

## *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?

## Inferring direct DNA binding from ChIP-seq

Timothy L. Bailey<sup>1,\*</sup> and Philip Machanick<sup>2</sup>

<sup>1</sup>Institute for Molecular Bioscience, The University of Queensland, Brisbane 4072, Queensland, Australia and

<sup>2</sup>Department of Computer Science, Rhodes University, Grahamstown 6140, South Africa

Received November 10, 2011; Revised April 2, 2012; Accepted April 23, 2012

# Chromatin Conformation Prediction

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

## Inferring direct DNA binding from ChIP-seq

Timothy L. Bailey<sup>1,\*</sup> and Philip Machanick<sup>2</sup>

<sup>1</sup>Institute for Molecular Bioscience, The University of Queensland, Brisbane 4072, Queensland, Australia and

<sup>2</sup>Department of Computer Science, Rhodes University, Grahamstown 6140, South Africa

Received November 10, 2011; Revised April 2, 2012; Accepted April 23, 2012

- **Strategy:** Model ChIPseq signal using Mixture Models to cluster the direct and indirect bindings.

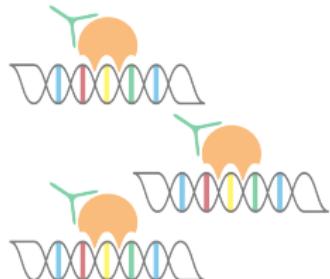
## What is ChIPseq?

# CHIP-SEQ



CHROMATIN  
IMMUNOPRECIPITATION

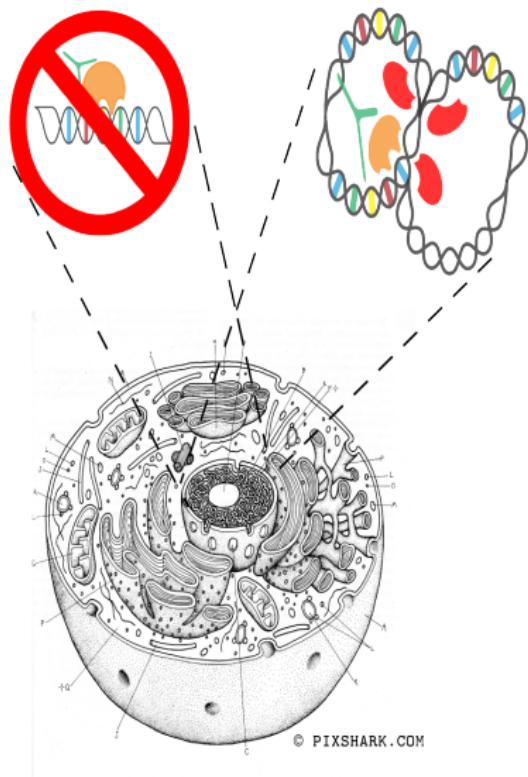
SEQUENCING  
**CAGTTACGGCTAAGGCCA**



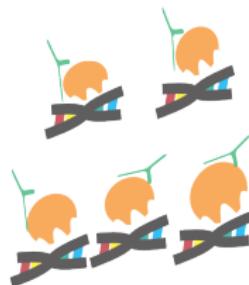
GENE  
REFERENCE GENOME

A diagram illustrating the reference genome step of CHIP-SEQ. An orange triangle represents a gene, which is a segment of the reference genome. The reference genome is shown as a collection of smaller DNA molecules with various colored segments (blue, red, yellow, green), representing the full set of chromosomes used as a reference for sequencing.

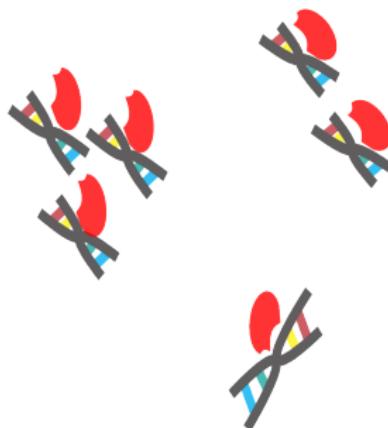
# CHROMATIN CONFORMATION



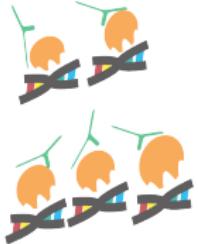
DIRECT BINDING SITES



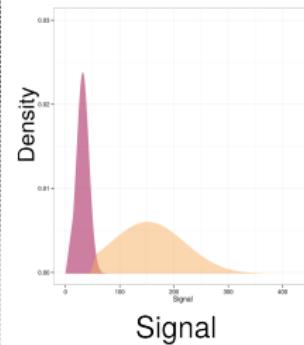
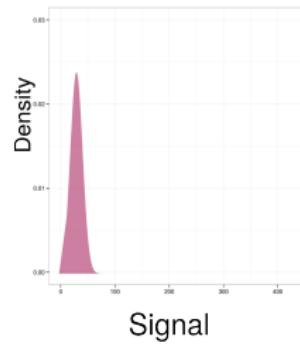
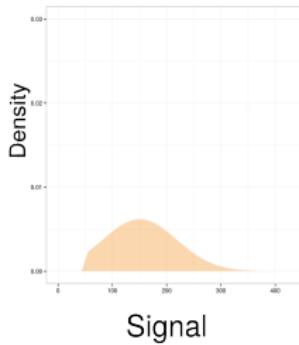
INDIRECT BINDING SITES



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:

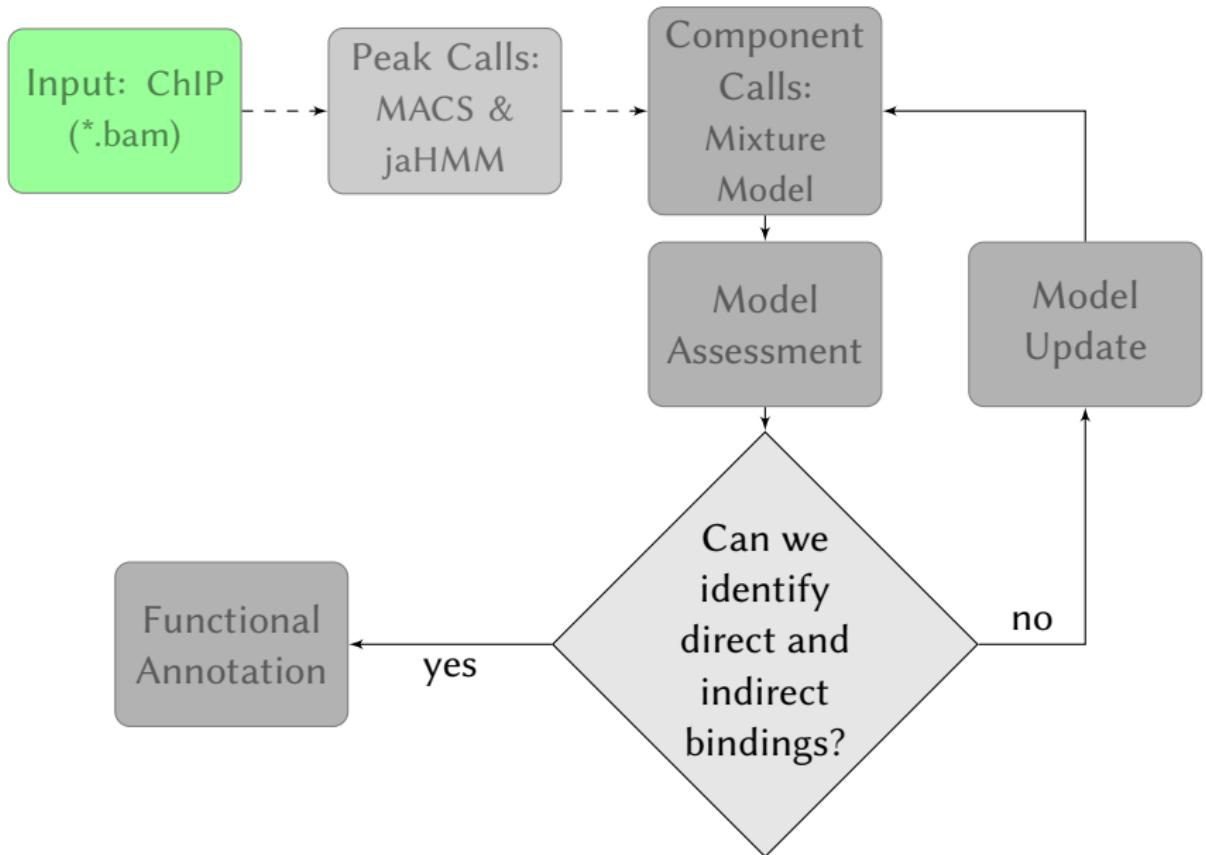
- Hard clustering: non-overlapping clusters
- Soft clustering: overlapping clusters

