

A Non-Homogeneous Hidden Markov Model on First Order Differences for Automatic Detection of Nucleosome Positions

Pei Fen Kuan¹, Dana Huebert², Audrey Gasch³, Sündüz Keleş^{1,4*}

¹Department of Statistics, University of Wisconsin,
Madison, WI 53706.

²Department of Cellular and Molecular Biology, University of Wisconsin,
Madison, WI 53706.

³Department of Genetics, University of Wisconsin,
Madison, WI 53706.

⁴Department of Biostatistics and Medical Informatics, University of Wisconsin,
Madison, WI 53706.

*E-mail: keles@stat.wisc.edu

May 12, 2008

Abstract

The heterogeneity of nucleosome densities across genomes and short linker regions are the two main challenges in mapping nucleosome occupancies based on chromatin

immunoprecipitation on microarrays (ChIP-chip) data. Previous works rely on heuristic detrending and careful visual examination to detect low density nucleosomes, which may exist in subpopulation of cells. We propose a non-homogeneous hidden Markov model based on first order differences of experimental data along genomic coordinates that bypasses the need for local detrending and can automatically detect nucleosome positions of various occupancy levels. Our proposed approach is applicable to both ChIP-chip and ChIP-Seq (Chromatin Immunoprecipitation and Sequencing) data, and is able to map nucleosome-linker boundaries accurately. This automated algorithm is also computationally efficient and only requires a simple preprocessing step. We provide several examples illustrating the pitfalls of existing methods, the difficulties of detrending the observed hybridization signals and demonstrate the advantages of utilizing first order differences in detecting nucleosome occupancies via simulations and case studies involving ChIP-chip and ChIP-Seq data on nucleosome occupancy in yeast.

Keywords: nucleosomes; ChIP-chip; ChIP-Seq; non-homogeneous hidden Markov model; first order differences.

1 Introduction

Nucleosomes consist of approximately 146 base pairs of DNA wrapped around a histone octamer (Chakravarty et al.; 2006). The positioning of nucleosomes along the genome has been implicated in the regulation of gene expression. Packaging DNA into nucleosomes may prevent DNA binding proteins from accessing their sites, recruit transcriptional activators or repressors, and bring distant DNA sequences into close proximity to promote transcription (Millar and Grunstein; 2006). A high percentage of the *S.cerevisiae* genome is known to be occupied by nucleosomes, however there exists substantial variation in nucleosome density across the genome. In particular, relatively higher density of nucleosomes is observed at

transcribed regions and lower density is found in intergenic regions (Lee et al.; 2004; Bernstein et al.; 2004).

Numerous chromatin immunoprecipitation experiments have been carried out to map nucleosome occupancy in yeast via tiling arrays (ChIP-chip) (Liu et al.; 2005; Yuan et al.; 2005; Lee et al.; 2007; Shivaswamy and Iyer; 2008) and more recently, a high resolution whole genome nucleosome map for yeast genome was developed using chromatin immunoprecipitation sequencing technology (ChIP-Seq) (Albert et al.; 2007; Shivaswamy et al.; 2008). In both technologies, the sample input consists of mono-nucleosomes prepared via micrococcal nuclease digestions. The digested sample is sequenced in ChIP-Seq, or competitively hybridized against a control sample using tiling arrays in ChIP-chip. For expository purposes, we limit our detailed discussion to the case of ChIP-chip data and provide an extension of our proposed method to ChIP-Seq data in the case study. Positions of nucleosomes across the whole genome are therefore characterized by a stretch of consecutive probes encompassing approximately 146 base pairs with higher signals than the background. Two nucleosomes are connected by linker DNA, which is digested by the enzyme. An interesting feature observed in many of the ChIP-chip experiments for mapping nucleosome positions is that the magnitude of log base 2 ratios for regions occupied by nucleosomes exhibit large variabilities. Specifically, some regions of the genome thought to be occupied by nucleosomes actually show log base 2 ratios below the baseline. Yuan et al. (2005) provided substantial evidence of this problem and referred to this phenomena as unpredictable trends in hybridization. The variability in the magnitudes of nucleosome occupancy is also observable from the ChIP-Seq data in Shivaswamy et al. (2008). This trend in hybridization is attributed to the heterogeneity of nucleosome densities across the whole genome, resulting in both stable and unstable nucleosome occupancies. Unstable or low-signal nucleosomes are nucleosome peaks having low maxima and may correspond to nucleosomes found only in subpopulation of cells (Yuan

et al.; 2005). We will refer to these as “low-signal nucleosomes” in our subsequent discussion.

Previous works in identifying nucleosome positions in ChIP-chip data include using a hidden Markov model (HMM) (Yuan et al.; 2005) or a hierarchical generalized hidden Markov model (HGHMM) (Gupta; 2007) on the observed log base 2 ratios. Yuan et al. (2005) proposed an HMM that takes into account the length of nucleosomal DNA and allows for one emission distribution for each of the nucleosome and linker states, respectively. To account for a global trend, Yuan et al. (2005) applied the HMM to a sliding window of 40 probes and averaged the estimated model parameters and posterior probabilities over all the windows covering a fixed probe to compute the most likely hidden state path. They also adopted a heuristic procedure to further detrend the data locally by comparing the median intensities of the peak and trough within a window size of 7 probes in order to detect low-signal nucleosomes and finally postprocessing the results so that the window boundaries do not cut across nucleosomes. In addition, potential low nucleosomes missed by the sliding window HMM were hand picked via close visual inspection. This heuristic approach becomes tedious when one needs to map nucleosome occupancy in a larger genomic region.

In contrast, Gupta (2007) proposed a HGHMM that modeled the length of nucleosomal DNA explicitly and allowed for probe specific emission distribution in a hierarchical Bayesian framework. The proposed algorithm is computationally intensive and some parameters were chosen based on simulation results. In addition, the HGHMM approach does not address the trend issues which could potentially miss the low-signal nucleosomes. To accommodate the serious drawbacks of existing methods, we propose an alternative approach which automatically identifies nucleosome occupancy and incorporates the length of nucleosomal DNA and the observed trends in hybridization signals. At the core of our methodology is a non-homogeneous HMM architecture tailored for ChIP-chip data measuring nucleosome

occupancies. By designing the architecture for first order differences of log base 2 ratios, we bypass the problem of unpredictable trends in the log base 2 ratios. An additional unique feature of our approach is its applicability to the more recent ChIP-Seq data. We illustrate the methodology and benchmark its performance against other available methods in simulations and a case study involving yeast ChIP-chip nucleosome occupancy data. We also provide an illustration of its applicability to yeast ChIP-Seq nucleosome occupancy data. Two consecutive nucleosomes are separated by a linker of variable length. Therefore, a good methodology for mapping nucleosome occupancy should be able to identify nucleosome-linker boundaries accurately. This is usually challenging for the common tiling array design in which a linker is represented by one or two probes. Our proposed methodology carefully exploits the structure of nucleosomes and accurately maps nucleosome positions.

2 Motivation

We motivate the idea behind our methodology using the ChIP-chip data from Yuan et al. (2005). We use the normalized median log base 2 ratios of the 8 replicates for illustration. The top panel of Figure 1 shows the nucleosome profile for a region in chromosome 3 in which the nucleosomes identified by Yuan et al. (2005) are marked with black lines (each line representing a probe), and a stable nucleosome is represented by 6 to 8 probes. It is clear from the plot that the magnitude of log base 2 ratios of a nucleosome region exhibits large variabilities. Despite having heterogeneous hybridization signals, the plot suggests that a nucleosome is characterized by a peak in the local signal intensity, even if the log base 2 ratio is below the baseline. In other words, a nucleosome occupied region exhibit a “bump” shape irrespective of the actual strength in hybridization signal. In addition, the plot also suggests that using a single distribution for each of the nucleosome and linker/nucleosome depleted regions may fail to distinguish short linkers between stable or well-positioned nucleosomes,

(i.e., linkers between well-positioned nucleosomes have comparable hybridization strength to low-signal nucleosomes.)

Given the observed “bump” (or peak with low maxima) characteristic of annotated nucleosomes in the original data, we consider a simple smoothing by replacing the log base 2 ratios of probe i with the average values of probe $i - 1$, i and $i + 1$. As evident in the middle panel of Figure 1, the “bump” shape is enhanced in the smoothed data which enable easier mapping of the nucleosome positions. The “bumps” also suggest that a nucleosome occupied region is characterized by a series of decreasing positive slopes, followed by slopes of approximately zero in magnitude and then a series of increasing negative slopes. This observation forms the modeling framework of our proposed methodology. The first order differences automatically take care of the trend in hybridization and thereby bring both the low and stable nucleosomes to a comparable level.

