Proj20160515

hy

5/15/2016

#load packages  
library("ggplot2", lib.loc="~/R/x86\_64-pc-linux-gnu-library/3.2")  
library("reshape2", lib.loc="~/R/x86\_64-pc-linux-gnu-library/3.2")  
load("/opt/rstudio/rstudio1/rstudio1/proj/refdata/refhg19.RData")

rep.row<-function(x,n){  
 matrix(rep(x,each=n),nrow=n)  
}  
rep.col<-function(x,n){  
 matrix(rep(x,each=n), ncol=n, byrow=TRUE)  
}  
#rep.row(c(1,2,3,4),5)  
#rep.col(c(1,2,3,4),5)  
  
overlap<- function(rbed,cbed,FUN=sum,lap=TRUE){  
 rlap<- rep(0,nrow(rbed))  
 clap<- rep(0,nrow(cbed))  
 rl<- list()  
 cl<- list()  
 for(chr in intersect(levels(rbed[,1]),levels(cbed[,1]))){  
 rind<- which(rbed[,1]==chr)  
 cind<- which(cbed[,1]==chr)  
 if(length(rind)>0 && length(cind)>0){  
 rmts<- rep.col(as.integer(rbed[rind,2]),length(cind))  
 rmte<- rep.col(as.integer(rbed[rind,3]),length(cind))  
 cmts<- rep.row(as.integer(cbed[cind,2]),length(rind))  
 cmte<- rep.row(as.integer(cbed[cind,3]),length(rind))  
 if(lap){  
 s<- rmts<=cmte  
 e<- rmte>=cmts  
 se<- s==e  
 }else{  
 s<- rmts<=cmts  
 e<- rmte>=cmte  
 se<- s==e  
 }  
 rlap[rind]<- apply(se, 1, FUN = FUN)  
 clap[cind]<- apply(se, 2, FUN = FUN)  
 rl[[chr]]<- apply(se, 1, FUN = FUN)  
 cl[[chr]]<- apply(se, 2, FUN = FUN)  
 }  
 }  
 return(list(rlap,clap,rl,cl))  
}  
  
overlap.target<- function(rbed,cbed,col=4,lap=TRUE){  
 rlap<- rep(NA,nrow(rbed))  
 rlap.list<- rep(NA,nrow(rbed))  
 for(chr in intersect(levels(rbed[,1]),levels(cbed[,1]))){  
 rind<- which(rbed[,1]==chr)  
 cind<- which(cbed[,1]==chr)  
 cname<- cbed[cind,col]  
 if(length(rind)>0 && length(cind)>0){  
 rmts<- rep.col(as.integer(rbed[rind,2]),length(cind))  
 rmte<- rep.col(as.integer(rbed[rind,3]),length(cind))  
 cmts<- rep.row(as.integer(cbed[cind,2]),length(rind))  
 cmte<- rep.row(as.integer(cbed[cind,3]),length(rind))  
 if(lap){  
 s<- rmts<=cmte  
 e<- rmte>=cmts  
 se<- s==e  
 }else{  
 s<- rmts<=cmts  
 e<- rmte>=cmte  
 se<- s==e  
 }  
   
 rlap[rind]<- apply(se, 1, function(x){as.character(cname[x>0][1])})  
 rlap.list[rind]<- apply(se, 1, function(x){as.character(cname[x>0])})  
  
 }  
 }  
 return(list(rlap,rlap.list))  
}

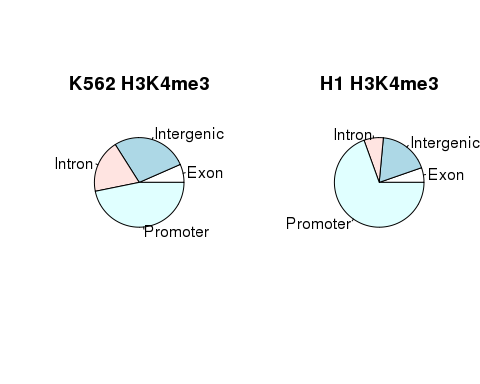
# Goal:This project is to study the non-promoter H3K4me3 in K562.

