Predicting the Chromatin Conformation From ChIPseq Dataset: part 2 (Motif Discovery Analysis)

Ricky Lim[1], Samuel Collombet[2], 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

benoukraf@nus.edu.sg

```
Filename: motifAnalysis.Rnw
Working directory: /home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/MotifAnalysis
```

```
source('/home/ricky/Rlim/ChromatinConformation/MotifCalls/MotifCalls.R')
work_dir= '/home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/'
```

1 Introduction

Goal: to infer the motif bindings of transcription factors and co-factors from ChIPseq experiment.

2 Local Clustering: Direct and Indirect Bindings

Here, we applied local clustering following the component calls by mixture models. The distance of 3Kb, 8Kb, 20Kb, and 50Kb was used for this clustering.

With local clustering, we found that most of the clusters are single peaks. Increasing the clustering distance from 3Kb to 50Kb gives rise to the decreasing coverage percentage of clusters with single peaks. Although the majority of the clusters remains single peaks (figure 1).

As we are interested in the chromatin conformation consisting of direct transcription factors and cofactors, we excluded the clusters with only single peaks.

2.1 Local Clustering: 4 Components-Mixture Negative Binomial Models

```
## Saving 7 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
```

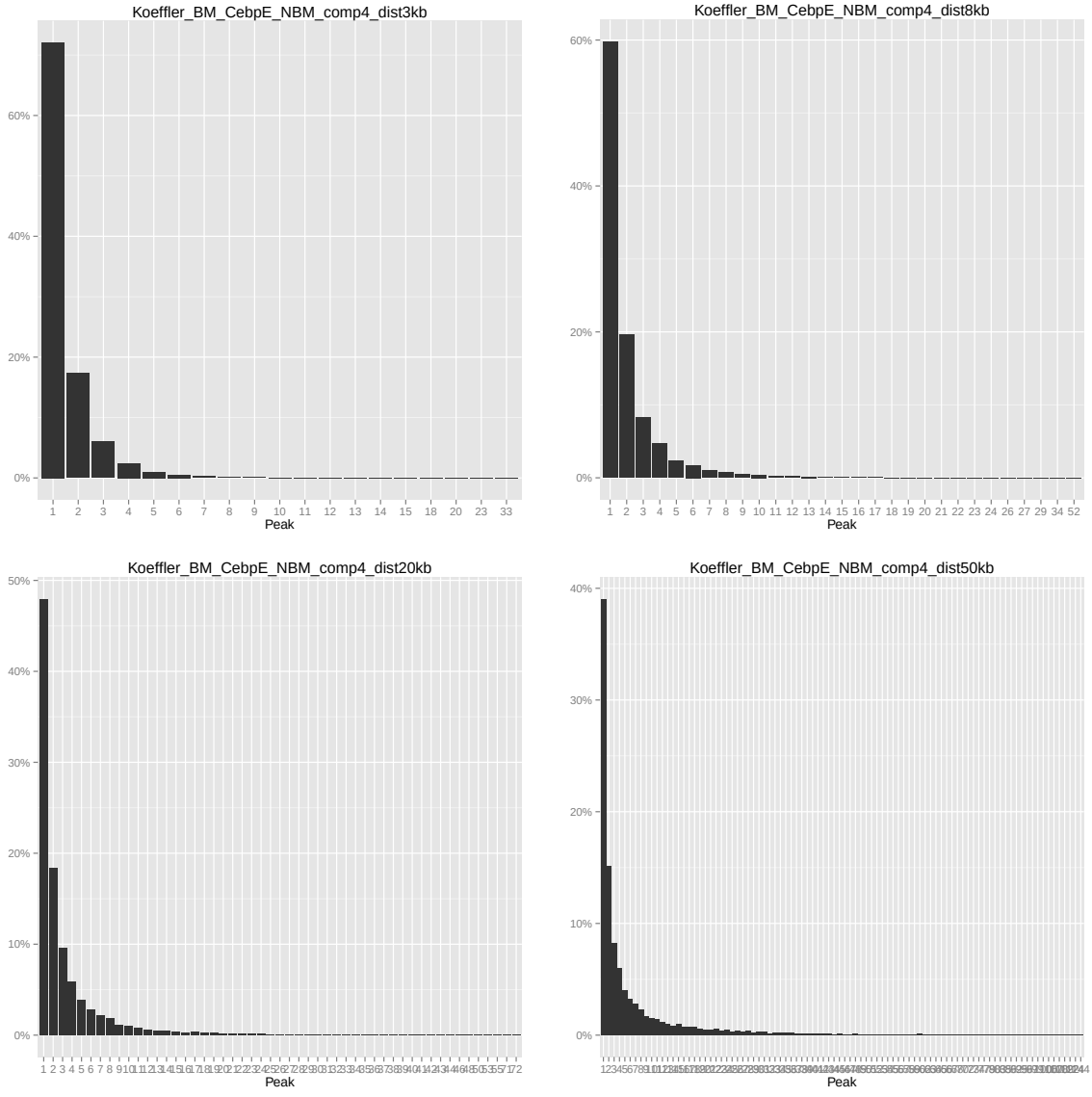


Figure 1: The Number of Peaks (percentage) in Clusters

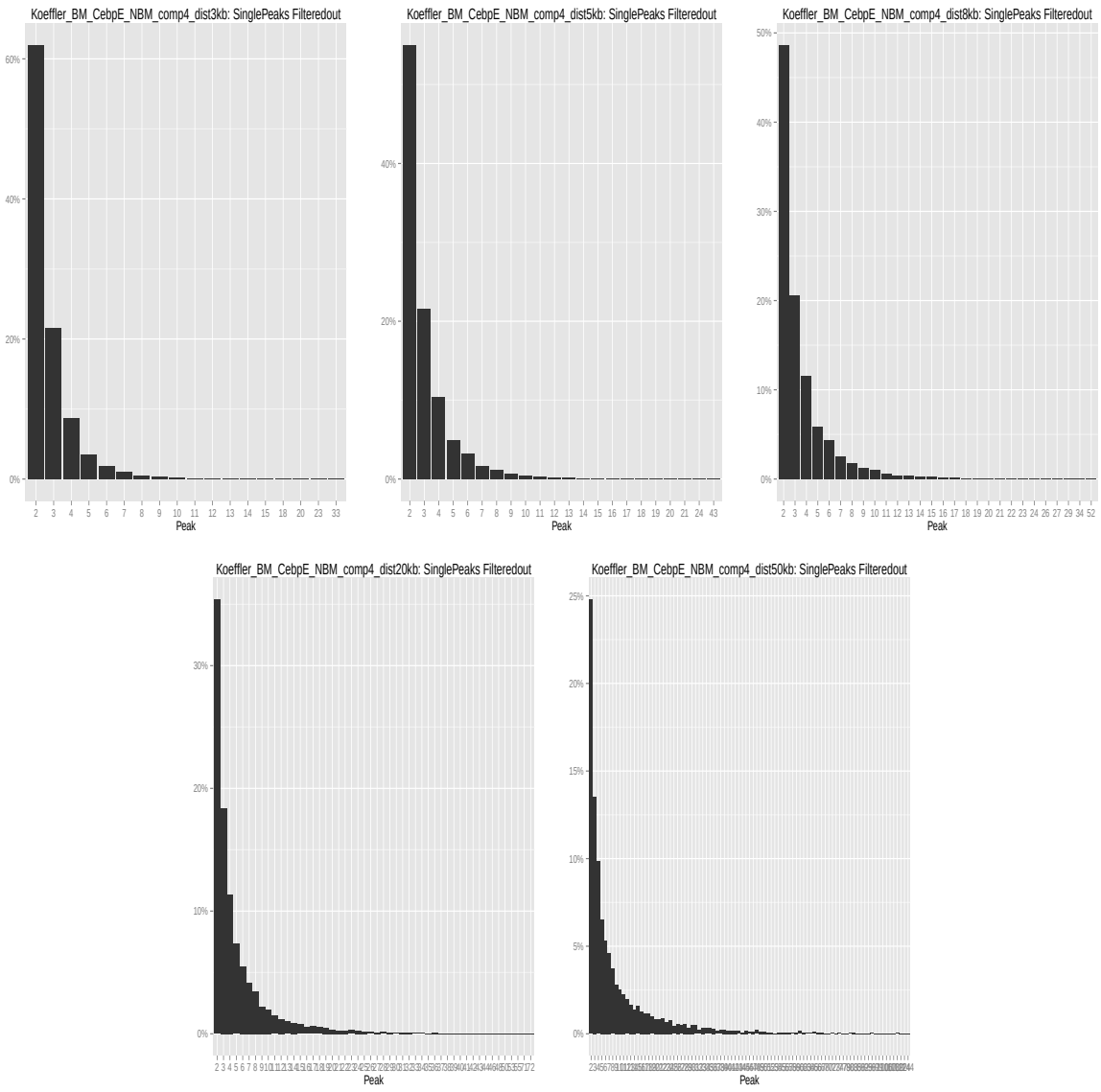


Figure 2: The Number of Peaks (percentage) in Clusters without Single Peaks

	Dist3kb	Dist5kb	Dist8kb	Dist20kb	Dist50kb
Direct	25572	23455	21315	16320	11604
Indirect	12933	16763	20655	29183	36760
Noise	12972	11259	9507	5974	3113
Total	51477	51477	51477	51477	51477

Table 1: Clustering Peaks in Different Cluster Distance

	Dist3kb	Dist5kb	Dist8kb	Dist20kb	Dist50kb
Direct	11542	12780	13420	12829	10440
Indirect	12933	16763	20655	29183	36760
Noise	1933	2021	2015	1677	1023
Total	26408	31564	36090	43689	48223

Table 2: Clustering Peaks in Different Cluster Distance: Single Peaks Filteredout

2.2 Local Clustering: Log-Transformed 4 Components-Gaussian Mixture Models

```
## Saving 7 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
```