3 Hidden Markov model for mapping nucleosome positions

As motivated in Section 2, to circumvent the problem of decoding nucleosome occupancy locally to accommodate for the observed local trends as in Yuan et al. (2005), we consider an alternative approach to infer nucleosome positions based on first order differences, O_t , which we defined as:

$$O_t = X_t - X_{t-1},$$

$$X_t = \sum_{j=t-w}^{t+w} Y_j,$$

where Y_j is the observed log base 2 ratio of probe j and X_t is the corresponding moving average statistic in a window size of $2w + 1$ probes. Substituting the log base 2 ratios by the corresponding moving average statistic X_t 's reduces the noise in the data and enhances the shape of peaks and troughs, but not at the expense of over smoothing the data as shown in the middle panel of Figure 1. A nucleosome occupied region is characterized by a series of positive followed by negative slopes or O_t 's, while the boundaries of nucleosomes-linker regions are characterized by steeper slopes. This is evident in the middle panel of Figure 1 and motivates the use of O_t 's to infer nucleosome positions. Detecting jumps in O_t 's via segmentation is a potential approach to map nucleosome occupancy but traditional segmentation approaches do not incorporate the length of nucleosomal DNA. In addition, since the data is obtained from tiling arrays, spatial correlations among observations of nearby probes are expected. To account for the length of nucleosomal DNA and the correlation structure, we propose a non-homogenous HMM (NHMM) based on first order differences O_t 's. Next, we give a detailed characterization of the NHMM architecture.

Consider the state transitions given in Figure 2(a) where N_i 's represent the nucleosome region states, L_i 's represent linker or nucleosome depleted region state and B_i 's represent nucleosome-linker boundaries. The self transitions of N_1 and N_3 is to account for less stable nucleosomes which span a larger region than well-positioned nucleosomes, termed "fuzzy" nucleosomes by Yuan et al. (2005). We introduce state duration $d(i)$ to capture the length of nucleosomal DNA explicitly. Assume that a well-positioned nucleosome (146 base pairs) is characterized by p probes, or equivalently $p - 1$ first order differences. We require

$$\sum_{i \in \{N_{2a}, N_{2b}, N_{2c}\}} d(i) + 2 = p - 1, \quad 0 \leq d(N_{2a}), d(N_{2b}) \leq p - 3,$$

since at least one probe is from N_1 and one is from N_3 out of $p - 1$ probes representing a

nucleosome.

In most cases, the “bump” shape of a nucleosome on tiling arrays is symmetrical, which implies that $d(N_{2a}) = d(N_{2c})$. Moreover, given the state duration constraint, the state transitions can be further simplified as in Figure 2(b) by tying states N_{2a} , N_{2b} and N_{2c} as N_2 with a trinomial duration density:

$$p_{N_2}(d_1, d_2, d_3) = \frac{(p-3)!}{d_1!d_2!d_3!} p_1^{d_1} p_2^{d_2} p_3^{d_3},$$

where $p_1 + p_2 + p_3 = 1$ and $d_1 + d_2 + d_3 = p - 3$.

Let $b_i(O_t)$ denote the emission distribution for observed value at probe $t = 1, \dots, T$ given unknown state $i \in \{N_i, L_i, B_i\}$. We model $b_i(O_t)$ with Gaussian distributions,

$$\begin{aligned} b_{B_N}(O_t) &\sim N(\mu_1, \sigma_{B_N}^2), & b_{N_1}(O_t) &\sim N(\mu_2, \sigma_{N_1}^2), \\ b_{N_3}(O_t) &\sim N(-\mu_2, \sigma_{N_3}^2), & b_{B_L}(O_t) &\sim N(-\mu_1, \sigma_{B_L}^2), \\ b_{L_1}(O_t) &\sim N(-\mu_2, \sigma_{L_1}^2), & b_{L_2}(O_t) &\sim N(0, \sigma_{L_2}^2), \\ b_{L_3}(O_t) &\sim N(\mu_2, \sigma_{L_3}^2), & b_{N_2}(O_{t:t+p-3}) &\sim N(\tilde{\mu}, \Sigma), \end{aligned}$$

where

$$\begin{aligned} \tilde{\mu} &= (\underbrace{\mu_2, \dots, \mu_2}_{d_1}, \underbrace{0, \dots, 0}_{d_2}, \underbrace{-\mu_2, \dots, -\mu_2}_{p-3-d_1-d_2}), \\ \Sigma &= \text{diag}(\underbrace{\sigma_{N_{2a}}^2, \dots, \sigma_{N_{2a}}^2}_{d_1}, \underbrace{\sigma_{N_{2b}}^2, \dots, \sigma_{N_{2b}}^2}_{d_2}, \underbrace{\sigma_{N_{2c}}^2, \dots, \sigma_{N_{2c}}^2}_{p-3-d_1-d_2}), \end{aligned}$$

and $0 < \mu_2 < \mu_1$. The constraint on the mean of emission distributions is to ensure the series

of decreasing positive slopes, zero slopes and followed by increasing negative slopes which characterize the “bump” shape of a nucleosome. In the case of symmetric “bump” shape, the duration density for N_2 reduces to univariate density $p(d_1)$ and

$$\tilde{\mu} = \left(\underbrace{\mu_2, \dots, \mu_2}_{d_1}, \underbrace{0, \dots, 0}_{p-3-2d_1}, \underbrace{-\mu_2, \dots, -\mu_2}_{d_1} \right).$$

The non-parametric discrete duration density assumption implies that the proposed non-homogeneous duration HMM is equivalent to a non-homogeneous HMM with a larger hidden state space. We can recast the state transition in Figure 2(a) as Figure 3 which have the same complexity by considering all possible uni-directional paths transiting from N_1 and incorporating the constraint $\sum_{i \in \{N_{2a}, N_{2b}, N_{2c}\}} d(i) + 2 = p - 1$. We can equivalently let $b_{N_{2a}}(O_t) \sim N(\mu_2, \sigma_{N_{2a}}^2)$, $b_{N_{2b}}(O_t) \sim N(0, \sigma_{N_{2b}}^2)$ and $b_{N_{2c}}(O_t) \sim N(-\mu_2, \sigma_{N_{2c}}^2)$. In scenarios where we have high resolution experiments for mapping nucleosome occupancy such as the 4 base pairs resolution ChIP-chip data of Lee et al. (2007) or 1 base pair resolution ChIP-seq data of Shivaswamy et al. (2008), the “bump” shape of nucleosome is relatively well characterized by a few positive slopes, followed by a plateau and a few negative slopes. In such cases, we can reduce the range of d_1 by removing some uni-directional paths in Figure 3 and thereby simplify the structure of the HMM state transitions.

Since high log base 2 ratios represent regions that are more likely to be occupied by nucleosomes and vice versa for low log base 2 ratios, we model the hidden state transitions as a function of observed log base 2 ratios X_t . Let $a_{i,j}(x) = P(q_{t+1} = j \mid q_t = i, X_{t+1} = x)$ be the transition probabilities from state i to j between probe t and probe $t + 1$ given covariate X_{t+1} . Here, q_t is the hidden state for probe t . To avoid overparametrization, only transitions $a_{B_L, \bullet}(X_{t+1})$, $a_{L_3, \bullet}(X_{t+1})$ and $a_{N_3, \bullet}(X_{t+1})$ are functions of X_{t+1} 's. Other transition probabilities are assumed to be time homogeneous. We employ a logistic regression model

to parametrize the hidden transitions for B_L , L_3 and N_3 :

$$a_{i,j}(X_{t+1}) = \frac{\exp(\gamma_{i,j} + \beta_j X_{t+1})}{\sum_{k=1}^N \exp(\gamma_{i,k} + \beta_k X_{t+1})}.$$

In cases where the data has been median centered at zero, we observe that a simpler version of the non-homogenous transition probabilities for these three hidden states performs well (see case study). That is, we consider

$$a_{i,j}(X_{t+1}) = \begin{cases} a_{i,j}^n, & \text{if } X_{t+1} < 0, \\ a_{i,j}^p, & \text{if } X_{t+1} \geq 0. \end{cases}$$

For instance, we can let $a_{L_3, B_N}(X_{t+1}) = a^{I(X_{t+1} < 0)}$ to impose transition into nucleosome states when $X_{t+1} \geq 0$. The details on model fitting are given in the Supplementary Materials.

4 Simulation studies

Yuan et al. (2005) attributed the heterogeneous nucleosome density to unpredictable trends in hybridization data. They applied the HMM to a sliding window of 40 consecutive probes to address this issue. Hidden states decoding via the Viterbi algorithm was based on average values of the model parameters and posterior probabilities of all windows containing a fixed probe. We referred to this method as sliding window HMM (SHMM). SHMM is computationally intensive and requires one to select the window size, which depends on the trend in hybridization. Yuan et al. (2005) also proposed detrending the data by comparing the magnitude of peak and trough locally to capture low-signal nucleosomes. In particular, for each probe, they considered a window size of 7 probes (\sim size of a nucleosome) centered at the probe and replaced the observed log base 2 ratio by the difference between the median of log base 2 ratios within the window and the minimum log base 2 ratio of the two probes adjacent to this window. They observed that the trend was effectively eliminated using this

procedure. We referred to this method as HMMD (detrending followed by usual HMM to infer nucleosome/linker states).

