

Chromatin Conformation Prediction from ChIPseq Signal



Ricky Lim

Cancer Science Institute of Singapore - NUS
Touati Benoukraf's Lab

CSI-Meeting
<mailto:csilr@nus.edu.sg>

Contents

① Introduction

Goal

What is ChIPseq?

What is Mixture Model?

② Our Strategy: Results

Pipeline

Summary: Peak Calls

Summary: Component Calls

Summary: Motif Calls

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

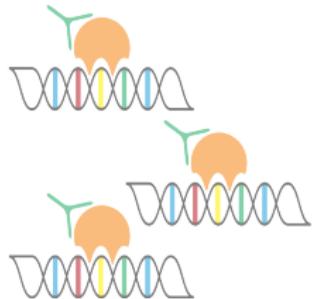
- **Main Question:** Can we use transcription factor (TF)-ChIPseq to predict protein complexes (direct and indirect bindings) on chromatin?
- **Strategy:** Model ChIPseq signal using Mixture Models to cluster the direct and indirect bindings.

What is ChIPseq?

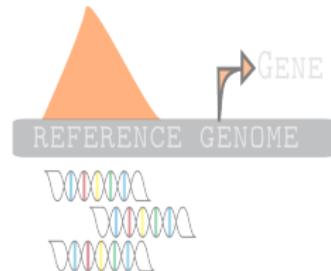
CHIP-SEQ



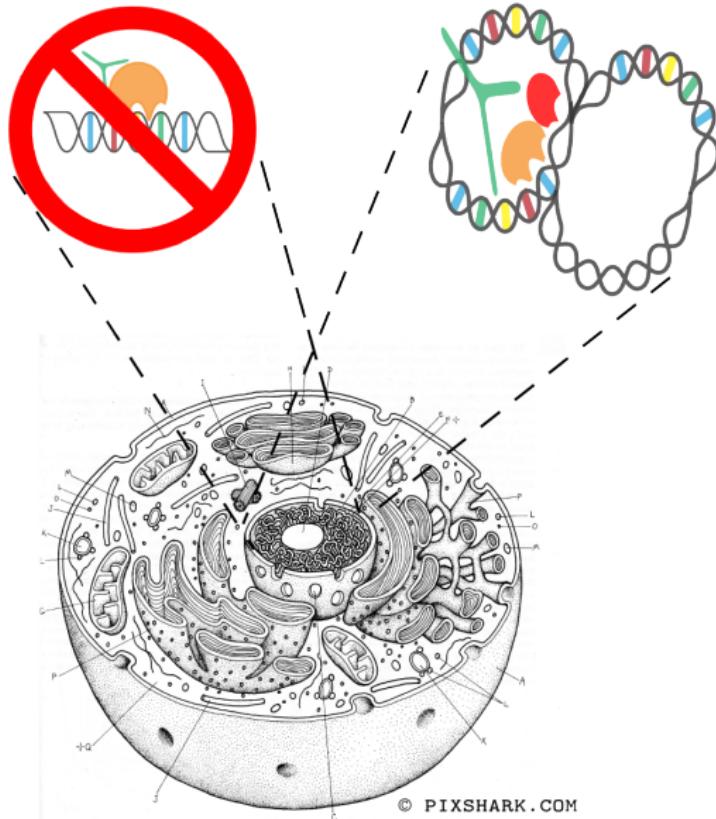
CHROMATIN
IMMUNOPRECIPITATION



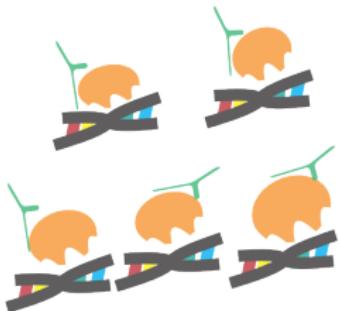
SEQUENCING
CAGTTAACGCTAAGGCCA



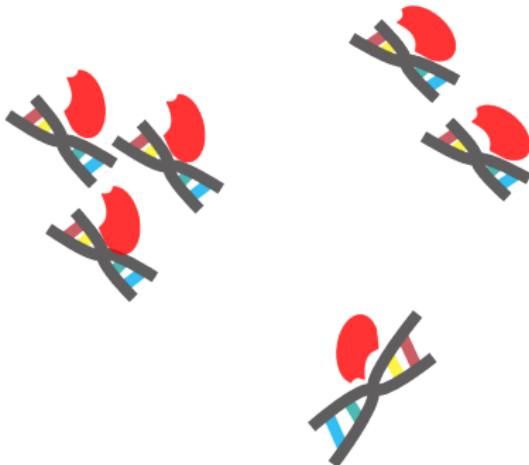
CHROMATIN CONFORMATION



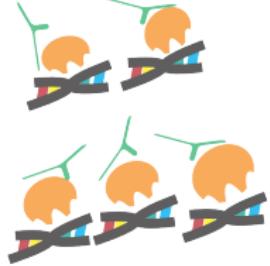
DIRECT BINDING SITES



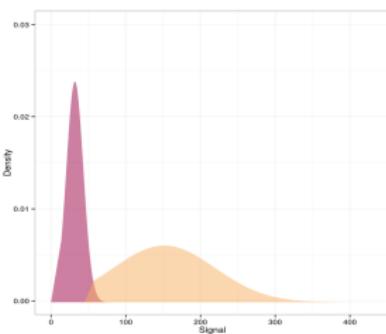
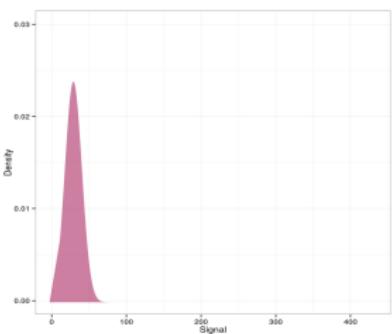
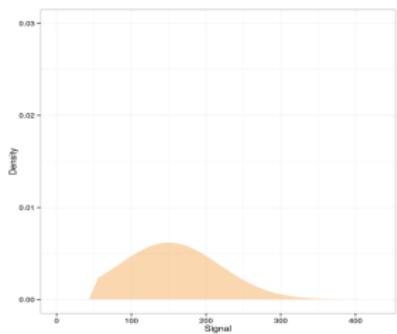
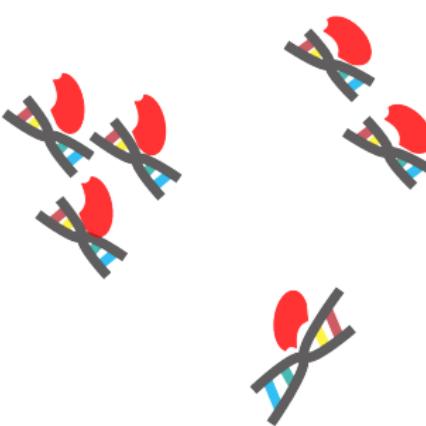
INDIRECT BINDING SITES



DIRECT BINDING SITES



INDIRECT BINDING SITES



What is Mixture Model (MM)?