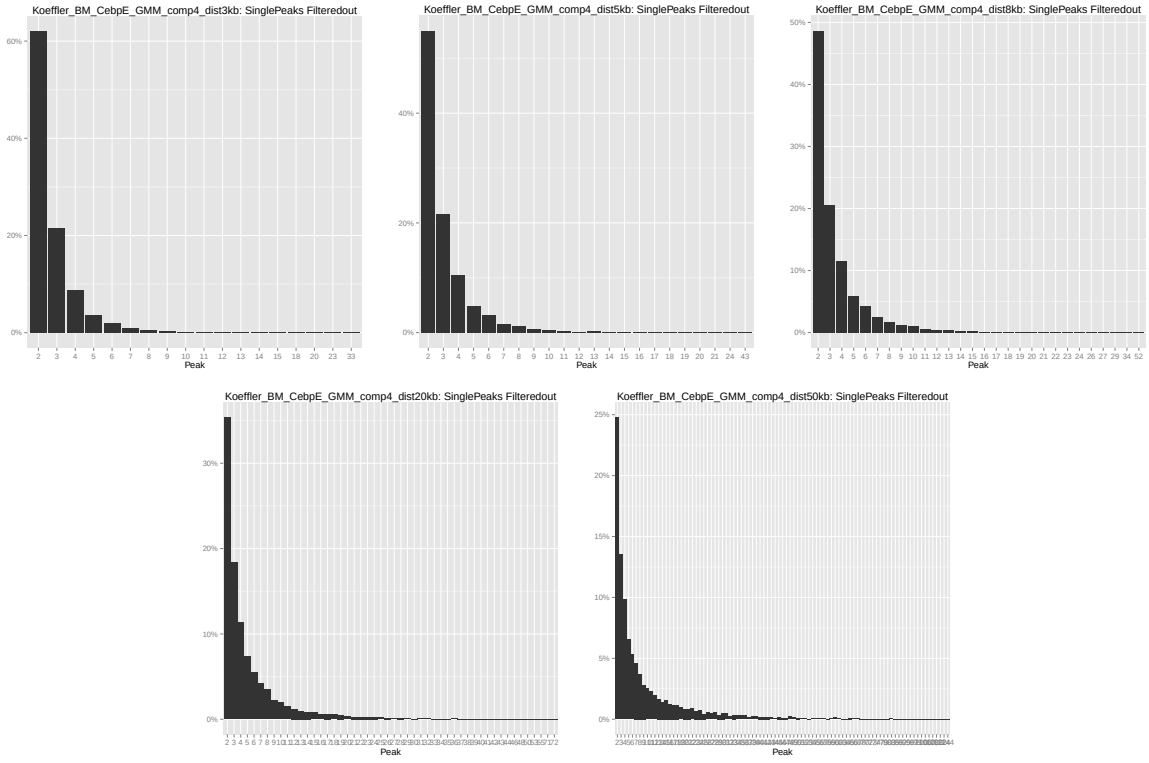


Figure 3: The Number of Peaks (percentage) in Clusters without Single Peaks: 4 Components-GMM

	Dist3kb	Dist5kb	Dist8kb	Dist20kb	Dist50kb
Direct	13609	15080	15712	14733	11150
Indirect	12235	15952	19828	28507	36740
Noise	564	532	550	449	333
Total	26408	31564	36090	43689	48223

Table 3: Clustering Peaks in Different Cluster Distance: 4 Comp-GMM

2.3 Local Clustering: Log-Transformed 5 Components-Gaussian Mixture Models

```
## Saving 7 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
## Saving 6.99 x 7 in image
```

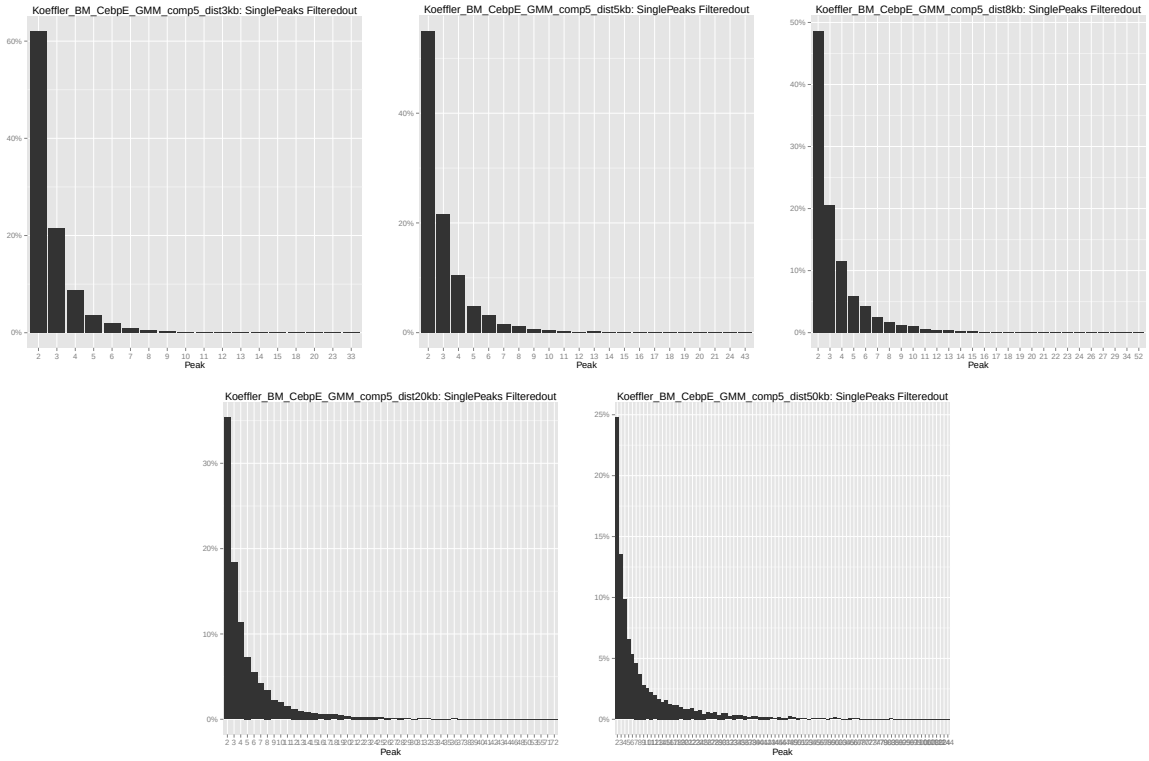


Figure 4: The Number of Peaks (percentage) in Clusters without Single Peaks: 5 Components-GMM

	Dist3kb	Dist5kb	Dist8kb	Dist20kb	Dist50kb
Direct	13044	14411	14920	13680	9909
Indirect	13067	16877	20896	29788	38141
Noise	297	276	274	221	173
Total	26408	31564	36090	43689	48223

Table 4: Clustering Peaks in Different Cluster Distance: 5 Comp-GMM

3 Database-based Motif Discovery: Centdist

```
# awk -F'\t' '{print $1"\t"$2-500"\t"$3+500"\t"$5}'
# Input/Koeffler_BM_CebpE_NBM_ModelAssignment_compSorted4.bed
# > Output/Koeffler_BM_CebpE_Peaks_1kb.bed
```

We run CentDist to search for the occurrances of database motifs (JASPAR and TRANSFAC) CentDist on the ChIPseq Peaks indirect and direct clusters. In all cluster distance, however the CEBP motifs were found in both direct and indirect clusters.

```
work_dir_centdist = '/home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/'
Koeffler_BM_ChIPseq_CebpE_Bicluster_compSorted_dist50kb_direct <-
    read.delim(paste0(work_dir,'Output/CentDist/',
    'Koeffler_BM_CebpE_NBM_BiclusterAssignment_SinglePeakFilteredOut_compSorted4_dist50kb_direct/',
    'result.txt'))
Koeffler_BM_ChIPseq_CebpE_Bicluster_compSorted_dist50kb_indirect <-
    read.delim(paste0(work_dir,'Output/CentDist/',
    'Koeffler_BM_CebpE_NBM_BiclusterAssignment_SinglePeakFilteredOut_compSorted4_dist50kb_indirect/',
    'result.txt'))
```

4 Centdist Motif Results for Direct and Indirect Clusters within 50kb: CebpE

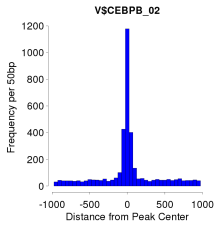
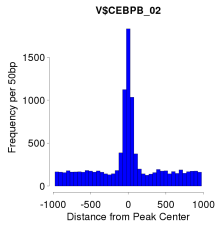
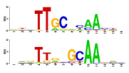
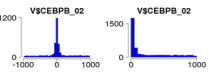
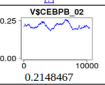
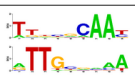
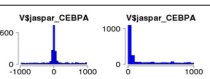
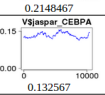
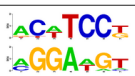
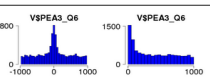
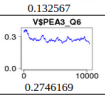
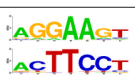
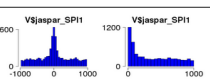
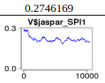
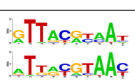
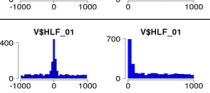
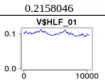
Direct		Indirect	
Info	Dist	Info	Dist
• Name: V-CEBPB-02 • Binding.Range: 80 • P.value: 0 • Rank: 1		• Name: V-CEBPB-02 • Binding.Range: 120 • P.value: 0 • Rank: 1	

Table 5: CebpE ChIPseq in BM: CentDist Motif Discovery

Results for Koeffler_BM_CebpE_NBM_BiclusterAssignment_SinglePeakFilteredOut_compSorted4_dist50kb_direct
VERSION: 2011.07.08

Try our De Novo Motif Finding Tool for ChIP-seq (SEME)

746 TFs
Show top 50 Families Go Download As Text

Rank [?]	Name [?]	Family [?]	Logo [?]	Score [?]	Distribution [?]	%Sequence with motif optimal setting [?]	%Sequence with motif 1e-4 fdr within +/- 200bp [?]	Binding Range [?]	PWM Score Cutoff [?]	Z0Score [?]	Z1Score [?]	P-value [?]
1	V$CEBPB_02	CEBP		121.024		 0.1460728	 0.2148467	80	2.9101	115.958	5.06592	0
2	V$jaspar_CEBPA	jaspar_Leucine_Zipper		100.261		 0.09224138	 0.132567	80	2.9262	95.3888	4.87201	0
3	V$PEA3_Q6	ETS		61.751		 0.1855364	 0.2746169	160	2.8742	55.1104	6.64066	0
4	V$jaspar_SPI1	jaspar_Ets		61.0182		 0.1469349	 0.2158046	160	3.5842	55.1445	5.87372	0
5	V$HLF_01	CREB		56.7262		 0.05651341	 0.09923372	80	3.0108	54.7479	1.97831	0