4.1 Simulation I: Hidden Markov model with trend line

In the first simulation, we generated the data using the HMM hidden states architecture in Supplementary Figure 1(a) (or Figure S1E of Yuan et al. (2005)), in which well-positioned nucleosomes were represented by 6 to 8 probes (N1-N8) and delocalized nucleosomes (D1-D9) covered at least 9 probes. Nucleosome regions were expected to have high log base 2 ratios whereas linker regions had lower values. The hidden state transitions in Yuan et al. (2005) allowed for linker regions (L) to have variable length. Conditioned on the hidden states, the observed log base 2 ratios were generated from Gaussian distributions, with mean 0.7, standard deviation (s.d.) 0.2 for nucleosome states and mean -0.7, s.d. 0.3 for linker state. We illustrated that although we were simulating the observed log base 2 ratios, and not the first order differences, our proposed NHMM was able to map nucleosome positions accurately.

To simulate heterogeneous nucleosome densities, we added a trend line to the simulated data following Yuan et al. (2005). Figure 1 suggests that the underlying trend line in the observed data resembles a curve. Therefore, instead of adding a linear trend line as in Yuan et al. (2005), we let the trend be a sinusoidal curve so that the synthetic data resembles the observed data to a larger extent (Figure 4 top right panel). The bottom left panel of Figure 4 plots the detrended data obtained by comparing peak to trough in a window size of 7 described above. Although this procedure was able to remove the trend in hybridization, it introduced artificial linkers within delocalized nucleosomes and spurious “bumps” within nucleosome depleted/long linker regions and resulting in data with higher noise level. This suggests that applying the same detrending procedure to the whole data is not desirable. On the other hand, a simple smoothing of the synthetic data preserved the “bump” shape that

characterizes a nucleosome (Figure 4 bottom right panel). We considered sinusoidal curves with different periodicity (Supplementary Figure 2) in this simulation study.

4.2 Simulation II: Hidden Markov model with mixture emission distributions

Although adding a trend line results in sythetic data that resembled the actual observed data, it may not be the most realistic model to describe the heterogeneity of nucleosome densities. We considered a more realistic simulation setup to generate nucleosomes with various occupancy levels by using mixture emission distributions for the hidden states. We enlarged the hidden state transitions (Supplementary Figure 4(b)) by introducing low and high (stable) nucleosome states. The stable nucleosomes (N1-N8, D1-D9) were generated from a Gaussian distribution with mean 0.7 and s.d. 0.2. Low-signal nucleosomes (NL1-NL8, DL1-DL8) were generated from a Gaussian distribution with mean 0.1 and s.d. 0.3 and the linker state was generated from a mixture of 3 Gaussian distributions with means -0.3, -0.5, -0.7 and constant s.d. 0.3 with equal mixing proportion. An example of simulated data is shown in Supplementary Figure 3. The middle panel again shows that detrending introduces a higher noise level to the original data.

We simulated observations for 1000 probes according to a tiling design of 50-mer probes overlapped by 30 base pairs covering a 20030 base pair region. In both simulations, we decoded the hidden states using the usual HMM with two emission distributions, one for linker and one for nucleosomes (without differentiating fuzzy/well-positioned, low/high), SHMM, HMMD (detrend first, then apply usual HMM) and our proposed NHMM (on first order differences). The most probable path for each method was decoded via the Viterbi algorithm (Supplementary Materials).

4.3 Results

We compared the performance of each method via the area under a receiver operating characteristic (AUROC) curve, by varying the posterior probabilities of declaring a probe to be in a nucleosome (well positioned and delocalized) state. In addition, we also evaluated the sensitivity and specificity at probe level of the most probably path for each method. The results, averaged over 50 simulated data sets of 1000 probes, are summarized in Table 1.

In both simulations, NHMM has a consistent result and outperforms other methods in both the sensitivity/specificity at the 0.5 posterior probability threshold and AUROC, since its main assumption is the “bump” shape that characterizes a nucleosome and this characteristic is preserved irrespective of the underlying trends in hybridization (Simulation I). HMM consistently tends to declare fewer nucleosomes, resulting in lower sensitivities. On the other hand, in cases where the trend line has larger periodicity, comparing the magnitudes of peaks and troughs is able to remove the trend effect and improves the performance of HMMD, although it is still worse than NHMM. The superior performance of SHMM in Simulation I with larger periodicity is not surprising. When the periodicity is large, the simulated data in each segment consisting of 40 probes is very close to the original hidden Markov model generator with scaled mean in the emission distributions, and therefore fitting a usual HMM to each segment in SHMM agrees with the underlying data generator. However, when the trend line oscillates more frequently (Simulation I) or unpredictable (Simulation II), the performance of SHMM decreases rapidly. This indicates that the sliding window size in SHMM depends heavily on the trend in hybridization. In the actual data analysis, it is hard to calibrate the window size since the exact trend is unknown, and a reasonable number of probes within the window size is required for obtaining reliable parameter estimates in an HMM fit.

Trend	Method	Sensitivity	Specificity	AUROC
$\sin(x/5)$	HMM	0.527 ± 0.095	0.937 ± 0.094	0.718 ± 0.056
	SHMM	0.671 ± 0.042	0.783 ± 0.039	0.821 ± 0.021
	HMMD	0.596 ± 0.064	0.903 ± 0.045	0.786 ± 0.037
	NHMM	0.874 ± 0.051	0.873 ± 0.031	0.962 ± 0.010
$\sin(x/10)$	HMM	0.501 ± 0.065	0.969 ± 0.081	0.727 ± 0.048
	SHMM	0.721 ± 0.048	0.886 ± 0.054	0.870 ± 0.022
	HMMD	0.788 ± 0.043	0.898 ± 0.028	0.949 ± 0.012
	NHMM	0.956 ± 0.028	0.927 ± 0.032	0.986 ± 0.005
$\sin(x/20)$	HMM	0.542 ± 0.100	0.909 ± 0.144	0.717 ± 0.077
	SHMM	0.989 ± 0.006	0.992 ± 0.012	0.997 ± 0.004
	HMMD	0.814 ± 0.032	0.917 ± 0.015	0.917 ± 0.015
	NHMM	0.966 ± 0.025	0.922 ± 0.028	0.988 ± 0.006
$\sin(x/50)$	HMM	0.542 ± 0.086	0.959 ± 0.112	0.738 ± 0.064
	SHMM	0.998 ± 0.003	0.998 ± 0.003	0.999 ± 0.001
	HMMD	0.817 ± 0.031	0.899 ± 0.024	0.963 ± 0.007
	NHMM	0.969 ± 0.025	0.943 ± 0.023	0.988 ± 0.005
mixture emission	HMM	0.564 ± 0.131	0.996 ± 0.010	0.731 ± 0.044
	SHMM	0.834 ± 0.036	0.967 ± 0.022	0.969 ± 0.007
	HMMD	0.571 ± 0.107	0.902 ± 0.080	0.751 ± 0.096
	NHMM	0.928 ± 0.055	0.967 ± 0.016	0.987 ± 0.005

Table 1: *Mean sensitivity, mean specificity and AUROC from the 50 simulations with the corresponding standard errors for each method.* Sensitivity and specificity calculations are based on the most probably path decoding in each method. AUROC illustrates the overall performance across the range of all posterior probabilities cut-offs.

5 Case studies

5.1 Mapping nucleosome occupancy in ChIP-chip data

We illustrated our proposed NHMM on the normalized median log base 2 ratios of the 8 replicates from Yuan et al. (2005). The data was generated from microarrays which consist of 50-mer oligonucleotides probes tiled at 20 base pairs resolution, covering approximately half megabase of the yeast genome. A moving average in a window size of 3 probes was first applied across the whole data as the smoothing step. A well positioned nucleosome (146 base pairs) is represented by at least 6 probes (Yuan et al.; 2005), which implies that $0 \leq d(N_{2a}), d(N_{2b}), d(N_{2c}) \leq 1$. We also assumed that $d(N_{2a}) = d(N_{2c})$. Therefore, the structure of state transitions in the HMM is simplified and given in Figure 5. For this case study, we considered the simpler non-parametric transition probabilities for B_L , L_3 and N_3 :

$$\begin{aligned} a_{N_3, B_L}(X_{t+1}) &= \begin{cases} 1, & \text{if } X_{t+1} < 0, \\ a_{N_3, B_L}^p, & \text{if } X_{t+1} \geq 0, \end{cases} \\ a_{B_L, B_N}(X_{t+1}) &= \begin{cases} a_{B_L, B_N}^n, & \text{if } X_{t+1} < 0, \\ 1, & \text{if } X_{t+1} \geq 0, \end{cases} \\ a_{L_3, B_N}(X_{t+1}) &= \begin{cases} a_{L_3, B_N}^n, & \text{if } X_{t+1} < 0, \\ 1, & \text{if } X_{t+1} \geq 0. \end{cases} \end{aligned}$$

This transition structure implies that if the current state is in a linker region, a positive log base 2 ratios observed in the next probe imposes transition into a nucleosome state. Similarly, if the current state is in N_3 nucleosome state, a negative log base 2 ratios observed in the next probe imposes transition into a linker state. This transition structure appears to be sufficient and works well on the data. We first illustrated that our proposed NHMM is able to detect low-signal nucleosomes in the HIS3 promoter region as shown in Figure 6. The horizontal black line between positions 721871 and 721971 is the low-signal nucleosome annotated in Figure 1(B) of Yuan et al. (2005) which was only identified via ‘‘SHMM &

detrend”. For “SHMM & detrend”, Yuan et al. (2005) first applied SHMM to decode nucleosome positions, followed by further detrending to detect low-signal nucleosomes. This low-signal nucleosome was also identified by others according to Yuan et al. (2005) and in the ChIP-Seq experiment of Shivaswamy et al. (2008), therefore it is not likely to be an artifact of hybridization. Our proposed NHMM is able to map this low-signal nucleosome automatically and accurately without any detrending. We also showed that the duration constraint in nucleosome states in our NHMM architecture is able to distinguish real “bumps” which characterize a nucleosome from spurious small “bumps” at positions 103400 (between nucleosomes 1 and 2) and 104400 (between nucleosomes 6 and 7) in the top panels of Figure 7. The problem with detrending the data by comparing peak and trough within a window size of 7 probes is also visible in this region. As evident in the bottom left panel of Figure 7, detrending introduced more noise to the original data and diminished the distinction between linker and nucleosomes.

