# Motif enrichment

Vi Dang

2022-09-11

```
#Load libraries
library(tidyverse)
library(ggplot2)
setwd("D:/PhD/TSS cluster/H99/TRASS_d17/")
```

```
#read data
# TC motif
TC_ESharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TC motif/TC E30S.tsv",sep = "\t",hea
TC_ESharp<-TC_ESharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])   #252

TC_EBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TC motif/TC E30B.tsv",sep = "\t",hea
TC_EBroad<-TC_EBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])   #421

TC_SSharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TC motif/TC S30S.tsv",sep = "\t",hea
TC_SSharp<-TC_SSharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #282

TC_SBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TC motif/TC S30B.tsv",sep = "\t",hea
TC_SBroad<-TC_SBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #432

TC_E37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TC motif/TC E37S.tsv",sep = "\t",he
TC_E37Sharp<-TC_E37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #247

TC_E37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TC motif/TC E37B.tsv",sep = "\t",he
TC_E37Broad<-TC_E37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])   #346
```

```
TC_S37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TC motif/TC S37S.tsv",sep = "\t",he
TC_S37Sharp<-TC_S37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #221

TC_S37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TC motif/TC S37B.tsv",sep = "\t",he
TC_S37Broad<-TC_S37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #388

#GG motif
GG_ESharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GG motif/GG E30S.tsv",sep = "\t",hea
GG_ESharp<-GG_ESharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GG")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #23

GG_EBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GG motif/GG E30B.tsv",sep = "\t",hea
GG_EBroad<- GG_EBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GG")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #27

GG_SSharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GG motif/GG S30S.tsv",sep = "\t",hea
GG_SSharp<-GG_SSharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GG")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #21

GG_SBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GG motif/GG S30B.tsv",sep = "\t",hea
GG_SBroad<-GG_SBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GG")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #64

GG_E37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GG motif/GG E37S.tsv",sep = "\t",he
GG_E37Sharp<- GG_E37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GG")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #23

GG_E37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GG motif/GG E37B.tsv",sep = "\t",he
GG_E37Broad<-GG_E37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GG")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #30

GG_S37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GG motif/GG S37S.tsv",sep = "\t",he
GG_S37Sharp<-GG_S37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GG")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #13

GG_S37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GG motif/GG S37B.tsv",sep = "\t",h
GG_S37Broad<-GG_S37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GG")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #49

#TATA motif
TATA_ESharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TATA motif/TATA E30S.tsv",sep = "\
TATA_ESharp<- TATA_ESharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TATA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #46

TATA_EBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TATA motif/TATA E30B.tsv",sep = "\
TATA_EBroad<- TATA_EBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TATA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #58

TATA_SSharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TATA motif/TATA S30S.tsv",sep = "\
TATA_SSharp<- TATA_SSharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TATA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #45

TATA_SBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TATA motif/TATA S30B.tsv",sep = "\
TATA_SBroad<- TATA_SBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TATA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #77