Mixture Model: Revisited

Types of clustering methods:

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

Mixture Model: Revisited

Types of clustering methods:

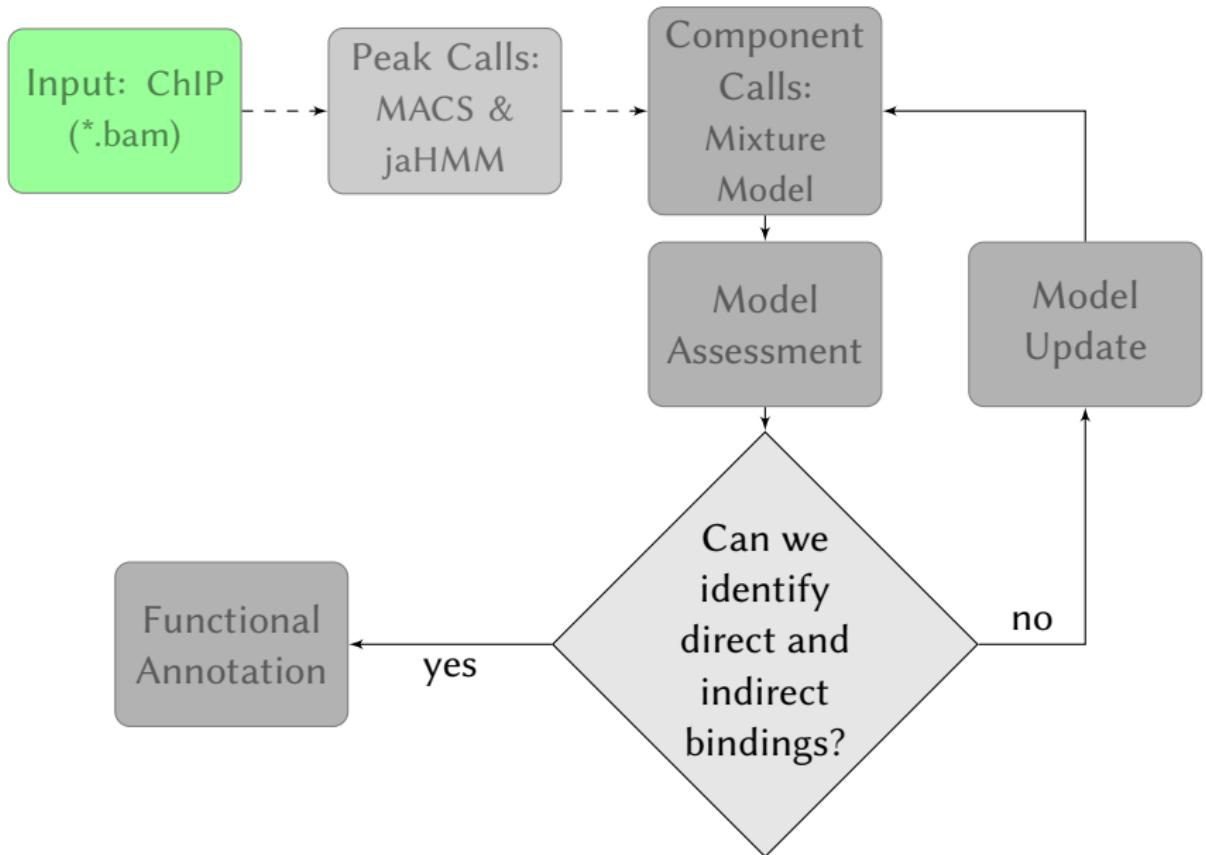
- Hard clustering: non-overlapping clusters
- Soft clustering: overlapping clusters

Mixture model is a probabilistic way of soft clustering. Each cluster is a generative mixture model with its parameters.

Gaussian Mixture Model

Key Assumption:

