Chromatin Conformation Prediction from ChIPseq

Ricky Lim¹, Samuel Collombet², Agus Salim³, Touati Benoukraf¹

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

CSI-Meeting mailto:rlim.email@gmail.com

Contents

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

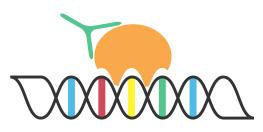
 Main Question: Can we use transcription factor (TF)-ChIPseq to predict protein complexes (direct and indirect bindings) on chromatin?

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.

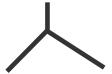
What is ChIPseq?

ChIP-Seq



Chromatin ImmunoPrecipitation

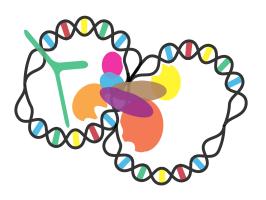




Sequencing ATCGTTAACGCATTAGCAGT...



Chromatin Conformation



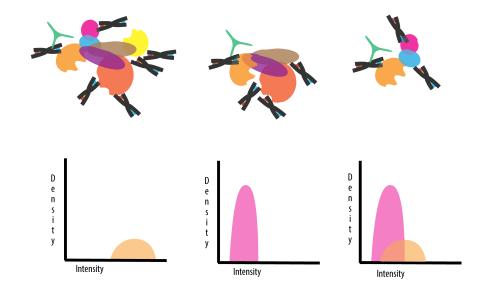
Direct binding sites



Indirect binding sites



Mixture of Chromatin Conformations



What is Mixture Model (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:

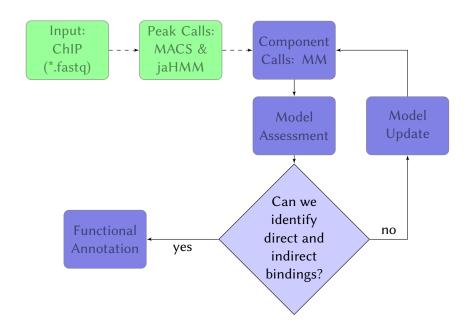
- Hard clustering: non-overlapping clusters
- Soft clustering: overlapping clusters

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.



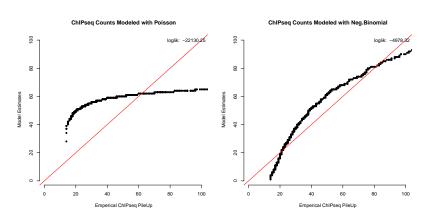
Input: ChIP-seq of Cebp ϵ from Koeffler-BM

##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

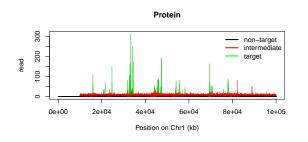
Principles

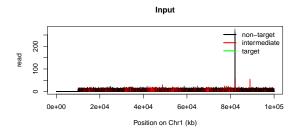
- MACS2: poisson model-based analysis of Peak calls MACS reference
- **jaHMM**: *negative binomial* model-based analysis of Peak calls jaHMM reference

Why jaHMM is better?

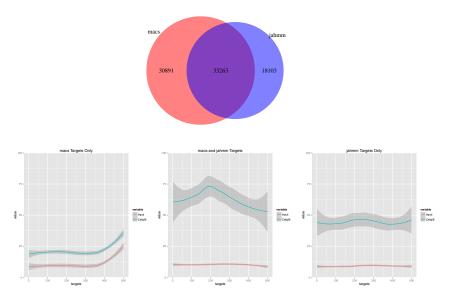


Targets Identified by jahmm





Targets Identified by MACS2 vs jahmm



Why jaHMM is better than MACS2?

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

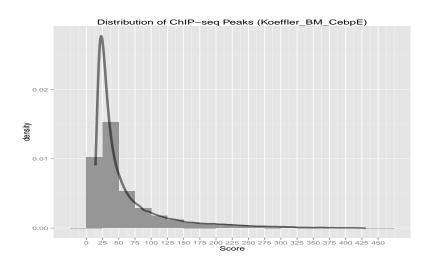
Why jaHMM is better than MACS2?

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

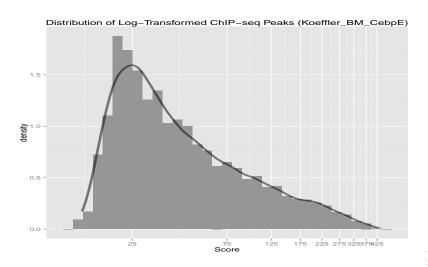
Pipeline Summary: Peak Calls Summary: Component Calls Summary: Motif Calls

Can we model ChIPseq Peaks using components of MMs?

