

expression divergence

Changfu Jia

2022-07-19

```
#knitr::opts_chunk$set(echo = TRUE)
#library(knitr)
knitr::opts_chunk$set(fig.width=12, fig.height=8,
                      echo=TRUE, warning=FALSE, message=FALSE)

#kable(data)
options(warn=-1)

library(tidyverse)
setwd("F:/dir")

#at_tpm <- read.table("input-data/expression_dta/Atha_more_tissue2.tsv2", header = T, sep = "\t")

orth <- read.table("input-data/wgdi/dup_gene.tsv"
                  ,header = T, sep = "\t", na.strings = "-")
orth <- orth %>% filter(class=="wgdi") %>% select(-OG , -N1 , -N11)

six_tis<-read.table("input-data/expression_data/Atha_Ov_sixtis.tsv",
                   header = T, sep = " ")
six_tis<-six_tis %>% select(-TPC,-TPCA)
#c( "flower" , "leaf" , "root" , "stem" , "silique", "seed" )
#cbind(orth,six_tis)

kaks<-read.table("input-data/wgdi/ov.kaks", header=T, sep="\t")

kaks[, "id1"] <- gsub("\\.t.*", "", kaks[, "id1"])
kaks[, "id2"] <- gsub("\\.t.*", "", kaks[, "id2"])

#kaks %>%
# mutate(id3=paste(id1,id2,sep=""), id4=paste(id2,id1,sep=""))

kaks_bind<-data.frame()
for(i in 1:nrow(orth)){

  dup1<-orth[i, "dup1"]
  dup2<-orth[i, "dup2"]
  dup1_ks<-kaks[which(kaks$id1 %in% dup1),]
  if ( any((dup1_ks$id2%in% dup2) == "TRUE" ) ){
    dup2_ks<-dup1_ks[which(dup1_ks$id2%in% dup2),]
  }else{
    dup2_ks<- t(as.data.frame(rep(NA,6)))
  }
}
```

```

#kaks_bind<-cbind(kaks_bind,orth[i,],dup2_ks)
if (i==1){
  kaks_bind<-cbind(orth[i,],dup2_ks)
  colnames(kaks_bind)<-1:12
}else{
  LL<-cbind(orth[i,],dup2_ks)
  colnames(LL)<-1:12
  kaks_bind<-rbind(kaks_bind,LL)
}
}

```

```

Tau<-function(x){
  aa<- apply(x ,1,function(x){x/max(x)})
  tau<-as.data.frame(colSums(1-aa)/(ncol(x)-1))
  return(tau)
}

```

#Filter genes with all six tissues TPM 0; #calculate Tau and Ka Ks; #calculate Pearson cor and EI between each homeologous gene pairs

```

orth_tbl<-
cbind(orth,six_tis,kaks_bind[,9:10]) %>%
  rename(c("ka"="9", "ks"="10")) %>%
  mutate(kaks= ka/ks) %>%
  na.omit() %>%
  mutate(EI = 0, Pearson=0, pvalue=0, Tau_dup1=0, Tau_dup2=0)

for (i in 1:nrow(orth_tbl)){
  dup1<-orth_tbl[i,7:12]
  dup2<-orth_tbl[i,13:18]
  rownames(dup1)<-orth_tbl[i,2]
  rownames(dup2)<-orth_tbl[i,3]
  mat<-log(rbind(as.matrix(dup1),as.matrix(dup2))+ 0.01 )
  #EI calculate
  orth_tbl[i,"EI"]<-dist( mat)[1]
  #cor(t(mat))[1,2]
  #Pearson
  pears<-cor.test(as.matrix(log2(dup1+0.01)), as.matrix(log2(dup2+0.01)))
  orth_tbl[i,"Tau_dup1"]<-Tau(dup1)[1,1]
  orth_tbl[i,"Tau_dup2"]<-Tau(dup2)[1,1]
  orth_tbl[i,"Pearson"]<-pears$estimate
  orth_tbl[i,"pvalue"]<-pears$p.value
}

#write.table(orth_tbl,file="orth_tbl_intra.txt", quote = F, sep="\t", row.names = F )

```

```

EI_kaks<-
orth_tbl %>%
  select(EI,kaks) %>%
  mutate(EIclass=ifelse(EI<=2,"0-2",
                        ifelse(EI<=4,"2-4" ,
                              ifelse(EI<=6, "4-6", ifelse(EI<=8, "6-8", "8") ) ) ) ) %>%
  mutate(KAKSclass= ifelse(kaks<=0.1,"0.1",
                          ifelse(kaks<=0.2,"0.1-0.2" ,
                                ifelse(kaks<=0.3, "0.2-0.3",
                                      ifelse(kaks<=0.4, "0.3-0.4", "0.4") ) ) ) ) %>%

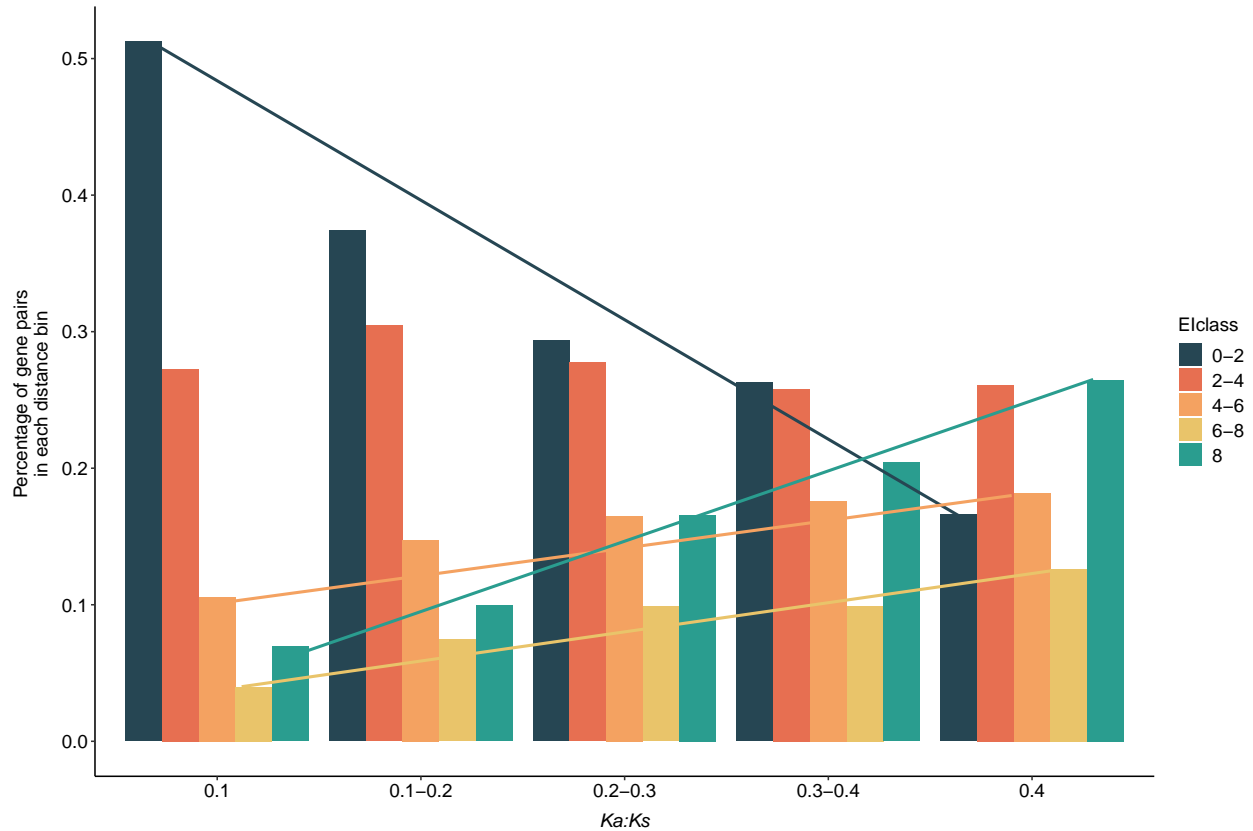
  group_by(EIclass,KAKSclass) %>%
  summarise(n=n() ) %>%
  ungroup() %>%
  group_by(KAKSclass) %>%
  summarise(all=sum(n))

orth_tbl %>%
  select(EI,kaks) %>%
  mutate(EIclass=ifelse(EI<=2,"0-2",
                        ifelse(EI<=4,"2-4" ,
                              ifelse(EI<=6, "4-6", ifelse(EI<=8, "6-8", "8") ) ) ) ) %>%
  mutate(KAKSclass= ifelse(kaks<=0.1,"0.1",
                          ifelse(kaks<=0.2,"0.1-0.2" ,
                                ifelse(kaks<=0.3, "0.2-0.3",
                                      ifelse(kaks<=0.4, "0.3-0.4", "0.4") ) ) ) ) %>%

  group_by(EIclass,KAKSclass) %>%
  summarise(n=n()) %>%
  ungroup() %>%
  left_join(EI_kaks) %>%
  mutate(prop=n/all) %>%
  ggplot(aes(x=KAKSclass,y=prop, fill=EIclass))+
  geom_col( position='dodge') +
  theme_classic() +
  theme(
    axis.title.x=element_text(size=13,color="black",hjust=0.5, face="italic", vjust = -1),
    axis.text.x=element_text(size=13,color="black"),
    axis.text.y=element_text(size=13,color="black"),
    legend.text=element_text(size=13,color="black"),
    legend.title=element_text(size=13,color="black"),
    plot.title=element_text(size=13,color="black",face="italic",hjust=0.5)) +
  ylab(paste("Percentage of gene pairs ", "in each distance bin ", sep="\n" ))+
  xlab("Ka:Ks") +
  scale_fill_manual(values=c("0-2"= "#264653",
                             "2-4"="#e76f51", "4-6"="#f4a261", "6-8"="#e9c46a", "8"="#2a9d8f" ) ) +
  annotate("segment", x=0.7, xend=4.7, y=0.51, yend=0.16,size=1, color="#264653" ) +
  annotate("segment", x=0.95, xend=4.9, y=0.1, yend=0.18,size=1, color="#f4a261" ) +
  annotate("segment", x=1.12, xend=5.1, y=0.04, yend=0.125,size=1, color="#e9c46a" ) +