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  $\lambda$  parameter.
- Each gaussian corresponds to a cluster of signals with  $\mu$  and  $\sigma$  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

---

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

<sup>2</sup>Filion 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 <sup>1</sup>

---

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

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

# Peak Calls: MACS2 vs jaHMM

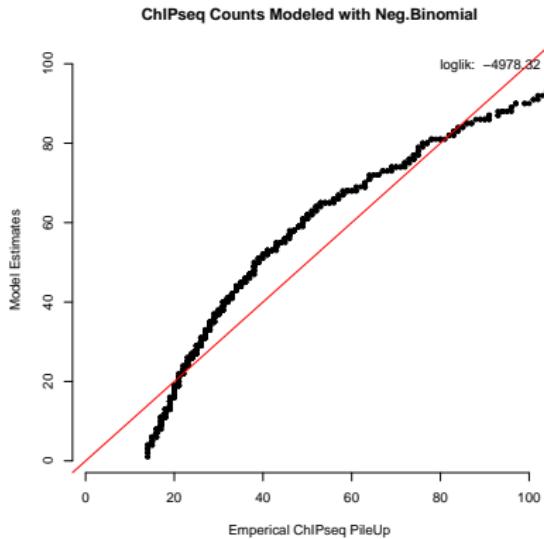
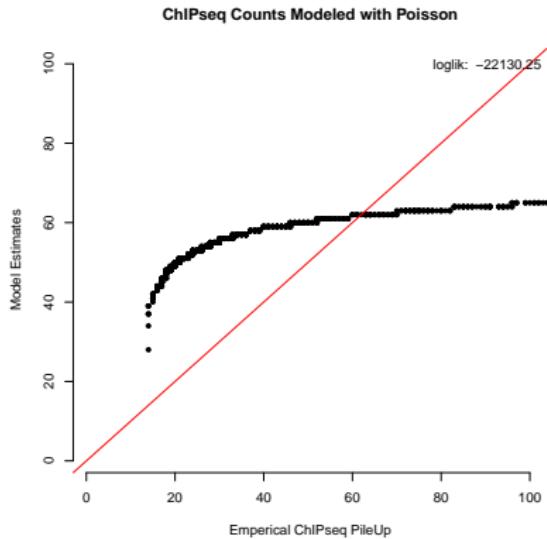
- **MACS2:** *poisson* model-based analysis of Peak calls <sup>1</sup>
- **jaHMM:** *negative binomial* model-based analysis of Peak calls <sup>2</sup>

---

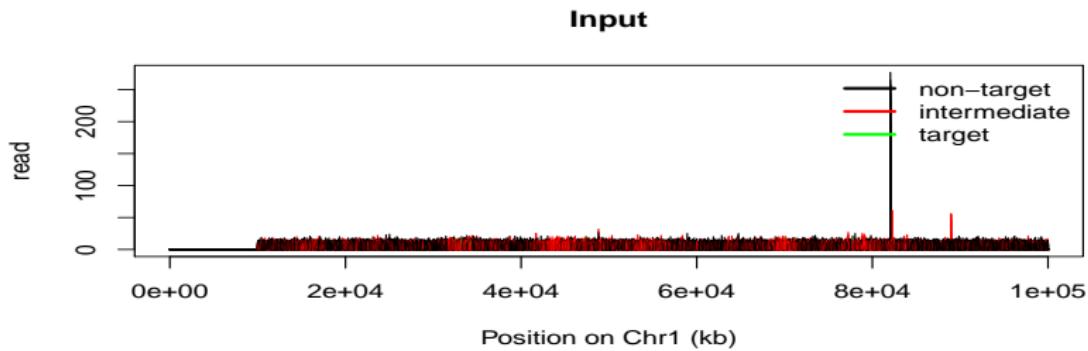
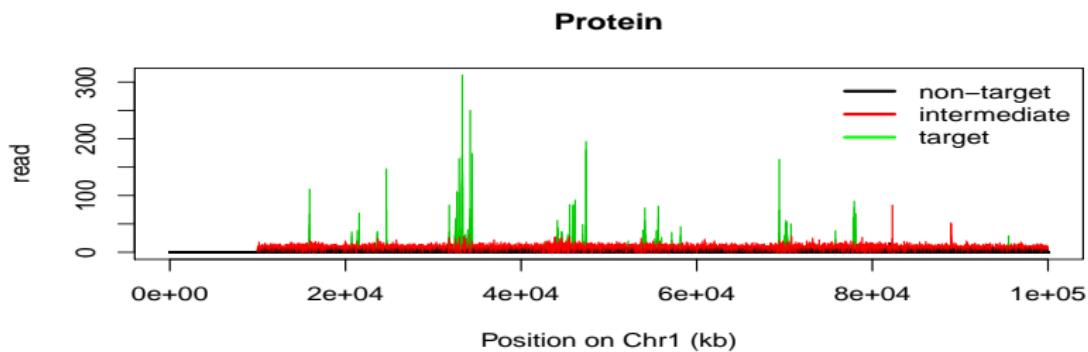
<sup>1</sup>Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137

