DPNuc: Identifying Nucleosome Positions Based on the Dirichlet Process Mixture Model

Huidong Chen, Jihong Guan, and Shuigeng Zhou

Abstract—Nucleosomes and the free linker DNA between them assemble the chromatin. Nucleosome positioning plays an important role in gene transcription regulation, DNA replication and repair, alternative splicing, and so on. With the rapid development of ChIP-seq, it is possible to computationally detect the positions of nucleosomes on chromosomes. However, existing methods cannot provide accurate and detailed information about the detected nucleosomes, especially for the nucleosomes with complex configurations where overlaps and noise exist. Meanwhile, they usually require some prior knowledge of nucleosomes as input, such as the size or the number of the unknown nucleosomes, which may significantly influence the detection results. In this paper, we propose a novel approach *DPNuc* for identifying nucleosome positions based on the Dirichlet process mixture model. In our method, Markov chain Monte Carlo (MCMC) simulations are employed to determine the mixture model with no need of prior knowledge about nucleosomes. Compared with three existing methods, our approach can provide more detailed information of the detected nucleosomes and can more reasonably reveal the real configurations of the chromosomes; especially, our approach performs better in the complex overlapping situations. By mapping the detected nucleosomes to a synthetic benchmark nucleosome map and two existing benchmark nucleosome maps, it is shown that our approach achieves a better performance in identifying nucleosome positions and gets a higher F-score. Finally, we show that our approach can more reliably detect the size distribution of nucleosomes.

1 Introduction

Nucleosomes are the basic subunits of chromatins in eukaryotes. A nucleosome is composed of the 146 bp core DNA twined about 1.65 turns around the histone octamer and a 10∼90 bp of linker DNA between nucleosomes [1]. The nucleosomes are arranged at a regular interval of ∼200 bp along the chromosomes and appear like *beads* on a string. The core particle consists of two copies of each histone protein (e.g., H2A, H2B, H3 and H4) [2]. The linker histone H1 does not make up the *beads*, but it helps stabilize the structure. Nucleosomes package DNA and further compact chromosomes into higher order structure so that chromosomes can be stored in the limited space of a cell. The presence of nucleosomes blocks transcription factors' access to DNA so that the nucleosome-free DNA can be transcribed more easily [3], [4]. But the positions of nucleosomes are not static on chromosomes and they can shift in a certain range in terms of time or cell types [5]. As a result, nucleosome positioning plays an key role in the direct or indirect regulation of DNA replication, transcription, repair and alternative slicing [6], [7], [8], [9]. Thus, the accurate identification of nucleosome positions has a great significance in understanding the mechanisms of biological processes [10], [11].

With the rapid development of next generation sequencing (NGS), massive amounts of data can be generated in a relatively short time, which makes nucleosome mapping to chromosomes at a high resolution possible. Based on chromatin immunoprecipitation (ChIP), the ChIP-seq technique of low cost and high efficiency was developed to analyze genome-wide interactions between DNA and proteins. In the process of ChIP, micrococcal nuclease (MNase) is used to digest nucleosome-free DNA [12]. It cuts the chromatin into DNA fragments and isolates the DNA wrapped around histones after MNase digestion [13]. For single-end sequencing that is still adopted widely, DNA reads are obtained by sequencing 20∼50 base pairs starting from the 5′ end, and then are aligned to the reference genome. DNA fragments can be sequenced only from one direction, and the mapped reads aggregate on both sides of the potential nucleosomes, thus presenting the bimodal pattern. For paired-end sequencing, a single DNA fragment can be sequenced from both 5′ end and 3′ end, so both ends of the DNA fragment can be sequenced. A mapped paired-end fragment indicates the position of nucleosomal DNA of a certain size.

In the past years, some nucleosome maps across the genomes of various model organisms were reported [14], [15], [16], [17]. As experimentally determining of nucleosome positions is of high cost and low efficiency, which limits the subsequent analysis to some extent [18], [19], more and more effort has been put to establish computational models based on sequencing data to predict the positions of nucleosomes. This is also our focus in this paper.

Existing computational approaches for identifying nucleosome positions based on sequenced reads can be roughly classified into three types: peak calling, template matching and data mining based approaches.

MACS [20], QuEST [21] and PeakSeq [22] are representatives of peak calling methods. They usually consist of the following basic components: the generated signal profile, background noise model, peak calling criteria and ranking

- *H. Chen and J. Guan are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China. E-mail: jhguan@tongji.edu.cn, gjlmile@gmail.com.*
- *S. Zhou is with the Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai 200433, China. E-mail: sgzhou@fudan.edu.cn.*

Manuscript received 7 Mar. 2015; accepted 17 Apr. 2015. Date of publication 25 May 2015; date of current version 4 Dec. 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2430350

of the called peaks. They are good at detecting well-positioned nucleosomes, but cannot handle the nucleosomes with complex configurations well, where nucleosomes do not keep phased and show fuzzy or overlapped in the same gene across a cell population [23], [24], [25]. So they are usually applied to the analysis of transcription factor binding sites, which have shorter length and much simpler configurations. In addition, peak calling methods require the estimation of read shift distance or read extension length, which has a significant influence on the final detection result [26]. However, it is difficult to do such estimation accurately in advance. What is worse, the information of discrepancy between forward and reverse strands after read shift or extension will be lost. As a result, these methods cannot even provide additional information of nucleosomes, such as the nucleosome size. And some of them cannot effectively filter two types of false peaks from the preliminary peaks generated at the peak calling step: the false peaks based on a single strand and the false peaks formed by duplicate accumulation of only one or a few reads at one site [27].

Typical template matching approaches include TF [28] and Nucleofinder [29]. They usually first introduce a series of representative templates (also called *models*) to depict the distribution of reads. Then, they calculate the defined statistical indicator of each template or each possible combination of templates for each possible nucleosome. Finally, the nucleosomes are identified by capturing the templates with the largest statistical indicator.

The TF method generates seven representative templates, which correspond to possible read distributions flanking the potential nucleosomes [30]. It then constructs the *heat map* of correlation coefficient through various offsets between forward and reverse templates. The nucleosomes of a certain width are identified according to the peak points in the heat map and under the constraint of a prespecified overlap threshold, they are finally arranged on the chromatin by a greedy strategy. TF can provide the size of nucleosomes, but its greedy strategy limits the accuracy of size estimation [31]. Furthermore, it does not design any specific strategy for filtering noise, so it cannot handle the background noise well. The Nucleofinder method introduces eight models including background and enriched areas. The model, denoted as *background, enriched, background*, is defined as a well-positioned nucleosome. According to the assumed distributions of reads, it calculates the marginal likelihood in the 150 bp sliding region for each model. If the *background, enriched, background* model has the largest marginal likelihood among the eight models, the corresponding region is recognized as a nucleosome. This method performs well in eliminating experimental bias and has a high specificity by introducing the control samples. But it obtains the center position of nucleosome according to a fixed nucleosome size, which makes it lose the accurate information of nucleosome size and thus be not suitable for detecting nucleosomes in complex conditions. It provides confidence score only for well-positioned nucleosomes and does not consider overlapping configurations. Finally, both TF and Nucleofinder cannot provide quantified fuzziness of nucleosome positions.

A representative method based on data mining is NOrMAL [31]. This method builds a parametric probabilistic model for the mapped reads and uses expectation maximum (EM) to calculate the parameters of the gaussian mixture model. For each component of the mixture, the mean and variance are defined as the center and fuzziness of nucleosomes respectively, while the weight is regarded as the confidence score of nucleosomes. NOrMAL can extract the nucleosome size from the parameters. Currently, it provides the most detailed information of nucleosomes. NOrMAL requires the prior size of nucleosomes to choose the cluster number. However, the nucleosome size varies substantially under different experiment conditions, so it is difficult to accurately estimate in advance. Meanwhile, NOrMAL is sensitive to input parameters, some small fluctuations of the parameters may lead to significant difference in the final result. In addition, it does not provide any mechanism to effectively handle the background noise introduced by experiments.

In this study, we present a novel approach for identifying nucleosome positions based on the Dirichlet process (DP) mixture model. We call the new approach ***DPNuc***, which is the abbreviation of ***D***irichlet ***P***rocess mixture model based ***Nuc***leosome positioning. Our approach can overcome the shortcomings of existing methods mentioned above. Here, Markov chain Monte Carlo (MCMC) simulations are applied to estimating the parameters of the mixture model with an unknown number of components. The $5'$ end positions of the mapped reads are directly taken as the input of our method without read shift or extension. We identify the clusters piled up by reads on the forward and reverse strands, respectively. Then the concept *support reads* is introduced for each cluster, and the number of *support reads* is calculated. The background noise is modeled as a Poisson distribution where a dynamic parameter λ_{local} is adopted [20]. Only the clusters that have a significant number of *support reads* (p-value $\leq 10^{-5}$ by default) will be kept. In such a way, we can filter the background noise well. After the identification of nucleosome borders on each strand, we merge the borders. Then, we match the resulting borders within a specified range according to a proposed matching strategy. Finally, the detected nucleosomes are merged if the overlap between them is above a prespecified overlap threshold.