To compare the annotation based on our proposed NHMM against the annotation based on HMM, SHMM, SHMM & detrend and HMMD, we used the “hand picked” annotation in Yuan et al. (2005) as the gold standard. The annotations based on SSHM and SSHM & detrend were also from Yuan et al. (2005). Hand picked annotation was based on careful visual inspection (Yuan et al.; 2005), and thus formed a reliable nucleosome map. However, it is inevitable that there may still exist some uncertainties in mapping nucleosome-linker boundaries even by careful visual inspection as shown in Supplementary Figure 4. To account for the one/two probes boundary uncertainties in the “hand picked” annotation, we allowed for one probe margin in defining sensitivity and specificity. That is, suppose that the underlying state for probe i based on “hand picked” annotation is a nucleosome, we declared this probe to be correctly inferred if either one of the probes $i - 1$, i or $i + 1$ was annotated as a nucleosome for each of the method under comparison. To measure the sensitivity of our

proposed method in detecting low-signal nucleosomes, we considered two possible sets of true positives. The first set was defined by using probes annotated as nucleosomes (both low and high signals) in the “hand picked” annotation. The second set was defined by using probes categorized as low-signal nucleosomes by “hand picked” annotation according to Yuan et al. (2005) (that is corresponding to score 0.25 and 0.5 in Yuan et al. (2005)).

Table 2 summarizes the sensitivity and specificity for these methods using “hand picked” annotation as the gold standard. The performance of our proposed method is comparable to SHMM & detrend in Yuan et al. (2005) in terms of sensitivity and specificity when the gold standard includes all annotated nucleosomes. “Sensitivity(both)” was obtained using all annotated nucleosomes as true positives while “Sensitivity(low)” was obtained using annotated low-signal nucleosomes as true positives. SHMM misses a very large fraction of the low-signal nucleosomes, and thereby has extremely poor sensitivity. HMM has a higher sensitivity than SHMM, but a much lower specificity. Although SHMM & detrend is able to capture a significant number of low-signal nucleosomes, it is still significantly less sensitive (with a 0.806 sensitivity) compared to our proposed NHMM (with a sensitivity of 0.909). The methods are comparable in terms of their specificities, except for HMM. We provided an example of a low-signal nucleosome that was still missed by further detrending (i.e., SHMM & detrend) in Figure 8. This low-signal nucleosome was also annotated in high resolution data of Shivaswamy et al. (2008) and this provides evidence against it being a hybridization artifact. The sensitivity analysis illustrates that the proposed NHMM based on first order differences is able to bypass the need for local detrending and automatically map nucleosome positions accurately. HMMD is the worst among all, which again illustrates that detrending the data is a difficult procedure and could potentially distort the signals in the observed data.

We also compared the performance of our proposed NHMM, HMM, HMMD and SHMM

(from Yuan et al. (2005)) via ROC curves, by varying the posterior probabilities of declaring a probe to be in a nucleosome (well positioned and delocalized) state using the low-signal nucleosomes as true positive set. The annotation based on SHMM & detrend in Yuan et al. (2005) was not compared since there is no probabilistic model to describe the detrending and therefore an analogue of posterior probability thresholding is not feasible. The results are shown in Figure 9, which demonstrates that the proposed NHMM based on first order differences performs better than all the other methods.

Method	Sensitivity(both)	Sensitivity(low)	Specificity
HMM	0.905	0.547	0.784
SHMM	0.849	0.231	0.965
SHMM & detrend	0.943	0.806	0.946
HMMD	0.654	0.519	0.753
NHMM	0.937	0.909	0.934

Table 2: *Sensitivity/specificity for the case study.* Sensitivity and specificity are computed by treating the “hand picked” annotation of Yuan et al. (2005) as the gold standard.

5.2 Extension to ChIP-Seq data

Next, we will illustrate the applicability of our proposed NHMM in mapping nucleosome occupancy on ChIP-Seq data from Shivaswamy et al. (2008). Since our modeling framework utilizes first order differences which capture the “bump” shape of a nucleosome and not the observed log base 2 ratios in the emission distribution, it can be applied to first order differences on tag counts/reads in ChIP-Seq data. In Shivaswamy et al. (2008), 514803 uniquely aligned reads were generated for the normal cells via the sequencing technology. We considered the following strategy for mapping nucleosome positions on ChIP-Seq data. Since each of the 27 base pairs Solexa sequencing read corresponds to a mono-nucleosome of size 150-200 base pairs, we first extended these reads to 150 base pairs according to the sequence orientation for both the plus and minus strands. The total reads for each genomic position

is then taken to be the sum of all extended reads at the position, as shown in Supplementary Figure 5. Therefore, the total reads at every 50 base pairs on the genome is analogous to the observed log base 2 ratios in ChIP-chip data of 50 base pairs resolution.

We demonstrated the utility of our proposed NHMM in annotating Chr3:206500-208500 region (Supplementary Figure S1A in Shivaswamy et al. (2008)) using 5 base pairs resolution. Ideally, any two inferred consecutive nucleosomes should be separated by a linker region. The analysis of Shivaswamy et al. (2008) was based on 1 base pair resolution. However, based on the formula in Supplementary Note on ChIP-Seq read requirement in Mikkelsen et al. (2007), the number of sequence reads required for a reasonable sensitivity/specificity is much larger than the actual reads sequenced in Shivaswamy et al. (2008) if we were to analyze the data using a 1 base pair resolution. Despite the analysis based on 1 base pair resolution, some of the nucleosomes inferred by Shivaswamy et al. (2008) were overlapping. For example, in Figure 10 (or Supplementary Figure S1A in Shivaswamy et al. (2008)), the boundaries for nucleosomes 4, 5 and 6 overlap by 5 and 16 base pairs respectively. On the other hand, our proposed NHMM is able to identify the linker region between nucleosomes 4 and 5, but it misses the linker region between nucleosome 5 and 6. The larger extent of overlap between the boundaries of nucleosomes 5 and 6 in Shivaswamy et al. (2008) suggests that inferring them together as a fuzzy nucleosome in our proposed NHMM is reasonable.

6 Discussion

We introduced a non-homogeneous hidden Markov (NHMM) model that automatically maps nucleosome positions and is computationally efficient. The modeling framework utilizes first order differences which capture the “bump” shape that characterize a nucleosome and enable accurate mapping of nucleosome-linker boundaries. The NHMM bypasses the need for local

detrending, which is not a statistically justified procedure (SHMM & detrend) and could still potentially miss low-signal nucleosomes (Figure 8). We also demonstrated the pitfalls of detrending the data with a simple method of comparing peak and trough within a window size covering a nucleosome (HMMD). Such a detrending introduced higher noise levels to the data in both the simulations and a case study on yeast nucleosome occupancies. Modeling the emission distribution on first order differences allows our method to be applicable to both the ChIP-chip and ChIP-Seq data, since the defining characteristic of a nucleosome in both cases is the “bump” shape.

The only preprocessing step required before applying our proposed NHMM in detecting nucleosome positions is data smoothing. We have illustrated in the case studies that simple smoothing such as moving averages in a window size of 3 is generally sufficient. The window size can be adjusted provided it does not over-smooth the nucleosome-linker boundaries. Alternatively, simple local smoothing can be applied to noisy regions, i.e., regions with zigzag/jagged pattern instead of a global smoothing using a larger window size to avoid over-smoothing.

The numerous examples and extensive simulations provided in this paper demonstrate that our proposed method is able to detect linker regions that are represented by only one/two probes, low-signal nucleosomes (Figures 6 and 8) and outperforms currently available methods. Although the underlying architecture of our non-homogeneous HMM is simple, it is effective in detecting nucleosome occupancies in both ChIP-chip and ChIP-Seq data.

Acknowledgements

Supplementary Materials are available at

http://www.stat.wisc.edu/~keles/nucleosome_NHMM_sm.pdf. This research has been supported in part by a PhRMA Foundation Research Starer Grant in Informatics (P.K. and S.K.) and the NIH grant HG003747 (S.K.).

