{

library(BSgenome.Hsapiens.UCSC.hg19)

library(SomaticSignatures)

library(GenomicRanges)

library(readr)

library(ggplot2)

library(dplyr)

library(VariantAnnotation)

library(tidyr)

library(csv)

library(readxl)

}

{

b<-'Lung'

inputfile <- paste0('/home/san/halinejad/Desktop/Masroor Bayati/DeepCancer project/Data/ICGC/',b,'/simple\_somatic\_mutation.open.tsv')

Cancer<- data.frame(read\_tsv(inputfile,col\_names = TRUE))

gene\_list <- read\_excel('/home/san/halinejad/Desktop/Masroor Bayati/DeepCancer project/Data/annotation/FANTOM5\_gene\_list.xlsx')

annot <- as.data.frame(gene\_list[,c(1,3,4,5,6,8)])

colnames(annot) <- c('id','chromosome','start','end','strand','geneClass')

gene<- subset(annot, annot$geneClass =='coding\_mRNA')

lncRNA <- subset(annot, annot$geneClass =='lncRNA')

pseudogene<- subset(annot, annot$geneClass =='pseudogene')

sense\_overlap\_RNA<- subset(annot, annot$geneClass =='sense\_overlap\_RNA')

short\_ncRNA<- subset(annot, annot$geneClass =='short\_ncRNA')

small\_RNA<- subset(annot, annot$geneClass =='small\_RNA')

structural\_RNA<- subset(annot, annot$geneClass =='structural\_RNA')

uncertain\_coding<- subset(annot, annot$geneClass =='uncertain\_coding')

tmp<- as.data.frame(Cancer[,c(1,5,9,10,11,12,16,17)])

tmp1<- data.frame(subset(tmp, tmp$icgc\_mutation\_id !="NA" ))

tmp3<- data.frame(subset(tmp1, nchar(tmp1$mutated\_from\_allele)==1 & nchar(tmp1$mutated\_to\_allele)==1

& tmp1$mutated\_from\_allele !='-'& tmp1$mutated\_to\_allele != "-"

& tmp1$mutated\_from\_allele !='\_'& tmp1$mutated\_to\_allele != "\_"

& tmp1$chromosome != "MT" & tmp1$chromosome != "M") )

tmp2<- tmp3[!duplicated(tmp3),]

tmp2$chromosome\_strand[which(tmp2$chromosome\_strand == '1')] <- '+'

tmp2$chromosome\_strand[which(tmp2$chromosome\_strand == '2')] <- '-'

tmp2$chromosome <- paste0("chr",tmp2$chromosome)

gr <- makeGRangesFromDataFrame(tmp2,keep.extra.columns = T,

seqnames.field = 'chromosome',

start.field='chromosome\_start',

end.field = 'chromosome\_end',

strand.field = 'chromosome\_strand')

idx <- match(c('mutated\_from\_allele','mutated\_to\_allele','icgc\_sample\_id'),names(mcols(gr)))

mcols(gr) <- cbind(mcols(gr)[idx],mcols(gr)[-idx])

vr <- makeVRangesFromGRanges(gr,ref.field='mutated\_from\_allele',

alt.field='mutated\_to\_allele',

sampleNames.field = 'icgc\_sample\_id',

keep.extra.columns = T)

vr <- mutationContext(vr,Hsapiens)

variations <- data.frame(icgc\_samlpe\_id = mcols(gr)$icgc\_sample\_id,

chromosome = as.character(seqnames(vr)),

position = start(vr),

strand = as.character(strand(gr)),

motif = paste0(as.character(mcols(vr)$alteration),

'-',as.character(mcols(vr)$context)))

clear\_var <- subset( variations, !grepl("n", variations$motif))

x<- as.data.frame(clear\_var[,c(1,5,2,3)],header=T)