```

```
annotate("segment", x=1.32, xend=5.3, y=0.06, yend=0.265, size=1, color="#2a9d8f" )
```



```
#dir.create("result")
#ggsave("result/1.EI_KAKS.pdf",width = 12,height = 7)

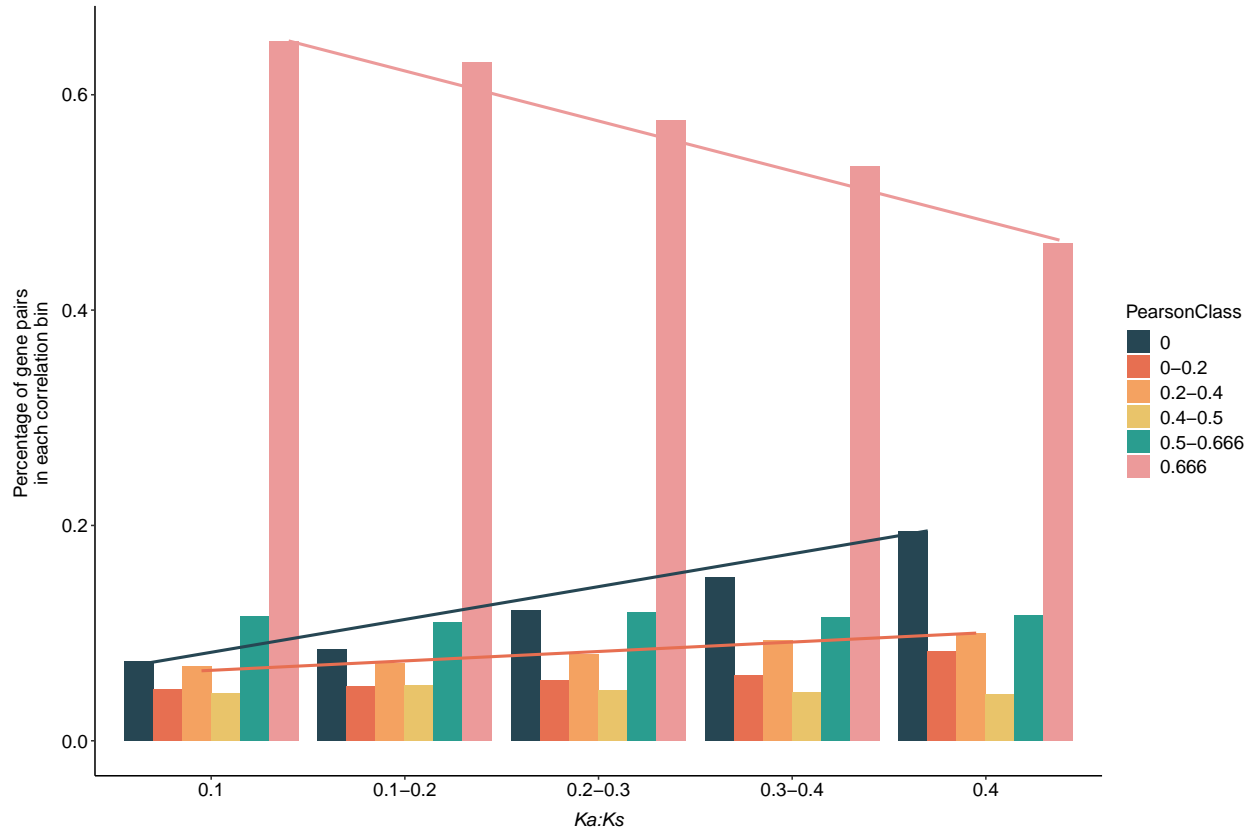
Pr_kaks<-
orth_tbl %>%
  select(Pearson,kaks) %>%
  mutate(PearsonClass=ifelse(Pearson<=0,"0",
    ifelse(Pearson<=0.2,"0-0.2",
    ifelse(Pearson<=0.4,"0.2-0.4",
    ifelse(Pearson<=0.5,"0.4-0.5",
    ifelse(Pearson<=0.666,"0.5-0.666","0.666")))) %>%
  mutate(KAKSclass= ifelse(kaks<=0.1,"0.1",
    ifelse(kaks<=0.2,"0.1-0.2",
    ifelse(kaks<=0.3,"0.2-0.3",
    ifelse(kaks<=0.4,"0.3-0.4","0.4")))) %>%
  group_by(PearsonClass,KAKSclass) %>%
  summarise(n=n()) %>%
  ungroup() %>%
  group_by(KAKSclass) %>%
  summarise(all=sum(n))

orth_tbl %>%
  select(Pearson,kaks) %>%
```

```

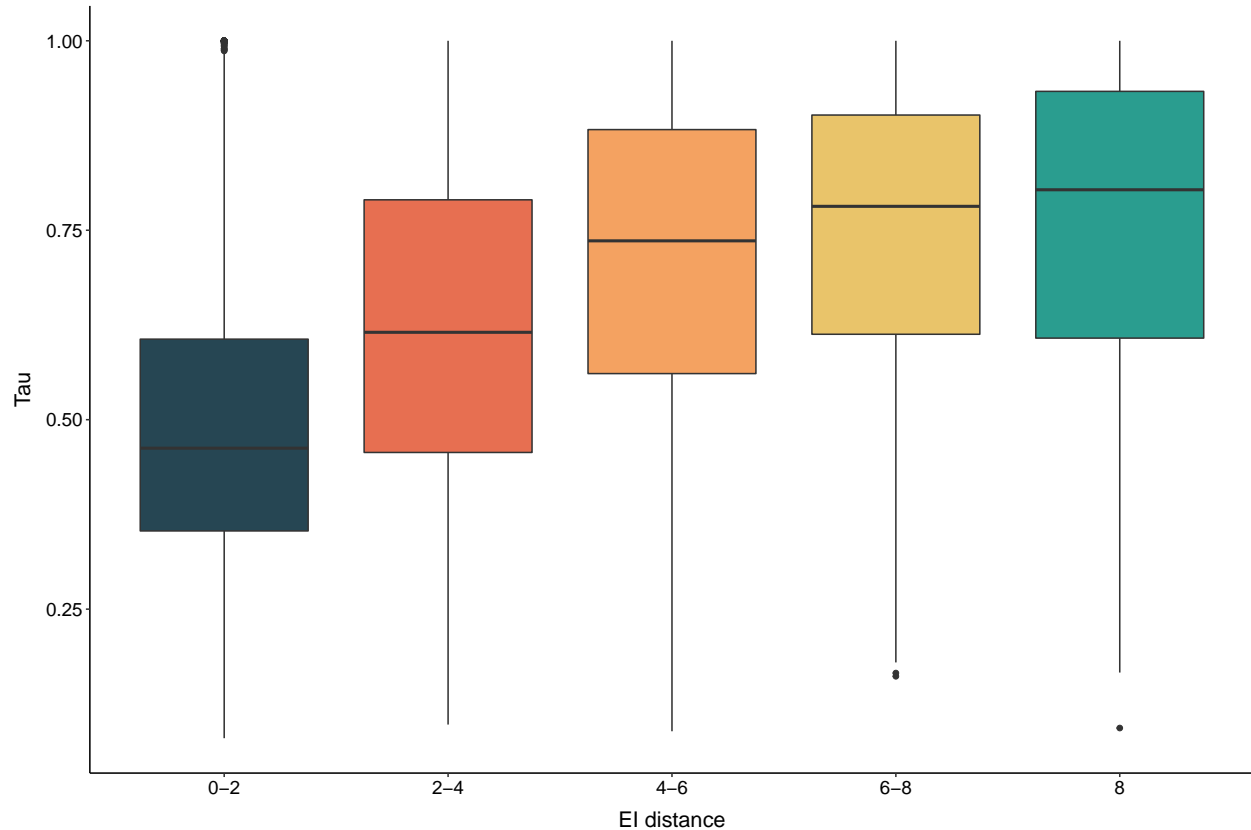
mutate(PearsonClass=ifelse(Pearson<=0,"0",
  ifelse(Pearson<=0.2,"0-0.2" ,
    ifelse(Pearson<=0.4, "0.2-0.4",
      ifelse(Pearson<=0.5, "0.4-0.5",
        ifelse(Pearson<=0.666, "0.5-0.666", "0.666") ) ) ) ) ) %>%
mutate(KAKSclass= ifelse(kaks<=0.1,"0.1",
  ifelse(kaks<=0.2,"0.1-0.2" ,
    ifelse(kaks<=0.3, "0.2-0.3",
      ifelse(kaks<=0.4, "0.3-0.4", "0.4") ) ) ) ) %>%
group_by(PearsonClass,KAKSclass) %>%
summarise(n=n() ) %>%
ungroup() %>%
left_join(Pr_kaks)%>%
mutate(prop=n/all) %>%
ggplot(aes(x=KAKSclass,y=prop, fill=PearsonClass))+
geom_col( position='dodge') +
theme_classic() +
theme(
axis.title.x=element_text(size=13,color="black",hjust=0.5, face="italic", vjust = -1),
axis.text.x=element_text(size=13,color="black"),
axis.text.y=element_text(size=13,color="black"),
legend.text=element_text(size=13,color="black"),
legend.title=element_text(size=13,color="black"),
plot.title=element_text(size=13,color="black",face="italic",hjust=0.5)) +
ylab(paste("Percentage of gene pairs ", "in each correlation bin ", sep="\n" ))+
xlab("Ka:Ks")+
scale_fill_manual(values=c("0"= "#264653",
  "0-0.2"="#e76f51", "0.2-0.4"="#f4a261",
  "0.4-0.5"="#e9c46a", "0.5-0.666"="#2a9d8f", "0.666"="#ec9a9a" ) ) +
annotate("segment", x=0.6, xend=4.7, y=0.07, yend=0.195,size=1, color="#264653" ) +
annotate("segment", x=0.95, xend=4.95, y=0.065, yend=0.1,size=1, color="#e76f51" ) +
annotate("segment", x=1.4, xend=5.38, y=0.065, yend=0.465,size=1, color="#ec9a9a" )

