BioSystems 185 (2019) 104033

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

DNMHMM: An approach to identify the differential nucleosome regions in multiple cell types based on a Hidden Markov Model

Jiahao Xie[a,1], Yiran Cai[a,1], Huamei Li[a], Jiahui Wu[a], Xinlei Zhao[a], Kun Luo[b], Amit Sharma[c], Jianming Xie[a], Xiao Sun[a], Hongde Liu[a,*]

[a] State Key Laboratory of Bioelectronics, School of Biological Science & Medical Engineering, Southeast University, Nanjing 210096, China
[b] Department of Neurosurgery, Xinjiang Evidence-Based Medicine Research Institute, First Affiliated Hospital of Xinjiang Medical University, Urumqi 830054, China
[c] Department of Ophthalmology, University Hospital Bonn, Germany

ARTICLE INFO

Keywords:
Differential nucleosome regions (DNRs)
Hidden Markov Model (HMM)
10–11 bp periodicities
Gene expression

ABSTRACT

Nucleosome occupancy changes across cell types and environmental conditions and such changes often have profound influence in transcription. It's of importance to identify the differential nucleosome regions (DNRs) where the nucleosome occupancy level differs across cell types. Here we developed DNMHMM, a Hidden Markov Model (HMM) based algorithm, to detect the DNRs with nucleosomal DNA sequenced dataset. The performance evaluation indicates that DNMHMM is advisable for multi-cell type comparison. Upon testing this model in yeast mutants, where the modifiable histone residues were mutated into alanine, we found that DNA sequences of the dynamic nucleosomes lack 10–11 bp periodicities and harbor binding motifs of the nucleosome remodelling complex. Moreover, the highly expressed genes have more dynamic nucleosomes at promoters. We further compared nucleosome occupancy between resting and activated human $CD4^+$ T cells with this model. It was revealed that during the activation of $CD4^+$ T cells, dynamic nucleosomes are enriched at regulatory sites, hence, up to some extent can affect the gene expression level. Taken together, DNMHMM offers the possibility to access precise nucleosome dynamics among multiple cell types and also can describe the closer association between nucleosome and transcription.

1. Introduction

Nucleosome, a basic unit of eukaryotic chromatin, is composed with 147-bp DNA wrapping on the surface of histone octamer (Luger et al., 1997). By controlling the accessibility to transcription factor binding sites on DNA, nucleosome plays an important role in transcription regulation (Jiang and Pugh, 2009). Like other type of the epigenetic marks, nucleosomes are highly dynamic in different cell types and can undergo sliding, unwrapping and even dissociation from the DNA. Therefore, the nucleosome position decides the fate of many cellular processes. It has been shown that in aging yeast cells, nucleosome occupancy decreased by 50% across the whole genome, and genes that are normally repressed by promoter nucleosomes simply up-regulated the transcriptional expression (Hu et al., 2014). In embryonic stem cells, where the substantial chromatin related changes occur very frequently and extensively throughout the course of differentiation, there again, nucleosome positions act as a key factor to maintain the genome integrity. During differentiation of embryonic stem cells, nucleosome exhibits a lower occupancy level in pluripotent cells than in somatic cells at enhancers (West et al., 2014). Furthermore, nucleosome complexes, e.g. PBAF/BAF complexes tightly co-associate with the nucleosome related changes. This can be evident from the fact that > 20% of human cancers with loss-of-function mutations in PBAF or BAF complex results into cell-type-specific abnormal nucleosome structure and further misregulation in the transcription (Henikoff, 2016; Wilson and Roberts, 2011). Moreover, nucleosome occupancy is found to link with mutation rate (Makova and Hardison, 2015).

Since next generation sequencing (NGS) technologies, such as MNase-Seq (Schones et al., 2008), genome-wide nucleosome profiles at a single-nucleotide resolution can be determined. From a computing perspective, it is necessary to identify genomic regions where nucleosomes change across the cell types. As for nucleosome comparison

* Corresponding author.
E-mail addresses: 610225668@qq.com (J. Xie), charlottecaiir@gmail.com (Y. Cai), li_hua_mei@163.com (H. Li), jiahuiwuu@163.com (J. Wu), 2674587032@qq.com (X. Zhao), luokun_2822@sohu.com (K. Luo), Amit.Sharma@ukbonn.de (A. Sharma), xiejm@seu.edu.cn (J. Xie), xsun@seu.edu.cn (X. Sun), liuhongde@seu.edu.cn (H. Liu).
[1] The authors contributed equally.

https://doi.org/10.1016/j.biosystems.2019.104033
Received 20 March 2019; Received in revised form 7 August 2019
Available online 18 September 2019
0303-2647/