We conduct experiments on the datasets of Saccharomyces cerevisiae of different sequence depths and different experiment conditions, and the datasets of paired-end reads of mouse embryonic stem cells (ESCs). Compared with three state-of-the art methods, including TF [28], NOrMAL [31] and Nucleofinder [29], our method can provide more detailed information about the reported nucleosomes, and performs better in identifying nucleosomes with complex overlapping configurations. Comparing the identified nucleosomes against a synthetic benchmark nucleosome map of mouse ESC and two existing benchmark nucleosome maps of Saccharomyces cerevisiae, our method gets the highest F-score, which demonstrates that our method does better than the existing methods in identifying nucleosome positions. We also show that our method can detect the changes of experiment conditions and estimate the size distribution of nucleosomes more reliably.

For a quick understanding of the difference between our approach and the existing methods, we present a general and qualitative comparison of our approach with three

TABLE 1
A General and Qualitative Comparison of Our Approach with Major Existing Approaches

Approach	Need to input prior size of nucleosomes?	Can output nucleosome size?	Can output nucleosome fuzziness?	Applicable to complex configurations?	Performance (F-score)
TF [28]	No	Yes	No	Yes	High
Nucleofinder [29]	Yes	No	No	No	Middle
NOrMAL [31]	Yes	Yes	Yes	Yes	Low
DPNuc (this paper)	No	Yes	Yes	Yes	Highest

major existing methods in Table 1 from four aspects: whether or not the method need prior size of nucleosomes as input? whether or not the method can output the size and fuzziness of nucleosomes? whether or not the method can detect nucleosomes with complex configurations? and how does the method perform? Among all the compared methods, our method is the only one that holds all the four merits: need no prior size of nucleosomes as input, can output the size and fuzziness of nucleosomes, can detect nucleosomes with complex configurations, and has the highest F-score.

2 Methods

In this section, we present the proposed method DPNuc. Fig. 1 shows the pipeline of DPNuc, which consists of the following functional modules: *preprocessing*, *normalization*, *modeling read distribution based on the Dirichlet process mixture model*, *MCMC simulations*, *kernel density estimation (KDE)*, *background noise filtering*, *border merging*, *border matching*, *nucleosome identification and merging*. Here, MCMC simulations, KDE, background noise filtering and border merging constitute the major steps of *nucleosome border identification*; border identification and matching, nucleosome identification and merging constitute the main steps of the *nucleosome detection* process.

In what follows, we first introduce the data used in this study, and then elaborate the major techniques of the DPNuc approach in detail.

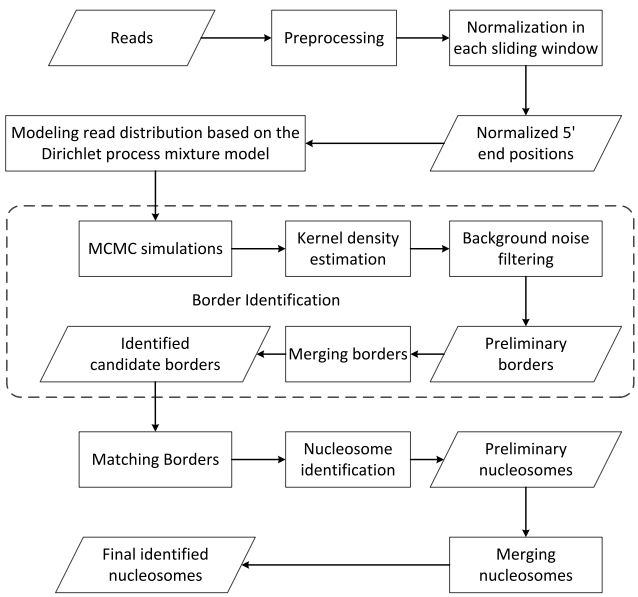


Fig. 1. The pipeline of DPNuc. Here, the rectangles indicate functional modules, and the parallelograms indicate input/output modules. The modules in the dashed box are for *border identification*.

2.1 Data

In this study, we use both MNase-based single-end sequencing data of Saccharomyces cerevisiae and MNase-seq paired-end reads of mouse ESC to identify nucleosome positions. All the data used are summarized in Table 2.

The first dataset is from [32], which is termed *Dataset-1*. It has six biological replicates grown in YPD medium, including four non-crosslinked and two crosslinked. The MNase-seq data is considered the most deeply sequenced among the published datasets of yeast [36]. Here, we choose YPD_-NOCL_R4, which has the largest sequence depth among the six replicates. Thus, it is possible for us to get more accurate and detailed nucleosome map.

The second dataset is from [28], which is called *Dataset-2*. Here, the gel-purified mononucleosomal DNA reads are generated at two different titration levels, that is, typical digestion (10 μL) and overdigested (15 μL).

The third dataset is from [33], which is named *Dataset-3*. It consists of the mapped paired-end nucleosomal DNA from the mouse ESCs. The paired-end reads, which can provide the length of nucleosome DNA, are obtained after digesting the linker DNA using MNase.

Besides the above datasets, two precise nucleosome maps of Saccharomyces cerevisiae [34], [35] are used as

TABLE 2
A Summary of the Data Used In This Paper

Name	Authors	Source	Description
Dataset-1	Kaplan et al. [32]	GEO accession: GSM351492	the most deeply sequenced dataset of yeast
Dataset-2	Weiner et al. [28]	GEO accession: GSM461562	typical digested MNase-seq genomic DNA
		GEO accession: GSM461563	overdigested MNase-seq genomic DNA
Dataset-3	Teif et al. [33]	GEO accession: GSM1004653	the mapped paired-end nucleosomal DNA from mouse embryonic stem cells
Map 2008	Mavrich et al. [34]	http://genome.cshlp.org/content/early/2008/06/12/gr.078261.108/suppl/DC1	the nucleosome map was generated according to the widely accepted barrier model
Map 2012	Brogaard et al. [35]	http://www.nature.com/nature/journal/v486/n7404/full/nature11142.html	the most precise and detailed nucleosome map at present

benchmark to verify the nucleosomes identified by our method and the three compared methods.

A preprocessing step is performed as in [20]. The duplicate reads whose counts exceed the sequence depth are removed (the p-value of binomial distribution is set to 10^{-5} by default). We take the $5'$ end position of each read without shift or extension as the initial input of our method, because we do not know the exact average size of nucleosomes in advance.

2.2 The Dirichlet Process Mixture Model

In our approach, we model DNA read distribution by the Dirichlet process mixture model. After the MNase-seq process, we obtain a large number of mapped nucleosomal reads. For the ith ($i \leq N$) DNA fragment (or read) in the forward (reverse) strand of a chromosome (here N denotes the total number of reads in the forward or reverse strand), we define x_i as the $5'$ end position of the ith read, and get a position set $X = \{x_1, \ldots, x_N\}$ in the forward (reverse) strand. Assume that each $x_i \in X$ is modeled as a mixture of K parametrized normal distributions where K is unknown. The set $X = \{x_1, \ldots, x_N\}$ have a corresponding set of latent variables $Z = \{z_1, \ldots, z_N\}(z_i \in [1, K])$, where z_i indicates that the ith read belongs to the left (right) border of the z_ith possible nucleosome. The K normal distributions are parametrized by $\Theta = \{\theta_1, \ldots, \theta_K\}$, where $\theta_{z_i} = \{\mu_{z_i}, \sigma^2_{z_i}\}$, μ_{z_i} and $\sigma^2_{z_i}$ represent respectively the mean value and the variance of the left (right) border of the z_ith possible nucleosome. The set $\Pi = \{\pi_1, \ldots, \pi_K\}$ is applied to denoting the weights of the K distributions and $\sum_{k=1}^{K} \pi_k = 1(k \in [1, K])$. Thus, the mixture model for DNA reads in the forward (reverse) strand can be expressed as follows:

$$p(x_i|\Theta, \Pi, K) = \sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \sigma_k^2). \quad (1)$$

We calculate the parameters Θ by Bayesian random sampling. The Dirichlet process is taken as a prior distribution for the primary parameter μ_{z_i} of normal distribution to build the Dirichlet process mixture model. In the DP mixture model, the prior distribution function G is uncertain and drawn from a Dirichlet process $G \sim DP(\alpha, G_0)$, where α is the concentration parameter and G_0 is the base distribution. Each μ_{z_i} is drawn independently and identically from G while $\sigma_{z_i}^{-2}$ has the Gamma distribution parametrized by (a, b). The Dirichlet process mixture model based on the position data X can be expressed as below:

$$\begin{aligned} x_i &\sim N(x_i|\mu_{z_i}, \sigma^2_{z_i}), \\ \mu_{z_i} &\sim G, \\ G &\sim DP(\alpha, N(0, 1)), \\ \sigma_{z_i}^{-2} &\sim Gamma(a, b). \end{aligned}$$

Above, we use a Bayesian nonparametric method that employs MCMC simulations to solve the conjugate normal-normal DP mixture model with an uncertain number of components [37], [38].

