Projreport20160708

hy

7/10/2016

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

# Data discription

There are ChIP-Seq data(H3K4me3, H3K4me1, and H3K27ac) of multiple cell lines from ENCODE,including K562 and H1hesc, There are 6 paired-end RNA-Seq data, including 2 K562 and 4 H1hesc.

# Preprocessing

run Macs 1.3.7(default) to get peaks for ChIP-Seq data use DESeq2 to normalize these sample's read count and calculate the log2-tranformed change of gene expression form K562 to H1hesc

# Why are there more non-promoter H3K4me3 peaks in K562 than in H1hesc?

In previous analysis, we have found the resluts, showing that cancer cell lines seem to have more H3K4me3 in non-promoter region than normal cell lines. next, we use K562 and H1hesc to explore the charaterisitic of these non-promoter peaks.

**There are three parts in analyzing K562 and H1hesc H3K4me3 peaks.**

First, H3K4me3 peaks classification and annotation classification: promoter,non-promoter, confusion How, balabala It has already been done.

Second, the character of these groups of H3K4me3 peaks

data for the analysis K562 and H1hesc

positive control: DNase, H3K4me3, H3K4me1, H3K27ac,,,active TF

negtive control: H3K27me3, PRC2

Third, D score Using GRO-seq data (the reads of minus strand - the reads of plus strand) / total reads in the region

# H3K4me3 peaks classification and annotation

According to the M value of H3K4me3 between K562 and H1hesc, peaks are classified into: B(big,M>1),M(middle,-1<=M<=1),S(small,M<-1). Then, peaks are classified into three groups (promoter,non-promoter,confusion) by genomic feature. Here, there are knowngene's promoters(78827) and the merged promoters(254406) defined by merging hg19\_Ensembl\_Genes, hg19\_GENCODE\_V14,knownGene, and refGene, together. Actually, the merged promoters highly overlap with the knowngene's promoters.

p1

in the figure, the K562-unique non-promoter H3K4me3 peaks(4437) have more than twice the K562-unique promoter H3K4me3 peaks(2020). The H1hesc-unique non-promoter H3K4me3 peaks(2011) are only about a half the number of H1hesc-unique promoter H3K4me3 peaks(4784). the result means that K562 have more H3K4me3 in non-promoter region than H1hesc.

we removed the confusion group of peaks(1312 k562-unique[B],648 common[M],846 H1hesc-unique[S]).

we investigated the relationship between the six groups of H3K4me3 and H3K4me1 peaks in K562 and H1hesc, respectively.

p2

K562 H3K4me1 peaks tend to co-occur in K562-unique H3K4me3 peaks. the result means that unique peaks are highly cell-type specific. there are quite many H1hesc H3K4me1 peaks overlap with K562-unique H3K4me3 peaks, indicating that there seems to be a shift from H3K4me1 to H3K4me3 in tumorigenesis.

p3333

we also investigated H3K27ac and got similar results.

K562-unique H3K4me3 peaks appear to be active, like enhancer, because they are highly co-occur with H3K4me1 and H3K27ac.

# The character of these groups of H3K4me3 peaks

next, we studied their character, such as enhancer, open chromatin, and eRNA expression, as positive control. we chose H3K27me3 as negtive control, which means H3K27me3 signal in K562-unique non-promoter peaks should be lower than the signal in H1hesc-unique non-promoter peaks.

these results confirmed that K562-unique non-promoter H3K4me3 peaks tend to enrich active signal, meaning these region probably perform some kinds of active function in K562. We will do more detail downstream analysis about that, such as GO analysis for their target genes.

# D score

the expression of most enhancer have no direction. we want to check the D score distribution in K562-unique non-promoter H3K4me3 peaks.