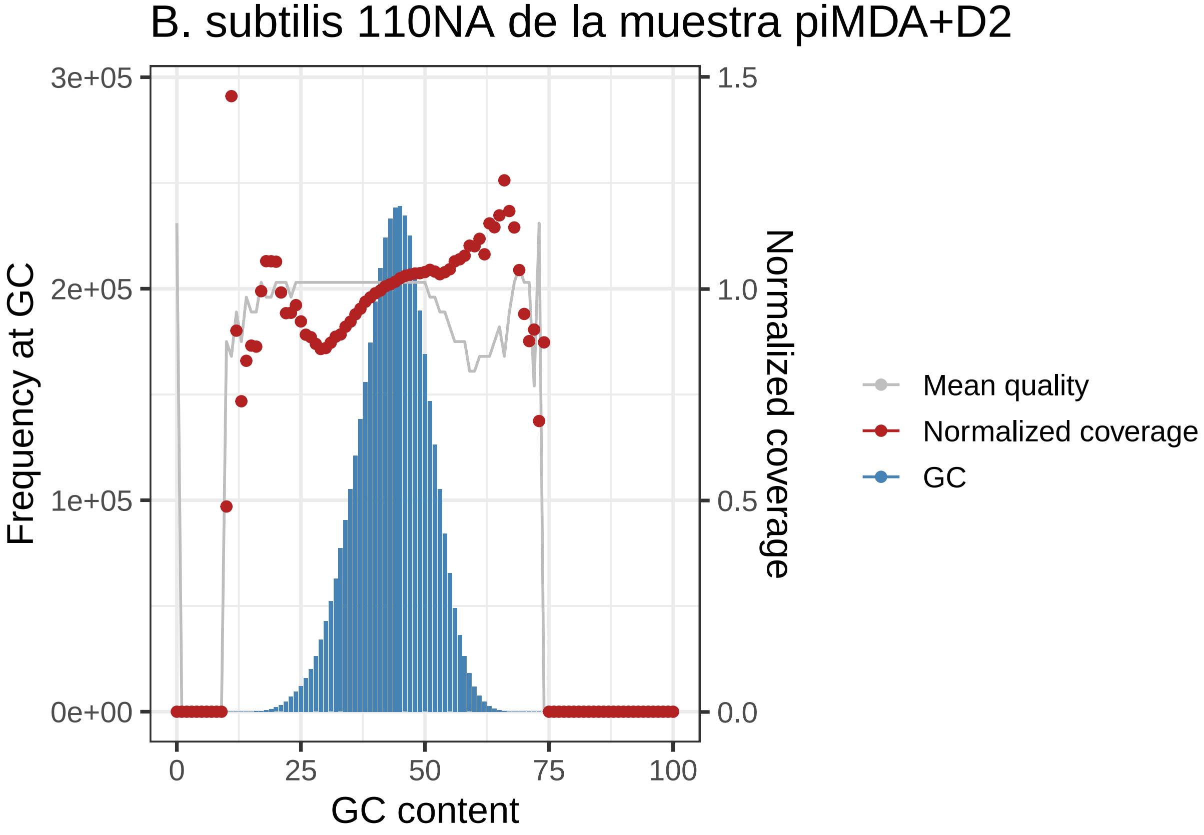

E. coli de la muestra piMDA+D
Frequency at GC
3e+05
2e+05
1e+05
0e+00
Normalized coverage
1.5
1.0
0.5
0.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC

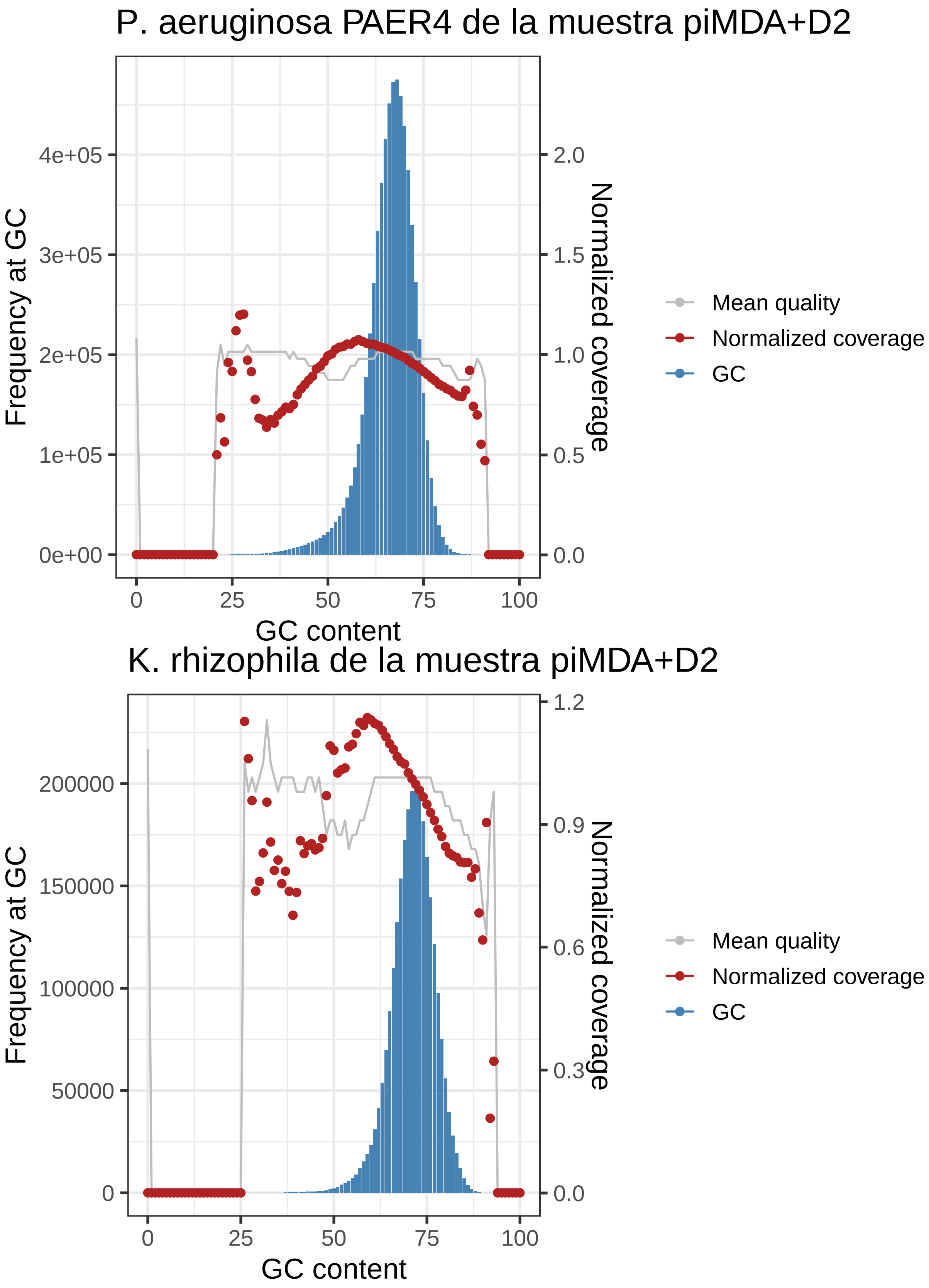

B. subtilis 110NA de la muestra piMDA+D
Frequency at GC
2e+05
1e+05
0e+00
Normalized coverage
1.0
0.5
0.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC


P. aeruginosa PAER4 de la muestra piMDA+D


K. rhizophila de la muestra piMDA+D


E. coli de la muestra piMDA+D2
Frequency at GC
3e+05
2e+05
1e+05
0e+00
Normalized coverage
1.5
1.0
0.5
0.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC


B. subtilis 110NA de la muestra piMDA+D2
Frequency at GC
3e+05
2e+05
1e+05
0e+00
Normalized coverage
1.5
1.0
0.5
0.0
0
25
50
75
100
GC content
Mean quality
Normalized coverage
GC

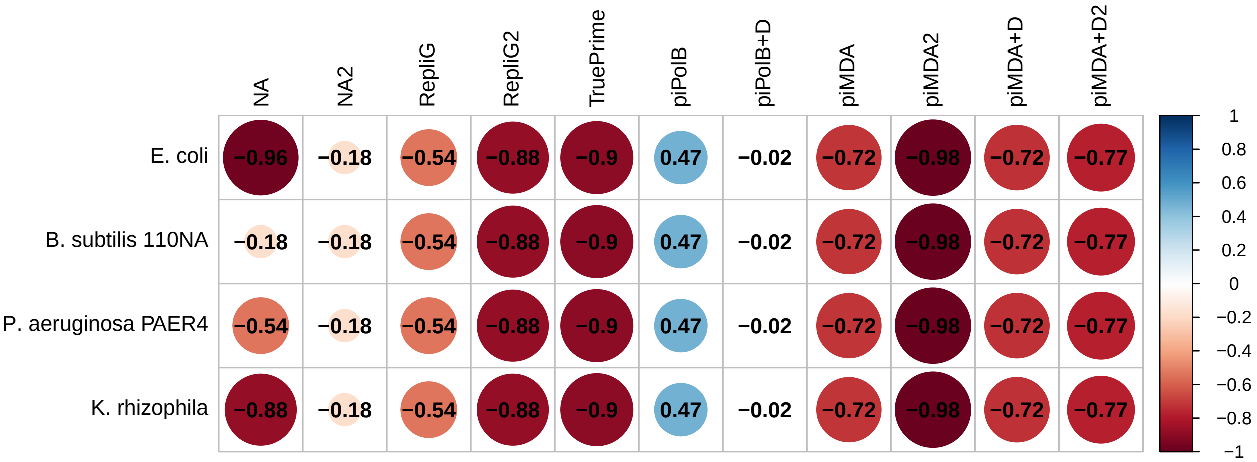


Correlations 1: GC content vs. Normalized Coverage

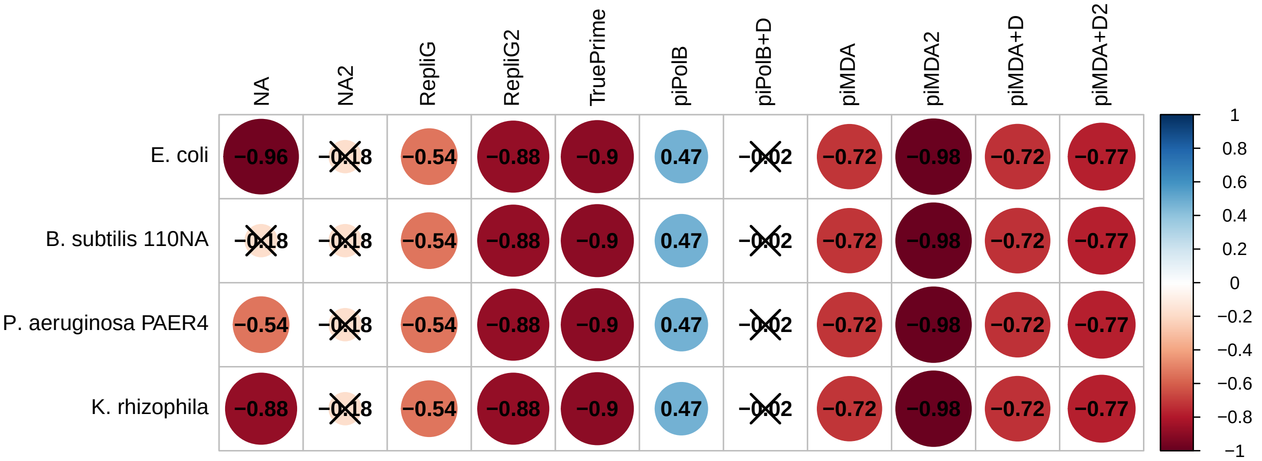
```r
cor_matrix <- data.frame(44,3)
for (i in 1:length(gc)){
  gc[[i]] <- gc[[i]][gc[[i]]$WINDOWS>5,] #remove 0 values
  cor_matrix[i,1] <- gc_picard[i]
  tmp <- cor.test(gc[[i]]$GC,gc[[i]]$NORMALIZED_COVERAGE)
  cor_matrix[i,2] <-tmp$estimate
  cor_matrix[i,3] <-tmp$p.value
}

corr_frame <- data.frame(matrix(ncol = 11, nrow = 4))

#split the data to genome vs sample
#GC vs Normalized Coverage
for (i in 1:ncol(corr_frame)){
  if (i==1){
    corr_frame[1:4,i] <- cor_matrix[1:4,2]
  }else{
    corr_frame[1:4,i] <- cor_matrix[(4*i)-3:4*i,2]
  }
}
corr_frame <- sapply(corr_frame, as.numeric)
colnames(corr_frame) <- c("NA","NA2","RepliG","RepliG2","TruePrime","piPolB","piPolB+D","piMDA","piMDA2
row.names(corr_frame) <- c("E. coli","B. subtilis 110NA", "P. aeruginosa PAER4","K. rhizophila")
#plot
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
corrplot(as.matrix(corr_frame), tl.col="black", addCoef.col = 1, number.cex = 1)
```



