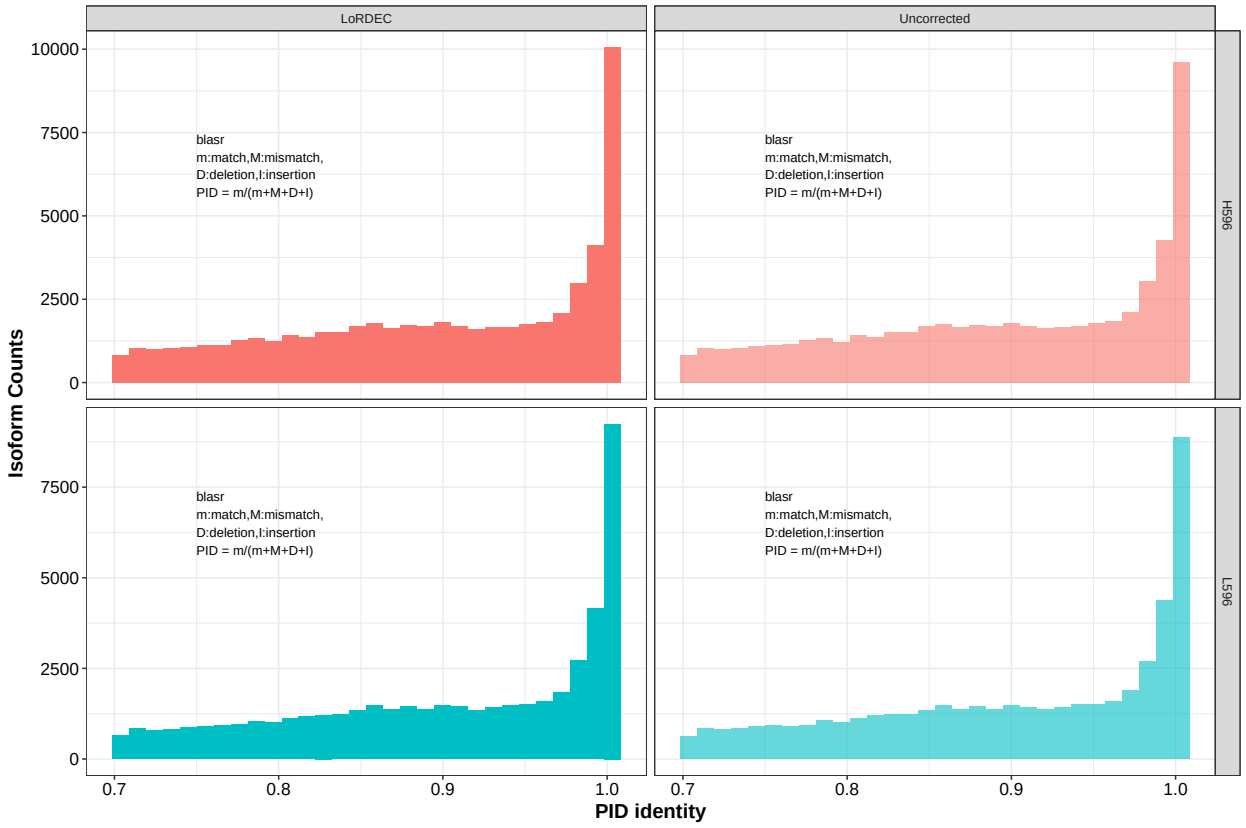
Iso-seq structral analysis

Changfu Jia

2022-09-10

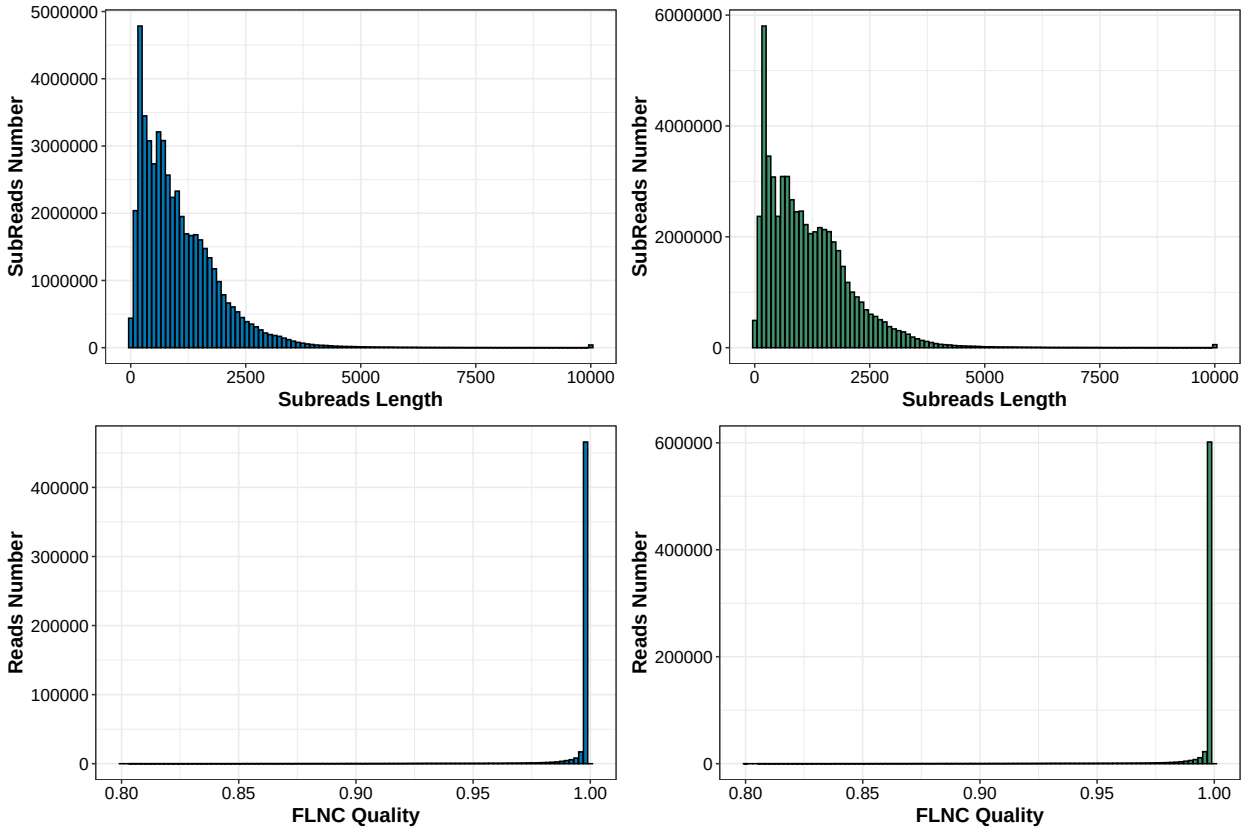
```
setwd("F:/dir")

library(tidyverse)

L596_cor<-read.table("m64032_191231_031514.subreads.6--6_75.ccs.lima.refine.cluster.hq..corrected.blasr
L596<-read.table("m64032_191231_031514.subreads.6--6_75.ccs.lima.refine.cluster.hq.blasr.out55.pid",sep

H596_cor<-read.table("m64032_191231_031514.subreads.7--7_75.ccs.lima.refine.cluster.hq..corrected.blasr
H596<-read.table("m64032_191231_031514.subreads.7--7_75.ccs.lima.refine.cluster.hq.blasr.out55.pid",sep



all<-rbind(H596_cor %>% mutate(type="LoRDEC", cond="H596"),
          H596 %>% mutate(type="Uncorrected", cond="H596"),
          L596_cor %>% mutate(type="LoRDEC", cond="L596"),
           L596 %>% mutate(type="Uncorrected", cond="L596"))

ggplot(all, aes(x=V2, fill=cond, alpha=type))+
  geom_histogram(bins = 30) +
  facet_grid(rows = c("cond" ,"type"), scales = "free_y")+
  theme_bw()+
  theme(legend.position ="none",
        axis.title.x=element_text(size=14,color="black",vjust=1,face = "bold"),

        axis.title.y=element_text(size=14,color="black",face = "bold" ),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=10,face = "bold")) +
  scale_alpha_manual(values=c(1,0.6))+
  annotate("text",   x=0.75,y= 6500, hjust=0, color= "black",
          label=paste("blasr", "m:match,M:mismatch,","D:deletion,I:insertion",
                      "PID = m/(m+M+D+I)" ,sep="\n"  ),size=3)+
  ylab("Isoform Counts")+
  xlab("PID identity")
```




```
setwd("F:/dir")

#library(tidyverse)


frag <- read.table("m64032_191231_031514.subreads.6--6_75.fragment", sep="\t")
frag <- read.table("6.fragment.length.txt", sep="\t")



bin <- seq(0,10000,length.out=100)
#?seq
out<-data.frame( bin, 0 )

for (i in 1:nrow(frag)) {
  num<-frag[i,]$V1
  freq<-frag[i,]$V2
  ind<-length(which(num-bin >0))
  out[ind,2]<-out[ind,2]+freq
}

options(scipen=200)

L596_subreads<-ggplot(out, aes(x=bin,y=X0)) +
  geom_col(fill="#0077b6", color="black") +
  theme_bw()  +
  theme(legend.position ="none",
```

