Genome Biology

METHOD

Open Access

NucHMM: a method for quantitative modeling of nucleosome organization identifying functional nucleosome states distinctly associated with splicing potentiality

Kun Fang[1], Tianbao Li[2], Yufei Huang[3] and Victor X. Jin[2*]

* Correspondence: jinv@uthscsa.edu
[2]Department of Molecular Medicine, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA
Full list of author information is available at the end of the article

Abstract

We develop a novel computational method, NucHMM, to identify functional nucleosome states associated with cell type-specific combinatorial histone marks and nucleosome organization features such as phasing, spacing and positioning. We test it on publicly available MNase-seq and ChIP-seq data in MCF7, H1, and IMR90 cells and identify 11 distinct functional nucleosome states. We demonstrate these nucleosome states are distinctly associated with the splicing potentiality of skipping exons. This advances our understanding of the chromatin function at the nucleosome level and offers insights into the interplay between nucleosome organization and splicing processes.

Keywords: Nucleosome organization, Hidden Markov model, Splicing potentiality

Background

A nucleosome is the fundamental structural unit of eukaryotic chromatin and nucleosome core is formed by the wrapping of 146-bp DNA in 1.75 left-handed superhelices around a histone octamer [1–3]. Nucleosome organization, described as nucleosomal phasing, spacing, and positioning, is determined by the interplay among nucleosome, nucleosome-binding factors such as DNA-binding factors, histone chaperones, and ATP-dependent chromatin remodelers [4, 5]. Several models, supported by substantial experimental findings, have been proposed for determining nucleosome organization: (1) DNA-binding factors or ATP-dependent chromatin remodelers forcing nucleosome depletion in certain genomic regions [6–8]; (2) the intrinsic DNA sequence patterns preferring histone binding [9–11]; and (3) a barrier statistically favoring deposition of a well-positioned nucleosome and forcing the periodic positioning of all other nucleosomes [12]. Despite of these elegant models, there still lacks a quantitative model to determine the combinational effects of the different influencing factors on nucleosome organization. For example, can nucleosome organization be quantitatively classified into distinct nucleosome states? How many nucleosome states are there in an epigenome? How many characteristic features are there in a particular nucleosome state? What are the relationships among these features? Are nucleosome states cell type-specific and/or genomic regional-specific?

Many studies have revealed that nucleosome organization plays a key role in the regulation of gene expression [4, 5, 13–15]. Genome-wide nucleosome mapping has also provided structural and mechanistic links among nucleosome, wrapped DNA, and nucleosome-binding factors [16–18] and elucidated novel functionalities of organized nucleosomal arrays in an unbiased way [19–21]. Recent studies have found that chromatin structure, in terms of nucleosome organization and specific histone modifications, acts as key regulators of alternative splicing. These studies provided evidence that there exists crosstalk between chromatin and splicing [22–24]. Among these studies, genome-wide mapping of nucleosomes has clearly illustrated the enrichment of nucleosomes at intron-exon junctions [25–27]. Other works, including ours, has revealed a strong correlation between several histone modifications across the alternatively spliced regions and splicing outcome [28, 29]. However, these findings are mostly correlative and observational. Therefore, it is imperative to develop a computational model to examine their relationship quantitatively.

Although many computational methods were developed to determine epigenetic states [30–60], several limitations include that (1) some supervised learning methods such as ChromaSig [60] cannot find de novo information, and (2) some unsupervised learning methods such as HMMSeg [31], ChomHMM [35], Segway [39], and T-cep [59] cannot optimally capture spatial patterns of the epigenetic marks on the nucleosomes, and they were not designed with modeling nucleosome organization. Thus, none of the above methods can define functional nucleosome states, i.e., states encoding combinatorial histone marks and nucleosome organization features that perform specific functions and respond to the different environment and intercellular signaling. Our knowledge at the quantitative aspect is very limited about the phasing of a nucleosome array, the spacing between two dyads of the nucleosomes, the degree of nucleosome positioning, as well as the extent to which the combinatorial epigenetic pattern influences nucleosome organization. There is a lack of quantitative measures on the association of functional nucleosome states with the splicing potentiality of skipping exons (SEs).

In this study, we develop a novel computational method, NucHMM, which integrates a hidden Markov model (HMM) with the characteristics of nucleosome organization (phasing, spacing, positioning), to identify the nucleosome states associated with cell type-specific combinatorial histone marks. We test it on publicly available MNase-seq and ChIP-seq of H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K79me2, H3K9me3, and H3K27me3 data in MCF7, H1, and IMR90 cells [61] and identify cell type-specific functional nucleosome states. We further quantitatively measure the association of functional nucleosome states with the splicing potentiality of SEs. Our work advances our understanding of chromatin function at the nucleosome level and further offers mechanistic insight into the interplay between nucleosome organization and splicing process.

Results

An overview of NucHMM

To quantitatively modeling the nucleosome organization, we have developed a novel algorithm, NucHMM, to identify functional nucleosome states. NucHMM is composed of three consecutive modules: (1) initialization, (2) training, and (3) functioning (Fig. 1 and the "Methods" section). Briefly, the initialization module pre-processes the raw sequencing data into the readable data input for the training module including converting fastaq into bam, calling the peaks for ChIP-seq data by MACS2 [62] or EPIC2 [63], identifying the positioned nucleosomes from MNase-seq by iNPS [64], and binning the genome based on positioned nucleosomes where each nucleosome-bin is assigned with an observation symbol from an alphabet list of 2^n observations symbols representing each possible combination of the number (n) of histone marks. The training module is composed of two rounds of HMM training. The first round is to train multiple HMMs for 300 iterations and to select the best HMM based on the smallest BIC score. The second round is to retrain the best HMM for another 200 iterations (Additional file 1: Fig. S1) after revising the input as aborting the states with very few bins (lower than 0.5% of the total nucleosomes or a user-defined cutoff) and evenly redistributing the transition probabilities of the aborted states to the remaining states. The resulting HMM further uses the Viterbi decoding algorithm to obtain the HMM states at the