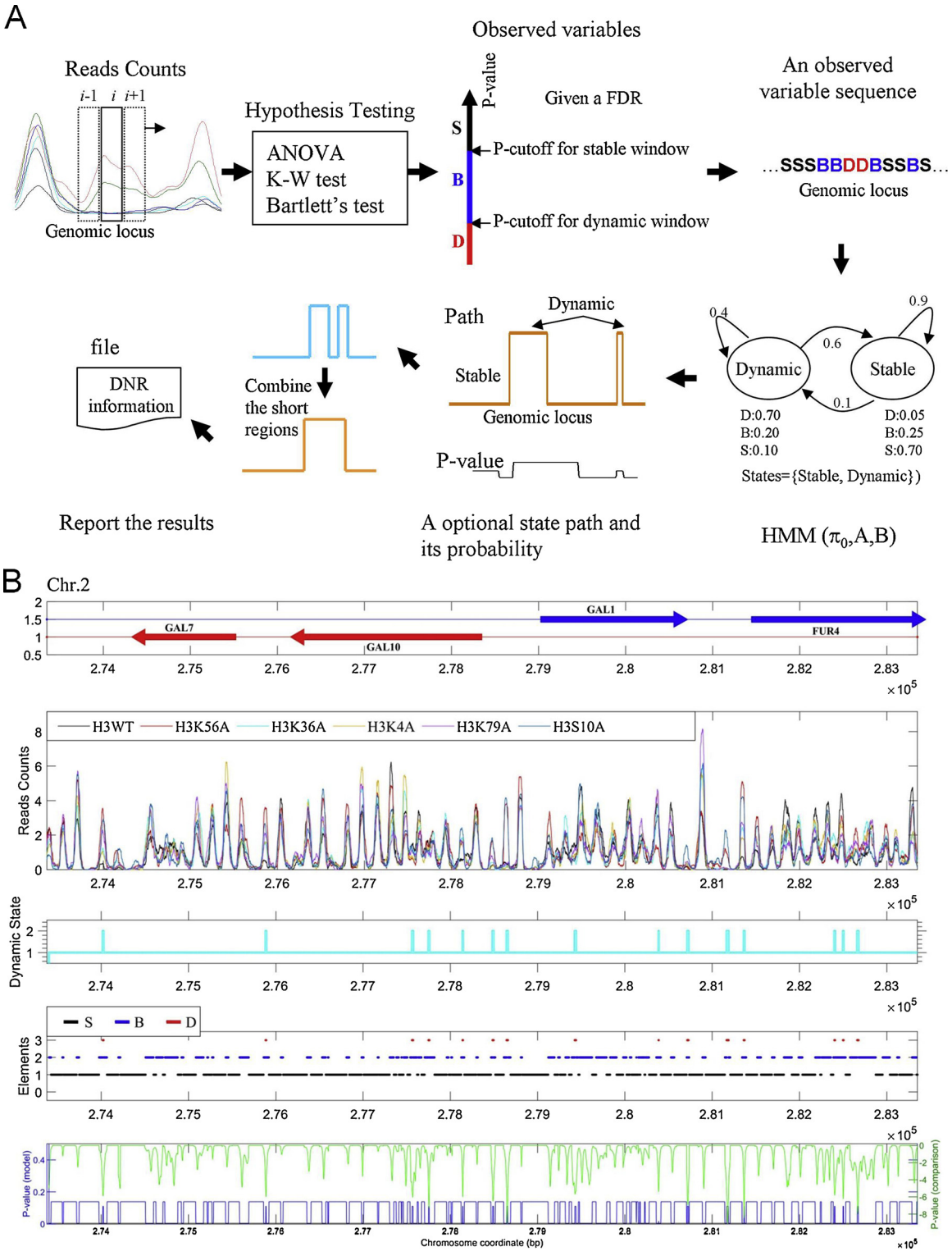


Fig. 1. Illustration of method DNMHMM to identify differential nucleosome regions in multiple samples.

A, Shown is the calculation procedure of DNMHMM. Prior to being inputted into the model, reads counts were calculated by averaging the reads depth of each genomic locus per sample. The first step is to calculate an observed variable sequence. In a sliding window, a hypothesis test is used to test the difference of reads counts in the window among the cell types with a P-value. According to the P-value, an observed variable is determined for the window. If P-value < P-cutoff-D, the observed variable will be "D"; if P-value > P-cutoff-S, the variable will be "S"; if P-value is between the two cutoffs, the variable will be "B". The two cutoffs are estimated with a given false discovery ratio (FDR). The second step is to infer an optional hidden state sequence using the observed sequence with Viterbi algorithm. The hidden state sequence indicates the genomic regions with differential nucleosomes. The parameters of the HMM are trained from a dataset in which we manually marked the differential nucleosome regions. In the end, the differential nucleosome regions which are close to each other (distance < 15 bp) are combined. The DNMHMM reports the position, FDR and p-value of each differential nucleosome region.

B, Identification of the differential nucleosome regions in six cell types of *Saccharomyces cerevisiae* with DNMHMM. Shown is for one segment of chr. 2. The second panel shows the normalized reads counts (nucleosome occupancy profile). The third panel indicates the DNRs identified by DNMHMM (FDR = 1×10^{-10}, window width = 25 bp, hypothesis test = ANOVA). The fourth panel is the observed variables at each genomic locus. The probability of the observed variables sequence on the HMM (left axis) and the probability of the hypothesis test (right axis) are also plotted in the bottom panel.

between two cell types, methods that based on basic statistical model have been proposed (Polishko et al., 2012; Becker et al., 2013). Shivaswamy S et al. introduced a nucleosome positioning periodicity (NPP) score to evaluate the stability of nucleosome positioning, which is sequencing-depth independent (Shivaswamy et al., 2008). Kai Fu et al. developed an approach DiNuP that uses two-sample Kolmogorov–Smirnov (K-S) test to identify regions of differential nucleosome positioning (Fu et al., 2012). In order to give a robust cutoff, DiNuP estimates false discovery ratio (FDR) by re-sampling reads, thus reducing the influencing of sequencing depth and noise in data (Fu et al., 2012). Chen K et al.'s method DANPOS provided another strategy in which the nucleosome differential signal between control and treatment samples was calculated based on Poisson or binomial distributions (Chen et al., 2013). One advantage of DANPOS is that it can detect all three categories of nucleosome dynamics, position shift, fuzziness change, and occupancy change (Chen et al., 2013). Also, DANPOS is comprehensive and provides options in both normalization and difference test.

The methods mentioned above are mainly suitable for two-sample (cell types) comparison. However, in a cell, nucleosome change is a continuous biological process, namely a multiple-step process. In studies about cell tumorigenesis and differentiation, it would be critical to understand the details of each middle step. In this situation, a comparison of nucleosomes among multiple cell types is needed. Several models are the candidates to serve to this kind of task at present. MultiGPS is designed to detect the ChIP-seq enrichment across multiple cell types (Mahony et al., 2014). NucTools is nucleosome-specific software package, also allowing comparison of nucleosome occupancy landscapes between different conditions. It identifies "stable" and "unstable (fuzzy)" nucleosome through calculating the relative error of nucleosome occupancy signal in the replicates (conditions) (Vainshtein et al., 2017). In our previous work, we developed a Chi-square test based method, called Dimnp, to detect nucleosome difference among multiple cell types (Liu et al., 2017). West et al employed negative binomial distribution to identify regions of differential in nucleosome occupancy (West et al., 2014). To accurately find the boundaries and centers of the differential nucleosome regions, researchers agreed to have strategy with a hard cutoff for differential significance (often P-value) like used in DANPOS, DiNuP, Dimnp and NucTools (West et al., 2014; Fu et al., 2012; Chen et al., 2013; Vainshtein et al., 2017; Liu et al., 2017). Alternatively, to incorporate a computational frame into the raw nucleosome signal can act like a "soft cutoff". In this regards, Hidden Markov Model (HMM) appeared to be advantageous mainly to locate nucleosome position in Microarray and DNA sequence signal by setting the linker state and well positioned/ delocalized nucleosome states (Yuan and Liu, 2008; Yuan et al., 2005).

