

# A Fully Bayesian Hidden Ising Model for ChIP-Seq Data Analysis

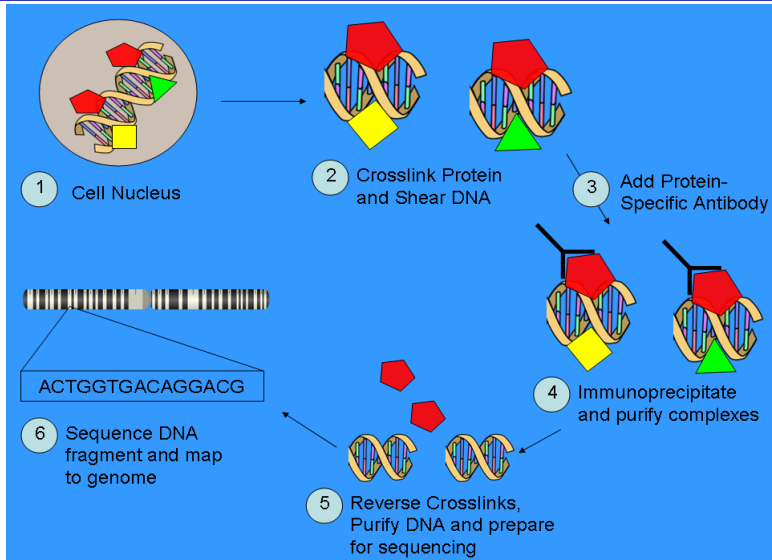
**Qianxing Mo**

**Department of Epidemiology and Biostatistics  
Memorial Sloan-Kettering Cancer Center**

**January 27, 2011**

- **An introduction to ChIP-seq experiments**
- **Literature review - the existing methods**
- **The proposed model - A fully Bayesian hidden Ising model**
- **Case study - Analysis of NRSF, CTCF and STAT1 ChIP-seq data**
- **Simulation Study**
- **Conclusions**

# An introduction to ChIP-seq experiments



[http://en.wikipedia.org/wiki/File:Chip\\_sequencing2.png](http://en.wikipedia.org/wiki/File:Chip_sequencing2.png)





## An example of ChIP-seq data

Chromosome	Start	End	Strand
chr1	42806148	42806175	R
chr5	53516508	53516535	R
chr11	100399671	100399698	F
chr12	28623360	28623387	R
chr9	27547010	27547037	F

## Typical steps for ChIP-seq data analysis

1. Build signal profiles for ChIP-seq data.
2. Find regions (peaks) enriched with sequence tags.
3. Remove artifactual enriched regions.

## Window-based tag counting

- **XSET** – Robertson et al. (2007) *Nat. Methods* 4, 651-657.
- **CisGenome** – Ji et al. (2008) *Nat. Biotechnol.* 26, 1293-1300.
- **SISSRs** – Jothi et al. (2008) *Nucleic Acids Res.* 36, 5221-5231.
- **SPP** – Kharchenko, Tolstorukov and Park. (2008) *Nat. Biotechnol.* 26, 1351-1359.
- **MACS** – Zhang et al. (2008) *Genome Biology* 9, R137.1-R137.9
- **PeakSeq** – Rozowsky et al. (2009) *Nat. Biotechnol.* 27, 66-75.
- **BayesPeak** – Syrou et al. (2009) *BMC Bioinformatics* 10:299.
- **HPeak** – Qin et al. (2010) *BMC Bioinformatics* 11:369.
- **PICS** – Zhang et al. (2010) *Biometrics* in press.

## Kernel density estimation

- **QuEST** – Valouev et al. (2007) *Nat. Methods* 5,829-8347.
- **F-Seq** – Boyle et al. (2008) *Bioinformatics* 24, 2537-2538.

## The simplest approach

- Set a threshold, and then define enriched regions if their tag counts exceed the threshold value ([Johnson et al., 2007](#); [Fejes et al., 2008](#)).

## One-sample analysis - only ChIP sample available

- The background (null) distribution used to calculate p-value is usually modeled with a Poisson or negative binomial distribution ([Robertson et al., 2007](#); [Ji et al., 2008](#); [Zhang et al., 2008](#); [Jothi et al., 2008](#)).

## Two-sample analysis - both ChIP and control sample available

- Binomial distribution is often used to estimate the p-values for the enriched regions ([Ji et al., 2008](#); [Rozowsky et al., 2009](#)).
- Estimate empirical FDR, which is defined as the ratio of the number of control peaks to the number of ChIP peaks ([Valouev et al., 2008](#); [Zhang et al., 2008](#)).