References

- Albert, I., Mavrich, T., Tomsho, L., Qi, J., Zanton, S., Schuster, S. and Pugh, B. (2007). Translational and rotational settings of h2a.z nucleosomes across the saccharomyces cerevisiae genome, *Nature* **446**: 572C576.
- Bernstein, B., Liu, C., abd E.O. Perlstein, E. H. and Schreiber, S. (2004). Global nucleosome occupancy in yeast, *Genome Biology* **5**(62).
- Chakravarthy, S., Park, Y., Chodaparambil, J., Edayathumangalam, R. and Luger, K. (2006). Structure and dynamic properties of nucleosome core particles, *FEBS Letters* **579**(4): 895–898.
- Gupta, M. (2007). Generalized hierarchical markov models for the discovery of length-constrained sequence features from genome tiling arrays, *Biometrics* **63**: 797–805.
- Lee, C., Shibata, Y., Rao, B., Strahl, B. and Lieb, J. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide, *Nature Genetics* .
- Lee, W., Tillo, D., Bray, N., Morse, R., Davis, R., Hughes, T. and Nislow, C. (2007). A high-resolution atlas of nucleosom occupancy in yeast, *Nature Genetics* .
- Liu, C., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S., Friedman, N. and Rando, O.

- (2005). Single-nucleosome mapping of histone modifications in *s.cerevisiae*, *PLoS Biol* **3**(10): 1753–1769.
- Mikkelsen, T., Ku, M., Jaffe, D., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. and Bernstein, B. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *Nature* **448**: 653–560.
- Millar, C. and Grunstein, M. (2006). Genome-wide patterns of histone modifications in yeast, *Nature Reviews Molecular Cell Biology* **7**: 657–666.
- Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M. and Iyer, V. (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation, *PLOS Biology* **6**(3): 618–630.
- Shivaswamy, S. and Iyer, V. (2008). Stress-dependent dynamics of global chromatin remodeling in yeast: dual role for swi/snf in the heat shock stress response, *Molecular and cellular biology* **28**(7): 2221–2234.
- Yuan, G., Liu, Y., Dion, M., Slack, M., Wu, L., Altschuler, S. and Rando, O. (2005). Genome-scale identification of nucleosome positions in *s.cerevisiae*, *Science* **309**: 626–630.

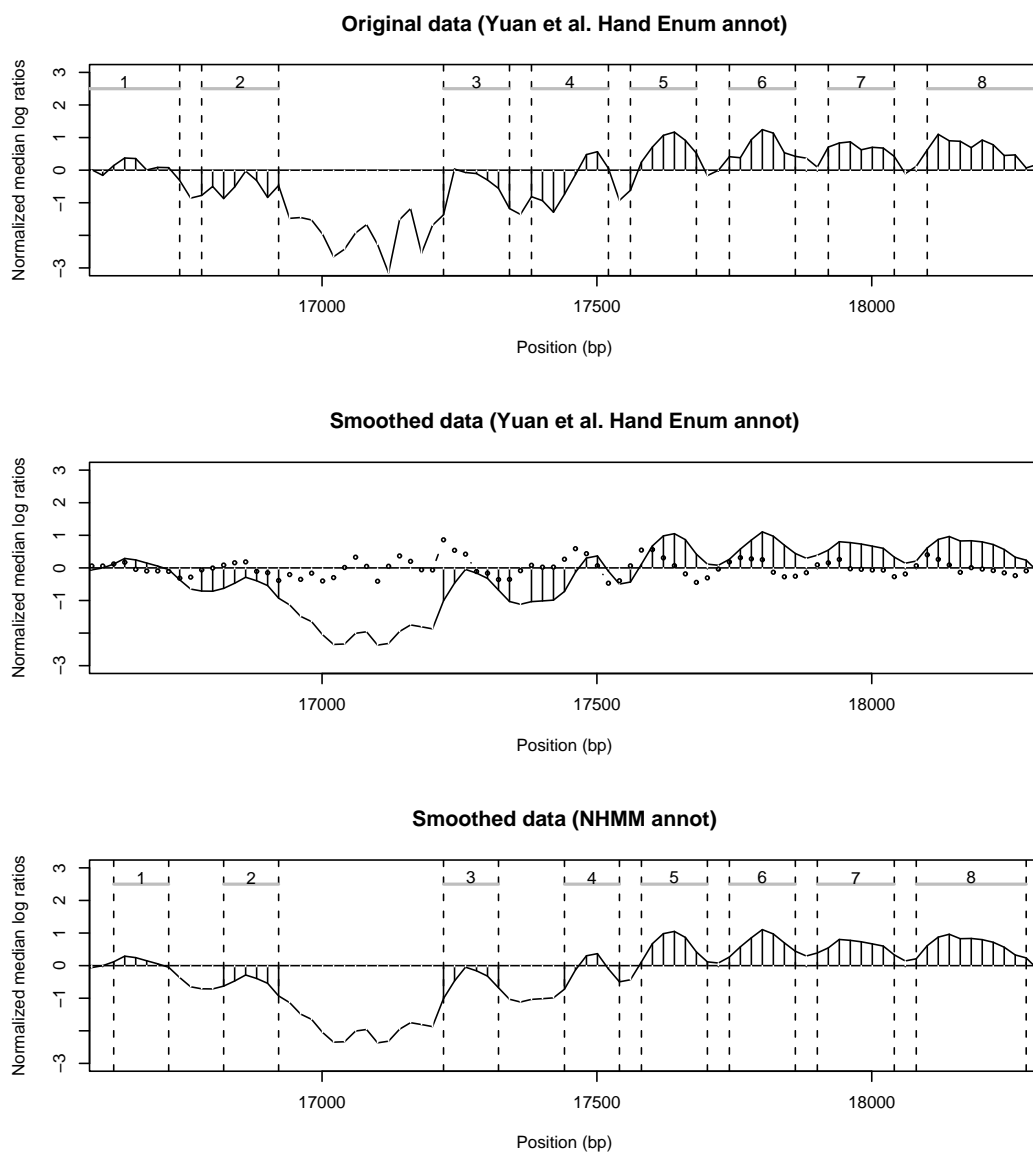
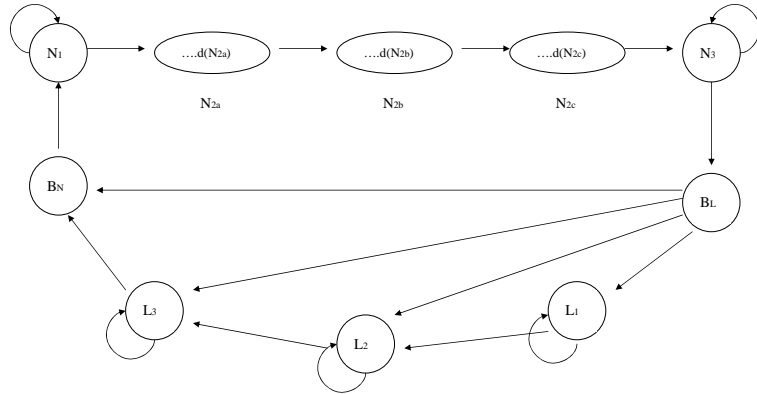
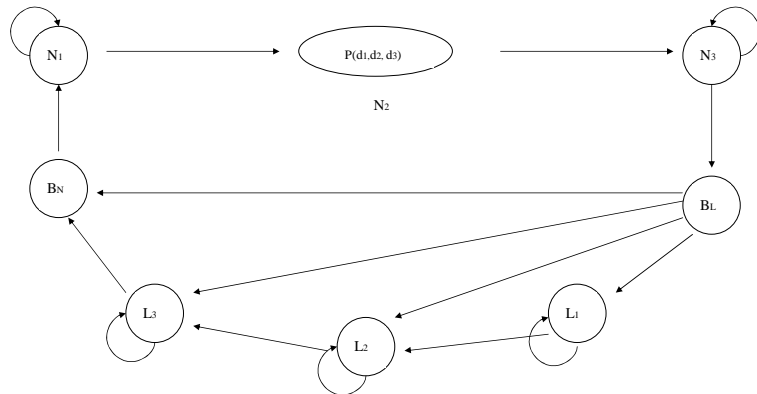


Figure 1: *Typical characteristics of ChIP-chip nucleosome occupancy data from Yuan et al. (2005).* Top panel is the original normalized data tiling a region in chromosome 3. The vertical black solid lines represent probes identified as nucleosome state according to “hand picked” annotation in Yuan et al. (2005). The vertical dotted lines are boundaries separating nucleosome-linker states. Gray horizontal lines at $y=2.5$ are the nucleosomes inferred. Middle panel is the corresponding smoothed data by taking moving averages in a window size of 3 probes and the dots are the first order differences. Bottom panel is based on annotation from our proposed NHMM.



(a)



(b)

Figure 2: *State transition representation in NHMM.* N_i represents nucleosome states, L_i represents linker states, B_N and B_L represent linker-nucleosome and nucleosome-linker boundaries, respectively.