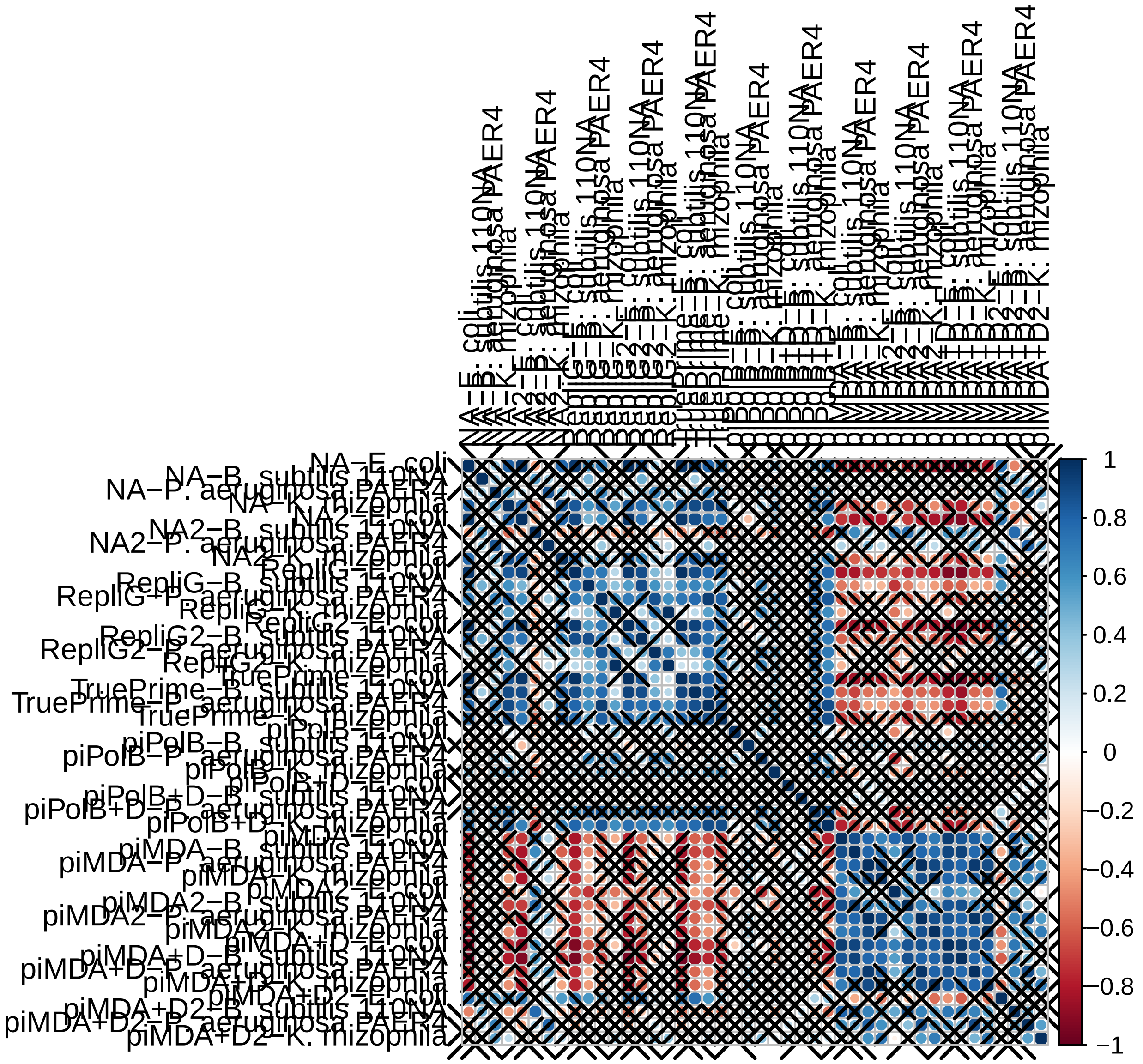
```r
#include the p-values
corr_frame2 <- data.frame(matrix(ncol = 11, nrow = 4))

for (i in 1:ncol(corr_frame2)){
  if (i==1){
    corr_frame2[1:4,i] <- cor_matrix[1:4,3]
  }else{
    corr_frame2[1:4,i] <- cor_matrix[(4*i)-3:4*i,3]
  }
}
corr_frame2 <- sapply(corr_frame2, as.numeric)
colnames(corr_frame2) <- colnames(corr_frame)
row.names(corr_frame2) <- row.names(corr_frame)

#plot
corrplot(as.matrix(corr_frame), tl.col="black", p.mat=corr_frame2, addCoef.col = 1,
         number.cex = 1)
```



```
library(gdata)
```