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



Input: ChIP-seq of Cebpe 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
```

Peak Calls: MACS2 vs jaHMM

¹Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

²Filion et al. jahmm: A tool for discretizing multiple ChIP seq profiles. arXiv (2014)

Peak Calls: MACS2 vs jaHMM

- **MACS2:** *poisson* model-based analysis of Peak calls ¹

¹Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

²Filion et al. jahmm: A tool for discretizing multiple ChIP seq profiles. arXiv (2014)

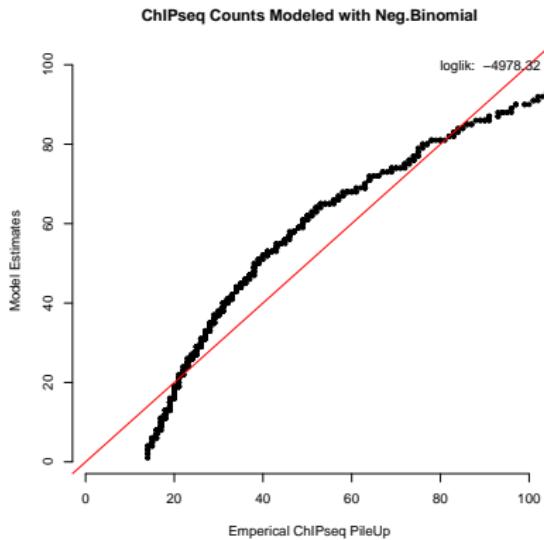
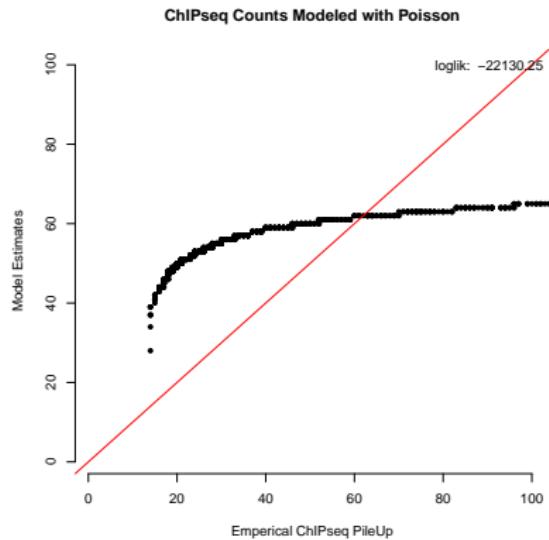
Peak Calls: MACS2 vs jaHMM

- **MACS2:** *poisson* model-based analysis of Peak calls ¹
- **jaHMM:** *negative binomial* model-based analysis of Peak calls ²

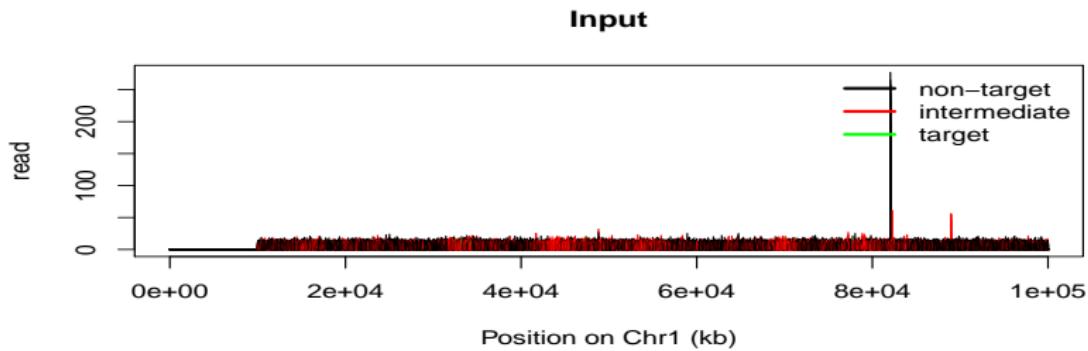
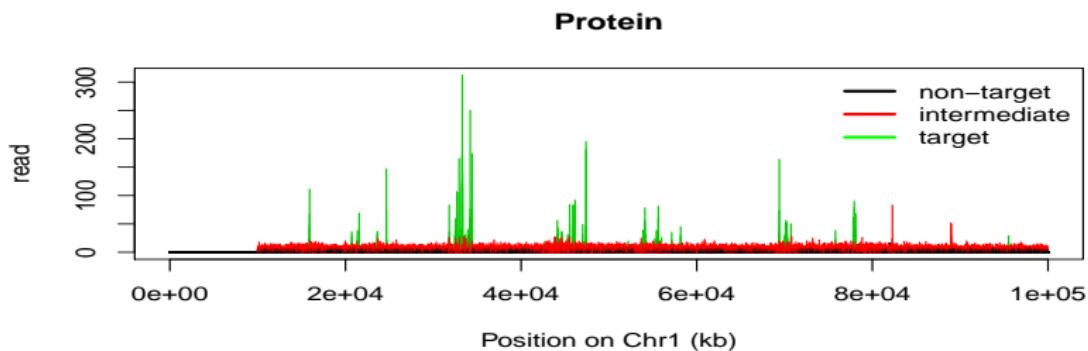
¹Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

²Filion et al. jahmm: A tool for discretizing multiple ChIP seq profiles. arXiv (2014)