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

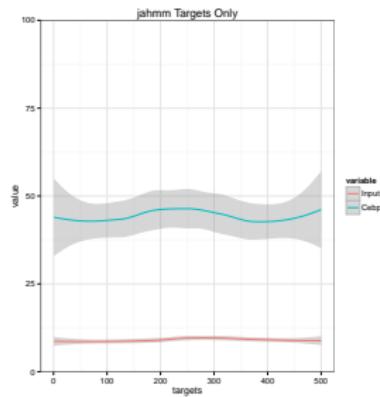
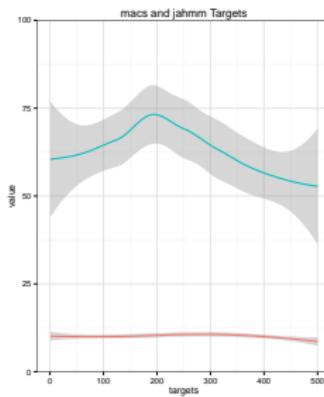
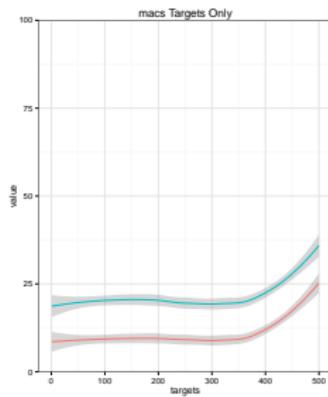
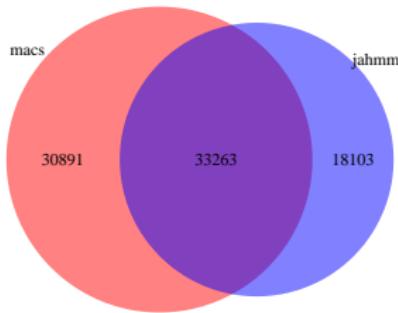
# jaHMM fits ChIPseq Signals better than MACS2



# Peaks Called by jahmm



# Peaks Called by MACS2 vs jahmm



# Summary: Peak Calls

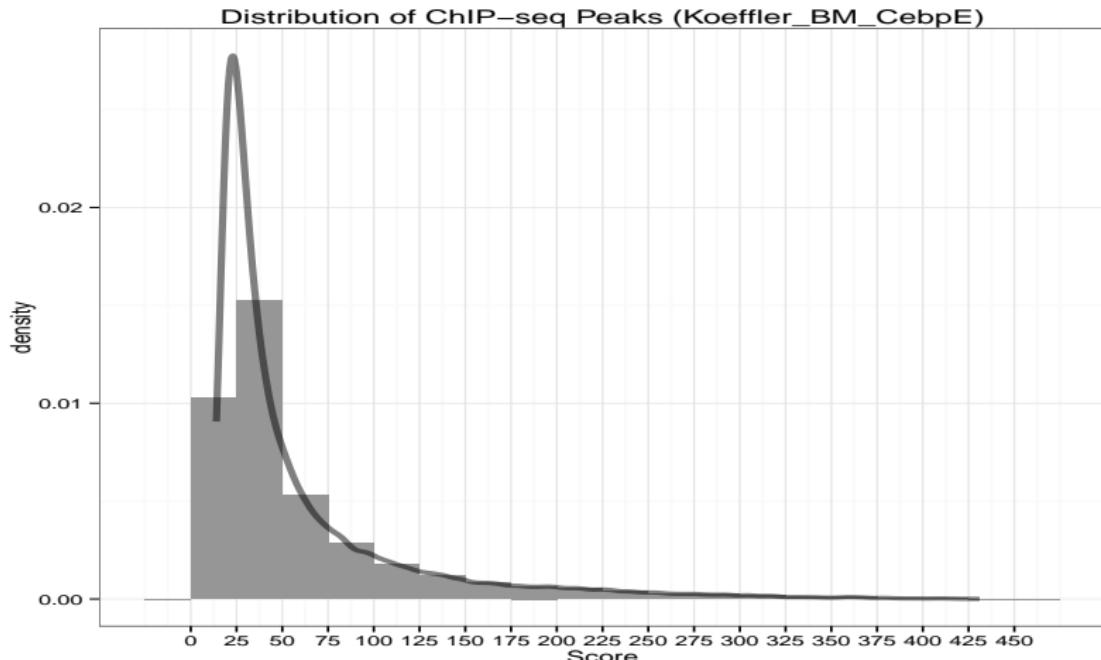
- Given our dataset, MACS2 is able to call peaks however, the estimated scores are less fit than JAHMM

# Summary: Peak Calls

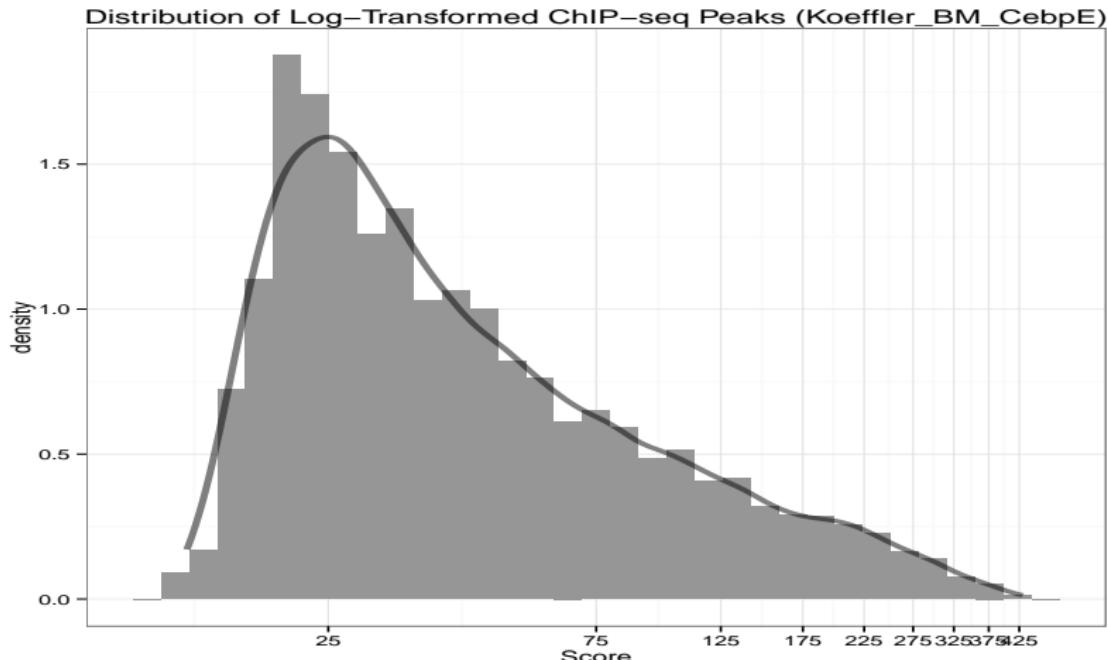
- Given our dataset, MACS2 is able to call peaks however, the estimated scores are less fit than JAHMM
- Peaks identified solely by jaHMM have scores higher with respect to their input (higher ratio) than solely by MACS2

Can we model ChIPseq using components of MMs?

