ChIP-Seq data analysis workshop

Exercise 1. ChIP-Seq peak calling

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 fastq.gz and bam files to the working directory, then de-compress the file.

```
mkdir /workdir/myUserID

cd /workdir/myUserID

cp /shared_data/peak_annotation/ test_peak.txt ./
```

2. Install packages by Bioconductor

```
source("http://bioconductor.org/biocLite.R")
biocLite("RSQLite")
biocLite("biomaRt")
biocLite("ShortRead")
biocLite("GenomicFeatures")
biocLite("ChIPseeker")
biocLite("org.Ce.eg.db")
biocLite("clusterProfiler")
biocLite("rGADEM")
biocLite("ChIPpeakAnno")
biocLite("BSgenome.Celegans.UCSC.ce10")
library("rGADEM")
library("ChIPpeakAnno")
library("RSQLite");
library("biomaRt");
library("GenomicFeatures");
```

```
library("ChIPseeker");
library("org.Ce.eg.db");
library("clusterProfiler")
library("ChIPpeakAnno")
```

3. Peak annotation

```
##########peak distribution#########
peak<- readPeakFile("test_peak.txt", as="GRanges")
pdf(file="distribution.pdf")
plotChrCov(peak, weightCol=5, xlab = "Chromosome Size (bp)", ylab = "", title = "ChIP
Peaks over Chromosomes")
dev.off()
############# peak annotation#########

listMarts()
species_version<-useMart ("ensembl")
listDatasets(species_version)
transcriptsDb <- makeTranscriptDbFromBiomart(biomart="ensembl",
circ_seqs=NULL ,dataset="celegans_gene_ensembl")
metadata(transcriptsDb)
Txpts <- transcripts(transcriptsDb)
saveDb(transcriptsDb,file="Cel235DB.sqlite")
txdb<-loadDb('Cel235DB.sqlite')
peakAnno <- annotatePeak(peak, tssRegion=c(-3000,3000), TranscriptDb = txdb,
annoDb="org.Ce.eg.db")
write.table(as.data.frame(peakAnno),file="peak_annotation.txt",sep="\t")
############ peak features plotting###############
pdf(file="pie_plot.pdf")
plotAnnoPie(peakAnno)
dev.off()
pdf(file="bar_plot.pdf")
plotAnnoBar(peakAnno)
dev.off()
pdf (file="TSS_distance_distribution.pdf")
TSS_distance = peakAnno $distanceToTSS [!is.na(peakAnno $distanceToTSS)]
hist(TSS_distance, xlab = "Distance To Nearest TSS",prob=T, breaks = 20, xlim = c(-
10000,10000),col=rainbow(12))
dev.off()
promoter <- getPromoters(TranscriptDb=txdb, upstream=3000, downstream=3000)
tagMatrix <- getTagMatrix(peak, weightCol=NULL, windows=promoter)
pdf(file="tag_heatmap.pdf")
tagHeatmap(tagMatrix, xlim=c(-3000, 3000), color="red")
dev.off()
```

```
pdf(file="tag_average_density.pdf")
plotAvgProf(tagMatrix, xlim=c(-3000, 3000),xlab="Genomic Region (5->3)", ylab = "Read
Count Frequency")
dev.off()
####################GO term ####################
ensemblID<- as.list(org.Ce.egENSEMBL2EG)
ensemblID_subset<-ensemblID [unique(as.data.frame(peakAnno)$geneId)]
entrez_ID<-matrix(unlist(ensemblID_subset),ncol=1,byrow=TRUE)
gene<-as.vector(entrez_ID)
ego<-enrichGO(gene,organism ="worm", ont = "CC", pvalueCutoff = 0.01,pAdjustMethod =
"BH",qvalueCutoff = 0.05, minGSSize = 5,readable = TRUE)
write.table(summary(ego),file="ego_CC.txt",sep="\t")
```