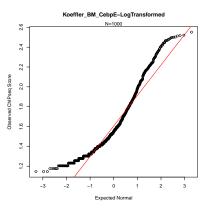
Input: ChIP-seq of Cebp ϵ from Koeffler-BM

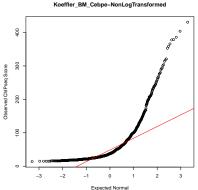


Log Transformation of ChIP-seq Input

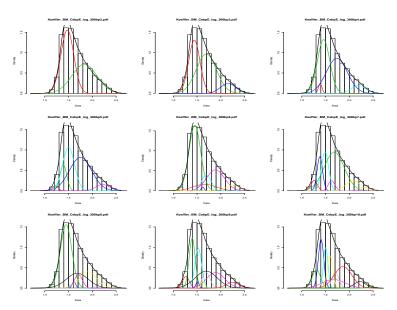


Check the Normality

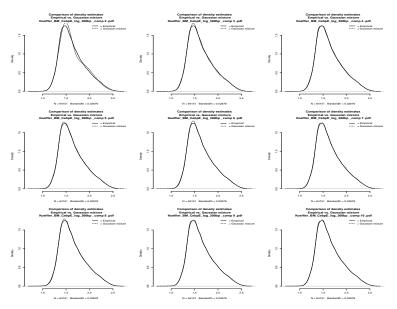




ComponentCalls: Fit ChIPseq Peaks with GMMs



GMM-ModelAssessment: Overfit¹



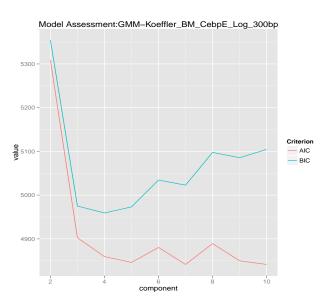
Model Assessment: BIC-AIC

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

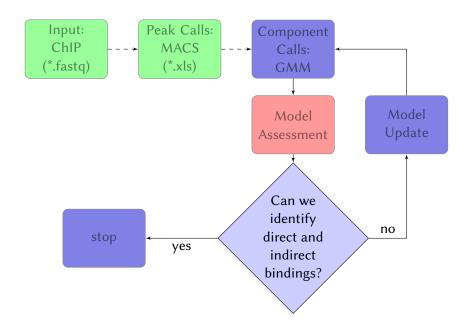
 Can we model ChIPseq using several components of MMs?

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.



Pipeline Summary: Peak Calls Summary: Component Calls Summary: Motif Calls

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 Ax Text

Rank 121	Name III	Family 121	Logo 121	Score 121	Distribution [2]	%Sequence with motif optimal setting	within +/- 200bp	Binding Range	PWM Score Cutoff	Z0Score	Z1Score
1	V\$jaspar_MZF1_1_4	jaspar_BetaBetaAlpha_zinc_finger	GGGGA TCCCC	12.2743	1400 VSjamper_MZF(_1_4 VSjamper_MZF(_1_4	0.4 MANAYAMAN 0.0 0.3096857	0.2864102		2.7671	6.19578	6.07853
2	VSjaspar_SP1	jaspar BetaBetaAlpha zinc finger	CCCc CCccc	11.5458	V\$jespar_SP1 V\$jespar_SP1	0.5 VS maper_SP1	0.3048479		3.0083	8.28603	3.25976
3	V\$SP1_01	SP1	GGc	11.3454	VSSPI_91 1500 VSSPI_91 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.25 VSSP1_01 0.25 VSSP1_01 0.25 0.25 0.0000	0.0 0.1500389		2.7192	8.56304	2.78238
4	V\$SP1_Q2_01	SP1	-cc-CCc-	9.69061	VSSP1_O2_01 VSSP1_O2_01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.4 M. Junior 0.0 0.2746372	0.0 0.2415846 00000		3.2844	7.55059	2.14002
5	VSMAZR_01	SP1	CCCC-CC-	9.67933	VSMAZR_01 VSMAZR_01	0.0 0.2536283 20000	0.2355628		2.9471	5.14373	4.5356
6	V\$MUSCLE_INI_B	MINI	Le. 2012 . T. 1000000000000000000000000000000000	9.64468	YSMUSCLE_INLB 1500 1500 YSMUSCLE_INLB 1500 0 0 500	0.25 MUSCLE INI B 0.25 Muscle INI B 0.25 0.25 20000 0.1862715	0.0 0.1611354		2.8998	7.04083	2.60384