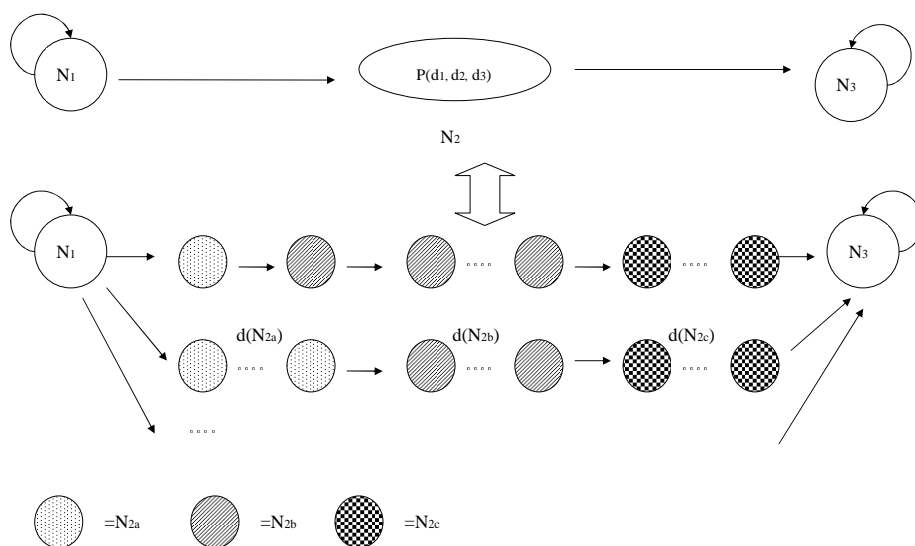


Figure 3: *State transition representation in NHMM.* An equivalent representation of the non-parametric duration HMM of Figure 2(a).

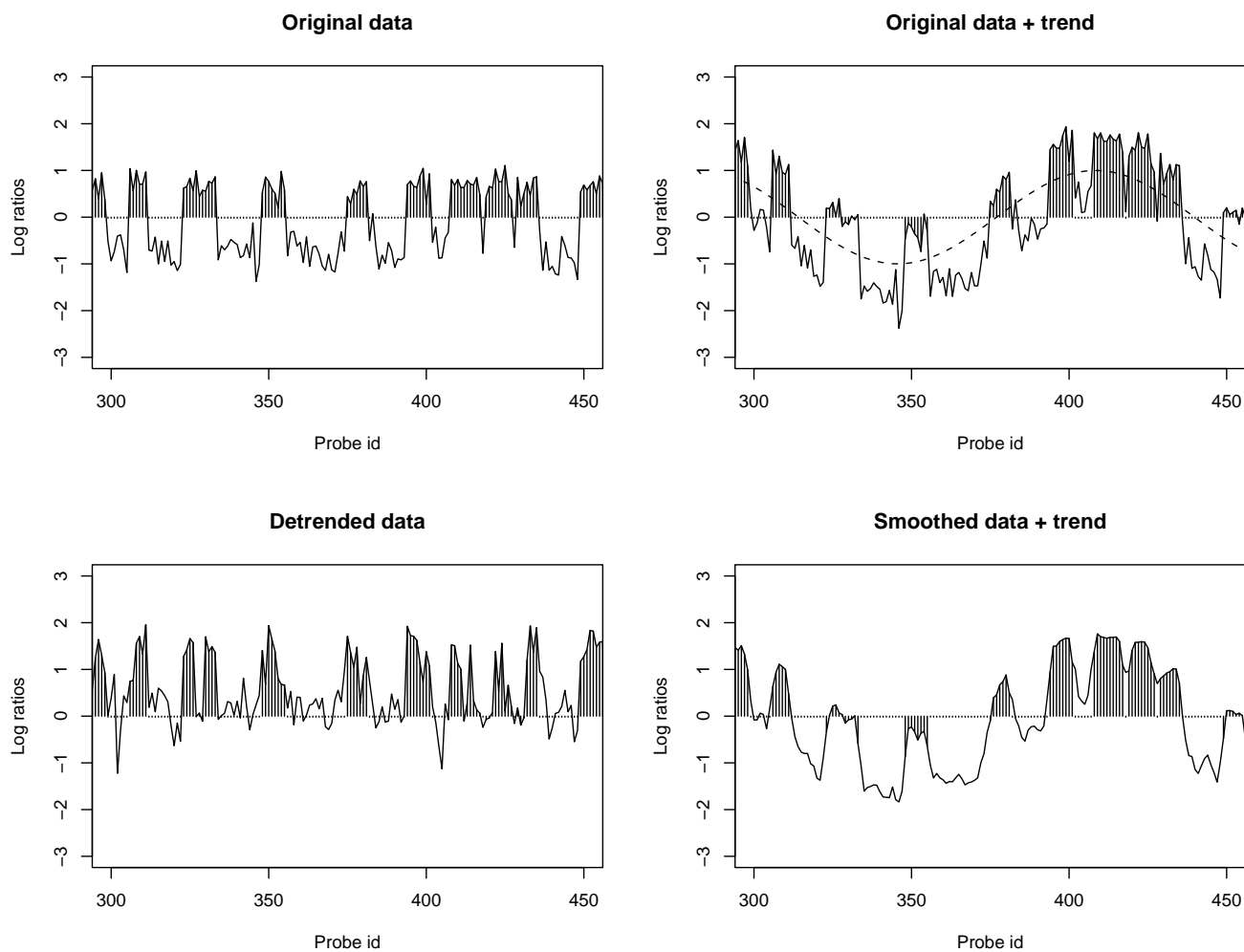


Figure 4: *An example of simulated data from Simulation I.* The dotted line in the top right panel is the trend line. Bottom left panel is the data detrended by comparing peak and trough within a window size of 7 probes. Bottom right panel is the smoothed data. Black vertical lines represent true nucleosome probes.

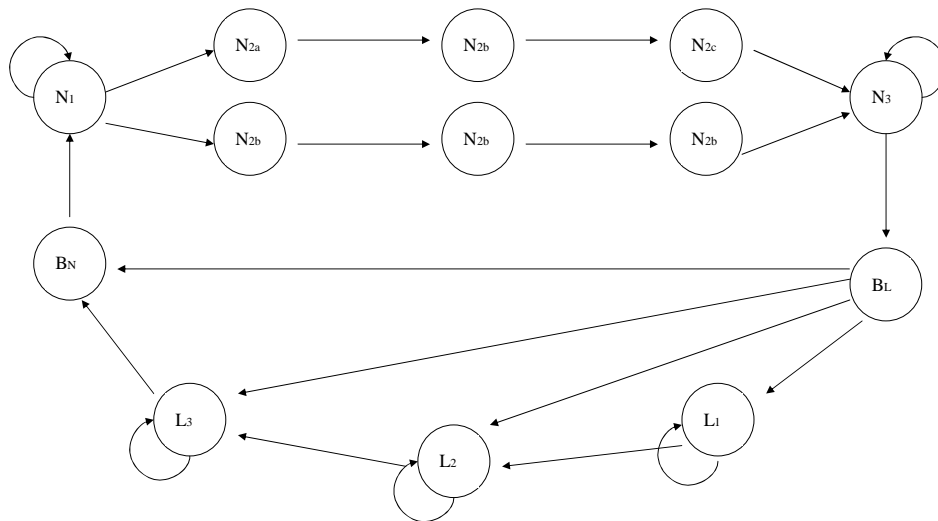


Figure 5: *Simplified state transition representation in NHMM for ChIP-chip data of Yuan et al. (2005). We assume that $d(N_{2a}) = d(N_{2c})$.*