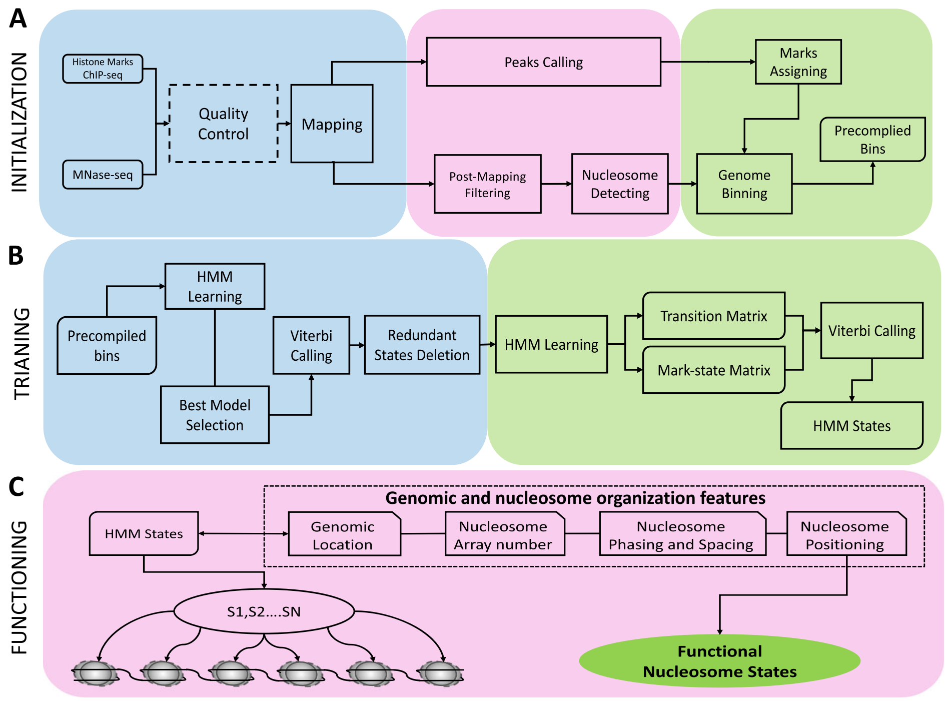
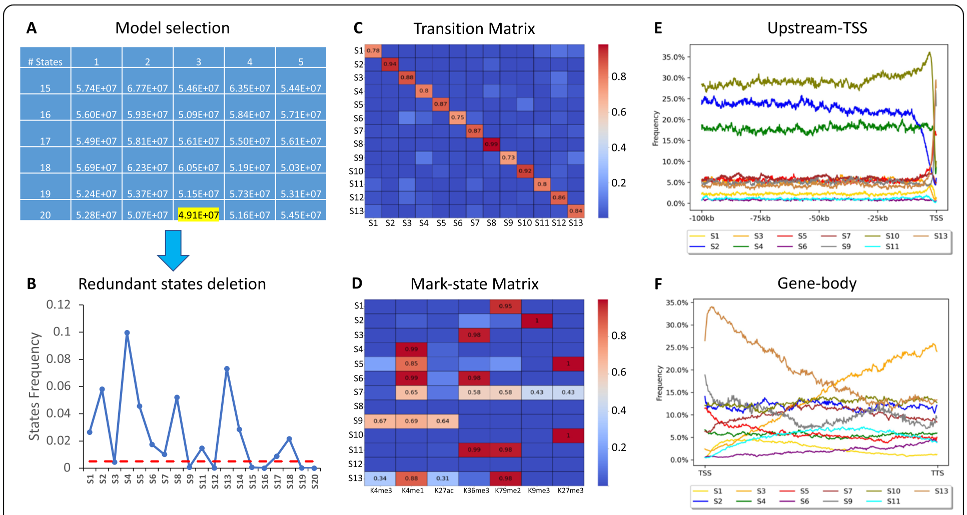


Fig. 1 An overview of NucHMM workflow. **A** The initialization step combines several existing tools to construct nucleosome-level HMM training sequences. **B** The training step includes a selection of the best model and two rounds of HMM training for the best selected model. BW algorithm is applied to acquire the transition probability matrix and the mark-state matrix derived from the emission probability matrix. **C** The functioning step performs functional screening on the nucleosomes based upon the genomic location, nucleosome array number, nucleosome spacing, phasing, and positioning and finally identifies the functional nucleosome states

nucleosome level. The functioning module defines the functional nucleosome states (NucSs) by associating each of HMM states with genomic and nucleosome organization features, including (1) genomic location—identifying the most enriched genomic regions for each of HMM states; (2) an average number (Ave No.) of nucleosomes—identifying an average of the number of nucleosomes in a nucleosome array for each of HMM states, where a nucleosome array is defined as a set of nucleosomes that have the same state and the distance between two adjacent nucleosomes is less than 350 bp (Additional file 1: Fig. S2); (3) nucleosome phasing and spacing—determining nucleosome phasing score and average (Ave) spacing for each of HMM states by filtering out those nucleosomes if their spacing is out of a defined range (see the "Methods" section—Eq. 6); and (4) nucleosome positioning—determining nucleosome positioning score for each of HMM states by firstly building the group containing all nucleosome positioning scores for each state and then filtering out the nucleosomes with user-defined nucleosome positioning cutoff (by default, NucHMM will use 0.05 and 0.95 quantile values of the group as the cutoff).

Selecting the best HMM and determining the genomic location

We tested NucHMM in publicly available MNase-seq and ChIP-seq of H3K4me1, H3K4me3, H3K27ac, H3K9me3, H3K27me3, H3K36me3, H3K79me2 data in MCF7, H1, and IMR90 cell types (Additional file 1: Tables S1-2). We used iNPS to identify 11.6, 11.9, and 12.7 million genome-wide positioned nucleosomes in MCF7, H1, and IMR90 cell types, respectively. Since the functional nucleosomes are likely located in close to 5′transcription start site (5TSS), we chose a gene-centric genomic region for training HMM ranging from − 100Kb upstream to 5TSS (Upstream-TSS), gene body (Gene-body), and + 10Kb downstream of transcription terminal site (TTS) (Downstream-TTS) (Additional file 1: Suppl. Notes). Thus, only around 7.2, 7.4, and 7.3 million positioned nucleosomes for MCF7, H1, and IMR90 cell types were used for the first round training. We trained a total of 50 HMMs with five initial states ranging from 15 to 25 each repeated by five times and selected the best model with 20 initial states based on its smallest BIC score, 4.91E+07 (Fig. 2A and Additional file 1: Tables S3-5). We found seven states in the best model that are redundant (Fig. 2B) and thus removed them before the second round of training (Additional file 1: Suppl. Notes and Fig. S3). We finally achieved a model with 13 HMM states with a transition matrix showing the transition probabilities among states (Fig. 2C) and a mark-state matrix showing the emission probabilities for each of the seven marks in each of the 13 HMM states (Additional file 1: Fig. 2D and Fig. S4). We compared our HMM states to ChromHMM/Segway states and confirmed that our HMM is capable of capturing the chromatin states with the improved nucleosome level (Additional file 1: Suppl. Notes and Figs. S5-8). We further performed genomic location analysis and observed state 2 with H3K9me3 mark, state 4 with H3K4me1, and state 10 with H3K27me3 mark were highly enriched in the Upstream-TSS, particularly in Distal/Proximal (− 100Kb to − 1Kb upstream to 5TSS), and state 5 with H3K27me3/K4me1 marks, state 7 with H3K4me1/K36me3/K79me2/K9me3/K27me3 marks, and state 9 with H3K4me1/K4me3/K27ac marks were modestly enriched in the same region (Fig. 2E). As expected, states with H3K36me3 or H3K79me2 marks including state 1 with H3K79me2, state 3 with H3K36me3, state 6



# States	1	2	3	4	5
15	5.74E+07	6.77E+07	5.46E+07	6.35E+07	5.44E+07
16	5.60E+07	5.93E+07	5.09E+07	5.84E+07	5.71E+07
17	5.49E+07	5.81E+07	5.61E+07	5.50E+07	5.61E+07
18	5.69E+07	6.23E+07	6.05E+07	5.19E+07	5.03E+07
19	5.24E+07	5.37E+07	5.15E+07	5.73E+07	5.31E+07
20	5.28E+07	5.07E+07	4.91E+07	5.16E+07	5.45E+07

Fig. 2 Selecting the best HMM and defining genomic regions for each of HMM states. **A** We trained 50 HMM models (other 25 models were shown in Additional file 1: Table S4) with different numbers of initial states and select the best model with the smallest BIC score (the highlighted model). **B** A line plot showed seven states are redundant in the current "best" model. We applied the second round HMM training by removing those seven redundant states. **C** The transition probabilities of the final 13-states HMM. The transitions were from states on *y*-axis to the *x*-axis. **D** The mark-state probabilities that derived from the emission probabilities. Each column represents a histone mark and each row represents a HMM state. **E** A distribution of each of HMM states in 100Kb Upstream TSS. **F** A distribution of each of HMM states in the gene body

with H3K4me1/K36me3 marks, state 11 with H3K36me3/K79me2 marks, and state 13 with H3Kme1/K4me3/K27ac/K79me2 were highly enriched in the Gene-body and Downstream-TTS (Fig. 2F and Additional file 1: Figs. S9-11). States 8 and 12 were not enriched with any known marks, thus not included for further functional characterization.

