Chromatin Conformation Prediction from ChIPseq

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

[1]CSI-NUS [2]Ecole Normale Superieur [3]La Trobe University

CSI-Meeting

mailto:rlim.email@gmail.com

Contents

1. Introduction
   - Goal
     - What is ChIPseq?
     - What is Mixture Model (MM)?
2. Preliminary results
   - Pipeline
     - Input Data
     - Peak Calls: MACS2 vs jaHMM
   - Summary: Peak Calls
     - Component Calls
   - Summary: Component Calls
     - Motif Calls
   - Summary: Motif Calls
     - Model Update: Biclustering
3. Summary and Future

Chromatin Conformation Prediction

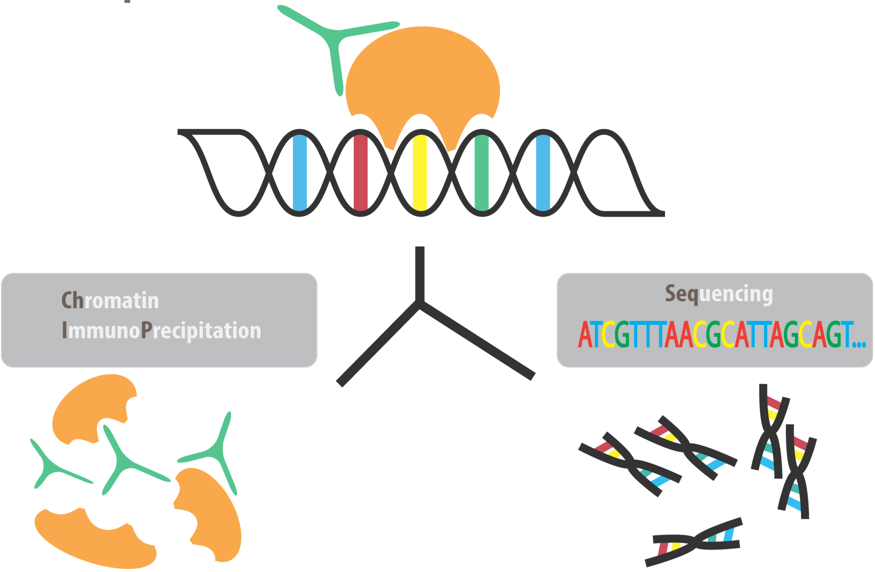
- **Main Question**: Can we use transcription factor (TF)-ChIPseq to predict protein complexes (direct and indirect bindings) on chromatin?

Chromatin Conformation Prediction

- **Main Question**: Can we use transcription factor (TF)-ChIPseq to predict protein complexes (direct and indirect bindings) on chromatin?
- **Strategy**: Model ChIPseq signal using Mixture Models to cluster the direct and indirect bindings.

What is ChIPseq?

ChIP-Seq

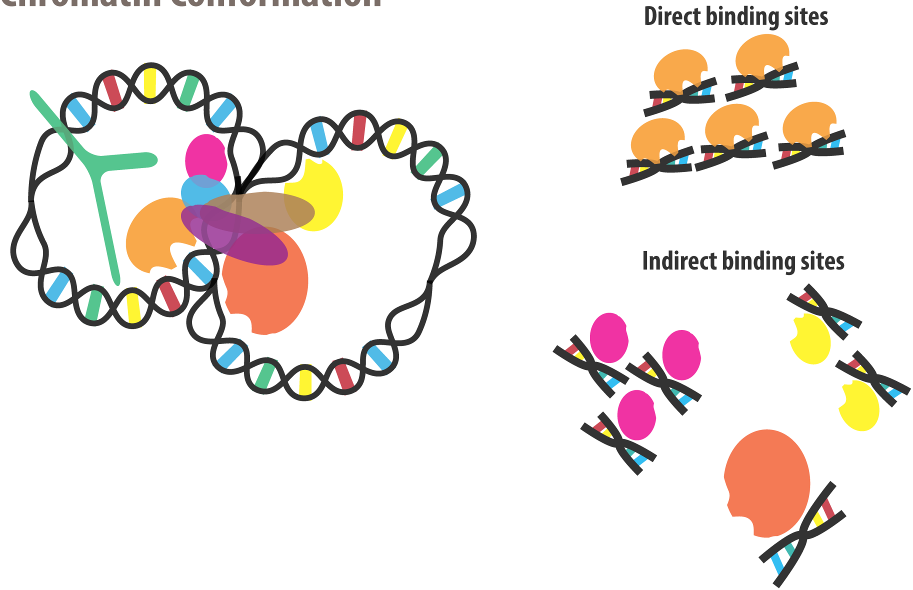


Chromatin ImmunoPrecipitation

Sequencing

ATCGTTTAACGCATTAGCAGT...