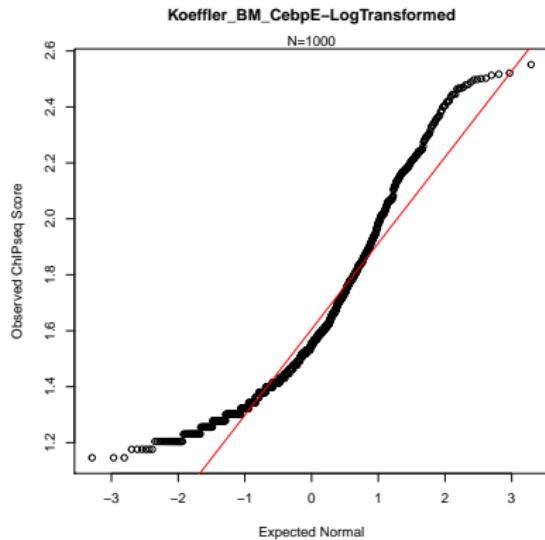
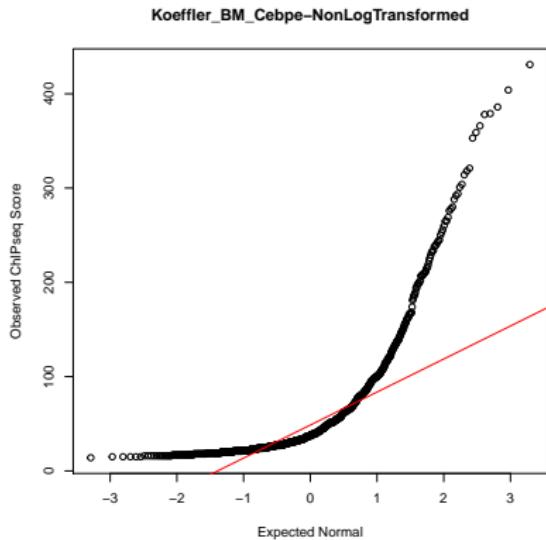
# Input: ChIP-seq of Cebp $\epsilon$ from Koeffler-BM



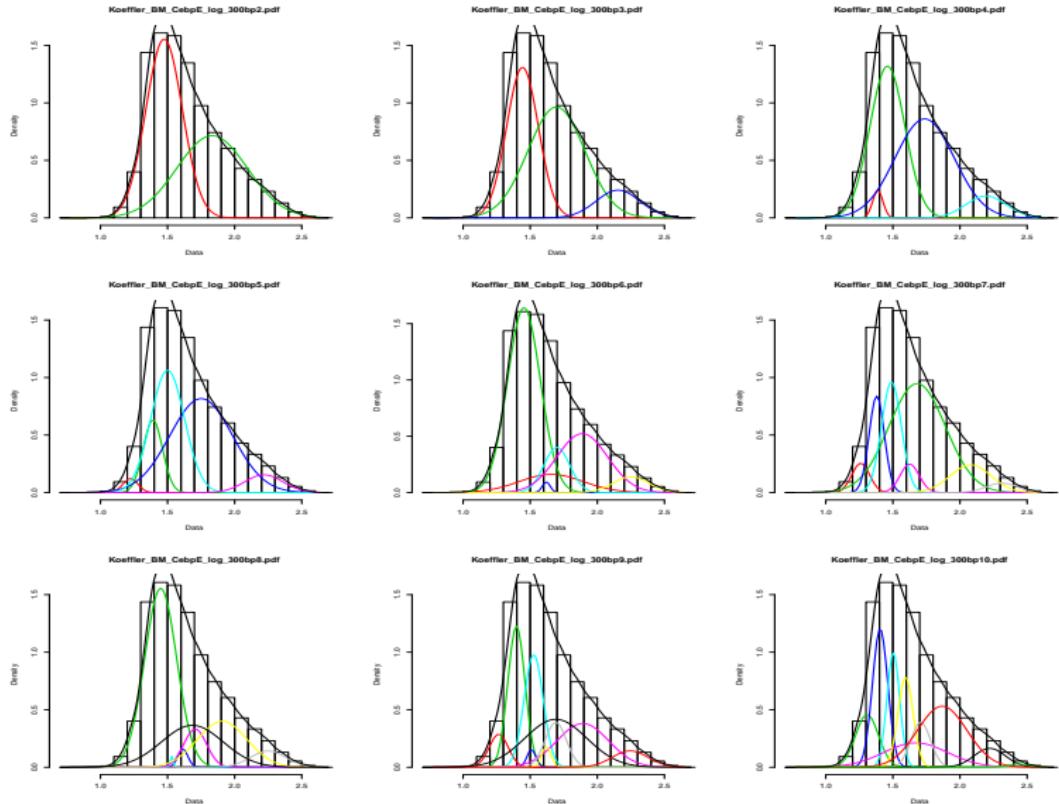
# Log Transformation of ChIP-seq Input



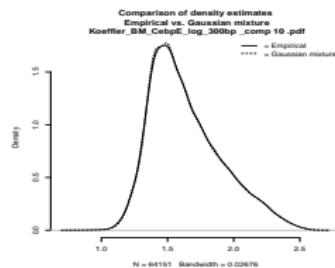
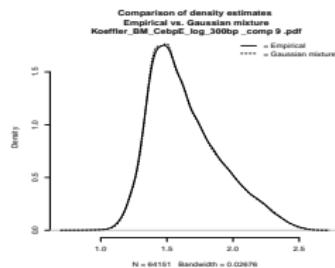
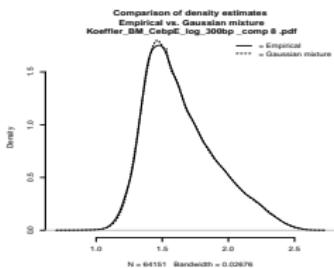
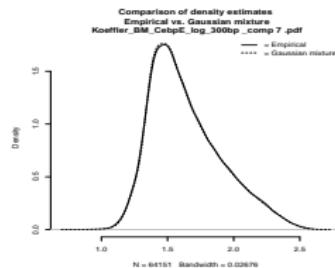
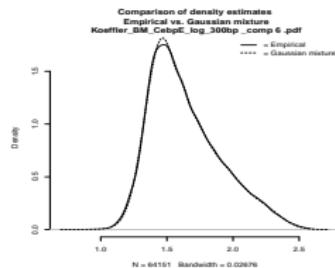
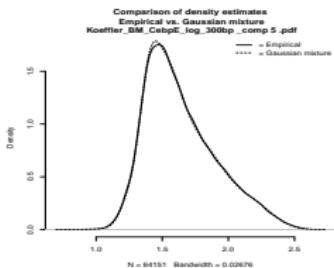
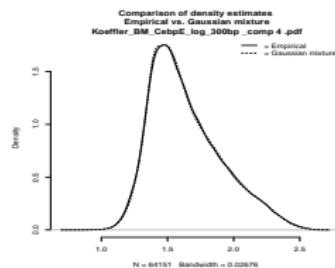
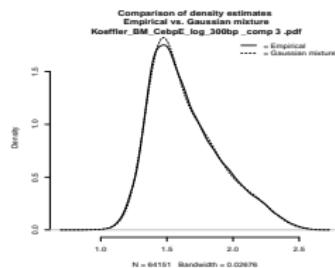
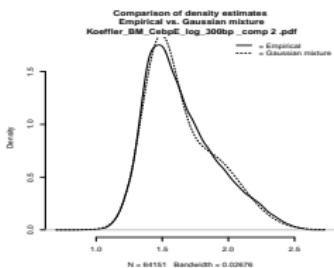
# Check the Gaussian Normality