```



```
#ggsave("result/1.PR_KAKS.pdf",width = 12)
```

```
orth_tbl %>%
  select(EI, Tau_dup1, Tau_dup2 ) %>%
  gather(key=dup, value= Tau, 2:3) %>%
  mutate(EIclass=ifelse(EI<=2,"0-2", ifelse(EI<=4,"2-4" ,
    ifelse(EI<=6, "4-6", ifelse(EI<=8, "6-8", "8") ) ) ) ) %>%
  ggplot(aes(x=EIclass, y=Tau, fill=EIclass)) +
  geom_boxplot() +
  # geom_jitter()+
  theme_classic() +
  theme(legend.position = "none",
    axis.title.x=element_text(size=15,color="black",hjust=0.5, vjust = -1),
    axis.text.x=element_text(size=13,color="black"),
    axis.text.y=element_text(size=13,color="black"),
    legend.text=element_text(size=13,color="black"),
    legend.title=element_text(size=13,color="black"),
    plot.title=element_text(size=13,color="black",hjust=0.5)) +
  ylab("Tau")+
  xlab("EI distance") +
  scale_fill_manual(values=c("0-2"= "#264653",
    "2-4"="#e76f51", "4-6"="#f4a261", "6-8"="#e9c46a", "8"="#2a9d8f" ) )
```

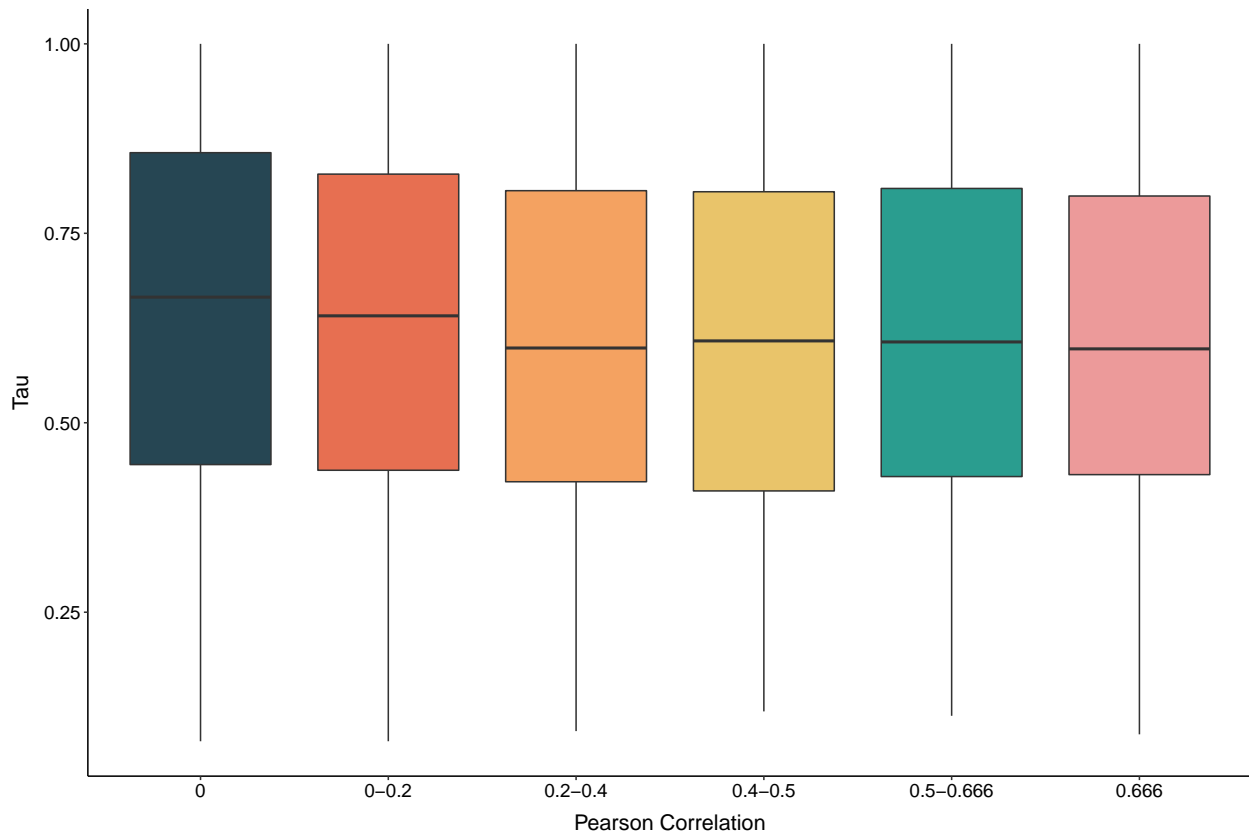


```
#ggsave("result/1.EI_Tau.pdf",width = 12,height = 7)
```

```
library(ggthemes)
```

```
orth_tbl %>%
  select(Pearson, Tau_dup1, Tau_dup2 ) %>%
  gather(key=dup, value= Tau, 2:3) %>%
  mutate(PearsonClass=ifelse(Pearson<=0,"0",
    ifelse(Pearson<=0.2,"0-0.2" , ifelse(Pearson<=0.4, "0.2-0.4",
ifelse(Pearson<=0.5, "0.4-0.5",
      ifelse(Pearson<=0.666, "0.5-0.666", "0.666") ) ) ) ) ) %>%
  ggplot(aes(x=PearsonClass, y=Tau, fill=PearsonClass)) +
  geom_boxplot() +
  #geom_jitter()+
  theme_classic() +
  theme(legend.position = "none",
axis.title.x=element_text(size=15,color="black",hjust=0.5, vjust = -1),
axis.text.x=element_text(size=13,color="black"),
axis.text.y=element_text(size=13,color="black"),
legend.text=element_text(size=13,color="black"),
legend.title=element_text(size=13,color="black"),
plot.title=element_text(size=13,color="black",hjust=0.5)) +
ylab("Tau")+
xlab("Pearson Correlation") +
```

```
scale_fill_manual(values=c("0"= "#264653",
  "0-0.2"="#e76f51", "0.2-0.4"="#f4a261",
  "0.4-0.5"="#e9c46a", "0.5-0.666"="#2a9d8f", "0.666"="#ec9a9a" ) )
```



```
#ggsave("result/1.Pearson_Tau_box.pdf",width = 12,height = 7)
```

```
library(pheatmap)
all_tpm_tbl<-orth_tbl[7:24]
rownames(all_tpm_tbl)<-orth_tbl$ref

dist.obs.tis<-as.dist(1-cor(t(all_tpm_tbl)))
dist.obs.tis.tre<- hclust(dist.obs.tis, method = "ward.D")

all_tpm_tbl_plot<-cbind(orth_tbl[19:24],orth_tbl[7:12],orth_tbl[13:18], orth_tbl[19:24] )

#all_tpm_tbl_plot<-log2(all_tpm_tbl_plot+0.01)

for (i in 1:6) {
  all_tpm_tbl_plot[,24+i]<- (all_tpm_tbl_plot[,i]+ all_tpm_tbl_plot[,6+i])
}

all_tpm_tbl_plot<-log2(all_tpm_tbl_plot+0.01)

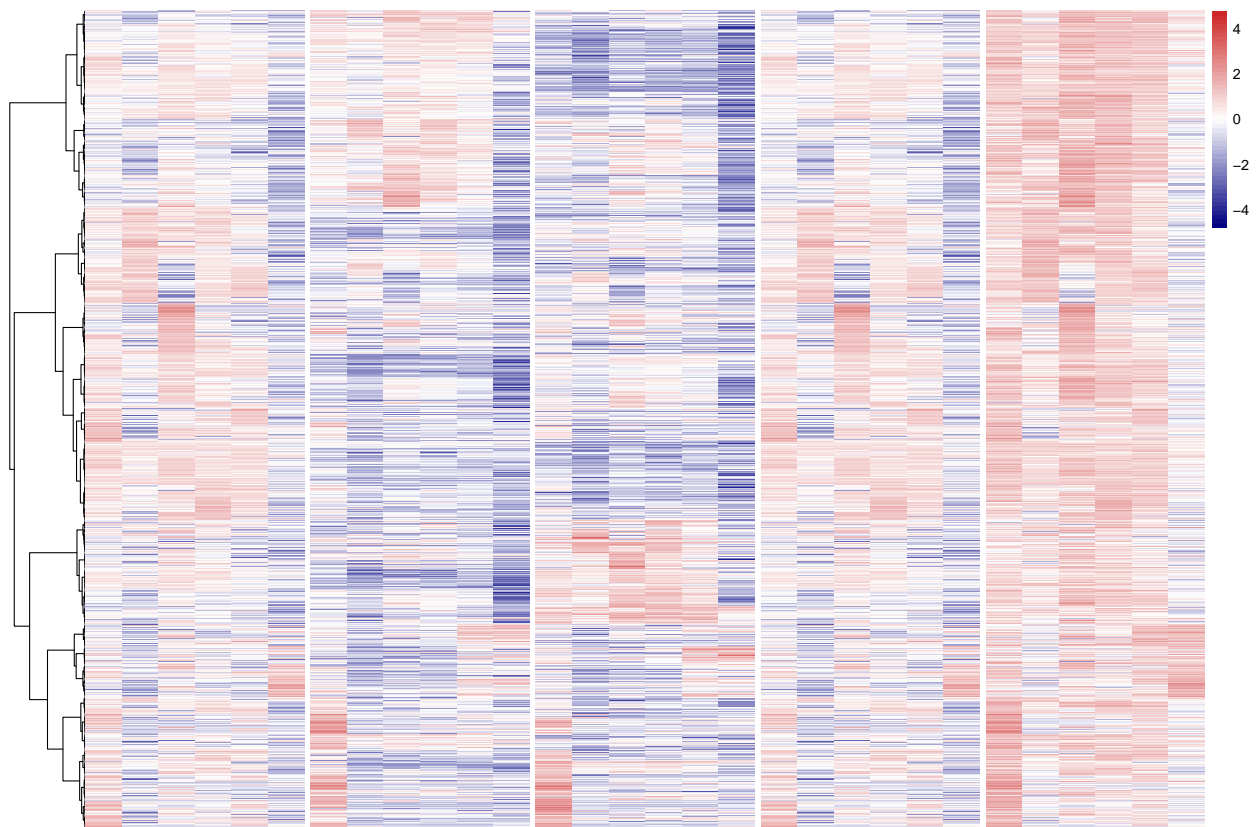
pheatmap(mat=as.matrix(all_tpm_tbl_plot),gaps_col = c(6,12,18,24),
```



```

scale = "row",
show_rownames = FALSE,
# color = colorRampPalette(c("blue","white","red"))(10),
color = colorRampPalette(c("navy", "white", "firebrick3"))(50),
#treeheight_row = 200,
#treeheight_col = 100,
cluster_rows = dist.obs.tis.tre,
cluster_cols = F,
#annotation_row=annotation_row,
annotation_legend = FALSE,
annotation_names_row = F,
show_colnames = F
#annotation_colors = ann_colors
#cutree_cols = c(6,12,18,24)
)