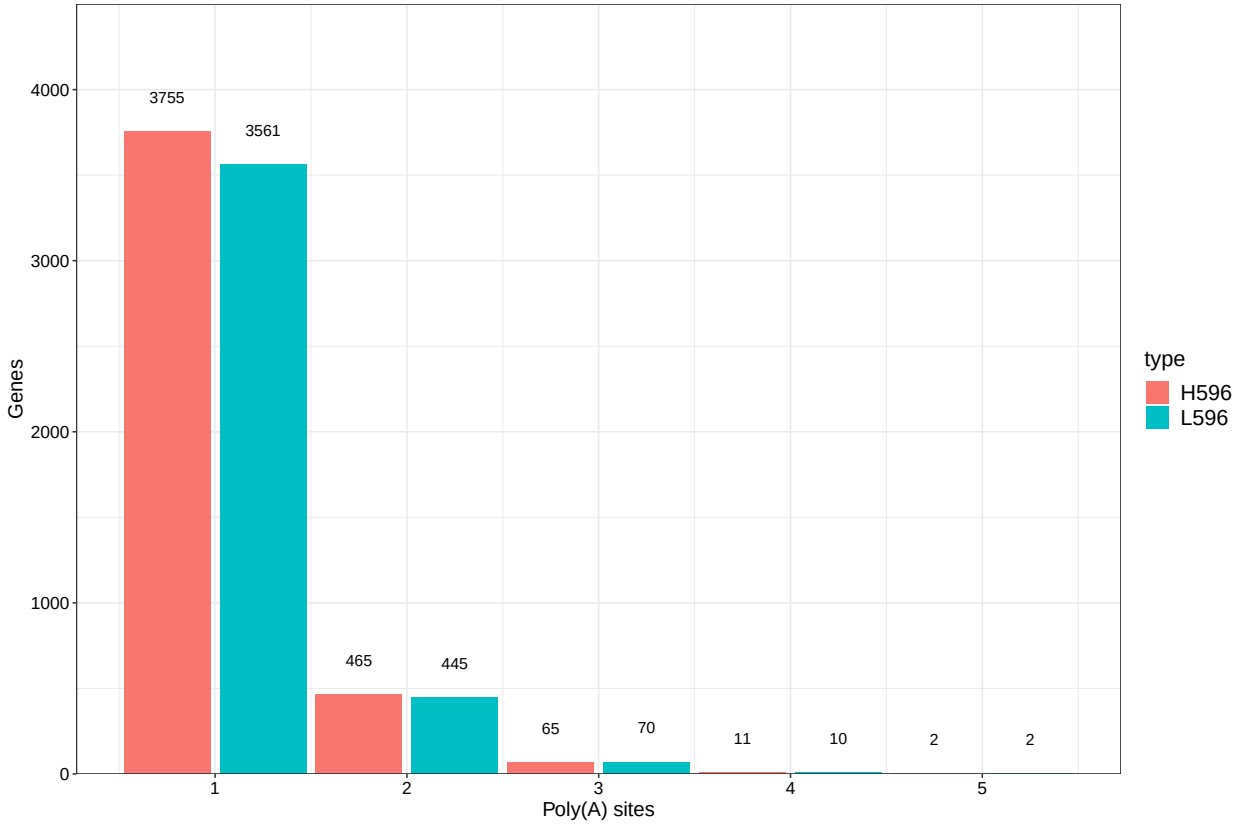
```
                    axis.title.x=element_text(size=14,color="black",vjust=1,face = "bold"),

                    axis.title.y=element_text(size=14,color="black",face = "bold" ),
                    axis.text.x=element_text(size=12,color="black"),
                    axis.text.y=element_text(size=12,color="black"),
                    legend.text=element_text(size=15,color="black"),
                    legend.title=element_text(size=15,color="black"),
                    plot.title=element_text(size=15,color="black",hjust=10,face = "bold")) +
  ylab("SubReads Number")+
  xlab("Subreads Length")

dir.create("1.reads_quality/")



frag<- read.table("7.fragment.length.txt", sep="\t")


bin <- seq(0,10000,length.out=100)
#?seq
out<-data.frame( bin, 0 )

for (i in 1:nrow(frag)) {
  num<-frag[i,]$V1
  freq<-frag[i,]$V2
  ind<-length(which(num-bin >0))
  out[ind,2]<-out[ind,2]+freq
}

H596_subreads<-ggplot(out, aes(x=bin,y=X0)) +
  geom_col(fill="#40916c", color="black") +
  theme_bw()  +
  theme(legend.position ="none",
        axis.title.x=element_text(size=14,color="black",vjust=1,face = "bold"),

        axis.title.y=element_text(size=14,color="black",face = "bold" ),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=10,face = "bold")) +
  ylab("SubReads Number")+
  xlab("Subreads Length")




#qv
frag<-read.table("H596_flnc.qv.txt")

bin <- seq(0.8,1,length.out=100)
out<-data.frame( bin, 0 )
```

```
for (i in 1:nrow(frag)) {
  num<-frag[i,]$V1
  freq<-frag[i,]$V2
  ind<-length(which(num-bin >0))
  out[ind,2]<-out[ind,2]+freq
}

H596_flnc_quality<-ggplot(out, aes(x=bin,y=X0)) +
  geom_col(fill="#40916c", color="black") +
  theme_bw()  +
  theme(legend.position ="none",
        axis.title.x=element_text(size=14,color="black",vjust=1,face = "bold"),

        axis.title.y=element_text(size=14,color="black",face = "bold" ),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=10,face = "bold")) +
  ylab("Reads Number")+
  xlab("FLNC Quality")


frag<-read.table("L596_flnc.qv.txt")

bin <- seq(0.8,1,length.out=100)
out<-data.frame( bin, 0 )

for (i in 1:nrow(frag)) {
  num<-frag[i,]$V1
  freq<-frag[i,]$V2
  ind<-length(which(num-bin >0))
  out[ind,2]<-out[ind,2]+freq
}


L596_flnc_quality<-ggplot(out, aes(x=bin,y=X0)) +
  geom_col(fill="#0077b6", color="black") +
  theme_bw()  +
  theme(legend.position ="none",
        axis.title.x=element_text(size=14,color="black",vjust=1,face = "bold"),

        axis.title.y=element_text(size=14,color="black",face = "bold" ),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=10,face = "bold")) +
  ylab("Reads Number")+
  xlab("FLNC Quality")

ggpubr::ggarrange(L596_subreads,H596_subreads,L596_flnc_quality,H596_flnc_quality)
```



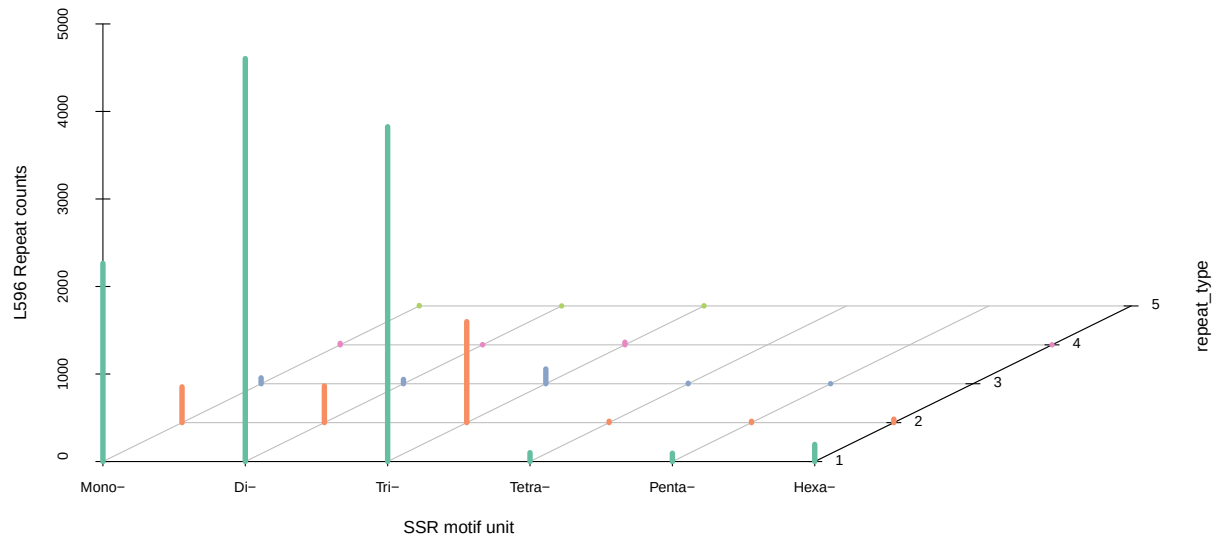

```
setwd("F:/dir/2.Structral_analysis/4.PolyA")

library(tidyverse)

H596_poly<-read.table("H596_polyA_summary.txt", sep="\t",header = T)
L596_poly<-read.table("L596_polyA_summary.txt", sep="\t",header = T)


rbind(L596_poly %>% mutate(type="L596")  , H596_poly %>% mutate(type="H596")   )  %>%
  filter(num.sites>0) %>%
  group_by(num.sites, type) %>%
  mutate(number=n()) %>%
  ungroup( ) %>%
  select(num.sites, number, type) %>%
  unique(.) %>%
  ggplot(aes(x=num.sites, y=number, fill=type))+
  geom_col(position = position_dodge(width = 1) ) +
  theme_bw() +
  theme(legend.position="right",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("Poly(A) sites")+
```

```
  ylab("Genes") +
  scale_y_continuous(limits=c(0,4500   ), expand = c(0,0)) +
  geom_text( aes( x=ifelse( type == "H596", num.sites-0.25, num.sites+0.25  ) ,
                  y = number +200 , label=number ))
```



