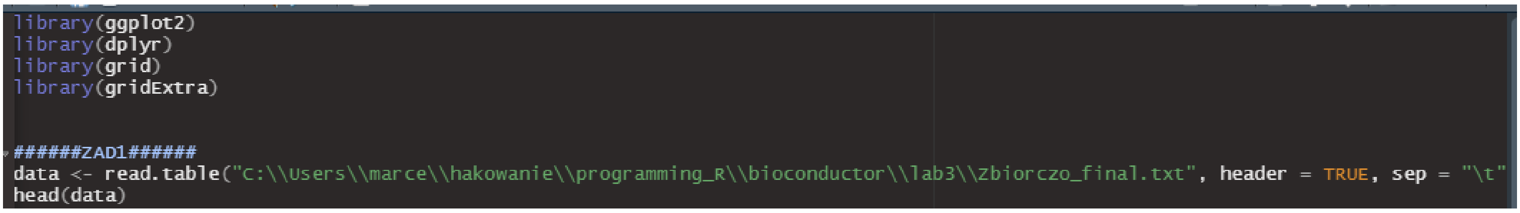
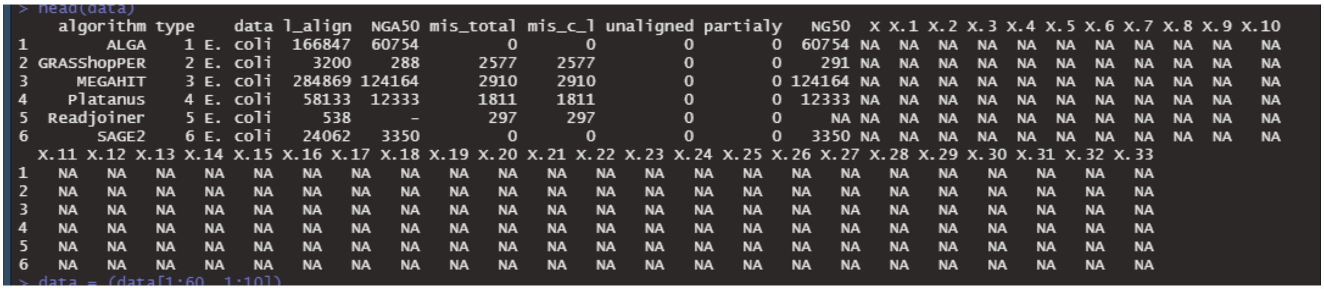
Marcel Buczkowski 145159

Bioconductor lab 3 - raport

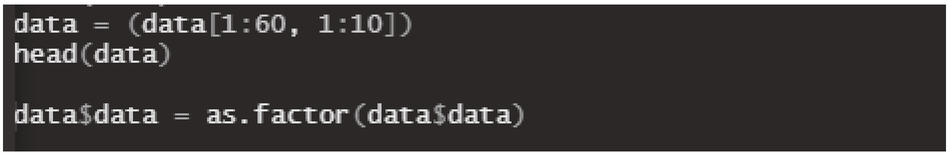
Z1:



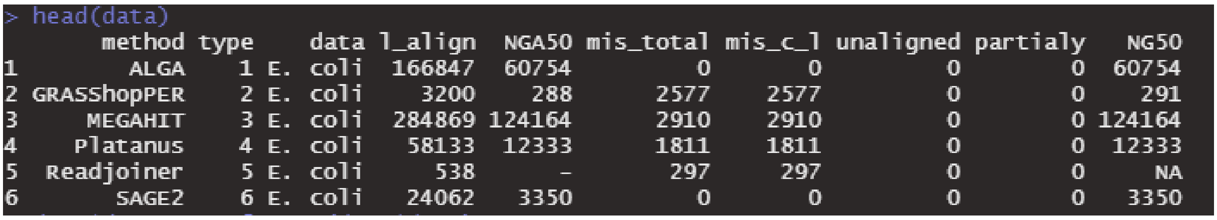
```
library(ggplot2)
library(dplyr)
library(grid)
library(gridExtra)


######ZAD1######
data <- read.table("C:\\Users\\marce\\hakowanie\\programming_R\\bioconductor\\lab3\\zbiorczo_final.txt", header = TRUE, sep = "\t"
head(data)
```



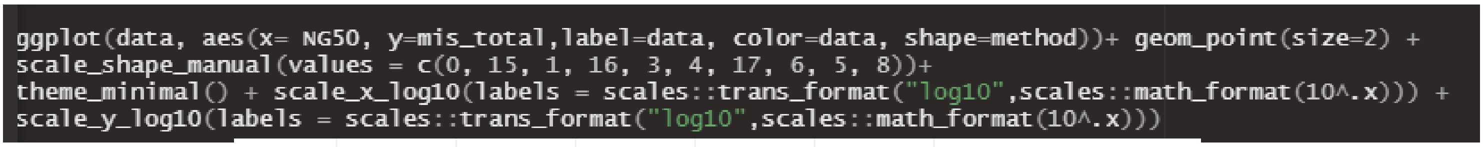
```
> head(data)
    algorithm type    data l_align  NGA50 mis_total mis_c_l unaligned partialy   NG50  X X.1 X.2 X.3 X.4 X.5 X.6 X.7 X.8 X.9 X.10
1        ALGA    1 E. coli  166847  60754         0       0         0        0  60754 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA
2 GRASShopPER    2 E. coli    3200    288      2577    2577         0        0    291 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA
3     MEGAHIT    3 E. coli  284869 124164      2910    2910         0        0 124164 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA
4    Platanus    4 E. coli   58133  12333      1811    1811         0        0  12333 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA
5  Readjoiner    5 E. coli     538      -       297     297         0        0     NA NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA
6       SAGE2    6 E. coli   24062   3350         0       0         0        0   3350 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA
  X.11 X.12 X.13 X.14 X.15 X.16 X.17 X.18 X.19 X.20 X.21 X.22 X.23 X.24 X.25 X.26 X.27 X.28 X.29 X.30 X.31 X.32 X.33
1   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
2   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
3   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
4   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
5   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
6   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
> data = (data[1:60, 1:10])
```



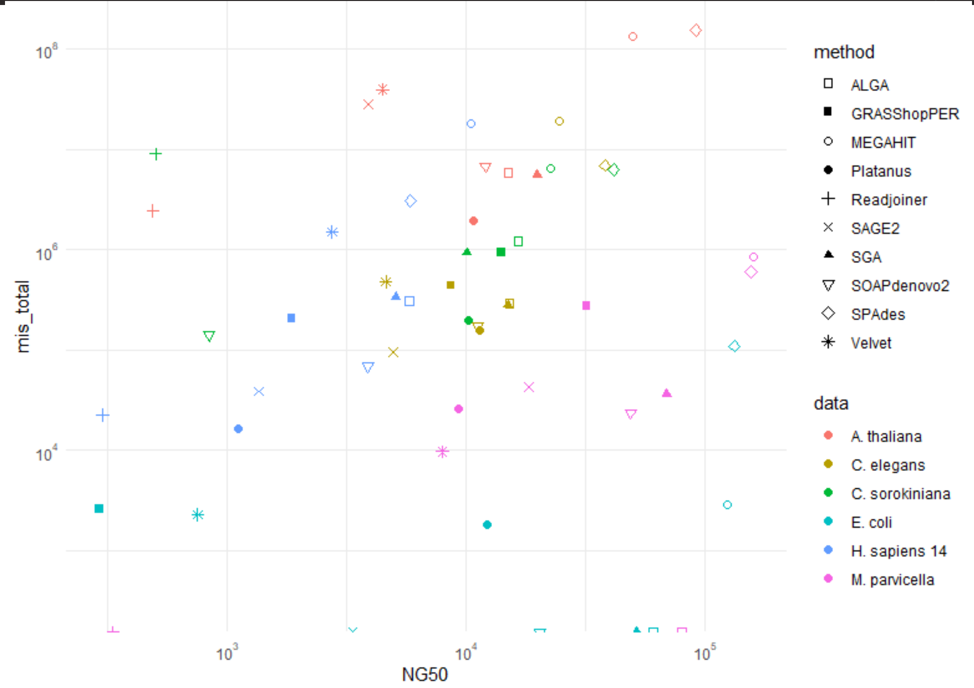
```
data = (data[1:60, 1:10])
head(data)

data$data = as.factor(data$data)
```