```



```

methy_orth_tbl<-
orth_tbl %>%
  select(EI,kaks,dup1, dup2,Pearson) %>%
  mutate(EIclass=ifelse(EI<=2,"0-2",
    ifelse(EI<=4,"2-4",
      ifelse(EI<=6, "4-6", ifelse(EI<=8, "6-8", "8") ) ) ) ) %>%
  mutate(KAKSclass= ifelse(kaks<=0.1,"0.1",
    ifelse(kaks<=0.2,"0.1-0.2", ifelse(kaks<=0.3, "0.2-0.3",
      ifelse(kaks<=0.4, "0.3-0.4", "0.4") ) ) ) ) %>%
  mutate(PearsonClass=ifelse(Pearson<=0,"0", ifelse(Pearson<=0.2,"0-0.2",
    ifelse(Pearson<=0.4, "0.2-0.4", ifelse(Pearson<=0.5, "0.4-0.5",

```

```

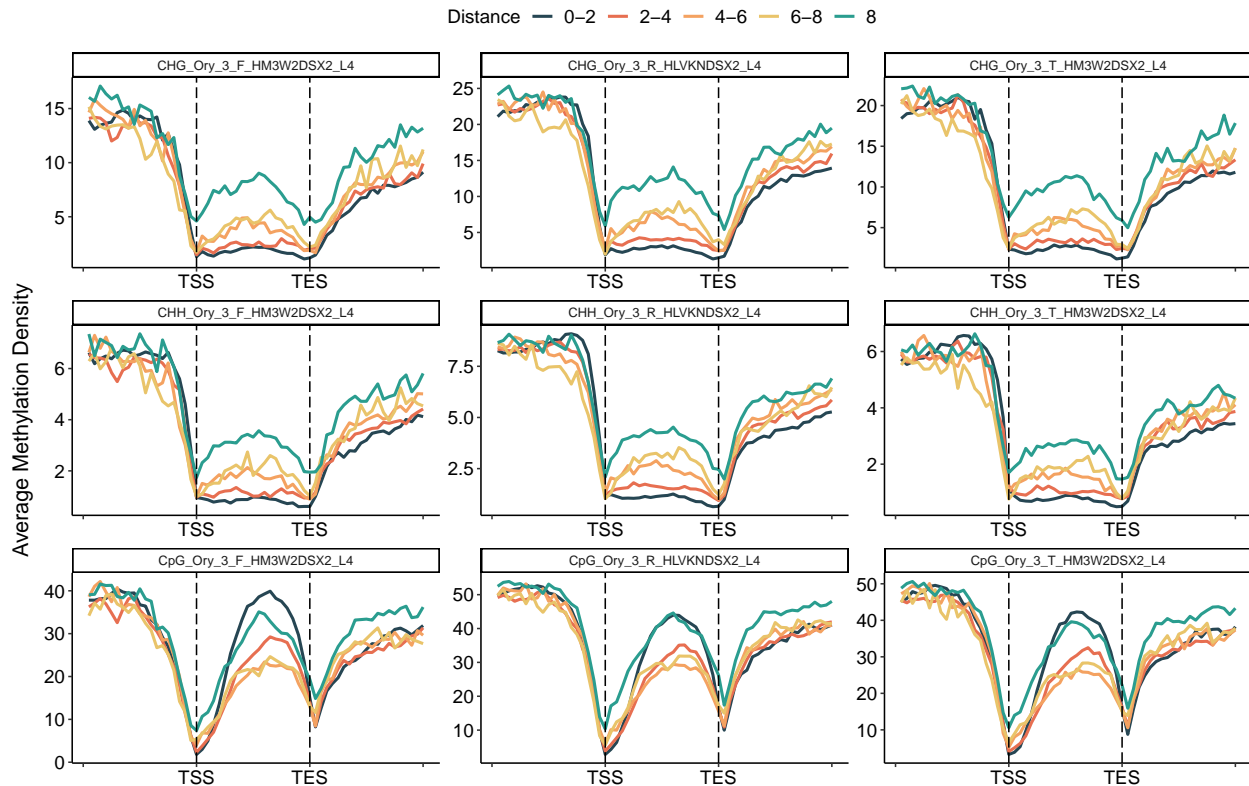
        ifelse(Pearson<=0.666, "0.5-0.666", "0.666") ) ) ) ))

methy_EI<-read.table("input-data/methy/EI.plot")

methy_EI$V4<-gsub("context_", "", methy_EI$V4 )
methy_EI$V4<-gsub(".flank.bin.meth", "", methy_EI$V4 )

methy_EI %>%
  ggplot(aes(x=V1,y=V2,color=V3))+
  geom_line(size=1) +
  facet_wrap(~ V4, scales = "free")+
  scale_fill_manual(values=c("0-2"= "#264653", "2-4"="#e76f51",
                             "4-6"="#f4a261", "6-8"="#e9c46a", "8"="#2a9d8f" ) )+
  scale_color_manual(values=c("0-2"= "#264653", "2-4"="#e76f51",
                             "4-6"="#f4a261", "6-8"="#e9c46a", "8"="#2a9d8f" ) )+
  theme_classic() +
  theme(legend.position = "top",
        axis.title.x=element_text(size=15,color="black",hjust=0.5, vjust = -1),
        axis.text.x=element_text(size=13,color="black"),
        axis.text.y=element_text(size=13,color="black"),
        legend.text=element_text(size=13,color="black"),
        legend.title=element_text(size=13,color="black"),
        plot.title=element_text(size=13,color="black",hjust=0.5)) +
  geom_vline(xintercept = 20, linetype = "longdash" ) +
  geom_vline(xintercept = 40, linetype = "longdash" ) +
  ylab("Average Methylation Density") +
  xlab("")+
  scale_x_continuous(labels = c("", "TSS", "TES", ""))+
  guides(color=guide_legend(title = "Distance"))