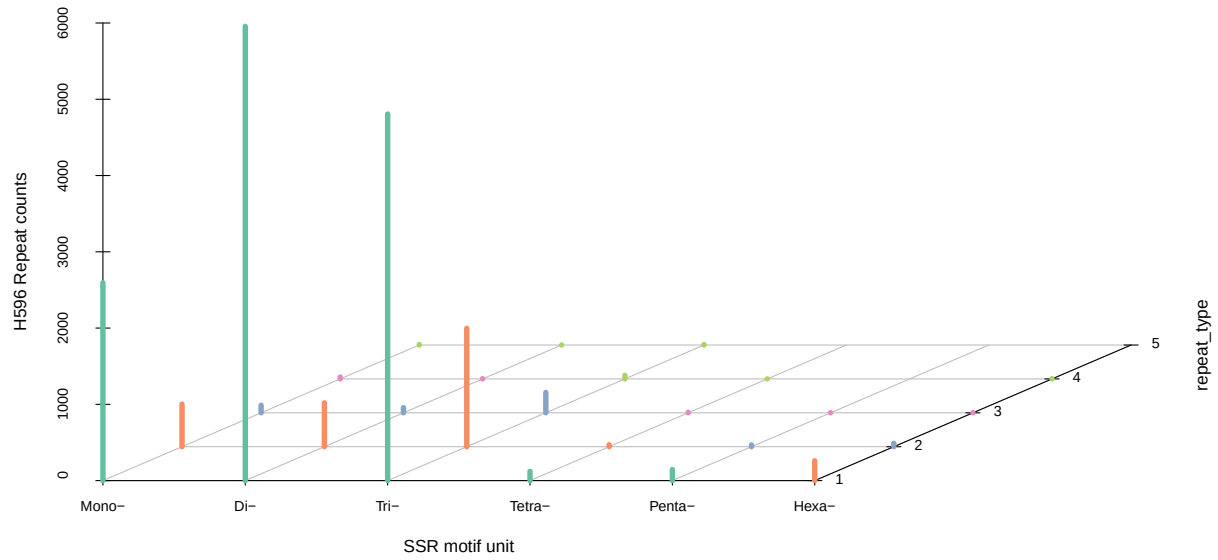

```
#ggsave("PolyA_sites_stat.pdf", height = 7, width = 8)

rm(H596_poly, L596_poly)



library(scatterplot3d)
setwd("F:/dir/2.Structral_analysis/6.misa")

H596_misa<-read.table("H596.fasta.misa", sep="\t",header = T)
L596_misa<-read.table("L596.fasta.misa", sep="\t",header = T)



L596_misa_s<-
L596_misa %>% filter( !grepl("c",SSR.type )  ) %>%
  group_by(SSR.nr., SSR.type) %>%
  mutate(count=n()) %>%
  select(SSR.nr., SSR.type, count) %>%
  unique(.) %>%
  ungroup() %>%
  filter(SSR.nr.!=6)
```

```
H596_misa_s<-
  H596_misa %>% filter( !grepl("c",SSR.type )  ) %>%
  group_by(SSR.nr., SSR.type) %>%
  mutate(count=n()) %>%
  select(SSR.nr., SSR.type, count) %>%
  unique(.) %>%
  ungroup() %>%
  filter(SSR.nr.!=6)


#scatterplot3d(dat,type='h',lwd=5,pch='',color=rbc,box=F, xlab = "V-gene", ylab = "J-gene", zlab = "Per


scatterplot3d( L596_misa_s[,c(2,1,3)] %>%  arrange(SSR.nr., SSR.type)  ,
               type='h', lwd=5,pch='',box=F, xlab = "SSR motif unit",
               ylab = "repeat_type", zlab = "L596 Repeat counts",
               color =  c(rep("#65BEA2",6) ,rep("#F78D63",6), rep("#89A3C8",5) ,
                          rep("#E689C2", 4) ,rep("#ACD265", 3)   ),
               x.ticklabs= c("Mono-","Di-","Tri-","Tetra-","Penta-","Hexa-")   )
```



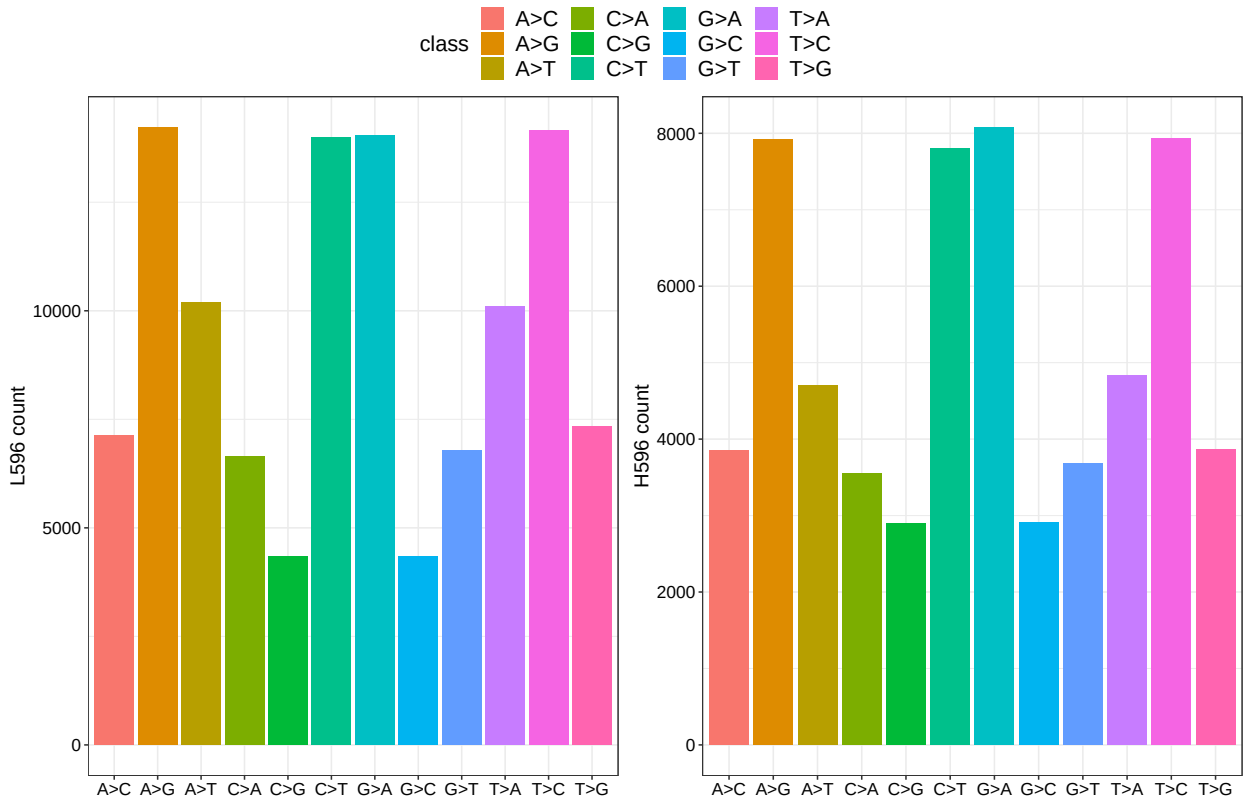

```
            #   color =c(rep("#65BEA2",20) ,rep("#F78D63",4)    ),  x.ticklabs= c("Mono-","Di-","Tri","

scatterplot3d( H596_misa_s[,c(2,1,3)] %>%arrange(SSR.nr., SSR.type) ,
               type='h', lwd=5,pch='',box=F, xlab = "SSR motif unit",
```

```
              ylab = "repeat_type", zlab = "H596 Repeat counts",
              color =  c(rep("#65BEA2",5) ,rep("#F78D63",5), rep("#89A3C8",5) ,
                         rep("#E689C2", 5) ,rep("#ACD265", 6)   ),
              x.ticklabs= c("Mono-","Di-","Tri-","Tetra-","Penta-","Hexa-")   )
```