Determining nucleosome phasing and spacing

Nucleosome phasing and spacing are two main features to characterize nucleosome organization (Fig. 3A). We mathematically defined the nucleosome phasing score and spacing value based on the distribution signals of the nucleosome arrays (see the "Methods" section). We first plotted nucleosome array frequency and clearly observed distinct coverage patterns associated with each of HMM states (Fig. 3B). We then calculated the phasing score for each of HMM states by Welch's method and found that states 5, 9, and 10 have the highest phasing score (Fig. 3C and Additional file 1: Fig. S12), suggesting that H3K4me1 and H3K27me3 marks may be capable of imposing a better organized nucleosome array. We then derived the average of nucleosome spacing for each of HMM states after averaging four nucleosome spacing values within the 1Kb nucleosome array (Fig. 3D). Interestingly, we found states 2, 3, and 10 with two repressive marks H3K9me3 and H3K27me3 and one elongation mark H3K36me3 tend to have larger nucleosome spacing values, while states 5, 6, and 9 associated with active marks H3K4me1 and H3K27ac have smaller spacing values. We further verified the reliability of our methods for calculating the phasing score and spacing value by using a simulated nucleosome array coverage signal (Additional file 1: Suppl. Notes and Fig.

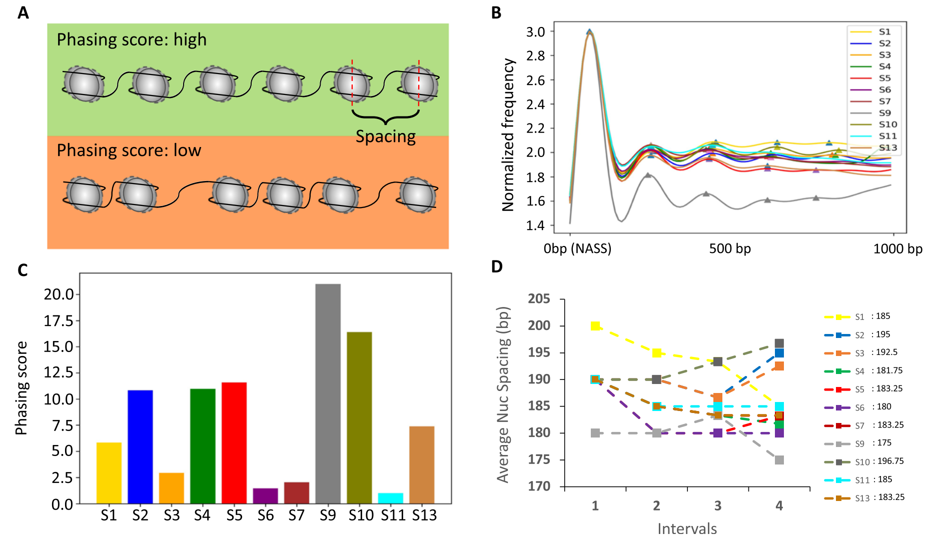


Fig. 3 Nucleosome phasing and spacing of each of HMM states. **A** A schematic diagram shows the definition of nucleosome phasing and spacing. The upper nucleosome array has a higher phasing score than the bottom nucleosome array does. Nucleosome spacing is the distance between two dyads of the nucleosomes. **B** A nuc-array-coverage plot showed the normalized nucleosome array frequency coverage of each state in the 1 kb range. 0 is the start position of the array with a certain HMM state. **C** A bar plot showed the phasing score of each HMM state calculated by Welch's method. **D** A line plot of nucleosome spacing. The interval is the distance between two peaks in panel B which is also the nucleosome spacing. We averaged 4 nucleosome spacing in the 1Kb nucleosome array as the nucleosome spacing value. The final nucleosome spacing values for each state were shown next to the legend

S13). Taken together, our results strongly suggest nucleosome phasing and spacing are intimately correlated with distinct functionality of different histone marks.

Determining nucleosome positioning and defining functional nucleosome states

Nucleosome positioning is the most important characteristic of nucleosome organization and is often qualitatively classified as well-positioned or fuzzy (Fig. 4A—upper). We first defined a reference nucleosome positioning (rNP) score based on the pile-up of raw reads (see the "Methods" section—Eq. 7 and Fig. 4A—lower). We then tested nine empirical equations (Additional file 1: Suppl. Notes) on the positioned nucleosomes by the Pearson correlation with rNP to derive a final equation to determine the NP score or the degree of nucleosome positioning (see the "Methods" section—Eq. 8 and Fig. 4B, Additional file 1: Fig. S14). We found the distributions of the nucleosome positioning scores showed a slightly difference among HMM states (Fig. 4C) and defined the mean of each distribution as the nucleosome positioning score. An IGV visualization of a genomic region for the nucleosome reads distribution and the positioning score calculated by Eq. 8 was shown in Fig. 4D.

After examining the HMM states with four genomic regions and nucleosome organization features, we defined 11 functional nucleosome states (NucSs) (Table 1) with a detailed description and visualization of each of 11 NucSs in Additional file 1: Suppl. Notes and Fig. S15. Cell type-specific NucSs-genes analysis and the lists were found in Additional file 1: Suppl. Notes, Fig. S16 and Additional files 2, 3 and 4, respectively.

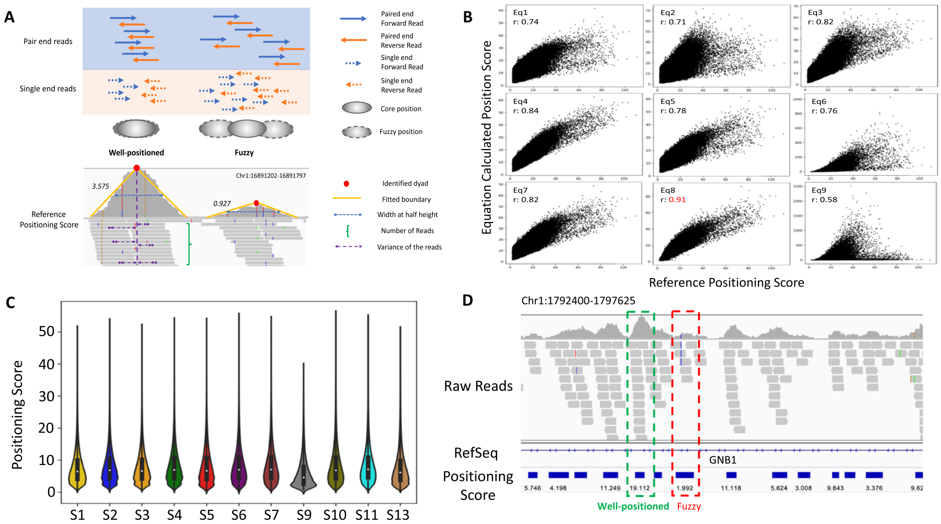


Fig. 4 Nucleosome positioning of each of HMM states. **A** The illustration of nucleosome positioning definition and reference positioning score calculation. The upper panel showed the definition of well-positioned as well as fuzzy nucleosome. The bottom panel schematically drew the elements that were used to calculate the reference nucleosome positioning: width at half height restricted by fitted boundary, the number of reads and the variance of the reads centered by the identified dyad. **B** The correlation between each iNPS-derived positioning equation and reference positioning score equation in MCF7 cell type. Equation 8 showed the best correlation with the reference positioning score. **C** The violin plots showed the distribution of the nucleosome positioning score of each HMM states. **D** IGV visualized the nucleosome reads distribution and the positioning score calculated by Eq. 8. Green box shows the well-positioned nucleosome with a high positioning score, and red box shows the fuzzy nucleosome with a low positioning score