# ComponentCalls: Fit ChIPseq Peaks with GMMs



## GMM-ModelAssessment: Overfit



# Model Assessment: BIC-AIC

AIC<sup>1</sup> and BIC<sup>2</sup> 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

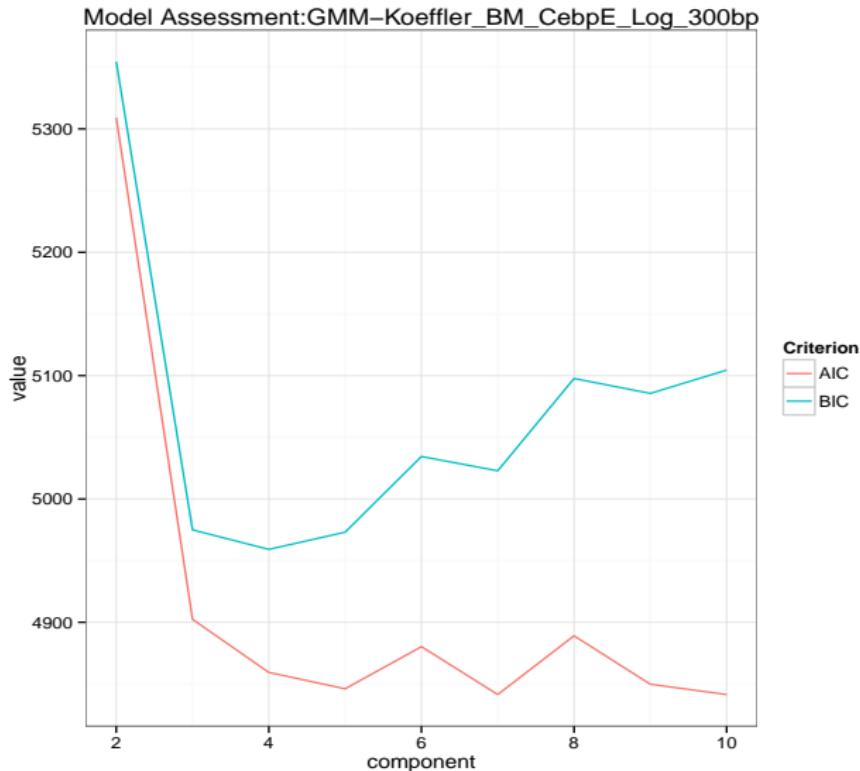
$P$  is the number of parameters

---

<sup>1</sup>Akaike information criterion

<sup>2</sup>Bayesian 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)

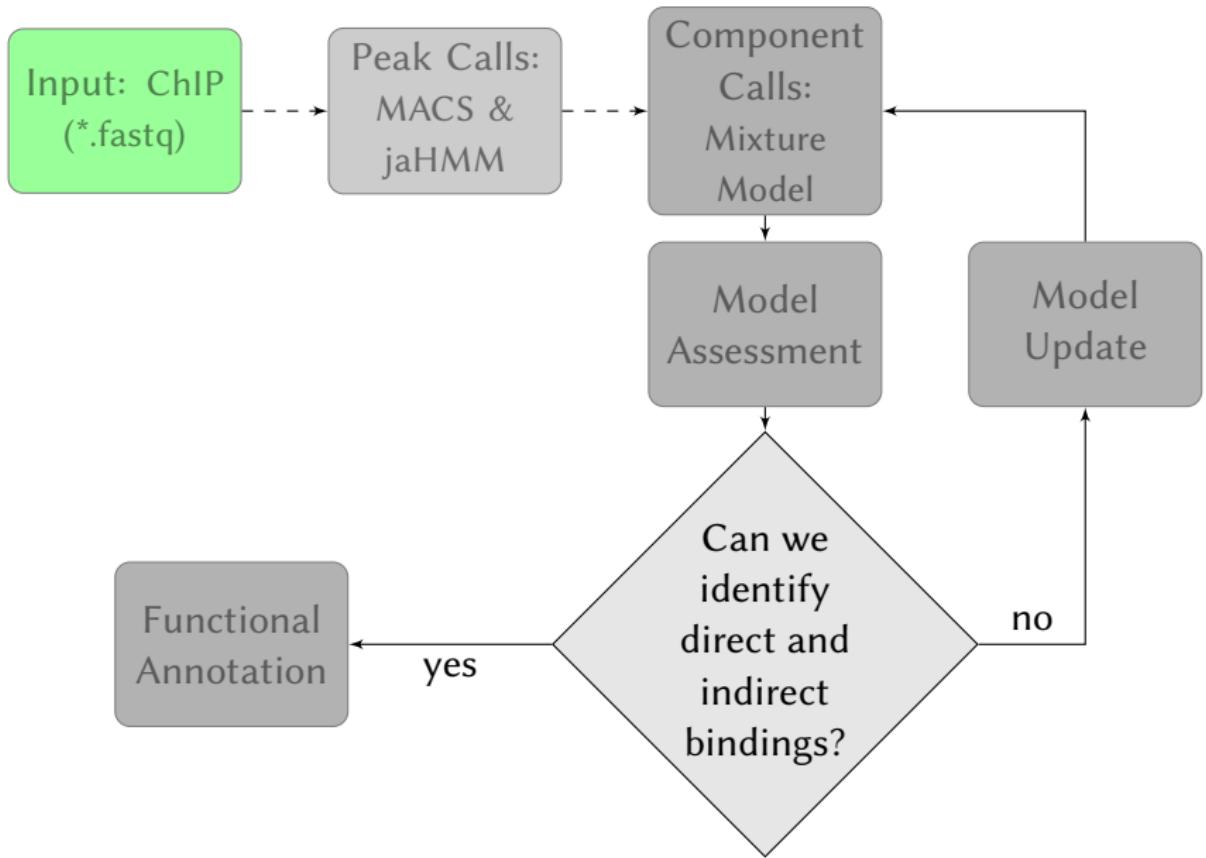
- **Can we model ChIPseq using several components of MMs?**

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.