}

{

df1<-structure(x)

df2<-structure(as.data.frame(gene[,c(1,2,3,4)]) )

q<- df1 %>% inner\_join(df2, "chromosome") %>%

mutate(geneID\_motif = paste(id, motif, sep = ","),

n = if\_else(position >= start & position <= end, 1, 0)) %>%

select(icgc\_samlpe\_id, geneID\_motif, n) %>%

group\_by(icgc\_samlpe\_id, geneID\_motif) %>%

summarise(n = sum(n)) %>%

spread(key = geneID\_motif, value = n, fill = 0)

q0= q[,!grepl("N",colnames(q))]

row\_name<- q0$icgc\_samlpe\_id

q0 <- q0[,-1]

rownames(q0)<- row\_name

col\_name<-colnames(q0)

q0<-colSums(q0)

q0<-as.data.frame(q0)

k<-vector(max(q0), mode='list')

for(i in 1:max(q0)){

ttt<-subset(q0,q0>=i)

print(i)

w<- paste0(b," gene\_bigger than ",as.character(i))

barplot(t(ttt), main = w) # create the barplot

k[[i]] <- recordPlot()

}

graphics.off()

outputfile <- paste0('/home/san/halinejad/Desktop/Masroor Bayati/DeepCancer project/Dashti/',b,'\_barplot\_gene.pdf')

pdf(outputfile, onefile=TRUE)

for (my.plot in k) {

replayPlot(my.plot)

}

graphics.off()

}

{

df1<-structure(x)

df2<-structure(as.data.frame(lncRNA[,c(1,2,3,4)]) )

q<- df1 %>% inner\_join(df2, "chromosome") %>%

mutate(geneID\_motif = paste(id, motif, sep = ","),

n = if\_else(position >= start & position <= end, 1, 0)) %>%

select(icgc\_samlpe\_id, geneID\_motif, n) %>%

group\_by(icgc\_samlpe\_id, geneID\_motif) %>%

summarise(n = sum(n)) %>%

spread(key = geneID\_motif, value = n, fill = 0)

q0= q[,!grepl("N",colnames(q))]

row\_name<- q0$icgc\_samlpe\_id

q0 <- q0[,-1]

rownames(q0)<- row\_name

col\_name<-colnames(q0)

q0<-colSums(q0)

q0<-as.data.frame(q0)

k<-vector(max(q0), mode='list')

for(i in 1:max(q0)){

ttt<-subset(q0,q0>=i)

print(i)

w<- paste0(b," gene\_bigger than ",as.character(i))

barplot(t(ttt), main = w) # create the barplot

k[[i]] <- recordPlot()

}

graphics.off()

outputfile <- paste0('/home/san/halinejad/Desktop/Masroor Bayati/DeepCancer project/Dashti/',b,'\_barplot\_lncRNA.pdf')

pdf(outputfile, onefile=TRUE)

for (my.plot in k) {

replayPlot(my.plot)

}

graphics.off()

}

{

df1<-structure(x)

df2<-structure(as.data.frame(pseudogene[,c(1,2,3,4)]) )

q<- df1 %>% inner\_join(df2, "chromosome") %>%

mutate(geneID\_motif = paste(id, motif, sep = ","),

n = if\_else(position >= start & position <= end, 1, 0)) %>%

select(icgc\_samlpe\_id, geneID\_motif, n) %>%

group\_by(icgc\_samlpe\_id, geneID\_motif) %>%

summarise(n = sum(n)) %>%

spread(key = geneID\_motif, value = n, fill = 0)

q0= q[,!grepl("N",colnames(q))]

row\_name<- q0$icgc\_samlpe\_id

q0 <- q0[,-1]

rownames(q0)<- row\_name

col\_name<-colnames(q0)

q0<-colSums(q0)

q0<-as.data.frame(q0)

k<-vector(max(q0), mode='list')

for(i in 1:max(q0)){

ttt<-subset(q0,q0>=i)

print(i)

w<- paste0(b," gene\_bigger than ",as.character(i))

barplot(t(ttt), main = w) # create the barplot

k[[i]] <- recordPlot()