# Data discription

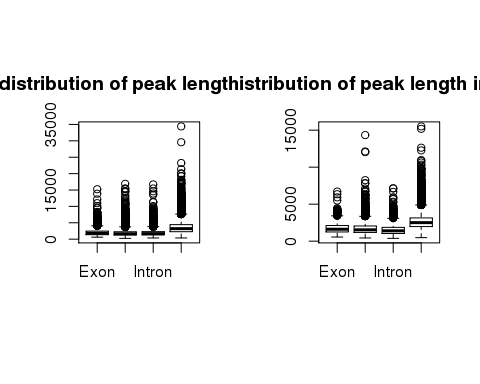
ChIP-Seq data(H3K4me3, H3K4me1, and H3K27ac) from ENCODE, run Macs 1.3.7(default) to get peaks RNA-Seq data(K562 and H1hesc)

## First, let's check the H3K4me3 peak distribution in K562 and H1 cell lines

k562.me3<- read.table("/mnt/local-disk1/rsgeno2/MAmotif/macs.20151224/H3K4me3/K562.H3k4me3.Rep2.macs.20151224\_peaks.xls",header = TRUE)  
h1.me3<- read.table("/mnt/local-disk1/rsgeno2/MAmotif/macs.20151224/H3K4me3/H1.H3k4me3.Rep2.macs.20151224\_peaks.xls",header = TRUE)  
Peaks.Distribution<- function(peaks,refhg19=refhg19){  
 peaks$state<- "Intergenic"  
 tp.promoter<- overlap(peaks,refhg19$PROMOTER)  
 peaks$state[tp.promoter[[1]]>0]<- "Promoter"  
 tp.exon<- overlap(peaks[peaks$state=="Intergenic",],refhg19$EXON)  
 peaks$state[peaks$state=="Intergenic"][tp.exon[[1]]>0]<- "Exon"  
 tp.intron<- overlap(peaks[peaks$state=="Intergenic",],refhg19$INTRON)  
 peaks$state[peaks$state=="Intergenic"][tp.intron[[1]]>0]<- "Intron"  
 return(peaks)  
}  
k562.me3<- Peaks.Distribution(k562.me3,refhg19)  
h1.me3<- Peaks.Distribution(h1.me3,refhg19)  
op <- par(mfrow = c(1, 2), pty = "s")  
pie(table(k562.me3$state),main="K562 H3K4me3")  
pie(table(h1.me3$state),main = "H1 H3K4me3")



par(op)  
op <- par(mfrow = c(1, 2), pty = "s")  
boxplot(k562.me3$end-k562.me3$start~k562.me3$state,main="the distribution of peak length in K562")  
boxplot(h1.me3$end - h1.me3$start ~ h1.me3$state, main="the distribution of peak length in H1hesc")



par(op)

In K562 cells, intergenic peaks(or non-promoter peaks) are higher than intergenic peaks in H1hesc. **1.next, we can check other normal and cancer cell lines to inverestage whether this is common phenomenon between them.** **2.we can also look into other histone modification(such as H3K27ac, H3K4me1, and H3K27me3) change from normal cell lines to cancer cell lines.** the length of H3K4me3 peaks in promoter are longer than other peaks.

# Questions

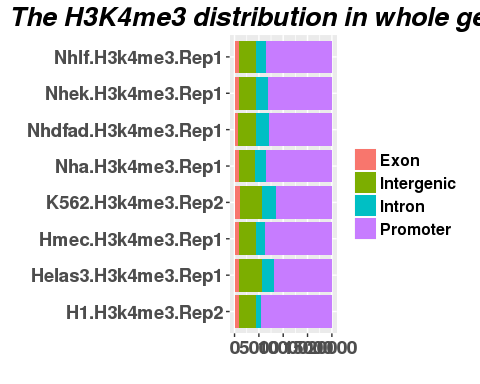
1. why are intergenic peaks in K562 cells higher than intergenic peaks in H1hesc? 1.1 Do cancer cell lines have more H3K4me3 histone modification in non-promoter region than normal? 1.2 what about the distribution of other histone modification between cancer cell lines and normal cell lines?
2. what is the relationship between H3K4me3 intergenic peaks and enhancer? 2.1 what is the relationship between H3K4me3 intergenic peaks and the expression of target gene?