Determining the splicing potentiality of SEs

H3K79me2 mark has been reported to be functionally associated with elongation and splicing processes [28, 29]; we were thus particularly interested in understanding the functional relationship of SEs with NucS1 (elongation accelerator), NucS7 (elongation processor), NucS10 (elongation speeder,) and NucS11 (elongation initiator), four nucleosome states enriched with H3K79me2 mark in the gene body. Interestingly, we observed NucS10 with both H3K79me2 and H3K36me3 marks showed the highest enrichment in exons for all three cell types (Fig. 5A). We then defined a NucS-SE affinity, a ratio of SEs associated with a NucS vs randomized SEs associated with that NucS, to semi-quantitatively determine the association between nucleosome states and SE events. We found that NucS10 again showed a higher SE affinity among all three cell types (Fig. 5B). To further determine the splicing potentiality of SEs, we also developed an empirical equation to quantify the splicing potentiality for each of four nucleosome states, where we assessed the splicing potentiality from three following aspects: (1) Fréchet distance between the nucleosome distribution of reliable SE (rSE) and unreliable SE (urSE) (Additional file 1: Fig. S17); (2) the difference of nucleosome positioning between nucleosomes in rSE and urSE (Additional file 1: Fig. S18); and (3) the normalized counts coefficient of each H3K79me2 related NucS (see the "Methods" section and Eq. 9). Remarkably, we found the potentiality score of NucS10 is the highest among all four H3K79me2 related NucSs (Fig. 5C). Together, our results suggest nucleosomes modified with H3K36me3 and H3K79me2 histone tails might play an important role in influencing the skipping exon processing due to its lowest phasing and a higher degree of positioning.

Table 1 The definition of functional nucleosome states

Functional nucleosome states (HMM states)	Histone marks	Ave. no. of nucleosomes	Genomic location	Ave. spacing	Positioning score	Phasing score
Elongation accelerator NucS1 (S1)	K79me2	3.99	5′-gene body	185.00	7.97	5.84
Compacting organizer NucS2 (S2)	K9me3	8.49	Distal	195.00	8.48	10.82
Elongation stabilizer NucS3 (S3)	K36me3	5.82	3′-gene body	192.50	8.06	2.93
Accessible booster NucS4 (S4)	K4me1	4.13	Distal	181.75	8.47	10.96
Primed intermediator NucS5 (S5)	K4me1/K27me3	4.56	Distal/ promoter	183.25	8.75	11.58
Elongation terminator NucS6 (S6)	K4me1/K36me3	3.74	3′-gene body	180.00	8.61	1.46
Elongation processor NucS7 (S7)	K4me1/K36me3/ K79me2/K9me3/ K27me3	5.32	Gene body	183.25	8.90	2.04
Transcriptional stimulator NucS8 (S9)	K4me3/K4me1/ K27ac	2.09	Promoter	175.00	6.07	21
Crowding controller NucS9 (S10)	K27me3	7.37	Distal	196.75	8.82	16.4
Elongation speeder NucS10 (S11)	K36me3/K79me2	4.55	Gene body	185.00	8.72	1
Elongation initiator NucS11 (S13)	K4me3/K4me1/ K27ac/K79me2	4.67	Promoter/ 5′-gene body	183.25	7.81	7.39

Discussion

Despite several existing computational methods for determining epigenetic states, none of them is able to quantitatively examine the relationship of nucleosome organization, histone marks, and genomic regions at a finer nucleosome resolution level. To the best of our knowledge, our NucHMM is the first computational algorithm and tool to identify functional nucleosome states associated with cell type-specific combinatorial histone marks and nucleosome organization. We rigorously trained and tested it on all publicly available MNase-seq and ChIP-seq data of various histone marks in MCF7, H1 and IMR90 cells. We were able to identify 11 cell type-specific functional nucleosome states, each encoded with specific biological meanings (Table 1). Importantly, NucHMM is applicable to train MNase-seq and ChIP-seq of various histone marks in many different cell types.

To test the reliability of NucHMM results, we first compared “Training” module of NucHMM with ChromHMM and Segway to evaluate its performance. We found that both NucHMM “Training” module and ChromHMM/Segway produced similar results in terms of HMM states with distinct combinatorial histone marks (Additional file 1: Figs. S5-6). We then used a simulated nucleosome array coverage signal to verify the fidelity of our methods for calculating the phasing score and spacing value (Additional file 1: Fig. S13). Remarkably, the spacing value calculated directly from the simulation sine function is consistent with the one calculated from NucHMM. The phasing score from the simulated signal is also in line with our knowledge. Finally, we constructed

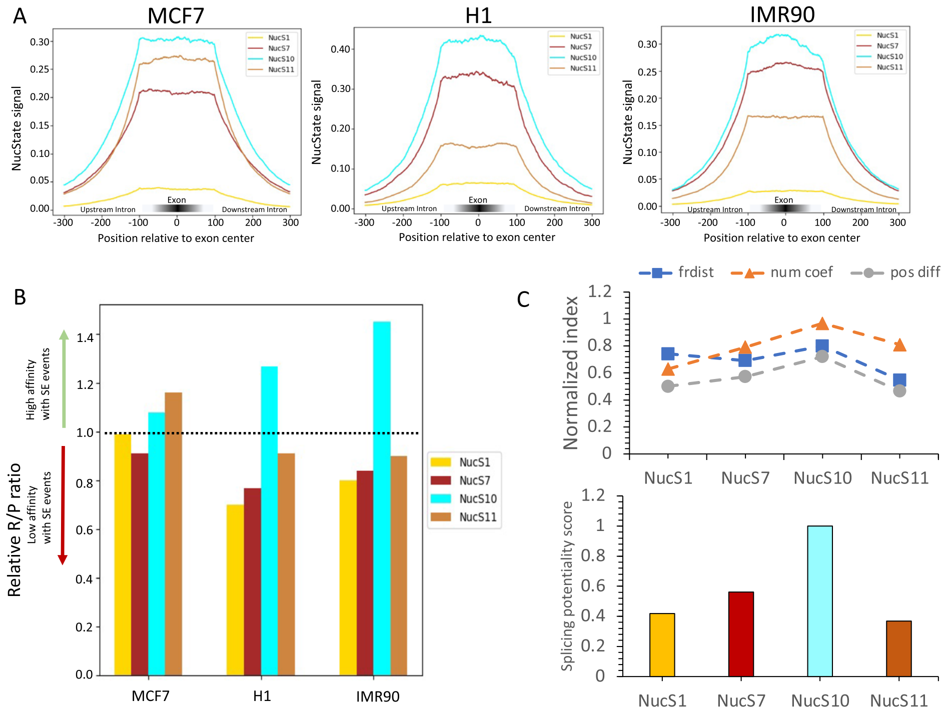


Fig. 5 The functional nucleosome states associated with the splicing potentiality of SEs. **A** H3K79me2-related NucSs signals center on exons. NucS10 showed the highest enrichment in exons. **B** A bar plot showed different SE affinity of each H3K79me2-related NucSs in each cell type. Briefly, we counted the raw number of SE events for each H3K79me2-related NucSs in each cell type and then assumed SE event was randomly associated with NucSs to get the predicted SE events. Finally, we used the ratio of raw vs predicted number as the SE event affinity for each H3K79me2-related NucS. NucS10 showed a higher SE affinity among all three cell types. **C** Semi-quantifying the SE potentiality of each H3K79me2-related NucSs. The NucS10 consistently had the highest SE potentiality score. We assessed SE potentiality from three aspects: the Frechet distance between the nucleosome distribution of SE and no-SE; the nucleosome positioning difference between nucleosomes in SE and no-SE region; the normalized SE events of each NucS

the equation for measuring nucleosome positioning with the validation by reference positioning score (Fig. 4) and Nucleosome Dynamics [65] (Additional file 1: Suppl. Notes and Figs. S19-20).