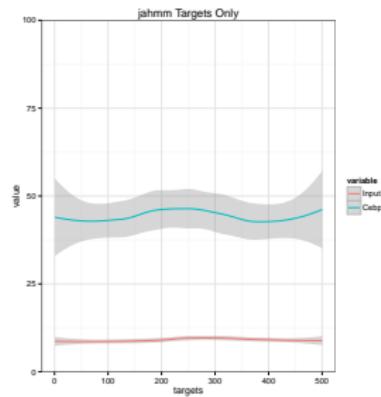
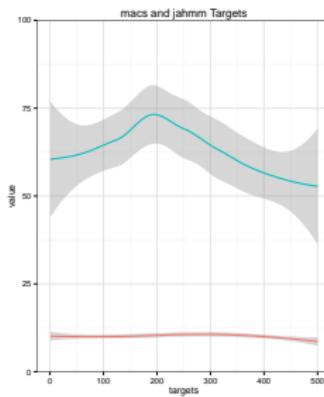
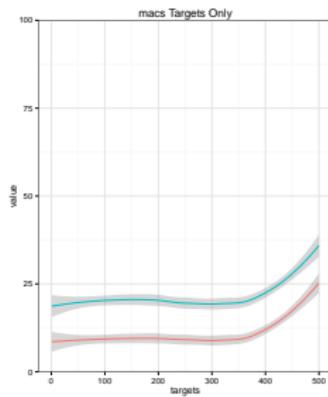
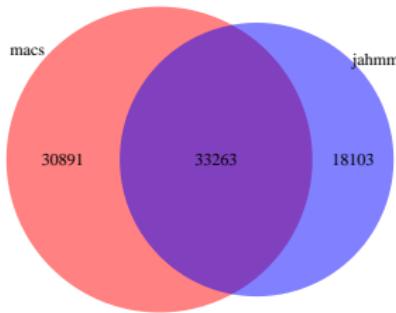
jaHMM fits ChIPseq Signals better than MACS2



Peaks Called by jahmm



Peaks Called by MACS2 vs jahmm



Summary: Peak Calls

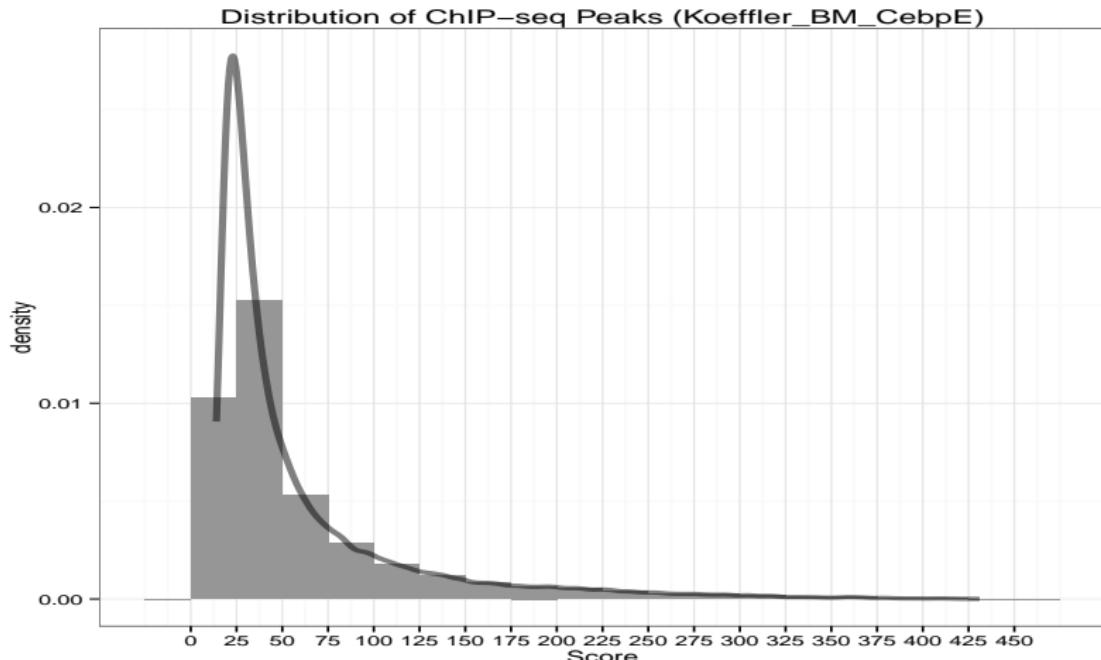
- Given our dataset, MACS2 is able to call peaks however, the estimated scores are less fit than JAHMM

Summary: Peak Calls

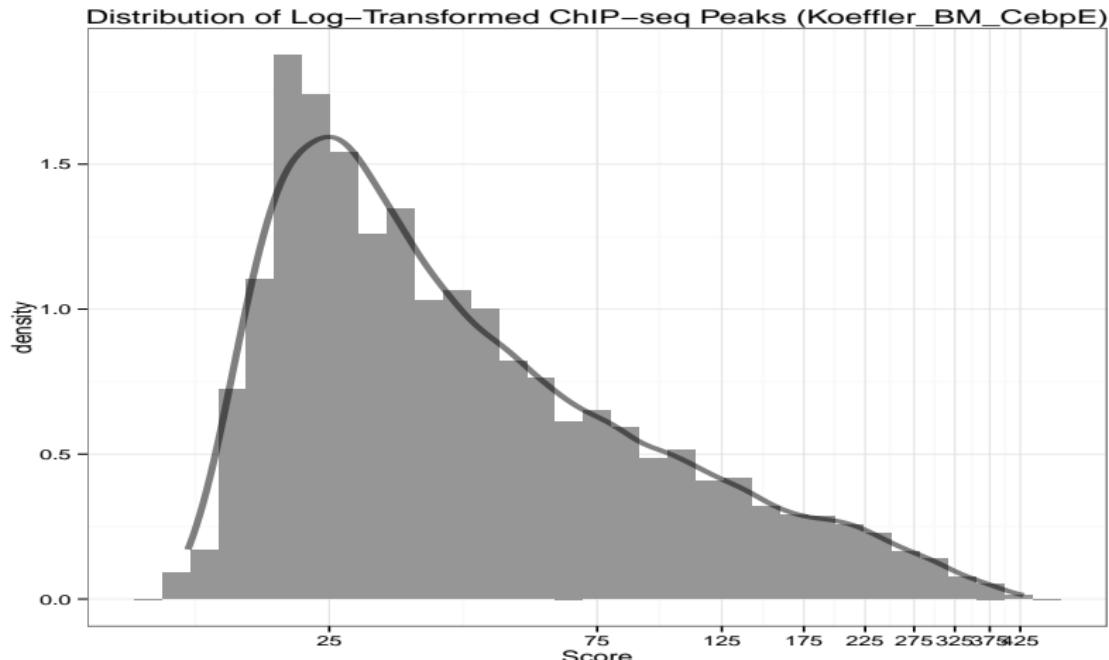
- Given our dataset, MACS2 is able to call peaks however, the estimated scores are less fit than JAHMM
- Peaks identified solely by jaHMM have scores higher with respect to their input (higher ratio) than solely by MACS2

Can we model ChIPseq using components of MMs?

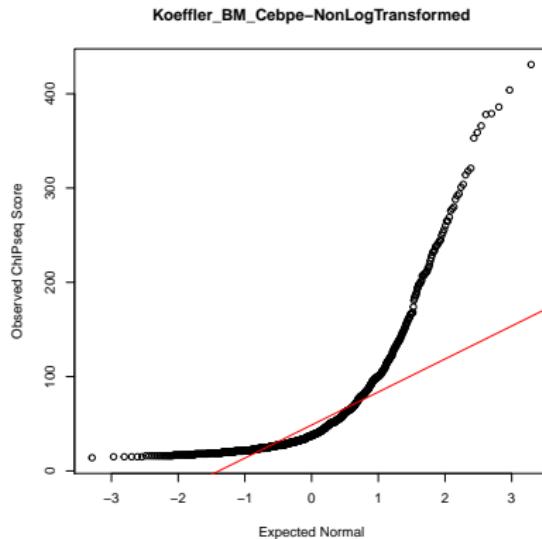
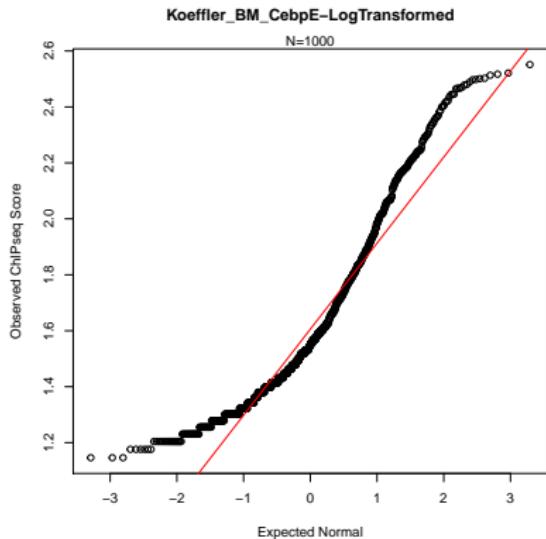
Input: ChIP-seq of Cebp ϵ from Koeffler-BM



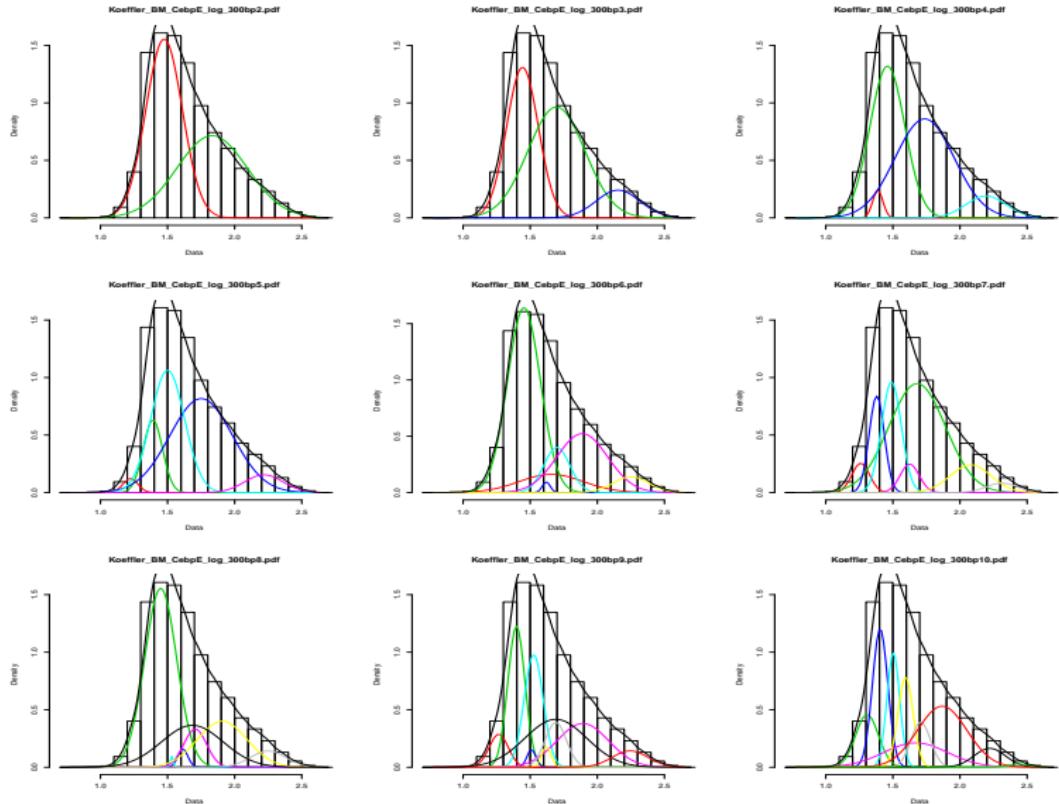
Log Transformation of ChIP-seq Input