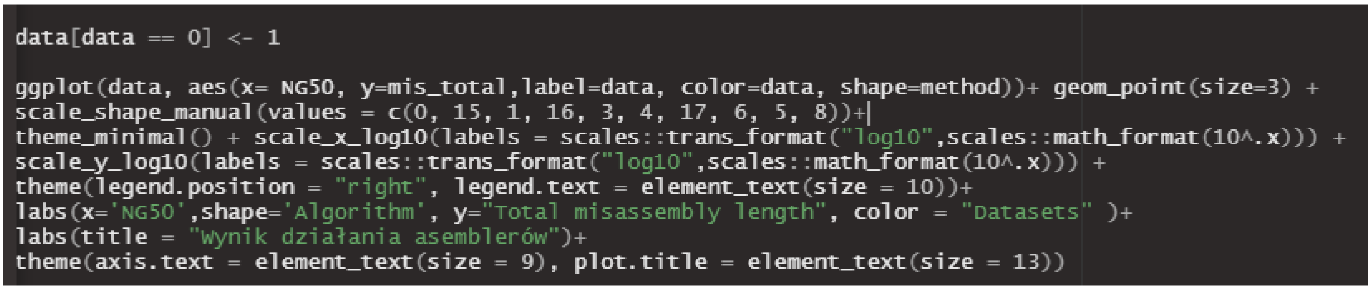


```
> head(data)
       method type    data l_align  NGA50 mis_total mis_c_l unaligned partialy   NG50
1        ALGA    1 E. coli  166847  60754         0       0         0        0  60754
2 GRASShopPER    2 E. coli    3200    288      2577    2577         0        0    291
3     MEGAHIT    3 E. coli  284869 124164      2910    2910         0        0 124164
4    Platanus    4 E. coli   58133  12333      1811    1811         0        0  12333
5  Readjoiner    5 E. coli     538      -       297     297         0        0     NA
6       SAGE2    6 E. coli   24062   3350         0       0         0        0   3350
```

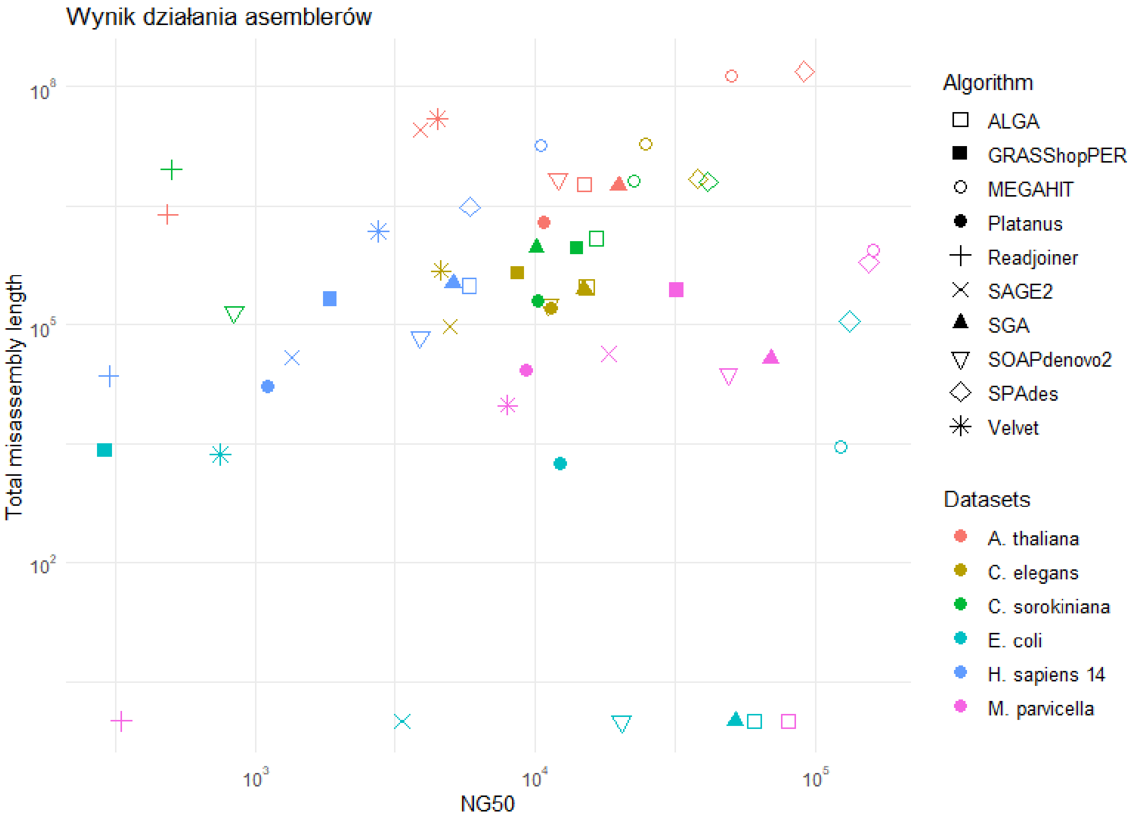


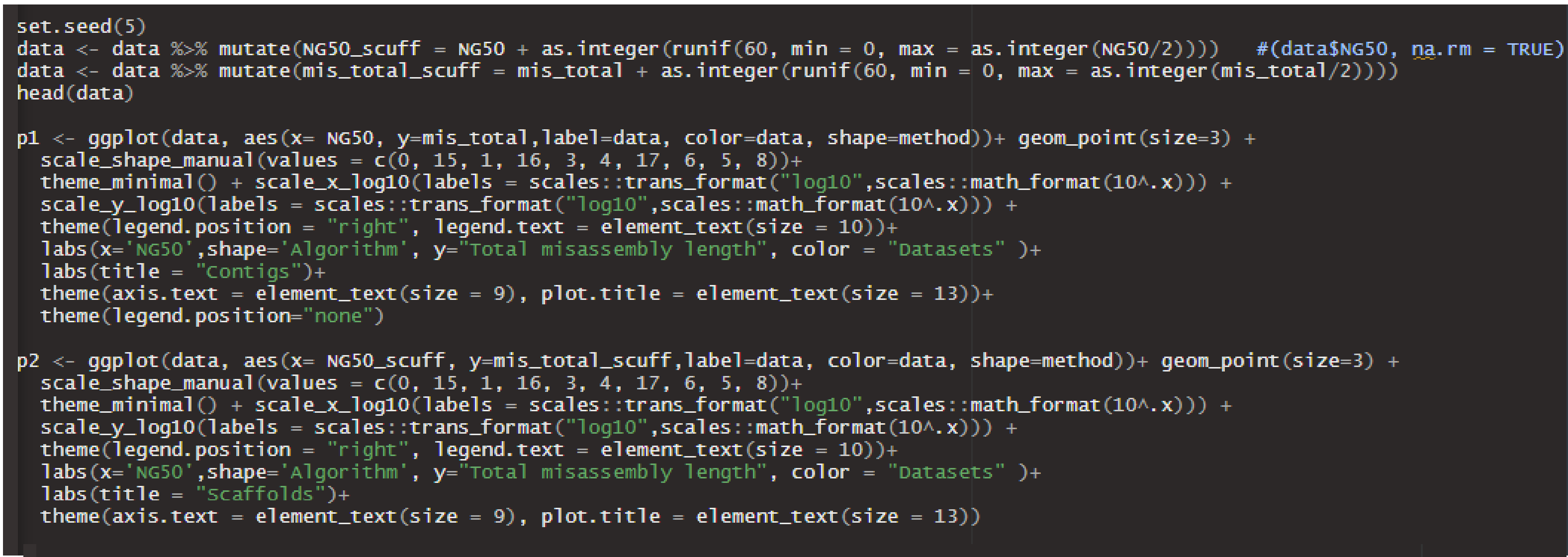
```
ggplot(data, aes(x= NG50, y=mis_total,label=data, color=data, shape=method))+ geom_point(size=2) +
scale_shape_manual(values = c(0, 15, 1, 16, 3, 4, 17, 6, 5, 8))+
theme_minimal() + scale_x_log10(labels = scales::trans_format("log10",scales::math_format(10^.x))) +
scale_y_log10(labels = scales::trans_format("log10",scales::math_format(10^.x)))
```