Results for Koeffler_BM_CebpE_NBM_BiclusterAssignment_SinglePeakFilteredOut_compSorted4_dist50kb_indirect
VERSION: 2011.07.08

Try our De Novo Motif Finding Tool for ChIP-seq (SEME)

746 TFs
Show top 50 Families Go Download As Text

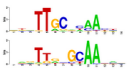
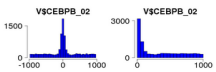
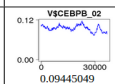
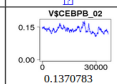
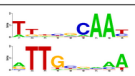
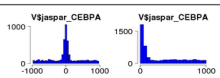
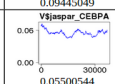
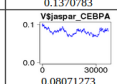
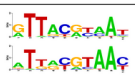
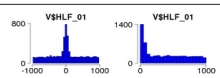
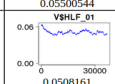
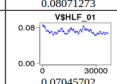
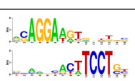
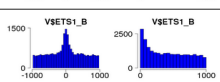
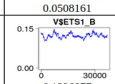
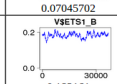
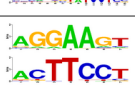
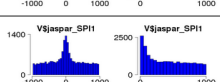
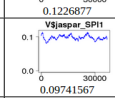
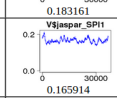
Rank [?]	Name [?]	Family [?]	Logo [?]	Score [?]	Distribution [?]	%Sequence with motif optimal setting [?]	%Sequence with motif 1e-4 fdr within +/- 200bp [?]	Binding Range [?]	PWM Score Cutoff [?]	Z0Score [?]	Z1Score [?]	P-value [?]
1	V$CEBPB_02	CEBP		129.73		 0.09445049	 0.1370783	120	2.9101	118.441	11.2893	0
2	V$jaspar_CEBPA	jaspar_Leucine_Zipper		102.866		 0.05500544	 0.08071273	120	2.9262	94.9489	7.91746	0
3	V$HLF_01	CREB		72.0175		 0.0508161	 0.07045702	160	3.0108	61.3525	10.6649	0
4	V$ETS1_B	ETS		63.8321		 0.1226877	 0.183161	200	3.1802	55.1567	8.67531	1.11E-16
5	V$jaspar_SPI1	jaspar_Ets		62.342		 0.09741567	 0.165914	160	3.5842	58.3031	4.03891	3.33E-16

Figure 5: Top 5 Motifs Discovered from Direct and Indirect Clusters.

5 Centrimo Motif Results for Log-Transformed ChIPseq

5.1 Centrimo Motif Results from 4-Component GMM

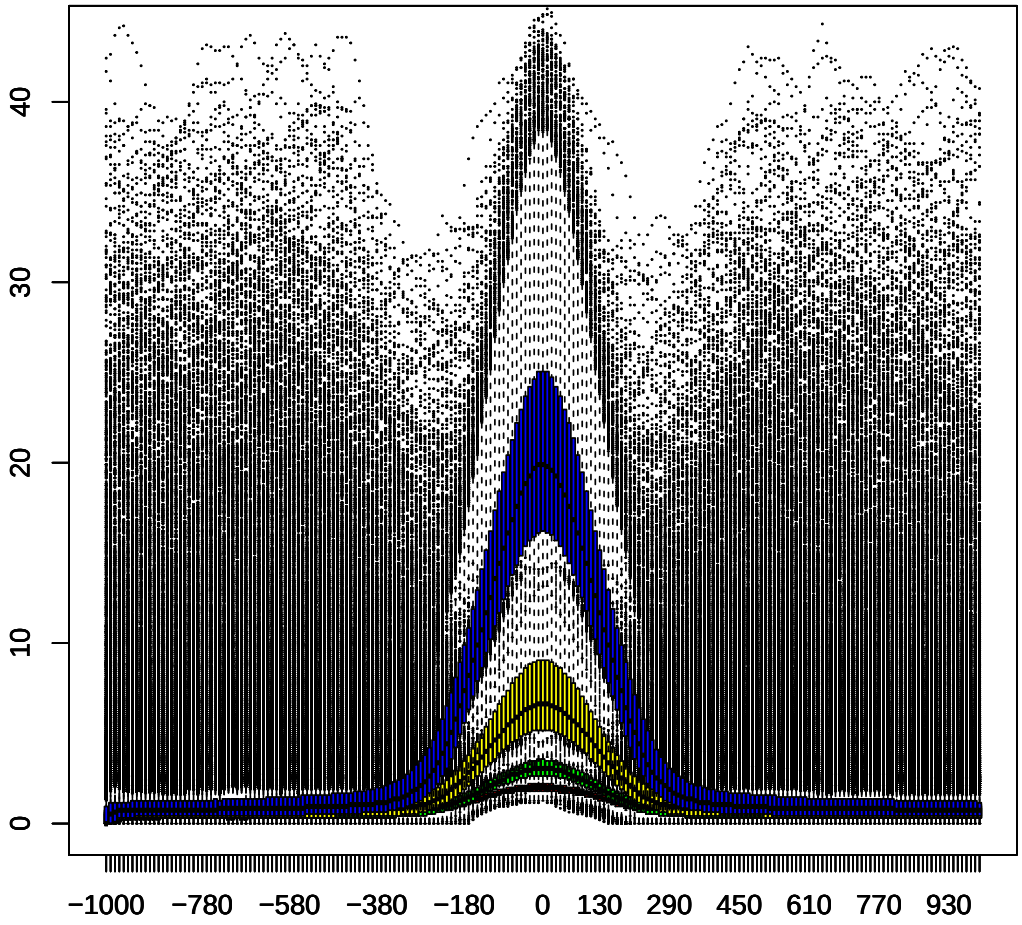


Figure 6: Boxplots of ChIP Profiles in 4-component GMM group1 (Red), group2(Green), group3 (Yellow) and group4 (Blue)

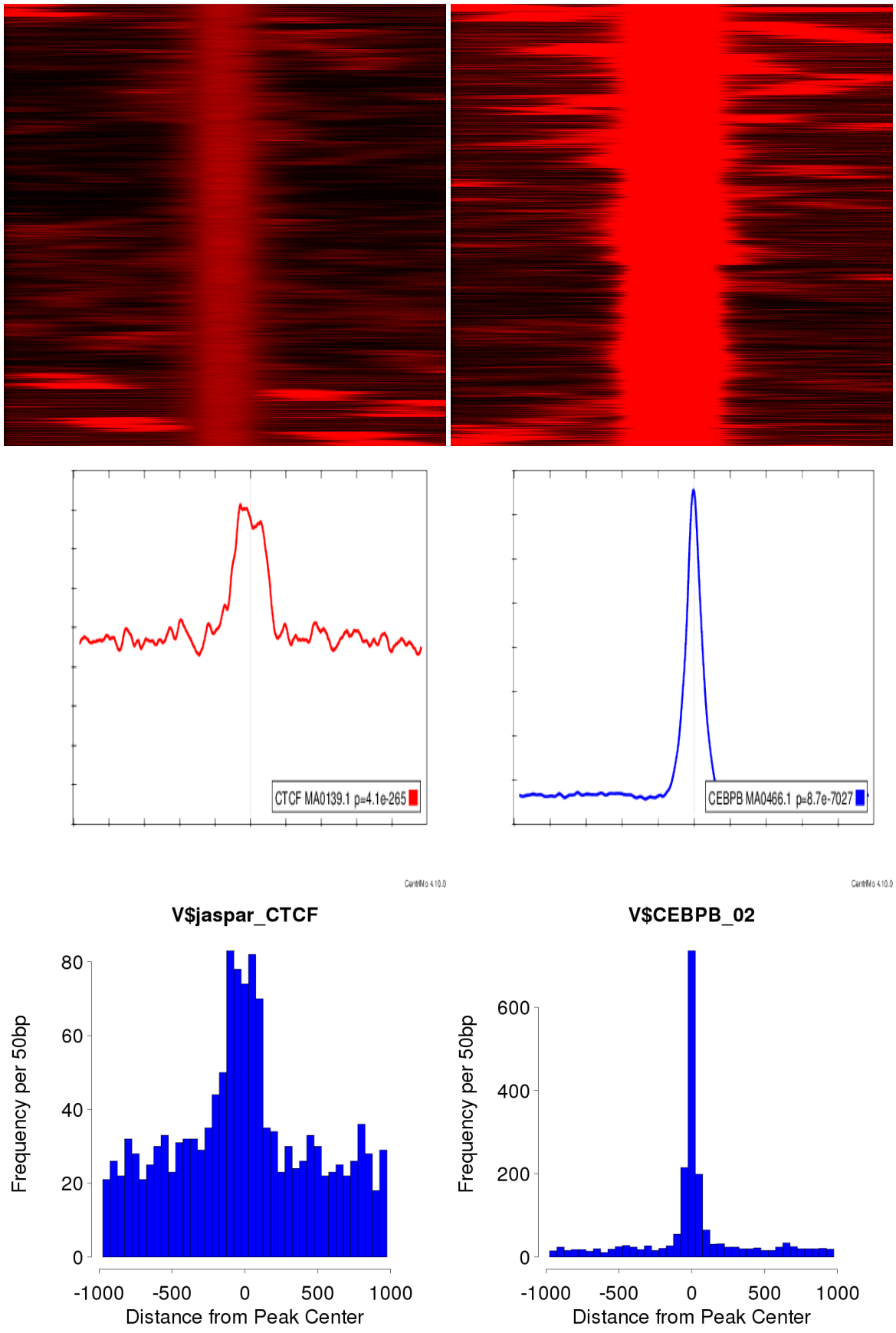


Figure 7: Heatmaps of ChIP Profiles in 4-component GMM group1 (Top Left), and group4 (Top Right). Red and Black represents positive and zero values,respectively. Below are the motif distributions for CTCF(Left) and CEBPB(Right) discovered by ChIPmeme.