In present work, we introduced another model, called DNMHMM, for the identification of differential nucleosome regions (DNRs) in multiple cell types or conditions. Here, DNRs means the regions where nucleosome occupancy levels change across cell types or during the changing conditions. Using sequenced reads of nucleosome DNA as inputs, DNMHMM can call DNRs through a HMM. By calculating each genomic window with a hypothesis test, DNMHMM acquires observed variable sequence to infer the most likely path of hidden states, which indicates the nucleosome occupancy difference. We showed that the model can capture the differential nucleosome regions in multiple cell types, especially in locating centers of the differential nucleosomes. With the model, in yeast mutants where the modifiable histone residues were mutated into alanine, we found that DNA sequences of the dynamic nucleosomes lack 10–11 bp periodicities and harbor the motifs of the nucleosome remodelling complex (ABF1 and CBF1). During the activation of human $CD4^+$ T cells, the dynamic nucleosomes are around regulatory sites and partly account for changes of gene expression.

2. Methods

2.1. An approach to identify differential nucleosome regions in multiple samples based on Hidden Markov Mode (DNMHMM)

We proposed DNMHMM to identify the differential nucleosome regions (DNRs) where nucleosome occupancy levels usually change. DNMHMM is designed to perform through three basic steps (Fig. 1A). First, it begins with the normalized reads count data of n cell types. In normalization, reads count at each genomic locus is divided by the average reads count of the chromosome. The first step is to call observed variables for each genomic locus (window), which forms an observed variable sequence as input for HMM. Observed variable for locus i depends on the difference of reads count across the n cell types. Significance (p-value) of the difference is calculated with hypothesis test which is based on Analysis of Variance (ANOVA), Wilcoxon test and Bartlett's test. Based on the p-value, an observed variable is assigned to the locus. Generally, the observed variable can be "S" for a great p-value, and "D" for a small p-value, and "B" for a moderate p-value, corresponding to three grades of the reads count difference. By sliding the window, the observed variables are assigned to each genomic locus to form the observed variable sequence. Two cutoffs are used to determine which observed variable should be assigned to the locus according the p-value. In order to find appropriate cutoffs, DNMHMM randomly chooses one thousand genomic regions and call p-value of the reads count difference in each region across cell types, resulting into one thousand p-values. With the p-values, given a false discovery ratio (FDR) which is defined by user, DNMHMM directly estimates the cutoff that classifies the p-value for the observed variables "B" and "D". The cutoff classifying for variables "B" and "S" is mainly a value which is $> 35\%$ of the p-values.

In the second step, DNMHMM calls DNRs with a HMM using the observed variable sequence of three consecutive windows ($[i\text{-}1, i, i+1]$) as the inputs. Here, for the HMM parameters (A, B, π), two hidden states "dynamic" and "stable", corresponding to the DNR and stable nucleosome region are applied [A: transition probability matrix (2×2) between the hidden states; B: 3×2 matrix, represents distribution of the observed variables in the hidden states; π: the initial probability of each hidden state]. Under "dynamic" state, the observed variable "D" has a higher proportion than the observed variable "S"; inversely, under the "stable" state the observed variable "S" has a higher proportion. HMM model will give a most likely path of the hidden states for the observed variable sequence with Viterbi algorithm as well as the probability of the path, thus assigning one of the hidden states to each genomic locus. The genomic regions marked with state hidden "dynamic" are then considered as DNRs. The parameters A, B and π are estimated with a reads-count dataset in which the hidden states are manually identified and then empirically adjusted. DNMHMM also allows to update the parameters by maximum likelihood estimation (MLS) on the genomic regions for which the hidden states have been inferred by itself.

Finally, the adjacent DNRs with distance less than 15 base pair (bp) is combined together as final DNRs. DNMHMM exclusively defines the position and p-value for each DNR. The source code of the model is available at website http://bioseu.seu.edu.cn/matweb/index. It should be noted that the purpose of adding HMM onto the hypothesis test is to suppress the noise in case of a low sequencing depth and to provide a precise location of differential nucleosome.

2.2. Enrichment analysis

Enrichment analysis was carried out with KEGG pathway and GO term data using the functional annotation table module of DAVID (http://david.abcc.ncifcrf.gov/).

2.3. Matching percentage between differential nucleosome regions (DNRs) identified by different methods

A matching percentage was calculated to obtain the matching degree between the two methods. Given the deviation (d) from 1 bp to 100 bp, the matching percentage (P) represented the number of matched DNRs between two methods relative to the total number of identified peaks in the first method. That is: $P = n/N_1$, where n is the matching DNRs given d, N_1 is number of DNRs calculated by the first method.

2.4. Identifications of nucleosomal dyad sites

Nucleosomal dyad coordinates were identified from the normalized nucleosome occupancy profile with a wavelet transformation-based peak-finding algorithm under the setup of (i) peak height > 1.2 and (ii) full width at half height ≥73 bp (the "mspeaks" function in bioinformatics toolbox in MATLAB [R2015]). Function "mspeaks" returns local maxima of a signal by incorporating wavelet transformation-based denoising method. Here, we set Mexican hat function as mother wavelet, and 3 as decomposing level.

2.5. Nucleosome occupancy profile

The raw reads were mapped to genome using Bowtie (Langmead et al., 2009). Prior to call reads count at each genomic locus, each read was extended to 73 bp in the 3′ direction and shifted by 73 bp towards the 3′ direction. The reads count at a genomic locus represents the count of the reads covering the locus. The reads counts were then normalized by dividing the average reads count of the chromosome.

2.6. Correlation functions

Correlation functions measure the probability to find certain nucleotide pair at a distance of k bp (Eq. (1)).

$$C_{xy}(k) = \frac{N_{xy}(k)/L}{P_x \times P_y} \quad (1)$$

Where, C represents the correlation of nucleotides x and y at a distance of k bp; $N_{xy}(k)$ is number of x–y pairs in a distance k bp in a DNA sequence with length of L bp; P_x and P_y are the probabilities of the nucleotides x and y, respectively.

2.7. Motifs in nucleosomal DNA sequence

We extracted 500 DNA sequences from the most dynamic and most stable nucleosomes with each sequence of 147 base pairs in length. Considering that the more transcription factor binding sites would be found in the dynamic nucleosome DNA than in the stable one, we therefore, employed a bioinformatics tool Homer to find the motifs in these two kinds of DNA sequences (Heinz et al., 2010).

3. Datasets

We determined nucleosome occupancy in 22 mutant strains of *S. cerevisiae,* listed in Table S1 (Liu et al., 2015). In each mutant, a modifiable histone residue was mutated into alanine (A). The reads count at each genomic site was counted to indicate the nucleosome occupancy. The data of reads count of each strain is available at website http://bioinfo.seu.edu.cn/Nu_dynamics_data_public/ (Liu et al., 2015). Also, we retrieved nucleosome occupancy data and gene expression data of human resting and activated $CD4^+$ T cells from the literature (https://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.aspx) (Schones et al., 2008). Data of transcription frequency for wild-type strains of *S. cerevisiae* was also retrieved from the literature (Holstege et al., 1998).

