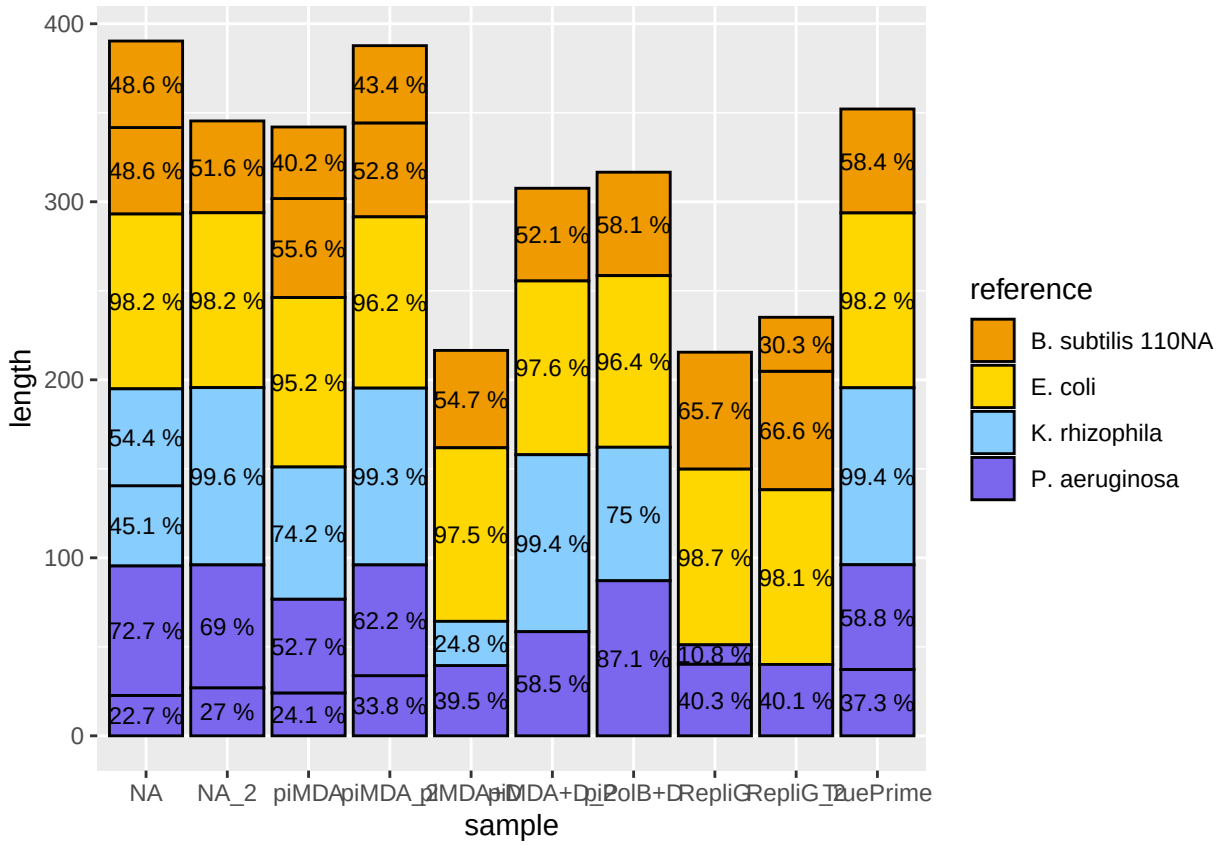
Binning

Modesto Redrejo Rodríguez

2022-04-22

```
dist <- read.csv(file = 'all_distances.csv', sep=";",header=T)

dist$reference <- gsub("../refs/NC_000913.3.fasta","E. coli",dist$reference)
dist$reference <- gsub("../refs/NC_000964.3.fasta","B. subtilis 110NA",dist$reference)
dist$reference <- gsub("../refs/CP013113.1_PAER4.fasta","P. aeruginosa",dist$reference)
dist$reference <- gsub("../refs/Kocuria_rhizophila_ATCC_9341.fasta","K. rhizophila",dist$reference)


dist$sample <- substr(dist$bin,1,5)
dist$sample <- gsub("Ctrl/","NA",dist$sample)
dist$sample <- gsub("Ctrl2","NA_2",dist$sample)
dist$sample <- gsub("A2/bi","piMDA+D_2",dist$sample)
dist$sample <- gsub("B2/bi","piMDA_2",dist$sample)
dist$sample <- gsub("F2/bi","piPolB+D",dist$sample)
dist$sample <- gsub("C3/bi","TruePrime",dist$sample)
dist$sample <- gsub("D3/bi","RepliG_2",dist$sample)
dist$sample <- gsub("N2/bi","piMDA",dist$sample)
dist$sample <- gsub("N4/bi","RepliG",dist$sample)
dist$sample <- gsub("N6/bi","piMDA+D",dist$sample)

dist$bin <- gsub(".fasta","",dist$bin)
dist$bin <- substr(dist$bin,nchar(dist$bin)-5,nchar(dist$bin))
dist$bin <- gsub("/","",dist$bin)
dist$length <- substr(dist$length,1,nchar(dist$length)-5)
dist$length <- as.integer(dist$length)/10




#select the bins
selection <- dist[dist$length>5,]
library(ggplot2)
#remotes::install_github("coolbutuseless/ggpattern")

library(ggpattern)
g <- ggplot(selection, aes(fill=reference, y=length, x=sample, label=paste(length,"%"))) +     geom_bar
  # geom_bar_pattern(position="stack",stat="identity",
                 #  mapping=aes(pattern=bin), alpha=0.6)
g+ geom_text(size = 3, position = position_stack(vjust = 0.5))
```



Try the same but with different MetaCoAG parameters: –mg_threshold 0.4 –bin_mg_threshold 0.2

```