5.2 Motif ChipSeq Distribution of CEBPB and CTCF: 4-Component GMM

```
source('/home/ricky/Rlim/ChromatinConformation/ComponentCalls/ComponentCalls.R')

visualizeBedPeaks(bed_f = paste0(work_dir,
                    'Output/ChipMeme_Koeffler_BM_CebpE_GMM_ModelAssignment_log_compSorted4_2kb/',
                    'Motif/CEBPB_CTCF_Motif.bed'),
                    chrom='chr14', title_f = 'Koeffler_BM_CebpE_GMM_ModelAssignment_log_CEBPB_CTCF',
                    output_f = 'figs/Koeffler_BM_CebpE_GMM_ModelAssignment_log_CEBPBCTCF',
                    max_peak=50)

## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
## use score as y by default
## Saving 15 x 7 in image
```

```
##   Comp Freq Total Percentage
## 1    1 1705 10186      16.74
## 2    2 4985 18419      27.06
```

```
## 3    3 6000 17555      34.18
## 4    4 2222  5317      41.79
```

Comp	Freq	Total	Percentage
1	1476	10186	14.49
2	2884	18419	15.66
3	2665	17555	15.18
4	760	5317	14.29

Table 6: CTCF Discovered by Centrimo in Groups of 4-Component Log-Transformed ChIPseq

Comp	Freq	Total	Percentage
1	1964	10186	19.28
2	5808	18419	31.53
3	6964	17555	39.67
4	2595	5317	48.81

Table 7: CEBPB Discovered by Centrimo in Groups in 4-Component Log-Transformed ChIPseq

Comp	Freq	Total	Percentage
1	1217	10186	11.95
2	2061	18419	11.19
3	1701	17555	9.69
4	387	5317	7.28

Table 8: CTCF Discovered by Centrimo in Groups of 4-Component Log-Transformed ChIPseq:no duplicate

Comp	Freq	Total	Percentage
1	1705	10186	16.74
2	4985	18419	27.06
3	6000	17555	34.18
4	2222	5317	41.79

Table 9: CEBPB Discovered by Centrimo in Groups in 4-Component Log-Transformed ChIPseq:no duplicate

5.3 Centrimo Motif Results from 5-Component GMM

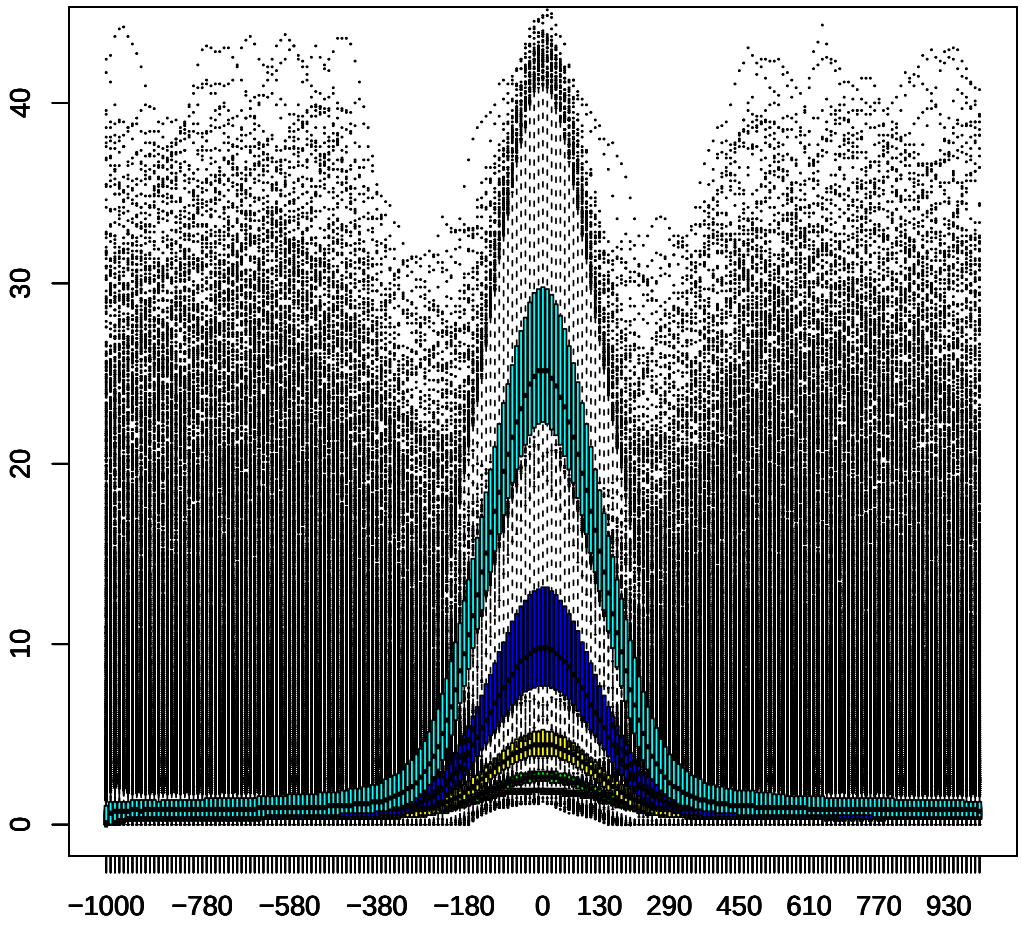


Figure 8: Boxplots of ChIP Profiles in 4-component GMM group1 (Red), group2(Green), group3 (Yellow), group4 (Blue), group5 (Cyan)

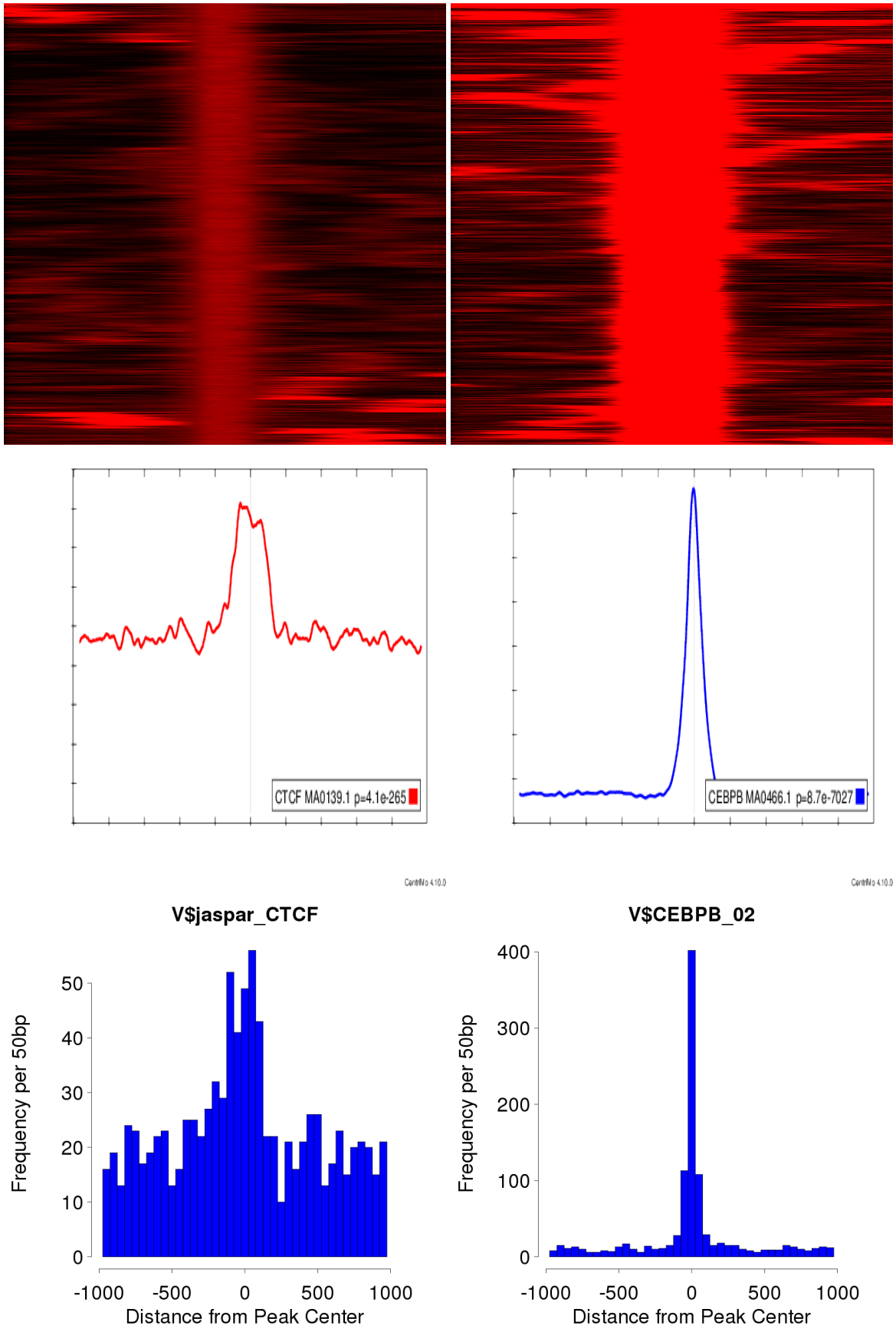


Figure 9: Heatmaps of ChIP Profiles in 5-component GMM group1 (Top Left), and group5 (Top Right). Red and Black represents positive and zero values, respectively. Below are the motif distributions for CTCF(Left) and CEBPB(Right) discovered by ChIPmeme.

```
## [1]  6964 15777 13812 12311  2613
##   Comp Freq Total Percentage
## 1    1  848  6964     12.177
## 2    2 1795 15777     11.377
## 3    3 1430 13812     10.353
## 4    4 1124 12311      9.130
## 5    5  169  2613      6.468
```

Comp	Freq	Total	Percentage
1	848	6964	12.18
2	1795	15777	11.38
3	1430	13812	10.35
4	1124	12311	9.13
5	169	2613	6.47

Table 10: CTCF Discovered by Centrimo in Groups of 5-Component Log-Transformed ChIPseq:no duplicate