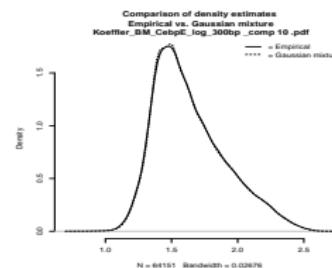
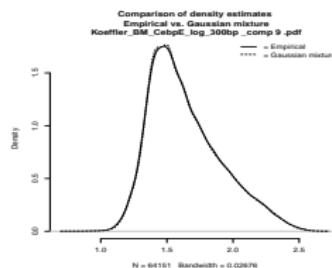
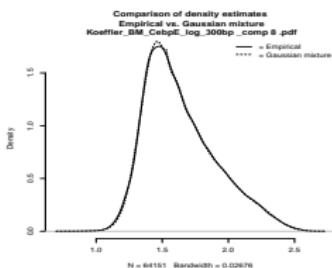
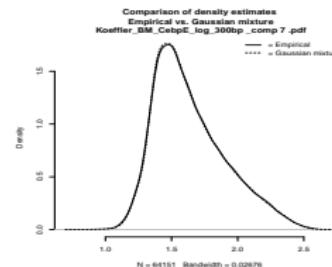
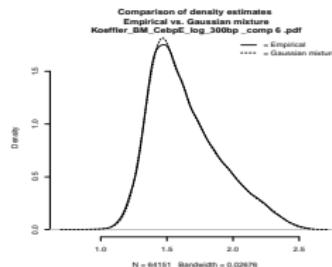
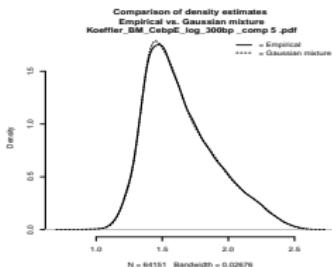
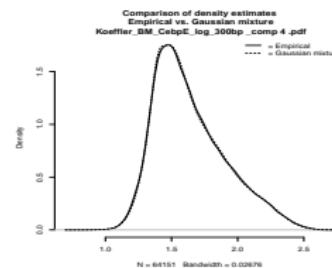
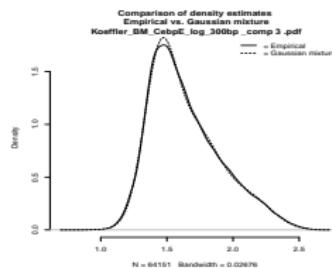
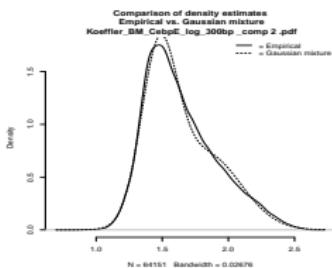
Check the Gaussian Normality



ComponentCalls: Fit ChIPseq Peaks with GMMs



GMM-ModelAssessment: Overfit



Model Assessment: BIC-AIC

AIC¹ and BIC² is based on Occam's razor principle, i.e, the simplest the better.

$$\text{AIC} = -2 \times \log L + 2 * P$$

$$\text{BIC} = -2 \times \log L + \log(n) * P$$

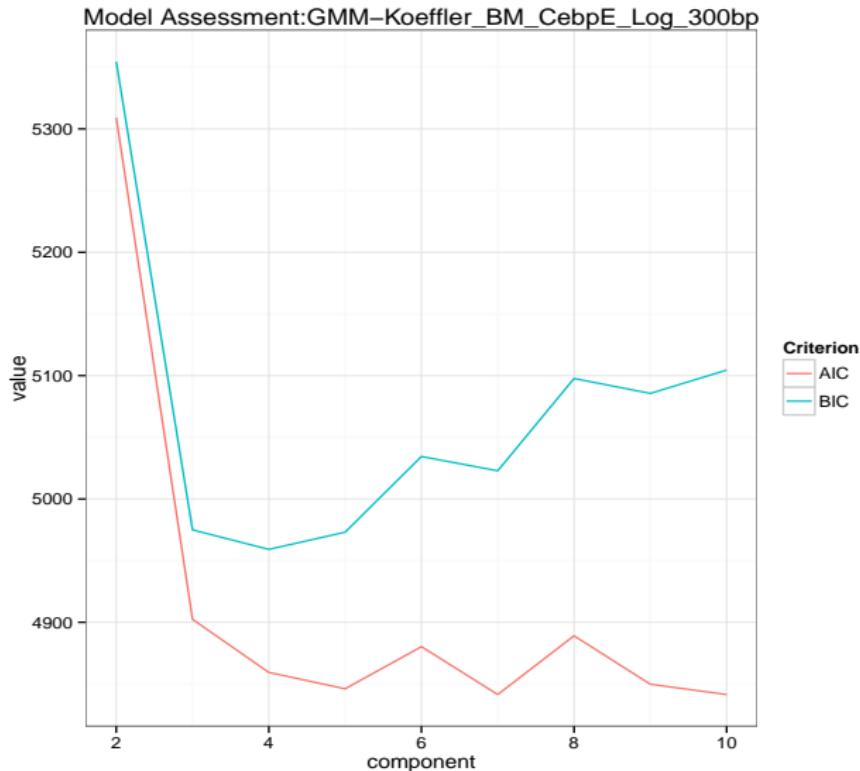
L is likelihood

P is the number of parameters

¹Akaike information criterion

²Bayesian information criterion

Model Assessment: BIC-AIC



Summary: Gaussian Mixture Models (GMMs)

- **Can we model ChIPseq using several components of MMs?**