2.3 The Sliding Window

To use the DP mixture model, we scale the initial position data X to the range $\{0, 1\}$ as a preprocessing step of MCMC simulations. The position data in the whole chromosome have a large span, and if the data are directly normalized into the range $\{0, 1\}$, the difference between any two positions x_i and x_j may be smaller even than the specified threshold of read distribution variance used in nucleosome identification (see Eq. (1)). So we apply a sliding window along either strand of a chromosome and normalize the position data in each sliding window.

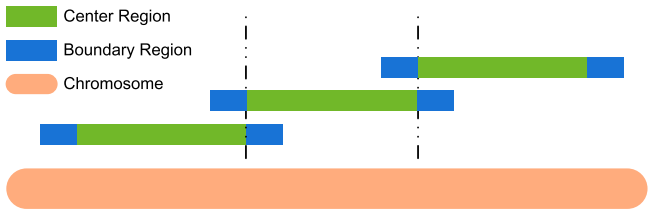


Fig. 2. Three consecutive sliding windows on a chromosome. The sliding window consists of three parts: the center region and two flanking boundary regions. The length of center region is selected as the step size of sliding windows.

Each sliding window consists of a center region and two flanking boundary regions, as shown in Fig. 2. The length of center region is selected as the step size of sliding windows. So for two consecutive windows, their overlap is a right boundary region plus a left boundary region, and the right boundary region of the first window falls in the center region of the second window. We apply the Dirichlet process to the whole region of each sliding window to cluster data. But for all the resulting clusters, we choose only those falling in the center region. The introduction of flanking boundary regions ensures that the clusters detected in the center regions are reliable. So we consider all the clusters in each center region. As the center regions of all windows cover almost the whole chromosome (except for the left boundary region of the first window and the right boundary region of the last window), we can detect almost all possible clusters along the chromosome. Finally, we discard the duplicate clusters detected in neighboring windows. The default sliding window size is 1,000 in this study. A too large window size may make the distance between two adjacent positions be smaller than the prespecified threshold of read distribution variance. As we carry out MCMC simulations in each sliding window, a too small window will increase the computation burden.

2.4 Nucleosome Detection

2.4.1 *Identifying Borders*

In each sliding window, we conduct a number (1,500 in this study) of iterations of MCMC simulations, and keep the results of the last m ($m = 1000$ in this study) iterations. For each iteration, we can get a set $M = \{\mu_1, \ldots, \mu_{K_i}\}$ and a set $\Sigma^2 = \{\sigma_1^2, \ldots, \sigma_{K_i}^2\}$, where K_i represents the number of distributions (corresponding to K_i clusters) in the ith iteration. Note that the value of K_i may be different in different iterations. μ_j and σ_j^2 ($j \in [1, K_i]$) represents the mean value and the variance of the jth distribution generated in the ith iteration of MCMC simulation. Combining the parameter sets M and Σ^2 obtained in the last m iterations, we get the resulting sets $M' = \{M_1, \ldots, M_m\}$ and $\Sigma^{2'} = \{\Sigma_1^2, \ldots, \Sigma_m^2\}$.

We then apply the Gaussian kernel density estimation to M' as follows [39]:

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{l=1}^{n}\kappa\left(\frac{x-\mu_l}{h}\right) \tag{2}$$

where $\kappa(.)$ is the Gaussian kernel function as below:

$$\kappa(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}, \tag{3}$$

n is the total number of elements in the set M', and h is the bandwidth that is selected empirically so that h can keep consistent in different sliding windows.

With the result of KDE above, we take the locations of peaks above the averaged density of the resulting KDE curve, those locations constitute a set of final mean values of read distributions: $\mu = \{\mu_1, \ldots, \mu_K\}$. Here, K is the number of peaks above the averaged density of the KDE curve. In what follows, we try to find the K corresponding variance values of read distributions.

For each $\mu_p \in \mu$ $(p \in [1, K])$, we search the set M' for a set of μ_p's neighboring mean values $M_{neighbor}$ such that each element of $M_{neighbor}$ is within a specified threshold T_μ (by tuning, we set $T_\mu = 0.005$ in our experiments) away from μ_p. For each mean value in $M_{neighbor}$, we can search a corresponding variance value in $\Sigma^{2'}$, these searched variance values constitute a set of standard deviation values $\Sigma_{neighbor}$. We then apply KDE to $\Sigma_{neighbor}$, and the location of the maximum density of the resulting KDE curve is identified as the standard deviation σ_p, corresponding to μ_p. In such a way, we can get the set of K final standard deviation values $\sigma = \{\sigma_1, \ldots, \sigma_K\}$, which corresponds to $\mu = \{\mu_1, \ldots, \mu_K\}$. Finally, the K final mean values $\mu = \{\mu_1, \ldots, \mu_K\}$ are mapped to the original chromosome, and the corresponding locations on the chromosome are taken as the mean values of distributions of reads' $5'$ end positions on each strand, which are regarded as borders of possible nucleosomes.

To illustrate the process above, we give an example in Fig. 3. Fig. 3B shows the coverage profile of short reads (36 bp) in a sliding window on chromosome II with the forward strand on top and the reverse strand at bottom. The accumulated reads on top indicate the potential left borders of nucleosomes, and similarly those at bottom indicate the potential right borders of nucleosomes. Figs. 3A and 3C describe respectively the kernel density estimates (KDEs) constructed from M' on the forward (Green Curve) and the reverse strand (Red Curve) in the same sliding window. As we scale the position data in the sliding window to the range $\{0, 1\}$ during the preprocessing step, all elements in the resulting M' also fall into $\{0, 1\}$.

In Fig. 3, we can see that the peaks of the green density curve (in Fig. 3A) conform well to the green peaks of coverage profile (top of Fig. 3B), and similarly the peaks of the red density curve (Fig. 3C) conform well to the red peaks of coverage profile (bottom of Fig. 3B).

Now we introduce the concept of *support reads*. For the ith identified normal distribution parameterized by μ_i and σ_i, the cumulative distribution function (CDF) is evaluated by

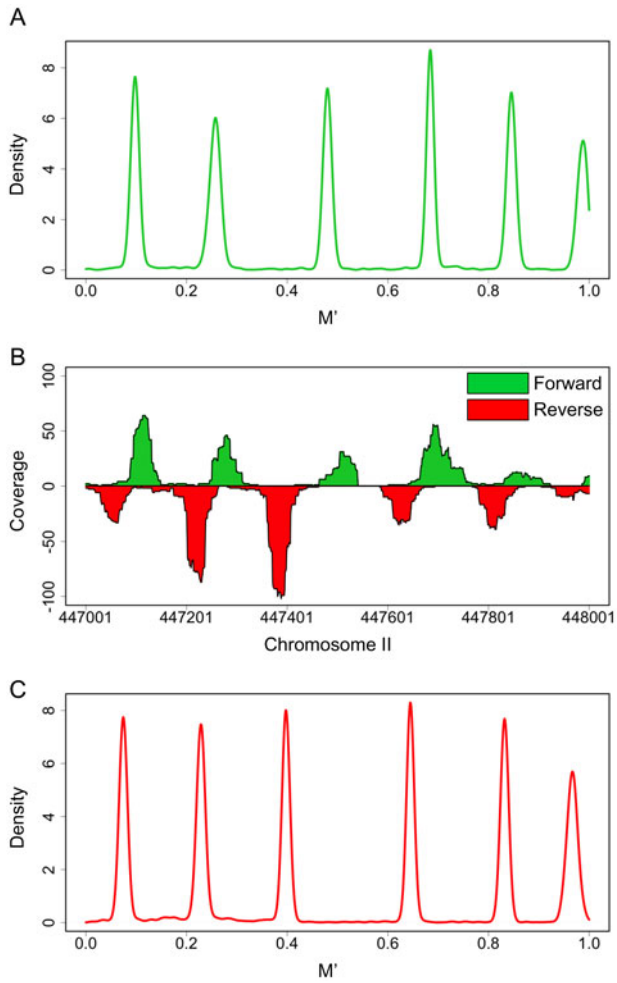


Fig. 3. The results of MCMC simulations in one sliding window. (A) The KDE of M' on the forward strand. The default bandwidth is set to 0.006. (B) The coverage profile formed by mapped reads. The length of reads is 36 bp. (C) The KDE of M' on the reverse strand. The default bandwidth is set to 0.006.

$$F_i(x; \mu_i, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}}\int_{-\infty}^{x} exp\left(-\frac{(t-\mu_i)^2}{2\sigma_i^2}\right)dt. \tag{4}$$

We calculate the quantiles Q_1 and Q_2 where $F_i(Q_1) = 0.05$ and $F_i(Q_2) = 0.95$, respectively. The reads whose $5'$ end positions fall in the region between Q_1 and Q_2 are regarded as *support reads*. They are considered to contribute to the ith peak of the coverage profile.