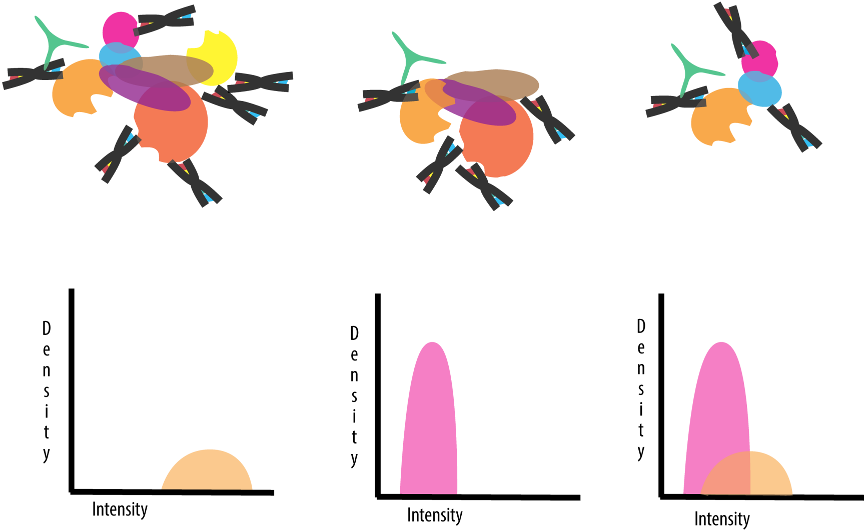
Chromatin Conformation



Direct binding sites

Indirect binding sites

Mixture of Chromatin Conformations



What is MM?

Mixture Model (GMM): Revisited

Types of clustering methods:

- Hard clustering: non-overlapping clusters
- Soft clustering: overlapping clusters

Mixture Model (GMM): Revisited

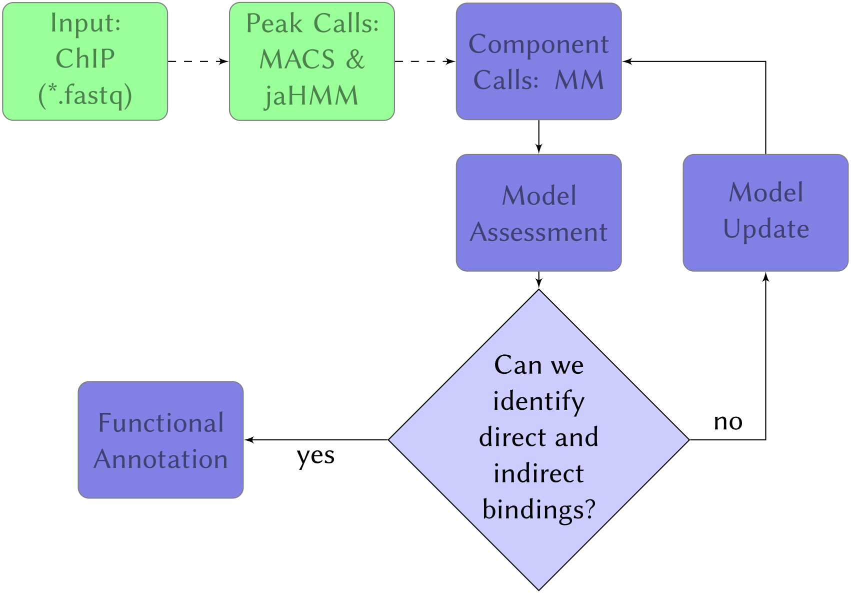
Types of clustering methods:

- Hard clustering: non-overlapping clusters
- Soft clustering: overlapping clusters

MM is a probabilistic way of soft clustering. Each cluster is a generative mixture model (pdf) with its parameters.

Mixture Gaussian pdf:
Key Assumption:

- ChIP-seq peaks are drawn from a finite set of gaussian distributions.
- ChIPseq peaks are fit with gaussian mixture models, with mixing λ parameter.
- Each gaussian corresponds to a cluster of peaks with μ and σ parameters.



Input: ChIP (*.fastq)

Peak Calls: MACS & jaHMM

Component Calls: MM

Model Assessment

Model Update

Can we identify direct and indirect bindings?

no

yes

Functional Annotation

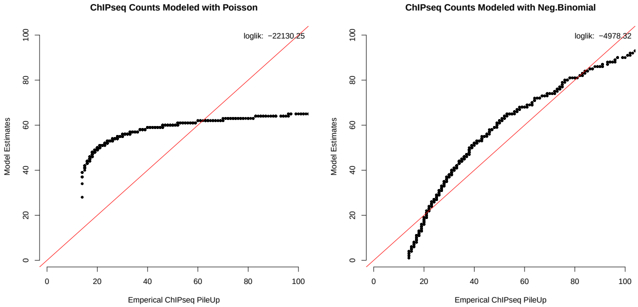
Input: ChIP-seq of Cebpϵ from Koeffler-BM

```
##FastQC 0.10.1
>>Basic Statistics pass
#Measure Value
Encoding Illumina 1.5
Total Sequences 41586141
Sequence length 40
#Summary
PASS Basic Statistics
PASS Per base sequence quality
PASS Per sequence quality scores
WARN Per base sequence content
PASS Per base GC content
PASS Per sequence GC content
PASS Per base N content
PASS Sequence Length Distribution
PASS Sequence Duplication Levels
PASS Overrepresented sequences
WARN Kmer Content
```

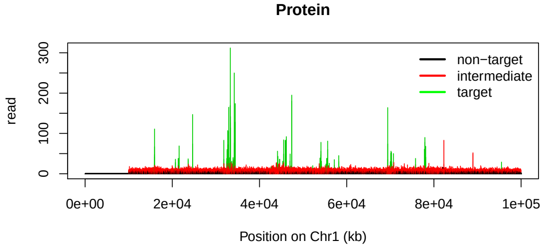
Principles

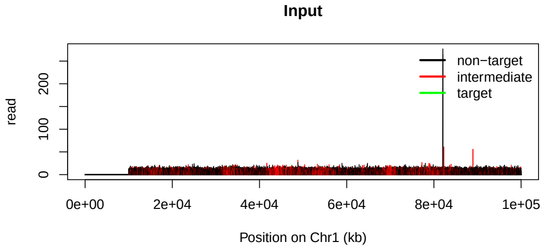
- **MACS2**: *poisson* model-based analysis of Peak calls MACS reference
- **jaHMM**: *negative binomial* model-based analysis of Peak calls jaHMM reference

Why jaHMM is better?