# Bayesian and HMM-based methods

- BayesPeak ([Spyrou et al., 2009](#)) — A fully Bayesian hidden Markov model.
- HPeak ([Qin et al., 2010](#)) — A two-state HMM; Viterbi training is used for parameter estimation.
- PICS ([Zhang et al., 2010](#)) — A Bayesian hierarchical t-mixture model. EM algorithm is used for parameter estimation.

# Modeling ChIP-seq data through hidden Ising models

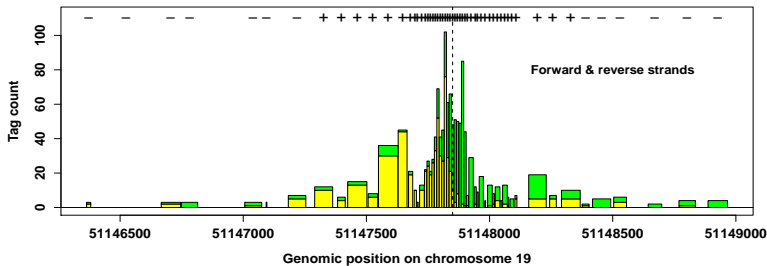
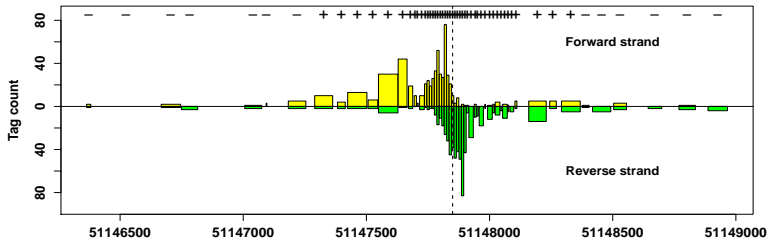
This work is an extension of Mo and Liang's work on ChIP-chip data analysis (Mo and Liang, 2010).



$$\pi(\mathbf{x}|\kappa) = \frac{1}{Z(\kappa)} \exp \left( \sum_{i=1}^n \left( \kappa \sum_{j \in W(i)} \delta(x_i, x_j) \right) \right)$$

# Building dynamic signal profiles for ChIP-seq data

Bin size: {80, 40, 20, 10}; Threshold of triggering size change: 10 tags.



# The proposed model for ChIP-seq data analysis

## The model

- Let  $\mathbf{y} = (y_1, \dots, y_n)$  be a realization of tag counts  $(Y_1, \dots, Y_n)$  in  $n$  genomic bins along a chromosome. Let each bin associate with a binary latent variable  $X_i \in \{-1, 1\}$ , where  $X_i = 1$  or  $X_i = -1$  denotes that the bin belongs to an enriched region or non-enriched region, respectively.
- We assume, conditional on  $X_i = x_i$ ,

$$y_i | x_i \sim \begin{cases} \text{Poisson}(\lambda_a) & \text{if } x_i = -1, \\ \text{Poisson}(\lambda_b) & \text{if } x_i = 1. \end{cases}$$

- Conditional on  $\mathbf{X} = (X_1, \dots, X_n)$ , we assume  $Y_1, \dots, Y_n$  are independent.

# The priors

The prior for the latent vector  $\mathbf{x}$ :

$$\pi(\mathbf{x}|\kappa) = \frac{1}{Z(\kappa)} \exp\left(\kappa \sum_{i=1}^{n-1} x_i x_{i+1}\right),$$

where  $\kappa$  is the interaction parameter, and  $Z(\kappa) = 2^n (\cosh(\kappa))^{n-1}$  is the normalizing constant of the distribution. When  $\kappa > 0$ , it is a ferromagnetic model.