4. Results and discussion

4.1. DNMHMM

Schematic overview of DNMHMM is shown in Fig. 1A. This HMM-based model consists of three steps: calling observed variable sequence as HMM's input, inferring hidden state sequence with Viterbi algorithm as DNRs, and by combining the short DNRs to make final differential regions. The observed variable is calculated with the concrete hypothesis tests for each genomic locus. There are two hidden states, "dynamic" and "stable", in the HMM, corresponding to the stable and dynamic nucleosomes. The further parameters like transition matrix (A) and distribution matrix (B) are derived from an empirical estimation.

In order to test the capacity of DNMHMM, we arbitrarily chose six cell types of yeasts carrying the histone mutations (Table S2), namely, H3WT, H3K56A, H3K36A, H3K4A, H3K79A and H3S10A, and identified their respective DNRs. The nucleosome occupancy profiles of the cell types are shown in the second panel of Fig. 1B. The normalized reads data was first evaluated through an ANOVA which generated an observed variable sequence (the fourth panel, Fig. 1B) according the p-value cutoffs which were obtained by FDR estimation (1×10^{10} as default, see Method section). HMM then calculated a most likely path of the hidden states that indicated the nucleosome difference across the cell types (the third panel, Fig. 1B).

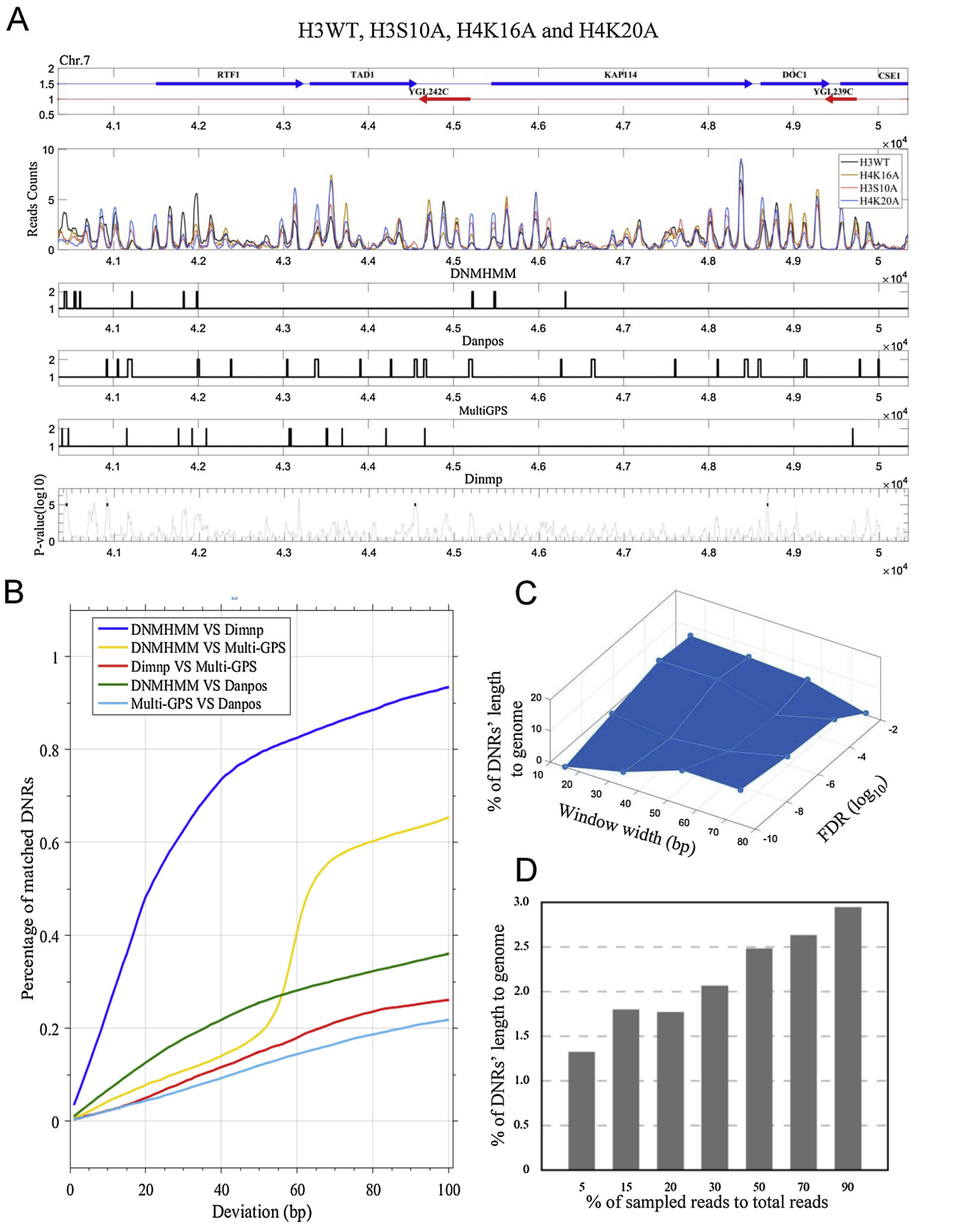
As expected, the DNRs correspond to the loci where the nucleosome occupancy varies greatly (Fig. 1B). Hence, the DNRs contained more dynamic observed variables (D), and the centers of DNRs appeared to be almost like a nucleosome dyad sites (Fig. 1B). Moreover, we found that DNRs were enriched in the intergenic regions. In the bottom panel of Fig. 1B, the model p-values (left axis) represents the probabilities of the observed variables given by the model and the right axis indicates the p-values of ANOVA for each locus. One additional example supporting our findings is also shown in Fig. S1. Briefly, the results indicate that DNMHMM is a reliable algorithm to identify DNRs among multiple cell types.

4.2. Performance evaluation

The DNRs identified by DNMHMM were compared with DNRs predicted by three independent models, namely DANPOS, MultiGPS and Dimnp. We used datasets from 4 cell types (H3WT, H3S10A, H4K16A and H4K20A) and the default parameters were set to the tools for multiple comparisons. Fig. 2A exhibits the identified DNRs for a randomly selected genomic region with all models. The DNRs by DNMHMM showed great variation of the reads counts in the four cell types. MultiGPS and DANPOS identify much more DNRs. But some of the DNRs have very low reads counts (Fig. 2A), which were mainly at the nucleosome-depleted regions in all cell types. Dimnp only used a p-value cutoff to identify DNRs, thus excluded some DNRs that can be detected by other three models. To make a comprehensive evaluation, we calculated the matching percentage between the DNRs identified by these methods. DNMHMM and Dimnp in case of multiple cell types showed a very high matching percentage (75%, deviation ≤ 40 bp) (Fig. 2B). The other comparisons showed a quite low matching percentage (Fig. 2B), which might be due to the fact that DNAPOS is stringent for comparing of two cell types while MultiGPS is very specific for ChIP-Seq data. The matching percentage between DNMHMM and MultiGPS sharply increases when the deviation is > 60 bp (Fig. 2B). In short, the results suggested DNMHMM among all models is suitable and reliable for the identification of DNRs in multiple cell types.