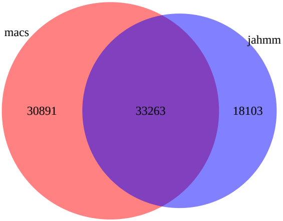


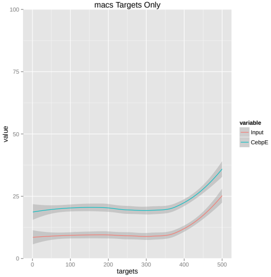
Targets Identified by jahmm

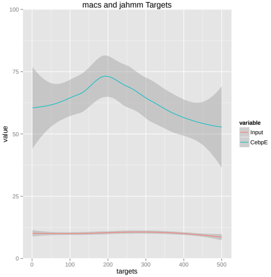


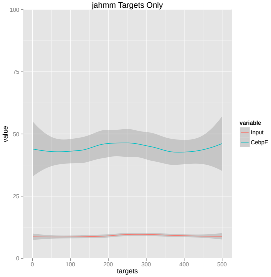


Targets Identified by MACS2 vs jahmm




macs Targets Only


macs and jahmm Targets


jahmm Targets Only

Why jaHMM is better than MACS2?

- Given our dataset, negative binomial model assumed by jaHMM fits better than poisson model assumed by MACS2

Why jaHMM is better than MACS2?

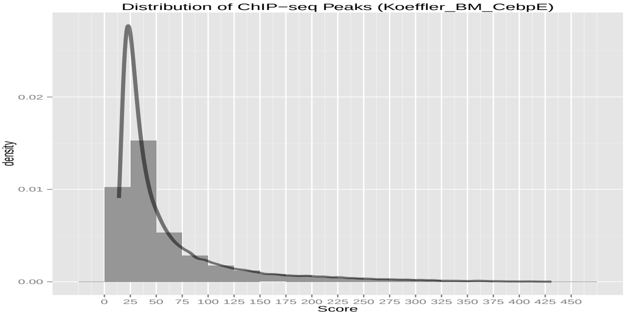
- Given our dataset, negative binomial model assumed by jaHMM fits better than poisson model assumed by MACS2
- jaHMM identified more peaks than MACS2

Why jaHMM is better than MACS2?

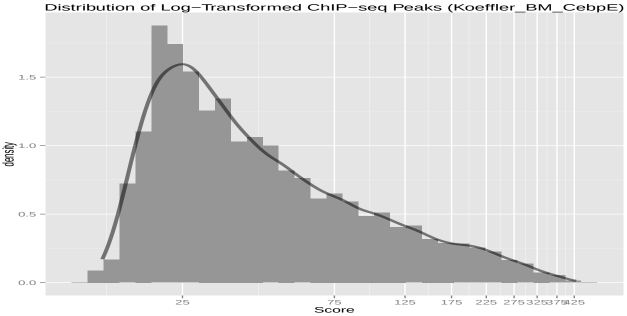
- Given our dataset, negative binomial model assumed by jaHMM fits better than poisson model assumed by MACS2
- jaHMM identified more peaks than MACS2
- Peaks identified solely by jaHMM have scores higher with respect to their input than solely by MACS2

Can we model ChIPseq Peaks using components of MMs?

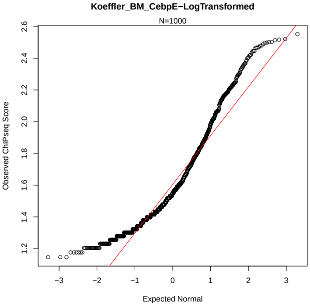
Input: ChIP-seq of Cebpϵ from Koeffler-BM