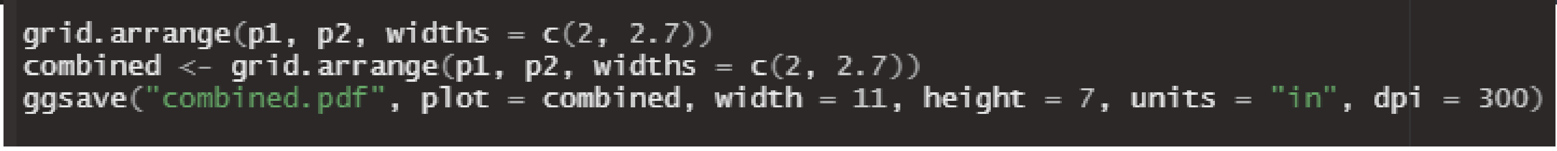


```
data[data == 0] <- 1

ggplot(data, aes(x= NG50, y=mis_total,label=data, color=data, shape=method))+ geom_point(size=3) +
scale_shape_manual(values = c(0, 15, 1, 16, 3, 4, 17, 6, 5, 8))+|
theme_minimal() + scale_x_log10(labels = scales::trans_format("log10",scales::math_format(10^.x))) +
scale_y_log10(labels = scales::trans_format("log10",scales::math_format(10^.x))) +
theme(legend.position = "right", legend.text = element_text(size = 10))+
labs(x='NG50',shape='Algorithm', y="Total misassembly length", color = "Datasets" )+
labs(title = "Wynik działania asemblerów")+
theme(axis.text = element_text(size = 9), plot.title = element_text(size = 13))
```

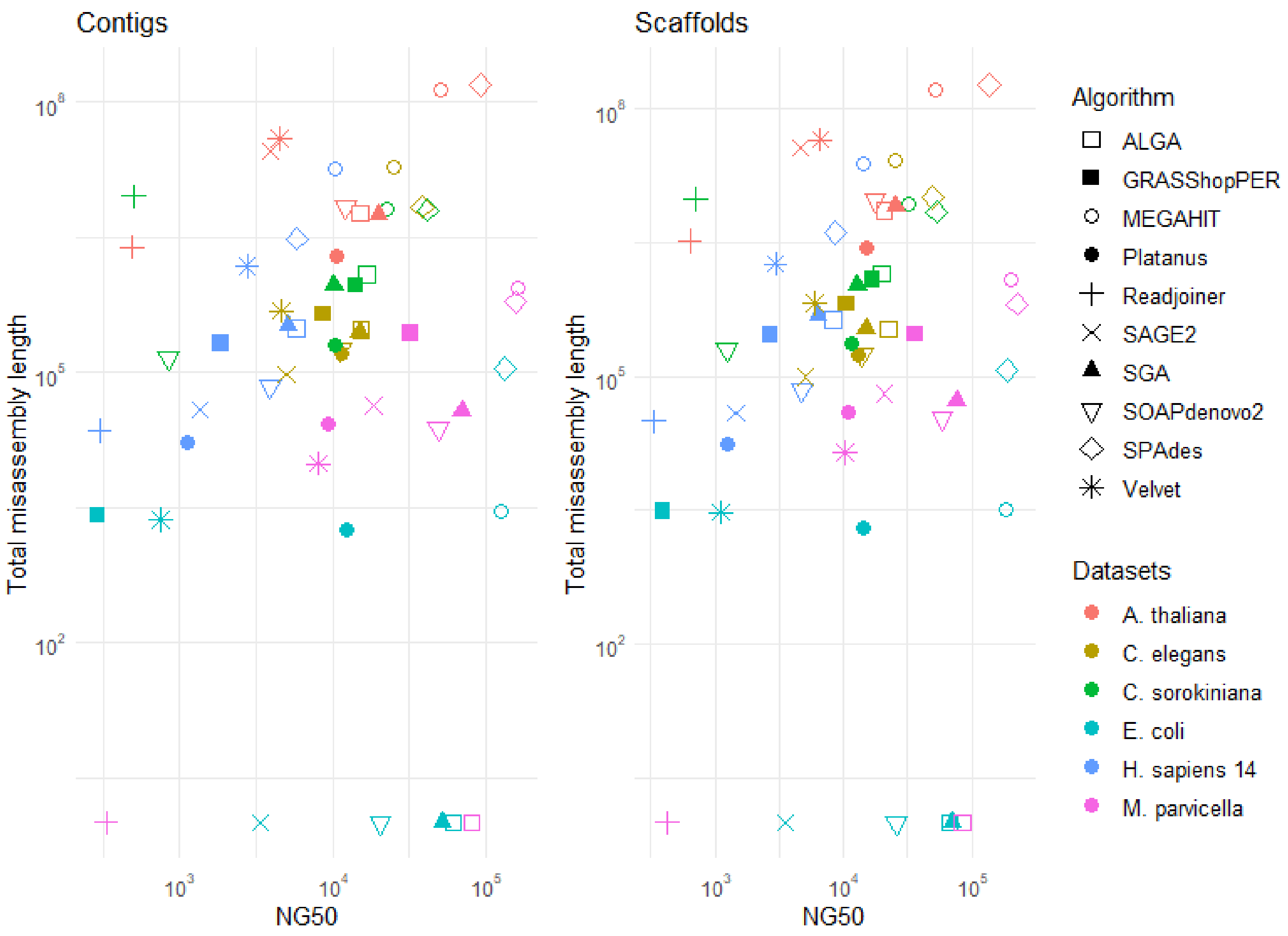




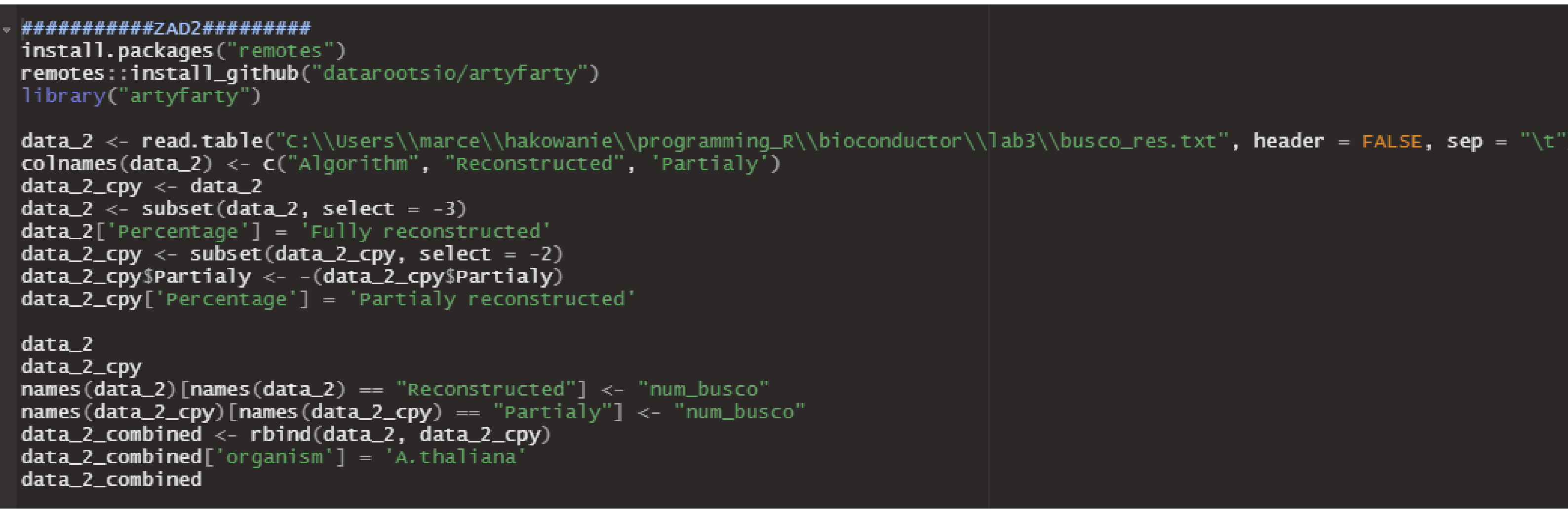
```
set.seed(5)
data <- data %>% mutate(NG50_scuff = NG50 + as.integer(runif(60, min = 0, max = as.integer(NG50/2))))   #(data$NG50, na.rm = TRUE)
data <- data %>% mutate(mis_total_scuff = mis_total + as.integer(runif(60, min = 0, max = as.integer(mis_total/2))))
head(data)

p1 <- ggplot(data, aes(x= NG50, y=mis_total,label=data, color=data, shape=method))+ geom_point(size=3) +
  scale_shape_manual(values = c(0, 15, 1, 16, 3, 4, 17, 6, 5, 8))+
  theme_minimal() + scale_x_log10(labels = scales::trans_format("log10",scales::math_format(10^.x))) +
  scale_y_log10(labels = scales::trans_format("log10",scales::math_format(10^.x))) +
  theme(legend.position = "right", legend.text = element_text(size = 10))+
  labs(x='NG50',shape='Algorithm', y="Total misassembly length", color = "Datasets" )+
  labs(title = "Contigs")+
  theme(axis.text = element_text(size = 9), plot.title = element_text(size = 13))+
  theme(legend.position="none")

p2 <- ggplot(data, aes(x= NG50_scuff, y=mis_total_scuff,label=data, color=data, shape=method))+ geom_point(size=3) +
  scale_shape_manual(values = c(0, 15, 1, 16, 3, 4, 17, 6, 5, 8))+
  theme_minimal() + scale_x_log10(labels = scales::trans_format("log10",scales::math_format(10^.x))) +
  scale_y_log10(labels = scales::trans_format("log10",scales::math_format(10^.x))) +
  theme(legend.position = "right", legend.text = element_text(size = 10))+
  labs(x='NG50',shape='Algorithm', y="Total misassembly length", color = "Datasets" )+
  labs(title = "Scaffolds")+
  theme(axis.text = element_text(size = 9), plot.title = element_text(size = 13))
```



