Predicting Chromatin Conformation From ChIPseq Transcription Factor

Ricky Lim[1], Samuel Collombet[2], Nicolas Bertin[1], Agus Salim[3], Touati Benoukraf[1]

[1]Cancer Science Institute of Singapore, National University of Singapore

[2]Institut de Biologie de l', Ecole Normale Superieur de Paris

[3]Department of Mathematics and Statistics, La Trobe University

benoukraf@nus.edu.sg

1 Introduction

The Idea:

ChIPseq could identify not only the direct binding sites of the chromatin protein of interest, but also the indirect ones. This owes to the existance of chromatin conformations or complexes. The formation of such complexes will cluster co-factors in close proximity.
Here, we are interested in the identification of not only the *direct* binding sites but also the *indirect* ones. With such knowledge, we could predict the chromatin conformation of the protein of interest.

Question:

- Could we cluster the direct and indirect bindings of chromatin proteins from ChIP-seq experiment using Mixture Models (MMs) ?
- Could we cluster the direct and indirect bindings of chromatin proteins from ChIP-seq experiment using Mixture Models (MMs), followed by local clusterings (distance proximity) ?

2 Pipeline Overview

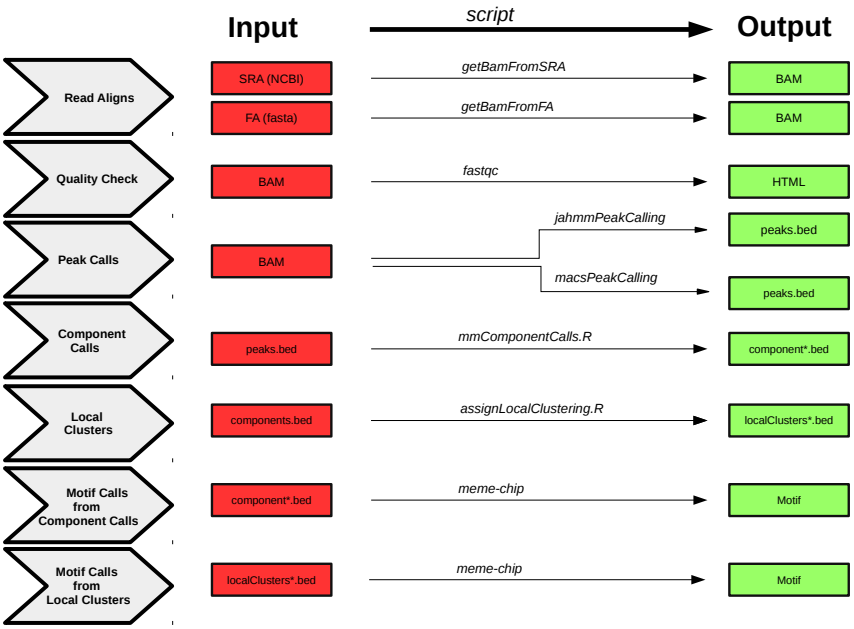

Figure 1: Pipeline for Chromatin Conformation Prediction

3 Implementation and Test

```
# export PATH for scripts required
PATH=$PATH:/home/ricky/Rlim/ChromatinConformation/Script;export PATH;
```

For more details information concerning the datasets please contact Nicolas Bertin for NF-κB Zhao dataset and Cebpα with Samuel Collombet

3.1 Read Aligns

NF-κB ChIPseq Transcription factor from Zhao, et al 2014

The read dataset was obtained and aligned using bowtie by Nicolas Bertin (Fullwood Lab). The dataset was in BAM format and sorted accordingly, as follows:

```
ls *.bam | parallel -j 3 'samtools sort {} {.}.sorted' &
```

Only the sorted BAM files were stored.

Cebpα ChIPseq in different cell lines from Samuel

The read dataset was obtained and aligned by Samuel. The dataset was downloaded from tblab-server at the following directory:

`|/DAS/TBlab/Samuel/raw/new/|`

The dataset was in BAM format and sorted accordingly, as follows:

```
ls *.bam | parallel -j 3 'samtools sort {} {.}.sorted' &
```

Only the sorted BAM files were stored.