As there exists experimental background noise, we need still to filter *fake* peaks caused by noise. Here, we model background reads along a chromosome as a Poisson distribution with the parameter λ (the p-value of Poisson distribution is 10^{-5} by default) [27]. As in [20], λ is not estimated from the whole chromosome, instead it is dynamic. For the ith identified distribution, λ is the largest value among λ_{bg}, λ_{5k} and λ_{10k}, where λ_{bg} is estimated from the whole chromosome, λ_{5k} and λ_{10k} are estimated respectively from the mean value-centered 5 and 10 kilobases regions. If the number of support reads from the ith distribution meets the statistical

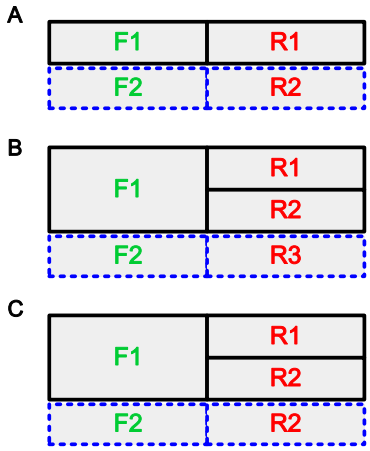


Fig. 4. The three situations of our border matching strategy to handle. The black solid boxes represent the borders in the memory, which are being processed. The blue dashed boxes represent the borders to be processed. Boxes with green label denote the borders on the forward strand and boxes with red label denote the borders on the reverse strand.

significance, we recognize this distribution as a candidate border. In such a way, the distribution caused by background noise can be removed.

Given the prespecified overlap ratio threshold ξ between two adjacent nucleosomes and the range of nucleosome size $[ns_{min}, ns_{max}]$, two adjacent candidate borders are merged if the distance between them is less than $(1-\xi)\cdot ns_{min}$. We then combine the support reads of both candidate borders as the support reads of the resulting border; and the mean value of the merged support reads is calculated as the resulting border's position. This process is repeated till no more candidate borders can be merged.

2.4.2 Matching Borders

After identifying all candidate left and right borders of nucleosomes, we select paired borders for each possible nucleosome on the forward and the reverse strands. For each left border on the forward strand, we search for the right borders on the reverse strand in the range of 80∼200 bp away from the left border. Except the borders that cannot be paired, for one border on the forward (reverse) strand, there may be more than one suitable border on the reverse (forward) strand, just as [28] showed that the borders of nucleosomes may exhibit bimodal shapes.

Fig. 4 illustrates our border matching strategy. In implementation, we consider three situations. In Fig. 4, the left boxes with green label represent identified borders on the forward strand and the right boxes with red label represent the right borders identified on the reverse strand. Each rectangle is labeled by "F/R+ID", where "F" means the forward strand, "R" indicates the reverse strand, and *ID* is the identifier of an identified border. We assume that each time we match only one pair of borders in the memory. The *dashed* rectangles indicates the borders to be matched next time, so they may be not yet in the memory.

Fig. 4A illustrates the simple situation, where one border $F1$ on the forward strand matches only one border $R1$ on the reverse strand. The matched border pair ($F1$, $R1$) makes up a nucleosome and is moved out of the memory. Following that, ($F2$, $R2$) is moved to the memory for matching.

Fig. 4B illustrates the situation where there are more than one suitable border on the reverse strand matching a border on the forward strand. Here, $F1$ matches both $R1$ and $R2$. So $R1$ and $R2$ are merged to a new right border, which assembles with $F1$ to form a nucleosome. Then, $F1$, $R1$ and $R2$ are moved out of the memory, and ($F2$, $R3$) is moved to the memory for matching.

Fig. 4C illustrates the situation where multiple borders on the forward strand match one border on the reverse strand. Here, both $F1$ and $F2$ match $R2$. First, $F1$ combines $R1$ to constitute a nucleosome. Then, $F1$ and $R1$ are moved out, while $R2$ is kept in the memory to match $F2$ to form another nucleosome.

2.4.3 Merging Nucleosomes

By matching borders, we get a number of candidate nucleosomes, in which those nucleosomes having too much overlap should be merged. As the nucleosomes identified have different sizes about 80∼200 bp by default, for a pair of adjacent candidate nucleosomes NC_1 and NC_2, we compute their overlap ratios with each other: ξ_{12} and ξ_{21}. Suppose the lengths of the two adjacent nucleosomes and their overlap are L_1, L_2 and L_o, then $\xi_{12} = L_o/L_1$ and $\xi_{21} = L_o/L_2$. If ξ_{12} and ξ_{21} are both larger than the prespecified overlap ratio threshold ξ, we merge the two adjacent nucleosomes NC_1 and NC_2 into one nucleosome.

During the mergence of nucleosomes, we combine the support reads of the left borders and the right borders respectively for the two merged nucleosomes. Then, the mean value of $5'$ end positions of support reads from the left borders is defined as the new left border; and similarly, the mean value of $5'$ end positions of support reads from the right borders is defined as the new right border. This process is repeated along the chromosome till the number of nucleosomes identified does not decrease any more. For either border of each finally determined nucleosome, the number of support reads is defined as the weight, and the standard deviation of $5'$ end positions of support reads is defined as the fuzziness. Combining the left and the right borders, we can determine the size of each identified nucleosome as the DNA length between the left and the right borders, and then we calculate the confidence score ($conf\_score$) for each nucleosome as follows:

$$conf\_score = size \cdot \frac{w_l + w_r}{max(sp_r) - min(sp_l)}, \tag{5}$$

where $size$ is the size of nucleosome, w_l and w_r represent the weights of the left border and the right border respectively, sp_l and sp_r are the set of $5'$ end positions of support reads on the left border and the right border, respectively. The confidence score is equivalent to the average coverage of the support reads of both borders. To evaluate confidence score, those support reads are first extended to the size of each detected nucleosome. A larger *score* indicates that the corresponding nucleosome covers more support reads, so it is more possible that the detected nucleosome is a real one.

3 Experimental Results

To validate the performance of the proposed approach, we first conduct local analysis of the identified nucleosomes to

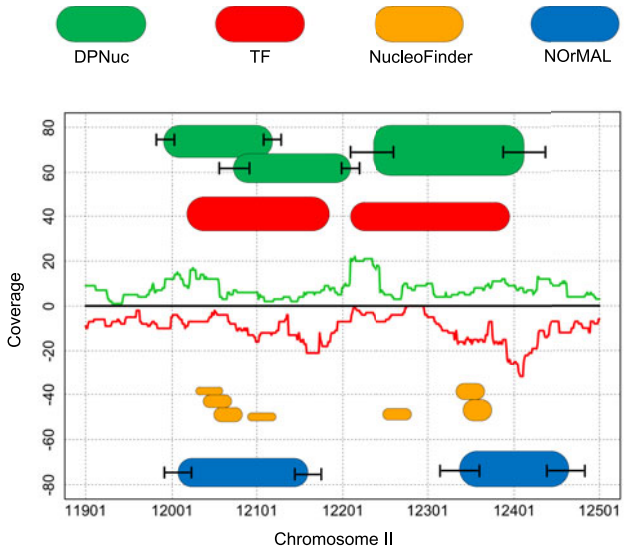


Fig. 5. The results of local analysis on the identified nucleosomes on chromosome II. The rounded rectangles of different colors represent the detected nucleosomes by different methods (the yellow rounded rectangles represent only the 30 bp center regions of nucleosomes detected by Nucleofinder). The green curve and red curve denote the coverage profiles on the forward strand and the reverse strand, respectively. The height and length of each rounded rectangle depict the confidence score and the size of the corresponding identified nucleosome.

show that our approach can provide detailed information of the detected nucleosomes, then check the positions of identified nucleosomes against a synthetic benchmark nucleosome map and two existing benchmark nucleosome maps to prove that our approach performs well in identifying nucleosomes, and finally analyze the size distribution of identified nucleosomes to demonstrate that the results output by our approach are reliable. We compare our approach with three state-of-the art nucleosome identification methods: TF, Nucleofinder and NOrMAL. TF and Nucleofinder are two representative template-based methods, and NOrMAL is a typical data mining based method. TF, NOrMAL and Nucleofinder were all proposed in 2010, 2012 and 2013 respectively.

3.1 Local Analysis

To check the detailed information of the identified nucleosomes, we carry out local analysis on the detected nucleosomes on Chromosome II from Datatset-1. We set the range of nucleosome size as [80, 200] and the overlap threshold ξ as 0.35, respectively, as most existing methods did. Other parameters are taken by default according to the experimental settings reported in these methods' original papers.

Fig. 5 shows the results of local analysis. Here, the coverage profile shows the pileup of short reads (36 bp in length) on each strand: the green curve represents the coverage profile on the forward strand, and the red curve depicts the coverage profile on the reverse strand; the rounded rectangles represent the identified nucleosomes, with different colors to describe the identified nucleosomes by different methods; the height of each rounded rectangle indicates the confidence score of the identified nucleosome; the error bars adhered to the borders of each rounded rectangle mean the fuzziness provided by the corresponding method.