Yes, our ChIPseq Peaks identified by jaHMM can be fit with GMMs.

Summary: Gaussian Mixture Models (GMMs)

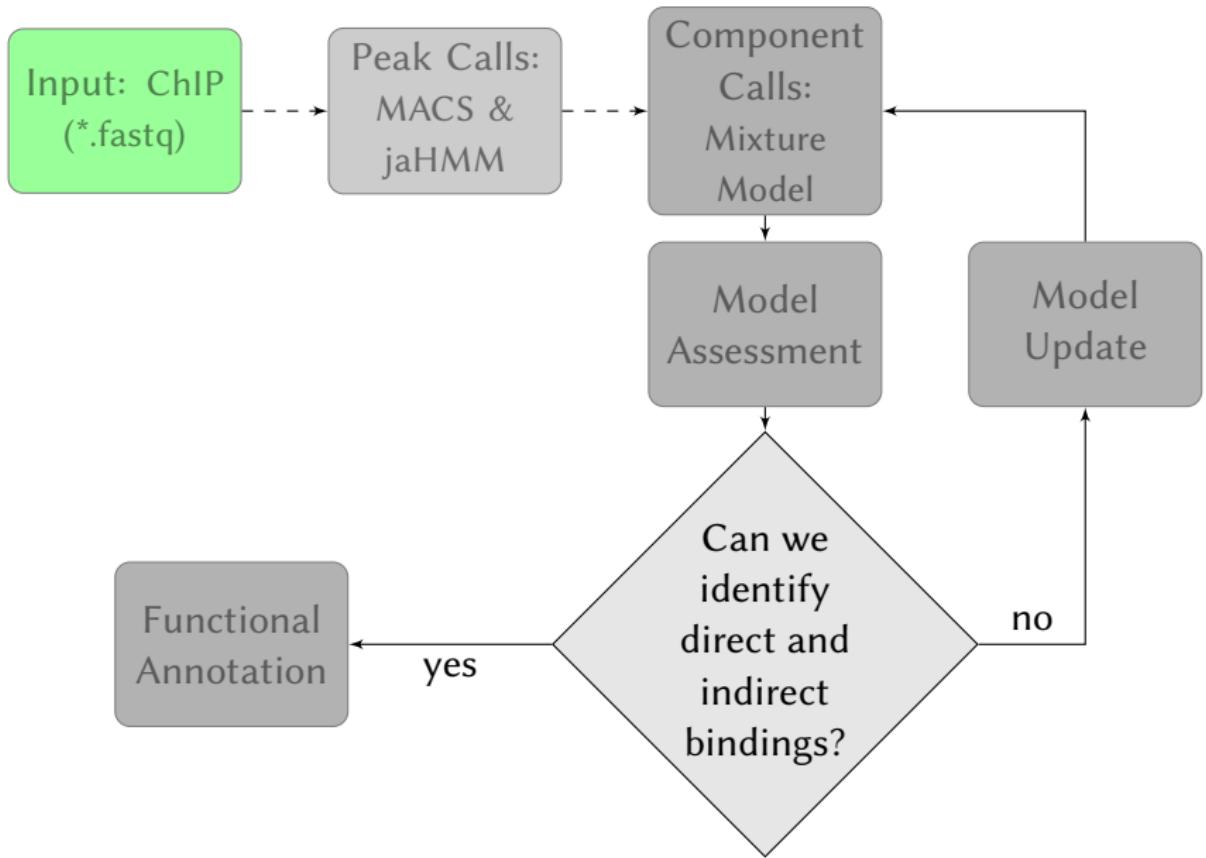
- **Can we model ChIPseq using several components of MMs?**

Yes, our ChIPseq Peaks identified by jaHMM can be fit with GMMs.

- **How many components are required?**

From AIC-BIC model assessment, 3 components are sufficient to fit ChIPseq signals.

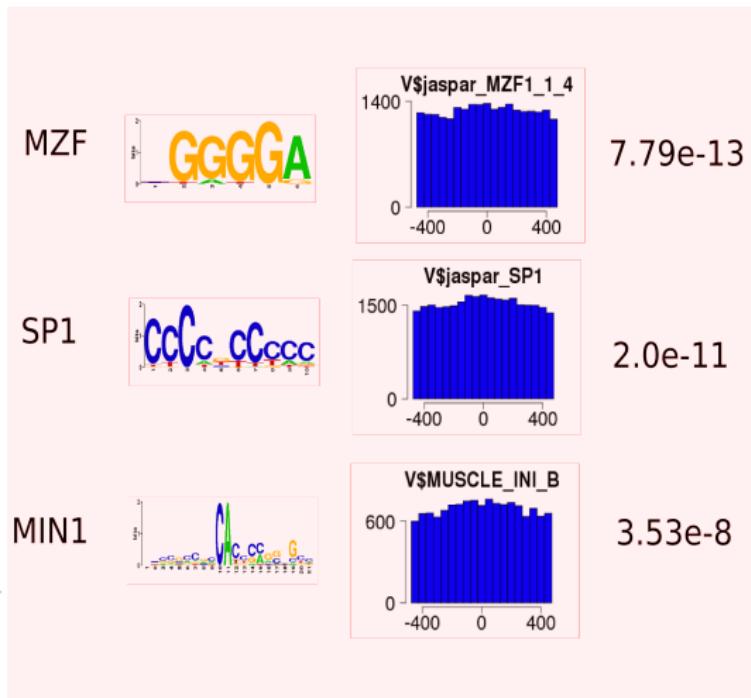
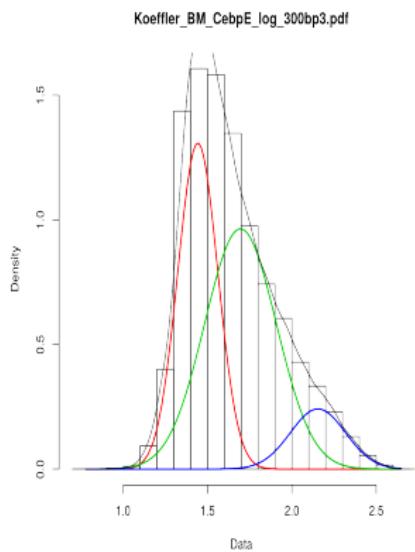
Note: the lower the AIC and BIC values, the better the fitting.



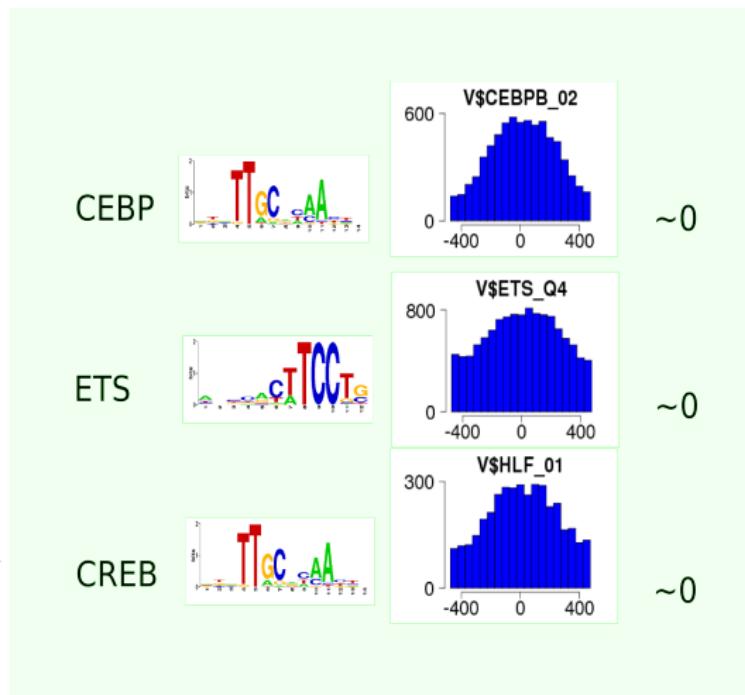
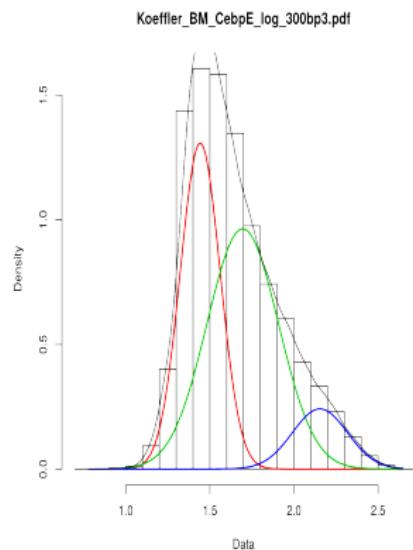
Motif Calls using Centdist ¹

¹Zhang et al. CENTDIST: discovery of co-associated factors by motif distribution. Nucleic Acids (2011)

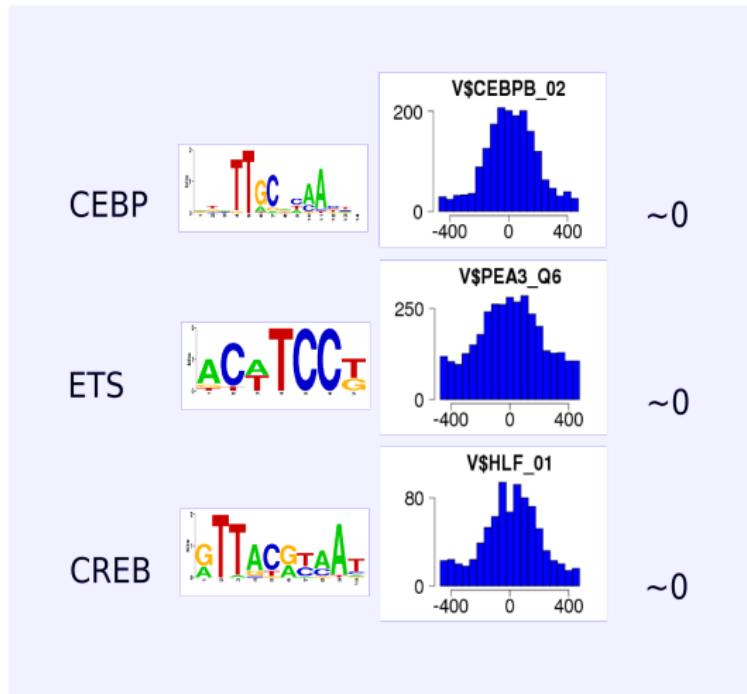
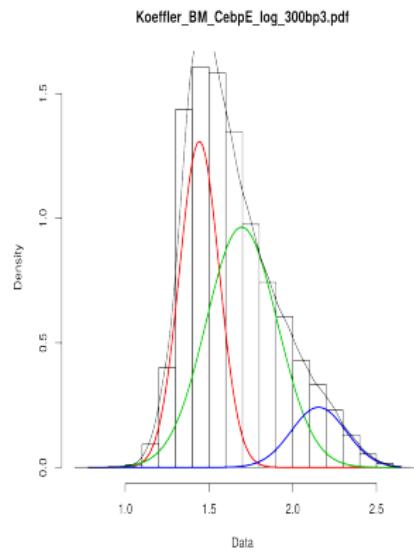
Component 1: low peak score (29559 peaks)



Component 2: intermediate peak score (28851 peaks)



Component 3: high peak score (5741 peaks)