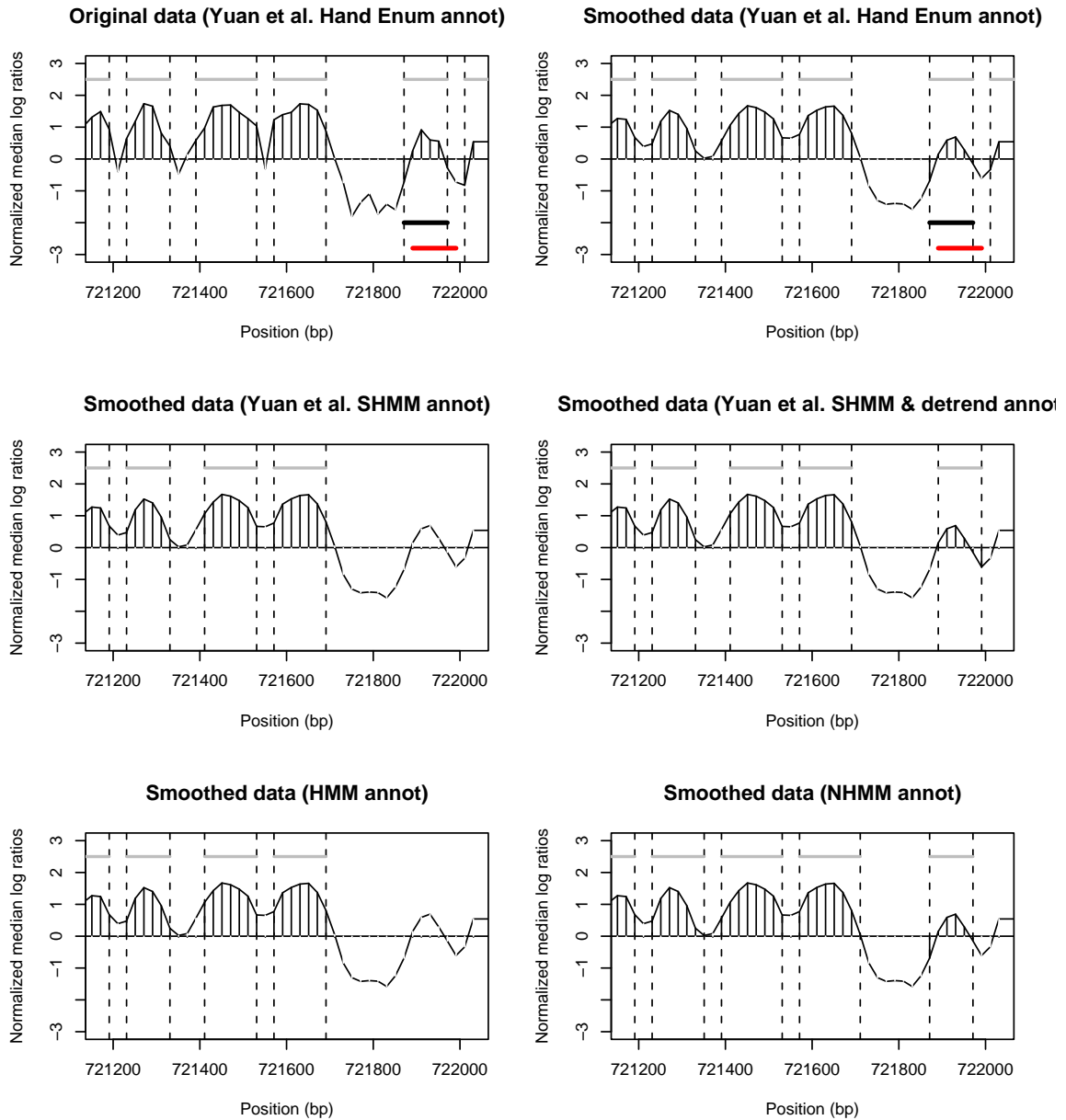


Figure 6: *Nucleosome occupancy in HIS3 promoter*. Top left panel is the original normalized data tiling *HIS3* promoter region and using annotation based on “hand picked” nucleosomes in Yuan et al. (2005). Top right panel is similar to top left panel except that we plot the corresponding smoothed data by taking moving averages in a window size of 3 probes, and annotation is based on “hand picked” nucleosomes in Yuan et al. (2005). Middle left panel is based on SHMM annotation in Yuan et al. (2005). Middle right panel is based on SHMM & detrend annotation in Yuan et al. (2005). Bottom left panel is based on ordinary HMM annotation. Bottom right panel is based on annotation from our proposed NHMM. The black horizontal line between positions 721871 and 721971 in each panel is the low nucleosome identified by Yuan et al. (2005) after further detrending (SHMM & detrend). The red horizontal line is the nucleosome region independently identified by Shivaswamy et al. (2008) using CHIP-Seq technology.

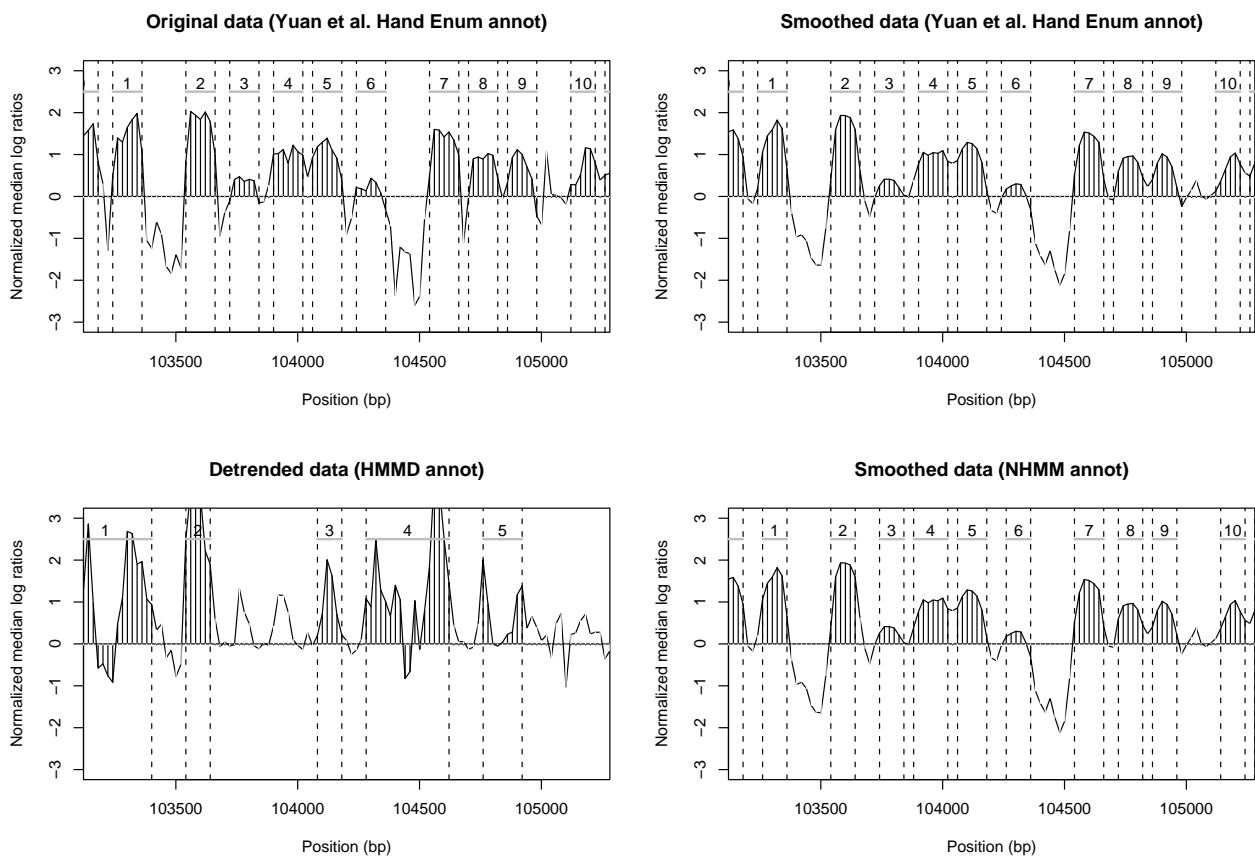


Figure 7: *Nucleosome occupancy for a region in chromosome 3 in Yuan et al. (2005)*. Top panels are based on “hand picked” annotation. Bottom left panel is the detrended data by comparing peak and trough within a window size of 7 probes. Bottom right panel is based on annotation from our proposed model. The spurious “bumps” at positions 103400 (between nucleosomes 1 and 2) and 104400 (between nucleosomes 6 and 7) in the top panels are not picked up by our model. The annotation based on HMMD deviates significantly from the “hand picked” annotation.