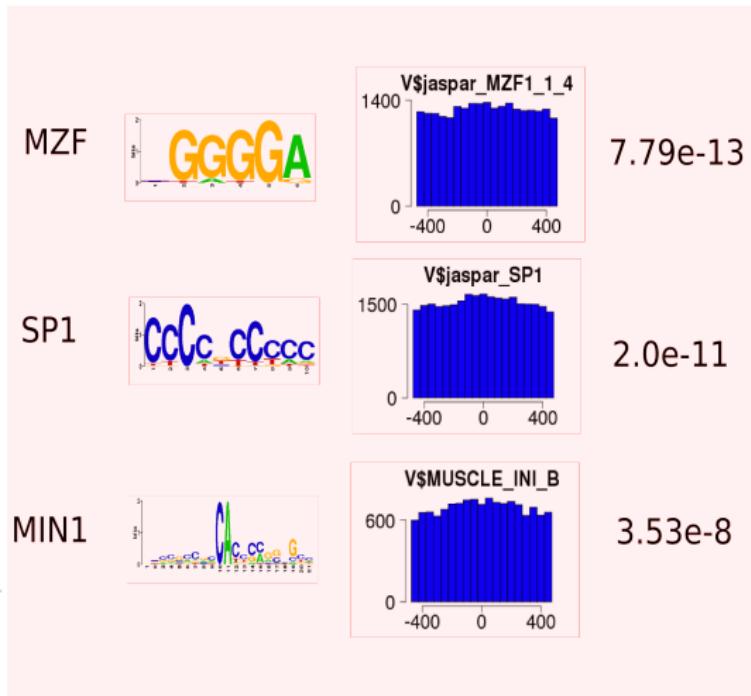
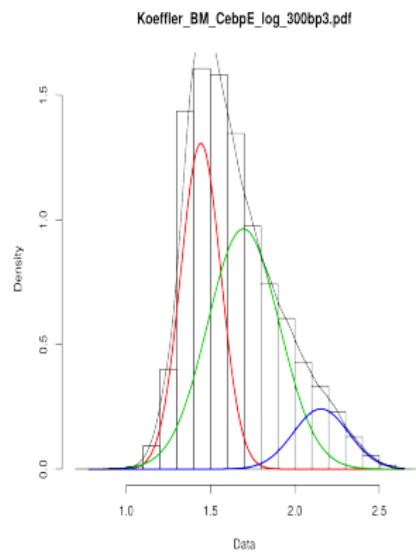


## Motif Calls using Centdist <sup>1</sup>

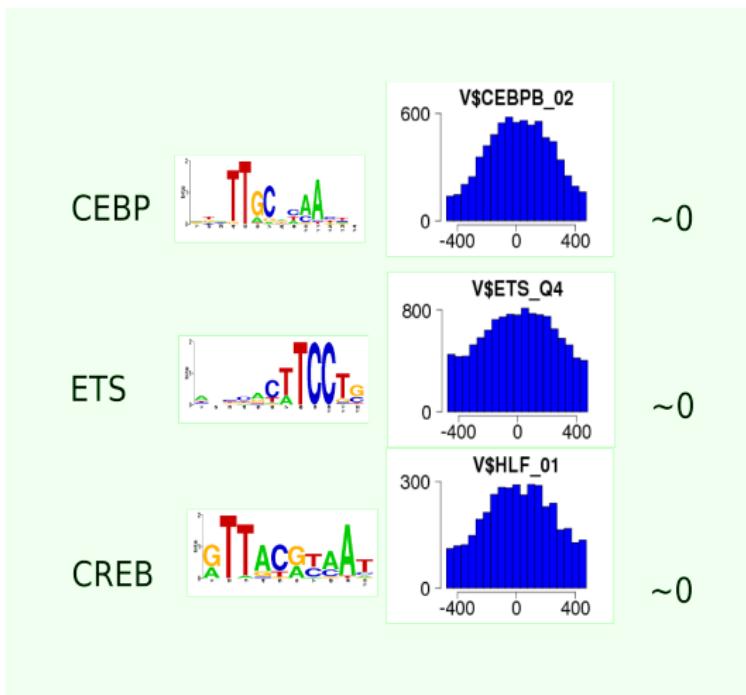
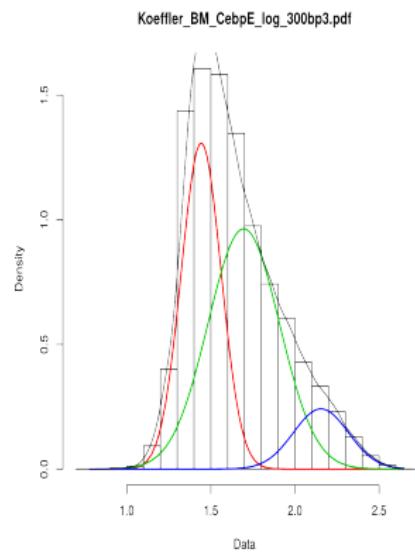
---

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

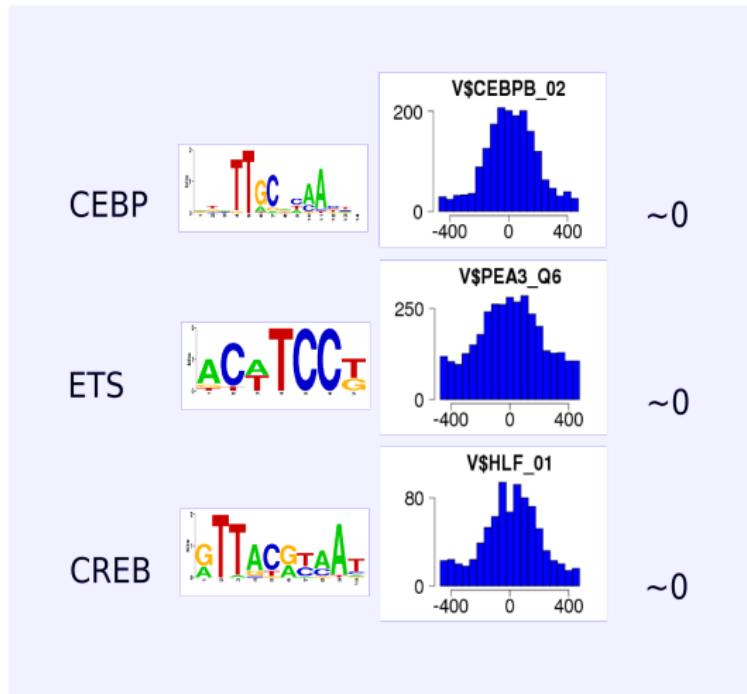
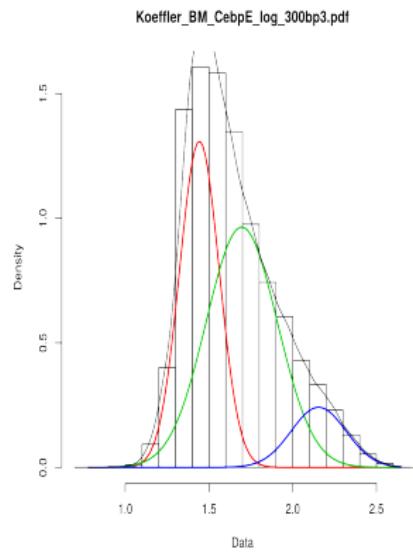
# Component 1: low peak score (29559 peaks)



# Component 2: intermediate peak score (28851 peaks)



# Component 3: high peak score (5741 peaks)



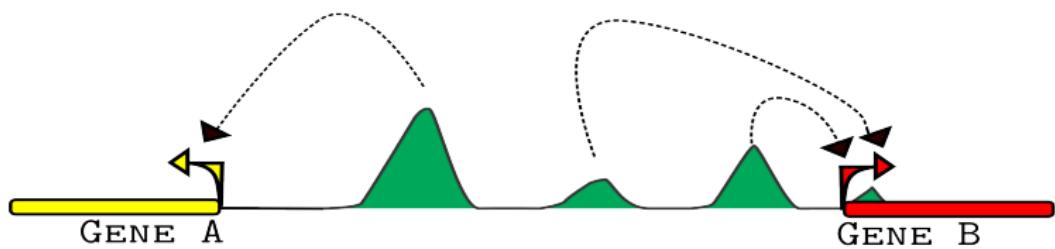
# Summary: Motif Calls

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

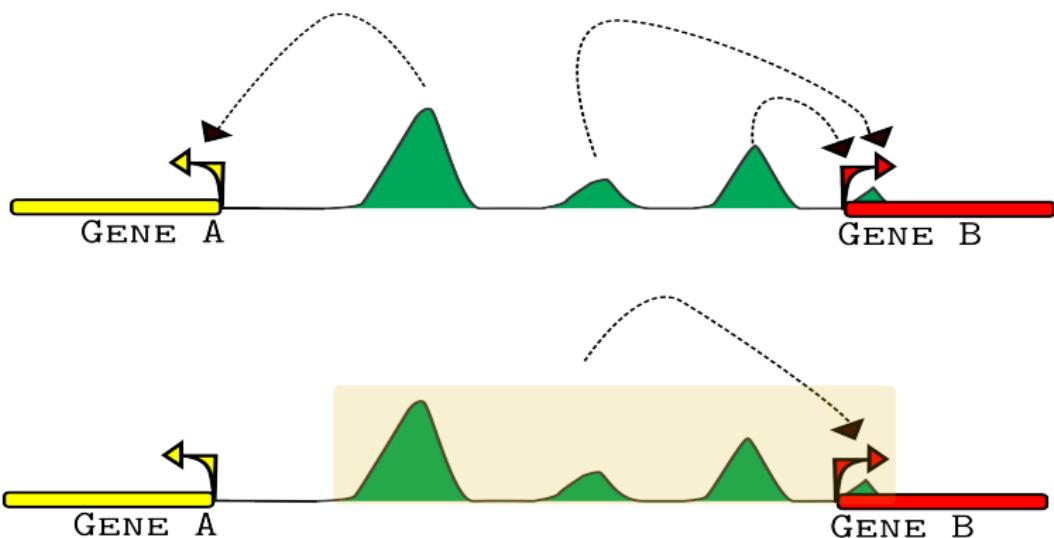
# Summary: Motif Calls

- Cebp motif is found in component 2 and 3 in 3-component GMMS using centdist
- Next, can we further segregate these groups into direct and indirect bindings?

# Peak Annotation

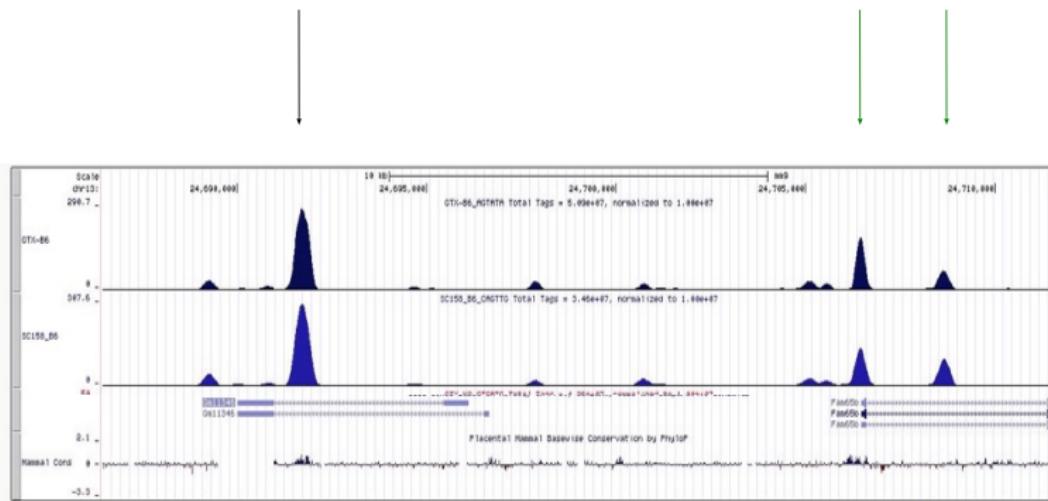


# Peak Annotation: Clustering



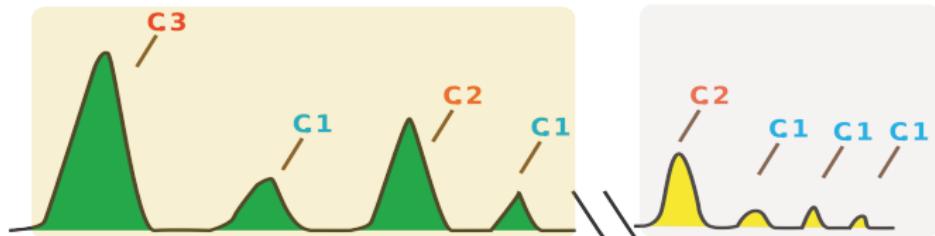
*What problems the invention solves and advantages over existing methods?*