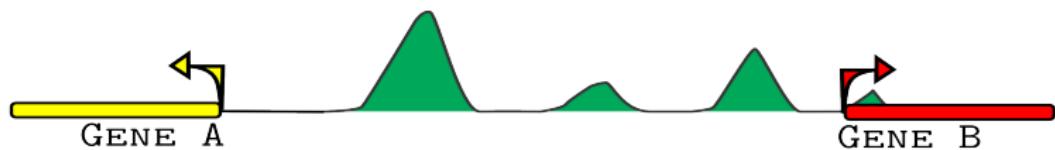
Summary: Motif Calls

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

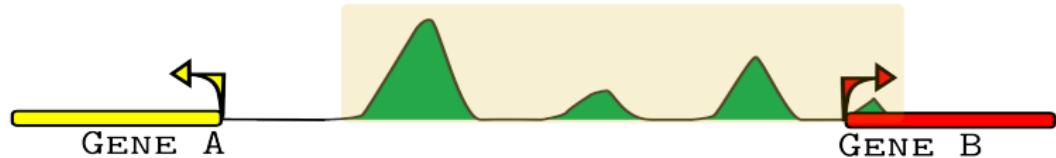
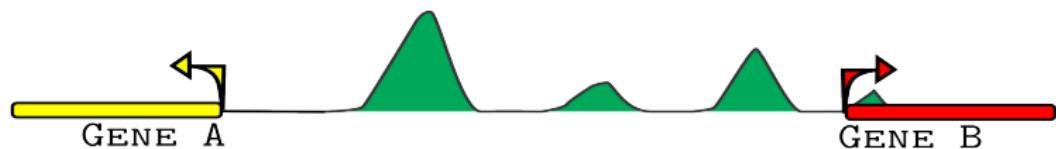
Summary: Motif Calls

- 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?

Peak Annotation

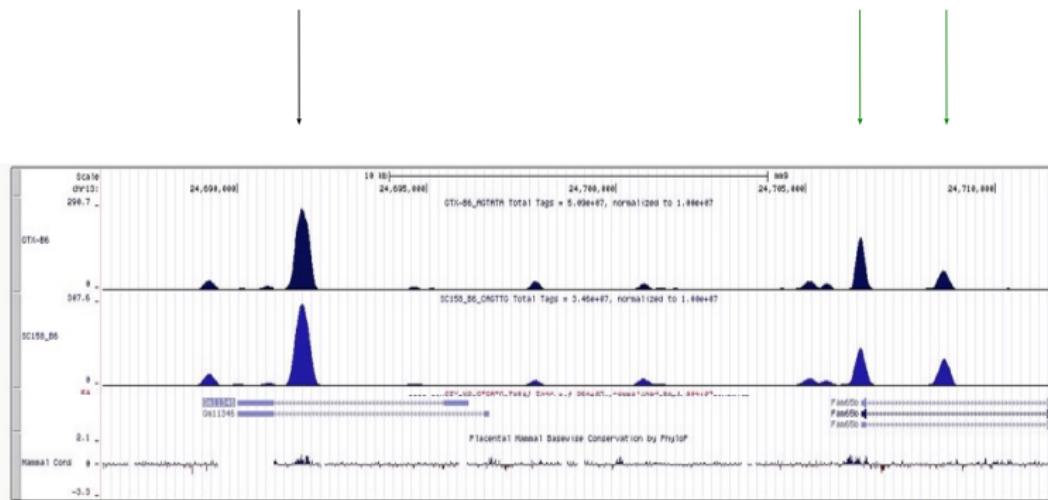


Peak Annotation: Clustering



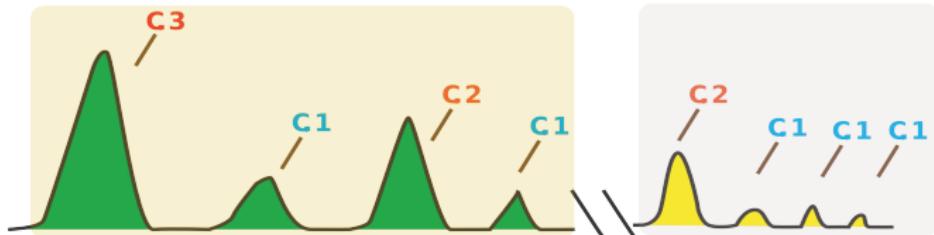
What problems the invention solves and advantages over existing methods?

An Example:

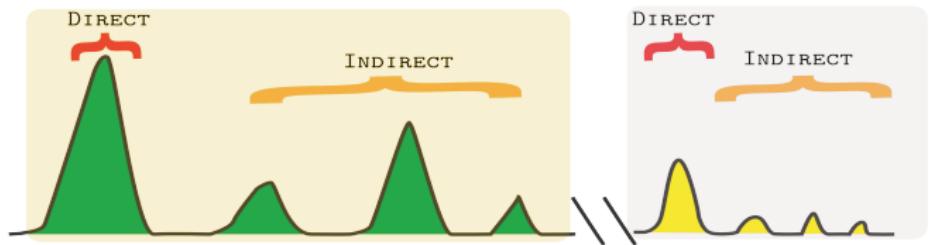


Peak Clustering

COMPONENT CLUSTERING

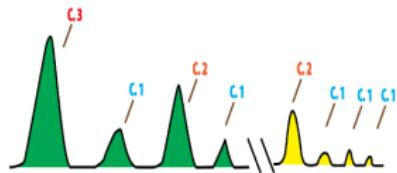


LOCAL CLUSTERING

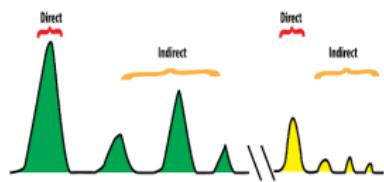


Direct: 24948 peaks

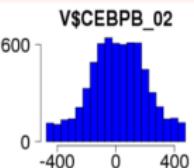
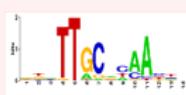
3 Component-Mixture Model



Local Clustering

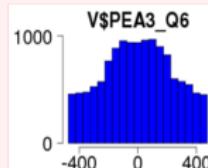


CEBP



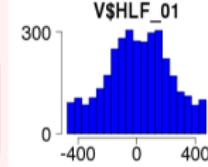
~0

ETS



~0

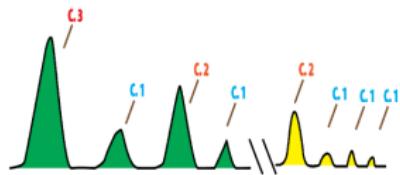
CREB