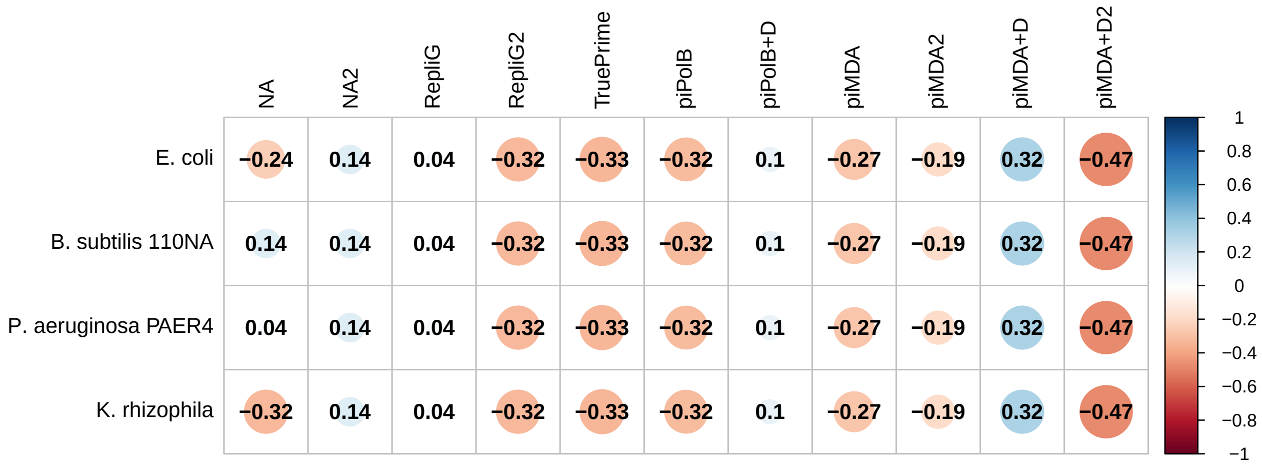
```
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

##

## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.

##
## Attaching package: 'gdata'

## The following object is masked from 'package:stats':
##
##     nobs

## The following object is masked from 'package:utils':
##
##     object.size

## The following object is masked from 'package:base':
##
##     startsWith
```

```
#multiple correlation
cor_matrix2 <- gc[[1]]$NORMALIZED_COVERAGE
for (i in 2:length(gc)){
  gc[[i]] <- gc[[i]][gc[[i]]$WINDOWS!=0,] #remove 0 values
  cor_matrix2 <-cbindX(as.data.frame(cor_matrix2),as.data.frame(gc[[i]]$NORMALIZED_COVERAGE))
}
colnames(cor_matrix2) <-paste0(genomas[,2],"-",genomas[,1])
testRes = cor.mtest(cor_matrix2, conf.level = 0.95)


corrplot(cor(cor_matrix2, use="complete.obs"), tl.col="black", p.mat = testRes$p)
```