}

graphics.off()

outputfile <- paste0('/home/san/halinejad/Desktop/Masroor Bayati/DeepCancer project/Dashti/',b,'\_barplot\_pseudogene.pdf')

pdf(outputfile, onefile=TRUE)

for (my.plot in k) {

replayPlot(my.plot)

}

graphics.off()

}

{

df1<-structure(x)

df2<-structure(as.data.frame(sense\_overlap\_RNA[,c(1,2,3,4)]) )

q<- df1 %>% inner\_join(df2, "chromosome") %>%

mutate(geneID\_motif = paste(id, motif, sep = ","),

n = if\_else(position >= start & position <= end, 1, 0)) %>%

select(icgc\_samlpe\_id, geneID\_motif, n) %>%

group\_by(icgc\_samlpe\_id, geneID\_motif) %>%

summarise(n = sum(n)) %>%

spread(key = geneID\_motif, value = n, fill = 0)

q0= q[,!grepl("N",colnames(q))]

row\_name<- q0$icgc\_samlpe\_id

q0 <- q0[,-1]

rownames(q0)<- row\_name

col\_name<-colnames(q0)

q0<-colSums(q0)

q0<-as.data.frame(q0)

k<-vector(max(q0), mode='list')

for(i in 1:max(q0)){

ttt<-subset(q0,q0>=i)

print(i)

w<- paste0(b," gene\_bigger than ",as.character(i))

barplot(t(ttt), main = w) # create the barplot

k[[i]] <- recordPlot()

}

graphics.off()

outputfile <- paste0('/home/san/halinejad/Desktop/Masroor Bayati/DeepCancer project/Dashti/',b,'\_barplot\_sense\_overlap\_RNA.pdf')

pdf(outputfile, onefile=TRUE)

for (my.plot in k) {

replayPlot(my.plot)

}

graphics.off()

}

{

df1<-structure(x)

df2<-structure(as.data.frame(uncertain\_coding[,c(1,2,3,4)]) )

q<- df1 %>% inner\_join(df2, "chromosome") %>%

mutate(geneID\_motif = paste(id, motif, sep = ","),

n = if\_else(position >= start & position <= end, 1, 0)) %>%

select(icgc\_samlpe\_id, geneID\_motif, n) %>%

group\_by(icgc\_samlpe\_id, geneID\_motif) %>%

summarise(n = sum(n)) %>%

spread(key = geneID\_motif, value = n, fill = 0)

q0= q[,!grepl("N",colnames(q))]

row\_name<- q0$icgc\_samlpe\_id

q0 <- q0[,-1]

rownames(q0)<- row\_name

col\_name<-colnames(q0)

q0<-colSums(q0)

q0<-as.data.frame(q0)

k<-vector(max(q0), mode='list')

for(i in 1:max(q0)){

ttt<-subset(q0,q0>=i)

print(i)

w<- paste0(b," gene\_bigger than ",as.character(i))

barplot(t(ttt), main = w) # create the barplot

k[[i]] <- recordPlot()

}

graphics.off()

outputfile <- paste0('/home/san/halinejad/Desktop/Masroor Bayati/DeepCancer project/Dashti/',b,'\_barplot\_uncertain\_coding.pdf')

pdf(outputfile, onefile=TRUE)

for (my.plot in k) {

replayPlot(my.plot)

}

graphics.off()