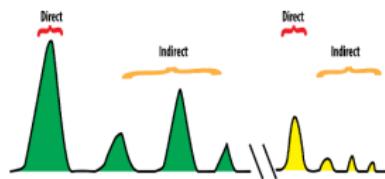
~0

Indirect: 26547 peaks

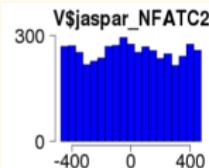
3 Component-Mixture Model



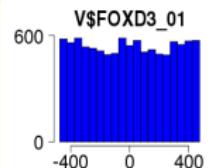
Local Clustering



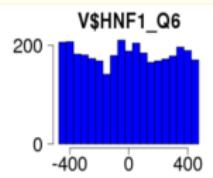
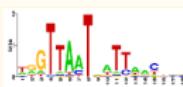
REL



FOX

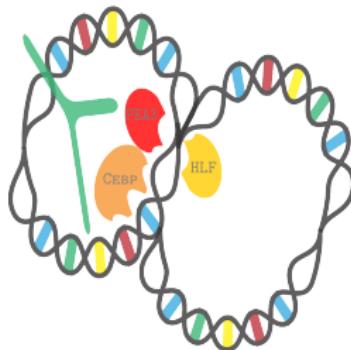


HNF1

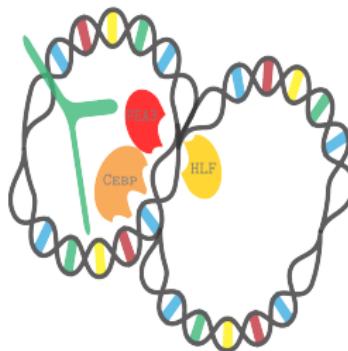


- Our current method could separate direct and indirect bindings

- Our current method could separate direct and indirect bindings
- Given Cebp ϵ ChIPseq, we could predict the chromatin conformation of Cebp-Pea3-HLF



- Our current method could separate direct and indirect bindings
- Given Cebp ϵ ChIPseq, we could predict the chromatin conformation of Cebp-Pea3-HLF



- Next, how this chromatin conformation functions?

Acknowledgement

- Touati Benoukraf (CSI-NUS)
- Samuel Collombet (Ecole Normale Superieur)
- Agus Salim (La Trobe University)
- Tong Yin (CEDARS SINAI Hospital)
- Phillip Koeffler (CSI-NUS)