There are several notable strengths of NucHMM. Firstly, we built directional nucleosome-based observations in the "Initialization" module and used it for the univariate HMM "Training" module. The nucleosome-level observations allow us to annotate the combinatorial histone modifications on the nucleosomes (Additional file 1: Suppl. Notes and Figs. S6A, S7B-D, S8A and S8C) and also to capture the 5′ TSS more accurately (Additional file 1: Fig. S8B). While univariate HMM enumerates each possible combination of histone marks as the possible output of HMM, it more straightforwardly determines whether a particular histone mark occurs in a state compared to a multivariate HMM, which also enhances NucHMM ability to precisely annotate HMM states on a nucleosome. Furthermore, the directionality information provides a more realistic model of the underlying epigenetic patterns and their transitions. Secondly, we employed the "Functioning" module to convert HMM states to functional nucleosome states (NucSs), which are associated with not only combinatorial histone modifications, but also with nucleosome organization features, including nucleosome phasing, spacing and positioning (Additional file 1: Fig. S21). This extra layer of nucleosome organization information expands the features space of genomic states from one dimensional (traditional chromatin states) into two dimensional (functional nucleosome states), which, for the first time, offers an opportunity to genome-wide study the interplay of epigenetic marks-nucleosome organization. But there are few limitations of NucHMM: (1) the initial number of HMM states needs to be estimated at the beginning of NucHMM training, (2) the increased number of states and number of nucleosomes requires more computational and memory resources, (3) the initial assignment of a histone mark within a nucleosome bin may not be very accurate since the overlapping criteria between a nucleosome bin and a histone mark peak is a little bit subjective, and (4) the cutoff threshold of the emission probability in the mark-state matrix is arbitrary for determining whether the histone marks should be included into a state. To mitigate these limitations, future improvements may be focused on implementing a parallel computing framework, optimizing the assignment of histone marks and using a statistical method to devise the initial number of HMM states and to define a cutoff threshold of the emission probability.

Importantly, we were able to associate gene body functional nucleosome states with publicly available RNA-seq to quantitatively measure the splicing potentiality. Our quantitative comparison of the influence of four gene body nucleosome states on SE events revealed that NucS10 has the highest SE potentiality (Fig. 5C). This might due to its higher distribution at the middle gene body (Fig. 2F), its lowest nucleosome phasing (Fig. 3C), and its higher degree of positioning (Fig. 4C), as well as its most enrichment at internal exons (Fig. 5A). Most of the previous studies showed either H3K79me2 or H3K36me3 has a role in regulating alternative splicing [28, 29, 66]. However, our analyses clearly showed that the nucleosomes with both H3K36me3 and H3K79me2 marks might have the most effective influence in co-regulating the skipping exon processing. Our finding may offer new opportunities to interrogate the mechanisms of the functional crosstalk between H3K36me3 and H3K79me2 marked nucleosomes and the skipping exon processing.

Conclusion

In summary, we developed a novel computational method, NucHMM, for identifying cell type-specific nucleosome states. With NucHMM, we identified 11 distinct functional nucleosome states for MCF7, H1, and IMR90 cell types. We further demonstrated that these functional nucleosome states can be used to quantitatively determine the splicing potentiality of SEs. Our work advances our understanding of chromatin function at the nucleosome level and further offers mechanistic insight into the interplay between nucleosome organization and splicing process.

Methods

NucHMM initialization

To remove background noises and decrease the false positive rate of called positioned nucleosomes positioning and peaks of histone marks, we performed quality control (QC) for both MNase-seq and ChIP-seq data by using trim-galore [67]. We used bowtie or bowtie2 to uniquely map the reads to human HG19 reference genome. For MNase-seq data, we used Deeptools [68] to keep fragments within the range of 130–180 bp because of the length of the wrapped DNA of nucleosome plus the linker histone is within this range. We applied iNPS, which smoothed the MNase-seq wave profile with Laplacian of Gaussian convolution, to detect the borders of the nucleosome peaks, and then use a Poisson approximation filtering process to locate the final nucleosomes. We used MACS2 to identify narrow peaks for ChIP-seq of H3K4me1, H3K4me3, and H3K27ac but used EPIC2 to identify broad peaks for ChIP-seq of H3K9me3, H3K27me3, H3K36me3, and H3K9me3 with parameters -bin 100, -fdr 0.05, and -g 2 (or -g 5).

The entire genome was then binned based on detected nucleosomes. An alphabet of 128 (2^7) observation notations was built by enumerating each possible combination of marks (Additional file 1: Table S5) including no marks. For example, observation 9 (0b0001001) corresponds to the presence of H3K4me3 (1 = 0b0000001) and H3K79me2 (8 = 0b0001000) and the absence of all other marks. We then assigned the converted notations to the bins based on the degree of overlapping between the histone mark's peak and the nucleosome position. We limited the trained genomic region ranging from − 100Kb upstream to 5TSS (Upstream-TSS), gene body (Gene-body), and + 10Kb downstream of transcription terminal site (TTS) (Downstream-TTS) (Additional file 1: Suppl. Notes) and compiled a set of 19,189 protein-coding genes with the unique 5′TSS from UCSC RefSeq Genes.

NucHMM training

NucHMM training included two rounds of HMM learning process. In the first round, we empirically chose initial states ranging from 15 to 25 and ran five first-order HMMs for each of them. Each HMM was trained for 300 iterations to ensure the convergence using the Baum-Welch algorithm [69]. We then selected the HMM with the lowest Bayesian Information Criterion (BIC) score. Before the second round training, we removed those states with less than 0.5% of the total nucleosomes in the model from the transition probability and emission probability matrices. To simplify the HMM and maximize its states' the descriptive power, we used the modified transition probability and emission probability matrices for the second HMM learning process. The resulting HMM was trained with the Baum-Welch algorithm (Additional file 1: Suppl. Notes) for another 200 iterations to achieve the final HMM. The log-likelihood of HMM after each iteration was calculated to ensure to reach the local minimum. We found that 200 iterations were sufficient for this second round HMM to approach the convergence. The Viterbi algorithm was applied to decode HMM states on each nucleosome (Additional file 1: Suppl. Notes). The probabilities of an individual histone mark were calculated by marginalization among all output combinations of marks probabilities. The individual emission probability follows

$$Pr_{id} = \sum_{x=1}^{2^n} \left\{ P_{(x)}\ (x\&id > 0) \middle|\ 0\ (x\&id = 0) \right\}, \tag{Eq.1}$$

where n is the number of histone marks, & is bitwise AND operator, and x is the output number.

Nucleosome phasing and spacing

We first processed the Gaussian smoothed nucleosome signals from iNPS to nucleosome state-specific nucleosome array signals. We then averaged nucleosome array signals by the sum of all nucleosome array signals within the nucleosome state divided by the number of the nucleosome arrays. As the resolution of the Gaussian smoothed signal is 10 bp/point in the iNPS result, the initial sample rate of the nucleosome state array signal is 100 (10 bp/point). In order to keep the fidelity and more precisely convert the signals from the genome domain to the frequency domain, we first interpolated the signals and increased the sample rate to 1000 (1 bp/point), and then implemented Welch's method [70] to make the conversion based on the periodogram spectrum estimates, which was used to calculate the nucleosome phasing score.

