Exercise 2. Functional annotation of ChIP-peaks

1. Using Putty (Windows) or Terminal (Mac) to connect to your assigned computer.

Create a directory /workdir/myUserID (replace myUserID with you BioHPC ID), copy the test regions, and test region sequences to the working directory.

```
mkdir  /workdir/myUserID

cd  /workdir/myUserID

cp  /shared_data/ChIP_seq_workshop2_2016/*.fas  ./

cp  /shared_data/ChIP_seq_workshop2_2016/ *.bed ./
```

2. Install Bioconductor packages and make a TxDb object from transcript annotations available on a BioMart database

```
R

source("http://bioconductor.org/biocLite.R")

biocLite("biomaRt")

biocLite("GenomicFeatures")

biocLite("ChIPseeker")

biocLite("org.At.tair.db")

biocLite("ChIPpeakAnno")

library("biomaRt")

library("GenomicFeatures")

library("ChIPseeker")

library("org.At.tair.db")

library("ChIPpeakAnno")

# head(listMarts(host = "www.ensembl.org"), 10)

listMarts(host="plants.ensembl.org")
```

```
listDatasets(useMart(biomart="plants_mart",host="plants.ensembl.org"))

arabidopsis <-
useDataset("athaliana_eg_gene",mart=useMart("plants_mart",host="plants.ensembl.org"))

transcriptsDb <-
makeTxDbFromBiomart(biomart="plants_mart",host="plants.ensembl.org" ,dataset="athaliana_eg_gene")

metadata(transcriptsDb)

saveDb(transcriptsDb,file="Arabidopsis.sqlite")

txdb<-loadDb("Arabidopsis.sqlite")

##############genomic feature distribution#######################

peak<- readPeakFile("test_results_peaks.narrowPeak.bed", as="GRanges")

aCR<-assignChromosomeRegion(peak, nucleotideLevel=FALSE,
precedence=c("Promoters", "immediateDownstream", "fiveUTRs", "threeUTRs",
"Exons", "Introns"), TxDb=txdb)

pdf(file="pie.pdf")

pie(aCR$percentage,labels =paste(names(aCR$percentage),
round(aCR$percentage,
digits=2),"%"),col=rainbow(length(aCR$percentage)),main="Genomic Feature
Distribution")

dev.off()

###############Distance to TSS regions################

tx_by_gn <- transcriptsBy(txdb, by="gene")

unlisted <- unlist(tx_by_gn)

TSS <- ifelse(strand(unlisted) == "+", start(unlisted), end(unlisted))

TSS <- GRanges(seqnames(unlisted), IRanges(TSS, width=1), strand(unlisted))

TSS_by_gn <- relist(TSS, tx_by_gn)

mcols(TSS) <- mcols(unlisted)
```

```
TSS_by_gn <- relist(TSS, tx_by_gn)
unlisted_TSS <- unlist(TSS_by_gn)
macs.anno <- annotatePeakInBatch(peak, AnnotationData=unlisted_TSS)
pdf(file="TSS_hist.pdf")
hist(macs.anno$distancetoFeature,xlab="Distance To TSS", main="",
xlim=c(-10000,10000),breaks=20,prob=T,col="red")
dev.off()
###############heatmap and average density distribution##########
promoter <- getPromoters(TxDb=txdb, upstream=3000, downstream=3000)
tagMatrix <- getTagMatrix(peak, weightCol=NULL, windows=promoter)
pdf(file="heat_average_plot.pdf")
tagHeatmap(tagMatrix, xlim=c(-3000, 3000), color="red")
plotAvgProf(tagMatrix, xlim=c(-3000, 3000), xlab="Genomic Region (5'->3')",
ylab = "Read Count Frequency")
dev.off()
```