4.3. Parameters for DNMHMM

The two parameters of DNMHMM (the width of the sliding window



(caption on next page)

(Win-width) and the FDR cutoff) affect the number of the DNRs. Therefore, we tested the effect of these parameters by calculating the percentage of length of the DNRs to the genome. Given a smaller window width ($\leq$ 35 bp) and a greater FDR cutoff ($\geq 10^{-4}$), more DNRs were identified; while a greater window width ($\geq$ 55 bp) and a smaller FDR cutoff ($\leq 10^{-7}$) resulted into few DNRs (Fig. 2C and Fig. S2). The percentage did not vary sharply with the change of the parameters, thus indicating the robustness of the model. Sequencing depth

Fig. 2. Performance comparisons of DNMHMM and other existing methods.
A, The differential nucleosome regions identified using different methods for randomly selected genomic region. DNMHMM, FDR $< 1.0 \times 10^{-6}$, window width = 30 bp; Danpos, the DNRs with p-value $\geq 1.0 \times 10^{-6}$ were filtered; and the DNRs of every two cell types were combined and compared with results by other methods; MultiGPS, q-value < 0.1, "diffp" = 0.05; Dinmp, p-value cutoff $< 1.0 \times 10^{-5}$, window width = 10 bp, and also with global normalization method.
B, The matching percentages of the DNRs between the models under different deviations. Given a deviation *d* (x-axis), the matching percentage (*P*) represented the number of matched DNRs between two models relative to the total number of the DNRs in the first method. The parameters are same as in Fig. 1A.
C, Effect of window width and false discovery ratio on the identification of DNRs. The test was evaluated on the data of the four cell types (H3WT, H3S10A, H4K16A and H4K20). Percentage of DNRs (z-axis) means that the percentage of length of the DNRs to total length of the genomes.
D, Influence of sequencing depth on identification of DNRs. The test was also on the data of the four cell types (Table S2). Window width and FDR are fixed at 35 bp and 1.0×10^{-6}, respectively.

is another factor that affects the identification. In general, the number of identified DNRs increases with the number of the mapped reads (Fig. 2D).

4.4. Application I, analysis for nucleosome alteration in the yeast strains of histone mutation

DNMHMM was applied to the study of histone mutations in yeast which could influence the nucleosome alteration. Initially, we selected four groups of cell types (Table S2) and identified the DNRs for each group. Near transcription start sites (TSSs), nucleosomes appear to be altered in both occupancy levels and their positions (Fig. 3A and Fig. S3). The +1 nucleosome downstream of TSSs showed variations in occupancy level (Fig. 3A and Fig. S3). The first (–1) nucleosome upstream of TSSs were mainly dynamic at the position (Fig. S3). The +1 nucleosome was thought as a barrier and was directly positioned (Mavrich et al., 2008; Mobius and Gerland, 2010). Its position appears to be less susceptible/ changed to the TSS. The –1 nucleosome is statistically positioned via a nucleosome-repelling DNA region (Mobius and Gerland, 2010) and it would be shifted or removed when TFs bind to promoter, which leads to a great changes in position of the nucleosome.

Previous studies indicated that the 10–11 bp periodicities of certain dinucleotides (such as AA, TT, TA and GC) also play an important role in positioning nucleosomes (Segal et al., 2006; Liu et al., 2011). Here, we speculated that the periodicities of DNA sequences in stable nucleosomes are more pronounced than that in dynamic nucleosomes (in DNRs). Thus, we tested this hypothesis by calculating a correlation function of the dinucleotides AA, TT, TA and GC in both stable as well as in dynamic nucleosomes (Fig. 3B and Fig. S4). Obviously, the occurrence of the dinucleotides oscillates with ~10-bp periodicity in stable nucleosomes, and the phase difference between TA and AA/GC/TA was ~5 bp. Conversely, in dynamic nucleosomes, the oscillating signal was very weak (Fig. 3B and Fig. S4).

Considering the impact of TF binding sites/motifs at the nucleosomal DNA, which could potentially remodeled the nucleosomes, we therefore checked the enrichment of such motifs in the stable and dynamic nucleosomes. As expected, several TFs motifs were found in the dynamic nucleosome, including the motifs of TFs like ABF1 and CBF1, both of which are previously known to participate in nucleosome repositioning processes (Fig. 3C and Fig. S5), thus suggesting that the dynamic nucleosomal DNA sequences do harbor the motifs of chromatin-remodeling complex. However, we did not found a significant enrichment of such motifs in the stable nucleosomes. We also noticed the four groups show a different enrichment in terms of TFs.

If the nucleosome alteration associates with TFs' binding, percentage of the length of DNRs should be related to gene transcription. To this end, we plotted transcription frequency of wild type versus percentage of the DNRs at promoters and found a positive correlation between them ($r \geq 0.19$) (Fig. 3D). This also suggests that the highly expressed genes have more of dynamic nucleosomes at the promoters. Interestingly, we found that the genes with dynamic nucleosomes (DNRs) at promoters are common among four sets of comparison (Fig. S6). These results were consistent with our one previous study (Liu et al., 2015).

Taken together, with DNMHMM, some new characteristics are revealed for the dynamic nucleosomes, such as weak 10–11 bp periodicities, enriched TFs' motifs and the link to gene transcription, hence, suggesting the strong validity of DNMHMM method.