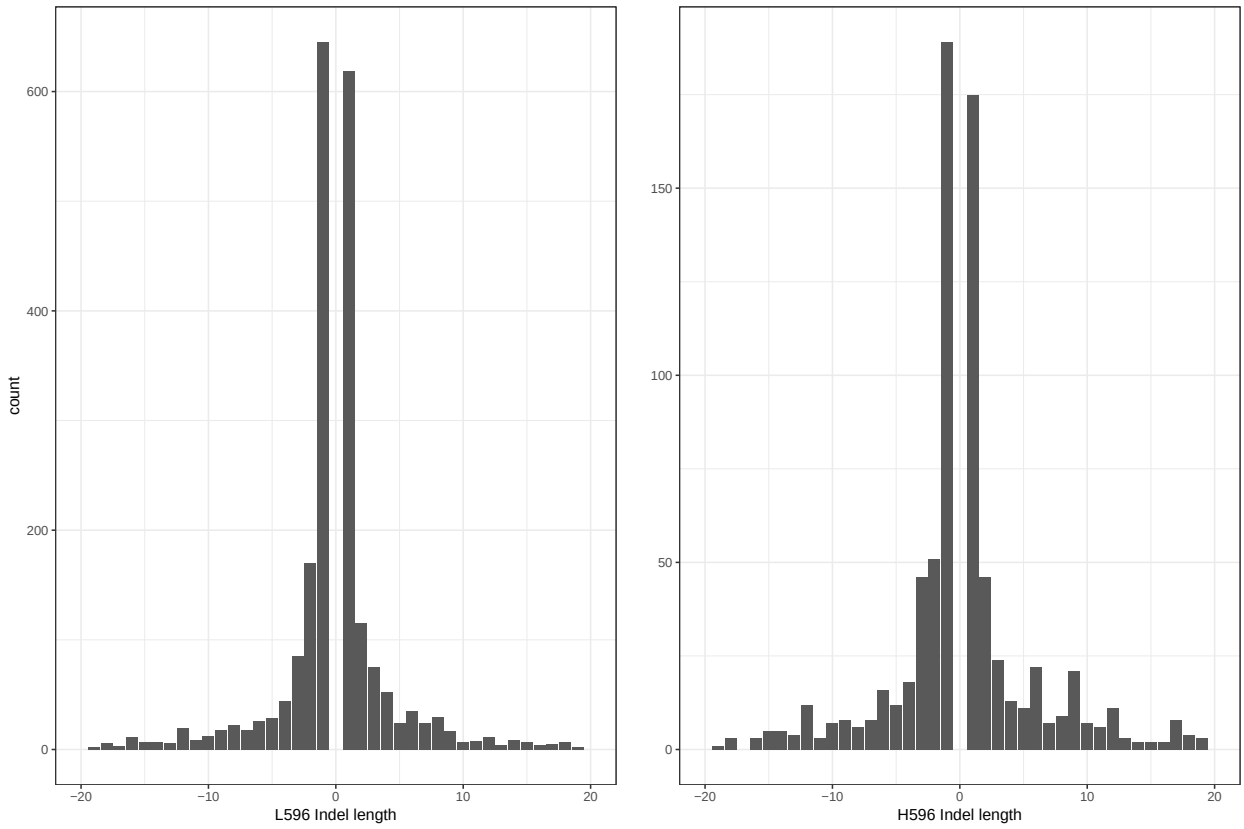

```
rm(H596_misa,H596_misa_s, L596_misa,L596_misa_s, p1,p2 )

setwd("F:/dir/2.Structral_analysis/8.SNP_indel")


H596_vcf<-read.table("H596.vcf",sep="\t")
L596_vcf<-read.table("L596.vcf",sep="\t")


p1<-
L596_vcf %>%
  filter(!grepl("INDEL", V8)) %>%
  mutate(class = paste(V4,">",V5, sep="")) %>%
  select(class) %>%
  filter(!grepl(",",class)) %>%
  ggplot(aes(x=class,fill=class))+
  geom_bar()+
  theme_bw() +
  theme(legend.position="right",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
```

```
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("")+
  ylab("L596 count")

p2<-
H596_vcf %>%
  filter(!grepl("INDEL", V8)) %>%
  mutate(class = paste(V4,">",V5, sep="")) %>%
  select(class) %>%
  filter(!grepl(",",class)) %>%
  ggplot(aes(x=class,fill=class))+
  geom_bar()+
  theme_bw() +
  theme(legend.position="right",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("")+
  ylab("H596 count")

ggpubr::ggarrange(p1,p2,common.legend = T)
```



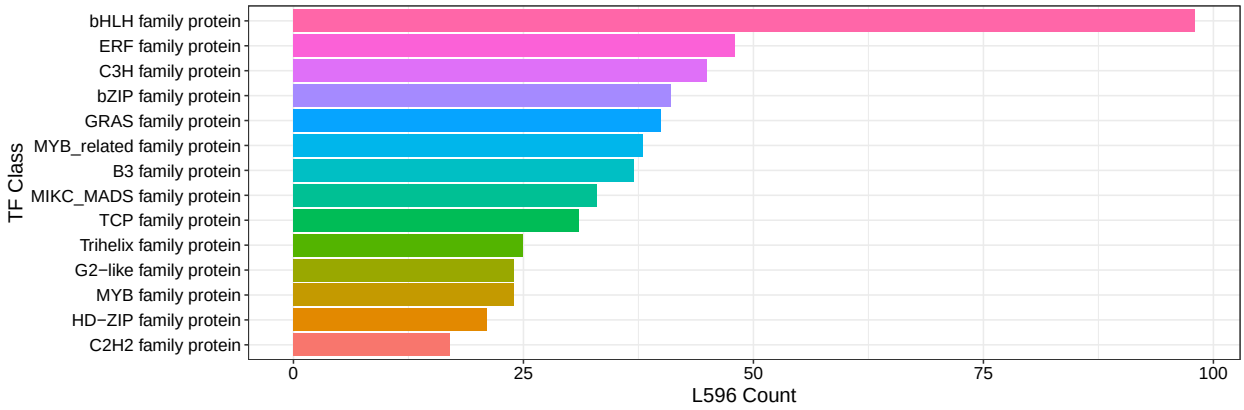

```
#ggsave("SNP_freq.pdf", width = 20, height = 13)



#rbind( L596_vcf %>%mutate(class="L596"),   H596_vcf %>%mutate(class="H596")  ) %>%
L596_vcf_ind<-
L596_vcf %>%
  filter(grepl("INDEL", V8)) %>%
  mutate(ind_len= str_length(V4) -  str_length(V5) ) %>%
  select(ind_len) %>%
  ggplot(aes(x=ind_len))+
  geom_bar()+
  theme_bw() +
  xlim(-20,20)+
  xlab("L596 Indel length")

H596_vcf_ind<-
H596_vcf %>%
  filter(grepl("INDEL", V8)) %>%
  mutate(ind_len= str_length(V4) -  str_length(V5) ) %>%
  select(ind_len) %>%
  ggplot(aes(x=ind_len))+
  geom_bar()+
  theme_bw() +
  xlim(-20,20) +
  xlab("H596 Indel length") +
  ylab("")
```

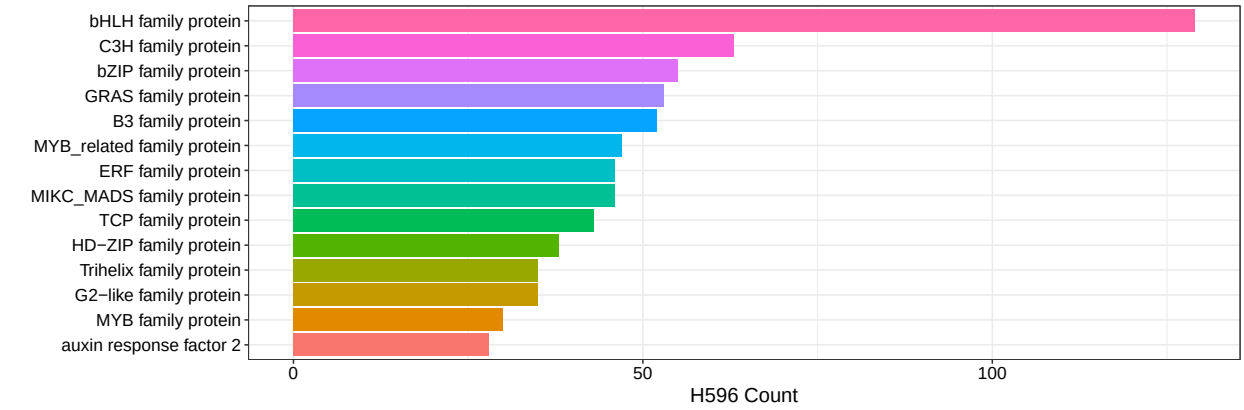
```
ggpubr::ggarrange(L596_vcf_ind,H596_vcf_ind)
```




```
#ggsave("Indel_freq.pdf", width = 8)

rm(H596_vcf, H596_vcf_ind, L596_vcf,L596_vcf_ind,p1,p2 )

setwd("F:/dir/2.Structral_analysis/9.TF")

H596<- read.csv("H596.TF.csv" ,header = F )
L596<- read.csv("L596.TF.csv", header = F)

H596_TF<-
H596 %>%
  select(V5) %>%
  group_by(V5)%>%
  mutate(count= n()) %>%
  unique(.) %>%
  arrange(desc(count)) %>%
  ungroup() %>%
  filter(V5!="")%>%
  top_n(n=14) %>%
  ggplot(aes(y=factor(V5,levels = rev(V5)), x=count,
             fill= factor(V5,levels = rev(V5)))) +
  geom_col()+
  theme_bw() +
  theme(legend.position="none",
```