We assume that  $\kappa, \lambda_a, \lambda_b$  have the following prior distributions:

$$\kappa \sim U(0, 10)$$

$$\lambda_a \sim \Gamma(\alpha_0, \beta_0)$$

$$\lambda_b \sim \Gamma(\alpha_1, \beta_1)$$

# The full conditional distributions

$$\lambda_a | \cdot \sim \Gamma(s_0 + \alpha_0, n_0 + \beta_0),$$

$$\lambda_b | \cdot \sim \Gamma(s_1 + \alpha_1, n_1 + \beta_1),$$

$$\pi(\kappa | \cdot) \propto \frac{1}{Z(\kappa)} \exp\left(\kappa \sum_{i=1}^{n-1} x_i x_{i+1}\right) I(0 < \kappa < 10),$$

$$\pi(x_i = 1 | \cdot) = \left(1 + \frac{\lambda_a^{y_i}}{\lambda_b^{y_i}} \exp\left(\lambda_b - \lambda_a - 2\kappa(x_{i-1} + x_{i+1})\right)\right)^{-1},$$

where  $i = 1, \dots, n$ ,  $x_0 = x_{n+1} = 0$ ,  $s_0 = \sum y_i I(x_i = -1)$ ,  
 $n_0 = \sum I(x_i = -1)$ ,  $s_1 = \sum y_i I(x_i = 1)$ , and  $n_1 = \sum I(x_i = 1)$ .

# The NRSF, CTCF and STAT1 ChIP-seq Data

T.Factor	Authors	# of uniquely mapped tags
CTCF	<a href="#">Barski et al., 2007</a>	~2.95 M ChIP tags
NRSF	<a href="#">Johnson et al., 2007</a>	~5.36 M control tags ~6.16 M ChIP tags
STAT1 (no Chr Y)	<a href="#">Rozowsky et al., 2009</a>	~26.73 M control tags ~23.44 M ChIP tags

- Illumina/Solexa sequencing technology was used for the three experiments.
- The lengths of the sequence tags: 25– to 36– bp.
- Only the uniquely mapped tags (up to two mismatches) were used.

# The algorithm for ChIP-seq data analysis

## A. Built dynamic signal profiles for ChIP-seq data.

- Built the signal profiles using bin size = {80, 40, 20, 10 bp}.
- Used 10 tags for the NRSF and CTCF profiles and 20 tags for the STAT1 profiles as the thresholds for triggering size change.

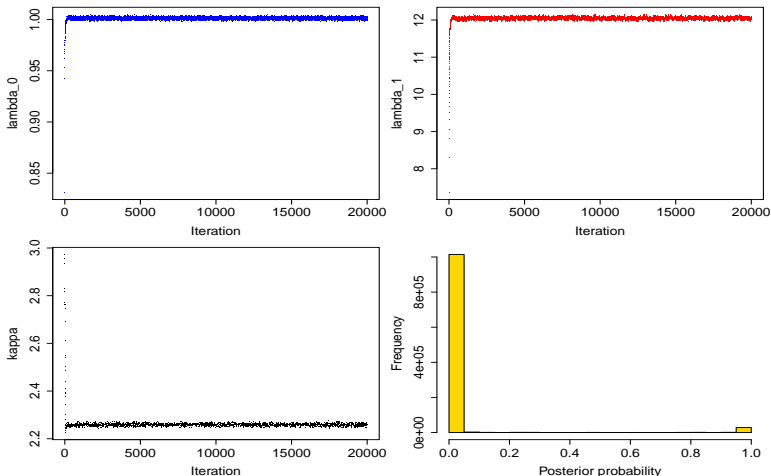
## B. Modeled the signal profiles using the proposed model.

## C. Called enriched regions and defined the predicted exact binding sites (PEBSs).

- The posterior probabilities of the hidden states being enriched were used to call enriched regions.
- The counts and directionality of tags were used to specify the PEBSs.