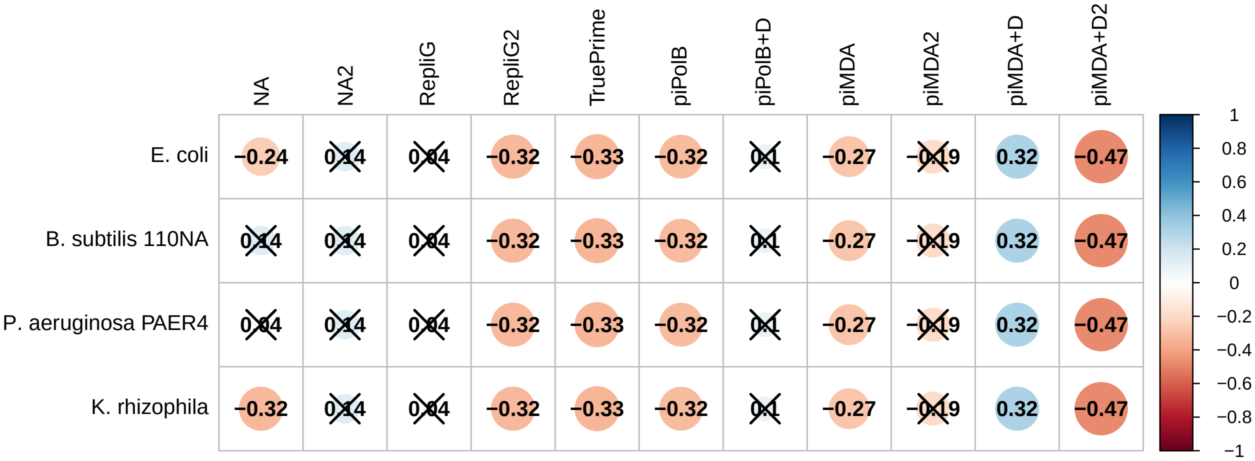


Correlations 1: Windows vs. Normalized Coverage

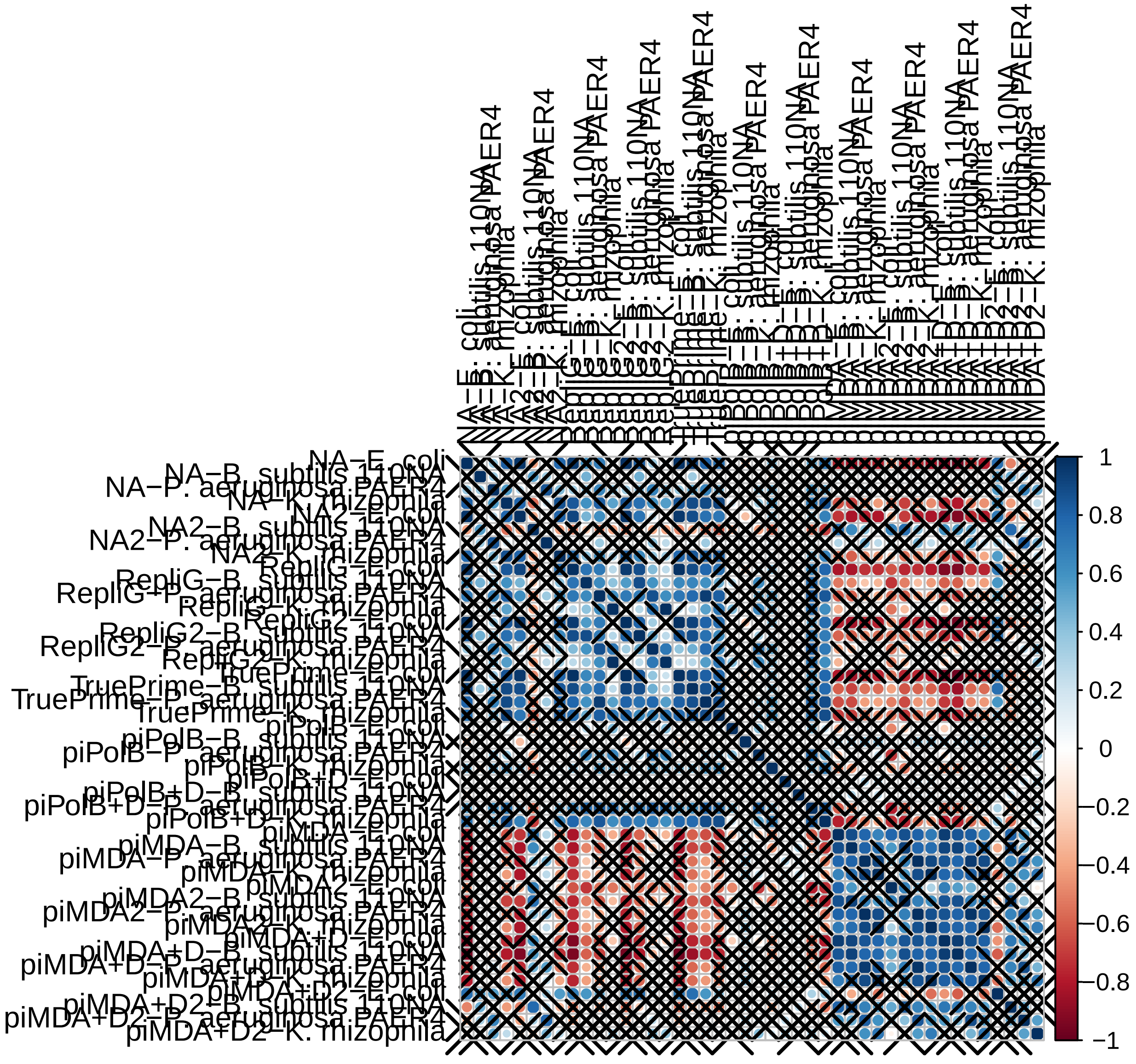
```
cor_matrix <- data.frame(44,3)
for (i in 1:length(gc)){
  gc[[i]] <- gc[[i]][gc[[i]]$WINDOWS>5,] #remove 0 values
  cor_matrix[i,1] <- gc_picard[i]
  tmp <- cor.test(gc[[i]]$WINDOWS,gc[[i]]$NORMALIZED_COVERAGE)
  cor_matrix[i,2] <-tmp$estimate
  cor_matrix[i,3] <-tmp$p.value
}

corr_frame <- data.frame(matrix(ncol = 11, nrow = 4))

#split the data to genome vs sample
#GC vs Normalized Coverage
for (i in 1:ncol(corr_frame)){
  if (i==1){
    corr_frame[1:4,i] <- cor_matrix[1:4,2]
  }else{
    corr_frame[1:4,i] <- cor_matrix[(4*i)-3:4*i,2]
  }
}
corr_frame <- sapply(corr_frame, as.numeric)
colnames(corr_frame) <- c("NA","NA2","RepliG","RepliG2","TruePrime","piPolB","piPolB+D","piMDA","piMDA2
row.names(corr_frame) <- c("E. coli","B. subtilis 110NA", "P. aeruginosa PAER4","K. rhizophila")
#plot
library(corrplot)

corrplot(as.matrix(corr_frame), tl.col="black", addCoef.col = 1, number.cex = 1)