For a detailed implementation, we firstly used the Hanning window function $w(n)$ to divide the nucleosome state array signal x into K available frames with M points in each frame. Each frame is represented by

$$x_m(n)\triangleq w(n)x(n+mR), n=1,2,\cdots,M-1, m=1,2,\cdots,K-1 \qquad \text{(Eq.2)}$$

where R is the window hop size.

Then, the periodogram of the m^{th} frame is given by

$$P_{x_{m,M}}(w_k)=\frac{1}{M}\left|FFT_{N,k}(x_m)^2\right|\triangleq\frac{1}{M}\left|\sum_{n=0}^{N-1}x_m(n)e^{-j2\pi nk/N}\right|^2 \qquad \text{(Eq.3)}$$

We then averaged the periodograms across the genome. The Welch estimate of power spectral density is given by

$$\hat{S}_x^W(w_k)\triangleq\frac{1}{K}\sum_{m=0}^{K-1}P_{x_m,M}(w_k) \qquad \text{(Eq.4)}$$

The simplified conversion equation between the genome domain and the frequency domain is given by:

$$freqc=\frac{fs}{\text{genome length}} \qquad \text{(Eq.5)}$$

where fs is the sample rate of the signal.

We finally focused on the power spectrum density within frequency 4–10 Hz, which corresponds to the genome domain range 100–250 bp. We used the highest spectral density value of each nucleosome state in the window and multiplied 1000 as the nucleosome phasing score.

The calculation of the nucleosome spacing value utilizes the distribution of a nucleosome state-specific nucleosome array. We computed all local maxima of the array distribution by the following two rules: (1) for sharp peaks, the local maximum is defined as any sample point whose two direct neighbors have a smaller amplitude, and (2) for flat peaks, the middle point index is considered as the local maximum. We then calculated the average distances between the maxima of peaks as the nucleosome spacing value. To determine the spacing range for a nucleosome within a NucS-specific nucleosome array, we used Eq. 6:

$$\left[\mathrm{Spacing}_{\mathrm{NucS}}-\mathrm{Interval}\times\left(5+\mathrm{Rank}*\mathrm{Coef}_{\mathrm{range}}\right),\mathrm{Spacing}_{\mathrm{NucS}}+\mathrm{Interval}\times\left(5+\mathrm{Rank}*\mathrm{Coef}_{\mathrm{range}}\right)\right] \tag{Eq.6}$$

where $\mathrm{Spacing}_{\mathrm{NucS}}$ is the average nucleosome spacing of a NucS; Interval is the order of the nucleosome minus one, e.g., for the second nucleosome in the array, its Interval is one; Rank refers to the rank for each of 11 NucSs based on their phasing scores; $\mathrm{Coef}_{\mathrm{range}}$ is a user-defined parameter that used to adjust the range with 1 bp as the default.

Nucleosome positioning

We used two inter-correlated approaches, the "raw reads" reference approach and the "iNPS-derived" approach, to determine the nucleosome positioning (NP) score. We defined the well-positioned nucleosomes would have higher positioning score than fuzzy nucleosomes in both methods. Both approaches were applied with the idea that nucleosome positioning is the geometric-mean of the nucleosome fuzziness and nucleosome occupancy. In the "raw-reads reference" approach, we measured three features: (1) the standard deviation of raw reads, (2) the enrichment of raw reads, and (3) the full width at half maximum of reads peak. The equation of this approach can be described as:

$$rNP_t=\frac{\mathrm{norm}(\mathrm{enrich}_t)}{\mathrm{norm}(\mathrm{fwhm}_t)+\ \mathrm{norm}(\mathrm{std}_t)} \tag{Eq7}$$

where $t\in\{1,2,\cdots,\mathrm{T}\}$ = nucleosome population set, and norm represents the interquartile range normalization.

For example, the numerator should be relatively small for a fuzzy nucleosome while the denominator should be large and make the nucleosome positioning score small. In the "iNPS-derived" approach, we first empirically created nine equations based on iNPS results to calculate the nucleosome positioning. We then used the Pearson correlation method to determine which equation has the highest correlation with the "raw-reads reference" approach. The final determined equation is given by:

$$NP_t=\frac{\mathrm{height}+\ \log_2\left(\mathrm{pval}_{\mathrm{peak}}\times\mathrm{pval}_{\mathrm{valley}}+1\right)+\mathrm{area}}{3\times\mathrm{width}} \tag{Eq.8}$$

where height, width, area, $\mathrm{pval}_{\mathrm{peak}}$, and $\mathrm{pval}_{\mathrm{valley}}$are all from iNPS. Generally, the numerator in the Eq. 8 reflected the occupancy measurements and denominator reflected the fuzziness measurement. Besides, we noticed that the $\mathrm{pval}_{\mathrm{valley}}$ is abnormally high at the end of the nucleosome array regardless of the shape of the real nucleosome. Thus, we manually replaced all $\mathrm{pval}_{\mathrm{valley}}$ of the last nucleosome in the array with the median value of the whole $\mathrm{pval}_{\mathrm{valley}}$ set. All elements in Eq. 8 are also applied interquartile range normalization.

Splicing potentiality of SE

We assessed SE's the splicing potentiality associated with each of four NucSs with H3K79me2 mark by measuring the difference of nucleosome organization between the reliable SE group and the unreliable SE group. We first used MISO [71] to identify the potential SE events. The reliable SE events result from applying two rules on the identified potential SE events. Rule 1: $X+Y\geq N\ and\ Y\geq 1$, where X, Y are integer counts corresponding to the number of reads in each of these categories, (1,0):X, (0,1):Y. Class

(1,0) are reads consistent with the first isoform in the annotation but not the second while class (0,1) are reads consistent with the second but not the first. N was the cutoff value derived from X + Y frequency distribution. Rule 2: CI-width > median of CI-width, where CI is the confidence intervals outputted by MISO for each estimate of Ψ. The rest of the potential SE events are then defined as unreliable SE group. We then extracted nucleosome distribution from iNPS results based on the coordinates of the rSE and urSE groups. The difference of nucleosome organization between rSE and urSE groups was then measured by Fréchet distance [72] and nucleosome positioning population (Additional file 1: Suppl. Notes—the pseudocode for calculating Fréchet distance). The following equation calculates splicing potentiality of SE (SPSE):

$$\mathrm{SPSE}_{S_o} = \mathrm{norm}(\mathrm{frdist}) \times \mathrm{norm}\left(\mathrm{abs}\left(\mathrm{diff}_{\mathrm{nucpos}}\right) \times \mathrm{coef}_{\mathrm{event-counts}} \right. \tag{Eq.9}$$

where norm means the results scaling to range [0, 1], frdist is the acronym of Fréchet distance, abs is the acronym of absolute function, $\mathrm{diff}_{\mathrm{nucpos}}$ represents the difference of median values of NucS rSE and urSE group, and $\mathrm{coef}_{\mathrm{event-counts}}$ is the normalized event count coefficient.

More specifically, the Fréchet Distance (norm(frdist)) is used to measure the difference of the averaged NucS array signal (containing both nucleosome spacing and phasing measurements) between rSE and urSE group (Additional file 1: Fig. S17). The $\mathrm{norm}(\mathrm{abs}(\mathrm{diff}_{\mathrm{nucpos}})$ measured the different of the nucleosome positioning between rSE and urSE group (Additional file 1: Fig. S18). The larger the Fréchet distance and nucleosome positioning implied a higher SE potentiality of the NucS. The last $\mathrm{coef}_{\mathrm{event-counts}} = \mathrm{norm}\left(\frac{\text{number of NucS}_{\mathrm{rSE}}}{\text{number of NucS}}\right)$ is used to measure the 'abundance' of the NucS in the rSE group.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02465-1.