```



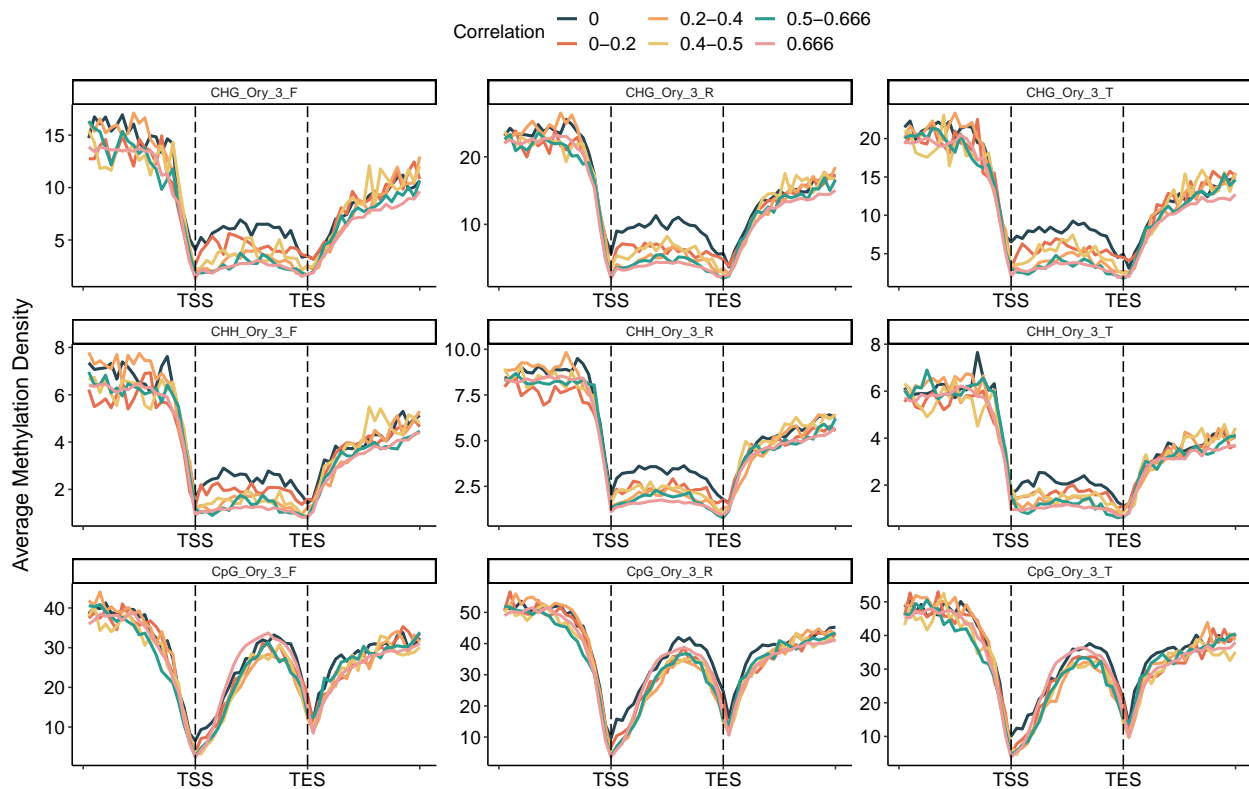
```
#ggsave("Methy_EI_density.pdf")
```

```
methy_PR<-read.table("input-data/methy/PR.plot")
```

```
methy_PR$V4<-gsub( "H.*", "", gsub("_context", "", methy_PR$V4 ))
```

```
methy_PR %>%
  ggplot(aes(x=V1,y=V2,color=V3))+
  geom_line(size=1) +
  facet_wrap(~ V4 , scales = "free") +
  scale_color_manual(values=c("0"= "#264653",
    "0-0.2"="#e76f51", "0.2-0.4"="#f4a261", "0.4-0.5"="#e9c46a",
    "0.5-0.666"="#2a9d8f", "0.666"="#ec9a9a" ) )+
  theme_classic() +
  theme(legend.position = "top",
    axis.title.x=element_text(size=15,color="black",hjust=0.5, vjust = -1),
    axis.text.x=element_text(size=13,color="black"),
    axis.text.y=element_text(size=13,color="black"),
    legend.text=element_text(size=13,color="black"),
    legend.title=element_text(size=13,color="black"),
    plot.title=element_text(size=13,color="black",hjust=0.5)) +
  geom_vline(xintercept = 20, linetype = "longdash" ) +
  geom_vline(xintercept = 40, linetype = "longdash" ) +
  ylab("Average Methylation Density") +
  xlab("")+
```

```
scale_x_continuous(labels = c("", "TSS", "TES", ""))+
guides(color=guide_legend(title = "Correlation"))
```



```
#ggsave("Methy_PR_density.pdf")
```

```
methy_EI<-read.table("input-data/methy/EI_intra.plot")
```

```
methy_EI$V4<-gsub("context_", "", methy_EI$V4 )
```

```
methy_EI$V4<-gsub(".flank.bin.meth", "", methy_EI$V4 )
```

```
methy_EI$V4<-gsub( "_H.*", "", gsub("_context", "", methy_EI$V4 ))
```

```
methy_EI<-
```

```
methy_EI %>%
```

```
mutate(Type=V5)
```

```
methy_EI %>%
```

```
ggplot(aes(x=V1,y=V2,color=V3,linetype=Type))+
```

```
geom_line(size=1) +
```

```
facet_wrap(~ V4, scales = "free")+
```

```
scale_fill_manual(values=c("0-2"= "#264653", "2-4"= "#e76f51",
```

```
"4-6"= "#f4a261", "6-8"= "#e9c46a", "8"= "#2a9d8f" ) )+
```

```
scale_color_manual(values=c("0-2"= "#264653", "2-4"= "#e76f51",
```

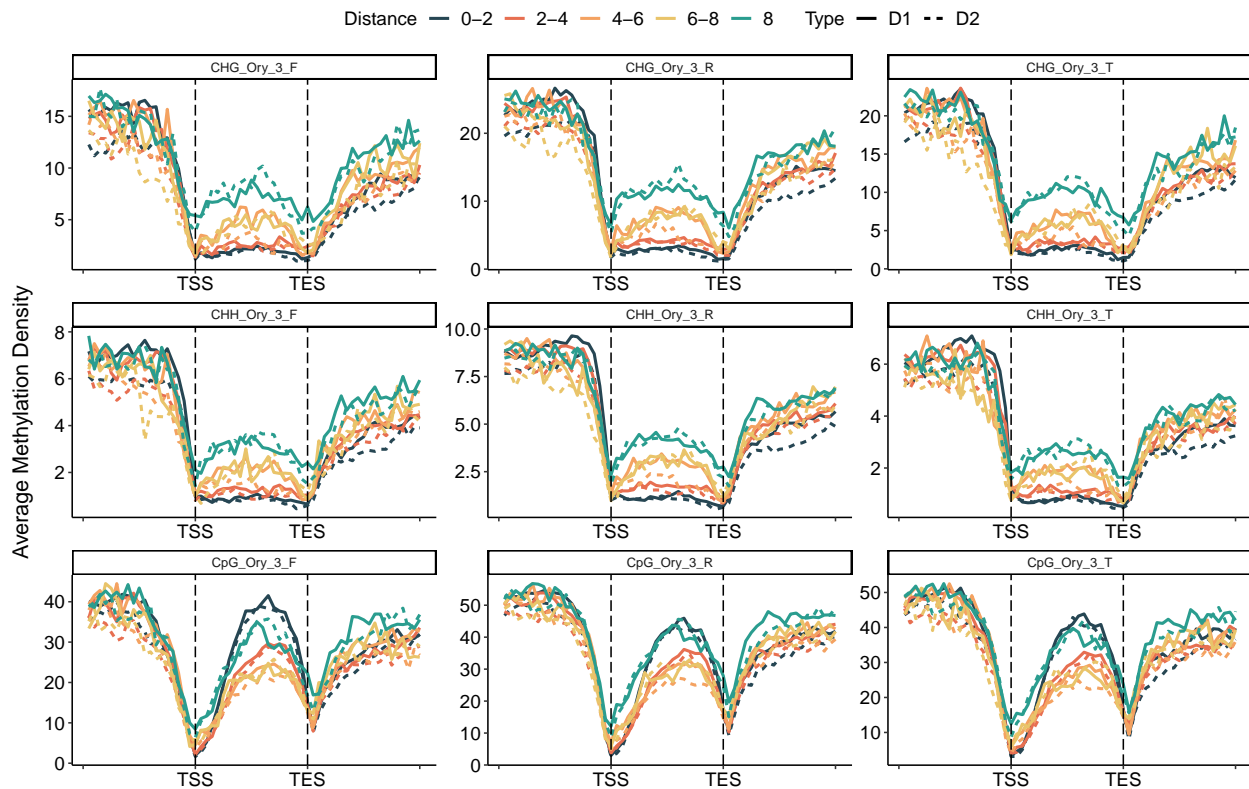
```
"4-6"= "#f4a261", "6-8"= "#e9c46a", "8"= "#2a9d8f" ) )+
```

```
theme_classic() +
```

```

theme(legend.position = "top",
axis.title.x=element_text(size=15,color="black",hjust=0.5, vjust = -1),
axis.text.x=element_text(size=13,color="black"),
axis.text.y=element_text(size=13,color="black"),
legend.text=element_text(size=13,color="black"),
legend.title=element_text(size=13,color="black"),
plot.title=element_text(size=13,color="black",hjust=0.5)) +
geom_vline(xintercept = 20, linetype = "longdash" ) +
geom_vline(xintercept = 40, linetype = "longdash" ) +
ylab("Average Methylation Density") +
xlab("")+
scale_x_continuous(labels = c("", "TSS", "TES", ""))+
guides(color=guide_legend(title = "Distance"))

```



```

#ggsave("Methy_EI_intra_density.pdf")

```

```

methy_PR<-read.table("input-data/methy/PR_intra.plot")

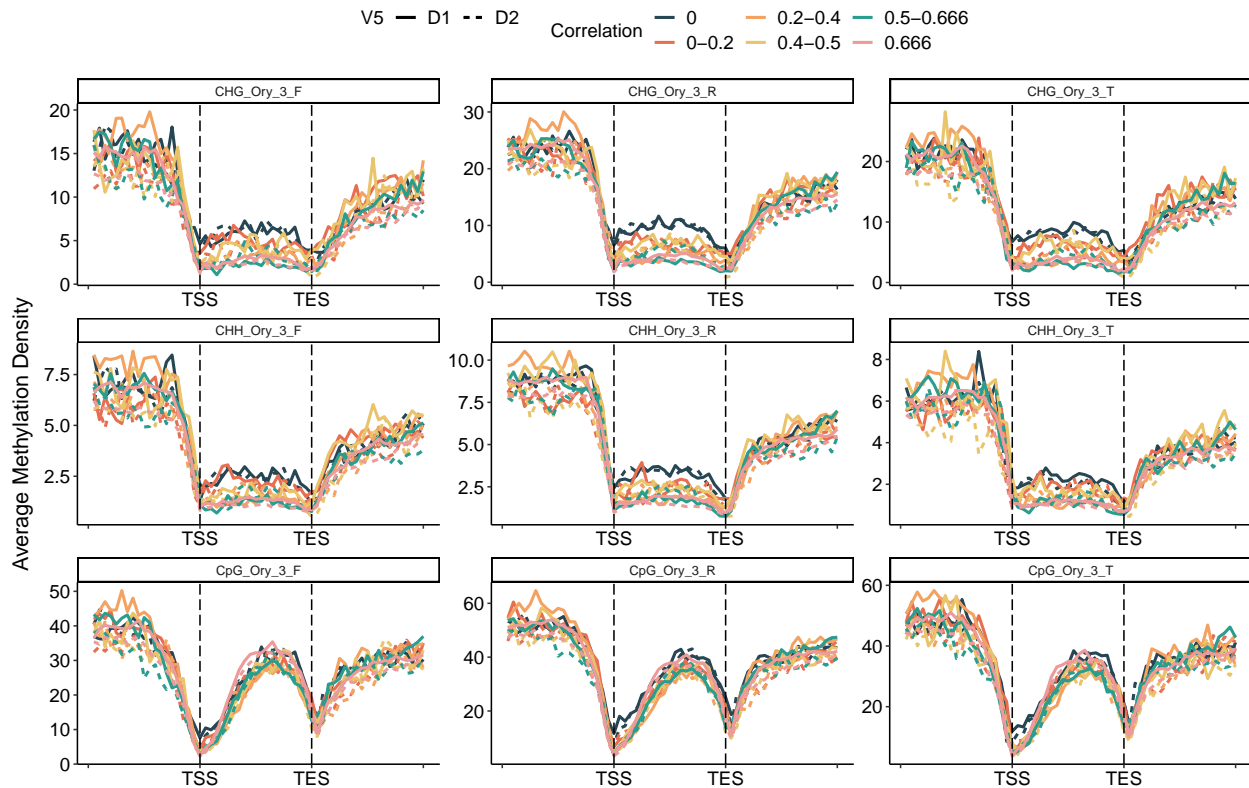
methy_PR$V4<-gsub( "H.*", "",      gsub("_context", "", methy_PR$V4 ))
methy_PR<- methy_PR %>%
  mutate(Type=V5)