```



```
#include the p-values
corr_frame2 <- data.frame(matrix(ncol = 11, nrow = 4))

for (i in 1:ncol(corr_frame2)){
  if (i==1){
    corr_frame2[1:4,i] <- cor_matrix[1:4,3]
  }else{
    corr_frame2[1:4,i] <- cor_matrix[(4*i)-3:4*i,3]
  }
}
corr_frame2 <- sapply(corr_frame2, as.numeric)
colnames(corr_frame2) <- colnames(corr_frame)
row.names(corr_frame2) <- row.names(corr_frame)
#plot


corrplot(as.matrix(corr_frame), tl.col="black", p.mat=corr_frame2, addCoef.col = 1,
         number.cex = 1)
```



```
library(gdata)
#multiple correlation
cor_matrix2 <- gc[[1]]$NORMALIZED_COVERAGE
for (i in 2:length(gc)){
  gc[[i]] <- gc[[i]][gc[[i]]$WINDOWS!=0,] #remove 0 values
  cor_matrix2 <-cbindX(as.data.frame(cor_matrix2),as.data.frame(gc[[i]]$NORMALIZED_COVERAGE))
}

testRes = cor.mtest(cor_matrix2, conf.level = 0.95)
colnames(cor_matrix2) <-paste0(genomas[,2],"-",genomas[,1])

corrplot(cor(cor_matrix2, use="complete.obs"),tl.col="black",  p.mat = testRes$p)
```