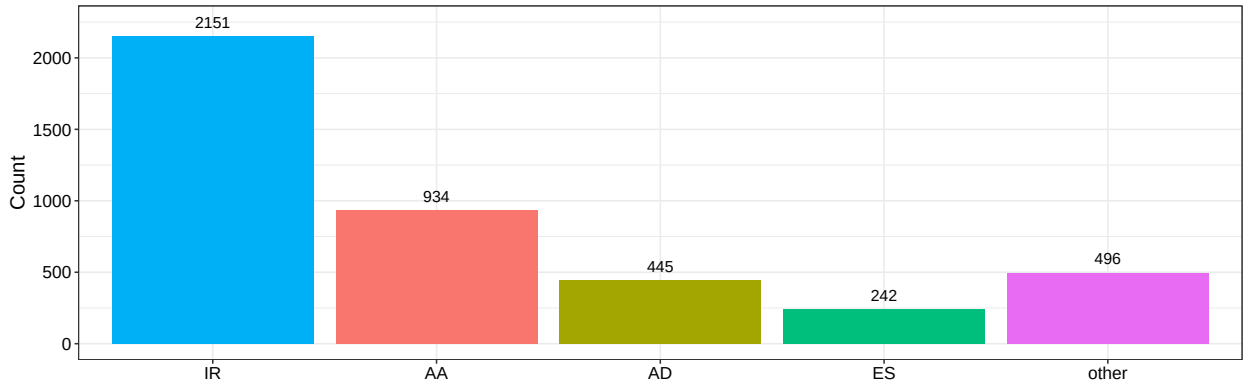
```
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  ylab("")+
  xlab("H596 Count")


L596_TF<-
L596 %>%
  select(V5) %>%
  group_by(V5)%>%
  mutate(count= n()) %>%
  unique(.) %>%
  arrange(desc(count)) %>%
  ungroup() %>%
  filter(V5!="")%>%
  top_n(n=14) %>%
  ggplot(aes(y=factor(V5,levels = rev(V5)), x=count,
             fill= factor(V5,levels = rev(V5)))) +
  geom_col()+
  theme_bw() +
  theme(legend.position="none",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  ylab("TF Class")+
  xlab("L596 Count")


ggpubr::ggarrange(L596_TF,H596_TF, ncol = 1)
```

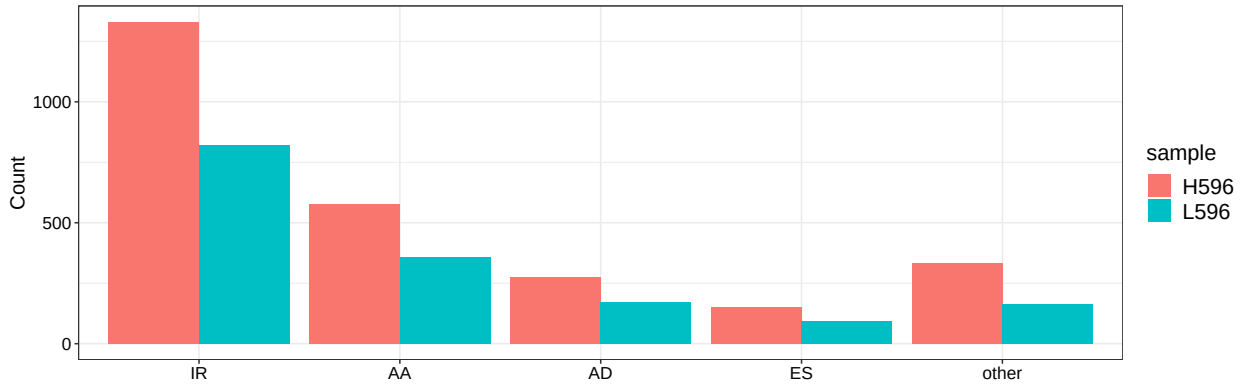





```
#ggsave("TF_class_stat.pdf", width = 25, height = 15)

rm(H596,L596,H596_TF, L596_TF)

setwd("F:/dir/2.Structral_analysis/2.AS")

AS<-read.table("AS.list", header = F, sep="\t")
H596_AS<-read.table("H596_AS.list", header = F, sep = "\t")
L596_AS<-read.table("L596_AS.list", header = F, sep = "\t")

all_AS<-
AS %>%
  right_join(rbind(H596_AS,L596_AS)   ,by=c("V1"="V2")) %>%
  mutate(class=ifelse(is.na(V2), "other", V2   )) %>%
  select(V1.y, class) %>%
  group_by(class) %>%
  #mutate(number= sum()  ) %>%
  summarize(number = sum(V1.y))  %>%
  select(class,number) %>%
  ungroup() %>%
#  unique(.)
  ggplot(aes(x=factor(class, levels = c("IR", "AA", "AD", "ES",
                                "other")),y=number,fill=class)) +
  geom_col()+
  theme_bw() +
  theme(legend.position="none",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
```

```
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("")+
  ylab("Count") +
  geom_text(aes(y=number+100,  label = number ))

sample_AS<-
  AS %>%
  right_join(rbind(H596_AS%>% mutate(sample="H596") ,
              L596_AS%>% mutate(sample="L596")  )   ,by=c("V1"="V2")) %>%
  mutate(class=ifelse(is.na(V2), "other", V2   )) %>%
  select(V1.y, class,sample) %>%
  group_by(class, sample) %>%
  #mutate(number= sum()  ) %>%
  summarize(number = sum(V1.y))  %>%
  select(class,number, sample) %>%
  ungroup() %>%
  #  unique(.)
  ggplot(aes(x=factor(class, levels = c("IR", "AA", "AD", "ES", "other")),
             y=number,fill=sample)) +
  geom_col(position = position_dodge() )+
  theme_bw() +
  theme(legend.position="right",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("")+
  ylab("Count")
  #geom_text(aes(y=number+50 ,label = number ))

ggpubr::ggarrange(all_AS, sample_AS, ncol = 1)
```

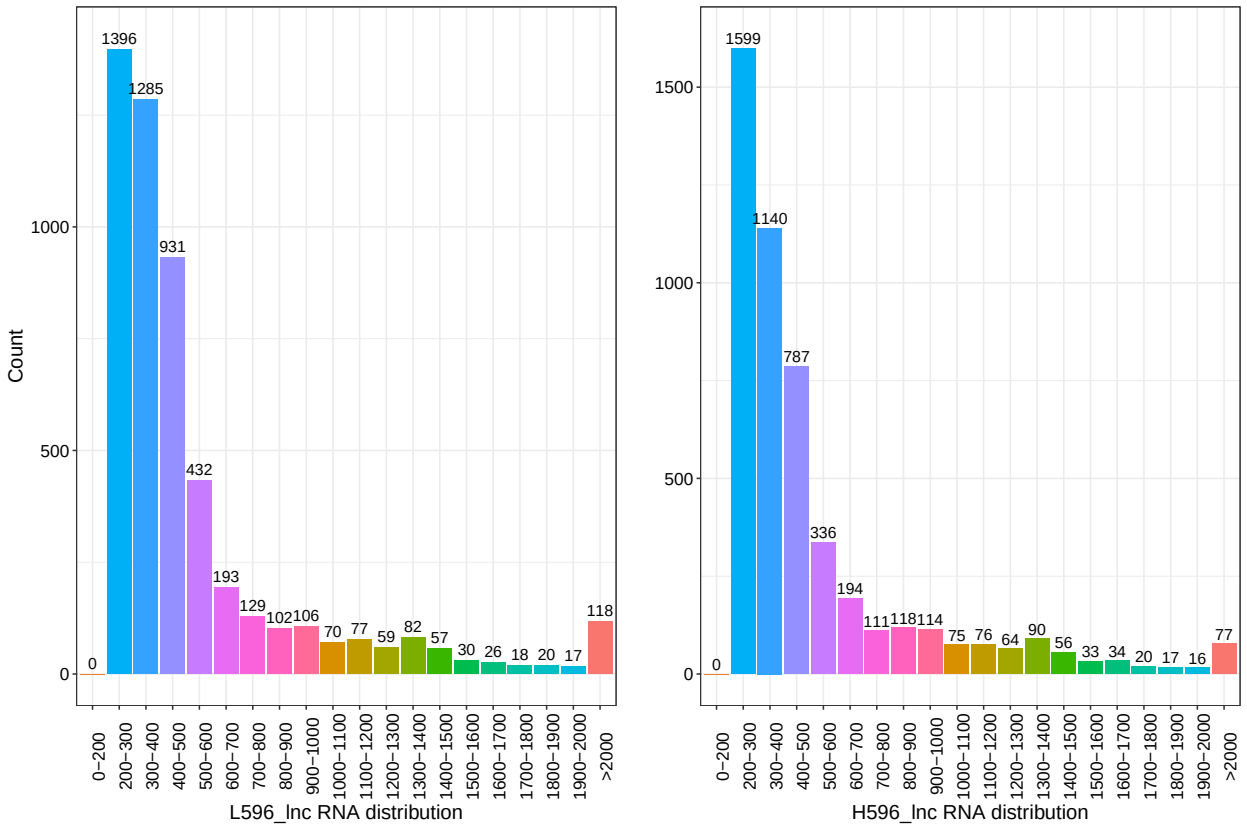





```
AS_sam<-
AS %>%
  right_join(rbind(H596_AS%>% mutate(sample="H596") ,L596_AS%>%
                     mutate(sample="L596")  )    ,by=c("V1"="V2")) %>%
  mutate(class=ifelse(is.na(V2), "other", V2   )) %>%
  select(V1.y, class,sample) %>%
  group_by(class, sample) %>%
  #mutate(number= sum()  ) %>%
  summarize(number = sum(V1.y))  %>%
  select(class,number, sample) %>%
  ungroup()

AS_al<-
  AS %>%
  right_join(rbind(H596_AS,L596_AS)    ,by=c("V1"="V2")) %>%
  mutate(class=ifelse(is.na(V2), "other", V2   )) %>%
  select(V1.y, class) %>%
  group_by(class) %>%
  #mutate(number= sum()  ) %>%
  summarize(number = sum(V1.y))  %>%
  select(class,number) %>%
  ungroup() %>%
  mutate(sample="all")


#write.csv(rbind(AS_al,AS_sam) %>%
#  spread(key = sample, value =number), row.names = F, quote = F , "AS_stat.csv")
```

```
#ggsave("AS_stat.pdf", height = 10, width = 7)


setwd("F:/dir/2.Structral_analysis/5.lncRNA")

L596_lnc<-read.table("m64032_191231_031514.subreads.6--6_75.ccs.lima.refine.cluster.hq..corrected.fastq
H596_lnc<-read.table("m64032_191231_031514.subreads.7--7_75.ccs.lima.refine.cluster.hq..corrected.fastq




L596_lnc$class<-cut(L596_lnc$V2,c( 0, seq(200,2000,100), Inf) ,
                    labels = c( paste(seq(0,1900,100)[-2],
                                      seq(200,2000,100) ,sep="-"), ">2000" ))
p1<-
L596_lnc %>%
  group_by(class) %>%
  mutate(num=n()) %>%
  arrange(V2) %>%
  select(class,num) %>%
  unique(.) %>%
  rbind( data.frame(class = c("0-200") , num=0   ), .) %>%
  ggplot(aes(x=  factor(class, levels = class), y=num, fill=class)) +
  geom_col()+
  theme_bw() +
  theme(legend.position="none",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black",angle = 90),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("L596_lnc RNA distribution")+
  ylab("Count") +
  geom_text(aes(label=num,y = num+25))



H596_lnc$class<-cut(H596_lnc$V2,c( 0, seq(200,2000,100), Inf) ,
                    labels = c( paste(seq(0,1900,100)[-2],
                                      seq(200,2000,100) ,sep="-"), ">2000" ))
p2<-
H596_lnc %>%
  group_by(class) %>%
  mutate(num=n()) %>%
  arrange(V2) %>%
  select(class,num) %>%
  unique(.) %>%
  rbind( data.frame(class = c("0-200") , num=0   ), .) %>%
  ggplot(aes(x=  factor(class, levels = class), y=num, fill=class)) +
  geom_col()+
  theme_bw() +
```