Except for Nucleofinder, all the other three methods can output nucleosome size. Nucleofinder regards that all nucleosomes have similar size of 150 bp and applies a 150 bp sliding window with a step size of 10 bp. It detects only 30 bp enriched center regions of nucleosomes and calculates the Bayesian factor as confidence score. So we use short yellow rounded rectangles of length 30 bp to indicate the center regions of nucleosomes detected by Nucleofinder. Compared with the other three methods, it provides the least information about the identified nucleosomes. Although it can detect independent nucleosomes relatively accurately, it cannot deal with overlapping nucleosomes well. So the results of Nucleofinder cannot reveal the intricate configurations of nucleosomes in a reasonable way.

Although NOrMAL can output detailed information of identified nucleosomes, including size and fuzziness, the precision of nucleosome positions is lower than the other methods (we delay the detailed discussion on prediction accuracy of NOrMAL to next section). Here, we can see that the first nucleosome identified by NOrMAL is roughly similar to the first ones detected by the other methods, but its second detected nucleosome is unreasonable. NOrMAL requires a prior size of nucleosomes as input, and the change of input prior nucleosome size will lead to instable results. So the size distribution of identified nucleosomes relies on the input prior size to a great extent. In other words, the resulting nucleosome size by NOrMAL cannot reflect the real nucleosome size accurately. We will show detailed results in the following section. Compared with our method, it can provide only the fuzziness of nucleosomes as a whole (the error bars on both sides are the same), while our method can provide separate fuzziness for the two borders of each identified nucleosome (the error bars on both sides are independent).

TF cannot output fuzziness values of identified nucleosomes, which leads to the loss of detailed nucleosome configuration information. In the first half of the coverage profile, TF recognizes one nucleosome while our method detects two overlapping nucleosomes with small fuzziness. In the second half of the coverage profile, both methods report one nucleosome, while our method generates wide error bars and a larger confidence score, which indicates that the identified nucleosome is probably formed by merging two overlapping nucleosomes. However, TF cannot reveal such information and cannot distinguish the second detected nucleosome much in configuration from the first one, except that the latter has a smaller confidence score.

By local analysis and comparison with three existing methods, we can see that our method can provide more detailed and accurate information of the reported nucleosomes, and the nucleosomes identified by our method reveal the real and intricate configurations better on the chromosomes.

3.2 Comparing Against Nucleosome Maps

To further validate the performance of our method, we compare the detected result of our method against a synthetic benchmark nucleosome map and existing benchmark nucleosome maps.

For the paired-end reads of Dataset-3 we select all the 2,000,000 nucleosomal DNA fragments, which fall into the area 3,001,422 $\sim$ 22,086,379 on Chr14 of mouse ESCs. We regard the centers of nucleosomal DNA fragments as

TABLE 3
Results When Aligned to the Synthetic Benchmark Nucleosome Map

Approach	Precision	Recall	F-score
TF	0.7189	0.6851	0.7016
Nucleofinder	0.7289	0.4011	0.5175
NOrMAL	0.6856	0.3440	0.4582
DPNuc	**0.9096**	**0.7791**	**0.8393**

the initial nucleosome positions. Then, we merge the adjacent overlapping nucleosomes whose overlapping amount is above the given maximum overlap into a new nucleosome. The position of the new nucleosome is determined by the weighted average position of the overlapping nucleosomes, with the number of nucleosomal DNA fragments of each nucleosome as weight. The mergence process is repeated till no more adjacent nucleosomes can be merged. Finally, the resulting nucleosome positions are set as the synthetic benchmark nucleosome map. Here, we convert the original paired-end reads into single-end reads, and process them by all the four methods.

Because of the difference in experiment conditions and biological materials, existing maps of nucleosomes in Saccharomyces cerevisiae are not always consistent. Nevertheless, most nucleosomes in Yeast are arranged in an well-positioned way [40]. To comprehensively compare the performance of the four methods, we also use two nucleosome maps [34], [35] of Saccharomyces cerevisiae as benchmark. One genome-wide map (called Map 2008 in this paper) was presented according to the widely accepted barrier model for nucleosome statistical positioning, and the other map (denoted as Map 2012 in this paper) at the base-pair resolution is regarded as the most precise and detailed at present. Here, we use Dataset-1 because it generates the most deeply sequenced data of Yeast up to now and can help us analyze the nucleosomes more accurately and detailedly.

We align the nucleosomes identified by the four methods to the synthetic nucleosome map and the two existing maps respectively, and evaluate the F-score (F in short). We use the F measure because it is a combination of *precision* and *recall*, which can more comprehensively evaluate the prediction performance.

The number of nucleosomes successfully aligned to the maps (The distance between their centers is no larger than 30 bp) is regarded as *true positive* (tp), and the number of nucleosomes that cannot be aligned to the map is defined as *false positive* (fp). The number of the remaining nucleosomes on the map is defined as *false negative* (fn). Then, F is evaluated as follows:

$$precision = \frac{tp}{tp + fp}, \tag{6}$$

$$recall = \frac{tp}{tp + fn}, \tag{7}$$

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{8}$$

Table 3 shows the precisions, recalls and F scores by the four approaches on the selected area of Chromosome 14 from mouse ESCs. We can see that our method achieves the highest values (bold numbers) of precision, recall and F score.

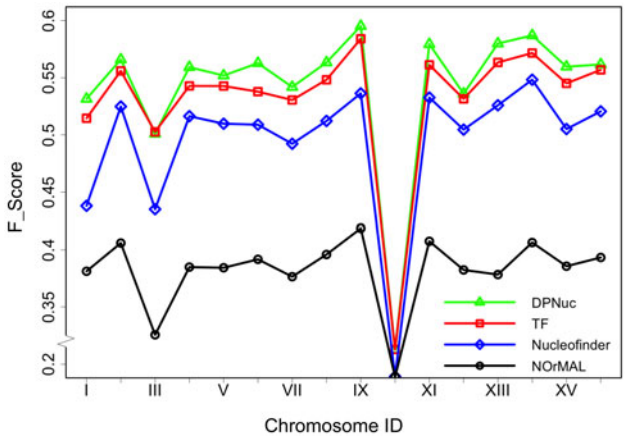


Fig. 6. The F scores of the four methods on the 16 chromosomes of Saccharomyces cerevisiae (aligning to Map 2012).

Figs. 6 and 7 show the F scores of the four methods on the 16 chromosomes of Saccharomyces cerevisiae, by aligning the identified nucleosomes to Map 2012 and Map 2008, respectively. We can see that our method achieves a relatively higher F score than the other three methods. For Map 2012, our method obtains the highest F score on 14 of 15 chromosomes (except Chromosome 3). Note that the F score on Chromosome 10 shows an abnormally sharp down for all methods on both maps, we think this may be due to the original dataset, so we do not take Chromosome 10 into consideration in the following analysis. As for Map 2008, our method also has the highest F score on 14 of 15 chromosomes (except Chromosome 14).

Compared with the other methods, NOrMAL performs poorly and has a significantly lower F score. This may be caused by its aggressive merging strategy. By contrast, our method achieve highest F score on the synthetic map and most chromosomes of both existing maps. So our method can position the nucleosomes more satisfactorily.

Table 4 presents the time cost of the four methods. As MCMC simulation is computationally intensive, the processing time of our method is more than the other three methods. However, the processing time is still acceptable.

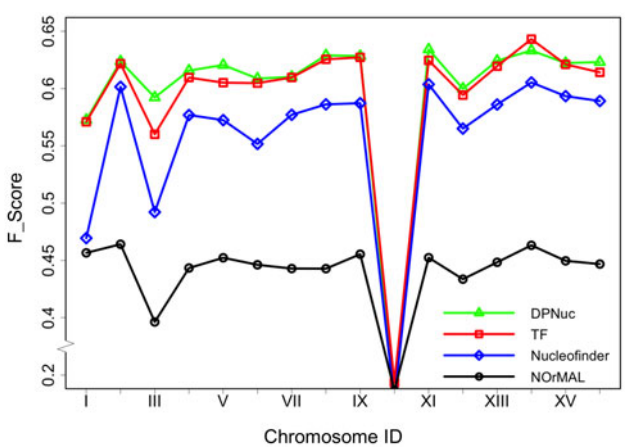


Fig. 7. The F scores of the four methods on the 16 chromosomes of Saccharomyces cerevisiae (aligning to Map 2008).

TABLE 4
Time-Cost Comparison of the Four Methods

Approach	Dataset-1 (s) (Chr1-Chr16)	Dataset-3 (s) (the selected area on Chr14)
TF	544	7,759
Nucleofinder	273	438
NOrMAL	117	4,951
DPNuc	3,918	7,842

For processing Dataset-3, our method consumes 7,842 seconds, while TF, NOrMAL and Nucleofinder need 7,759, 4,951 and 438 seconds, respectively.