4.5. Application II, analysis for nucleosome alteration between human resting and activated CD4+ T cells

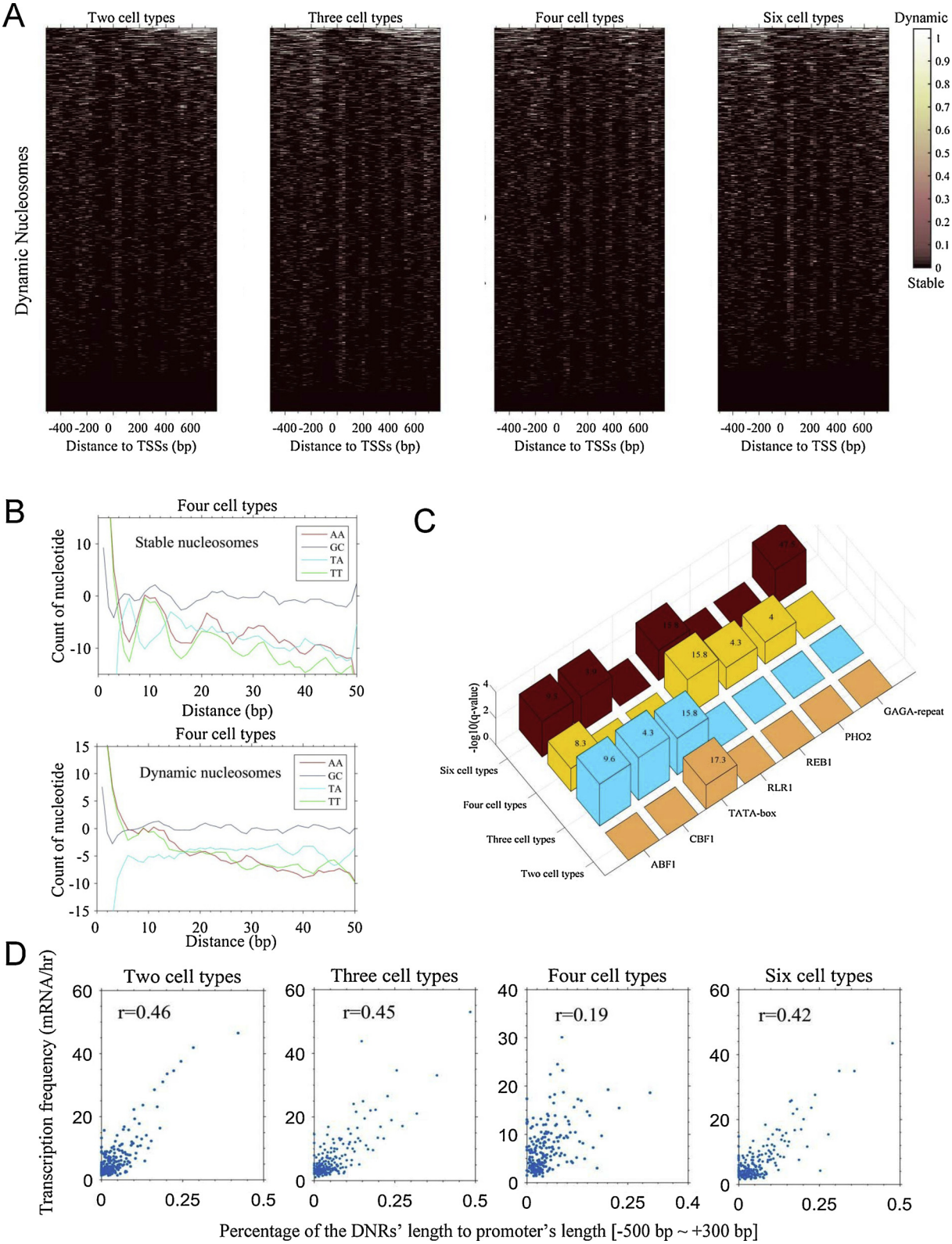
We also identified the DNRs between the resting and activated $CD4^+$ T cells. Additionally, we calculated the distributions of DNRs around these regulatory sites (Fig. 4A). Here again, we observed that the nucleosomes were dynamic around enhancer sites, TSSs and junction sites between introns and exons; while nucleosomes were stable in the intron and around the 3′ end of genes (Fig. 4A). These findings together with the results from application I, suggests that the nucleosome occupancies are different at the regulatory sites in the different cell types.

We then further investigated the association between gene expression changes and the nucleosomes dynamics. For this, we divided human 15,944 genes into 1000 groups (bins) according to the percentages of the length of the DNRs at promoters, and then calculated the percentages of the differentially expressed genes in each bin. Fig. 4B shows the percentage of the differentially expressed genes against the percentage of the length of the DNRs to the promoter. The result suggested that a gene with a high percentage of DNRs at its promoter tends to associate with a great change in expression (blue line in Fig. 4B), indicating an effect of nucleosome alterations on transcription. We also noticed that some of the genes with the dynamic nucleosomes at promoters do not exhibit a significant change in the expression (square dots close to 0). This might be due to two factors, 1) false positive of the identifications; 2) nucleosomes alterations just link to the transcription and chromatin regulators, not directly to the changes of expression, as previously discussed (Karlic et al., 2010; Huebert et al., 2012). In some cases, nucleosome alterations closely associates with expression change (Fig. 4C). For instance, at promoter of gene DUSP1, nearly 40% (10/25) of nucleosomes were dynamic and the expression of the gene was drastically reduced during activation of $CD4^+$ T cells (Fig. 4C). The product of DUSP1 can dephosphorylate MAP kinase MAPK1/ERK2, thus could impact the cellular response to stress and promote the negative regulation of cellular proliferation (Robitaille et al., 2017; Lopes et al., 2017). The enrichment analysis also suggested that genes that are related to GTPase activity and PI3K-Akt signaling pathway have a high percentage of DNRs at promoters (Fig. 4D).

Briefly, we could show that the DNMHMM can provide deeper insights into the association between nucleosome and transcription.

4.6. Web server

Finally, we built up a web server to provide online demonstration service of DNMHMM (http://bioseu.seu.edu.cn/matweb/index). This web server is built with Django (https://www.djangoproject.com/) on apache 2.4. Users' request through internet and apache is transferred to the Django through mod_wsgi-3.0 (a python package to deal with communication between Django and Apache) (Fig. S7). Django recognizes different requests and call different function by Matlab Engine



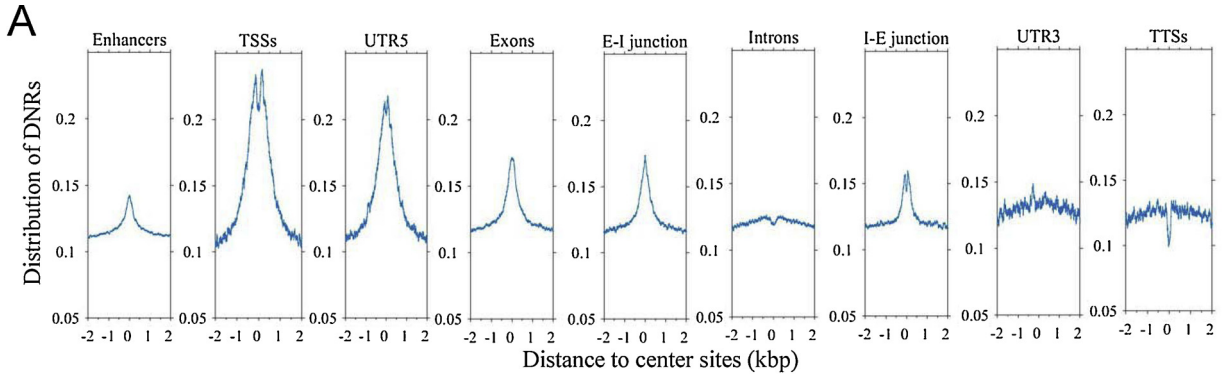
(caption on next page)