```

```

methy_PR %>%
  ggplot(aes(x=V1,y=V2,color=V3,linetype=V5))+
  geom_line(size=1) +
  facet_wrap(~ V4 , scales = "free") +
  scale_color_manual(values=c("0"= "#264653", "0-0.2"="#e76f51",
    "0.2-0.4"="#f4a261", "0.4-0.5"="#e9c46a", "0.5-0.666"="#2a9d8f",
    "0.666"="#ec9a9a" ) )+
  theme_classic() +
  theme(legend.position = "top",
    axis.title.x=element_text(size=15,color="black",hjust=0.5, vjust = -1),
    axis.text.x=element_text(size=13,color="black"),
    axis.text.y=element_text(size=13,color="black"),
    legend.text=element_text(size=13,color="black"),
    legend.title=element_text(size=13,color="black"),
    plot.title=element_text(size=13,color="black",hjust=0.5)) +
  geom_vline(xintercept = 20, linetype = "longdash" ) +
  geom_vline(xintercept = 40, linetype = "longdash" ) +
  ylab("Average Methylation Density") +
  xlab("")+
  scale_x_continuous(labels = c("", "TSS", "TES", ""))+
  guides(color=guide_legend(title = "Correlation"))

```



```

blast<-read.table("input-data/ov.blast",header = F,sep = "\t")
blast$V1<- gsub("\\.t.*", "", blast$V1)
blast$V2<- gsub("\\.t.*", "", blast$V2)

```

#orth_tbl

```

orth_tbl<-orth_tbl %>%mutate(cDNA=0)

for (i in 1:nrow(orth_tbl)){
  #dup1 <- orth_tbl[i,"dup1"]
  #dup2 <- orth_tbl[i,"dup2"]
  L1 <- which(blast$V1 %in% orth_tbl[i,"dup2"])
  L2<- which(blast$V2 %in% orth_tbl[i,"dup1"])
  #L1[L1 %in% L2]
  if ( ( TRUE %in% (L1 %in% L2 ) ) ){
    orth_tbl[i,"cDNA"]<-blast[intersect(L1,L2 ),3]

  }else{
    L1 <- which(blast$V1 %in% orth_tbl[i,"dup1"])
    L2<- which(blast$V2 %in% orth_tbl[i,"dup2"])
    if ( ( TRUE %in% (L1 %in% L2 ) ) ){
      orth_tbl[i,"cDNA"]<-blast[intersect(L1,L2 ),3]

    }else{
      orth_tbl[i,"cDNA"]<-0
    }
  }
  #cc<-blast[intersect(L1,L2 ),3]
  #orth_tbl[i,"cDNA"]<-cc
}

blast<-
blast %>%
  mutate(media = paste(V1,V2,sep="_")) %>%
  select(media,V3)

orth_tbl<-
orth_tbl %>%
  mutate(media= paste(dup1,dup2,sep = "_"), media2= paste(dup2,dup1,sep = "_") ) %>%
  left_join(blast, by = c("media"="media")) %>%
  left_join(blast, by = c("media2"="media")) %>%
  mutate(cDNA_identity=ifelse( (!is.na(V3.x))&(!is.na(V3.y)),
    V3.x, ifelse( is.na(V3.x)&!is.na(V3.y),
    V3.y, ifelse( !is.na(V3.x)&is.na(V3.y), V3.x, 0 ) ) ) ) %>%
  select( -media,-media2,-V3.x,-V3.y )

cDNA_iden<-
orth_tbl %>%

```

```

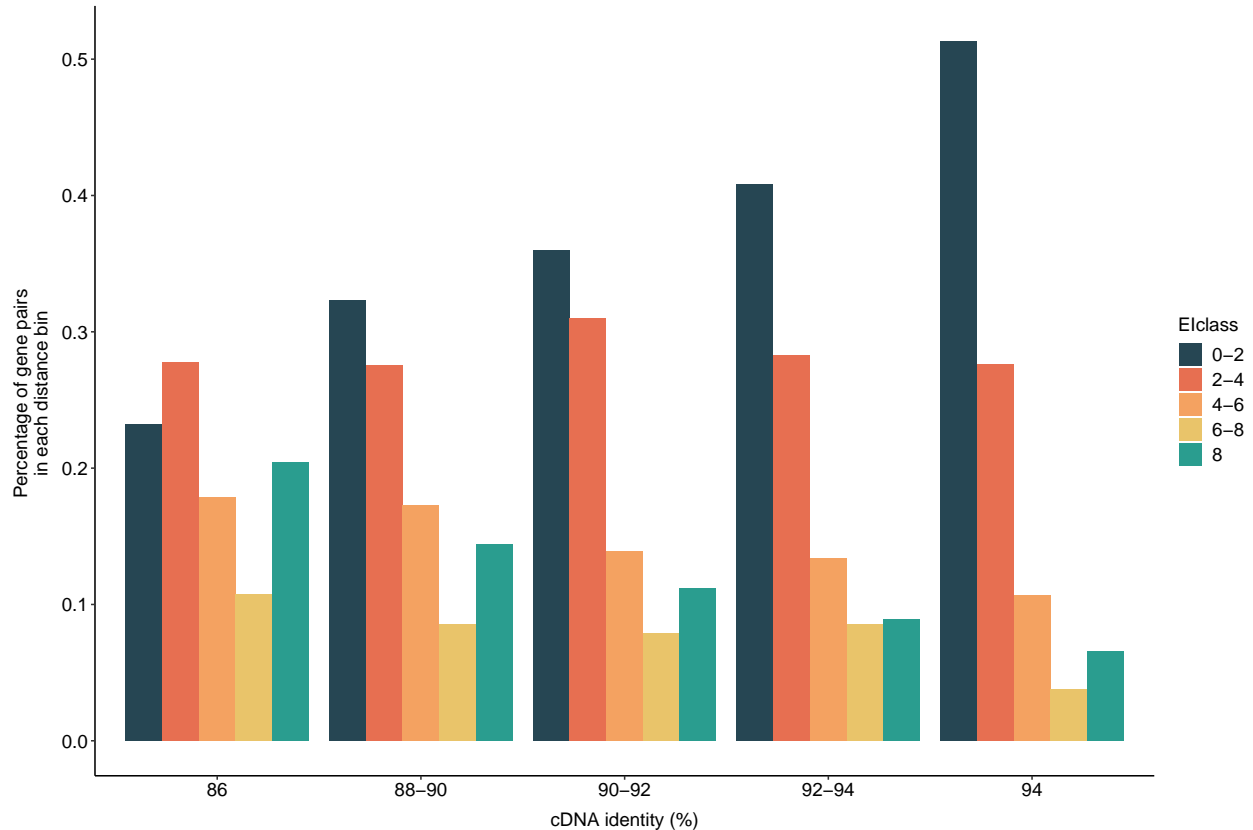
select(EI,cDNA_identity) %>%
mutate(EIclass=ifelse(EI<=2,"0-2", ifelse(EI<=4,"2-4" ,
      ifelse(EI<=6, "4-6", ifelse(EI<=8, "6-8", "8") ) ) ) ) %>%
mutate(cDNA_identityclass= ifelse(cDNA_identity>=94,"94",
      ifelse(cDNA_identity>=92,"92-94" , ifelse(cDNA_identity>=90, "90-92",
      ifelse(cDNA_identity>=88, "88-90", "86") ) ) ) ) %>%

group_by(EIclass,cDNA_identityclass) %>%
summarise(n=n() ) %>%
ungroup() %>%
group_by(cDNA_identityclass) %>%
summarise(all=sum(n))

orth_tbl %>%
  #select( -media,-media2,-V3.x,-V3.y ) %>%
select(EI,cDNA_identity) %>%
mutate(EIclass=ifelse(EI<=2,"0-2", ifelse(EI<=4,"2-4" ,
      ifelse(EI<=6, "4-6", ifelse(EI<=8, "6-8", "8") ) ) ) ) %>%
mutate(cDNA_identityclass= ifelse(cDNA_identity>=94,"94",
      ifelse(cDNA_identity>=92,"92-94" ,
      ifelse(cDNA_identity>=90, "90-92",
      ifelse(cDNA_identity>=88, "88-90", "86") ) ) ) ) %>%

group_by(EIclass,cDNA_identityclass) %>%
summarise(n=n()) %>%
ungroup() %>%
left_join(cDNA_iden) %>%
mutate(prop=n/all) %>%
ggplot(aes(x=cDNA_identityclass,y=prop, fill=EIclass))+
geom_col( position='dodge') +
theme_classic() +
theme(
axis.title.x=element_text(size=13,color="black",hjust=0.5, vjust = -1),
axis.text.x=element_text(size=13,color="black"),
axis.text.y=element_text(size=13,color="black"),
legend.text=element_text(size=13,color="black"),
legend.title=element_text(size=13,color="black"),
plot.title=element_text(size=13,color="black",face="italic",hjust=0.5)) +
ylab(paste("Percentage of gene pairs ", "in each distance bin ", sep="\n" ))+
xlab("cDNA identity (%)") +
scale_fill_manual(values=c("0-2"= "#264653", "2-4"="#e76f51",
      "4-6"="#f4a261", "6-8"="#e9c46a", "8"="#2a9d8f" ) )