3.3 Size Distribution of Identified Nucleosomes

To further validate the performance of our approach, we examine the size distribution of identified nucleosomes from data generated under different experimental conditions.

In ChIP-seq experiments, DNA attached to nucleosomes is protected from digestion of micrococcal nuclease, MNase mainly digests the linker DNA and isolates the nucleosomes. Through reverse crosslink, DNA wrapping around histones is released. The lengths of DNA fragments are related to nuclease titration level [41], [42]. With a higher level of nuclease titration, the proportion of mononucleosomal DNA gets larger, and DNA fragments become shorter [28]. Here, we adopt two titration levels, i.e., the typical 10 μL MNase and the overdigested 15 μL MNase, and examine the size distribution of nucleosomes predicted by our method, TF and NOrMAL. Because Nucleofinder cannot provide nucleosome size, it is not considered here. Considering that the input prior size of nucleosomes is crucial to the final size distribution predicted by NOrMAL, we take two different prior sizes: 110 bp and the default 140 bp, respectively. Here, we identify the nucleosomes on Chromosome II of Saccharomyces cerevisiae of Dataset-2, the results are shown in Fig. 8.

In Fig. 8, the vertical axis *Frequency* means the number of detected nucleosomes of a certain size; the subfigures on the left and the right illustrate the distributions of nucleosome size in the range 80$\sim$100 bp for 10 and 15 μL MNase, respectively. Our method and TF detect a significant change of size distribution at different nuclease titration levels. As the nuclease titration level increases, the size of identified nucleosomes turns smaller clearly. This conforms to experimental observation. In Figs. 8A and 8B, we can see that our method detects roughly the same distribution change as TF does. That is, the distribution of nucleosome size concentrates from around 125 bp to around 90 bp when the MNase titration level grows from 10 to 15 μL. However, NOrMAL does not detect such change. From Figs. 8C and 8D, we can see that for NOrMAL, the concentration center (where the distribution density is the largest) of size distribution for the two different nuclease titration levels are quite close (the shift is about 10 bp), whether the input prior size is set to 140 or 110 bp. Surprisingly, at the same titration level, when the input prior size decreases, the concentration center of size distribution follows the same down trend. For example,

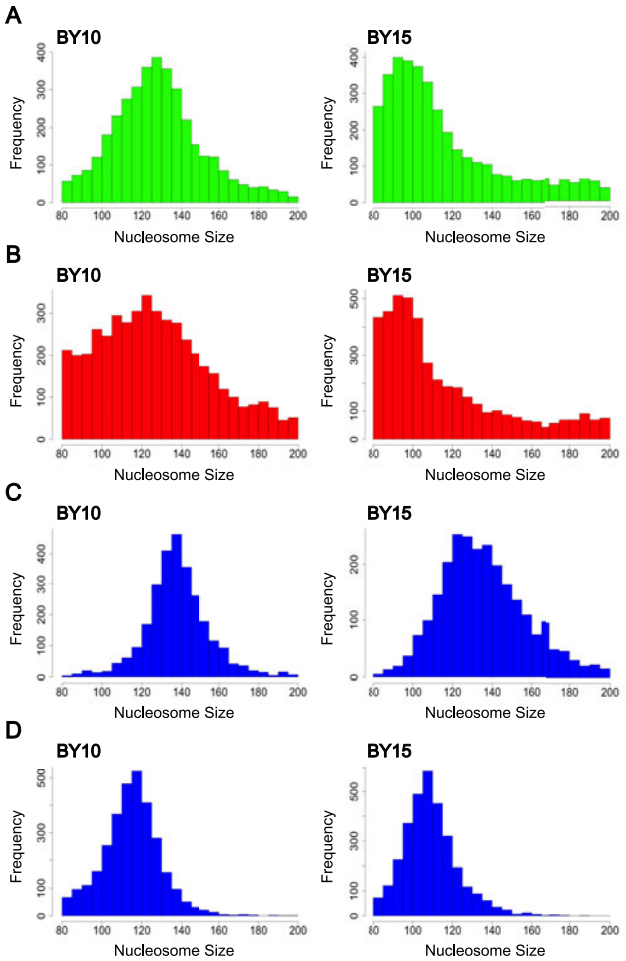


Fig. 8. The size distribution of nucleosomes identified by different methods over Chromosome II of Dataset-2 at different MNase titration levels. The vertical axis *Frequency* means the number of detected nucleosomes of a certain size; the left and the right figures illustrate the distributions when the MNase titration level is 10 μL and 15μL, respectively. (A) DPNuc; (B) TF; (C) NOrMAL with the input prior size being 140 bp; (D) NOrMAL with the input prior size being 110 bp.

at 10 μL MNase titration level, when the input prior size decreases from 140 to 110 bp, the concentration center of the size distribution changes from around 140 to around 110 bp correspondingly. Thus, the input prior size of nucleosomes has a significant influence on the final results of NOrMAL. However, it is difficult to get an accurate prior size of nucleosomes, so NOrMAL cannot output accurate nucleosome size as our method and TF can do.

3.4 Reliability of Nucleosome Size Distribution

Here, we demonstrate the reliability of the size distribution of nucleosomes identified by our method. For this end, on the one hand, we compare the size distribution of paired-end sequenced nucleosomal DNA with the with size distributions of nucleosomes identified by different methods over Chromosome 14 of Dataset-3; on the other hand, we check the sensitivity of our method to random read permutation along chromosomes.

We identify nucleosomes from the single-end reads transformed from paired-end reads of Chromosome 14 of Dataset-3, by the four methods. An ideal result is that the size distribution of detected nucleosomes keeps consistent

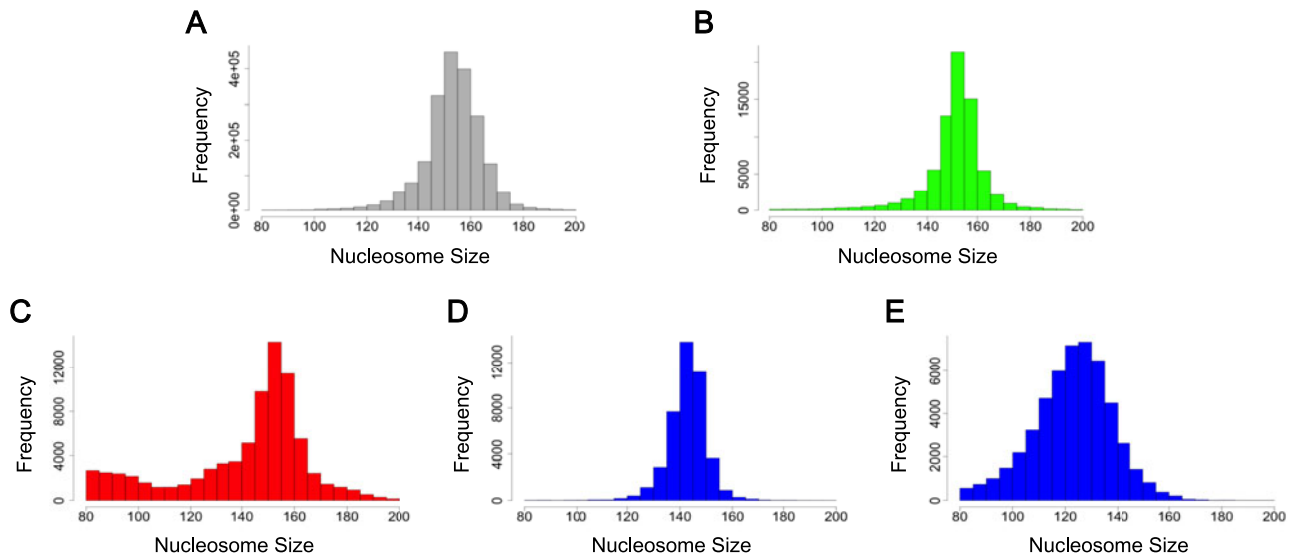


Fig. 9. Comparison of size distribution of paired-end sequenced nucleosomal DNA with size distributions of nucleosomes identified by different methods over Chromosome 14 of Dataset-3. The vertical axis *Frequency* means the number of detected nucleosomes of a certain size. (A) The paired-end sequenced nucleosomal DNA; (B) DPNuc; (C) TF; (D) NOrMAL with the input prior size being 140 bp; and (E) NOrMAL with the input prior size being 110 bp.

with the size distribution of paired-end sequenced nucleosomal DNA. A method that gets the ideal size distribution is considered reliable.