}

{

df1<-structure(x)

df2<-structure(as.data.frame(short\_ncRNA[,c(1,2,3,4)]) )

q<- df1 %>% inner\_join(df2, "chromosome") %>%

mutate(geneID\_motif = paste(id, motif, sep = ","),

n = if\_else(position >= start & position <= end, 1, 0)) %>%

select(icgc\_samlpe\_id, geneID\_motif, n) %>%

group\_by(icgc\_samlpe\_id, geneID\_motif) %>%

summarise(n = sum(n)) %>%

spread(key = geneID\_motif, value = n, fill = 0)

q0= q[,!grepl("N",colnames(q))]

row\_name<- q0$icgc\_samlpe\_id

q0 <- q0[,-1]

rownames(q0)<- row\_name

col\_name<-colnames(q0)

q0<-colSums(q0)

q0<-as.data.frame(q0)

k<-vector(max(q0), mode='list')

for(i in 1:max(q0)){

ttt<-subset(q0,q0>=i)

print(i)

w<- paste0(b," gene\_bigger than ",as.character(i))

barplot(t(ttt), main = w) # create the barplot

k[[i]] <- recordPlot()

}

graphics.off()

outputfile <- paste0('/home/san/halinejad/Desktop/Masroor Bayati/DeepCancer project/Dashti/',b,'\_barplot\_short\_ncRNA.pdf')

pdf(outputfile, onefile=TRUE)

for (my.plot in k) {

replayPlot(my.plot)

}

graphics.off()

}

{

df1<-structure(x)

df2<-structure(as.data.frame(small\_RNA[,c(1,2,3,4)]) )

q<- df1 %>% inner\_join(df2, "chromosome") %>%

mutate(geneID\_motif = paste(id, motif, sep = ","),

n = if\_else(position >= start & position <= end, 1, 0)) %>%

select(icgc\_samlpe\_id, geneID\_motif, n) %>%

group\_by(icgc\_samlpe\_id, geneID\_motif) %>%

summarise(n = sum(n)) %>%

spread(key = geneID\_motif, value = n, fill = 0)

q0= q[,!grepl("N",colnames(q))]

row\_name<- q0$icgc\_samlpe\_id

q0 <- q0[,-1]

rownames(q0)<- row\_name

col\_name<-colnames(q0)

q0<-colSums(q0)

q0<-as.data.frame(q0)

k<-vector(max(q0), mode='list')

for(i in 1:max(q0)){

ttt<-subset(q0,q0>=i)

print(i)

w<- paste0(b," gene\_bigger than ",as.character(i))

barplot(t(ttt), main = w) # create the barplot

k[[i]] <- recordPlot()

}

graphics.off()

outputfile <- paste0('/home/san/halinejad/Desktop/Masroor Bayati/DeepCancer project/Dashti/',b,'\_barplot\_small\_RNA.pdf')

pdf(outputfile, onefile=TRUE)

for (my.plot in k) {

replayPlot(my.plot)

}

graphics.off()

}

{

df1<-structure(x)

df2<-structure(as.data.frame(structural\_RNA[,c(1,2,3,4)]) )

q<- df1 %>% inner\_join(df2, "chromosome") %>%

mutate(geneID\_motif = paste(id, motif, sep = ","),

n = if\_else(position >= start & position <= end, 1, 0)) %>%

select(icgc\_samlpe\_id, geneID\_motif, n) %>%

group\_by(icgc\_samlpe\_id, geneID\_motif) %>%

summarise(n = sum(n)) %>%

spread(key = geneID\_motif, value = n, fill = 0)

q0= q[,!grepl("N",colnames(q))]

row\_name<- q0$icgc\_samlpe\_id

q0 <- q0[,-1]

rownames(q0)<- row\_name

col\_name<-colnames(q0)

q0<-colSums(q0)

q0<-as.data.frame(q0)

k<-vector(max(q0), mode='list')

for(i in 1:max(q0)){

ttt<-subset(q0,q0>=i)

print(i)

w<- paste0(b," gene\_bigger than ",as.character(i))

barplot(t(ttt), main = w) # create the barplot

k[[i]] <- recordPlot()

}

graphics.off()

outputfile <- paste0('/home/san/halinejad/Desktop/Masroor Bayati/DeepCancer project/Dashti/',b,'\_barplot\_structural\_RNA.pdf')

pdf(outputfile, onefile=TRUE)

for (my.plot in k) {

replayPlot(my.plot)

}

graphics.off()

}