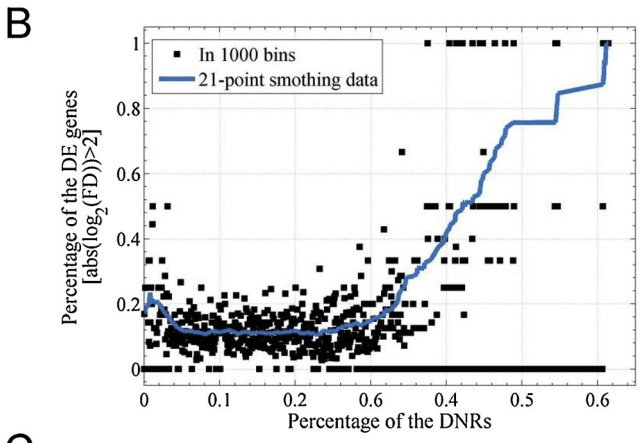
API for Python that code for Matlab as a computational engine from Python.

5. Discussion

Nucleosome alteration profoundly affects the gene transcription and closely associates with many cellular processes during development and tumorigenesis. Identifying such differential nucleosomes is a prime interest of the researchers, especially to define in the multiple cell types. In this work, we reported DNMHMM, an approach to identify DNRs, and demonstrated its two applications. DNMHMM performs with three key steps calculating the observed variable sequence based on

Fig. 3. Characteristics of dynamic nucleosomes in the yeast cells with mutations in modifiable histone residues.
A, Distribution of the DNRs around transcription start sites of 5419 genes of yeast in four groups of the multiple cell types comparison by DNMHMM; "dynamic" and "stable" states respectively represent with 1 to 0 in heating maps; the cell types of each group are listed in Table S2.
B, The correlation functions (see method) of dinucleotides AA, TT, TA and GC in 500 most stable nucleosomes (top panel) and in 500 most dynamic nucleosomes (bottom panel). Shown is for the comparison of four cell types (Table S2). Definition of the correlation function is in Eq.1 in method section.
C, The most common motifs that were found in dynamic nucleosomes; x-axis indicates the comparison group; y-axis shows the transcription factors corresponding to the motifs; z-axis is the enrichment significance ($-\log_{10}$ (q-value)) of the motifs in the top 500 dynamic nucleosomes (FDR $<$ 0.01) (also see Table S3). We did not find any motif enriched in the top 500 stable nucleosomes. The number at the top of each pillar is the percentage of motif-containing nucleosomes to total nucleosomes (500).
D, Genes with a high transcription level in wild type will have more of dynamic nucleosomal regions at promoters (–0.5 kbp ~ +0.3 kbp) with the mutation at the modifiable histone residues. Shown is the plotting of the transcription frequency (mRNA/hour) of each gene in wild type to the percentage of DNRs to promoters. The transcription frequency was sorted and then averaged for every 20 genes.





D

Terms	P-value (FDR ≤0.2)
GO: positive regulation of GTPase activity (17)	2×10^{-5}
hsa04151: PI3K-Akt signaling pathway (11)	2×10^{-4}
GO:transcriptional repressor activity, RNA polymerase II core promoter proximal region sequence-specific binding (7)	4×10^{-4}

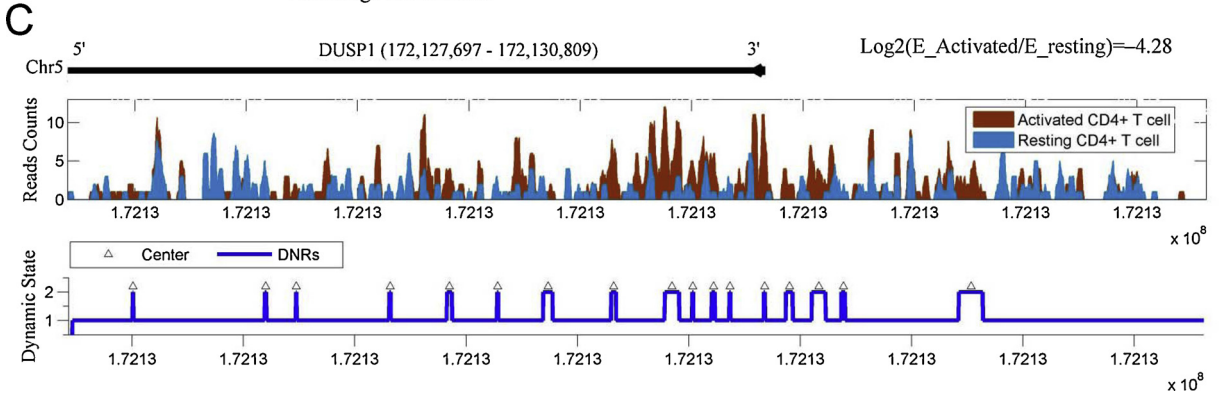


Fig. 4. The dynamic nucleosomes in activation of human $CD4^+$ T cell types.
A, Percentages of DNRs around regulatory sites. The DNRs are identified with FDR ≤ 1.0×10^{-6}, window length is 35 bp.
B, Percentage of DNRs at promoters partly accounts for gene expression changes. Shown is plotting of the percentage of DNRs at promoters (–2k bp ~ +0.5k bp relatively to TSSs) against the percentage of the differentially expressed (DE) genes (square dots). The percentage of DNRs was calculated at promoter for each gene and then was sorted out. The sorted percentages were divided into 1000 bins. In each bin, the percentages of the DE genes were calculated. A DE gene means the expression of the gene with a fold change (FD) of either ≥ 4 or ≤ 1/4 in activation of CD4 + T cells. The blue line shows the data after a 21-point smoothing.
C, The DNRs for gene DUSP1, top panel shows reads counts in resting and activated $CD4^+$ T cells; bottom panel shows the DNRs identified (FDR = 1.0×10^{-6}, window width = 35 bp;); The gene exhibits a more than eight fold decrease in the resting cells than in the activated cells.
D, Enrichment analysis for the top 200 genes with the most DNRs.

hypotheses test, inferring the DNRs with HMM and finally combining the DNRs.