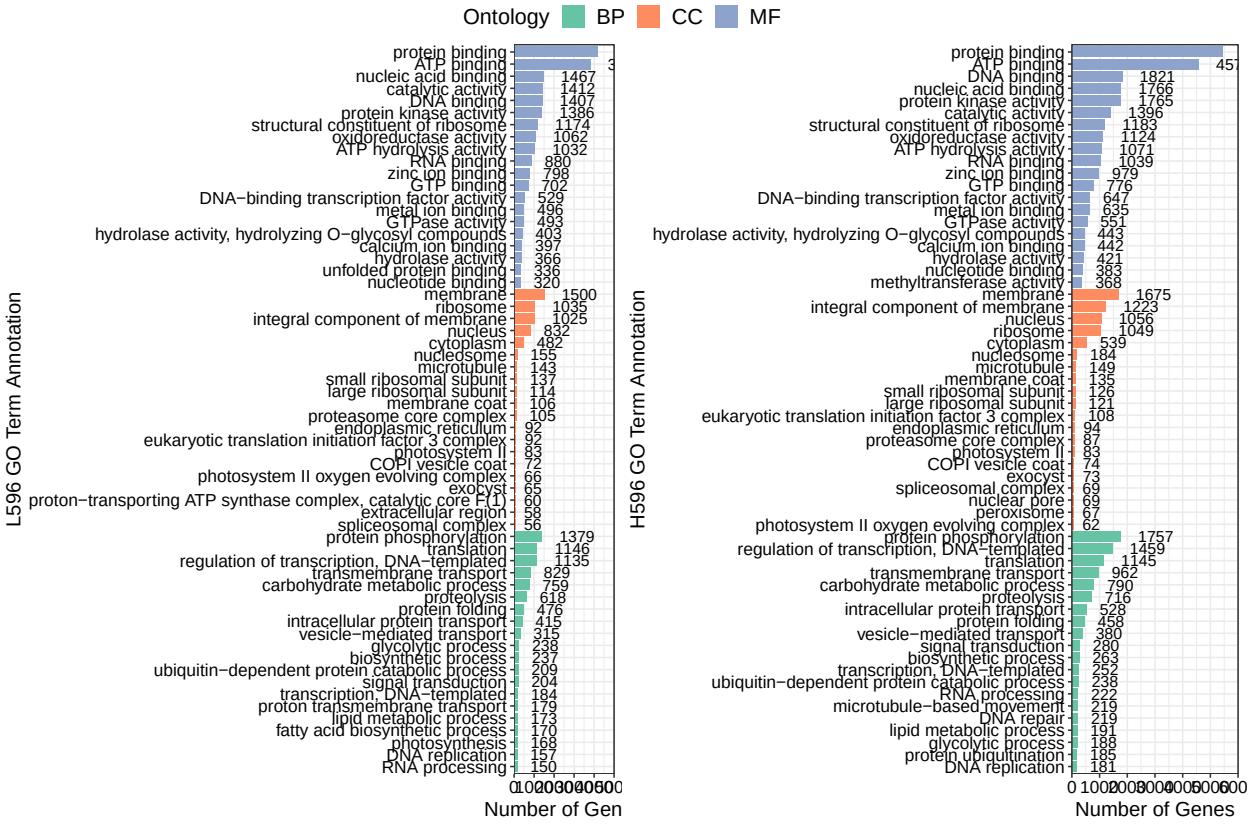
```
  theme(legend.position="none",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black",angle = 90),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("H596_lnc RNA distribution")+
  ylab("") +
  geom_text(aes(label=num,y = num+25))


ggpubr::ggarrange(p1,p2)
```




```
setwd("F:/dir/2.Structral_analysis/function_annonation/")
library(tidyverse)
library("clusterProfiler")


#term<-go2term(res1_go$V2)
#GO

L596<-read.table(file= "F:/dir/2.Structral_analysis/function_annonation/5.GO_KEGG/m64032_191231_031514.s

H596<-read.table("F:/dir/2.Structral_analysis/function_annonation/5.GO_KEGG/m64032_191231_031514.subread
```

```
all_term<-unique(rbind(go2term(L596$V2),go2term(H596$V2)))
all_class<-unique(rbind(go2ont(L596$V2), go2ont(H596$V2)))



L596_GO<-left_join(L596, all_term, by=c("V2"="go_id"))
H596_GO<-left_join(H596, all_term, by=c("V2"="go_id"))

L596_GO<-left_join(L596_GO, all_class, by=c("V2"="go_id"))
H596_GO<-left_join(H596_GO, all_class, by=c("V2"="go_id"))

#write.csv(L596_GO,"L596_go_anno.csv")
#write.csv(H596_GO,"H596_go_anno.csv")


L596_anno<-
na.omit(L596_GO) %>%
  group_by(Term) %>%
  mutate(count=n()) %>%
  ungroup() %>%
  group_by(Ontology) %>%
  arrange(desc(count)) %>%
  select(Term, count) %>%
  ungroup() %>%
  unique(.) %>%
  group_by(Ontology) %>%
  top_n(n=20,wt=count) %>%
  ungroup() %>%
  arrange(Ontology, count) %>%
  #ggplot(  aes(x=  factor(Term,levels = .$Term), y=count, fill=Ontology ))+
  ggplot(  aes(y=  factor(Term,levels =Term), x=count, fill= Ontology ))+
  geom_col() +
  scale_x_continuous(limits=c(0,5000   ), expand = c(0,0)) +
  #scale_y_continuous( breaks=c(0.5, 1.0, 1.5,2.0,2.5) ,limits=c(0,3), expand = c(0,0)) +
  geom_text(aes(label=count),size=4,hjust=-0.5)+
  scale_fill_manual(values=c("#66C3A4", "#FD8D61", "#8DA3CB")) +
  theme_bw() +
  theme(legend.position="right",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("Number of Genes") +
  ylab("L596 GO Term Annotation")



H596_anno<-
  na.omit(H596_GO) %>%
  group_by(Term) %>%
```