Cebp ChIPseq in from Porse

The read dataset was obtained from NCBI with GEO GSE42321. The download link was stored in:
`/home/ricky/Rlim/ChromatinConformation/Input/SRA/Porse/sampleDownloadLinks.txt`
The dataset was stored in the following directory:
`/home/ricky/Rlim/ChromatinConformation/Input/ReadAligns/Porse`

```
Script/getBamFromSRA -c 2 -g /home/ricky/Rlim/Biotools/Genomes/mm10_bowtie2_index/bowtie2\
-a /home/ricky/Rlim/ChromatinConformation/Input/ReadAligns/Porse/annot.txt
-i /home/ricky/Rlim/ChromatinConformation/Input/ReadAligns/Porse\
-o /home/ricky/Rlim/ChromatinConformation/Output/ReadAligns/Porse 2> Log/getBamFromSRA_Porse.txt
```

Cebp ChIPseq in from Benner

The read dataset was obtained from NCBI with GEO GSM537984. The download link was stored in:
`/home/ricky/Rlim/ChromatinConformation/Input/FA/Benner/sampleDownloadLinks.txt`
The dataset was stored in the following directory:
`/home/ricky/Rlim/ChromatinConformation/Input/ReadAligns/Benner`

```
Script/getBamFromFA -c 2 -g /home/ricky/Rlim/Biotools/Genomes/mm10_bowtie2_index/bowtie2\
-a Input/ReadAligns/Benner/annot.txt -i Input/ReadAligns/Benner/
-o Output/ReadAligns/Benner/ 2> Log/getBamFromFA_Benner.txt
```

3.2 Quality Check

NF-κB ChIPseq Transcription factor from Zhao, et al 2014

```
ls Input/Bam/ZhaoB_etal.CellRep2014/*.bam |\
parallel -j 3 'fastqc -o Output/QC/ZhaoB_etal.CellRep2014/ {}' &\
```

Cebpα ChIPseq in different cell lines from Samuel

```
ls Input/Bam/Samuel/*.bam |\
parallel -j 2 'fastqc -o Output/QC/Samuel {}' &\
```

Cebp ChIPseq in from Porse

```
ls Input/QC/Porse/*.bam | parallel -j 2 'fastqc -o Output/QC/Porse/ {}'
```

Cebp ChIPseq in from Benner

```
ls Input/QC/Benner/*.bam | parallel -j 2 'fastqc -o Output/QC/Benner/ {}'
```

3.3 Peak Calls

We used jaHMM for peak callings and macs2 peak caller. jahmm library was downloaded and stored locally at
`/home/ricky/Rlim/ChromatinConformation/Library/jahmm`

NF-κB ChIPseq Transcription factor from Zhao, et al 2014

```
Script/jahmmPeakCalling.sh -c 4 -g hg19 -r 300\
-i Input/PeakCalls/ZhaoB_etal.CellRep2014/\
-o Output/PeakCalls/Jahmm/ZhaoB_etal.CellRep2014/\
2> Log/jahmmPeakCall_ZhaoB.txt
```

```
   8599068 300bin-ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_cREL_Rep1.sorted_peaks.bed
   9382822 300bin-ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_cREL_Rep2.sorted_peaks.bed
   6144560 300bin-ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p50_Rep1.sorted_peaks.bed
   5832975 300bin-ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p50_Rep2.sorted_peaks.bed
   8622964 300bin-ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p52_Rep1.sorted_peaks.bed
   8627622 300bin-ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p52_Rep2.sorted_peaks.bed
   9384884 300bin-ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p65_Rep1.sorted_peaks.bed
   9363618 300bin-ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p65_Rep2.sorted_peaks.bed
   9398329 300bin-ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_RELB_Rep0.sorted_peaks.bed
```

```
Script/macsPeakCalling.sh -c 4 -g hs\
-i Input/PeakCalls/ZhaoB_etal.CellRep2014/\
-o Output/PeakCalls/ZhaoB_etal.CellRep2014\
2> Log/macsPeakCall_Zhao.txt
```

Number of peaks called by macs2 from Zhao datasets in majority are in millions.

```
  1138227 ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_cREL_Rep1.sorted_summits_peaks.bed
  1045436 ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_cREL_Rep2.sorted_summits_peaks.bed
   445594 ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p50_Rep1.sorted_summits_peaks.bed
   465586 ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p50_Rep2.sorted_summits_peaks.bed
   984604 ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p52_Rep1.sorted_summits_peaks.bed
   385493 ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p52_Rep2.sorted_summits_peaks.bed
  3591979 ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p65_Rep1.sorted_summits_peaks.bed
  3406538 ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_p65_Rep2.sorted_summits_peaks.bed
  3706260 ZhaoB_etal.CellRep2014.ChIP-Seq_GM12878_RELB_Rep0.sorted_summits_peaks.bed
```

Number of peaks called by jahmm and macs from Zhao datasets are in millions. As there are many peaks being called, the subsequent pipelines were not run for this dataset.