Comp	Freq	Total	Percentage
1	1009	6964	14.49
2	3972	15777	25.18
3	4283	13812	31.01
4	4496	12311	36.52
5	1152	2613	44.09

Table 11: CEBPB Discovered by Centrimo in Groups in 5-Component Log-Transformed ChIPseq:no duplicate

6 De Novo Motif Discovery

6.1 PeakMotif: RSAT

```
Command: nice -n 19 /home/rsat/rsat/perl-scripts/peak-motifs
-v 1 -title 'Koeffler_BM_CebpE_Peak-Motifs'
-i /home/rsat/rsat/public_html/tmp/wwwrun/2014/11/03/
peak-motifs.2014-11-03.083858_2014-11-03.083858_XPVNnt/peak-motifspeak_seq
-max_seq_len 1000 -markov auto -disco oligos,dyads,positions,local_words
-nmotifs 10  -minol 6 -maxol 8  -no_merge_lengths -2str
-origin center  -motif_db jaspar_core_vertebrates tf
/home/rsat/rsat/public_html/data/motif_databases/JASPAR/jaspar_core_vertebrates_2013-11.tf
-scan_markov 1 -task purge,seqlen,composition,disco,collect_motifs,
motifs_vs_motifs,timelog,archive,synthesis,small_summary,motifs_vs_db,scan
-prefix peak-motifs -noov -img_format png
-outdir /home/rsat/rsat/public_html/tmp/
wwwrun/2014/11/03/peak-motifs.2014-11-03.083858_2014-11-03.083858_XPVNnt
```

```
# pwd: /home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/Output/
#      Rsat/peak-motifs.2014-11-03.083858_2014-11-03.083858_XPVNnt/results/sites

# skip the lines starting with '#' and ';'
# awk 'NF && $1!~/^#|^;/' peak-motifs_all_motifs_seqcoord.tab > peak_motifs_sites.tab

# for the jaspar db
# pwd: /home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/Output/Rsat/
#      peak-motifs.2014-11-03.083858_2014-11-03.083858_XPVNnt/results/discovered_vs_db

# skip the lines starting with '#' and ';'
# awk 'NF && $1!~/^#|^;/' peak-motifs_motifs_vs_db_jaspar_core_vertebrates.tab >
#   peak-motifs_db_jaspar.tab
```

```
#motif_dir = '/home/ricky/Rlim/ChromatinConformation/MotifCalls/'

Koeffler_BM_CebpE_Peaks_MotifSites <- read.table(paste0(work_dir,
    'Output/Rsat/peak-motifs.2014-11-03.083858_2014-11-03.083858_XPVNnt/results/',
    'sites/peak_motifs_sites.tab'))

# remove the duplicated sites
Koeffler_BM_CebpE_Peaks_MotifSites_Unique <- Koeffler_BM_CebpE_Peaks_MotifSites[
            !duplicated(Koeffler_BM_CebpE_Peaks_MotifSites$V1),]
Koeffler_BM_CebpE_Peaks_MotifSites_Unique <-
            Koeffler_BM_CebpE_Peaks_MotifSites_Unique[, c('V1', 'V3')]

Koeffler_BM_CebpE_Peaks_MotifSites_Unique <-
            addChrCoordinates(Koeffler_BM_CebpE_Peaks_MotifSites_Unique)

head(Koeffler_BM_CebpE_Peaks_MotifSites_Unique)

##                               V1                 V3  chr      end
## 1    mm10_chr1_6214908_6215908_+ oligos_6nt_mkv4_m1 chr1  6215908
## 5    mm10_chr1_7088298_7089298_+ oligos_6nt_mkv4_m1 chr1  7089298
```

```
## 10   mm10_chr1_9544786_9545786_+ oligos_6nt_mkv4_m1 chr1  9545786
## 11   mm10_chr1_9748015_9749015_+ oligos_6nt_mkv4_m1 chr1  9749015
## 12 mm10_chr1_10037332_10038332_+ oligos_6nt_mkv4_m1 chr1 10038332
## 14 mm10_chr1_10231356_10232356_+ oligos_6nt_mkv4_m1 chr1 10232356
```

```
# get the peak score
Koeffler_BM_CebpE_Peaks_1kb  <- read.table(paste0(work_dir,
                                        'Input/Koeffler_BM_CebpE_Peaks_1kb.bed'))
colnames(Koeffler_BM_CebpE_Peaks_1kb) <- c('chr', 'start', 'end', 'score')
Koeffler_BM_CebpE_Motifs<-
    merge(Koeffler_BM_CebpE_Peaks_MotifSites_Unique, Koeffler_BM_CebpE_Peaks_1kb,
          by=c('chr', 'end'))
head(Koeffler_BM_CebpE_Motifs)
```

```
##     chr       end                                 V1                 V3     start score
## 1 chr10 100439227 mm10_chr10_100438227_100439227_+ local_words_6nt_m2 100438227    29
## 2 chr10  10094466   mm10_chr10_10093466_10094466_+ oligos_6nt_mkv4_m8  10093466    25
## 3 chr10  10177046   mm10_chr10_10176046_10177046_+ oligos_6nt_mkv4_m9  10176046    19
## 4 chr10 102162675 mm10_chr10_102161675_102162675_+ oligos_6nt_mkv4_m6 102161675    20
## 5 chr10  10225546   mm10_chr10_10224546_10225546_+ oligos_6nt_mkv4_m7  10224546    20
## 6 chr10  10254385   mm10_chr10_10253385_10254385_+ local_words_6nt_m9  10253385    24
```

```
Koeffler_BM_CebpE_Motifs <- Koeffler_BM_CebpE_Motifs[,
          c('chr', 'start', 'end', 'score', 'V3')]
colnames(Koeffler_BM_CebpE_Motifs) <-
    c('chr', 'start', 'end', 'score', 'motifId')

# get the motif name from jaspar database
Koeffler_BM_CebpE_Jaspar <- read.table(paste0(work_dir,
    'Output/Rsat/peak-motifs.2014-11-03.083858_2014-11-03.083858_XPVNnt/results/',
    '/discovered_vs_db/peak-motifs_db_jaspar.tab'))
Koeffler_BM_CebpE_Jaspar <- Koeffler_BM_CebpE_Jaspar[, c('V1', 'V4')]
colnames(Koeffler_BM_CebpE_Jaspar) <- c('motifId', 'motifName')
Koeffler_BM_CebpE_Motifs <- merge(Koeffler_BM_CebpE_Motifs, Koeffler_BM_CebpE_Jaspar,
                        by=c('motifId','motifId'))
Koeffler_BM_CebpE_Motifs <- Koeffler_BM_CebpE_Motifs[,
                        c('chr', 'start', 'end', 'score', 'motifId', 'motifName')]

# get only CEBP motif
CEBP_motif <- Koeffler_BM_CebpE_Motifs$motifName %in% c('CEBPA', 'CEBPB', 'Cebpa')
Koeffler_BM_CebpE_CEBP_motif <- Koeffler_BM_CebpE_Motifs[CEBP_motif,
                                                         c('motifName', 'score')]
Koeffler_BM_CebpE_CEBP_motif[, 'motifName'] <- rep('CEBP',
                                                   nrow(Koeffler_BM_CebpE_CEBP_motif))
Koeffler_BM_CebpE_nonCEBP_motif <- Koeffler_BM_CebpE_Motifs[!CEBP_motif,
                                                         c('motifName', 'score')]
```

```
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
```

```
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
```

```
## Saving 7 x 7 in image
```