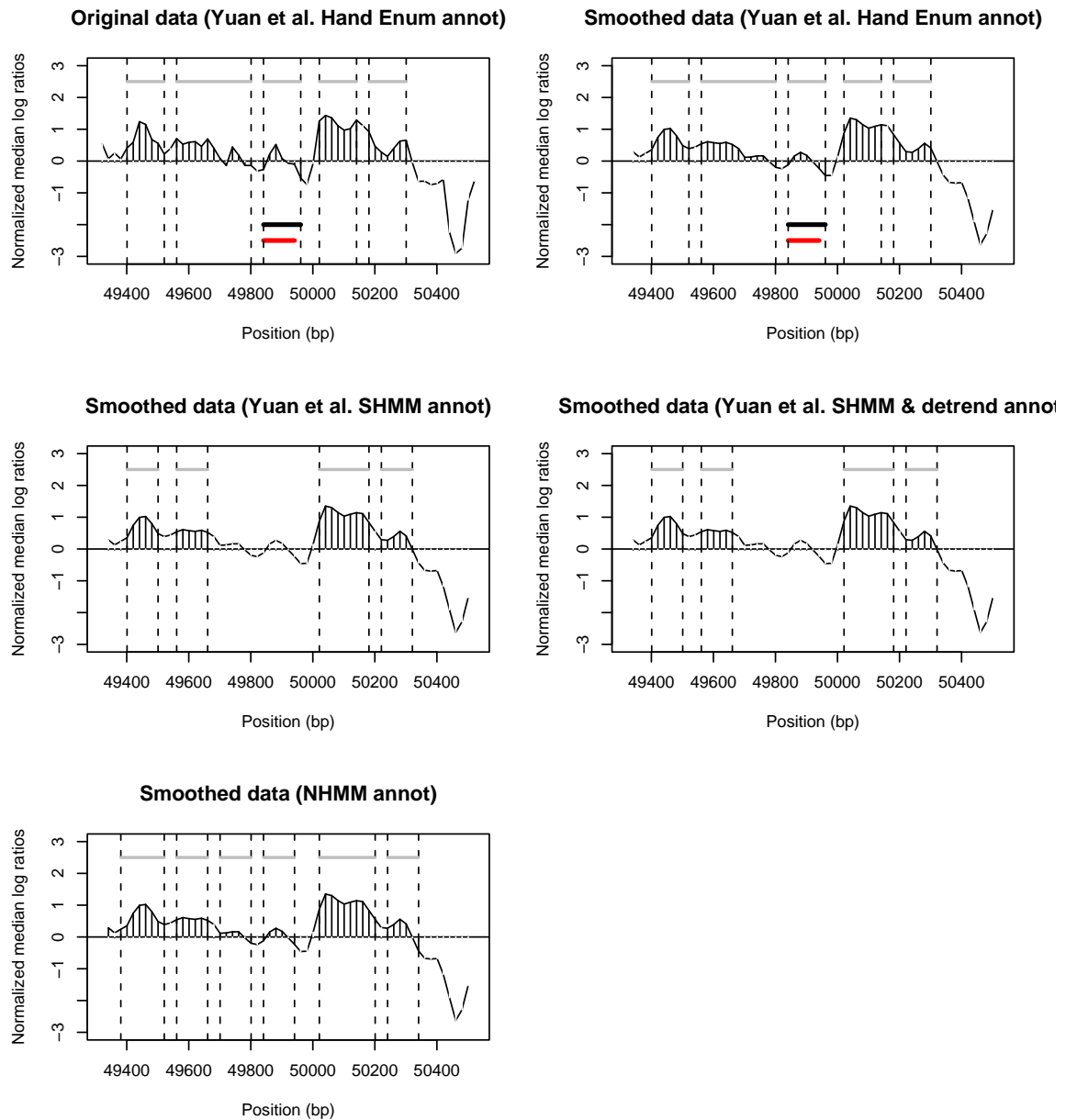


Figure 8: *An example of “hand picked” low-signal nucleosome for a region in chromosome 3.* The black horizontal line between positions 49841 and 49961 is an example of “hand picked” low-signal nucleosome by Yuan et al. (2005). The red horizontal line is the nucleosome region identified by Shivaswamy et al. (2008). SHMM & detrend still misses some of the low-signal nucleosomes, but NHMM is able to capture them.

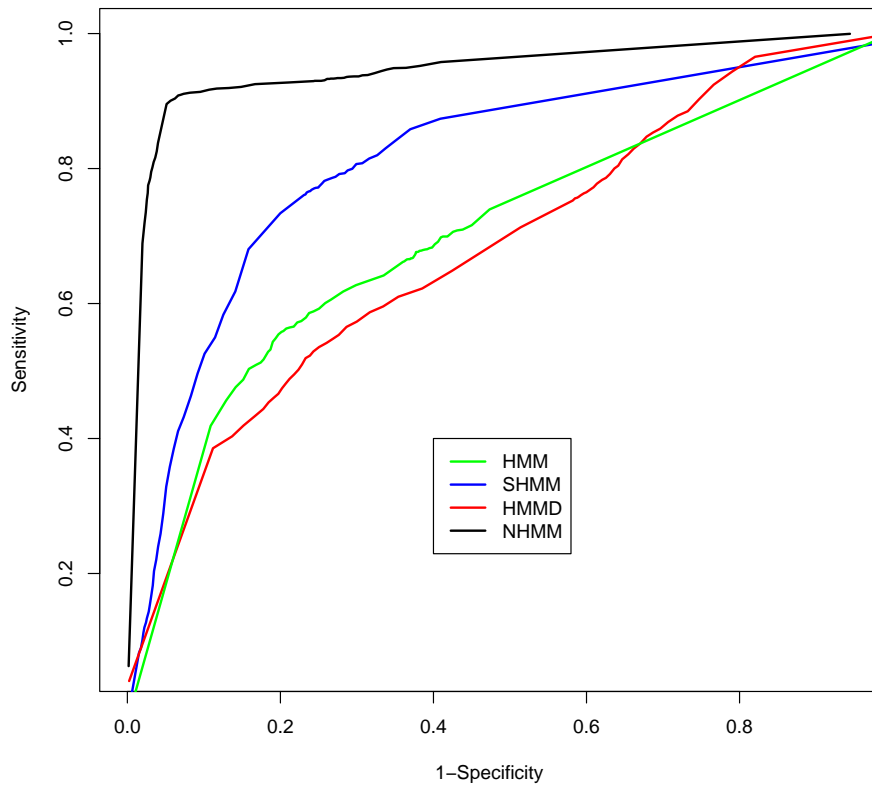


Figure 9: *Receiver operating characteristic (ROC) curve.* Comparison of various methods on ChIP-chip data from Yuan et al. (2005) using the set of “hand picked” annotated low-signal nucleosomes as the true positive set.

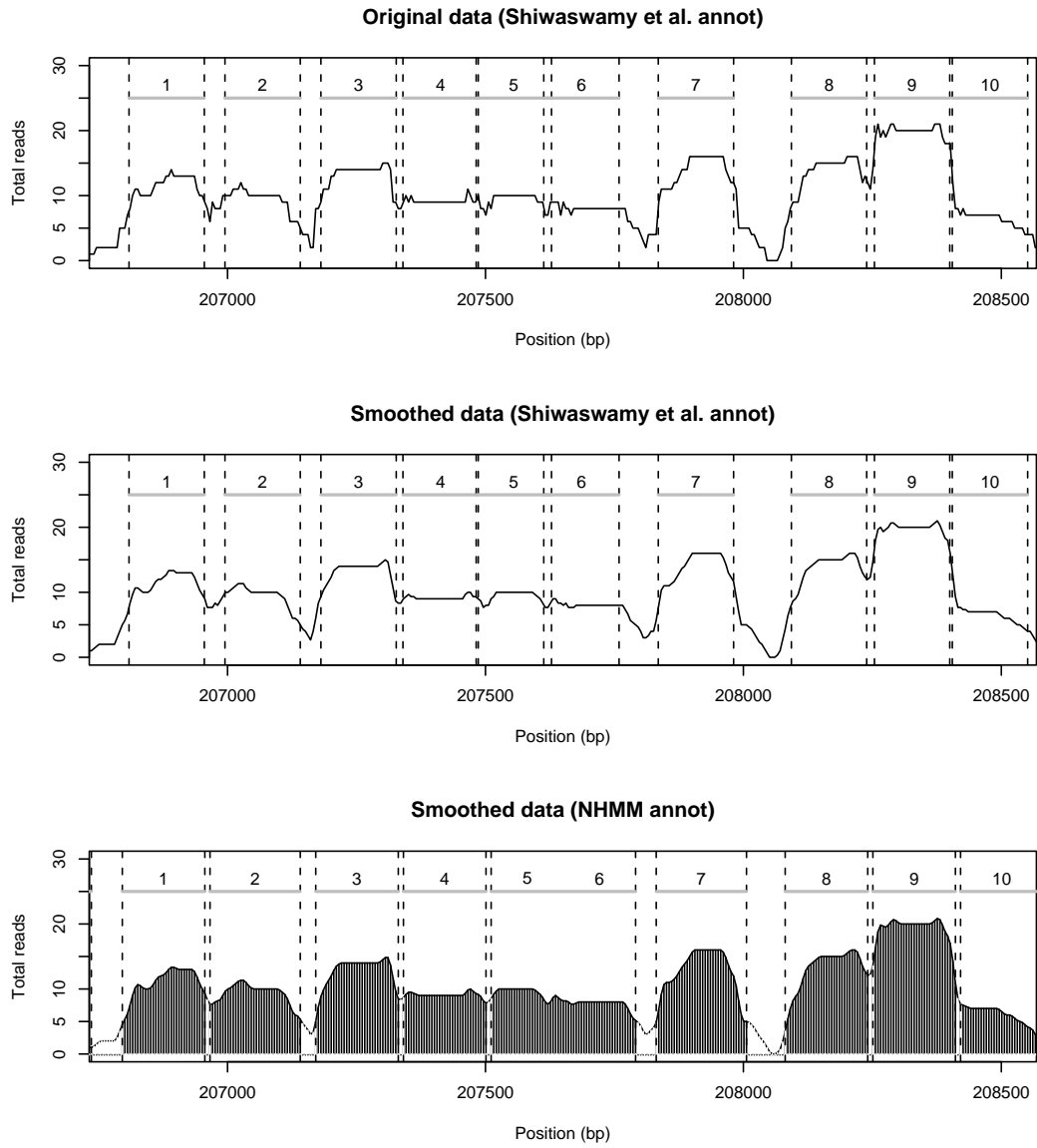


Figure 10: *Nucleosome occupancy for region Chr3:206500-208500 based on the ChIP-Seq data of Shivaswamy et al. (2008).* The top two panel is the total reads at every 5 base pairs from ChIP-Seq data. The middle panel is the corresponding smoothed reads using a window size of 3 positions. The annotation in these two panels is based on Shivaswamy et al. (2008), while the bottom panel is based on annotation from our proposed model. The boundaries for nucleosomes 4, 5 and 6 in the top two panels overlap. NHMM is able to identify the linker between nucleosomes 4 and 5.