Cebp from Samuel

Note that we used macs2 as the dataset from Samuel has not input and jahmm could call peaks without an input sample.

```
Script/macsPeakCalling.sh -c 4 -g hs\
-i Input/PeakCalls/Samuel/\
-o Output/PeakCalls/Macs/Samuel/\
2> Log/macsPeakCall_Samuel.txt
```

Cebp ChIPseq in from Porse

```
Script/jahmmPeakCalling.sh -c 2 -g mm10 -r 300 -i Input/PeakCalls/Porse/\
 -o Output/PeakCalls/Jahmm/Porse/ 2> Log/jahmmPeakCall_Porse.txt
```

Cebp ChIPseq in from Benner

```
Script/jahmmPeakCalling.sh -c 2 -g mm10 -r 300 -i Input/PeakCalls/Benner/\
-o Output/PeakCalls/Benner/ 2> Log/jahmmPeakCall_Benner.txt
```

3.4 Component Calls

Cebp from Samuel

```
Script/mmComponentCalls.R --model='GMM' --ncomp=3 --oneComp=TRUE\
```

```
--input_dir='/home/ricky/Rlim/ChromatinConformation/Input/ComponentCalls/Macs/Samuel/'\
--output_dir='/home/ricky/Rlim/ChromatinConformation/Output/ComponentCalls/Macs/Samuel/'\
2> Log/mmComponentCall_Samuel.txt\
```

Cebp from Porse

```
Rscript Script/mmComponentCalls.R --model='GMM' --ncomp=5 --oneComp=TRUE\
--input_dir='Input/ComponentCalls/Jahmm/Porse/'\
--output_dir='Output/ComponentCalls/Jahmm/Porse/' 2> Log/mmComponentCall_Porse.txt
```

Cebp from Benner

```
Script/mmComponentCalls.R --model='GMM' --ncomp=3 --oneComp=TRUE\
--input_dir='Input/ComponentCalls/Jahmm/Benner/'\
--output_dir='Output/ComponentCalls/Jahmm/Benner/' 2> Log/mmComponentCall_Benner.txt
```

3.5 Local Clusters: Direct and Indirect

Cebp from Samuel

```
Rscript Script/assignLocalClustering.R --distance=3000 --filter=TRUE\
--input_dir='/home/ricky/Rlim/ChromatinConformation/Input/LocalClusters/Macs/Samuel/'\
--output_dir='/home/ricky/Rlim/ChromatinConformation/Output/LocalClusters/Macs/Samuel/'\
2> Log/assignLocalCluster_Samuel.txt
```

Cebp from Porse

```
Script/assignLocalClustering.R --distance=3000 --filter=TRUE\
--input_dir='/home/ricky/Rlim/ChromatinConformation/Input/LocalClusters/Jahmm/Porse/'\
--output_dir='/home/ricky/Rlim/ChromatinConformation/Output/LocalClusters/Jahmm/Porse/'
```

Cebp from Benner

```
Script/assignLocalClustering.R --distance=3000 --filter=TRUE\
--input_dir='/home/ricky/Rlim/ChromatinConformation/Input/LocalClusters/Jahmm/Benner/'\
--output_dir='/home/ricky/Rlim/ChromatinConformation/Output/LocalClusters/Jahmm/Benner/'\
2> Log/assignLocalCluster_Benner.txt
```

3.6 Motif Calls from Component Calls

Cebp from Samuel

```
# extend bed peak summit from 1 bp resolution to 300 bp bin (150 bp in both strand directions)
# as samuel peaks were called using macs peak summit with 1 bp resolution
ls Input/MotifCalls/Macs/Samuel/ComponentCalls/*.bed |\
parallel -j 2 ``bedtools slop -i {}\
-g /home/ricky/Rlim/Biotools/Genomes/mm10/refGenome/mm10.genome\ -b 150 >{.}_150bp_extended.bed''

# bed to fasta
ls Input/MotifCalls/Macs/Samuel/ComponentCalls/*extended.bed |\
parallel -j 2 ``bedtools getfasta\
-fi /home/ricky/Rlim/Biotools/Genomes/mm10/refGenome/mm10.fa -bed {} -fo {.}.fa''

ls Input/MotifCalls/Macs/Samuel/ComponentCalls/*.fa |\
parallel -j 2 ``meme-chip\
-db /home/ricky/Rlim/Biotools/motif_databases/JASPAR_CORE_2014_vertebrates.meme\
```

```
-oc /home/ricky/Rlim/ChromatinConformation/Output/MotifCalls/Macs/Samuel/ComponentCalls/{/.}\
-index-name {/.} -meme-mod zoops -meme-minw 4 -meme-maxw 10 -meme-nmotifs 10 {}''
```

Cebp from Porse

```
# convert the bed to fasta
ls Input/MotifCalls/Jahmm/Porse/ComponentCalls/*.bed |\
parallel -j 2 "bedtools getfasta\
-fi /home/ricky/Rlim/Biotools/Genomes/mm10/refGenome/mm10.fa -bed {} -fo {.}.fa"

# meme-chip
ls Input/MotifCalls/Jahmm/Porse/ComponentCalls/*.fa |\
parallel -j 2 "meme-chip -db ~/Rlim/Biotools/motif_databases/JASPAR_CORE_2014_vertebrates.meme\
-oc Output/MotifCalls/Jahmm/Porse/ComponentCalls{/.}\
-index-name {/.} -meme-mod zoops -meme-minw 4 -meme-maxw 10 -meme-nmotifs 10 {}"
```