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

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

Show top 50 Factors * Go Download As Text

Rank 121	Name [2]	Family I21	Logo D	Score [2]	Distribution [21	%Sequence with motif optimal setting	within +/- 200bp	Binding Range	PWM Score Cutoff	Z0Score	Z1Score	P- value
1	V\$CEBPB_02	CERP	TI- 90AA	36.4541	900 V9CEBPB_02 1200 V9CEBPB_02	0.25 VSCESPB_02 0.25 VSCESPB_02 0.00 0.1615542	0,1398565		2.9101	33.7596	2.69456	0
2	V\$CEBP_Q2_01	CERP	TT_C	30.0046	900 V\$CEBP_02_01 V\$CEBP_02_01		0.15 VSCEBP 02 01		3.1246	27.234	2.7706	0
3	VSjaspar_CEBPA	jaspæ Leucine Zipper	T- CAA-	29.099	700 V\$ saper_CEBPA 700 V\$ saper_	0.00	0.1 V\$jampar_CEBPA 0.1 0.00 0.00283942		2.9262	26.4199	2.67911	0
4	V\$CEBP_Q2	CERP	· TTGA	27.1049	VICEBP_02 000 VICEBP_02	0.15 V\$CEBP_02 0.00 0 3000 0.1106028	0.12 VSCEBP_02 0.12 0.00 0.09573325		2.9207	23.2332	3.87169	0
5	V\$CEBPA_01	CERP		27.0551	V9CEBPA_01 1200 V9CEBPA_01		0.15 VSCEBPA 01		2.8306	24.272	2.78314	0
6	V\$ETS_Q4	ETS	- CTTCCT-	26.0123	900 VSETS_Q4 1500 VSETS_Q4	0.25 WSETS_Q4	VSETS_G4 0.2\(\sqrt{\sq}}}}\sqrt{\sq}}}}}}}}\sqrt{\sqrt{\sqrt{\sq}}}}}}}}}}}}}}}}}}}}}}}}}}}}}}}}}}}}		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)

2/9/2015

CENTDIST:Koeffler_BM_CebpE_GMM_ModelAssignment_log_300_group3_compSorted3.bed

Results for Koeffler_BM_CebpE_GMM_ModelAssignment_log_300_group3_compSorted3.bed

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

Show top 50 Factors * Go Download As Text

Rank [2]	Name (2)	Family I21	Logo D	Score [2]	Distribution [2]	%Sequence with motif optimal setting	within +/- 200bp	Binding	PWM Score Cutoff	Z0Score	Z1Score	P- value
1	V\$CEBPB_02	CERP	TI- 90AA	32.9624	V9CEBPB_02 V9CEBPB_02	900 02 000 000 000 000 000 000 000 000 0	0.25 V9CEBPB_02	320	2.9101	29.8398	3.12262	0
2	V\$CEBP_Q2_01	CERP	TT_C	28.6415	VSCEEP_Q2_01 VSCEEP_Q2_01	0.15 0.00 0.1684376	0.2 VSCEBP 02 01	360	3.1246	24.4458	4.19579	0
3	VSPEA3_Q6	<u>ets</u>	AÇATCC± aGGAAG±	28.3666	VSPEA3_G6 VSPEA3_G6	V\$PEA3_G6 0.4 0.0 0.3097021	0.4 V\$PEA3_G6	440	2.8742	21.9021	6.46444	0
4	V\$jaspar_CEBPA	iespar Leucine Zipper	T. CAA.	27.798	V\$1 seper_CEBPA	0.15 CEBPA 0.00 0 5000 0.1301167	0.15 0.00 0 5000 0.1381292	360	2.9262	24.5191	3.2789	0
5	V\$CEBPB_01	CERP	TG_AA	27.6113	V9CEBPE_01 V9CEBPE_01	0.2 0.0 0.1863787	0.2 0.0 0.1975266	360	3.1659	23.722	3.88938	0
6	V\$CEBPA_01	CEBP	TT	26.0984	VSCEBPA_01 SSO VSCEBPA_01	0.0 0.1750566	0.2 0.0 0.1865529	360	2.8312	21.6892	4.40913	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

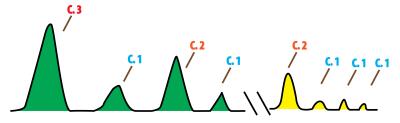
Pipeline Summary: Peak Calls Summary: Component Calls Summary: Motif Calls

 Cebp motif is found in group 2 and 3 in 3-component GMMS using centdist

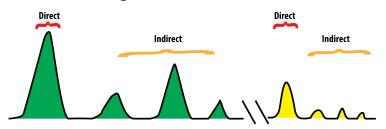
Pipeline Summary: Peak Calls Summary: Component Calls Summary: Motif Calls

- Cebp motif is found in group 2 and 3 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