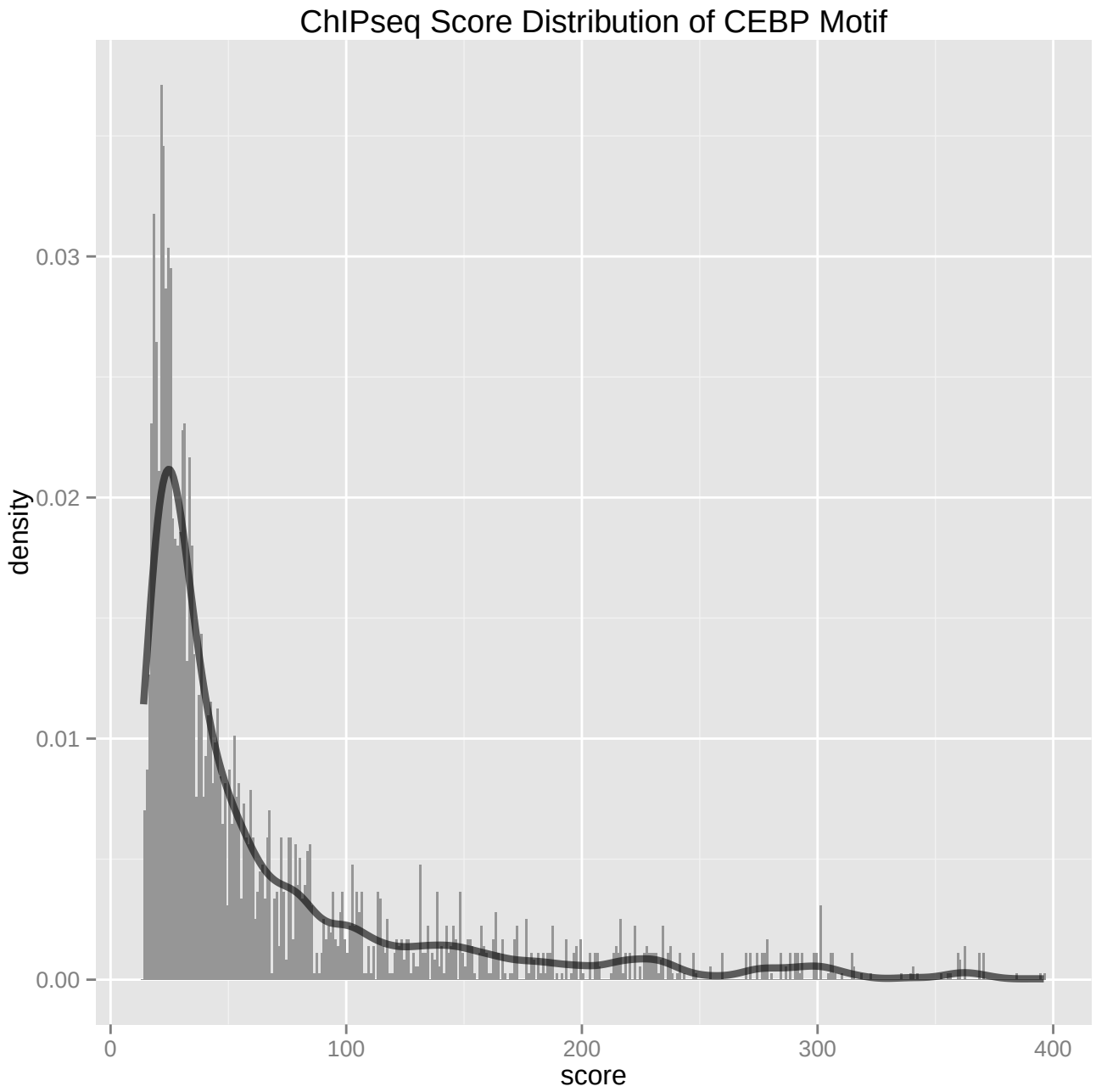


Figure 10: Distribution of ChIPseq Scores over all CEBP Motifs

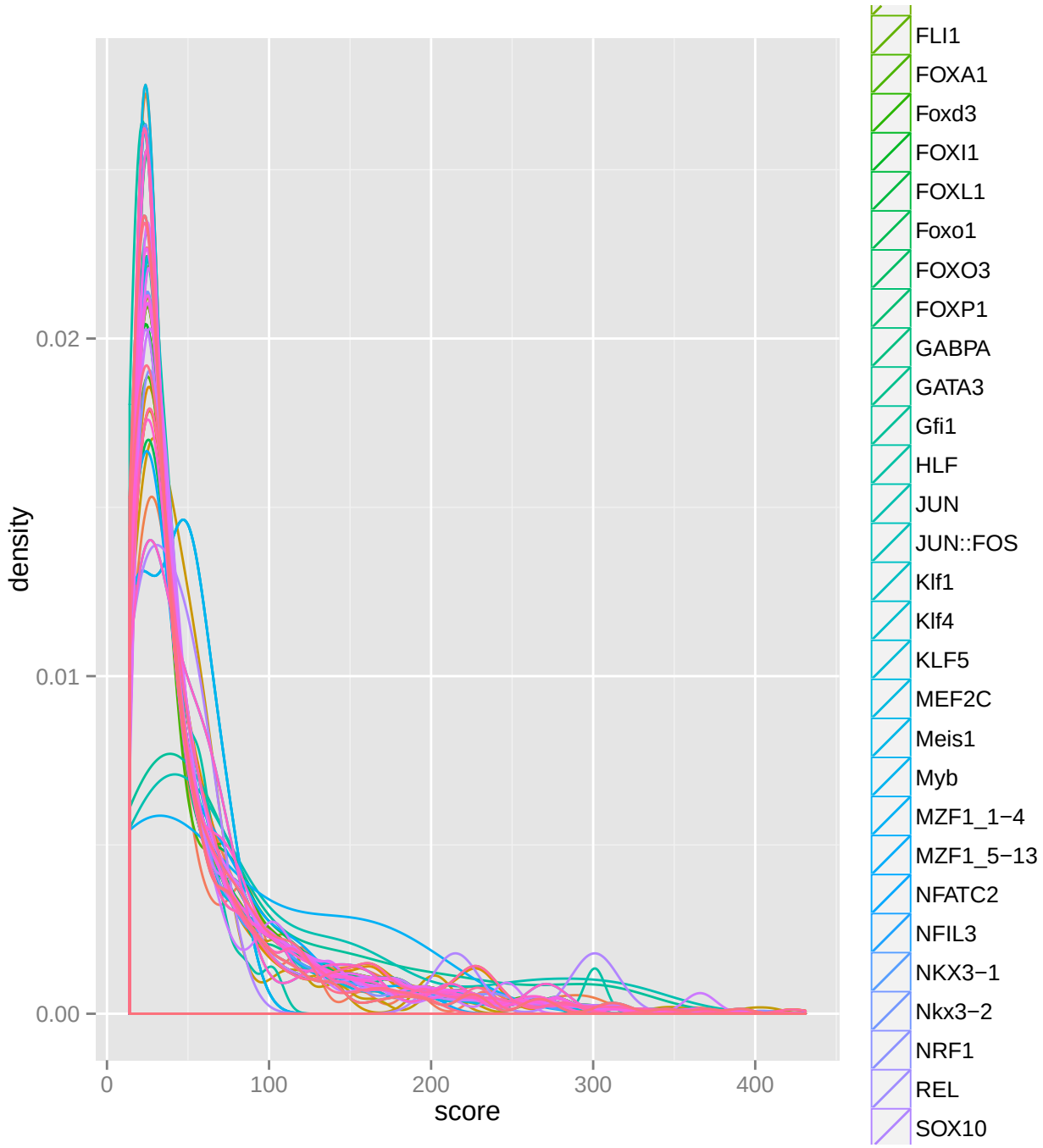


Figure 11: Distribution of ChIPseq Scores over all Non-CEBP Motifs

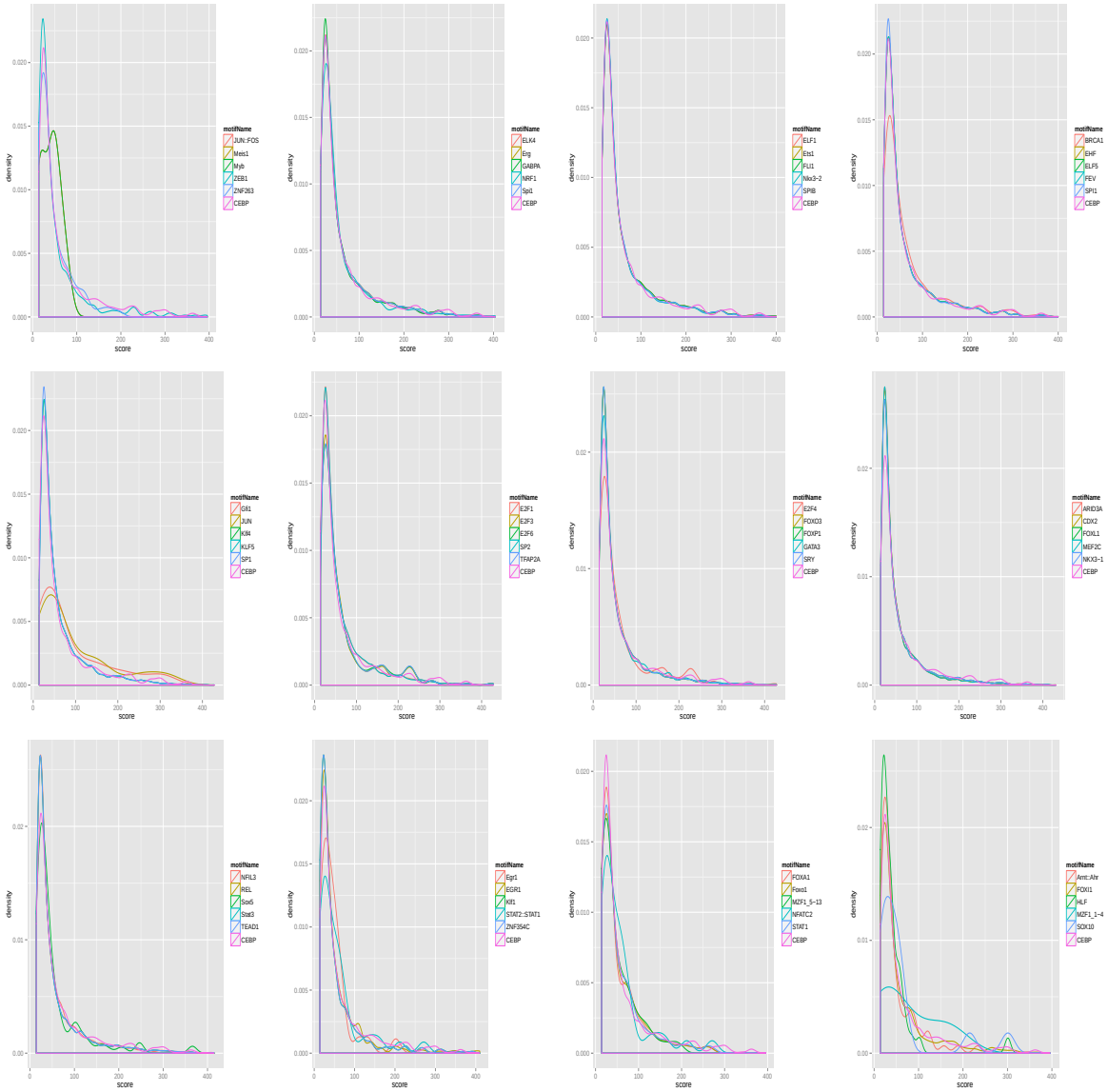


Figure 12: Distribution of ChIPseq Scores over all Non-CEBP Motifs in Chunks

6.2 Meme-Chip

We run meme-chip with the input of ChIPseq peaks of width 0.5kb. Below is the command-line that we run.

```
meme-chip -db ~/Rlim/Biotools/motif_databases/JASPAR_CORE_2014_vertebrates.meme
-oc MotifCalls/Output/ChipMeme
-index-name Koeffler_BM_CebpE_Peaks_ChipMeme1000bp
-meme-mod zoops -meme-minw 4 -meme-maxw 10 -meme-nmotifs 10
MotifCalls/Input/Koeffler_BM_CebpE_Peaks_0.5kb.fasta
```