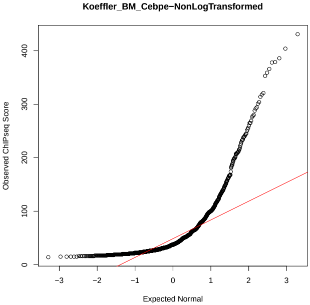


Log Transformation of ChIP-seq Input

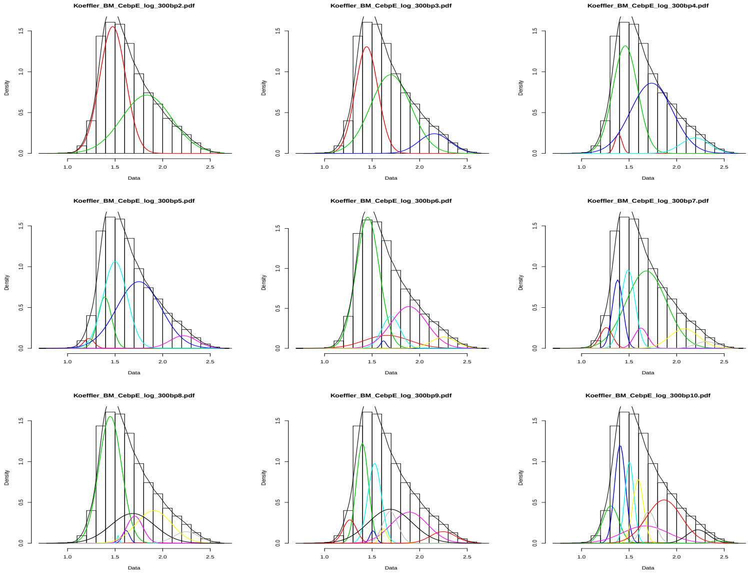


Check the Normality





ComponentCalls: Fit ChIPseq Peaks with GMMs



Koeffler_BM_CebpE_log_300bp2.pdf

Koeffler_BM_CebpE_log_300bp3.pdf

Koeffler_BM_CebpE_log_300bp4.pdf

Koeffler_BM_CebpE_log_300bp5.pdf

Koeffler_BM_CebpE_log_300bp6.pdf

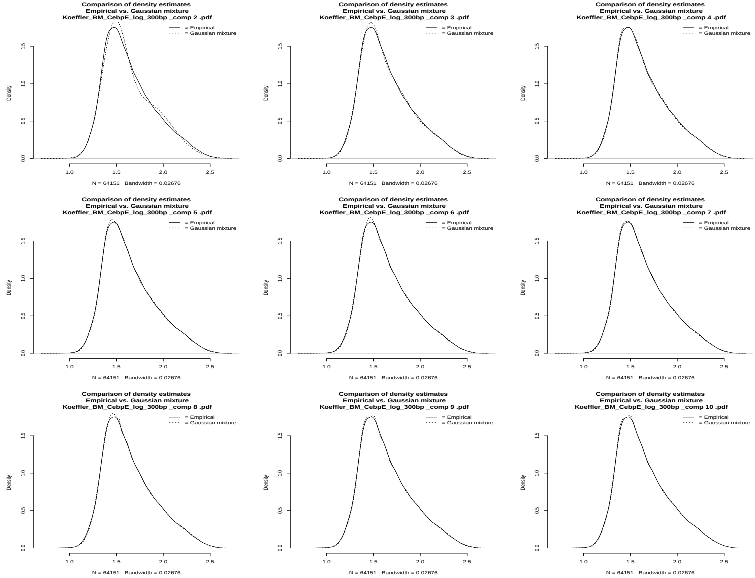
Koeffler_BM_CebpE_log_300bp7.pdf

Koeffler_BM_CebpE_log_300bp8.pdf

Koeffler_BM_CebpE_log_300bp9.pdf

Koeffler_BM_CebpE_log_300bp10.pdf

GMM-ModelAssessment: Overfit[1]

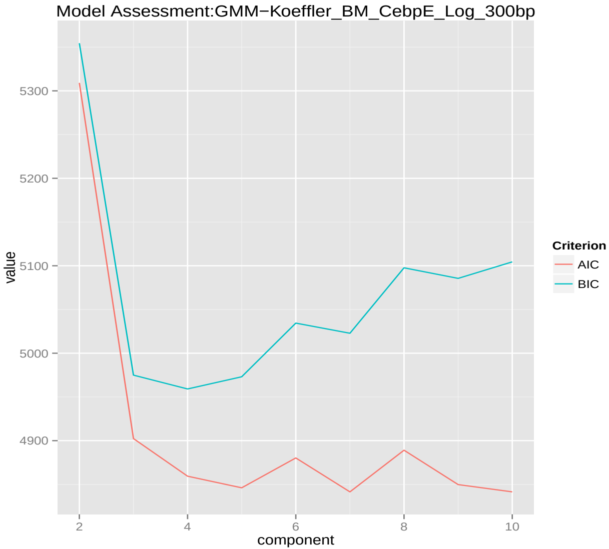


Model Assessment: BIC-AIC

AIC and BIC is based on Occam's razor principle, i.e, the simplest the better.

AIC = $-2 \times \log L + 2 * P$
BIC = $-2 \times \log L + \log(n) * P$
L is likelihood
P is the number of parameters

Model Assessment: BIC-AIC

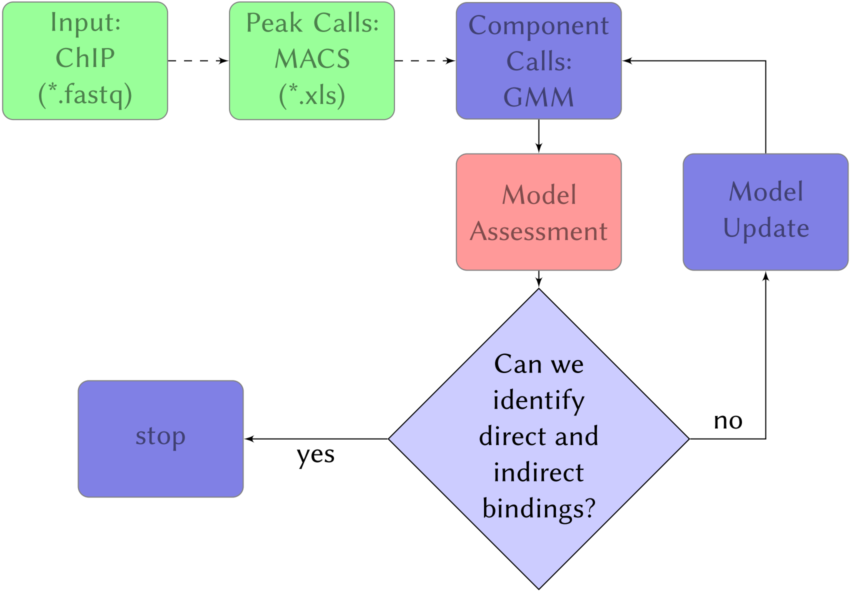


Summary

- **Can we model ChIPseq using several components of MMs?**
  Yes, our ChIPseq Peaks identified by jaHMM can be fit with GMMs.