Fig. 9 shows the size distribution of paired-end sequenced nucleosomal DNA and the size distributions of nucleosomes identified by our method, TF, NOrMAL respectively. The size distribution of paired-end sequenced nucleosomal DNA is considered the ground truth, which is shown in Fig. 9A. We can see that its concentration center is located around 150 bp. The result of our method is shown in Fig. 9B, which is quite similar to Fig. 9A. Although the size distribution obtained by TF (see Fig. 9C) concentrates at around 150 bp, it also aggregates on the left boundary, that is near 80 bp, which is not reasonable. As for the size distributions achieved by NOrMAL with the input prior size 140 and 110 bp (see Figs. 9D and 9E), they concentrate at around 140 and 125 bp respectively, and their distribution shapes are significantly different from that of the ground truth.

Now we examine the the sensitivity of our method to random read permutation. Before permutation, the size distribution of nucleosomes should appear like Gaussian distribution and shows a significant concentration at a certain

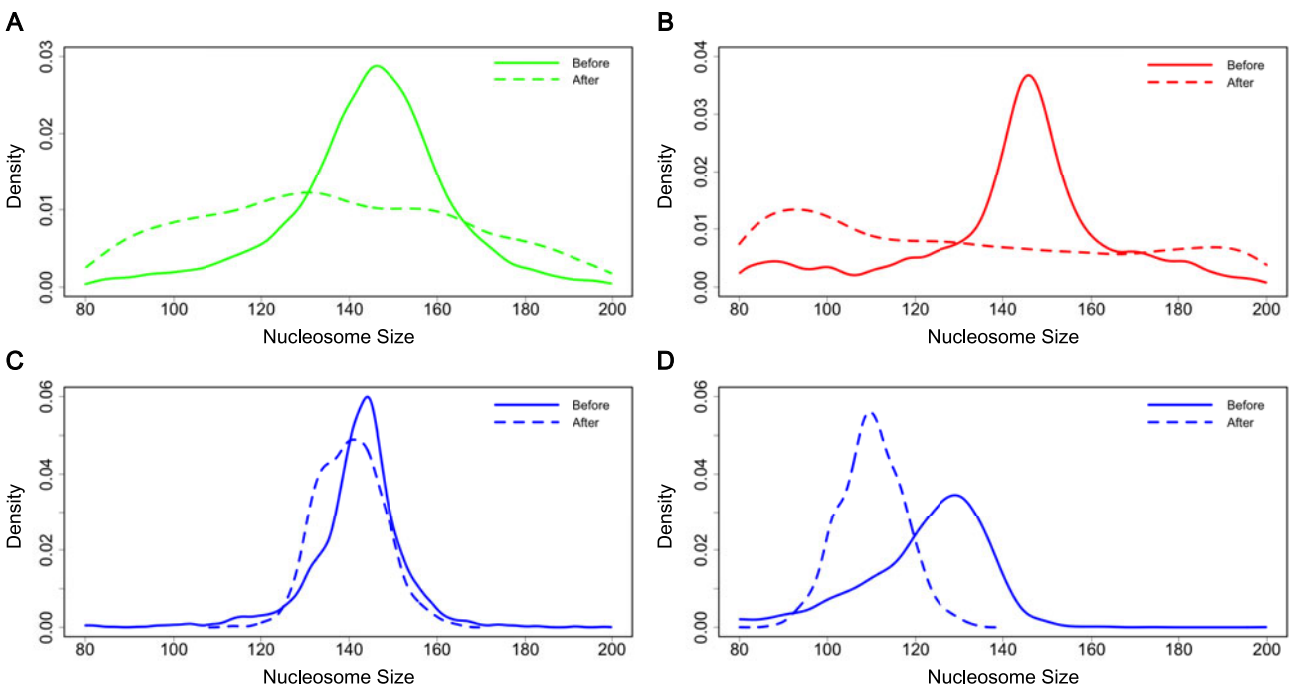


Fig. 10. The size distribution of nucleosomes identified by different methods on Chromosome II of Dataset-1. The solid curves and the dashed curves illustrate respectively the nucleosome size distributions before and after permutation. (A) DPNuc; (B) TF; (C) NOrMAL with the input prior size being 140 bp; and (D) NOrMAL with the input prior size being 110 bp.

size. This has special biological meanings. Once we randomly permute the positions of reads on the chromosomes, such permuted reads cannot be acquired in real biological experiments, so nucleosomes detected on such data should make no sense. We expect that the size distribution of nucleosome detected on permuted data is more uniform than the size distribution of nucleosome detected on the original data. If our method meets this expectation, we think that its result is reliable.

In Fig. 10, we present the size distributions of the identified nucleosomes by our method, TF and NOrMAL, before and after permutation on Chromosome II of Dataset-1. The solid curves depict the size distributions before permutation and the dashed curves depict the size distributions after permutation. From Figs. 10A and 10B, we can see that both our method and TF obtain a much more uniform size distribution after permutation than before permutation. After permutation, the size distribution by our method shows a slightly higher density (the largest density is 0.0122) in the central region and a relatively lower density in the boundary regions. This result is reasonable. A little differently, the size distribution by TF concentrates in the boundary regions (the largest density is 0.0134), which may be caused by its nucleosome assembly strategy. TF determines the final nucleosomes from preliminary nucleosomes using a greedy approach; and for overlapping nucleosomes, it keeps only the nucleosome with the largest correlation score in a certain range, while ignoring the one overlapping with it.

On the contrary, as shown in Figs. 10C and 10D, NOrMAL does not show much distribution change after permutation, the size distribution it obtains still has an obvious concentration around the input prior size. We can see that different input prior size results in quite different size distributions after permutation. So we can infer that nucleosome size obtained by NOrMAL is not reliable.

In summary, in comparison with the size distribution of paired-end sequenced nucleosomal DNA, our method can detect the nucleosome size distribution consistent with the ground truth; in the testing of sensitivity to random read permutation, our method and TF both obtain more uniform size distributions after read permutation, and the distributions do not significantly concentrate at any specific size, but the size distributions obtained by TF have a little higher density near the boundary after permutation. Quite differently, NOrMAL is much less sensitive to read permutation than our approach and TF. So the nucleosomes identified by our method are more reliable than those identified by TF and NOrMAL.

4 Conclusion

In this paper, we present a new approach called DPNuc for nucleosome positioning, which is based on the Dirichlet process mixture model. Compared with three state of the art methods, our method can provide more detailed and accurate information of the detected nucleosomes. The local analysis of identified nucleosomes shows that our method can detect nucleosomes with complex configurations and can reveal more intricate information of nucleosomes on chromosomes. By comparing against a synthetic nucleosome map and two existing nucleosome maps, our method performs better in identifying nucleosomes and obtains a higher F-score. Furthermore, by detecting nucleosomes from data generated at two different nuclease titration levels, the results show that our method can successfully detect the change of experimental conditions, which means that the sizes of nucleosomes obtained by our approach are accurate. Finally, we compare the size distributions of nucleosomes detected by different methods with the size distribution of paired-end sequenced nucleosomal DNA, and check the size distributions of nucleosomes identified before and after random read permutation along chromosomes, the results show that our method can obtain nucleosome size distribution quite consistent with that of paired-end sequenced nucleosomal DNA, and quite different nucleosome size distributions before and after read permutation, which indicates that the size of nucleosomes obtained by our approach is more reliable.

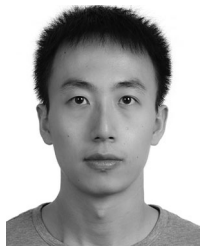
Acknowledgments

This work was partially supported by National Natural Science Foundation of China (NSFC) under grants No. 61173118 and No. 61272380. Jihong Guan is the corresponding author.

References

[1] T. J. Richmond and C. A. Davey, "The structure of DNA in the nucleosome core," *Nature*, vol. 423, no. 6936, pp. 145–150, 2003.
[2] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 å resolution," *Nature*, vol. 389, no. 6648, pp. 251–260, 1997.
[3] J. Mellor, "Dynamic nucleosomes and gene transcription," *Trends Genetics*, vol. 22, no. 6, pp. 320–329, 2006.
[4] J. Workman and R. Kingston, "Alteration of nucleosome structure as a mechanism of transcriptional regulation," *Annu. Rev. Biochemistry*, vol. 67, no. 1, pp. 545–579, 1998.
[5] S. Shivaswamy, A. Bhinge, Y. Zhao, S. Jones, M. Hirst, and V. R. Iyer, "Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation," *PLoS Biol.*, vol. 6, no. 3, pp. 618–630, 2008.
[6] X. Chen, Z. Chen, H. Chen, Z. Su, J. Yang, F. Lin, S. Shi, and X. He, "Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes," *Science*, vol. 335, no. 6073, pp. 1235–1238, 2012.
[7] H. Tilgner, C. Nikolaou, S. Althammer, M. Sammeth, M. Beato, J. Valcárcel, and R. Guigó, "Nucleosome positioning as a determinant of exon recognition," *Nat. Struct. Molecular Biol.*, vol. 16, no. 9, pp. 996–1001, 2009.
[8] Z. Zhang, C. J. Wippo, M. Wal, E. Ward, P. Korber, and B. F. Pugh, "A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome," *Science*, vol. 332, no. 6032, pp. 977–980, 2011.
[9] C. Jiang and B. F. Pugh, "Nucleosome positioning and gene regulation: advances through genomics," *Nat. Rev. Genetics*, vol. 10, no. 3, pp. 161–172, 2009.
[10] Z. Zhang and B. F. Pugh, "High-resolution genome-wide mapping of the primary structure of chromatin," *Cell*, vol. 144, no. 2, pp. 175–186, 2011.
[11] T. Raveh-Sadka, M. Levo, U. Shabi, B. Shany, L. Keren, M. Lotan-Pompan, D. Zeevi, E. Sharon, A. Weinberger, and E. Segal, "Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast," *Nat. Genetics*, vol. 44, no. 7, pp. 743–750, 2012.
[12] D. J. Clark, "Nucleosome positioning, nucleosome spacing and the nucleosome code," *J. Biomolecular Struct. Dynamics*, vol. 27, no. 6, pp. 781–793, 2010.
[13] N. Kaplan, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, "Contribution of histone sequence preferences to nucleosome organization: Proposed definitions and methodology," *Genome Biol.*, vol. 11, no. 11, p. 140, 2010.