Following the run, we analyzed the distribution of ChIPseq scores for each discovered motif.

```
pwd: /home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/Output/ChipMeme

# File of origin where we downloaded the fasta files:
meme_out/meme.html
For CebpE, only two motifs that are discovered.

# For centrimo we downloaded the gtf file for JUN motif:
file:///home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE
Output/ChipMeme/Koeffler_BM_Cebpe_Peaks_ChipMemeHomeHeteroCebpE
file stored: /home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/
Output/ChipMeme/jun_sites.gtf
```

```
## for a list of motifs (10 motifs output of meme)
# for i in $(ls motif_*_fasta.txt); do awk '/^>/' $i | sed 's/^>//g' | awk '{print $1}'>
# "`basename $i .txt`.site"; done

## add filename without _fasta.site to each line
# for i in $(ls *.site);do awk '{print $1"\t"FILENAME}' $i | sed 's/_fasta.site//g' >
# "`basename $i .site`.msite"; done

## concatenate all the motif sites
# cat *.msite > allmotif.msite

## get the peak with width of 0.5kb
# awk -F"`\t`" '{print $1"`\t`"$2-250"`\t`"$3+249"`\t`"$5}'
#  Koeffler_BM_CebpE_NBM_ModelAssignment_compSorted4.bed>Koeffler_BM_CebpE_Peaks_0.5kb.bed
# awk -F"\t" '{print $1"\t"$2-500"\t"$3+499"\t"$5}'
#  Koeffler_BM_CebpE_NBM_ModelAssignment_compSorted4.bed>Koeffler_BM_CebpE_Peaks_1kb.bed
```

```
# note that ChIP meme uses ChIP regions of 0.5kb
peak_0.5kb <- read.table(paste0(work_dir,
                                'Input/Koeffler_BM_CebpE_Peaks_0.5kb.bed'))
colnames(peak_0.5kb) <- c('chr', 'start', 'end', 'score')


all_motifs <- read.table(paste0(work_dir,
                                'Output/ChipMeme/MotifMeme/allmotif.msite'))
head(all_motifs)
```

```
##                                         V1      V2
## 1 mm10_chr11_108696540_108697040_+_site_1 motif_1
## 2  mm10_chr6_116416102_116416602_+_site_1 motif_1
## 3 mm10_chr12_100170230_100170730_+_site_1 motif_1
## 4   mm10_chr15_97590718_97591218_+_site_1 motif_1
## 5    mm10_chr3_67576180_67576680_+_site_1 motif_1
## 6  mm10_chr4_155666850_155667350_+_site_1 motif_1
```

```
cebp_motif <- all_motifs[all_motifs$V2 == 'motif_2',]

allmotif_score <- getPeakScore(all_motifs, peak_0.5kb)
cebpmotif_score <- getPeakScore(cebp_motif, peak_0.5kb)
colnames(allmotif_score) <- c('chr', 'start', 'end', 'id', 'score')
colnames(cebpmotif_score) <- c('chr', 'start', 'end', 'id', 'score')

#motif_1_score <- getPeakScore(motif_1, peak_0.5kb)
#motif_2_score <- getPeakScore(motif_2, peak_0.5kb)
head(allmotif_score)
```

```
##     chr     start       end      id score
## 1 chr10 115527961 115528461 motif_1    21
## 2 chr10 121571515 121572015 motif_2    56
## 3 chr10 128307932 128308432 motif_1    21
## 4 chr10 128491793 128492293 motif_2    51
## 5 chr10 128491793 128492293 motif_1    51
## 6 chr10  21146646  21147146 motif_2    27
```

```
head(cebpmotif_score)
```

```
##     chr     start       end      id score
## 1 chr10 121571515 121572015 motif_2    56
## 2 chr10 128491793 128492293 motif_2    51
## 3 chr10  21146646  21147146 motif_2    27
## 4 chr10   3274017   3274517 motif_2    50
## 5 chr10  66937028  66937528 motif_2   111
## 6 chr10  67712563  67713063 motif_2   271
```

```
## Saving 7 x 7 in image
## Saving 7 x 7 in image
```

1	MEME		13	MA0474.1 (Erg), MA0475.1 (FLI1), MA0098.2 (Ets1), MA0473.1 (ELF1), MA0156.1 (FEV), MA0598.1 (EHF), MA0080.3 (Spi1), MA0136.1 (ELF5), MA0520.1 (Stat6), MA0076.2 (ELK4)
2	MEME		5	MA0466.1 (CEBPB), MA0102.3 (CEBPA), MA0488.1 (JUN), MA0025.1 (NFIL3), MA0043.1 (HLF)
3	MEME		1	MA0472.1 (EGR2)
4	MEME		8	MA0079.3 (SP1), MA0162.2 (EGR1), MA0516.1 (SP2), MA0599.1 (KLF5), MA0472.1 (EGR2), MA0469.1 (E2F3), MA0039.2 (Klf4), MA0528.1 (ZNF263)
5	MEME		5	MA0506.1 (NRF1), MA0006.1 (Arnt::Ahr), MA0259.1 (HIF1A::ARNT), MA0162.2 (EGR1), MA0464.1 (Bhlhe40)
6	MEME		0	
7	MEME		6	MA0162.2 (EGR1), MA0469.1 (E2F3), MA0516.1 (SP2), MA0472.1 (EGR2), MA0470.1 (E2F4), MA0471.1 (E2F6)
8	MEME		6	MA0155.1 (INSM1), MA0516.1 (SP2), MA0139.1 (CTCF), MA0162.2 (EGR1), MA0079.3 (SP1), MA0469.1 (E2F3)
9	MEME		2	MA0162.2 (EGR1), MA0512.1 (Rxra)
10	MEME		7	MA0162.2 (EGR1), MA0516.1 (SP2), MA0506.1 (NRF1), MA0469.1 (E2F3), MA0472.1 (EGR2), MA0146.2 (Zfx), MA0076.2 (ELK4)

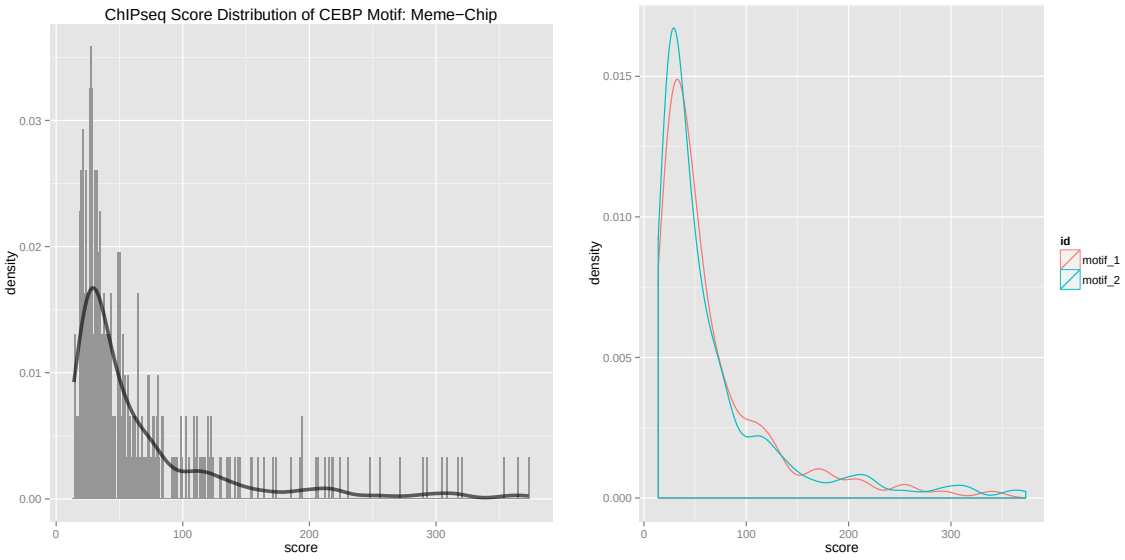


Figure 13: Distribution of ChIPseq Scores over Motif1: Meme-Chip

6.3 JunMotif from CentriMo

```
## convert jun sites in gtf to bed
# ./gtfMeme2Bed.py jun_sites.gtf > jun_sites.bed

## intersect the bed files with the ChIPseq bed
# bedtools intersect
# -b /home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/Output/
#    ChipMeme/jun_sites.bed
# -a /home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/Input/
#    Koeffler_BM_CebpE_Peaks_0.5kb.bed
# -u > /home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/Output/
#    ChipMeme/MotifMeme/jun_ChIPSites.bed
```

```
jun_chipScore <- read.table(paste0(work_dir, 'Output/ChipMeme/MotifMeme/jun_ChIPSites.bed'))
head(jun_chipScore)
```

```
##     V1       V2       V3  V4
## 1 chr1  7737308  7737808  22
## 2 chr1  9782091  9782591  24
## 3 chr1  9805843  9806343 138
## 4 chr1  9825750  9826250  36
## 5 chr1 10261386 10261886 265
## 6 chr1 10268226 10268726  81
```

```
colnames(jun_chipScore) <- c('chr', 'start', 'end', 'score')
p <- ggplot(jun_chipScore, aes(x=score))+
        geom_histogram(aes(y = ..density..), binwidth=1, fill='#969696') +
        geom_line(stat='density', alpha=0.6, size=1.5) +
        ggtitle('ChIPseq Score Distribution of JUN Motif: Meme-Chip')
ggsave(file='figs/Koeffler_BM_CebpE_Score_JUN_MemeChip.pdf')
```

```
## Saving 7 x 7 in image
```

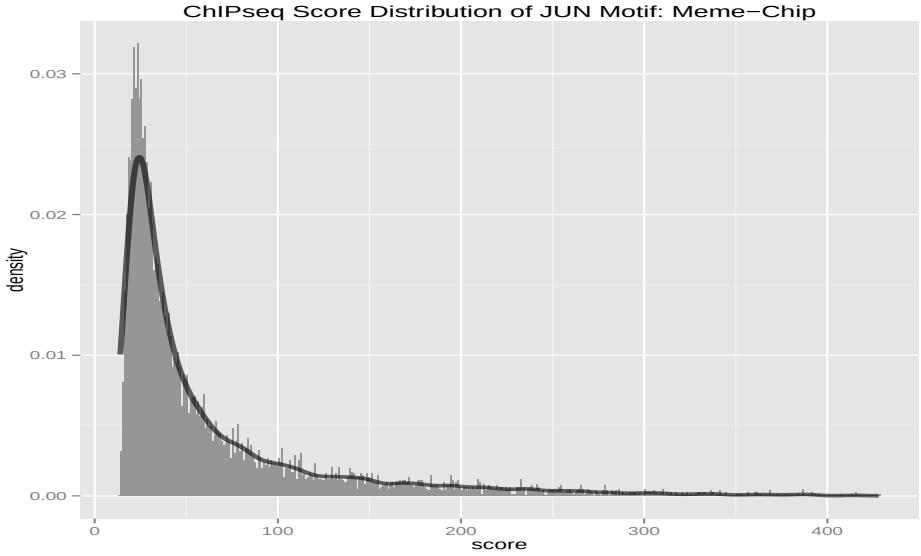


Figure 14: Distribution of ChIPseq Scores over Motifs: Centrimo-Chip

7 Perfect Homodimer and Perfect Heterodimer

Currently, Cebp has been identified to exist in two forms, i.e., homodimer and heterodimer. The homodimer motif of Cebp: ATTG __ CAAT.
The heterodimer motif of Cebp: TGA __ CAAT.

The motif sites of perfect homodimer and heterodimer of Cebp are selected from centrimo output.
file:///home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/Output/
ChipMeme/centrimo_out/centrimo.html

The matching sequences (for perfect homo and heterodimer) are then stored at:
/home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/Output/ChipMeme/MotifMeme/HomoHetero

```
## grep only unique sites for homo and heterodimer motif sites
# grep -v -x -f cebpPerfectHeterodimer.site cebpPerfectHomodimer.site >
# cebpPerfectHomodimerOnly.site
# grep -v -x -f cebpPerfectHomodimer.site cebpPerfectHeterodimer.site >
# cebpPerfectHeterodimerOnly.site

## add the filename to the 2nd column
# for i in $(ls *.site);do awk '{print $1"\t"FILENAME}' $i | sed 's/.site//g' > "`basename
$i .site`.msite"; done

## combine the homo and heterodimer motif sites
# cat cebpPerfectHeterodimerOnly.msite cebpPerfectHomodimerOnly.msite >
# cebpPerfectHomoHeterodimerOnly.msite
```

```
homohetero_dir=paste0('/home/ricky/Rlim/ChromatinConformation/MotifCalls/CebpE/',
                      'Output/ChipMeme/MotifMeme/HomoHetero/')

homoheterodimer_msite <- read.table(paste0(homohetero_dir,
                                  'cebpPerfectHomoHeterodimerOnly.msite' ))

homodimer_motif <- homoheterodimer_msite[
                       homoheterodimer_msite$V2 == 'cebpPerfectHomodimerOnly',]
nHomodimer <- nrow(homodimer_motif)
heterodimer_motif <- homoheterodimer_msite[
                       homoheterodimer_msite$V2 == 'cebpPerfectHeterodimerOnly',]
nHeterodimer <- nrow(heterodimer_motif)


homoheterodimer_score <- getPeakScore(homoheterodimer_msite, peak_0.5kb)
homodimer_score <- getPeakScore(homodimer_motif, peak_0.5kb)
heterodimer_score<- getPeakScore(heterodimer_motif, peak_0.5kb)
colnames(homoheterodimer_score) <- c('chr', 'start', 'end', 'id', 'score')
colnames(homodimer_score) <- c('chr', 'start', 'end', 'id', 'score')
colnames(heterodimer_score) <- c('chr', 'start', 'end', 'id', 'score')
```

```
## Saving 7 x 7 in image
## Saving 7 x 7 in image
## Saving 7 x 7 in image
```

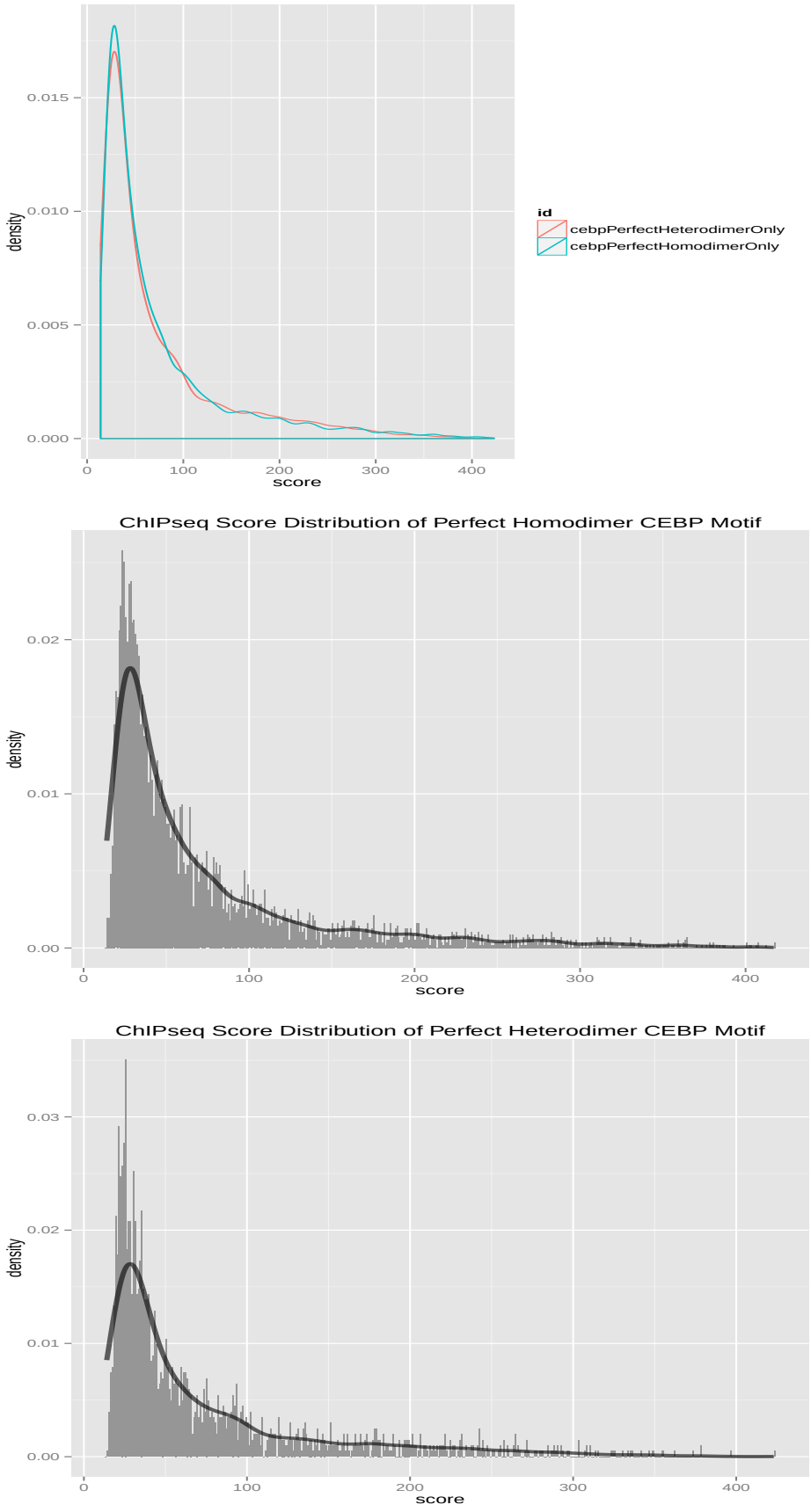


Figure 15: Distribution of ChIPseq Scores over Homo and Heterodimer of Cebp Motifs.

8 Summary

Our preliminary results from motif discovery analysis show:

- The presence of CEBP motif in direct and indirect clusters (CentDist).
- CEBP motif is distributed across the range of ChIPseq scores.
- Homodimer CEBP motif (5587) occurs more frequently than its heterodimer form (2024).
- CEBP homo and heterodimer motifs show overlapping distribution of ChIPseq scores.

The presence of CEBP motif in all clusters (direct and indirect) and across ChIPseq scores could be due to the heterogeneity of cells in our dataset which originated from bone marrow.

9 Metainfo

```
sessionInfo()

## R version 3.0.2 (2013-09-25)
## Platform: x86_64-pc-linux-gnu (64-bit)
##
## locale:
##  [1] LC_CTYPE=en_SG.UTF-8       LC_NUMERIC=C               LC_TIME=en_SG.UTF-8
##  [4] LC_COLLATE=en_SG.UTF-8     LC_MONETARY=en_SG.UTF-8    LC_MESSAGES=en_SG.UTF-8
##  [7] LC_PAPER=en_SG.UTF-8       LC_NAME=C                  LC_ADDRESS=C
## [10] LC_TELEPHONE=C             LC_MEASUREMENT=en_SG.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] grid      parallel  stats     graphics  grDevices utils     datasets  methods
## [9] base
##
## other attached packages:
##  [1] wordcloud_2.5        rtracklayer_1.22.7   reshape2_1.4         RColorBrewer_1.0-5
##  [5] mixtools_1.0.2       segmented_0.4-0.0    MASS_7.3-34          gridExtra_0.9.1
##  [9] ggbio_1.10.16        BSgenome_1.30.0      GenomicRanges_1.14.4 broom_0.2
## [13] boot_1.3-11          Biostrings_2.30.1    XVector_0.2.0        IRanges_1.20.7
## [17] BiocGenerics_0.8.0   plyr_1.8.1           xtable_1.7-3         scales_0.2.4
## [21] ggplot2_1.0.0        knitr_1.6
##
## loaded via a namespace (and not attached):
##  [1] AnnotationDbi_1.24.0     Biobase_2.22.0           biomaRt_2.18.0
##  [4] biovizBase_1.10.8        bitops_1.0-6             cluster_1.15.2
##  [7] codetools_0.2-9          colorspace_1.2-4         DBI_0.2-7
## [10] dichromat_2.0-0          digest_0.6.4             evaluate_0.5.5
## [13] formatR_1.0              Formula_1.1-2            GenomicFeatures_1.14.5
## [16] gtable_0.1.2             highr_0.3                Hmisc_3.14-4
## [19] labeling_0.3             lattice_0.20-29          latticeExtra_0.6-26
## [22] munsell_0.4.2            proto_0.3-10             Rcpp_0.11.2
## [25] RCurl_1.95-4.3           Rsamtools_1.14.3         RSQLite_0.11.4
## [28] slam_0.1-32              splines_3.0.2            stats4_3.0.2
## [31] stringr_0.6.2            survival_2.37-7          tools_3.0.2
## [34] VariantAnnotation_1.8.13 XML_3.98-1.1             zlibbioc_1.8.0
```

```
library(knitr)
purl("motifAnalysis.Rnw" ) # compile to tex
purl("motifAnalysis.Rnw", documentation = 0) # extract R code only
knit2pdf("motifAnalysis.Rnw")

## Error:  duplicate label 'setup'
```