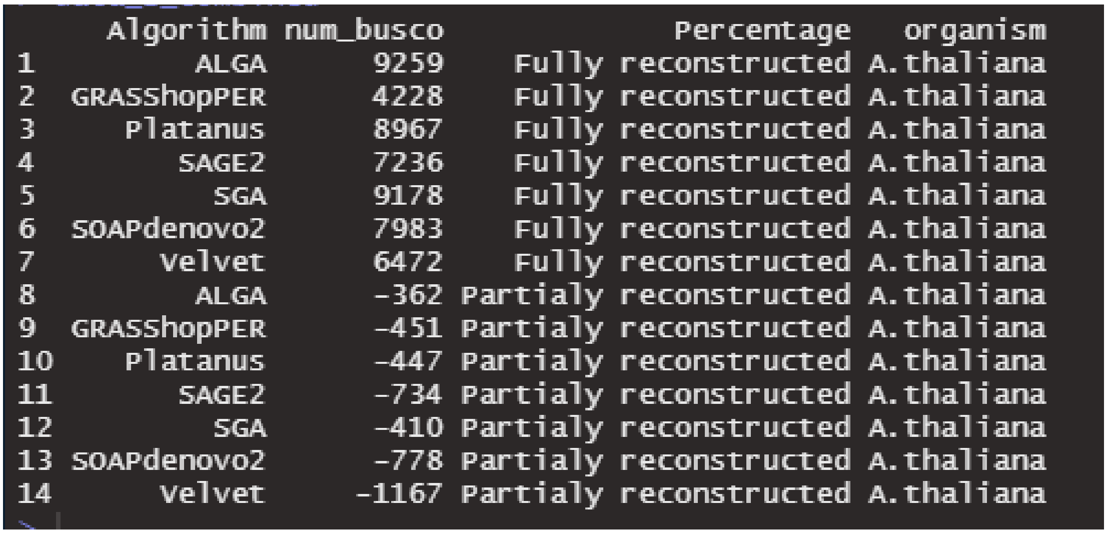
```
grid.arrange(p1, p2, widths = c(2, 2.7))
combined <- grid.arrange(p1, p2, widths = c(2, 2.7))
ggsave("combined.pdf", plot = combined, width = 11, height = 7, units = "in", dpi = 300)
```



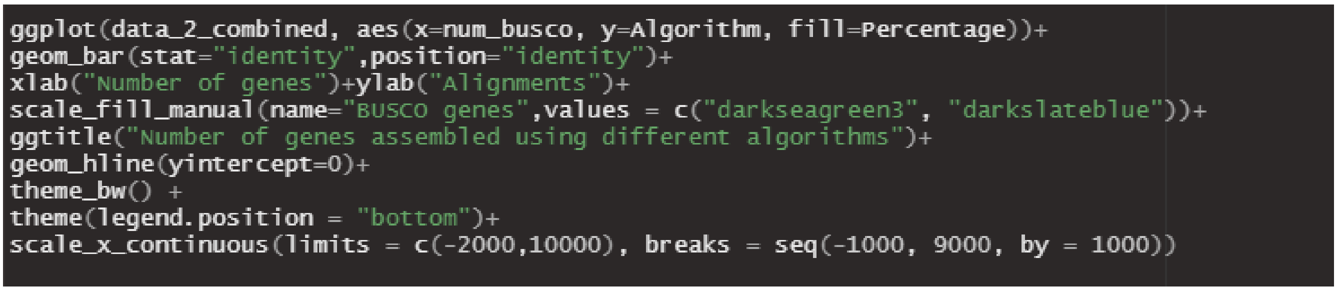
Z2:



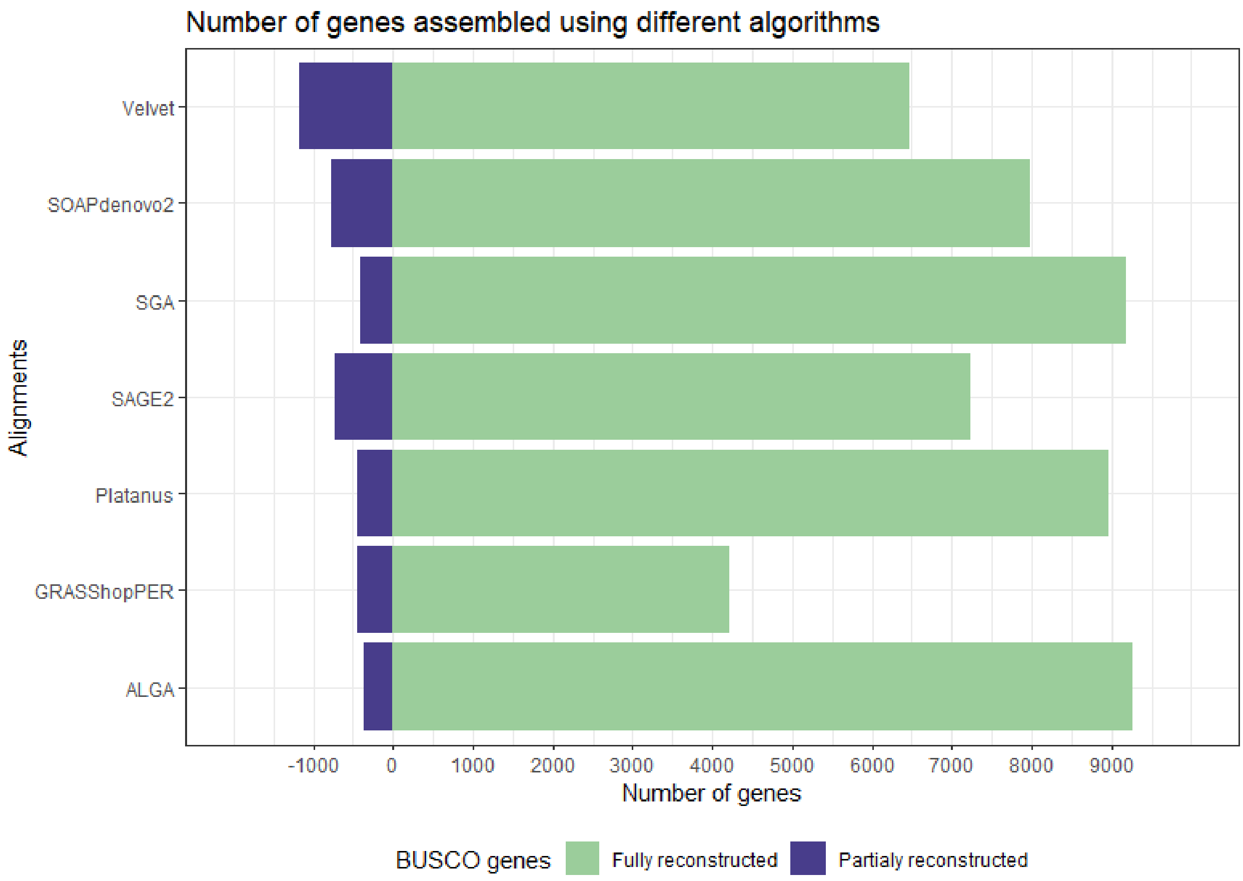
```
##########ZAD2#########
install.packages("remotes")
remotes::install_github("datarootsio/artyfarty")
library("artyfarty")

data_2 <- read.table("C:\\Users\\marce\\hakowanie\\programming_R\\bioconductor\\lab3\\busco_res.txt", header = FALSE, sep = "\t"
colnames(data_2) <- c("Algorithm", "Reconstructed", 'Partialy')
data_2_cpy <- data_2
data_2 <- subset(data_2, select = -3)
data_2['Percentage'] = 'Fully reconstructed'
data_2_cpy <- subset(data_2_cpy, select = -2)
data_2_cpy$Partialy <- -(data_2_cpy$Partialy)
data_2_cpy['Percentage'] = 'Partialy reconstructed'

data_2
data_2_cpy
names(data_2)[names(data_2) == "Reconstructed"] <- "num_busco"
names(data_2_cpy)[names(data_2_cpy) == "Partialy"] <- "num_busco"
data_2_combined <- rbind(data_2, data_2_cpy)
data_2_combined['organism'] = 'A.thaliana'
data_2_combined
```