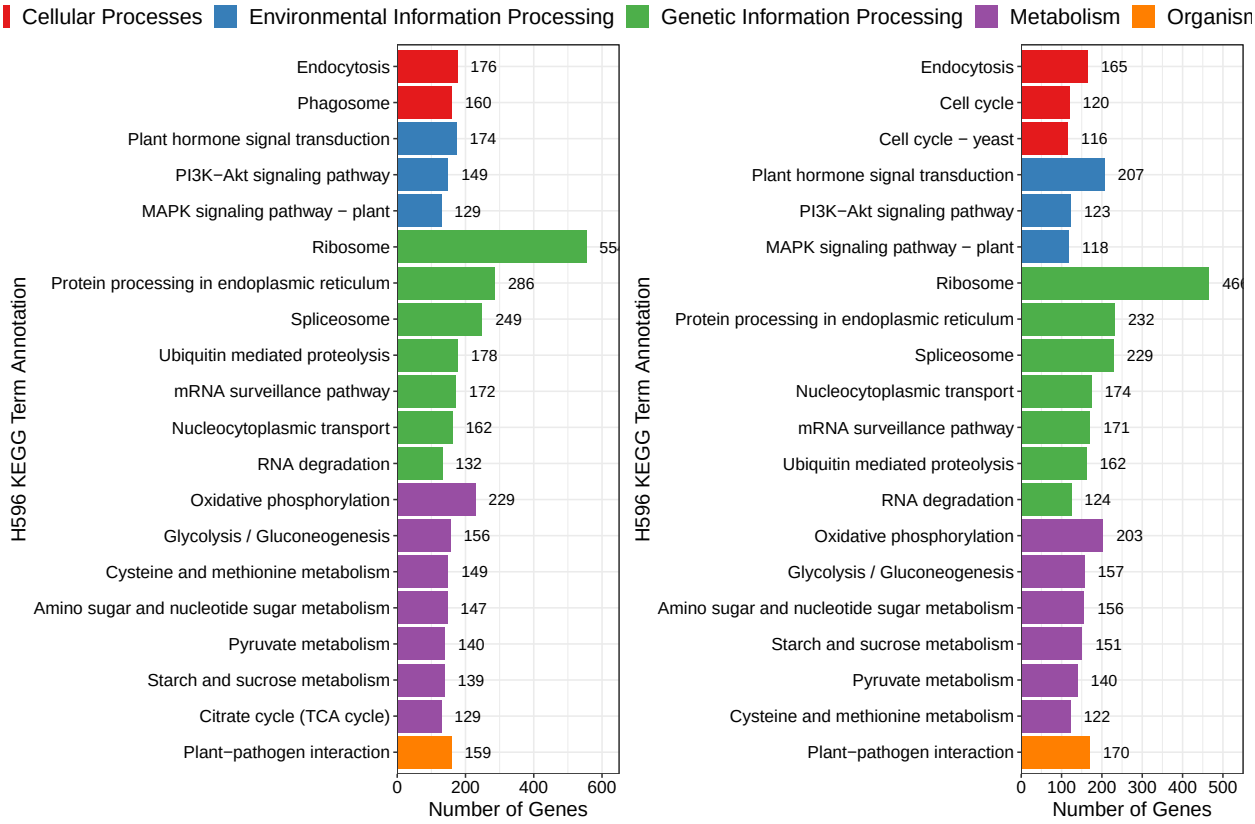
```
  mutate(count=n()) %>%
  ungroup() %>%
  group_by(Ontology) %>%
  arrange(desc(count)) %>%
  select(Term, count) %>%
  ungroup() %>%
  unique(.) %>%
  group_by(Ontology) %>%
  top_n(n=20,wt=count) %>%
  ungroup() %>%
  arrange(Ontology, count) %>%
  #ggplot(  aes(x=  factor(Term,levels = .$Term), y=count, fill=Ontology ))+
  ggplot(  aes(y=  factor(Term,levels =Term), x=count, fill= Ontology ))+
  geom_col() +
  scale_x_continuous(limits=c(0,6000   ), expand = c(0,0)) +
  #scale_y_continuous( breaks=c(0.5, 1.0, 1.5,2.0,2.5) ,limits=c(0,3), expand = c(0,0)) +
  geom_text(aes(label=count),size=4,hjust=-0.5)+
  scale_fill_manual(values=c("#66C3A4", "#FD8D61", "#8DA3CB")) +
  theme_bw() +
  theme(legend.position="right",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("Number of Genes") +
  ylab("H596 GO Term Annotation")



ggpubr::ggarrange(L596_anno,H596_anno ,common.legend = T)
```




```
#ggsave("5.GO_KEGG/GO_annotation.pdf", width = 23, height = 10)


L596_ko<-read.table("F:/dir/2.Structral_analysis/function_annonation/5.GO_KEGG/6.KEGGpathway.annot", sep
H596_ko<-read.table("F:/dir/2.Structral_analysis/function_annonation/5.GO_KEGG/7.KEGGpathway.annot", sep

khiA<-read.table("F:/dir/2.Structral_analysis/function_annonation/5.GO_KEGG/khier.tsv",header = T)

khiA<-
khiA %>%
  separate(col = pathway, sep = " ", into = c("ko","pathway")) %>%
  select(category, ko) %>%
  mutate(ko_id= paste("ko",ko,sep=""))


L_ko<-
L596_ko %>%
  left_join(khiA, by=c("V4"= "ko_id")) %>%
  mutate(class= ifelse(is.na(category), "Unknown", category)) %>%
  #filter(!is.na(category) ) %>%
  select(V1,V7,class) %>%
  unique(.) %>%
  select(V7,class) %>%
  group_by(V7) %>%
  mutate(count=n()) %>%
  arrange(desc( count)) %>%
  unique(.) %>%
```

```
  ungroup() %>%
  filter(class!="Human Diseases", class!="Drug Development", class!="Unknown") %>%
#  group_by(class) %>%
  top_n(n=20,wt=count) %>%
  mutate( Pcount=ifelse(count>1000, 800,count)) %>%
  arrange(class) %>%
  #group_by(class) %>%
  ggplot(  aes(y=  factor(V7,levels = rev(V7)) , x=Pcount , fill=class ))+
  geom_col() +
  scale_x_continuous(limits=c(0,650   ), expand = c(0,0)) +
  #scale_y_continuous( breaks=c(0.5, 1.0, 1.5,2.0,2.5) ,limits=c(0,3), expand = c(0,0)) +
  geom_text(aes(label=count),size=4,hjust=-0.5)+
  scale_fill_manual(name="Category",values=c("#E41A1C", "#377EB8", "#4DAF4A", "#984EA3", "#FF7F00")) +
  theme_bw() +
  theme(legend.position="right",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("Number of Genes") +
  ylab("H596 KEGG Term Annotation")



H_ko<-
H596_ko %>%
  left_join(khiA, by=c("V4"= "ko_id")) %>%
  mutate(class= ifelse(is.na(category), "Unknown", category)) %>%
  #filter(!is.na(category) ) %>%
  select(V1,V7,class) %>%
  unique(.) %>%
  select(V7,class) %>%
  group_by(V7) %>%
  mutate(count=n()) %>%
  arrange(desc( count)) %>%
  unique(.) %>%
  ungroup() %>%
  filter(class!="Human Diseases", class!="Drug Development", class!="Unknown") %>%
  #  group_by(class) %>%
  top_n(n=20,wt=count) %>%
  mutate( Pcount=ifelse(count>1000, 800,count)) %>%
  arrange(class) %>%
  #group_by(class) %>%
  ggplot(  aes(y=  factor(V7,levels = rev(V7)) , x=Pcount , fill=class ))+
  geom_col() +
  scale_x_continuous(limits=c(0,550   ), expand = c(0,0)) +
  #scale_y_continuous( breaks=c(0.5, 1.0, 1.5,2.0,2.5) ,limits=c(0,3), expand = c(0,0)) +
  geom_text(aes(label=count),size=4,hjust=-0.5)+
  scale_fill_manual(name="Category",values=c("#E41A1C", "#377EB8", "#4DAF4A", "#984EA3", "#FF7F00")) +
```

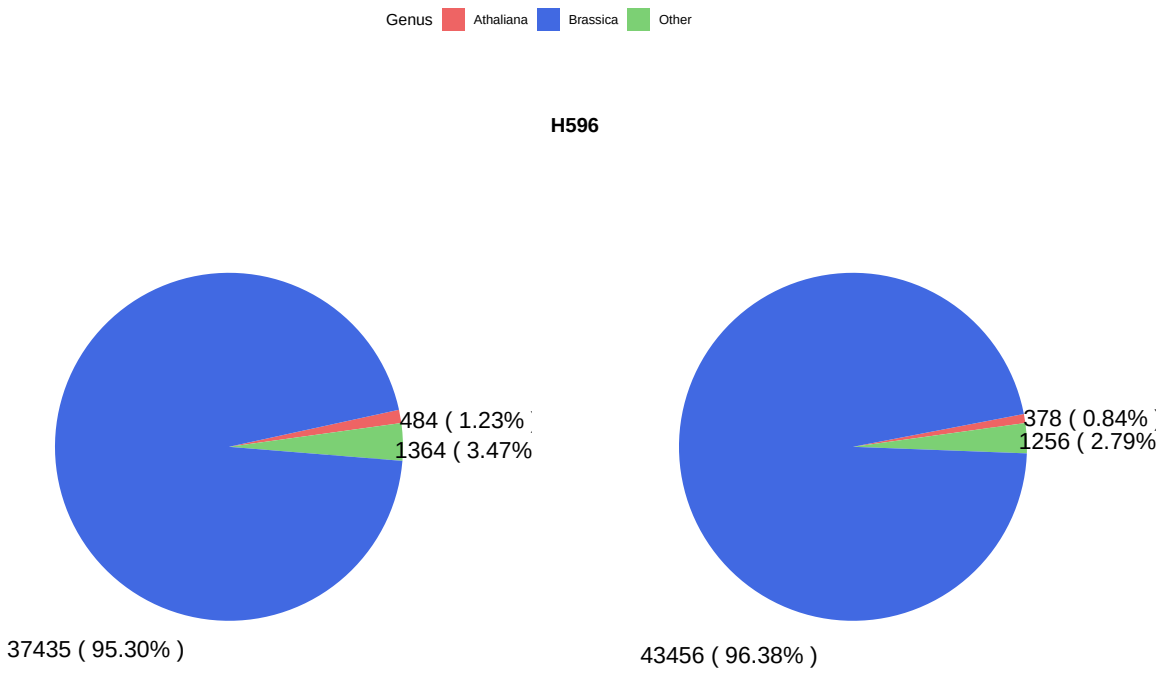
```
  theme_bw() +
  theme(legend.position="right",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5))+
  xlab("Number of Genes") +
  ylab("H596 KEGG Term Annotation")

ggpubr::ggarrange(L_ko, H_ko, common.legend = T)
```




```
#ggsave("5.GO_KEGG/KEGG_annotation.pdf", width = 23, height = 10)

rm(H_ko)
rm(L_ko)
rm(H596_anno)
rm(L596_anno)


H596_NR<-read.table("F:/dir/2.Structral_analysis/function_annonation/3.NR/m64032_191231_031514.subreads
L596_NR<-read.table("F:/dir/2.Structral_analysis/function_annonation/3.NR/m64032_191231_031514.subreads


L596_NR$Subject_annotation<-gsub("\\]", "",gsub(".*\\[", "", L596_NR$Subject_annotation))
```