Cebp from Benner

```
# convert the bed to fasta
ls Input/MotifCalls/Jahmm/Benner/ComponentCalls/*.bed |\
parallel -j 2 "bedtools getfasta\
-fi /home/ricky/Rlim/Biotools/Genomes/mm10/refGenome/mm10.fa -bed {} -fo {.}.fa"

# meme-chip
ls Input/MotifCalls/Jahmm/Benner/ComponentCalls/*.fa |\
parallel -j 2 "meme-chip -db ~/Rlim/Biotools/motif_databases/JASPAR_CORE_2014_vertebrates.meme\
-oc Output/MotifCalls/Jahmm/Benner/ComponentCalls/{/.}\
-index-name {/.} -meme-mod zoops -meme-minw 4 -meme-maxw 10 -meme-nmotifs 10 {}"
```

3.7 Motif Calls from Local Clusters

Cebp from Samuel

```
# extend bed peak summit from 1 bp resolution to 300 bp bin (150 bp in both strand directions)
ls Input/MotifCalls/Macs/Samuel/LocalClusters/*.bed |\
parallel -j 2 ``bedtools slop -i {}\
-g /home/ricky/Rlim/Biotools/Genomes/mm10/refGenome/mm10.genome\ -b 150 >{.}_150bp_extended.bed''

# bed to fasta
ls Input/MotifCalls/Macs/Samuel/LocalClusters/*extended.bed |\
parallel -j 2 ``bedtools getfasta\
-fi /home/ricky/Rlim/Biotools/Genomes/mm10/refGenome/mm10.fa -bed {} -fo {.}.fa''

ls Input/MotifCalls/Macs/Samuel/LocalClusters/*.fa |\
parallel -j 2 ``meme-chip\
-db /home/ricky/Rlim/Biotools/motif_databases/JASPAR_CORE_2014_vertebrates.meme\
-oc /home/ricky/Rlim/ChromatinConformation/Output/MotifCalls/Macs/Samuel/LocalClusters/{/.}\
-index-name {/.} -meme-mod zoops -meme-minw 4 -meme-maxw 10 -meme-nmotifs 10 {}''
```

Cebp from Porse

```
# convert the bed to fasta
ls Input/MotifCalls/Jahmm/Porse/LocalClusters/*.bed |\
parallel -j 2 "bedtools getfasta\
-fi /home/ricky/Rlim/Biotools/Genomes/mm10/refGenome/mm10.fa -bed {} -fo {.}.fa"

# meme-chip
ls Input/MotifCalls/Jahmm/Porse/LocalClusters/*.fa |\
```

```
parallel -j 2 "meme-chip -db ~/Rlim/Biotools/motif_databases/JASPAR_CORE_2014_vertebrates.meme\
-oc Output/MotifCalls/Jahmm/Porse/LocalClusters{/.}\
-index-name {/.} -meme-mod zoops -meme-minw 4 -meme-maxw 10 -meme-nmotifs 10 {}"
```

Cebp from Benner

```
# convert the bed to fasta
ls Input/MotifCalls/Jahmm/Benner/LocalClusters/*.bed |\
parallel -j 2 "bedtools getfasta\
-fi /home/ricky/Rlim/Biotools/Genomes/mm10/refGenome/mm10.fa -bed {} -fo {.}.fa"

# meme-chip
ls Input/MotifCalls/Jahmm/Benner/LocalClusters/*.fa |\
parallel -j 2 "meme-chip -db ~/Rlim/Biotools/motif_databases/JASPAR_CORE_2014_vertebrates.meme\
-oc Output/MotifCalls/Jahmm/Benner/LocalClusters/{/.}\
-index-name {/.} -meme-mod zoops -meme-minw 4 -meme-maxw 10 -meme-nmotifs 10 {}"
```

```
  Filename: chromatinConformation.Rnw
  Working directory: /home/ricky/Rlim/ChromatinConformation
```

4 Metainfo

```
sessionInfo()

## R version 3.2.1 (2015-06-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.1 LTS
##
## locale:
##  [1] LC_CTYPE=en_SG.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_SG.UTF-8        LC_COLLATE=en_SG.UTF-8
##  [5] LC_MONETARY=en_SG.UTF-8    LC_MESSAGES=en_SG.UTF-8
##  [7] LC_PAPER=en_SG.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_SG.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] knitr_1.10.5
##
## loaded via a namespace (and not attached):
## [1] magrittr_1.5  formatR_1.2   tools_3.2.1   stringi_0.5-2 highr_0.5
## [6] digest_0.6.8  stringr_1.0.0 evaluate_0.7
```