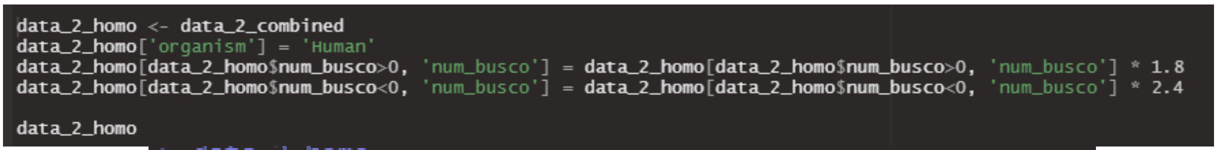


```
     Algorithm num_busco            Percentage   organism
1         ALGA      9259    Fully reconstructed A.thaliana
2   GRASShopPER     4228    Fully reconstructed A.thaliana
3      Platanus     8967    Fully reconstructed A.thaliana
4         SAGE2     7236    Fully reconstructed A.thaliana
5           SGA     9178    Fully reconstructed A.thaliana
6   SOAPdenovo2     7983    Fully reconstructed A.thaliana
7        Velvet     6472    Fully reconstructed A.thaliana
8         ALGA      -362 Partialy reconstructed A.thaliana
9   GRASShopPER     -451 Partialy reconstructed A.thaliana
10     Platanus     -447 Partialy reconstructed A.thaliana
11        SAGE2     -734 Partialy reconstructed A.thaliana
12          SGA     -410 Partialy reconstructed A.thaliana
13 SOAPdenovo2      -778 Partialy reconstructed A.thaliana
14       Velvet    -1167 Partialy reconstructed A.thaliana
```

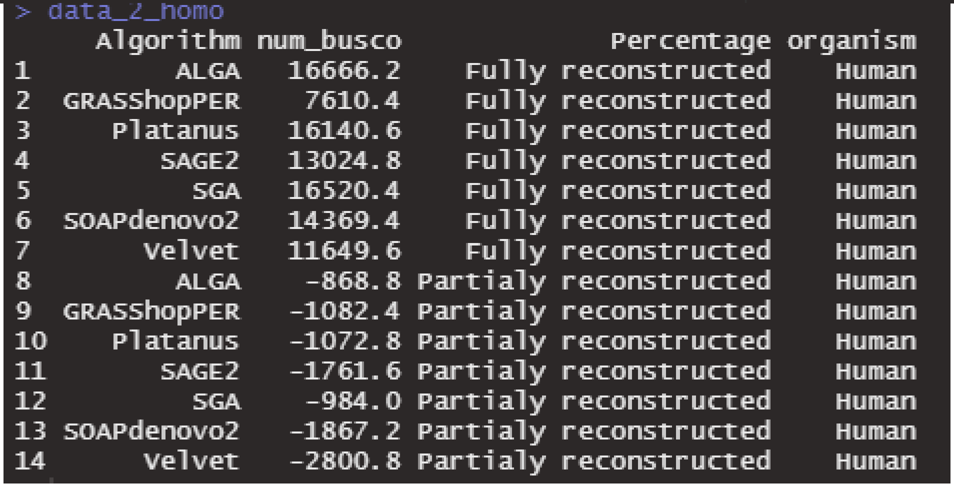


```
ggplot(data_2_combined, aes(x=num_busco, y=Algorithm, fill=Percentage))+
geom_bar(stat="identity",position="identity")+
xlab("Number of genes")+ylab("Alignments")+
scale_fill_manual(name="BUSCO genes",values = c("darkseagreen3", "darkslateblue"))+
ggtitle("Number of genes assembled using different algorithms")+
geom_hline(yintercept=0)+
theme_bw() +
theme(legend.position = "bottom")+
scale_x_continuous(limits = c(-2000,10000), breaks = seq(-1000, 9000, by = 1000))
```

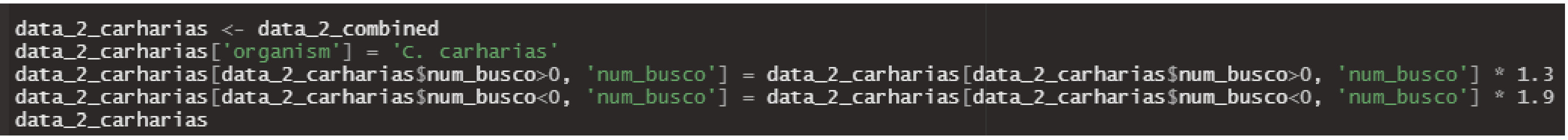




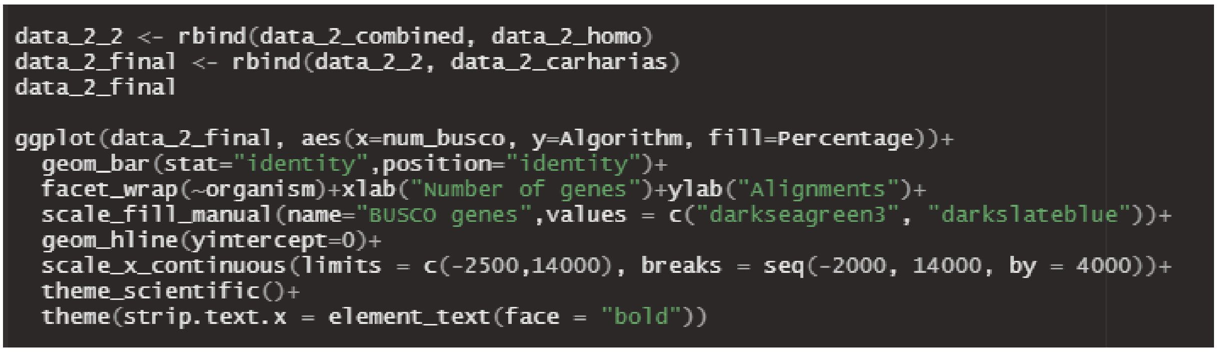
```
data_2_homo <- data_2_combined
data_2_homo['organism'] = 'Human'
data_2_homo[data_2_homo$num_busco>0, 'num_busco'] = data_2_homo[data_2_homo$num_busco>0, 'num_busco'] * 1.8
data_2_homo[data_2_homo$num_busco<0, 'num_busco'] = data_2_homo[data_2_homo$num_busco<0, 'num_busco'] * 2.4

data_2_homo
```



