Inspect a list of indexes

This notebook is designed to inspect a list of index.

### Loading package

library(stringr)  
library(vwr)  
library(reshape2)  
library(ggplot2)   
library(viridis)  
library(dplyr)

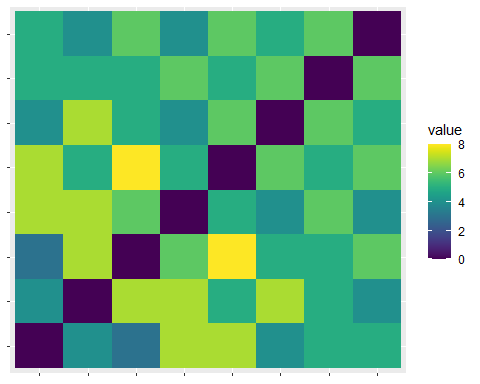
### Loading data

User should prepare a file with no header, all capitalized, same length, with A T C or G, only as the input.

input.file.path <- params$input\_path  
input.file <- read.table(input.file.path,quote="\"", comment.char="")  
input.index.seq <- input.file$V1

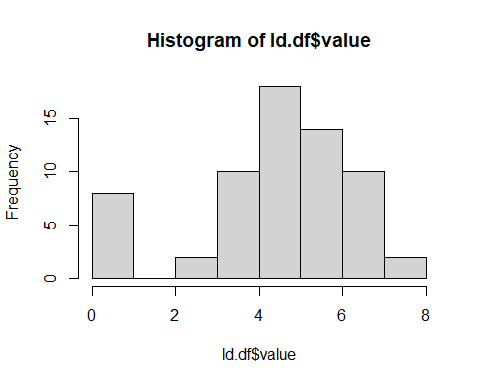
### Generating Heatmap

ld\_mtx <- sapply(input.index.seq,levenshtein.distance,targets = input.index.seq)  
ld.df <- melt(ld\_mtx)  
plot.heatmap <- ggplot(ld.df,aes(Var1,Var2))  
plot.heatmap + geom\_tile(aes(fill=value)) +  
 theme(axis.text.x = element\_blank(),axis.text.y = element\_blank(),axis.title.x = element\_blank(),axis.title.y = element\_blank())+  
 scale\_fill\_viridis()



### Statistic on the edit distance

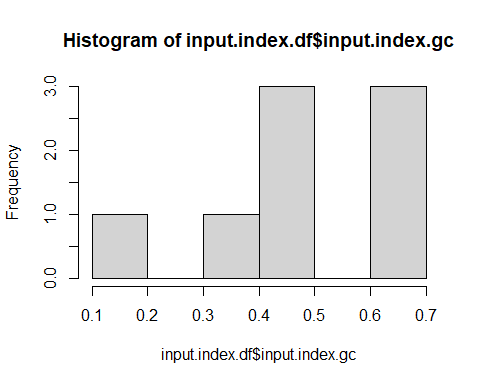
hist(ld.df$value)



ld.df %>% filter(value == 0) %>% nrow() -> num.ld.zero   
ld.df %>% filter(value == 1) %>% nrow() -> num.ld.one   
ld.df %>% filter(value == 2) %>% nrow() -> num.ld.two  
ld.df %>% filter(value == 3) %>% nrow() -> num.ld.three

### Other quality metrics on the index, GC, repeats

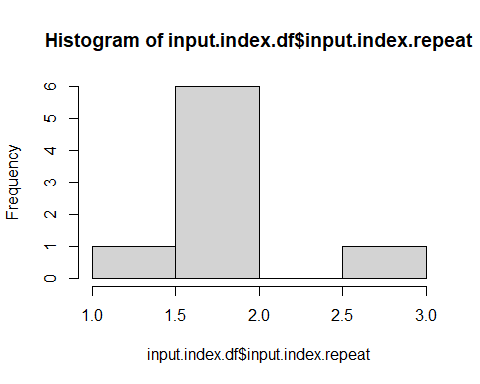
check\_gc <- function(inputString){  
 full.length = str\_length(inputString)   
 G.count <- str\_count(inputString, "G")  
 C.count <- str\_count(inputString,"C")  
 return((G.count+C.count)/full.length)  
}  
check\_repeat <- function(inputString) {  
   
 possible\_choice <- c("A", "T", "C", "G")  
 full.length = str\_length(inputString)  
 chr.list = str\_split(inputString, pattern = "", simplify = TRUE)  
 max.counter =1  
 for (base in possible\_choice) {  
 counter = 1  
 for (i in seq(length(chr.list) - 1)) {  
 if (chr.list[i] != chr.list[i + 1]) {  
 next  
 } else if(base == chr.list[i]){  
 counter = counter + 1  
   
 }  
 if(counter >= max.counter){  
 max.counter = counter  
 }  
 }  
  
 }  
 return(max.counter)  
}  
input.index.gc <- sapply(input.index.seq, check\_gc)  
input.index.repeat <- sapply(input.index.seq, check\_repeat)  
input.index.df <-data.frame(input.index.gc,input.index.repeat)  
  
hist(input.index.df$input.index.gc)



summary(input.index.df$input.index.gc)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1250 0.4688 0.5000 0.4844 0.6250 0.6250

hist(input.index.df$input.index.repeat)



summary(input.index.df$input.index.repeat)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 2 2 2 2 3

input.index.df %>% filter(input.index.repeat >=4)

## [1] input.index.gc input.index.repeat  
## <0 rows> (or 0-length row.names)