Summary

- **Can we model ChIPseq using several components of MMs?**
  Yes, our ChIPseq Peaks identified by jaHMM can be fit with GMMs.
- **How many components are required?**
  From AIC-BIC and cross-validation, with 3 components are sufficient to fit the ChIPseq.
  Note: the lower the AIC and BIC values, the better the fitting.



Input:
ChIP
(*.fastq)

Peak Calls:
MACS
(*.xls)

Component
Calls:
GMM

Model
Assessment

Can we
identify
direct and
indirect
bindings?

yes

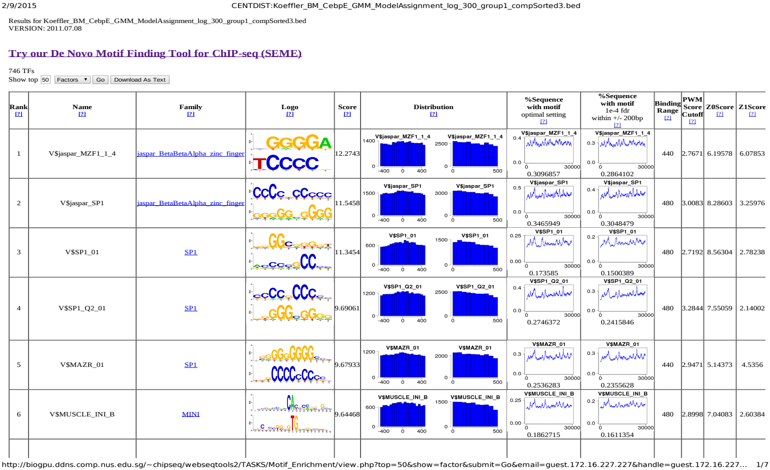
stop

no

Model
Update

Motif Calls using Centdist

Group1: low peak score (29559 peaks)



2/9/2015 CENTDIST:Koeffler_BM_CebpE_GMM_ModelAssignment_log_300_group1_compSorted3.bed

Results for Koeffler_BM_CebpE_GMM_ModelAssignment_log_300_group1_compSorted3.bed
VERSION: 2011.07.08

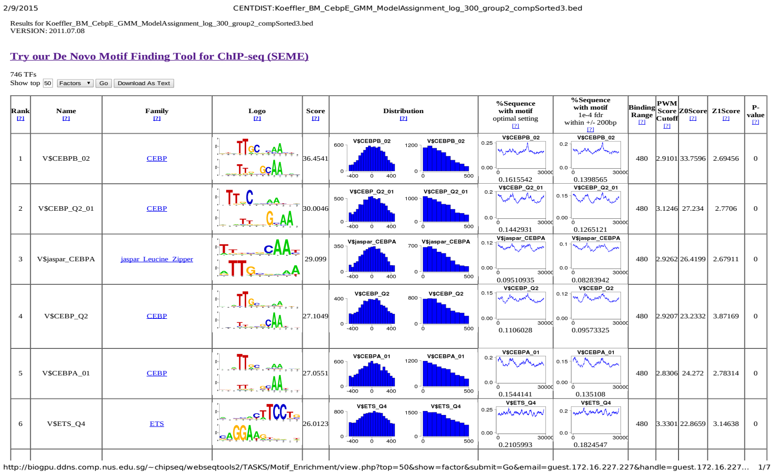
Try our De Novo Motif Finding Tool for ChIP-seq (SEME)

746 TFs
Show top 50 Factors ▾ 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** [?]
1	V$jaspar_MZF1_1_4	jaspar_BetaBetaAlpha_zinc_finger		12.2743	V$jaspar_MZF1_1_4 V$jaspar_MZF1_1_4	V$jaspar_MZF1_1_4 0.3096857	V$jaspar_MZF1_1_4 0.2864102	440	2.7671	6.19578	6.07853
2	V$jaspar_SP1	jaspar_BetaBetaAlpha_zinc_finger		11.5458	V$jaspar_SP1 V$jaspar_SP1	V$jaspar_SP1 0.3465949	V$jaspar_SP1 0.3048479	480	3.0083	8.28603	3.25976
3	V$SP1_01	SP1		11.3454	V$SP1_01 V$SP1_01	V$SP1_01 0.173585	V$SP1_01 0.1500389	480	2.7192	8.56304	2.78238
4	V$SP1_Q2_01	SP1		9.69061	V$SP1_Q2_01 V$SP1_Q2_01	V$SP1_Q2_01 0.2746372	V$SP1_Q2_01 0.2415846	480	3.2844	7.55059	2.14002
5	V$MAZR_01	SP1		9.67933	V$MAZR_01 V$MAZR_01	V$MAZR_01 0.2536283	V$MAZR_01 0.2355628	440	2.9471	5.14373	4.5356
6	V$MUSCLE_INI_B	MINI		9.64468	V$MUSCLE_INI_B V$MUSCLE_INI_B	V$MUSCLE_INI_B 0.1862715	V$MUSCLE_INI_B 0.1611354	480	2.8998	7.04083	2.60384

http://biogpu.ddns.comp.nus.edu.sg/~chipseq/webseqtools2/TASKS/Motif_Enrichment/view.php?top=50&show=factor&submit=Go&email=guest.172.16.227.227&handle=guest.172.16.227… 1/7

