

Genomic Computing Evaluation

Assignment 2: GMQL

Fabrizio Frasca

April 6, 2017

1 Activator and Repressive epigenetic signals

Among the different epigenetic signals, H3K27ac and H3K27me3 are generally associated with active and repressed chromatin regions, respectively. Considering H3K4me1 in cell line A549 and the aforementioned signals (H3K27ac and H3K27me3) in broadPeak format, under ethanol treatment (EtOH):

1.1

Using GMQL, select the required ENCODE ChIP-seq data for the reference human genome hg19 (originally from the @UCSC website), i.e. the H3K4me1, H3K27ac and H3K27me3 under ethanol treatment (EtOH) in broadPeak format, and the promoter region annotation for the same reference human genome, by writing the required GMQL statements.

insert code here

Then, compute the following:

a. Active H3K4me1 regions (i.e. overlapping with H3K27ac regions)

insert code here

b. Repressed H3K4me1 regions (i.e. overlapping with H3K27me3 regions)

insert code here

c. Poised H3K4me1 regions (i.e. being simultaneously active and repressed)

insert code here

d. Active H3K4me1 regions in promoters

insert code here

e. For each region in (d) find the closest H3K4me1 region farther than 10 kb

insert code here

f. Store the last result (e)

insert code here

g. Run the created GMQL query and report: running time, obtained number of samples and number of regions in each sample. Can the result include replicated regions? Why?

insert code here

2 Differential binding

Consider the JUN Transcription Factor (TF) in cell line K562 in two different conditions: treatment with interferon alpha 30 minutes (IFNa30) and untreated. Considering the untreated sample as the baseline:

2.1

Using GMQL, select the required ENCODE ChIP-seq data for the reference human genome hg19 (originally from the @UCSC website), i.e. for JUN antibody_target with none, or IFNa30 treatment in narrowPeak format, and the RefSeq promoter region annotation for the same reference human genome, by writing the required GMQL statements. Combine replicas if needed.

insert code here

Then, compute the following:

a. DNA regions in common with the baseline and the treatment

insert code here

b. Considering the regions identified in (a), find the promoters in which they are present

insert code here

c. Store the last result (b)

insert code here