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

Show top 50 Factors • Go Download As Text

Rank 121	Name 121	Family IZI	Logo (2)	Score III	Distribution (2)	%Sequence with motif optimal setting	% Sequence with motif 1e-4 fdr within +/- 2000pp	Binding Range	PWM Score Cutoff	Z0Score	Z1Score	P. value
1	V\$CEBPB_02	CERP	TT- OCAA	58.3326	VSCERPE_02 VSCERPE_02	0.25 0.00 0.1833414	0.25 0.25 0.00 0.1734007		2.9101	47.6979	10.6347	0
2	V\$jaspar_CEBPA	jaspar Leucine Zipper	T- CAA-	46.312	600 VS(maper_CEBPA 600 VS)maper_CEBPA 600 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.00 0.1111111	0.1045775		2.9262	39.0362	7.27579	0
3	V\$PEA3_Q6	ETS	ACATCCE AGGAAGE	41.8737	1000 VSPEA3_OE 2000 VSPEA3_G6		0.2477152		2.8742	33.5307	8.34297	0
4	VSCEBP_Q2_01	CERP	TT_CAA	41.8022	V\$CEBP_G2_01 V\$CEBP_G2_01	0.15 V9CEBP 02 01 0.15 V3V3V3V3V3V3V3V3V3V3V3V3V3V3V3V3V3V3V3	0.09 0.1499519 25000		3.1246	39.0291	2.77308	0
5	VSjaspar_SPI1	jaspar Ets	AGGAAGT ACTTCCT	40.3973	7500 V\$jaspar_SPH 1500 V\$jaspar_SPH 1500 0 500	0.25 VS[aspar_SPH 0.25 VS 25000 0.2049864	0.25 VS[mpar_SPH 0.25 0.1925204		3.5842	32.6871	7.71024	0
6	V\$CEBPB_01	СЕВР	L.Tx.G.AA	40.0647	000 VSCEBPB_01 1300 VSCEBPB_01	0.00 0.1501924	0.2 V9CEBPB_01		3.1658	36.5447	3.52005	0
												г

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_BickasterAssignment_SinglePeakFilteredOut_log_300_compSorted3_dist3kb_indirect.bed VERSION: 2011.07.08

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

746 TPs
Show top 50 Factors • Go Download Ax Text

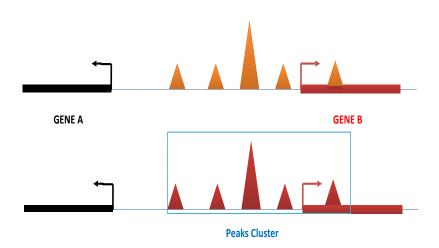
Rank (21	Name [21	Family (2)	Logo Di	Score IZI	Distribution III	%Sequence with motif optimal setting	within +/- 200bp	Binding Range	PWM Score Cutoff	Z0Score	Z1Score	P-vi
1	V\$jaspar_NFATC2	jaspar_Rel	TTTCC.	9.86315	2003 VS(imper_NFATC2 VS(imper_NFATC2 2003 400 0 0 0 0 500	VSjaspar NFATC2 0.09 0.00 0 25000 0.06211625	0.1 VSjaspar NFATC2 0.0 0 25000 0.07575244	320	3.4936	3.21144	6.65171	0.001
2	V\$FOXD3_01	FOX	LAXISTTTATE	9.20345	0 460 0 460 0 0 560	0.00 0 25000	0.18 VSFOXDS_01 0.18 VALANA 0.00 0.1397145	120	3.1121	2.05819	7.14525	0.001
3	V\$HNF1_Q6	HNE1	Luction III	8.25904	VSHNF1_O6 400 VSHNF1_O6	0.03 0.03 0.03 0.03 0.03 0.03 0.03 0.03	0.09 V9HNF1_G6 0.09 0.0592534	200	3.1667	2.46373	5.7953	0.00
4	V\$SRY_01	FOX	AAACA LETTT	8.24092	0 -60 0 400 0 0 500	0.00 0.05352771	0.15 VSSNY_01 0.00 0.1202396	160	2.7795	0.858234	7.38269	0.00
5	VSPAX4_04	PAX	bether	7.70944	VSPAX4_04 600 VSPAX4_04	0.0 0.07827626	0.14 V\$PAX4_04 0.00 0 0 25000 0.1067917	280	3.0252	2.41782	5.29162	0.00
6	VSFOXP1_01	FOX	*MATTISTSTEAMAN *AutomobileCAMAN	7.70615	250 VSFOXP1_01 400 VSFOXP1_01	0.00 0.02512525	0.14	160	2.2301	1.69717	6.00898	0.00

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

· 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?

Find the targeted genes



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