Additional file 1. NucHMM: a method for quantitative modeling of nucleosome organization identifying functional nucleosome states distinctly associated with splicing potentiality: Suppl. Notes, Figures and Tables.

Additional file 2. The lists of cell type-specific NucSs-genes for MCF7.

Additional file 3. The lists of cell type-specific NucSs-genes for H1.

Additional file 4. The lists of cell type-specific NucSs-genes for IMR90.

Additional file 5. Review history.

Acknowledgements

We are grateful to the committee members (Drs. Jason Ernst, Yufei Huang, Wouter Koek, and Yogesh Gupta) for providing suggestive comments.

Review history

The review history is available as additional file 5.

Peer review information

Anahita Bishop was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

V.X.J. conceived the project. K.F developed the NucHMM program and performed the data analyses. V.X.J. and K.F. wrote the manuscript, with all authors including T.L. and Y.H. contributing to writing and providing the feedback. The author(s) read and approved the final manuscript.

Funding

This project was partially supported by grants from NIH R01GM114142 and U54CA217297.

Availability of data and materials

The datasets analyzed during the current study are available in the GEO repository: GSM1238700 [73], GSM1194220 [74], and GSE21823 [75] and in the ENCODE repository [61] : listed in Additional file 1: Table S1.

Our method is implemented as a C++/python package and is freely available at Under GNU General Public License (GPL-v3.0) (https://github.com/KunFang93/NucHMM) [76] with the version used to generate data in this manuscript deposited in zenodo (https://zenodo.org/record/4581548) [77].

Declarations

Ethics approval and consent to participate

Ethics approval is not applicable for this study

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Molecular Medicine, UTHSA-UTSA Joint Biomedical Engineering Program, San Antonio, TX 78229, USA. [2]Department of Molecular Medicine, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA. [3]Department of Medicine, UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA 15232, USA.

Received: 1 November 2020 Accepted: 12 August 2021
Published online: 26 August 2021

References

1. Kornberg RD, Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell. 1999;98(3):285–94. https://doi.org/10.1016/S0092-8674(00)81958-3.
2. Malik HS, Henikoff S. Phylogenomics of the nucleosome. Nat Struct Biol. 2003;10(11):882–91. https://doi.org/10.1038/nsb996.
3. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature. 1997;389(6648):251–60. https://doi.org/10.1038/38444.
4. Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol. 2013;20(3):267–73. https://doi.org/10.1038/nsmb.2506.
5. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell. 2008;132(5):887–98. https://doi.org/10.1016/j.cell.2008.02.022.
6. Sadeh R, Allis CD. Genome-wide “re”-modeling of nucleosome positions. Cell. 2011;147(2):263–6. https://doi.org/10.1016/j.cell.2011.09.042.
7. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. Determinants of nucleosome organization in primary human cells. Nature. 2011;474(7352):516–20. https://doi.org/10.1038/nature10002.
8. Yen K, Vinayachandran V, Batta K, Koerber RT, Pugh BF. Genome-wide nucleosome specificity and directionality of chromatin remodelers. Cell. 2012;149(7):1461–73. https://doi.org/10.1016/j.cell.2012.04.036.
9. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 2009;458(7236):362–6. https://doi.org/10.1038/nature07667.
10. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, et al. A genomic code for nucleosome positioning. Nature. 2006;442(7104):772–8. https://doi.org/10.1038/nature04979.
11. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, et al. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol. 2008;4(11):e1000216. https://doi.org/10.1371/journal.pcbi.1000216.
12. Kornberg R. The location of nucleosomes in chromatin: specific or statistical. Nature. 1981;292(5824):579–80. https://doi.org/10.1038/292579a0.
13. Radman-Livaja M, Rando OJ. Nucleosome positioning: how is it established, and why does it matter? Developmental biology. 2010;339(2):258–66. https://doi.org/10.1016/j.ydbio.2009.06.012.
14. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet. 2009;10(3):161–72. https://doi.org/10.1038/nrg2522.
15. Zhou K, Gaullier G, Luger K. Nucleosome structure and dynamics are coming of age. Nat Struct Mol Biol. 2019;26(1):3–13. https://doi.org/10.1038/s41594-018-0166-x.
16. Henikoff S, Henikoff JG, Sakai A, Loeb GB, Ahmad K. Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome Res. 2009;19(3):460–9. https://doi.org/10.1101/gr.087619.108.
17. Hodges C, Bintu L, Lubkowska L, Kashlev M, Bustamante C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. Science. 2009;325(5940):626–8. https://doi.org/10.1126/science.1172926.
18. Choi JK, Kim YJ. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nat Genet. 2009;41(4):498–503. https://doi.org/10.1038/ng.319.
19. Lieleg C, Ketterer P, Nuebler J, Ludwigsen J, Gerland U, Dietz H, et al. Nucleosome spacing generated by ISWI and CHD1 remodelers is constant regardless of nucleosome density. Mol Cell Biol. 2015;35(9):1588–605. https://doi.org/10.1128/MCB.01070-14.
20. He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, et al. Nucleosome dynamics define transcriptional enhancers. Nat Genet. 2010;42(4):343–7. https://doi.org/10.1038/ng.545.
21. Ioshikhes IP, Albert I, Zanton SJ, Pugh BF. Nucleosome positions predicted through comparative genomics. Nature genetics. 2006;38(10):1210–5. https://doi.org/10.1038/ng1878.
22. Baldi P, Brunak S, Chauvin Y, Krogh A. Naturally occurring nucleosome positioning signals in human exons and introns. J Mol Biol. 1996;263(4):503–10. https://doi.org/10.1006/jmbi.1996.0592.