Group2: intermediate peak score (28851 peaks)



2/9/2015 CENTDIST:Koeffler_BM_CebpE_GMM_ModelAssignment_log_300_group2_compSorted3.bed

Results for Koeffler_BM_CebpE_GMM_ModelAssignment_log_300_group2_compSorted3.bed
VERSION: 2011.07.08

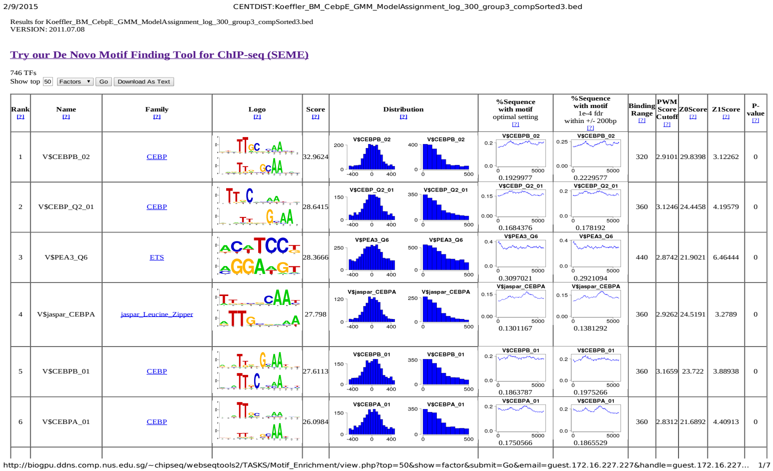
Try our De Novo Motif Finding Tool for ChIP-seq (SEME)

746 TFs
Show top 50 Factors ▾ 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$CEBPB_02	CEBP		36.4541		0.1615542	0.1398565	480	2.9101	33.7596	2.69456	0
2	V$CEBP_Q2_01	CEBP		30.0046		0.1442931	0.1265121	480	3.1246	27.234	2.7706	0
3	V$jaspar_CEBPA	jaspar_Leucine_Zipper		29.099		0.09510935	0.08283942	480	2.9262	26.4199	2.67911	0
4	V$CEBP_Q2	CEBP		27.1049		0.1106028	0.09573325	480	2.9207	23.2332	3.87169	0
5	V$CEBPA_01	CEBP		27.0551		0.1544141	0.135108	480	2.8306	24.272	2.78314	0
6	V$ETS_Q4	ETS		26.0123		0.2105993	0.1824547	480	3.3301	22.8659	3.14638	0

http://biogpu.ddns.comp.nus.edu.sg/~chipseq/webseqtools2/TASKS/Motif_Enrichment/view.php?top=50&show=factor&submit=Go&email=guest.172.16.227.227&handle=guest.172.16.227... 1/7

Group3: high peak score (5741 peaks)



Results for Koeffler_BM_CebpE_GMM_ModelAssignment_log_300_group3_compSorted3.bed
VERSION: 2011.07.08

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

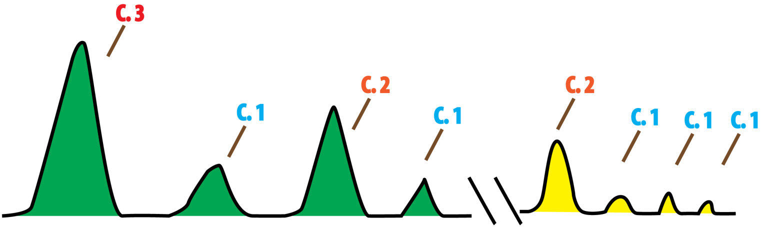
746 TFs
Show top 50 Factors ▾ Go Download As Text

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$CEBPB_02	CEBP		32.9624	V$CEBPB_02 V$CEBPB_02	V$CEBPB_02 0.1929977	V$CEBPB_02 0.2229577	320	2.9101	29.8398	3.12262	0
2	V$CEBP_Q2_01	CEBP		28.6415	V$CEBP_Q2_01 V$CEBP_Q2_01	V$CEBP_Q2_01 0.1684376	V$CEBP_Q2_01 0.178192	360	3.1246	24.4458	4.19579	0
3	V$PEA3_Q6	ETS		28.3666	V$PEA3_Q6 V$PEA3_Q6	V$PEA3_Q6 0.3097021	V$PEA3_Q6 0.2921094	440	2.8742	21.9021	6.46444	0
4	V$jaspar_CEBPA	jaspar_Leucine_Zipper		27.798	V$jaspar_CEBPA V$jaspar_CEBPA	V$jaspar_CEBPA 0.1301167	V$jaspar_CEBPA 0.1381292	360	2.9262	24.5191	3.2789	0
5	V$CEBPB_01	CEBP		27.6113	V$CEBPB_01 V$CEBPB_01	V$CEBPB_01 0.1863787	V$CEBPB_01 0.1975266	360	3.1659	23.722	3.88938	0
6	V$CEBPA_01	CEBP		26.0984	V$CEBPA_01 V$CEBPA_01	V$CEBPA_01 0.1750566	V$CEBPA_01 0.1865529	360	2.8312	21.6892	4.40913	0