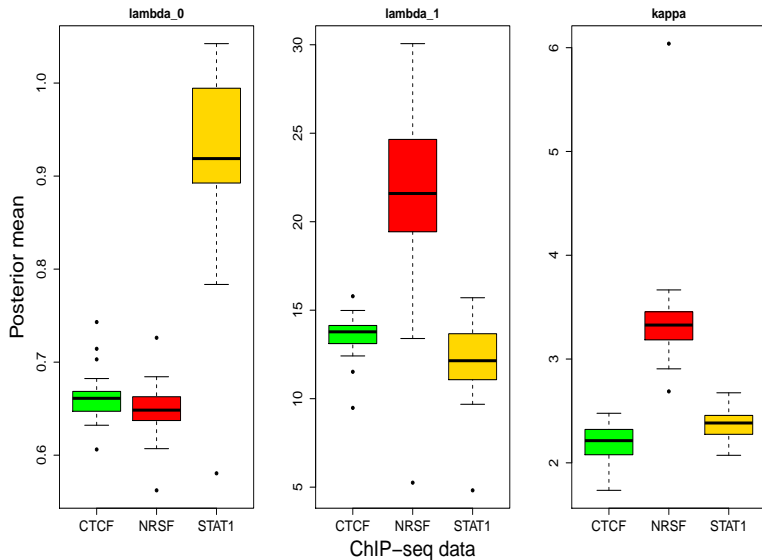
# Model diagnosis

Trace plots and posterior probabilities for the STAT1 chr1 profile

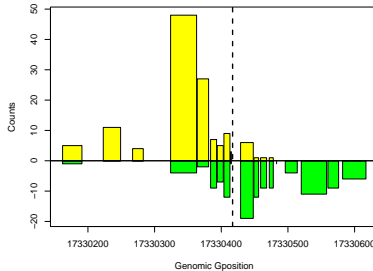
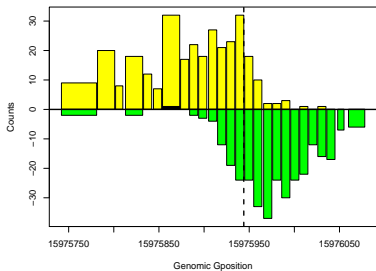
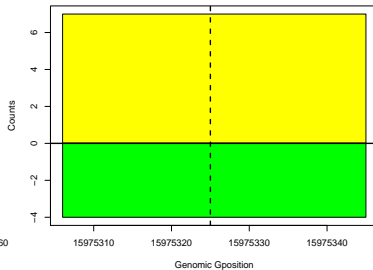
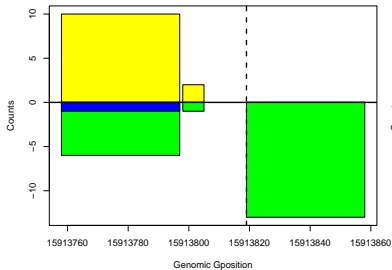


All the chains passed the Heidelberger and Welch's test ( $p < 0.05$ ).

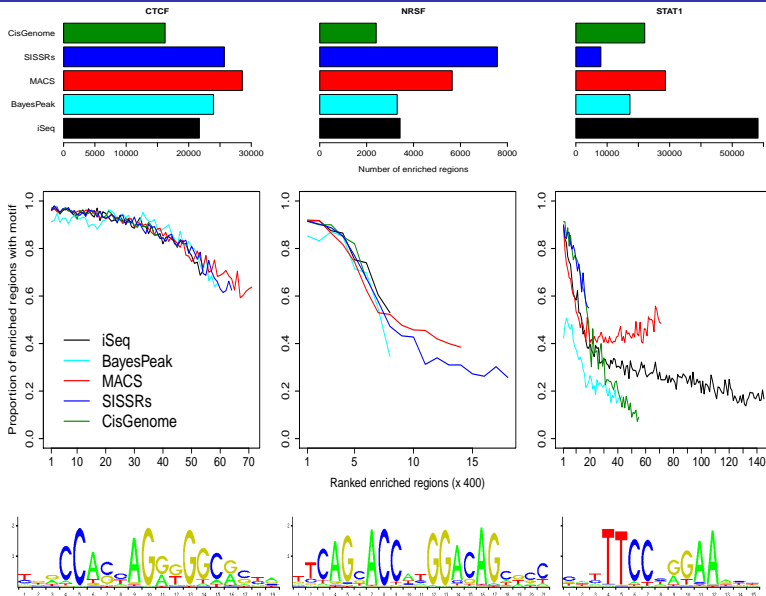
# Box plots of the posterior means of the model parameters



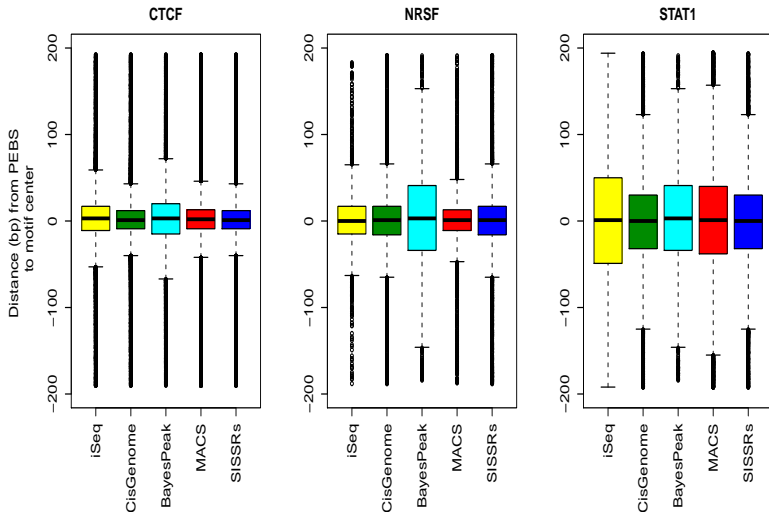
# Typical enriched regions detected by iSeq