TATA_E37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TATA motif/TATA E37S.tsv",sep =
TATA_E37Sharp<- TATA_E37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TATA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #43

TATA_E37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TATA motif/TATA E37B.tsv",sep =
TATA_E37Broad<- TATA_E37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TATA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #67

TATA_S37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TATA motif/TATA S37S.tsv",sep =
TATA_S37Sharp<- TATA_S37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TATA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #36

TATA_S37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TATA motif/TATA S37B.tsv",sep =
TATA_S37Broad<- TATA_S37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TATA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #91


#AC motif
AC_ESharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/AC motif/AC E30S.tsv",sep = "\t",hea
AC_ESharp<- AC_ESharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="AC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])   #9

AC_EBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/AC motif/AC E30B.tsv",sep = "\t",hea
AC_EBroad<- AC_EBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="AC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #61

AC_SSharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/AC motif/AC S30S.tsv",sep = "\t",hea
AC_SSharp<- AC_SSharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="AC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #13

AC_SBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/AC motif/AC S30B.tsv",sep = "\t",hea
AC_SBroad<- AC_SBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="AC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #78

AC_E37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/AC motif/AC E37S.tsv",sep = "\t",he
AC_E37Sharp<- AC_E37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="AC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #19

AC_E37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/AC motif/AC E37B.tsv",sep = "\t",he
AC_E37Broad<- AC_E37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="AC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #57

AC_S37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/AC motif/AC S37S.tsv",sep = "\t",he
AC_S37Sharp<- AC_S37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="AC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #16

AC_S37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/AC motif/AC S37B.tsv",sep = "\t",he
AC_S37Broad<- AC_S37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="AC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #52
```

```
#TTAC motif
TTAC_ESharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TTAC motif/TTAC E30S.tsv",sep = "\
TTAC_ESharp<- TTAC_ESharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TTAC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])   #11


TTAC_EBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TTAC motif/TTAC E30B.tsv",sep = "\
TTAC_EBroad<- TTAC_EBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TTAC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #50


TTAC_SSharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TTAC motif/TTAC S30S.tsv",sep = "\
TTAC_SSharp<- TTAC_SSharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TTAC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #15


TTAC_SBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TTAC motif/TTAC S30B.tsv",sep = "\
TTAC_SBroad<- TTAC_SBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TTAC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #48


TTAC_E37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TTAC motif/TTAC E37S.tsv",sep =
TTAC_E37Sharp<- TTAC_E37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TTAC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #15


TTAC_E37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TTAC motif/TTAC E37B.tsv",sep =
TTAC_E37Broad<- TTAC_E37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TTAC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #37


TTAC_S37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TTAC motif/TTAC S37S.tsv",sep =