Furthermore, DNMHMM harbors two major advantages: 1) it is suitable for multiple cell types (n ≥ 2) which makes it unique among other previously known methods. It can reveal some interesting characteristics of dynamics nucleosomes, which we have showed successfully by using histone mutants of yeast and activation of human $CD4^+$ T cells. 2) The robustness and convenience of this approach with high efficiency just only by using two basic parameters (FDR of 1.0×10^{-6} and window width of 35 bp). Furthermore, DNMHMM can work on the dataset with a poor sequencing depth. This was evident from our analysis in application II, where the average reads coverage were ~1 bp in $CD4^+$ T cells, still the approach worked efficiently. The distribution of the DNRs and the association between the DNRs and the expression justify reasonability of the identification.

Although, this model cannot describe the types of nucleosome alteration such as position shift, occupancy alteration or fuzzy positioning, still, this can be optimized by pair comparison after DNRs' identifications in multiple cell types. However, due to step of the sliding window the analysis will take bit longer, which can be further improved by using a bigger step.

Taken together, DNMHMM provides all the necessities which are required to identify the dynamic nucleosomes in multiple cell types.

Authors' contributions

HDL conceived the project and completed the core program. JHX designed the web server. HDL, YRC and HML performed the computational analysis. HDL, YRC and AS wrote the manuscript. All authors analyzed and discussed the results.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (No. 31371339, No. 61972084 and No. 81660471).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.biosystems.2019.104033.

References

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., Richmond, T.J., 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389 (6648), 251–260.
Jiang, C., Pugh, B.F., 2009. Nucleosome positioning and gene regulation: advances through genomics. Nat. Rev. Genet. 10 (3), 161–172.
Hu, Z., Chen, K., Xia, Z., Chavez, M., Pal, S., Seol, J.H., Chen, C.C., Li, W., Tyler, J.K., 2014. Nucleosome loss leads to global transcriptional up-regulation and genomic instability during yeast aging. Genes Dev. 28 (4), 396–408.
West, J.A., Cook, A., Alver, B.H., Stadtfeld, M., Deaton, A.M., Hochedlinger, K., Park, P.J., Tolstorukov, M.Y., Kingston, R.E., 2014. Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. Nat. Commun. 5, 4719.
Henikoff, S., 2016. Mechanisms of nucleosome dynamics in vivo. Cold Spring Harb. Perspect. Med. 6 (9).
Wilson, B.G., Roberts, C.W., 2011. SWI/SNF nucleosome remodellers and cancer. Nat. Rev. Cancer 11 (7), 481–492.
Makova, K.D., Hardison, R.C., 2015. The effects of chromatin organization on variation in mutation rates in the genome. Nat. Rev. Genet. 16 (4), 213–223.
Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G., Zhao, K., 2008. Dynamic regulation of nucleosome positioning in the human genome. Cell 132 (5), 887–898.
Polishko, A., Ponts, N., Le Roch, K.G., Lonardi, S., 2012. NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model. Bioinformatics 28 (12), i242–249.
Becker, J., Yau, C., Hancock, J.M., Holmes, C.C., 2013. NucleoFinder: a statistical approach for the detection of nucleosome positions. Bioinformatics 29 (6), 711–716.
Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M., Iyer, V.R., 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol. 6 (3), e65.
Fu, K., Tang, Q., Feng, J., Liu, X.S., Zhang, Y., 2012. DiNuP: a systematic approach to identify regions of differential nucleosome positioning. Bioinformatics 28 (15), 1965–1971.
Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., Li, W., 2013. DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. Genome Res. 23 (2), 341–351.
Mahony, S., Edwards, M.D., Mazzoni, E.O., Sherwood, R.I., Kakumanu, A., Morrison, C.A., Wichterle, H., Gifford, D.K., 2014. An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding. PLoS Comput. Biol. 10 (3), e1003501.
Vainshtein, Y., Rippe, K., Teif, V.B., 2017. NucTools: analysis of chromatin feature occupancy profiles from high-throughput sequencing data. BMC Genomics 18 (1), 158.
Liu, L., Xie, J., Sun, X., Luo, K., Qin, Z.S., Liu, H., 2017. An approach of identifying differential nucleosome regions in multiple samples. BMC Genomics 18 (1), 135.
Yuan, G.C., Liu, J.S., 2008. Genomic sequence is highly predictive of local nucleosome depletion. PLoS Comput. Biol. 4 (1).
Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., Rando, O.J., 2005. Genome-scale identification of nucleosome positions in S. cerevisiae. Science 309 (5734), 626–630.
Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (3), R25.
Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38 (4), 576–589.
Liu, H., Wang, P., Liu, L., Min, Z., Luo, K., Wan, Y., 2015. Nucleosome alterations caused by mutations at modifiable histone residues in Saccharomyces cerevisiae. Sci. Rep. 5, 15583.
Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., Young, R.A., 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95 (5), 717–728.
Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I., Pugh, B.F., 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res. 18 (7), 1073–1083.
Mobius, W., Gerland, U., 2010. Quantitative test of the barrier nucleosome model for statistical positioning of nucleosomes up- and downstream of transcription start sites. PLoS Comput. Biol. 6 (8).
Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., Widom, J., 2006. A genomic code for nucleosome positioning. Nature 442 (7104), 772–778.
Liu, H., Lin, S., Cai, Z., Sun, X., 2011. Role of 10-11bp periodicities of eukaryotic DNA sequence in nucleosome positioning. Bio Systems 105 (3), 295–299.
Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K., Vingron, M., 2010. Histone modification levels are predictive for gene expression. Proc. Natl. Acad. Sci. U. S. A. 107 (7), 2926–2931.
Huebert, D.J., Kuan, P.F., Keles, S., Gasch, A.P., 2012. Dynamic changes in nucleosome occupancy are not predictive of gene expression dynamics but are linked to transcription and chromatin regulators. Mol. Cell. Biol. 32 (9), 1645–1653.
Robitaille, A.C., Caron, E., Zucchini, N., Mukawera, E., Adam, D., Mariani, M.K., Gelinas, A., Fortin, A., Brochiero, E., Grandvaux, N., 2017. DUSP1 regulates apoptosis and cell migration, but not the JIP1-protected cytokine response, during Respiratory Syncytial Virus and Sendai Virus infection. Sci. Rep. 7 (1), 17388.
Lopes, L.J.S., Tesser-Gamba, F., Petrilli, A.S., de Seixas Alves, M.T., Garcia-Filho, R.J., Toledo, S.R.C., 2017. MAPK pathways regulation by DUSP1 in the development of osteosarcoma: potential markers and therapeutic targets. Mol. Carcinog. 56 (6), 1630–1641.