### *An Example:*

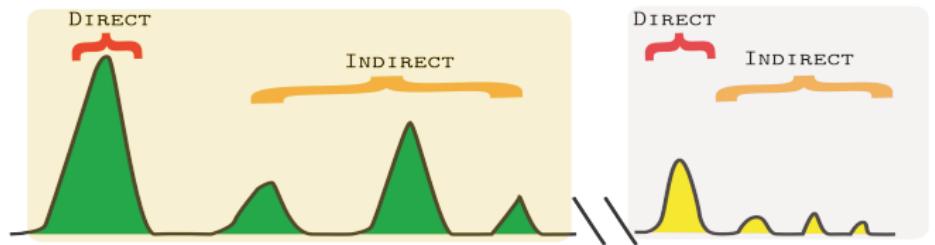


# Peak Clustering

## COMPONENT CLUSTERING

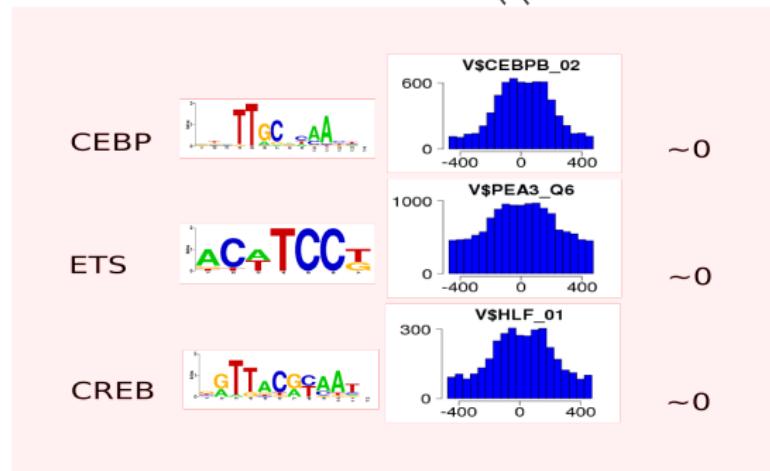


## LOCAL CLUSTERING



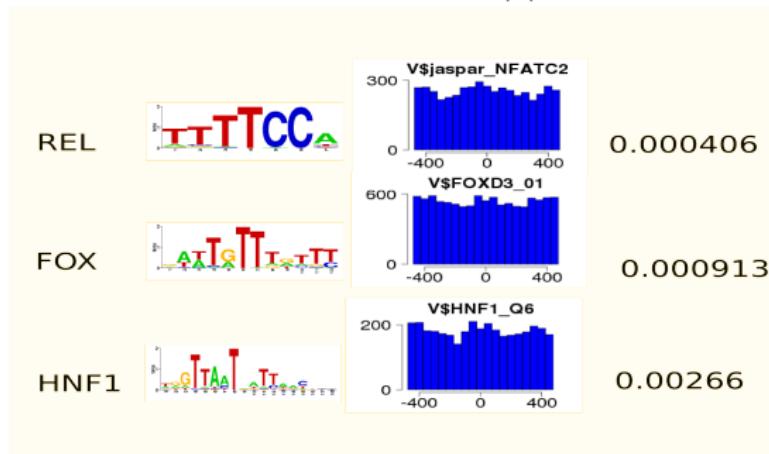
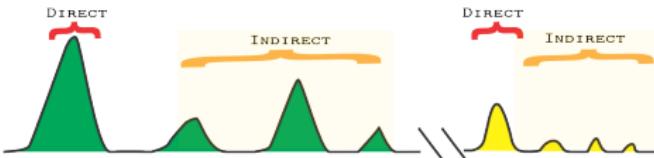
# Direct: 24948 peaks

## LOCAL CLUSTERING



# Indirect: 26547 peaks

## LOCAL CLUSTERING



## Motif Discovery using MEME-ChIP<sup>1</sup> and STAMP<sup>2</sup>

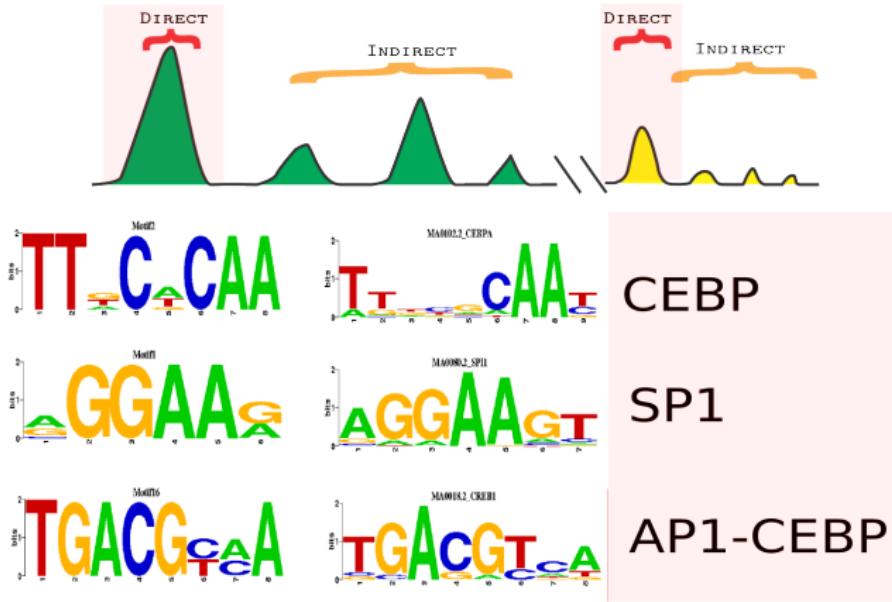
---

<sup>1</sup>Machanick, P and Bailey, T. MEME-ChIP: motif analysis of large DNA datasets. Nucleic Acids (2011)

<sup>2</sup>Mahony, S and Benos, P. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids (2007)

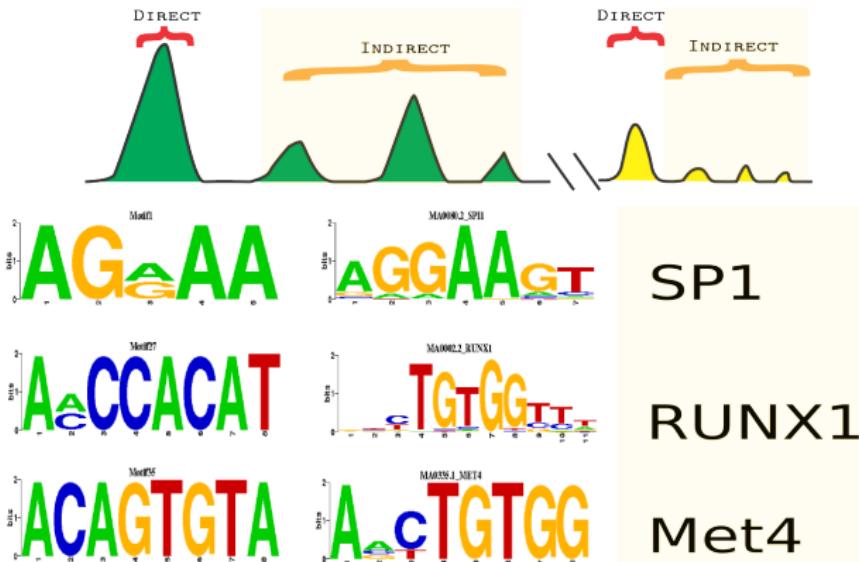
# Direct: 24948 peaks

## LOCAL CLUSTERING



# Indirect: 26547 peaks

## LOCAL CLUSTERING

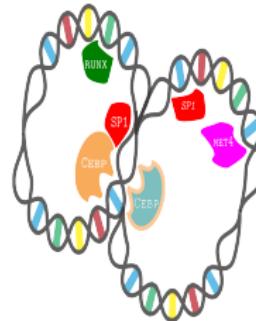


## Summary: Chromatin Conformation Prediction from ChIPseq Signal

- Our current method could separate direct and indirect bindings

## Summary: Chromatin Conformation Prediction from ChIPseq Signal

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



# Future Directions

- Dry-lab work

# Future Directions

- Dry-lab work
  - ChIPseq signals normalized mainly on GC and mappability

# Future Directions

- Dry-lab work
  - ChIPseq signals normalized mainly on GC and mappability
  - Motif validated with other methods

# Future Directions

- Dry-lab work
  - ChIPseq signals normalized mainly on GC and mappability
  - Motif validated with other methods
  - Method tested using other **high-quality** ChIPseq

# Future Directions

- Dry-lab work
  - ChIPseq signals normalized mainly on GC and mappability
  - Motif validated with other methods
  - Method tested using other **high-quality** ChIPseq
  - Chromatin conformation validated with other methods, such as ChIA-PET, and 5c

# Future Directions

- Dry-lab work
  - ChIPseq signals normalized mainly on GC and mappability
  - Motif validated with other methods
  - Method tested using other **high-quality** ChIPseq
  - Chromatin conformation validated with other methods, such as ChIA-PET, and 5c
- Wet-lab work

# Future Directions

- Dry-lab work
  - ChIPseq signals normalized mainly on GC and mappability
  - Motif validated with other methods
  - Method tested using other **high-quality** ChIPseq
  - Chromatin conformation validated with other methods, such as ChIA-PET, and 5c
- Wet-lab work
  - Chromatin conformation validated by 3C and ChIP

# Acknowledgement

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