NA−E. coli
NA−B. subtilis 110NA
NA−P. aeruginosa PAER4
NA−K. rhizophila
NA2−E. coli
NA2−B. subtilis 110NA
NA2−P. aeruginosa PAER4
NA2−K. rhizophila
RepliG−E. coli
RepliG−B. subtilis 110NA
RepliG−P. aeruginosa PAER4
RepliG−K. rhizophila
RepliG2−E. coli
RepliG2−B. subtilis 110NA
RepliG2−P. aeruginosa PAER4
RepliG2−K. rhizophila
TruePrime−E. coli
TruePrime−B. subtilis 110NA
TruePrime−P. aeruginosa PAER4
TruePrime−K. rhizophila
piPolB−E. coli
piPolB−B. subtilis 110NA
piPolB−P. aeruginosa PAER4
piPolB−K. rhizophila
piPolB+D−E. coli
piPolB+D−B. subtilis 110NA
piPolB+D−P. aeruginosa PAER4
piPolB+D−K. rhizophila
piMDA−E. coli
piMDA−B. subtilis 110NA
piMDA−P. aeruginosa PAER4
piMDA−K. rhizophila
piMDA2−E. coli
piMDA2−B. subtilis 110NA
piMDA2−P. aeruginosa PAER4
piMDA2−K. rhizophila
piMDA+D−E. coli
piMDA+D−B. subtilis 110NA
piMDA+D−P. aeruginosa PAER4
piMDA+D−K. rhizophila
piMDA+D2−E. coli
piMDA+D2−B. subtilis 110NA
piMDA+D2−P. aeruginosa PAER4
piMDA+D2−K. rhizophila
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1