TTAC_S37Sharp<- TTAC_S37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TTAC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #11


TTAC_S37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/TTAC motif/TTAC S37B.tsv",sep =
TTAC_S37Broad<- TTAC_S37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="TTAC")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #47


#GA motif
GA_ESharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GA motif/GA E30S.tsv",sep = "\t",hea
GA_ESharp<- GA_ESharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])   #12


GA_EBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GA motif/GA E30B.tsv",sep = "\t",hea
GA_EBroad<- GA_EBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #39


GA_SSharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GA motif/GA S30S.tsv",sep = "\t",hea
GA_SSharp<- GA_SSharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2]) #10


GA_SBroad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GA motif/GA S30B.tsv",sep = "\t",hea
GA_SBroad<- GA_SBroad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #33


GA_E37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GA motif/GA E37S.tsv",sep = "\t",h
GA_E37Sharp<- GA_E37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #11


GA_E37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GA motif/GA E37B.tsv",sep = "\t",h
GA_E37Broad<- GA_E37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #27


GA_S37Sharp <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GA motif/GA S37S.tsv",sep = "\t",h
GA_S37Sharp<- GA_S37Sharp%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #6
```

```
GA_S37Broad <- read.table("D:/PhD/TSS cluster/H99/TRASS_d17/MEME all/GA motif/GA S37B.tsv",sep = "\t",he
GA_S37Broad<- GA_S37Broad%>%filter(p.value<0.0005)%>%
  mutate(position_to_TSS=start-100)%>%
  mutate(motif="GA")%>%
  mutate(Id=str_match(sequence_name,"(\\b.+)::")[,2])  #36
```

```
#Position relative to TSS
#position of motif in Expo Sharp cluster
ESharp_motif<-bind_rows(TC_ESharp,GG_ESharp,TATA_ESharp,AC_ESharp,TTAC_ESharp,GA_ESharp)

#remove gene as identify in Percentage small set All cluster
gene_reject_E30<-read.delim("D:/PhD/TSS cluster/H99/TRASS_d17/gene_reject_E30.txt")
ESharp_motif<-ESharp_motif%>%filter(!(Id%in%gene_reject_E30$gene_reject_E30))


neworder <- c("AC","TTAC","GG","GA","TATA","TC")
table(ESharp_motif$motif)
```

```
##
##   AC   GA   GG TATA   TC TTAC
##    9   10   21   41  241   11
```

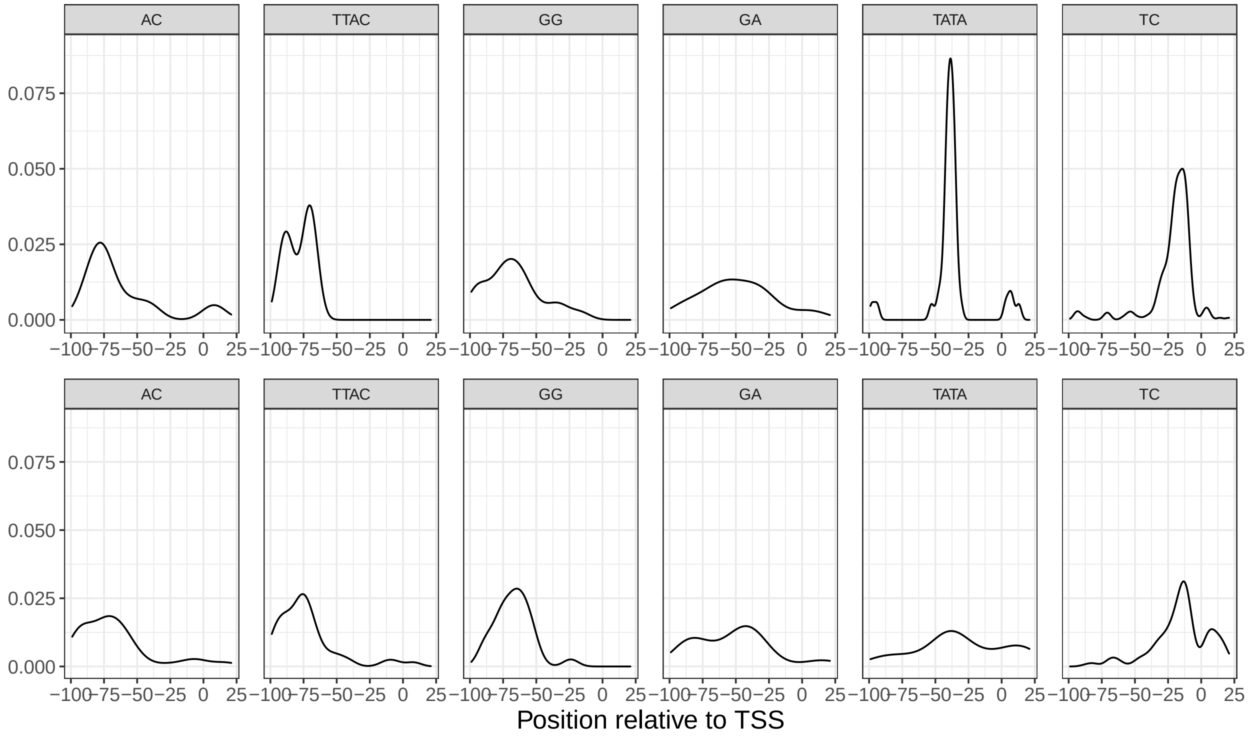
```
ESharp_motif<-ESharp_motif%>%
  mutate(motif=factor(motif,levels=neworder))%>%
  arrange(motif)

Sharp_plot<-ESharp_motif%>%ggplot(aes(x=position_to_TSS))+
  geom_density()+
  facet_wrap(vars(motif),nrow=1)+
  scale_y_continuous(limits = c(0,0.09))+
  theme_bw()+
  theme(panel.spacing = unit(0.5, "cm"))+
  theme(plot.title = element_text(size=23,vjust=4),
        axis.title.x.bottom = element_blank(),
        axis.title.y.left = element_blank(),
        axis.text.x.bottom = element_text(size=11),
        axis.text.y.left = element_text(size=11))



#Position of motif in Expo Broad cluster

EBroad_motif<-bind_rows(TC_EBroad,GG_EBroad,TATA_EBroad,AC_EBroad,TTAC_EBroad,GA_EBroad)

#remove gene as identify in Percentage small set All cluster
EBroad_motif<-EBroad_motif%>%filter(!(Id%in%gene_reject_E30$gene_reject_E30))

table(EBroad_motif$motif)
```

```
##
##   AC   GA   GG TATA   TC TTAC
##   60   39   27   55  392   47
```

```r
neworder <- c("AC","TTAC","GG","GA","TATA","TC")

EBroad_motif<-EBroad_motif%>%
  mutate(motif=factor(motif,levels=neworder))%>%
  arrange(motif)


Broad_plot<-EBroad_motif%>%ggplot(aes(x=position_to_TSS))+
  geom_density()+
  facet_wrap(vars(motif),nrow=1)+
  scale_y_continuous(limits = c(0,0.09))+
  xlab("Position relative to TSS")+
  theme_bw()+
  theme(panel.spacing = unit(0.5, "cm"))+
  theme(plot.title = element_text(size=23,vjust=4),
        axis.title.x.bottom = element_text(size=15),
        axis.title.y.left = element_blank(),
        axis.text.x.bottom = element_text(size=11),
        axis.text.y.left = element_text(size=11))


library(gridExtra)
```

```
##
## Attaching package: ’gridExtra’

## The following object is masked from ’package:dplyr’:
##
##     combine
```

```r
grid.arrange(Sharp_plot,Broad_plot, nrow=2)
```


AC
TTAC
GG
GA
TATA
TC
0.075
0.050
0.025
0.000
−100
−75
−50
−25
0
25
Position relative to TSS