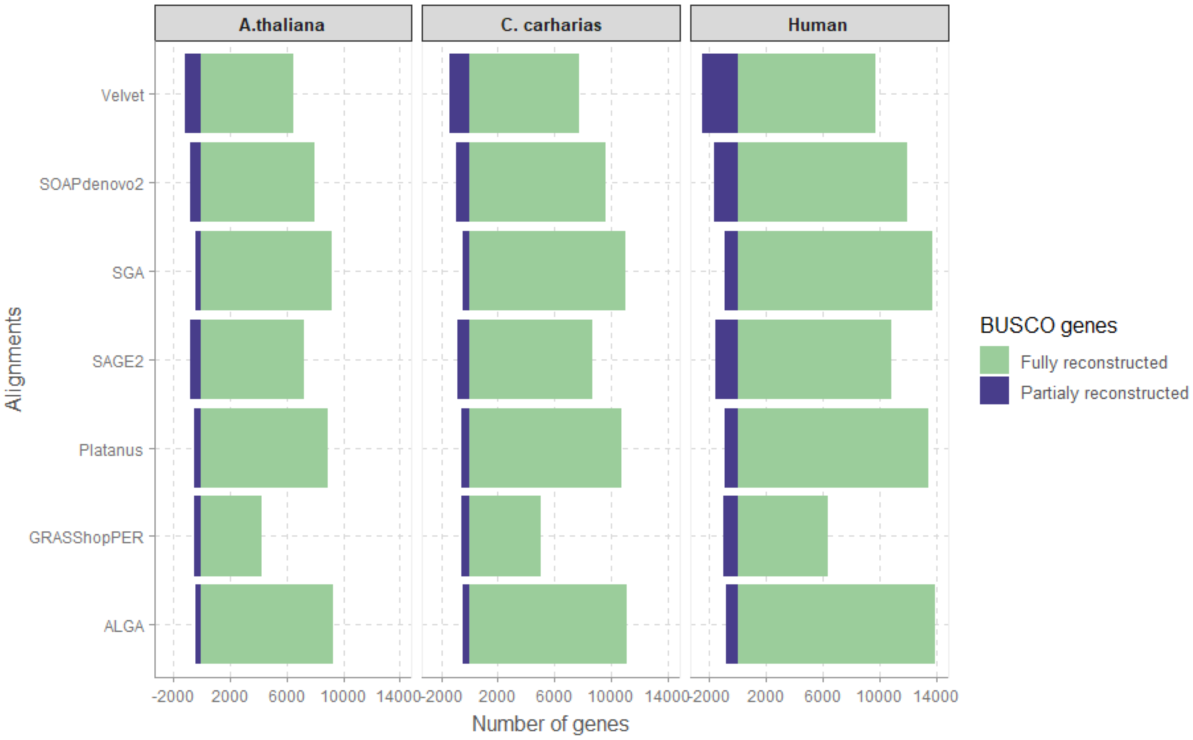
```
> data_2_homo
     Algorithm num_busco            Percentage organism
1         ALGA   16666.2   Fully reconstructed    Human
2  GRASShopPER    7610.4   Fully reconstructed    Human
3     Platanus   16140.6   Fully reconstructed    Human
4        SAGE2   13024.8   Fully reconstructed    Human
5          SGA   16520.4   Fully reconstructed    Human
6   SOAPdenovo2  14369.4   Fully reconstructed    Human
7       Velvet   11649.6   Fully reconstructed    Human
8         ALGA    -868.8 Partialy reconstructed    Human
9  GRASShopPER   -1082.4 Partialy reconstructed    Human
10    Platanus   -1072.8 Partialy reconstructed    Human
11       SAGE2   -1761.6 Partialy reconstructed    Human
12         SGA    -984.0 Partialy reconstructed    Human
13 SOAPdenovo2   -1867.2 Partialy reconstructed    Human
14      Velvet   -2800.8 Partialy reconstructed    Human
```



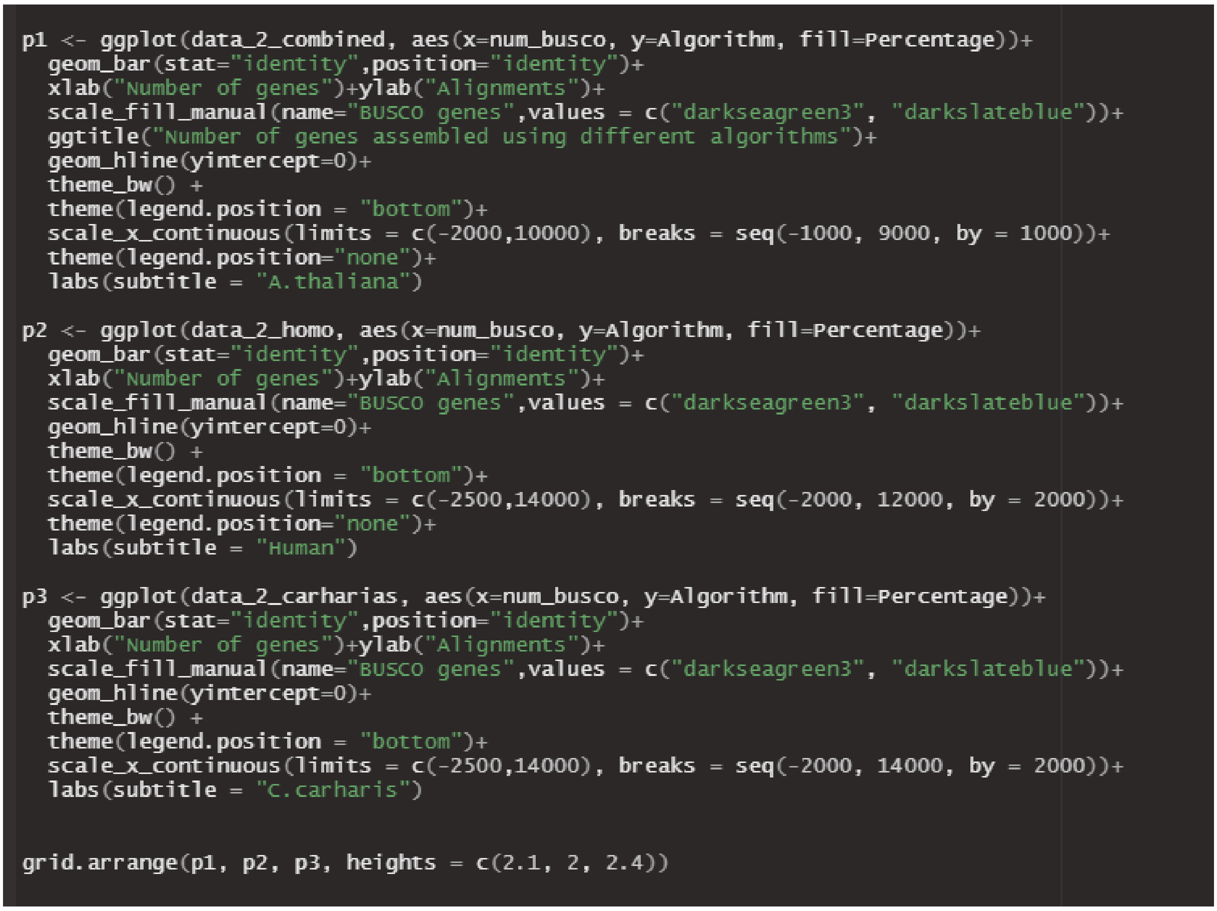
```
data_2_carharias <- data_2_combined
data_2_carharias['organism'] = 'C. carharias'
data_2_carharias[data_2_carharias$num_busco>0, 'num_busco'] = data_2_carharias[data_2_carharias$num_busco>0, 'num_busco'] * 1.3
data_2_carharias[data_2_carharias$num_busco<0, 'num_busco'] = data_2_carharias[data_2_carharias$num_busco<0, 'num_busco'] * 1.9
data_2_carharias
```



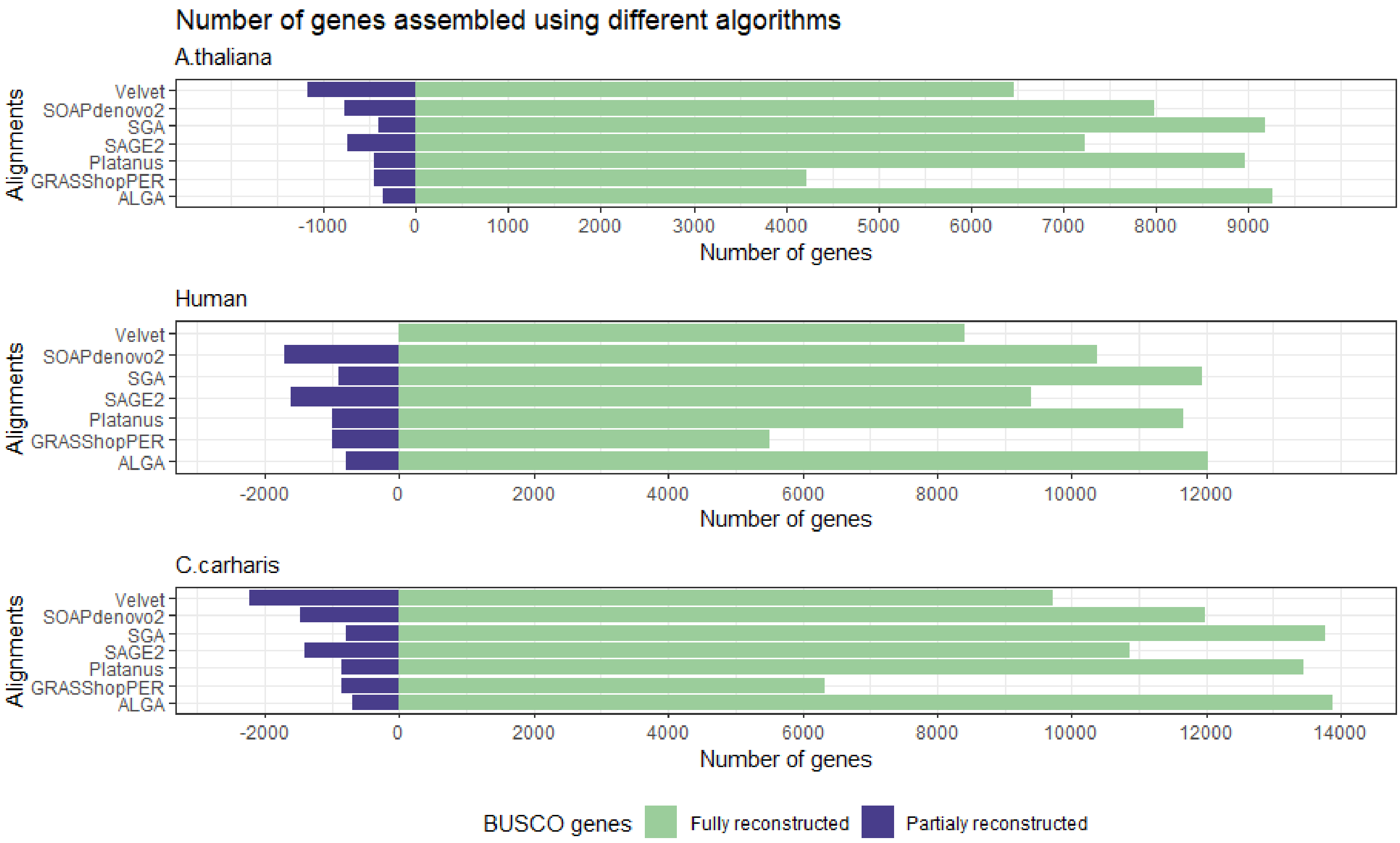
```
data_2_2 <- rbind(data_2_combined, data_2_homo)
data_2_final <- rbind(data_2_2, data_2_carharias)
data_2_final

ggplot(data_2_final, aes(x=num_busco, y=Algorithm, fill=Percentage))+
  geom_bar(stat="identity",position="identity")+
  facet_wrap(~organism)+xlab("Number of genes")+ylab("Alignments")+
  scale_fill_manual(name="BUSCO genes",values = c("darkseagreen3", "darkslateblue"))+
  geom_hline(yintercept=0)+
  scale_x_continuous(limits = c(-2500,14000), breaks = seq(-2000, 14000, by = 4000))+
  theme_scientific()+
  theme(strip.text.x = element_text(face = "bold"))
```



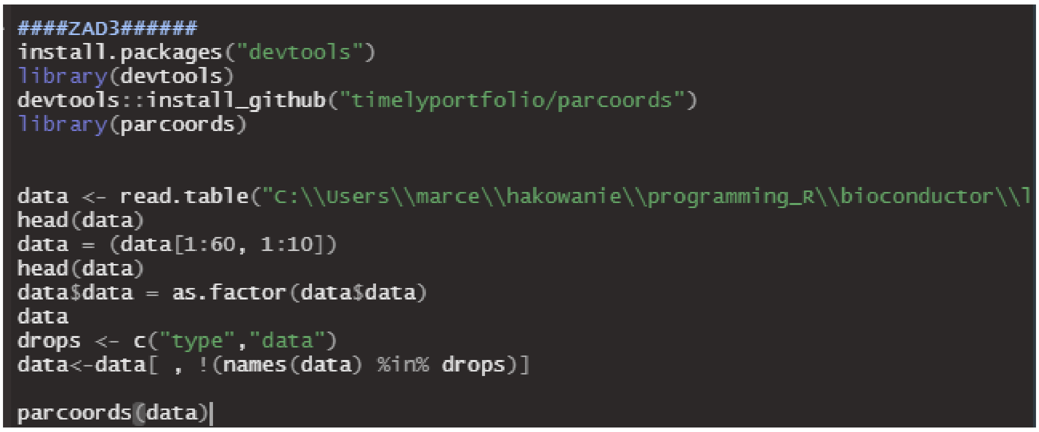
2 sposób:



```
p1 <- ggplot(data_2_combined, aes(x=num_busco, y=Algorithm, fill=Percentage))+
  geom_bar(stat="identity",position="identity")+
  xlab("Number of genes")+ylab("Alignments")+
  scale_fill_manual(name="BUSCO genes",values = c("darkseagreen3", "darkslateblue"))+
  ggtitle("Number of genes assembled using different algorithms")+
  geom_hline(yintercept=0)+
  theme_bw() +
  theme(legend.position = "bottom")+
  scale_x_continuous(limits = c(-2000,10000), breaks = seq(-1000, 9000, by = 1000))+
  theme(legend.position="none")+
  labs(subtitle = "A.thaliana")

p2 <- ggplot(data_2_homo, aes(x=num_busco, y=Algorithm, fill=Percentage))+
  geom_bar(stat="identity",position="identity")+
  xlab("Number of genes")+ylab("Alignments")+
  scale_fill_manual(name="BUSCO genes",values = c("darkseagreen3", "darkslateblue"))+
  geom_hline(yintercept=0)+
  theme_bw() +
  theme(legend.position = "bottom")+
  scale_x_continuous(limits = c(-2500,14000), breaks = seq(-2000, 12000, by = 2000))+
  theme(legend.position="none")+
  labs(subtitle = "Human")

p3 <- ggplot(data_2_carharias, aes(x=num_busco, y=Algorithm, fill=Percentage))+
  geom_bar(stat="identity",position="identity")+
  xlab("Number of genes")+ylab("Alignments")+
  scale_fill_manual(name="BUSCO genes",values = c("darkseagreen3", "darkslateblue"))+
  geom_hline(yintercept=0)+
  theme_bw() +
  theme(legend.position = "bottom")+
  scale_x_continuous(limits = c(-2500,14000), breaks = seq(-2000, 14000, by = 2000))+
  labs(subtitle = "C.carharis")


grid.arrange(p1, p2, p3, heights = c(2.1, 2, 2.4))
```



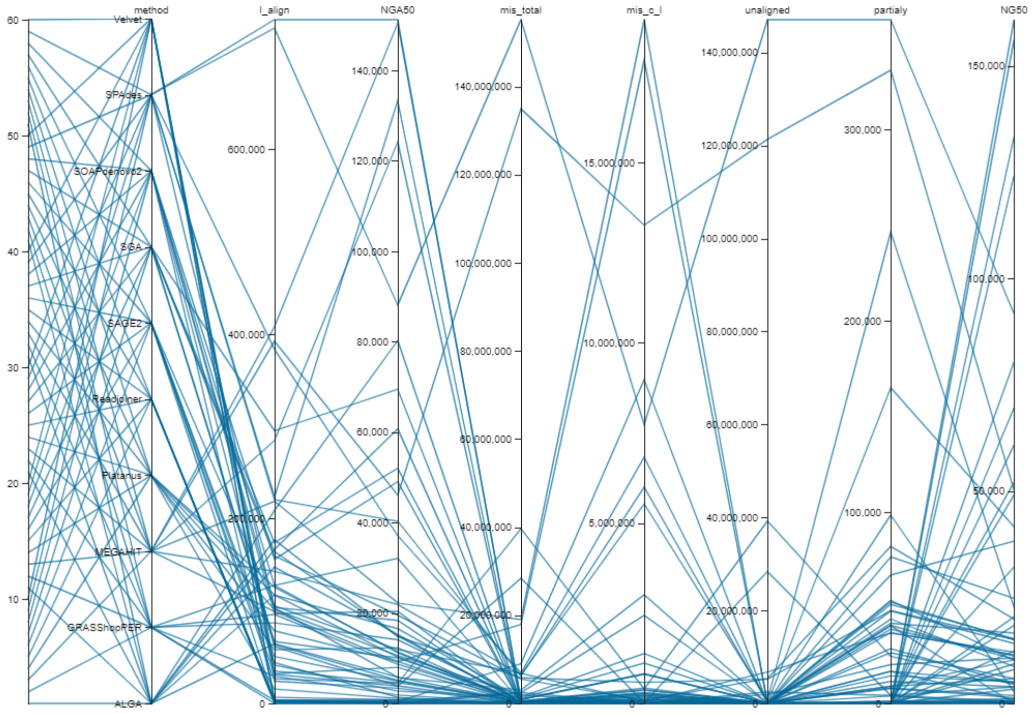
Zad 3.

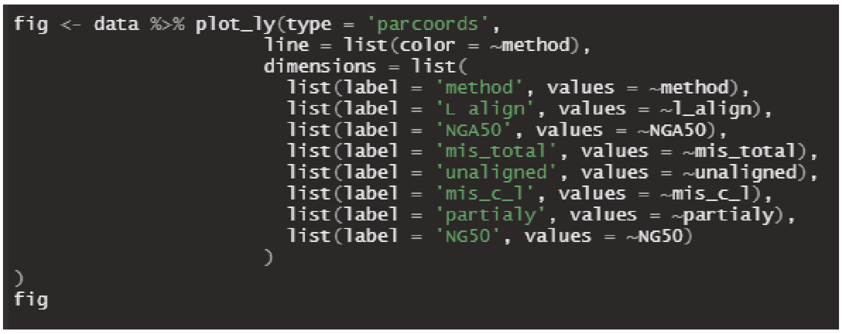


```
####ZAD3######
install.packages("devtools")
library(devtools)
devtools::install_github("timelyportfolio/parcoords")
library(parcoords)


data <- read.table("C:\\Users\\marce\\hakowanie\\programming_R\\bioconductor\\l
head(data)
data = (data[1:60, 1:10])
head(data)
data$data = as.factor(data$data)
data
drops <- c("type","data")
data<-data[ , !(names(data) %in% drops)]

parcoords(data)
```

Z uwagi na wygląd wykresu i dodatkową warstwę której nie dało się usunąć mimo manualnej konstrukcji warstw, zdecydowano się na użycie pakietu ggparcoord() (niżej)

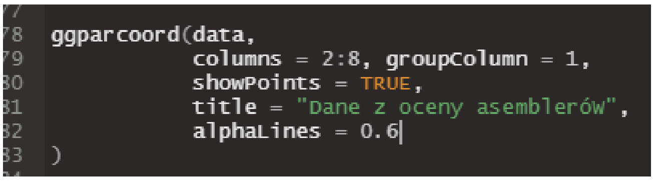
Parcoord:



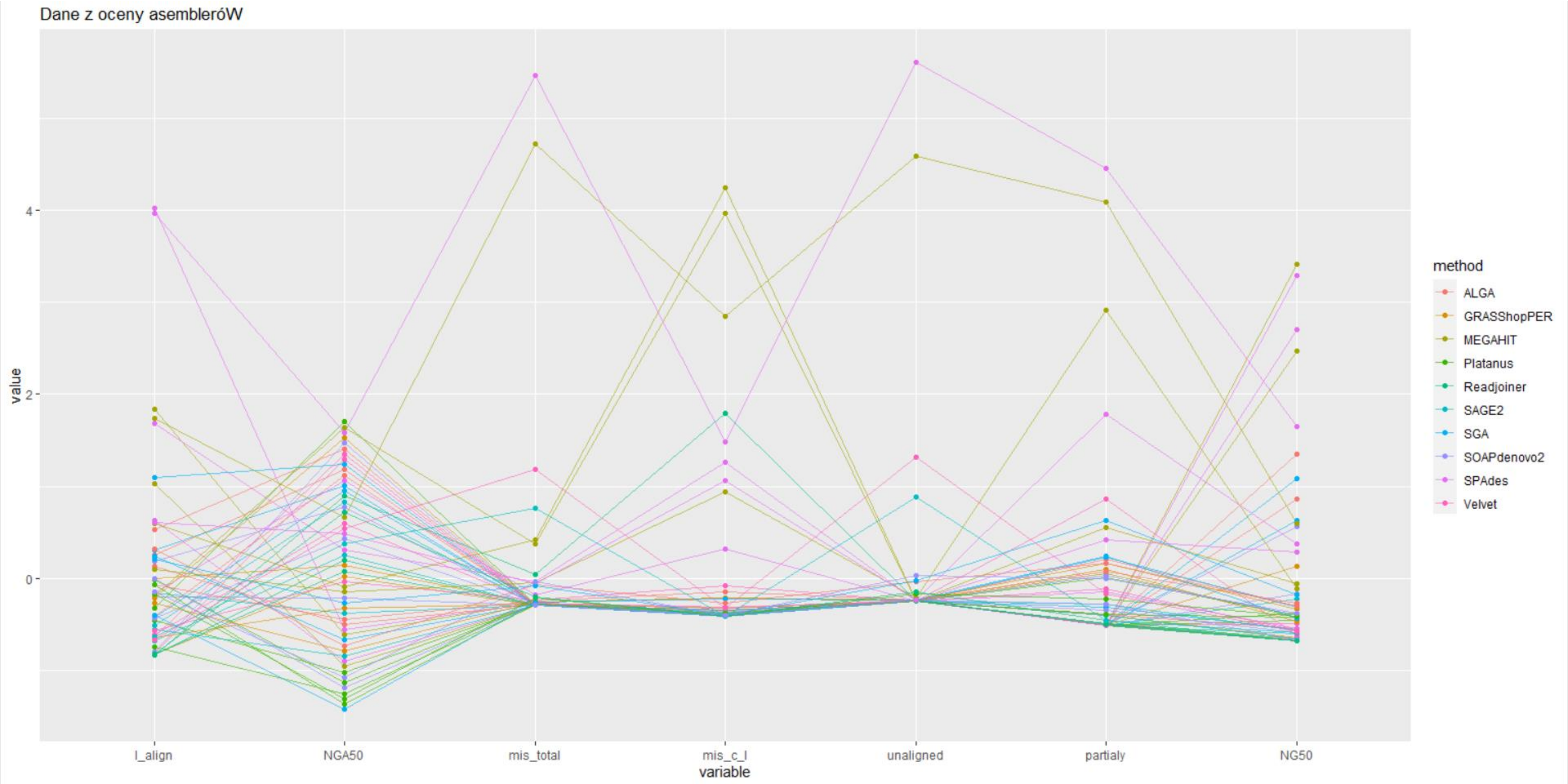


```
fig <- data %>% plot_ly(type = 'parcoords',
                        line = list(color = ~method),
                        dimensions = list(
                          list(label = 'method', values = ~method),
                          list(label = 'L align', values = ~l_align),
                          list(label = 'NGA50', values = ~NGA50),
                          list(label = 'mis_total', values = ~mis_total),
                          list(label = 'unaligned', values = ~unaligned),
                          list(label = 'mis_c_l', values = ~mis_c_l),
                          list(label = 'partialy', values = ~partialy),
                          list(label = 'NG50', values = ~NG50)
                        )
)
fig
```

Ggparcoord:



```
ggparcoord(data,
           columns = 2:8, groupColumn = 1,
           showPoints = TRUE,
           title = "Dane z oceny asemblerów",
           alphaLines = 0.6
)
```


Dane z oceny asemblerów
value
l_align
NGA50
mis_total
mis_c_l
unaligned
partialy
NG50
variable
method
ALGA
GRASShopPER
MEGAHIT
Platanus
Readjoiner
SAGE2
SGA
SOAPdenovo2
SPAdes
Velvet