23. Batsche E, Yaniv M, Muchardt C. The human SWI/SNF subunit Brm is a regulator of alternative splicing. Nat Struct Mol Biol. 2006;13(1):22–9. https://doi.org/10.1038/nsmb1030.
24. Sims RJ 3rd, Millhouse S, Chen CF, Lewis BA, Erdjument-Bromage H, Tempst P, et al. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell. 2007;28(4):665–76. https://doi.org/10.1016/j.molcel.2007.11.010.
25. Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res. 2009;19(10):1732–41. https://doi.org/10.1101/gr.092353.109.
26. Dhami P, Saffrey P, Bruce AW, Dillon SC, Chiang K, Bonhoure N, et al. Complex exon-intron marking by histone modifications is not determined solely by nucleosome distribution. PLoS One. 2010;5(8):e12339. https://doi.org/10.1371/journal.pone.0012339.
27. Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nat Struct Mol Biol. 2009;16(9):990–5. https://doi.org/10.1038/nsmb.1659.
28. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. Science. 2010;327(5968):996–1000. https://doi.org/10.1126/science.1184208.
29. Li T, Liu Q, Garza N, Kornblau S, Jin VX. Integrative analysis reveals functional and regulatory roles of H3K79me2 in mediating alternative splicing. Genome Med. 2018;10(1):30. https://doi.org/10.1186/s13073-018-0538-1.
30. Libbrecht MW, Chan RC, Hoffman MM: Segmentation and genome annotation algorithms. arXiv preprint arXiv: 210100688 2021.
31. Day N, Hemmaplardh A, Thurman RE, Stamatoyannopoulos JA, Noble WS. Unsupervised segmentation of continuous genomic data. Bioinformatics. 2007;23(11):1424–6. https://doi.org/10.1093/bioinformatics/btm096.
32. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215–6. https://doi.org/10.1038/nmeth.1906.
33. Libbrecht MW, Ay F, Hoffman MM, Gilbert DM, Bilmes JA, Noble WS. Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. Genome Res. 2015;25(4):544–57. https://doi.org/10.1101/gr.184341.114.
34. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell. 2010;143(2):212–24. https://doi.org/10.1016/j.cell.2010.09.009.
35. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010;28(8):817–25. https://doi.org/10.1038/nbt.1662.
36. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013;41(2):827–41. https://doi.org/10.1093/nar/gks1284.
37. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
38. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12(12):2478–92. https://doi.org/10.1038/nprot.2017.124.
39. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9(5):473–6. https://doi.org/10.1038/nmeth.1937.
40. Chan RCW, Libbrecht MW, Roberts EG, Bilmes JA, Noble WS, Hoffman MM. Segway 2.0: Gaussian mixture models and minibatch training. Bioinformatics. 2018;34(4):669–71. https://doi.org/10.1093/bioinformatics/btx603.
41. Libbrecht MW, Rodriguez OL, Weng Z, Bilmes JA, Hoffman MM, Noble WS. A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. Genome Biol. 2019;20(1):180. https://doi.org/10.1186/s13059-019-1784-2.
42. Biesinger J, Wang Y, Xie X: Discovering and mapping chromatin states using a tree hidden Markov model. BMC Bioinformatics. 2013;14(Suppl 5):S4.
43. Song J, Chen KC. Spectacle: fast chromatin state annotation using spectral learning. Genome Biol. 2015;16(1):33. https://doi.org/10.1186/s13059-015-0598-0.
44. Sohn KA, Ho JW, Djordjevic D, Jeong HH, Park PJ. Kim JH: hiHMM: Bayesian non-parametric joint inference of chromatin state maps. Bioinformatics. 2015;31(13):2066–74. https://doi.org/10.1093/bioinformatics/btv117.
45. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. Genome Biol. 2015;16(1):56. https://doi.org/10.1186/s13059-015-0621-5.
46. Mammana A, Chung HR. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. Genome Biol. 2015;16(1):151. https://doi.org/10.1186/s13059-015-0708-z.
47. Libbrecht M, Hoffman M, Bilmes J, Noble W: Entropic graph-based posterior regularization. In International Conference on Machine Learning. PMLR; 2015: 1992-2001.
48. Zhang Y, An L, Yue F, Hardison RC. Jointly characterizing epigenetic dynamics across multiple human cell types. Nucleic Acids Res. 2016;44(14):6721–31. https://doi.org/10.1093/nar/gkw278.
49. Zhang Y, Hardison RC. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. Nucleic Acids Res. 2017;45(17):9823–36. https://doi.org/10.1093/nar/gkx659.
50. Zacher B, Michel M, Schwalb B, Cramer P, Tresch A, Gagneur J. Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. PLoS One. 2017;12(1):e0169249. https://doi.org/10.1371/journal.pone.0169249.
51. Marco E, Meuleman W, Huang J, Glass K, Pinello L, Wang J, et al. Multi-scale chromatin state annotation using a hierarchical hidden Markov model. Nat Commun. 2017;8(1):15011. https://doi.org/10.1038/ncomms15011.
52. Girimurugan SB, Liu Y, Lung PY, Vera DL, Dennis JH, Bass HW, et al. iSeg: an efficient algorithm for segmentation of genomic and epigenomic data. BMC Bioinformatics. 2018;19(1):131. https://doi.org/10.1186/s12859-018-2140-3.
53. Coetzee SG, Ramjan Z, Dinh HQ, Berman BP, Hazelett DJ: Statehub-statepaintr: rapid and reproducible chromatin state evaluation for custom genome annotation. F1000Research. 2020;7:214.
54. Poulet A, Li B, Dubos T, Rivera-Mulia JC, Gilbert DM, Qin ZS. RT States: systematic annotation of the human genome using cell type-specific replication timing programs. Bioinformatics. 2019;35(13):2167–76. https://doi.org/10.1093/bioinformatics/bty957.

55. Arneson A, Ernst J. Systematic discovery of conservation states for single-nucleotide annotation of the human genome. Commun Biol. 2019;2(1):248. https://doi.org/10.1038/s42003-019-0488-1.
56. Benner P, Vingron M. ModHMM: a modular supra-Bayesian genome segmentation method. J Comput Biol. 2020;27(4): 442–57. https://doi.org/10.1089/cmb.2019.0280.
57. Wang Y, Zhang Y, Zhang R, van Schaik T, Zhang L, Sasaki T, et al. SPIN reveals genome-wide landscape of nuclear compartmentalization. Genome Biology. 2021;22:1–23.
58. Mendez M, Scott MS, Hoffman MM: Unsupervised analysis of multi-experiment transcriptomic patterns with SegRNA identifies unannotated transcripts. bioRxiv 2020.
59. Liu Q, Bonneville R, Li T, Jin VX. Transcription factor-associated combinatorial epigenetic pattern reveals higher transcriptional activity of TCF7L2-regulated intragenic enhancers. BMC Genomics. 2017;18(1):375. https://doi.org/10.1186/s12864-017-3764-9.
60. Hon G, Ren B, Wang W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biol. 2008;4(10):e1000201. https://doi.org/10.1371/journal.pcbi.1000201.
61. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74. https://doi.org/10.1038/nature11247.
62. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137. https://doi.org/10.1186/gb-2008-9-9-r137.
63. Stovner EB. Saetrom P: epic2 efficiently finds diffuse domains in ChIP-seq data. Bioinformatics. 2019;35(21):4392–3. https://doi.org/10.1093/bioinformatics/btz232.
64. Chen W, Liu Y, Zhu S, Green CD, Wei G, Han JD. Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. Nat Commun. 2014;5(1):4909. https://doi.org/10.1038/ncomms5909.
65. Buitrago D, Codo L, Illa R, de Jorge P, Battistini F, Flores O, et al. Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning. Nucleic Acids Res. 2019;47(18):9511–23. https://doi.org/10.1093/nar/gkz759.
66. Ye Z, Chen Z, Lan X, Hara S, Sunkel B, Huang TH, et al. Computational analysis reveals a correlation of exon-skipping events with splicing, transcription and epigenetic factors. Nucleic Acids Res. 2014;42(5):2856–69. https://doi.org/10.1093/nar/gkt1338.
67. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17(1): 10–2. https://doi.org/10.14806/ej.17.1.200.
68. Ramirez F, Dundar F, Diehl S, Gruning BA. Manke T: deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42(W1):W187–91. https://doi.org/10.1093/nar/gku365.
69. Rabiner L, Juang B. An introduction to hidden Markov models. ieee assp magazine 1986;3:4-16.
70. Welch P. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Transactions on audio and electroacoustics. 1967;15(2):70–3. https://doi.org/10.1109/TAU.1967.1161901.
71. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010;7(12):1009–15. https://doi.org/10.1038/nmeth.1528.
72. Eiter T, Mannila H: Computing discrete Fréchet distance. Citeseer; 1994.
73. Shimbo T, Du Y, Grimm SA, Dhasarathy A, Mav D, Shah RR, Shi H, Wade PA: MBD3 localizes at promoters, gene bodies and enhancers of active genes. PLoS Genet. 2013;9:e1004028.
74. Yazdi PG, Pedersen BA, Taylor JF, Khattab OS, Chen YH, Chen Y, Jacobsen SE, Wang PH: Nucleosome Organization in Human Embryonic Stem Cells. PLoS One. 2015;10:e0136314.
75. Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA: Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res. 2012;22:2497-2506.
76. Kun F, Tianbao L, Yufei H, Victor XJ. NucHMM: a method for quantitative modeling of nucleosome organization identifying functional nucleosome states distinctly associated with splicing potentiality. In Github; 2021.
77. Kun F, Tianbao L, Yufei H, Victor XJ. NucHMM: a method for quantitative modeling of nucleosome organization identifying functional nucleosome states distinctly associated with splicing potentiality.: Zenodo; 2021.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.