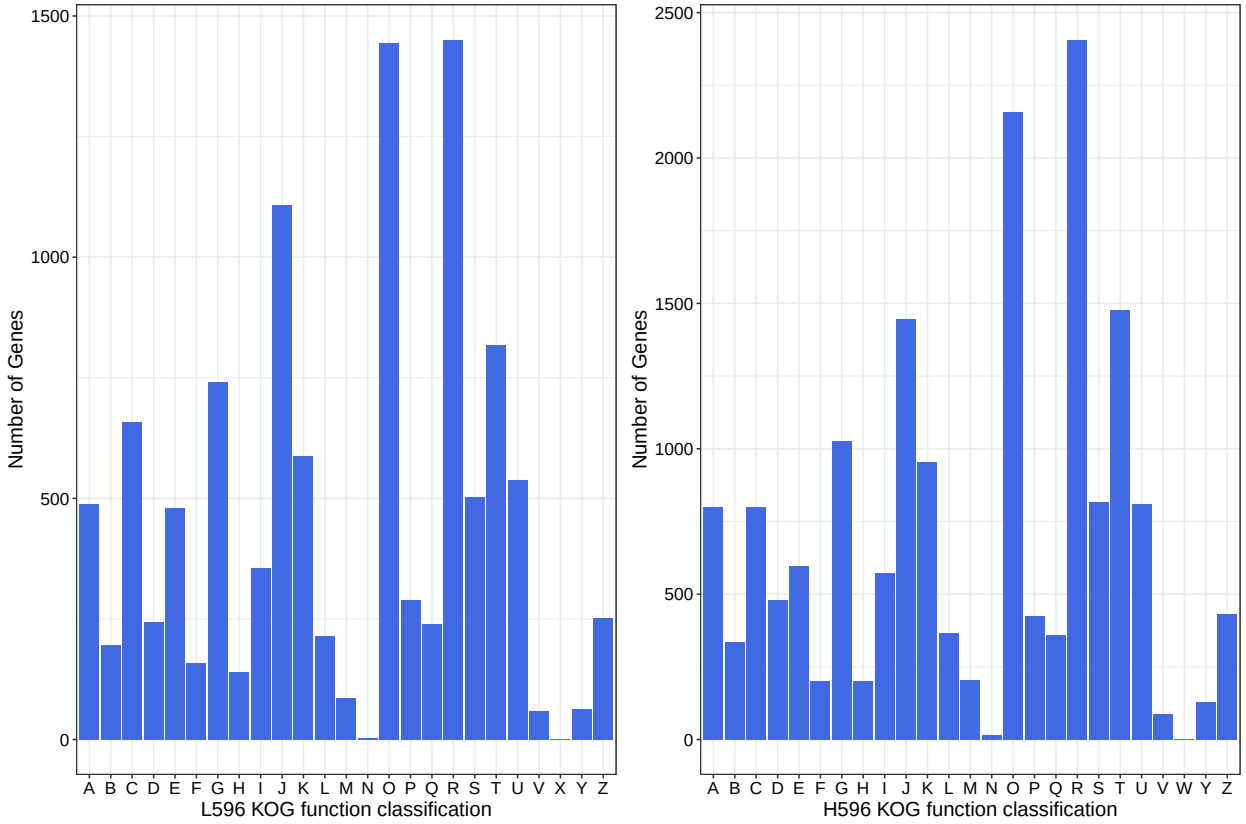
```
H596_NR$Subject_annotation<-gsub("\\]", "",gsub(".*\\[", "", H596_NR$Subject_annotation))


blank_theme <- theme_minimal()+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.border = element_blank(),
    panel.grid=element_blank(),
    axis.ticks = element_blank(),
    plot.title=element_text(size=14, face="bold")
  )

L596_NR_p<-
L596_NR %>%
  mutate(Genus=ifelse(grepl("Brassica", Subject_annotation), "Brassica", ifelse(grepl("Arabidopsis", Su
  group_by(Genus) %>%
  mutate(count=n(),c="c") %>%
  select(Genus, count,c)%>%
  unique(.) %>%
  arrange(count)  %>%
  mutate(prop= count/sum(.$count), per=scales::percent(prop,accuracy = 0.01) ,lab= paste(count,"(",per,
  ungroup() %>%
  ggplot(aes(y=prop,x=c, fill=Genus)) +
  geom_bar(stat = 'identity',position = 'stack')+
  coord_polar(theta = 'y',start = 14) + labs(x = '', y = '', title = 'L596')+
  #theme(axis.text = element_blank()) +
  geom_text(aes(x=1.8,y = prop/3 + c(0, cumsum(prop)[-length(prop)]),
                label = lab), size=5.5) +
  blank_theme +
  theme(axis.text = element_blank()) +
  scale_fill_manual(values=c("#EE6464","#4169E2",  "#7CD074"))


H596_NR_p<-
H596_NR %>%
  mutate(Genus=ifelse(grepl("Brassica", Subject_annotation), "Brassica", ifelse(grepl("Arabidopsis", Su
  group_by(Genus) %>%
  mutate(count=n(),c="c") %>%
  select(Genus, count,c)%>%
  unique(.) %>%
  arrange(count)  %>%
  mutate(prop= count/sum(.$count), per=scales::percent(prop,accuracy = 0.01) ,lab= paste(count,"(",per,
  ungroup() %>%
  ggplot(aes(y=prop,x=c, fill=Genus)) +
  geom_bar(stat = 'identity',position = 'stack')+
  coord_polar(theta = 'y',start = 14) + labs(x = '', y = '', title = 'H596')+
  #theme(axis.text = element_blank()) +
  geom_text(aes(x=1.8,y = prop/3 + c(0, cumsum(prop)[-length(prop)]),
                label = lab), size=5.5) +
  blank_theme +
  theme(axis.text = element_blank()) +
  scale_fill_manual(values=c("#EE6464","#4169E2",  "#7CD074"))
```

```
ggpubr::ggarrange(L596_NR_p,H596_NR_p,common.legend = T, legend = "top")
```




```
ggsave("3.NR/Genus_Distribution.pdf", height = 12, width = 15)

rm(H596_NR,H596_NR_p,L596_NR,L596_NR_p)

L596_KOG<-read.table("F:/dir/2.Structral_analysis/function_annonation/1.COG_KOG/m64032_191231_031514.6-
H596_KOG<-read.table("F:/dir/2.Structral_analysis/function_annonation/1.COG_KOG/m64032_191231_031514_7-

H596_KOG_p<-
H596_KOG %>%
  ggplot(aes(x=V4)) +
  geom_bar(fill= "#4169E2")+
  theme_bw() +
  theme(legend.position="right",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5)) +
  xlab("H596 KOG function classification") +
  ylab("Number of Genes")

L596_KOG_p<-
```

```
L596_KOG %>%
  ggplot(aes(x=V4)) +
  geom_bar(fill= "#4169E2")+
  theme_bw() +
  theme(legend.position="right",
        axis.title.x=element_text(size=14,color="black",hjust=0.5),
        axis.title.y=element_text(size=14,color="black"),
        axis.text.x=element_text(size=12,color="black"),
        axis.text.y=element_text(size=12,color="black"),
        legend.text=element_text(size=15,color="black"),
        legend.title=element_text(size=15,color="black"),
        plot.title=element_text(size=15,color="black",hjust=0.5)) +
  xlab("L596 KOG function classification") +
  ylab("Number of Genes")

ggpubr::ggarrange(L596_KOG_p,H596_KOG_p)
```




```
#ggsave("1.COG_KOG/KOG_anno.pdf", width = 16, height = 8)
```