dist <- read.csv(file = 'all_distances_final.csv', sep=";",header=T)

dist$reference <- gsub("../refs/NC_000913.3.fasta","E. coli",dist$reference)
dist$reference <- gsub("../refs/NC_000964.3.fasta","B. subtilis 110NA",dist$reference)
dist$reference <- gsub("../refs/CP013113.1_PAER4.fasta","P. aeruginosa",dist$reference)
dist$reference <- gsub("../refs/Kocuria_rhizophila_ATCC_9341.fasta","K. rhizophila",dist$reference)



dist$sample <- substr(dist$bin,1,5)
dist$sample <- gsub("Ctrlb","NA",dist$sample)
dist$sample <- gsub("Ctrl2","NA_2",dist$sample)
dist$sample <- gsub("A2b/b","piMDA+D_2",dist$sample)
dist$sample <- gsub("B2b/b","piMDA_2",dist$sample)
dist$sample <- gsub("F2b/b","piPolB+D",dist$sample)
dist$sample <- gsub("C3b/b","TruePrime",dist$sample)
dist$sample <- gsub("D3b/b","RepliG_2",dist$sample)
dist$sample <- gsub("N2b/b","piMDA",dist$sample)
dist$sample <- gsub("N4b/b","RepliG",dist$sample)
dist$sample <- gsub("N6b/b","piMDA+D",dist$sample)

dist$bin <- gsub(".fasta","",dist$bin)
dist$bin <- substr(dist$bin,nchar(dist$bin)-5,nchar(dist$bin))
dist$bin <- gsub("/","",dist$bin)
dist$length <- substr(dist$length,1,nchar(dist$length)-5)
```

```
dist$length <- as.integer(dist$length)/10




#select the bins
dist <- dist[-184,]
selection <- dist[dist$length>5,]

#plot
library(ggplot2)
#remotes::install_github("coolbutuseless/ggpattern")
#library(ggpattern)
g <- ggplot(selection, aes(fill=reference, y=length, x=sample, label=paste(length,"%"))) +     geom_bar
  # geom_bar_pattern(position="stack",stat="identity",
                #  mapping=aes(pattern=bin), alpha=0.6)
g <- g+ggtitle("Assembly binning by reference") +
  xlab("Sample") + ylab("Cumulative bin length") + guides(fill=guide_legend(title="Reference genome")) +
g+ geom_text(size = 3, position = position_stack(vjust = 0.5))+ theme_linedraw()
```