Practical7: Downstream Analysis of ChIP-seq data

Shamith Samarajiiwa

September 2017

Contents

0.1 Peak Analysis

0.1.1 Set working directory

bash

```
cd /home/participant/Course_Materials/ChIPSeq/Materials/Practicals/

mkdir motifs

cd motifs
```

0.1.2 Extract region around peak summit

R

```
# read in the peak summit file

setwd("~/Course_Materials/ChIPSeq/Materials/Practicals/motifs")
peakfile <-read.table
("~/Course_Materials/ChIPSeq/Preprocessed/Peaks/TAp73beta_r2.fastq_trimmed.fastq_sorted_summits.bed")
# get a 500bp window around summit (-250 and +250)
Motif_Peaks <- data.frame((peakfile$V1), (peakfile$V2-250), (peakfile$V2+250), stringsAsFactors = FALSE)
# create bed file
write.table(Motif_Peaks, file= "Peaks_for_motif_detection.bed",
            row.names=FALSE,col.names = FALSE, sep="\t", quote =FALSE)

dir()
```

0.2 ChIP peak annotation

R

```
library("ChIPpeakAnno")
library("rtracklayer")

bed <- import("Peaks_for_motif_detection.bed")
```

```
gr1 <- GRanges(seqnames=bed$V1, ranges=IRanges(start=bed$V2, end=bed$V3))

#annotation

library("EnsDb.Hsapiens.v86")
#source("https://bioconductor.org/biocLite.R")
#biocLite("EnsDb.Hsapiens.v86")


library("EnsDb.Hsapiens.v86"") ##(hg38)
## create annotation file from EnsDb or TxDb
annoData <- toGRanges(EnsDb.Hsapiens.v86, feature="gene")
annoData[1:2]

anno.gr1 <- annotatePeakInBatch(gr1,
AnnotationData=annoData,
output="nearestBiDirectionalPromoters",
bindingRegion=c(-2000, 500))

#annotate with Entrez IDs

library("org.Hs.eg.db")
anno.gr1 <- addGeneIDs(anno.gr1,
                           "org.Hs.eg.db",
                           IDs2Add = "entrez_id")
head(anno.gr1)
```

0.3 Motif detection

MEME-ChIP performs several motif analysis steps on a set of user provided DNA sequences in fasta format. It is especially useful for analyzing peaks from ChIP-seq experiments. MEME-ChIP can 1. discover novel DNA-binding motifs (MEME and DREME) 2. analyze them for similarity to known binding motifs (TOMTOM) 3. visualize the arrangement of the predicted motif sites in your input sequences 4. identify motif enrichment relative to a background model

0.3.1 Get fasta for peak regions

bash

```
bedtools getfasta -fi ~/Course_Materials/Introduction/SS_DB/Preprocessed/Alignment/Bowtie/hg38_chr3.fa \
-bed Peaks_for_motif_detection.bed -fo Tp73_peaks.fa

# If strand information is known use -s option as well
```

0.3.2 Meme and Dreme

bash

```
meme
```

```
dreme
```

Meme is an EM algorithm and Dreme uses pattern matching to finds de novo motifs.

The fasta file should be 500bp for each peak, Meme looks at a 100bp window around the peak center for motifs. Repeat masking is recommended. * TOMTOM compares motifs discovered with known motifs from PWM databases. *Centrimo looks for motif enrichment.* MAST can visualize the location of the discovered motifs within a sequence. * AMA measures how strongly a motif is associated with each sequence * AME discovers subtly enriched known bindng motifs in the sequences. * FIMO detects motif locations

0.4 Motif Enrichment Analysis

Download meme motif PWM databases onto your working folder

PWM databases

bash

```
wget "http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.15.tgz"
chmod 755 motif_databases.12.15.tgz
#unzip them

tar -zxvf motif_databases.12.15.tgz


mkdir meme

ls
```

0.4.1 MemeChIP

bash

```
meme-chip -oc meme/ \
-db motif_databases/EUKARYOTE/jolma2013.meme \
-db motif_databases/JASPAR/JASPAR_CORE_2016_vertebrates.meme \
-meme-mod zoops -meme-minw 6 -order 1 \
-meme-maxw 30 -meme-nmotifs 10 -dreme-e 0.05 -centrimo-score 5.0 -centrimo-ethresh 10.0 Tp73_peaks.fa
```

or upload fasta file meme-chip

0.4.2 PScanChIP

Demonstration on web version of PScanChIP

PScanChIP

- Upload your bed file

0.5 GREAT for ontology enrichment

Genes co-regulated by TFs or associated with certain epigenomic modifications may belong to common functional groups. Gene Ontology enrichment analysis, GSEA, pathway and interaction analysis helps identify these enriched biological features and their connectivity.

STRING

GENEMANIA

Ontology enrichment analysis may be performed on the sets of genes with peaks associated to them. ChIPpeakAnno can map peaks to genes. In this example we will consider genes with peaks within 5000bp of a gene's TSS.

R

```
library("rGREAT")


# use LiftOver tool to conver your
# sequence to hg19 and create a bed file object in Granges.


job = submitGreatJob(bed,
rule = "oneClosest",adv_oneDistance = 5000,version = "3.0")
```

R

```
job

tb = getEnrichmentTables(job)
names(tb)
```

R

```
job

tb[[1]][1:2, ]
tb = getEnrichmentTables(job, ontology = c("GO Molecular Function", "BioCyc Pathway"))
tb = getEnrichmentTables(job, category = c("GO"))


availableOntologies(job, category = "Pathway Data")


par(mfrow = c(1, 3))
res = plotRegionGeneAssociationGraphs(job)

res[1:2, ]
```

- Use liftOver tool at UCSC convert your hg38 bed file to hg19 coordinates.
- Download the converted file and name it "hg19_Peaks_for_motif_detection.bed"
- upload this to web version of GREAT