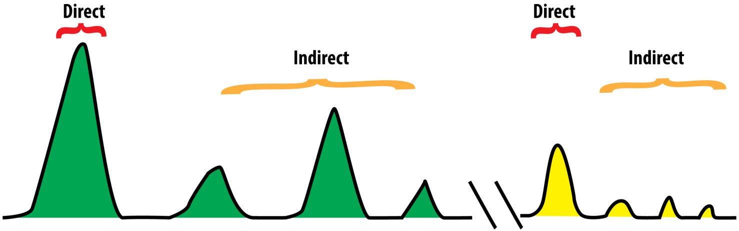
- Cebp motif is found in group3 only in 3-component GMMS using centdist

- Cebp motif is found in group3 only in 3-component GMMS using centdist
- Next, can we further segregate these groups into direct and indirect bindings?

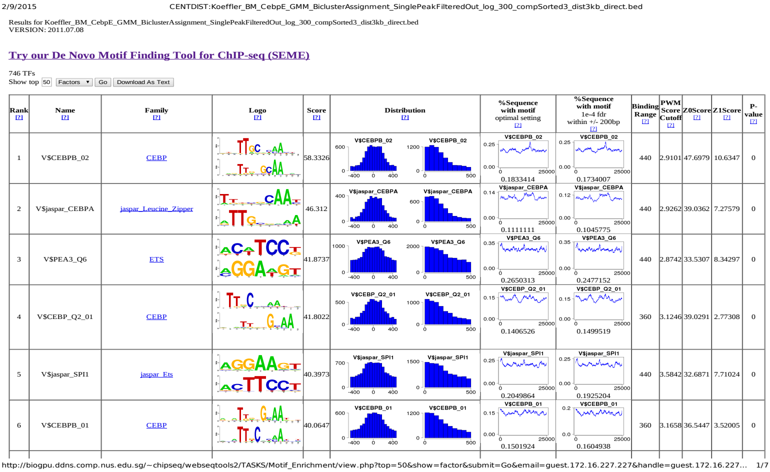
3 Component-Mixture Model



Local Clustering



Direct: 24948 peaks



2/9/2015 CENTDIST:Koeffler_BM_CebpE_GMM_BiclusterAssignment_SinglePeakFilteredOut_log_300_compSorted3_dist3kb_direct.bed

Results for Koeffler_BM_CebpE_GMM_BiclusterAssignment_SinglePeakFilteredOut_log_300_compSorted3_dist3kb_direct.bed
VERSION: 2011.07.08

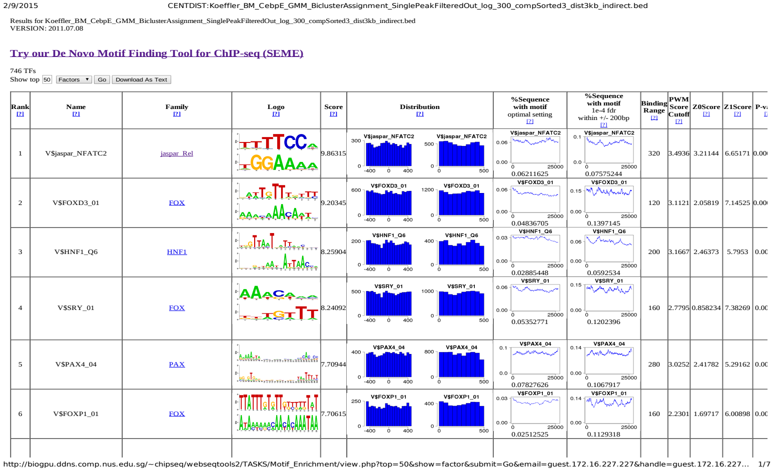
Try our De Novo Motif Finding Tool for ChIP-seq (SEME)

746 TFs
Show top 50 Factors ▾ 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$CEBPB_02	CEBP		58.3326	V$CEBPB_02 V$CEBPB_02	V$CEBPB_02 0.1833414	V$CEBPB_02 0.1734007	440	2.9101	47.6979	10.6347	0
2	V$jaspar_CEBPA	jaspar_Leucine_Zipper		46.312	V$jaspar_CEBPA V$jaspar_CEBPA	V$jaspar_CEBPA 0.1111111	V$jaspar_CEBPA 0.1045775	440	2.9262	39.0362	7.27579	0
3	V$PEA3_Q6	ETS		41.8737	V$PEA3_Q6 V$PEA3_Q6	V$PEA3_Q6 0.2650313	V$PEA3_Q6 0.2477152	440	2.8742	33.5307	8.34297	0
4	V$CEBP_Q2_01	CEBP		41.8022	V$CEBP_Q2_01 V$CEBP_Q2_01	V$CEBP_Q2_01 0.1406526	V$CEBP_Q2_01 0.1499519	360	3.1246	39.0291	2.77308	0
5	V$jaspar_SPI1	jaspar_Ets		40.3973	V$jaspar_SPI1 V$jaspar_SPI1	V$jaspar_SPI1 0.2049864	V$jaspar_SPI1 0.1925204	440	3.5842	32.6871	7.71024	0
6	V$CEBPB_01	CEBP		40.0647	V$CEBPB_01 V$CEBPB_01	V$CEBPB_01 0.1501924	V$CEBPB_01 0.1604938	360	3.1658	36.5447	3.52005	0

http://biogpu.ddns.comp.nus.edu.sg/~chipseq/webseqtools2/TASKS/Motif_Enrichment/view.php?top=50&show=factor&submit=Go&email=guest.172.16.227.227&handle=guest.172.16.227... 1/7