```

```
#annotate("segment", x=0.7, xend=4.7, y=0.51, yend=0.16, size=1, color="#264653" ) +
#annotate("segment", x=0.95, xend=4.9, y=0.1, yend=0.18, size=1, color="#f4a261" ) +
#annotate("segment", x=1.12, xend=5.1, y=0.04, yend=0.125, size=1, color="#e9c46a" ) +
#annotate("segment", x=1.32, xend=5.3, y=0.06, yend=0.265, size=1, color="#2a9d8f" )
#ggsave("result/1.EI_cDNA.pdf", width = 12, height = 7)
```

```
cDNA_iden<-
orth_tbl %>%
  select(Pearson, cDNA_identity) %>%
  mutate(PearsonClass=ifelse(Pearson<=0, "0", ifelse(Pearson<=0.2, "0-0.2",
    ifelse(Pearson<=0.4, "0.2-0.4", ifelse(Pearson<=0.5, "0.4-0.5",
    ifelse(Pearson<=0.666, "0.5-0.666", "0.666") ) ) ) ) ) %>%
  mutate(cDNA_identityclass= ifelse(cDNA_identity>=94, "94",
    ifelse(cDNA_identity>=92, "92-94", ifelse(cDNA_identity>=90, "90-92",
    ifelse(cDNA_identity>=88, "88-90", "86") ) ) ) ) %>%

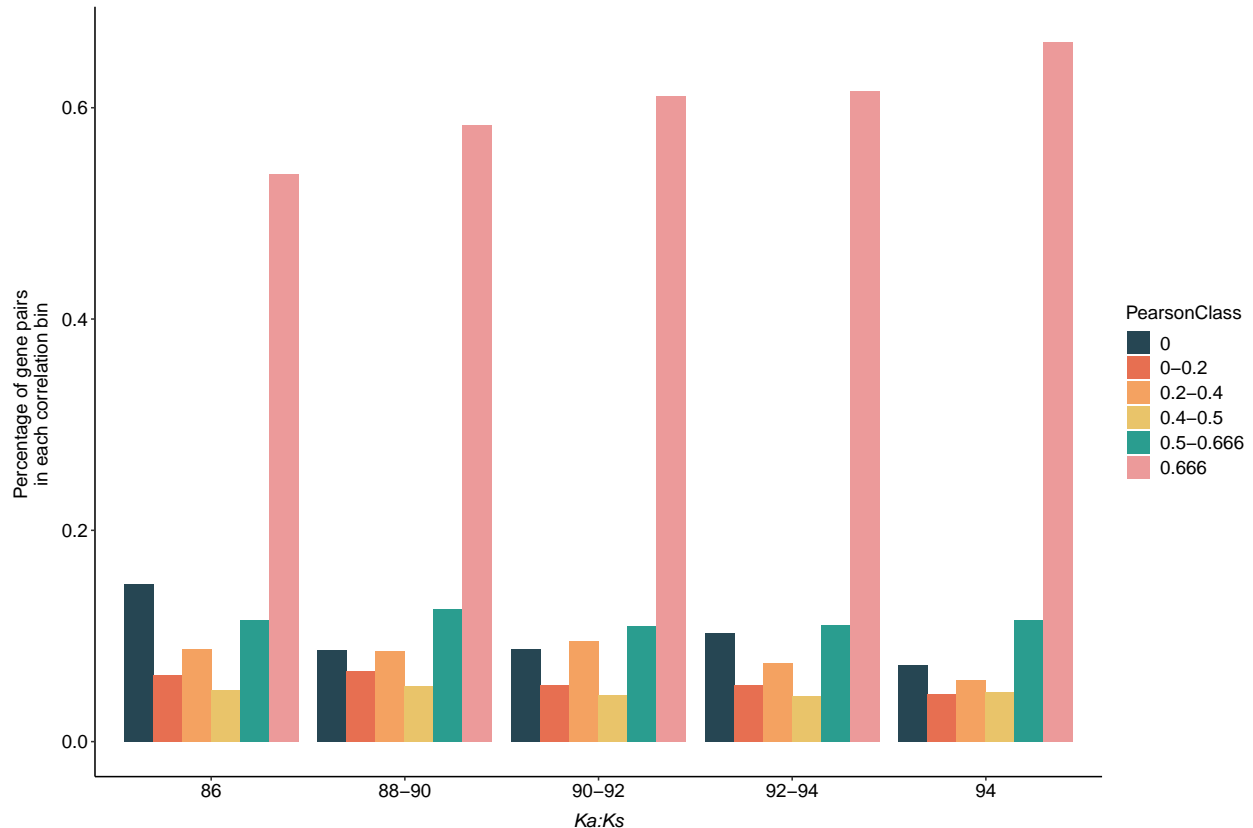
  group_by(PearsonClass, cDNA_identityclass) %>%
  summarise(n=n()) %>%
  ungroup() %>%
  group_by(cDNA_identityclass) %>%
  summarise(all=sum(n))

orth_tbl %>%
  select(Pearson, cDNA_identity) %>%
  mutate(PearsonClass=ifelse(Pearson<=0, "0", ifelse(Pearson<=0.2, "0-0.2",
```

```

    ifelse(Pearson<=0.4, "0.2-0.4", ifelse(Pearson<=0.5, "0.4-0.5",
      ifelse(Pearson<=0.666, "0.5-0.666", "0.666") ) ) ) ) %>%
mutate(cDNA_identityclass= ifelse(cDNA_identity>=94,"94",
  ifelse(cDNA_identity>=92,"92-94" ,
  ifelse(cDNA_identity>=90, "90-92",
    ifelse(cDNA_identity>=88, "88-90", "86") ) ) ) ) %>%
group_by(PearsonClass,cDNA_identityclass) %>%
summarise(n=n() ) %>%
ungroup() %>%
left_join(cDNA_iden)%>%
mutate(prop=n/all) %>%
ggplot(aes(x=cDNA_identityclass,y=prop, fill=PearsonClass))+
geom_col( position='dodge') +
theme_classic() +
theme(
axis.title.x=element_text(size=13,color="black",hjust=0.5, face="italic", vjust = -1),
axis.text.x=element_text(size=13,color="black"),
axis.text.y=element_text(size=13,color="black"),
legend.text=element_text(size=13,color="black"),
legend.title=element_text(size=13,color="black"),
plot.title=element_text(size=13,color="black",face="italic",hjust=0.5)) +
ylab(paste("Percentage of gene pairs ", "in each correlation bin ", sep="\n" ))+
xlab("Ka:Ks")+
scale_fill_manual(values=c("0= "#264653", "0-0.2"="#e76f51",
  "0.2-0.4"="#f4a261", "0.4-0.5"="#e9c46a", "0.5-0.666"="#2a9d8f",
  "0.666"="#ec9a9a" ) )

```



```
#ggsave("result/1.PR_cDNA.pdf",width = 12, height = 7)
```

```
ATAC<-list()
```

```
for (i in dir("input-data/ATAC")){  
  file<-paste("input-data/ATAC", i,sep="/")  
  ID<- gsub(".geneDistance.bed", "", i)  
  ATAC[[ID]]<-read.table(file, header = F, sep="\t")  
}
```

```
#rm(ATAC_plot)
```

```
ATAC_EI_plot<-list()
```

```
#names(ATAC)
```

```
for (i in names(ATAC)) {
```

```
  out<-paste("EI",i,sep="_")  
  ATAC_EI_plot[[out]]<-  
  orth_tbl %>%  
    #select( -media,-media2,-V3.x,-V3.y ) %>%  
    select(dup1,dup2,EI) %>%  
    mutate(EIclass=ifelse(EI<=2,"0-2", ifelse(EI<=4,"2-4",  
      ifelse(EI<=6, "4-6", ifelse(EI<=8, "6-8", "8") ) ) ) %>%  
    left_join( ATAC[[i]], by=c("dup1"="V9") ) %>%  
    select(dup1,dup2,EI, EIclass,V11) %>%  
    left_join( ATAC[[i]], by=c("dup2"="V9") ) %>%  
    select(EIclass,V11.x,V11.y) %>%  
    gather(key=EI,value=distance,2:3) %>%  
    na.omit() %>%  
    select(-EI) %>%  
    mutate(distance_kb=distance/1000) %>%  
    ggplot(aes(x=log10(distance_kb+0.0001),color=EIclass )) +  
    geom_density(size=2) +  
    geom_vline(xintercept = log10(2+0.0001), linetype = "longdash", color= "red")+  
    theme_bw() +  
    theme(#legend.position = c(.4, 0.8),  
      legend.position = "none",  
      axis.ticks = element_line(color = "black", linetype = "solid", size = 1),  
      axis.ticks.length.x=unit(0.2, "cm"),  
      panel.grid=element_blank(),  
      axis.title.x=element_text(size=20,color="black",hjust=0.5, vjust = -1),  
      axis.text.x=element_text(size=15,color="black"),  
      axis.text.y=element_text(size=15,color="black"),  
      #legend.text=element_text(size=20,color="black"),  
      #legend.title=element_blank(),  
      #legend.key.size = unit(1, "cm"),  
      plot.title=element_text(size=13,color="black",face="italic",hjust=0.5)) +  
    #ylab("Relative Density")+  
    ylab("")+  
    #xlab("Distance between ACR and nearest gene (kb)") +  
    xlab("")+  
    scale_color_manual(values=c("0-2"= "#264653", "2-4"="#e76f51",  
      "4-6"="#f4a261", "6-8"="#e9c46a", "8"="#2a9d8f" ) ) +  
    scale_x_continuous( breaks = c(-4,-1,0,1,2,3), limits = c(-4,3.1) ,
```

```

      labels = c("0", "0.1", "1", "10", "100", "1000")) +
      annotate("text", x=0.4, y= 0.02, hjust=0, color= "red", label="2kb", size=5 )

      #xlim(0,1000)
    }

  ATAC_pr_plot<-list()
  for (i in names(ATAC)) {

    out<-paste("Pearson", i, sep="_")
    ATAC_pr_plot[[out]]<-
    orth_tbl %>%
      select(dup1, dup2, Pearson) %>%
      mutate(PearsonClass=ifelse(Pearson<=0, "0", ifelse(Pearson<=0.2, "0-0.2",
        ifelse(Pearson<=0.4, "0.2-0.4", ifelse(Pearson<=0.5, "0.4-0.5",
          ifelse(Pearson<=0.666, "0.5-0.666", "0.666")))) %>%
      left_join( ATAC[[i]], by=c("dup1"="V9")) %>%
      select(dup1, dup2, Pearson, PearsonClass, V11) %>%
      left_join( ATAC[[i]], by=c("dup2"="V9")) %>%
      select(PearsonClass, V11.x, V11.y) %>%
      gather(key=Pearson, value=distance, 2:3) %>%
      na.omit() %>%
      select(-Pearson) %>%
      mutate(distance_kb=distance/1000) %>%
      ggplot(aes(x=log10(distance_kb+0.0001), color=PearsonClass )) +
      geom_density(size=2) +
      geom_vline(xintercept = log10(2+0.0001), linetype = "longdash", color= "red")+
      guides(colour = guide_legend(nrow = 1))+
      theme_bw() +
      theme(#legend.position = c(.4, 0.8),
        legend.position = "none",
        axis.ticks = element_line(color = "black", linetype = "solid", size = 1),
        axis.ticks.length.x=unit(0.2, "cm"),
        panel.grid=element_blank(),
        axis.title.x=element_text(size=20, color="black", hjust=0.5, vjust = -1),
        axis.text.x=element_text(size=15, color="black"),
        axis.text.y=element_text(size=15, color="black"),
        #legend.text=element_text(size=20, color="black"),
        #legend.title=element_blank(),
        #legend.key.size = unit(1, "cm"),
        plot.title=element_text(size=13, color="black", face="italic", hjust=0.5)) +
      #ylab("Relative Density")+
      ylab("")+
      #xlab("Distance between ACR and nearest gene (kb)") +
      xlab("")+
      scale_color_manual(values=c("0"= "#264653", "0-0.2"="#e76f51",
        "0.2-0.4"="#f4a261", "0.4-0.5"="#e9c46a", "0.5-0.666"="#2a9d8f",
        "0.666"="#ec9a9a" ))+
      #scale_color_manual(values=c("0-2"= "#264653", "2-4"="#e76f51", "4-6"="#f4a261", "6-8"="#e9c46a", "8-10"="#2a9d8f"))+
      scale_x_continuous( breaks = c(-4, -1, 0, 1, 2, 3), limits = c(-4, 3.1) ,
        labels = c("0", "0.1", "1", "10", "100", "1000")) +
      annotate("text", x=0.4, y= 0.02, hjust=0, color= "red", label="2kb", size=5 )
  }

```

```

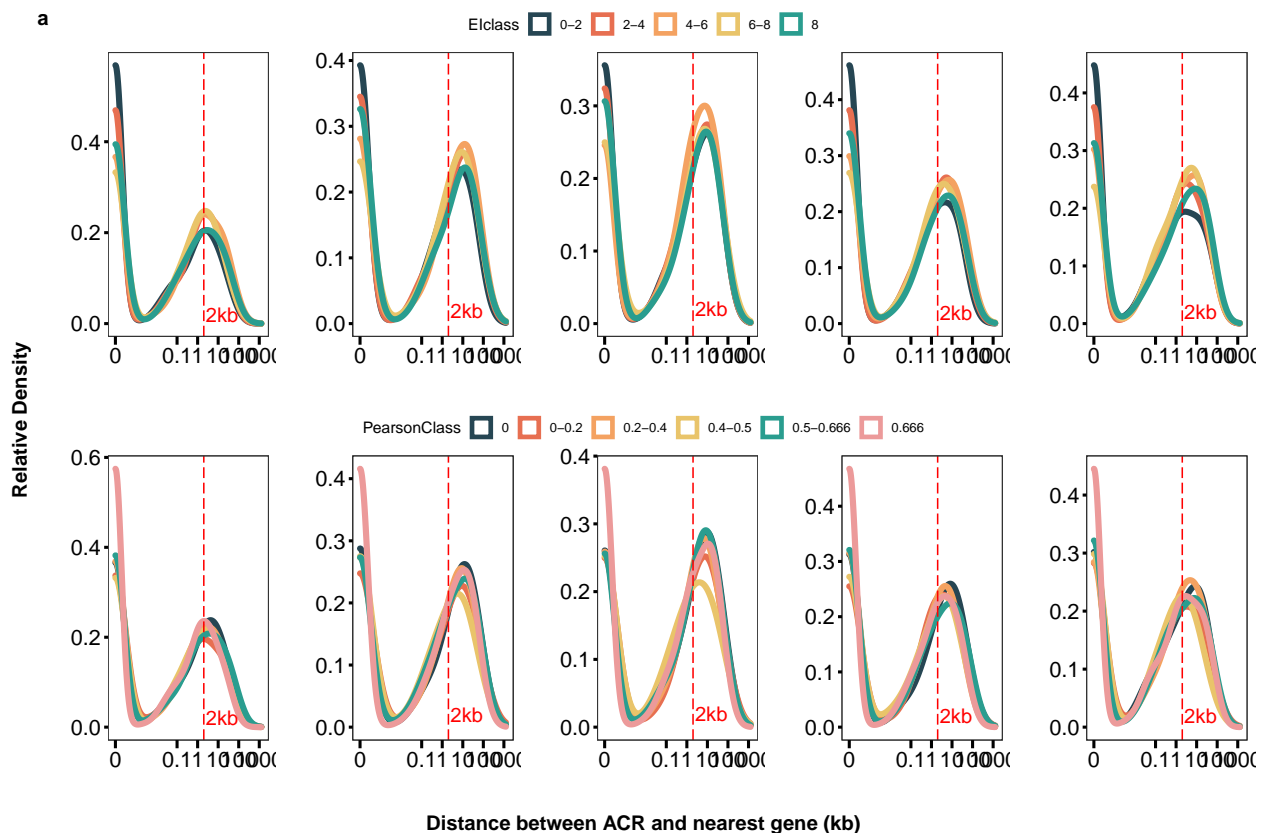
}

library(ggpubr)
#plt.legend(ncol=6)
p1<-ggarrange(
ggarrange(plotlist = ATAC_EI_plot,ncol = 5, nrow=1, heights = 2.5,widths = 10 , common.legend = T),
ggarrange(plotlist = ATAC_pr_plot,ncol = 5, nrow=1, heights = 2.5,widths = 10 , common.legend = T),
labels = "a", ncol = 1, nrow=2

)

annotate_figure(p1, bottom = text_grob("Distance between ACR and nearest gene (kb)",
color = "black", face = "bold", size = 14),
left= text_grob("Relative Density", color = "black", face = "bold",
size = 14, rot = 90) )

```



```

#ggsave("result/ATAC_distance.pdf",width = 25, height = 10)

rm(ATAC_EI_plot)
rm(ATAC_pr_plot)

#ATAC number do not have any sign for different levels divergent gene pairs in expression

do.call( rbind,

```

```

lapply(names(ATAC), function(x){
orth_tbl %>%
  #select( -media, -media2, -V3.x, -V3.y ) %>%
  select(dup1,dup2,EI) %>%
  mutate(EIclass=ifelse(EI<=2,"0-2", ifelse(EI<=4,"2-4",
    ifelse(EI<=6, "4-6", ifelse(EI<=8, "6-8", "8")) )) %>%
  left_join( ATAC[[x]], by=c("dup1"="V9") ) %>%
  select(dup1,dup2,EI, EIclass,V11) %>%
  left_join( ATAC[[x]], by=c("dup2"="V9") ) %>%
  select(dup1,dup2,EIclass,V11.x,V11.y) %>%
  gather(key=key, value=distance, 4:5) %>%
  na.omit() %>%
  select(-key) %>%
  group_by(dup1,dup2,EIclass) %>%
  count()
})
) %>%
  group_by(dup1,dup2,EIclass) %>%
  summarize( sum_count=sum(n) ) %>%
  filter(sum_count<40) %>%
  ggplot(aes(x=EIclass,y=sum_count, fill=EIclass))+
  geom_jitter(alpha=0.5, color="grey") +
  geom_boxplot()+
  #ylim(0,40)+
  stat_compare_means(method = 'wilcox.test',
    comparisons = list(c("0-2","2-4"), c("0-2","8")) )+
  theme_bw() +
  theme(#legend.position = c(.4, 0.8),
    legend.position = "none",
    axis.ticks = element_line(color = "black", linetype = "solid", size = 1),
    axis.ticks.length.x=unit(0.2, "cm"),
    panel.grid=element_blank(),
    axis.title.x=element_text(size=20,color="black",hjust=0.5, vjust = -1),
    axis.text.x=element_text(size=15,color="black"),
    axis.text.y=element_text(size=15,color="black"),
    #legend.text=element_text(size=20,color="black"),
    #legend.title=element_blank(),
    #legend.key.size = unit(1,"cm"),
    plot.title=element_text(size=13,color="black",face="italic",hjust=0.5)) +
  #ylab("Relative Density")+
  ylab("")+
  #xlab("Distance between ACR and nearest gene (kb)") +
  xlab("")+
  scale_fill_manual(values=c("0-2"= "#264653", "2-4"="#e76f51",
    "4-6"="#f4a261", "6-8"="#e9c46a", "8"="#2a9d8f" ) )

```