Fisrt, Let's try to study questions 1.1 and 1.2 ##we check the H3K4me3 peak distribution in different cell lines preprocess h3k4me3 data, get top 20k peaks by decreaseing p-vale and each state in different cell lines

load("/opt/rstudio/rstudio1/rstudio1/proj/IronMan/data/h3k4me3.Rdata")  
Toppeaks<- function(x,top=20000){  
 pk<- x[order(-x[,7]),]  
 return(pk[1:min(top,nrow(x)),])  
}  
h3k4me3.top<- lapply(h3k4me3,Toppeaks)  
len.k4me3<-melt(lapply(h3k4me3.top,function(x)return(x$end - x$start)))  
state.k4me3<- melt(lapply(h3k4me3.top,function(x)x$state))  
len.state.h3k4me3<- data.frame(Len=len.k4me3$value,State=state.k4me3$value,Name=len.k4me3$L1)

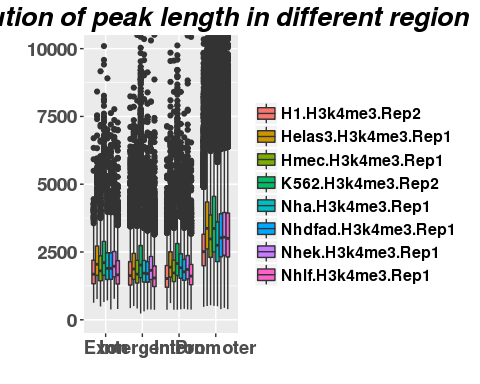
look into the H3K4me3 distribution in the whole genome to inverstigate the difference of H3K4me3 between normal and cancer cell lines

ggplot(len.state.h3k4me3)+geom\_bar(aes(x=Name,fill=State))+  
 labs(x = "",y = "",title="The H3K4me3 distribution in whole genome",fill="") +  
 coord\_flip()+  
 theme(plot.title = element\_text(color="black", size=20, face="bold.italic"),  
 axis.title.x = element\_text( face="bold",size=14),  
 #axis.title.y = element\_text(color="#993333", size=14, face="bold"),  
 legend.title =element\_text(face = "bold", size = 14, color = "black"),  
 legend.text = element\_text(face = "bold", size = 12),  
 axis.text.x = element\_text(face="bold",size=14),  
 axis.text.y = element\_text(face="bold", size=14)  
 )

 cancer cell lines, except Hmec, have more H3K4me3 in intergenic region than normal cell lines. the result of Hmec may be caused by the quality of the data. **I will check the other replicate.**

check the length distribution of H3K4me3 peaks in different cell lines

ggplot(len.state.h3k4me3)+geom\_boxplot(aes(x=State,y=Len,fill=Name))+  
 labs(x = "",y = "",title="The distribution of peak length in different region",fill="") +  
 theme(plot.title = element\_text(color="black", size=20, face="bold.italic"),  
 axis.title.x = element\_text( face="bold",size=14),  
 #axis.title.y = element\_text(color="#993333", size=14, face="bold"),  
 legend.title =element\_text(face = "bold", size = 14, color = "black"),  
 legend.text = element\_text(face = "bold", size = 12),  
 axis.text.x = element\_text(face="bold",size=14),  
 axis.text.y = element\_text(face="bold", size=14)  
 )+coord\_cartesian(ylim = c(0, 10000))

 the length of H3K4me3 peaks in promoter region are higher than the other.