Indirect: 26547 peaks



2/9/2015 CENTDIST:Koeffler_BM_CebpE_GMM_BiclusterAssignment_SinglePeakFilteredOut_log_300_compSorted3_dist3kb_indirect.bed

Results for Koeffler_BM_CebpE_GMM_BiclusterAssignment_SinglePeakFilteredOut_log_300_compSorted3_dist3kb_indirect.bed
VERSION: 2011.07.08

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

746 TFs
Show top 50 Factors ▾ 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-va
1	V$jaspar_NFATC2	jaspar_Rel		9.86315		0.06211625	0.07575244	320	3.4936	3.21144	6.65171	0.00
2	V$FOXD3_01	FOX		9.20345		0.04836705	0.1397145	120	3.1121	2.05819	7.14525	0.00
3	V$HNF1_Q6	HNF1		8.25904		0.02885448	0.0592534	200	3.1667	2.46373	5.7953	0.00
4	V$SRY_01	FOX		8.24092		0.05352771	0.1202396	160	2.7795	0.858234	7.38269	0.00
5	V$PAX4_04	PAX		7.70944		0.07827626	0.1067917	280	3.0252	2.41782	5.29162	0.00
6	V$FOXP1_01	FOX		7.70615		0.02512525	0.1129318	160	2.2301	1.69717	6.00898	0.00

http://biogpu.ddns.comp.nus.edu.sg/~chipseq/webseqtools2/TASKS/Motif_Enrichment/view.php?top=50&show=factor&submit=Go&email=guest.172.16.227.227&handle=guest.172.16.227... 1/7

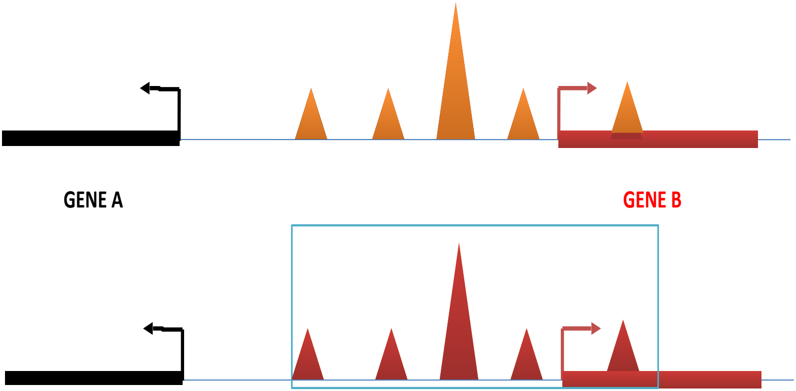
- Our current method could separate direct and indirect bindings

- Our current method could separate direct and indirect bindings
- Next, can we further using peak clusters increase functional annotation?

- Our current method could separate direct and indirect bindings
- Next, can we further using peak clusters increase functional annotation?
- Working on DNA methylation review on region to single base resolution DNA methylation research

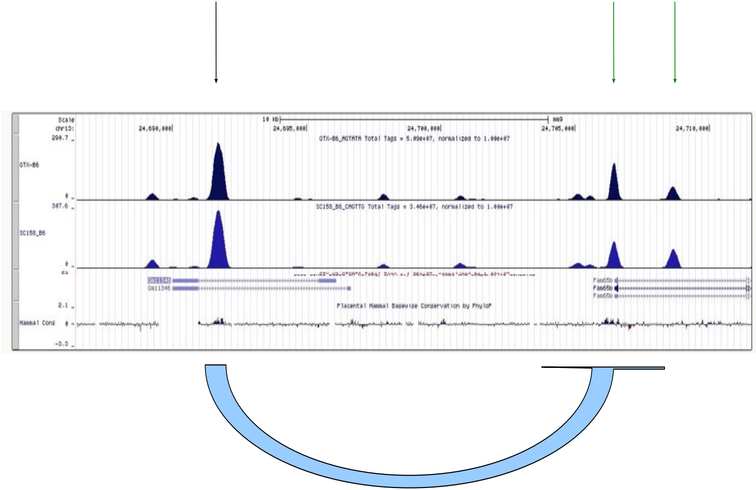
- Our current method could separate direct and indirect bindings
- Next, can we further using peak clusters increase functional annotation?
- Working on DNA methylation review on region to single base resolution DNA methylation research
- TCGA methylation on 19 cancer patients

Find the targeted genes



Peaks Cluster

What problems the invention solves and advantages over existing methods?
An Example:


Scale
chr13:
10 kb
mm9
GTX-B6
GTX-B6_AGTATA Total Tags = 5.09e+07, normalized to 1.00e+07
SC158_B6
SC158_B6_CAGTTG Total Tags = 3.46e+07, normalized to 1.00e+07
Gm11346
Gm11346
Fam65b
Fam65b
Fam65b
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons