Challenge

Lavinia Lavin

6/29/2022

Import od csv data sets

```
install.packages("data.table")
library(data.table)
setwd("~/Desktop/Ejemplos R/Genomica/Challenge")
files<- list.files(pattern = ".csv")
temp<- lapply(files, fread, sep=",")
data <- rbindlist(temp, fill = TRUE)
write.csv(data, file="", row.names=F)
```

Import vcf files

```
install.packages("vcfR")
library(vcfR)
install.packages("data.table")
library(data.table)
setwd("~/Desktop/Ejemplos R/Genomica/Challenge")
files<- list.files(pattern = ".vcf")
vcf1<- lapply(files, fread, sep=",")
vcf <- rbindlist(temp)
read.vcfR(vcf, file="", row.names=F)

install.packages("dplyr")
library(dplyr)
```

Combine csv and vcf data sets in one

```
data_set<- rbind(data, vcf)
data_set
```

Upload data from coordination tsv data set

```
coor<-read_tsv("coordination.txt")
```

Merge data set with coor

```
df<-rbind(data_set, coor, fill=TRUE)
view(df)
```

Call libraries to perform k-mean cluster

```
install.packages("stats")
install.packages("dplyr")
install.packages("gglopt2")
install.packages("ggfortify")
library(stats)
library(dplyr)
library(ggplot2)
library(ggfortify)
View(df)
```

Determine the number of columns used for clusters

```
mydata = select(df,c(1, 2, 3, 4, 5, 6, 7))
View(mydata)
```

wssplot plot generated to illustrate the number of cluster needed for data base

```
wssplot<-function(df, nc=15 , seed=1234)
{
  wss<-(nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i]<-sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of clusters", ylab="Within groups sum of squares")
}
```

Plot to know the number of cluster following the trend in the graph

```
wssplot(mydata)
```