[14] W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow, "A high-resolution atlas of nucleosome occupancy in yeast," *Nat. Genetics*, vol. 39, no. 10, pp. 1235–1244, 2007.
[15] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, et al., "A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning," *Genome Res.*, vol. 18, no. 7, pp. 1051–1063, 2008.
[16] T. N. Mavrich, C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, et al., "Nucleosome organization in the drosophila genome," *Nature*, vol. 453, no. 7193, pp. 358–362, 2008.
[17] D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao, "Dynamic regulation of nucleosome positioning in the human genome," *Cell*, vol. 132, no. 5, pp. 887–898, 2008.
[18] P. J. Park, "Chip–seq: Advantages and challenges of a maturing technology," *Nat. Rev. Genetics*, vol. 10, no. 10, pp. 669–680, 2009.
[19] B. E. Bernstein, A. Meissner, and E. S. Lander, "The mammalian epigenome," *Cell*, vol. 128, no. 4, pp. 669–681, 2007.
[20] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, et al., "Model-based analysis of chip-seq (macs)," *Genome Biol*, vol. 9, no. 9, p. R137, 2008.
[21] A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow, "Genome-wide analysis of transcription factor binding sites based on chip-seq data," *Nat. Methods*, vol. 5, no. 9, pp. 829–834, 2008.
[22] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein, "Peakseq enables systematic scoring of chip-seq experiments relative to controls," *Nat. Biotechnol.*, vol. 27, no. 1, pp. 66–75, 2009.
[23] K. Struhl and E. Segal, "Determinants of nucleosome positioning," *Nat. Struct. Molecular Biol.*, vol. 20, no. 3, pp. 267–273, 2013.
[24] K. Chen, Y. Xi, X. Pan, Z. Li, K. Kaestner, J. Tyler, S. Dent, X. He, and W. Li, "Danpos: Dynamic analysis of nucleosome position and occupancy by sequencing," *Genome Res.*, vol. 23, no. 2, pp. 341–351, 2013.
[25] R. Schöpflin, V. B. Teif, O. Müller, C. Weinberg, K. Rippe, and G. Wedemann, "Modeling nucleosome position distributions from experimental nucleosome positioning maps," *Bioinformatics*, vol. 29, no. 19, pp. 2380–2386, 2013.
[26] A. Mammana, M. Vingron, and H.-R. Chung, "Inferring nucleosome positions with their histone mark annotation from chip data," *Bioinformatics*, vol. 29, no. 20, pp. 2547–2554, 2013.
[27] S. Pepke, B. Wold, and A. Mortazavi, "Computation for chip-seq and rna-seq studies," *Nat. Methods*, vol. 6, pp. S22–S32, 2009.
[28] A. Weiner, A. Hughes, M. Yassour, O. J. Rando, and N. Friedman, "High-resolution nucleosome mapping reveals transcription-dependent promoter packaging," *Genome Res.*, vol. 20, no. 1, pp. 90–100, 2010.
[29] J. Becker, C. Yau, J. M. Hancock, and C. C. Holmes, "Nucleofinder: A statistical approach for the detection of nucleosome positions," *Bioinformatics*, vol. 29, no. 6, pp. 711–716, 2013.
[30] G. L. Turint, "An introduction to matched filters," *IRE Trans. Inf. Theory*, vol. 6, no. 3, pp. 311–329, 1960.
[31] A. Polishko, N. Ponts, K. G. Le Roch, and S. Lonardi, "Normal: Accurate nucleosome positioning using a modified gaussian mixture model," *Bioinformatics*, vol. 28, no. 12, pp. i242–i249, 2012.
[32] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, et al., "The DNA-encoded nucleosome organization of a eukaryotic genome," *Nature*, vol. 458, no. 7236, pp. 362–366, 2009.
[33] V. B. Teif, Y. Vainshtein, M. Caudron-Herger, J.-P. Mallm, C. Marth, T. Höfer, and K. Rippe, "Genome-wide nucleosome positioning during embryonic stem cell development," *Nat. Struct. Molecular Biol.*, vol. 19, no. 11, pp. 1185–1192, 2012.
[34] T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, S. C. Schuster, I. Albert, and B. F. Pugh, "A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome," *Genome Res.*, vol. 18, no. 7, pp. 1073–1083, 2008.
[35] K. Brogaard, L. Xi, J.-P. Wang, and J. Widom, "A map of nucleosome positions in yeast at base-pair resolution," *Nature*, vol. 486, no. 7404, pp. 496–501, 2012.
[36] K. Fu, Q. Tang, J. Feng, X. S. Liu, and Y. Zhang, "Dinup: A systematic approach to identify regions of differential nucleosome positioning," *Bioinformatics*, vol. 28, no. 15, pp. 1965–1971, 2012.
[37] A. F. da Silva and M. A. F. da Silva, dpmixsim: Dirichlet process mixture model simulation for clustering and image segmentation [Online]. Available: http://cran.r-project.org/package=dpmixsim
[38] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *J. Comput. Graphical Statist.*, vol. 9, no. 2, pp. 249–265, 2000.
[39] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, 1962.
[40] M. Radman-Livaja and O. J. Rando, "Nucleosome positioning: How is it established, and why does it matter?" *Develop. Biol.*, vol. 339, no. 2, pp. 258–266, 2010.
[41] J. Allan, R. M. Fraser, T. Owen-Hughes, and D. Keszenman-Pereyra, "Micrococcal nuclease does not substantially bias nucleosome mapping," *J. Molecular Biol.*, vol. 417, no. 3, pp. 152–164, 2012.
[42] H.-R. Chung, I. Dunkel, F. Heise, C. Linke, S. Krobitsch, A. E. Ehrenhofer-Murray, S. R. Sperling, and M. Vingron, "The effect of micrococcal nuclease digestion on nucleosome positioning data," *PloS One*, vol. 5, no. 12, p. e15754, 2010.



Huidong Chen is currently working toward the PhD degree in the Department of Computer Science and Technology, Tongji University, Shanghai, China. His research interest is algorithm design for identifying nucleosome positions.

Jihong Guan received the bachelor's degree from Huazhong Normal University in 1991, the master's degree from the Wuhan Technical University of Surveying and Mapping (merged into Wuhan University since Aug. 2000) in 1991, and the PhD degree from Wuhan University in 2002. She is currently a professor in the Department of Computer Science and Technology, Tongji University, Shanghai, China. Before joining Tongji University, she served in the Department of Computer, Wuhan Technical University of Surveying and Mapping, from 1991 to 1997, as an assistant professor and an associate professor (since August 2000), respectively. She was an associate professor (Aug. 2000-Oct. 2003) and a professor (Since Nov. 2003) in the School of Computer, Wuhan University. Her research interests include databases, data mining, distributed computing, bioinformatics, and geographic information systems (GIS). She has published more than 100 papers in domestic and international journals and conferences.

Shuigeng Zhou received the bachelor's degree from the Huazhong University of Science and Technology (HUST) in 1988, the master's degree from the University of Electronic Science and Technology of China (UESTC) in 1991, and the PhD degree of computer science from Fudan University in 2000. He is currently a professor at the School of Computer Science, Fudan University, Shanghai, China. He served in Shanghai Academy of Spaceflight Technology from 1991 to 1997, as an engineer and a senior engineer (since August 1995), respectively. He was a postdoctoral researcher in the State Key Lab of Software Engineering, Wuhan University, from 2000 to 2002. His research interests include data management, data mining, machine learning, and bioinformatics. He has extensively published in domestic and international journals (including *IEEE TKDE*, *IEEE TPDS*, *IEEE TCBB*, *IEEE TGRS*, *DKE*, *NAR* and *Bioinformatics* etc.) and conferences (including SIGMOD, SIGKDD, SIGIR, VLDB, ICDE, AAAI, IJCAI, SODA and RECOMB etc.). Currently, he is a member of the IEEE, ACM, and the IEICE.