# Comparison: number of identified enriched regions and motif occurrence rates



# Comparison: the distance from the predicted exact binding site to the motif center

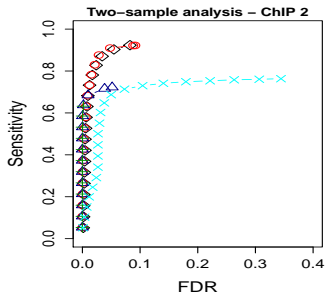
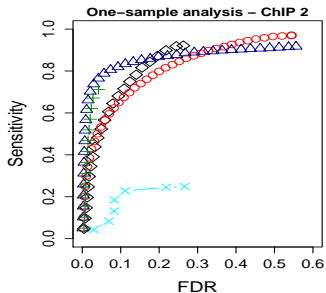
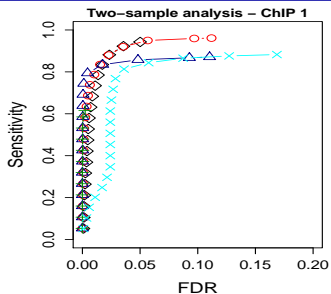
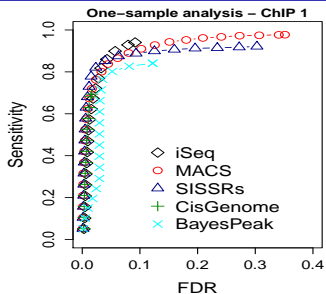


**The mixed ChIP-seq data:** The mixed datasets were made of the STAT1 control tags and 372,662 ChIP tags that formed 9,395 enriched regions with high confidence CTCF motif ( $p\text{-value} < 10^{-6}$ ).

Dataset	Total # of tags	# of ChIP tags	# of ER
ChIP data 1	~4.41 M	372,662	9,395
Control data 1	~4.51 M		
ChIP data 2	~8.33 M	372,662	9,395
Control data 2	~8.47 M		

- (A) One-sample analysis of ChIP data 1.
- (B) Two-sample analysis of ChIP data 1 + Control data 1.
- (C) One-sample analysis of ChIP data 2.
- (D) Two-sample analysis of ChIP data 2 + Control data 2.

# Sensitivity and FDR of the detected enriched regions



## Characteristics of the proposed method – iSeq

- **Dynamic profiles**
- **Simple and flexible**
- **Computationally efficient**
- **High sensitivity and low FDR**
- **Dichotomized posterior probabilities**
- **Freely available**

The screenshot shows a Mozilla Firefox browser window with the address bar displaying `http://www.bioconductor.org/packages/2.7/bioc/html/iSeq.html`. The page title is "iSeq". The main heading is "iSeq" in a large, bold font, followed by the subtitle "Bayesian Hierarchical Modeling of ChIP-seq Data Through Hidden Ising Models". A green box contains the text: "This package uses Bayesian hidden Ising models to identify IP-enriched genomic regions from ChIP-seq data. It can be used to analyze the ChIP-seq data with or without controls." Below this, the author and maintainer are listed as "Qianxing Mo". A white box contains the installation instructions: "To install this package, start R and enter:" followed by the R code: 

```
source("http://bioconductor.org/biocLite.R")
biocLite("iSeq")
```

 The "Documentation" section includes links for "iSeq", "PDF", "R Script", and "Reference Manual". The "Details" section shows "biocViews" as "ChIP-seq, next generation sequencing, massively parallel sequencing, bioinformatics" and "Depends" as "R".

**iSeq**

**Bayesian Hierarchical Modeling of ChIP-seq Data Through Hidden Ising Models**

This package uses Bayesian hidden Ising models to identify IP-enriched genomic regions from ChIP-seq data. It can be used to analyze the ChIP-seq data with or without controls.

Author    Qianxing Mo  
Maintainer    Qianxing Mo

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("iSeq")
```

**Documentation**

[iSeq](#) [PDF](#) [R Script](#)  
[Reference Manual](#)

**Details**

biocViews    [ChIP-seq](#), [next generation sequencing](#), [massively parallel sequencing](#), [bioinformatics](#)  
Depends    R